Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis

Jean Fan[1], Neeraj Salathia[2], Rui Liu[3], Gwendolyn E Kaeser[4], Yun C Yung[4], Joseph L Herman[1], Fiona Kaper[2], Jian-Bing Fan[2,6], Kun Zhang[3], Jerold Chun[4] & Peter V Kharchenko[1,5]

The transcriptional state of a cell reflects a variety of biological factors, from cell-type-specific features to transient processes such as the cell cycle, all of which may be of interest. However, identifying such aspects from noisy single-cell RNA-seq data remains challenging. We developed pathway and gene set overdispersion analysis (PAGODA) to resolve multiple, potentially overlapping aspects of transcriptional heterogeneity by testing gene sets for coordinated variability among measured cells.

Single-cell transcriptome measurements[1,2] provide an unbiased approach for studying the complex cellular compositions of healthy and diseased tissues[3–9]. High levels of technical noise[10] and a strong dependence on expression magnitude pose difficulties for principal-component analysis (PCA) and other dimensionality reduction approaches such as the Gaussian process latent variable model (GP-LVM)[11] and *t*-distributed stochastic neighbor embedding (t-SNE)[12]. Even when cell-to-cell differences expose prominent biological processes taking place in the measured cells, such as cell cycle or metabolic processes, these processes might not be of primary interest[6]. Such cross-cutting transcriptional features represent alternative ways to classify cells and pose a challenge for the commonly used clustering approaches that aim to reconstruct a single subpopulation structure[4–6,13]. Partitioning methods such as *k*-means clustering and the specialized BackSPIN algorithm[5] may, for example, classify cells first on the basis of cell cycle phase, rather than tissue-specific signaling state, if the cell cycle differences are more pronounced.

Here we describe PAGODA, an alternative approach for analyzing transcriptional heterogeneity that aims to detect all statistically significant ways in which measured cells can be classified. PAGODA (available at http://pklab.med.harvard.edu/scde/ and as **Supplementary Software**) evaluates the coordinated expression variability of genes in both annotated pathways and automatically detected gene sets. Gene set testing with methods such as GSEA[14] has been widely used for 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, whereas cell-to-cell variability in the expression of a single neuronal differentiation marker such as *Neurod1* may be considered noisy and inconclusive, coordinated upregulation of many genes associated with neuronal differentiation in the same subset of cells could provide a prominent signature distinguishing a subpopulation of differentiating neurons. We used PAGODA with published data sets to recover both new and known subpopulations and suggest their likely functional roles.

Transcriptional diversity in mouse neural progenitor cells (NPCs) is likely to depend on a variety of intrinsic and external factors that include programmed cell death[15], genomic mosaicism[16,17] and exposure to signaling lipids[18]. Using single-cell RNA-seq (scRNA-seq) to assess a cohort of cortical NPCs from an embryonic mouse, we found that PAGODA recovered the known neuroanatomical and functional organization of NPCs. Our approach allowed us to identify multiple aspects of transcriptional heterogeneity in the developing mouse cortex that are difficult to discern with existing heterogeneity-analysis approaches.

To characterize significant aspects of transcriptional heterogeneity, PAGODA uses a series of steps (**Fig. 1** and Online Methods). First, the effective sequencing depth, drop-out rate and amplification noise of each cell are estimated via a previously described mixture-model approach[19] with minor enhancements (step 1; **Fig. 1**). Using these models, the observed expression variance of each gene is renormalized on the basis of the expected genome-wide variance at the appropriate expression magnitude (step 2). Batch correction is also performed at this stage. The resulting residual variance, modeled by the χ^2 statistic, effectively distinguishes subpopulation-specific genes (**Supplementary Notes 1** and **2**) and determines the contribution of each gene to 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 in a given data set (*de novo* gene sets). The prevalent transcriptional signature of each gene set is captured by its first principal component (PC), with weighted PCA used to adjust for technical noise. If the amount of variance explained by the first PC of a given gene set is significantly higher than expected (step 4, correcting for

[1]Department of Biomedical Informatics, Harvard Medical School, Boston, Massachusetts, USA. [2]Illumina Inc., San Diego, California, USA. [3]Department of Bioengineering, University of California, San Diego, California, USA. [4]Department of Molecular and Cellular Neuroscience, Dorris Neuroscience Center, The Scripps Research Institute, La Jolla, California, USA. [5]Harvard Stem Cell Institute, Cambridge, Massachusetts, USA. [6]Present address: AnchorDx Corporation, International Biotech Island, Guangzhou, Guangdong, China. Correspondence should be addressed to P.V.K. (peter_kharchenko@hms.harvard.edu).

