Data-driven detection of subtype-specific and differentially expressed genes

Lulu Chen[1], Yingzhou Lu[1], Guoqiang Yu[1], Robert Clarke[2], Jennifer E. Van Eyk[3], David M. Herrington[4,*], and Yue Wang[1,*]

[1]Department of Electrical and Computer Engineering, Virginia Polytechnic Institute and State University, Arlington, VA 22203, USA; [2]Lombardi Comprehensive Cancer Center, Georgetown University, Washington, DC 20057, USA; [3]Advanced Clinical Biosystems Research Institute, Cedars Sinai Medical Center, Los Angeles, CA 90048, USA; [4]Department of Internal Medicine, Wake Forest University, Winston-Salem, NC 27157, USA

Article Format:	bioRxiv
Running title:	Subtype-specific DEG detection
Open source software:	A Bioconductor R package of OVESEG-test is available at http://bioconductor.org/packages/OVESEG

[*]Authors for correspondence:

Yue Wang, Ph.D.
Virginia Tech
E-mail: yuewang@vt.edu
David M. Herrington, M.D.
Wake Forest University
E-mail: dherring@wakehealth.edu

Tissue or cell subtype-specific and differentially-expressed genes (SDEGs) are defined as being differentially expressed in a particular tissue or cell subtype among multiple subtypes. Detecting SDEGs plays a critical rolse in molecularly characterizing and identifying tissue or cell subtypes, and facilitating supervised deconvolution of complex tissues. Unfortunately, classic differential analysis assumes a convenient null hypothesis and associated test statistic that is subtype-non-specific and thus, resulting in a high false positive rate and/or lower detection power with respect to particular subtypes. Here we introduce One-Versus-Everyone Fold Change (OVE-FC) test for detecting SDEGs. To assess the statistical significance of such test, we also propose the scaled test statistic OVE-sFC together with a mixture null distribution model and a tailored permutation scheme. Validated with realistic synthetic data sets on both type 1 error and detection power, OVE-FC/sFC test applied to two benchmark gene expression data sets detects many known and de novo SDEGs. Subsequent supervised deconvolution results, obtained using the SDEGs detected by OVE-FC/sFC test, showed superior performance in deconvolution accuracy when compared with popular peer methods.

Molecular characterization (e.g. gene expression profile) of a complex biologic system often includes features that are ubiquitously expressed by all cell or tissue types in the system (e.g. housekeeping genes)[1], and expressed features that are specific to one or more cell or tissue subtypes present in the system (marker genes or differentially-expressed genes)[2-4]. An important but frequently underappreciated issue is how a “cell or tissue subtype-specific expression pattern” is defined. Ideally, a specific expression pattern would be composed of individual features that are specifically and differentially expressed in the cognate cell or tissue subtype of interest in relation to every others – so-called subtype-specifc and differentially expressed genes (SDEGs)[5-8].

SDEGs play a critical role in molecularly characterizing and identifying tissue or cell subtypes, and facilitating the supervised deconvolution of complex tissues[5,8,9]. However, detecting SDEGs using tissue or cell-specific molecular expression profiles remains a challenging task[10]. For example, the most frequently used methods rely on an extension of ANOVA that identifies genes differentially expressed across any of the relevant cell or tissue subtypes. In this case, the null hypothesis is that samples in all subtypes are drawn from the same population, resulting in the selection of genes that may not conform to the SDEG definition. One-Versus-Rest Fold Change (OVR-FC) is another popular method based on the ratio of the average expression in a particular subtype to that of the average expression in the rest samples[10-12], and OVR t-test is occasionally used to assess the statistical significance of detected genes[13]. However, a gene with low average expression in the rest is not necessarily expressed at a low level in every subtype in the rest. Expectedly, simulation studies show that Marker Gene Finder in Microarray data (MGFM) outperforms OVR t-test[14]. Alternative strategies include One-Versus-One (OVO) t-test and Multiple Comparisons with the Best (MCB)[15] that use additional pairwise significance testing or the confidence intervals of OVO statistics[2,16].

To address the critical problem of the absence of a detection method explicitly matched to the definition of SDEGs, here we introduce One-Versus-Everyone Fold Change (OVE-FC) test to detect SDEGs among many subtypes. OVE-FC test has previously been proposed to detect SDEGs to improve multiclass classification, where the selection is based on whether the mean of one subtype is significantly higher or lower than the mean from each of the other subtypes[5,6]. To assess the statistical significance of such test, we also propose the scaled test statistic OVE-sFC together with a mixture null distribution model. Because the expression patterns of non-SDEGs can be highly complex, a tailored permutation scheme is used to estimate the corresponding distribution under the null hypothesis.

We first validate the performance of OVE-sFC test on extensive simulation data, in terms of type 1 error rate and False Discovery Rate (FDR) control. We then demonstrate the detection power of the OVE-FC/sFC test over a comprehensive set of scenarios, in terms of partial area under the receiver operating characteristic curve (pAUC), and in comparison with top peer methods. We present the utility of OVE-FC/sFC test by applying it to benchmark public data, and assessing the performance by comparing with known SDEGs and by the accuracy of supervised deconvolution that uses the expression patterns of *de novo* SDEGs detected by OVE-FC/sFC test.

Results

Validation of OVE-sFC test on type 1 error using simulation data sets

To test whether our OVE-sFC test can detect SDEGs at appropriate significance levels, we assessed the type 1 error using simulation studies under the null hypothesis (Methods). Accuracy of type 1 error is crucial for any hypothesis testing methods that detect SDEGs based on their p-values because if the type 1 error is either too conservative or too liberal, the p-value is inflated by either too many false positive or false negative estimates, loses its intended meaning, and becomes difficult to interpret.

The simulation data contain 10,000 genes whose baseline expression levels are sampled from the real benchmark microarray gene expression data with replicates of purified cell subtypes (GSE19380[8]). Using realistic simulation data sets with various parameter settings, we show that in all scenarios the empirical type 1 error produced by OVE-sFC test closely approximates the expected type 1 error (**Figure 1a, Figure S2, Figure 2a-b**). The p-values associated with OVE-sFC test statistics exhibit a uniform distribution as expected. Moreover, even with unbalanced sample sizes among the subtypes, the mixture null distribution estimated by our posterior-weighted permutation scheme produces the expected empirical type 1 error rate (**Figure S2** and **Figure 2a**). In contrast, the empirical type 1 error produced by OVR t-test and OVO t-test either over-estimates or under-estimates the expected type 1 error. The p-values associated with OVR t-test and OVO t-test deviate from a uniform distribution (**Figure 1b**). We also evaluate the type 1 error associated with individual subtypes under high noise levels and small sample sizes. For each of the subtypes, experimental results show that the empirical type 1 error produced by OVE-sFC test closely matches the expected type 1 error (**Figure 1b** and Supplementary Information).

We conducted similar validation studies involving five subtypes over a wide range of simulation scenarios (**Error! Reference source not found.**). Experimental results again show that OVE-sFC test produces the empirical type 1 error rates that match the expected type 1 error rates. Furthermore, subtype-specific p-value estimates effectively balance the uneven type 1 error rates among the subtypes with different numbers of upregulated genes (Methods, **Figure 2b**, and Supplementary Information).

Comparative assessment of OVE-FC/sFC test on power of detecting SDEGs using simulation data sets

Using realistic simulation data sets, we simulated a comprehensive set of scenarios to compare the power of OVE-sFC test and peer methods in detecting SDEGs (Method and Supplementary Information). Simulation data are generated again by modifying the expression levels of real gene expression data, where a portion of the genes are designated as SDEGs that are upregulated specifically in one of the participating subtypes, with fold change drawn in certain ranges (Supplementary Information). To recapitulate the characteristics of real expression data, various parameter settings are considered including unbalanced sample sizes or diverse mixture null distribution cross subtypes, each with 20 replications.

In large-scale multiple testing, False Discovery Rate (FDR) control is an important issue when assessing the detection power. For a well-designed significance test, the objective is to maximize power while assuring FDR below an allowable level. To test whether the q-value reflects the actual FDR, 'fdrtool' package is used to estimate the q-value for each gene[17], where the empirical FDR with an estimated q-value of 0.05 is expected to be around 0.05. Another informative criterion is the pAUC that emphasizes the leftmost partial area under the receiver operating characteristic curve, focusing on the sensitivity at lower False Positive Rates (FPR)[18].

Experimental results show that both overall and subtype-specific OVE-sFC test achieve a well-controlled FDR that matches the q-value cutoff (**Figure S5-S6**). In contrast, OVR t-test underestimates, while OVO t-test overestimates, the FDR (Supplementary Information). Moreover, subtype-specific OVE-sFC test exhibits a more balanced FPR for SDEGs across subtypes, while peer methods produce higher FPRs in the subtypes of small sample size.

