

## ABSTRACT

Title of dissertation: MULTIVARIATE METHODS FOR HIGH-  
THROUGHPUT BIOLOGICAL DATA WITH  
APPLICATION TO COMPARATIVE GENOMICS

Chiao-wen Joyce Hsiao, Doctor of Philosophy, 2015

Directed By: Dr. Héctor Corrada Bravo  
Department of Computer Science

Phenotypic variation in multi-cellular organisms arises as a result complex gene regulation mechanisms. Modern development of high-throughput technology opens up the possibility of genome-wide interrogation of aspects of these mechanisms across molecular phenotypes. Multivariate statistical methods provide convenient frameworks for modeling and analyzing data obtained from high-throughput experiments probing these complex aspects. This dissertation presents multivariate statistical methods to analyze data arising from two specific high-throughput molecular assays: (1) ribosome footprint profiling experiments, and (2) flow cytometry data.

Ribosome footprint profiling describes an in vivo translation profile in a living cell and offers insights into the process of post-transcriptional gene regulation. Translation efficiency (TE) is a measure that quantifies the rate at which active translation is occurring for each gene – defined as the ratio of ribosome protected fragment count to mRNA fragment count. We introduce pairedSeq, an empirical covariance shrinkage method for differential testing of translation efficiency from sequencing data. The method draws on variance decomposition techniques in mixed-effect modeling and analysis of variance. Benchmark tests comparing to the existing methods reveals that pairedSeq effectively detects signals in genes with high variation in expression measurements across samples due to high co-variability between ribosome occupancy and transcript abundance. In contrast, existing methods tend to mistake genes with negative co-variability as signals, as a result of variance underestimation when not accounting for negative co-variability. We then present a genome-wide survey of primate species divergence at the translational and post-translational layer of gene regulation.

FCM is routinely employed to characterize cellular characteristics such as mRNA and protein expression at the single-cell level. While many computational methods have been developed that focus on identifying cell populations in

individual FCM samples, very few have addressed how the identified cell populations can be matched across samples for comparative analysis. FlowMap-FR can be used to quantify the similarity between cell populations under scenarios of proportion differences and modest position shifts, and to identify situations in which inappropriate splitting or merging of cell populations has occurred during gating procedures. It has been implemented as a stand-alone R/Bioconductor package easily incorporated into current FCM data analytical workflows.

MULTIVARIATE METHODS FOR HIGH-THROUGHPUT BIOLOGICAL  
DATA WITH APPLICATION TO COMPARATIVE GENOMICS

By

Chiao-wen Joyce Hsiao

Dissertation submitted to the Faculty of the Graduate School of the  
University of Maryland, College Park, in partial fulfillment  
of the requirements for the degree of  
Doctor of Philosophy  
2015

Advisory Committee:

Professor Héctor Corrada Bravo, Chair

Dr. Zia Khan

Dr. Steve Mount

Dr. Larry Washington

Dr. Mei-Ting Lee, Dean's representative

© Copyright by  
Chiao-wen Joyce Hsiao  
2015

Dedication

*To my father 蕭毓欽*

## Acknowledgements

I am grateful for the many colleagues and collaborators who have made this thesis possible and supported me in reaching this point of my career.

First and foremost, I am grateful for my advisor Héctor Corrada Bravo who brought me into the field of bioinformatics and computational biology and shared his experiences and insights with me in the course of this thesis. I am also grateful for Paul Smith who not only served as an invaluable statistics mentor, but also supported my transition from a career in clinical psychology to statistics. I would also like to thank my collaborators for their intellectual and moral support in the research that led to this thesis: Yu Max Qian and Richard Scheuermann at the J. Craig Venter Institute, Zia Khan at the University of Maryland, and Sidney Wang at the University of Chicago.

Friends and family have also provided invaluable intellectual and moral support along the way. I am thankful for all of my colleagues at the Center for Bioinformatics and Computational Biology who have made science fun and exciting. Special thanks to Kwame Okrah, Joseph Paulson, and Hisham Talukder for making the lab a fun place to work 24-7, and for the tremendous support you have given me throughout my graduate career. I would like to extend my appreciation to my friends in the mathematics department who have made

Graduate Hills a fun place to live: Jeff Frazier, Eddie Kim, Cameron Mozafari, Edward Phillips, Sean Rostami, Domingo Ruiz, and Christian Sykes. I would also like to thank my dear girlfriends Anna Ghambaryan and Sayomi for their unconditional love and support, despite my having to work all the time. I am grateful for my father Yu-Chin Hsiao and mother Chuen-mei Chien along with the rest of my family for their continued support and tolerance of my frequent absence in the past few years. Finally, my partner in crime, Andrew Michael Sanders, for whom I am grateful for making sure I have a life outside of this thesis.

## Preface

Below is a summary of the collaborative efforts in each main chapter of this thesis.

Chapter 2 presents a statistical framework for detecting differential translation efficiency. The idea grew out of my work on comparative genomics analysis between primates with Sidney Wang at University of Chicago and Zia Khan at the University of Maryland.

Chapter 3 resulted from a collaboration with the Gilad lab at the Department of Human Genetics in the University of Chicago. The idea of investigating the translational landscape in primates was initially proposed by Sidney Wang, who also designed the Ribosome Profiling experiment. My contribution was statistical analysis presented in this paper. I would like thank Sidney Wang for invaluable discussion on comparative genomics analysis.

Chapter 4 resulted from a three months' visit to J. Craig Venter Institute in Summer 2013. Menya Liu and Monnie McGee at the Southern Methodist University proposed the idea of applying the Friedman-Rafsky (FR) test to mapping cell populations in flow cytometry data. Yu Qian and Richard Scheuermann proposed the use of the FR test under scenarios of biological and

technical variations between flow cytometry samples. My contribution was designing an efficient downsampling scheme, implementing the method in an R package (FlowMap-FR, in Bioconductor since 2013), and evaluating the performance of the method.

## Table of Contents

List of Figures .....	ix
1 Introduction.....	1
2 Empirical Bayes analysis of intra-sample variability improves power in differential analysis of translation efficiency .....	4
2.1 Background .....	4
2.2 Methods.....	7
2.2.1 Variance-mean dependency .....	8
2.2.2 Model.....	9
2.2.3 Empirical Bayes shrinkage for intra-sample variability.....	11
2.2.4 Comparative benchmarks .....	20
2.3 Results and discussion .....	24
2.3.1 Variance and mean dependency .....	24
2.3.2 Co-variability between ribosome occupancy and transcript abundance .....	25
2.3.3 Shrunk covariance estimates and differential translation efficiency .....	27
2.4 Conclusion .....	30
2.5 Figures.....	32
3 Comparative genomic study between primates.....	44
3.1 Abstract.....	44
3.2 Background .....	45
3.3 Methods.....	48
3.3.1 Ribosome footprint profiling.....	48
3.3.2 Statistical analysis.....	50
3.3.4 Functional enrichment analysis.....	51
3.4 Results.....	52

3.4.1	Ribosome profiling captures variations in protein translation among primates.....	52
3.4.2	Variations in transcription across primates are mostly propagated to translation .....	54
3.4.3	Differences between RNA and protein level divergence across primates can partly be explained by translational regulation.....	57
3.4.4	Variations in translation across primates are frequently attenuated in protein levels.....	61
3.4.5	Post-translational buffered genes are under stabilizing selection in protein level .....	64
3.6	Discussion .....	67
3.7	Figures.....	73
4	Mapping cell populations in flow cytometry data for cross-sample comparison using the Friedman-Rafsky test statistic as a distance measure.....	80
4.1	Abstract.....	80
4.2	Background .....	82
4.3	Methods.....	88
4.3.1	Terminology .....	88
4.3.2	Overview of FlowMap-FR.....	89
4.3.3	Simulation study.....	97
4.3.4	Real flow cytometry samples .....	101
4.4	Results.....	104
4.4.1	Simulation study .....	104
4.4.2	Mapping cell populations across multiple samples in real data... ..	114
4.5	Discussion.....	117
4.6	Figures.....	124
5	Discussion .....	143
	Bibliography.....	145

## List of Figures

Figure 1. Gene count distributions vary between ribosome occupancy data and RNA-seq data.....	32
Figure 2. Variance-mean dependency in 3 real data sets. ....	33
Figure 3. Correlation between ribosome occupancy and transcript abundance.....	35
Figure 4. Per gene correlation between ribosome occupancy and transcript abundance across species.....	36
Figure 5. Variance and mean trend among genes with high and with low correlation between ribosome occupancy and transcript abundance.....	37
Figure 6. Species difference in translation efficiency is larger for genes with positive correlation between ribosome occupancy and transcript abundance.....	38
Figure 7. pairedSeq shrinks both positive and negative correlations between ribosome occupancy and transcript abundance.....	39
Figure 8. pairedSeq detects significant fold change despite high intra-sample variability.....	40
Figure 9. Benchmark sensitivity in pairedSeq, pairedSeq-r, DESeq2, voom, and LMM at FDR < .01.....	41
Figure 10. Benchmark precision in pairedSeq, pairedSeq-r, DESeq2, voom, and LMM at FDR < .01.....	42
Figure 11. Benchmark false discovery rate in pairedSeq, pairedSeq-r, DESeq2, voom, and LMM. ....	43
Figure 12. Ribosome profiling captures species differences in protein translation levels. ....	73
Figure 13. Translation propagates majority of transcriptional divergence.....	75
Figure 14. Translational regulations contribute to differences in human-chimpanzee divergence between protein and mRNA.....	76
Figure 15. Translational divergence leads to stable protein level and divergent protein turnover rate.....	77
Figure 16. Post-translational gene expression buffering is prominent. ....	78
Figure 17. Post-translational gene expression buffering is conserved in primates. ...	79

Figure 18. Multivariate run calculation.....	125
Figure 19. Data simulations and test scenarios.....	127
Figure 20. Matching cell populations that differ in proportions between samples.	129
Figure 21. Matching cell populations with shifted marker distributions between samples.....	132
Figure 22. Matching cell populations inappropriately divided into two populations in one sample. ....	135
Figure 23. Matching cell populations using SKL divergence measure.....	137
Figure 24. Matching cell populations across the real FCM data set #1. ....	140
Figure 25. Matching cell populations across the real FCM data set #2. ....	141

## 1 Introduction

Phenotypic variation in multi-cellular organisms arises as a result of complex gene regulation mechanisms. Modern development of high-throughput technology opens up the possibility of genome-wide interrogation of aspects of these mechanisms across molecular phenotypes. A wide array of high-throughput biological assays are now available for characterizing DNA structure (e.g., DNA-seq and transposon sequencing <sup>1,2</sup>), DNA-protein interactions (e.g., ChIP-seq <sup>3</sup> and DNase-seq <sup>4,5</sup>), global transcriptome profile (RNA-seq <sup>6</sup>), and more recently, genome wide translational level of gene expression control (Ribosome footprint profiling <sup>7</sup>). Joint collection and analysis of genetic materials can provide valuable insights into regulatory patterns that shape phenotype variations <sup>8-12</sup>.

Gene expression profiles are central to understanding protein expression patterns that underlie phenotype variation <sup>13-17</sup>. When a functional group of genes is being expressed in the cell, a chain of reactions is activated to translate encoding genetic materials in the DNAs to instructions directing protein synthesis in the RNAs, and to modulate protein functions. This chain of gene regulatory actions can be summarized in three processes: transcriptional and translational control of gene expression and post-translational regulation of protein expression.

Much of the literature focuses on transcriptional profiling and transcription binding factors and on protein expression patterns. We now know transcriptional gene expression correlates poorly with protein expression patterns and much work is needed to investigate post-transcriptional control of gene expression. For example, synonymous mutation is a genetic variant that alters one or more nucleotide bases in an encoding gene but does not change the produced amino acid sequences. We now know synonymous mutations are instrumental in shaping phenotypic variation along with non-synonymous mutations and in modulating protein abundance variations via their control of translation initiation and codon usage bias.

Ribosome footprint profiling is a high-throughput biological assay that characterizes the *in vivo* translational profile of a living cell. The joint collection of ribosome footprints and transcriptional expression profile can shed light on the post-transcriptional control of gene expression. In addition, in conjunction with protein expression data, we can begin to investigate in high-resolution the post-translational modification of amino acid sequences. Multivariate statistical methods provide convenient frameworks for modeling and analyzing data obtained from high-throughput experiments probing these complex aspects. This dissertation presents multivariate statistical methods to analyze data arising from two specific

high-throughput molecular assays: (1) ribosome footprint profiling experiments, and (2) flow cytometry data.

This thesis begins with a conceptual framework for empirical Bayes methods for correlated data. Chapter 2 presents a statistical method developed to identify differences in translation efficiency between biological conditions. Chapter 3 investigates divergence between primate species using data collected at transcriptional, translational and protein synthesis layers. Chapter 4 presents a statistical method that is developed to compare cell populations homogeneous in protein expression profiles.

## 2 Empirical Bayes analysis of intra-sample variability improves power in differential analysis of translation efficiency

### 2.1 Background

Ribosome footprint profiling experiments allow for the first time a high-resolution survey of the sites on mRNA molecules undergoing active translation <sup>7</sup>. Ribosome footprinting has been adopted to survey translational landscapes <sup>18</sup>, to delineate mechanisms underlying translation initiation and elongation <sup>19,20</sup>, and to study evolutionary pressure at translational and post-translational regulation <sup>21</sup>. Initial steps of ribosome footprinting involve capturing and isolating ribosome-protected mRNA molecules from the mRNA molecules not protected by ribosomes. Ribosomal RNAs are then immobilized and depleted from the ribosome-footprint complexes. Each ribosome protected fragment is about 28 to 30 nucleotide bases. These ribosome protected fragments are size-selected and ligated using the same procedure in a typical RNA sequencing experiment. Sequencing reads generated in a ribosome footprint profiling experiment are mapped onto the genome using alignment tools that are sensitive to spliced mRNA fragments <sup>22</sup>.

Ribosome occupancy data, as measured by the number of ribosome protected fragment reads, is count based and can be analyzed with statistical tools

developed for other sequencing data. Translation efficiency (TE) is a measure that quantifies the rate at which active translation is occurring for each gene. It is defined as the ratio of ribosome protected fragment count to mRNA fragment count. Because the quantity of ribosome protected fragments is dependent on the transcriptional expression level of the gene, an overall high variation in expression measurements across ribosome footprofiling and RNA-seq data is expected when positive co-variation between ribosome occupancy data and transcript abundance is present. The unique feature of correlated count data sets the analysis of translation efficiency apart from standard RNA-seq analysis.

Few methods are available for differential analysis of translation efficiency<sup>23,24</sup>. Olshen et al. (2013) propose a count based statistical framework that proceeds in two steps. First, per gene per sample measurements are analyzed to quantify the degree to which ribosome occupancy level differs from what would be expected given its transcript abundance. An errors-in-variables regression model is performed on ribosome occupancy measurements using corresponding sample transcript abundance as a covariate. The P-value associated with the transcript abundance covariate is used to quantify the magnitude of translational regulation. Next, P-values within each biological condition are transformed to Z-scores. These Z-scores are then compared between conditions to determine significance of condition

differences in translational regulation. Zhong et al. <sup>24</sup> applies the DESeq2 framework and compares translation efficiency using an interaction design. The design matrix has four Boolean-valued columns: a column indicating data type (ribosome footprinting or RNA-seq), a column corresponding to biological condition (1 or 2), a column for interaction (data type by condition), and finally, a column of 1's describing the model intercept. This model assumes the independence of ribosome protected fragment count and transcript abundance.

We describe a conceptual framework for the analysis of translation efficiency, using an established approach in the analysis of DNA microarray where probe intensities are modeled on a log scale. This approach can be easily incorporated into complex experimental designs such as split-plot and time series experiments. Recently, Law et al. (2014) presented *voom*, a statistical framework for RNA-seq data in which the standard deviation of log<sub>2</sub> counts per million (CPM) is proved to be mathematically equivalent to the dispersion parameter in a negative binomial distribution, given large library size <sup>25</sup>. The *voom* algorithm computes an observational-level precision weight to adjust for heterogeneity of variances commonly observed in RNA-seq data.

Small sample size in sequencing experiments is a core challenge in statistical analysis of RNA-seq data. Empirical Bayes shrinkage is routinely applied to

stabilize variance estimates across genes. In differential analysis of translation efficiency, we are also concerned with co-variability between sample measurements obtained from ribosome profiling experiments and RNA-seq experiments. We introduce pairedSeq, an empirical covariance shrinkage method for differential testing of translation efficiency from sequencing data.

## 2.2 Methods

In a typical experiment designed to compare translation efficiency between biological conditions, both ribosome occupancy and mRNA transcript abundance are collected for  $n$  cell lines. A total of  $2n$  samples are prepared for each experiment. Per gene log2-cpm values  $(\mathbf{Y}_{g,i}, \mathbf{Z}_{g,i})'$  are computed for ribosome occupancy and transcript abundance, respectively, in cell line  $i$ . The  $n$  pairs of samples are assumed to be independent and identically distributed and follow a multivariate normal distribution with mean  $\boldsymbol{\mu}_g$  and covariance matrix  $\boldsymbol{\Sigma}_g$ , denoted by  $N_2(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$ .

The main interest in this paper is to describe and model  $\boldsymbol{\Sigma}_g$ . Later on, we describe various covariance structures that are evaluated in this paper. For now, the covariance matrix of  $(\mathbf{Y}_{g,i}, \mathbf{Z}_{g,i})'$  is described as unstructured:

$$\text{Cov}(\mathbf{Y}_g, \mathbf{Z}_g) = \boldsymbol{\Sigma}_g \otimes \mathbf{I}_n = \begin{pmatrix} \tau_{g,y}^2 & \tau_{g,b}^2 \\ \tau_{g,b}^2 & \tau_{g,z}^2 \end{pmatrix} \otimes \mathbf{I}_n,$$

where  $\tau_{g,y}^2$  and  $\tau_{g,z}^2$  denote the variability within each data type across samples in log2-cpm values of ribosome occupancy and transcript abundance, respectively, and  $\tau_{g,b}^2$  represents intra-sample variability in gene  $g$  across  $n$  cell lines – covariation between ribosome occupancy and transcript abundance for each gene. The quantities  $\mathbf{Y}_g$  and  $\mathbf{Z}_g$  denote gene  $g$ 's vector of log2-cpm values of ribosome occupancy and transcript abundances, respectively, and are match-ordered according to sample's cell line membership,

$$\mathbf{Y}_g = \begin{bmatrix} Y_{g,1} \\ \vdots \\ Y_{g,n} \end{bmatrix}, \text{ and } \mathbf{Z}_g = \begin{bmatrix} Z_{g,1} \\ \vdots \\ Z_{g,n} \end{bmatrix}.$$

### 2.2.1 Variance-mean dependency

We adopt *voom*, a log2-cpm based approach for variance-mean dependency across ribosome occupancy data and RNA-seq data <sup>25</sup>. A nonparametric function of per gene log2-cpm values of expression is fitted across ribosome occupancy and transcript abundance measurements across genes to describe the per gene observed squared standard deviation of the log2-cpm measurements. The fitted nonparametric trend measures the degree of variance-mean dependency across data types. The nonparametric fit computes predicted variances of expression values at

the observational level, which represents the magnitude of variance-mean dependency in the data. The inverse of the predicted variances are used as observation weights to remove the variance-mean dependency for all subsequent analyses. The *voom* precision weights are proven to downweight inflated variances among weakly expressed genes in data sets of small to large numbers of biological replicates. See Law et al. <sup>25</sup> for a comparison of *voom* with the count-based methods.

Denote  $\mathbf{W}_g = \text{diag}(w_{g,1}, \dots, w_{g,N})$  as gene  $g$ 's *voom* precision weights, for a total of  $N = 2n$  samples of expression values. The unstructured covariance matrix

$\Sigma_g$  is scaled by  $\mathbf{W}_g$  and converted to a weighted covariance matrix,

$$\Delta_g = \mathbf{W}_g \Sigma_g \mathbf{W}_g' = \begin{pmatrix} \sigma_g^2 & \sigma_{g,b}^2 \\ \sigma_{g,b}^2 & \sigma_g^2 \end{pmatrix} \otimes \mathbf{I}_n,$$

where we assume homogeneity of variances across data types.

### 2.2.2 Model

The log2-cpm values of ribosome occupancy and transcript abundance are described in a linear model,

$$\begin{pmatrix} \mathbf{Y}_g \\ \mathbf{Z}_g \end{pmatrix} = \mathbf{X}\boldsymbol{\beta}_g + \begin{pmatrix} \boldsymbol{\eta}_{g,y} \\ \boldsymbol{\eta}_{g,z} \end{pmatrix},$$

Error! Reference source not

found.

where  $\mathbf{X}\boldsymbol{\beta}_g$  models the fold change in translation efficiency between biological conditions,  $\mathbb{E}(\mathbf{Y}'_g, \mathbf{Z}'_g)' = \mathbf{X}\boldsymbol{\beta}_g$ , and  $\mathbf{X}$  represents a  $2n \times 4$  interaction design matrix that indicates sample membership in each data type and each biological condition under study. For example, suppose there are two biological conditions in the comparison. The columns of  $\mathbf{X}$  includes a main effect column of condition (1 = Condition 1, 0 = Condition 2), a main effect column of data type (1 = ribosome occupancy, 0 = RNA-seq), and an interaction effect column (condition column multiplied by the data type column componentwise), with a column of 1's for the intercept.  $\boldsymbol{\eta}'_{g,y}$  and  $\boldsymbol{\eta}'_{g,z}$  represent the residuals in ribosome occupancy and transcript abundance after weighting the data for variance-mean dependency, and

$$\text{Cov}(\boldsymbol{\eta}_{g,y}, \boldsymbol{\eta}_{g,z}) = \Delta_g = \begin{pmatrix} \sigma_g^2 & \sigma_{g,b}^2 \\ \sigma_{g,b}^2 & \sigma_g^2 \end{pmatrix} \otimes \mathbf{I}_n.$$

**Error! Reference source not found.**

The coefficient vector  $\boldsymbol{\beta}_g$  consists of four effects corresponding to the columns in the interaction design matrix. Each of the coefficients is estimated by a linear combination of four average sample expression values of samples: Ribosome occupancy in condition 1 and 2, and RNA-seq in condition 1 and 2. Let  $\bar{\mathbf{Y}}_{g,l}$  and  $\bar{\mathbf{Z}}_{g,l}$  denote gene  $g$ 's average log2-cpm values of ribosome occupancy and transcript

abundance in condition  $l$ . We let  $\hat{\beta}_g^{TE}$  as the fold change in translation efficiency and compute

$$\hat{\beta}_g^{TE} = \bar{\mathbf{Y}}_{g,l} - \bar{\mathbf{Z}}_{g,l} - (\bar{\mathbf{Y}}_{g,k} - \bar{\mathbf{Z}}_{g,k}),$$

between condition  $k$  and  $l$ . Following the linearity of  $\hat{\beta}_g^{TE}$ , the variance of the fold change estimate is computed a weighted sum of the two sources of within-data-type variability and intra-sample variability,

$$\text{Var}(\hat{\beta}_g^{TE}) = 2 \left\{ (\sigma_g^2 - \sigma_{g,b}^2) \left( \frac{1}{\mathbf{n}_k} + \frac{1}{\mathbf{n}_l} \right) \right\},$$

where  $\mathbf{n}_k$  and  $\mathbf{n}_l$  are the number of cell lines collected for condition  $k$  and  $l$ , respectively.

### 2.2.3 Empirical Bayes shrinkage for intra-sample variability

We draw on variance decomposition techniques in the statistical literature of mixed effect modeling and analysis of variance and suggest an empirical shrinkage method that can work with both negatively and positively valued intra-sample co-variability. Briefly, the intra-sample variability or co-variability between log2-cpm values of ribosome occupancy and transcript abundance is transformed to a linear combination of independent random variables following chi-squared distributions. The conjugacy of the chi-square distribution with the inverse-gamma

distribution offers closed form solutions for the prior and posterior estimates. The posterior estimates of the independent chi-squared random variables are inverse-transformed to obtain a shrunken estimate of intra-sample variability. This section is divided into two parts. First, we describe variance decomposition techniques borrowed from mixed-effect modeling and restate our covariance matrix accordingly. Then, the linear transformation for obtaining independent variance components is delineated, followed by the covariance shrinkage method.

*Variance decomposition using mixed effect modeling*

Under the nested design of ribosome occupancy experiment, each gene’s variability in expression levels across samples are of varying magnitude from variability across cell lines or random perturbation in expression measurements. We explicitly model cell line variability as a random effect and restate the normal linear model in **Error! Reference source not found.** as a mixed-effect model,

$$\begin{pmatrix} \mathbf{Y}_g \\ \mathbf{Z}_g \end{pmatrix} = \mathbf{X}\boldsymbol{\beta}_g + \mathbf{U}\boldsymbol{\gamma}_g + \boldsymbol{\epsilon}_g, \quad \text{Error! Reference source not found.}$$

where  $\mathbf{U}$  is an  $N \times n$  design matrix consisting of Boolean-valued vectors that

indicates each sample’s cell line membership,  $\boldsymbol{\gamma}_g$  is a length- $n$  coefficient vector that correspond to the random effect design matrix  $\mathbf{U}$ , and  $\boldsymbol{\epsilon}_g$  contains residuals

estimating random perturbation in expression measurements across samples that is

independent of cell line variability. We assume that  $\boldsymbol{\gamma}_g$  and  $\boldsymbol{\epsilon}_g$  are independent and distributed as  $N(0, \sigma_{g,b}^2 \mathbf{I}_n)$  and  $N(0, \sigma_{g,e}^2 \mathbf{I}_N)$ , respectively. Thus, the covariance matrix of  $\boldsymbol{\epsilon}_g$ ,

$$\text{Cov}[\boldsymbol{\epsilon}_g, \boldsymbol{\epsilon}_g] = \sigma_{g,b}^2 \mathbf{U}\mathbf{U}' + \sigma_{g,e}^2 \mathbf{I}_N,$$

is equal to the weighted covariance matrix in **Error! Reference source not found.**. The equivalence between these two covariance matrices

$$\Delta_g = \begin{pmatrix} \sigma_g^2 & \sigma_{g,b}^2 \\ \sigma_{g,b}^2 & \sigma_g^2 \end{pmatrix} \otimes \mathbf{I}_n + \begin{pmatrix} \sigma_{g,e}^2 + \sigma_{g,b}^2 & 0 \\ 0 & \sigma_{g,e}^2 + \sigma_{g,b}^2 \end{pmatrix} \otimes \mathbf{I}_n$$

states that the explicit modeling of variability across cell lines leads to the

decomposition of within-data-type variability  $\sigma_g^2$  into variability generated from (1)

$\sigma_{g,b}^2$  the variability across cell lines between log2-cpm values of ribosome occupancy

and transcript abundance, or equivalently, the intra-sample variability or co-

variability between the two data types, and (2)  $\sigma_{g,e}^2$  the variability across sample

measurements that estimates random perturbation in expression measurements, or

equivalently, within-data-type expression variability.

Our main objective in applying mixed-effect modeling is to obtain a representation of covariance matrix that delineates variation from intra-sample variability and random perturbation in expression measurements. Therefore, we

require the covariance matrix  $\Delta_g = \sigma_{g,b}^2 \mathbf{U}\mathbf{U}' + \sigma_{g,e}^2 \mathbf{I}_N$  to be positive semidefinite, rather than putting constraints on the individual variance components.

### *Empirical Bayes shrinkage in Analysis of variance components*

Analysis of variance (ANOVA) is a statistical tool for complex data structures such as nested designs that decomposes observation into orthogonal components and offers empirical estimates of variance components with easy-to-work-with distributional properties. Under the assumption of normality and balance of the data structure, they are independent and follow scaled chi-squared distributions <sup>26</sup>. The conjugate prior of scaled chi-squared distribution is inverse-gamma, which offers closed form solution of posterior estimate. Most importantly, ANOVA estimation permits negatively valued shrunken intra-sample covariance estimates.

Typical Bayesian analysis of variance components  $\sigma_{g,b}^2$  and  $\sigma_{g,e}^2$  assumes conjugate prior of inverse-gamma, which constrains the posterior parameter space to be strictly non-negative <sup>27</sup>. Our approach puts the parameter space  $\sigma_{g,b}^2$  on a real line. This section describes a linear transformation of variance components to ANOVA components and the covariance shrinkage method <sup>28,29</sup>.

From the analysis of variance perspective, the mixed-effect model in **Error! eference source not found.** gives rise to two sources of variation that have a

one-to-one relationship with the variance components  $\sigma_{g,b}^2$  and  $\sigma_{g,e}^2$ . The total intra-sample sum-of-squares (SSB) computes the sum of squared deviations of expression measurements in each cell line; this is a function of co-variability of expression values between data types and also random perturbation of measurements within each data type. The within-sample sum-of-squares (SSW) computes the sum of squared deviation of each sample's expression value from the mean gene expression level. ANOVA components are the averaged sum-of-squares divided by corresponding degrees of freedom. Let  $AC_{g,b}$  denote the per gene ANOVA component corresponding to SSB. Then  $AC_{g,b} = SSB/df_{SSB}$ , where degrees of freedom of intra-sample sum-of-squares is the number of cell lines minus 1 ( $n - 1$ ). Let  $AC_{g,e}$  denote the per gene ANOVA component corresponding to SSW. Then  $AC_{g,e} = SSW/df_{SSW}$ , where degrees of freedom of within-sample sum-of-squares is the total number of samples minus the number of cell lines ( $N - n$ ).

Each of the ANOVA components follows a chi-squared distribution

$$AC_{g,b} \sim \frac{2\sigma_{g,b}^2 + \sigma_{g,e}^2}{n - 1} \chi_{n-1}^2,$$

$$AC_{g,e} \sim \frac{\sigma_{g,e}^2}{n} \chi_n^2,$$

where

$$\mathbb{E}AC_{g,b} = 2\sigma_{g,b}^2 + \sigma_{g,e}^2 \quad \text{and} \quad \mathbb{E}AC_{g,e} = \sigma_{g,e}^2.$$

Given the above relationship, we can estimate the per gene ANOVA components as

$$\widehat{AC}_{g,b} = 2\widehat{\sigma}_{g,b}^2 + \widehat{\sigma}_{g,e}^2 \quad \text{and} \quad \widehat{AC}_{g,e} = \widehat{\sigma}_{g,e}^2, \quad \text{Error! Reference source not found.}$$

where  $\widehat{\sigma}_{g,b}^2$  and  $\widehat{\sigma}_{g,e}^2$  denote the estimated variance components based on observed data, and  $\widehat{AC}_{g,b}$  and  $\widehat{AC}_{g,e}$  denote the corresponding sample estimates of ANOVA components. Conversely, suppose we only have information on the ANOVA components. Then, we can compute

$$\widehat{\sigma}_{g,b}^2 = (\widehat{AC}_{g,b} - \widehat{AC}_{g,e})/2 \quad \text{and} \quad \widehat{\sigma}_{g,e}^2 = \widehat{AC}_{g,e}. \quad \text{Error! Reference source not found.}$$

Shrunken estimates of  $\sigma_{g,b}^2$  and  $\sigma_{g,e}^2$  are computed in three steps. First, we estimate gene-wise component estimates using restricted maximum likelihood estimation method (REML). The variance component REML estimates are then transformed to ANOVA components<sup>28</sup>. Second, we estimate prior parameters of the likelihood ANOVA components following the chi-squared-and-inverse-gamma family conjugacy. Finally, we combine the likelihood and the prior density to obtain posterior estimates of the ANOVA components. The posterior ANOVA

estimates are then inverse-transformed to obtain our posterior estimates of variance components.

