MEGA-ANALYSIS OF GENE EXPRESSION PATTERNS ACROSS TISSUES IN HUMAN AND MOUSE

by

Min Feng

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

The Faculty of Graduate and Postdoctoral Studies

(Genome Science and Technology)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

August 2018

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, a thesis/dissertation entitled:

Mega-analysis of gene expression patterns across tissues in human and mouse

Submitted by Min Feng in partial fulfillment of the requirements for the degree of Master of Science in Genome Science and Technology

Examining Committee:

Paul Pavlidis, Department of Psychiatry and Centre for Brain Health
Supervisor

Joerg Bohlmann, Department of Forest Science and Botany
Supervisory Committee Member

Leonard Foster, Department of Biochemistry and Molecular Biology
Supervisory Committee Member

Carolyn Brown, Department of Medical Genetics
Additional Examiner

Abstract

Expression patterns across tissues are a primary indicator of gene function. High-throughput technology created many cross-tissue data sets on a transcriptomic level (tissue panel data sets). However, the existence of multiple tissue panel data sets creates a challenge for the scientific community to decide if these data sets are equally valid or decide which data set to choose. To date, the multiple tissue panel data sets have not been well compared, nor fully evaluated.

In my Master's thesis, I collected a large number of public-available tissue panel data sets, harmonized them, integrated the data sets into a tissue expression atlas including human data and mouse data, compared and contrasted the data sets across the atlas, evaluated each data set preliminarily with a gene-specific disagreement index that I developed. I found in general, these data sets had a good agreement. However, in certain data sets the amount of disagreement was high, which indicated the qualities of these data sets were suspect.

Applying the disagreement index, I was able to offer a summarized expression pattern in the tissue expression atlas with either consensus or disagreements outlined. I also developed a web-based prototype to access to this atlas.

Furthermore, I explored the range of changes in gene expression patterns that may be caused by experimental conditions, such as diseases or drug treatments. I found most of the changes could not be as dramatic as a change from unexpressed to highly expressed, even though these changes were reported as statistically significant in literatures. Only a couple of conditions such as cancer or inflammation could cause an unexpressed-to-highly-expressed change, because tissue composition in those conditions were changed substantially.

Lay Summary

A gene can be selectively expressed in different tissues. Knowing where a gene is expressed helps us learn about gene function.

This thesis gathered a number of published data sets. I found that existing cross-tissue transcriptomic data sets (tissue panel data sets) did not always agree with each other. The consistency level among published results has not been well measured.

I quantified the disagreements between any two tissue panel data sets to evaluate the existing tissue panel data sets. Additionally, I integrated existing data sets into a *tissue expression atlas* instead of trusting any single data set. The atlas offers summarized expression patterns from multiple data sets for human and mouse genes. The atlas is presented as a prototype web-based application.

Preface

The idea of building a tissue expression atlas to offer summarized tissue specific gene expression patterns from multiple data sets was initiated and designed by Dr. Paul Pavlidis in September 2016. With his supervision, I implemented the project, performed the analysis and wrote this thesis. Dr. Pavlidis also helped review the thesis. I wish to acknowledge and thank Dr. Pavlidis for his mentorship and his patience.

During my implementation, the curation team from Dr. Pavlidis' lab helped me with general data curation. Three of the group members from Dr. Pavlidis' lab helped me with data collection: Marjan Farahbod helped me collect GTEx RNA-sequencing (RNA-seq) data; Nathaniel Lim helped me download data sets from an in-house data base. He also offered me differential expression data for the exploration on conditional expression. Manuel Belmadani helped me collect three human RNA-seq data sets. In addition, Marjan Farahbod also offered useful suggestions about the experimental design and writing. I received assistance in English editing from Patrick Bruskiewich and Nathan Kingston.

Table of Contents

List of Tables

List of Figures

List of Abbreviations and Glossary

cDNA: complementary DNA

Cpt.Panel: computational-acquired tissue panel data set

DDX3Y: DEAD-Box Helicase 3

dNTP: deoxynucleotide

Exp.Panel: experiment-acquired tissue panel data set

GTEx: Genotype-Tissue Expression

HPA: Human Protein Atlas

KDM5D:Lysine-specific demethylase 5D

mRNA: messenger RNA

NA: missing value

PCR: polymerase chain reaction

RPS4Y : Ribosomal Protein S4

RNA-seq: RNA sequencing

rRNA: ribosome RNA

UBC: Ubiquitin C

XIST: X-Inactive Specific Transcript

Acknowledgements

My heart is filled with gratitude. All the help I have received I have wired and netted into a mist vest that I wear each morning when I greet the first beam. All the days and nights in Vancouver have been colored into an agate by the wind in fall that I will treasure in my future.

My first thanks goes to my supervisor, Dr. Paul Pavlidis. I cannot thank him enough for his guidance, prioritization of students and his surprising patience.

I sincerely appreciate my thesis committee, Dr. Joerg Bohlmann and Dr. Leonard Foster, for their support throughout my journey. Special thanks to Dr. Carolyn Brown, my examination chair. My learning would not have been fulfilled without your help.

Special thanks to PhD candidates Marjan Farahbod and Nathaniel Lim, who have directly helped with my thesis project. Thanks to both of them for their valuable advice and hands-on demonstrations in each stage of my project. I would also like to thank all of the Pavlidis lab's members for their unconditional support and company.

Last but not the least, I would like to thank all the faculty and staff members in Michael Smith Laboratories, my rotation labs, and the funding agencies. All your help made my master's experience excellent.

Dedication

For smiles

Chapter 1: Introduction

1.1 Background

In mammalian functional genomics, one of the most fundamental and common questions about a gene is "where and when is it expressed?" One of the most basic and important versions of this question is the distribution of transcripts across tissues. Tissue specificity offers a major clue to gene function. For example, if a gene is exclusively expressed in liver, it is fair to hypothesize that this gene has liver-specific functions. In contrast, if a gene is expressed widely, it is more reasonable to think that the gene has a general or basic cellular function (i.e. the traditional definition of housekeeping genes). Other functional factors such as developmental stage or cell type are also important, but cross-tissue transcriptome is relatively easily measured, and forms a foundation upon which questions of other factors can be built (J. Yang et al. 2015; Lin et al. 2014). Therefore, the cross-tissue gene expression profile (cross-tissue transcriptomics) is a key piece of information about a gene's expression pattern and having high-quality data is of major importance. The subject of this thesis is on the development of improved resources for assessing gene expression profiles across tissues with the goal of generating a high-quality and easy-to-access "atlas".

The importance of cross-tissue transcriptomics is reflected in the existence of large data sets that sample multiple tissues for transcriptome profiling (tissue panel data sets), and the prominence of tissue distribution in web sites and databases that offer information on genes (Figure 1).The existence of multiple tissue panel data sets creates a challenge in deciding which is the best, and surprisingly there has been little attempt to evaluate the level of agreement between them. Here, I emphasize how important it is to have high-quality tissue panel data sets.

These tissue panel data sets have had a big impact on the scientific community, which makes the data quality very important. For example, the BioGPS data set is used to present transcriptomics information on human genes in Wikipedia (Figure 1). The BioGPS data set includes 79 tissues for human and 61 tissues for mouse (Su et al. 2002, 2004). For each tissue, there are two replicates. The human gene pages in Wikipedia have been "viewed over 50 million times per year and edited over 15,000 times per year" in total (https://en.wikipedia.org/wiki/Portal:Gene_Wiki). The second example is the GTEx RNA-seq data set (Figure 1). The data set includes 1641 postmortem samples from 55 human tissues which were collected from 175 individuals. The GTEx RNA-seq data set has been cited more than a thousand times since its launch in 2013 (Lonsdale et al. 2013). The third example is the FANTOM5 data set (Yu et al. 2015). The FANTOM5 Project sampled 68 human tissues. One important reuse of the FANTOM5 data set and the GTEx RNA-seq data set is being part of tissue "atlases" (Figure 1), such as the European Bioinformatics Institute (EMBL-EBI) tissue atlas or the Human Protein Atlas (HPA) (Uhlén et al. 2015a). Thus, where there is inaccurate expression data, the reuse of the well-known tissue panel data sets can prove problematic.

The existence of multiple cross-tissue transcriptomic data sets creates a challenge in decision making. In an ideal situation, all the cross-tissue transcriptomic data sets agree with each other. However, the expression pattern of genes across different tissues might differ across data sets (Figure 2). This creates a challenge in deciding which data set has higher quality and validity and determining what caused the inconsistencies. Surprisingly, there has been few attempts to evaluate the level of agreement between cross-tissue transcriptomic data sets.

In this introduction, I present the existence of disagreements between the data sets, describe the potential causes of the challenges in comparing and evaluating those tissue panel data sets, assess

previous atlases, and introduce tools that help me to undertake my task of comparing data sets, evaluating data sets, and building a new atlas.

In this thesis, a *data set* refers to a matrix (table) in which studies represent their transcriptomic-level gene expression profiles: genes are *rows* and samples are *columns*. A *tissue panel data set* refers to a *data set* that samples multiple (generally >10) tissues. There may be replicates for each tissue (samples taken from more than one person).

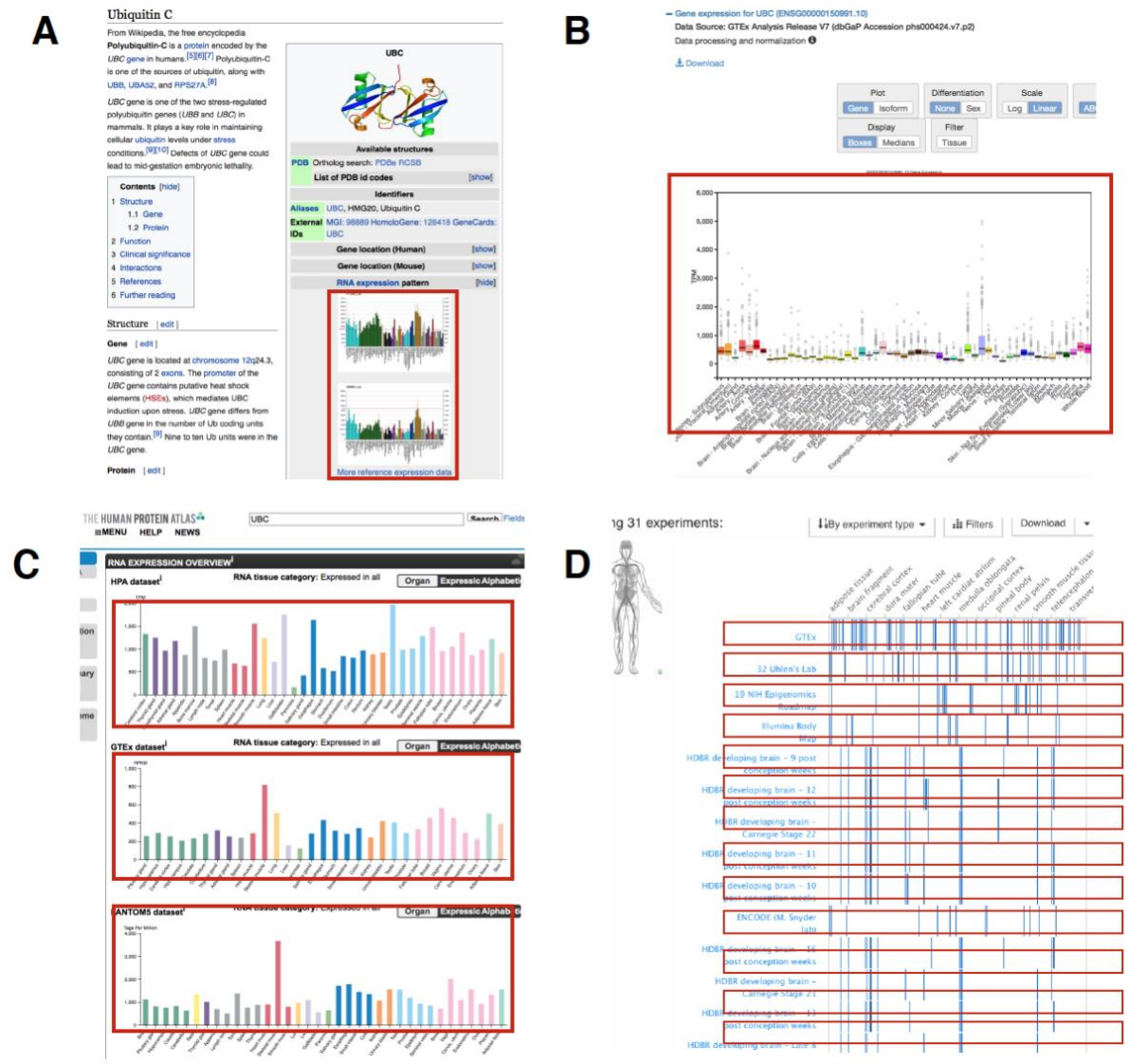

Figure 1 Examples of tissue panel data sets and tissue atlases.