For pAUC, the OVE strategy in OVE-FC/sFC test achieved the highest power in detecting SDEGs (**Figure 3**, **Figure S7-S8**, **Table S1-S3**). More specifically, for more realistic SDEGs (with sufficiently large fold change), OVE-sFC test shows the best performance; for the ideal SDEGs (i.e., marker genes with significantly large fold change[8,19]), both OVE-FC and OVE-sFC achieve the best performance with slight outperformance by OVE-FC. In comparison with the peer methods, OVE-sFC test consistently outperforms OVO t-test in these challenging experiments involving more subtypes and using RNAseq data. More importantly, the outperformance of OVE-FC/sFC over the peer methods at a stringent FPR range in ROC analysis is significant, because the related FDR is problematic in many real-world applications where large scale multiple comparisons are involved. In contrast, all three OVR methods exhibit lower detection power, and ANOVA has the lowest detection power (Supplementary Information).

It is worth mentioning that when sample size is small, OVE-sFC test statistic borrows information across genes in estimating *a priori* variance via the limma method, thus stabilizing variance estimate for each gene. Furthermore, OVE-sFC test statistic estimates the parameters of the limma model from all subtypes, producing better results than that applying t-test independently with the limma model for each subtype pair. Indeed, for small sample size cases, our experimental results show that OVE-sFC test clearly outperforms OVO t-test (**Figure 3**, **Figure 5c** and **Tables S2-S3**). Note that when a large number of genes is involved, a more stringent multiple comparison correction or FPR/FDR control is applied.

Application of OVE-sFC test on two benchmark gene expression data sets detects SDEGs (human immune cells)

We applied OVE-sFC test to two real microarray gene expression data sets, GSE28490 (Roche) and GSE28491 (HUG), to detect SDEGs associated with human immune cells[20]. In these data sets, the constituent subtypes are composed of seven different human immune cells that were isolated from healthy human blood: B cells, CD4+ T cells, CD8+ T cells, NK cells, monocytes, neutrophils, and eosinophils. Because Roche and HUG used the same protocols for cell isolation and sample processing from two independent panels of donors, the derived gene expression profiles enable use of a cross-validation strategy.

With an FDR control of q-value < 0.05 applied to both data sets, OVE-sFC test detects n=28 CD4+ T cell marker genes, n=7 CD8+ T cell marker genes, and multiple marker genes for other more distinctive cell types (**Table S4-S6**). Between the two data sets, we obtain a Jaccard index (intersection over union) of 36.8% for all SDEGs across all seven cell types. Overlap of monocyte and neutrophil marker genes detected from the two datasets is >40% (**Figure 4**). The number of SDEGs accounts for about one-third of all probesets (Roche: 39%, HUG: 34%). This result is expected because these subtypes are pure cell types and so more distinctive than would be seen with samples from multicellular complex tissues[9,21,22]. We also applied a Bonferroni multiple testing correction and a more stringent p-value < 0.001; the number of SDEGs account for 10.7% and 2.7% of all probesets in Roche and HUG data sets, respectively (**Table S4**), with only one common CD4+ T cell marker gene (FHIT) and one common CD8+ T cell marker gene (CD8B).

To present all kinds of combinatorial upregulation patterns among cell types occurred under the null hypothesis (**Figure S9**), probeset-wise posterior probabilities of component hypotheses in the null mixtures (Eq. 4) are accumulated and normalized to estimate the counterpart probabilities of the alternative hypotheses (Eq. S10), where the patterns of upregulation in B cells, monocytes, or neutrophils rank the top in both data sets, followed by upregulation in lymphoid cells (B cells, CD4+ T cells, CD8+ T cells, NK cells) and T cells (CD4+ T cells, CD8+ T cells) in the Roche dataset.

Evaluation of ideal SDEGs detected by OVE-FC/sFC test via supervised deconvolution

Accurate and reliable detection of ideal SDEGs has significant impact on the performance of many supervised deconvolution methods that use the expression patterns of ideal SDEGs to score constituent subtypes in heterogeneous samples[21,23,24]. We adopted a Convex Analysis of Mixtures (CAM) score calculated from ideal SDEGs-guided supervised deconvolution to quantify the proportional aboundance of each subtype (Supplementary Information). The correlation coefficient between estimated scores and true proportions was used to assess the accuracy of various SDEGs selection methods.

Both OVE-FC and OVE-sFC tests are applied to three independent data sets acquired from the purified subtype expression profiles (GSE28490 Roche), purified subtype RNAseq profiles (GSE60424), and classified single-cell RNAseq profiles (GSE72056), respectively. The idea SDEGs are detected by six different methods including OVE-FC, OVE-sFC, OVR-FC, OVR t-stat, OVR t-test, and OVO t-test, and then used to supervise the deconvolution of realistically synthesized mixtures with ground truth.

The proportions of constituent subtypes are estimated by the CAM scores derived from expression levels of top-ranked SDEGs for each subtype. Supervised deconvolution results show that OVE-sFC test, OVE-FC test and OVO t-test methods achieved the highest correlation coefficients between CAM score and true proportions when compared with the performance of other methods (**Figure 5a, Figure S10**).

As a more biologically realistic case involving higher between-sample variations, we synthesize a set of n=50 *in silico* mixtures by combining the subtype expression profiles from bootstrapped samples in the RNAseq data set according to pre-determined proportions. Again, supervised deconvolution results show that the ideal SDEGs detected by OVE-FC test or OVE-sFC test or OVO t-test achieved superior deconvolution performance (**Figure 5b, Figure S11**).

As a more challenging case of RNAseq data with lower signal-to-noise ratio and small sample size, we repeated the simulations where *in silico* mixtures were synthesized by combining subtype mean expressions (GSE28491 HUG) and ideal SDEGs were detected from downsampled RNAseq profiles (GSE60424, n=3). Three purified samples were randomly selected for each subtype and analyzed by the six methods. The experimental results show that, in terms of ideal SDEG-guided deconvolution performance, OVE-sFC test strongly defeats OVO t-test. Moreover, OVE-sFC test outperforms OVE-FC test for phenotypically closer cell types (CD4+ T and CD8+ T cell types) (**Figure 5c).**

Across the varying number of ideal SDEGs (5~200) being selected, **Figure 5** shows the impact of SDEGs (both at a fixed number and the corresponding content) selected by different methods on the performance of supervised deconvolution. While different subtypes are expected to have different number of ideal SDEGs practically and biologically, e.g., B cell or monocyte versus CD4+ T cell or CD8+ T cell, the fundamental working principle of various tissue deconvolution methods is that there is a proper small number of ideal SDEGs to be expressed exclusively in only one particular subtype. Thus applying a stringent OVE-sFC test p-value threshold, e.g., < 0.001 after correction (**Table S4**) is a good option, because suitable number of ideal SDEGs for CD4+ or CD8+ T cell is 5~20, while B cells or monocytes often allows more ideal SDEGs to be used in supervised deconvolution.

Discussions

Interpreting an expression profile of complex tissues requires both the knowledge of the relative abundance of the different cell or tissue subtypes and their individual expression patterns. Understanding the relative contribution of individual cell or tissue subtypes in individual samples may illuminate pathophysiologic mechanisms, biologic responses to various stimuli, or transitions in tissue phenotype - especially when the cell-cell and cell-matrix interactions in a complex system are necessary conditions for biological relevance. Expression patterns of SDEGs for the relevant cells or tissues can be used to support supervised deconvolution to estimate the relative prevalence of the contributing cell or tissue subtypes. Our present work is focused on SDEGs, *i.e.* restricted to the SDEG definition widely adopted[7,8,19,25]. Indeed, the work here is motivated by the need to obtain SDEGs for supervised *in silico* tissue deconvolution[21] and/or tissue subtype characterization[9], where the measured data are the mixtures of the expressions from multiple underlying subtypes and the SDEGs are used to estimate both the proportions of each subtype in individual heterogeneous samples as well as the averaged subtype-specific expression profiles.

Though ideal SDEGs are defined as being exclusively and consistently expressed in a particular tissue or cell subtype across varying conditions, biological reality dictates a more relaxed definition that allows SDEGs of a particular tissue or cell subtype having low or insignificant expressions in all other subtypes. Experimental results show that SDEGs detected by OVE-FC/sFC test with high thresholds or small p-values can accurately estimate both subtype proportions and expression profiles, serving as effective molecular markers (**Figure 5c**, **Figure S10 and S11**). Accuracy of OVE-FC/sFC test based SDEG detection may be affected by batch effect, normalization, and outliers; and reliability of OVE-sFC would depend on the variance estimate particularly when sample size is small.