*Gene-wise ANOVA component estimates.* We start with the model in **Error!**

**reference source not found.** and compute the per gene REML estimates of

variance components, denoted as  $\hat{\sigma}_{g,b}^2$  and  $\hat{\sigma}_{g,e}^2$ . The weighted variance-covariance

matrix  $\mathbf{\Delta}_g$  is required to be positive semi-definite.  $\hat{\sigma}_{g,b}^2$  and  $\hat{\sigma}_{g,e}^2$  are the converted

to the per gene ANOVA estimates  $\hat{\mathbf{AC}}_{g,b}$  and  $\hat{\mathbf{AC}}_{g,e}$  using **Error! Reference**

**source not found.**

*ANOVA prior estimates.* Recall that each ANOVA component follows a chi-

squared distribution, and denote the parameter of  $\mathbf{AC}_{g,b}$  as  $\sigma_{g,AC_b}^2 = 2\sigma_{g,b}^2 + \sigma_{g,e}^2$

and the parameter of  $\mathbf{AC}_{g,e}$  as  $\sigma_{g,AC_e}^2 = \sigma_{g,e}^2$ . Then,  $\sigma_{g,AC_b}^2$  and  $\sigma_{g,AC_e}^2$  each follows

a scaled-inverse-chi-squared distribution

$$\frac{1}{\sigma_{g,AC_b}^2} \sim \frac{1}{d_{0,b} \mathbf{AC}_{0,b}} \chi_{d_{0,b}}^2,$$

$$\frac{1}{\sigma_{g,AC_e}^2} \sim \frac{1}{d_{0,e} \mathbf{AC}_{0,e}} \chi_{d_{0,e}}^2,$$

where  $\mathbf{AC}_{0,b}$  and  $\mathbf{AC}_{0,e}$  are prior parameter for the gene-wise ANOVA components,

and  $d_{0,b}$  and  $d_{0,e}$  are their corresponding degrees of freedom.

Prior parameters are estimated using a computational efficiency approach that is implemented in Smyth's *limma* algorithm<sup>30</sup>. Briefly, take  $AC_{g,e}$  as an example. The ANOVA components are log2 transformed, and the two equations are set up equating log2-ANOVA sample moments with population moments

$$\mathbb{E}[\log AC_{g,e}] = \log AC_{0,e} + \psi(d_{g,e}/2) - \psi(d_{0,e}/2) + \log(d_{0,e}/d_{g,e})$$

$$\text{Var}[\log AC_{g,e}] = \psi'(d_{g,e}/2) + \psi'(d_{0,e}/2)$$

where  $\psi(\cdot)$  and  $\psi'(\cdot)$  are the digamma and trigamma function, respectively. The hyperparameters are computed in two steps. First, using delta method, we approximate the value of  $d_{0,e}$ . Given the estimate for  $d_{0,e}$ , we obtain the estimate for  $AC_{0,e}$ . Details of the computational steps for each estimator can be found in Smyth's *limma* paper.<sup>30</sup>

*ANOVA shrunken estimates.* Posterior densities of the ANOVA components distribute as scaled-inverse-chi-squared distributions, provided that  $AC_{g,b}$  spans the positive real line<sup>31</sup>. The shrunken estimates are the values at which the posterior density is maximized. Thus, posterior estimates are also referred to as maximum *a posteriori* estimates. The shrunken estimates  $\widetilde{AC}_{g,b}$  and  $\widetilde{AC}_{g,e}$  for the intra-sample and within-sample ANOVA components are

$$\widetilde{\text{AC}}_{g,b} = \frac{\text{AC}_{0,b} * d_{0,b} + \text{AC}_{g,b} * d_{g,b}}{d_{0,b} + d_{g,b}}$$

$$\widetilde{\text{AC}}_{g,e} = \frac{\text{AC}_{0,e} * d_{0,e} + \text{AC}_{g,e} * d_{g,e}}{d_{0,e} + d_{g,e}}$$

with  $d_{0,b} + d_{g,b}$  and  $d_{0,e} + d_{g,e}$  degrees of freedom, respectively.

The final shrunken variance estimates are obtained by inverse-transforming  $\widetilde{\text{AC}}_{g,b}$  and  $\widetilde{\text{AC}}_{g,e}$ :

$$\widetilde{\sigma}_{g,b}^2 = (\widetilde{\text{AC}}_{g,b} - \widetilde{\text{AC}}_{g,e})/2 \quad \text{and} \quad \widetilde{\sigma}_{g,e}^2 = \widetilde{\text{AC}}_{g,e}.$$

### *Hypothesis testing*

A moderated t-statistic is used to test the null hypothesis whether there is a difference between log2 translation efficiency in one condition versus another, defined as

$$\tilde{t}_g^{TE} = \frac{\hat{\beta}_g^{TE}}{\tilde{\sigma}_g \sqrt{(\mathbf{c}' \mathbf{X}' (\tilde{\sigma}_{g,b}^2 \mathbf{U} \mathbf{U}' + \tilde{\sigma}_{g,e}^2 \mathbf{I}_N)^{-1} \mathbf{X} \mathbf{c})}}$$

where  $\mathbf{c} = \mathbf{c}(0,0,0,1)'$ ,  $\tilde{\sigma}_g = \sqrt{(\tilde{\sigma}_{g,b}^2 + \tilde{\sigma}_{g,e}^2)}$ ,  $\hat{\beta}_g^{TE}$  is the estimated log2 fold change of translation efficiency between the two conditions<sup>32</sup>, and  $\mathbf{X}$  is the interaction design matrix. For degrees of freedom, we need to consider the posterior density of  $\tilde{\sigma}_{g,b}^2$  and  $\tilde{\sigma}_{g,e}^2$ . The posterior density of  $\tilde{\sigma}_{g,e}^2$  is known and equivalent to the posterior

density of  $\widetilde{AC}_{g,e}$ , which is distributed as an inverse-scaled-chi-squared distribution with  $\mathbf{d}_{0,e} + \mathbf{d}_{g,e}$  degrees of freedom.  $\widetilde{\sigma}_{g,b}^2$  is effectively a linear combination of two inverse-scaled-chi-squared random variables. The density of  $\widetilde{\sigma}_{g,b}^2$  is mathematically intractable and difficult to compute<sup>33</sup>. Therefore, we arbitrarily chose two values of degrees of freedom for  $\widetilde{\sigma}_g$  and compared the performance of our method under each setting: 1) sum of the degrees of freedom of the two posterior ANOVA components ( $\mathbf{d}_{0,b} + \mathbf{d}_{g,b} + \mathbf{d}_{0,e} + \mathbf{d}_{g,e}$ ), 2) the degrees of freedom of the within-sample ANOVA component ( $\mathbf{d}_{0,b} + \mathbf{d}_{g,b}$ ). These two settings represent our belief in the reliability of sample information. The first setting is lenient and assumes all sample information is usable; the second setting is conservative and assumes a small effective sample size based on the number of cell lines.

#### 2.2.4 Comparative benchmarks

Benchmark tests follow the procedures described in Love et al.<sup>34</sup>. The performance of pairedSeq is evaluated and compared with LMM (linear mixed model), *voom*<sup>25</sup>, and DESeq2<sup>34</sup>. LMM and *voom* are normal-based models that perform differential testing on log2-cpm values of read counts in ribosome occupancy data and in transcript abundance. LMM assumes a gene-wise intra-sample variability and reports results without empirical shrinkage of variance components, while *voom*

models a gene-wide intra-sample variability measure and performs empirical shrinkage on within-sample variability in expression measures. DESeq2 is a count-based model that performs shrinkage on both fold change and within-sample variability, but the method does not support intra-sample variability specification.

### *Data sets*

Three real data sets from ribosome profiling experiments are discussed and presented for features of ribosome occupancy data. In addition, we use these data sets to construct simulated data sets for comparative benchmark tests.

*Yoruba40*: This data set includes lymphoblastoid cell lines (LCLs) derived from 40 unrelated female Yoruba individuals. The RNA-seq data was collected in Pickrell et al. <sup>15</sup>, and ribosome occupancy data was collected in Battle et al. <sup>35</sup>.

*Primate5*: This data set includes LCLs derived from 6 Yoruba individuals and 5 rhesus monkey individuals. The RNA-seq part of the data was collected in Khan et al. <sup>36</sup>, and ribosome occupancy data was collected in Wang et al. <sup>37</sup>.

*Yeast4*: McManus et al. <sup>21</sup> performed both RNA-seq and ribosome footprint profiling on two strains of yeast (*S. cerevisiae* and *S. paradoxus*). Four samples were collected for each strain of yeast.

### *Sensitivity and precision*

The power performance of each differential test is evaluated using *Primate5*, following the design delineated in Love et al. <sup>38</sup>. The basic idea is to assess the sensitivity and precision of each differential test where the true calls of genes with differential translation efficiency are constructed from the comparison tests. The data set in which the true calls are made is referred to as the target set, and the data set in which the true calls are assessed is called the evaluation set. The samples of ribosome occupancy and transcript abundance are matched on the cell lines from which they are derived. The cell lines are randomly assigned to the target set and the evaluation set, with 3 human versus 2 rhesus samples in the target set and 2 human versus 2 rhesus samples in the evaluation set. Each simulation constitutes a random assignment of target set and evaluation set. The number of possible random assignments is determined using the number of cell lines in each condition. *Primate5* consists of 6 human versus 5 rhesus samples; thus there can be a maximum of 10 pairs of target set versus evaluation set. For each pair of target versus evaluation set, we apply each of the 5 differential tests to establish true calls at a series of false discovery rate less than .05 to .5 with .5 increments. The same procedure is applied to the evaluation set to identify the significant genes.

Sensitivity is defined as the proportion of true positives (among the genes that passed the FDR threshold as true call in the target set, the proportion of those that are identified as significant in the evaluation set). Precision is defined as the proportion of true positives among genes identified as significant in the evaluation set (among the genes passed the FDR threshold in the evaluation set, the proportion of those that are identified as a true call in the target set). At each nominal adjusted p-value threshold, we computed the mean and standard error of power and precision for each differential test of translation efficiency.

#### *False discovery rate*

Simulated data sets are constructed from *Yoruba40*, a data set in which the samples do not represent known biological conditions. False discovery rate is computed for two experimental settings with 10 and 30 cell lines across 50 simulated data sets. In each simulated data set, we matched samples of ribosome occupancy and transcript abundance that are derived from the same cell line. We then randomly selected cell lines from the 40 cell lines in the original data without replacement. Next, two sets of arbitrary condition labels are assigned to the selected cell lines. These cell lines are not expected to differ in translation efficiency between the two arbitrary assigned conditions. We then performed differential testing of translation efficiency using pairedSeq and the comparison methods.

Significant call of fold change in translation efficiency was set at false discovery rate less than .005, .01, .05, and .1. For each differential test of translation efficiency, we compute the proportion of significant genes out of the total number of genes in each simulated data set and averaged across the 50 simulation experiments.

## 2.3 Results and discussion

### 2.3.1 Variance and mean dependency

The trend of variance dependency on mean in log<sub>2</sub>-cpm values of read counts is evaluated in all three real data sets. [Figure 2](#) shows for each real data set, the fitted nonparametric relationship between variance and mean using locally weighted scatter plot smoothing (loess). The trend fit is similar between ribosome occupancy data and in RNA-seq data, especially among the highly expressed genes. Among the lowly expressed genes, there is a stronger dependency of the variances on the mean of the log<sub>2</sub>-cpm values in the ribosome occupancy data in *Yoruba40* and *Primate5*. However, in the yeast data *Yeast4*, the variance-mean dependency is similar across all levels of mean log<sub>2</sub>-cpm values of read counts. In all subsequent analyses shown here, we modeled the variance-mean trend across the two data. The *voom* precision weights were computed using the cross-data-type variance-

mean trend fit and used to weight covariance matrices of each gene in differential testing of translation efficiency.

### 2.3.2 Co-variability between ribosome occupancy and transcript abundance

We characterized co-variability between ribosome occupancy and transcript abundance in data sets before correcting for variance-mean dependency. [Figure 3](#) presents per gene correlations between ribosome occupancy levels and transcript abundances in *Yoruba40*. Across the genes, there is a positive relationship between log2-cpm values of ribosome occupancy and transcript abundance for a single sample and also for averaged expression values across samples ([Figure 3a,b](#)). This phenomenon is frequently cited as one of the supports for the quality of ribosome footprinting experiments. However, we observed that this positive relationship does not hold at the gene level. The correlations between ribosome occupancy and transcript abundance vary from -1 to 1, with the median suggesting a positive correlation ([Figure 3c](#)). This variation in gene-level co-variability is consistent across levels of translation efficiency and also gene expression level as in log2-cpm values ([Figure 3d,e](#)).

Next, we ask whether the co-variability may differ as a function of biological conditions in *Primate5* and *Yeast 4*. *Primate5* consisted of human and rhesus

monkey samples and are expected to exhibit high biological variability in expression measurements. *Yeast4* is a data set of low biological variability which includes two closely related yeast strains. [Figure 4a](#) and [Figure 4c](#) display cross-species co-variability between ribosome occupancy and transcript abundance. The correlations spans -1 to 1 in both data sets, with distribution skewed to the left, indicating a large number of genes with positive co-variability between ribosome occupancy and transcript abundance. This result is not surprising, given that samples with high biological variability present a wide dynamic range of possible data values, and hence are likely to present high co-variability between samples.

In [Figure 5](#), genes are ranked by co-variability between ribosome occupancy and transcript abundance and plotted for their log<sub>2</sub>-cpm values of variance versus mean gene expression values. We observed a tendency for genes with high co-variability to exhibit high variance in the data and vice versa for genes with low co-variability. This trend is consistent across differing levels of gene expression and hence persists after correlation for variance-mean dependency in the data. [Figure 4b](#) and [d](#) further compare per gene correlations between species and shows that co-variability is not a function of species difference in *Primate5* or *Yeast4*.

[Figure 6a,b](#) demonstrate that among genes with large co-variability, there tends to be a larger fold change or species difference in translation efficiency. This

dependency cannot be accounted for by the *voom* variance-mean dependency correction, and persists after adjusting for different levels of variances in log2-cpm values ([Figure 6c,d](#)).

The intra-sample co-variability between ribosome occupancy and transcript abundance can be negatively valued in a ribosome occupancy experiment because it corresponds to covariation between log2-cpm values of ribosome occupancy and transcript abundance. We note that a negative variance component violates the assumption of mixed-effect modeling under the hierarchical framework where intra-sample co-variability represents the variance parameter of a normal random variable. Suppose we are interested in making inferences about sample measurements at the cell line level, a negatively valued variance component would suggest that cell line variability follows a normal distribution with negative variance parameter. Since we are not interested in making inferences at the cell line level, a negative variance component is not so much of a concern.

### 2.3.3 Shrunken covariance estimates and differential translation efficiency

[Figure 7](#) compares correlation estimates observed in *Primate5* with the shrunken correlation estimates. `pairedSeq` shrinks correlation in both genes with positive and negative correlation between log2-cpm values of ribosome occupancy

and transcript abundance. The shrinkage is stronger among genes with positive co-variability. [Figure 8](#) compares significance finding of pairedSeq with *voom* and DESeq2 in *Primate5* at false discovery rate of .01. When compared to *voom*, we observed that pairedSeq is able to pick up genes with high fold change in translation efficiency, especially for genes with high and positive co-variability between ribosome occupancy and transcript abundance. On the other hand, *voom* tends to pick up genes with lower fold change in translation efficiency and with negative co-variability. In *voom*, a single positive correlation is used to model co-variability in the observed expression values without any empirical shrinkage. While, pairedSeq allows the observed co-variability to vary across genes and also performs empirical shrinkage on covariance estimates. Because co-variability is associated with high variability in the expression values, by shrinking co-variability in significance testing, pairedSeq is able to uncover signals that would be otherwise difficult to detect. Similar findings were observed when comparing pairedSeq to DESeq2, which provide shrinkage on fold change and variance estimates but not on covariance estimates. pairedSeq picks up signals with higher fold change than DESeq and these signals are genes with positive co-variability.

Negative co-variability between ribosome occupancy and transcript abundance indicate that the samples have a differing order of magnitude in

ribosome occupancy than in transcript abundance data. We believe that such a difference, if it were observed between species, would be the result of biological variance between samples. However, we observed many of the genes with negative co-variability to have differing order of magnitude between data types within species. Such phenomenon may be due to experimental artifacts in the ribosome footprinting experiment. A sensible differential test of translation efficiency should avoid calling genes in which expression values are confounded by experimental artifacts. Compared to both *voom* and DESeq2, pairedSeq is able to pick up genes with strong biological variation in the data instead of genes with experimental artifacts.

We performed a series of benchmark tests based on simulated data generated from Primate5 and compared pairedSeq's sensitivity, precision, and false discovery rate with *voom*, LMM, and DESeq2. [Figure 9](#) and [Figure 10](#) illustrate the power and sensitivity of pairedSeq using a nominal adjusted p-value threshold of .01. pairedSeq has higher sensitivity than DESeq2, *voom*, and LMM. pairedSeq-r is a version of pairedSeq with the moderated t-statistic computed at a conservative degrees of freedom. pairedSeq-r is also comparable in sensitivity with DESeq2. Precision of *voom* is the highest among all methods, followed by DESeq2, pairedSeq and LMM. In summary, the benchmark tests illustrate that pairedSeq is

effective in controlling false discovery rate and in detecting true positives of differential translation efficiency.

## 2.4 Conclusion

We introduce pairedSeq, a statistical method designed for the integrated analysis of ribosome footprint profiling and RNA-seq to identify phenotype differences in translation efficiency. The method explicitly models co-variability in ribosome occupancy and transcript abundance, a source of variation that directly affects the dynamic range of expression measurements in the analysis of translation efficiency. Positive co-variability indicates that a strong biological variation in the data, while negative co-variability suggests potential technical artifacts in the experiment. Empirical Bayes shrinkage is proposed for co-variation between ribosome occupancy and transcript abundance. Results indicate that pairedSeq effectively shrinks covariance estimates and is able to identify differences in translation efficiency in samples with high measurement variability. Future work can extend the proposed covariance shrinkage method to the joint analysis of sequencing data and quantitative spectrometry data.



## 2.5 Figures

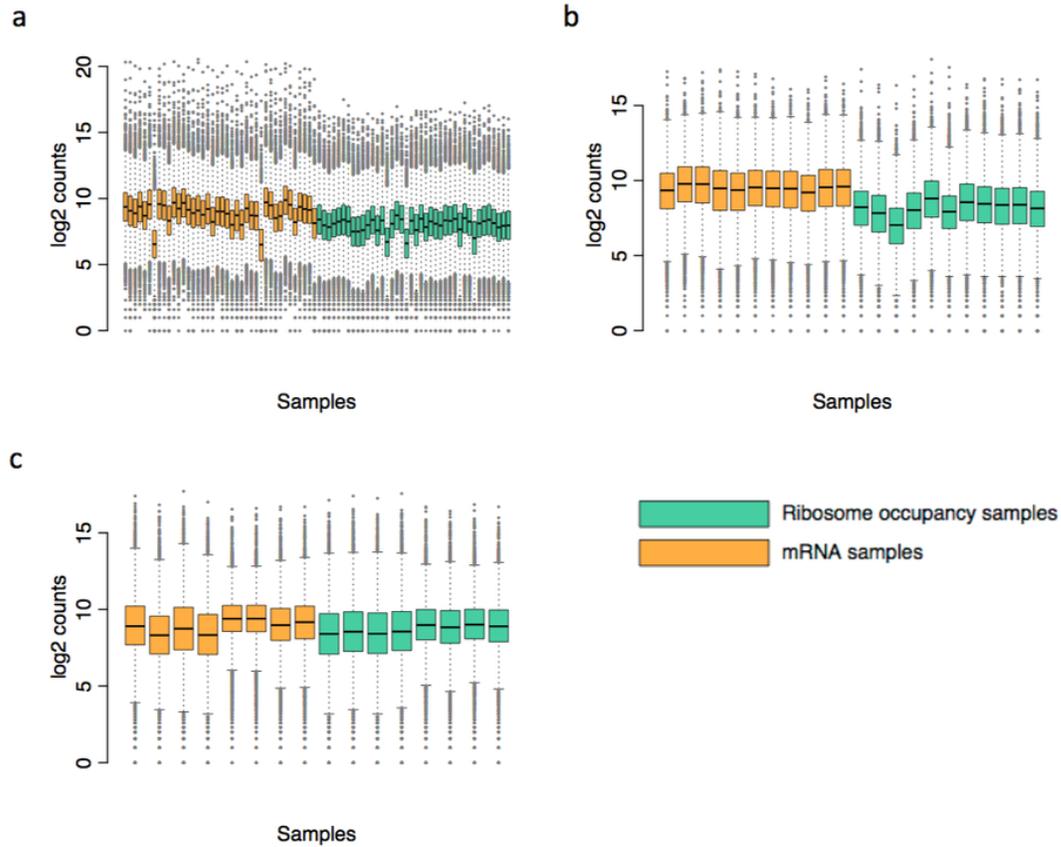


Figure 1. Gene count distributions vary between ribosome occupancy data and RNA-seq data.

Three real data sets are presented here. (a) Lymphoblastoid cell lines (LCL) derived from 40 unrelated female Yoruba individuals. (b) LCLs derived from 6 human Yoruba individuals and 5 rhesus monkey individuals. (c) Four *S. cerevisiae* and 4 *S. paradoxus* yeast strains.

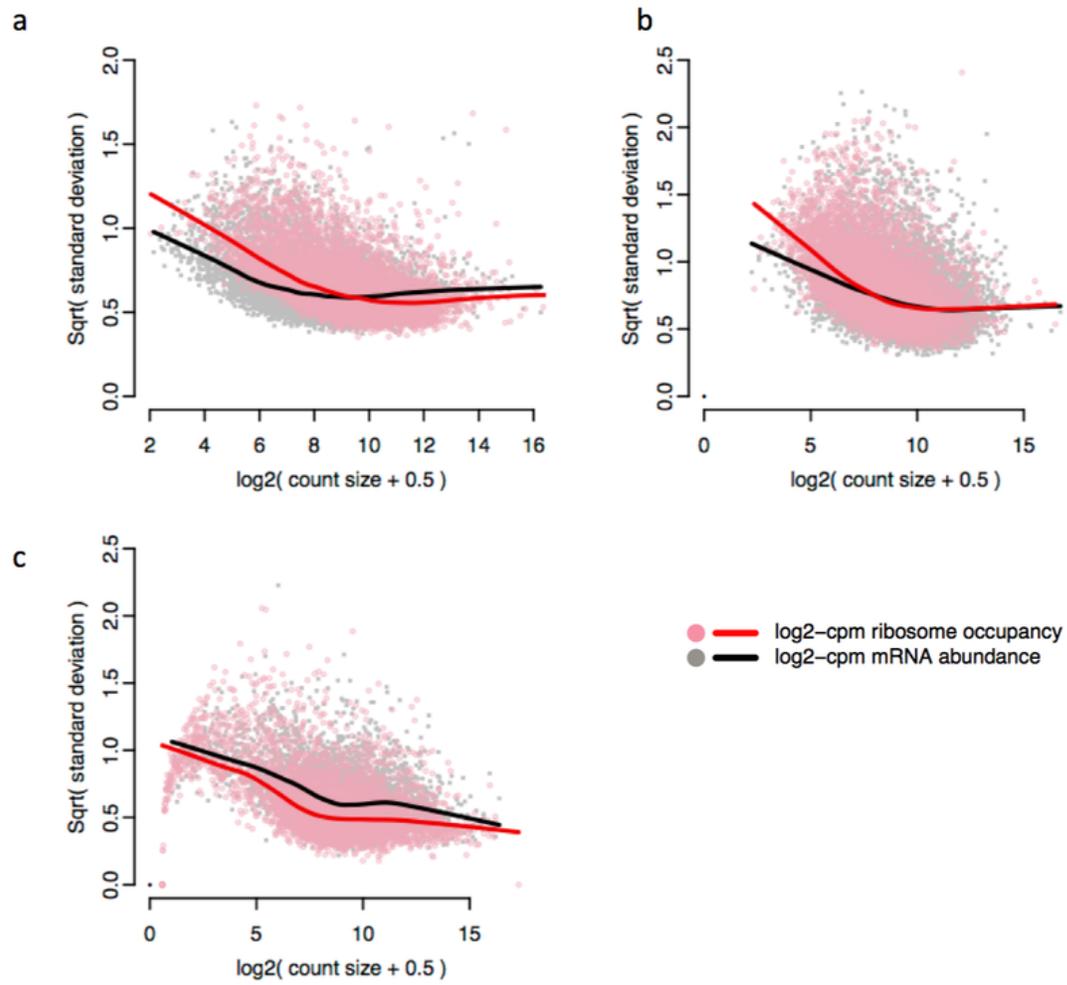


Figure 2. Variance-mean dependency in 3 real data sets.  
 (a) LCL from 40 female individuals. (b) 6 human individuals and 5 rhesus monkey individuals. (c) Four *S. cerevisiae* and 4 *S. paradoxus* yeast strains.

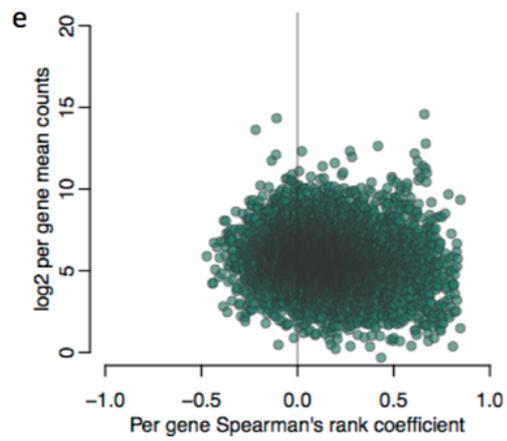
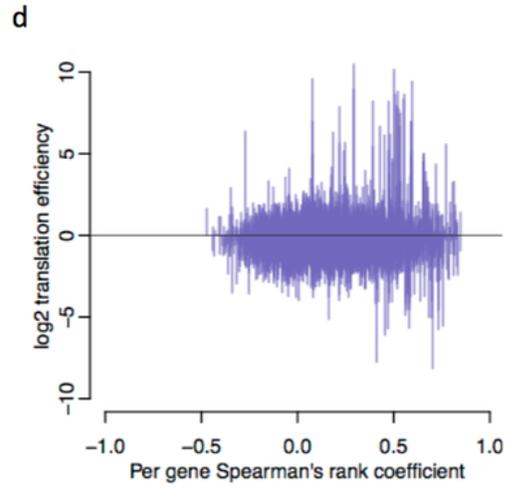
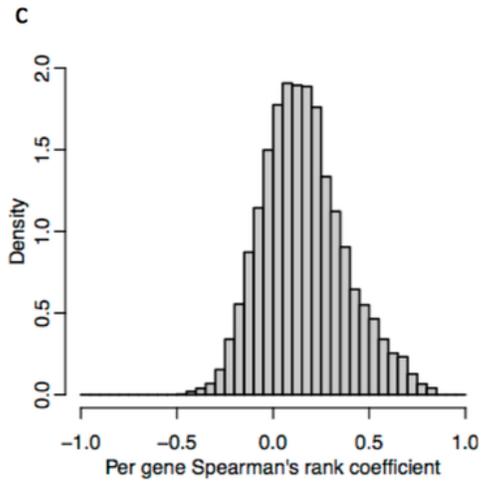
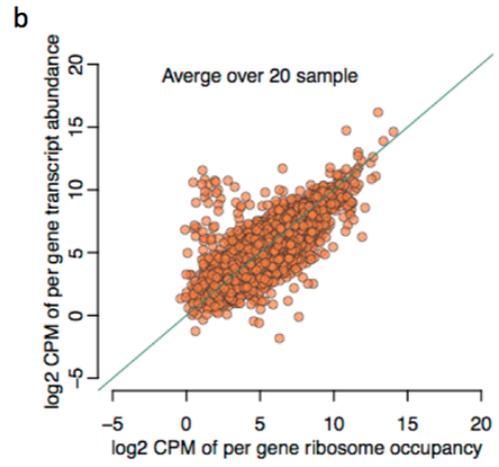
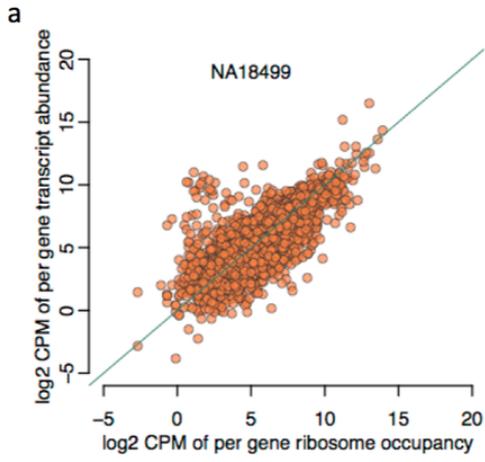


Figure 3. Correlation between ribosome occupancy and transcript abundance.

Data set includes 40 unrelated Yoruba female individuals. (a) Plotted is each gene's log<sub>2</sub>-cpm values of ribosome occupancy versus transcript abundance in NA18499. (b) Plotted is each gene's average value across samples in log<sub>2</sub>-cpm of ribosome occupancy and transcript abundance. (c) Frequency distribution per gene rank correlation between log<sub>2</sub>-cpm ribosome occupancy and transcript abundance. (d) log<sub>2</sub> translation efficiency (TE) was computed for each gene as difference between sample averages of log<sub>2</sub>-cpm value of ribosome occupancy and transcript abundance. Each purple line represents a gene: the length indicates the size of the log<sub>2</sub> TE and the x-coordinate of the line indicates the magnitude and directionality of the gene's rank correlation between log<sub>2</sub>-cpm values of ribosome occupancy and transcript abundance. (e) Plotted is each gene's mean log<sub>2</sub>-cpm values across samples of both data types and the corresponding rank correlation between samples.

