



Published in final edited form as:

*Nat Methods*. 2016 March ; 13(3): 241–244. doi:10.1038/nmeth.3734.

## Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis

Jean Fan<sup>1</sup>, Neeraj Salathia<sup>2</sup>, Rui Liu<sup>3</sup>, Gwendolyn E. Kaeser<sup>4</sup>, Yun C. Yung<sup>4</sup>, Joseph L. Herman<sup>1</sup>, Fiona Kaper<sup>2</sup>, Jian-Bing Fan<sup>2,3</sup>, Kun Zhang<sup>4</sup>, Jerold Chun<sup>5</sup>, and Peter V. Kharchenko<sup>1,6</sup>

<sup>1</sup>Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA

<sup>2</sup>Illumina Inc., San Diego, CA, USA

<sup>3</sup>Present address: AnchorDx Corporation, International Biotech Island, Guangzhou, Guangdong, China

<sup>4</sup>Department of Bioengineering, University of California, San Diego, CA, USA

<sup>5</sup>Department of Molecular and Cellular Neuroscience, Dorris Neuroscience Center, The Scripps Research Institute, La Jolla, CA, USA

<sup>6</sup>Harvard Stem Cell Institute, Cambridge, MA, USA

### Abstract

The transcriptional state of a cell reflects a variety of biological factors, from persistent cell-type specific features to transient processes such as cell cycle. Depending on biological context, all such aspects of transcriptional heterogeneity may be of interest, but detecting them from noisy single-cell RNA-seq data remains challenging. We developed PAGODA to resolve multiple, potentially overlapping aspects of transcriptional heterogeneity by testing gene sets for coordinated variability amongst measured cells.

### Introduction

Single-cell transcriptome measurements provide an unbiased approach for studying the complex cellular compositions inherent to multicellular organisms. Increasingly sensitive single-cell RNA-sequencing (scRNA-seq) protocols<sup>1,2</sup> have been used to examine both healthy and diseased tissues<sup>3–14</sup>. Nevertheless, analysis of scRNA-seq data remains

---

Users may view, print, copy, and download text and data-mine the content in such documents, for the purposes of academic research, subject always to the full Conditions of use:[http://www.nature.com/authors/editorial\\_policies/license.html#terms](http://www.nature.com/authors/editorial_policies/license.html#terms)

Correspondence should be addressed to PVK (; Email: Peter\_Kharchenko@hms.harvard.edu)

**Author Contributions.** K.Z., J.C. and P.V.K. conceived the study. N.S., R.L., G.E.K., Y.C.Y., F.K. and J.-B.F. carried out the single-cell purification and RNA-seq measurements. G.E.K. and J.C. carried out RNAscope *in situ* validation. J.F. and P.V.K. designed and implemented the statistical analysis approach, with the help of J.L.H. P.V.K. and J.F. wrote the manuscript with the help of J.C. and K.Z.

**Competing Financial Interests Statement.** N.S. and F.K. are a current employees and shareholders of Illumina, Inc. The authors declare no competing financial interest.

challenging, as measurements expose numerous differences between cells, only some of which may be relevant for system-level functions.

High levels of technical noise<sup>15</sup> and strong dependency on expression magnitude pose difficulties for principal component analysis (PCA) and other dimensionality reduction approaches. Because of this, application of PCA as well as more flexible approaches such as GP-LVM<sup>16</sup> or tSNE<sup>17</sup> is often restricted to highly expressed genes<sup>11,12,18</sup>. Even when cell-to-cell variation captures prominent biological processes taking place within the measured cells, these processes may not be of primary interest. For example, differences in metabolic state or cell cycle phase may be common to multiple cell types, and can mask more subtle cell-to-cell variability associated with the biological processes being studied<sup>11</sup>. Such cross-cutting transcriptional features represent alternative ways to classify cells, posing a challenge for the commonly-used clustering approaches that aim to reconstruct a single subpopulation structure<sup>5,8,9,11</sup>. Partitioning methods, such as k-means clustering or the specialized BackSPIN algorithm<sup>9</sup> may, for example, choose to classify cells first based on the cell cycle phase instead of tissue-specific signaling state, if the cell cycle differences are more pronounced.

Here, we describe an alternative approach for analyzing transcriptional heterogeneity called PAGODA that aims to detect all statistically-significant ways in which measured cells can be classified. PAGODA is based on statistical evaluation of coordinated expression variability of previously-annotated pathways as well as automatically-detected gene sets. Gene set testing with methods such as GSEA<sup>19</sup> has been extensively utilized in the context of differential expression analysis to increase statistical power and uncover likely functional interpretations. A similar rationale can be applied in the context of heterogeneity analysis. For example, while cell-to-cell variability in expression of a single neuronal differentiation marker such as *Neurod1* may be too noisy and inconclusive, coordinated upregulation of many genes associated with neuronal differentiation in the same subset of cells would provide a prominent signature distinguishing a subpopulation of differentiating neurons. Examining previously published datasets, we illustrate that PAGODA recovers known subpopulations and reveals additional subsets of cells in addition to providing important insights about the relationships amongst the detected subsets.

The extent of transcriptional diversity in mouse NPCs is likely to be influenced by a variety of unexamined factors that include programmed cell death<sup>20</sup>, genomic mosaicism<sup>21-23</sup> as well as a variety of “environmental” influences such as changes in exposure to signaling lipids<sup>24-26</sup>. We therefore used scRNA-seq to assess a cohort of cortical NPCs from an embryonic mouse. We demonstrate that PAGODA effectively recovers the known neuroanatomical and functional organization of NPCs, identifying multiple aspects of transcriptional heterogeneity within the developing mouse cortex that are difficult to discern by the existing heterogeneity analysis approaches.

## Results

### Pathway and Gene Set Overdispersion Analysis (PAGODA)

To characterize significant aspects of transcriptional heterogeneity in a scRNA-seq dataset, PAGODA relies on a series of statistical and computational steps (Fig. 1). First, the measurement properties of each cell, such as effective sequencing depth, drop-out rate and amplification noise are estimated using a previously described mixture model approach<sup>27</sup> with minor enhancements (Step 1, Fig. 1). Using these models, the observed expression variance of each gene is renormalized based on the genome-wide variance expectation at the appropriate expression magnitude (Step 2). Batch correction is also performed at this stage. The resulting residual variance, modeled by the  $\chi^2$  statistic, effectively distinguishes subpopulation-specific genes (Supplementary Notes 1,2), and determines the contribution of each gene to the subsequent PCA calculations.

PAGODA then examines an extensive panel of gene sets to identify those showing a statistically significant excess of coordinated variability (Step 3). The gene sets include annotated pathways, such as Gene Ontology (GO) categories, as well as clusters of transcriptionally-correlated genes found within a given dataset (*de novo* gene sets). The later allows PAGODA to detect aspects of transcriptional heterogeneity driven by processes that are not represented in the pathway annotation. The prevalent transcriptional signature of each gene set is captured by its first principal component (PC), using weighted PCA to adjust for technical noise contributions. If the amount of variance explained by the first PC of a given gene set is significantly higher than expected (Step 4, correcting for multiple hypotheses), the gene set is said to be *overdispersed*, and is included in the subsequent analysis.

