**Title**
Gene expression deconvolution and co-expression methods



**Author**
Cai, Chaochao

**Publication Date**
2013

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

# Gene expression deconvolution and co-expression methods

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in Human Genetics

by

Chaochao Cai

2013

ABSTRACT OF THE DISSERTATION

Gene expression deconvolution and co-expression methods

by

Chaochao Cai

Doctor of Philosophy in Human Genetics

University of California, Los Angeles, 2013

Professor Steve Horvath, Chair

Gene expression analysis provides the link between genome information and phenotype, and is widely used in biomedical research. With the rapid advance of high-throughput technology, it is feasible to measure global mRNA expression in multiple samples at low cost. Over the past decade, many computational and statistical methods have been developed to interpret large-scale gene expression data. However, two questions still have not been thoroughly investigated: 1) how to study gene expression preservation across different tissues, like between brain and blood; and 2) how to analyze the gene expression data generated from heterogeneous tissues comprised of many cell types?

Blood samples are an important surrogate to study neurological diseases due to the limited access of brain samples. My dissertation first investigated the gene expression preservation between

brain and blood by cross-referencing three brain expression data sets (from cortex, cerebellum and caudate nucleus) with two large blood data sets. While previous studies have focused on the preservation of individual gene expression levels across the two tissues, I utilized a systems biology approach to study the preservation of gene co-expression modules. Since oligodendrocytes, astrocytes, and neurons are not present in blood, it is not surprised that only a handful of human brain modules showed evidence of preservation in human blood while global transcriptome organization is poorly preserved. These shared relationships characterized here may aid future efforts to identify blood biomarkers for neurological and neuropsychiatric diseases when brain tissue samples are unavailable.

For the second question, several previous publications have proposed gene expression deconvolution methods, including estimating cell abundances or cell type-specific gene expression (CTSE) values, for admixed samples comprised of distinct cell types. These methods have not yet been widely adopted since comprehensive empirical evaluations are needed to assess their reliability. Here I evaluated different types of expression deconvolution methods in four empirical data sets, including a neuro-scientific application. Since cell type-specific estimation of the mean value for individual genes is sometimes problematic, we propose to consider sets of genes (as opposed to individual genes) and show that this can increase the accuracy of CTSE estimation. Furthermore, comprehensive simulation studies are used to evaluate the effect of mis-specifying cell types. Our simulations indicated that erroneously omitting cell types from the analysis only has an adverse effect on CTSE estimation if the omitted cell type has a high abundance. We also present two R functions, *proportionsInAdmixture* and *populationMeansInAdmixture*, which implement cell abundance estimation and CTSE estimation, respectively.

The dissertation of Chaochao Cai is approved.

Daniel H. Geschwind

Eleazar Eskin

Beth D. Jamieson

Steve Horvath, Committee Chair

University of California, Los Angeles

2013

This thesis is dedicated to my family:

Rui Luo, Minghua Cai, and Juxiang Bao

for their unconditional love and support

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First and foremost, I would like to express my sincere gratitude to my mentor Dr. Steve Horvath for his continuous support over the past five years. I could not have imagined having a better mentor for my Ph.D study.

I sincerely thank my committee members, Drs. Daniel H. Geschwind, Eleazar Eskin, Beth D. Jamieson, and Paul S. Mischel for their time and useful suggestions. I really appreciate their kindly help and that they made it very easy for me to schedule meetings. I also thank Peter Langfelder, Michael C. Oldham, and Jeremy Miller for helpful technical assistance.

Last but not least, I am also thankful for the ACCESS, Human Genetics, and Biostatistics programs for admitting me into these great programs and providing the excellent curriculum.

VITA

## EDUCATION

| | |
|---|---|
| 2009-2012 | Master of Sciences, Biostatistics<br>Department of Biostatistics,<br>University of California, Los Angeles |
| 2004-2008 | Bachelor of Science<br>College of Life Sciences<br>Zhejiang University |

## PUBLICATIONS

**Cai C**, Langfelder P, Fuller TF, Oldham MC, Luo R, van den Berg LH, Ophoff RA, Horvath S. Is human blood a good surrogate for brain tissue in transcriptional studies? BMC Genomics. 2010 Oct 20; 11:589

**Cai C,** Miller JA, Lein ES, Kurian SM, Salomon DR, Horvath S. Empirical evaluation of expression deconvolution methods. *Submmited*

Miller JA, **Cai C**, Langfelder P, Geschwind DH, Kurian SM, Salomon DR, Horvath S. Strategies for aggregating gene expression data: the collapseRows R function. BMC Bioinformatics. 2011 Aug 4; 12:322

Wang A, Huang K, Shen Y, Xue Z, **Cai C**, Horvath S, Fan G. Functional modules distinguish human induced pluripotent stem cells from embryonic stem cells. Stem Cells Dev. 2011 Nov; 20(11):1937-50

Leuchter AF, Cook IA, Hunter AM, **Cai C**, Horvath S. Resting-state quantitative electroencephalography reveals increased neurophysiologic connectivity in depression. PLoS One. 2012; 7(2): e32508

Luo R, Sanders SJ, Tian Y, Voineagu I, Huang N, Chu SH, Klei L, **Cai C**, Ou J, Lowe JK, Hurles ME, Devlin B, State MW, Geschwind DH. Genome-wide transcriptome profiling reveals the functional impact of rare de novo and recurrent CNVs in autism spectrum disorders. The American Journal of Human Genetics. 2012 Jul 13; 91(1):38-55

Horvath S, Nazmul-Hossain AN, Pollard RP, Kroese FG, Vissink A, Kallenberg CG, Spijkervet FK, Bootsma H, Michie SA, Gorr SU, Peck AB, **Cai C**, Zhou H, Wong DT. Systems analysis of primary Sjögren's syndrome pathogenesis in salivary glands identifies shared pathways in human and a mouse model. Arthritis Research & Therapy. 2012 Nov 1; 14(6): R238

# CHAPTER 1: IS HUMAN BLOOD A GOOD SURROGATE FOR BRAIN TISSUE IN TRANSCRIPTIONAL STUDIES?

## Background

There is no clear consensus regarding the use of blood-based gene expression data for addressing neurological and neuroscientific research questions. On the one hand, gene expression levels in whole blood are only weakly correlated with those in brain tissue [1, 2]. On the other hand, blood gene expression profiles have been used to study neuropsychiatric diseases such as bipolar disorder and schizophrenia [3-6], as well as neurological diseases such as Amyotrophic Lateral Sclerosis [7], Huntington's disease [8] and Alzheimer's disease [9]. There are at least two major reasons why the relationship between human brain and human blood expression profiles remains poorly understood. The first reason concerns data quality and quantity: it is notoriously difficult to measure human brain tissue expression levels because of potential biases from post-mortem effects and relatively low sample sizes. The second reason is that most previous studies have focused on the preservation of mean expression levels, as opposed to the preservation of co-expression relationships. Given that the human brain transcriptome is organized into biologically meaningful co-expression modules [10], it is important to study the preservation of this organization in blood.

Because human brain expression data is derived from post-mortem brain tissue, special attention must be paid to RNA quality, post-mortem interval, and pH. To minimize the influence of these factors, we used highly reproducible and validated brain gene expression data sets from a recent meta-analysis of publicly available brain expression data [10]. Data set 1 (referred to as CTX)

consisted of 67 control samples from 67 individuals representing four cortical areas [11-13]. Data set 2 (referred to as CN) consisted of 27 control samples from 27 individuals taken from the head of the caudate nucleus [14]. Data set 3 (referred to as CB) consisted of 24 control samples from 24 individuals taken from cerebellar hemisphere [11].

By applying weighted gene co-expression network analysis (WGCNA) [15-17] to these data sets, Oldham et al. (2008) identified 19 cortex (CTX) modules, 23 caudate nucleus (CN) modules, and 22 cerebellum (CB) modules. These modules were defined as branches of a hierarchical clustering tree and were labeled by different colors. Many modules were highly preserved across the three brain regions, which was why they received the same color label. For example, 45% ($p = 2.8 \times 10^{-53}$) of genes overlapped between the yellow cortex module (labeled yellow/CTX) and yellow caudate nucleus module (labeled yellow/CN). Similarly, 46% ($p = 1.1 \times 10^{-54}$) of genes overlapped between the blue/CTX and the blue/CN modules [10]. By considering cell type-specific markers, several brain modules were found to contain genes that are preferentially expressed in oligodendrocytes, astrocytes or neurons [10].

Here we report the results of a comprehensive statistical analysis by cross-referencing the brain expression data with two large blood data sets (comprising a total of 1463 individuals). While previous studies have focused on the preservation of individual gene expression levels across the two tissues, we also investigated the preservation of gene co-expression modules. Since oligodenrocytes, astrocytes, and neurons are not present in blood, we were not surprised that only a handful of human brain modules showed evidence of preservation in human blood. Furthermore, we determined that these preserved modules could be combined into a single large module in blood. We also found that preserved intramodular hub genes tended to have heritable blood expression levels and were highly correlated with a small set of cluster of differentiation

(CD) genes.

## Results

### Blood gene expression data

We used whole blood gene expression data from healthy controls in a Dutch data set (n = 405) and published lymphocyte gene expression data (n = 1240), herein referred to as the San Antonio Family Heart Study (SAFHS) data set [18]. The Dutch data set originally consisted of 405 peripheral blood samples from healthy individuals (50.4% men and 49.6% women, mean age 56.4 and range from 19-88). This data set was analyzed with Illumina Human HT-12 microarrays. The SAFHS data set originally consisted of 1240 lymphocytes samples obtained from 1240 individuals (40.8% men and 59.2% women, mean age 39.3 and range from 15-94). This data set was analyzed with Illumina WG-6 microarrays. Using hierarchical clustering with inter-array correlations as a distance measure, we identified potential outlying arrays in an unbiased fashion. Since outlying arrays showed relatively low correlations with the other arrays (across the genes), they were deemed suspicious. To err on the side of caution, we removed these suspicious arrays from the analysis. Potential batch effects (due to different hybridization dates) were also removed using ComBat [19]. These are the same data pre-processing steps that Oldham et al. (2008) used in the brain data analysis. More sample pre-processing information can be found in Additional file 1-1.

After these pre-processing steps, 380 samples remained in the Dutch data set and 1084 samples remained in the SAFHS data set. Multiple probes corresponding to one gene (symbol) were

combined into one measurement. Next, we merged the Affymetrix (brain) data with the Illumina (blood) data by gene symbol, which resulted in 8799 genes in each data set.

**Preservation of mean expression levels and connectivity**

We first studied the preservation of mean gene expression levels of the 8799 genes between brain and blood. The pairwise scatterplots in Additional file 1-2 related mean expression values in the three brain regions to mean expression values in the two blood data sets. We found significant but weak correlations (*r* range: [0.24,0.32]) between mean expression in brain and mean expression in blood.

Next we investigated the extent to which co-expression patterns were preserved between brain and blood. For each gene, the network connectivity (also known as degree) is defined as the sum of its connection strengths with all other genes in the network. Thus, connectivity measures how correlated a gene is with all other genes (see Methods). Genes with high connectivity are informally referred to as "hub" genes. Overall, we found that gene connectivity was even less preserved (*r* range: [0.021, 0.11], Additional file 1-3) in blood than mean expression levels. These results show that global co-expression relationships are poorly preserved between brain and blood. However, Additional file 1-3 also shows some genes with high connectivity in both data sets. These genes may be part of sets of genes (co-expression modules) that are preserved between the two tissues. A more focused analysis that considered individual modules did reveal some evidence of preservation between the two tissues, as described below.

**Preservation of brain co-expression modules in blood**

Oldham et al. (2008) applied rigorous gene filters to the brain data set to ensure that transcripts were present and had high connectivity in the brain data (see the Supplemental Information of Oldham et al. 2008). These filters reduced the number of probe sets in each network to 5549 (CTX), 4050(CN), and 4029 (CB). After combining probes into single measures for each gene symbol and merging the data with the blood data sets, the CTX network contained 2640 genes, CN network contained 2063 genes and the CB network contained 2001 genes.

To determine whether a module found in a reference data set (e.g. human cortex) can also be found in a test data set (e.g. the Dutch blood data set), we used a powerful module preservation statistic approach implemented in the R software function modulePreservation [16] (desribed in Methods). This permutation test procedure evaluates whether module genes show significant evidence of network connectivity preservation in the test data. This module preservation test results in a statistic (referred to as preservation Z statistic or Zsummary statistic) for each module. The higher the preservation Z statistic is for a given brain module, the stronger the evidence that the brain module is preserved in a given blood data set. Under the null hypothesis of no module preservation, the preservation Z statistic follows an approximately standard normal distribution. Comprehensive simulation studies led to the following thresholds: a module shows no evidence of preservation if its Z statistic is smaller than 2; a Z statistic larger than 5 (or 10) indicates moderate (strong) module preservation.

We started out by evaluating the preservation of CTX modules in the Dutch and SAFHS blood data sets. The horizontal barplots in Fig. 1-1a show that the preservation Z statistics of the blue, yellow, and green CTX modules were above the threshold of 10 in both blood data sets, i.e. these

modules showed strong evidence of preservation. Similarly, Fig. 1-1b presents the module preservation results for the CN modules identified by Oldham et al. (2008). Only the yellow CN module was strongly preserved in both blood data sets. Fig. 1-1c shows that only the blue CB module was strongly preserved in both blood data sets. In total, we find that five brain modules were strongly preserved in human blood. More details and numeric values are presented in Additional file 1-4.

The preserved modules tended to be relatively large: Out of 2640 CTX network genes (from merging the CTX data with the blood data), 690 were part of the blue module, 421 were part of the green module and 658 were part of the yellow module. The preserved (yellow) CN module contained 254 genes out of 2063 CN network genes. The preserved CB (blue) module contained 819 out of 2001 CB network genes. Thus, 67% of genes in the cortex network, 12% of genes in the caudate nucleus network, and 41% of genes in the cerebellum network were part of a preserved module.

One can also visualize the evidence of module preservation by clustering the genes in the blood data sets. Since the brain modules were defined as branches of a hierarchical clustering tree (dendrogram), we used the identical approach to define modules in the blood gene expression data. Additional file 1-5 shows dendrograms of the blood gene expression data. As described in the Methods section, blood modules were defined as branches of the dendrogram [16, 17]. The first color-band underneath each dendrogram encodes blood module colors. The remaining color bands display the overlap with the preserved modules from each respective brain region. Visual inspection of these dendrograms revealed that genes from the preserved modules (based on the permutation test) tended to cluster together in the blood data. The fact that some colors were not contiguous shows that the preservation is not perfect. Below, we define measures of module

membership to identify the subsets of genes inside each of the five preserved modules that showed the strongest evidence of preservation.

**Relationships among preserved modules in blood**

While the brain modules were clearly distinct in the brain data sets, their preserved counterparts no longer appeared distinct in the blood data sets. To explore the relationships among preserved modules in blood, we summarized module expression profiles by forming the first principal component, which is referred to as the module eigengene (ME) [20]. For example, ME(blue/CTX) denotes the module eigengene of the blue cortex module. The ME can be considered a weighted average of the gene expression profiles in a module. If the ME of one module is highly correlated with that of another module in the blood data, then the genes inside the two modules have similar blood expression patterns, i.e. the two modules cannot be distinguished.

For the Dutch data set and the SAFHS data set, Fig. 1-2 shows that ME(blue/CTX), ME(blue/CB), ME(yellow/CTX), and ME(yellow/CN) had highly significant positive correlations ($r >= 0.95$, $p <= 10^{-40}$) with each other, but highly significant negative correlations ($r <= -0.95$, $p <= 10^{-40}$) with ME(green/CTX). This result indicates that the five preserved brain modules can hardly be distinguished in an unsigned gene coexpression network in blood, as they coalesce into one large preserved module. It is natural to ask whether these five modules were distinct in the original brain data sets. Additional file 1-6 shows that the three preserved CTX modules (blue/CTX, green/CTX, and yellow/CTX) were only moderately correlated in the CTX data: the correlation between ME(blue/CTX) and ME(yellow/CTX) was 0.52; the correlation

between ME(blue/CTX) and ME(green/CTX) was -0.09; the correlation between ME(yellow/CTX) and ME(green/CTX) was -0.69. The brain data did not allow us to correlate MEs of different brain regions, since the data consisted of samples from different individuals.

**Preservation of module membership between brain and blood**

We defined a measure of module membership (MM) by correlating the ME with each gene expression profile [21]. For example, MMbluei = cor (xi, MEblue) measures how correlated the expression profile of the i-th gene is with the blue ME. If MMbluei is close to 0, the i-th gene is not part of the blue module. But if MMbluei is close to 1 (or -1), it is highly positively (or negatively) correlated with the blue module genes. The module membership measure is highly related to intramodular connectivity [21]; thus, a gene with high absolute value MMbluei turns out to be a highly connected hub gene inside the blue module.

For each of the five preserved modules, we defined module membership measures in the respective brain data set and the two blood data sets (Additional file 1-7, 1-8, and 1-9). Fig. 1-3 shows that MM measures were highly correlated between the two blood data sets, indicating that the MM measure can be robustly defined in blood.

The extremely significant correlation test p-values in the scatterplots reflect the large sample size, i.e. numbers of genes. It may be more meaningful to consider the correlation coefficient value, e.g. a correlation value of 0.76 indicates a strong (but not perfect) linear relationship. We combined the MM measures for the Dutch and SAFHS data to arrive at a summary measure for human blood, which was referred to as "Blood MM measure". Additional file 1-10 reports the

correlations between the summary blood MM measure and the two individual blood MM measures.

Fig. 1-4 shows that MM values for the three preserved CTX modules (yellow/CTX, green/CTX, and blue/CTX) were highly correlated in the blood data sets, which reflects what we already know from our eigengene-based analysis (Fig. 1-2): these modules are indistinguishable in blood. Specifically, MM of yellow/CTX was positively correlated with MM of blue/CTX ($|r| = 1$, $p < 10^{-200}$, Fig. 1-4a), while MM of green/CTX was negatively correlated with MM of both yellow/CTX and blue/CTX (Fig. 1-4b-c). Given the exceptionally high correlations between the individual MM measures, it made sense to combine them by forming a weighted average, which flipped the sign of the negatively related green CTX module. We refer to the weighted average MM (across the modules) as the "combined MM measure". Additional file 1-11 shows that the combined MM value was highly correlated with the original MM value from the three modules.

Although the three modules were distinct in the cortex data, their MM measures also showed high correlations in cortex (Fig. 1-4d-f), which allowed us to define a combined MM measure for the CTX data. The combined cortex MM measure was significantly correlated ($r = 0.69$, $p < 10^{-200}$, Fig. 1-5a) with the combined blood MM measure. At the same time, the CN MM measure and the CB MM measure also showed significant correlations with the blood MM measure ($r = 0.45$, $p < 2.9\times10^{-107}$, Fig. 1-5b; $r = 0.28$, $p < 7.6\times10^{-38}$, Fig. 1-5c). These results support the original finding that the five co-expression modules (blue/CTX, green/CTX, yellow/CTX, yellow/CN and blue/CB) exhibit significant preservation in blood.