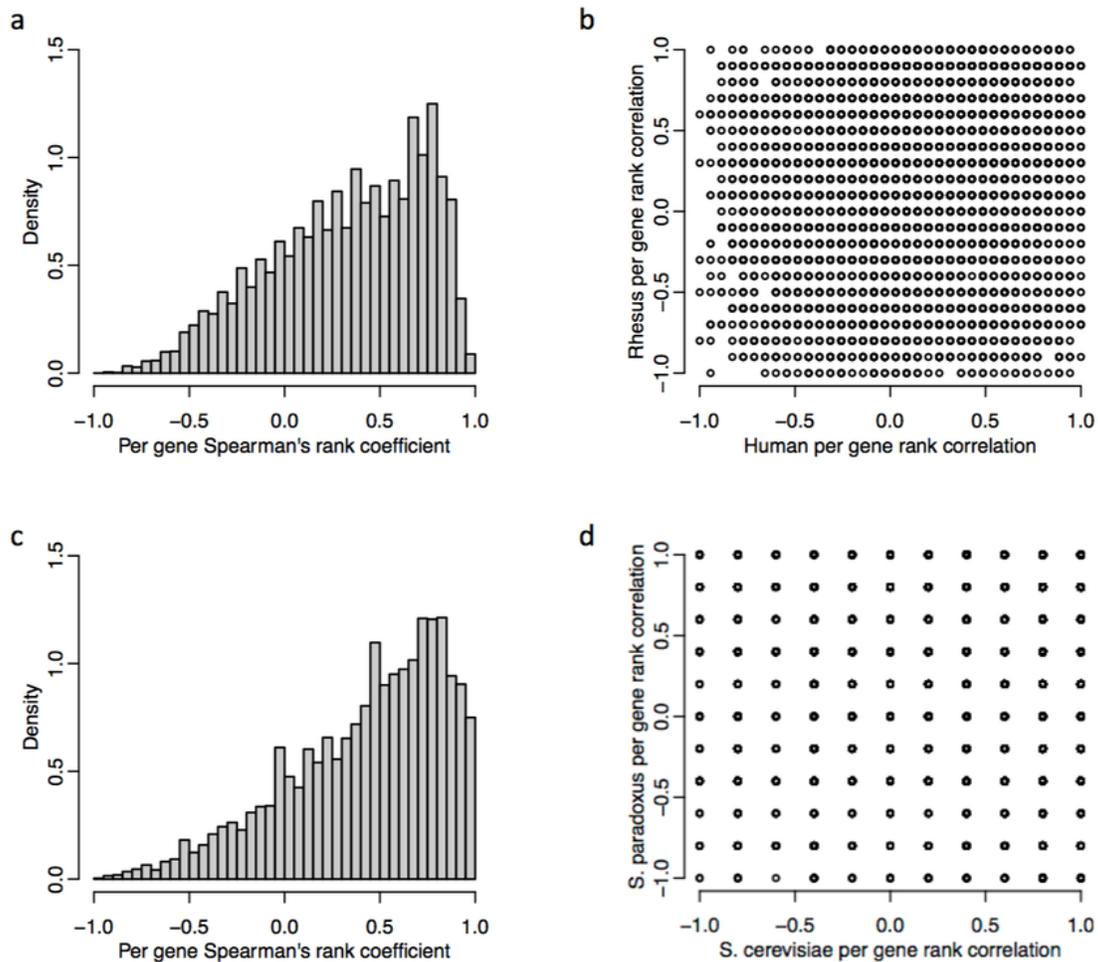


Figure 4. Per gene correlation between ribosome occupancy and transcript abundance across species. Plots (a) and (b) are based on a data set of 6 human individuals and 5 rhesus monkey individuals. (a) Frequency of per gene rank correlation between log<sub>2</sub>-cpm values of ribosome occupancy and transcript abundance across samples (b) Plotted is each gene's correlation between ribosome occupancy and transcript abundance in human individuals versus its correlation in rhesus monkey individuals. (c) and (d) are based on a data set of 2 yeast strains (4 samples each). (c) Frequency of per gene rank correlation between log<sub>2</sub>-cpm values of ribosome occupancy and transcript abundance across samples (d) Plotted is each gene's correlation between ribosome occupancy and transcript abundance in *S. paradoxus* versus correlation in *S. cerevisiae* individuals.

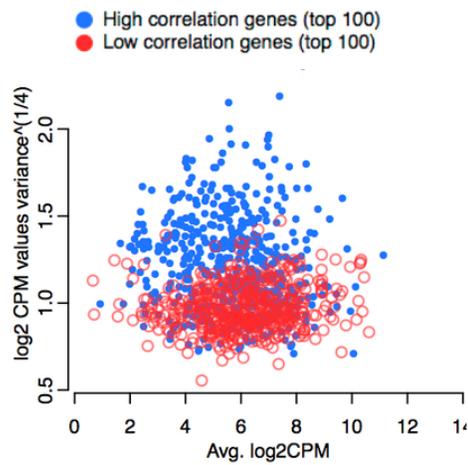


Figure 5. Variance and mean trend among genes with high and with low correlation between ribosome occupancy and transcript abundance. Based on a data set of 6 human individuals and 5 rhesus monkey individuals. We computed Spearman's rank correlation between log2-cpm values of ribosome occupancy and transcript abundance. Two sets of genes are of particular interest: the top 100 genes and the bottom 100 genes in correlation. Plotted is per gene log2-cpm values variance across samples and the average log-cpm values across samples. Blue dots indicate genes with top 100 correlations. Red dots indicate genes with bottom 100 correlations.

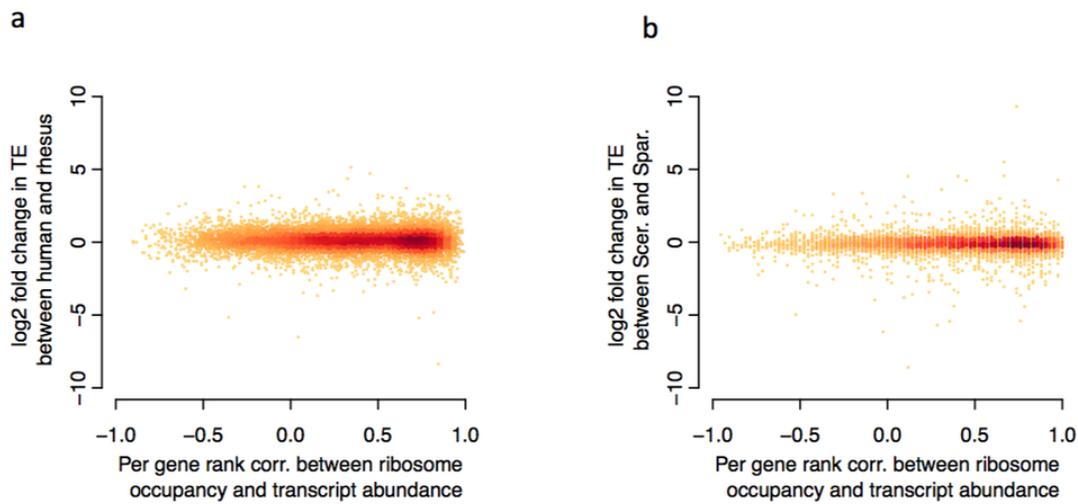


Figure 6. Species difference in translation efficiency is larger for genes with positive correlation between ribosome occupancy and transcript abundance. Plotted is per gene log<sub>2</sub> fold change in translation efficiency versus per gene's rank correlation between log<sub>2</sub>-cpm ribosome occupancy and transcript abundance across all samples. (a) based on a data set of 6 human individuals and 5 rhesus monkey individuals. (b) based on two yeast species (4 samples each).

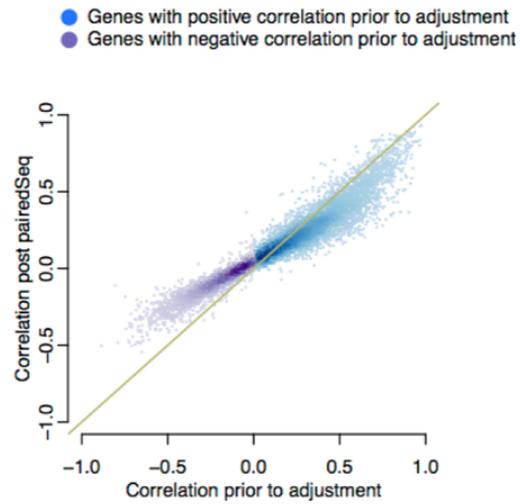


Figure 7. pairedSeq shrinks both positive and negative correlations between ribosome occupancy and transcript abundance.

Plotted is each gene's shrunken correlation coefficient based on the pairedSeq method and the observed sample rank correlation coefficient. Blue dots represent genes with positive observed correlation. Purple dots depict genes with negative observed correlation.

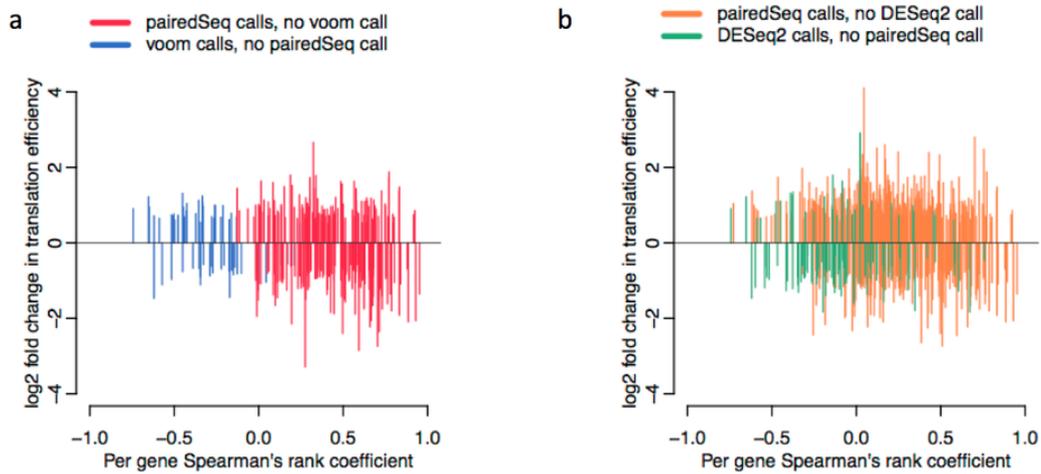


Figure 8. pairedSeq detects significant fold change despite high intra-sample variability.

Results are based on a real data set of 6 human individuals and 5 rhesus monkey individuals. We performed tests of differential translation efficiency at false discovery rate  $< .01$ . (a) Compares significant calls in *voom* and in pairedSeq. Red line represents genes identified as significant in pairedSeq but non-significant in *voom*. Blue lines depicts genes identified as significant in *voom* but non-significant in pairedSeq. (b) Compares significant calls in DESeq2 and in pairedSeq. Red line represents genes identified as significant in pairedSeq but non-significant in DESeq2. Blue lines depicts genes identified as significant in DESeq2 but non-significant in pairedSeq.

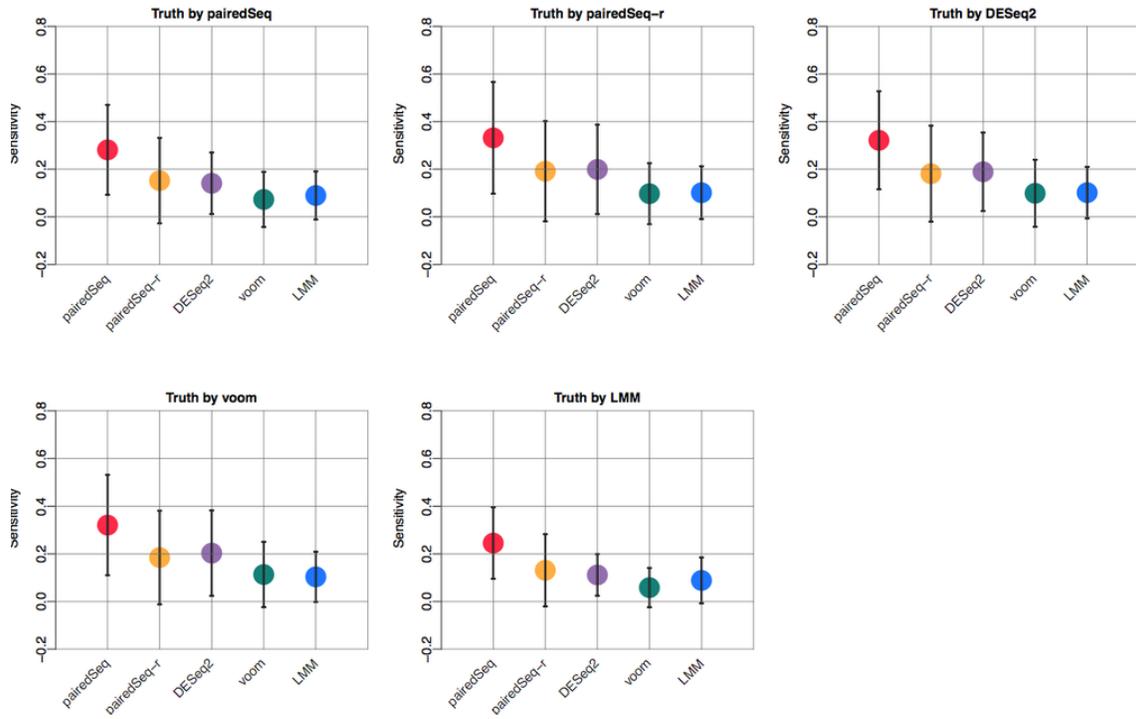


Figure 9. Benchmark sensitivity in pairedSeq, pairedSeq-r, DESeq2, *voom*, and LMM at FDR < .01.

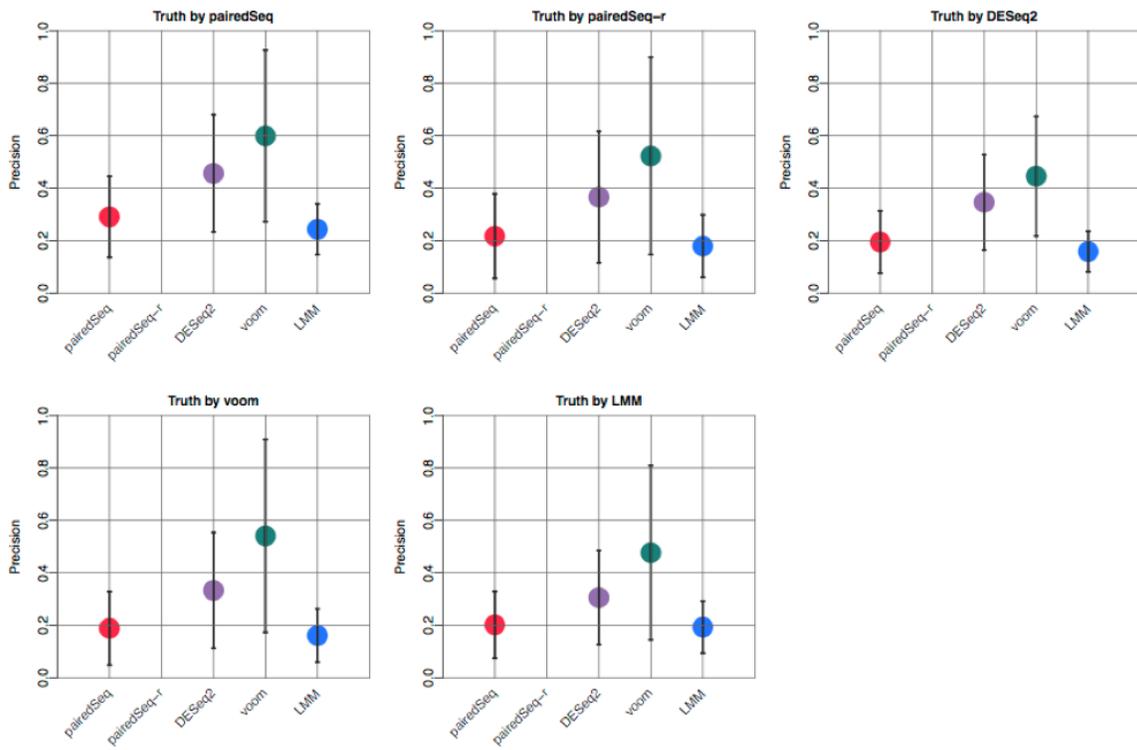


Figure 10. Benchmark precision in pairedSeq, pairedSeq-r, DESeq2, *voom*, and LMM at FDR < .01

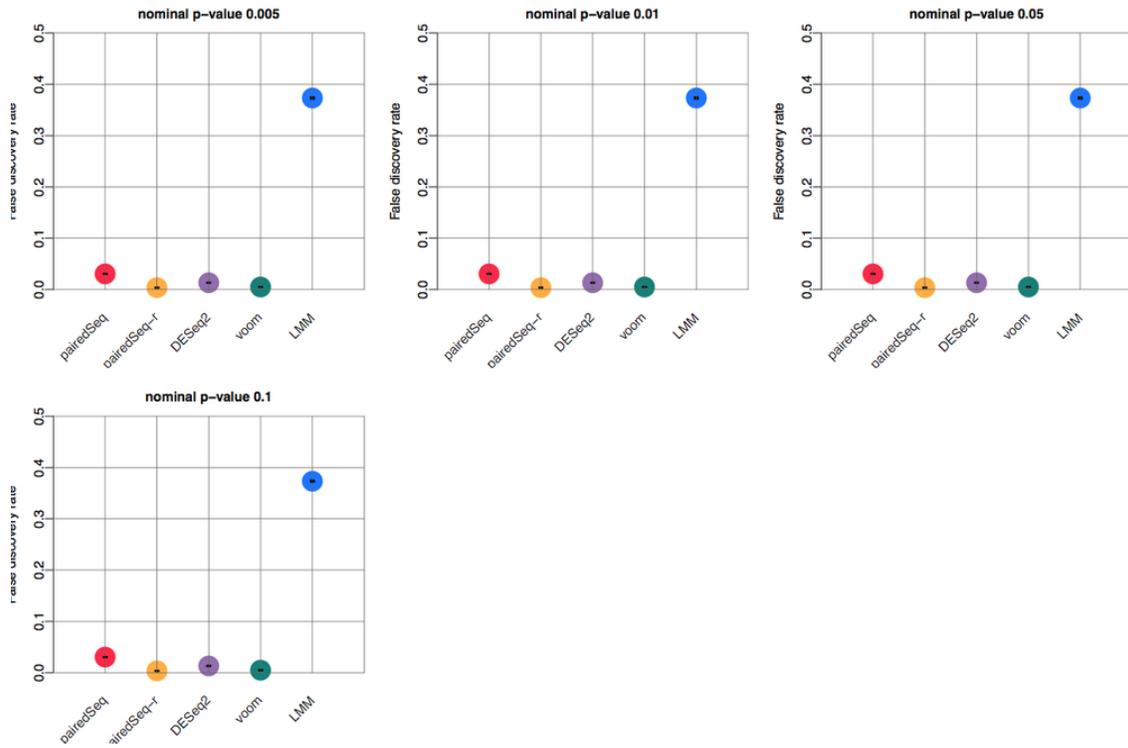


Figure 11. Benchmark false discovery rate in pairedSeq, pairedSeq-r, DESeq2, *voom*, and LMM.

## 3 Comparative genomic study between primates

### 3.1 Abstract

Mechanistic descriptions of gene regulation differences between homo sapiens and its closely related species, the chimpanzee, is likely to reveal principles determining phenotypes that are distinctly human. Although gene expression is a multi-step process, the bulk of the literature concerning gene expression in primates is focused on transcript level and transcription factor binding differences. This transcription centered view of gene regulation disregards more than half of the process and is therefore potentially misleading. To gain a comprehensive view on gene regulation, we surveyed genome wide ribosome occupancy levels in five lymphoblastoid cell lines (LCL) each derived from humans, chimpanzees and Rhesus macaques to estimate translation level divergence in primates. We performed integrative analysis on gene regulation, combining mRNA level and protein level measurements we previously collected from matching cell lines. We found that translational level divergence often propagates existing divergence at the transcript level between humans and chimpanzees. For genes that show differences in divergence between RNA and protein levels, we estimate about 30% are mediated through translational regulation. The majority of the differences in

protein-RNA divergence, however, are explained by post-translational attenuation, likely mediated through post-translational modifications. This attenuation is conserved across the three primate species studied here. It appears that the divergence attenuation could evolve under stabilizing selection of protein expression levels to serve as a buffering mechanism maintaining homeostasis against perturbations from environment or genetic variations. Finally, we observed increased variations in mRNA levels for buffered genes in a human population, indicating relaxation of selective constrain on transcript levels for buffered genes in recent human evolution.

## 3.2 Background

Differences in gene regulation are the major factors determining phenotypic variations between closely related species <sup>39,40</sup>. Alterations of tissue specific gene expression patterns of essential genes is more likely to survive natural selection than coding substitutions, which would result in ubiquitous effects that often sum up to a negative net impact on the fitness level of an organism <sup>41</sup>. Almost half a century ago, based on examination of coding substitutions between human and chimpanzee, King and Wilson <sup>42</sup> postulated that gene regulation differences are

more likely to be the major factor driving the phenotypic divergence. Over the past decade, ample studies surveying gene expression variation in primates reached conclusions concurring with this hypothesis <sup>17,43-45</sup>. Consistently, population genomics studies looking for signatures of recent selection also highlighted the key roles of regulatory variations in human adaptation <sup>46,47</sup>.

Even though it is clear that protein expression levels are the relevant quantities for coding genes, most studies investigating variations in gene expression among primates focused on comparing expression levels of mRNA <sup>17,43,48,49</sup>. While the general efficacy of using mRNA level as a proxy for estimating protein levels is still up for debate <sup>50</sup>, it is clear that in some instances translational and/ or posttranslational regulation of gene expression resulted in protein levels that far diverged from the mRNA levels upstream <sup>51</sup>. In fact, it has been shown that protein expression levels are far more conserved across diverse taxa than mRNA levels <sup>52</sup>. How this conservation of protein levels is achieved given the apparent divergence at the mRNA level is still unclear.

To estimate the divergence between mRNA and protein levels in primates, we previously collected RNA-sequencing and quantitative mass spectrometry data from a set of 15 primates (5 humans, 5 chimpanzees and 5 Rhesus monkeys) Lymphoblastoid Cell Lines (LCL). Consistent with the earlier observation across

wider taxa, we found extensive variations in mRNA levels between human and chimpanzee, which were attenuated by some downstream processes and resulted in a conserved protein expression level <sup>53</sup>. While a stronger evolutionary constraint at the protein expression level is in line with the fact that proteins form the machineries that execute biological functions, it remains unclear whether the attenuation occurred translationally or post translationally.

Ribosome profiling is a technique utilizing next generation sequencing to survey ribosome footprints in a massively parallel fashion <sup>7</sup>. It has been shown that ribosome occupancy levels estimated from counting number of ribosome footprint sequencing reads are good approximations for the level of protein translation <sup>54</sup>. We have recently applied this technique in a panel of HapMap cell lines to identify genetic variants regulating translation and to estimate relative contribution of translational and posttranslational regulation to steady state protein levels <sup>55</sup>. We found that even within the human population, protein level tend to be less variable than mRNA level. Interestingly, variations of translation levels tend to be in concordance with mRNA rather than protein, indicating a major role of post-translational processes in variation attenuation.

To further investigate this property across primates, we performed ribosome profiling experiments on the same five chimpanzee and five rhesus cell lines as

reported in Khan et al. <sup>53</sup>. Through integrative analysis of ribosome profiling data and quantitative mass spectrometry data, we estimated the contribution of translational regulation in attenuating gene expression variations to achieve a conserved protein levels across primates. Our results suggest that a post-translational gene expression buffering mechanism evolved under stabilizing selection of protein expression levels. To our knowledge, our dataset provides the first global view on translational landscape across primates. This dataset also allowed us to interrogate the regulatory relationships across different layers of gene expression phenotypes and their roles in primate evolution.

### 3.3 Methods

#### 3.3.1 Ribosome footprint profiling

We collected ribosome occupancy data from Epstein-Barr virus (EBV)-transformed lymphoblastoid cell lines (LCLs) derived from five human (Coriell YRI, NIGMS Human Genetics Cell Repository, GM18585, GM18507, GM18516, GM19193, GM19204) and five chimpanzee (*Pan troglodytes*) individuals (New Iberia Research Center: Min 18358, Min 18359; Coriell/IPBIR: NS03659, NS04973, Arizona State University, Pt91). Data also collected from rhesus *Herpesvirus papio*

transformed LCLs from five rhesus macaque (*Macaca mulatta*) individuals (Harvard Medical School, NEPRC: 150-99, R181-96, R249-97, 265-95, R290-96). Cell culture work followed the same procedure as previously described<sup>55</sup>. Ribosome profiling experiments were performed using the ARTseq<sup>TM</sup> Ribosome Profiling kit for mammalian cells (RPHMR12126) following vendor's instructions, with minor modifications that we described in Battle et al.<sup>55</sup>. Flash frozen pellets of 30 to 50 million cells were used for each experiment. Preprocessing of sequencing data followed the procedure used by Ingolia et al.<sup>56</sup>. Mapping of processed reads and counting for estimating orthologous gene expression followed the same steps for RNA-sequencing data as described previously<sup>53</sup>. Only uniquely mapped reads were included for downstream analysis. All rRNA, tRNA and snoRNA reads were discarded.

Aggregate plots for codon periodicity were generated by aggregating read counts across annotated start codons on the plus strand of the human genome that have average PhostCons scores greater than 0.9 in the flanking 100 bp window. Only 5' end position of each read is counted. For chimpanzee and rhesus macaque, start codon annotation was converted from the human annotation using liftOver (<https://genome.ucsc.edu/cgi-bin/hgLiftOver>).

### 3.3.2 Statistical analysis

#### *Gene expression data*

We performed a series of analyses that focused on detecting differential divergence between primates across molecular phenotypes using RNA-seq, Ribosome Profiling, and quantitative Mass Spectrometry data. Sequencing read counts for RNA-seq and ribosome profiling data are converted to log<sub>2</sub> Reads Per Kilobase per Million mapped reads (RPKM) for each orthologous gene. To test gene expression differences in translational level between primate species, we adopted a nested linear model approach and computed a likelihood ratio statistic to quantify statistical significance of species differences in ribosome occupancy levels. To test species difference in translational divergence, we performed a joint analysis of RNA-seq and ribosome profiling data. We adopted a nested linear model approach, in which the likelihood ratio statistic quantifies the statistical significance of differences between species divergence at each molecular layer. Similarly, we performed joint analysis of ribosome profiling data and mass spectrometry data to test species differences in post translational divergence. Since protein expression level was measured by the SILAC labeling technique which computes a  $\log_2(\text{sample} / \text{standard})$  ratio of peptide abundance for each individual

<sup>53</sup>. Ribosome profiling data were adjusted accordingly in the joint analysis to assimilate the quantitative range between data types.

#### *Subsampling for variance comparison*

Subsampling for variance comparisons was performed using the sample function in R <sup>57</sup>. Background genes were classified by expression level into 20 categories. The probability for each category to be drawn is proportional to the expression level distribution of the buffered genes. For each iteration, sampling is performed without replacement to avoid drawing the same gene twice.

#### 3.3.4 Functional enrichment analysis

Enrichment analyses were performed as previously described in <sup>53</sup> with the following exceptions. Gene ontology analysis was performed using GOstats <sup>58</sup>. For each gene list tested, the full set of 3188 quantified genes was used as the background list. To calculate the number of missense mutations for each gene we converted the human-chimpanzee coding sequence alignment fasta file (downloaded from UCSC genome browser, 2011) to a VCF file and used SnpEff <sup>59</sup> to annotate the functional consequences of each human chimpanzee substitution. We then used SnpSift and bedtools <sup>60</sup> to intersect and count number of missense mutations for each gene. For UTR PhastCons scores, we computed average scores for each

annotated UTR exons (downloaded from Ensembl, 2014) by averaging per base PhastCons scores from the 100-vertebrate alignment (downloaded from UCSC, date) using the UCSC tool `bigWigAverageOverBed` <sup>61</sup>. For each gene, we identified the median value among the average PhastCons scores from all associated UTR exons as the representative PhastCons score for the 5' and 3' UTR, respectively.

## 3.4 Results

### 3.4.1 Ribosome profiling captures variations in protein translation among primates

To compare levels of protein translation genome-wide in primates, we used high throughput sequencing technology to profile ribosome protected fragments of mRNA <sup>7</sup> in five lymphoblastoid cell lines for each of human, chimpanzee and rhesus macaque that we previously collected mRNA level and protein level measurement <sup>53</sup>. After removing sequencing reads from rRNA and other contaminating sources, we obtained a median coverage of ~12 million uniquely mapped reads per sample. High sequencing quality and high technical replications for each cell line between sequencing runs was observed. Across all samples, we found median footprint length of 29 nt, and consistent codon periodicity patterns

among species ([Figure 12a](#)), both features indicative of footprints of elongating ribosomes.

Ribosome occupancy level has previously been shown to be an effective approximation for the level of translation<sup>54</sup>. We computed translational levels using normalized read counts for Ribosome Protected Fragments (RPF) that aligned to exons orthologous across the three primate species being studied (<http://giladlab.uchicago.edu/orthoExon/>). The major signal in the ribosome occupancy data is reflective of species differences ([Figure 12b](#)). As indicated in the results of Principal Component Analysis, we observed clear separation of the data points by species in the first two principal components. Interestingly, when ribosome profiling and RNA sequencing data are analyzed jointly, the data points are also separated by technology type (in addition to species) in the first two principal components ([Figure 12c](#)). This result highlights the fact that ribosome profiling captures biological signals different from those of an RNA sequencing experiment.

Between human and chimpanzee, we found 1,700 genes differentially translated (likelihood ratio test at 1% FDR) out of 10,050 genes quantified by ribosome profiling. Significantly more genes are differentially translated between human and rhesus, at the same FDR cut-off (3,295 genes,  $p < .0001$ , Chi-squared

test). Similar results are observed between chimpanzee and rhesus. A higher number of differentially expressed genes is expected between human (or chimpanzee) and rhesus, given the greater evolutionary distance.

Taken together, our analysis indicates that the major signal in the set of ribosome profiling data captures the essence of biological variations in translation amongst primate species.

#### 3.4.2 Variations in transcription across primates are mostly propagated to translation

Previously we reported that the divergence of gene expression in mRNA level between human and chimpanzee is largely attenuated at the protein level <sup>53</sup>. To gain a mechanistic understanding, we evaluated the relative contribution of translation. To do so, we analyzed ribosome profiling data in conjunction with RNA-seq and quantitative mass spectrometry data. We focus on a set of 3,188 genes that are quantified across all three layers of molecular phenotype.

Overall, strong correlation across genes is observed among mRNA level, translation level and protein level. While level of translation (ribosome occupancy) appear to correlate better with mRNA levels than with iBAQ estimates of protein levels, we reason that the difference could be driven by similarity in technology

used to estimate mRNA and translation levels. To better account for technology differences and take advantage of the more quantitative estimate of protein level with SILAC labeling, we estimated relative expression levels by standardizing over a reference line (GM19238) for each type of the gene expression level estimates independently. We then compared correlation between data type per gene across species. Interestingly, in this comparison, we still found significantly higher correlation between mRNA and ribosome occupancy (median Spearman's .38) comparing to between protein and mRNA (median Spearman's .32) or between protein and ribosome occupancy (median Spearman's .28) ([Figure 13a](#)). Higher correlation between mRNA and ribosome occupancy in this context indicates that for genes differentially expressed among species, translation levels tend to share more variation with transcript levels than with protein levels. Similarly, in a replication test across layers of molecular phenotypes for genes showing differential divergence between human and chimpanzee, we found higher replication rate between mRNA level divergence and ribosome occupancy level divergence (64%) than replication rates of either by protein level divergence (~45%).