RECEIVED 26 MAY 2015; ACCEPTED 16 DECEMBER 2015; PUBLISHED ONLINE 18 JANUARY 2016; DOI:10.1038/NMETH.3734

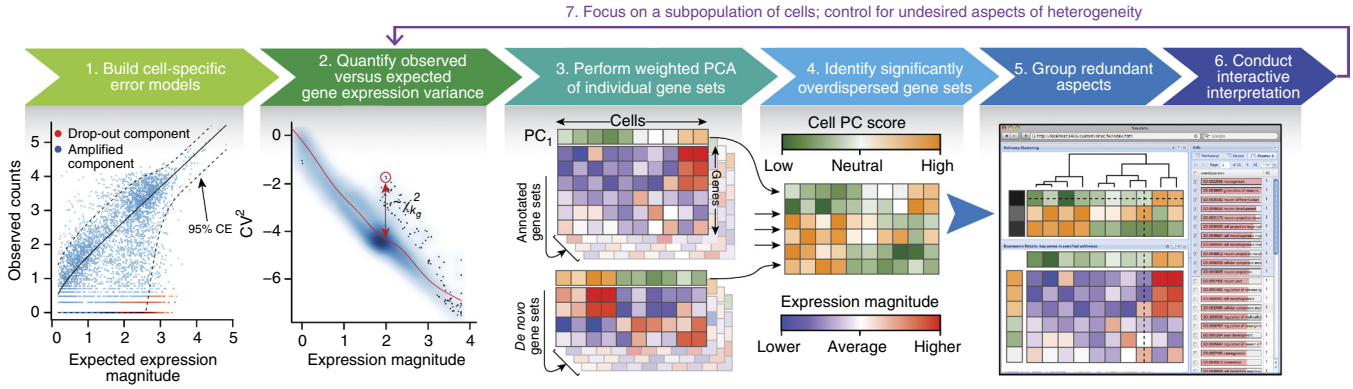

Figure 1 | Overview of PAGODA. Transcriptional heterogeneity is analyzed in seven steps. (1) Error models are fit for each cell[19]. A model fit for a cell is shown, separating drop-out and amplified components with the 95% confidence envelope (CE) of the amplified component. (2) The residual expression variance for each gene is determined relative to the transcriptome-wide expectation model (red curve), taking into account the uncertainty in the variance estimate for each gene by determining the effective degrees of freedom (k_g) for the χ^2 distribution. CV, coefficient of variation. (3) Weighted PCA is performed on annotated gene sets and on *de novo* gene sets determined on the basis of correlated expression in the current data set. (4) Cell PC scores of overdispersed gene sets (those with PC variance significantly higher than expected) are identified as significant aspects of heterogeneity. (5) Redundant aspects are grouped to provide a succinct overview of heterogeneity. (6) A web interface is used to navigate the identified aspects of heterogeneity, associated gene sets and gene expression patterns. (7) Aspects of heterogeneity deemed artifactual or extraneous with respect to the biological question can be controlled for in a subsequent iteration.

multiple hypotheses), the gene set is said to be 'overdispersed' and is included in the subsequent analysis.

Many PCs will separate cells in a similar way, either because the same genes drive them or because multiple biological processes distinguish the same subsets of cells. To provide a nonredundant view of transcriptional heterogeneity, 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**). Major aspects of transcriptional heterogeneity can be explored numerically or through an interactive web browser interface (step 6). As we illustrate below, examining individual aspects and their relationships can provide insights and functional clues that are not apparent with the most prominent cell classification. Finally, if one or more aspects of transcriptional heterogeneity are determined to be extraneous to the biological context, there is an option to control for them explicitly (step 7).

To illustrate the use of PAGODA on a complex cell population, we re-examined scRNA-seq data for 3,005 cells from mouse cortex and hippocampus[5]. This extensive data set covered a variety of cell types with distinct expression signatures. PAGODA revealed nine major aspects of heterogeneity that distinguish the seven top-level classes and two lower-level subpopulations originally identified by BackSPIN[5], a recursive partitioning method (**Fig. 2**). The functional interpretation of the identified aspects was evident from the identity of the overdispersed GO categories. The most significant aspect separated oligodendrocytes, which are easily distinguished by strong overdispersion of myelination-related pathways. Similarly, overdispersion of immune, vascular and muscle-associated GO-annotated gene sets identified microglia, vascular endothelial and mural subpopulations, respectively. Other cell types, such as ependymal cells and different types of neurons, were distinguished by *de novo* gene set signatures, with most overdispersed genes revealing their identity (for example, *Gad1*, *Tbr1* and *Gabra5*).

Aspects distinguishing many of the cell types seemed to overlap, most frequently with the myelination signature. For instance, a subset of 35 cells showed prominent expression of both immune-response genes characteristic of microglia and genes responsible for production of the myelin sheath (**Fig. 2**; interactive PAGODA results can be found at http://pklab.med.harvard.edu/scde/pagoda.links.html). Similarly, a myelin-associated expression signature was observed for a subset of vascular cells, astrocytes, pyramidal neurons and interneurons. These hybrid signatures are most likely to correspond to cases in which two different cells were captured together (co-occurrence frequencies are presented in **Supplementary Fig. 2**). BackSPIN and other partitioning methods would need to classify such cells on the basis of a single signature or isolate them as a separate class without exposing their relationship to other groups. In contrast, PAGODA can expose multiple alternative classifications of a given cell.