The PC of each overdispersed gene set separates cells along a certain axis (PC scores). Many PCs will show very similar patterns, either because the same genes drive them, or because multiple biological processes distinguish the same subsets of cells. To provide a non-redundant view of the transcriptional heterogeneity within the dataset, PCs from significantly overdispersed gene sets are clustered, and those with similar gene loadings or cell separation patterns are combined to form a single 'aspect' of heterogeneity (Step 5, Supplementary Fig. 1). The resulting major aspects of transcriptional heterogeneity can be explored numerically or through an interactive web browser interface<sup>28</sup> (Step 6). As we illustrate below, examination of individual aspects and their relationships to each other can provide insights and functional clues not apparent from the most prominent cell classification. Finally, if upon further interpretation one or more aspects of transcriptional heterogeneity are determined to be extraneous to the biological context, PAGODA provides an option to control explicitly for such aspects (Step 7).

### PAGODA captures alternative annotations of individual cells

To illustrate PAGODA on a complex cell population, we re-examined scRNA-seq data for 3,005 cells from the mouse cortex and hippocampus from a recent publication by Zeisel *et al.*<sup>9</sup>. This extensive dataset covers a variety of cell types, some of which exhibit very distinct expression signatures. Zeisel *et al.* also introduced a novel heterogeneity analysis method

called BackSPIN<sup>9</sup> that performs recursive partitioning. Applying PAGODA revealed nine major aspects of heterogeneity that distinguish the seven top-level classes and two lower-level subpopulations identified by BackSPIN (Fig. 2). The functional interpretation of the identified aspects is evident from the identity of overdispersed GO categories. The most significant aspect separates oligodendrocytes, the most numerous cell type in the dataset, which are easily distinguished by strong overdispersion of myelination-related pathways. Similarly, overdispersion of immune, vascular, and muscle-associated GO-annotated gene sets identify microglia, vascular endothelial, and mural subpopulations respectively. Other cell types, such as ependymal cells, or different types of neurons are distinguished by *de novo* gene set signatures, with most overdispersed genes revealing their identity (*e.g. Gad1, Tbr1, Gabra5*).

We noted that aspects distinguishing many of the cell types appear to overlap, most frequently with the myelination signature. For instance, a subset of 35 cells exhibits prominent expression of both immune response genes characteristic of microglia as well as genes responsible for myelin sheath (Fig. 2). Similarly, myelin-associated expression signature is observed for a subset of vascular cells, astrocytes, pyramidal neurons and interneurons. These hybrid signatures most likely correspond to cases in which two cells of different cell types were captured together (see Supplementary Fig. 2 for the analysis of cell type co-occurrence frequencies). The occurrence and functional interpretation of such ambiguous cases where a given cell exhibits multiple alternative signatures are apparent from PAGODA analysis. In contrast, BackSPIN, as well as other partitioning methods, would need to classify such cells based on one of the signatures or isolate them as a separate class without exposing their relationship to other groups.

We further evaluated PAGODA performance by re-analyzing datasets that were used to present alternative methods of heterogeneity analysis<sup>8,11,29</sup>, recovering previously identified subpopulations and identifying additional biologically-relevant features (Supplementary Note 3). In particular, PAGODA's ability to associate with a given cell multiple, potentially independent aspects of transcriptional heterogeneity, allows one to focus on biologically-relevant subpopulations that may be distinguished by relatively subtle transcriptional variation. For instance, in reanalyzing data for mouse CD4<sup>+</sup> T that was used to present an elegant GP-LVM approach by Buettner *et al*<sup>11</sup>, PAGODA successfully recovered *Il4ra-Il24* response and a closely aligned glycolysis aspect in addition to a prominent mitosis-associated signature, without requiring explicit correction steps. Furthermore, PAGODA revealed a prominent subpopulation of cells exhibiting an expression signature typical of dendritic cells that was not previously observed.

### PAGODA reveals multiple aspects of heterogeneity in mouse NPCs

As heterogeneity amongst NPCs may influence downstream neural diversity, we performed Smart-Seq<sup>30</sup> on 65 NPCs isolated from the cerebral cortex of 13.5-day embryonic mouse brain. The most significant aspect of heterogeneity identified by PAGODA within the isolated NPCs reflects gradual induction of the genes associated with neuronal maturation and growth (Fig. 3a, top aspect). Approximately half of the cells express *Dcx*, *Sox11*, and other known markers of neuronal maturation, with the most mature subset expressing genes

involved in neuronal maturation and growth cones (*Neurod6*, *Gap43*). Such cells maintain expression of some progenitor markers (e.g., vimentin) and therefore likely represent developing, committed neurons. In contrast, the set of early NPCs exhibits strong M- and S-phase signatures that are absent from the more mature NPCs, as well as up-regulation of genes characteristic of early progenitor state<sup>31</sup> (*Sox2*, *Notch2*, *Hes1*) captured by the “negative regulation of neuronal differentiation” and “neural tube development” GO categories.

Maturation of neuronal progenitors is closely tied to the spatial organization of the developing cortex<sup>32</sup>. We used spatial expression patterns<sup>33</sup> of genes differentially expressed between the early and maturing NPCs to reconstruct the most likely spatial distribution of these cells within the mouse brain (Fig. 3b, Online Methods). As expected, we found early NPCs localize close to ventricular zone (VZ). We also used *in situ* RNA-FISH (Online Methods) to examine two genes, *Rpa1* and *Ndn*, of unknown relationship to the embryonic cerebral cortex (Fig. 3c). Consistent with their predicted pattern, *Rpa1* was most prominent in proliferative regions. *Ndn* localized in the post-mitotic regions (especially the cortical plate), as well as rare cells within the subventricular zone (SVZ, Supplementary Fig. 3).

An additional subset of NPCs was distinguished by expression of *Eomes*, *Neurod1*, and other genes localized to the SVZ region and thought to distinguish basal progenitors<sup>31,34</sup>. The *Eomes* signature mark cells that express intermediate levels of genes associated with neuronal maturation as well as a subset of mature NPCs and subset of early NPCs undergoing DNA replication, likely representing neuronally-committed NPCs maturing in the SVZ, and dividing basal NPCs, respectively. These dividing cells express notch signaling genes (*Dll1*, *Notch2*, *Mfng*) concurrently with *Eomes* and therefore likely represent nascent basal progenitors<sup>31</sup>.

Two other aspects cut across the main NPC maturation axis. The first is driven by prominent expression of *Ndn* (Fig. 3a). *Ndn*, initially noted for high expression in mature neurons<sup>35</sup>, has also been shown to be expressed in the VZ<sup>36</sup>, and to restrict both proliferation and apoptosis rates in NPCs<sup>36-38</sup>. In combination with RNAscope analyses (Supplementary Fig. 3), we found *Ndn* to be expressed within a subset of NPCs, approximately a quarter of which exhibit pronounced mitotic signatures and are likely localized in the SVZ. The second cross-cutting aspect is coordinated expression of *Dlx* homeodomain transcription factors. *Dlx* genes mark tangentially-migrating NPCs, which originate in the ganglionic eminence (GE) and migrate to the cortical areas, giving rise to the GABAergic neurons<sup>39,40</sup>. The *Dlx*-positive cells express other markers of tangentially migrating NPCs, most notably Sp9 and Sp8 transcription factors<sup>41</sup>. Indeed, spatial localization of these cells was predicted to be in the GE region, where tangentially-migrating NPCs are expected to originate (Fig. 3b). In agreement with earlier observations of such NPCs undergoing mitosis in the cortical VZ/SVZ areas, two of ten *Dlx*-positive NPCs were captured in S-phase and one in M-phase.