**Definition of preserved intramodular hub genes**

We refer to genes with high module membership in a preserved module as a preserved intramodular hub gene. Preserved hub genes show highly significant evidence of being centrally located inside a preserved module. Specifically, we defined preserved CTX module hub genes as having consistently high positive or negative combined MM in both cortex and blood. Toward this end, we thresholded the combined MM measures in both blood and cortex at a value of +0.35 and -0.35 (corresponding to a correlation test $p$-value < $5x10^{-13}$ in blood). These thresholds resulted in 357 preserved CTX hub genes, which are colored in red in Fig. 1-5a. For the preserved yellow CN module and preserved blue CB module, we found 305 preserved CN hub genes (colored yellow in Fig. 1-5b) and 277 preserved CB hub genes (colored blue in Fig. 1-5c) using the same threshold.

In summary, only 357 genes (13.5%) from the CTX network, 305 genes (14.8%) from the CN network and 277 genes (13.8%) from the CB network are preserved intramodular hub genes. These preserved intramodular hub genes exhibited the following overlap: the sets of preserved CTX (357) genes and preserved CN (305) genes shared 123 genes (Fisher's exact $p$-value < $2.2x10^{-16}$). The sets of preserved CTX (357) genes and preserved CB (277) genes shared 109 genes (Fisher's exact $p$-value < $2.2x10^{-16}$). The sets of preserved CN (305) genes and preserved CB (277) genes shared 64 genes (Fisher's exact $p$-value < $1.8x10^{-15}$). All three sets of preserved intramodular hub genes shared 36 genes. The names of these preserved hub genes and their MM values can be found in Additional file 1-12. The biological role of the 36 genes is discussed below.

The union of the three sets of preserved intramodular hub genes contains 678 genes. A functional enrichment analysis of the 678 genes reveals that some of these preserved hub genes play a role in infectious disease and infection mechanism ($p = 8.6\times10^{-10}$), post-translational modification ($p$

= $2.4\times10^{-8}$), and RNA post-transcriptional modification ($p$ = $2.9\times10^{-8}$) (Fig. 1-6). A more detailed functional enrichment analysis for each set of preserved CTX, CN, and CB module genes can be found in Additional file 1-13.

**Preserved intramodular hub genes have more heritable expression levels**

In the original publication of the SAFHS data, the authors calculated the heritability of each gene expression level and created a heritability table [18]. The gene expression heritability measures the proportion of expression trait variance attributable to genetic variance. These data allowed us to test whether preserved intramodular hub genes have more highly heritable expression levels than non-preserved intramodular hub genes. The red, yellow and blue bars in Fig. 1-5d-f show the mean heritability (y-axis) for the preserved CTX, CN and CB hub genes, respectively. To facilitate a comparison, we also report the mean heritability for all genes in heritability table (from Goring et al. 2008, grey bars) and for all genes in the merged blood and brain data set (black bars).

Fig. 1-5d shows that that the preserved CTX hub genes (n=357, red bar) have a significantly (analysis of variance test p = $10^{-108}$) higher mean heritability (32%) than all genes in heritability table (n=18525, mean heritability: 23%) and all genes in the CTX network (n=2640, mean heritability: 29%). Analogous results were observed for the CN data set ($p$ = 8.7 x $10^{-92}$, Fig. 1-5e) and the CB data set ($p$ = 8.1 x $10^{-93}$, Fig. 1-5f). These differences in heritability did not reflect differences in mean expression levels, as can be seen from Fig. 1-5g-i, which report mean blood expression values (y-axis) across the different groups of genes. While preserved intramodular hub genes and brain network genes had significantly higher mean expression values

than all genes in the heritability table (grey bar), preserved intramodular hub genes did not have higher mean expression levels than brain network genes (black bars).

**Relationships between preserved modules and cluster of differentiation genes**

We also investigated the relationships between the preserved modules and a special class of cell surface markers: cluster of differentiation (CD) genes, which are routinely used to characterize blood cell types. If a module is enriched with cell type-specific genes, then its module eigengene should have a strong correlation with the expression values of CD genes that are specific to that cell type. A high positive correlation would therefore suggest that a particular cell type might be related to the module. We found that the MEs of the five preserved modules had highly significant ($p < 10^{-40}$) positive correlations with the following CD genes: *CD58*, *CD47*, *CD48*, *CD53* and *CD164*. Statistical details for the individual modules are presented in Additional file 1-14.

In the following, we briefly describe what is known about the products of these CD genes while Additional file 1-15 presents more detailed gene information (adapted from http://www.genecards.org and http://pathologyoutlines.com).

*CD58* (present on Antigen Presenting Cells) is known to be a ligand of the T lymphocyte CD2 protein, and functions in adhesion and activation of T lymphocytes.

*CD47* (present on leukocyte, neuroblast, glial cell and other cells) is a membrane protein, which is involved in the increase in intracellular calcium concentration that occurs upon cell adhesion to extracellular matrix.

*CD48* (present on lymphocyte and other cells) is an activation-associated cell surface glycoprotein, and involved in facilitating interaction between activated lymphocytes.

*CD53* (present on leukocyte, glial cell and other cells) is cell surface glycoprotein and involved in the regulation of development, activation, growth and motility.

*CD164* (present on leukocyte, glial cell and other cells) is a type I integral transmembrane sialomucin that functions as an adhesion receptor. It is involved in hematopoiesis, migration of umbilical cord blood, prostate cancer metastasis, infiltration of bone marrow, myogenesis and myoblast migration.

**Module preservation between different brain regions**

As mentioned in the introduction, many brain modules were found to be highly preserved across the three brain regions, which is why they received the same color label. Here we use a more powerful approach for measuring module preservation (based on the modulePreservation R function) than then one used in the original analysis by Oldham et al. Therefore, we use the modulePreservation function to re-analyze brain module preservation across brain regions. For example, we evaluate which CTX brain modules are preserved in the CN and CB data. Detailed results of this analysis can be found in Additional file 1-16. For CTX brain modules, we find that 11 out of 19 CTX module show at least moderate evidence of preservation (Preservation Z statistic > 5) in both CN and CB data. For CN brain modules, we find that 12 out of 23 CN modules also show at least moderate evidence of preservation (Preservation Z statistic > 5) in both CTX and CB data. For CB brain modules, we find that 12 out of 22 CB modules show at least moderate evidence of preservation (Preservation Z statistic > 5) in both CTX and CN data.

In summary, 55% modules showed preservation cross the different brain regions. These results are congruent with those presented in the original analysis by Oldham et al. It is particularly interesting to study which of our 5 preserved blood/brain modules are preserved in other brain regions.

For the 3 preserved CTX/blood modules (blue, green and yellow), we find that all 3 of them showed very high evidence of preservation in both CN (Preservation Z statistic >= 16.6) and CB data (Preservation Z statistic >= 8.7).

For the preserved (yellow) CN/blood module, we find very high evidence of preservation in CTX data (Preservation Z statistic = 19.1) , but only moderate/weak evidence preservation in CB data (Preservation Z statistic = 3.8).

For the preserved (blue) CB/blood module, we find very high evidence of preservation in both CN (Preservation Z statistic = 20.8) and CB data (Preservation Z statistic > 16.0). Further, details can be found in Additional File 1-16.

Overall, we find strong evidence that the preserved brain/blood modules are also preserved in multiple brain regions.

**Discussion**

Few studies are able to access human neural tissue for studying diseases [22]. Given the difficulty of procuring human brain tissue versus the relative ease of measuring blood expression levels, a question of great practical importance is to determine to what extent blood is a reasonable surrogate for brain in gene expression studies. Here we relate highly reproducible

brain expression data from a recent meta-analysis of human brain data sets to two large blood data sets. Overall, we find that mean expression levels are weakly preserved between three brain regions and blood ($r$ range [0.24, 0.32]). Since gene expression profiles in human brain regions are organized into highly reproducible co-expression modules [10], it is important to determine which of these modules show evidence of preservation in blood. Only three out of 19 cortex modules, one out of 23 caudate nucleus modules and one out of 22 cerebellum modules show strong evidence of preservation. In blood, these five modules exhibit very similar expression patterns as can be seen from the very high absolute correlations ($|r| > 0.96$) between their respective eigengenes (Fig. 1-2).

Although few modules were preserved, they tended to be relatively large. 67% of genes in the cortex network were part of one of the three preserved modules; 41% of genes in the cerebellum network and 12% of the caudate nucleus network genes were part of their respective preserved modules. Intramodular hub genes inside preserved modules are centrally located in both modules. The number of intramodular hubs depends on the threshold used for the module membership measures in brain and blood. 13.5% (357) of genes in the cortex network, 14.8% (305) of genes in the caudate nucleus network, and 13.8% (277) of genes in the cerebellum network were defined as preserved intramodular hub genes. Using our posted data and R software code, the reader can change the thresholds used for defining these hub genes. Our biological characterization of preserved intramodular hub genes is highly robust with respect to the chosen threshold values.

In mice, mean expression levels of heritable genes have been found to be highly correlated between mouse hippocampus and spleen [23]. We do not find that heritable genes exhibit highly

correlated mean expression levels between brain and blood (Additional file 1-17). However, we find that the preserved intramodular hub genes tend to be more heritable (Fig. 1-5).

The preserved CTX blue, green, and yellow modules were found to be enriched with neuronal markers, glutamatergic synapse genes, and metabolism-related genes, respectively. The preserved CN yellow module was also found to be enriched with metabolism-related genes, while the preserved CB blue module was enriched with neuronal markers and genes encoding synaptic proteins [10]. In blood, studying the enrichment with regard to brain cell type markers is not meaningful. However, one can classify blood cell types using human clusters of differentiation (CD) genes. Interestingly, the following CD molecules consistently have significant positive correlation with genes inside the preserved modules: *CD58*, *CD47*, *CD48*, *CD53* and *CD164*.

A functional enrichment analysis of brain module preservation reveals basic functional pathways preserved between the two tissues. Fig. 1-6 shows that these preserved intramodular hub genes are significantly enriched for genes that play a role in infectious disease and infection mechanism, post-translational modification and RNA post-transcriptional modification. Other categories include Cell Death, Energy Production, Nucleic Acid Metabolism, Molecular Transport and Protein Trafficking (Fig. 1-6). The 36 intramodular hub genes that were preserved in all three sets exhibit several common functional themes. First, nearly 20% of these genes, including *ASF1A*, *ATF2*, *DR1*, *HCFC1R1*, *HMGN4*, *MBD3*, and *RAD21*, are known to play roles in modifying chromatin structure. Some of these modifications have been shown to induce transcription (e.g. *ATF2*, *DR1*, *HMGN4*), while others produce repressive effects (e.g. *MBD3*). A number of other genes in the group of 36 encode signalling proteins that are thought to play roles in a wide variety of cellular processes, including *ARPP-19*, *CSNK1G3*, *MAP4K5*, *PPP1CB*,

and *YWHAQ*. A third category of genes relates to protein trafficking and includes *RAB1A*, *SNX2*, *SNX3*, while a fourth category consists of genes involved in mitochondrial function, including *DLAT*, *SUCLA2*, and *YME1L1*. Some of the proteins encoded by these 36 genes may physically interact, such as *ATP6AP2*, which associates with the transmembrane sector of vacuolar ATPases (proton pumps), and *ATP6V1C1*, which is a subunit of the vacuolar ATPase protein complex. Intriguingly, for a number of other genes in this group, biological functions remained to be elucidated (e.g. *FAM3C*, *FLJ20254*, *LANCL1*, *PRNP*, *RABGGTB*, and *WRB*). We note that many of these 36 preserved intramodular hub genes are expressed ubiquitously. Therefore, it is possible, perhaps even probable, that these genes are also co-expressed in other tissue types beyond brain and blood. Their co-expression may therefore help serve to maintain differentiated cells in a particular state (e.g. chromatin modifying genes) in response to a particular environment (e.g. signalling genes), as well as enable other shared, basic cellular processes (e.g. protein trafficking, energy metabolism).

Our study has several strenghts including the use of multiple large data sets, carefully validated brain co-expression modules from Oldham et al 2008, and a powerful statistical approach for evaluating module preservation.

But our study also has several limitations including the following. First, the brain expression data were measured using the Affymetrix platform, while the blood expression data were measured using the Illumina platform. Since platform differences bias our results towards the null hypothesis of no preservation, we can be confident about preservation, but less confident about lack of preservation. The weak correlations between mean expression profiles may reflect platform differences. A second limitation is that we studied the preservation of brain modules in blood (and not vice versa). Our goal was to determine the preservation of robustly defined and

well annotated brain modules. Defining blood modules and studying their preservation in brain tissue is beyond the scope of this article. A third limitation is the relatively small set of genes considered for the co-expression module preservation study. Oldham et al. had applied stringent filtering criteria to construct the brain network, which greatly reduced the number of probes considered in that study. After combining probes by gene symbol and merging the brain and blood data, the co-expression module preservation study focused on 2604 CTX, 2001 CB, and 2063 CN network genes. We focused on this relatively small set of genes since their connectivity pattern in brain was found to be highly reproducible across array platforms and independent data sets (Oldham et al 2008). But we should point out that our study of mean expression preservation involved 8799 genes. A fourth limitation is that we only use correlation network methodology. Many alternative co-expression network methods have been proposed in the literature [22, 24-26]. We focus on WGCNA since this method was used in Oldham et al (2008). An exploration of alternative procedures is beyond our scope but we encourage the reader to apply their method to our posted data.

## Conclusions

In summary, we find that transcriptome organization is poorly preserved between brain and blood and only a handful of large brain co-expression modules that exhibit strong overall evidence of preservation in blood. However, these modules are not preserved whole cloth. Instead, only certain aspects of these modules (i.e. subsets of genes appear to be involved in basic cellular processes, such as metabolism) exhibit strong preservation of gene co-expression relationships. The subset of preserved co-expression relationships characterized here may aid

future efforts to identify blood biomarkers for neurological and neuropsychiatric diseases when brain tissue samples are unavailable.

## Methods

### Weighted gene co-expression network analysis and preservation visualization

The statistical analysis software (WGCNA R package) and R tutorials for constructing a weighted gene co-expression network can be found in [16]. The WGCNA package first calculates all pairwise Pearson correlations coefficients across all samples. In a weighted network, the resulting Pearson correlation matrix is transformed into a adjacency matrix ($aij = |cor(xi , xj )|^{\beta}$), which represent the pairwise connection strengths. The power β facilitates a soft-thresholding approach that preserves the continuous nature of the co-expression relationships [17]. As a power we chose the default value of 6. An advantage of weighted networks is that they are highly robust with regard to the choice of the soft threshold parameter value. As a network dissimilarity measure we used 1 - the topological overlap measure as input for average linkage hierarchical clustering. The topological overlap measure is a highly robust measure of interconnectedness [27]. We used the dynamic branch cutting method to define modules as branches of the hierarchical clustering tree [28]. Unassigned background genes, outside of each of the modules, were denoted with the color grey.

### Connectivity and module membership measures

Whole network connectivity for a certain gene is defined as the sum of its connection strengths

with all other genes in the network. Mathematically, it can be calculated easily as the sum of a given column in the adjacency matrix. Intramodular connectivity is defined as the sum of the connection strengths between a particular gene and all other genes in the same module. Module membership (MM), or eigengene-based connectivity (kME), is another measure of connectivity. It is defined as MMqi = cor(xi, MEq), where xi is the expression profile of i-th gene and MEq is the eigengene of q-th module. The larger the absolute values of MM, the greater the similarity between the i-th gene and the q-th ME. If the absolute value of MM is 0, it means that this gene is not part of the module. Although the MM measure is highly correlated with intramodular connectivity [21], the MM measure is preferred since it can be easily extended to genes outside the original module, and the statistical significance (*p*-value) of MM can be calculated for every gene with respect to every module.

**Correlation tests**

To measure the relationship between brain tissue connectivity and blood tissue connectivity (and for relating mean expressions), we used a robust estimator of the correlation (the biweight midcorrelation implemented in the WGCNA R package) to protect against outliers. Simulation studies show that the biweight midcorrelation is more robust than the Pearson correlation but often more powerful than the Spearman correlation.

**Correcting *p*-values for multiple comparison tests**

To protect against false positives due to multiple testing, we also report Bonferroni corrected *p*-

values. The Bonferroni correction method is the most conservative approach for correcting for multiple comparisons. The corrected *p*-value is defined by the product of the uncorrected *p*-value times the number of tests. Since we carried out 50 correlation tests in this article, a Bonferroni corrected *p*-value is defined by multiplying the uncorrected *p*-values by 50.

**Module Preservation analysis**

Our module preservation analysis is based on the modulePreservation R function implemented in the WGCNA R package. The modulePreservation function implements several powerful network based statistics for evaluating module preservation. These statistics are summarized into the composite preservation called Zsummary. For each module in the reference data (e.g. brain data) one observed a value Zsummary in the test data (e.g. a blood data set). An advantage of the preservation Z statistic is that it makes few assumptions regarding module definition and module properties. Traditional cross-tabulation based statistics are inferior for the purposes of our study. While cross-tabulation approaches are intuitive, they have several disadvantages. To begin with, they are only applicable if the module assignment in the test data results from applying a module detection procedure to the test data. Even when modules are defined using a module detection procedure, cross-tabulation based approaches face potential pitfalls. A module found in the reference data set will be deemed non-reproducible in the test data set if no matching module can be identified by the module detection approach in the test data set. Such non-preservation may be called the weak non-preservation: "the module cannot be found using the current parameter settings of the module detection procedure". On the other hand, one is often interested in strong non-preservation: "the module cannot be found irrespective of the parameter settings of the

module detection procedure". Strong non-preservation is difficult to establish using cross-tabulation approaches that rely on module assignment in the test data set. A second disadvantage of a cross-tabulation based approach is that it requires that for each reference module one finds a matching test module. This may be difficult when a reference module overlaps with several test modules or when the overlaps are small. A third disadvantage is that cross-tabulating module membership between two networks may miss that the fact that the patterns of connectivity between module nodes are highly preserved between the two networks.

The correlation network based statistics implemented in the modulePreservation function do not require the module assignment in the test network but require the user to input the gene expression data.

**Functional Enrichment Analysis**

The Ingenuity Pathways Analysis (Ingenuity® Systems, http://www.ingenuity.com) software was used to determine whether sets of genes (e.g. preserved intramodular hub genes) were significantly enriched with known gene ontologies. This software ranks the pathways by their Fisher exact test p-value of functional enrichment. We chose the default background gene list (here all all human genes) for the analysis. Ingenuity only reports uncorrected p-values. The gene lists published in our Additional files allow the reader to choose alternative backgrounds or software tools.