Red boxes highlight the data set parts in the websites. (A) A screen shot of a Wikipedia web page showing the expression data from the BioGPS data set. *URL:*https://en.wikipedia.org/wiki/Ubiquitin_C (B) A screen shot from the GTEx website showing the GTEx data set. *URL:* https://www.gtexportal.org/home/gene/UBC (C) A screen shot form the Human Protein Atlas website showing multiple data sets collected by the atlas. *URL:*https://www.proteinatlas.org/ENSG00000150991-UBC/tissue (D) A screen shot from the EBI website showing a collection of multiple data sets. *URL:*

https://www.ebi.ac.uk/gxa/search?geneQuery=%5B%7B"value"%3A"HIST1H4C","category"%3A"symbol"%7D%5D&organism=&conditionQuery=%5B%5D&bs=%7B"homo sapiens"%3A%5B"ORGANISM_PART"%5D%7D&ds=%7B"kingdom"%3A%5B"animals"%5D%7D-baseline

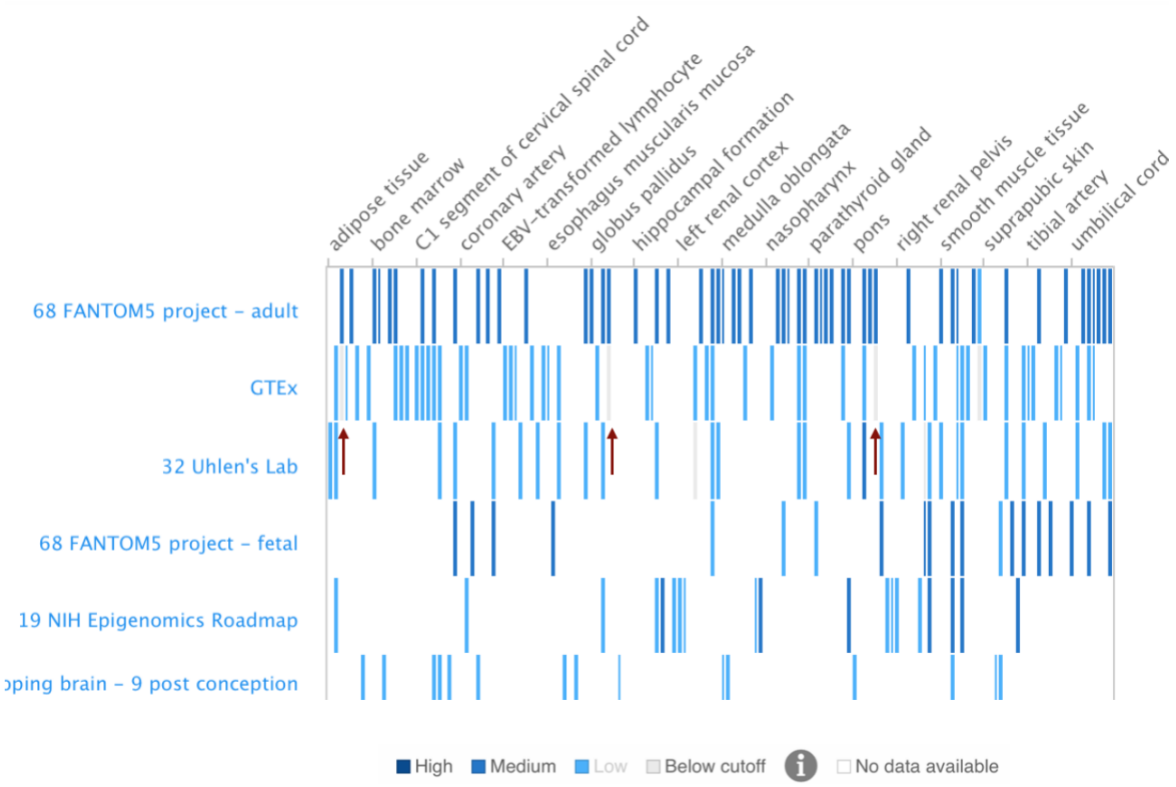

Figure 2 Demonstration of inconsistencies between some data sets

A screen shot from the EBI tissue atlas showing the cross-tissue expression patterns for human gene *HIST1H4* from different data sets. *Dark blue* means *high expression level*. *Blue* means *medium expression level*. *Light blue* means *low expression level*. *Grey* means *below cut off*. *White* means *missing value*. Red arrows added by me highlight disagreements: for a certain gene, at a certain tissue, two different data sets give opposite expression level (high vs. not expressed) *URL:* *https://www.ebi.ac.uk/gxa/search?geneQuery=%5B%7B"value"%3A"HIST1H4C","category"%3A"symbol"%7D%5D&orga*

1.2 Challenges in comparing and evaluating tissue panel data sets

Comparing and evaluating tissue panel data sets proves challenging because of the lack of standards for evaluation, and because differences among data sets are a result of both random and systematic variations. First, there is no gold standard or "true expression pattern" to compare each data set with. Second, studies differ in terms of technical factors such as platforms, quality control processes, and how a tissue was described and labeled in meta data. Third, the studies used samples from different individuals, which contributes to part of the random variation. In this section, the potential differences among individual data sets are described.

1.2.1 Different data sets had different scales and distributions

The different scales and distributions among data sets are caused by technical effects, such as platform effects and study effects (Head et al. 2014). For example, data sets generated by RNA-seq consist of integers as counts of mapped reads, which are then normalized as counts per million (CPM) or fragments per kilobase million (FPKM). In contrast, the data generated by microarray ranges continuously as representation of the relative intensity of the fluorescence signals. A log transformation was commonly performed to adjust the range difference between RNA-seq data and microarray data. Still, to make a comparison of them could result in artifacts, which may mask the meaningful biological impacts. For example, if we look at the expression value of gene G in a tissue T, the absolute values reported by an RNA-seq data set can be very different from the absolute values reported by a microarray data set. However, the expression levels across tissues could have the same pattern in two data sets, even though the actual values might be different.

This thesis project includes two types of high-throughput data: microarray data and RNA-seq data. Knowing the technical mechanisms behind these different data types helps me understand and explained the disagreements among studies.

In general, microarray technology and RNA-seq technology have similar procedures, though each step of the procedure is implemented differently. A typical procedure includes RNA isolation, library preparation, quantification of cDNA abundance with sequencer/microarray scanner, data quality control and normalization, and eventually generation of expression data sets.

Taking the most used microarray platform in this thesis (the Affymetrix chip) as an example, the library of an Affymetrix chip is prepared by Poly(A) selection. This step means to avoid the impact from rRNA on sequencing, which consists of more than 90% of the total RNA. Then purified mRNA is reverse transcribed to cDNA and labeled with a fluorescent tag as a sample target. The library is then put on the array to hybridize with complimentary capture probes which are predesigned and immobilized on the array. For the output, an Affymetrix chip typically returns a data matrix in which some target genes are represented by multiple probe sets. If there is more than one probe set per gene, one probe set will be picked based on the purpose of each study, or a combination of all the probe sets will be chosen as a representation of the gene.

Take the most used RNA-seq platform in this thesis (the Illumina sequencing) as an example, the library is prepared by either poly(A) selection or rRNA depletion to separate rRNA from mRNA. Then size selection is used to pick out fragments around 200bp for conversion of double-strand DNA, followed by adding sequencing adaptors and PCR amplification. Then each of the molecules is sequenced by synthesis, base by base. For the output, a gene level expression matrix is returned where probes are mapped to gene levels directly during assembly in data processing.

Another example of the RNA-seq protocol used in this thesis is the CAGE protocol (Kawaji et al. 2014a; Yu et al. 2015). The CAGE protocol was developed to identify and quantify the 5' ends of capped mRNAs based on cap-trapping (Kawaji et al. 2014b). Then the library is sequenced by a second generation sequencer. By avoiding PCR, ligation and poly(A) selection, Kawaji's study found that the CAGE protocol is less biased compared to the standard RNA-seq protocol (Kawaji et al. 2014b).

1.2.2 Samples tissues labels in meta data were inconsistent among data sets

For my project, the ideal tissue description is that all the studies obtain identical cell composition for a certain tissue and use exactly the same term to describe that tissue. However, several reasons make tissue descriptions heterogeneous. I will discuss two of these causes.

The first cause is the different descriptions for the same tissue and the different resolutions at which a certain tissue is studied. For example, heart tissue is labeled with "heart" in the BioGPS data set, but in the GTEx RNA-seq data it is represented by two parts "Heart-Atrial Appendage" and "Heart-Left Ventricle".

The second cause for heterogeneity is the complexity in the cellular and sub-cellular composition of a tissue (King and Sinha 2001). This can be affected by two factors.

The first factor is a combined effect of an imprecise dissection and the complex nature of bulk tissue. For example, a single punch biopsy of a gland can include epithelium, fat, glands, ducts, blood vessels, nerves, etc. The cells of each of the tissue types listed above (epithelium, fat, glands, ducts, blood vessels, nerves, etc.) could be at various developmental stages and various levels of activation.

The second factor is the unclear distinction between healthy and diseased tissues (King and Sinha 2001). The unclear distinction is not just due to the inaccurate diagnoses but also complicated disease backgrounds. Taking into account disease backgrounds, seemingly unrelated diseases may influence the gene expression of a distant tissue site. For example, a healthy heart tissue biopsy from patients with a certain lung disease may be different from a healthy heart tissue biopsy from a patient with normal lungs.

In this project, I develop approaches to address the potential differences among individual data sets to make building a tissue expression atlas feasible.

1.3 Assessment of previous tissue atlases

To date there are at least six web sites presenting their atlases. However, no evaluation or quantified comparison was conducted in any of these existing atlases. In this section, I compare and assess the following atlases in detail (Table 1):

- "Human Protein Atlas" (HPA)
- "Expression Atlas" from the European Bioinformatics Institute (EMBL-EBI)
- "All RNA-seq and CHIP-Seq Signature Search Space" (ARCHS4)
- "RNA-seq Atlas" (Krupp et al. 2012)
- "TISSUES 2".0 (Palasca et al. 2018)
- The National Center for Biotechnology Information (NCBI) gene resource (Agarwala et al. 2018).

The NCBI and the HPA collected multiple data sets, but the data sets are presented separately, which requires users to scroll down the web pages in order to see all the data. Thus, the NCBI and

the HPA make it difficult, if not impossible, to compare among the other data sets or evaluate an individual data set.

The EBI tissue atlas collects and further integrates the data sets into a single user-friendly plot but does not summarize the multiple different expression patterns from different data sets into a consensus expression pattern (Figure 2). When users inquire the EBI atlas for the expression pattern of a specific gene, they will see the pattern from different data sets are different from each other (Figure 2). Those disagreements presented in Figure 2 very likely confuse the users with questions like: which data set should I trust?

The ARCHS4 atlas offers only summarized expression patterns for each gene, without presenting data from each individual data set. In this case, variation across data sets is obscured.

All the four atlases mentioned above collected tissue panel data sets that were generated with RNA-seq. No data sets generated with microarray were included. However, microarray data sets are still useful, as they are a good technical complement for RNA-seq data sets.

The Tissue 2.0 atlas has a slightly different aim from the above mentioned atlases. It summarizes multiple data sets to demonstrate distinctions between tissue-specific expression and ubiquitous expression. However, the summarization in the Tissue 2.0 atlas is apparently based on the assumption that all the data sets are correct and consistent, without any evaluation of individual data sets.

In my project, both microarray data sets and RNA-seq data sets are included. I also offer information on individual data sets after integration, and summarized gene expression patterns with either consensus or disagreement.

1.4 Introduction of the tools and data used in this thesis

To select the best data set for the scientific community, I integrated as many tissue panel data sets as possible into a tissue atlas in order to make these tissue panel data sets comparable. I also evaluated collected data sets by the consistent level among the data sets. In this section, I give basic background on the tools (resources and methods) that I used. More details are given in the next chapter.

This thesis project is only feasible with the support of the Gemma system (Zoubarev et al. 2012). Gemma is a versatile system offering not only access to a big collection curated published expression studies, but also downstream results from functional analyses including differential expression and co-expression analyses. With Gemma, all the data sets were processed with a unified in-house pipeline to reduce biases.

To harmonize the inconsistent usage of tissue labels in different studies, a widely-used standard-the UBERON Ontology-is chosen for my project. The UBERON Ontology is an integrated cross-species ontology covering anatomical structures in animals (Mungall et al. 2012). This ontological system helps organize tissues based on different relationships (i.e. “part_of” relationship or “is_a” relationship). Most tissues can be aggregated to a higher level (parent level) or be broken down to a lower level (children level). For example, a parent term “lung” has two children terms: “left lung” and “right lung”. Based on this information, “left lung” can go to a higher level by being replaced by the parent term “lung”. On the other hand, if there is enough information in the meta data indicating that all the samples labeled with “lung” are actually acquired from only the left side, then the term “lung” can be specified into a children term “left lung”. This harmonization process needs manual work sometimes due to the lack of information or the ambiguous description of tissues in meta data.

Notably, there are only a small number of tissue panel data sets, meanwhile there are many small data sets with a few tissues in each. The small data sets are valuable if we can find a way to combine them into a multi-tissue transcriptomic data set.

I integrated the numerous small transcriptomics data sets computationally into a big data set with multiple tissues. This new combined tissue panel data set is referred to as the "Cpt.Panel" data set. Other tissue panel data sets that I collected directly from literature are the data sets generated by experiments on the bench. As a contrast to my Cpt.Panel data set, I referred to the data sets generated by experiments as "Exp.Panel" data sets. Although, my computational combination is also challenged by the differences among individual data sets, I argue that the Cpt.Panel data sets are worthy to be added to Exp.Panel data sets.

I combined both Cpt.Panel data sets and Exp.Panel data sets into one *tissue expression atlas* for each species. I refer to my tissue expression atlas as the "Gemma Atlas".

In this thesis, I undertook collecting, harmonizing, comparing and evaluating the chosen tissue panel data sets. Furthermore, I integrated these tissue panel data sets into a new tissue expression atlas. For each gene, the atlas summarized disagreements and variations of those cross-tissue transcriptomes. Additionally, I created a prototype website for the public to access to the atlas. I argue that this atlas provides a more realistic view of cross-tissue transcriptomics by allowing for variation across data sets, while also offering the possibility of identifying a consensus.

Table 1 Assessment of previous tissue atlases

Data set integration makes different data sets comparable and shows across-data set variation. Summarized expression pattern is a direct accessible gene expression pattern summarized across all the data sets with consensus and variation. The two websites in green columns are two examples of single data set resources. The six web sites in blue are six websites presenting atlases (a collection of data sets). The Human Protein Atlas and the EBI atlas are also showed in Fig.1-C and Fig.1-D as examples. Denotation: HS: human. MM: mouse. Y for yes. N for No. NA for no applicable answer.

Name of Web sites	BioGPS Project	GTEx Project	NCBI	Human Protein Atlas	RNA-seq Atlas	EBI Tissue Atlas	ARCHS4	TISSUES 2.0
# Data sets	1	1	4	3	4	30	4	14
Species	HS, MM	HS	Multiple	HS	HS	Multiple	HS, MM	Multiple
Uniform processing	Y	Y	Y	Y	Y	Y	Y	Y
Data set integration	NA	NA	N	N	N	Y	N	Y
Summarized expression pattern	Y	Y	N	N	N	N	Y	Y
Evaluation of each data set	NA	NA	N	N	N	N	N	N

Chapter 2: Materials and Methods

Data analyses in this project were performed using the statistical application program R version 3.3.3 on a computer platform x86_64-redhat-linux-gnu (64-bit) running under CentOS Linux 7 (Core). Scripts used in this project were peer validated and are available from the GitHub (https://github.com/PavlidisLab/Mega-analysis-on-gene-expression-patterns-across-tissues-in-human-and-mouse.git).

