Data-driven robust detection of tissue/cell-specific markers

Lulu Chen[1], David M. Herrington[2], Robert Clarke[3], Guoqiang Yu[1], and Yue Wang[1,#]

[1]Department of Electrical and Computer Engineering, Virginia Polytechnic Institute and State University, Arlington, VA 22203, USA; [2]Department of Internal Medicine, Wake Forest University, Winston-Salem, NC 27157, USA; [3]Lombardi Comprehensive Cancer Center, Georgetown University, Washington, DC 20057, USA

Article Format:	bioRxiv
Running title:	Marker detection
Open source software:	R package of OVESEG-test is available at https://github.com/Lululuella/OVESEG

[#]Author for correspondence:

Yue Wang, Ph.D.
Virginia Tech Research Center - Arlington
900 N. Glebe Road, Arlington, VA 22203
E-mail: yuewang@vt.edu

Tissue/cell-specific marker genes (MGs) are defined as being exclusively and consistently expressed in a particular tissue/cell subtype across varying conditions. Detecting MGs plays a critical role in molecularly characterizing and conferring tissue/cell subtypes. Unfortunately, classic differential analysis assumes a convenient statistical distribution for the null hypothesis that however does not enforce MG definition and thus results in high false positives. Here we describe a statistically-principled method, One Versus Everyone Subtype Exclusively-expressed Genes (OVESEG) test, and propose a mixture null distribution model estimated via novel permutation schemes. Validated with realistic synthetic data sets on both type 1 error and detection power, OVESEG-test applied to two benchmark gene expression data sets detects many known and de novo subtype-specific markers. The subsequent supervised deconvolution results, obtained using MGs detected by OVESEG-test, show superior performance as compared with that by peer methods.

In molecular characterization of biological systems, members of molecular profiles (*e.g.*, gene expressions) can be divided into three major categories: house-keeping genes (constantly expressed genes – CEGs) [1], differentially expressed genes (DEGs) [2, 3], and tissue/cell-specific marker genes (MGs) [4]. In particular, MGs are defined as being exclusively and consistently expressed in a particular tissue/cell subtype across varying conditions [5-8]. Moreover, when MGs are known *a priori*, they are often used to not only define molecular subtypes [9, 10] but also support supervised data deconvolution [5-8].

Detecting MGs using tissue/cell-specific molecular expression profiles is an important yet challenging task [11]. Even with the conceptual yet subtle differences between MGs and DEGs, classical differential analysis methods have been conveniently extended to detecting MGs. For example, one-way 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 [12]. However, ANOVA uses the null hypothesis that samples in all subtypes are drawn from same population thus detects DEGs of any forms across subtypes, not truly enforcing the definition of MGs. One-Versus-Rest Fold Change (OVR-FC) is another popular method that is based on the ratio of the average expression in a particular subtype to the average expression in the rest [11, 13, 14]. OVR t-test is occasionally used to assess the statistical significance [15]. Nevertheless, a gene with low average expression in the rest is not necessarily low-expressed in every subtype in the rest, clearly violating the definition of MGs. On the other hand, One-Versus-Everyone Fold

Change (OVE-FC) [16, 17] has been proposed to specifically detect MGs that has led to some novel discoveries [9, 18] and much improved classification [16, 17]. OVE-FC checks whether the mean of one subtype is significantly higher/lower than the highest/lowest mean among every others, thus is consistent with the definition of MGs [7, 9]. Supportively, simulation studies show that Marker Gene Finder in Microarray data (MGFM), a method similar to OVE-FC, outperforms OVR t-test [19]. Other similar strategies include Multiple Comparisons with the Best (MCB) [20] and OVE t-test that use additional pairwise significance test or confidence intervals of OVE statistics [4, 21], while without rigorous modeling the null distribution in relation to the definition of MGs.

To address the critical problem of the absence of detection method explicitly matched the definition of MGs, we developed a statistically-principled method (One Versus Everyone Subtype Exclusively-expressed Genes – OVESEG-test) that can detect tissue/cell-specific MGs among many subtypes. OVESEG-test is based on our earlier work on detecting One Versus Everyone Phenotype Unregulated Genes – OVEPUG [9, 16, 17]. To assess the statistical significance of MGs, OVESEG-test uses a specifically designed test statistics that mathematically matches the definition of MGs, and employs a specifically designed novel permutation scheme to estimate the corresponding distribution under null hypothesis where the expression patterns of non-MGs can be highly complex (Methods).

We validate the performance of OVESEG-test on extensive simulation data, in terms of type 1 error rate, False Discovery Rate (FDR), partial area under the receiver operating characteristic curve (pAUC), and comparisons with top peer methods. We demonstrate the utility of OVESEG-test by applying it to benchmark public data, and assess the performance by comparisons with the known MGs reported in literature and by the accuracy of supervised deconvolution that uses the de novo MGs detected by OVESEG-test.

Results

Validation of OVESEG-test statistics on type 1 error using simulated datasets

To test whether our OVESEG-test statistics can detect MGs at the right significance levels, we assessed the type 1 error via simulation studies under the null hypothesis (Methods). Accuracy of type 1 error is crucial for any hypothesis testing methods that detect MGs based on their p-values, because if the type 1 error is either too conservative or too liberal, the p-value loses its intended meaning and fails to reflect the actual false positives.

The simulation data contain 10,000 genes whose baseline expression levels are sampled from the real benchmark microarray gene expression data with purified replicates (GSE19380 [7]). Using the realistic simulation data sets with various parameter settings, we show that in all scenarios the empirical type 1 error produced by OVESEG-test statistics closely approximates the expected type 1 error (**Fig. 1a, Fig. 2a-b, Fig. S2**), and the p-values associated with OVESEG-test statistics expectedly exhibit a uniform distribution. Specifically, 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 (**Fig. 2a** and **Fig. S2**). In contrast, the empirical type 1 error produced by OVR t-test and OVE t-test either over-estimates or under-estimates the expected type 1 error, and the p-values associated with OVR t-test and OVE t-test clearly deviate from a uniform distribution (**Fig. 1b**). We also evaluate the type 1 error under high noise levels and small sample sizes using subtype-specific p-value estimates. For each of the subtypes, experimental results again show that the empirical type 1 error produced by OVESEG-test statistics matches very well the expected type 1 error (**Fig. 1b**, Supplementary Information).