**Figure 1-1**

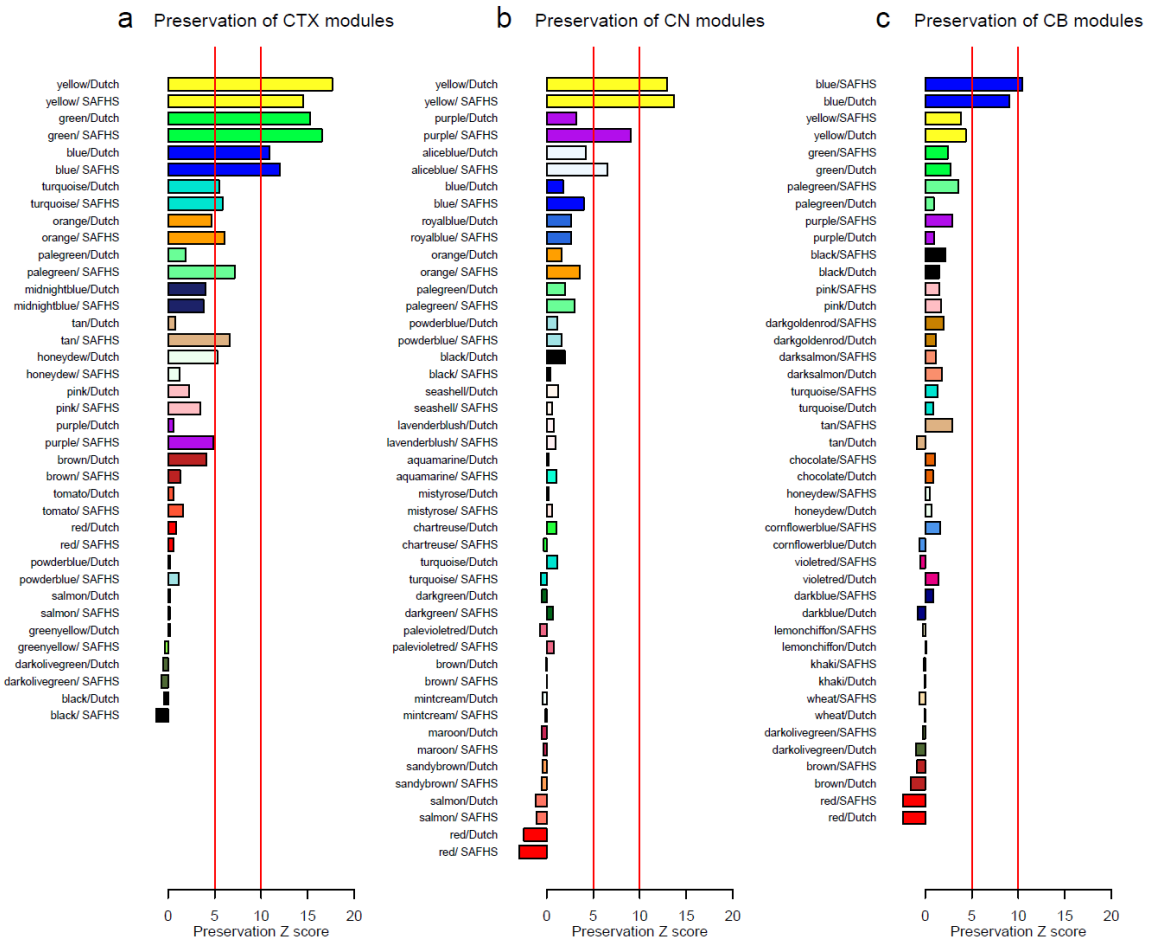


**Figure 1-1. Studying the brain module preservation in human blood**

The row bars correspond to brain co-expression modules found by Oldham et al. (2008). Modules are labeled by a color. For each module color, there are two horizontal bars which correspond to the module preservation Z statistics in the Dutch blood data and the SAFHS blood data, respectively. The two red vertical lines correspond to thresholds of moderate preservation (5) and strong preservation (10). Panel (a) shows that only three (yellow, green, and blue) out of 19 cortex modules showed strong preservation in both blood data sets. Panel (b) shows that only

one (yellow) out of 23 caudate nucleus modules was strongly preserved in both blood data sets. Panel (c) shows that only one (blue) out of 22 cerebellum modules was strongly preserved in both blood data sets. In summary, only five modules from Oldham et al. (2008) show strong evidence of preservation in human blood.

**Figure 1-2**

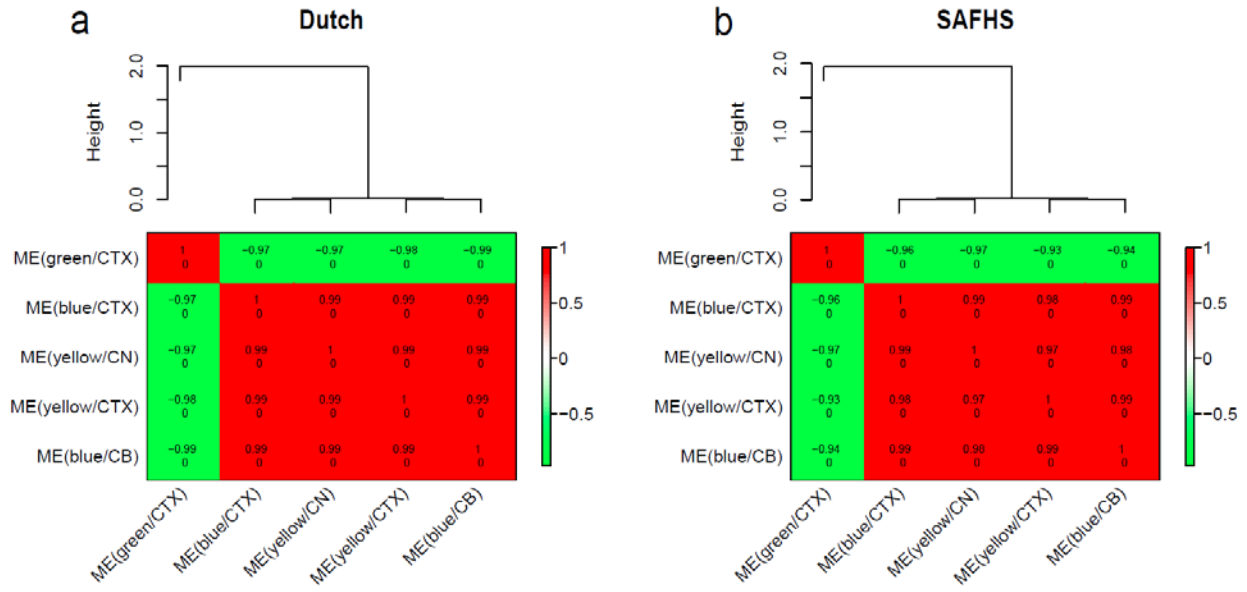


**Figure 1-2. Relationships between the five preserved modules in the two blood data sets**

The expression profiles of each preserved module were summarized by the respective module eigengene (defined as the first principal component). The correlations between the module eigengenes can be used to measure relationships between the modules (Langfelder and Horvath 2007). The hierarchical cluster tree shows the correlation relationships between the module eigengenes in the Dutch blood data (a) and the SAFHS blood data (b). The tables show the

pairwise correlation coefficients (upper number) between the eigengenes and the correlation test p-values (lower number). The colors of the table entries color code the values of the correlations (green and red correspond to negative and positive correlations). Note that the four modules ME(blue/CTX), ME(yellow/CN), ME(yellow/CTX), and ME(blue/CB) were highly positively correlated with each other but negatively correlated with ME(green/CTX).

**Figure 1-3**

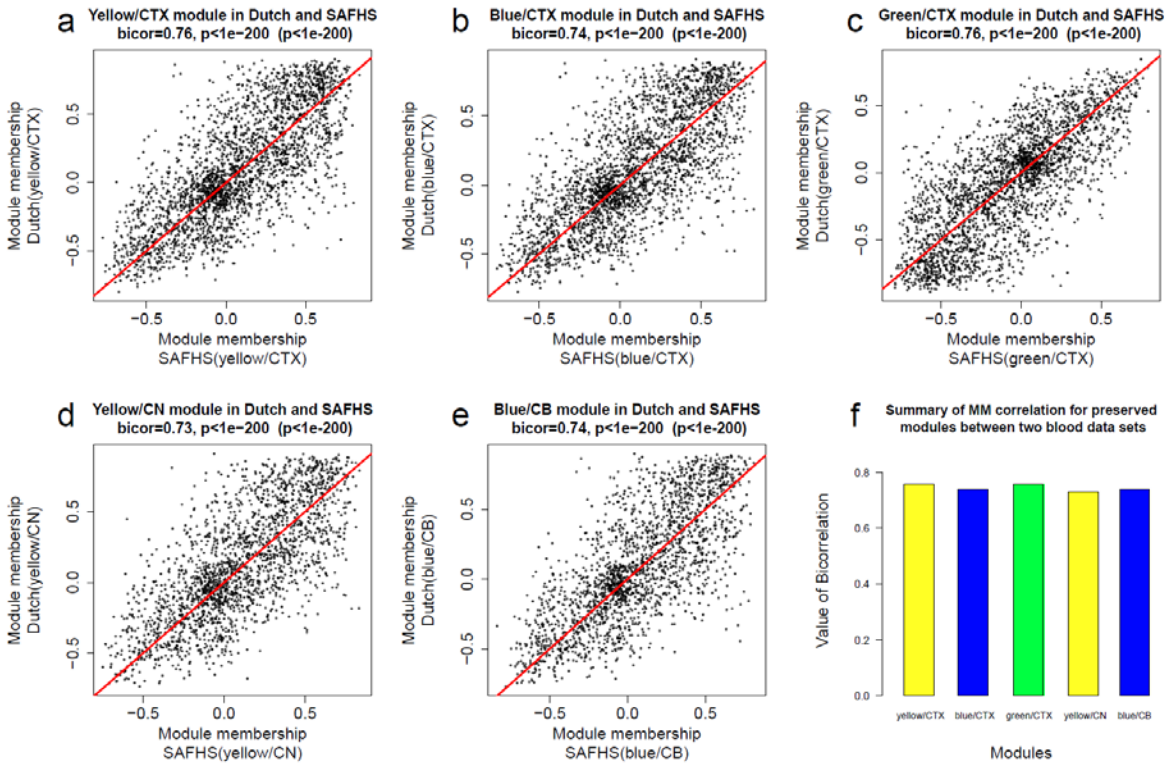


**Figure 1-3. Module membership measure of preserved modules is highly reproducible between the two blood data sets**

For each of the five preserved brain modules, the scatterplots show that module membership measure was reproducible between the SAFHS blood data (x-axis) and the Dutch blood data (y-axis). Each dot corresponds to a gene. The red diagonal line corresponds to y=x. Results are shown for the following preserved modules: (a) yellow/CTX, (b) blue/CTX, (c) green/CTX, (d) yellow/CN module, and (e) blue/CB. We report both uncorrected correlation test p-values and Bonferroni corrected p-value (inside of brackets). The extremely significant correlation test p-values reflect the large sample size (number of gene). It may be more meaningful to focus on the correlation coefficients. Overall, we find that the module membership measures are highly reproducible.

**Figure 1-4**

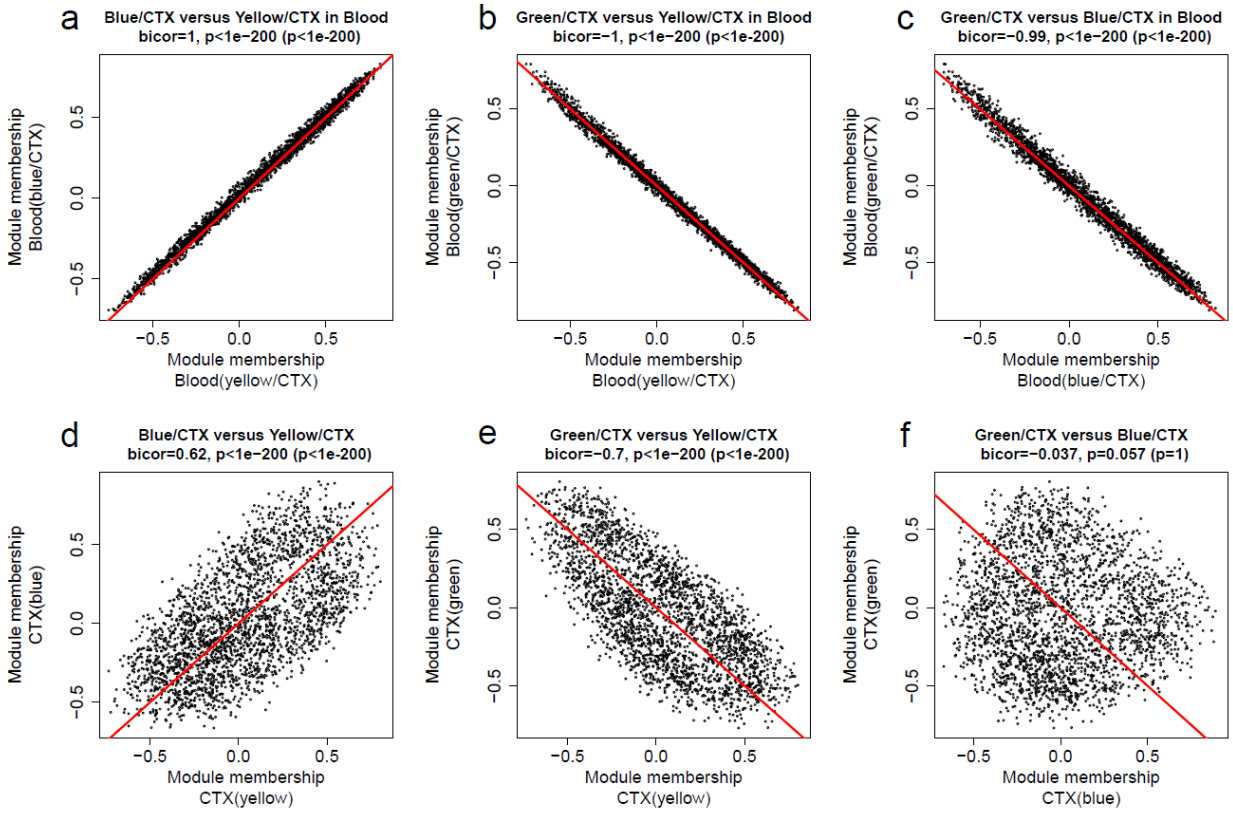

**Figure 1-4. Relationships between the module membership measures of the three preserved cortex modules**

Relationships between the module membership of blue/CTX, yellow/CTX and green/CTX modules in blood (a-c) and in the original cortex brain samples (d-f). Panel (a) shows that the correlation between MMyellow and MMblue equaled 1 in the blood data, which reflects the fact that these modules were indistinguishable in blood. In contrast, panel (d) shows that correlation between MMyellow and MMblue equaled 0.62 in cortex. Figure (b-c) shows that MMgreen had a correlation close to -1 with MMyellow (b) and MMblue (c) in blood. We report both uncorrected correlation test p-values and Bonferroni corrected p-value (inside of brackets).

Given the very high correlations between MMgreen, MMyellow, and MMblue in blood, we combined the three measures in an overall module membership measure referred to as MM.combined.Blood. Analogously, we combined the three MM measures for the cortex network (referred to as MM.combined.CTX).

**Figure 1-5**

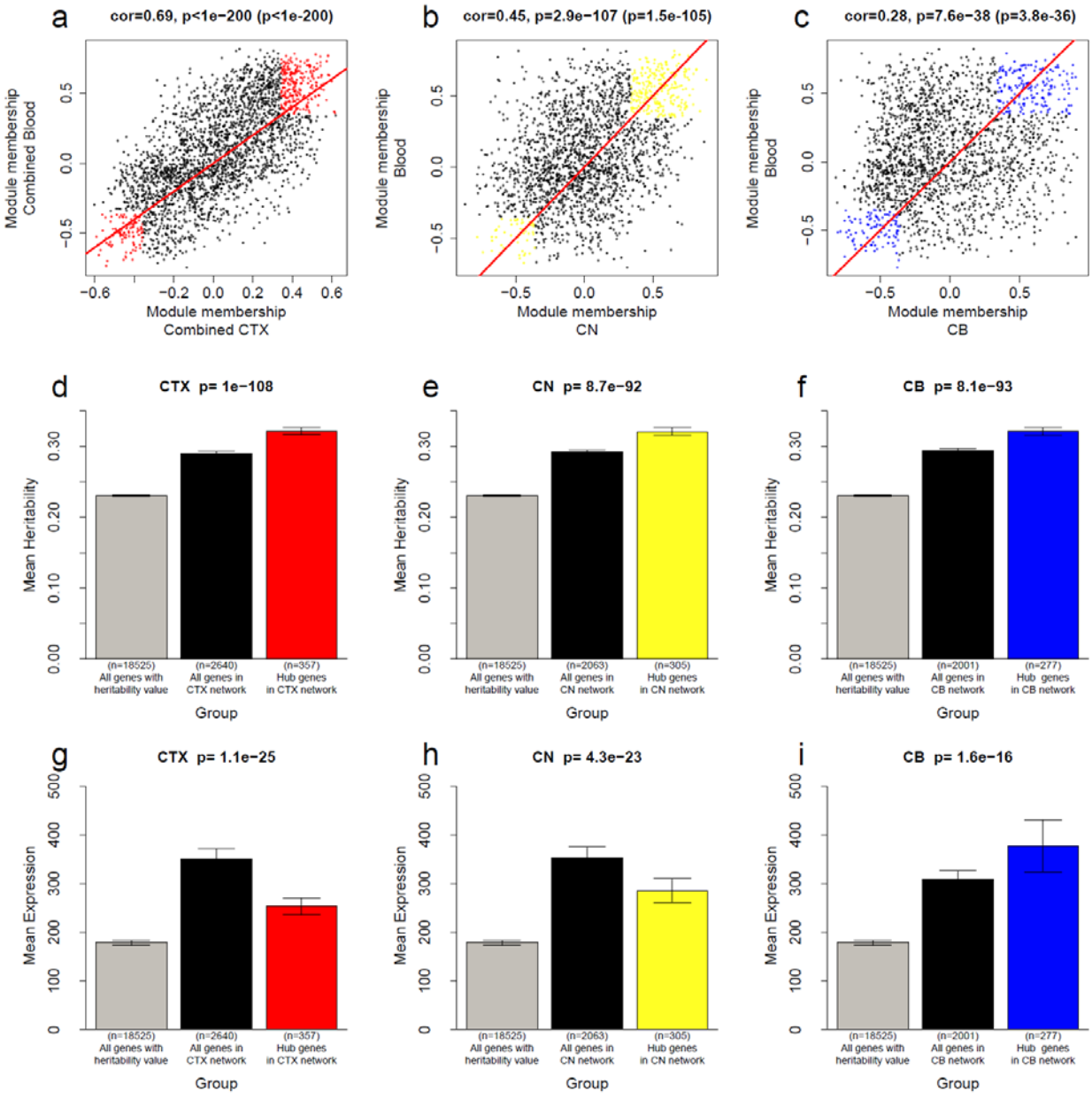


**Figure 1-5. Definition and characterization of preserved intramodular hub genes**

The scatterplots show how the combined measure of module membership in blood MM.combined.Blood (or MM.Blood, y-axis) related to the analogous measure in cortex (a),

caudate nucleus (b), and cerebellum (c). Preserved intramodular hub genes were defined as those genes whose combined MM measure in blood and brain tissue is larger than +0.35 (or smaller than -0.35), since these genes showed highly significant evidence of being part of the preserved modules in both tissues. For the CTX, CN, and CB networks, the roughly 300 preserved intramodular hub genes are colored in red (a), yellow (b), and blue (c). We report both uncorrected correlation test p-values and Bonferroni corrected p-value (inside of brackets). Barplots (d-f) show the mean heritability of the blood expression profiles (y-axis) across preserved intramodular hub genes (blue bars), across genes in the CTX (d), CN (e), and CB (f) networks (black bars), and across all genes on the blood array (grey bar). Note that the preserved intramodular hub genes have significantly higher mean heritability than non-preserved intramodular hub genes. Barplots (g-i) show the mean blood expression values (y-axis) across the same groups of genes. Note that the preserved intramodular hub genes and the brain network genes have significantly higher mean expression values in blood than all genes on the array (grey bar). However, preserved intramodular hub genes do not have higher mean expression levels than the (roughly 2500) genes that form the brain network (black bars).