To further investigate this observation, we ask in general how correlated are the cross species divergence in translation levels and transcript levels. We found that between human and chimpanzee, divergence in transcript levels tends to

correlate well with the divergence in translation level (Spearman  $\rho = .5$  among 3,188 genes, [Figure 13b](#)). This correlation increased when only genes that are differentially expressed between human and chimpanzee at either transcriptional or translational layer were considered (Spearman  $\rho = .64$ ), indicating a sizable dilution of the correlation by noise in the data. Similar results were seen within a larger set of 8,572 genes for which we have quantification for both RNA level and ribosome occupancy level. This observation is in contrast to previous observations on differences in divergence between RNA and protein levels<sup>53</sup>, indicating that translational regulation is likely either less common or smaller in effect size than post-translational regulation between human and chimpanzee LCLs.

To further substantiate this inference we compared the significance level between differential divergence for each of the transcript-ribosome occupancy level comparison and ribosome occupancy-protein level comparison ([Figure 13c](#)) and the effect size of differential divergence ([Figure 13d](#)). We found larger effect size ( $p < .0005$ , Wilcoxon Rank Sum) and more significant p values in ribosome occupancy-protein level divergence. Both results support the inference of less regulation at the translational levels. We therefore conclude that in general translation tends to propagate variations in transcript levels with minor modulations.

3.4.3 Differences between RNA and protein level divergence across primates can partly be explained by translational regulation

While translation in general propagates variations in transcript level, there is a subset of genes that show translation specific regulation. To evaluate the contribution of translational regulations to differential divergence between RNA and protein level across primates, we examined the divergence level of ribosome occupancy relative to mRNA and protein. We found that ribosome occupancy shows intermediate levels of divergence for genes that are differentially divergent between RNA and protein levels ([Figure 14a-h](#)). While we found more similarity between ribosome occupancy and RNA level divergence ([Figure 14c, g](#)), it is clear that translation contributes to some extent the differences in divergence at the protein level ([Figure 14a, d, e, h](#)). To further estimate the level of contribution, we calculated the replication rate of differences in protein level divergence from RNA by ribosome occupancy. We found that for genes that show enhanced divergence at the protein level, 31 % can be explained by an enhanced divergence at the ribosome occupancy level. Similarly, for genes that show attenuated divergence at the protein level, 24 % can be explained by attenuation at ribosome occupancy (likelihood ratio test, 1% FDR). Interestingly, for genes that show no significant differential divergence between RNA and protein levels ([Figure 14i-l](#)), some

differential divergence is observed at ribosome occupancy levels when comparing to RNA divergence levels ([Figure 14k](#)). These differences are then attenuated post translationally ([Figure 14l](#)), indicating that certain regulation at the level of translation only lead to changes in turnover rate of the protein but not the steady state level.

To formally test the differences specifically at the translation level (i.e. not propagated from transcript level), we developed a model that can be used for differential testing of translation efficiency — an estimate for the amount of translation per transcript as a function of the mRNA expression level. After taking into account the inherent variations at the mRNA level, we found 478 genes that show translation specific regulation out of 3,188 at 1% FDR. Of those 291 genes show translation-specific enhanced variation and 187 genes show attenuation of transcript level variation ([Figure 15a](#)).

We next ask if there are common features shared by genes that are differentially regulated in translation between human and chimpanzee. We found these genes to have significantly lower GC content ( $p < 10^{-12}$ , Wilcoxon rank sum) and more reported protein-protein interactions ( $p < 10^{-4}$ , Wilcoxon rank sum) ([Figure 15b](#)) suggesting that sequence content in the coding region could play a role in functional regulation of differential translation. In addition, we found that

these genes tend to have longer 3' UTR ( $p < .0002$ , Wilcoxon rank sum) and more predicted miRNA binding sites ( $p < .005$ , Wilcoxon rank sum) (Figure 15b). This observation is consistent with a role for regulatory elements in UTR regions in regulating translation. Interestingly, we also found that the UTR regions of these genes are more conserved across vertebrates (5'UTR,  $p < 0.005$ ; 3'UTR,  $p < 0.0005$ ; Wilcoxon rank sum) (Figure 15b), which could indicate that the divergence in translational regulation may be specific to the primate lineage. Taken together, we observed primate specific translational regulations that could be mediated through differential miRNA binding in the 3' UTR sequences.

To further investigate the functional implications of translational regulation in primate evolution. We estimated the proportion of differential translational regulation between human and chimpanzee that gets percolated to the protein expression levels. We found that among 478 genes that show differential regulation at the level of translation (compare to the transcript level) at 1% FDR, only 79 genes maintain the same direction of variation at the protein level. The majority showed no differential divergence between protein level and mRNA level (Figure 15c). Intriguingly, when we categorize genes by direction of translational divergence, we found that translationally enhanced genes tend to have low percolation rate (8%) (Figure 15d) This is in sharp contrast to the 30% percolation

rate for translationally attenuated genes ([Figure 15e](#)). The estimated proportion of percolation varies with the significance cutoff choices due to power issues. However the asymmetry is robust to the cutoffs used for the differential divergence tests. It should be noted that here we are focusing on the extreme tails of the translational regulation distribution. Taken together with the analysis on the general distribution, these observations indicate that translational regulations may affect the turnover rate of proteins in addition to the steady state level. While translational enhancements of divergence mainly contribute to divergence in the protein turnover rate, translational attenuations of divergence contribute more to stabilizing protein levels. Interestingly, gene ontology analysis for genes that are translationally enhanced in divergence (divergent in protein turnover rates) ([Figure 15d](#)) showed enrichment of genes involved in negative regulation of cAMP metabolism, cell division and translation. While in contrast, for genes that show translational attenuation of divergence ([Figure 15d](#)), we found enrichment of negative regulation of catabolic activity and response to chemical or cytokine. These results indicate that species divergence in protein turnover rate is likely reflecting species differences in signal transduction pathways. On the other hand, divergence attenuation by translational regulation likely reflects a mechanism maintaining homeostasis in response to environmental stimuli.

### 3.4.4 Variations in translation across primates are frequently attenuated in protein levels

By jointly analyzing ribosome profiling data and mass spectrometry measurement of protein levels, we could estimate the contribution of posttranslational gene regulation to steady state protein levels. Between human and chimpanzee, we found 381 genes that are differentially divergent between translation level and protein level (FDR 1%), indicating significant post-translational regulation on expression of these genes. More genes were attenuated at the protein level than enhanced (325 vs. 56, Fisher's Test  $p < .005$ ). This asymmetry indicates a prevalence of variation attenuation in posttranslational regulation of protein levels ([Figure 16a](#)). In fact, using a more relaxed significance cutoff (5% FDR), we found 1937 genes (out of 3188 genes quantified) to be significantly less divergent (fold difference between human and chimpanzee) at protein level than at the translational level. The observed prevalence of post-translational attenuation of divergence is not simply reflecting an increase in technical noise in protein measurement. We have previously shown that technical variation in quantitative mass spectrometry data is actually smaller than sequencing data <sup>53</sup>.

While the above analysis indicated a relatively strong contribution of post-translational attenuation in gene regulation, it remains possible that fundamental differences in data distribution between the two data types could obscure the results. To address this issue, we coerced both data types into the same quantitative ranges and compared their effect sizes. We found similar patterns in divergence attenuation at the protein level reinforcing the conclusion that post-translational attenuation should be considered a major gene regulation process contributing to human and chimpanzee gene expression differences.

To further investigate the potential mechanisms that could account for this observed post-translational gene expression attenuation, we seek common features shared among genes in this group (1% FDR). Post-translational attenuation could result from protein sequence divergence between human and chimpanzee that led to differences in protein stability. However, we found no enrichment of non-synonymous substitution between human and chimpanzee in the group of post-translationally attenuated genes ([Figure 16b](#)). In contrast, we found enrichment of reported phosphorylation sites ( $p < .002$ , Wilcoxon rank sum) and ubiquitination sites ( $p < .003$ , Wilcoxon rank sum) after standardizing by protein length ([Figure 16b](#)). Enrichment of known post-translational modification sites suggests that the attenuation of translational differences could be mediated through differential post-

translational modification mechanisms between human and chimpanzee that buffers the variation in gene expression levels between species. In addition, we found that post-translationally buffered genes tend to be expressed in more tissues ( $p < .006$ , Wilcoxon rank sum) and have more reported protein-protein interactions ( $p < .0002$ , Wilcoxon rank sum) ([Figure 16b](#)), suggesting importance of post-translational buffering in multiple biological processes. We next ask if post-translationally buffered genes are enriched for functions in certain biological processes. We found enrichment of mainly housekeeping functions such as translation, nucleic acid metabolic processing and protein targeting to ER ([Figure 16c](#)). It appears that post-translational attenuation of species divergence in gene expression levels functions mainly in maintaining stable protein level for genes involved in fundamental biological processes. Unexpectedly, we found enrichment of associative learning and neuronal ontology terms for genes that are post-translationally enhanced for the protein level divergence between human and chimpanzee. While it is interesting to speculate how protein level divergence driven by post-translational regulation could contribute to cognitive differences between human and chimpanzee, the observation made here are based on experiments performed using immortalized B cells. Although gene regulation modules are known

to be shared across tissue types<sup>62,63</sup>, it remains unclear whether this enrichment would be replicated in neurons.

#### 3.4.5 Post-translational buffered genes are under stabilizing selection in protein level

The observed post-translational buffering between human and chimpanzee pointed to a mechanism in place to maintain stable protein expression levels across species. To ask if this buffering mechanism is conserved across primate lineages, we ask if the buffered genes are under stabilizing selection at protein levels in primates. We found that buffered genes (1% FDR) are 2.36 times as likely to be under stabilizing selection as genes that are not buffered, when stabilizing selection is defined as the top 300 genes that have the lowest protein expression level

variance among primates<sup>53</sup> ([Figure 17a](#)). This enrichment is not simply driven by a mean variance relationship, as the mean expression levels between buffered and non-buffered genes are comparable across all three species. Since the current framework for estimating stabilizing selection could suffer from issues resulted from significance cutoff selections, we choose to further examine the enrichment of buffered genes in the group of genes under stabilizing selection defined by various cutoffs ([Figure 17a](#)). It appears that even for top 1000 stabilizing selected genes, we

still observed a  $\sim 1.8$  fold enrichment of buffered genes. A similar result is observed with a relaxed cutoff for defining the set of buffered genes. To further substantiate this conclusion, we identified buffered genes from each of the three pairwise species comparisons independently and then tested if the overlap is more than expected by chance. We found that around half of human-chimpanzee buffered genes are also buffered in human-rhesus comparison ( $p < 10^{-6}$ ). In addition, roughly a third of human-chimpanzee buffered genes are buffered in chimpanzee-rhesus comparison ( $p < 10^{-15}$ ) ([Figure 17b](#)). Significant overlaps across each group of buffered genes supports the conclusion that post-translational buffering of protein expression is likely a conserved mechanism.

Since protein expression levels are likely the main relevant quantities for gene function, we expect that the presence of a post-translational buffering mechanism would permit relaxation of selection constraints on the transcript level. Consistent with this hypothesis, we found higher cross species variations in mRNA levels for buffered genes. This result, however, is harder to interpret given that ribosome occupancy and mRNA levels are highly correlated and here we are examining variations in mRNA levels across samples that are highly variable at the ribosome occupancy level to begin with. In addition, species divergence that are not buffered also contribute to variance in the background set, which leads to an

underestimate of differences in the variance level between buffered and background genes. To better test this hypothesis, we examine signs of constrained relaxation at mRNA levels in the context of recent human evolution. We compare variations at the mRNA levels in the Yoruba population<sup>15</sup> between buffered genes (identified across human and chimpanzee) and the background. We found slightly higher variance for buffered genes across individuals. This difference is significant ( $p < 0.02$ ) when the selected background set is adjusted according to expression levels of the buffered genes ([Figure 16c](#)) to account for the known mean-variance correlation in sequencing data<sup>25,64</sup>. It appears that the buffered genes identified across human and chimpanzee have a relaxed constraint at the mRNA levels among the Yoruba individuals. We reason that high variation at the mRNA level in the population will allow us to identify more eQTL with greater effect size. Counter intuitively we found no apparent enrichment of eQTL among buffered genes and on average smaller effect size. Higher variation at mRNA expression levels accompanied by lower power in detecting eQTL indicates that there are more linked variants in the regulatory regions of buffered genes that are acting in opposite directions. This interpretation is in line with the hypothesized relaxation in selection constraint at the regulatory regions for buffered genes. Following the same line of logic we compared expression variations between lincRNA (functional molecule) and mRNA

(information intermediate) of matching expression levels. We found lower variations among individuals in lincRNAs relative to mRNA, supporting the idea of a general relaxation in selection constraint at the mRNA level. This observation supports the hypothesized general post-translational buffering mechanism acting in humans that stabilize the protein level despite variations upstream.

### 3.6 Discussion

To determine the contribution of translational regulation in gene expression level differences between human and chimpanzee, we generated new data using ribosome profiling to estimate translation levels. Close inspection of the features of ribosome protected fragments (sequencing reads) and global analysis on ribosome occupancy levels across species lead us to conclude that our data captures the footprints of translating ribosomes, which have been shown to be the closest measurement to approximate the level of translational activities<sup>54</sup>. This dataset in conjunction with the data described in Battle et al.<sup>55</sup> and Cenik et al.<sup>65</sup> provides a unique opportunity to explore recent evolution of translational regulation in humans.

Through joint analysis with RNA-seq measurement of transcript levels and quantitative mass spectrometry measurement of protein levels, we provided an integrated view of gene regulation. We found that variations at the transcript level tend to propagate to level of translation (ribosome occupancy), which suggests minor regulatory variations between human and chimpanzee that directly impact the level of translation. This observation is in contrast to previous reports on pervasive translational buffering observed in F1 hybrids between *S. cerevisiae* and *S. paradoxus*<sup>21,66</sup>. Interestingly a report focusing on the same process in hybrids between laboratory and wild isolated strains of budding yeast<sup>67</sup> and a follow up reanalysis of the Artieri dataset<sup>68</sup> did not confirm the pervasive translational buffering. Instead, their results were more in line with our observation in primates. Scarcity in translational regulation differences between human and chimpanzee is unexpected, given the amount of substitutions present in the UTR regions. It is possible that these genetic variations are cryptic in the environment tested and response to perturbations could reveal species divergence in translational regulation

<sup>69</sup>.

While we observed general concordance between transcript level and ribosome occupancy level divergence between human and chimpanzee, we did find contributions of translational regulation to RNA-protein differential divergence

when focusing on extreme tails of the distribution. We found roughly a third of protein level divergence deviated from mRNA level could be replicated in ribosome profiling data at 1% FDR. Although the proportion replicated by translational regulation is determined by the choice of cutoff, it appears that the major force driving protein level divergence is downstream of translation. More surprisingly, among the limited number of genes that show differential regulation in translation, translational regulation appears to have minor impact on gene expression differences at the protein level. Only about 15% of the genes that show differential human-chimpanzee divergence between translation and transcription (signal of translational regulation) show the same direction of effect percolated to the protein level. The observed percolation rate is highly asymmetric. While close to 30% of translational attenuation percolated to a stable protein level, only about 8% of translationally enhanced divergence is percolated (1% FDR). Although potential power issues prevented us from obtaining a precise estimate for percentage of translational regulation that has persistent impact on steady state protein levels. It is clear that only a minor proportion of translationally enhanced divergence has impact on protein levels. While we do find enrichment of miRNA binding sites amongst the group of genes that are differentially regulated at the level of translation, which could potentially explain the mechanism for translational

divergence <sup>70</sup>, the biological significance at this layer of regulation remains unclear. It is possible that differential turnover rate, instead of steady state protein levels, between human and chimpanzee in this group of genes has biological relevance. Consistent with this notion, we found enrichment of genes in the cAMP pathway, which is fundamental to signal transduction of cellular processes <sup>71</sup>, from the group of genes that show translationally enhanced divergence. We speculate that the differential turnover rate of this group of genes could determine distinct features distinguishing human and chimpanzee.

In sharp contrast, posttranslational gene regulation appeared to have a much broader impact on protein levels. We found that regulation at this layer tend to attenuate variations created upstream. At 5% FDR, we estimated that as much as 60% of genes are under regulation of this post-translational buffering. Although precise estimation of this proportion is technically challenging, it appears that post-translational buffering is quite common in the system inspected. Pervasive buffering of transcriptional and translational variation of gene expression levels has broad implications, especially in the context of evolution. For most genes, protein executes the cellular function. Variations in gene expression that failed to percolate to protein level are therefore less likely to cause any organismal phenotype. In fact, we found evidence for relaxation of selection constrain on the regulation of mRNA

levels in YRI population for buffered genes identified between human and chimpanzee. It appears that further investigating this property in context of population genetics would likely to provide valuable insights on how selection might act on the regulatory variants associated with genes possess such properties. We found similarities between effects of the observed post-translational buffering on gene expression variation and HSP90 chaperone action on rectifying mis-folding caused by missense mutations <sup>72,73</sup>. We speculate that similar to HSP90, post-translational gene expression buffering could confer phenotypic robustness by stabilizing protein expression levels.

We found that genes regulated by post-translational gene expression buffering between human and chimpanzee tend to also be under post-translational buffering between human and rhesus (and between chimpanzee and rhesus). This observation suggested that post-translational buffering is likely a mechanism evolved under stabilizing selection for protein levels in primates. It remains unclear how post-translational buffering is achieved. We found enrichment of reported post-translational modifications amongst this group of genes but no significant enrichment of coding substitutions. One interpretation of this observation would suggest that divergence in post-translational modification mechanisms, between human and chimpanzee, instead of divergence in the coding sequence drives the

buffering. This interpretation makes intuitive sense, since substitutions in the coding sequence that affects turn over rate could not serve as a regulatory buffering mechanism while post-translational modifications are usually reversible and regulatory. Further investigation to identify factors involved in maintaining this mechanism would provide insights to advance our understanding of both how natural selection acts on gene regulation and how to better predict gene expression levels given genetic variations.

Taken together, our results provided the first integrative view on gene express variations across primates that allows a separation between translational and post translational events. We found extensive post translational buffering of gene expression variations that lead to a stable protein level across primate species. We propose a scenario where buffering evolved under stabilizing selection of protein levels that removes negative effects on organismal fitness from protein level variations and allows the transcript level to diverge for quick adaptation to environmental changes. Given the energy cost of translation <sup>74</sup>, it remains puzzling to us that the stabilizing selection appears to act on post translational level instead of at the translational level. We reason that evolution of post translational buffering mechanism is probably the more parsimonious path and speculate a

*trans*-acting mechanism to achieve this pervasive buffering in a relatively short evolutionary time.

### 3.7 Figures

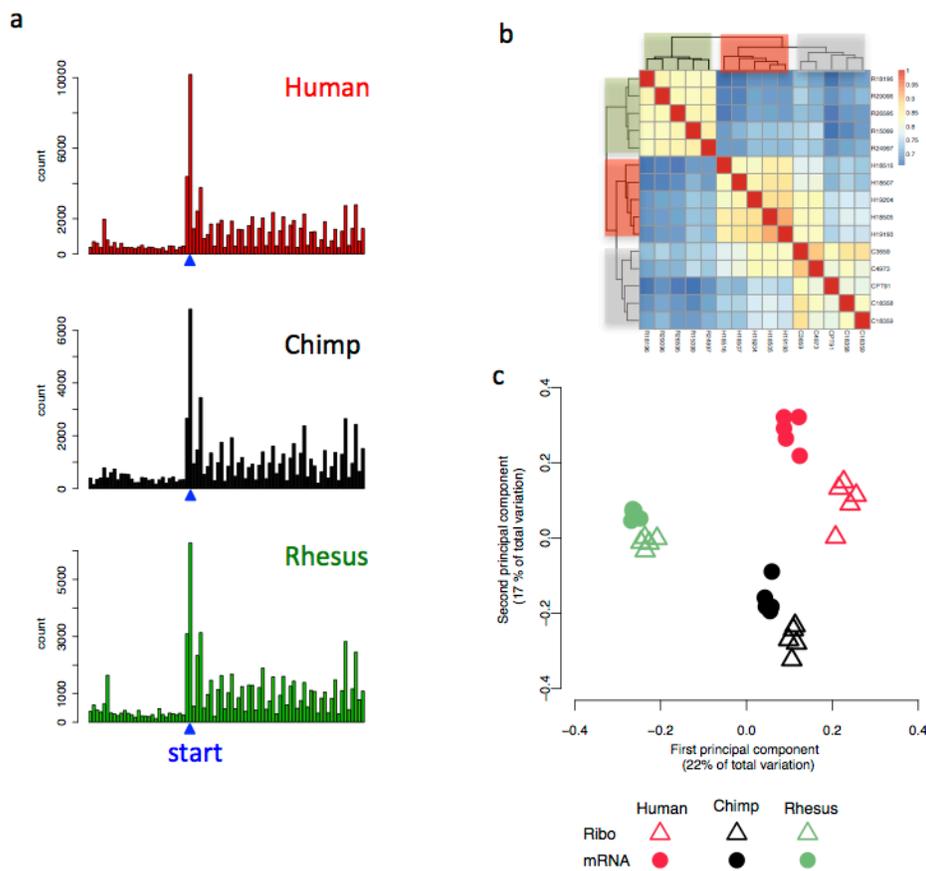


Figure 12. Ribosome profiling captures species differences in protein translation levels.

(a) Aggregate plots of 5' position from all sequence reads that fall in each 80 base pair window flanking conserved start codons for each species. (b) Heat map presenting Spearman's rank correlations for ribosome occupancy levels between individuals. Species label for each individual is color coded at the top and the side of the plot. (c) Joint principal component analysis of ribosome profiling data and

RNA-seq data. The first two principal components contributed to a combined 39% of variance across individual samples. The amount of variance in each sample is projected onto the first two principal components in the scatter plot.

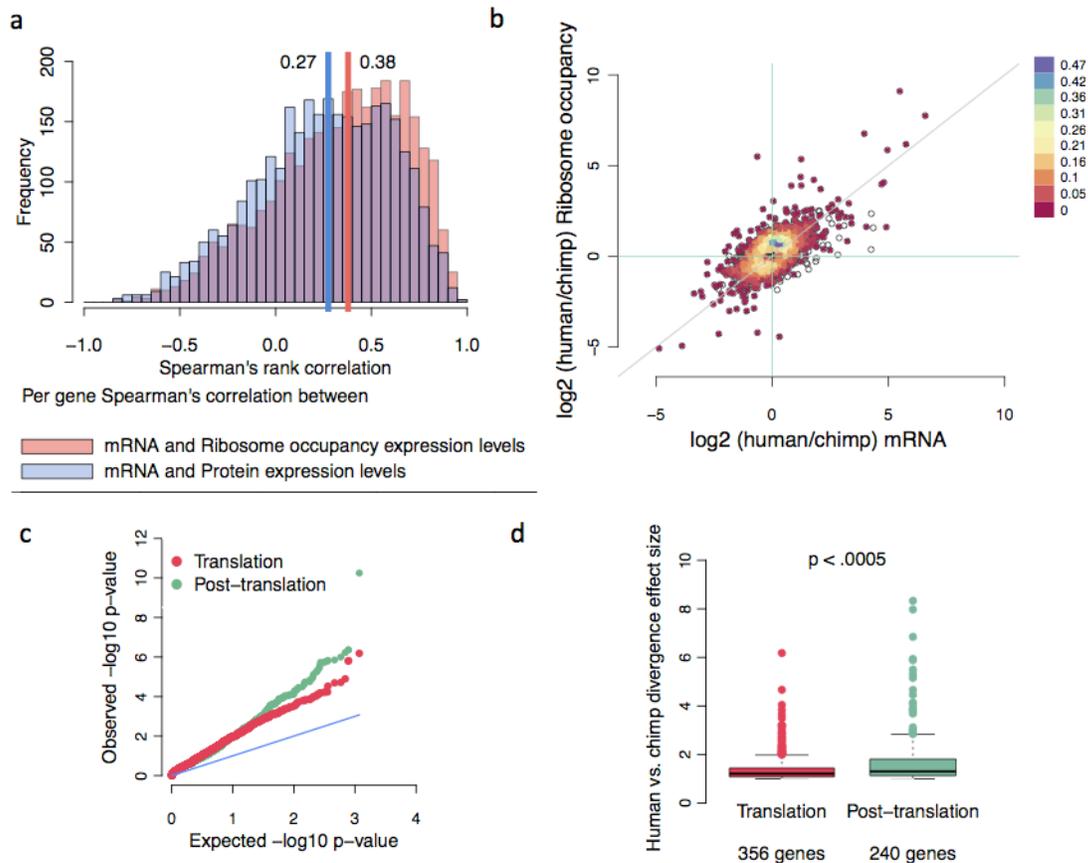


Figure 13. Translation propagates majority of transcriptional divergence. (a) Frequency of gene-wise Spearman's rank correlation coefficients for the pairwise association of the three data types. The median per gene correlation is labeled by the vertical line. (b) Correlation between human/chimp species difference at transcriptional and translational level of gene expression. Plotted for each gene is the estimated fold change (human/chimp) of log<sub>2</sub> RPKM values of ribosome occupancy levels against transcript abundance. Each gray circle represents a gene, Density of data points plotted are color coded as indicated in the key. (c) Statistical significance value of post-translational divergence for each gene exceeds the significance value of translational divergence. For each gene, the estimated significance level of divergence is plotted against the expected significance level. Red represents translational divergence, and green indicates post-translational divergence. (d) Effect sizes of post-translational divergence are larger than the effect sizes of translational divergence. Box plots display absolute value of divergence effect size among genes that are at least 2-fold divergent between species in either translational (red) or post-translational level (green).

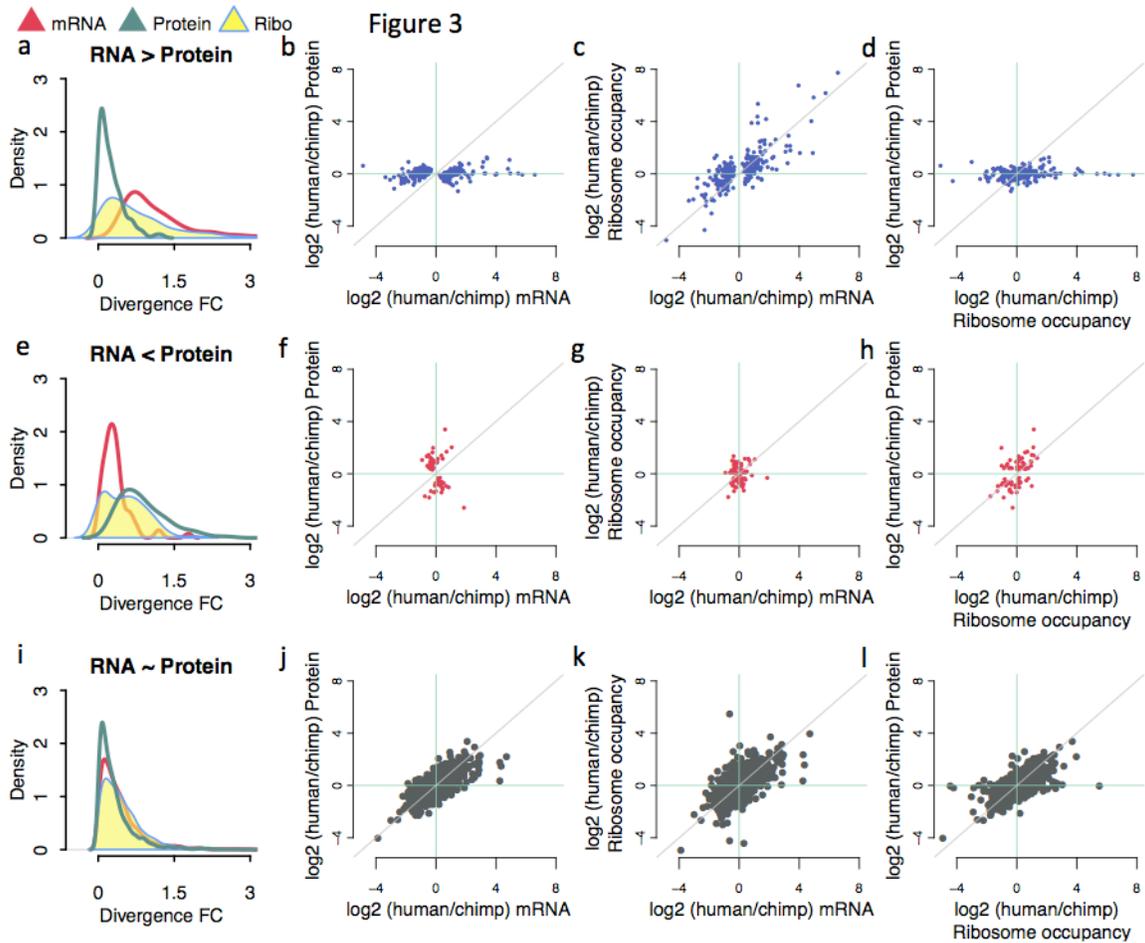


Figure 14. Translational regulations contribute to differences in human-chimpanzee divergence between protein and mRNA.

(a, e, i) Density plots of  $\log_2$  transformed absolute divergence between human and chimpanzee for each layer of molecular phenotype. (b~d, f~h, j~l) Scatter plots of mean  $\log_2$  fold differences between human and chimpanzee for each data type specified. Each data point represents a gene. (a~d) Showing data from genes with absolute RNA divergence greater than absolute protein divergence at FDR 1%. (e~h) Showing data from genes with absolute protein divergence greater than absolute RNA divergence at FDR 1%. (i~l) Showing data from genes with no significant differences between absolute RNA divergence and absolute protein occupancy at FDR 1%.