We conduct similar validation studies involving five subtypes over a wide range of simulation scenarios. The experimental results show that OVESEG-test statistics produces the empirical type 1 error rates that match well the expected type 1 error rates, where subtype-specific p-value estimates effectively balance the uneven type 1 error rates among the subtypes with different numbers of upregulated genes (**Fig. 2b**, Supplementary Information).

Comparative assessment of OVESEG-test statistics on power of detecting subtype-specific markers using simulated datasets

For power considerations, we simulated a comprehensive set of scenarios to examine the power of OVESEG-test statistics and peer methods in detecting subtype-specific MGs (Method and Supplementary Information). The simulation data are generated, similarly, by modifying the expression levels of real gene expression data, where about 20% of the genes are designated as MGs with exclusively and consistently upregulations in each of the participating subtypes, following a uniform distribution. 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.

When assessing the detection power involving true MGs, FDR control is an important factor because for a well-designed significance test, the objective is to maximize power while controlling FDR below the allowable level. To test whether the q-value can reflect true FDR, ‘fdrtool’ package is used to estimate the q-value for each gene [22], where the FDR with estimated q-value of 0.05 is expected to be around 0.05. Another informative criterion is the pAUC that emphasizes the leftmost portion of the receiver operating characteristic curve, focusing on the sensitivity at low FDR.

Experimental results show that both overall and subtype-specific OVESEG-test statistics achieves well-controlled FDR that matches nicely the q-value cutoff (**Fig. 3a**, **Fig. S5**), while OVR t-test underestimates and pairwise OVE t-test overestimates the FDR (Supplementary Information). Moreover, subtype-specific OVESEG-test statistics attains well-balanced false positive MGs across subtypes while peer methods produce higher false positive MGs in the subtypes of small sample size.

In terms of pAUC, experimental results show that OVESEG-test strategy achieves nearly the highest power in detecting true MGs (**Fig. 3b**, **Table S1**), and subtype-specific OVESEG-test further improves the power especially when null hypothesis composition is unbalanced. Furthermore, independent pairwise OVE t-test shows comparable yet slightly less competitive detection power, OVE-FC exhibits lower detection power in highly noisy cases. In contrast, all three OVR methods show much weaker detection power, and ANOVA attains expectedly the lowest detection power (Supplementary Information).

Application of OVESEG-test statistics on two benchmark gene expression data sets detects subtype-specific markers (human immune cells)

We apply OVESEG-test statistics to two real microarray gene expression data sets, GSE28490 (Roche) and GSE28491 (HUG), to detect subtype-specific markers associated with human immune cells [23]. In these data sets, the constituent subtypes are seven human immune cells isolated from healthy human blood that are phenotypically very similar to each other: 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 panel of donors, the derived gene expression profiles serve well for the purpose of analytics cross-validation.

With FDR control of q-value < 0.05 applied to both data sets, OVESEG-test detects 28 CD4+ T cell markers, 7 CD8+ T cell markers, and numerous markers for other more distinctive cell types (**Table S2-S4**). Between the two data sets, a Jaccard index (intersection over union) reaches 36.8% for all MGs across all cell types, and in particular, the overlap of monocyte markers, as well as neutrophil markers, detected from the two datasets is over 40%

(**Fig. 4**). The number of subtype-specific MGs account for about one third of all probesets (Roche: 39%, HUG: 34%), and this is expected because these subtypes are pure cell types, more distinctive as compared with multicellular tissue types [9, 18, 24]. Moreover, we employ Bonferroni multiple testing correction and a more stringent p-value < 0.001, the number of MGs account for 10.7% and 2.7% of all probesets in Roche and HUG data sets, respectively (**Table S2**), with only one common CD4+ T cell marker (FHIT) and one common CD8+ T cell marker (CD8B).

To portrait the upregulation patterns among cell types (**Fig. S6**), probeset-wise posterior probabilities of component hypotheses in the null mixtures (Eq. 4) are accumulated and normalized to estimate the counterpart probabilities of alternative hypotheses (Eq. S10), where the upregulations in B cells, monocytes, or neutrophils ranks the top in both datasets, followed by the upregulations in lymphoid cells (B cells, CD4+ T cells, CD8+ T cells, NK cells) and T cells (CD4+ T cells, CD8+ T cells) in Roche dataset.

Evaluation of MGs detected by OVESEG-test statistics via supervised deconvolution

Accurate and reliable detection of MGs has significant impact on the performance of many supervised deconvolution methods that use the expression patterns of MGs to score constituent subtypes in heterogeneous samples [25-27]. In our experiments, CAM score derived from MGs-guided supervised deconvolution is adopted to quantify each subtype (Supplementary Information), and correlation coefficient between estimated scores and true proportions is used to assess the accuracy of various MGs selection methods.

OVESEG-test statistics is 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 subtype-specific MGs are detected by five different methods including OVR-FC, OVR t-stat, OVR t-test, OVE-FC, and OVESEG-test, and are used to supervise the deconvolution of realistically synthesized mixtures with ground truth.

Supervised deconvolution results show that, measured by CAM score derived from expression levels of top-ranked markers for each subtype, OVESEG-test and OVE-FC achieve the highest correlation coefficients between CAM score and true proportions as compared with the performance produced by other methods (**Fig. 5a**). More importantly, the subtype-specific MGs detected by OVESEG-test or OVE-FC have led to the significantly improved deconvolution of hard-to-separate molecularly-similar subtypes (**Fig. S7**), demonstrating the key advantage of OVESEG-test or OVE-FC for detecting true MGs.

As a more challenging and biologically realistic case involving between-sample variations, we synthesize a set of 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 subtype-specific MGs detected by OVESEG-test or OVE-FC achieved superior deconvolution performance (**Fig. 5b, Fig. S8**). Moreover, stringent OVESEG-test p-value threshold, e.g., < 0.001 after correction (**Table S2**) is a good option, because suitable number of MGs for CD4+ or CD8+ T cell is 5~20, while B cells or monocytes often allows more MGs.

Discussions

While ideal MGs are defined as being exclusively and consistently expressed in a particular tissue/cell subtype across varying conditions [5-8], biological reality dictates a more relaxed definition that allows MGs of a particular tissue/cell subtype having low or insignificant expressions in all other subtypes. Experimental results show that MGs detected by OVESEG-test with small p-values can accurately estimate both subtype proportions and expression profiles (**Fig. S7 and S8**).

It is noted that the accuracy of OVESEG-test based MG detection may be affected by batch effect and normalization methods, and the reliability would depend on the variance estimate particularly when sample size is small. One solution adopted in our method is to leverage the ability of "limma-zoom" that can borrow information across genes and model the mean-variance relationship.