**Figure 1-6**

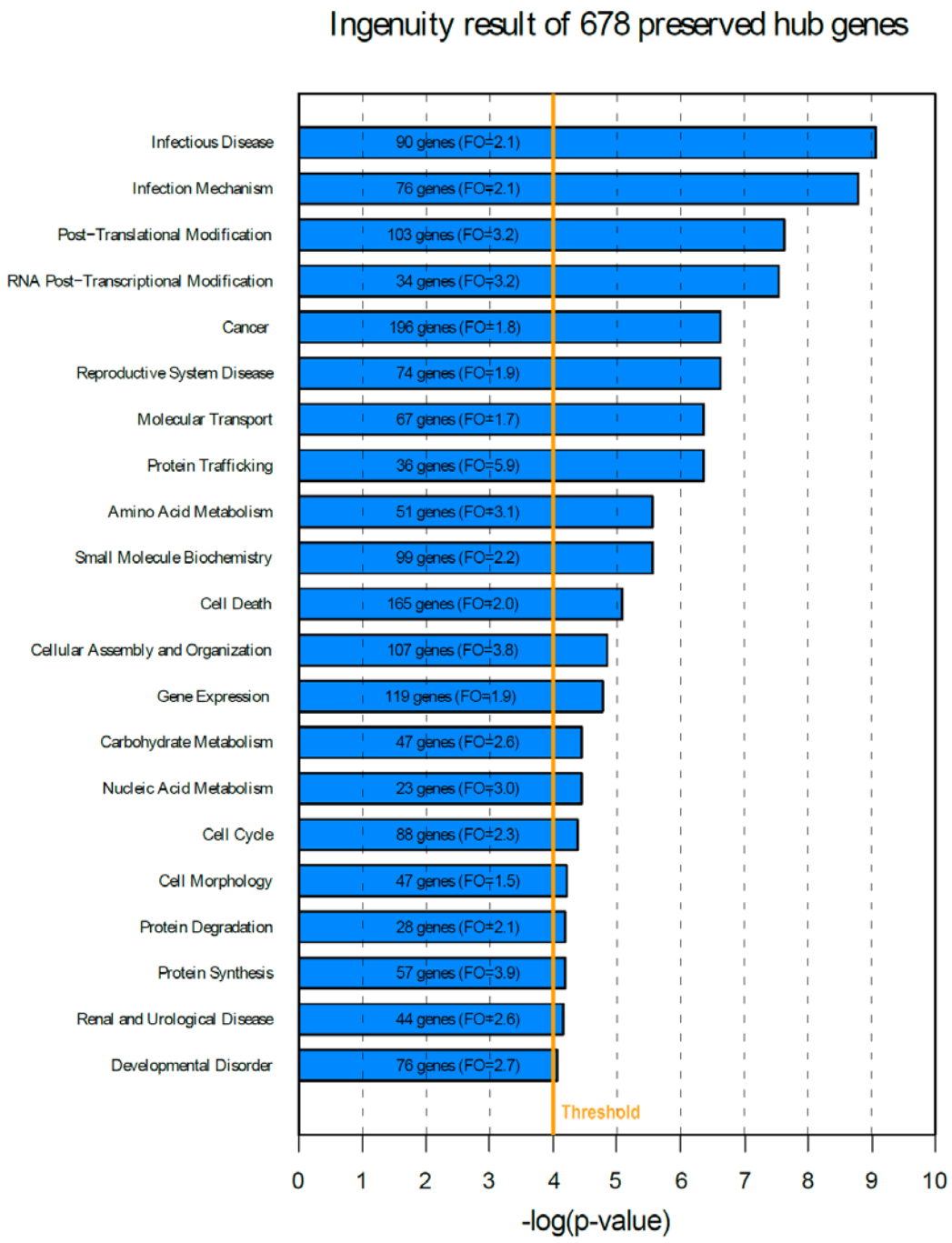
Ingenuity result of 678 preserved hub genes
Infectious Disease
Infection Mechanism
Post−Translational Modification
RNA Post−Transcriptional Modification
Cancer
Reproductive System Disease
Molecular Transport
Protein Trafficking
Amino Acid Metabolism
Small Molecule Biochemistry
Cell Death
Cellular Assembly and Organization
Gene Expression
Carbohydrate Metabolism
Nucleic Acid Metabolism
Cell Cycle
Cell Morphology
Protein Degradation
Protein Synthesis
Renal and Urological Disease
Developmental Disorder
90 genes (FO=2.1)
76 genes (FO=2.1)
103 genes (FO=3.2)
34 genes (FO=3.2)
196 genes (FO=1.8)
74 genes (FO=1.9)
67 genes (FO=1.7)
36 genes (FO=5.9)
51 genes (FO=3.1)
99 genes (FO=2.2)
165 genes (FO=2.0)
107 genes (FO=3.8)
119 genes (FO=1.9)
47 genes (FO=2.6)
23 genes (FO=3.0)
88 genes (FO=2.3)
47 genes (FO=1.5)
28 genes (FO=2.1)
57 genes (FO=3.9)
44 genes (FO=2.6)
76 genes (FO=2.7)
Threshold
0
1
2
3
4
5
6
7
8
9
10
-log(p-value)

**Figure 1-6. Ingenuity analysis result for 678 preserved hub genes**

Functional enrichment analysis of the preserved intramodular hub genes found in CTX, CN or CB data (selected in Figure 1-5) (678 genes). Only gene categories with significant enrichment p-values are presented. FO (inside brackets) denotes fold overrepresentation (defined as observed counts divided by expected counts under the null hypothesis). To calculate the Fisher's exact p-values and FOs, we use the Ingenuity default background (here all human genes).

**Additional files:**

The corresponding addtional files for this section can be downloaded from the following linkage:

http://labs.genetics.ucla.edu/horvath/htdocs/CoexpressionNetwork/Bloodbrain/Additional/

# CHAPTER 2: EMPIRICAL EVALUATION OF EXPRESSION DECONVOLUTION METHODS

## Background

Gene expression studies often assess the mRNA abundances in biological samples that are comprised of multiple cell populations, which often have distinct gene expression patterns [29, 30]. For example, whole blood expression levels may reflect expression levels from different cell populations such as T cells, B cells, monocytes, eosinophils or nature killer cells [30-34]. Since cell counts vary greatly across different individuals [35, 36], analysis methods that ignore the varying cell population composition may have reduced statistical power to detect disease related genes and pathways [34, 37, 38].

To address the challenge of sample heterogeneity due to cell types, one can use experimental methods for separating out expression patterns from distinct cell populations or computational methods (referred to as expression deconvolution methods). Commonly used experimental methods include fluorescent-activated cell sorting (FACS) [39-41] and laser capture micro (LCM)-dissection [42]. For example, FACS has been used to cleanly separate populations of neurons, astrocytes, and oligodendrocytes in the mouse brain based on gene expression patterns [43]. Similarly, selective cell sorting using a combination of fluorescent labeling and aspiration was performed to transcriptionally profile 12 distinct neuronal subpopulations in the adult mouse forebrain [44]. Furthermore, LCM has been effective at isolating small brain regions for analysis in human tissues, where genetic toll are not available. For example, it has been used to extract

RNA from small groups of pyramidal neurons in CA1 and CA3 in order to transcriptionally distinguish these two regions of the human hippocampus [42, 45].

Experimental methods are sometimes not applicable for isolating pure cell populations, e.g. they may not yield sufficient quantities of mRNA which is why computational approaches for "deconvoluting" expression data are very appealing. Venet et al. (2001) appear to be the first to propose a computational deconvolution method that can both estimate cell abundances and mean cell type-specific (CTSE) expression values [46] in mixed samples. Many subsequent articles describe related approaches. For example, Lu et al. (2003) computationally estimated the abundance of yeast cells in different stages of the cell cycle [47]. A similar approach was proposed by Bar-Joseph et al. (2004) [48]. We recently described a method which predicted the cell abundance based on cell markers [49], referred to as the cell marker (CM) method. Apart from estimating the abundances of different cell types, one can also aim to estimate the mean expression value of each gene in a "pure" cell type using expression data from mixed samples [34, 50, 51].

Although expression deconvolution methods have shown promise in individual applications, there is a need to provide a comprehensive and unbiased evaluation across a range of real data applications. Here we compare several types of deconvolution methods in four different gene expression data sets (involving blood and brain tissues). Using simulation studies we evaluate the effect of mis-specifying or omitting certain cell types from the analysis. Since cell type-specific estimation of the mean value for individual genes is sometimes problematic, we propose to consider sets of genes (as opposed to individual genes) and show that this can increase the accuracy of the CTSE estimation. We also present two R functions that implement these methods.

## Results

### Overview of expression deconvolution methods

There is a rich literature on expression deconvolution methods [31, 34, 46-48, 52-54] reviewed in Zhao et al. (2010) [36]. We distinguish four categories of deconvolution methods (described in Table 1). The first category aims to estimate cell abundances [47], e.g. it allows one to estimate the proportion of CD8 T cells in a blood sample. The second category estimates mean cell type-specific expression (CTSE) level of a gene in a given cell type based on known cell abundance measures [34]. For example, it would allow one to estimate the mean expression value of a gene in CD8 T cells based on whole blood gene expression data as long as cell count data are available. The third category estimates mean CTSE levels of individual genes based on estimated cell abundances, e.g. such a method was described in [51]. The fourth category estimates mean CTSE levels of sets of genes (as opposed to individual genes) based on estimated cell abundances.

### Estimating the cell type abundances (population proportions)

Here we evaluate 2 statistical methods for estimating the abundances of cell populations in a mixed sample (comprised of several distinct cell populations). Since the first method is based on a multivariate linear regression model (LRM), it is referred to as LRM abundance estimation method. As described in Methods, the method assumes that the expression value of a given gene in a mixed sample is a linear combination of its mean expression values in the pure cell types [46]. The abundance (proportion) of a given cell type is estimated based on its coefficient value in the LRM. We have implemented the LRM abundance method in the R function proportionsInAdmixture, which as indicated by the name, allows one to estimate the proportions

of populations (e.g. T cells, B cells) in an admixture (e.g. whole blood). The LRM method assumes that the mean expression values in the pure cell populations are known for a subset of genes. While this assumption is restrictive, the method is attractive since it allows one to directly estimate the proportions via the coefficient values.

The second method, referred to as cell marker (CM) abundance estimation method, cannot be used to estimate proportions directly. Instead, it provides an abundance estimate that will be correlated with the true proportion. This method is attractive since it makes fewer assumptions than the LRM method. It only requires that the user specify the cell marker genes, i.e. genes that are expected to be highly correlated with the true proportions. To find cell marker genes for a cell type, one can use genes that have a high expression value in this cell type but have low expressions value in other cell types that are part of the mixture. Once a set of cell markers has been identified, one can simply average their expression values to arrive at an abundance estimate (which is expected to have a linear relationship with the true abundance). While the average CM marker expression value is often highly correlated with the true cell abundance, it cannot be used to estimate the absolute number (or proportion) of the cell type without making further assumptions. In this paper, the fraction $f_q$ of population $q$ in a mixed sample is given by $f_q = x_q / \sum_1^p x_i$ (Equation 1), where xi denotes the mean expression value of the marker genes in population i. For example, if there are three cell populations in the admixed samples and the mean expression values of marker genes for these three cell populations are 100, 400, 500, then the fractions are estimated to be 100/(1000) = 0.1, 400/(1000) = 0.4, 500/(1000) = 0.5 respectively.

**Cell Type-Specific Expression (CTSE) methods**

If successful, CTSE deconvolution methods would allow one to estimate the mean value of a given gene in a given cell type based on expression values collected from a cell mixture. For example, these methods promise to allow one to calculate the mean value of a gene in CD8 T cells when only whole blood expression data are available.

Here we consider regression-based methods that assume that the abundances of cell types vary across the admixed samples [34, 51]. We distinguish approaches that assume cell abundances in each (admixed) sample are known from approaches that use estimated abundances (e.g. based on the LRM or CM based methods). Shen-Orr et al, describe a CTSE estimation method that assumes that the cell type abundances are known [34], e.g. complete blood count data often provide this information. Kuhn et al proposed an approach that does not require that the user specify cell proportions [51]. Instead, this approach estimates the abundances.

In this article, we address the following two questions about CTSE methods. i) How well do CTSE estimates predict the true observed mean expression values in pure cell types? ii) Do CTSE estimation methods based on estimated cell abundances perform as well as those that use true observed abundances? In other words, is it worth the trouble to collect cell count data on each individual?

**Measures of estimation accuracy: correlation and MSE**

To measure the performance of expression deconvolution methods we used two accuracy measures: i) the correlation, r, between true value and the predicted (estimated) value and ii) the mean square error, MSE, between the true value and its estimated value. Thus, a good estimation method will have a high correlation value and a low MSE. These measures capture different

properties of an estimation method. For example, the CM abundance estimate can be highly correlated with the true cell type abundance but its MSE will typically be high since it is poorly calibrated.

**Empirical evaluation of abundance estimation methods**

We applied the LRM and CM based abundance methods to three expression data sets with known cell type abundance measures, which served as gold standard to evaluate the estimation methods. The data are described in Methods. The transformed blood cell data set [31] and the rat tissue data set [34] were generated by experimentally mixing multiple cell populations in pre-defined ratios. We first applied the two abundance estimation methods to the transformed blood cell data set, which is comprised of four blood cell lines (interpreted here as 4 distinct cell populations): Raji, IM-9 (both from B cells), Jurkat (from T cells), and THP-1 (from monocyte). The data set contained 12 mixed samples for which the true proportions of the underlying pure cell types were known (according to the mixture ratios).

Figure 2-1a shows that the LRM method leads to highly accurate abundance estimates (correlations r range from 0.75 to 0.99). Figure 2-1b shows that the CM method performs as well as that of the LRM when one uses the correlation between predicted and observed abundances as accuracy measure (r ranges from 0.83 to 0.96). But as expected, the CM method is inferior to the LRM when it comes to the mean square error, MSE, since the CM method will typically be poorly calibrated (mean MSE of the CM method is $3.2\times10^{-2}$ versus MSE = $2.1\times10^{-2}$ for the LRM method). The disparity between the two accuracy measures (correlation versus MSE) is

particularly pronounced for the THP1 cell population where the CM method returns higher correlation value (0.96 versus 0.75) while it gives a larger MSE ($4.9 \times 10^{-2}$ versus $3.1 \times 10^{-2}$).

Figure 2-1c-d shows how the abundance estimation methods performed on the kidney transplantation data set comprised of blood expression data from kidney transplantation patients [35] for whom cell count data were also available. Here the LRM method is superior to the CM irrespective of the accuracy measure: mean $r$ = 0.73 versus mean $r$ = 0.53, MSE = $2.0 \times 10^{-2}$ versus MSE = $2.8 \times 10^{-2}$. Figure 2-1e-f shows the results for the rat tissue data set. Here the LRM and CM method have similar performance when it comes to correlating observed and predicted abundances (mean $r$ = 0.94 versus mean $r$ = 0.82) but the LRM method tends to superior when it comes to MSE (MSE = $1.4 \times 10^{-2}$ versus MSE = $1.9 \times 10^{-2}$).

The performance of the expression deconvolution methods are summarized in Table 2 and Table 3, which present correlation values and MSE values, respectively. When we average our results across all the considered data sets we find: i) that the mean correlation of the LRM method (mean $r$ = 0.85) is slightly higher than that of the CM method (mean $r$ = 0.74) and ii) that mean MSE of the LRM method (mean MSE = $1.9 \times 10^{-2}$) is significantly lower than that of the CM method (mean MSE = $2.7 \times 10^{-2}$). In conclusion, the LRM method is superior to the CM method when it comes to abundance estimation.

**Empirical evaluation of cell type-specific expression (CTSE) estimation methods**

Here we evaluate the CTSE methods that aim to estimate the mean expression value of a gene in a given cell type based on mixed samples. These methods assume that the cell abundances vary across the mixed samples and that estimates of the cell abundances are available for each mixed

sample. In some tissues (such as blood), the true proportions of the underlying cell populations can be readily measured but in other heterogeneous tissues it can be desirable to estimate cell abundances, e.g. using the LRM or CM estimation method. To test whether estimated (predicted) abundances can be used as surrogates for measured abundances, we studied whether CTSE estimate based on true measured abundances are more accurate than those estimated via the LRM or CM method.

We first estimated the CTSE for each cell population based on measured abundances, and correlated the predicted mean values with the observed mean expression values. As shown in Table 2 (Additional File 2-1 contains corresponding scatter plots), a high correlation is observed between observed and estimated CTSE: mean $r$ = 0.91, 0.82, and 0.96 in the transformed blood cell data set, kidney transplantation data set, and rat tissue data set, respectively. This result illustrates the utility of Shen-Orr's CTSE method based on true observed abundances. Besides, this result can also serve as a benchmark for evaluating CTSE methods based on estimated abundance measures.

Strikingly, we find that the CTSE results based on estimated abundances are almost as good as those based on the true, measured abundances. Figures 2-2a-b show the results for the transformed blood data set where the mean r based on LRM estimates is 0.94, the mean r for CM based abundances is 0.91 which compares favorably with the mean r based on true, measured abundances is 0.91 (Additional File 2-1a). The results for the kidney transplantation data set (Figure 2-2c-d) and the rat tissue data (Figure 2-2e-f) also show that the LRM and CM based abundance return similar CTSE estimates as those based on the measured abundances.

In summary, the performance of predicted abundances is close to measured abundances during CTSE estimation, and can serve as surrogate when measured abundances are not accessible (the mean r from LRM CTSE is 0.89 which is the same value as that observed based on measured abundances). The results from LRM based abundances generally have better performance than the ones from CM based abundances (mean $r$ = 0.89 versus mean r = 0.83, Table 2), which is consistent with our previous results regarding abundance estimation.

**Set based CTSE analysis (SB-CTSE)**

Here, we generalize the idea of cell type-specific expression from individual genes to entire gene sets. We hypothesized that the averaging afforded by considering sets of genes (as opposed to individual genes) averages out the noise and may therefore increase the accuracy of the CTSE method. While having estimates for individual genes will always be more valuable than having such estimates for entire sets of genes, there will be important applications where set specific estimates are sufficient. For example, it could be useful to estimate the mean expression value of a defined set of T cell activation genes when studying how the adaptive immune system changes between states (e.g. acute rejection vs. stable kidney transplants). Many approaches and resources exist for selecting sets of genes. One could specifically identify such genes in the process of a biomarker discovery project using gene co-expression modules [16, 17, 28] or use empirical approaches based on literature-based gene ontology categories, gene sets from the Molecular Signatures Database [10, 55], or pathways from the KEGG data base.