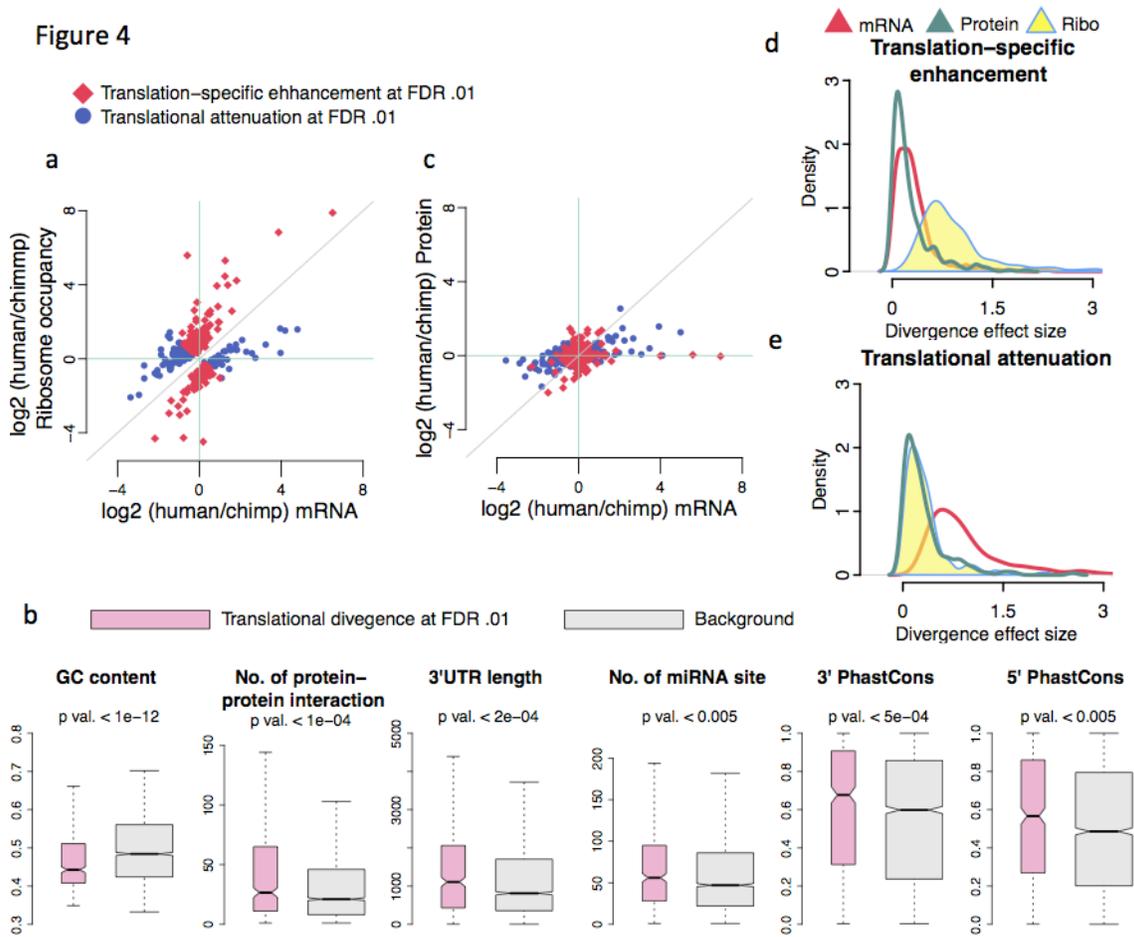


Figure 15. Translational divergence leads to stable protein level and divergent protein turnover rate.

(a) Scatter plot of mean log<sub>2</sub> fold differences between human and chimpanzee for mRNA and ribosome occupancy levels. Only genes that are differentially divergent at 1% FDR are shown. Each data point represents a gene and is colored by direction of effects. (b) Boxplots comparing enrichment of molecular features between genes that are shown in (a) and the background (all the remainder). (c) The same as (a), except that mean log<sub>2</sub> protein divergence is shown on the Y-axis instead of ribosome occupancy. (d) Density plots of log<sub>2</sub> transformed absolute divergence between human and chimpanzee for each layer of molecular phenotype for genes that show translational enhancement of gene expression. (e) The same as (d), but for genes that show translational attenuation of gene expression.

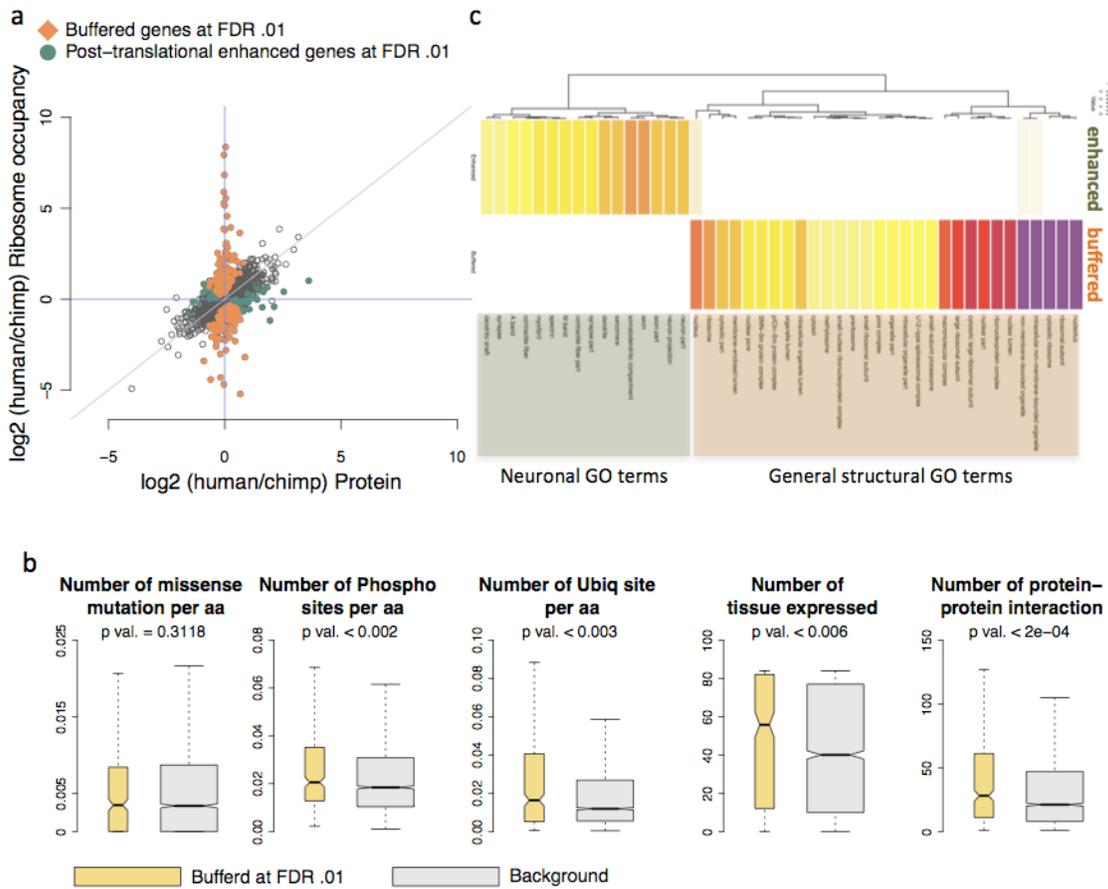


Figure 16. Post-translational gene expression buffering is prominent. (a) Scatter plot of mean  $\log_2$  fold differences between human and chimpanzee for protein and ribosome occupancy levels. Each data point represents a gene and is colored by direction of differential divergence at 1% FDR. (b) Boxplots comparing enrichment of molecular features between genes that are buffered [orange in (a)] and the background [open circle and green in (a)]. (c) Heat map showing p-values of gene ontology enrichment for genes that are post-translationally buffered or enhanced.

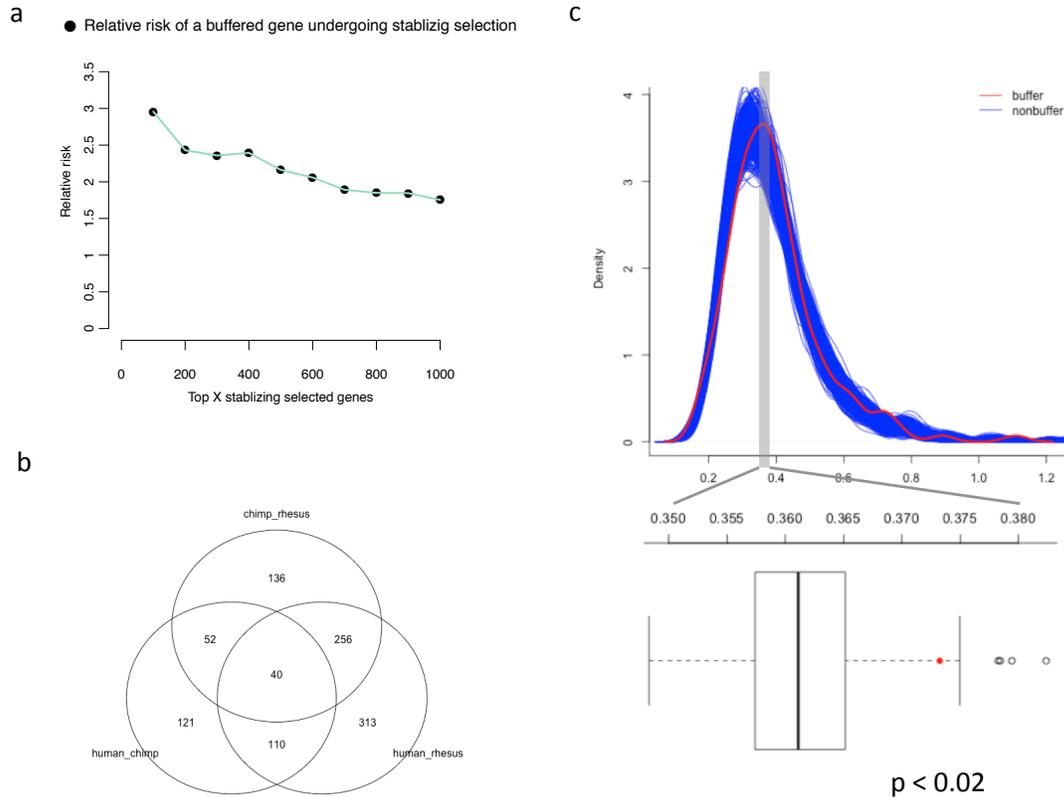


Figure 17. Post-translational gene expression buffering is conserved in primates. (a) Scatter plot showing odds for human-chimpanzee buffered gene to also be top stabilizing selected genes in primates. (b) Venn diagram showing overlaps of buffered genes at 1% FDR amongst human-chimpanzee, human-rhesus and rhesus-chimpanzee comparisons out of total 3188 genes quantified (c) Top: density plot of standard deviations of mRNA levels from YRI individuals comparing between buffered genes (red) and 500 sets of subsampled background genes of matching expression levels (blue). Bottom: Box plot summarizing the median of each subsampled standard deviation distribution. The red dot labels the median of standard deviation of buffered genes. The empirical p-value represents the probability of observing a median standard deviation greater than or equal to the buffered gene median standard deviation by chance.

4 Mapping cell populations in flow cytometry data for cross-sample comparison using the Friedman-Rafsky test statistic as a distance measure

## 4.1 Abstract

Flow cytometry (FCM) is a fluorescence-based single cell experimental technology that is routinely applied in biomedical research for identifying cellular biomarkers of normal physiological responses and abnormal disease states. While many computational methods have been developed that focus on identifying cell populations in individual FCM samples, very few have addressed how the identified cell populations can be matched across samples for comparative analysis. This Chapter presents FlowMap-FR, a novel method for cell population mapping across FCM samples. FlowMap-FR is based on the Friedman-Rafsky nonparametric statistic (FR statistic), which tests the equivalence of multivariate distributions. As applied to FCM data by FlowMap-FR, the FR statistic objectively quantifies the similarity between cell populations based on the shapes, sizes, and positions of fluorescence data distributions in the multi-dimensional feature space. To test and

evaluate the performance of FlowMap-FR, we simulated the kinds of biological and technical sample variations that are commonly observed in FCM data. The results show that FlowMap-FR is able to effectively identify equivalent cell populations between samples under scenarios of proportion differences and modest position shifts. As a statistical test, FlowMap-FR can be used to determine whether the expression of a cellular marker is statistically different between two cell populations, suggesting candidates for new cellular phenotypes by providing an objective statistical measure. In addition, FlowMap-FR can indicate situations in which inappropriate splitting or merging of cell populations has occurred during gating procedures. We compared the FR statistic with the symmetric version of Kullback-Leibler divergence measure used in a previous population matching method with both simulated and real data. The FR statistic outperforms the symmetric version of KL-distance in distinguishing equivalent from nonequivalent cell populations. FlowMap-FR was also employed as a distance metric to match cell populations delineated by manual gating across thirty FCM samples from a benchmark FlowCAP data set. An F-measure of .88 was obtained, indicating high precision and recall of the FR-based population matching results. FlowMap-FR has been implemented as a stand-alone R/Bioconductor package so that it can be easily incorporated into current FCM data analytical workflows.

## 4.2 Background

As the most mature single cell analysis technology, flow cytometry (FCM) has been widely applied in the diagnosis and characterization of cancers, infectious diseases, neurological disorders, immune system diseases, and hematological disorders <sup>75</sup>. In a typical FCM study, tens to thousands of blood or tissue samples are processed to quantify cellular characteristics (e.g., protein expression levels) in individual cells. A modern polychromatic flow cytometer can measure up to 27 cellular characteristics for millions of cells in each sample <sup>76</sup>. To characterize and differentiate FCM samples from different experimental conditions/perturbations, cell populations need to be identified and their variations across samples need to be quantified and assessed. For example, regulatory T cells are known to suppress a variety of pathological and physiological immune responses. In peripheral blood of individuals with autoimmune disease, regulatory T cells tend to exist in smaller proportions than in healthy controls <sup>77</sup>. Cell populations may also differ because of an individual's inherited biological traits. For example, immunoglobulin E (IgE) is an antibody that is elevated when the immune system overreacts to environmental allergens, such as pollen. In individuals that are predisposed to allergic responses,

elevated numbers of circulating B cells expressing the high affinity IgE receptor (CD23) can be found <sup>78</sup>.

Historically, manual gating has been used as the methodology of choice to delineate cell populations sharing common characteristics in FCM data. This graphically driven approach relies on the sequential application of manually drawn boundaries (i.e., gates) to distinguish cells on uni- or bi-axial data plots. The placement of manual gating boundaries is subjective and depends on the experience of the data analyst. In recent years, computational gating methods have made significant advances in identifying cell populations at the individual sample level <sup>79</sup>. Model-based computational gating approaches, such as Gaussian and multivariate skew-*t* mixture model fitting <sup>80-82</sup>, employ statistical assumptions on the shape and location of cell population distributions. Non-model based methods, such as grid-based density clustering <sup>83</sup> and spectral clustering <sup>84</sup> algorithms, group cells into homogeneous populations based on unsupervised data clustering.

After cell populations are identified in individual samples, the next step is to map cell populations between samples so that cell population characteristics, such as marker expression levels and proportions, can be compared across the sample set. In manual gating approaches, the gating boundaries drawn on one sample are often directly applied to another sample. However, marker expression

levels of equivalent cell populations can shift between different samples due to technical artifacts and natural biological variability. Technical artifacts can be unintentionally introduced during data acquisition, especially in multicenter clinical studies where samples are prepared at several sites, with slight differences in sample preparation procedures, staining protocols, and instrument settings.

Biological variability in marker expression can occur due to the complex interplay of genome sequence polymorphisms, especially in outbred human populations.

Indeed, the effects of technical artifacts and biological variation can be difficult to distinguish. These inherent sources of variability in marker expression make the cell population mapping step using direct application of manual gating boundaries problematic for cross-sample comparisons.

To our knowledge, there is no stand-alone method implementation focused solely on cell population matching. Probability binning <sup>85</sup> is able to compare multivariate distributions between FCM samples but it remains unclear how it can be adapted to compare population-level data as cell populations frequently shift expression distributions across samples. Finak et al. <sup>86</sup> compared sample level variability in cell population marker expression among fluorescent channel transformation methods (e.g., bi-exponential or generalized Box-Cox). Variation between cell population locations is defined as the sum of squared deviations in the

cell population locations (mean marker expression levels) across FCM samples.

Small inter-sample variations in cell population locations are associated with low population misclassification rates.

Other existing approaches, including FLAME<sup>80</sup>, HDPGMM<sup>82</sup>, JCM<sup>81</sup>, and flowMatch<sup>87</sup> bundle the cell population identification method and cross-sample mapping function together, with the mapping component operating under the principle of global template finding. In FLAME<sup>80</sup>, mapping cell populations across samples is the last step of their computational gating method. Each sample is modeled as a mixture of cell populations, each with a multivariate skew- $t$  distribution. The modes of cell population distributions are pooled together across samples to establish a global template of cell populations, marked by their mode locations. The sample cell populations are then matched to the global populations based on the respective mode locations. In both HDPGMM<sup>82</sup> and JCM<sup>81</sup>, a multilevel modeling approach is applied to simultaneously identify cell populations and map populations across samples. A global template is generated based on shared location and shape characteristics among cell populations across samples in the same cohort. JCM ascribes multivariate skew- $t$  distributions to the cell populations as in FLAME, while the HDPGMM assumes Gaussian distributions for the cell populations. Both methods perform the mapping step automatically while

the population is being identified, which precludes their implementation with other data clustering methods. When a new sample is added to the data set, HDPGMM needs to be re-run on all samples to generate a new hierarchy that could be different from the original one even for the same sample. JCM directly compares the new sample with the population location and shape parameters at the cohort-level using the Kullback-Leibler divergence measure (KL distance).

In flowMatch <sup>87</sup>, samples are organized into a hierarchy based on overall shape similarity between sample cell populations. The KL distance is also used to quantify the multivariate similarity between cell populations. Two samples are merged together under the hierarchy when the total between-sample KL distance is minimized. The root of the hierarchy is the global template of cell populations. These existing methods all require the composition of a global template, which can be error-prone without careful selections of mapping thresholds at each comparison. The construction of the template is also very sensitive to the clustering or gating procedures. Some of these methods do not calculate the degree of similarity between cell populations, limiting the ability to map heterogeneous cell populations across samples.

Here we describe a novel method, FlowMap-FR, that uses a data-driven approach for cell population mapping. FlowMap-FR directly compares cell

populations between samples using the Friedman-Rafsky (FR) test statistic (FR statistic, <sup>88</sup>) – a nonparametric multivariate statistical measure utilizing a minimum spanning tree approach to describe the “ordering” of values in multidimensional space. The FR statistic has been used as a similarity measure in statistical pattern recognition <sup>89</sup>, image retrieval <sup>90</sup>, and image registration <sup>91</sup>. The basic principle is to “sort” the events from any two merged cell populations (e.g. cell populations in different samples being tested to determine if they are equivalent) based on edge connections in a minimum spanning tree constructed from the marker expression levels of each cell. The cell populations being compared are considered to be equivalent if their respective member events are randomly dispersed in the tree, and are different if the events of the same membership tend to congregate in different branches of the tree. Thus, FlowMap-FR evaluates cell population similarity by computing a statistical distance measure for every possible population pair in a cross-sample mapping problem.

FlowMap-FR is a stand-alone method that can be applied to mapping cell populations delineated by manual gating or computational clustering procedures.

We evaluated the performance of FlowMap-FR in simulation experiments designed to mimic commonly observed scenarios of sample variability for mapping cell populations in which differences in cell population proportions occur between

samples, slight differences in marker expression levels in equivalent populations occur between samples, and a cell population in one sample is inappropriately divided into two by over-partitioning in comparison with another sample. We also compared the performance of FlowMap-FR with the symmetric version of KL distance used in flowMatch <sup>87</sup> using both simulated and real data, and applied it to match gated populations from a FlowCAP benchmark data set [5].

## 4.3 Methods

### 4.3.1 Terminology

In a given FCM experiment, the levels of a number of different quantitative markers (features) are measured in individual cells. Each cell can then be represented as a feature vector of marker levels in  $d$ -dimensional space. A cell population is defined as a homogenous group of cells sharing similar quantitative levels for all markers measured, and can be delineated by manual or computational gating methods as a feature vector cluster in multidimensional space. The number of features evaluated in the FCM experiment is equivalent to the number of dimensions of the multivariate vector. When comparing two cell populations, the events from the different populations can be combined to form pooled data. A

graph can be constructed on the pooled data, where the nodes represent the cell events and the edges represent the Euclidean distance between the multivariate feature vectors.

### 4.3.2 Overview of FlowMap-FR

[Figure 18](#) shows a hypothetical FCM assay of 4 expression markers (CD4, CD45RA, SLP76, ZAP70) for two different biological samples. The goal of cross sample comparison is to determine if either Cell Population (CP)red or CPgreen in sample B is equivalent to the CPblue reference cell population in sample A. The bi-axial plots indicate similarity between CPblue and CPgreen in all expression marker levels except for CD4, while CPblue and CPred are similar in all markers and would therefore be considered to be equivalent. The goal of any quantitative method for cross sample comparison would be to accomplish cell population mapping by objectively making this distinction using cell populations delineated by any data transformation procedure or gating method, including manual gating or algorithmic clustering.

The FlowMap-FR method for cross sample mapping described here utilizes the Friedman-Rafsky multivariate generalization of the Wald-Wolfowitz run statistic for comparing two data distributions to determine if they have been

sampled from the same global data population. Wald and Wolfowitz<sup>92</sup> described a statistical procedure to compare univariate non-parametric distributions by merging the values from two different data sets into an ordered list and quantifying the number of runs that connect values derived from the same data set. The number of runs is thus associated with the tendency of the values to cluster together according to their respective membership in the data sets. A small number of runs connecting values from the same data set suggests that the values have been sampled from more than one distribution. Friedman and Rafsky<sup>88</sup> proposed a generalization of this approach for multivariate data in which value order is determined based on proximity in a minimum spanning tree constructed in multivariate space.

The basic idea is to connect the events across the two cell populations to be compared according to their similarity in expression of all  $d$  markers using a minimum spanning tree. The individual cell events are represented as nodes on the tree. The distance between two nodes is calculated as the Euclidean distance in  $d$ -dimensional space. The FR statistic quantifies the multivariate similarity of nodes from any two underlying distributions in the minimum spanning tree (MST). The FR statistic also controls for the size of the MST across comparisons and the topological structure of the MST. We calculate the FR statistic comparing each

pair of cell populations. For example, a comparison of two biological samples with  $n_1$  and  $n_2$  cell populations would involve  $n_1 n_2$  total comparisons.

FlowMap-FR estimates the FR statistic based on controlled statistical sampling of the events in data pooled from the two cell populations being compared. Each controlled statistical sample taken from this pool is comprised of events sampled to be proportional to those in the original cell population pair (above some minimum number of events). This controlled statistical sampling approach is employed because the computing time for calculating a minimum spanning tree is dependent on the number of nodes, that is, total number of events involved in a cell population comparison pair.

#### *Finding the Minimum Spanning Tree*

We begin by mixing events from two cell populations under comparison while keeping track of their population membership. The mixture of events is henceforth referred to as the *pooled data*. We then take  $S$  controlled statistical samples of  $N$  events randomly selected from the pooled data, without replacement. Each controlled sample maintains a constant ratio of events from the two cell populations, calculated from the pooled data before any event selection. For every controlled statistical sample, we compute the Euclidean distance between every

pair of events based on the expression level of all markers. A complete weighted undirected graph is constructed based on the  $N$ -by- $N$  distance matrix. We then use Prim's algorithm [21,22] to find the minimum spanning tree (MST) on the graph. The distance between events on the minimum spanning tree corresponds to the dissimilarity of their marker expressions in  $d$ -dimensional space. Therefore, events with similar marker expression levels are placed near each other on the MST branches.

#### *Friedman-Rafsky Statistic Computation*

Central to the Friedman-Rafsky (FR) statistic are the multivariate “runs” in the Euclidean MST. The multivariate runs are the set of subtrees in the MST consisting of connected events from a single cell population. For each controlled sample MST based on  $N$  events, we remove the edges connecting events derived from different cell populations. Because any removal of an edge in an MST breaks the tree into two disjoint subtrees, the number of subtrees in an MST is equal to the number of removed edges ( $G$ ) plus 1. Thus, the number of multivariate “runs”  $R$  is equal to  $G + 1$ .

The FR statistic compares the observed with the expected number of multivariate runs in a given MST from two equivalent population distributions,

then standardizes the difference by the variance of the multivariate runs. Given two cell populations  $X$  and  $Y$  of event sizes  $m$  and  $n$ , respectively, with  $N$  total events, the expected number of multivariate runs  $\mathbb{E}[R]$  is equal to one plus the expected number of edges. For the  $N - 1$  edges of the given MST, the probability that an arbitrarily selected edge connecting  $X$  and  $Y$  (or  $Y$  and  $X$ ) is the proportion of events in  $N$  belonging to  $X$ ,  $m/N$  (or  $n/N$  if considering edges connecting  $Y$  and  $X$ ) multiplied by the probability that the edge connects to a node in  $Y$  (i.e.,  $n/(N - 1)$  or  $m/(N - 1)$  if considering  $Y - X$  edges). Hence, the expected number of edges is

$$\mu = E(R) = \frac{2mn}{N} + 1$$

The variance of the number of runs in a given MST is dependent on the corresponding topological feature – the total number of edge pairs sharing common nodes ( $C$ ), which is  $\binom{N-1}{2}$  in a graph of  $N$  nodes. Hence, the variance reflects the range of runs possible given the composition of membership events in a cell population pair comparison, and

$$\sigma^2 = \frac{2mn}{N - 1} \left\{ \frac{2mn - N}{N} + \frac{(C - N + 2)(m + N(N - 1) - 4mn + 2)}{(N - 2)(N - 3)} \right\}.$$

Details of the derivation can be found in Friedman and Rafsky<sup>88</sup>. The FR statistic

( $w$ ) is then defined as

$$w = \frac{R - \mu}{\sigma}$$

The median FR statistic from the  $S$  controlled statistical samples of the pooled data is taken as the estimated measure for similarity between the two cell populations. The estimated FR statistic is multiplied by  $-1$  to compute an FR-based distance measure, where a small value indicates high degree of similarity and a large value indicates high degree of dissimilarity. This FR-based distance measure can then be used with various clustering methods to group cell populations across samples (e.g., by hierarchical clustering).

#### *Hypothesis testing using the FR statistic*

We can also use the estimated FR statistic to perform a statistical test where the null hypothesis is that the two cell populations follow the same distribution. The p-value of the FR statistical test is computed under the assumption that the standardized FR statistic follows a normal distribution<sup>88</sup>. A large p-value is evidence that the cell populations in the comparison are probably similar in their distributions, while a small p-value is evidence that the cell populations are probably different in their distributions. In addition to cell population similarity, the p-value of the FR statistic also depends on the number of controlled statistical samples of the pooled data ( $S$ ) and the size of each

controlled sample ( $N$ ). For the analyses in the current study, we chose a value of  $10^{-7}$  for the p-value threshold of the FR statistic where the number of true positive matched cell population pairs is maximized and the number of false positive cell population pairs is minimized. The threshold was chosen based on sampling parameters  $N = 200$  and  $S = 200$ . Details of choosing the sampling parameters are described in the following section.

### *Sampling Parameters*

In order to both reduce runtime and provide for consistent statistics, data sampling is necessary before the FR statistic is applied. With the controlled sampling approach, the FR statistic value in a pairwise comparison between two cell populations depends on the number of controlled samples ( $S$ ), and the number of events in the pooled controlled sample ( $N$ ). We assessed the precision (reflecting the extent of reproducibility of the controlled sampling procedure) and accuracy (indicating the biasedness of the estimated statistic as a function of the sampling procedure) of the FR statistic under  $N = 100, 200, 400$  and  $S = 100, 200, 400$ .

While increasing  $S$  results in a larger range of the FR statistics, the ranks of the population pairs remain the same, and the FR statistic value increases as  $N$  increases. Nonetheless, when mapping cell populations across two samples, the ranks of the population comparisons remain the same across varying  $N$  and  $S$ . The

time complexity of the controlled sampling approach is assessed under  $N = 50, 100, 200, 400, 600, 800,$  and  $1000$  and  $S = 100, 200,$  and  $400$ . The time complexity increased quadratically in the number of the events, but did not change across the number of controlled samples. FlowMap-FR computes an adjacency matrix of similarity between  $N$  events based on Euclidean distance between the  $d$ -dimensional measurement vectors. Prim's algorithm is then employed to find the MST of each controlled sample. The computing time of MST finding is known to increase quadratically in  $N$  in the Prim's algorithm when the similarity between the events are represented in an adjacency matrix. In summary,  $N$  and  $S$  are chosen to preserve the ranks of the FR statistics with balanced precision and accuracy as well as minimized computing time of MST.

### *Data preprocessing*

FlowMap-FR is designed as a stand-alone algorithm that can be applied to mapping cell populations derived from any gating procedures or any normalization methods. The input data contains ASCII files with each cell population's labels and marker expression levels derived from manual gating or automated gating method. Before applying FlowMap-FR, the user needs to choose a transformation method to transform the raw data into equivalent quantitative ranges based on their data formats and use cases. We employed FCSTrans<sup>95</sup> for all the analyses in this paper.

### 4.3.3 Simulation study

#### *Design*

In order to assess the performance of FlowMap-FR under a variety of different population mapping scenarios, a simulated data set was constructed to closely mimic real FCM data. Data from FCM experimental samples vary in distributional shape depending on the sample's biological characteristics. In order to conduct a fair performance evaluation of our cell population mapping method, we sought to mimic possible sources of experimental sample variability in cell population characteristics. We selected a real data set that includes cell populations possessing features inherent to FCM data: sparseness of some populations but not others, skewness in the distribution of some cell population markers, high correlation between expression levels for a subset of markers, and flat density distributions for some markers. The real data were derived from an FCM experiment in which human peripheral blood was assayed with a four marker panel: CD14, CD23, CD3, and CD19 <sup>96</sup>. The FLOCK clustering algorithm <sup>83</sup> was used to identify nine distinct cell populations (CP1:CP9) in an FCS data file from one sample totaling 20,000 events.

Multivariate skew- $t$  distributions were employed to extract location, variance and skewness parameters of each cell population in the real data. The estimated parameters of the fitted skew- $t$  distribution were then used to simulate a new data set that mimics the marker distributions observed for each reference cell population. [Figure 19A](#) and [Figure 19B](#) show cell population distributions for selected markers in the real sample and in the simulated sample, respectively. An important FCM data feature is that some cell populations may overlap in some marker channels while being well separated in other dimensions. This can be illustrated by CP3 and CP9. Their marker expression levels are overlapping and correlated in the two-dimensional scatter of CD23 and CD14 and also between CD3 and CD14. However, they are well separated based on CD19 marker expression levels.

### *Evaluation Scenarios*

In FCM data, cell populations may exhibit slight shifts in marker levels between biological samples or vary in the percent composition per sample between individuals or cohorts (i.e., varying proportions). The simulated data set was used to construct a series of test samples designed to mimic these real scenarios in cross sample comparison challenges to test the cell population mapping performance of FlowMap-FR, as follows:

*Scenario 1. Differences in cell population proportions between biological samples*

([Figure 19C](#)). Test samples were constructed in which cell proportions were changed to 1%, 10%, 25%, 50%, 75%, 125% and 150% of the original simulated cell population. Location and shape of the changed cell populations was maintained as in the original simulated cell populations. Each changed cell population was a statistical sample of events randomly generated with the same location and shape parameters as the original simulated cell population.

*Scenario 2. Differences in cell population numbers between biological samples.* Test samples were constructed in which one of the simulated cell populations from one biological sample was removed. The resulting test sample containing eight cell populations was compared with the original simulated sample containing nine cell populations.

*Scenario 3. Shifts in marker expression levels between biological samples* ([Figure](#)

[19D](#)). Test samples were constructed in which the simulated cell population from one biological sample is shifted along each marker channel one at a time. The unit of location shift is standardized for each cell population and defined as the

interquartile range ( $IQR_{o,i}$ ) of the original simulated cell population along channel

$i$ , where  $IQR_{o,i} = 75^{\text{th}}$  percentile –  $25^{\text{th}}$  percentile of the original simulated cell

population in channel  $i$ . For each of the nine cell populations, we simulated 1, 2, 3,

4 and 5  $IQR_{o,i}$  shifts along each marker channel. Denote  $\mu_{o,i}$  as the original location of a cell population along channel  $i$  and  $\mu_{2,i}$  as the shifted location after 2 units of interquartile range shift. Then,  $\mu_{2,i} = \mu_{o,i} + 2 * IQR_{o,i}$ .