To illustrate the methodological advantage of PAGODA, we re-examined our NPC data using alternative analysis methods, including PCA, ICA, tSNE<sup>12,17</sup>, GP-LVM<sup>16</sup>, and BackSPIN<sup>9</sup> (Supplementary Figs. 4,5). While none of the methods were able to recover all of the identified subpopulations, BackSPIN provided the most compelling results, capturing

heterogeneity involving expression of *Dlx* and *Prdx4/Mest*. However, the reported clustering grouped only some of the cells associated with each signature, illustrating limitations of partitioning-based interpretation in a complex biological context.

## Discussion

Just like organisms as a whole, individual cells can be classified according to a variety of meaningful criteria. For example, tangentially migrating NPCs that, despite being a distinct progenitor subtype, go through the same neuronal maturation process as other NPCs. By identifying significantly overdispersed gene sets, PAGODA is able to effectively recover such complex heterogeneity structures. The potential ambiguity of classification illustrated by the NPCs is likely to be present in many biological contexts. In such cases, an optimal partition or clustering of cells is unlikely to be fully informative, and the analysis can benefit from concurrent interpretation. The gene-set-based approach and interactive interface implemented by PAGODA aims to identify and facilitate interpretation of significant transcriptional features separating cells within the population.

## Methods

### Isolation and single-cell RNA-seq of mouse neural progenitor cells (NPC) and astrocytes (ASC)s

Single NPCs were isolated from C57BL/6J embryonic day 13.5 cortices for RNA-sequencing. Timed-pregnant mice were sacrificed by deep anesthesia followed by cervical dislocation. The embryos were quickly removed and cortical hemispheres were isolated, ganglionic eminences removed, and all pups brains were pooled. All animal protocols were approved by the Institutional Animal Care and Use Committee at The Scripps Research Institute (La Jolla, CA) and conform to the National Institutes of Health guidelines.

Single cells were isolated by gentle trituration in ice-cold phosphate buffered saline containing 2 mM EGTA (PBSE) using P1000 tips with decreasing bore diameter. Cells were then filtered through a 40  $\mu$ M nylon cell strainer and stained with propidium iodide (PI), a live-dead stain, and fluorescence activated single cell sorting (FACS) was performed selecting for PI negative cells. Samples remained on ice throughout the process and total processing time from cervical dislocation to sorting was limited to 2 hours. Single cells were sorted directly into cell lysis buffer provided in the Clontech SMARTer® Ultra™ Low RNA Kit for Illumina® Sequencing (cat # 634936), and sequencing libraries were generated using the manufacturer's protocol. Resulting libraries were sequenced on the Illumina® HiSeq™ 2000 sequencing platform.

### Gene validation using *in situ* hybridization with RNA-scope

Mouse E13.5 embryos were removed from timed pregnant mice and prepared according to RNAscope instructions for paraffin embedded tissue. RNAscope probes (Advanced Cell Diagnostics) were designed by the manufacturer (Cat. # : GINS2 435891, RPA1 435911) and sections were processed using RNAscope 2.0 High Definition Reagent Kit - BROWN (Cat. #:310035) according to the manufacturer's instructions. Sections were imaged on a Zeiss Axioimager at 20 $\times$  magnification.

## Previously published single-cell RNA-seq data

For the mixture of cultured human neuronal progenitor cells (NPCs) and primary cortical samples from Pollen *et al.*<sup>29</sup>, SRA files for each study were downloaded from the Sequence Read Archive (<http://www.ncbi.nlm.nih.gov/sra>) and converted to FASTQ format using the SRA toolkit (v2.3.5). FASTQ files were aligned to the human reference genome (hg19) using Tophat (v2.0.10) with Bowtie2 (v2.1.0) and Samtools (v0.1.19). Gene expression counts were quantified using HTSeq (v0.5.4). Read counts for the Th2 data by Buettner *et al.*<sup>11</sup> were downloaded from the supplementary site ([http://github.com/PMBio/scLVM/blob/master/data/Tcell/data\\_Tcells.Rdata](http://github.com/PMBio/scLVM/blob/master/data/Tcell/data_Tcells.Rdata)). Read (or UMI) count matrices for other two datasets were downloaded from GEO: GSE60361 for Zeisel *et al.*<sup>9</sup>; GSE59739 for Usoskin *et al.*<sup>8</sup>.

**Fitting single-cell error models**—Following the approach described in Kharchenko *et al.*<sup>27</sup>, the read count for a gene  $g$  in a cell  $i$  was modeled as a mixture of a negative binomial (signal) and Poisson (drop-out) components:

$c_g^i \sim p_i^d(e_g) \text{Poisson}(\lambda_{bg}) + (1 - p_i^d(e_g)) \text{NB}(\alpha_i e_g, \theta_i(e_g))$ , where  $p_i^d(e_g)$  is the probability of encountering a drop-out event in a cell  $i$  for a gene with population-wide expected expression magnitude  $e_g$  (FPKM);  $\lambda_{bg} = 0.1$  is the low-level signal rate for the dropped-out observations;  $\theta_i(e_g)$  is the negative binomial size parameter (see functional form below); and  $\alpha_i$  is the library size of cell  $i$ , as inferred by the fitting procedure. The single-cell error models were fitted using the approach described in Kharchenko *et al.*<sup>27</sup>, with the following modifications. **1.** Rather than estimating expected expression magnitudes of genes using all pairwise comparisons between all other cells, each cell was compared to its  $k$  most similar cells (based on Pearson linear correlation of genes detected in both cells for any pair of cells). The value of  $k$  was chosen to approximate the complexity of the dataset (1/3<sup>rd</sup> of the cells for mouse and human NPC datasets, 1/5<sup>th</sup> for the larger Zeisel *et al.*<sup>9</sup> and Usoskin *et al.*<sup>8</sup> datasets). **2.** The count dependency on the expected expression magnitude was estimated on the linear scale with zero intercept. **3.** To improve fit, the drop-out probability was modeled using logistic regression on both expression magnitude (log scale) and its square value. **4.** Instead of fitting a constant value for the negative binomial size parameter  $\theta$ , it was fit as a function of expression magnitude, using the following functional form:  $\log(\theta) = a + h/(1+10^{(x-m)*s})^r$ , where  $x$  is the expression magnitude (log scale), and  $a, h, m, s, r$  are parameters of the fit. This functional form provides a more flexible fit than the  $\theta = (a_0 + a_1/x)^{-1}$  form used in DESeq<sup>42</sup>, while allowing for stable asymptotic behavior.