As indicated by the name, the observed SB-CTSE of a particular gene set is simply defined by averaging the cell-specific expression levels of the genes inside the set. To predict the set

specific CTSE, one can simply average the predicted gene based-CTSEs. Apart from the mean value, one could also use alternative approaches for summarizing the expression levels of a gene set, e.g. the median or other approaches implemented in the collapseRows R function [49]. In the following, we report several examples of predicting SB-CTSEs.

**Using sets defined via co-expression modules.**

First, we applied it to the transformed blood cell data set. Here gene sets were defined based on the pure cell types (Jurkat, THP-1, IM9, and Raji) expression data with signed WGCNA approach [16, 17, 28]. The scatter plots in Figure 2-3a-b show how the measured (observed) SB-CTSE values relate to the estimated SB-CTSE values. As expected, estimating CTSE for sets of genes is more accurate than estimating CTSE for individual genes (mean correlation across multiple cell types is $r = 0.99$ for SB-CTSE versus mean $r = 0.92$ for individual genes). Next, we also applied SB-CTSE to the kidney transplantation data set based on co-expression modules. Again the modules were defined based on included pure cell types (CD8, CD4, CD19, and CD14) with signed WGCNA. The scatterplots in Figure 2-3c-d show that the observed SB-CTSE measure correlates highly with the predicted SB-CTSE (mean $r = 0.91$) which compares favorably to the results for gene based CTSE (mean $r = 0.76$). Finally, we evaluated SB-CTSE applied to modules in the rat tissue data set (Figure 2-3e-f). Here the modules were defined based on expression data from Lung, Liver and Brain. And the accuracy of SB-CTSE is also better (mean $r = 0.99$) than gene based CTSE estimate (mean $r = 0.92$).

**Using sets defined as pathways from the Molecular Signatures Database**

We applied SB-CTSE method to the transformed blood cell data set based on the Molecular Signatures Database V.3 (6769 gene sets). The results are shown in Additional File 2-3. The

scatter plots show that CTSE for sets of genes (pathways) is more accurate than estimating CTSE for individual genes (mean correlation across multiple cell types is $r = 0.95$ for SB-CTSE versus mean $r = 0.86$ for individual genes).

We also applied SB-CTSE to the kidney transplantation data where gene sets were defined using 64 curated immune pathways reported in (Additional File 2-5) [56]. Again the prediction accuracy for SB-CTSE (mean $r = 0.92$) is higher than that of a gene based analysis (mean $r = 0.76$).

In conclusion, we find that SB-CTSE outperforms gene based CTSE in 7 comparisons which indicates that the averaging afforded by SB-CTSE analysis increases the estimation accuracy.

**Accurate estimation of regional expression levels in heterogeneous brain regions**

Given the high complexity and relative inaccessibility of brain tissue, methods for separating out expression patterns from distinct cell populations have particular importance. Many experimental separation methods, e.g. laser capture micro-dissection, can be used to isolate different cell populations before expression profiling [45]. While pure cell assays provide invaluable resources to the neuro-science community, they are often not possible due to cost, tissue degradation, experimental design, or ethical considerations.

To evaluate whether computational deconvolution methods can be useful in neuro-scientific applications, we applied them to data from the NIH Blueprint Non-Human Primate Atlas (http://www.blueprintnhpatlas.org/). This atlas includes hundreds of microarrays run on five regions of the rhesus macaque brain across postnatal development. In particular, samples were

collected from hippocampus and striatum at both the gross anatomical level, as well as for specific subdivisions of these same structures (Figure 2-4a-b). While in our other applications, we were interested in cell types and cell type-specific expression patterns, this neuro-scientific application did not involve cell types but rather brain regions (structures). We wanted to evaluate whether macro-dissected samples from hippocampus (interpreted as mixed samples) could be used to estimate mean expression levels of genes in eight subdivisions (interpreted as pure populations in the admixture). The following sub-divisions of the hippocampus (Sub, CA1sr, CA1so, CA1sp, CA2sp, CA3sp, DGpo, and DGgcl; CA4sp and DGsgz were omitted due to high similarity to CA3sp and DGgcl, respectively) play a role analogous to that of the cell types in our previous deconvolution examples.

We briefly digress to mention that the reason why we use the term "population" instead of cell type in the names of our R functions (proportionsInAdmixture and populationMeansInAdmixture) is that we anticipate that they can be used in many applications that do not involve cell types, such as the example presented here.

In order to apply the LRM abundance estimation method, we first applied Bayes ANOVA to eight "pure" expression samples corresponding to sub-regions in the laser micro-dissection (LMD) microarray data. By definition, Bayes ANOVA allows one to identify genes that vary significantly across the sub-regions. The 320 genes with the lowest p-value were used as input of the LRM abundance method (implemented in proportionsInAdmixture) to estimate the proportion of each hippocampal sub-region present in each macro-dissected (mixed) sample. The estimated proportions were used as input of the CTSE approach (implemented in the R function populationMeansInAdmixture) to estimate the hippocampal subregion-specific mean expression values based on the macro-dissected gene expression patterns (for example, in CA1-so: Figure 2-

4c). The proportion for DGpo and Subiculum were both calculated to be 0, and therefore we could not estimate the expression patterns of genes from these areas. Overall, we found moderate correlations (0.46, 0.55, 0.64, 0.80, and 0.84) between predicted and observed expression levels (Figure 2-4d), suggesting that the CTSE method is applicable to deconvoluting expression patterns in distinct brain regions, in addition to pure cell populations.

We hypothesized that the estimation accuracy could be improved if similar pure samples (i.e. sub-regions comprised of relatively similar cell types) would be combined. Figure 2-4e shows a multidimensional scaling plot of all the data from young adult macaque (T = 48). Note that samples from all sub-regions (except subiculum) clustered into three distinct groups based on the predominant cell type in that area: CA1-sp, CA2-sp, CA3-sp, and CA4-sp are pyramidal cell layers; DGgcl and DGsgz contain mostly granule cells; and CA1-so, CA1-sr, and DGpo contain a mix of GABAergic neurons and glial cells. Interestingly, the macro-dissected samples appear between these three groups on the plot, further suggesting that these samples could be a mix of many of these cell populations. Using predominant cell type as our classifying variable, we repeated the procedure described above to compare the estimated (predicted) versus observed mean expression value of genes for each cell class. We were able to predict these expression levels with very high accuracy (Figure 2-4f; r >= 0.97 in each case).

To further test this method in the context of heterogeneous brain tissue, we repeated our analysis using mRNA from different compartments of striatum. Striatum tissue could be split into five distinct groups based on composition of similar cell types and expression patterns: Acb/Tu, Cd/Pu, GPe/GPi, ic, and isl (Additional File 2-6a). As with hippocampus, the macro-dissected samples from striatum have expression patterns between these five groups on an MDS plot, suggesting a heterogeneous make-up of these samples. Once again, we find high correlations of

estimated versus actual mean expression levels of genes for each striatal sub-compartment (Additional File 2-6b-f): in Acb/Tu $r$ = 0.98, in Cd/Pu $r$ = 0.91, in GPe/GPi $r$ = 0.72, in ic $r$ = 0.95, and in isl $r$ = 0.65. Overall, these results show that the CTSE method has merit for gene expression deconvolution of brain region-specific expression levels assuming the regions are sufficiently distinct.

**Simulations for evaluating the effect of mis-specifying cell populations**

A critical issue of expression deconvolution methods is to specify the cell populations that make up a mixed sample. However, in real data applications one may not know all of the cell types. Or one may not even be interested in estimating CTSE for each possible cell type. Therefore, it is important to determine whether the expression deconvolution methods perform well when the cell populations are mis-specified, e.g. when one cell type is omitted from the analysis. We first assumed that one cell population is erroneously omitted from the analysis and used simulations to study the consequences of this omission.

We simulated the gene expression data using the previously mentioned linearity assumption (Equation 2), i.e. the expression level of a gene in a mixed sample is a weighted average of its expression value in the respective pure cell populations. 400 genes, 4 cell populations, and 50 mixed samples are simulated. The mean values of the 400 genes in each of the four pure cell populations are given by the matrix X400×4. Thus, these simulated values serve as known gold standard for evaluating the estimates of the CTSE method. Each cell population had varying levels of abundances across the 50 samples. The cell counts of the 4 cell types were drawn from a Poisson distribution. The abundances of the 4 cell types differed greatly: on average each

mixed sample contained 53% P1 cells, 27% P2 cells, 13% P3 cells and 7% P4 cells. We simulated dependencies between cell abundances, e.g. the proportion of P3 and P4 were highly correlated ($r$ = 0.44) across the 50 mixed samples. The simulated proportions of the 4 cell types in the 50 mixed samples were given by the matrix P4×50. Under the linearity assumption (Equation 2), the expression data of the 400 genes in the 50 mixed samples is given by X.mixed400×50 = X400×4 ×P4×50 + noise where random Gaussian noise was added.

First, we used the simulation data to evaluate the LRM and CM abundance estimation methods. We assumed that X400×4 is known but P4×50 is unknown. The LRM abundance method (implemented in the proportionsInAdmixture R function) based on the known pure mean expression values X400×4 allowed us to estimate the unknown proportions P4×50. The resulting abundance for each cell population is compared with true abundances in Figure 2-5a. Ignoring the least abundant cell population (P4) barely has an effect on the LRM method when it comes to estimating the abundances of the remaining three cell pupations (Figure 2-5b) (mean $r$ = 0.84 compared to that where all cell types are correctly specified mean $r$ = 0.86). Even ignoring the more abundant cell populations (such as P1) has a negligible effect on the abundance estimation accuracy (Figures 2-6c-e) (mean $r$ = 0.79).

Next we evaluated how omitting a cell population affects the performance of CTSE estimation. Here we assume that the user inputs the true abundance estimates of all but 1 cell population the R function populationMeansInAdmixture. Figure 2-6 shows that the CTSE estimation is adversely affected if cell types with large abundance are erroneously omitted from the analysis: when the most abundant cell type, P1, is omitted, the CTSE estimation accuracy is greatly diminished (mean $r$ = 0.52, Figure 2-6e). However, omitting low abundance cell types had only a negligible effect when it came to estimating mean CTSE levels in abundant cell types.

In conclusion, we find that erroneously omitting cell types from the analysis has a minor effect on estimating cell type abundances but one should not omit high abundance (i.e. >50%) cell types when it comes to estimating mean expression values in cell types (i.e. CTSE estimation methods).

## Conclusions

We carried out comprehensive evaluations to assess the reliability of several expression deconvolution methods (Table 1) in 4 empirical datasets. We started out by comparing two different approaches for estimating cell abundances: i) a multivariate linear regression model (LRM) method, which assumes that the mean expression values in each cell population are known and ii) the cell type marker (CM) method, which assumes only that cell markers are known.

However, the abundances from the CM method are calculated based on the mean expression level of marker genes and are on a different scale (depending on the normalization procedure) from that of the true cell proportion (ranging from 0 to 1). We have found it useful to calibrate the CM estimate as described in Equation 1. While this calibration step greatly improves the performance of the CM method, our studies show that the resulting estimate is inferior to that of the LRM method; the CM approach has a larger mean square error (CM MSE = $2.7\times10-2$ versus LRM MSE = $1.9\times10-2$). Our empirical studies show that both methods (abundance estimate, CTSE estimate) work well for estimating cell abundances. Although the LRM method is more accurate, the CM method remains attractive for the following reasons: i) it makes fewer assumptions (in particular it does not require that the user know the mean expression values of

marker genes in the cell type under consideration), ii) it is easily implemented, and iii) it can be readily used for estimating the cell abundances of a single cell type of interest.

Next we use these abundance estimation approaches for estimating CTSE values. Strikingly, the accuracy of CTSE estimates based on the LRM abundances is similar to CTSE estimates based on true measured abundances, which obviates the need to measure cell abundances directly. Our empirical data show that CTSE estimates based on LRM abundance estimates are far more accurate than those based on CM abundance estimates (LRM mean $r = 0.89$ versus CM mean $r = 0.83$). Thus, we recommend using abundance estimates based on the LRM method as surrogates for measured abundances if the latter are not available. Simulations indicate that erroneously omitting cell types from the analysis only has an adverse effect on CTSE estimation if the omitted cell type has a high abundance.

Since cell type-specific estimation of the mean value for individual genes is probably not accurate enough for many real data applications, we propose to focus on gene sets instead of individual genes. Set based CTSE (SB-CTSE) can be carried out with a variety of different gene sets, e.g. literature based pathways or co-expression modules. In our applications we find that SB-CTSE often outperforms gene based CTSE analyses: SB-CTSE based on co-expression modules leads to a mean $r = 0.95$, SB-CTSE based on pathways from the Molecular Signatures Database leads to a mean $r = 0.96$ while gene based CTSE leads to a mean $r = 0.87$.

While most expression deconvolutions will involve cell populations, we show that these methods are also useful in other settings, e.g. to study brain region-specific expression values based on macro-dissected brain expression data.

## Methods

### Transformed blood cell data set (GEO accession number: GSE11058):

Four pure cell line samples were obtained from the American Type Culture Collection (ATCC) as following: Jurkat (T cell leukemia), THP-1 (acute monocytic leukemia), IM9 (B lymphoblastoid multiple myeloma) and Raji (Burkitt B-cell lymphoma). They were selected since they show similar but distinguishable expression profiles. These cell lines provided the abundant sources of pure cells necessary to support experimental mixing of different types of cells in different ratios. Finally, 4 mixed samples were created based on defined proportions. Both pure cell lines and mixed samples were profiled with Affymetrix HGU133 expression microarrays in triplicate [31].

### Kidney transplantation data set (GEO accession number: GSE24223):

Whole peripheral blood from 10 kidney transplant patients was collected into PaxGene (Qiagen) tubes immediately prior to administration of immunosuppression and transplantation and at weeks 1, 2, 4, 8, and 12. Blood Samples from 5 healthy controls (2 males and 3 females, 25–45 years of age) were collected following the same protocol at a single time point. In parallel, magnetic beads were used to isolate the following four cell subset populations from whole blood: CD14 (monocyte/macrophage), CD19 (B cells), CD4, and CD8 T cells. These samples were analyzed with Affymetrix HGU133 Plus 2.0 GeneChips, following standard protocols. Finally, expression profiling from 64 whole blood samples, 7 CD14 samples, 6 CD19 samples, 10 CD4 samples, and 9 CD8 samples were included [35].

**Rat tissue data set (GEO accession number: GSE19830):**

Three pure rat tissues, including brain, liver, and lung, were experimentally mixed at 11 different proportions. Besides these 11 mixed samples, there are also 3 pure tissue samples. All these samples are analyzed with rat-specific RAE230_2 whole-genome expression arrays (Affymetrix), and each sample was analyzed in triplicate. So, there were totally 42 expression profiles in this data set [34].

**Brain region data set (Unpublished):**

A technical white paper describing LMD sample preparation, microarray profiling and preprocessing in detail is available at http://www.-blueprintnhpatlas.org/nhp/docs.html. Brain tissue was collected from five brain regions (including hippocampus and striatum) for a total of 12 rhesus macaque monkeys (N = 3 at each of four post-natal developmental time points: T = 0, 3, 12, and 48 months). Microarrays from T = 0 were excluded from these analyses, as the expression levels from this time point show considerable differences from T = 3, 12, and 48 (Unpublished observations). LMD was performed on a Leica LMD6000 (Leica Microsystems, Inc., Bannockburn, IL), using the Nissl stain as a guide to identify target brain regions. Microdissected tissue was collected directly into RLT buffer with ß-mercaptoethanol, and processed following the manufacturer's directions for the RNeasy Micro kit (Qiagen Inc., Valencia, CA) to isolate RNA. Samples passing RNA quality control (QC) were amplified, labeled, and hybridized to catalog GeneChip Rhesus Macaque Genome Arrays from Affymetrix

containing 52,803 probe sets/sequences. Macro-dissected samples were collected from macaque brain using manual dissection and processed following a similar procedure.

**Linear regression based cell abundance estimate and CTSE estimate**

As its name says, our R function proportionsInAdmixture implements methods for estimating the proportions of different cell populations in a mixture (based on multivariate regression). In this function, the expression profile of a mixed sample is modeled as a linear combination of the expression profile of each cell population comprising that sample. The function takes as input the expression profiles of mixed tissue samples and known expression values in pure cell populations. It outputs estimates of the proportions of the various cell types. The aim is to estimate the proportion of each cell population for all the mixed samples.

The methods implemented in this function were motivated by the gene expression deconvolution approach described by Abbas et al (2009), Lu et al (2003), Wang et al (2006), and Kuhn et al (2011). Similar to the notation from Kuhn et al (2011), the expression value y of gene g in a given mixed sample is modeled by

where xpg is the expression profile of gene g in the pure cell population p, fp is the fraction of population p in the mixed sample, and eg is measure error for gene g.

All of the considered expression deconvolution approaches start with Equation 2. When estimating the fractions fp one assumes that the values xpg are known. (This implemented in the

function proportionsInAdmixture) When estimating the values xpg, one assumes that the fp values are known (Implemented in the function populationMeansInAdmixture).

However, the two deconvolution approaches differ in the genes that are being considered. When estimating fp (proportions) one does not need to consider all genes. Instead, it is advisable to focus on a subset of genes whose mean values vary greatly across the cell populations (e.g. cell type markers that are over expressed in specific cell types). Cell markers can be found based on published data or by finding genes that are differentially expressed between pure cell populations (using a T-test or ANOVA across the pure populations). When estimating cell type-specific expressions, (i.e. xpg) any set of genes could be used (e.g. all available genes or genes that form a pathway). In order to arrive at stable estimates of cell type-specific expression one need to analyze mixture samples whose composition with respect to the cell types varies. In other words, if a cell type has a constant abundance (e.g. fp = 0.10) then one cannot estimate the corresponding cell-specific expression levels.

A third variant of the deconvolution approach is to combine these two approaches in the sense that one first estimates the abundances to arrive at estimates  and then to use it as input for cell type-specific expression estimation.

**Simulation of expression profiles**

In the following simulations we assume that the mixed sample is composed of four cell populations denoted by P1, P2, P3, and P4. We simulate 50 mixed samples that contain different proportions of the 4 cell types. Each simulated gene is a marker gene in one of the 4 populations.

For each population there are 100 marker genes, i.e. we consider a total of 400 genes. Specifically, the expression profiles of a mixed sample are simulated in the following steps.

First, each gene is assumed to be highly expressed in exactly 1 of the 4 pure populations where its expression value is drawn from a normal distribution with mean 1 and variance .04. In the remaining 3 populations, its expression value is drawn from a N (0, 0.22) distribution. This step results in a matrix of pure cell expression values X400×4 whose 400 rows correspond to the genes and whose 4 columns correspond to the 4 pure cell populations.

Second, we simulate the counts of the pure cells in the mixed sample. The number of pure cells (of a given type) in a mixed sample is drawn from a Poisson distribution whose mean value is given by the lambda parameter. The lambda values for populations P1 and P3 are given by 16 and 4, respectively. To simulate a dependence between the cell counts for populations 1 and 2, the lambda value for population P2 was set to be half that of population P1. Similarly, the lambda of P4 is half that of P3. To turn counts into proportions, each count was divided by the total count (sum). This step resulted in a matrix F4×50 of proportions whose 4 rows correspond to the 4 pure cell types and whose 50 columns correspond to the 50 mixed samples. On average, population P1, P2, P3, P4 had a prevalence of 54%, 29%, 12%, 5%, respectively.