*Scenario 4. A discrete cell population in the reference sample inappropriately divided into two in the test sample by over-partitioning (Figure 19E).* Test samples were constructed in which the single simulated cell populations from the reference sample were divided into two partitions along the CD23 channel. Two sets of partition samples were simulated accordingly, that include the upper and lower partitions above and below the 90<sup>th</sup>, 80<sup>th</sup>, 70<sup>th</sup>, 60<sup>th</sup>, 50<sup>th</sup>, 40<sup>th</sup>, 30<sup>th</sup>, 20<sup>th</sup>, and 10<sup>th</sup> percentile of the corresponding CD23 levels.

#### 4.3.3.3 Comparison with the Kullback-Leibler Divergence Measure

The results using the simulated scenarios was also evaluated using the symmetric version of Kullback-Leibler (referred to be SKL distance to distinguish from the original KL distance) divergence measure used in flowMatch to compare cell populations across multiple samples<sup>87</sup>. The KL distance is known as an asymmetric distance measure between two distributions such that the values comparing CP1 to CP2 and CP2 to CP1 can differ. In flowMatch, a symmetric version of the KL distance (SKL) is employed under which the cell populations are

assumed to follow multivariate normal distributions. Using this version, the KL distance is a function of means and variances of the cell populations. The SKL distance is achieved by averaging of the two possible KL values in a single comparison. Given two cell populations  $i$  and  $j$  in  $d$ -dimensional feature space, the KL value of comparing  $i$  against  $j$  is:

$$\frac{1}{2} \left\{ \log \frac{|\Sigma_j|}{|\Sigma_i|} + \text{Tr}(\Sigma_j^{-1}\Sigma_i) + (\mu_j - \mu_i)^T \Sigma_j^{-1} (\mu_j - \mu_i) - d \right\},$$

where  $\mu_i$  and  $\mu_j$  are  $d$ -dimensional mean vectors of the expression markers of cell populations  $i$  and  $j$ , respectively, and  $\Sigma_i$  and  $\Sigma_j$  are  $d$ -dimensional variance-covariance matrices of the markers for cell populations  $i$  and  $j$ , respectively. We computed sample means and variances to approximate  $\mu$ 's and  $\Sigma$ 's to calculate the SKL distance values of the simulated cell populations as in the flowMatch implementation.

#### 4.3.4 Real flow cytometry samples

FlowMap-FR was applied to two real flow cytometry data sets to evaluate its performance in mapping cell populations across multiple real flow cytometry samples. The first evaluation investigated the ability of FlowMap-FR to map cell populations that are known to be biological replicates across FCM samples. The second evaluation applied FlowMap-FR to FCM samples collected from thirty

different healthy individuals. In this data set the individual cell populations within each sample and their equivalence between samples were delineated by expert manual gating as part of the FlowCAP challenges <sup>79</sup>.

*Real FCM data set #1.* The first real data set evaluation included four FCM samples of peripheral blood mononuclear cells (PBMC) collected from two healthy individuals <sup>97</sup>; the blood sample from each individual was divided into two biological replicates. Each sample was stained with four fluorophore-labeled antibodies (marker panel: CD3, CD4, CD8, and CD19). Four cell populations were identified in each of the FCM samples using K-means clustering (parameter setting: minimum 4 and maximum 20 clusters). The K-means convergence criteria were set to minimize within-cluster sum of squares while maximizing between-cluster sum of squares. In order to perform cell population mapping across multiple samples, we computed the estimated FR statistics and FR-based distance metric (FR multiplied by minus one) for all population pairs across the four FCM samples. The FR-based distances were used as a similarity measure to group and map equivalent cell populations across samples. Hierarchical clustering with complete linkage was employed as the clustering method of choice. The cell populations were arranged in a hierarchy according to the FR distance to the other cell populations.

The FR-based mapping approach was compared to flowMatch<sup>87</sup>. flowMatch performs agglomerative clustering that repeatedly merges samples to form a template sample until there are no more samples to be added (meta-clustering). First, a template sample is created from merging the two most similar samples, and the matched cell populations are combined to form a cell population. Then, a template sample is compared to all the other samples and merged with the sample that is most similar. This step continues until there are no more samples to be compared with. At each step, cell populations are mapped across samples when the two samples are merged into one template sample. A bipartite graph algorithm is employed to match two samples or to match a sample with a template sample when the sum of distances between cell populations is minimized. The performance of the FR statistic was compared with the SKL divergence metric within the flowMatch algorithm.

*Real FCM data set #2:* This normal donor data set was one of the benchmark data sets included as part of the FlowCAP-I challenge<sup>79</sup> and contains manually gated cell populations that can be used as a benchmark for evaluating the performance of FlowMap-FR. A total of thirty FCM samples from normal healthy donors are included in the data set. Each sample was stained with a cocktail of ten fluorochrome reagents, interrogating both cell surface and intracellular proteins.

Expert manual gating by the data providers delineated 8 cell populations in each sample. To perform cell population mapping across multiple samples, we computed the estimated FR statistics for all population pairs across the thirty FCM samples (28,680 comparisons). Similar to real data set #1, hierarchical clustering with complete linkage was employed in order to organize the cell populations in a hierarchy according to FR distance. Based on the FR similarity hierarchy, cell populations were classified into eight sets of equivalent cell populations. The F-measure approach was then used to evaluate the combined precision and recall performance of the new cell population labels in comparison with the cross sample equivalence determined by the original data providers.

## 4.4 Results

### 4.4.1 Simulation study

*Matching with differences in cell population proportions between samples.*

The goal of cross-sample comparison is to match equivalent cell populations across multiple samples. In some circumstances, a given cell population can exhibit dramatic differences in proportions in different biological samples, especially cell populations that have been observed to be predictive cellular biomarkers of

immunological responses, disease states or therapeutic responses. In order to determine how robust the FR statistic would be to matching cell populations in scenarios in which the proportions of the population differ between biological samples, a total of eight test samples were generated to contain from 1% to 150% of the population's cell count in the original simulated reference sample for each of the nine cell populations separately. The original cell counts ranged from 324 events in CP9 to 7,380 events in CP7. Thus, in the case of CP9, the 1% proportion simulation contained as few as 3 events to be matched to the original cell population's 324 events. For CP7, the 150% proportion simulation contained over 11,000 events to be matched.

[Figure 20A](#) - I shows the estimated FR statistics for each population comparison. An FR value closer to zero indicates a higher degree of similarity between the cell populations being compared. [Figure 20A](#) shows the comparison across changing proportions of CP1. The changed CP1 in the test sample is consistently rated as more similar to the equivalent CP1 population in the reference sample than to the other cell populations, even when the proportion of CP1 in the test sample is 1% of the equivalent cell population in the reference sample. The results are similar for all cell populations. Given a selected cell population comparison, the FR statistics are fairly stable across changed

proportions. The situations in which the differences in the FR statistics between correctly matched and incorrectly matched populations are the smallest appear to occur when comparing the most rare populations (CP3 and CP9) with the most abundant population (CP7) ([Figure 20C](#) and [Figure 20I](#), respectively). But even in these situations, differences of approximately 3 FR units are observed between correct and incorrect matching, with the correct mapping still rated as most similar based on the FR statistic. We also computed the p-values of the FR statistics for each population comparison. A small p-value of the FR statistic indicates a potential mismatched pair of cell populations, while a large p-value of the FR statistic suggests a potential matched pair of cell populations. The cutoff for p-value was fixed across all three simulation scenarios to be  $10^{-7}$ . Similar to the results using the FR statistics, the p-values also distinguish between correctly matched and incorrectly matched populations. At p-value cutoff of  $10^{-7}$ , the FR test correctly matches the changed cell populations to their equivalent parent cell populations in the reference sample.

We also compared cell populations across the test samples of CP5 changed proportions to demonstrate the utility of FR statistic in a multiple-sample comparison scenario. In [Figure 20J](#), a total of 72 x 72 FR statistics are displayed in a heat map (comparing 8 test samples of 9 cell populations) and listed in order of

population proportions within each cell population block (e.g., CP5 block consists of 1% (#1.5), 10% (#2.5), 25% (#3.5), 50% (#4.5), 75% (#5.5), 100% (#6.5), 125% (#7.5), and 150% (#8.5)) The cell populations that were not changed in proportions are mapped to each other (i.e., FR statistics closer to zero) as expected. Although the 1% CP5 (#1.5) is slightly more similar to the other cell populations compared to the other changed CP5s, the 1% CP5 is still ranked as more similar to other CP5s under changed proportions than to any other cell population.

*Matching with differences in cell population numbers between samples.*

In some cross-sample comparison scenarios, differences in the numbers of cell populations detected in different samples might be expected. This could occur when comparing samples from normal healthy subjects with samples from diseased subjects in which a new abnormal cell population might be present (e.g. in leukemia or lymphoma patients, or in situations where stimulated and unstimulated samples are compared). [Figure 20A](#) - I show the estimated FR statistics for each population comparison. In this scenario there would be no comparison for one of the cell populations (e.g. CP1) in the reference sample since that population has been removed from the test sample. Because the comparison performed by FlowMap-FR occurs on a population-by-population basis, the FR

statistic values for comparisons between the incorrect cell populations in the test sample and the extra population in the reference sample is essentially the same as if the extra population was still present in the test sample. Thus, for a missing CP1 population, all FR statistics values would correspond to a run value of 1 and the curves would be located at the bottom of the graph in the comparison depicted in [Figure 20A](#). While all of the correct pairwise comparisons would give FR statistic values close to 0 for eight out of eight comparisons, the extra population would only give low FR values (e.g.  $< -10$  in most cases), similar to what is observed for incorrect comparisons. Thus, establishing a lower threshold of  $\sim -5$  would indicate that any population without a value above  $-5$  would indicate a unique population in one of the samples.

*Matching with shifts in marker expression levels between samples.*

Shifts in marker expression between cell populations across a set of experimental samples can occur due to natural biological variability in genetically diverse populations, cell differentiation response to some perturbation, or technical variability associated with differences in staining procedures or reagent lots. In order to determine how the FR statistic would respond to shifts in marker expression, the nine cell populations were mapped to themselves and other cell populations in simulated scenarios in which one population (CP4) was shifted in

position along each of the four dimensions. Shifting was standardized with respect to the range (IQR) of the cell populations' distributions along the selected dimension. A total of 6 test samples were generated with shifting positions along each dimension. The results of CP4 mapping are shown in [Figure 21](#). The distribution of CP4 is narrow along CD19 and is wide along the other three dimensions. As the position shift increases along CD19 (with the distribution in all three other dimensions kept the same), the FR statistic for matching with the original CP4 initially drops linearly with the degree of shifting and then plateaus ([Figure 21A](#)). Thus, FlowMap-FR could be used to determine statistically meaningful shifts in marker expression within a cell population.

In some cases, the shifted cell populations can also become more similar to other cell populations than to the corresponding population in the reference sample depending on the marker expression characteristics of the other populations. When CP4 is at a 2-IQR unit shift away from the original position in the CD19 dimension, the FR value in comparison with itself in the reference sample is about -10 ([Figure 21A](#)) and with CP5 is about -6 ([Figure 21B](#)). This indicates that CP4 in the test sample has become more similar to CP5 in the reference sample at a 2-unit shift. Indeed, the distribution of CP4 in the CD19 and CD23 dimensions overlaps closely with CP5 at a 2-unit shift (see the scatter plots in [Figure 21B](#)).

However, even when substantial overlap was achieved between the shifted CP4 and the original CP5 at a 2-unit shift along the CD19 dimension, the FR statistic was still below -5 due to differences in distributions between CP4 and CP5 in other dimensions. Similar observations were made for CP8 ([Figure 21C](#)), which also overlaps with CP4 at 2-unit shifts in the scatter plots. However, the comparison of CP8 with the shifted CP4 produced only a modest increase in the FR statistics to -11 since CP8 and CP4 are still quite different in shape and coverage of the feature space.

Complete pairwise comparison results of the 6 CP4-shifted samples are shown in [Figure 21F](#). With respect to relative shifts of 0 or 1 IQR units, CP4s are more similar to themselves than to the other cell populations. The 3-unit shift CP4 is more similar to the 2-unit shift CP4, and the 4-unit shift CP4 is more similar to the 3-unit shift and 5-unit shift CP4, etc. We computed the p-values of the FR statistic for the mapping of each cell population to its original parent cell population under shifts in marker expression levels. The cutoff for p-value was fixed at  $10^{-7}$  across all three simulation cases.  $-\log_{10}$  p-values of the FR statistics increase linearly as the shift in marker position increases. Based on these results, using a p-value threshold of  $10^{-7}$  would general provide robust mapping of

populations with slight shifts ( $<1$  IQR) in the expression of one of the cell surface markers between samples.

*Matching with over-partitioning of cell populations in some samples.*

During cell population identification, certain cell populations might be inappropriately divided into two (over-partitioning) depending on the method and configuration parameters used, even though there is no real evidence that the two partitions correspond to distinct cell populations. In order to determine how FlowMap-FR would handle over-and under-partitioning, we artificially partitioned each of the nine cell populations above and below a range of selected percentiles along the CD23 expression level axis. A total of 18 test samples were generated for each cell population, consisting of its corresponding partitions (9 samples each for partitions above and below the percentile cutoffs, from 10% to 90%; see Methods for details); the other cell populations remained unchanged.

[Figure 22A](#) - I shows the mapping results of the nine cell populations. In [Figure 22A](#), the two sets of estimated FR statistics computed when mapping the two partitioned CP1s in the test sample to the unchanged CP1 in the reference sample are significantly larger than those obtained in comparison with the other cell populations. The test correctly mapped both partitioned cell populations to the

unchanged reference cell population across varying partitions. Results also show FR statistics increases when the partition size increases, so the two lines of FR-statistics representing the two partitions cross as one partition increases size and the other decreases size. That the two lines are not completely symmetric is due to the non-symmetric distribution of skew- $t$  data simulation. The same pattern is found for the other eight cell populations in Figures 5B – I. Therefore, FlowMap-FR is able to quantify similarity of inappropriately partitioned subpopulations with the original cell population in the reference sample and could therefore be used to detect and correct over-partitioning that could arise from manual gating or algorithmic clustering. We also computed the p-values of the FR statistics for each cell population comparison. Similar to the results of FR statistics, selecting a p-value threshold of  $\sim 10^{-7}$  distinguishes between correctly matched and incorrectly matched population partitions. [Figure 22J](#) displays the complete pairwise comparisons of CP5 partitions above cutoffs along CD23 expression levels with itself and other cell populations. From this heat map, it is clear that all 10 partitions of CP5 are more similar to the unpartitioned CP5 in the reference sample based on the FR statistic since all of the squares in the central CP5 vs. CP5 box have a higher FR statistic (more blue) than any other comparisons of cell populations against CP5.

*Matching populations using the symmetric KL divergence (SKL) measure*

[Figure 23A - I](#) shows the results of matching CP1, CP2, and CP3 with all other cell populations using the SKL distance under scenarios in which differences in cell proportions occur between biological samples (A - C) and under scenarios in which a discrete cell population in one biological sample was inappropriately divided into two by over-partitioning of the data from another biological sample (D - I). Under the scenario in which differences in cell proportions occur between biological samples, the SKL distance value is always close to zero when matching cell populations of varying proportions in the test sample to corresponding cell populations in the reference sample. However, the differences between the SKL distance values of equivalent and nonequivalent cell populations are not as large as those between the FR statistics (compare [Figure 23 A - C](#) with [Figure 20](#)). Thus, it could be difficult to use the SKL distance value to distinguish mapping from non-mapping cases. In theory, the SKL distance could be used for population mapping by choosing the best matched cell population. But if the cell population to be mapped does not occur in the test sample, mapping to the best matched population without considering the similarity value would give an incorrect result. Similar phenomena were found under scenarios in which a cell population is inappropriately divided into two by over-partitioning ([Figure 21D - I](#)). The SKL

distance performs poorly when comparing CP1 partitions below 10th, 20th, and 30th percentiles in the test sample with CP1 in the reference sample ([Figure 23D](#) and G). In fact, the SKL distance values are smaller when comparing these CP1 partitions to CP4 in the reference sample than when comparing to the reference CP1.

For scenarios in which cell populations are shifted in marker expression levels, the SKL distance performs in a similar way to the FR statistics, i.e. linearly changing values with increasing shifts.

#### 4.4.2 Mapping cell populations across multiple samples in real data

##### *Real FCM data set #1*

[Figure 24A](#) – B shows the results of matching cell populations across four real FCM samples using the FR-based distance measure. Samples 1 and 2 are the biological replicates of the blood sample from the first subject, and Samples 3 and 4 are biological replicates of the blood sample from the second subject. In [Figure 24A](#), the FR-based distance measure ( $-1 * \text{FR statistic}$ ) was computed for each cell population pairwise comparison and displayed in a heatmap. The blocks colored in blue suggest matched population pairs, and the blocks colored in red suggest mismatched population pairs. We employed hierarchical clustering method with

complete linkage to match cell populations across samples, using the FR-based dissimilarity measure as the distance metric. The cell populations were grouped into four sets of equivalent cell populations. For example, in the first set of matched cell populations, CP1.2 (Sample 1, CP 2), CP2.4 (Sample 2, CP4), CP4.3 (Sample 4, CP3) and CP3.4 (Sample 3, CP4) were grouped and matched to each other. CP1.2 and CP2.4 belong to biological replicates of one blood sample, while CP4.3 and CP3.4 belong to biological replicates of the other blood sample. In the hierarchical relationship of the cell populations across samples, CP1.2 is more similar to CP2.4 than to CP4.3 or to CP3.4. We observed the same relationship in each set of equivalent cell populations, namely that cell populations belonging to the biological replicates of the same sample are more similar to each other. Similar to the results of FR statistics, selecting a p-value threshold of  $\sim 10^{-7}$  also distinguishes between correctly matched and incorrectly matched populations. We also performed cell population mapping using flowMatch with the FR-based distance measure and using flowMatch with symmetric KL divergence (SKL) as the distance measure. Both versions of flowMatch generated the same cell population mapping results as show in [Figure 24A](#). In [Figure 24B](#), the hierarchical relationship between the samples were computed using flowMatch with the FR-based distance measure. In both versions of flowMatch, biological replicates of the

same blood sample are matched and more similar to each other than the FCM samples that come from different subjects.

| *Real FCM data set #2*

[Figure 25](#) shows the results of matching cell populations across thirty real FCM samples using the FR-based distance measure. The FR-based distance measure was computed for each cell population pair comparison and displayed in a heatmap. The blue blocks with large FR-based distance (low FR statistic values) suggest matched population pairs, and the red blocks with small FR-based distance (high FR statistic values) suggest mismatched population pairs. The cell populations are arranged in a hierarchy based on the FR similarity to all other cell populations. The cell populations were then grouped according to their FR distance as reflected in the structure of the hierarchical clustering tree to generate eight sets of equivalent cell populations. F-measures were computed to evaluate the combined precision and recall of the FlowMap-FR classification method. We obtained an overall F-measure of 0.88, which indicates high agreement between manual gating/mapping and the cell population mapping derived from the FR-based similarity matrix.

## 4.5 Discussion

Mapping of equivalent cell populations across different samples is an essential component of comparative analysis pipelines for cell-based immunoprofiling and biomarker discovery in biomedical research to monitor disease progression and treatment responses. However, the ability to precisely match cell populations is complicated by natural and technical contributions to variation in marker expression values and their distributions. FlowMap-FR directly addresses the cell population mapping challenges that may arise during the FCM data processing workflow without *a priori* assumptions about the marker expression distributions in the different cell populations analyzed, and thus can be readily employed in comparison of skewed, non-parametric, and multi-modal distributions. The method is highly robust, as illustrated in matching cell populations of varying shapes, locations, and correlations between marker features under scenarios of differences in population proportions between samples and modest shifts in marker distributions. Because FlowMap-FR is a stand-alone cell population mapping method, it can be incorporated into any FCM analytical workflow that requires a cell population-matching step.

The mapping approach in FlowMap-FR provides a similarity measure of cell populations under various sample variation scenarios. This similarity measure can be converted into a probability measure assuming that the statistic follows a

normal distribution<sup>88</sup>. The statistic is an objective measure of similarity between data distributions, with values closer to zero reflecting similar data distributions (more precisely, that the two “samples” are derived from a single underlying global data distribution) and values approaching large negative values reflecting different data distributions (that the two “samples” are derived from different underlying global data distributions). However, the choice of whether two cell populations are “equivalent” is somewhat subjective and specific to the experiment in question. To deal with this experiment-specific decision, it is possible to choose a threshold for the statistical value across the comparison pairs to distinguish equivalent (matched) versus distinct (mismatched) cell populations. We have observed larger gaps in the FR statistic values for threshold selection compared with other existing methods, such as the SKL distance. When cell population marker distributions were similar between samples there was an obvious gap in the FR statistic values between matched versus mismatched cell population pairs such that the threshold was relatively easy to identify. Alternatively, an agglomerative clustering method, such as hierarchical clustering, could be applied to identify groupings of cell populations with similar expression profiles.

One challenging population mapping scenario occurs when one sample contains a cell population that is absent from another sample, as might occur when

a novel abnormal cell population arises in a particular disease setting. We found that when there were different numbers of cell populations between the test and reference samples, judicious selection of an FR statistic threshold could reveal the presence of a distinct cell population in one sample that was absent from the other. However, the selection of this threshold could be challenging in some cases. In this scenario, the ideal reference sample for comparison would be one that contains the union of all cell populations found in each of the individual test samples. This composite sample could be generated by concatenating the data from multiple FCS files and running the population identification methods on the concatenated file to identify all cell populations present in each of the individual samples. This composite sample could then be used as a reference for comparison and mapping.

While FlowMap-FR was relatively robust to moderate shifts in marker expression ( $<1$  IQR) that could result from natural biological variability or differences in staining protocols/reagents and instrument configuration settings between experiments and labs, we expect that its performance could be enhanced further by applying a sample alignment procedure to the data before the population mapping step in the FCM data processing workflow. Cell populations observed can be similar in shape and relative location in each sample but different in absolute marker expression levels across samples. Marker expression levels can

be normalized across samples on a per-channel basis <sup>98</sup> before cell population mapping to further improve the results. Many software and computer programs are available for this data transformation purpose, such as flowTrans <sup>86</sup>, FCSTrans (the method used in this paper <sup>95</sup>), FCS2CSV <sup>99</sup>, etc., before mapping cell populations in FlowMap-FR.

However, in some experimental scenarios marker expression shifts reflect important phenotypic changes in the cell population of interest, for example when activation marker expression increases in response to cell stimulation. As a statistical test, FlowMap-FR can be used to determine when the expression of a cellular marker has become significantly different from a comparison population (e.g., using FR values from known different cell types in control samples to determine thresholds). Although the FR statistic cannot determine whether one cell population is functionally different from the other, it provides an objective measure for scientists to identify candidate phenotypes for biological interpretation and validation.

FlowMap-FR was also found to be able to map cell populations that are inappropriately partitioned in a subset of samples. We observed that across varying partitions of cell populations in different samples, FlowMap-FR correctly mapped the partitions to the original cell populations in the reference sample and ranked

the partitions by degree of overlap with respect to the original cell population. The over- or under-partitioning of cell populations is a common artifact in many automatic gating methods. Thus, the FR statistic can also serve as a tuning metric for parameter adjustment during the automated gating process to prevent artificial population splitting or simply as a quality control metric on the gated samples.

Computational efficiency is a major consideration in the analytical workflow of FCM data processing because of the increasingly large quantities of samples, events and markers being evaluated. The bottleneck of FlowMap-FR computations lies in finding the minimum spanning tree (MST) in order to compute the FR statistic. We used Prim's algorithm to compute the MST, in which the computational complexity increases quadratically in the number of nodes, i.e., the number of events in the graph. To circumvent the runtime limitation, we implemented a controlled random sampling procedure to estimate the FR statistic for each cell population pair comparison. The random sample procedure achieved good precision and accuracy in estimating the true FR statistic. Moreover, we parallelized the estimation procedure so that the users may choose to perform the analysis on as many cores as their computing environment allows. For a single cell population comparison in FCM samples with four feature markers, the runtime on a 10 core system is ~10 times faster than the run time on a single core system. In

the future, the runtime can be further improved by parallelizing the sequential computations of multiple sample comparison of cell populations.

We have implemented FlowMap-FR in R as a Bioconductor package (<http://www.bioconductor.org/packages/devel/bioc/html/flowMap.html>). We are also in the process of implementing and incorporating FlowMap-FR into the GenePattern FCM suite <sup>100</sup> and the bioKepler workflow platform <sup>101</sup> so that it can be used along with other FCM data processing and analytical methods that have been deployed in these platforms. A common FCM computational workflow consists of four steps: data transformation and preprocessing, computational identification of cell populations, sample alignment, and cross-sample comparison of cell populations. While there have been a large number of methods developed for the transformation and identification steps, only a few methods are available for the sample alignment and cross-sample comparison steps. FlowMap-FR provides a robust non-parametric probability-based solution to these workflows, facilitating the move towards the next paradigm for result interpretation across samples and objective performance evaluation of workflows.

## 4.6 Figures

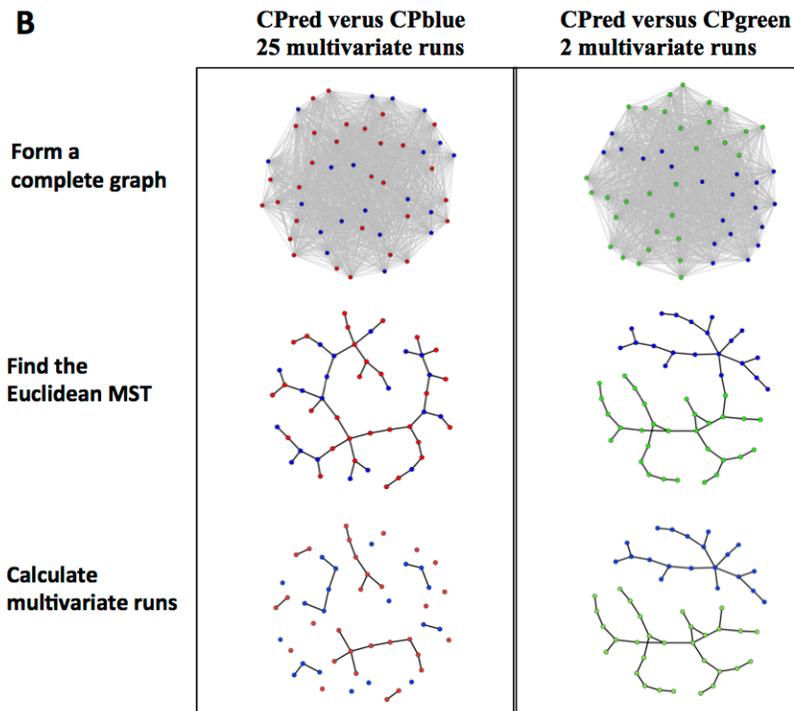
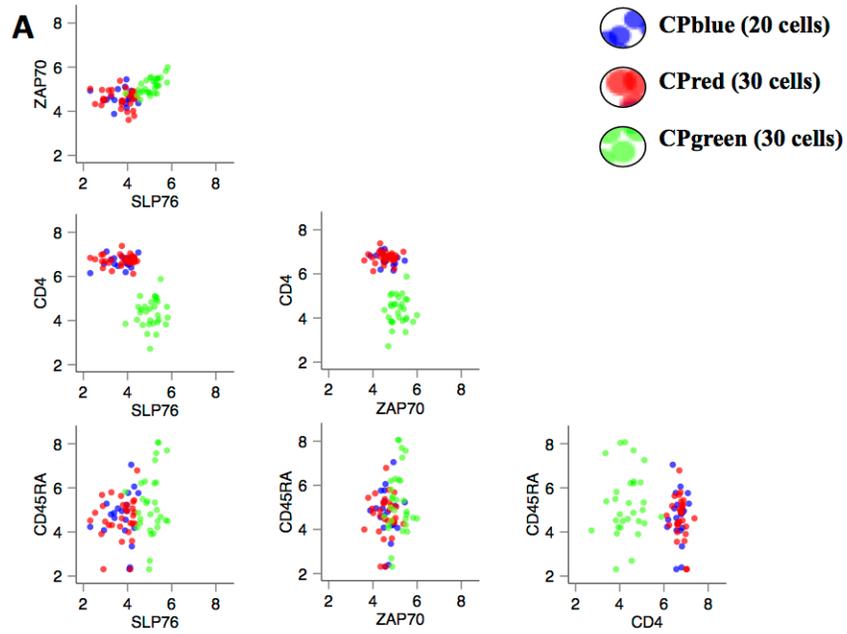


Figure 18. Multivariate run calculation.

This figure illustrates the general problem of mapping cell populations between samples using FCM data with four marker channels (CD4, CD45RA, ZAP70, and SLP76). In this example, we want to determine if Cell Population (CP)red in or CPgreen in Sample B corresponds to CPblue in Sample A. (A) Marker level distributions of CPblue in comparison with CPred and CPgreen. Note that the marker level distributions for CPblue and CPgreen are similar for ZAP70 and SLP76, but differ for CD4. Based on this difference we would infer that CPgreen in Sample B is different from CPblue in Sample A. On the other hand, the marker expression distributions for CPblue and CPred are similar for all four markers. Based on these similarities we would infer that CPred in Sample B is equivalent to CPblue in Sample A. (B) Multivariate run calculation for the CPblue/CPred and CPblue/CPgreen comparisons. The FlowMap-FR application of the Friedman-Rafsky test proceeds through the following steps separately for CPblue/CPred and CPblue/CPgreen comparisons: merge the cell event data from the reference (CPblue) and test (CPred or CPgreen) populations, calculate the pairwise Euclidean distances between all events (nodes) to form a complete Euclidean graph, find the minimum spanning tree that connects all nodes in the Euclidean graph, remove edges that connect nodes derived from different cell populations, and determine the number of subgraphs remaining (which equals the number of edges connecting nodes between the two different cell populations plus 1). In the case of the CPblue versus CPgreen comparison, the number of runs would equal 2. For the CPblue versus CPred comparison, the number of runs would equal 25. Relatively small run values indicate that the cell populations being compared are distinct because their events are segregated in multivariate space.