We further evaluated PAGODA's performance by reanalyzing data sets that were used to present alternative methods of heterogeneity analysis[4,6,20], recovering previously identified subpopulations and identifying additional biologically relevant features (**Supplementary Note 3**). In particular, PAGODA's ability to associate a given cell with multiple, potentially independent aspects of transcriptional heterogeneity allows one to focus on biologically relevant subpopulations that are distinguished by subtle transcriptional variation. For instance, in reanalyzing data for mouse CD4$^+$ T cells that were used to present an elegant GP-LVM approach[6], 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 had not been observed previously.

As heterogeneity among NPCs can influence downstream neural diversity, we performed Smart-Seq on 65 NPCs isolated from the cerebral cortex of mice at embryonic day 13.5 (E13.5; Online Methods). The most significant aspect of heterogeneity identified by PAGODA reflected gradual induction of the genes associated with neuronal maturation and growth (**Fig. 3a**). Approximately half of the cells expressed *Dcx*, *Sox11* and other known markers

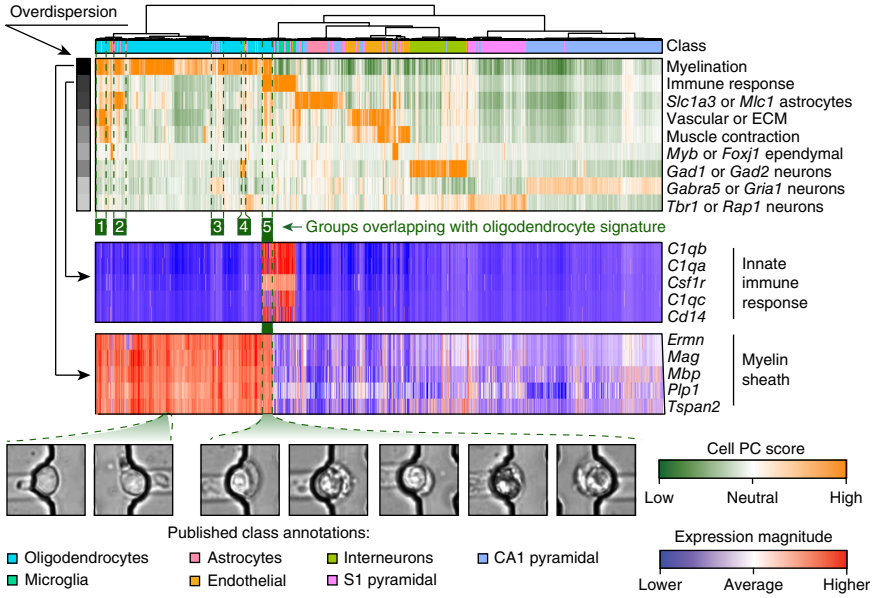

Figure 2 | PAGODA analysis of data from 3,005 mouse cortical and hippocampal cells[5]. The dendrogram shows overall clustering and the first row indicates group assignments from the original analysis[5]. The rows below reflect the top nine significant aspects of heterogeneity ($P < 0.05$) detected by PAGODA on the basis of gene sets defined by GO annotations. Aspect scores (Cell PC score) are oriented so that high values generally correspond to increased expression of associated gene sets. Row labels summarize key functional annotations of gene sets in each aspect. Also shown are expression patterns of top-loading genes for 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 probably represents capture of two associated cells of different types. The images at the bottom show the microfluidic traps corresponding to some of the dual-signature cells, along with cells exhibiting only the oligodendrocyte signature (leftmost two images). Green numbered boxes below the uppermost panel highlight cells showing a combination of signatures of oligodendrocytes and other cell types (1–5 denote, respectively, vascular endothelial cells, astrocytes, CA1 neurons, Gad1/2 interneurons and neuroglia).

of neuronal maturation, with the most mature subset expressing genes involved in neuronal maturation and growth cones (*Neurod6* and *Gap43*). Such cells maintain expression of some progenitor markers (for example, vimentin) and therefore probably 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 upregulation of genes characteristic of an early progenitor state[21] (*Sox2*, *Notch2* and *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[22]. We used spatial expression patterns[23] of genes differentially expressed between early and maturing NPCs to reconstruct the most likely spatial distribution of these cells in the mouse brain (**Fig. 3b** and Online Methods). As expected, we found that early NPCs localized close to the ventricular zone (VZ). We also used RNA-FISH (Online Methods) to examine two genes, *Rpa1* and *Ndn*, of unknown relationship to the embryonic cerebral cortex (**Supplementary Fig. 3**). Consistent with the predicted pattern, *Rpa1* was most prominent in proliferative regions. *Ndn* localized in postmitotic regions (especially the cortical plate), as well as in rare cells within the subventricular zone (SVZ; **Supplementary Fig. 3**).