Third, we simulated the expression levels of the 400 genes in 50 mixed samples. Toward this end, we used the formula Y400×50 = X400×4 F4×50 + noise, where the entries of the noise matrix were sampled from a normal distribution N(0,0.12). Note that Y400×50 is matrix of expression levels whose 400 rows correspond to the 400 genes and whose 50 columns correspond to mixed samples.

**Figure 2-1**

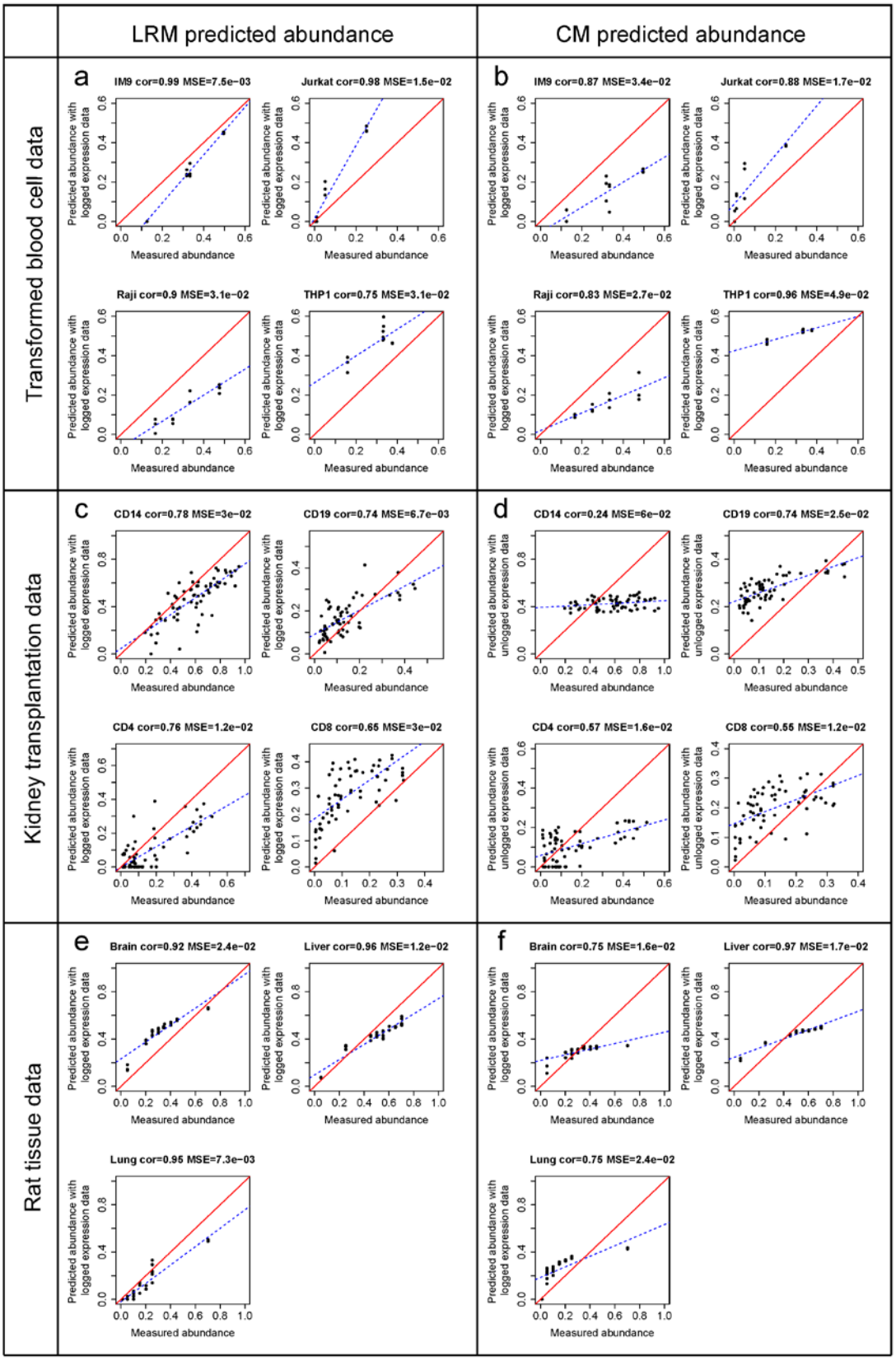
LRM predicted abundance
CM predicted abundance
Transformed blood cell data
Kidney transplantation data
Rat tissue data
a
IM9 cor=0.99 MSE=7.5e−03
Jurkat cor=0.98 MSE=1.5e−02
Raji cor=0.9 MSE=3.1e−02
THP1 cor=0.75 MSE=3.1e−02
b
IM9 cor=0.87 MSE=3.4e−02
Jurkat cor=0.88 MSE=1.7e−02
Raji cor=0.83 MSE=2.7e−02
THP1 cor=0.96 MSE=4.9e−02
c
CD14 cor=0.78 MSE=3e−02
CD19 cor=0.74 MSE=6.7e−03
CD4 cor=0.76 MSE=1.2e−02
CD8 cor=0.65 MSE=3e−02
d
CD14 cor=0.24 MSE=6e−02
CD19 cor=0.74 MSE=2.5e−02
CD4 cor=0.57 MSE=1.6e−02
CD8 cor=0.55 MSE=1.2e−02
e
Brain cor=0.92 MSE=2.4e−02
Liver cor=0.96 MSE=1.2e−02
Lung cor=0.95 MSE=7.3e−03
f
Brain cor=0.75 MSE=1.6e−02
Liver cor=0.97 MSE=1.7e−02
Lung cor=0.75 MSE=2.4e−02
Predicted abundance with logged expression data
Predicted abundance with unlogged expression data
Measured abundance

**Figure 2-1. Empirical comparison of abundance estimation methods using three different expression data sets**

Two abundance estimation methods (LRM and CM, corresponding to the columns) are compared in three independent data sets. Panels a + b report the findings for the transformed blood cell data, panels c + d for the kidney transplantation data, and e + f for the rat tissue data. Each dot is a mixed sample comprised of multiple cell types.

Each scatterplot shows how the estimated abundance (y-axis) of the cell population is related to the true observed, measured abundance (x-axis). The figure heading reports the cell population, the correlation between estimated and true value, and the mean square error of the estimation. The dashed blue line shows the regression line resulting from regressing y on x. The solid red line corresponds to the line y = x. The plots in the first column (panels a, c, e) shows that the LRM abundance estimates are highly correlated with the true values (mean $r = 0.85$). The plots in the second column (panels b, d, f) show that the CM estimation method leads to a high correlation as well (mean $r = 0.74$) but it tends to have a larger mean square error than the LRM method.

**Figure 2-2**

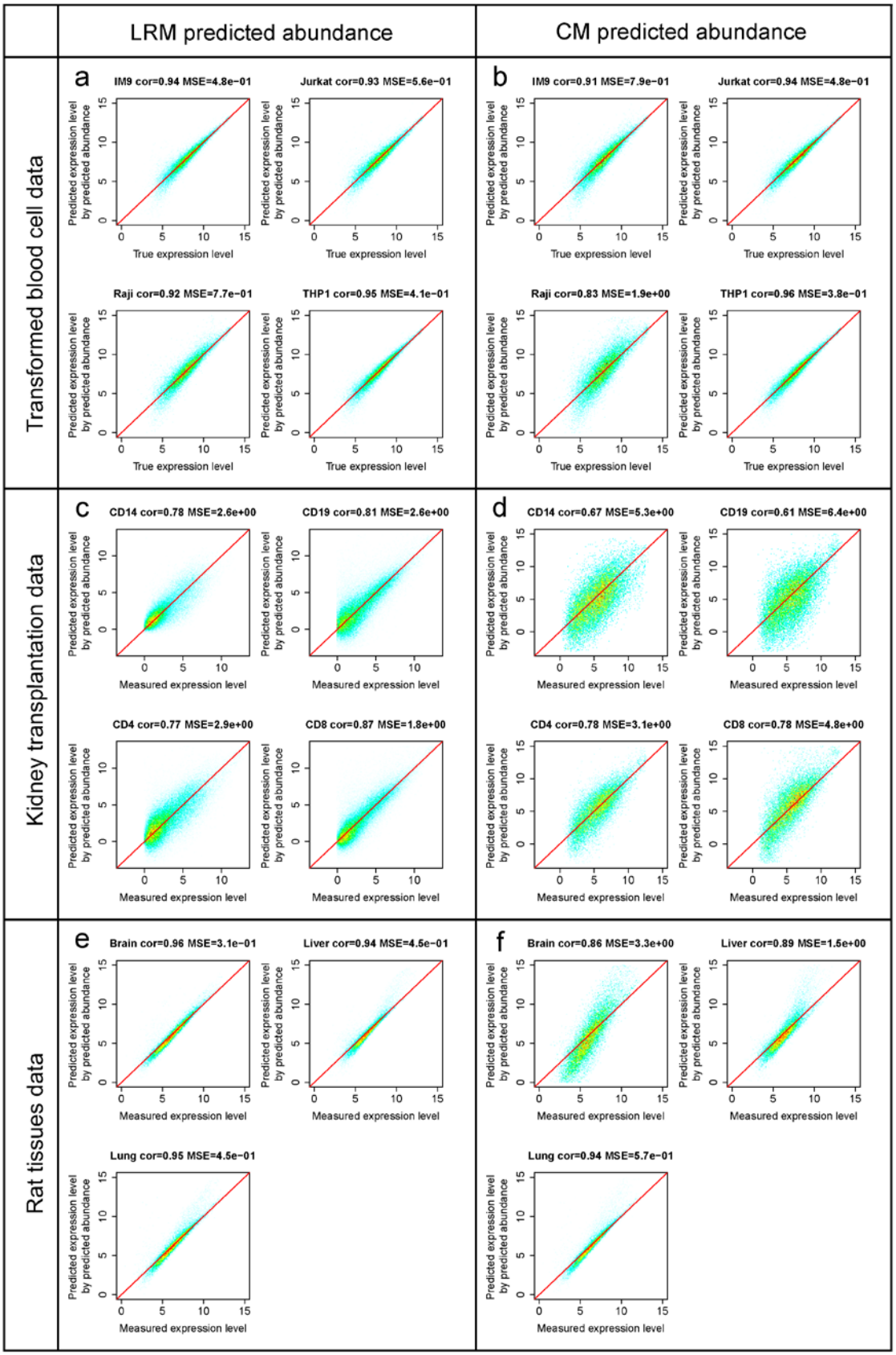
LRM predicted abundance
CM predicted abundance
Transformed blood cell data
Kidney transplantation data
Rat tissues data
a
IM9 cor=0.94 MSE=4.8e−01
Jurkat cor=0.93 MSE=5.6e−01
Raji cor=0.92 MSE=7.7e−01
THP1 cor=0.95 MSE=4.1e−01
b
IM9 cor=0.91 MSE=7.9e−01
Jurkat cor=0.94 MSE=4.8e−01
Raji cor=0.83 MSE=1.9e+00
THP1 cor=0.96 MSE=3.8e−01
c
CD14 cor=0.78 MSE=2.6e+00
CD19 cor=0.81 MSE=2.6e+00
CD4 cor=0.77 MSE=2.9e+00
CD8 cor=0.87 MSE=1.8e+00
d
CD14 cor=0.67 MSE=5.3e+00
CD19 cor=0.61 MSE=6.4e+00
CD4 cor=0.78 MSE=3.1e+00
CD8 cor=0.78 MSE=4.8e+00
e
Brain cor=0.96 MSE=3.1e−01
Liver cor=0.94 MSE=4.5e−01
Lung cor=0.95 MSE=4.5e−01
f
Brain cor=0.86 MSE=3.3e+00
Liver cor=0.89 MSE=1.5e+00
Lung cor=0.94 MSE=5.7e−01
Predicted expression level by predicted abundance
True expression level
Measured expression level

**Figure 2-2. Empirical comparison of cell type-specific expression (CTSE) values: gene based analysis**

To estimate the mean value of a gene in a given cell population one needs to input an estimate of the cell type abundance (LRM or CM corresponding to the 2 columns). The figure evaluates how the abundance estimate affects the accuracy of estimating the mean expression value of a gene in a specific cell population. Panels a + b report the findings for the transformed blood cell data, panels c + d for the kidney transplantation data, and e + f for the rat tissue data. Each dot corresponds to a gene.

Each scatterplot shows how the estimated expression value (y-axis) of the gene in the given cell population is related to the true observed, measured mean value (x-axis). The figure heading reports the cell population, the correlation between estimated and true value, and the mean square error of the estimation. The solid red line corresponds to the line y = x. Each figure shows the density of dots (red indicates high density while green indicates low density). Overall, the cell type-specific analyses lead to fairly high correlations between true and estimated mean expression values but the LRM based estimates are slightly better than those of the CM based method (mean $r$ = 0.89 for the LRM based method versus mean $r$ = 0.83 for the CM based methods).

**Figure 2-3**

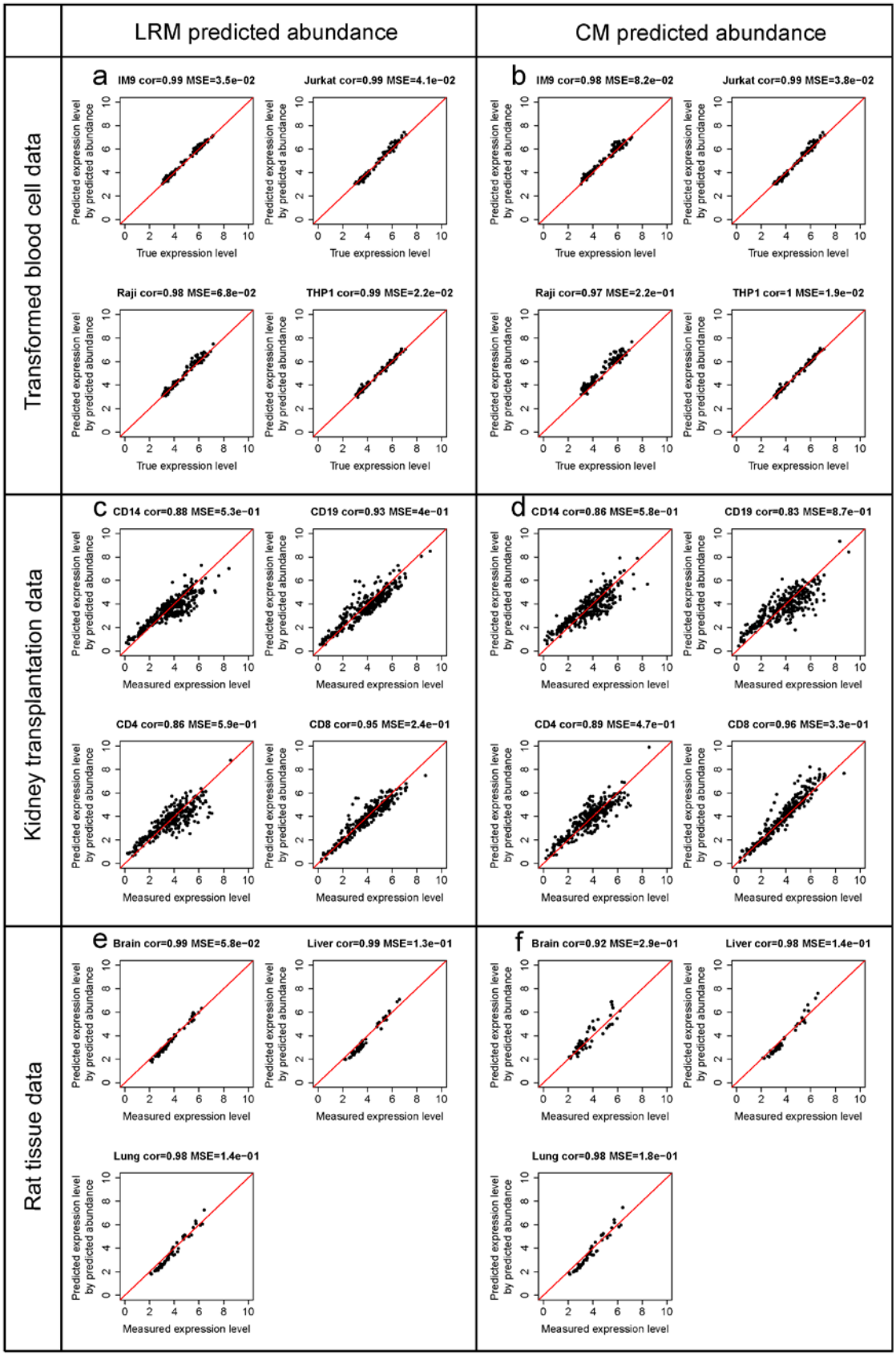
LRM predicted abundance
CM predicted abundance
Transformed blood cell data
Kidney transplantation data
Rat tissue data
a
IM9 cor=0.99 MSE=3.5e−02
Jurkat cor=0.99 MSE=4.1e−02
Raji cor=0.98 MSE=6.8e−02
THP1 cor=0.99 MSE=2.2e−02
b
IM9 cor=0.98 MSE=8.2e−02
Jurkat cor=0.99 MSE=3.8e−02
Raji cor=0.97 MSE=2.2e−01
THP1 cor=1 MSE=1.9e−02
c
CD14 cor=0.88 MSE=5.3e−01
CD19 cor=0.93 MSE=4e−01
CD4 cor=0.86 MSE=5.9e−01
CD8 cor=0.95 MSE=2.4e−01
d
CD14 cor=0.86 MSE=5.8e−01
CD19 cor=0.83 MSE=8.7e−01
CD4 cor=0.89 MSE=4.7e−01
CD8 cor=0.96 MSE=3.3e−01
e
Brain cor=0.99 MSE=5.8e−02
Liver cor=0.99 MSE=1.3e−01
Lung cor=0.98 MSE=1.4e−01
f
Brain cor=0.92 MSE=2.9e−01
Liver cor=0.98 MSE=1.4e−01
Lung cor=0.98 MSE=1.8e−01
Predicted expression level by predicted abundance
True expression level
Measured expression level

**Figure 2-3. Empirical comparison of set based cell type-specific expression (SB-CTSE) values: set based analysis**

This figure is analogous to Figure 2-2 except that dots represent sets of genes as opposed to individual genes. Thus, each dot corresponds to a gene set defined using the gene co-expression module. Panels a + b correspond to the transformed blood cell, panels c + d correspond to the kidney transplantation data, and e + f correspond to the rat tissue data. The two columns report the findings for different ways of estimating the cell population abundances (LRM versus CM based method). The dot here corresponds to one gene co-expression module. Overall, the results for the set based analysis are superior to those of the individual gene based analysis (reported in Figure 2). Further, the LRM based analysis leads to a similar estimation accuracy (mean $r$ = 0.96) as the CM based analysis (mean $r$ = 0.94).