Methods

OVESEG-test statistics. Consider the measured expression level $s_k(i,j)$ of gene j in sample i across $k = 1, \dots, \dots K$ subtypes. We assume that $\log s_k(i,j) \sim N(\mu_k(j), \sigma^2(j))$, where $\mu_k(j)$ and $\sigma^2(j)$ are the mean and variance of gene j logarithmic expressions in subtype k. We define OVESEG-test statistics for gene j that matches to the definition of MGs as [7, 16]

$$t_j = \max_{k=1,\dots,K} \left\{ \min_{l \neq k} \left\{ \frac{\mu_k(j) - \mu_l(j)}{\sigma(j)\sqrt{\frac{1}{N_k} + \frac{1}{N_l}}} \right\} \right\}, \tag{1}$$

where N_k and N_l are the numbers of samples in subtypes k and l, respectively. Conceptually, the null hypothesis about non-MGs and alternative hypothesis about MGs can be described as

$$\begin{aligned} H_{\text{non-MG}}&: d_j = 0; \\ H_{\text{MG}}&: d_j > 0; \end{aligned} \tag{2}$$

where $d_j = \max\limits_{k=1,\ldots,K}\left\{\min\limits_{l\neq k}\{\mu_k(j) - \mu_l(j)\}\right\}$ [17] (Supplementary Information).

Modeling OVESEG-test statistics under null hypothesis. It can be expected that, for more than two subtypes $K \geq 3$, modeling the null distribution of OVESEG-test statistics is challenging due to highly complex expression patterns of non-MGs. As aforementioned, under the null hypothesis, non-MGs include all the counterparts of MGs, i.e., CEGs, and DEGs of various combinatorial forms.

We propose the following mixture distribution to model OVESEG-test statistics under null hypothesis (**Fig. 6**)

$$f\{t|H_{non-MG}\} = \sum_{m=0}^{K-2} f\{t|H_{\text{non-MG}, m}\}P\{H_{\text{non-MG}, m}|H_{\text{non-MG}}\}, \tag{3}$$

where t is the OVESEG-test statistics, and $H_{\text{non-MG}, m}$ is the mth component of the mixture null hypothesis $H_{\text{non-MG}}$. We design a novel nested permutation scheme that not only approximates the complex null distribution but also preserves the definition of MGs. Principally, $H_{\text{non-MG}, m}$ 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 permutation. Note that $H_{\text{non-MG}, 0}$ corresponds to the same null distribution used in ANOVA where all samples participate in permutation. Specifically, the null distribution of OVESEG-test statistics under $H_{\text{non-MG}, m}$ is estimated based on permuted samples and aggregated from different genes with weights. These weights are the posterior probabilities of a component null hypothesis given the observation $\Pr\{H_{\text{non-MG}, m}|\boldsymbol{s}(j)\}$, estimated by the local FDR $\text{fdr}_{\text{non-MG}, m}(j)$ [28], given by

$$w_{\text{non-MG}, 0}(j) = \Pr\{H_{\text{non-MG}, 0}|\boldsymbol{s}(j)\} = \text{fdr}_{\text{non-MG}, 0}(j), \tag{4a}$$

$$\begin{aligned} w_{\text{non-MG}, m}(j) &= \Pr\{H_{\text{non-MG}, m}|\boldsymbol{s}(j)\} \\ &= \left\{1 - \sum_{n=0}^{m-1} w_{\text{non-MG}, n}(j)\right\}\text{fdr}_{\text{non-MG}, m}(j), 0 < m < K-2, \end{aligned} \tag{4b}$$

where $\mathrm{fdr}_{\text{non-MG},\,0}(j)$ is the local FDR associated with ANOVA on all subtypes, and $\mathrm{fdr}_{\text{non-MG},\,m}(j)$ is the local FDR associated with ANOVA on the top $(K-m)$ subtypes; estimated using R package "fdrtool" [22] (Supplementary Information).

Assessing statistical significance of candidate MGs. The p-values of candidate MGs are estimated using the learned 'mixture' null distribution

$$p\text{-}value = \Pr\{T > t_{obs} | H_{\text{non-MG}}\} = \sum_{m=0}^{K-2} \Pr\{T > t_{obs} | H_{\text{non-MG},\,m}\} P\{H_{\text{non-MG},\,m} | H_{\text{non-MG}}\}, \quad (5)$$

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

$$\Pr\{T > t_{obs} | H_{\text{non-MG},\,m}\} = \frac{\sum_{p=1}^{P} \sum_{j=1}^{J} w_{\text{non-MG},\,m}(j) I(T_{j,p} > t_{obs})}{P \sum_{j=1}^{J} w_{\text{non-MG},\,m}(j)}, \quad (6)$$

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 OVESEG-test statistics in the pth permution 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

$$P\{H_{\text{non-MG},\,m} | H_{\text{non-MG}}\} = \frac{\sum_{j=1}^{J} w_{\text{non-MG},\,m}(j)}{\sum_{j=1}^{J} \sum_{n=0}^{K-2} w_{\text{non-MG},\,n}(j)}. \quad (7)$$

Lastly, substituting (6) and (7) into (5), the p-value associated with gene j is calculated by:

$$p\text{-}value = \frac{\sum_{m=0}^{K-2} \sum_{p=1}^{P} \sum_{j=1}^{J} w_{\text{non-MG},\,m}(j) I(T_{j,p} > t_{obs})}{P \sum_{m=0}^{K-2} \sum_{j=1}^{J} w_{\text{non-MG},\,m}(j)}, \quad (8)$$

with a lower bound of $\min_{j}\{\sum_{m=0}^{K-2} w_{\text{non-MG},\,m}(j)\} / P \sum_{m=0}^{K-2} \sum_{j=1}^{J} w_{\text{non-MG},\,m}(j)$ (Supplementary Information).