An additional subset of NPCs was distinguished by their expression of *Eomes*, *Neurod1* and other genes localized to the

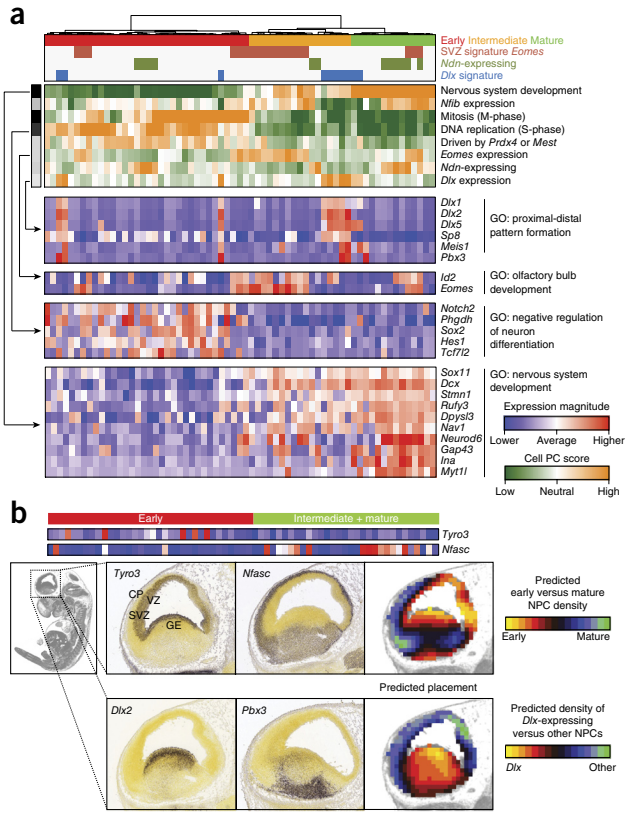

Figure 3 | Transcriptional heterogeneity of 65 NPCs in embryonic mouse cortex. (**a**) The top eight significant ($P < 0.01$) aspects of heterogeneity are shown, labeled by primary GO category or driving genes. The top aspect tracks the induction of neuronal maturation pathways, driving the overall subpopulation 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. The top panel summarizes key subpopulations of NPCs distinguished by the detected heterogeneity aspects. (**b**) Location of early versus maturing NPC classes within embryonic brain. *In situ* hybridizations in E13.5 mouse brain are shown for *Tyro3* and *Nfasc*, with the two heat map rows at the top showing scRNA-seq expression. Computational prediction (rightmost panels in image rows) based on the overall transcriptional profile placed early NPCs near the VZ and maturing ones in the SVZ and cortical plate (CP) regions. *In situ* images were generated by the Allen Institute for Brain Science[23]. The bottom row of images shows the anatomical placement of the *Dlx*-expressing NPCs and *in situ* images for the associated genes. GE, ganglionic eminence.

SVZ region and thought to distinguish basal progenitors[21,24]. The *Eomes* signature marks cells with intermediate levels of genes associated with neuronal maturation, as well as a subset of early NPCs undergoing DNA replication, probably representing neuronally committed NPCs maturing in the SVZ and dividing basal NPCs, respectively. These dividing cells expressed Notch signaling genes (*Dll1*, *Notch2* and *Mfng*) concurrently with *Eomes* and therefore were probably nascent basal progenitors[21].

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 its high expression in mature neurons[25], also has been shown to be expressed in the VZ[26] and to restrict both proliferation and apoptosis rates in NPCs[26,27]. Using PAGODA in combination with RNAscope analyses (**Supplementary Fig. 3**), we found that *Ndn* was expressed in a subset of NPCs, approximately a quarter of which exhibited pronounced mitotic signatures and probably represented cells 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 and migrate to the cortical areas, giving rise to GABAergic neurons[28,29]. *Dlx*-positive cells express other markers of tangentially migrating NPCs, most notably Sp9 and Sp8 transcription factors[30]. Indeed, spatial localization of these cells was predicted to be in the region of the ganglionic eminence, where tangentially migrating NPCs are expected to originate (**Fig. 3b**). In agreement with earlier observations of such NPCs undergoing mitosis in the cortical VZ and SVZ, 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, independent-component analysis, t-SNE[7,12], GP-LVM[11] and BackSPIN[5] (**Supplementary Figs. 4** and **5**). Although none of the methods recovered all of the identified subpopulations, BackSPIN provided the most compelling results, capturing heterogeneity involving expression of *Dlx* and *Prdx4* or *Mest*. However, the reported clustering grouped only some of the cells associated with each signature, illustrating the limitations of partitioning-based interpretation in a complex biological context.