2.1 Data collection and quality control

This thesis project aims to collect and analyze as many publicly-available tissue-related transcriptomics data sets for human and mouse as possible. Most of the data sets were obtained through an in-house-developed and validated platform (the Gemma database) which provides support for raw data reanalysis, quality control and annotation. The exception in data collection is the GTEx RNA-seq data (GTEx.seq) which was downloaded from the GTEx portal (www.broadinstitute.org/gtex). The last update of the input data was on June 20, 2017. The data sets I used for this project are summarized in Table 2 and Table 3.

Data sets and samples within data sets were both filtered from the data corpus to improve overall data quality. Filtering procedures were run based on two considerations:

1. Data quality. This could be broken down into three aspects:

- Two-color microarray data was removed.
- Samples marked as outliers by the Gemma curation pipeline were omitted.
- Samples whose tissue source cannot be determined with sufficient specificity were removed.

2. Tissue identity. I allowed the inclusion of a defined class of "abnormal" tissues (tissues from disease states and drug treatment conditions, etc.), but not tissues with tumors or inflammation. These were omitted because the cellular composition in these two conditions is changed. Data from immortalized cell lines was also removed.

2.2 Data classification

All the tissue related data sets were classed into two groups: tissue panel data sets that sampled many tissues in each and the remaining smaller data sets that sampled one or a few tissues in each. Specifically for this thesis, data sets were defined as *experiment-acquired tissue panel data sets* (Exp.Panel) if they had 10 or more tissue types. The remaining data sets (the majority of the data corpus) had only one or a few tissues included. For a broader presentation, and also for a broader tissue coverage, these smaller data sets were integrated into a single large data set (one for human, one for mouse). These integrated tissue panel data sets were named as *Computational-Acquired tissue panel data sets* (Cpt.Panel).

The integration for Cpt.Panel data sets was done in three steps:

First, within each small data set, all the samples from the same tissue were integrated into one expression value for each gene (a tissue vector) by taking the mean of all the samples with the same tissue label. For the data sets with only one tissue type, this step compressed the whole data set into one vector.

Second, all the tissue vectors from different data sets were combined by taking an union of all the genes. Taking the union caused many missing values (NA) due to differences in gene coverage among platforms. For example, if gene G was only measured in two data sets, taking the union would yield NAs in the other data sets.

Third, quantile normalization was performed with the Limma package (Ritchie et al. 2015) to reduce inter-data-set variation.

2.3 Data harmonization

This section describes processes conducted to deal with the heterogeneity within the data corpus. All data sets were preprocessed from raw data if available (e.g. from Affymetrix CEL files, or FASTQ files for RNA-seq) with a uniform in-house pipeline (the Gemma pipeline) to reduce artifacts in the preprocessing. In the Gemma pipeline, mm9 and hg38 were used as the reference genome for mouse and human, respectively, to assemble RNA-seq data. Then BOWTIE+RSEM was run for alignment to get BAM files (Teng et al. 2016). For microarray data, batch correction was performed if possible, leaving the confounding cases out (Li and Dewey 2011).

2.3.1 Harmonizing and aggregating tissue descriptions

The differences in tissue descriptions makes a computational comparison and integration of different data sets problematic. This challenge could be addressed with a unified tissue ontology the UBERON (Mungall et al. 2012). However, tissue descriptions that were not using the standard UBERON terms in the first place were required to be harmonized into equivalent UBERON ontology terms manually. Where necessary, multiple UBERON child terms were aggregated into a more inclusive parent term. This aggregation was continued until all the samples in one tissue reached a comparable resolution. Lastly, tissue labels at the final resolution were substituted again with public-familiar names. For example, the term “pair of lungs” was substituted with the word “lung”.

2.3.2 Harmonizing expression data at the gene level

A difference in expression data between technologies (microarray data sets and RNA-seq data sets) is that the microarray data sets are at a probe set level while RNA-seq data sets are quantified at the gene level. In order to compare and integrate data sets from different technologies, all microarray data sets were mapped to gene level using platform-specific gene annotations downloaded from Gemma; the method Gemma uses for this mapping is essentially as described in (Barnes et al. 2005). If a gene was represented by multiple probes, the highest expressed probe set (per data set) was picked to represent the gene. Probes that mapped to more than one gene were removed.

2.3.3 Harmonizing expression levels by a binning method

Different data sets have different distributions of expression levels. This difference needed to be harmonized in order to compare expression data of the same gene across data sets. Commonly used ranking methods like quantile normalization can get different distributions into the same shape. This ranking method results in a rank for each gene. However, when a huge number of genes is studied, there is no expectation that all the ranks will precisely agree. Instead, there might be an expectation that the expression level should agree on certain levels in a given tissue. In this project, the full expression range was described with three bins:

- Highly expressed
- Moderately expressed
- Unexpressed

The top 10% highest expression data points in the distribution of each data set were all counted as *highly expressed*. Any data point distributed lower than the top 10% but higher than background

noise was classed as *moderately expressed*. However, the threshold for background noises (*unexpressed*) could differ from data set to data set. Ideally, anything deemed unexpressed should be reported as zero value. However, in real microarray data, there is always a certain amount of background signal being captured.

To define the background noise for each data set, I used the expression values of gender specific genes in their opposite genders (Toker, Feng, and Pavlidis 2016). Take female-specific gene *X-inactive specific transcript (XIST)* as an example. All the expression values (data points) of *XIST* in a data set can be clustered into two groups based on the genetic gender. Each group has a center as the average expression value of that group. The group with the smaller center, if not zero value, is the male group, as gene *XIST* should not be detected as expressed in normal male samples (Cerase et al. 2015). The expression values of *XIST* detected in male group are just caused by background noises. Thus, the smaller center is set to be a predefined threshold for the background noise.

The same approach used for *XIST* was applied to the other three male specific genes: *Lysine-specific demethylase 5D* (*KDM5D*), *DEAD-Box Helicase 3* (*DDX3Y*) and *Ribosomal Protein S4* (*RPS4Y*). Together this resulted in a principled way to define thresholds for the background noise. I took the minimum value of all four predefined thresholds (acquired from *XIST*, *KDM5D*, *DDX3Y* and *RPS4Y)* as the threshold for the data set background noise. Any expression value lower than the data set background noise was classed as *unexpressed*.

2.3.4 Comparison between human data and mouse data

I compared between human data sets and mouse data sets for perfect agreements and data sets evaluation. To define homologues genes, I used an existing R function

(https://github.com/oganm/homologene). To use this R function, gene symbols or NCBI ids were needed as input, so were the species where the input genes came from and the species that was sought for. All the mapping information in this function was got from NCBI.

2.4 Building the Gemma Atlas

The across-data set comparison laid the foundation for the valuation of each data set. For comparison, I combined all the processed data sets into an atlas. In the atlas, data set-specific tissues and genes were not included as they cannot be compared with other data sets.

2.4.1 Combining data sets into an atlas

For each species, all the processed Exp.Panel data sets and the Cpt.Panel data set were combined by taking the union of all the genes and tissues. This process induced a number of missing values (NAs). However, the purpose of building this atlas, rather than use only one data set, is to offer summarized gene expression patterns across data sets, and to evaluate each data set based on across-data sets comparison. Thus, only genes or tissues that were reported by at least two data sets were included. The data set-specific tissues and genes were excluded from atlas. The combined atlas was named the *Gemma Atlas*.

2.4.2 Integrating data from sample level to tissue level

For ease of access to users, the Gemma Atlas offers one integrated expression value for each gene at each tissue. To achieve this tissue-level integration, the median expression value of all the samples in a certain tissue was used to represent the expression value in the specific tissue.

2.5 Comparison across data sets

In order to pick the best data set, I need to set the criteria for "best" first, as no one has attempted this before. The practical purpose of the project is to provide the scientific community with reliable expression patterns across as many tissues as possible, thus "best" is evaluated in two aspects:

- The reliability of data
- The tissue coverage (the number of tissues covered by an individual data set out of the number of tissues in the Gemma Atlas).

2.5.1 Definition of disagreement and perfect agreement

One goal of my study was to characterize differences among data sets in terms of the expression levels of genes in each tissue, in particular to identify disagreeing data points. Because of the differences of the data, there is no expectation that expression levels will precisely agree. Instead, there might be an expectation that data sets should agree on whether a gene is expressed in a given tissue. However, exploratory analysis indicated it is difficult to come up with a fully satisfying definition of "expressed", especially at low expression levels. Therefore, for this study I developed a compromise approach in which I prioritized the identification of what might be thought of as "*level 2 disagreement" (L2D)*: between a strictly "*not expressed*" and "*highly expressed*" (using the definitions given in the last section). While this is still an arbitrary and stringent definition of *"disagreement"*, it proved to yield a substantial number of potential cases to investigate. Similarly to L2D, the "*level 1 disagreement"(L1D)* is the disagreement either between *"highly expressed"* and *"moderately expressed"*, or between *"moderately expressed"* and "*not expressed*". Attempts to investigate the "*level 1 disagreement"* would be a topic for future study.

The per-gene, per-tissue level of extreme difference was also referred to as a *disagreement at point level*. For some genes, their expression patterns disagreed at multiple tissue sites between one pair of data sets. These genes were only counted once when unique disagreed genes were counted. These unique disagreed genes were also referred as *disagreed genes*. In this thesis, a by default denotes a disagreement at point level; the summarized gene level disagreements were specified as *disagreement at gene level*. In order to know how common disagreements are, all the data set pairs were evaluated with the criteria "*level 2 disagreement"* at both point level and gene level.

A *perfect agreement* is when the per-gene per-tissue level expression pattern are at the same expression level (same bin) across all the data sets in the atlas.

There are many cases with neither perfect agreement nor *level 2 disagreement*. They are more complicated and I leave them for future study.

2.5.2 Comparison of tissue coverages

The larger the tissue coverage of a data set (number of distinct tissues), the better. Tissues in the Gemma Atlas were selected to be able to compare across data sets. Thus, atlas tissues were the standards in the evaluation of other data sets' tissue coverage. Based on the earlier definition of the tissue coverage, an equation was used to calculate the tissue coverage for each individual data set.

Tissue coverage = # tissues that a data set contributed to the Gemma Atlas / # of tissues in Gemma Atlas.

2.6 Investigation of level 2 disagreement

Two thresholds used for defining *level 2 disagreements* were so stringent that I hypothesized the L2D were not caused by random noise, but technical or biological factors. However, disentangling the mixed effects is difficult, especially when multiple factors change simultaneously. For example, between the BioGPS data set and the GTEx.seq data set, there were three main differences (potential factors) existing simultaneously:

- Sample difference (a potential source of biological variation)
- Platform difference denoted as microarray vs. RNA-seq (a technical artifact)
- Lab difference (a technical artifact)

Another example is a pair of data sets both from the GTEx project. One of these two data set is the GTEx RNA-seq data set that we mentioned before; another data set from the GTEx project is the GTEx microarray data set. These two data sets shared samples, but profiled the transcriptome by different technologies. Thus this pair of GTEx data sets (the GTEx.array data set and the GTEx.seq data set) is a good technical control for my project.

2.7 Summarization across data sets for gene expression patterns

One goal of this project is to offer tissue specific expression patterns summarized from all the data sets in the atlas for each gene. In an ideal situation where all the data sets in the atlas are consistent with each other, the summarization is just a robust consensus. However, in most cases, it is typical to have inconsistencies across data sets and even have *level 2 disagreements* (identified in 2.5.1). I developed a strategy for these cases to have a summarized expression pattern for each gene in the atlas:

For each gene, the expression of a certain tissue will be summarized as a L2D as long as there is a point L2D between any two data sets. If there was no point L2D at that tissue, the three bins of expression level were assigned to arbitrary values 0, 1 and 2 (from low to high). Then the median expression of all the data sets was taken as the summarized value. In cases where the median was not an integer, they were rounded with the standard R function "round" to keep only three levels for expression value.

2.8 Exploration of conditional expression

The gene expression pattern in normal tissues could be changed with experimental conditions like disease and drug treatment. There are a lot of differential expression analyses reporting genes with statistically significant changes in expression level in certain conditions, but the relative significances cannot give information on how much the values are changed absolutely. Is the change big enough to cause a gene to switch status from "*expressed*" to "*unexpressed*", or vice versa (See previous section in method for the definition of "unexpressed")? However, for eukaryotes, the default state of gene expression is "unexpressed" rather than "expressed", and we care more about whether a gene is expressed ever in any given tissue. The genes that are expressed in "normal" tissues are counted as expressed in the tissue. However, the question is about the genes that are not expressed in the normal tissue. If the "*unexpressed*/*off*" genes could be "*turned on*" by certain conditions, these genes were defined as "*off-on gene*". Thus, in this study, I developed a definition for "*off-on gene*" to study the effect of conditional expression on gene expression patterns, as an extension of the Gemma Atlas.

For each normal tissue (no pathology), a gene was reported as "*off*" if it was "unexpressed" in more than one Exp.Panel data set. If an "off" gene could be differentially expressed in a certain (non-normal) condition, this gene was identified as *conditionally "on"*. Genes being able to switch

status like this were referred as *"off-on" genes*. A clear definition for an off-on gene was a gene that was not expressed (off) in a normal tissue, but was expressed in the same tissue under certain conditions, such as diseases, inflammation or drug treatment. Differential expression data was from the Gemma, in which each gene was given a raw p-value for each pairwise condition. I did a multiple hypothesis correction using the *Benjamini-Hochberg* Procedure for each comparison. Finally, a combined false discovery rate (FDR) < 0.05 and $\log_2$ fold change > 1 (fold change > 2) was used to pick differently expressed genes.

Chapter 3: Results

The multiple existence of cross-tissue transcriptomic data sets (tissue panel data sets) has created a challenge for decision making. In this thesis, I collected a large number of publicly-available tissue panel data sets, integrated them into a *tissue expression atlas* (the *Gemma Atlas*) to make across-data-set comparison feasible. I also developed an approach to each evaluate tissue panel data set. Furthermore, I explored the extension of normal expression patterns in conditional expression.