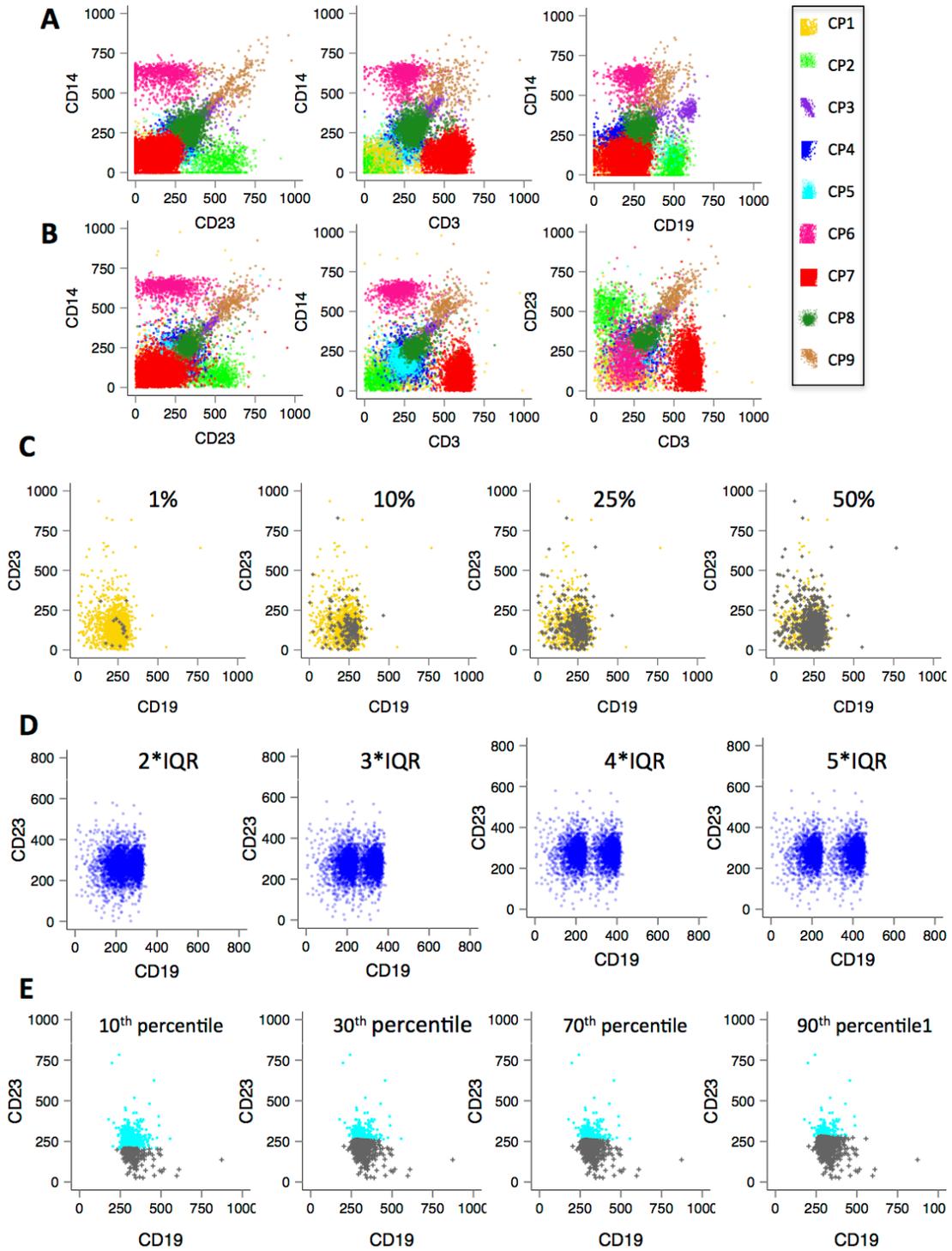
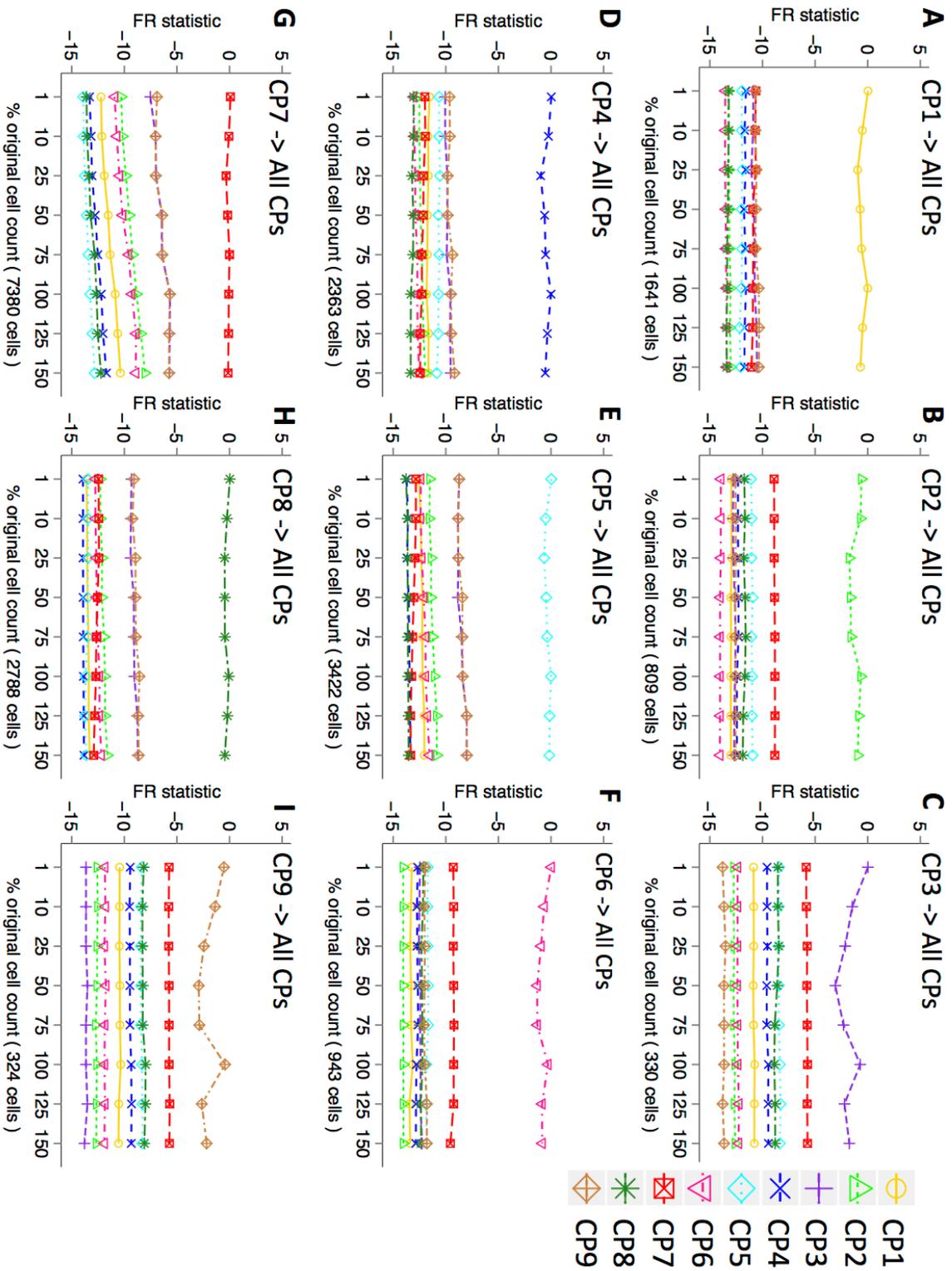


Figure 19. Data simulations and test scenarios.

(A) Selected bi-axial plots of the 9 cell populations from an experimentally measured data set in four marker channels (CD14, CD23, CD3, CD19). For each population, multivariate skew-t distributions were fitted and the corresponding distribution parameters determined. (B) Selected bi-axial plots of the 9 cell populations in the simulated data set. The skew-t distribution parameters derived from fitting the original data were used to simulate the nine populations shown. The simulated cell populations mimic the original cell population in the correlation between markers and also the marker distributions. These parameters were employed to simulate cell populations throughout the current study. (C) Scenario 1 - Differences in cell populations between samples. Overlap between CP4 changed in proportion to 1%, 10%, 25%, and 50% (colored in green) and the original 2363 events in population CP4 in the reference population (colored in cyan). Scenario 2, in which the test cell population was deleted, is not shown but would essentially correspond to the first plot without the green events. (D) Scenario 3 - Shifts in marker expression levels between samples. CP1 (colored in blue) shifted along CD19 to 2, 3, 4, and 5 units of interquartile range (IQR) of the CD19 distribution. (E) Scenario 4 - A discrete cell population in one sample inappropriately divided into two by over-partitioning in another sample. The original CP4 (colored in cyan) overlaid with the CP4 partitions below CD23 10th, 30th, 70th, and 90th percentile (colored in grey).



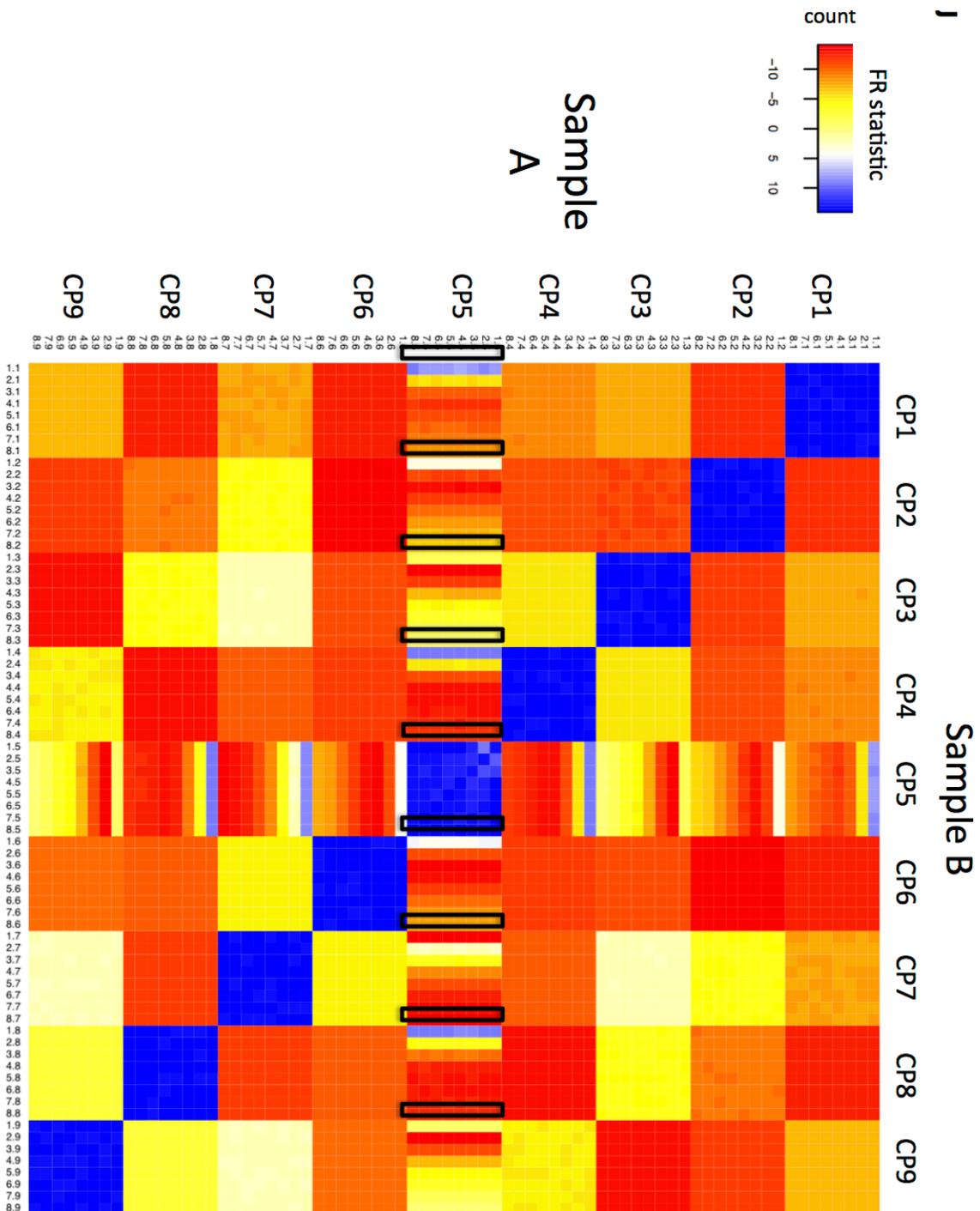
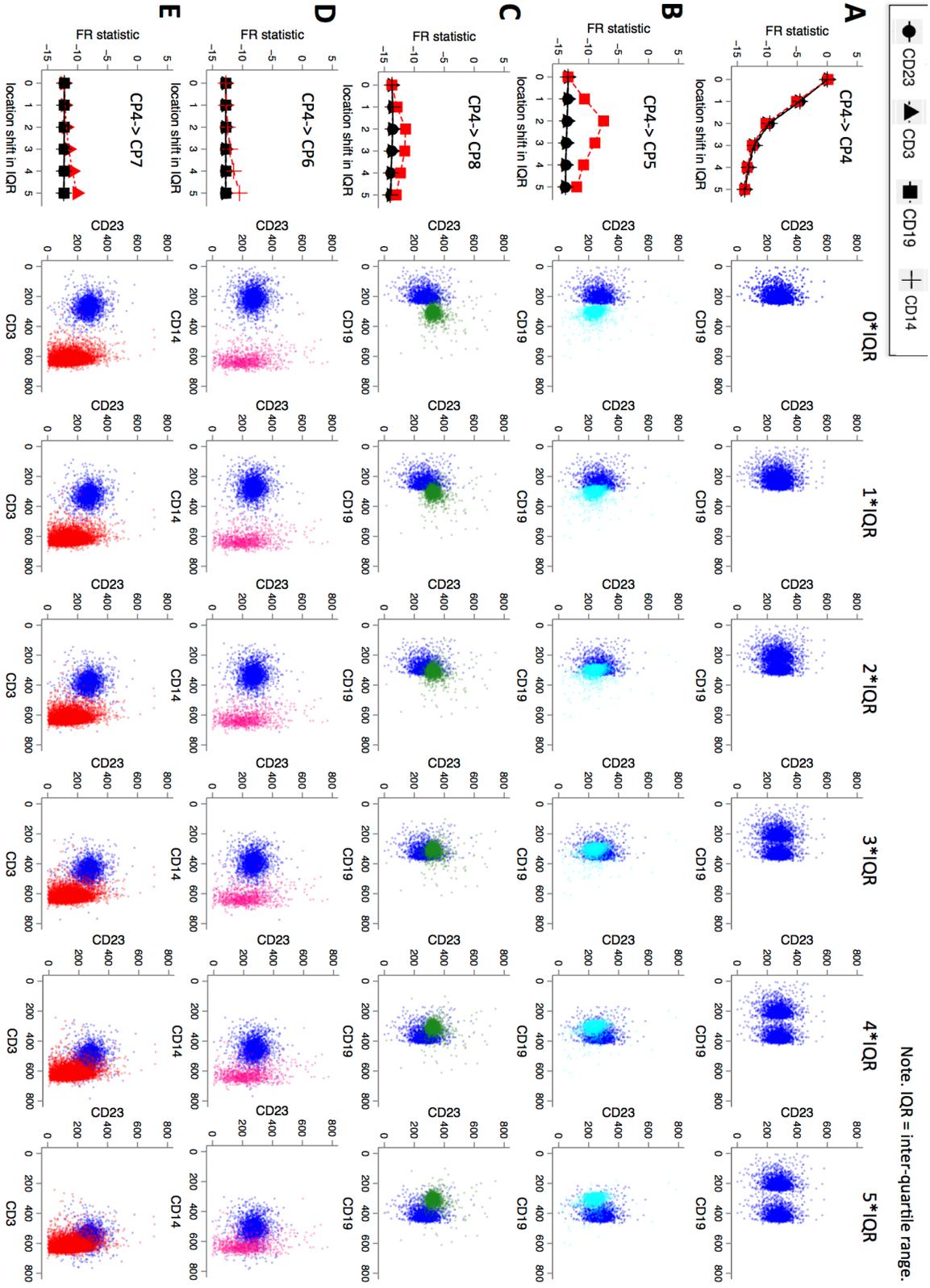


Figure 20. Matching cell populations that differ in proportions between samples.

FR statistics comparing each simulated cell population CP1 – CP9 (A – I, respectively) under varied proportions (1%, 10%, 25%, 50%, 75%, 100%, 125% and 150% of the original cell count) with all cell populations in the reference sample containing 100% of all cell events. In all 9 sets of analyses, the FR statistic is larger when comparing a changed cell population to the original cell population than when comparing it to the other cell populations across varying proportions of original cell counts. In other words, the changed cell population in the test sample can be determined to be most similar to the corresponding population in the reference sample based on the largest FR statistic value in a comparison against all nine populations in the reference sample. (J) Heat map for comparing all cell populations between the test samples (Sample Set A) and the reference samples (Sample Set B). We CP5 and changed its proportions in different test samples. The rows and columns are ordered by cell population IDs (CP1-CP9) and then by population proportion ID (1-8, with 1 being 1% and 8 being 150%). Squares colored in blue are the highly similar cell population pairs with an FR statistic close to 0, while yellow, orange and red squares are more dissimilar pairs with negative FR statistics. The regions of the heat map that correspond to the comparisons displayed in parts A – I are indicated with rectangles.



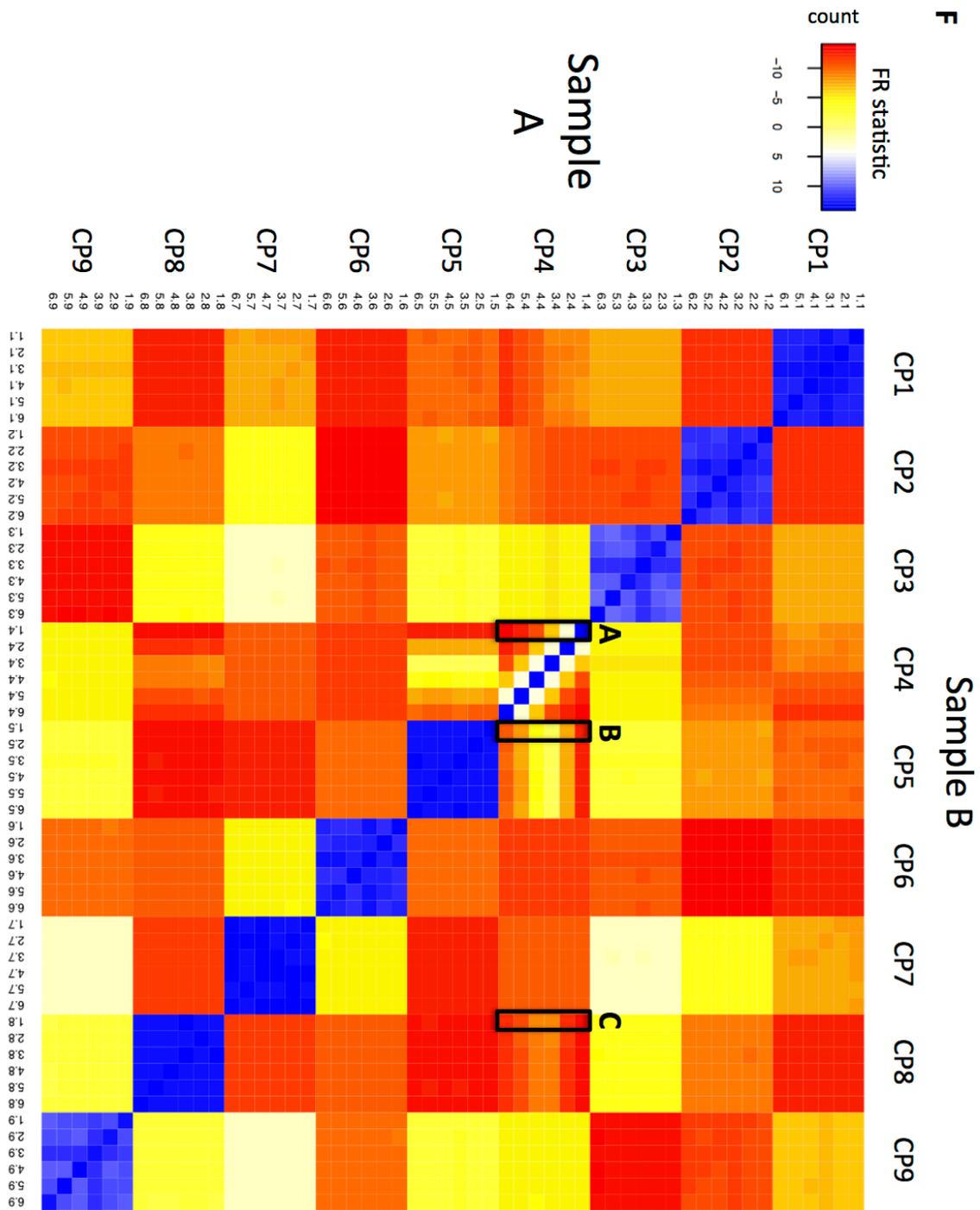
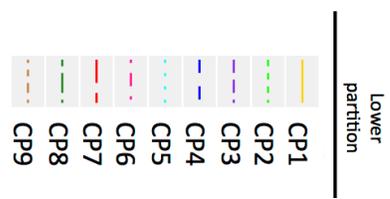
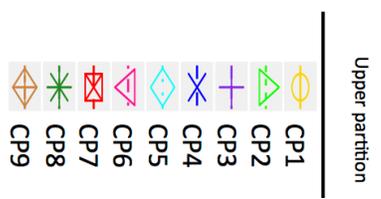
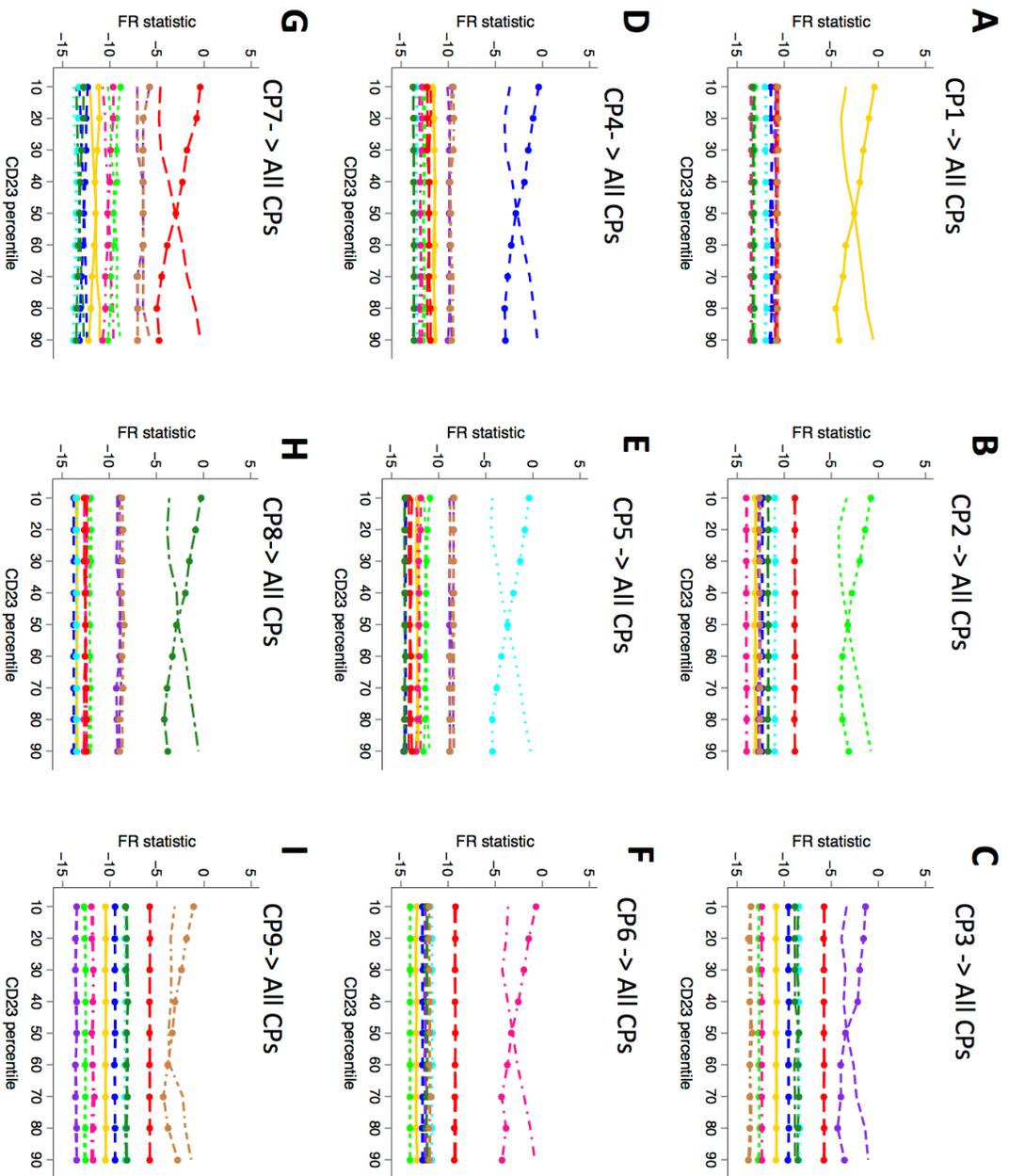


Figure 21. Matching cell populations with shifted marker distributions between samples. Shifted CP4 populations compared to the original CP4 (A) and other unchanged cell populations (B – E). The amount of shifting is quantified as units of the

calculated *interquartile range* (IQR) in each of the respective marker distributions (CD3, CD14, CD19, and CD23). The left-most column displays the FR statistics for the five sets of population comparisons with CP4 shifts of 0, 1, 2, 3, 4, and 5 IQR units of the corresponding original marker distribution in the indicated dimension. In (A), the shifted CP4 is compared to itself. As the amount of shifting increases, the dissimilarity grows between the changed CP4 and the original CP4 as indicated by the increasingly negative FR statistic in the left hand graph. The FR statistics are similar along the four marker distributions. The red line highlights the comparisons with the most pronounced FR statistics change over IQR shifts. For example, in (E), the red line corresponds to the FR statistics for comparing shifted CP4 along the CD3 axis against CP7, and the three black lines correspond to the comparisons of CP4 against CP7 along the CD23, CD19, and CD14 axis. The dot plots shown to the right illustrate locations of CP7 and the shifted CP4 along the CD3 axis. (F) Heat map for comparing all cell populations between the test samples (Sample Set A) and the reference samples (Sample Set B). We chose CP4 and shifted it along the CD19 axis in the test samples. The rows and columns are ordered by cell population IDs (CP1-9) and then by population shift ID (1-6, with 1 being no shift and 6 being a 5\*IQR shift). Squares colored in blue are the highly similar cell population pairs with an FR statistic close to 0, while yellow, orange and red squares are more dissimilar pairs with negative FR statistics. The regions of the heat map that correspond to the comparisons displayed in parts A – C are indicated with rectangles.



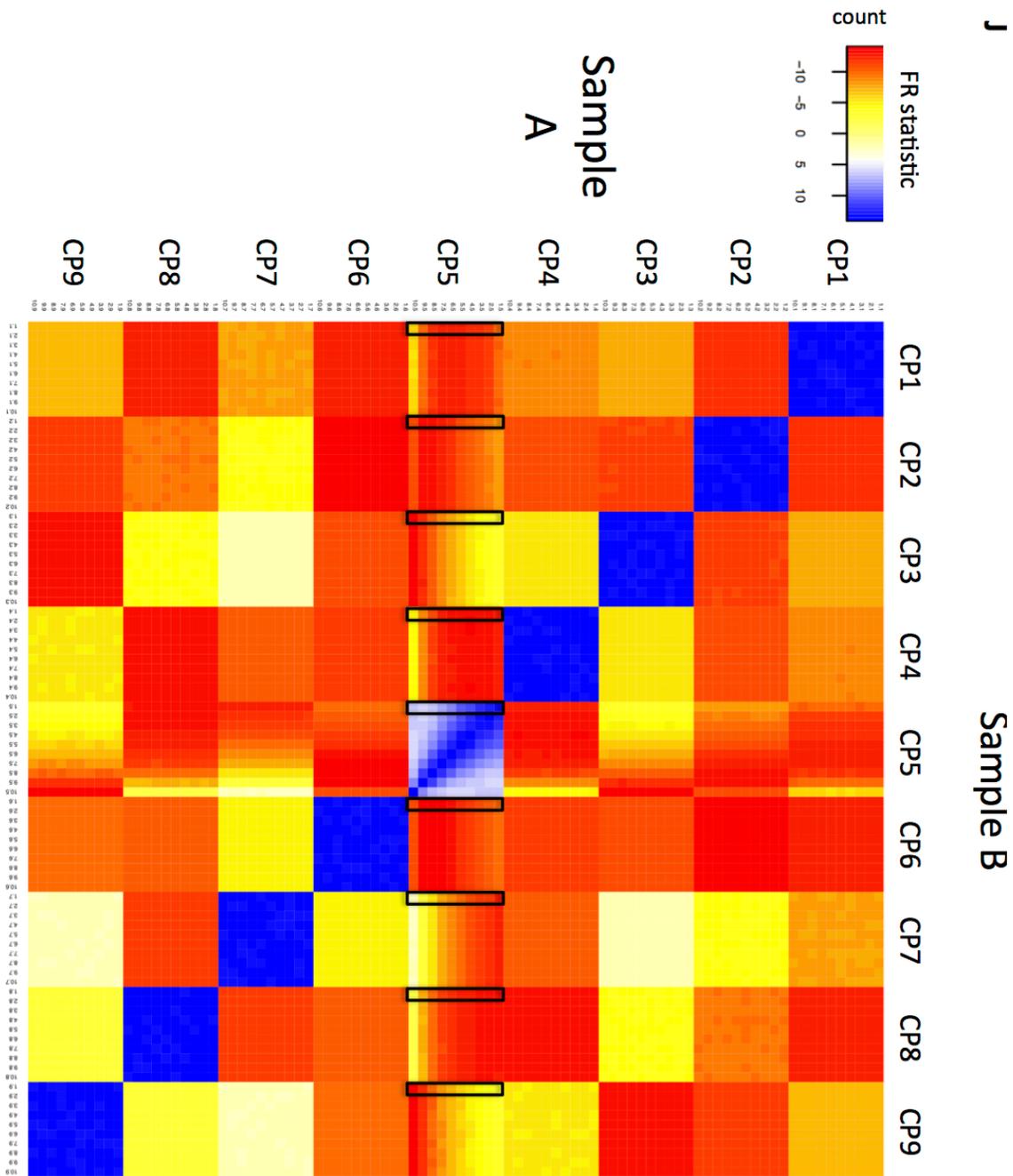


Figure 22. Matching cell populations inappropriately divided into two populations in one sample.

(A – I) FR statistics comparing the two partitioned cell populations in the test sample to all (intact) cell populations in the reference sample. The two partitioned populations are generated by dividing the indicated cell population with a discrete value for CD23 marker expression so that the two partitioned populations are

above and below the 10, 20, 30, 40, 50, 60, 70, 80, and 90<sup>th</sup> percentile in the CD23 marker distribution. Across the analyses, the two intersecting lines seen at the top of each graph show that the FR statistics are always the largest when the two partitioned populations in the test sample are compared to the intact parent cell population in the reference sample. (J) Heat map for comparing all cell populations between the test samples (Sample Set A) and the reference samples (Sample Set B). We chose CP5 to partition and compared the 10 testing samples with different CP5 upper partitions along the CD23 axis with the intact CP5 in the reference samples. The rows and columns are ordered by cell population IDs (CP1-9) and then by partition ID (1-10, with 1 being 10<sup>th</sup> percentile and 10 being no partition). Squares colored in blue are the highly similar cell population pairs with an FR statistic close to 0, while yellow, orange, and red squares are more dissimilar pairs with negative FR statistics.