OVE-sFC test makes a few assumptions and works best when all assumptions are valid. For example, while the proposed permutation scheme does not require the data to be normally distributed, under the null hypothesis, OVE-sFC test assumes that samples are drawn from the distribution of the same ‘shape’ for different genes, ensuring that the null distributions across genes can be aggregated together with variance-based standardization. Practically, when data distributions deviate significantly from a common shape, the limma-voom/vooma/voomaByGroup's variance models can be used to accommodate unequal variances by appropriate observational-level weights[26]; when data distributions deviate significantly from normality, a permutational ANOVA can be used to estimate the null hypothesis components of the mixture distribution. Our experimental results show that with the mean-variance relationship estimated by limma-voom on RNASeq data, the OVE-sFC test can maintain the expected type 1 error rates or specified FDR (**Figure S6**). For outliers and drop-out zero values in RNAseq data, when needed, state-of-the-art two-group test methods designed specifically for RNAseq will be exploited and adopted, *e.g.* edgeR[27], DESeq2[28].

While OVE-FC is the simpler version of our OVE strategy and drives the newer OVE-sFC test, here we have demonstrated that OVE-FC is also an effective and robust method for detecting SDEGs particularly when sample size is small. On the other hand, OVE-sFC test is a critical complement to OVE-FC in twofold: (1) OVE-FC does not assess statistical significance (producing p-values) while OVE-sFC test specifically aims to provide a significance assessment and to potentially improve FDR control; (2) OVE-sFC test improves detection power in some of more challenging situations. Theoretically, detecting SDEGs by evaluating the significance with accurate p-values is an attractive feature of OVE-sFC test that can help control FDR at the expected level. Indeed, our experimental results show that the OVE-sFC test outperforms OVE-FC in the more challenging cases involving nonideal SDEGs (**Figure S8**) or phenotypically closer cell types (**Figure 5c**). Nevertheless, OVE-sFC test may become unstable when the scaling factor is too small or inaccurately estimated, and OVE-FC will not work well when pre-exclusion of extremely lowly-expressed genes is not done properly.

Though ANOVA has been the most commonly used method to test differences among the means of multiple subtypes, often in conjunction with a post-hoc Tukey HSD comparing all possible pairs of means[29], it is not suitable for detecting SDEGs because the null hypothesis used by ANOVA does not truly enforce the definition of SDEGs. ANOVA detects differentially expressed genes rather than SDEGs and therefore produces too many false positives with respect to individual subtypes.

In addition to the SDEGs discussed here (genes uniquely up-regulated in a particular subtype), the counterpart subtype-specific down-regulated genes (genes uniquely down-regulated in a particular subtype) are also of biological interest[5]. OVE-FC/sFC test is principally applicable to detecting down-regulated SDEGs by reversing the comparison rule [5]. There are certainly other alternative definitions of 'informative genes' for different analytical purposes, *e.g.*, sample classification. In our earlier work on multiclass classification[5,6], we have shown that upregulated SDEGs selected by OVE-FC are sufficient to achieve multiclass classification and can often improve classifier performance over alternative informative gene subsets of the same size.

Lastly, when subtype-specific expression patterns are unknown, unsupervised deconvolution techniques (e.g., CAM[21]) are required. A theoretical advantage of unsupervised deconvolution is that it can identify both the cell/tissue subtype proportions and their specific expression patterns, albeit with possibly less fidelity than when neither is known a priori or measured from the same sample.

Methods

Basic OVE-FC test. Consider the measured expression level $s_k(i,j)$ of gene j in sample i across $k = 1, \dots, \dots K$ subtypes. We denote the mean and variance of the logarithmic expression levels $\log s_k(i,j)$ of gene j in subtype k by $\mu_k(j)$ and $\sigma^2(j)$, respectively. Accordingly, OVE-FC test after logarithm for gene j is defined as the gap between the highestly expressed subtype and the second highestly expressed subtype [5,14],

$$d_j = \min_{l \neq (K)} \{\mu_{(K)}(j) - \mu_l(j)\}, \tag{1}$$

where subscript (K) indicates the subtype with the maximum mean among all subtypes. Note that OVE-FC has previously been proposed for multiclass classification[13, 14], and matches well the definition of SDEGs [5,8,19,25]. Conceptually, the null hypothesis for non-SDEGs, and the alternative hypothesis for SDEGs, can be described as

$$\begin{gathered} H_{\text{non-SDEG}}\text{: } d_j = 0; \\ H_{\text{SDEG}}\text{: } d_j > 0. \end{gathered} \tag{2}$$

It is worth reiterating that the ideal SDEGs detected by OVE strategy with a stringent threshold are also termed the marker genes for supervised deconvolution[8,19], and are principally similar to what detected by the Convex Analysis of Mixtures (CAM) method for fully unsupervised deconvolution[9,21], i.e. the marker genes resided near the vertices of the scatter simplex.

OVE-sFC test statistic and null distribution modeling. To assess statistical significance of OVE-FC and to cross-fertilize the information among genes, we assume that $\log s_k(i,j) \sim N(\mu_k(j), \sigma^2(j)\ )$, and further define the scaled test statistic OVE-sFC as

$$t_j = \min_{l \neq (K)} \left\{ \frac{\mu_{(K)}(j) - \mu_l(j)}{\sigma(j)\sqrt{\frac{1}{N_{(K)}} + \frac{1}{N_l}}} \right\}, \tag{3}$$

where $N_{(K)}$ and N_l are the numbers of samples in subtypes (K) and l, respectively. However, for more than two subtypes $K \geq 3$, modeling the distribution of t_j under the null hypothesis is challenging because the expression patterns of non-SDEGs are highly complex, i.e., non-SDEGs include both housekeeping genes and various combinatorial forms of differentially-expressed genes among the subtypes.

We propose the following mixture distribution of OVE-sFC test statistic t under the null hypothesis (**Figure 6**)

$$f\{t|H_{\text{non-SDEG}}\} = \sum_{m=0}^{K-2} f\{t|H_{\text{non-SDEG},\,m}\}P\{H_{\text{non-SDEG},\,m}|H_{\text{non-SDEG}}\}, \tag{4}$$

where $H_{\text{non-SDEG},\,m}$ is the mth component of the mixture null hypothesis $H_{\text{non-SDEG}}$. We design a novel nested permutation scheme that approximates the complex null distribution and is consistent with the definition of SDEGs. Principally, $H_{\text{non-SDEG},\,m}$ is constructed by permuting the samples in the top $(K-m)$ subtypes with higher mean expressions; that is, the samples in the bottom m subtypes with lower mean expressions are removed from the permutation. Note that $H_{\text{non-SDEG},\,0}$ corresponds to the same null distribution used in ANOVA where all samples participate in the permutation.

This mixture null distribution model is proposed to model unknown yet potentially complex expression patterns of non-SDEGs under the null hypothesis. Accordingly, the specifically-designed permutation scheme(s) estimates such a mixture null distribution. The main advantage of the proposed permutation scheme(s) is its flexibility and comprehensiveness, which well match the mixture null distribution of various types and combinations. With varying the proportions of different non-SDEG types, OVE-sFC test is able to maintain the type 1 error rate close to the expected level with the help of the proposed permutation scheme(s) and the conditional probability of each non-SDEG type (**Figure S2**, Supplementary Information).

Note that $H_{\text{non-SDEG},m}$, $m = 0, \ldots, K-2$ represents $(K-1)$ different null hypotheses, each with an individualized null distribution that can be estimated by specific permutation scheme(s), *e.g.*, permuting samples in the top $(K-m)$ subtypes. Collectively, a mixture of null distribution is constructed via combinations of different null hypotheses in various proportions. In contrast, without conditioning on $H_{\text{non-SDEG},m}$, all null distributions are aggregated equally into the mixture null distribution in the same proportion. Consequently, this simpler permutation scheme produces an equal-weight mixture model that cannot represent the complexity of the null distribution. Thus, the null distribution of the OVE-sFC test statistic could become distorted, *e.g.*, a uniform distribution of p-values in null data is not guaranteed, and the observed False Discovery Rate may not match the expected level.

Specifically, the null distribution of OVESEG-test statistics under $H_{\text{non-SDEG},\,m}$ is estimated from permuted samples and aggregated from different genes with weights. Let $\boldsymbol{s}(j) = [s(1,j), \ldots, s(N,j)]$ denote the measured expression vector of gene j across samples, where N is the total number of samples. These weights are the posterior probabilities of a component null hypothesis given the observation $\Pr\{H_{\text{non-SDEG},\,m} \middle| \boldsymbol{s}(j)\}$, estimated by the local FDR $\text{fdr}_{\text{non-SDEG},\,m}(j)$ [30], given by

$$w_{\text{non-SDEG},\,0}(j) = \Pr\{H_{\text{non-SDEG},\,0} \middle| \boldsymbol{s}(j)\} = \text{fdr}_{\text{non-SDEG},\,0}(j), \qquad (5a)$$

$$w_{\text{non-SDEG},\,m}(j) = \Pr\{H_{\text{non-SDEG},\,m} \middle| \boldsymbol{s}(j)\}$$
$$= \left\{1 - \sum_{n=0}^{m-1} w_{\text{non-SDEG},\,n}(j)\right\} \text{fdr}_{\text{non-SDEG},\,m}(j), 0 < m < K-2, \qquad (5b)$$

where $\text{fdr}_{\text{non-SDEG},\,0}(j)$ is the local FDR associated with ANOVA on all subtypes, and $\text{fdr}_{\text{non-SDEG},\,m}(j)$ is the local FDR associated with ANOVA on the top $(K-m)$ subtypes, estimated using R package “fdrtool” [17] (Supplementary Information).