## Evaluating overdispersion of individual genes

For each gene, the approach estimates the ratio of observed to expected expression variance and the statistical significance of the observed deviation from the expected value. To illustrate the rationale, we start with a Poisson approximation. Let  $c_g^i$  be the number of reads observed for a gene  $g$  in a cell  $i$ . If such reads follow a Poisson distribution with the mean  $\mu_g$  and variance  $\nu_g$  (both equal to some Poisson rate  $\lambda_g$ ), then Fisher's index of dispersion

$D_g = \sum_{i=1}^k (c_g^i - \mu_g)^2 / \nu_g$  follows  $\chi_{k-1}^2$  distribution<sup>43</sup>. While for the Poisson case  $\nu_g = \mu_g$ , for negative binomial process,  $\nu_g = \mu_g + (\mu_g)^2/\theta$ , where  $\theta$  is the size parameter. As  $\theta$  decreases from very high values where the negative binomial is well approximated by a Poisson,  $D_g$

diverges from  $\chi^2_{k-1}$ . Analytical adjustments of  $D_g$  based on the negative binomial moments can improve  $\chi^2$  approximation<sup>44</sup>. For more accurate approximation we used a numeric correction of the  $\chi^2$  degrees of freedom, depending on the magnitude of  $\theta$ , so that  $D_g \sim \chi^2_{f(\theta)}$  (Supplementary Note 2, Figure SN2.2).

To account for the possibility of drop-out events, weighted sample variance estimates were

used, so that:  $D_g = \sum_{\text{cell } i} \left[ w_g^i (c_g^i - \mu_g^i)^2 \right] / \left[ \mu_g^i + (\mu_g^i)^2 / \theta_i(e_g) \right] \sim \chi^2_{k_g}$ , where  $w_g^i$  is the probability that the measurement in a cell  $i$  was not a drop-out event based on the error

model for cell  $i$ , and  $k_g = \sum_{i=1}^k w_g^i f(\theta_i(e_g))$  is the effective degrees of freedom for the gene  $g$ .  $\mu_g^i = e_g \alpha_i$ , where  $e_g$  is the expected expression magnitude of a gene  $g$  across the measured cells.

Since negative binomial (or NB/Poisson mixture) models do not fully capture the variability trends observed in the real scRNA-seq measurements,  $D_g$  estimates for the real data can systematically deviate from 1. To adjust for this non-centrality, we normalized  $D_g$  by its transcriptome-wide expectation value  $D_g^e$ , where  $D_g^e$  models the transcriptome-wide dependency of  $D_g$  on gene expression magnitude.  $D_g^e$  estimates were obtained using a general additive model (GAM, fit using the *mgcv* R package) as a smooth function of gene expression magnitude  $e_g$ . To improve smoothness, the GAM fit was performed on the corresponding squared coefficient of residual variance  $(D_g/E_g)^2$ . The fit is performed on all of the genes. The P value of overdispersion for a gene  $g$  was then be calculated as

$P_g^{od} = F_{\chi^2_{k_g}}(k_g D_g / D_g^e)$ , where  $F_{\chi^2_k}$  is CDF of  $\chi^2$  distribution with  $k$  degrees of freedom.

To improve stability of the estimates with respect to outliers, a Winsorization procedure<sup>45</sup> was applied to the read count matrix prior to the variance evaluation described above. To ensure that the outliers are trimmed in a manner independent of the total cell coverage, the Winsorization procedure was applied to the FPM matrix (i.e. normalizing counts by the library size), that were then translated back into the integer counts. A trim value of 3 was used for all datasets (i.e. observations from the three highest and tree lowest cells for each gene were Winsorized).

**Weighted PCA and significance of pathway overdispersion**—For PCA the data was transformed to better approximate the standard normal distribution. Specifically, PCA was carried out on a matrix of log-transformed read counts with a pseudocount of 1, normalized by the library size:  $x_g^i = \log(c_g^i / \alpha_i + 1)$ . The values for each gene (matrix row) were then scaled so that the weighted variance of a given gene matched the tail probabilities of the distribution for a standard normal process:  $y_g^i = x_g^i \sqrt{Q_N(P_g^{od}) / \text{var}_{w_g}(x_g)}$ , where  $Q_N$  is the quantile function of the standard normal distribution, and  $\text{var}_{w_g}(x_g)$  is the weighted variance of values  $x_g$ . As in our previous work<sup>27</sup>, the weight used for the clustering and PCA steps included an additional damping coefficient  $k = 0.9$ :  $w_g^i = 1 - k^* p_i^d(e_g) p^{bg}(c_g^i)$ , which



improved the stability of the subsequent cell clustering for noisy datasets ( $p^{bg}(c_g^i)$  is a probability of observing  $c_g^i$  counts in a drop-out event, evaluated from the Poisson PDF).

Weighted PCA was performed for each gene set as described by S. Bailey<sup>46</sup>, recording first (and optionally subsequent) principal components, the magnitude of the eigenvalue ( $\lambda_1$ ) and associated cell scores for each gene set. Statistical significance of the  $\lambda_1$  eigenvalues obtained for each gene set (overdispersion P value for a set  $s$ ,  $P_s^{od}$ ) was evaluated based on the Tracy-Widom  $F_1$  distribution<sup>47</sup>  $F_1(m, n_e)$ , where  $m$  is the number of genes in a given set  $s$ , and  $n_e$  is the effective number of cells, determined to fit the distribution of the randomly sampled gene sets (containing the same number of genes as the actual gene sets). The presented results used pathways annotated by Gene Ontology (GO), restricting evaluation to the GO terms that had between 1000 and 10 annotated genes.

**Identification and statistical treatment of *de novo* gene clusters**—Since some aspects of transcriptional heterogeneity can be driven by genes that are poorly represented or not at all described by the annotated pathways, PAGODA incorporates into the overall analysis *de novo* gene sets that group genes showing correlated patterns of expression across the cells measured in a particular dataset. By default, PAGODA, implements a straightforward clustering procedure: a hierarchical clustering is performed using Ward method (as implemented by the `hclust` package in R) using a Pearson correlation distance on the normalized expression matrix (that is used for the weighted PCA step described above). The resulting dendrogram is cut to obtain a pre-defined number of *de novo* gene clusters (the results shown use 150 clusters). As there are many alternative methods for clustering co-expressed genes, PAGODA implementation provides parameters to use alternative clustering procedures.

Since *de novo* gene clusters are by purposefully selected to contain genes with correlated expression profiles, the amount of variance explained by the first principal component (magnitude of  $\lambda_1$ ) will be higher than expected from random matrices, and cannot be modeled by the same Trace-Window  $F_1$  distribution as previously-annotated gene set. To evaluate statistical significance of overdispersion, a background distribution of  $\lambda_1$  was generated by performing the same hierarchical clustering and weighted PCA procedure on randomized matrices (where cell order was randomized for each gene independently, 100 randomizations). The  $\lambda_1$  values were normalized relative to Tracy-Widom  $F_1$  expectation as  $\lambda_1^s = [\lambda_1 - (a\lambda_1^{TW} + bn)] / \sqrt{\nu_1^{TW}}$ , where  $\lambda_1^{TW}$  and  $\nu_1^{TW}$  are the mean and variance of  $\lambda_1$  predicted by the Tracy-Widom  $F_1$  distribution, and coefficients  $a$  and  $b$  are determined by the linear model  $\lambda_1 \sim \lambda_1^{TW} + n$ . This standardized residual  $\lambda_1^s$  was modeled using Gumbel extreme value distribution, the parameters of which were fit using `extRemes` package in R. The overdispersion P value for each *de novo* gene set were determined from the tails of that distribution. The subsequent procedures treated *de novo* gene sets and annotated gene sets in the same way.