Just like whole organisms, individual cells can be classified according to a variety of meaningful criteria. For example, tangentially migrating NPCs, 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, a single 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 aim to identify significant transcriptional features distinguishing cells in a population and facilitate their interpretation.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. The scRNA-seq data and gene-count matrix for the NPCs are available in the Gene Expression Omnibus (GEO) under accession number GSE76005.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

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

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

The authors declare competing financial interests: details are available in the online version of the paper.

1. Islam, S. *et al. Nat. Methods* **11**, 163–166 (2014).
2. Picelli, S. *et al. Nat. Methods* **10**, 1096–1098 (2013).
3. Tang, F. *et al. PLoS ONE* **6**, e21208 (2011).
4. Usoskin, D. *et al. Nat. Neurosci.* **18**, 145–153 (2015).
5. Zeisel, A. *et al. Science* **347**, 1138–1142 (2015).
6. Buettner, F. *et al. Nat. Biotechnol.* **33**, 155–160 (2015).
7. Macosko, E.Z. *et al. Cell* **161**, 1202–1214 (2015).
8. Klein, A.M. *et al. Cell* **161**, 1187–1201 (2015).
9. Patel, A.P. *et al. Science* **344**, 1396–1401 (2014).
10. Grün, D., Kester, L. & van Oudenaarden, A. *Nat. Methods* **11**, 637–640 (2014).
11. Buettner, F. & Theis, F.J. *Bioinformatics* **28**, i626–i632 (2012).
12. van der Maaten, L.J.P. & Hinton, G.E. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
13. Jaitin, D.A. *et al. Science* **343**, 776–779 (2014).
14. Subramanian, A., Kuehn, H., Gould, J., Tamayo, P. & Mesirov, J.P. *Bioinformatics* **23**, 3251–3253 (2007).
15. Blaschke, A.J., Staley, K. & Chun, J. *Development* **122**, 1165–1174 (1996).
16. Rehen, S.K. *et al. Proc. Natl. Acad. Sci. USA* **98**, 13361–13366 (2001).
17. Peterson, S.E. *et al. J. Neurosci.* **32**, 16213–16222 (2012).
18. Herr, K.J., Herr, D.R., Lee, C.W., Noguchi, K. & Chun, J. *Proc. Natl. Acad. Sci. USA* **108**, 15444–15449 (2011).
19. Kharchenko, P.V., Silberstein, L. & Scadden, D.T. *Nat. Methods* **11**, 740–742 (2014).
20. Pollen, A.A. *et al. Nat. Biotechnol.* **32**, 1053–1058 (2014).
21. Kawaguchi, A. *et al. Development* **135**, 3113–3124 (2008).
22. Kriegstein, A., Noctor, S. & Martinez-Cerdeno, V. *Nat. Rev. Neurosci.* **7**, 883–890 (2006).
23. Lein, E.S. *et al. Nature* **445**, 168–176 (2007).
24. Englund, C. *et al. J. Neurosci.* **25**, 247–251 (2005).
25. Uetsuki, T., Takagi, K., Sugiura, H. & Yoshikawa, K. *J. Biol. Chem.* **271**, 918–924 (1996).
26. Minamide, R., Fujiwara, K., Hasegawa, K. & Yoshikawa, K. *PLoS ONE* **9**, e84460 (2014).
27. Huang, Z., Fujiwara, K., Minamide, R., Hasegawa, K. & Yoshikawa, K. *J. Neurosci.* **33**, 10362–10373 (2013).
28. Anderson, S.A., Eisenstat, D.D., Shi, L. & Rubenstein, J.L. *Science* **278**, 474–476 (1997).
29. Wonders, C.P. & Anderson, S.A. *Nat. Rev. Neurosci.* **7**, 687–696 (2006).
30. Ma, T. *et al. Cereb. Cortex* **22**, 2120–2130 (2012).

ONLINE METHODS

Isolation and single-cell RNA-seq of mouse NPCs and astrocytes. Single NPCs were isolated from C57BL/6J E13.5 mouse cortices for RNA-seq. Timed-pregnant mice were killed by deep anesthesia followed by cervical dislocation. The embryos were quickly removed, after which cortical hemispheres were isolated, ganglionic eminences were 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, California, USA) and conformed to the US National Institutes of Health guidelines.

Single cells were isolated by gentle trituration in ice-cold phosphate-buffered saline containing 2 mM EGTA using P1000 tips with decreasing bore diameter. Cells were then filtered through a 40-µm nylon cell strainer and stained with propidium iodide (PI), a live-dead stain, after which fluorescence-activated single-cell sorting was performed to select for PI-negative cells. Samples remained on ice throughout the process, and the total processing time from cervical dislocation to sorting was limited to 2 h. Single cells were sorted directly into the cell lysis buffer provided in the Clontech SMARTer Ultra Low RNA kit for Illumina sequencing (catalog no. 634936), and sequencing libraries were generated using the manufacturer's protocol. The resulting libraries were sequenced on the Illumina HiSeq 2000 sequencing platform.