Assessing statistical significance of candidate SDEGs. The p-values of candidate SDEGs are estimated using the learned ‘mixture’ null distribution

$$p\text{-}value = \Pr\{T > t_{obs} | H_{\text{non-SDEG}}\} = \sum_{m=0}^{K-2} \Pr\{T > t_{obs} \middle| H_{\text{non-SDEG},\,m}\} P\{H_{\text{non-SDEG},\,m} \middle| H_{\text{non-SDEG}}\}, (6)$$

where t_{obs} is the observed OVE-sFC test statistic, and T is the continuous dummy random variable. Specifically, $\Pr\{T > t_{obs} | H_{\text{non-SDEG}, m}\}$ is calculated by the weighted permutation scores

$$\Pr\{T > t_{obs} | H_{\text{non-SDEG}, m}\} = \frac{\sum_{p=1}^{P} \sum_{j=1}^{J} w_{\text{non-SDEG}, m}(j) I\left(T_{j,p} > t_{obs}\right)}{P \sum_{j=1}^{J} w_{\text{non-SDEG}, m}(j)}, \tag{7}$$

where P is the number of permutations, J is the number of participating genes, $I(.)$ is the indicator function, and $T_{j,p}$ is the OVE-sFC test statistic in the pth permutation on jth gene. Furthermore, the component weight in the mixture null distribution is estimated by the membership expectation of the posterior probabilities over all genes

$$P\{H_{\text{non-SDEG}, m} | H_{\text{non-SDEG}}\} = \frac{\sum_{j=1}^{J} w_{\text{non-SDEG}, m}(j)}{\sum_{j=1}^{J} \sum_{n=0}^{K-2} w_{\text{non-SDEG}, n}(j)}. \tag{8}$$

Lastly, substituting (7) and (8) into (6), the p-value associated with gene j is calculated by:

$$p\text{-}value = \frac{\sum_{m=0}^{K-2} \sum_{p=1}^{P} \sum_{j=1}^{J} w_{\text{non-SDEG}, m}(j) I\left(T_{j,p} > t_{obs}\right)}{P \sum_{m=0}^{K-2} \sum_{j=1}^{J} w_{\text{non-SDEG}, m}(j)}, \tag{9}$$

with a lower bound of $\min_j \left\{\sum_{m=0}^{K-2} w_{\text{non-SDEG}, m}(j)\right\} / P \sum_{m=0}^{K-2} \sum_{j=1}^{J} w_{\text{non-SDEG}, m}(j)$. Supplementary Information provides more details on the deviation of OVE-sFC test p-values when considering all subtypes together (Eq. 9) and when considering one subtype specifically (Eq. S7-S8).

Empirical Bayes moderated variance estimator of within-subtype expressions. The importance of an accurate estimator on pooled within-subtype variance $\sigma^2(j)$ is twofold - calculating the OVE-sFC test statistic t_j and determining the local false discovery rate $\text{fdr}_{\text{non-SDEG}, m}(j)$, particularly with small sample size. We assume a scaled inverse chi-square prior distribution $\sigma^2(j) \sim \nu_0 \sigma_0^2 / \mathcal{X}_{\nu_0}^2$, where ν_0 and σ_0^2 are the prior degrees of freedom and scaling parameter, respectively[31]. We then adopt the empirical Bayes moderated variance estimator $\tilde{\sigma}^2(j)$ that leverages information across all genes, as used in *limma* and given by

$$\tilde{\sigma}^2(j) = \frac{\nu_0 \hat{\sigma}_0^2 + (N - K)\hat{\sigma}^2(j)}{\nu_0 + N - K}, \tag{10}$$

where N is the total number of samples, and $\hat{\sigma}^2(j)$ is the pooled variance estimator, given by

$$\hat{\sigma}^2(j) = \frac{\sum_{k=1}^{K}\sum_{i=1}^{N_k}\left(\log s_k(i,j) - \mu_k(j)\right)^2}{N-K}. \tag{11}$$

The prior parameters ν_0 and σ_0^2 are estimated from the pooled variances. The moderated variances shrink the pooled variances towards the prior values depending on the prior degrees of freedom and the number of samples. Note that $t\text{-}stat(j)$ with moderated variance estimator $\tilde{\sigma}^2(j)$ follows a t-distribution with $\nu_0 + N - K$ degrees of freedom (Supplementary Information).

Brief review of the most relevant peer SDEG selection methods. The OVR-FC uses a simple test defined by

$$\text{OVR-FC}_k(j) = \frac{\bar{s}_k(j)}{\bar{s}_{\text{-}k}(j)}, \tag{12}$$

where $\bar{s}_k(j)$ and $\bar{s}_{\text{-}k}(j)$ are the geometric means of the jth gene expressions within subtype k and associated with the combined remaining subtypes, respectively. The OVR t-test uses a statistical test given by

$$\text{OVR t-stat}_k(j) = \frac{\hat{\mu}_k(j) - \hat{\mu}_{-k}(j)}{\sqrt{\frac{\hat{\sigma}_k(j)}{N_k} + \frac{\hat{\sigma}_{-k}(j)}{N-N_k}}}, \tag{13}$$

where $\hat{\mu}_k(j)$ and $\hat{\mu}_{-k}(j)$ are the sample means of the jth gene expressions within subtype k and associated with the combined remaining subtypes, respectively; and $\hat{\sigma}_k(j)$ and $\hat{\sigma}_{-k}(j)$ are the sample variance of the jth gene expressions within subtype k and associated with the combined remaining subtypes, respectively. The OVO t-test, conducts t-tests among all subtype pairs and selects genes upregulated in one subtype for all the involved tests, where the variances are estimated only from every pair of subtypes[16] (Supplementary Information). In contrast, the OVE-sFC test exploits all subtypes in estimating the variances. The benefit of using all subtypes for modeling is significant in challenging cases with higher variance, smaller sample size, and more subtypes (Supplementary Information).

Simulation study for validating OVE-sFC test statistics on type 1 error. Among the 10,000 simulated genes, a portion are housekeeping genes that take the baseline expression levels across all subtypes under $H_{\text{non-SDEG, 0}}$. Expression levels of the remaining genes are adjusted to exhibiting similar upregulations in at least two subtypes, mimicking all types of participating null hypotheses. The upregulations are modeled by uniform distribution(s) in scatter space, with variance following an inverse chi-square distribution $\sigma^2(j) \sim \nu_0 \sigma_0^2 / \mathcal{X}_{\nu_0}^2$, where the prior degree of freedom ν_0 takes 5 or 40, and σ_0 takes 0.2, 0.5, or 0.8 (Supplementary Information).

Gene expression data of human immune cells (GSE28490 and GSE28491). In these data sets, each cell subtype consists of at least five samples, excluding few outliers (**Table S7**). Following preprocessing of the raw measurements, 12,022 probesets in Roche and 11,339 probesets in HUG were retained and used in the analyses (Supplementary Information).

Realistic synthetic data for supervised deconvolution. Five subtypes (B cell, CD4+ T cell, CD8+ T cell, NK cell, monocytes) were included in synthesizing n=50 *in silico* mixtures, where purified subtype mean expressions (GSE28491 HUG) were combined according to pre-determined proportions with additive noise, simulating heterogeneous biological samples (Supplementary Information).

Availability of package and supporting data: A Bioconductor approved R package of OVE-sFC test is freely available at http://bioconductor.org/packages/OVESEG. A detailed user's manual and a vignette are provided within the package. In addition, public gene expression data analyzed in this paper are also available from the Gene Expression Omnibus Database under Accession Number GEO: GSE19380, GSE28490, GSE28491, GSE60424, and GSE72056.

REFERENCE

1 Wang, Y., Lu, J., Lee, R., Gu, Z. & Clarke, R. Iterative normalization of cDNA microarray data. *IEEE Trans Inf Technol Biomed* **6**, 29-37 (2002).

2 Abbas, A. R., Wolslegel, K., Seshasayee, D., Modrusan, Z. & Clark, H. F. Deconvolution of Blood Microarray Data Identifies Cellular Activation Patterns in Systemic Lupus Erythematosus. *PLoS ONE* **4**, e6098, doi:10.1371/journal.pone.0006098 (2009).

3 Montano, C. *et al.* Measuring cell-type specific differential methylation in human brain tissue. *Genome Biology* **14**, R94 (2013).

4 Shen-Orr, S. S. *et al.* Cell type-specific gene expression differences in complex tissues. *Nat Methods* **7**, 287-289, doi:10.1038/nmeth.1439 (2010).