3.1 Summary of collected data sets

In this project, I collected 1686 tissue-related data sets (79016 samples) from the Gemma (Zoubarev et al. 2012). Then I selected seven human data sets and four mouse data sets that sampled many tissues in each (tissue panel data sets). These data sets were referred as "experiment-acquired tissue panel data sets" (Exp.Panel). Additionally, I combined computationally the remaining data sets which sampled less tissues into a new tissue panel data set (Cpt.Panel) for each species. For human, 682 smaller human data sets were integrated into a new Cpt.Panel.HS data set. For mouse, 993 smaller data sets were integrated as the Cpt.Panel.MM data set. The human Exp.Panel data sets (Exp.Panel.HS) are summarized in Table 2, while mouse Exp.Panel data sets (Exp.Panel.MM) are summarized in Table 3. A summary of all the smaller data sets in both species is shown in

Table 4.

3.2 Comparable data sets

When I harmonized the tissue description, the number of tissue types shrank as some tissues were aggregated, and some other tissues were excluded if they appeared in only a single data set. These changes for each data set are tracked in Table 2 and Table 3.

Expression data from different platforms have different scales and distribution. I used two thresholds to bin each data set in to three bins: highly expressed, moderately expressed and unexpressed. An example showing how the binning method works is presented in Figure 3.

Table 2 Summary of Experimentally-acquired tissue Panel data sets for human

Rows in green are the statistics summarized from original data, the row in blue tracks the changes of tissue types in tissue aggregation and harmonization. The violet rows track the changes of tissue types and samples in building the atlas. NA: No available content.

Name	GTEx.Seq	HPA	BodyMap	FANTOM5	GTEx.Array	BioGPS.HS	Roth
Data types	RNA-seq	RNA-seq	RNA-seq	RNA-seq	MicroArray (Affymetrix)	MicroArray (Affymetrix)	MicroArray (Affymetrix)
library preparation	Poly(A)+	Poly(A)+	Poly(A)+	5'cap enrich	Poly(A)+	Poly(A)+	Poly(A)+
Platform	Illumina HiSeq 2000/X	Illumina HiSeq2000/2500	Illumina HiSeq 2000	CAGE+ HelioScope	HG 1.1 ST Array	HG-U133A	HG-U133 Plus 2.0 Array
#Genes original	18750	29972	29972	29972	17464	13150	20477
#Samples original	1989	200	500	96	837	158	677
#Tissues original	53	32	16+1	68	12	79	117
#Tissues processed	47	30	16	62	12	54	77
#Samples in Atlas	8555	200	14	94	837	80	574
#Tissues in Atlas	40	30	14	44	10	40	54
Ref	(Lonsdale et al. 2013)	(Uhlén et al. 2015)	(Barbosa-Morais et al. 2012)	(Yu et al. 2015)	(Lonsdale et al. 2013)	(Su et al. 2004)	NA

Table 3 Summary of Experimentally-acquired tissue Panel data sets for mouse

Rows in green are the statistics summarized from original data, the row in blue tracks the changes of tissue types in tissue aggregation and harmonization. The violet rows track the changes of tissue types and samples in building the atlas. NA: No available content.

Name	GSE36025	GSE24207	GSE24940	BioGPS.MM
Data types	RNA-seq	MicroArray (Affymetrix)	MicroArray (Affymetrix)	MicroArray (Affymetrix)
library preparation	Pol(A)+	Pol(A)+	Pol(A)+	Pol(A)+
Platforms	Illumina GAIIx/Hi-Seq	MG- 430 2.0 Array	MG 1.0 ST Array	GNF1M
# Genes original	25393	18602	21340	18483
# Samples original	30	73	40	122
# Tissues original	25	23	13	61
# Tissues processed	23	20	13	55
# Samples in Atlas	27	68	40	94
# Tissues in Atlas	21	18	13	42
Ref	Lin et al., 2014; Pervouchine et al., 2015	Thorrez et al., 2011	Thorrez et al., 2011	Su et al., 2004

Table 4 Summary of Computationally-acquired tissue Panel data sets for human and mouse

Cpt.Panel.HS/ Cpt.Panel.MM : the tissue panel data set that was integrated computationally from many data sets that studied a specific tissue type

Species	**Human**		**Mouse**	
Name of data sets	Cpt.Panel.HS		Cpt.Panel.MM	
Data sets types	Microarray	RNA-seq	Microarray	RNA-seq
# platforms	48	1	45	1
# Data sets	646	36	925	68
# Gene original	26257		24324	
# Samples original	44078		30216	
# Tissues original	189		270	
# Tissue after preprocessing	133		179	
# Samples in atlas	40853		26642	
# Tissues in atlas	52		45	

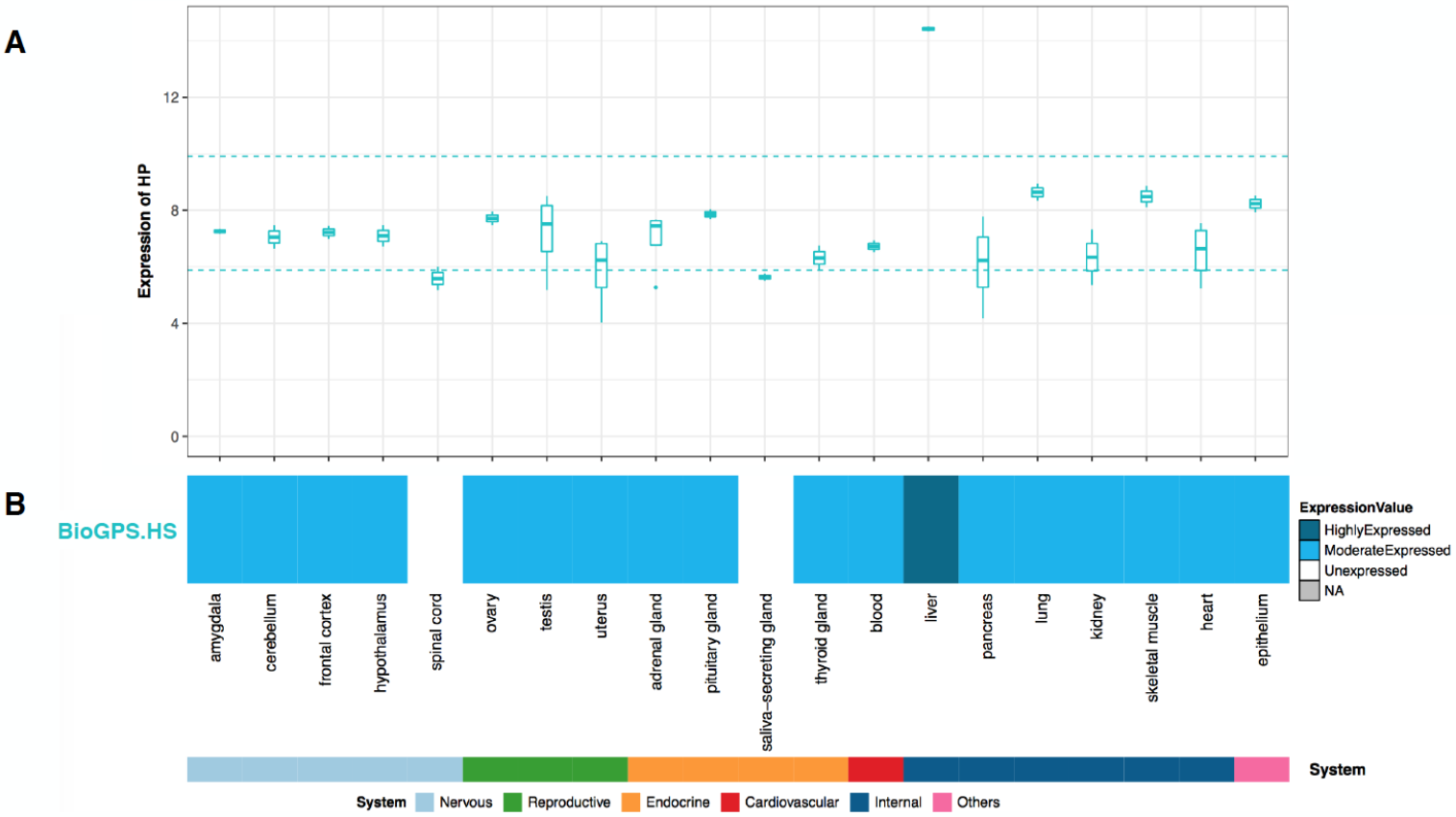

Figure 3 An example of binning method

The cross-tissue expression pattern of human gene *Haptoglobin (HP)* in the BioGPS data set.

(A) A box plot of expression profile of *HP*. Dash lines shows the value of two thresholds. Two thresholds binned the data into three levels. (B) Heatmap represent the three bins with three colors. Dark blue is for highly expressed, light blue is for moderately expressed, white is for unexpressed, grey is for missing data. The color bar named "system" beneath the heatmap shows the biological system where each tissue in the data set belongs to.

3.3 The Gemma Atlas

The Gemma Atlas for each species was constructed from Exp.Panel data sets and one Cpt.Panel data set. Only the genes and tissues that were presented in more than two tissue panel data sets were chosen for the Gemma Atlas. In total, the Gemma Atlas includes 30894 genes across 71 tissues for the human transcriptome, and 21978 genes across 51 tissues for the mouse transcriptome. The number of tissues and samples in each data set was reduced after the unique tissues in each data set were removed (Table 2, Table 3 and Table 4).

Once I got a list of tissues from the Gemma Atlas, I compared the contribution from each tissue panel data set to the whole atlas tissue types (Figure 4 and Figure 5).

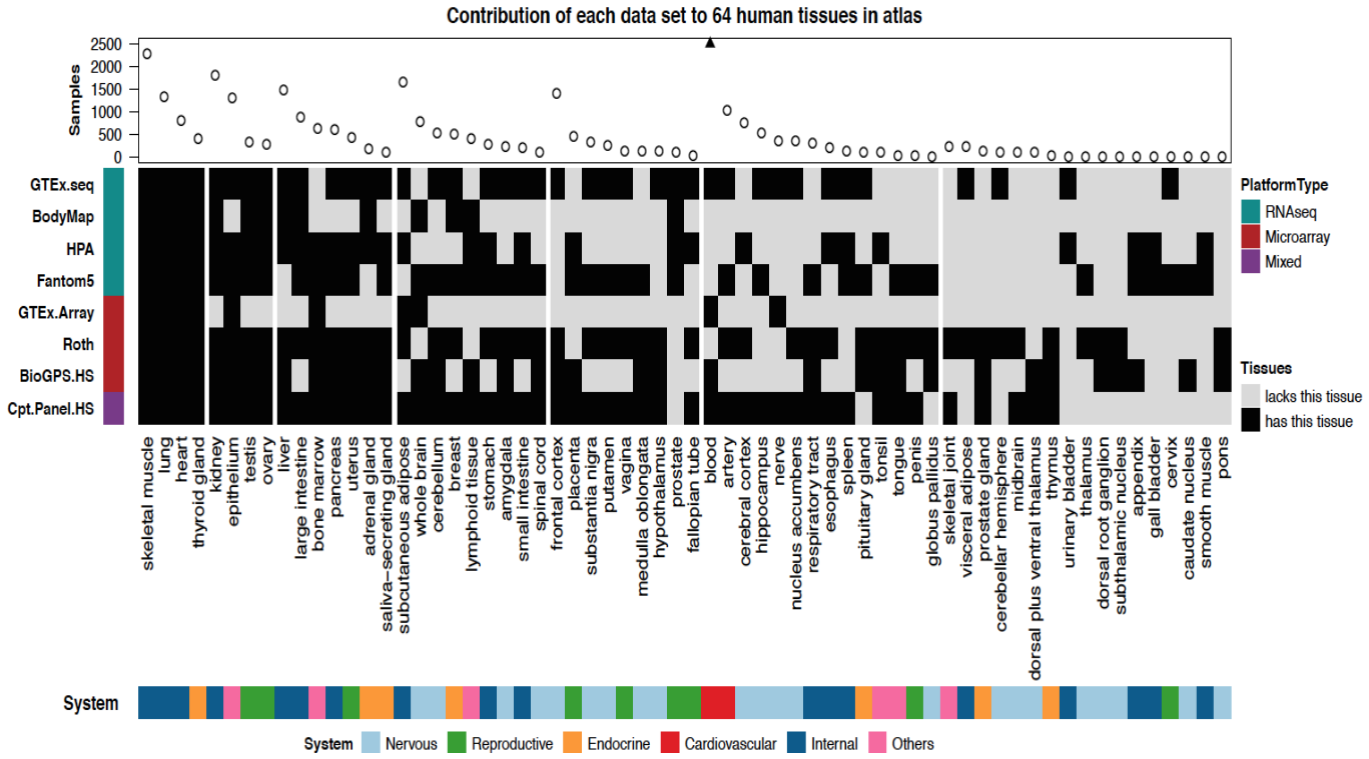

Figure 4 Comparison of tissue coverage among eight human tissue panel data sets

Upper lane: Dot plot of total sample number for each tissue summarized across data sets. If the sample size of a tissue was larger than y axis limitation 2500, it was plotted as a black triangle and was assigned arbitrarily to the maximum number 2500.

Lower lane: Binomial heatmap shows tissue coverage of each data set. Tissues are arranged and classed by the number of data sets they were studied in. Different classes were separated by white strips. The first group of tissues appear in all the data sets, the second group appear in all but one of the data sets, the third in all but two data sets and so on. The color bar named "system" beneath the heatmap shows the biological system where each tissue in the data set belongs to.