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 OVESEG-test statistics t_j and determining local false discovery rate $\mathrm{fdr}_{\text{non-MG},\,m}(j)$, 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 [29]. We then adopt the empirical Bayes moderated variance estimator $\tilde{\sigma}^2(j)$ that leverages information across all genes, used in *limma* and given by

$$\tilde{\sigma}^2(j) = \frac{\nu_0\hat{\sigma}_0^2 + (N-K)\hat{\sigma}^2(j)}{\nu_0 + N - K}, \tag{9}$$

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

$$\hat{\sigma}^2(j) = \frac{\sum_{k=1}^{K}\sum_{i=1}^{N_k}\left(\log s_k(i,j) - \mu_k(j)\right)^2}{N-K}. \tag{10}$$

The prior parameters ν_0 and σ_0^2 are estimated from the pooled variances. Then 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 on the most relevant peer MG selection methods. The OVR-FC uses a simple test defined by

$$\text{OVR-FC}_k(j) = \frac{\bar{s}_k(j)}{\bar{s}_{\text{-}k}(j)}, \tag{11}$$

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

$$\text{OVR t} - \text{stat}_k(j) = \frac{\hat{\mu}_k(j) - \hat{\mu}_{-k}(j)}{\sqrt{\frac{\hat{\sigma}_k(j)}{N_k} + \frac{\hat{\sigma}_{-k}(j)}{N - N_k}}}, \tag{12}$$

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 OVE-FC is defined as

$$\text{OVE-FC}_k(j) = \frac{\bar{s}_k(j)}{\max\limits_{l \neq k} \bar{s}_l(j)}, \tag{13}$$

with various variations [16, 17]. Additional peer methods include independent pairwise OVE t-test and dependent pairwise OVE t-test [21] (Supplementary Information).

Simulation study for validating OVESEG-test statistics on type 1 error. Among the 10,000 genes, a significant portion are the housekeeping genes that take the baseline expression levels across all subtypes under $H_{\text{non-MG, }0}$. The expression levels of the remaining genes are adjusted to exhibiting similar upregulations in at least two subtypes, mimicking all type 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 S5**). With proper preprocessing of raw measurements, 12,022 probesets in Roche and 11,339 probesets in HUG are retained 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) are included in synthesizing 50 in silico mixtures, where purified subtype mean expressions (GSE28491 HUG) are combined according to pre-determined proportions with additive noise, simulating heterogeneous biological samples (Supplementary Information).

ADDITIONAL INFORMATION

Supplementary information accompanies this paper at …

ACKNOWLEDGMENTS

This work was funded in part by the National Institutes of Health under Grants HL111362-05A1, HL133932, W81XWH-18-1-0723 (BC171885P1), and CA184902-01.

AUTHOR CONTRIBUTIONS

L.C. and Y.W. developed OVESEG-test framework and wrote manuscript; L.C. implemented OVESEG-test algorithm and performed real data analysis; D.M.H. 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.

1. Wang, Y., Lu, J., Lee, R., Gu, Z. & Clarke, R. Iterative normalization of cDNA microarray data. *IEEE transactions on information technology in biomedicine : a publication of the IEEE Engineering in Medicine and Biology Society* **6**, 29-37 (2002).
2. Shen-Orr, S.S. et al. Cell type-specific gene expression differences in complex tissues. *Nat Methods* **7**, 287-289 (2010).
3. Montano, C.M. et al. Measuring cell-type specific differential methylation in human brain tissue. *Genome Biol* **14**, R94 (2013).
4. 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 (2009).
5. 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 (2012).
6. Newman, A.M. et al. Robust enumeration of cell subsets from tissue expression profiles. *Nat Methods* **12**, 453-457 (2015).
7. 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 (2011).
8. 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 Comput Biol* **9**, e1003189 (2013).
9. Wang, N. et al. Mathematical modelling of transcriptional heterogeneity identifies novel markers and subpopulations in complex tissues. *Scientific Reports* **6**, 18909 (2016).
10. Avila Cobos, F., Vandesompele, J., Mestdagh, P. & De Preter, K. Computational deconvolution of transcriptomics data from mixed cell populations. *Bioinformatics* **34**, 1969-1979 (2018).
11. 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 (2015).
12. 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 (2008).
13. Zhang, Y. et al. An RNA-sequencing transcriptome and splicing database of glia, neurons, and vascular cells of the cerebral cortex. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **34**, 11929-11947 (2014).
14. Shoemaker, J.E. et al. CTen: a web-based platform for identifying enriched cell types from heterogeneous microarray data. *BMC Genomics* **13**, 460 (2012).
15. Chen, Z. et al. Inference of immune cell composition on the expression profiles of mouse tissue. *Sci Rep* **7**, 40508 (2017).

16. Yu, G. et al. PUGSVM: a caBIG analytical tool for multiclass gene selection and predictive classification. *Bioinformatics* **27**, 736-738 (2011).
17. Yu, G. et al. Matched Gene Selection and Committee Classifier for Molecular Classification of Heterogeneous Diseases. *J. Mach. Learn. Res.* **11**, 2141-2167 (2010).
18. Herrington, D.M. et al. Proteomic Architecture of Human Coronary and Aortic Atherosclerosis. *Circulation* **137**, 2741-2756 (2018).
19. 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 (2015).
20. Hsu, J.C. Multiple comparisons : theory and methods. (Chapman & Hall, London :; 1996).
21. Wang, M., Master, S.R. & Chodosh, L.A. Computational expression deconvolution in a complex mammalian organ. *BMC Bioinformatics* **7**, 328-328 (2006).
22. Strimmer, K. fdrtool: a versatile R package for estimating local and tail area-based false discovery rates. *Bioinformatics* **24**, 1461-1462 (2008).
23. 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 (2012).
24. Kuhn, A. et al. Cell population-specific expression analysis of human cerebellum. *BMC Genomics* **13**, 610 (2012).
25. Wang, N. et al. Mathematical modelling of transcriptional heterogeneity identifies novel markers and subpopulations in complex tissues. **6**, 18909 (2016).
26. Becht, E. et al. Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. *Genome Biology* **17**, 218 (2016).
27. Aran, D., Hu, Z. & Butte, A.J. xCell: digitally portraying the tissue cellular heterogeneity landscape. *Genome Biology* **18**, 220 (2017).
28. 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 (2005).
29. Smyth, G.K. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology* **3**, Article3 (2004).

FIGURE LEGENDS

Figure 1. Type I error rates and p-value distributions in null data sets with unbalanced sample size ($N_1 = 3, N_2 = 6, N_3 = 9$). (a) Bar chart for the mean and 95% confidence interval of type I error rates with p-value cutoff 0.05 in 30 simulated experiments; (b) Histograms of p-values from five test methods for simulation with 60% housekeeping genes, $\sigma_0 = 0.5$ and $d_0 = 40$.