5 Yu, G. *et al.* Matched Gene Selection and Committee Classifier for Molecular Classification of Heterogeneous Diseases. *J. Mach. Learn. Res.* **11**, 2141-2167 (2010).

6 Yu, G. *et al.* PUGSVM: a caBIG analytical tool for multiclass gene selection and predictive classification. *Bioinformatics* **27**, 736-738 (2011).

7 Qiao, W. *et al.* PERT: a method for expression deconvolution of human blood samples from varied microenvironmental and developmental conditions. *PLoS Comput Biol* **8**, e1002838, doi:10.1371/journal.pcbi.1002838
PCOMPBIOL-D-12-00905 [pii] (2012).

8 Kuhn, A., Thu, D., Waldvogel, H. J., Faull, R. L. & Luthi-Carter, R. Population-specific expression analysis (PSEA) reveals molecular changes in diseased brain. *Nat Methods* **8**, 945-947, doi:nmeth.1710 [pii]
10.1038/nmeth.1710 (2011).

9 Herrington, D. M. *et al.* Proteomic Architecture of Human Coronary and Aortic Atherosclerosis. *Circulation* **137**, 2741-2756, doi:10.1161/CIRCULATIONAHA.118.034365 (2018).

10 Chikina, M., Zaslavsky, E. & Sealfon, S. C. CellCODE: a robust latent variable approach to differential expression analysis for heterogeneous cell populations. *Bioinformatics* **31**, 1584-1591, doi:10.1093/bioinformatics/btv015 (2015).

11 Zhang, Y. *et al.* An RNA-sequencing transcriptome and splicing database of glia, neurons, and vascular cells of the cerebral cortex. *J Neurosci* **34**, 11929-11947, doi:10.1523/jneurosci.1860-14.2014 (2014).

12 Shoemaker, J. E. *et al.* CTen: a web-based platform for identifying enriched cell types from heterogeneous microarray data. *BMC Genomics* **13**, 460, doi:10.1186/1471-2164-13-460 (2012).

13 Chen, Z. *et al.* Inference of immune cell composition on the expression profiles of mouse tissue. *Sci Rep* **7**, 40508, doi:10.1038/srep40508 (2017).

14 Amrani, K. E., Stachelscheid, H., Lekschas, F., Kurtz, A. & Andrade-Navarro, M. A. MGFM: a novel tool for detection of tissue and cell specific marker genes from microarray gene expression data. *BMC Genomics* **16**, 645, doi:10.1186/s12864-015-1785-9 (2015).

15 Hsu, J. C. *Multiple comparisons : theory and methods*. xiv, 277 pages, [4] pages of plates : illustrations (some color) ; 24 cm (Chapman & Hall, 1996).

16 Wang, M., Master, S. R. & Chodosh, L. A. Computational expression deconvolution in a complex mammalian organ. *BMC Bioinformatics* **7**, 328-328, doi:10.1186/1471-2105-7-328 (2006).

17 Strimmer, K. fdrtool: a versatile R package for estimating local and tail area-based false discovery rates. *Bioinformatics* **24**, 1461-1462, doi:10.1093/bioinformatics/btn209 (2008).

18 McClish, D. K. Analyzing a portion of the ROC curve. *Medical decision making : an international journal of the Society for Medical Decision Making* **9**, 190-195, doi:10.1177/0272989x8900900307 (1989).

19 Zuckerman, N. S., Noam, Y., Goldsmith, A. J. & Lee, P. P. A Self-Directed Method for Cell-Type Identification and Separation of Gene Expression Microarrays. *PLoS computational biology* **9**, e1003189, doi:10.1371/journal.pcbi.1003189 (2013).

20 Allantaz, F. *et al.* Expression Profiling of Human Immune Cell Subsets Identifies miRNA-mRNA Regulatory Relationships Correlated with Cell Type Specific Expression. *PLoS ONE* **7**, e29979, doi:10.1371/journal.pone.0029979 (2012).

21 Wang, N. *et al.* Mathematical modelling of transcriptional heterogeneity identifies novel markers and subpopulations in complex tissues. *Scientific Reports* **6**, 18909, doi:10.1038/srep18909
https://www.nature.com/articles/srep18909#supplementary-information (2016).

22 Kuhn, A. *et al.* Cell population-specific expression analysis of human cerebellum. *BMC Genomics* **13**, 610 (2012).

23 Becht, E. *et al.* Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. *Genome Biology* **17**, 218, doi:10.1186/s13059-016-1070-5 (2016).

24 Aran, D., Hu, Z. & Butte, A. J. xCell: digitally portraying the tissue cellular heterogeneity landscape. *Genome Biology* **18**, 220, doi:10.1186/s13059-017-1349-1 (2017).

25 Newman, A. M. *et al.* Robust enumeration of cell subsets from tissue expression profiles. *Nature methods* **12**, 453-457, doi:10.1038/nmeth.3337 http://www.nature.com/nmeth/journal/v12/n5/abs/nmeth.3337.html#supplementary-information (2015).

26 Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* **43**, e47-e47, doi:10.1093/nar/gkv007 (2015).

27 McCarthy, D. J., Chen, Y. & Smyth, G. K. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res* **40**, 4288-4297, doi:10.1093/nar/gks042 (2012).

28 Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology* **15**, 550-550, doi:10.1186/s13059-014-0550-8 (2014).

29 Kao, L. S. & Green, C. E. Analysis of Variance: Is There a Difference in Means and What Does It Mean? *The Journal of surgical research* **144**, 158-170, doi:10.1016/j.jss.2007.02.053 (2008).

30 Guo, X. & Pan, W. Using weighted permutation scores to detect differential gene expression with microarray data. *Journal of Bioinformatics and Computational Biology* **03**, 989-1006, doi:10.1142/s021972000500134x (2005).

31 Smyth, G. K. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* **3**, Article3, doi:10.2202/1544-6115.1027 (2004).

ADDITIONAL INFORMATION

Supplementary information accompanies this paper at https://doi.org/

ACKNOWLEDGMENTS

This work was funded in part by the National Institutes of Health under Grants HL111362-05A1, HL133932, W81XWH-18-1-0723 (BC171885P1), and U01NS115658-01.

AUTHOR CONTRIBUTIONS

L.C. and Y.W. developed OVE-FC/sFC test framework and wrote the manuscript; L.C. implemented OVE-FCs/FC test algorithm; L.C and Y.L. performed real data analysis; D.M.H., J.E.V.E. and R.C. interpreted results and edited the manuscript; G.Y. provided statistical expertise support.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

FIGURE LEGENDS

Figure 1. Assessment on Type 1 error rates and p-value distributions using simulated data sets under the null hypothesis, involving three subtypes with unbalanced sample sizes ($N_1 = 3, N_2 = 6, N_3 = 9$). (a) Bar chart for the mean and 95% confidence interval of type I error rates with p-value cutoff at 0.05 over 30 simulated experiments, showing both overall and subtype-specific false-positive rates corresponding to different permutation schemes. (b) Histograms of p-value distributions associated with five SDEG detection methods, where simulation data consisted of 60% housekeeping genes, $\sigma_0 = 0.5$ and $\nu_0 = 40$. Note that subtype-specific p-values can be higher than 1.0 after multiple testing correction and thus will be truncated (indicated by the blue circle; see Supplementary Information for details).

Figure 2. Assessment on Type 1 error rates using simulated data sets involving five subtypes. The results are obtained using the p-value cutoff at 0.05 over 30 experiments. (a) Bar chart of the mean and 95% confidence interval of type I error rates with unbalanced sample sizes. (b) Bar chart of the mean and 95% confidence interval of type I error rates with unbalanced compositions of mixture null distribution.

Figure 3. Assessment on detection power (partial ROC curves, FPR < 0.01) using real data derived simulations (data distribution is consistent with the base real dataset under null hypothesis; variances are sampled from real microarray data GSE28490 or RNAseq data GSE60424 with keeping mean-variance trend) involving seven unbalanced subtypes with various parameter settings. (a)(b) partial ROC curves across different FPR points on microarray-derived data. (c)(d) partial ROC curves across different FPR points on RNAseq-derived data. (OVR-FC and OVR t-test are not shown here due to low pAUC; subtype-specific OVE-sFC test's performance is quite similar to OVE-sFC test; more complete ROC curves can be found in **Figure S7**.)

Figure 4. Percentile overlap of cell-type specific SDEGs between Roche and HUG datasets, quantified by Jaccard index (intersection over union). SDEGs are detected by subtype-specific OVE-sFC test with q-value < 0.05.

Figure 5. Correlation coefficients between CAM scores and ground truth proportions in simulated heterogeneous samples of mixed subtype mRNA expression profiles or RNAseq

counts (a-c based on three different real gene expression datasets). CAM scores are estimated using the detected SDEGs from independent dataset and reflect the proportions of subtypes. The mean and 95% confidence interval are computed over 20 repeated experiments. (OVR t-test results are not shown in (c) due to very poor performance.)