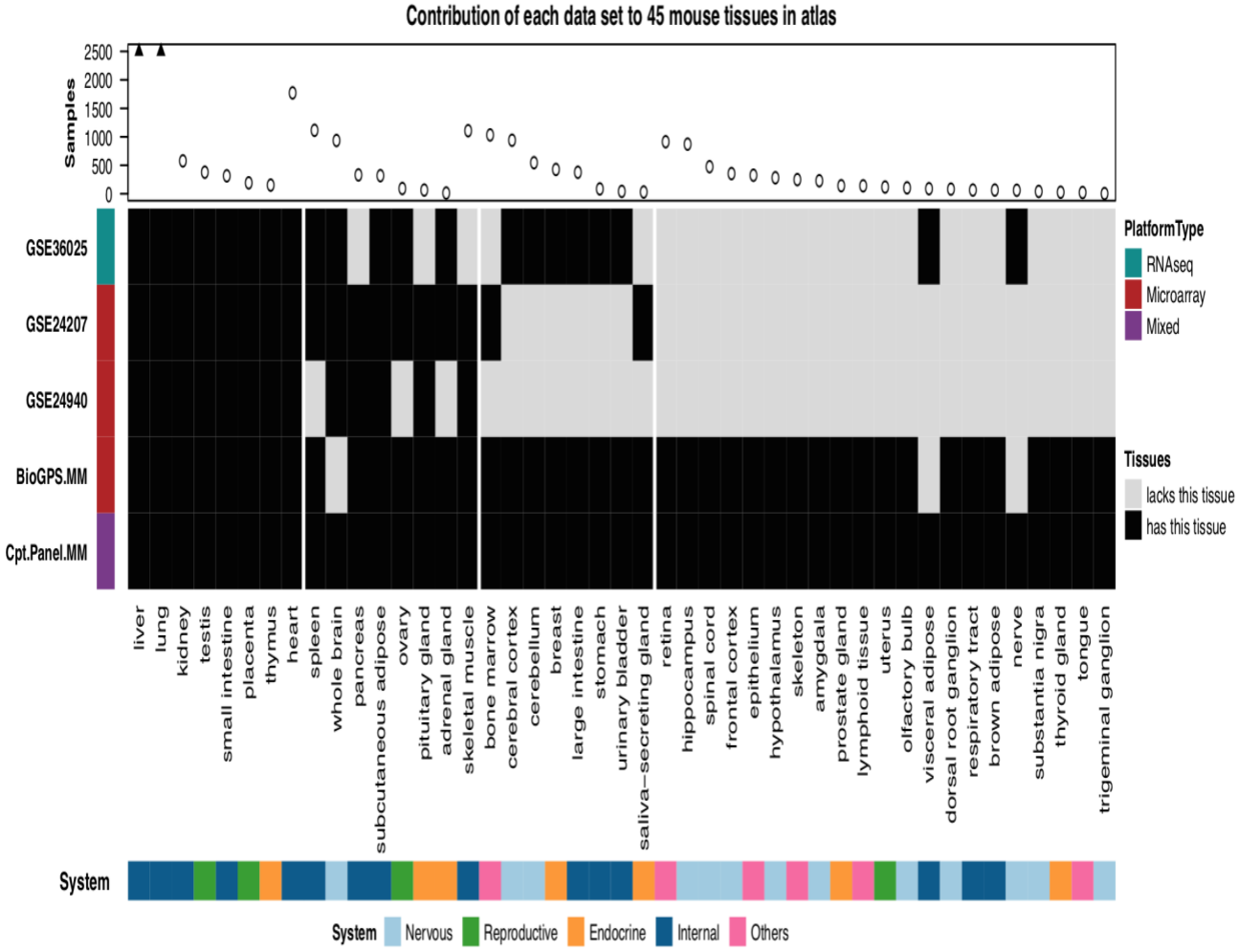

Figure 5 Comparison of tissue coverage among five mouse tissue panel data sets

Upper lane: Dot plot of total sample number for each tissue summarized across data sets. If the sample size of a tissue was larger than y axis limitation 2500, it was plotted as a black triangle and was assigned arbitrarily to the maximum number 2500.

Lower lane: Binary heatmap shows tissue coverage of each data set. Tissues are arranged and classed by the number of data sets they were studied in. Different classes were separated by white strips. The first group of tissues appear in all the data sets, the second group appear in all but one of the data sets, the third in all but two data sets and so on. The color bar named "system" beneath the heatmap shows the biological system where each tissue in the data set belongs to**.**

3.4 Comparison among data sets

Once all the data sets were harmonized and integrated into the Gemma Atlas, the data sets were comparable. I compared data sets from two aspects: the tissue coverage among data sets and the consistency among data sets.

3.4.1 Comparison of tissue coverages

Based on the content in Figure 4 and Figure 5, we can tell that among human tissue panel data sets, the Roth data set and the Cpt.Panel.HS both have the best coverage (49 tissues out of 64 atlas tissues) followed by the GTEx RNA-seq data set (39 tissues out of 64 atlas tissues). Among mouse tissue panel data sets, the BioGPS.MM data set and the Cpt.Panel.MM data set have the best tissue coverage.

3.4.2 Comparison of the agreement level among data sets

The Gemma Atlas was built after all the tissues and expression levels were harmonized into comparable status. The agreement level among data sets was measured at a per-gene per-tissue level between any two data sets. There were three levels of agreements: *perfect agreement,* "*level 1 disagreement" (L1D),* and "*level 2 disagreement" (L2D).* In this project, only L2D was used to evaluate each data set, agreement was showed as interesting exploration, L1D were left for future work.

3.4.2.1 Cases of perfect agreements

The most reliable expression data points should be a perfect agreement that a gene is expressed at the same level in a tissue across all the data sets in the atlas (Figure 6). Ideally there are three types

of perfect agreements corresponding to the three binning levels: *agreement on high expression level*, *agreement on moderate expression level*, and *agreement on unexpressed level*.

Only perfect agreements on the high expression level were found. More cases were that genes agreed with part of the atlas data sets, but not all the data sets. Figure 6 shows an example of perfect agreement using the expression pattern of human gene *Ubiquitin C* (*UBC*). All data sets that contain data for the gene *UBC* indicate a high expression level.

In total, I found nine genes perfectly agreed at the high expression level. These genes are consistently highly expressed not just across different tissues, but also across human and mouse. However it should be remembered that this gene list is defined by a stringent threshold of "high expression". There might be more than nine genes on the list. Names of these nine genes are listed below. If the gene name in mouse is the same with the name in human, only the human gene names will be listed. Otherwise, the gene names in mouse will be added behind the human gene names.

- *Eukaryotic Translation Elongation Factor 2 (EEF2)*
- *Guanine Nucleotide Binding Protein, beta polypeptide 2-like 1(GNB2L1)* in human & *guanine nucleotide-binding protein subunit beta-2-like 1 (LOC708526)* in mouse
- *Heat Shock Protein 90 Alpha Family Class B Member 1 (HSP90AB1)*
- *Integral Membrane Protein 2B (ITM2B)*
- *Prosaposin (PSAP)*
- *Ribosomal Protein L8 (RPL8)*
- *Ribosomal Protein S15 (RPS15)*
- *Ribosomal Protein S5 (RPS5)*
- *Ubiquitin C (UBC)*

3.4.2.2 Cases of level 2 disagreements

An expression pattern that varies across three bins between any two data sets was defined as a L2D. For example, the human gene *Histone Cluster 1 H4 Family Member C* (*HIST1H4C*) has multiple disagreed tissue sites between the BioGPS data set and the GTEx RNA-seq data set. This comparison was chosen as an example of L2Ds showed in Figure 7.

A L2D is a per-gene per-tissue level index between two data sets. Genes with multiple disagreed tissue sites are more likely to be caused by technical artifacts as the L2Ds are non-tissue specific.

The Gemma Atlas offers a comprehensive view of all potential L2Ds for a gene in all the tissue between any data set pairs (Figure 7). Gene *HIST1H4C* shows both disagreed site, as well as some partial agreements. For example, the GTEx RNA-seq data set, the Bodymap data set and Uhlen's data set agreed on most of the tissues with an "unexpressed" for human gene *HIST1H4C*.

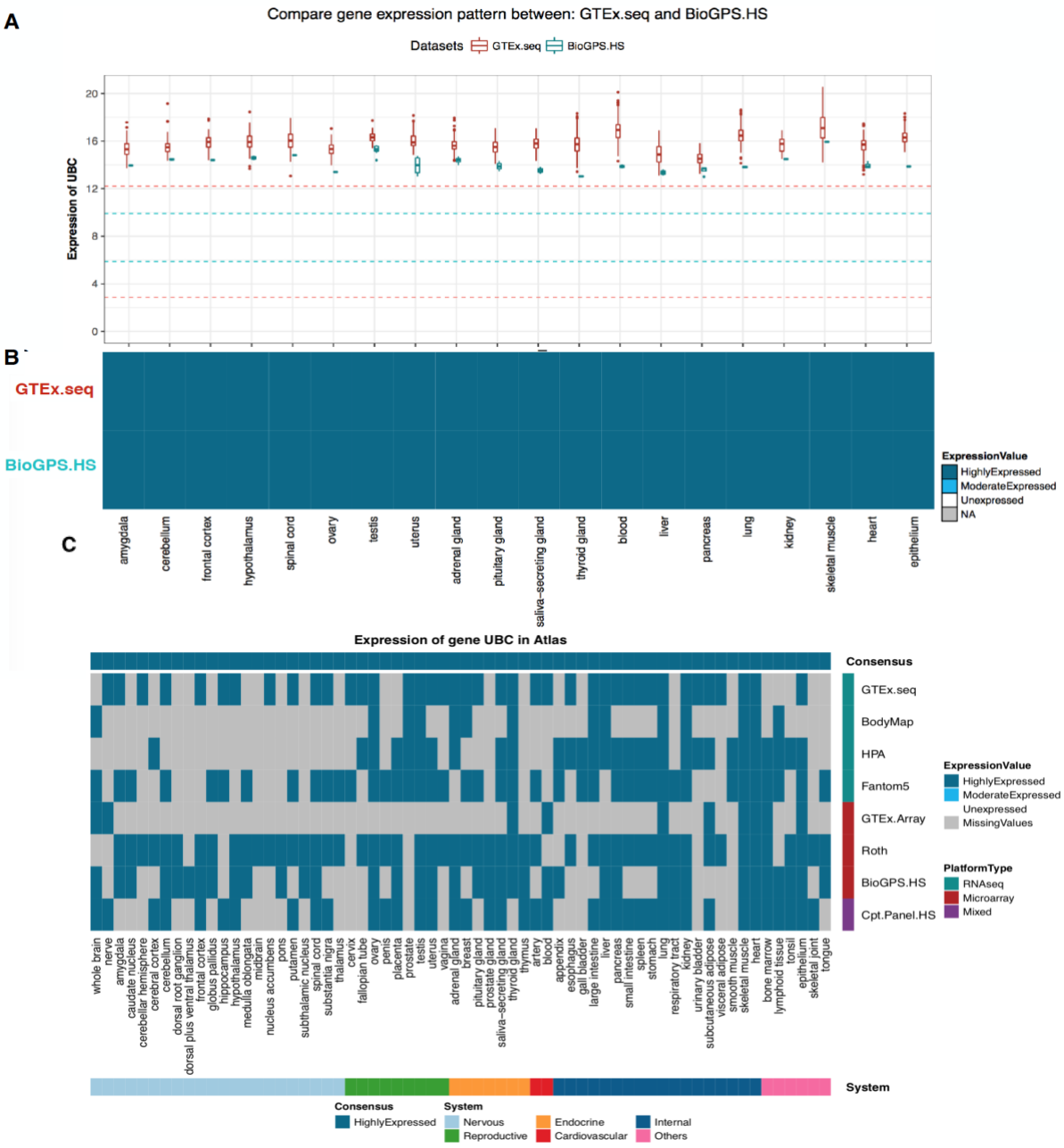

Figure 6 An example of perfect agreement across data sets

Gene *Ubiquitin C (UBC)* is expressed highly in all the tissues across all the human tissue panel data sets. (A) Box plot: *UBC* consistently highly in two Data sets. This is an example of one agreement between two data sets. (B) A heatmap for the same information as plot A. (C) Heatmap in the Gemma atlas

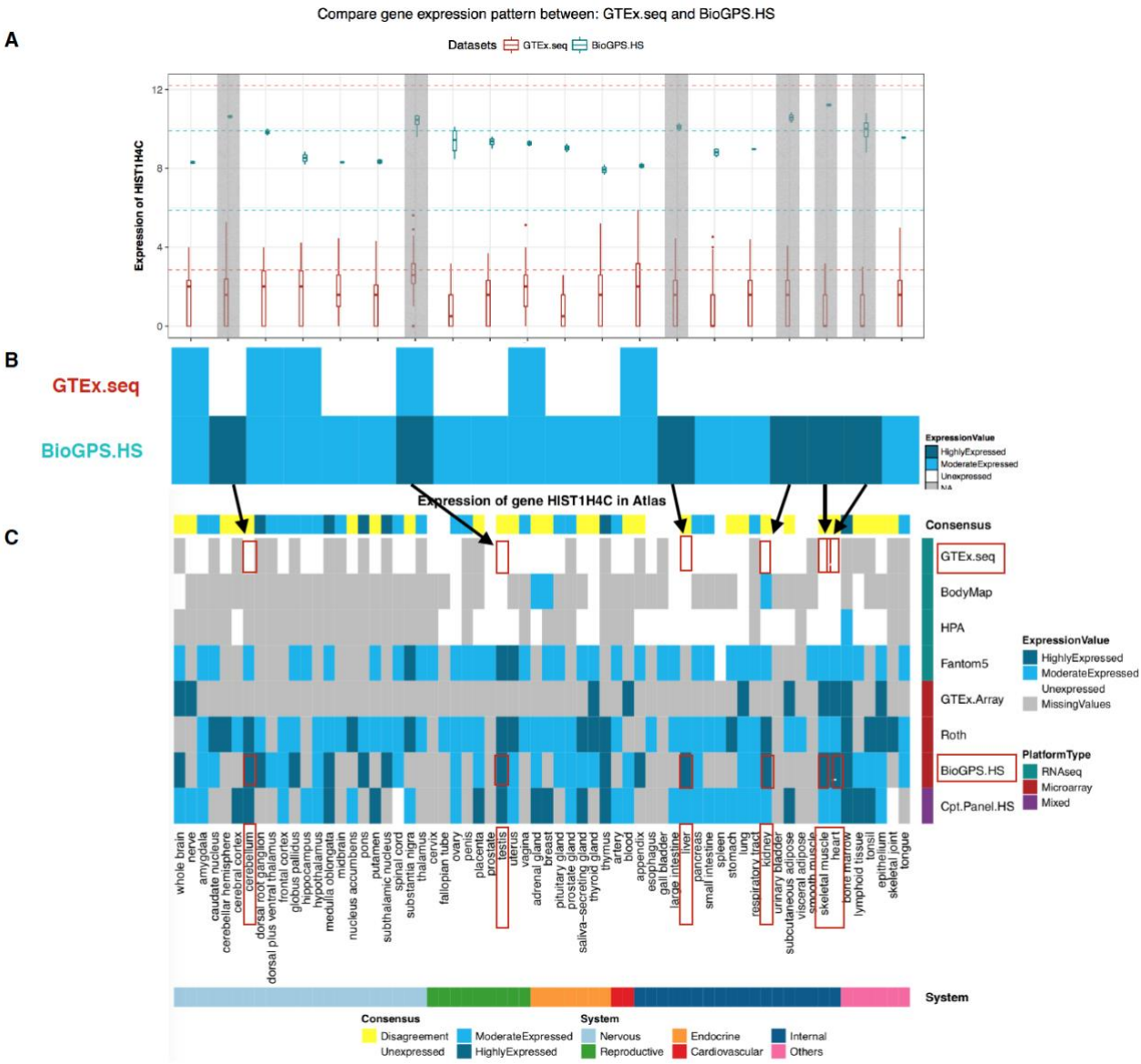

Figure 7 An example of disagreements across data sets

Gene *Histone Cluster 1 H4 Family Member C (HIST1H4C)* expressed highly in all the tissues across all the human tissue panel data sets (A) A box plot presenting expression level *HIST4H1C* in the BioGPS data set (red) and the GTEx data set (blue). The grey strips highlighted the disagreed tissues. The color of dash lines is consistent with the color of data sets. The upper dash line for each data set is the threshold differentiating the upper bin (**highly expressed**) and the middle bin (**moderate expressed**); the lower dash line for each data set is the threshold differentiating the middle bin (**moderate expressed**) and the lower bin (**unexpressed**). (B) heat map of expression pattern of gene *HIST1H4C* in atlas. Dark blue means **highly expressed,** light blue means **moderate expressed,** white means **unexpressed.** (C) disagreements in the Gemma Atlas.