**Clustering of redundant heterogeneity patterns**—To compile a non-redundant set of aspects, the PC cell scores (projections on the eigenvector) from each significantly

overdispersed (5% FDR, as estimated by the Benjamini-Hochberg method<sup>48</sup>) gene set were normalized so that the magnitude of their variance corresponds to the tail probability of the  $\chi^2$  distribution:  $\text{var}(s_i) = Q_{\chi_{n-1}^2}(P_i^{od}) / (n-1)$ , where  $Q_{\chi_n^2}$  is the quantile function of the  $\chi^2$  distribution with  $n$  degrees of freedom ( $n$  is the number of cells in the dataset). The redundant aspects of heterogeneity were reduced in two steps. First, aspects reflecting transcriptional variation of the same genes were grouped by evaluating similarity of the corresponding gene loading scores in combination with the pattern similarity using the

following distance measure between gene sets  $i$  and  $j$ :  $d_{ij} = \left(1 - \sqrt{|\text{cor}(l_i, l_j) * \text{cor}(s_i, s_j)|}\right)$ , where  $\text{cor}$  is Pearson linear correlation,  $l_i, l_j$  are the loading scores of genes found in both  $i$  and  $j$  sets, and  $s_i, s_j$  are the corresponding PC cell scores ( $d_{ij}$  was set to 1 if there were less than 2 genes in common between the gene sets  $i$  and  $j$ ). The distance  $d_{ij}$  was then used to cluster the aspects, using hierarchical clustering with complete-linkage. Clusters separated by a distance less than 0.1 were grouped. The cell scores of the grouped aspects were determined as cell scores of the first principal component of all aspects within a grouped cluster. The second step, aimed at grouping aspects showing similar patterns of cell separation, was accomplished by another round of hierarchical clustering using  $\text{cor}(s_i, s_j)$  distance measure with Ward clustering procedure. The similarity threshold for the final grouping of similar aspects varied between datasets depending on their complexity (0.5 for the human NPC data, 0.95 for the mouse cortical/hippocampal dataset, 0.9 for the T cell and the mouse NPC data).

**Batch correction**—To control for the effect of categorical covariates, such as presence of multiple batches in the data, the approach contrasted whole-population and batch-specific variance estimates. Specifically, for each gene  $g$ , a batch-specific average expression magnitude was estimated for each batch  $b$ :  $e_{g,b}$ . These batch-specific expression estimates were then used to obtain batch-adjusted values of  $D_g, w_g^i$  and  $k_g$  ( $D_{g,b}, w_{g,b}^i$  and  $k_{g,b}$  respectively). To identify genes showing batch-specific variation, the ratio of batch-specific and batch-adjusted variance was evaluated as  $\alpha_g = D_{g,b} / D_g$ . The residual variance of genes showing discrepant batch- and population-specific variance was taken to be

$$D_g^b = \min(\alpha_g, 1/\alpha_g) * D_{g,b} / D_g^e \text{ and } P_g^{od} = F_{\chi_{k_g}^2}(k_g D_g^b / D_g^e).$$

The procedure above ensures that batch-specific effects are not reflected in the magnitude of the adjusted variance. Batch effects also need to be controlled at the level of expression values on which weighted PCA is performed, as batch-specific expression patterns across a sufficiently large set of genes can still account for sufficiently high amount of total variance to be picked by the PCA analysis. The expression values,  $x_g^i = \log(c_g^i / \alpha_i + 1)$ , were adjusted in two steps, separating drop-out (0 read count) observations from the rest. To adjust for the disparity in the frequency of the drop-out observations between batches, the lower bound of the zero-count observation fraction ( $u$ ) was determined for each batch (assuming binomial process), and the weights  $w_g^i$  for each batch were multiplied by  $\min(1, \max(u) / Z_b)$ , where  $\max(u)$  is the maximum lower bound value amongst batches, and  $Z_b$  is the fraction of zero-count observations in a given batch. This procedure ensures that the expected number of

zero-count observations is equal amongst all of the batches. The second step adjusted the log expression magnitudes of non-zero observations so that the weighted means within each are each equal to the population-wide weighted mean. To further control for batch-specific effects, weighted PCA was performed using batch-specific centering (*i.e.* setting weighted mean of each batch to 0).

**Spatial placement of cell subpopulations**—To spatially place neuronal subpopulations identified by PAGODA, we used significantly differentially expressed genes (absolute corrected Z-score > 1.96) as relative gene expression signatures for each subpopulation of interest compared to all other NPCs. In situ hybridization (ISH) data for the developing 13.5 day embryonic mouse were downloaded from the Allen Developing Mouse Brain Atlas (Website: ©2013 Allen Institute for Brain Science. Allen Developing Mouse Brain Atlas: <http://developingmouse.brain-map.org>) for all available genes (n=2,194). ISH data are quantified as gene expression *energies*, defined as expression intensity times expression density, at a grid voxel level. Each voxel corresponds to a 100  $\mu\text{m}$  gridding of the original ISH stain images and corresponds to voxel level structure annotations according to the accompanying developmental reference atlas ontology. The 3-D reference model for the developing 13.5 day embryonic mouse derived from Feulgen-HP yellow DNA staining was also downloaded from the Allen Developing Mouse Brain Atlas for use as a higher resolution reference image. Energies for genes in each subpopulation's gene expression signature with corresponding ISH data available were weighted by expression fold change on a  $\log_2$  scale and summed to constitute a composite overlay of gene expression. Background signal and expression detection in regions not annotated as part of the mouse embryo in the reference model were removed by applying a minimum gene energy level threshold of 8 units. We focused on spatial placements within the developing mouse forebrain and thus restricted gene energies to voxels annotated as 'forebrain' or 'ventricles, forebrain' in the reference atlas ontology.

In contrast to more complex *in situ* landmark association methods as presented by Satija *et al.*<sup>49</sup> and Achim *et al.*<sup>50</sup>, the current method is focused on relative placement of mutually exclusive subpopulations. Because of this we are able to take advantage of both upregulated and downregulated gene sets in assigning the most likely spatial distribution of each identified subpopulation. For example, genes upregulated in the maturing NPCs relative to early NPCs can be used as indicators as to where the maturing NPC subpopulation is spatially localized. In addition, genes downregulated in maturing NPCs relative to early NPCs can also be used as indicators as to where maturing NPCs may be absent. Additionally, unlike Satija *et al.*<sup>49</sup>, we do not binarize the in situ data since we are particularly interested in gradients of expression across voxels or bins in our particular case. Likewise, due to the resolution limitations of our in situ data, where each voxel is much bigger than one cell, we are unable to precisely map individual cells to single locations as in Achim *et al.*'s method<sup>50</sup>.