Figure 6. Mixture null distribution of OVE-sFC test statistics for detecting SDEGs. The mixture distribution consists of $(K-1)$ null components, each estimated from permuting samples in the top $(K-m)$ subtypes of high mean expressions and weighted by the posterior probabilities of component null hypotheses.

Figure 1

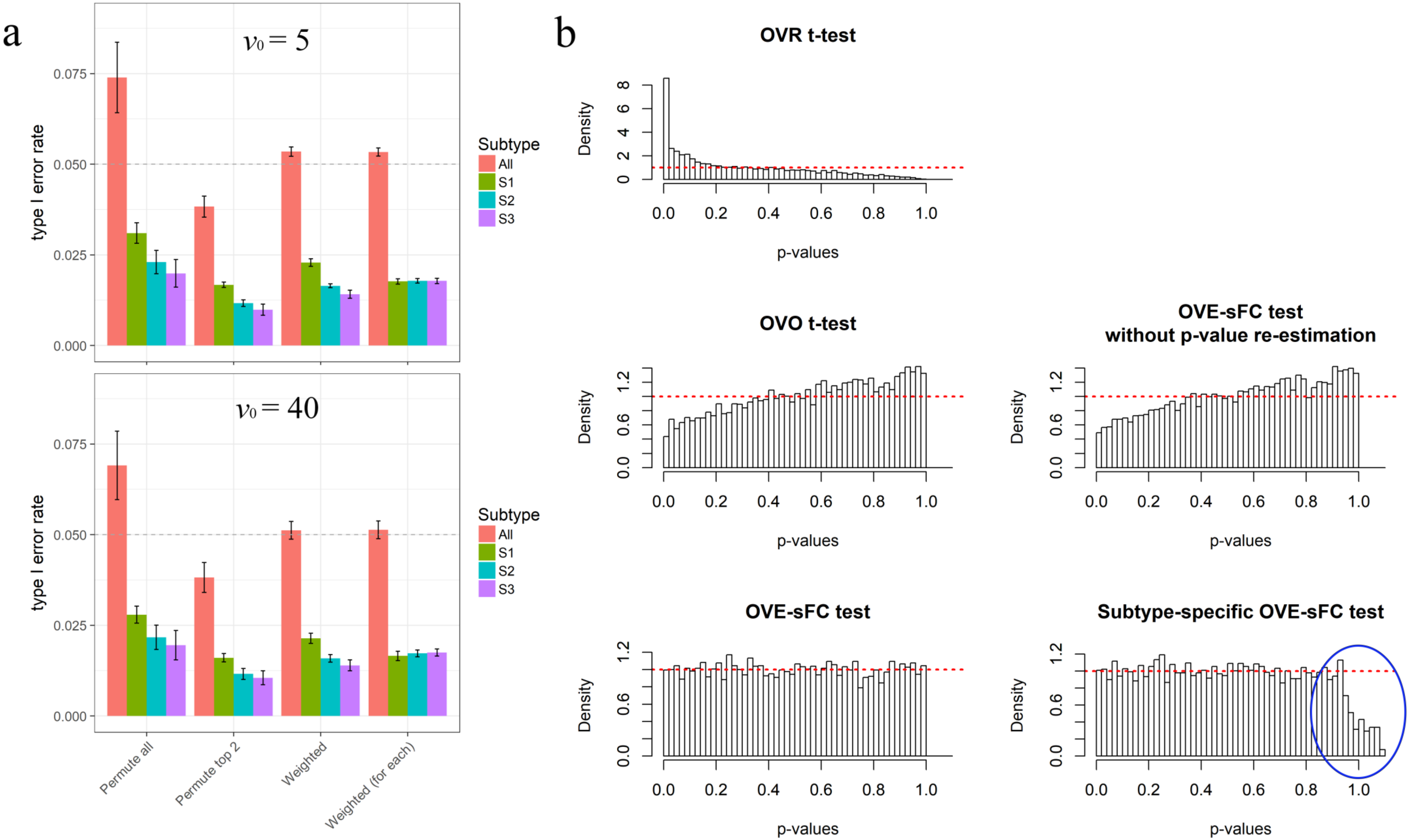
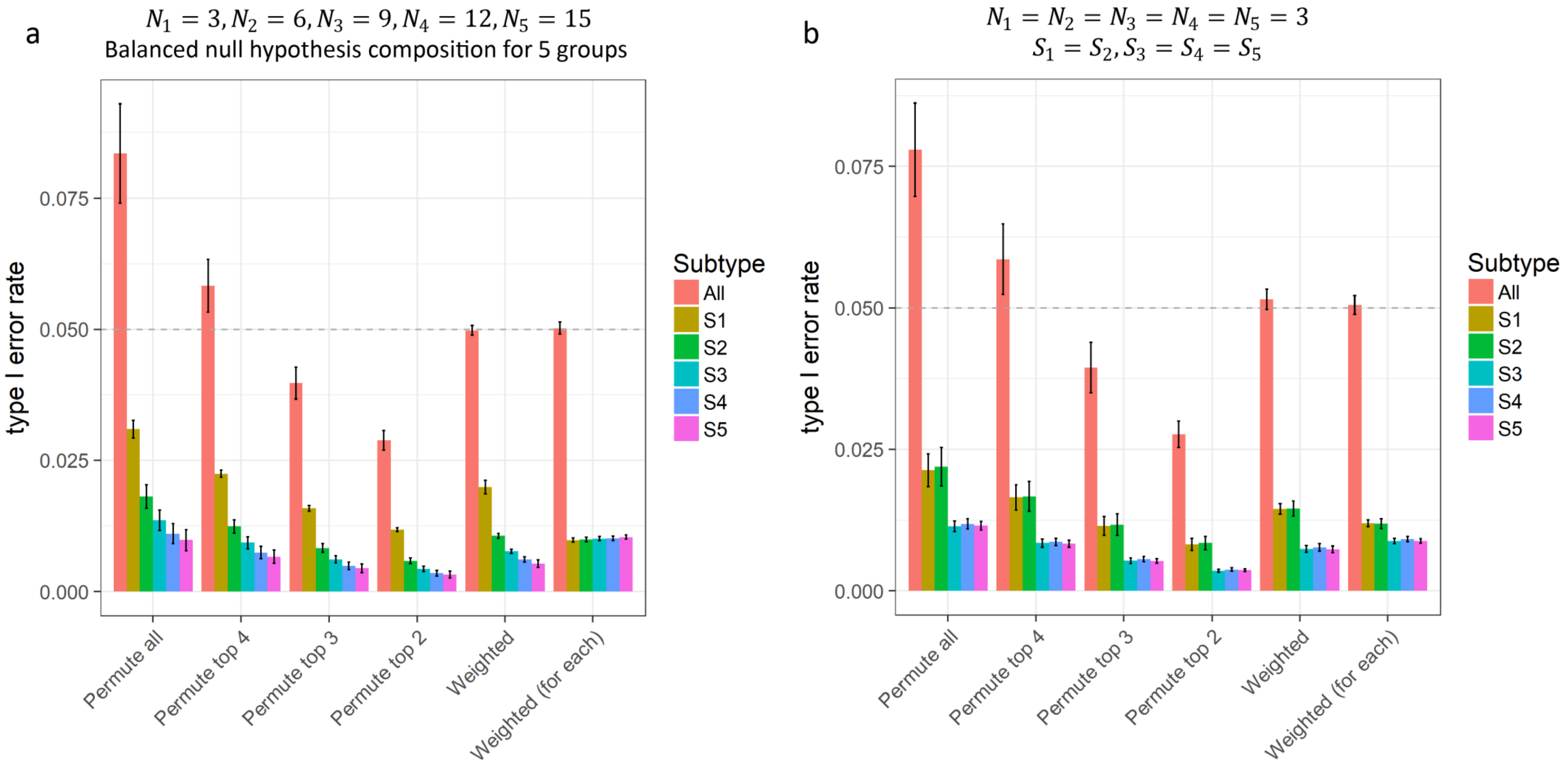

a
$\nu_0 = 5$
0.075
0.050
0.025
0.000
type I error rate
Subtype
All
S1
S2
S3
$\nu_0 = 40$
0.075
0.050
0.025
0.000
type I error rate
Subtype
All
S1
S2
S3
Permute all
Permute top 2
Weighted
Weighted (for each)
b
OVR t-test
Density
0 2 4 6 8
0.0 0.2 0.4 0.6 0.8 1.0
p-values
OVO t-test
Density
0.0 0.6 1.2
0.0 0.2 0.4 0.6 0.8 1.0
p-values
OVE-sFC test
without p-value re-estimation
Density
0.0 0.6 1.2
0.0 0.2 0.4 0.6 0.8 1.0
p-values
OVE-sFC test
Density
0.0 0.6 1.2
0.0 0.2 0.4 0.6 0.8 1.0
p-values
Subtype-specific OVE-sFC test
Density
0.0 0.6 1.2
0.0 0.2 0.4 0.6 0.8 1.0
p-values