3.5 Data sets were evaluated by the amount of level 2 disagreements

Quantification of all the pair-wise data sets shows that level 2 disagreements (L2Ds) in general are rare, but in certain data sets the amount of L2Ds is high.

The portion of disagreed genes in all common genes between any data sets within a species can be as large as 12.36% (in BioGOS data set). The median of disagreed genes in all common genes between any data sets within a species is 5% ; the median of the BioGPS human pairs is 8%; the median of integrated data sets was 5%; and the median of the remaining data sets was 0.5%. L2Ds from all pairwise comparisons were summarized at points level in Figure 8. There are some genes with L2Ds at multiple tissue sites; these L2Ds were summarized at gene level in Figure 9.

The existence of a L2D (an extreme inconsistency) indicates something really suspicious in at least one of the paired data sets. It is also very difficult to determine which side of the pairs is wrong. However, if a particular data set has numerous L2Ds regardless of which data set it was compared with, that data set could be considered suspect.

For instance, the BioGPS human data set had a relatively high number of L2Ds, no matter which data set it compared with. In contrast the human Roth data set and the human GTEx RNA-seq data sets had less L2Ds no matter which data set they compared with. This offered interesting direction for future study, but it is still not a strong conclusion, as the number of L2Ds was drawn from arbitrary thresholds.

In this part of the results, I plotted human data sets and mouse data sets together (Figure 8, Figure 9). When compared across species, we can tell that mouse data in general has fewer disagreements compared to human data.

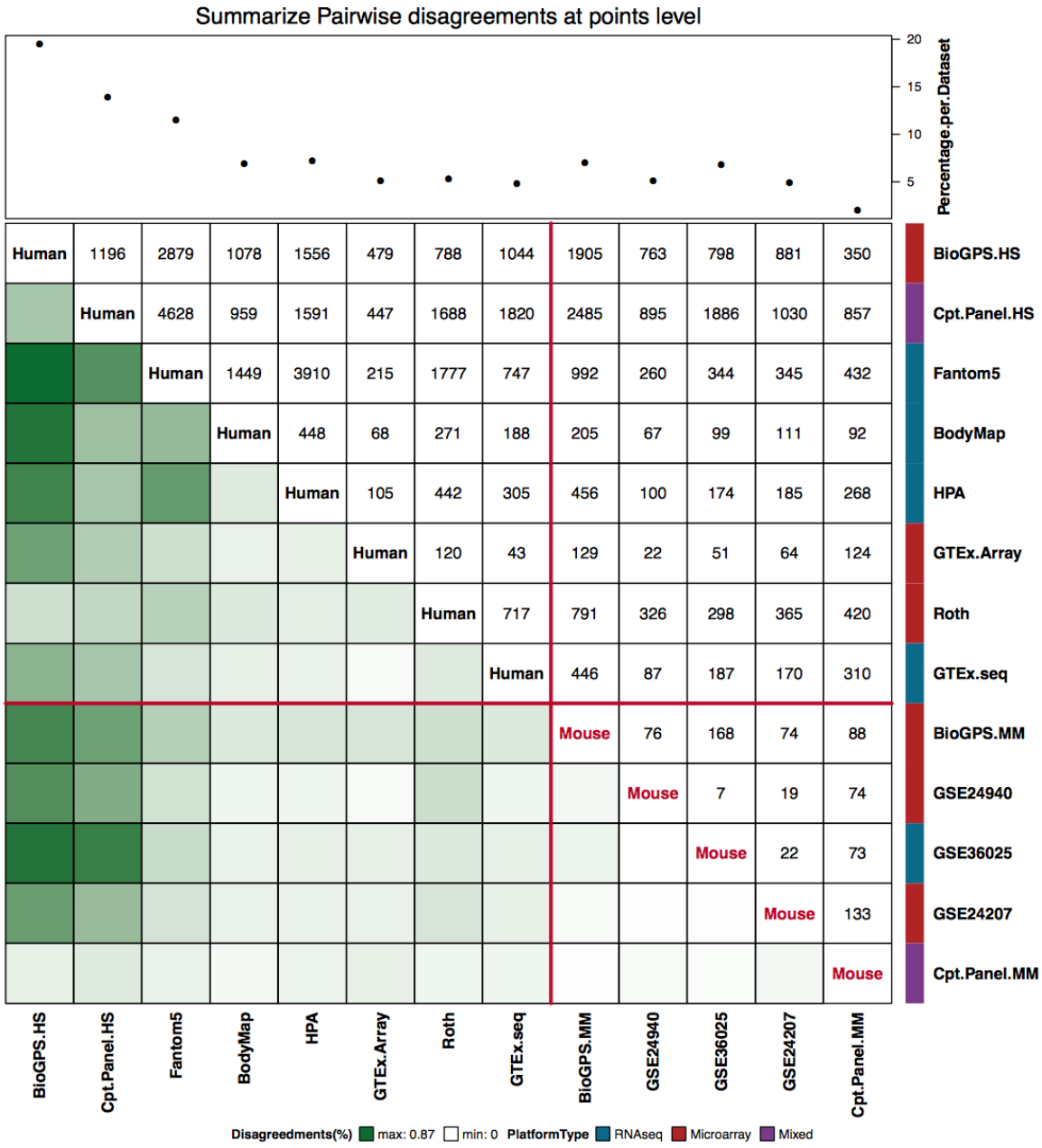

Figure 8 Quantified point-level disagreements between any data set pairs.

The number of disagreements was biased towards certain data sets. **Upper lane:** summarized disagreements from all the pairs compared with one data set. **Lower lane**: Numbers in the upper triangle of the heat map are the exact number of disagreements. The color in the lower triangle are the percentage of the disagreements adjusted by the total number of common genes between any two paired data sets.

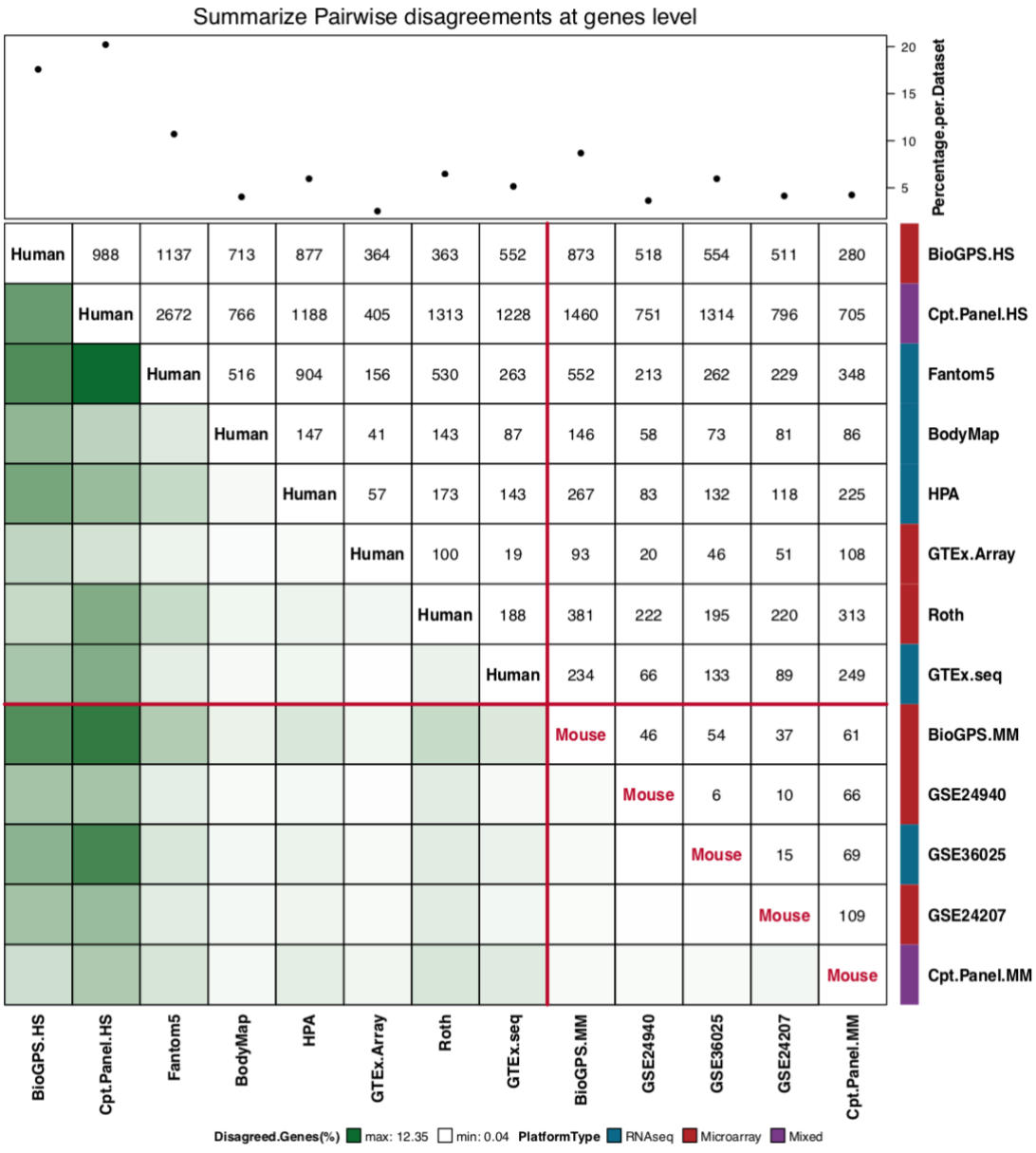

Figure 9 Quantified gene-level disagreements between any data set pairs.

The number of disagreements was biased towards certain data set. **Upper lane:** summarized disagreed gene from all the pairs compared with one data set. **Lower lane**: Numbers in the upper triangle of the heat map are the exact number of disagreed genes. The color in the lower triangle are the percentage of the disagreed genes adjusted by the total number of common genes between any two paired data sets.

3.6 Investigation of level 2 disagreements

3.6.1 Platform difference is not likely to be a major reason for level 2 disagreements

Most of the Level 2 Disagreements (L2Ds) were hard to explain, but there is a good start for my investigation-the GTEx.array data set and the GTEx.seq data set. As technical control, these two data sets narrow down the potential cause of their L2Ds to the only difference-the platform difference between RNA-seq and microarray. There were only 0.11% (19 L2Ds out 1780 common genes) of L2Ds between technical control. The 19 L2Ds were plotted in Figure 10. However, we still don't know why these 19 L2Ds exist, and further study is needed on these 19 genes.

3.6.2 Distribution of all level 2 disagreements varies among tissues

Besides technical artifacts, the biological artifact is another common difference between data sets that may cause L2Ds. However, there is no biological control for me to decode the effects of sample difference (no two data sets analyzed the same samples on the same platform). Still, an exploratory analysis was seeing the distribution of L2Ds among tissues (Figure 11). The number of L2Ds in the plot was corrected by total number of comparison in that tissue. So far, from this exploratory result, we can see the number of L2Ds is not distributed uniformly across the tissues. Some tissues like ovary had more L2Ds than the other tissues with a similar sample size. This may be caused by the differences in the menstrual cycle that cells are in. However, further study is still needed to explain the distribution variation.