**Implementation and data availability**—The PAGODA functions are implemented in version 1.99 of *scde* R package, available at <http://pklab.med.harvard.edu/scde/>. The source code is available on GitHub (<https://github.com/hms-dbmi/scde>). The spatial mapping of

neural cells based on the data generated by the Allen Institute for Brain Science has been implemented as a separate R package, called *brainmapr*, available from GitHub (<https://github.com/hms-dbmi/brainmapr>). The scRNA-seq data and gene count matrix for the NPC cells is available from Gene Expression Omnibus (GEO) under the GSE76005 accession number.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgments

We thank D. Usoskin, P. Ernfors and S. Linnarsson for helpful comments on the analysis approach. The work was supported by the Ellison Medical Foundation award and US National Science Foundation (NSF) CAREER award (NSF-14-532) to P.V.K, NSF Graduate Research Fellowship (DGE1144152) to J.F, US National Institutes of Health (NIH) grants U01 MH098977 to K.Z. and J.C., NIH R01 NS084398 to J.C. G.E.K. was supported by NIH T32 AG00216.

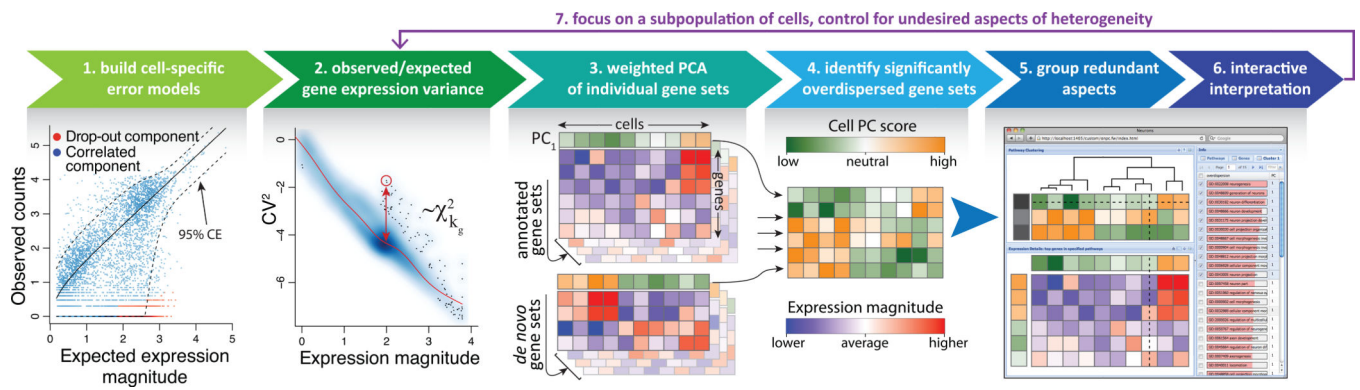
## References

1. Islam S, et al. *Nat Methods*. 2014; 11:163–166. [PubMed: 24363023]
2. Picelli S, et al. *Nat Methods*. 2013; 10:1096–1098. [PubMed: 24056875]
3. Tang F, et al. *PLoS One*. 2011; 6:e21208. [PubMed: 21731673]
4. Yan L, et al. *Nat Struct Mol Biol*. 2013; 20:1131–1139. [PubMed: 23934149]
5. Jaitin DA, et al. *Science*. 2014; 343:776–779. [PubMed: 24531970]
6. Dalerba P, et al. *Nat Biotechnol*. 2011; 29:1120–1127. [PubMed: 22081019]
7. Shalek AK, et al. *Nature*. 2014; 510:363–369. [PubMed: 24919153]
8. Usoskin D, et al. *Nat Neurosci*. 2015; 18:145–153. [PubMed: 25420068]
9. Zeisel A, et al. *Science*. 2015; 347:1138–1142. [PubMed: 25700174]
10. Deng Q, Ramskold D, Reinius B, Sandberg R. *Science*. 2014; 343:193–196. [PubMed: 24408435]
11. Buettner F, et al. *Nat Biotechnol*. 2015; 33:155–160. [PubMed: 25599176]
12. Macosko EZ, et al. *Cell*. 2015; 161:1202–1214. [PubMed: 26000488]
13. Klein AM, et al. *Cell*. 2015; 161:1187–1201. [PubMed: 26000487]
14. Patel AP, et al. *Science*. 2014; 344:1396–1401. [PubMed: 24925914]
15. Grun D, Kester L, van Oudenaarden A. *Nat Methods*. 2014; 11:637–640. [PubMed: 24747814]
16. Buettner F, Theis F. J. *Bioinformatics*. 2012; 28:i626–i632.
17. van der Maaten LJP, Hinton GE. *J Mach Learn Res*. 2008; 9:2579–2605.
18. Brennecke P, et al. *Nat Methods*. 2013; 10:1093–1095. [PubMed: 24056876]
19. Subramanian A, Kuehn H, Gould J, Tamayo P, Mesirov JP. *Bioinformatics*. 2007; 23:3251–3253. [PubMed: 17644558]
20. Blaschke AJ, Staley K, Chun J. *Development*. 1996; 122:1165–1174. [PubMed: 8620843]
21. Rehen SK, et al. *Proc Natl Acad Sci U S A*. 2001; 98:13361–13366. [PubMed: 11698687]
22. Kingsbury MA, Yung YC, Peterson SE, Westra JW, Chun J. *Cell Mol Life Sci*. 2006; 63:2626–2641. [PubMed: 16952055]
23. Peterson SE, et al. *J Neurosci*. 2012; 32:16213–16222. [PubMed: 23152605]
24. Herr KJ, Herr DR, Lee CW, Noguchi K, Chun J. *Proc Natl Acad Sci U S A*. 2011; 108:15444–15449. [PubMed: 21878565]
25. Mirendil H, et al. *Transl Psychiatry*. 2015; 5:e541. [PubMed: 25849980]
26. Yung YC, et al. *Sci Transl Med*. 2011; 3:99ra87.
27. Kharchenko PV, Silberstein L, Scadden DT. *Nat Methods*. 2014; 11:740–742. [PubMed: 24836921]

28. Interactive views of PAGODA results. <http://pklab.med.harvard.edu/scde/pagoda.links.html>.
29. Pollen AA, et al. *Nat Biotechnol.* 2014; 32:1053–1058. [PubMed: 25086649]
30. Ramskold D, et al. *Nat Biotechnol.* 2012; 30:777–782. [PubMed: 22820318]