Figure 2



a $N_1 = 3, N_2 = 6, N_3 = 9, N_4 = 12, N_5 = 15$
Balanced null hypothesis composition for 5 groups

b $N_1 = N_2 = N_3 = N_4 = N_5 = 3$
$S_1 = S_2, S_3 = S_4 = S_5$

Figure 3

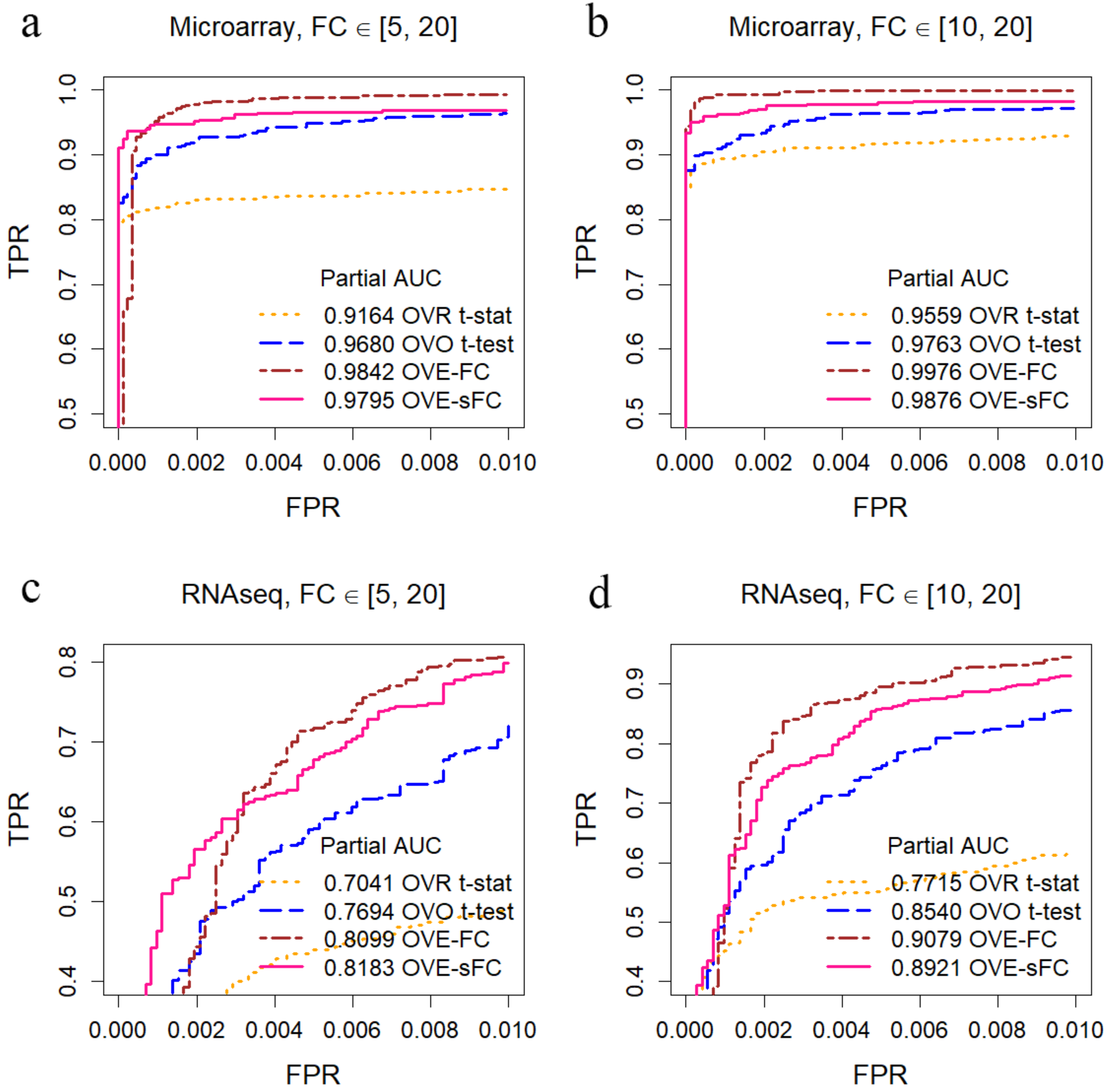
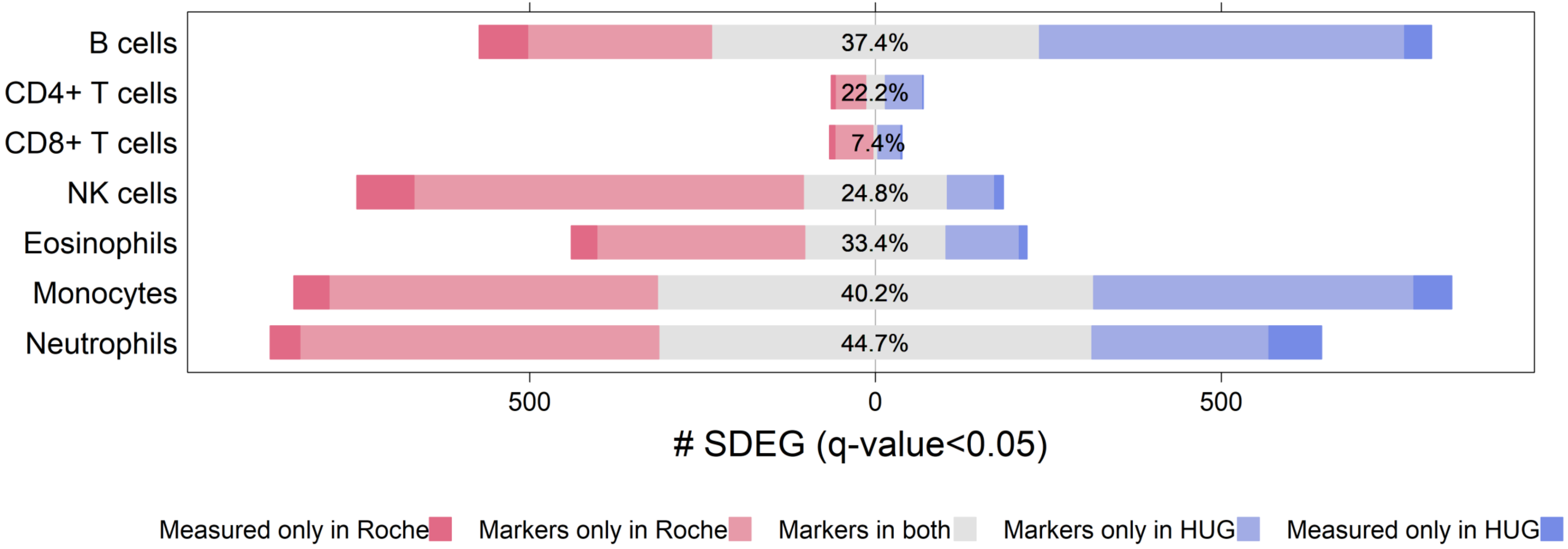

a
Microarray, FC ∈ [5, 20]
TPR
FPR
0.5
0.6
0.7
0.8
0.9
1.0
0.000
0.002
0.004
0.006
0.008
0.010
Partial AUC
0.9164 OVR t-stat
0.9680 OVO t-test
0.9842 OVE-FC
0.9795 OVE-sFC
b
Microarray, FC ∈ [10, 20]
TPR
FPR
0.5
0.6
0.7
0.8
0.9
1.0
0.000
0.002
0.004
0.006
0.008
0.010
Partial AUC
0.9559 OVR t-stat
0.9763 OVO t-test
0.9976 OVE-FC
0.9876 OVE-sFC
c
RNAseq, FC ∈ [5, 20]
TPR
FPR
0.4
0.5
0.6
0.7
0.8
0.000
0.002
0.004
0.006
0.008
0.010
Partial AUC
0.7041 OVR t-stat
0.7694 OVO t-test
0.8099 OVE-FC
0.8183 OVE-sFC
d
RNAseq, FC ∈ [10, 20]
TPR
FPR
0.4
0.5
0.6
0.7
0.8
0.9
0.000
0.002
0.004
0.006
0.008
0.010
Partial AUC
0.7715 OVR t-stat
0.8540 OVO t-test
0.9079 OVE-FC
0.8921 OVE-sFC

Figure 4



B cells
CD4+ T cells
CD8+ T cells
NK cells
Eosinophils
Monocytes
Neutrophils
37.4%
22.2%
7.4%
24.8%
33.4%
40.2%
44.7%
500
0
500
# SDEG (q-value<0.05)
Measured only in Roche
Markers only in Roche
Markers in both
Markers only in HUG
Measured only in HUG