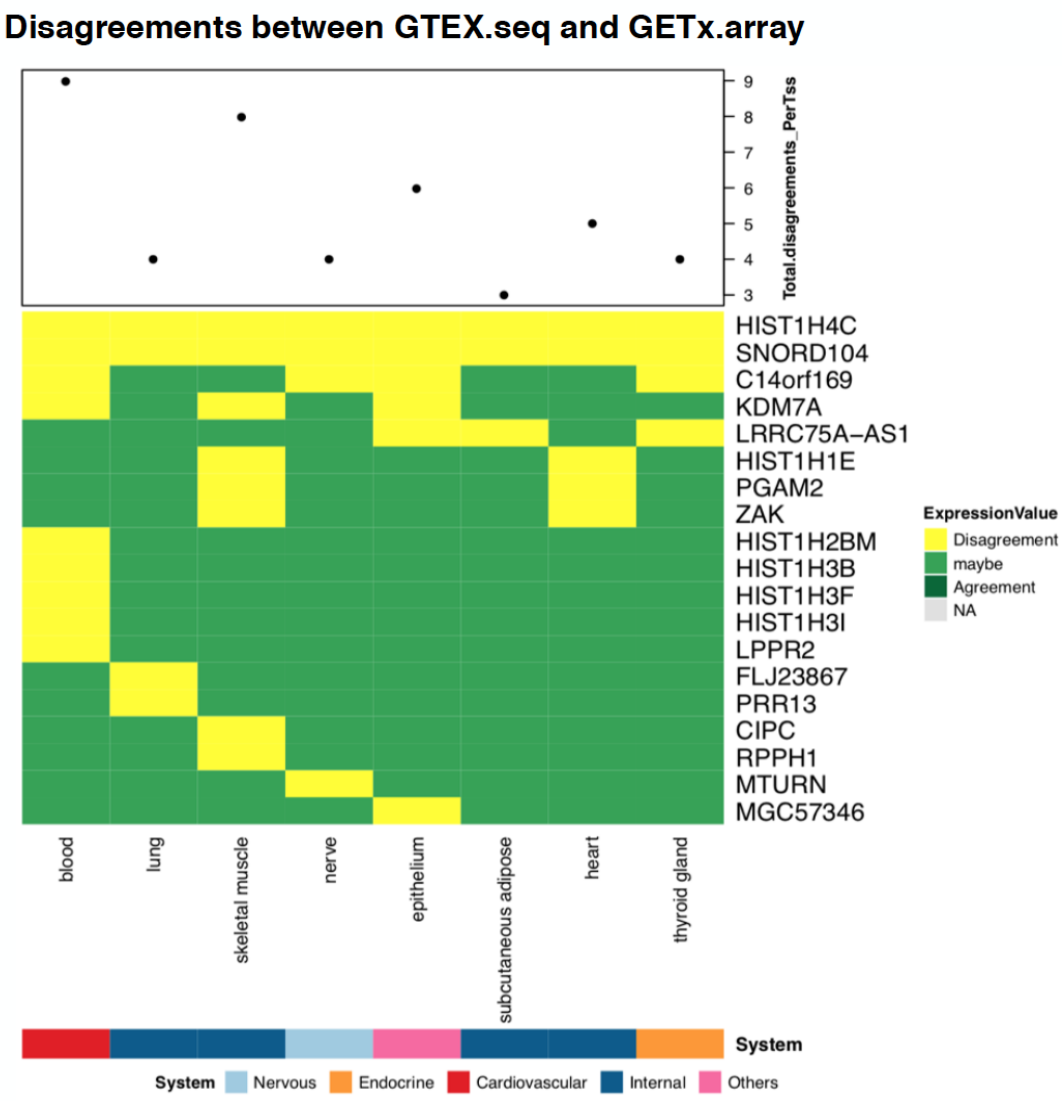

Figure 10 Case study on disagreements between GTEx technical replication data sets

GSE 45878 and GTEx shared the samples, but used different platforms (microarrays and RNA-seq respectively). Upper lane: total number of disagreements in that tissue.

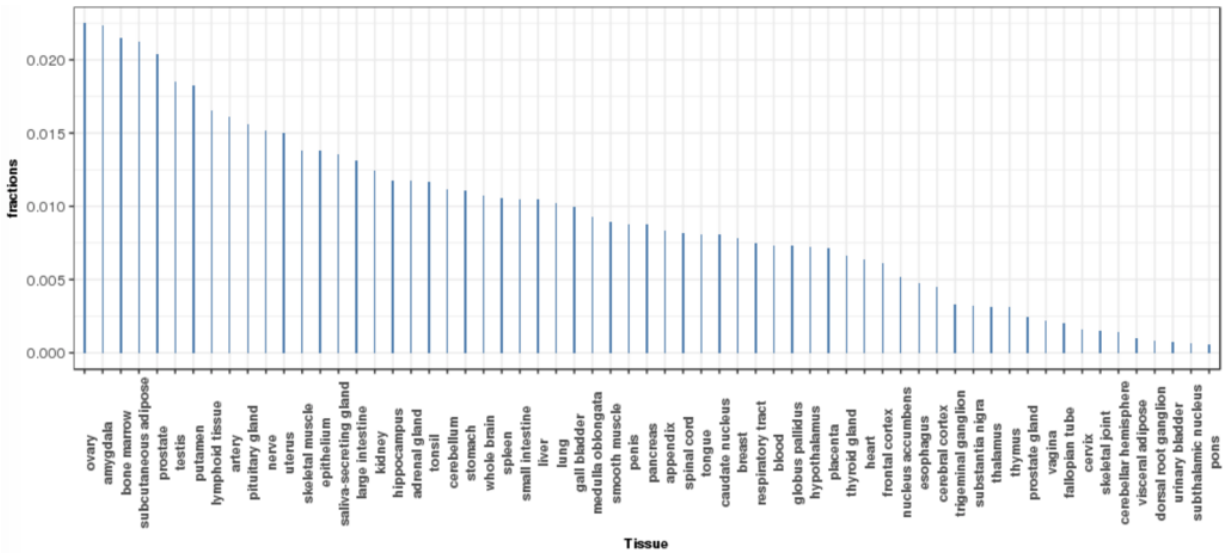

Figure 11 Distribution of all level 2 disagreements varies among tissues

The corrected amount of L2Ds in each tissue is plotted. The total number of L2Ds in a tissue is corrected by diving the total number of comparison in that tissue.

3.7 Exploring conditional expression

All the tissues in Exp.Panel data sets were normal tissues. Based on the expression patterns in normal tissues, I want to know further that if a gene is unexpressed, would it be expressed under certain conditions. I developed an approach "off-on gene" to detect if certain genes could change their expression status under some experimental conditions. With current data, I found that the only conditions that can turn an "off" gene to "on" are either cancer or inflammation related conditions. However, These two conditions are not interested for this project as the tissue identity or cellular composition in cancer or inflammation related tissues has been changed.

3.8 Web application prototype

I developed a web-based interface as a prototype of an integrated Gemma Atlas. Users can choose any gene in either human or mouse, the application offers a box plot of the cross-tissue expression patterns of this gene on the upper right side of the web page (Figure 12). This application also offers a heatmap in the right lower side of the web to present the expression patterns in the Gemma Atlas. However, it is still a prototype that needs to be refined.

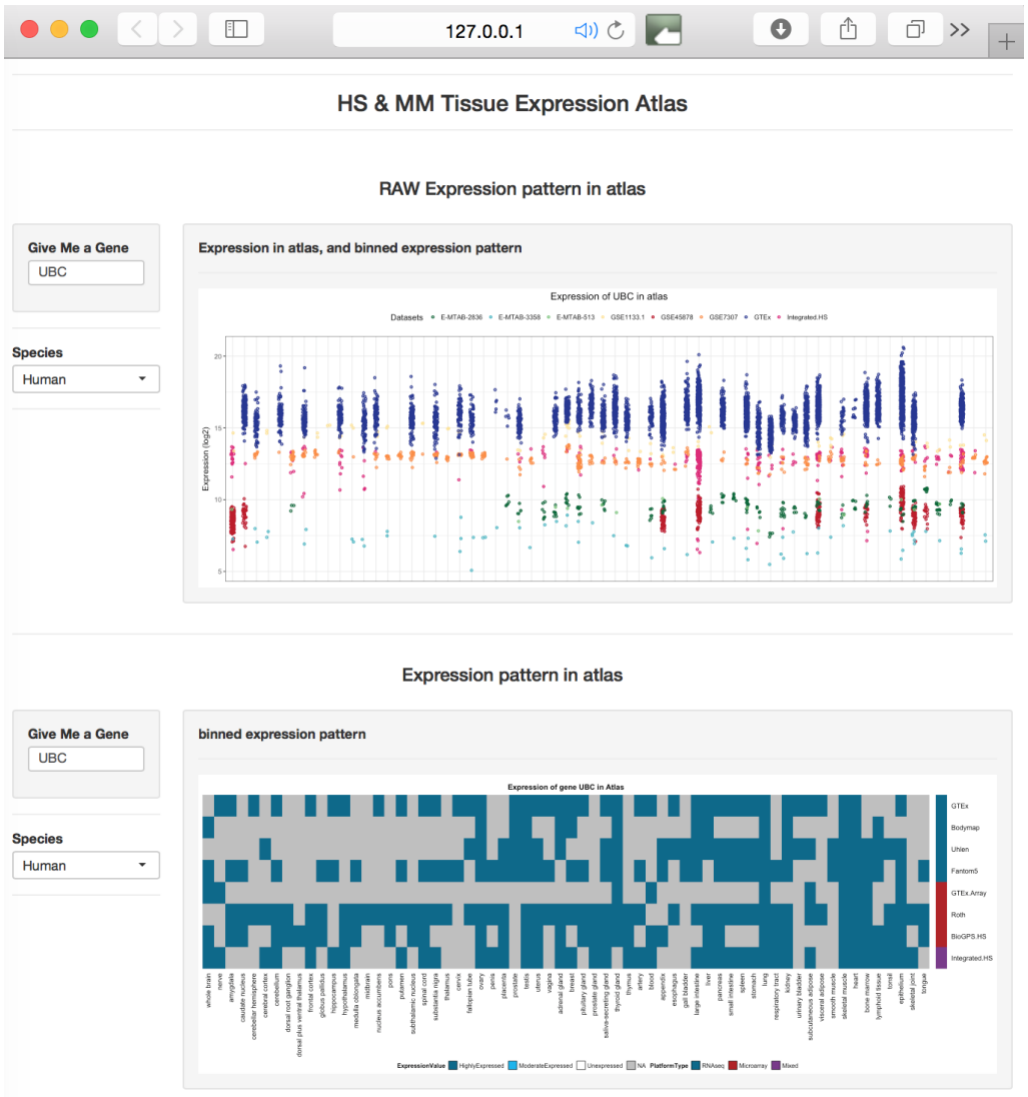

Figure 12 Prototype web application to explore the Gemma Atlas

Point disagreement at a tissue was presented as yellow in the summarized expression pattern. If there is not disagreement, a consensus was made (see methods) and the same color annotation was used to represent highly expressed (dark blue), moderately expressed (light blue), unexpressed (white) and missing value/ NA (grey) in both summarized color bar and supportive processed data.

Chapter 4: Discussion

I have built a new atlas of cross-tissue transcriptomics with the purpose of comparing data sets, indicating cases of consensus or "level 2 disagreement" (L2D), and offering a mechanism to help users understand variations. In the course of creating this atlas, I evaluated the overall comparative consistency of eight human data sets and five mouse data sets to provide a quantitative measurement of general agreement with other data sets. Most L2Ds are not easy to explain. I also did an exploratory analysis and found the expression patterns were stable in common experimental conditions, except in some abnormal cases involving cancer or inflammation.

4.1 Comparison between the Gemma Atlas and a single study

The Gemma Atlas developed in this thesis offered a more comprehensive view of expression patterns compared with any single data set.

First, the atlas has a better coverage for both tissues and genes than any single data set.

Second, the atlas offers more robust summarized expression patterns across data sets, as at least one more data set was included as replicates.

Third, the Gemma Atlas also offered general evaluation on data quality and reproducibility. For example, the L2Ds across data sets can be high (the number gene level L2Ds up to 12.36 % in Figure 9). However, the Gemma Atlas also has its own limitations, which includes but is not limited to first the difficulty of balancing between the amount of missing values and tissue coverage, second the lack of optimum normalization methods to avoid arbitrary thresholds, and third the lack of efficient normalization methods for plotting. These are discussed further in detail.

4.2 The balance between missing values and tissue coverage

One of the initial driving forces for building the tissue expression atlas is to gather data on as many tissues as possible. Thus, I undertook a union of genes and tissues to combine data sets to get a larger coverage as long as they were reported in no less than two data sets. Because of differences in studies, a trade-off to the comprehensiveness is the large amount of missing values (NAs) in the data. As a consequence, the power of analysis and conclusions was reduced for the cases with numerous NAs. For example, there were only three out of 64 human atlas tissues (skeleton muscles, lung and heart) and seven out of 54 mouse tissues (liver, lung, kidney, testis, small intestine, placenta thymus and heart) covered by all the atlas data sets. This fact effects the quantification of L2Ds, as well as downstream analysis like the investigation of L2Ds, the evaluation of data set quality, and the number of consistently highly expressed genes. However, this is a balance which had to be made and I leave this to the atlas users to make their own judgements. One solution for the future can be to give weights/confidence/statistical power on the number of missing values. Further research would build a more complete and robust atlas.

4.3 Impact of arbitrary thresholds used in binning methods

Another balance we need to pay attention to is how stringent the thresholds for binning should be. All conclusions through the whole study were drawn based on the number of L2Ds and *off-on genes*, which initially relied on the definition of "highly expressed" and "unexpressed". On the one hand, I need all my L2Ds to be valid. Thus the more stringent the thresholds are, the more confidence I have on the L2Ds; on the other hand, when it comes to quantification, the loss of true L2Ds could be misleading specifically when conclusions are drawn from the number of L2Ds. For example, what is clear are that "good" data sets like the Roth data set and the GTEx RNA-seq data sets may have more L2Ds than the BioGPS if different thresholds are used. Another impact is on

the investigation of L2Ds. I have a hypothesis that might be tested -- if a gene disagreed at all the tissues, then this gene level L2D is more likely caused by technical artifacts.

Another indication of the over-strict threshold is the number of highly expressed (a total of nine) and conservative genes that I found as examples of agreements. These nine genes are different from the Housekeeping (HK) genes. HK genes have, by definition, functions essential for cells no matter what other specific functions the cells may have. Thus they should be expressed universally and all the time across normal tissues. These nine genes are not only HK genes, they are also consistently expressed at a high level across tissues, across data sets, even across human and mouse these two species. Ribosomal genes are co-expressed. However, only three ribosomal genes were detected as expressed with the same pattern (highly expressed in all the tissues). I suspect that more ribosome genes might be missing. This needs to be studied in detail by someone to further this project. I hypothesize this is as a result of the strict threshold for highly expressed genes. Refining the thresholds is a question for future research and data analysis, together with the inclusion of less stringent *level 1 disagreements*.

4.4 Methods used to plot expression data from different platforms into one plot

Plotting raw data showing all the data points offers an insight from sample size. However, Fragments Per Kilobase of transcript per Million mapped reads (FPKM) from RNA-seq data sets have a higher dynamic range than microarray data and finding a satisfying method to transform both data types into the same range for plotting is not easy. Log_2 is commonly used in microarray normalization like RMA. In the atlas, I made expression plots of any given gene at a log_2 transformed scale. This method made most microarray data and RNA-seq data plot in the same range. However, there still are some genes that can't be plotted in the same range. Finding a better method to plot data from different platforms into one range is for future work.

4.5 Data set reproducibility

There are many previous studies comparing RNA-seq data and microarray data, and they reported an optimistic correlation of expression data between the two technologies (Mantione et al. 2014; Richard et al. 2014; Shi et al. 2006). Also a good correlation within RNA-seq technologies was reported in some studies. For example, the expression data generated with the CAGE protocol was reported highly correlated with the expression data generated with Illumine RNA-seq protocol (Kawaji et al. 2014b). However, evaluating consistency in general trends across genes with correlation may miss some inconsistencies at per-gene level which might be interesting for some researchers like the *HIST* family mentioned above. With the approach of studying disagreements in my study, I found the percentage of disagreed genes varied from 0.5% to 12% between data sets, a variation which in my opinion is high.