**Figure 2-4**

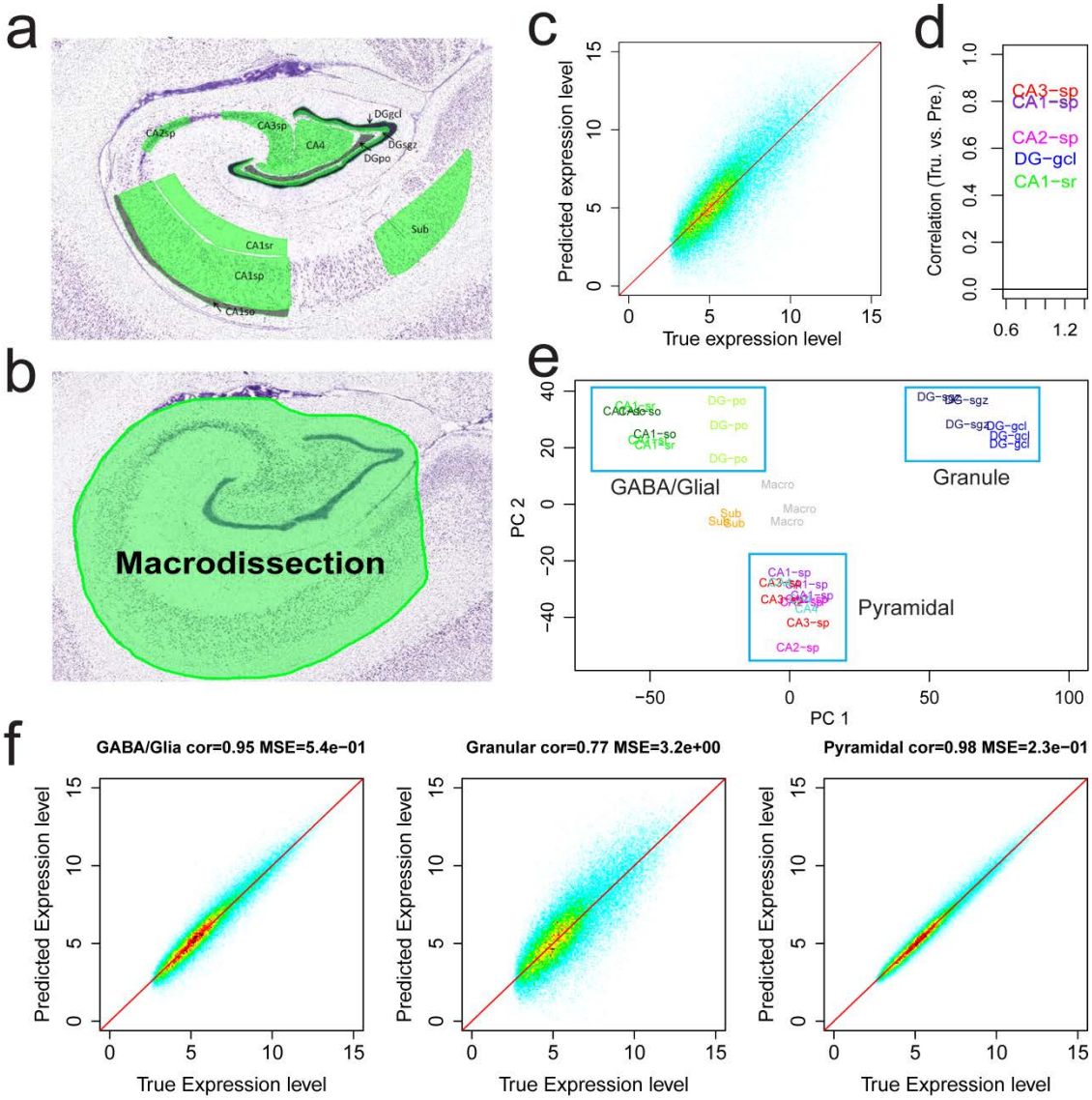
a
CA2sp
CA3sp
CA4
DGgcl
DGsgz
DGpo
CA1sr
CA1sp
CA1so
Sub
b
Macrodissection
c
Predicted expression level
True expression level
d
Correlation (Tru. vs. Pre.)
CA3−sp
CA1−sp
CA2−sp
DG−gcl
CA1−sr
e
PC 2
PC 1
GABA/Glial
Granule
Pyramidal
Macro
Sub
DG−po
DG−sgz
DG−gcl
CA1−sp
CA3−sp
CA2−sp
f
GABA/Glia cor=0.95 MSE=5.4e−01
Granular cor=0.77 MSE=3.2e+00
Pyramidal cor=0.98 MSE=2.3e−01
Predicted Expression level
True Expression level

**Figure 2-4. Expression of sufficiently distinct hippocampal cell types can be estimated from macro-dissected tissue**

a) Anatomical delineations of 10 hippocampal sub-regions in adult (48 months) macaque brain. mRNA from these sub-regions was collected using laser micro-dissection (LMD) and run on custom Affymetrix microarrays across four developmental time points (0, 3, 12, 48 months). b) mRNA from macro-dissected hippocampal samples was also collected. c) Expression levels of genes in the macaque CA1-so estimated from macro-dissected samples (x-axis) recapitulate true expression levels measured by LMD (y-axis) to a moderate degree. Each of the 52865 points represents a probe set on the microarray. d) Correlations between true and predicted expression levels (y-axis) are comparable across the other hippocampal sub-regions. e) Multidimensional scaling (MDS) using all genes groups samples based on the predominant cell types of the brain region (T = 48 months). X and y axes correspond to first two principal components (arbitrary units). Samples are plotted using text corresponding to their region of origin and are color coded for clarity. f) Grouping samples based on cell type leads to much more accurate estimations of gene expression profiles from heterogeneous tissue. Labeling as in c. Plots for GABA/glia (left), granular (center), and pyramidal (right) cells are show at the same scale.

**Figure 2-5**

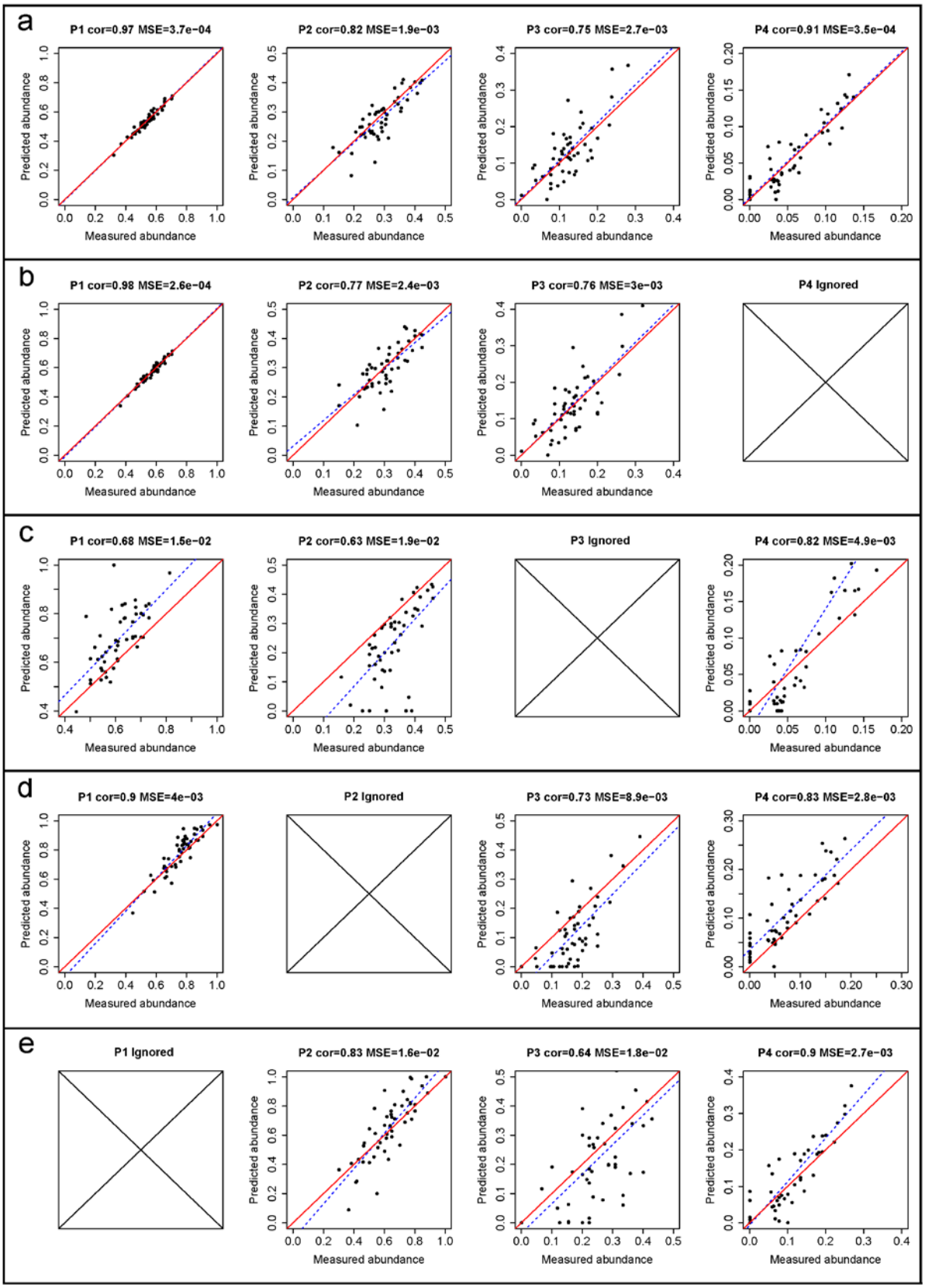
a
P1 cor=0.97 MSE=3.7e−04
P2 cor=0.82 MSE=1.9e−03
P3 cor=0.75 MSE=2.7e−03
P4 cor=0.91 MSE=3.5e−04
b
P1 cor=0.98 MSE=2.6e−04
P2 cor=0.77 MSE=2.4e−03
P3 cor=0.76 MSE=3e−03
P4 Ignored
c
P1 cor=0.68 MSE=1.5e−02
P2 cor=0.63 MSE=1.9e−02
P3 Ignored
P4 cor=0.82 MSE=4.9e−03
d
P1 cor=0.9 MSE=4e−03
P2 Ignored
P3 cor=0.73 MSE=8.9e−03
P4 cor=0.83 MSE=2.8e−03
e
P1 Ignored
P2 cor=0.83 MSE=1.6e−02
P3 cor=0.64 MSE=1.8e−02
P4 cor=0.9 MSE=2.7e−03
Predicted abundance
Measured abundance

**Figure 2-5. Simulation study to evaluate how the LRM abundance estimation method is affected if a cell population is erroneously omitted**

The 4 columns correspond to 4 simulated cell populations of decreasing average abundances (population 1 is most abundant, population 4 is least abundant). Each scatter plot reports the predicted abundance (y-axis) and the true abundance (x-axis), the correlation, and the mean square error. The dashed blue line shows the regression line resulting from regressing y on x. The solid red line corresponds to the line y = x. Overall, we observe high correlations and low MSEs when all 4 populations are used. Panel b shows the performance of the LRM abundance estimation method when population 4 is erroneously ignored. Note that ignoring the least abundant cell type P4 has only a negligible effect in this case. Panel c-e show the results when other cell types are ignored. Overall, we find that the LRM abundance estimation method is highly robust with respect to ignoring cell types even in case of ignoring the most abundant cell type (see panel e).

**Figure 2-6**

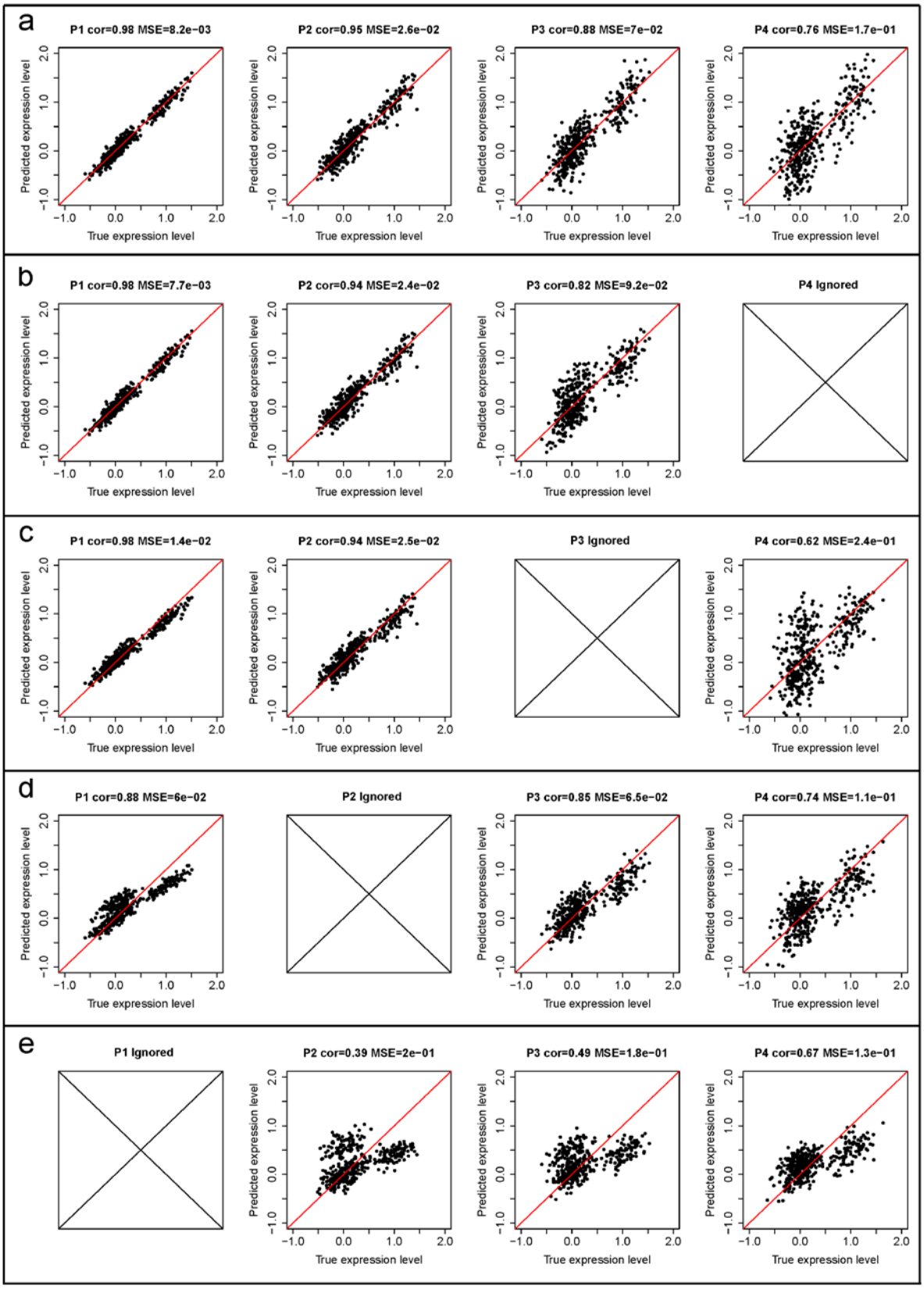
a
P1 cor=0.98 MSE=8.2e−03
P2 cor=0.95 MSE=2.6e−02
P3 cor=0.88 MSE=7e−02
P4 cor=0.76 MSE=1.7e−01
b
P1 cor=0.98 MSE=7.7e−03
P2 cor=0.94 MSE=2.4e−02
P3 cor=0.82 MSE=9.2e−02
P4 Ignored
c
P1 cor=0.98 MSE=1.4e−02
P2 cor=0.94 MSE=2.5e−02
P3 Ignored
P4 cor=0.62 MSE=2.4e−01
d
P1 cor=0.88 MSE=6e−02
P2 Ignored
P3 cor=0.85 MSE=6.5e−02
P4 cor=0.74 MSE=1.1e−01
e
P1 Ignored
P2 cor=0.39 MSE=2e−01
P3 cor=0.49 MSE=1.8e−01
P4 cor=0.67 MSE=1.3e−01
Predicted expression level
True expression level
−1.0
0.0
1.0
2.0

**Figure 2-6. Simulation study to evaluate how the performance of the CTSE estimation method is affected if a cell population is erroneously omitted**

This figure is analogous to Figure 5 but each point corresponds to a gene and we evaluate the performance of the CTSE based on the LRM abundance input. The 4 columns correspond to 4 simulated cell populations of decreasing average abundances (population 1 is most abundant, population 4 is least abundant). Each scatter plot reports the predicted mean expression value (y-axis) in the cell population and the true mean value (x-axis), the correlation, and the mean square error. The dashed blue line shows the regression line resulting from regressing y on x. The solid red line corresponds to the line y = x. Panel a shows a high predictive accuracy when all 4 populations are used. Note that the CTSE of the mean levels are most accurate for the most abundant cell population (P1). Panel b shows the performance of the CTSE estimation method when population 4 is erroneously ignored. Note that ignoring the least abundant cell type P4 has only a minor effect in this case. Panel c-e show the results when other cell types are ignored. Note that ignoring the most abundant cell type markedly degrades the prediction accuracy (see panel c).