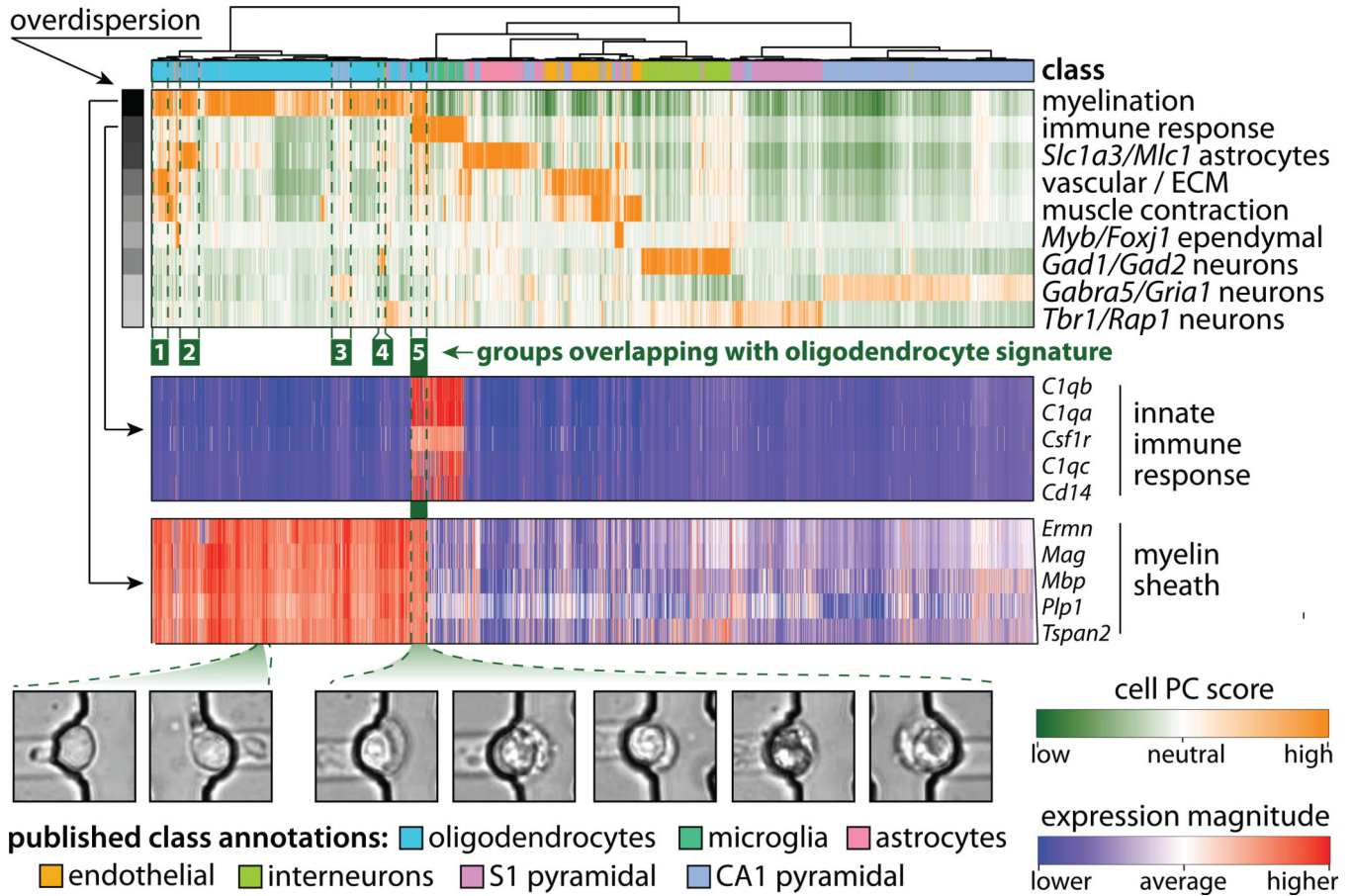
## Extended References

31. Kawaguchi A, et al. *Development.* 2008; 135:3113–3124. [PubMed: 18725516]
32. Kriegstein A, Noctor S, Martinez-Cerdeno V. *Nat Rev Neurosci.* 2006; 7:883–890. [PubMed: 17033683]
33. Lein ES, et al. *Nature.* 2007; 445:168–176. [PubMed: 17151600]
34. Englund C, et al. *J Neurosci.* 2005; 25:247–251. [PubMed: 15634788]
35. Uetsuki T, Takagi K, Sugiura H, Yoshikawa K. *J Biol Chem.* 1996; 271:918–924. [PubMed: 8557705]
36. Minamide R, Fujiwara K, Hasegawa K, Yoshikawa K. *PLoS One.* 2014; 9:e84460. [PubMed: 24392139]
37. Huang Z, Fujiwara K, Minamide R, Hasegawa K, Yoshikawa K. *J Neurosci.* 2013; 33:10362–10373. [PubMed: 23785149]
38. Kurita M, Kuwajima T, Nishimura I, Yoshikawa K. *J Neurosci.* 2006; 26:12003–12013. [PubMed: 17108174]
39. Anderson SA, Eisenstat DD, Shi L, Rubenstein JL. *Science.* 1997; 278:474–476. [PubMed: 9334308]
40. Wonders CP, Anderson SA. *Nat Rev Neurosci.* 2006; 7:687–696. [PubMed: 16883309]
41. Ma T, et al. *Cereb Cortex.* 2012; 22:2120–2130. [PubMed: 22021915]
42. Anders S, Huber W. *Genome Biol.* 2010; 11:R106. [PubMed: 20979621]
43. Fisher, RA. *Statistical Methods for Research Workers.* Hafner Publishing Company; 1970.
44. Abdel, HE. *Encyclopedia of Environmetrics.* 2nd. Wiley; 2012.
45. Hasings C, Mosteller F, Tukey JW, Winsor CP. *Ann. Math. Statist.* 1974:413–426.
46. Bailey S. 2012; 124:1023.
47. Johnstone IM. *Ann. Statist.* 2001; 29
48. Benjamini Y, Hochberg Y. *J Roy Stat Soc.* 1995; 57:289–300.
49. Satija R, Farrell JA, Gennert D, Schier AF, Regev A. *Nat Biotechnol.* 2015; 33:495–502. [PubMed: 25867923]
50. Achim K, et al. *Nat Biotechnol.* 2015; 33:503–509. [PubMed: 25867922]