4.6 Efforts to explain level 2 disagreements

Level 2 disagreements are conflicted expression status (highly expressed VS unexpressed) from two data sets. I hypothesized that big differences like L2Ds happen for a reason rather than random effect. However, a mixture of biological effects with technical effects made the detangle very difficult. Luckily, I have a pair of GTEx data sets that share samples. The 19 L2Ds between the GTEx microarray data set and the GTEx RNA-seq data set were only possible caused by technical factors (Figure 10). They took only 0.11% of the total common genes between the two data sets. It is also difficult to narrow down the causes for these 19 L2Ds. I tried to make a few efforts.

I noticed that they all were unexpressed in GTEx.seq data set. Compared with mis-detecting unexpressed as highly expressed, it is more likely that these 19 genes were missed by GTEx.seq data set. When I did case study on these 19 genes, I noticed that these 19 genes were in two

categories: noncoding genes and coding genes. Both GTEx microarray data set and the GTEx RNA-seq data set prepared their library by Poly(A) selection, how can coding gene be missed by RNA-seq data set?

In a previous study, some protein-coding genes, like some *HIST* genes, were found having no poly(A) tails (Zhang et al. 2014). I combined a list of genes that were non poly(A) modified with a list of genes that could conditionally lose their poly(A) tails (bimorphic genes) from Yang's study (L. Yang et al. 2011), coming up my own list of non-poly(A)- modified genes.

I found all the 19 genes were either non-coding genes or non-poly(A)–modified genes. In either case, these genes would be predicted to be omitted by poly(A) selection. Then how could GTEx microarray data set have these genes in its Ploy(A) selected library and with high expressions? There is a chance that 19 genes were omitted by poly(A) selection for microarray data set by chance, but this seems very unlikely. More evidence is need to solve the mystery.

With unexplained L2Ds, the evaluation of each data set might be misleading. Any L2D between two data sets could be caused by the poor data quality of either data set. For example, it is possible, though very unlikely, that all the L2Ds with the BioGPS data set are caused by the mistakes from other data sets, then the BioGPS data set will be the "perfect data set" with no mistakes, which is the opposite conclusion from my evaluation.

4.7 Off-on genes were rare

I explored the impact of conditional expression on cross-tissue gene expression patterns. I found that genes in abnormal tissues could only be switched from "*off*" to "*on*" when the tissues had cancer or inflammation. In another word, no gene expression could be changed as dramatic as from "*off*" to "*on*" in "normal" conditions. This suggests gene expression patterns are relatively stable,

in the sense that genes which are not expressed in a tissue tend to stay that way under a wide range of conditions. However, this conclusion might be limited by the amount of available data and investigation of more conditions might identify off-on genes.

4.8 Future work in improving the summarization method

The method used to summarize expression patterns across data sets in this thesis project is just a prototype. For the future work, an improved summarization method needs to be developed. The summarization should follow the three rules below:

Rule 1: If all the datasets agree perfectly on the expression level of a gene, the observed pattern is a "consensus".

Rule 2: If the expression pattern disagrees between any two data sets, this case is marked as an "L2D" in summarization. The assumption here is the L2D is caused by "mistakes" or technical artifacts that need to be investigated and explained rather than normal tolerable variation. Once the contributed "mistake" is identified and removed, the expression pattern can be summarized from the rest of the data sets.

Rule 3: If the expression pattern is neither perfectly agreed (variable) across data sets nor disagreed, an integrated value should be obtained to represent the majority of data sets. This is a "democratic" strategy, but the integrated value needs to be adjusted with confidence level.

In this thesis, my summarization strategy did not take missing values and confidence levels into account. There are also a number of unexplained L2Ds. These are future work for an improved atlas and could be a subject of further study.

Bibliography

Agarwala, Richa, Tanya Barrett, Jeff Beck, Dennis A. Benson, Colleen Bollin, Evan Bolton, Devon Bourexis, et al. 2018. "Database Resources of the National Center for Biotechnology Information." *Nucleic Acids Research* 46 (D1): D8–13. https://doi.org/10.1093/nar/gkx1095.

Barbosa-Morais, N. L., M. Irimia, Q. Pan, H. Y. Xiong, S. Gueroussov, L. J. Lee, V. Slobodeniuc, et al. 2012. "The Evolutionary Landscape of Alternative Splicing in Vertebrate Species." *Science (New York, N.Y.)* 338 (6114): 1587–93. https://doi.org/10.1126/science.1230612.

Barnes, Michael, Johannes Freudenberg, Susan Thompson, Bruce Aronow, and Paul Pavlidis. 2005. "Experimental Comparison and Cross-Validation of the Affymetrix and Illumina Gene Expression Analysis Platforms." *Nucleic Acids Research* 33 (18): 5914–23. https://doi.org/10.1093/nar/gki890.

Cerase, Andrea, Greta Pintacuda, Anna Tattermusch, and Philip Avner. 2015. "Xist Localization and Function: New Insights from Multiple Levels." *Genome Biology* 16 (August): 166. https://doi.org/10.1186/s13059-015-0733-y.

Head, Steven R., H. Kiyomi Komori, Sarah A. LaMere, Thomas Whisenant, Filip Van Nieuwerburgh, Daniel R. Salomon, and Phillip Ordoukhanian. 2014. "Library Construction for next-Generation Sequencing: Overviews and Challenges." *BioTechniques* 56 (2): 61–passim. https://doi.org/10.2144/000114133.

Kawaji, Hideya, Marina Lizio, Masayoshi Itoh, Mutsumi Kanamori-Katayama, Ai Kaiho, Hiromi Nishiyori-Sueki, Jay W. Shin, et al. 2014a. "Comparison of CAGE and RNA-Seq Transcriptome Profiling Using Clonally Amplified and Single-Molecule next-Generation Sequencing." *Genome Research* 24 (4): 708–17. https://doi.org/10.1101/gr.156232.113.

———. 2014b. “Comparison of CAGE and RNA-Seq Transcriptome Profiling Using Clonally Amplified and Single-Molecule next-Generation Sequencing.” *Genome Research* 24 (4): 708–17. https://doi.org/10.1101/gr.156232.113.

King, Hadley C., and Animesh A. Sinha. 2001. “Gene Expression Profile Analysis by DNA Microarrays: Promise and Pitfalls.” *JAMA* 286 (18): 2280–88. https://doi.org/10.1001/jama.286.18.2280.

Krupp, Markus, Jens U. Marquardt, Ugur Sahin, Peter R. Galle, John Castle, and Andreas Teufel. 2012. “RNA-Seq Atlas--a Reference Database for Gene Expression Profiling in Normal Tissue by next-Generation Sequencing.” *Bioinformatics (Oxford, England)* 28 (8): 1184–85. https://doi.org/10.1093/bioinformatics/bts084.

Li, Bo, and Colin N. Dewey. 2011. “RSEM: Accurate Transcript Quantification from RNA-Seq Data with or without a Reference Genome.” *BMC Bioinformatics* 12 (August): 323. https://doi.org/10.1186/1471-2105-12-323.

Lin, Shin, Yiing Lin, Joseph R. Nery, Mark A. Urich, Alessandra Breschi, Carrie A. Davis, Alexander Dobin, et al. 2014. “Comparison of the Transcriptional Landscapes between Human and Mouse Tissues.” *Proceedings of the National Academy of Sciences* 111 (48): 17224–29. https://doi.org/10.1073/pnas.1413624111.

Lonsdale, John, Jeffrey Thomas, Mike Salvatore, Rebecca Phillips, Edmund Lo, Saboor Shad, Richard Hasz, et al. 2013. “The Genotype-Tissue Expression (GTEx) Project.” Comments and Opinion. Nature Genetics. May 29, 2013. https://doi.org/10.1038/ng.2653.

Mantione, Kirk J, Richard M. Kream, Hana Kuzelova, Radek Ptacek, Jiri Raboch, Joshua M. Samuel, and George B. Stefano. 2014. “Comparing Bioinformatic Gene Expression Profiling

Methods: Microarray and RNA-Seq." *Medical Science Monitor Basic Research* 20 (August): 138–41. https://doi.org/10.12659/MSMBR.892101.

Melé, Marta, Pedro G. Ferreira, Ferran Reverter, David S. DeLuca, Jean Monlong, Michael Sammeth, Taylor R. Young, et al. 2015. "The Human Transcriptome across Tissues and Individuals." *Science* 348 (6235): 660–65. https://doi.org/10.1126/science.aaa0355.

Mungall, Christopher J., Carlo Torniai, Georgios V. Gkoutos, Suzanna E. Lewis, and Melissa A. Haendel. 2012. "Uberon, an Integrative Multi-Species Anatomy Ontology." *Genome Biology* 13 (January): R5. https://doi.org/10.1186/gb-2012-13-1-r5.

Palasca, Oana, Alberto Santos, Christian Stolte, Jan Gorodkin, and Lars Juhl Jensen. 2018. "TISSUES 2.0: An Integrative Web Resource on Mammalian Tissue Expression." *Database: The Journal of Biological Databases and Curation* 2018 (February). https://doi.org/10.1093/database/bay003.

Richard, Arianne C., Paul A. Lyons, James E. Peters, Daniele Biasci, Shaun M. Flint, James C. Lee, Eoin F. McKinney, Richard M. Siegel, and Kenneth G. C. Smith. 2014. "Comparison of Gene Expression Microarray Data with Count-Based RNA Measurements Informs Microarray Interpretation." *BMC Genomics* 15: 649. https://doi.org/10.1186/1471-2164-15-649.

Ritchie, Matthew E., Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, Wei Shi, and Gordon K. Smyth. 2015. "Limma Powers Differential Expression Analyses for RNA-Sequencing and Microarray Studies." *Nucleic Acids Research* 43 (7): e47–e47. https://doi.org/10.1093/nar/gkv007.

Shi, Leming, Laura H. Reid, Wendell D. Jones, Richard Shippy, Janet A. Warrington, Shawn C. Baker, Patrick J. Collins, et al. 2006. "The MicroArray Quality Control (MAQC) Project Shows Inter- and Intraplatform Reproducibility of Gene Expression Measurements." *Nature*

Biotechnology 24 (9): 1151–61. https://doi.org/10.1038/nbt1239.

Su, Andrew I., Michael P. Cooke, Keith A. Ching, Yaron Hakak, John R. Walker, Tim Wiltshire, Anthony P. Orth, et al. 2002. "Large-Scale Analysis of the Human and Mouse Transcriptomes." *Proceedings of the National Academy of Sciences* 99 (7): 4465–70. https://doi.org/10.1073/pnas.012025199.

Su, Andrew I., Tim Wiltshire, Serge Batalov, Hilmar Lapp, Keith A. Ching, David Block, Jie Zhang, et al. 2004. "A Gene Atlas of the Mouse and Human Protein-Encoding Transcriptomes." *Proceedings of the National Academy of Sciences of the United States of America* 101 (16): 6062–67. https://doi.org/10.1073/pnas.0400782101.

Teng, Mingxiang, Michael I. Love, Carrie A. Davis, Sarah Djebali, Alexander Dobin, Brenton R. Graveley, Sheng Li, et al. 2016. "A Benchmark for RNA-Seq Quantification Pipelines." *Genome Biology* 17: 74. https://doi.org/10.1186/s13059-016-0940-1.

"The Genotype-Tissue Expression (GTEx) Project." 2013. *Nature Genetics* 45 (6): 580–85. https://doi.org/10.1038/ng.2653.

Toker, Lilah, Min Feng, and Paul Pavlidis. 2016. "Whose Sample Is It Anyway? Widespread Misannotation of Samples in Transcriptomics Studies." *F1000Research* 5 (September): 2103. https://doi.org/10.12688/f1000research.9471.2.

Uhlén, Mathias, Linn Fagerberg, Björn M. Hallström, Cecilia Lindskog, Per Oksvold, Adil Mardinoglu, Åsa Sivertsson, et al. 2015a. "Tissue-Based Map of the Human Proteome." *Science* 347 (6220): 1260419. https://doi.org/10.1126/science.1260419.

———. 2015b. "Tissue-Based Map of the Human Proteome." *Science* 347 (6220): 1260419.

https://doi.org/10.1126/science.1260419.

Yang, Jialiang, Tao Huang, Francesca Petralia, Quan Long, Bin Zhang, Carmen Argmann, Yong Zhao, et al. 2015. “Synchronized Age-Related Gene Expression Changes across Multiple Tissues in Human and the Link to Complex Diseases.” *Scientific Reports* 5 (October): 15145. https://doi.org/10.1038/srep15145.

Yang, Li, Michael O. Duff, Brenton R. Graveley, Gordon G. Carmichael, and Ling-Ling Chen. 2011. “Genomewide Characterization of Non-Polyadenylated RNAs.” *Genome Biology* 12 (February): R16. https://doi.org/10.1186/gb-2011-12-2-r16.

Yu, Nancy Yiu-Lin, Björn M. Hallström, Linn Fagerberg, Fredrik Ponten, Hideya Kawaji, Piero Carninci, Alistair R. R. Forrest, et al. 2015. “Complementing Tissue Characterization by Integrating Transcriptome Profiling from the Human Protein Atlas and from the FANTOM5 Consortium.” *Nucleic Acids Research* 43 (14): 6787–98. https://doi.org/10.1093/nar/gkv608.

Zhang, Xiao-Ou, Qing-Fei Yin, Ling-Ling Chen, and Li Yang. 2014. “Gene Expression Profiling of Non-Polyadenylated RNA-Seq across Species.” *Genomics Data* 2 (August): 237–41. https://doi.org/10.1016/j.gdata.2014.07.005.

Zoubarev, Anton, Kelsey M. Hamer, Kiran D. Keshav, E. Luke McCarthy, Joseph Roy C. Santos, Thea Van Rossum, Cameron McDonald, et al. 2012. “Gemma: A Resource for the Reuse, Sharing and Meta-Analysis of Expression Profiling Data.” *Bioinformatics (Oxford, England)* 28 (17): 2272–73. https://doi.org/10.1093/bioinformatics/bts430.