## Table 2-1. Overview of expression deconvolution methods

Table 1. Overview of expression deconvolution methods

| Name | Purpose | Unit of observation | R function | Input | Output |
|---|---|---|---|---|---|
| LRM abundance | Abundance estimation | Sample | *proportionsInAdmixture* | Mean expression values in pure population; Expression profiles from mixed samples | estimate of proportion |
| CM abundance | Abundance estimation | Sample | *collapseRows* | Identifiers of cell markers; Expression profiles from mixed samples | measure that is correlated with the proportion |
| LRM CTSE | CTSE estimation | Gene | *populationMeansInAdmixture* | LRM based predicted abundances; Expression profiles from mixed samples | mean expression of genes in the cell type |
| CM CTSE | CTSE estimation | Gene | *populationMeansInAdmixture* | CM based predicted abundances; Expression profiles from mixed samples | mean expression of genes in the cell type |
| Measured abundance CTSE | CTSE estimation | Gene | *populationMeansInAdmixture* | Measured abundances; Expression profiles from mixed samples | mean expression of genes in the cell type |
| SB-CTSE | SB-CTSE estimation | Set of genes | *populationMeansInAdmixture* | Population abundances; Expression profiles from mixed samples; Gene sets information | mean expression of gene sets in the cell type |

**Table 2-2. Correlations between true and estimated values for different types of expression deconvolution methods**

| Table 2. Correlations between true and estimated values for different types of expression deconvolution methods | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Data set | Cell population | Abun. LRM | Abun. CM | CTSE LRM | CTSE CM | CTSE measured abun. | SB CTSE LRM (WGCNA) | SB CTSE CM (WGCNA) | SB CTSE measured abun. (WGCNA) | SB CTSE LRM (GSEA) | SB CTSE CM (GSEA) | SB CTSE measured abun. (GSEA) |
| Transformed blood cell data | IM9 | 0.99 | 0.87 | 0.94 | 0.91 | 0.93 | 0.99 | 0.98 | 0.99 | 0.98 | 0.98 | 0.97 |
| | Jurkat | 0.98 | 0.88 | 0.93 | 0.94 | 0.83 | 0.99 | 0.99 | 0.96 | 0.97 | 0.98 | 0.93 |
| | Raji | 0.9 | 0.83 | 0.92 | 0.83 | 0.96 | 0.98 | 0.97 | 0.99 | 0.98 | 0.95 | 0.98 |
| | THP1 | 0.75 | 0.96 | 0.95 | 0.96 | 0.9 | 0.99 | 1 | 0.99 | 0.99 | 0.99 | 0.97 |
| Kidney transplantation data | CD14 | 0.78 | 0.24 | 0.78 | 0.67 | 0.86 | 0.88 | 0.86 | 0.94 | 0.93 | 0.87 | 0.96 |
| | CD19 | 0.74 | 0.74 | 0.81 | 0.61 | 0.83 | 0.93 | 0.83 | 0.93 | 0.94 | 0.89 | 0.93 |
| | CD4 | 0.76 | 0.57 | 0.77 | 0.78 | 0.83 | 0.86 | 0.89 | 0.93 | 0.95 | 0.96 | 0.97 |
| | CD8 | 0.65 | 0.55 | 0.87 | 0.78 | 0.77 | 0.95 | 0.96 | 0.91 | 0.96 | 0.93 | 0.92 |
| Rat tissue data | Brain | 0.92 | 0.75 | 0.96 | 0.86 | 0.96 | 0.99 | 0.92 | 0.99 | 0.99 | 0.95 | 0.99 |
| | Liver | 0.96 | 0.97 | 0.94 | 0.89 | 0.94 | 0.99 | 0.98 | 0.98 | 0.99 | 0.97 | 0.98 |
| | Lung | 0.95 | 0.75 | 0.95 | 0.94 | 0.97 | 0.98 | 0.98 | 0.99 | 0.99 | 0.98 | 0.99 |
| Mean *r* | All | 0.85 | 0.74 | 0.89 | 0.83 | 0.89 | 0.96 | 0.94 | 0.96 | 0.97 | 0.95 | 0.96 |

**Table 2-3. Mean squar errors for the different types of expression deconvolution methods**

| Table 3. Mean squar errors for the different types of expression deconvolution methods | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Data set | Cell population | Abun. LRM | Abun. CM | CTSE LRM | CTSE CM | CTSE measured abun. | SB CTSE LRM (WGCNA) | SB CTSE CM (WGCNA) | SB CTSE measured abun. (WGCNA) | SB CTSE LRM (GSEA) | SB CTSE CM (GSEA) | SB CTSE measured abun. (GSEA) |
| Transformed blood cell data | IM9 | 7.5E-03 | 3.4E-02 | 4.8E-01 | 7.9E-01 | 6.0E-01 | 3.5E-02 | 8.2E-02 | 5.7E-02 | 1.8E-01 | 1.9E-01 | 2.2E-01 |
| | Jurkat | 1.5E-02 | 1.7E-02 | 5.6E-01 | 4.8E-01 | 1.9E+00 | 4.1E-02 | 3.8E-02 | 2.9E-01 | 2.4E-01 | 2.2E-01 | 7.0E-01 |
| | Raji | 3.1E-02 | 2.7E-02 | 7.7E-01 | 1.9E+00 | 3.3E-01 | 6.8E-02 | 2.2E-01 | 3.7E-02 | 2.0E-01 | 4.3E-01 | 1.4E-01 |
| | THP1 | 3.1E-02 | 4.9E-02 | 4.1E-01 | 3.8E-01 | 9.8E-01 | 2.2E-02 | 1.9E-02 | 1.1E-01 | 9.9E-02 | 9.0E-02 | 2.6E-01 |
| Kidney transplantation data | CD14 | 3.0E-02 | 6.0E-02 | 2.6E+00 | 5.3E+00 | 1.5E+00 | 5.3E-01 | 5.8E-01 | 3.2E-01 | 3.5E-01 | 5.8E-01 | 2.0E-01 |
| | CD19 | 6.7E-03 | 2.5E-02 | 2.6E+00 | 6.4E+00 | 2.3E+00 | 4.0E-01 | 8.7E-01 | 3.7E-01 | 3.5E-01 | 5.3E-01 | 3.8E-01 |
| | CD4 | 1.2E-02 | 1.6E-02 | 2.9E+00 | 3.1E-02 | 2.2E+00 | 5.9E-01 | 4.7E-01 | 3.1E-01 | 2.8E-01 | 2.2E-01 | 1.6E-01 |
| | CD8 | 3.0E-02 | 1.2E-02 | 1.8E+00 | 4.8E+00 | 3.1E+00 | 2.4E-01 | 3.3E-01 | 4.9E-01 | 2.6E-01 | 6.3E-01 | 3.8E-01 |
| Rat tissue data | Brain | 2.4E-02 | 1.6E-02 | 3.1E-01 | 3.3E+00 | 2.9E-01 | 5.8E-02 | 2.9E-01 | 4.5E-02 | 1.4E-01 | 1.2E+00 | 1.3E-01 |
| | Liver | 1.2E-02 | 1.7E-02 | 4.5E-01 | 1.5E+00 | 3.9E-01 | 1.3E-01 | 1.4E-01 | 1.6E-01 | 2.7E-01 | 5.1E-01 | 3.1E-01 |
| | Lung | 7.3E-03 | 2.4E-02 | 4.5E-01 | 5.7E-01 | 2.5E-01 | 1.4E-01 | 1.8E-01 | 1.3E-01 | 3.2E-01 | 3.9E-01 | 3.0E-01 |
| Mean MSE | All | 1.9E-02 | 2.7E-02 | 1.2E+00 | 2.3E+00 | 1.3E+00 | 2.0E-01 | 2.9E-01 | 2.1E-01 | 2.4E-01 | 4.5E-01 | 2.9E-01 |

**Additional files:**

The corresponding addtional files for this section can be downloaded from the following linkage:

http://labs.genetics.ucla.edu/horvath/htdocs/CoexpressionNetwork/Bloodbrain/Additional/

## References


1. Jasinska AJ, Service S, Choi OW, DeYoung J, Grujic O, Kong SY, Jorgensen MJ, Bailey J, Breidenthal S, Fairbanks LA *et al*: **Identification of brain transcriptional variation reproduced in peripheral blood: an approach for mapping brain expression traits**. *Human molecular genetics* 2009, **18**(22):4415-4427.
2. Sullivan PF, Fan C, Perou CM: **Evaluating the comparability of gene expression in blood and brain**. *American journal of medical genetics Part B, Neuropsychiatric genetics : the official publication of the International Society of Psychiatric Genetics* 2006, **141B**(3):261-268.
3. Gladkevich A, Kauffman HF, Korf J: **Lymphocytes as a neural probe: potential for studying psychiatric disorders**. *Progress in neuro-psychopharmacology & biological psychiatry* 2004, **28**(3):559-576.
4. Glatt SJ, Everall IP, Kremen WS, Corbeil J, Sasik R, Khanlou N, Han M, Liew CC, Tsuang MT: **Comparative gene expression analysis of blood and brain provides concurrent validation of SELENBP1 up-regulation in schizophrenia**. *Proceedings of the National Academy of Sciences of the United States of America* 2005, **102**(43):15533-15538.
5. Matigian NA, McCurdy RD, Feron F, Perry C, Smith H, Filippich C, McLean D, McGrath J, Mackay-Sim A, Mowry B *et al*: **Fibroblast and lymphoblast gene expression profiles in schizophrenia: are non-neural cells informative?** *PloS one* 2008, **3**(6):e2412.
6. Tsuang MT, Nossova N, Yager T, Tsuang MM, Guo SC, Shyu KG, Glatt SJ, Liew CC: **Assessing the validity of blood-based gene expression profiles for the classification of schizophrenia and bipolar disorder: a preliminary report**. *American journal of medical genetics Part B, Neuropsychiatric genetics : the official publication of the International Society of Psychiatric Genetics* 2005, **133B**(1):1-5.

7. Saris CG, Horvath S, van Vught PW, van Es MA, Blauw HM, Fuller TF, Langfelder P, DeYoung J, Wokke JH, Veldink JH *et al*: **Weighted gene co-expression network analysis of the peripheral blood from Amyotrophic Lateral Sclerosis patients**. *BMC genomics* 2009, **10**:405.

8. Borovecki F, Lovrecic L, Zhou J, Jeong H, Then F, Rosas HD, Hersch SM, Hogarth P, Bouzou B, Jensen RV *et al*: **Genome-wide expression profiling of human blood reveals biomarkers for Huntington's disease**. *Proceedings of the National Academy of Sciences of the United States of America* 2005, **102**(31):11023-11028.

9. Maes OC, Xu S, Yu B, Chertkow HM, Wang E, Schipper HM: **Transcriptional profiling of Alzheimer blood mononuclear cells by microarray**. *Neurobiology of aging* 2007, **28**(12):1795-1809.

10. Oldham MC, Konopka G, Iwamoto K, Langfelder P, Kato T, Horvath S, Geschwind DH: **Functional organization of the transcriptome in human brain**. *Nature neuroscience* 2008, **11**(11):1271-1282.

11. Hodges A, Strand AD, Aragaki AK, Kuhn A, Sengstag T, Hughes G, Elliston LA, Hartog C, Goldstein DR, Thu D *et al*: **Regional and cellular gene expression changes in human Huntington's disease brain**. *Human molecular genetics* 2006, **15**(6):965-977.

12. Iwamoto K, Bundo M, Kato T: **Altered expression of mitochondria-related genes in postmortem brains of patients with bipolar disorder or schizophrenia, as revealed by large-scale DNA microarray analysis**. *Human molecular genetics* 2005, **14**(2):241-253.

13. Ryan MM, Lockstone HE, Huffaker SJ, Wayland MT, Webster MJ, Bahn S: **Gene expression analysis of bipolar disorder reveals downregulation of the ubiquitin cycle and alterations in synaptic genes**. *Molecular psychiatry* 2006, **11**(10):965-978.

14. van der Merwe PA, McNamee PN, Davies EA, Barclay AN, Davis SJ: **Topology of the CD2-CD48 cell-adhesion molecule complex: implications for antigen recognition by T cells**. *Current biology : CB* 1995, **5**(1):74-84.

15. Horvath S, Zhang B, Carlson M, Lu KV, Zhu S, Felciano RM, Laurance MF, Zhao W, Qi S, Chen Z *et al*: **Analysis of oncogenic signaling networks in glioblastoma identifies ASPM as a molecular target**. *Proceedings of the National Academy of Sciences of the United States of America* 2006, **103**(46):17402-17407.

16. Langfelder P, Horvath S: **WGCNA: an R package for weighted correlation network analysis**. *BMC bioinformatics* 2008, **9**:559.

17. Zhang B, Horvath S: **A general framework for weighted gene co-expression network analysis**. *Statistical applications in genetics and molecular biology* 2005, **4**:Article17.

18. Goring HH, Curran JE, Johnson MP, Dyer TD, Charlesworth J, Cole SA, Jowett JB, Abraham LJ, Rainwater DL, Comuzzie AG *et al*: **Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes**. *Nature genetics* 2007, **39**(10):1208-1216.

19. Johnson WE, Li C, Rabinovic A: **Adjusting batch effects in microarray expression data using empirical Bayes methods**. *Biostatistics* 2007, **8**(1):118-127.

20. Langfelder P, Horvath S: **Eigengene networks for studying the relationships between co-expression modules**. *BMC systems biology* 2007, **1**:54.

21. Horvath S, Dong J: **Geometric interpretation of gene coexpression network analysis**. *PLoS computational biology* 2008, **4**(8):e1000117.

22. Torkamani A, Dean B, Schork NJ, Thomas EA: **Coexpression network analysis of neural tissue reveals perturbations in developmental processes in schizophrenia**. *Genome research* 2010, **20**(4):403-412.

23. Davies MN, Lawn S, Whatley S, Fernandes C, Williams RW, Schalkwyk LC: **To What Extent is Blood a Reasonable Surrogate for Brain in Gene Expression Studies: Estimation from Mouse Hippocampus and Spleen**. *Frontiers in neuroscience* 2009, **3**:54.

24. Huang Y, Li H, Hu H, Yan X, Waterman MS, Huang H, Zhou XJ: **Systematic discovery of functional modules and context-specific functional annotation of human genome**. *Bioinformatics* 2007, **23**(13):i222-229.

25. Keller MP, Choi Y, Wang P, Davis DB, Rabaglia ME, Oler AT, Stapleton DS, Argmann C, Schueler KL, Edwards S *et al*: **A gene expression network model of type 2 diabetes links cell cycle regulation in islets with diabetes susceptibility**. *Genome research* 2008, **18**(5):706-716.

26. Stuart JM, Segal E, Koller D, Kim SK: **A gene-coexpression network for global discovery of conserved genetic modules**. *Science* 2003, **302**(5643):249-255.

27. Yip AM, Horvath S: **Gene network interconnectedness and the generalized topological overlap measure**. *BMC bioinformatics* 2007, **8**:22.

28. Langfelder P, Zhang B, Horvath S: **Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R**. *Bioinformatics* 2008, **24**(5):719-720.

29. Luo R, Sanders SJ, Tian Y, Voineagu I, Huang N, Chu SH, Klei L, Cai C, Ou J, Lowe JK *et al*: **Genome-wide transcriptome profiling reveals the functional impact of rare de novo and recurrent CNVs in autism spectrum disorders**. *American journal of human genetics* 2012, **91**(1):38-55.

30. Palmer C, Diehn M, Alizadeh AA, Brown PO: **Cell-type specific gene expression profiles of leukocytes in human peripheral blood**. *BMC genomics* 2006, **7**:115.

31. Abbas AR, Wolslegel K, Seshasayee D, Modrusan Z, Clark HF: **Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus**. *PloS one* 2009, **4**(7):e6098.

32. Cai C, Langfelder P, Fuller TF, Oldham MC, Luo R, van den Berg LH, Ophoff RA, Horvath S: **Is human blood a good surrogate for brain tissue in transcriptional studies?** *BMC genomics* 2010, **11**:589.

33. Chaussabel D, Pascual V, Banchereau J: **Assessing the human immune system through blood transcriptomics**. *BMC biology* 2010, **8**:84.

34. Shen-Orr SS, Tibshirani R, Khatri P, Bodian DL, Staedtler F, Perry NM, Hastie T, Sarwal MM, Davis MM, Butte AJ: **Cell type-specific gene expression differences in complex tissues**. *Nature methods* 2010, **7**(4):287-289.

35. Grigoryev YA, Kurian SM, Avnur Z, Borie D, Deng J, Campbell D, Sung J, Nikolcheva T, Quinn A, Schulman H *et al*: **Deconvoluting post-transplant immunity: cell subset-specific mapping reveals pathways for activation and expansion of memory T, monocytes and B cells**. *PloS one* 2010, **5**(10):e13358.

36. Zhao Y, Simon R: **Gene expression deconvolution in clinical samples**. *Genome medicine* 2010, **2**(12):93.

37. Cobb JP, Mindrinos MN, Miller-Graziano C, Calvano SE, Baker HV, Xiao W, Laudanski K, Brownstein BH, Elson CM, Hayden DL *et al*: **Application of genome-wide expression analysis to human health and disease**. *Proceedings of the National Academy of Sciences of the United States of America* 2005, **102**(13):4801-4806.

38. Whitney AR, Diehn M, Popper SJ, Alizadeh AA, Boldrick JC, Relman DA, Brown PO: **Individuality and variation in gene expression patterns in human blood**. *Proceedings of the National Academy of Sciences of the United States of America* 2003, **100**(4):1896-1901.

39. Herzenberg LA, De Rosa SC: **Monoclonal antibodies and the FACS: complementary tools for immunobiology and medicine**. *Immunology today* 2000, **21**(8):383-390.

40. Herzenberg LA, Tung J, Moore WA, Herzenberg LA, Parks DR: **Interpreting flow cytometry data: a guide for the perplexed**. *Nature immunology* 2006, **7**(7):681-685.

41. Tung JW, Heydari K, Tirouvanziam R, Sahaf B, Parks DR, Herzenberg LA, Herzenberg LA: **Modern flow cytometry: a practical approach**. *Clinics in laboratory medicine* 2007, **27**(3):453-468, v.

42. Emmert-Buck MR, Bonner RF, Smith PD, Chuaqui RF, Zhuang Z, Goldstein SR, Weiss RA, Liotta LA: **Laser capture microdissection**. *Science* 1996, **274**(5289):998-1001.

43. Cahoy JD, Emery B, Kaushal A, Foo LC, Zamanian JL, Christopherson KS, Xing Y, Lubischer JL, Krieg PA, Krupenko SA *et al*: **A transcriptome database for astrocytes, neurons, and oligodendrocytes: a new resource for understanding brain development and function**. *The*

*Journal of neuroscience : the official journal of the Society for Neuroscience* 2008, **28**(1):264-278.

44. Sugino K, Hempel CM, Miller MN, Hattox AM, Shapiro P, Wu C, Huang ZJ, Nelson SB: **Molecular taxonomy of major neuronal classes in the adult mouse forebrain**. *Nature neuroscience* 2006, **9**(1):99-107.

45. Torres-Munoz JE, Van Waveren C, Keegan MG, Bookman RJ, Petito CK: **Gene expression profiles in microdissected neurons from human hippocampal subregions**. *Brain research Molecular brain research* 2004, **127**(1-2):105-114.

46. Venet D, Pecasse F, Maenhaut C, Bersini H: **Separation of samples into their constituents using gene expression data**. *Bioinformatics* 2001, **17 Suppl 1**:S279-287.

47. Lu P, Nakorchevskiy A, Marcotte EM: **Expression deconvolution: a reinterpretation of DNA microarray data reveals dynamic changes in cell populations**. *Proceedings of the National Academy of Sciences of the United States of America* 2003, **100**(18):10370-10375.

48. Bar-Joseph Z, Farkash S, Gifford DK, Simon I, Rosenfeld R: **Deconvolving cell cycle expression data with complementary information**. *Bioinformatics* 2004, **20 Suppl 1**:i23-30.

49. Miller JA, Cai C, Langfelder P, Geschwind DH, Kurian SM, Salomon DR, Horvath S: **Strategies for aggregating gene expression data: the collapseRows R function**. *BMC bioinformatics* 2011, **12**:322.

50. Gong T, Hartmann N, Kohane IS, Brinkmann V, Staedtler F, Letzkus M, Bongiovanni S, Szustakowski JD: **Optimal deconvolution of transcriptional profiling data using quadratic programming with application to complex clinical blood samples**. *PloS one* 2011, **6**(11):e27156.

51. Kuhn A, Thu D, Waldvogel HJ, Faull RL, Luthi-Carter R: **Population-specific expression analysis (PSEA) reveals molecular changes in diseased brain**. *Nature methods* 2011, **8**(11):945-947.

52. Lahdesmaki H, Shmulevich L, Dunmire V, Yli-Harja O, Zhang W: **In silico microdissection of microarray data from heterogeneous cell populations**. *BMC bioinformatics* 2005, **6**:54.

53. Repsilber D, Kern S, Telaar A, Walzl G, Black GF, Selbig J, Parida SK, Kaufmann SH, Jacobsen M: **Biomarker discovery in heterogeneous tissue samples -taking the in-silico deconfounding approach**. *BMC bioinformatics* 2010, **11**:27.

54. Wang M, Master SR, Chodosh LA: **Computational expression deconvolution in a complex mammalian organ**. *BMC bioinformatics* 2006, **7**:328.

55. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES *et al*: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles**. *Proceedings of the National Academy of Sciences of the United States of America* 2005, **102**(43):15545-15550.

56. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT *et al*: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium**. *Nature genetics* 2000, **25**(1):25-29.