Gene validation using *in situ* hybridization with RNAscope. Mouse E13.5 embryos were removed from timed-pregnant mice and prepared according to the RNAscope instructions for paraffin-embedded tissue. RNAscope probes (Advanced Cell Diagnostics) were designed by the manufacturer (catalog nos. GINS2 435891 and RPA1 435911), and sections were processed using the RNAscope 2.0 High Definition Reagent Kit—BROWN (catalog no. 310035) according to the manufacturer's instructions. Sections were imaged on a Zeiss Axioimager at 20× magnification.

Previously published scRNA-seq data. For the mixture of cultured human NPCs and primary cortical samples used by Pollen *et al.*[20], we downloaded SRA files for each study from the Sequence Read Archive (http://www.ncbi.nlm.nih.gov/sra) and converted them to FASTQ format using the SRA toolkit (v2.3.5). We aligned FASTQ files to the human reference genome (hg19) using Tophat (v2.0.10) with Bowtie2 (v2.1.0) and Samtools (v0.1.19). We quantified gene expression counts using HTSeq (v0.5.4). We downloaded read counts for the T_H2 data published by Buettner *et al.*[6] from their supplementary site (http://github.com/PMBio/scLVM/blob/master/data/Tcell/data_Tcells.Rdata). Read (or UMI) count matrices for other two data sets were downloaded from GEO (GSE60361 for Zeisel *et al.*[5] and GSE59739 for Usoskin *et al.*[4]).

Fitting single-cell error models. Following the approach described by Kharchenko *et al.*[19], the read count for a gene g in a cell i (c_g^i) was modeled as a mixture of a negative binomial (NB) (signal) and Poisson (drop-out) components:

$$c_g^i \sim p_i^d(e_g)\text{Poisson}\left(\lambda_{bg}\right)+\left(1-p_i^d(e_g)\right)\text{NB}\left(\alpha_i e_g,\theta_i(e_g)\right),$$

where $p_i^d(e_g)$ is the probability of encountering a drop-out event in a cell i for a gene with a population-wide expected expression magnitude e_g (fragments per kilobase of transcript per million mapped reads); $\lambda_{bg}=0.1$ is the low-level signal rate for the dropped-out observations; $\theta_i(e_g)$ is the NB size parameter (the functional form is described below); and α_i is the library size of cell i, as inferred from the fitting procedure. The single-cell error models were fit using the approach described by Kharchenko *et al.*[19], with the following modifications. (1) Instead of estimating the expected expression magnitudes of genes using all pairwise comparisons between all other cells, we compared each cell to its k most similar cells (on the basis of the 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 data set (one-third of the cells for mouse and human NPC data sets, and one-fifth for the larger data sets from Zeisel *et al.*[5] and Usoskin *et al.*[4]). (2) The count dependence on the expected expression magnitude was estimated on the linear scale with a zero intercept. (3) To improve fit, we modeled the drop-out probability using logistic regression on both the expression magnitude (log scale) and its square value. (4) Instead of fitting a constant value for the NB size parameter θ, we fit it as a function of the expression magnitude, using the following functional form:

$$\log(\theta)=a+\frac{h-a}{\left(1+10^{(x-m)*s}\right)^r},$$

where x is the expression magnitude (log scale) and a, h, m, s and 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[31] 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 will 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 mean μ_g and variance v_g (both equal to some Poisson rate λ_g), then Fisher's index of dispersion

$$D_g=\sum_{i=1}^{k}\left(c_g^i-\mu_g\right)^2\Big/v_g$$

follows a χ^2_{k-1} distribution[32]. For the Poisson case $v_g=\mu_g$, for a negative binomial process, $v_g=\mu_g+(\mu_g)^2/\theta$, where θ is the size parameter. As θ decreases from very high values where the NB is well approximated by a Poisson distribution, D_g diverges from χ^2_{k-1}. Analytical adjustments of D_g based on the NB moments can improve the χ^2 approximation[33]. For more accurate approximation, we used a numeric correction of the χ^2 degrees of freedom, depending on the magnitude of θ, so that $D_g\sim\chi^2_{f(\theta)}$ (**Supplementary Note 2**).

To account for the possibility of drop-out events, we used weighted sample variance estimates, so that

$$D_g=\sum_i\left[w_g^i\left(c_g^i-\mu_g^i\right)^2\right]/\left[\mu_g^i+(\mu_g^i)^2/\theta_i(e_g)\right]\sim\chi^2_{k_g},$$

where w_g^i is the probability that the measurement in a cell i was not a drop-out event on the basis of the error model for cell i, and

$$k_g=\sum_{i=1}^{k}w_g^i f\left(\theta_i(e_g)\right)$$

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.