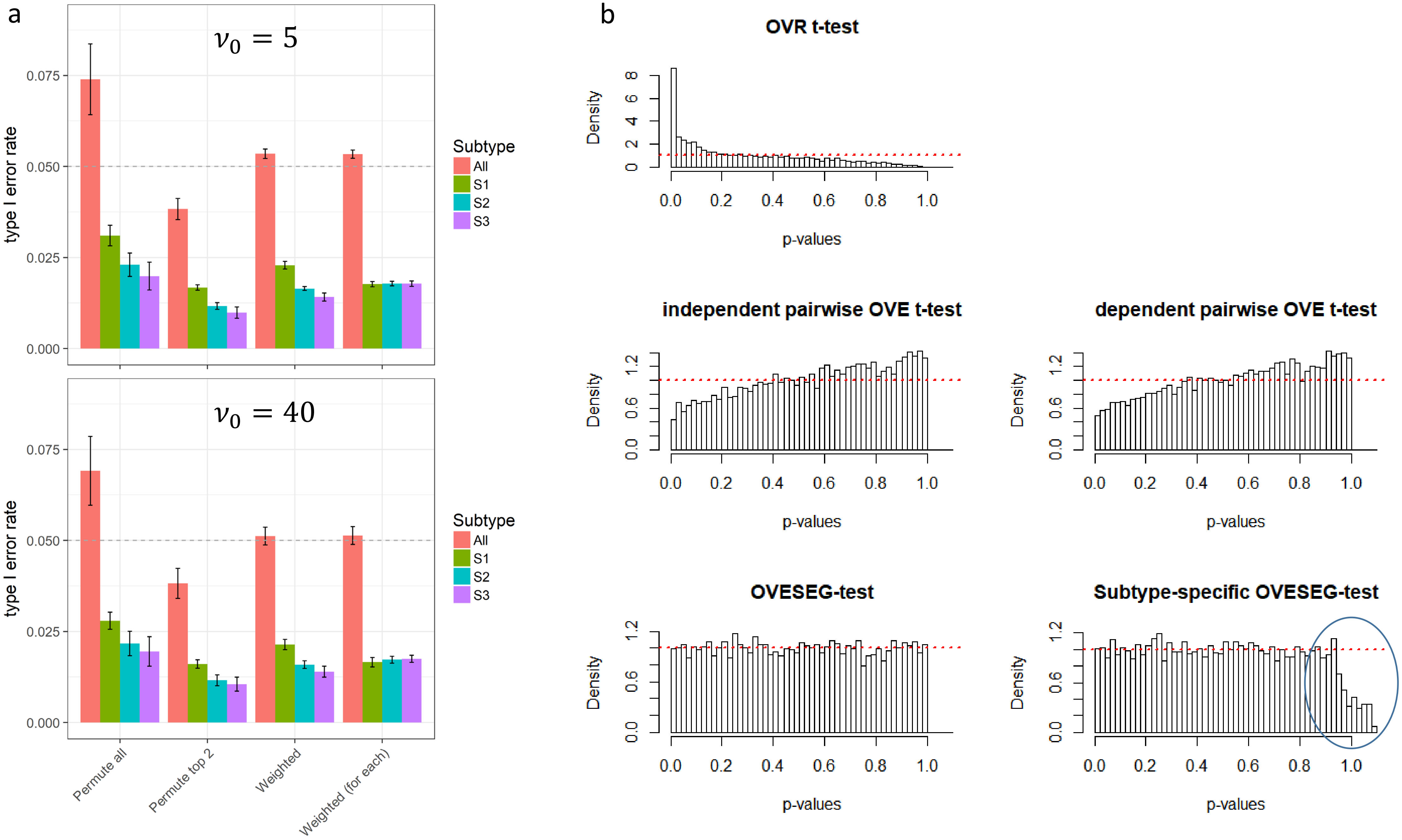
Figure 2. Bar chart of the mean and 95% confidence interval of type I error rates for 5 subtypes with unbalanced sample sizes (a) or unbalanced null hypothesis composition (b) in 30 simulated experiments.

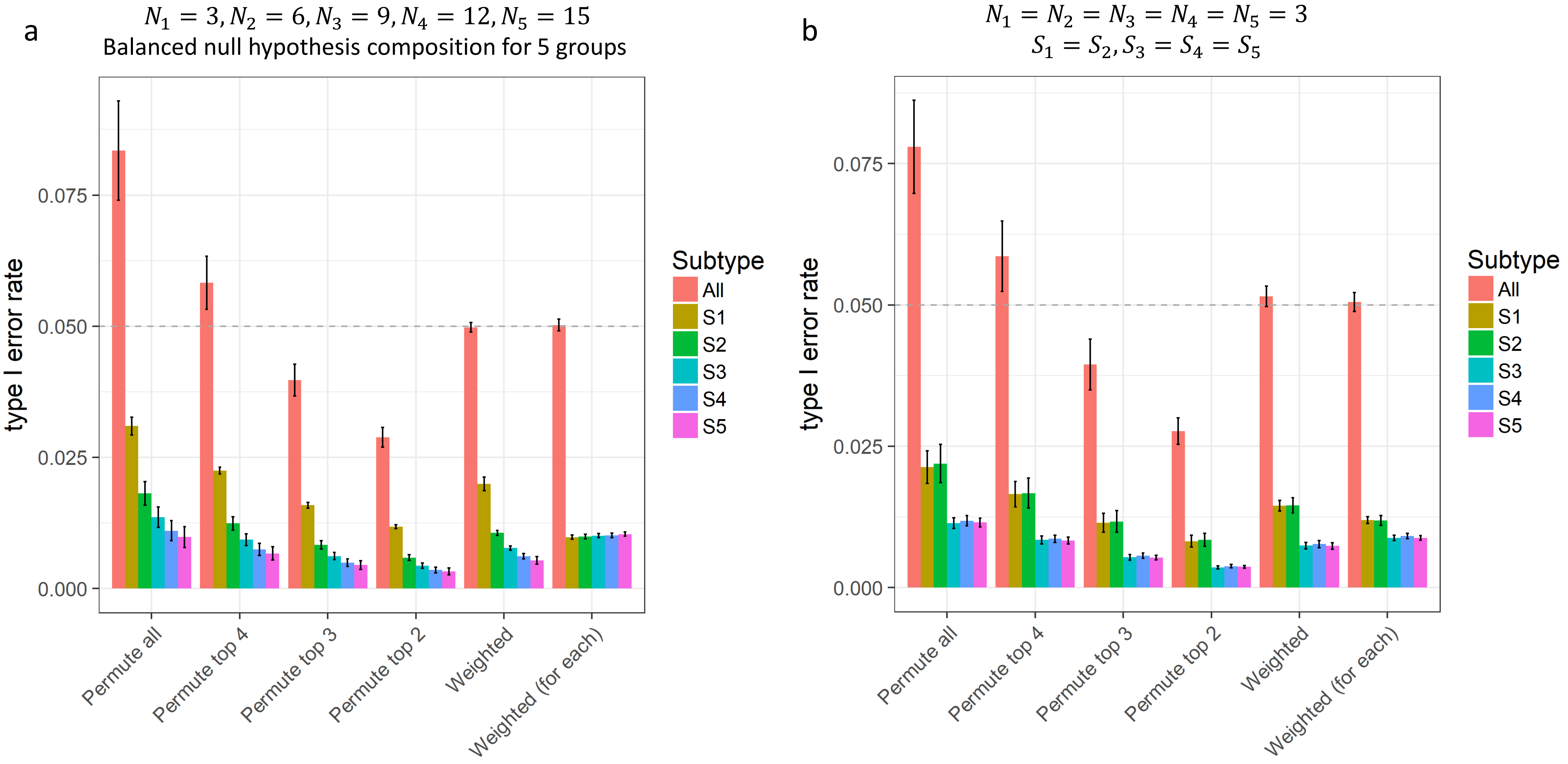
Figure 3. FDR control and pAUC performance under the multiple simulation settings with three unbalanced subtypes. (a) True FDR at q-value =0.05 across all subtypes or in each subtype (dash line is at 0.05/3). (b) pAUC when FPR<0.05 across all groups. (Dependent pairwise OVE t-test has the same pAUC as OVESEG-test.)