Figure 5

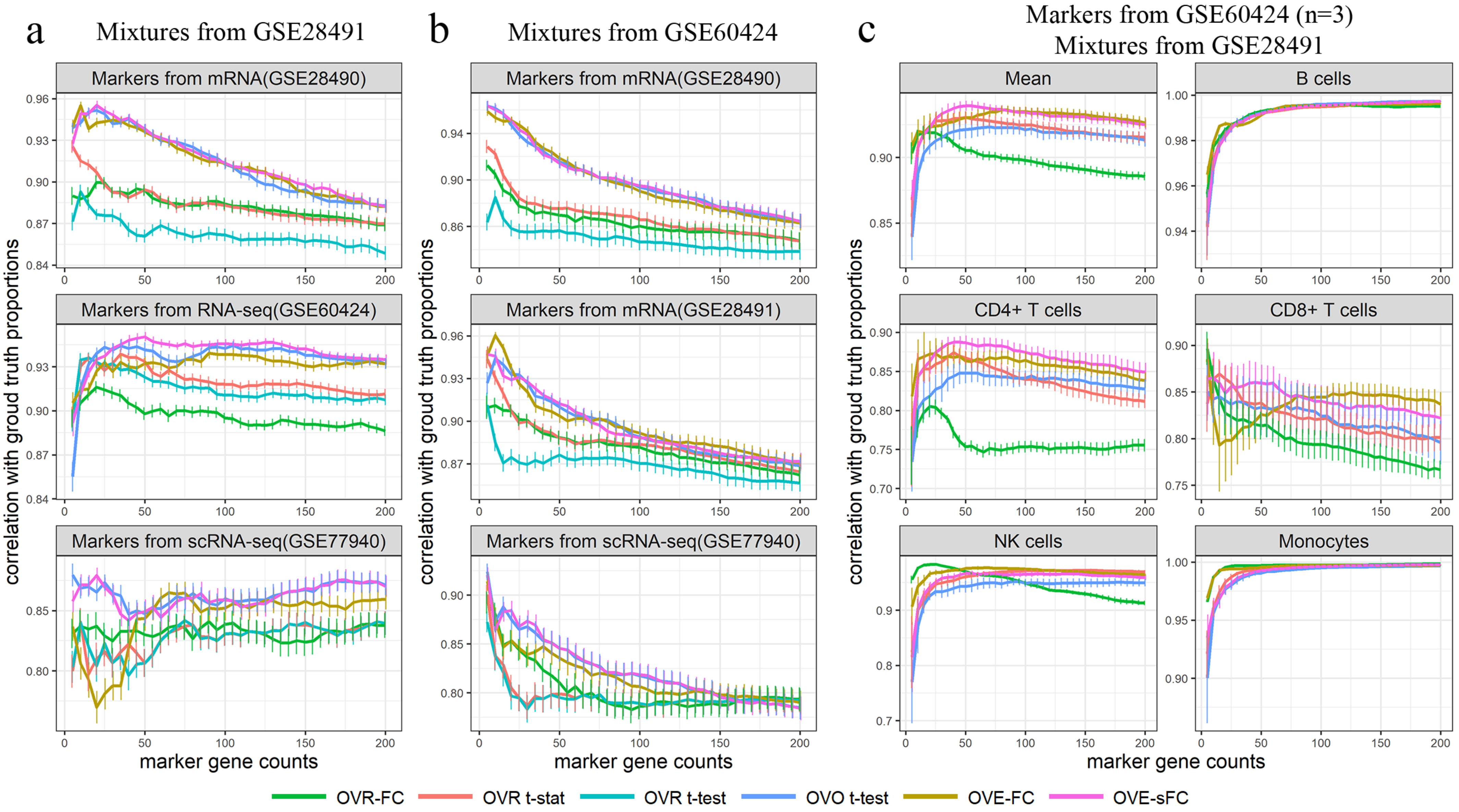
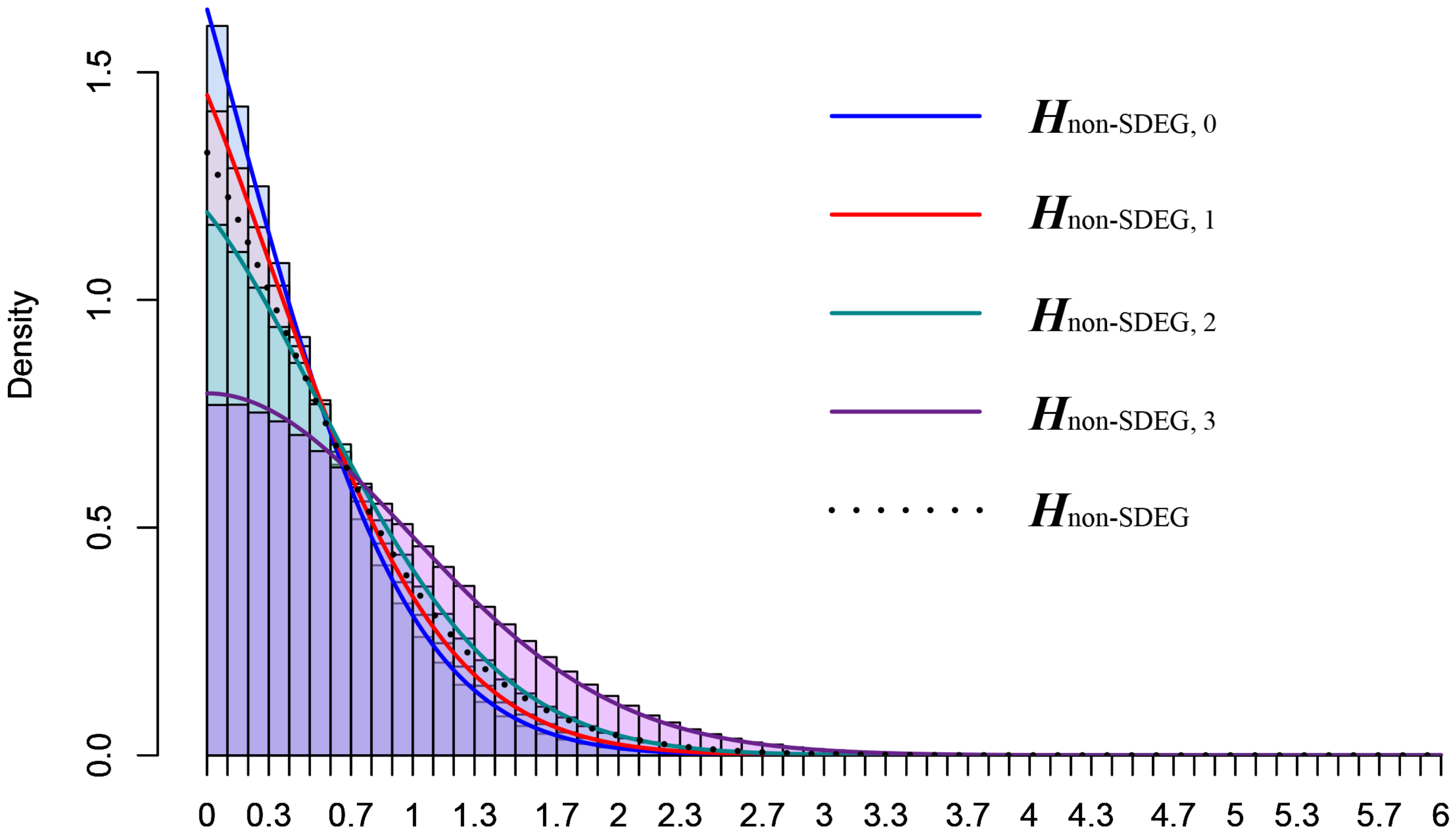

a
Mixtures from GSE28491
Markers from mRNA(GSE28490)
Markers from RNA-seq(GSE60424)
Markers from scRNA-seq(GSE77940)
b
Mixtures from GSE60424
Markers from mRNA(GSE28490)
Markers from mRNA(GSE28491)
Markers from scRNA-seq(GSE77940)
c
Markers from GSE60424 (n=3)
Mixtures from GSE28491
Mean
B cells
CD4+ T cells
CD8+ T cells
NK cells
Monocytes
correlation with groud truth proportions
marker gene counts
0
50
100
150
200
0.84
0.87
0.90
0.93
0.96
0.94
0.86
0.85
0.80
0.75
0.70
1.00
0.98
0.95
0.9
0.8
0.7
OVR-FC
OVR t-stat
OVR t-test
OVO t-test
OVE-FC
OVE-sFC

Figure 6


Density
0.0
0.5
1.0
1.5
0
0.3
0.7
1
1.3
1.7
2
2.3
2.7
3
3.3
3.7
4
4.3
4.7
5
5.3
5.7
6
$H_{\text{non-SDEG},\,0}$
$H_{\text{non-SDEG},\,1}$
$H_{\text{non-SDEG},\,2}$
$H_{\text{non-SDEG},\,3}$
$H_{\text{non-SDEG}}$