Because NB (or mixed NB-Poisson) 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 noncentrality, we normalized D_g by its transcriptome-wide expectation value D_g^e, where D_g^e models the transcriptome-wide dependence of D_g on gene expression magnitude. We obtained estimates of D_g^e using a general additive model (fit using the *mgcv* R package) as a smooth function of gene expression magnitude e_g. To improve smoothness, we fit the general additive model on the corresponding squared coefficient of residual variance $(D_g / e_g)^2$. This fit was performed on all of the genes. The P value of overdispersion for a gene g was then calculated as $P_g^{\mathrm{od}} = F_{\chi^2_{k_g}}(k_g D_g / D_g^e)$, where $F_{\chi^2_k}$ is the cumulative distribution function of the χ^2 distribution with k degrees of freedom.

To improve the stability of the estimates with respect to outliers, we applied a Winsorization procedure[34] to the read count matrix before the variance evaluation described above. To ensure that the outliers were trimmed independently of the total cell coverage, we applied the Winsorization procedure to the FPM matrix (i.e., normalizing counts by the library size; FPM, fragments per million) and then translated the resulting values back into the integer counts. We used a trim value of 3 for all data sets (i.e., observations from the three highest and three lowest cells for each gene were Winsorized).

Weighted PCA and significance of pathway overdispersion. For PCA, we transformed the data to better approximate the standard normal distribution. Specifically, we carried out PCA on a matrix of log-transformed read counts with a pseudocount of 1, normalized by the library size: $x_g^i = \log(c_g^i / \alpha_i + 1)$. We then scaled the values for each gene (matrix row) so that the weighted variance of a given gene matched the tail probabilities of the distribution for a standard normal process:

$$y_g^i = x_g^i \sqrt{Q_{\mathrm{N}}(P_g^{\mathrm{od}}) / \mathrm{var}_{w_g}(x_g)},$$

where Q_{N} is the quantile function of the standard normal distribution and $\mathrm{var}_{w_g}(x_g)$ is the weighted variance of values x_g. As in our previous work[19], 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 data sets ($p^{bg}(c_g^i)$ is the probability of observing c_g^i counts in a drop-out event, evaluated from the Poisson distribution).

We performed weighted PCA for each gene set as described by Bailey[35], recording first (and optionally subsequent) PCs, the magnitude of the eigenvalue (λ_1) and associated cell scores for each gene set. Statistical significance of the λ_1 eigenvalues obtained for each gene set (overdispersion P value for a set s, P_s^{od}) was evaluated on the basis of the Tracy-Widom F_1 distribution[36] $F_1(m, n_e)$, 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 were obtained with pathways annotated by GO, restricting evaluation to the GO terms that had between 10 and 1,000 annotated genes.

Identification and statistical treatment of *de novo* gene clusters. As some aspects of transcriptional heterogeneity can be driven by genes that are poorly represented or not 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 data set. By default, PAGODA implements a straightforward clustering procedure: hierarchical clustering is performed using the 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 predefined number of *de novo* gene clusters (the results shown here include 150 clusters). As there are many alternative methods for clustering coexpressed genes, PAGODA implementation provides parameters for using alternative clustering procedures.

As *de novo* gene clusters are purposefully selected to contain genes with correlated expression profiles, the amount of variance explained by the first PC (magnitude of λ_1) will be greater than that expected from random matrices and cannot be modeled by the same Tracy-Widom F_1 distribution used for the previously annotated gene set. To evaluate the statistical significance of overdispersion, we generated a background distribution of λ_1 by performing the same hierarchical clustering and weighted PCA procedure on randomized matrices (where cell order was randomized for each gene independently with 100 randomizations). The λ_1 values were normalized relative to the Tracy-Widom F_1 expectation as

$$\lambda_1^s = \left(\lambda_1 - (a\lambda_1^{\mathrm{TW}} + bn)\right) / \sqrt{v_1^{\mathrm{TW}}},$$

where λ_1^{TW} and v_1^{TW} are the mean and variance of λ_1 predicted by the Tracy-Widom F_1 distribution, and coefficients a and b are determined by the linear model $\lambda_1 \sim \lambda_1^{\mathrm{TW}} + n$. This standardized residual λ_1^s was modeled using Gumbel extreme-value distribution, the parameters of which were fit using the extRemes package in R. The overdispersion P value for each *de novo* gene set was determined from the tails of that distribution. In the subsequent procedures, *de novo* gene sets and annotated gene sets were treated in the same way.