Figure 4. Overlap of cell-type specific markers between Roche and HUG datasets, quantified by Jaccard index (intersection over union).

Figure 5. Correlation coefficients between CAM score and ground truth proportion, with score estimated by a fixed number of markers from independent dataset to quantify subpopulations in heterogeneous samples simulated by mixing purified mRNA expression levels (a) or by mixing purified RNAseq counts (b). Mean and 95% confidence interval are computed among 20 repeated experiments.

Figure 6. Null distribution under $H_{\text{non-MG}}$ as the mixture model of $H_{\text{non-MG},\, m}, m = 0,1,2,3$. ($K = 5,\ N_1 = N_2 = N_3 = N_4 = N_5 = 6$)


a
$\nu_0 = 5$
$\nu_0 = 40$
type I error rate
0.000
0.025
0.050
0.075
Subtype
All
S1
S2
S3
Permute all
Permute top 2
Weighted
Weighted (for each)
b
OVR t-test
independent pairwise OVE t-test
dependent pairwise OVE t-test
OVESEG-test
Subtype-specific OVESEG-test
Density
p-values
0
2
4
6
8
0.0
0.6
1.2
0.0
0.2
0.4
0.6
0.8
1.0



a

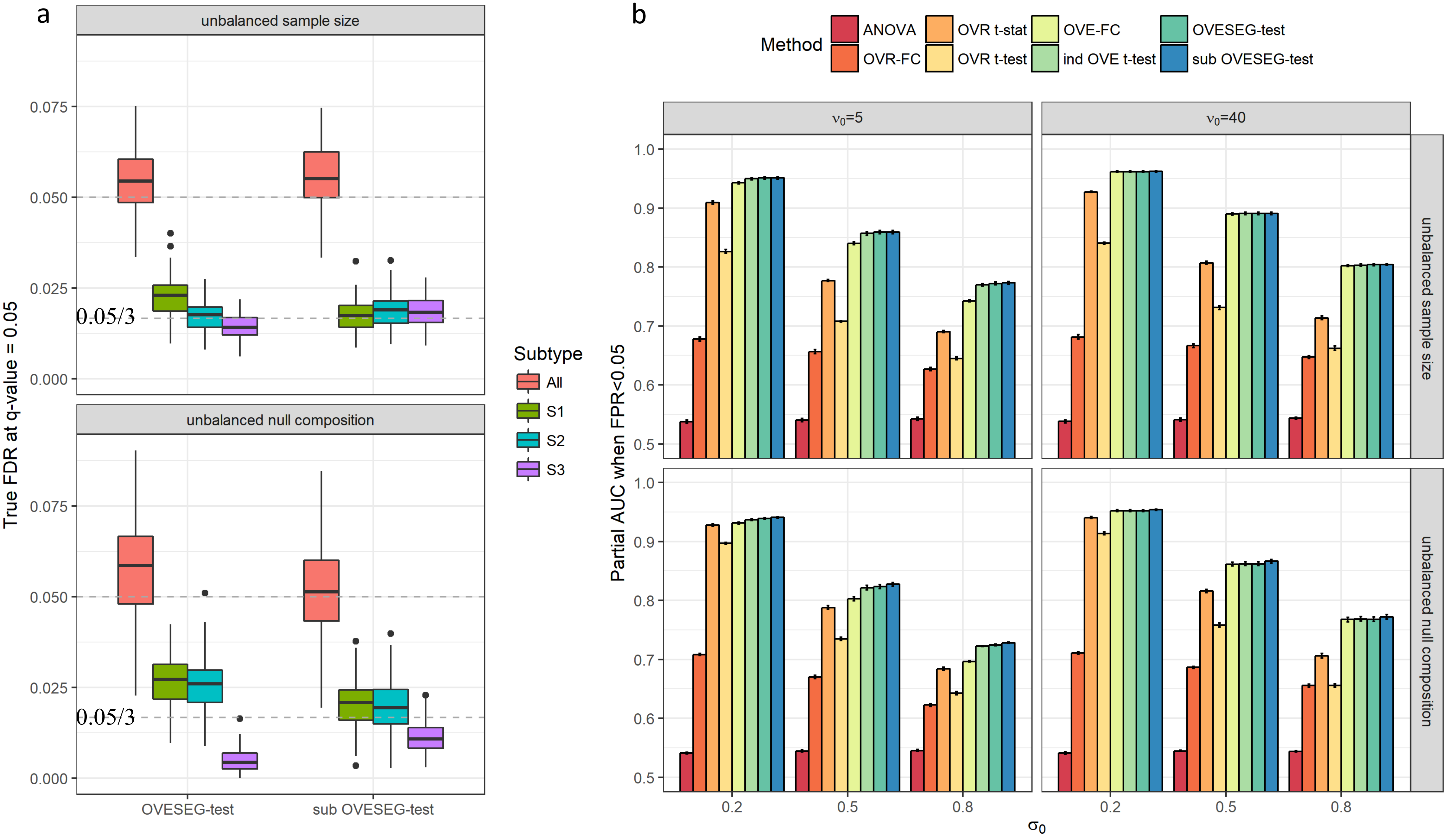
$N_1 = 3, N_2 = 6, N_3 = 9, N_4 = 12, N_5 = 15$