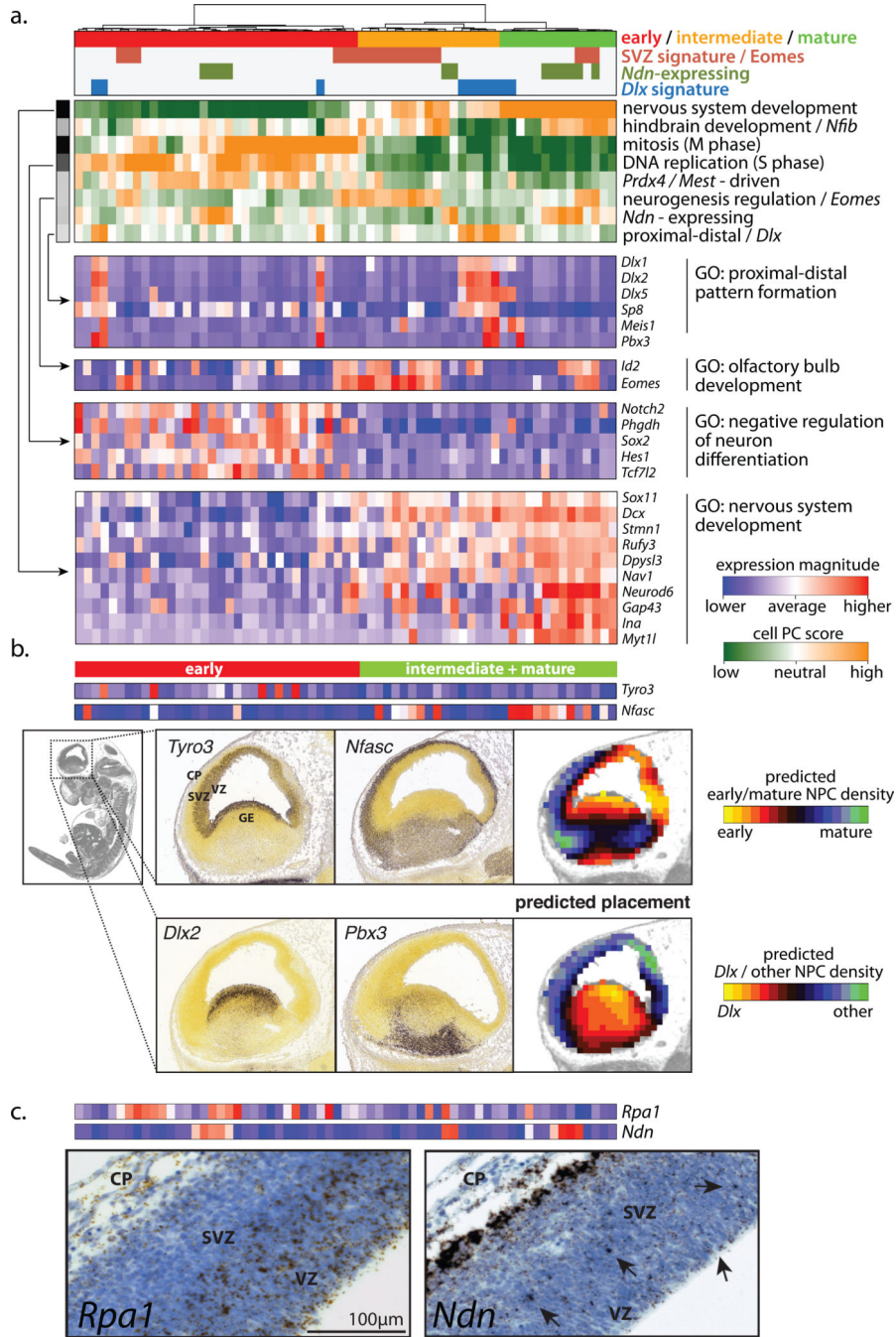
**Figure 1. Pathway and gene set overdispersion analysis (PAGODA)**

Transcriptional heterogeneity analyzed through the following key steps: **1.** Error models are fit for each cell to quantify the dependency of amplification noise and drop-out probabilities on the expression magnitude<sup>27</sup>. A model fit for a cell is shown, separating drop-out and amplified components, and the 95% confidence envelope of the amplified component; **2.** The residual expression variance magnitude for each gene is determined relative to the transcriptome-wide expectation model (red curve), taking into account the uncertainty in the variance estimates of each gene by determining effective degrees of freedom ( $k_g$ ) for the  $\chi^2$  distribution; **3.** Weighted PCA analysis is performed independently on functionally-annotated gene sets, as well as *de novo* gene sets determined based on correlated expression in the current dataset; **4.** If the amount of variance explained by a principal component of a gene set is significantly higher than expected, the gene set is called *overdispersed*, and the cell scores defined by that principal component (coded in orange-green gradient) are included as one of the significant aspects of heterogeneity; **5.** Redundant aspects that are driven by the same genes or show similar patterns of cell separation are grouped to provide succinct overview of heterogeneity; **6.** A web browser-based interface is used to navigate the identified aspects of heterogeneity, associated gene sets and gene expression patterns. **7.** Depending on the biological question, some of the detected aspects of heterogeneity may be deemed artifactual or extraneous, and can be actively controlled for in a subsequent iteration.



**Figure 2. PAGODA analysis of the 3,005 cells from mouse cortex and hippocampus measured by Zeisel *et al.***

The dendrogram shows the overall clustering of the cells, with the row immediately below specifying the group to which each cell was assigned in the original analysis by Zeisel *et al.* The main panel shows the top 9 significant aspects ( $P < 0.05$ ) of heterogeneity (rows) detected by PAGODA based on gene sets defined by GO annotations, with the orange/white/green gradient indicating high/neutral/low score of a cell with respect to a given aspect. The aspect scores are oriented so that high (orange) and low (green) values generally correspond, respectively, to increased and decreased expression of the associated gene sets. Row labels summarize the key functional annotations of the gene sets in each aspect. Two subsequent panels show expression patterns of top-loading genes innate immune response (from the aspect distinguishing neuroglia), and myelin sheath (distinguishing oligodendrocytes). A population of ~35 cells expressing both signatures is marked by a green bar, and most likely represents capture of two associated cells of different type. The bottom panel shows images of the microfluidic traps corresponding to some of the dual-signature cells, along with cells (leftmost two) exhibiting only the oligodendrocyte signature. Green boxes below the main panel highlight cells showing a combination of the oligodendrocyte signature with other cell types (numbered 1–5: vascular endothelial, astrocytes, CA1 neurons, *Gad1/2* interneurons and neuroglia). Detailed composition is available through an interactive online view<sup>28</sup>.



**Figure 3. Transcriptional heterogeneity of 65 neuronal progenitor cells in embryonic mouse cortex**

**a.** Top eight significant ( $P < 0.01$ ) aspects of heterogeneity are shown, labeled by their primary GO category or driving genes. Detailed are available through an online browser<sup>28</sup>. Top aspect tracks induction of neuronal maturation pathways, driving the overall subpopulation structure. Mitotic and S-phase signatures in early NPCs account for the next two most significant aspects, with the S-phase aspect incorporating closely matching



expression patterns of genes responsible for NPC maintenance. Color codes in the top panel summarize key subpopulations of NPCs distinguished by the detected heterogeneity aspects.

**b.** Anatomical placement of the early vs. maturing NPC classes within embryonic brain. *In situ* hybridization signals in E13.5 mouse brain are shown for *Tyro3* and *Nfasc*, with the two heatmap rows above showing their expression in the scRNA-seq. Computational prediction (third panel) based on the overall transcriptional profile places early NPCs near VZ, and maturing ones in SVZ (subventricular zone)/CP regions. *In situ* images were generated by Allen Institute for Brain Science<sup>33</sup>. The lower panel shows anatomical placement of the Dlx-expressing NPCs, and *in situ* images for the associated genes.

**c.** Validation of genes associated with specific subpopulations by *in situ* hybridization. Coronal E13.5 brain sections labeled using RNAscope probes for *Rpa1* (left) and *Ndn* (right). *Rpa1* showed high expression in the ventricular (VZ) and sub-ventricular zone (SVZ). *Ndn*, which is marks a distinct subpopulation of both mature and early NPCs, shows prominent expression throughout the CP, with rarer high expressing cells in the VZ and SVZ (black arrows).