Clustering of redundant heterogeneity patterns. To compile a nonredundant set of aspects, the PC cell scores (projections on the eigenvector) from each significantly overdispersed (5% false discovery rate, as estimated via the Benjamini-Hochberg method[37]) gene set were normalized so that the magnitude of their variance corresponded to the tail probability of the χ^2 distribution:

$$\mathrm{var}(s_i) = Q_{\chi^2_{n-1}}\left(P_i^{\mathrm{od}}\right) / (n-1),$$

where $Q_{\chi^2_n}$ is the quantile function of the χ^2 distribution with n degrees of freedom (n is the number of cells in the data set). The redundant aspects of heterogeneity were reduced in two steps. First, we grouped aspects reflecting transcriptional variation of the same genes by evaluating the similarity of the corresponding gene loading scores in combination with the pattern similarity using the following distance measure between gene sets i and j:

$$d_{ij} = \left(1 - \sqrt{\left|\mathrm{cor}(l_i, l_j) * \mathrm{cor}(s_i, s_j)\right|}\right),$$

doi:10.1038/nmeth.3734

where cor is the Pearson linear correlation, l_i,l_j are the loading scores of genes found in sets i and j, and s_i,s_j are the corresponding PC cell scores (d_{ij} was set to 1 if there were fewer than two genes in common between gene sets i and j). We then used the distance d_{ij} 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 PC of all aspects in a grouped cluster. The second step, aimed at grouping aspects showing similar patterns of cell separation, involved another round of hierarchical clustering using the cor(s_i,s_i) distance measure with the Ward clustering procedure. The similarity threshold for the final grouping of similar aspects varied between data sets depending on their complexity (0.5 for the human NPC data, 0.95 for the mouse cortical-hippocampal data set, and 0.9 for the T cell and mouse NPC data).

Batch correction. To control for the effect of categorical covariates, such as the 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, we evaluated the ratio of batch-specific to batch-adjusted variance as $\alpha_g = D_{g,b}/D_g$. The residual variance of genes showing discrepant batch- and population-specific variance was taken as $D_g^b = \min(\alpha_g, 1/\alpha_g) * D_{g,b}/D_g^e$ and $P_g^{\text{od}} = F_{\chi^2_{k_g}}(k_g D_g^b / D_g^e)$.

The procedure described 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 a sufficiently high amount of total variance to be picked by the PCA. 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 disparity in the frequency of the drop-out observations between batches, we determined the lower bound of the zero-count observation fraction (u) for each batch (assuming a binomial process) and multiplied the weights w_g^i for each batch by $\min(1,\max(u)/z_b)$, where $\max(u)$ is the maximum lower bound value among batches and z_b is the fraction of zero-count observations in a given batch. This procedure ensured that the expected number of zero-count observations was equal among all of the batches. The second step adjusted the log expression magnitudes of nonzero observations so that the weighted means in each were equal to the population-wide weighted mean. To further control for batch-specific effects, we performed weighted PCA using batch-specific centering (i.e., setting the 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 E13.5 mouse were downloaded from the Allen Developing Mouse Brain Atlas (http://developingmouse.brain-map.org) for all available genes (n = 2,194). ISH data were quantified as gene expression 'energies', defined as expression intensity times expression density, at a grid voxel level. Each voxel corresponded to a 100-µm gridding of the original ISH stain images and to voxel-level structure annotations according to the accompanying developmental reference atlas ontology. The 3D reference model for the developing E13.5 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. We removed background signal and expression detection in regions not annotated as part of the mouse embryo in the reference model by applying a minimum gene energy-level threshold of 8 units. We focused on spatial placements in 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, such as those presented by Satija *et al.*[38] and Achim *et al.*[39], the current method is focused on the relative placement of mutually exclusive subpopulations. Because of this, the user is 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 maturing NPCs relative to early NPCs can be used as indicators of where the maturing NPC subpopulation is spatially localized. In addition, genes downregulated in maturing NPCs relative to early NPCs can be used as indicators of where maturing NPCs may be absent. Additionally, unlike Satija *et al.*[38], we did not binarize the *in situ* data, because we were particularly interested in gradients of expression across voxels or bins in our particular case. Likewise, because of the resolution limitations of our *in situ* data, where each voxel is much bigger than one cell, we were unable to precisely map individual cells to single locations as in Achim *et al.*'s method[39].

Implementation and data availability. The PAGODA functions are implemented in version 1.99 of the *scde* R package, available at http://pklab.med.harvard.edu/scde/, and BioConductor. 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*, and is available from GitHub (https://github.com/hms-dbmi/brainmapr).

31. Anders, S. & Huber, W. *Genome Biol.* **11**, R106 (2010).
32. Fisher, R.A. *Statistical Methods for Research Workers* (Hafner, 1970).
33. Abdel, H.E. *Encyclopedia of Environmetrics* 2nd edn (Wiley, 2012).
34. Hasings, C., Mosteller, F., Tukey, J.W. & Winsor, C.P. *Ann. Math. Stat.* **18**, 413–426 (1974).
35. Bailey, S. *Publ. Astron. Soc. Pac.* **124**, 1023 (2012).
36. Johnstone, I.M. *Ann. Stat.* **29**, 295–327 (2001).
37. Benjamini, Y. & Hochberg, Y. *J. R. Stat. Soc. Series B Stat. Methodol.* **57**, 289–300 (1995).
38. Satija, R., Farrell, J.A., Gennert, D., Schier, A.F. & Regev, A. *Nat. Biotechnol.* **33**, 495–502 (2015).
39. Achim, K. *et al. Nat. Biotechnol.* **33**, 503–509 (2015).