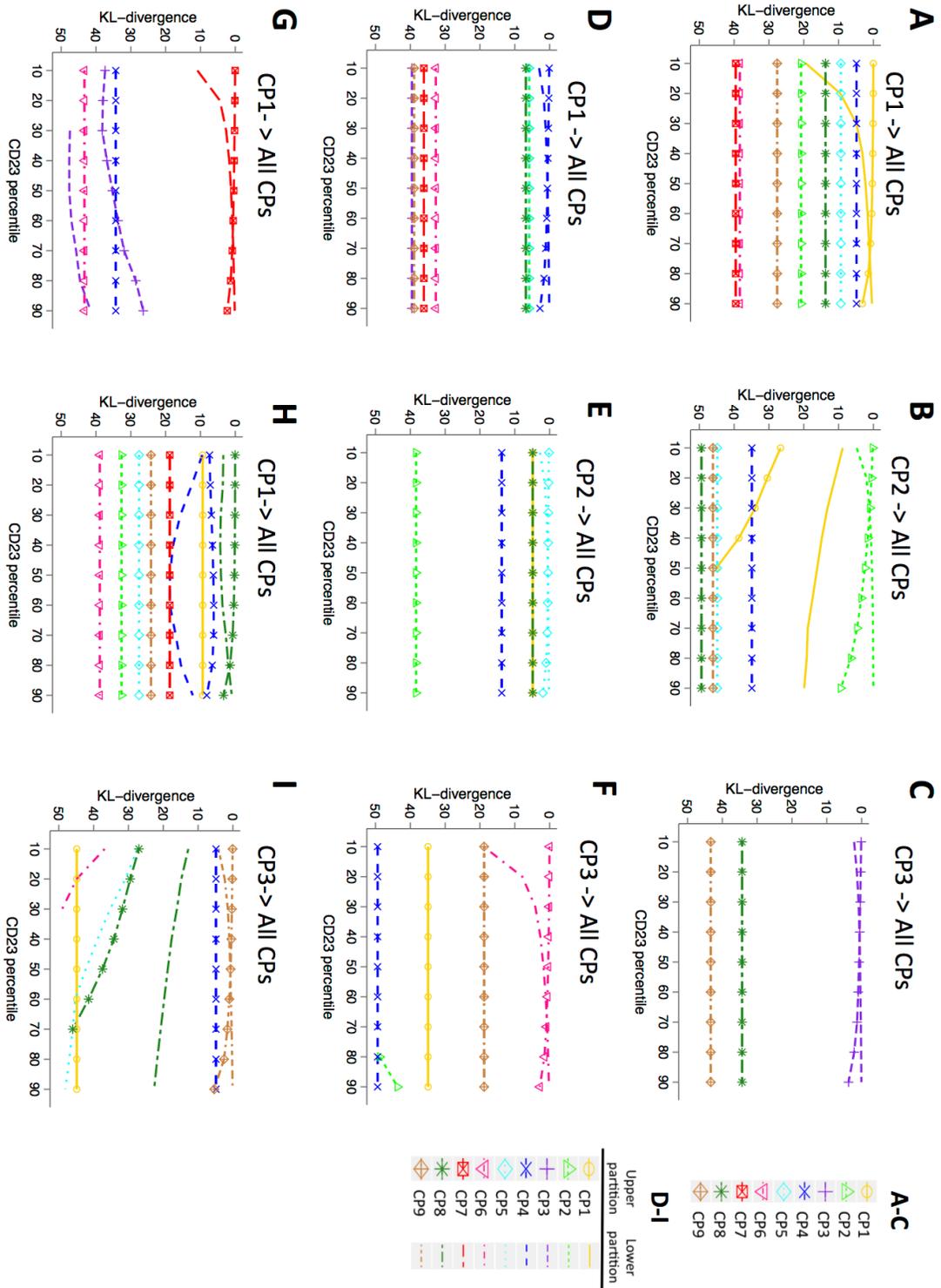


Figure 23. Matching cell populations using SKL divergence measure.

(A – C) SKL distance values (y-axis) from comparing CP1, CP2 and CP3 to all reference cell populations under varied proportions (x-axis showing 1%, 10%, 25%, 50%, 75%, 100%, 125% and 150% of the original cell counts of CP1-3, compared against reference cell populations with 100% of their cell events). For example, in (A), each line records the eight SKL distance values generated from comparing eight different proportions of CP1 to each of the original cell populations, including itself. (D - I) The SKL distance of comparing two partitions of a cell population (CP1-3) to all reference cell populations, including itself. D - F display SKL results in complete value ranges. G - I zoom in and show the top region of the D - F graphs with SKL values 0 - 50 so that the pattern of the lines can be seen.

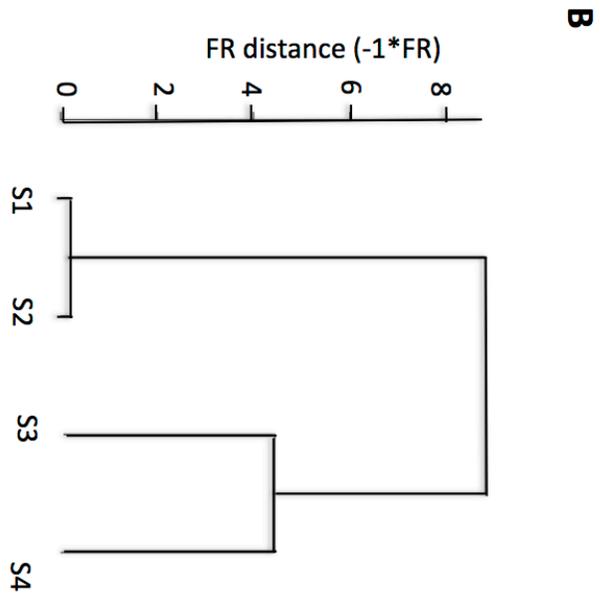
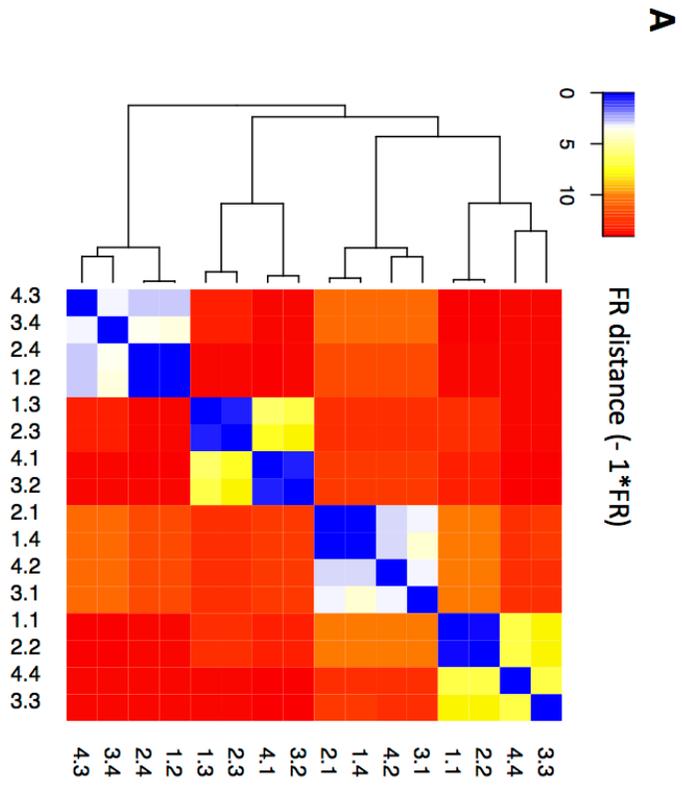


Figure 24. Matching cell populations across the real FCM data set #1.

(A) Heatmap of the FR distances ( $-1*FR$  statistics) for comparing all cell populations across four real flow cytometry samples. The cell populations in the samples were identified using K-means clustering with possible number of cell populations ranging from 4 to 20. The FR statistics were computed for all possible pairwise comparison of the cell populations. We computed a dissimilarity measure based on the FR statistic ( $-1*FR$ ) and employed the FR-based distance measure to organize the cell populations using hierarchical clustering with complete linkage.

(B) flowMatch results of matching FCM samples using the FR distance as the dissimilarity metric between cell populations across samples. The distance between FCM samples was computed based on the weighted sum of distance between cell populations across samples. y-axis represents the FR-based metric ( $-1*FR$ ) between individual FCM samples. These results are the same as [Figure 24A](#), where distance measure was only computed at the population level, and also the same as when applying the default flowMatch method with symmetric KL divergence measure as the distance metric.

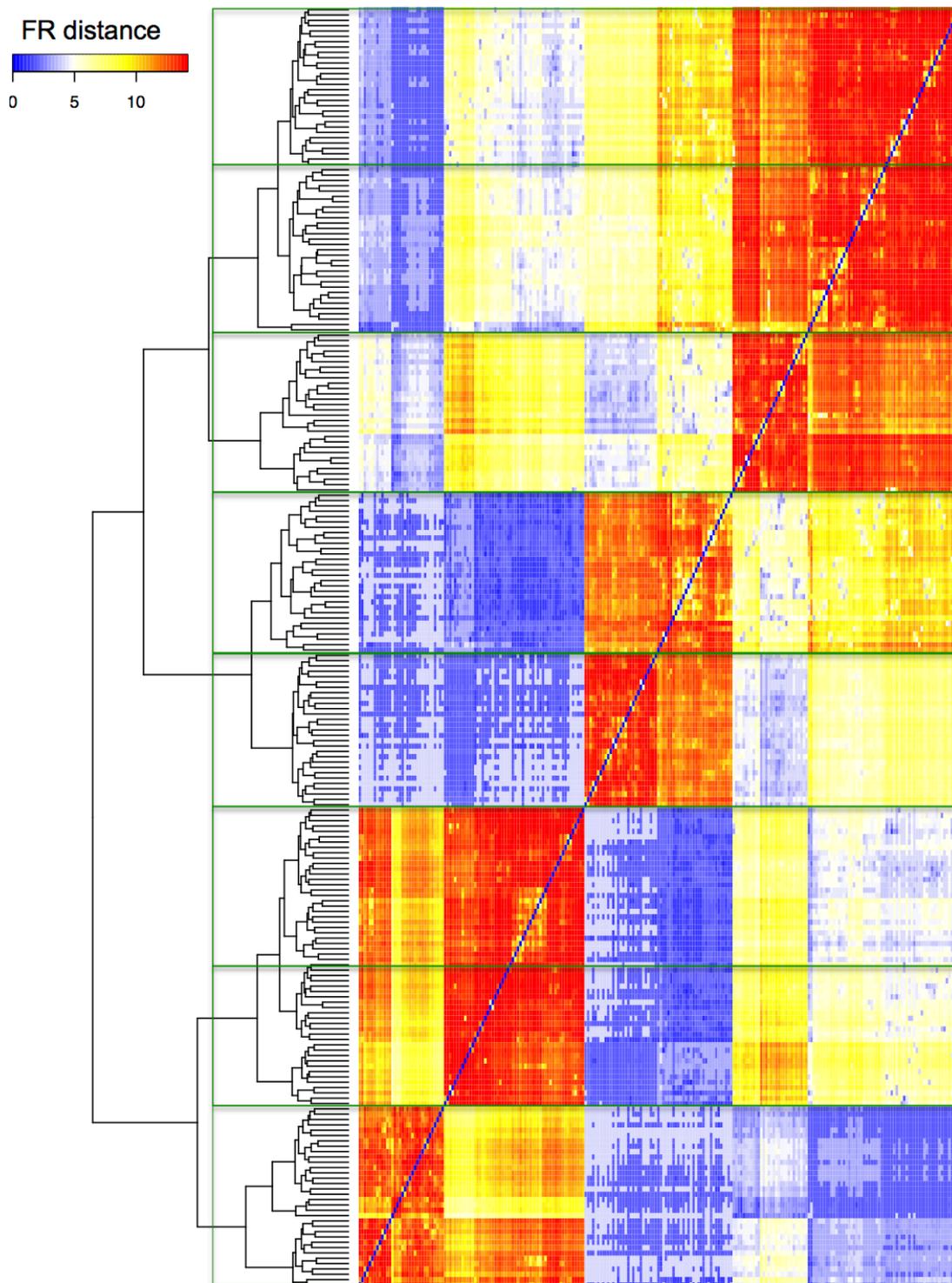


Figure 25. Matching cell populations across the real FCM data set #2.

Heatmap of the FR distance ( $-1*FR$ ) for clustering all cell populations across thirty real flow cytometry samples in the real FCM data set #2. The FR statistics were computed for all 28,680 possible pairwise cell population comparisons. We multiplied the FR statistics by -1 to obtain a dissimilarity measure. Blue boxes in the dissimilarity heatmap correspond to similar population pairs with small FR distance (large FR statistic); red boxes correspond to population pairs that are not similar to each other with large FR distance (small FR statistic). Each sample had 8 cell populations delineated by expert manual gating as part of the FlowCAP-I challenge. Cell populations are organized according to the value of the FR distance using hierarchical clustering with complete linkage. The green boxes on the plot delineate the eight cell populations deemed equivalent to each other across the thirty FCM samples in the experiment based on FR distance as reflected in the structure of the hierarchical clustering tree.

## 5 Discussion

Multivariate statistics provide a convenient framework for integrating layers of genetic information in the analysis of high-throughput biological assays. This dissertation introduces multivariate statistical methods for the analysis of ribosome footprinting profiling data and flow cytometry data.

Chapter 2 presents pairedSeq, an empirical covariance shrinkage method for differential testing of translation efficiency from sequencing data. The method explicitly models co-variability in ribosome occupancy and transcript abundance, a source of variation that directly affects the dynamic range of expression measurements in the analysis of translation efficiency. Results indicate that pairedSeq effectively shrinks covariance estimates and is able to identify differences in translation efficiency in samples with high measurement variability.

Chapter 3 surveys the contribution of translational regulation in gene expression level differences between human and chimpanzee. Through joint analysis of ribosome footprint profiling data with RNA-seq measurement of transcript levels and quantitative mass spectrometry measurement of protein levels, we provided the first integrative view on gene expression variations across primates that allows a separation between translational and post translational events. We found extensive

post translational buffering of gene expression variations that lead to a stable protein level across primate species. We propose a scenario where buffering evolved under stabilizing selection of protein levels that removes negative effects on organismal fitness from protein level variations and allows the transcript level to diverge for quick adaptation to environmental changes.

Chapter 4 describes FlowMap-FR, a statistical method developed to compare and match cell populations homogeneous in protein expression profiles in flow cytometry data. FlowMap-FR directly addresses the cell population mapping challenges that may arise during the FCM data processing workflow without *a priori* assumptions about the marker expression distributions in the different cell populations analyzed, and thus can be readily employed in comparison of skewed, non-parametric, and multi-modal distributions. The method is highly robust, as illustrated in matching cell populations of varying shapes, locations, and correlations between marker features under scenarios of differences in population proportions between samples and modest shifts in marker distributions. Because FlowMap-FR is a stand-alone cell population mapping method, it can be incorporated into any FCM analytical workflow that requires a cell population-matching step.

## Bibliography

1. Mardis, E. R. Next-generation DNA sequencing methods. *Annu. Rev. Genomics Hum. Genet.* **9**, 387–402 (2008).
2. Van Opijnen, T. & Camilli, A. Transposon insertion sequencing: a new tool for systems-level analysis of microorganisms. *Nat. Rev. Microbiol.* **11**, 435–42 (2013).
3. Solomon, M. J., Larsen, P. L. & Varshavsky, a. Mapping protein-DNA interactions in vivo with formaldehyde: evidence that histone H4 is retained on a highly transcribed gene. *Cell* **53**, 937–947 (1988).
4. Galas, D. J. & Schmit, A. DNAase footprinting a simple method for the detection of protein-DNA binding specificity. *Nucleic Acids Res.* **5**, 3157–3170 (1978).
5. Brenowitz, M., Senear, D. F., Shea, M. A. & Ackers, G. K. Quantitative DNase footprint titration: A method for studying protein-DNA interactions. *Methods Enzymol.* **130**, 132–181 (1986).
6. Morin, R. D. *et al.* Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. *Biotechniques* **45**, 81–94 (2008).
7. Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. S. & Weissman, J. S. Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling. **324**, 218–223 (2009).
8. Wu, J., Li, Y. & Jiang, R. Integrating Multiple Genomic Data to Predict Disease-Causing Nonsynonymous Single Nucleotide Variants in Exome Sequencing Studies. *PLoS Genet.* **10**, (2014).
9. West, M., Ginsburg, G. S., Huang, A. T. & Nevins, J. R. Embracing the complexity of genomic data for personalized medicine. *Genome Res.* **16**, 559–566 (2006).

10. Nevins, J. R. *et al.* Towards integrated clinico-genomic models for personalized medicine: combining gene expression signatures and clinical factors in breast cancer outcomes prediction. *Hum. Mol. Genet.* **12 Spec No**, R153–R157 (2003).
11. Monte, A. a *et al.* Improved drug therapy: triangulating phenomics with genomics and metabolomics. *Hum. Genomics* **8**, 16 (2014).
12. Emmert-Streib, F. Personalized medicine: Has it started yet? A reconstruction of the early history. *Front. Genet.* **3**, 1–4 (2013).
13. Lener, T. *et al.* Expression profiling of aging in the human skin. *Exp. Gerontol.* **41**, 387–397 (2006).
14. Morag, A. *et al.* Genome-wide expression profiling of human lymphoblastoid cell lines identifies CHL1 as a putative SSRI antidepressant response biomarker. *Pharmacogenomics* **12**, 171–184 (2011).
15. Pickrell, J. K. *et al.* Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464**, 768–772 (2010).
16. Veer, L. J. Van *et al.* Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**, (2002).
17. Gilad, Y., Oshlack, A., Smyth, G. K., Speed, T. P. & White, K. P. Expression profiling in primates reveals a rapid evolution of human transcription factors. *Nature* **440**, 242–245
18. Ingolia, N. T. *et al.* Ribosome Profiling Reveals Pervasive Translation Outside of Annotated Protein-Coding Genes. *Cell Rep.* **8**, 1365–1379 (2014).
19. Bazzini, A. a, Lee, M. T. & Giraldez, A. J. Ribosome profiling shows that miR-430 reduces translation before causing mRNA decay in zebrafish. *Science* **336**, 233–7 (2012).
20. Lee, S. *et al.* PNAS Plus: Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. *Proc. Natl. Acad. Sci.* **109**, E2424–E2432 (2012).

21. McManus, C. J., May, G. E., Spealman, P. & Shteyman, A. Ribosome profiling reveals post-transcriptional buffering of divergent gene expression in yeast. *Genome Res.* **24**, 422–430
22. Ingolia, N. T. Ribosome profiling: new views of translation, from single codons to genome scale. *Nat. Rev. Genet.* **15**, 205–13 (2014).
23. Olshen, A. B. *et al.* Assessing gene-level translational control from ribosome profiling. *Bioinformatics* **29**, 2995–3002 (2013).
24. Zhong, Y. *et al.* *RiboDiff: Detecting Changes of Translation Efficiency from Ribosome Footprints.* *bioRxiv* doi: 10.1101/017111 (2015).
25. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* **15**, R29 (2014).
26. Rencher, A. C. & Schaalje, G. B. *Linear Models in Statistics. Linear Models in Statistics* (John Wiley & Sons, Inc., 2008). doi:10.1002/9780470192610
27. Gelman, A. Prior distribution for variance parameters in hierarchical models. *Bayesian Anal.* **1**, 515–533 (2006).
28. Feng, S., Wolfinger, R. D., Chu, T. M., Gibson, G. C. & McGraw, L. a. Empirical Bayes analysis of variance component models for microarray data. *J. Agric. Biol. Environ. Stat.* **11**, 197–209 (2006).
29. Wolfinger, R. D. & Kass, R. E. Nonconjugate Bayesian analysis of variance component models. *Biometrics* **56**, 768–774 (2000).
30. Smyth, G. *Limma: linear models for microarray data. Bioinformatics and Computational Biology Solutions Using R and Bioconductor* (Springer, 2005).
31. Box, G. E. P. & Tiao, G. C. *Bayesian inference in statistical analysis.* (1973).
32. Smyth, G. K. Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray. **3**, 1–26 (2009).

33. Giron, F. J. & Del, C. C. A note on the convolution of inverted-gamma distributions with applications to the Behrens-Fisher distribution. *Rev. R. Acad. Cien. Ser. A. Mat.* **95**, 39–44 (2001).
34. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. *bioRxiv* 1–21 (2014). doi:10.1101/002832
35. Battle, A. *et al.* Impact of regulatory variation from RNA to protein. *Science* **347**, 664–667 (2015).
36. Kahn, Z. *et al.* Primate transcript and protein expression levels evolve under compensatory selection pressure. *Science* **342**, 1100–1104 (2013).
37. Sidney, H. Comparative genomics on translation levels between primates. 2013 (2013).
38. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. *bioRxiv* 1–21 (2014). doi:10.1101/002832
39. Romero, I. G., Ruvinsky, I. & Gilad, Y. Comparative studies of gene expression and the evolution of gene regulation. *Nat. Rev. Genet.* **13**, 505–516
40. Carroll, S. B. Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell* **134**, 25–36
41. Shapiro, M. D. *et al.* Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks. *Nature* **428**, 717–723
42. King, M. C. & Wilson, A. C. Evolution at two levels in humans and chimpanzees. *Science* **188**, 107–116
43. Blekhman, R., Oshlack, A., Chabot, A. E., Smyth, G. K. & Gilad, Y. Gene regulation in primates evolves under tissue-specific selection pressures. *PLoS Genet.* **4**, e1000271

44. Enard, W. *et al.* Intra- and interspecific variation in primate gene expression patterns. *Science* **296**, 340–343
45. Khaitovich, P. *et al.* Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees. *Science* **309**, 1850–1854
46. Capra, J. A., Erwin, G. D., McKinsey, G., Rubenstein, J. L. R. & Pollard, K. S. Many human accelerated regions are developmental enhancers. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **368**, 20130025
47. Lindblad-Toh, K. *et al.* A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478**, 476–482
48. Brawand, D. *et al.* The evolution of gene expression levels in mammalian organs. *Nature* **478**, 343–348
49. Khaitovich, P. *et al.* A neutral model of transcriptome evolution. *PLoS Biol.* **2**, E132
50. Li, J. J., Bickel, P. J. & Biggin, M. D. System wide analyses have underestimated protein abundances and the importance of transcription in mammals. *PeerJ* **2**, e270
51. Vogel, C. & Marcotte, E. M. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat. Rev. Genet.* **13**, 227–232
52. Laurent, J. M. *et al.* Protein abundances are more conserved than mRNA abundances across diverse taxa. *Proteomics* **10**, 4209–4212
53. Khan, Z. *et al.* Primate transcript and protein expression levels evolve under compensatory selection pressures. *Science* **342**, 1100–4 (2013).
54. Ingolia, N. T., Lareau, L. F. & Weissman, J. S. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* **147**, 789–802
55. Battle, A. *et al.* Genomic variation. Impact of regulatory variation from RNA to protein. *Science* **347**, 664–667

56. Ingolia, N. T., Brar, G. a, Rouskin, S., McGeachy, A. M. & Weissman, J. S. The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. *Nat. Protoc.* **7**, 1534–1550 (2012).
57. (2012) R Core Team. *R: A language and environment for statistical computing*. (R Foundation for Statistical Computing).
58. Falcon, S. & Gentleman, R. Using GOstats to test gene lists for GO term association. *Bioinformatics* **23**, 257–258
59. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)*. **6**, 80–92
60. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842
61. Kent, W. J., Zweig, A. S., Barber, G., Hinrichs, A. S. & Karolchik, D. BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics* **26**, 2204–2207
62. Flutre, T., Wen, X., Pritchard, J. & Stephens, M. A Statistical Framework for Joint eQTL Analysis in Multiple Tissues. *PLoS Genet.* **9**, (2013).
63. Nica, A. C. *et al.* The architecture of gene regulatory variation across multiple human tissues: the MuTHER study. *PLoS Genet.* **7**, e1002003
64. Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M. & Gilad, Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.* **18**, 1509–1517
65. Cenik, C. *et al.* Integrative analysis of RNA, translation and protein levels reveals distinct regulatory variation across humans. *bioRxiv* 018572  
doi:10.1101/018572
66. Artieri, C. G. & Fraser, H. B. Evolution at two levels of gene expression in yeast. *Genome Res.* **24**, 411–421

67. Albert, F. W., Muzzey, D., Weissman, J. S. & Kruglyak, L. Genetic influences on translation in yeast. *PLoS Genet.* **10**, e1004692
68. Bader, D. M. *et al.* Negative feedback buffers effects of regulatory variants. *Mol. Syst. Biol.* **11**, 785
69. Kirchner, S. & Ignatova, Z. Emerging roles of tRNA in adaptive translation, signalling dynamics and disease. *Nat. Rev. Genet.* **16**, 98–112
70. Fabian, M. R., Sonenberg, N. & Filipowicz, W. Regulation of mRNA translation and stability by microRNAs. *Annu. Rev. Biochem.* **79**, 351–379
71. Walsh, D. A. & Van Patten, S. M. Multiple pathway signal transduction by the cAMP-dependent protein kinase. *FASEB J. Off. Publ. Fed. Am. Soc. Exp. Biol.* **8**, 1227–1236
72. Rutherford, S. L. & Lindquist, S. Hsp90 as a capacitor for morphological evolution. *Nature* **396**, 336–342
73. Queitsch, C., Sangster, T. A. & Lindquist, S. Hsp90 as a capacitor of phenotypic variation. *Nature* **417**, 618–624
74. Wagner, A. Energy Constraints on the Evolution of Gene Expression. *Mol. Biol. Evol.* **22**, 1365–1374
75. Mellman, I., Coukos, G. & Dranoff, G. Cancer immunotherapy comes of age. *Nature* **480**, 480–9 (2011).
76. Chattopadhyay, P. *et al.* Toward 40+ Parameter Flow Cytometry. in *CYTO conference plenary presentation and abstract 388* (2014).
77. Akdis, C. A. & Akdis, M. Mechanisms and treatment of allergic disease in the big picture of regulatory T cells. *J. Allergy Clin. Immunol.* **123**, 735–46; quiz 747–8 (2009).
78. Casale, T. B. *et al.* Omalizumab pretreatment decreases acute reactions after rush immunotherapy for ragweed-induced seasonal allergic rhinitis. *J. Allergy Clin. Immunol.* **117**, 134–40 (2006).

79. Aghaeepour, N. *et al.* Critical assessment of automated flow cytometry data analysis techniques. *Nat. Methods* **10**, 228–38 (2013).
80. Pyne, S. *et al.* Automated high-dimensional flow cytometric data analysis. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 8519–24 (2009).
81. Pyne, S. *et al.* Joint modeling and registration of cell populations in cohorts of high-dimensional flow cytometric data. *PLoS One* **9**, e100334 (2014).
82. Cron, A. *et al.* Hierarchical modeling for rare event detection and cell subset alignment across flow cytometry samples. *PLoS Comput. Biol.* **9**, e1003130 (2013).
83. Qian, Y. *et al.* Elucidation of seventeen human peripheral blood B-cell subsets and quantification of the tetanus response using a density-based method for the automated identification of cell populations in multidimensional flow cytometry data. *Cytometry B. Clin. Cytom.* **78 Suppl 1**, S69–82 (2010).
84. Zare, H., Shooshtari, P., Gupta, A. & Brinkman, R. R. Data reduction for spectral clustering to analyze high throughput flow cytometry data. *BMC Bioinformatics* **11**, 403 (2010).
85. Roederer, M., Moore, W., Treister, A., Hardy, R. R. & Herzenberg, L. a. Probability binning comparison: A metric for quantitating multivariate distribution differences. *Cytometry* **45**, 47–55 (2001).
86. Finak, G., Perez, J.-M., Weng, A. & Gottardo, R. Optimizing transformations for automated, high throughput analysis of flow cytometry data. *BMC Bioinformatics* **11**, 546 (2010).
87. Azad, A., Pyne, S. & Pothen, A. Matching phosphorylation response patterns of antigen-receptor-stimulated T cells via flow cytometry. *BMC Bioinformatics* **13 Suppl 2**, S10 (2012).
88. Friedman, J. H. & Rafsky, L. C. Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests. *Ann. Stat.* **7**, 697–717 (1979).

89. Zhao, Ti., Soto, S. & Murphy, R. F. Improved comparison of protein subcellular location patterns. in *3rd IEEE international Symposium on Biomedical Imaging: Nano to Macro* 562–565 (2006).
90. Theoharatos, C., Laskaris, N., Economou, G. & Fotopoulos, S. A generic scheme for color image retrieval based on the multivariate Wald-Wolfowitz test. in *IEEE Transactions on Knowledge and Data Engineering* **17**, (2005).
91. Neemuchwala, H., A, H., Zabuawala, S. & Carson, P. Image registration methods in high dimensional space. *Int. J. Imaging Syst. Technol.* **16**, 130–145 (2007).
92. Wald, A. & Wolfowitz, J. An exact test for randomness in the non-parametric case based on serial correlation. *Ann. Math. Stat.* **14**, 378–388 (1943).
93. Moret, B. M. E. & Shapiro, H. D. in *Algorithms and Data Structure, Lecture Notes in Computer Science Volume 519* 400–411 (1991).
94. Prim, R. Shortest connection networks and some generalizations. *Bell Syst. Tech. J.* 1389–1401 (1957).
95. Qian, Y. *et al.* FCSTrans: an open source software system for FCS file conversion and data transformation. *Cytometry. A* **81**, 353–6 (2012).
96. Klunker, S. *et al.* Combination treatment with omalizumab and rush immunotherapy for ragweed-induced allergic rhinitis: Inhibition of IgE-facilitated allergen binding. *J. Allergy Clin. Immunol.* **120**, 688–95 (2007).
97. Azad, A. healthyFlowData : Healthy dataset used by the flowMatch package. R package version 1.3.1. (2013).
98. Lo, K., Hahne, F., Brinkman, R. R. & Gottardo, R. flowClust: a Bioconductor package for automated gating of flow cytometry data. *BMC Bioinformatics* **10**, 145 (2009).
99. [http://sourceforge.net/projects/flowcyt/files/GenePattern Flow Cytometry Suite/FCS2 CSV/](http://sourceforge.net/projects/flowcyt/files/GenePattern%20Flow%20Cytometry%20Suite/FCS2%20CSV/). Accessed February 02, 2012.

100. Spidlen, J. *et al.* GenePattern Flow Cytometry Suite. *Source Code Biol. Med.* **8**, (2013).
101. Altintas, I. Distributed workflow-driven analysis of large-scale biological data using biokepler. *Proc. 2nd Int. Work. Petascal data Anal. challenges Oppor. - PDAC '11* 41 (2011). doi:10.1145/2110205.2110215