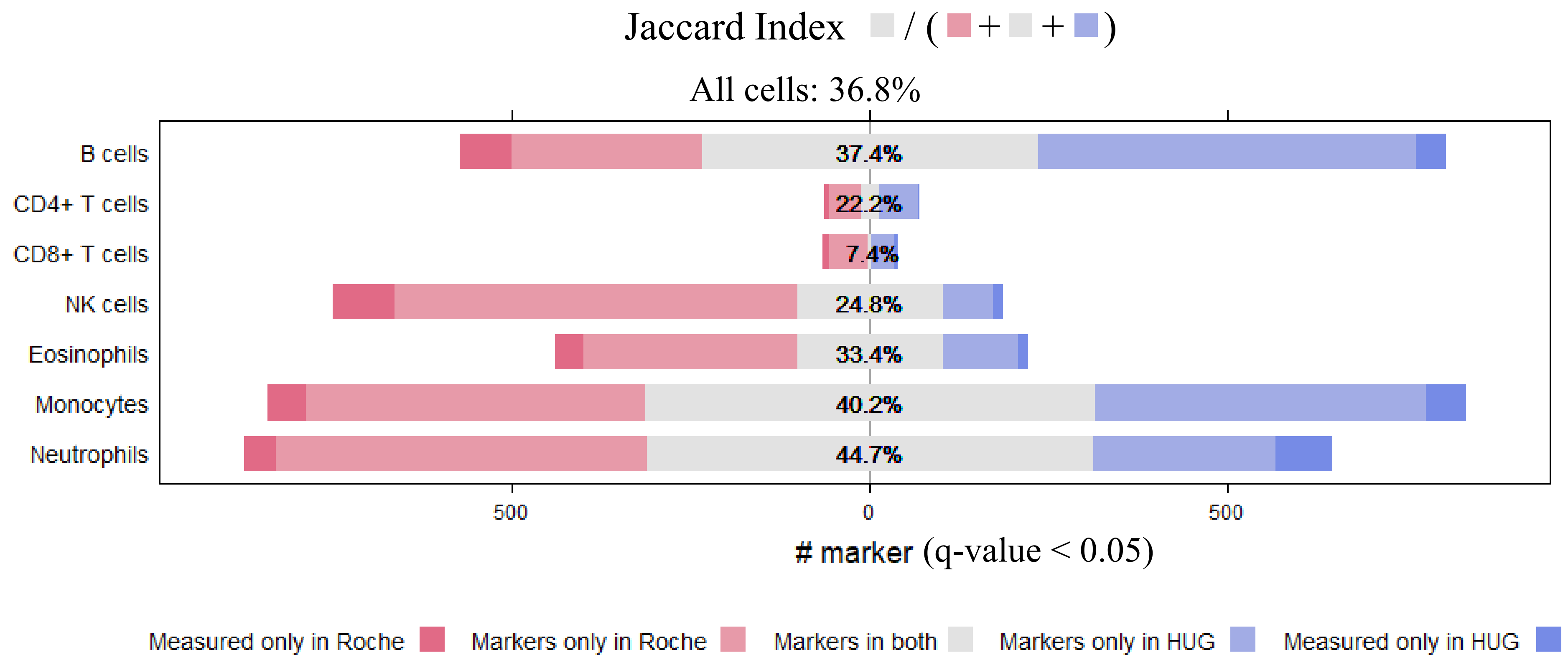
Balanced null hypothesis composition for 5 groups

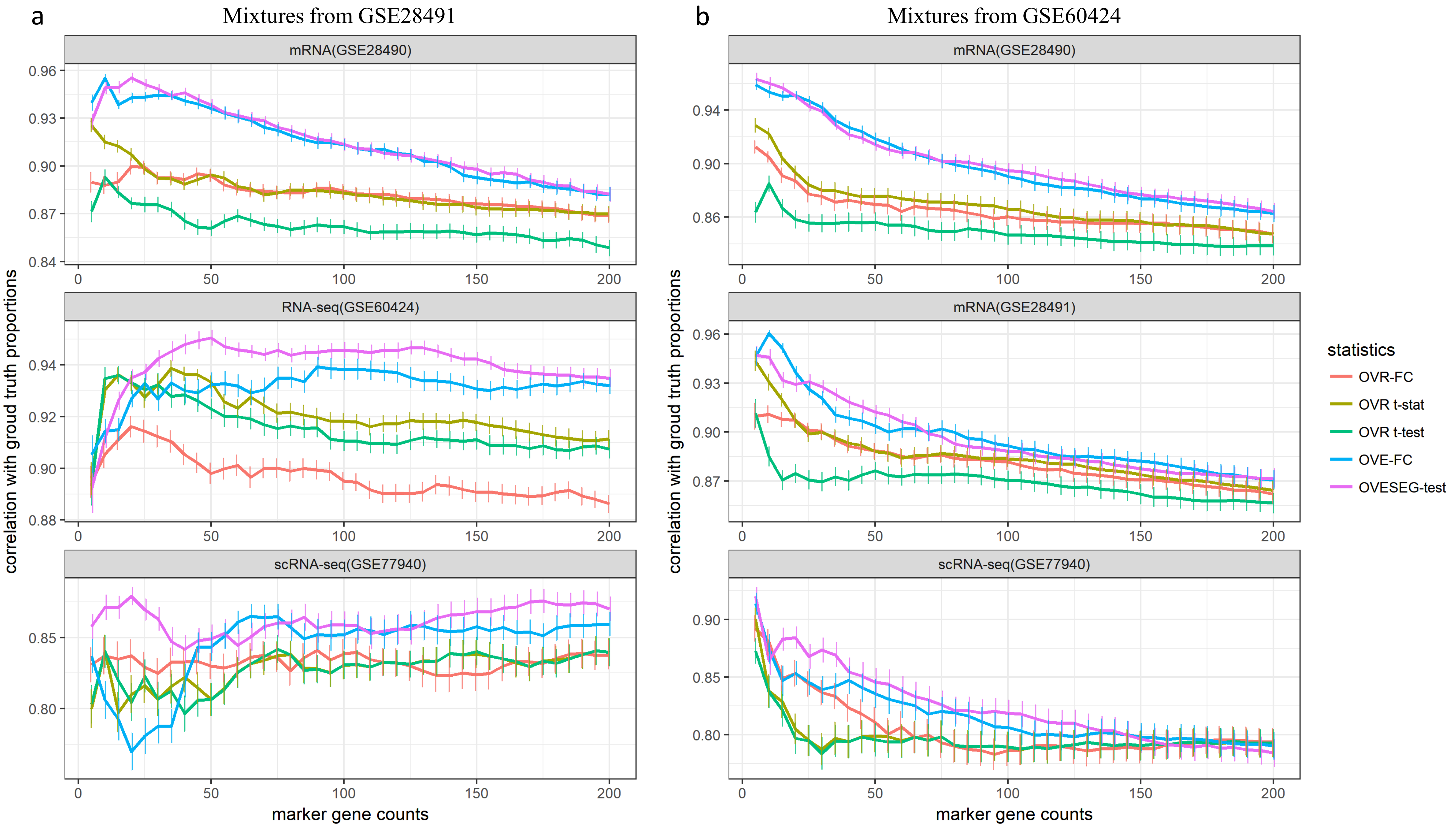
b

$N_1 = N_2 = N_3 = N_4 = N_5 = 3$

$S_1 = S_2, S_3 = S_4 = S_5$


a
unbalanced sample size
unbalanced null composition
True FDR at q-value = 0.05
0.075
0.050
0.025
0.000
0.05/3
OVESEG-test
sub OVESEG-test
Subtype
All
S1
S2
S3
b
Method
ANOVA
OVR-FC
OVR t-stat
OVR t-test
OVE-FC
ind OVE t-test
OVESEG-test
sub OVESEG-test
ν_0=5
ν_0=40
unbalanced sample size
unbalanced null composition
Partial AUC when FPR<0.05
1.0
0.9
0.8
0.7
0.6
0.5
0.2
0.5
0.8
σ_0

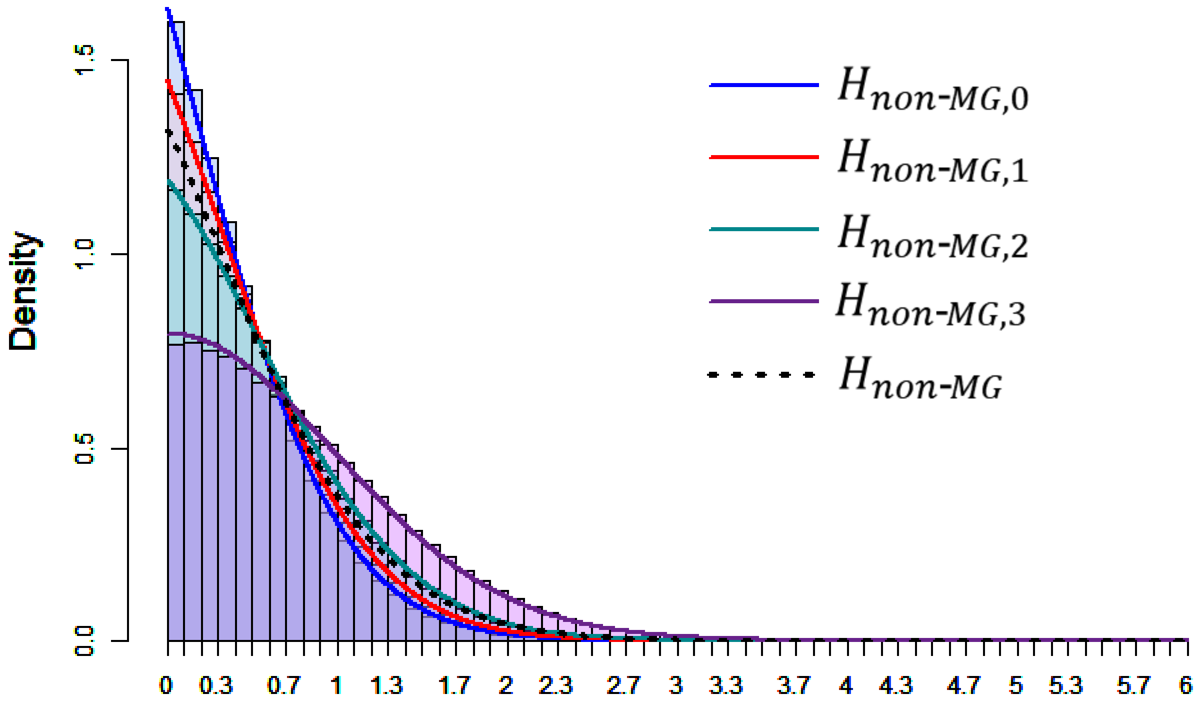

Jaccard Index ⬜ / (🟥 + ⬜ + 🟦)
All cells: 36.8%
B cells
CD4+ T cells
CD8+ T cells
NK cells
Eosinophils
Monocytes
Neutrophils
37.4%
22.2%
7.4%
24.8%
33.4%
40.2%
44.7%
500
0
500
# marker (q-value < 0.05)
Measured only in Roche
Markers only in Roche
Markers in both
Markers only in HUG
Measured only in HUG


a
Mixtures from GSE28491
b
Mixtures from GSE60424
mRNA(GSE28490)
0.96
0.93
0.90
0.87
0.84
0.94
0.90
0.86
0
50
100
150
200
RNA-seq(GSE60424)
0.94
0.92
0.90
0.88
mRNA(GSE28491)
0.96
0.93
0.90
0.87
scRNA-seq(GSE77940)
0.85
0.80
0.90
correlation with groud truth proportions
marker gene counts
statistics
OVR-FC
OVR t-stat
OVR t-test
OVE-FC
OVESEG-test

Null distribution under $H_{non\text{-}MG}$

as the mixture model of $H_{non\text{-}MG,\,m}, m = 0,1,2,3$

$(K = 5, N_1 = N_2 = N_3 = N_4 = N_5 = 6)$


Density
$H_{non\text{-}MG,0}$
$H_{non\text{-}MG,1}$
$H_{non\text{-}MG,2}$
$H_{non\text{-}MG,3}$
$H_{non\text{-}MG}$