M1 Quality control and preprocessing

Our method of outlier filtering for single-cell RNA-seq experiments is based on estimating genes expressed at background levels in each sample. Libraries with significantly high background levels are then discarded. If Y_{ij} are the raw fragment counts for the i^{th} sample $(i = 1..M)$, j^{th} gene $(j = 1..N)$, then first we compute $r_j = geomean\left(Y_{1j}, \ldots, Y_{Mj}\right)$ to generate a reference sample (Anders and Huber, 2010). We then sort r_j to compute its order-statistics, $r_{(j)}$. If $\sigma_{(j)} = m$ maps (j), to the index of $r_{(j)}$ in the original list, m, then we use $\sigma_{(j)}$ to re-order the tag counts in each sample as $Y_{i\sigma_{(1)}}, \ldots, Y_{i\sigma_{(N)}}$; i.e. the concomitant statistics of the Y_{ij}'s (Yang, 1977). The cumulative densities of these reordered samples, $F_i(k) = \sum_{j=1}^{k} Y_{i\sigma_{(j)}} / \sum_{j=1}^{N} Y_{i\sigma_{(j)}}$, and that of the reference, $R(k) = \sum_{j=1}^{k} r_{(j)} / \sum_{j=1}^{N} r_{(j)}$, yield "Lorenz-like" curves (Fig. S1A). A sample with low dynamic range (e.g. low rate Poisson noise, Fig. S1A black circles) will have a cumulative density that is nearly a straight line. A library that has deeply sampled its cell's transcriptome will have a deep bend at some threshold, K, separating actively transcribed genes from background noise. We compute K, for the reference sample r_j, as the point at which the difference from $R(k)$ to that of a Poisson noise model (a constant Poisson process with a rate given by the 10th percentile of counts across all samples) is maximized. The statistical significance of differences between individual samples and the reference can then be assessed via a score-test for binomial proportions (comparing $R(K)$ to $F_i(K)$), which we refer to as the *Lorenz statistic*. We filter samples whose background fraction is significantly high, via a threshold on the (Benjamini-Hochberg corrected) q-value, of the Lorenz statistic. In our tests, samples that have a small q-value have low complexity, as measured by Gini-Simpson index (Simpson, 1949), and they have low coverage, as estimated by the Good-Turing statistic (Good, 1953), (Figure S1B). Moreover, the Lorenz statistic correlates with the results live-dead staining (Pearson-correlation 0.7). This procedure also provides estimates of the relative differences in expressed RNA between cells. That is, $\frac{1-F_{i_1}(K)}{1-F_{i_2}(K)}$ estimates the fold-change in expressed RNA between cell i_1 and i_2.

The Good-Turing model of sample coverage, which dates back to the deciphering of the Enigma code (Good, 1953), still rates remarkably well and is widely used (Orlitsky *et al.*, 2003). More recently, the PRESEQ algorithm implemented a Padé series approximation to that model that has enabled even more accurate extrapolations (Daley and Smith, 2014). SCell reports library coverage estimates by comparing the number of genes sampled to the PRESEQ asymptotic estimate of genes in the library. SCell also reports the slope of the PRESEQ extrapolation curve as an estimate of marginal return, which can be used to prioritize libraries for re-sequencing.

M2 Normalization and feature selection

Single-cell data analysis depends critically on how samples are normalized; technical noise and cell cycle effects can be confounders to community detection and clustering (Buettner *et al.*, 2015). Generalized linear models have been used for some time to control for heterogeneity in gene expression measurements due to confounding factors (Leek and Storey, 2007). More recently, this approach (termed remove unwanted variation or RUV) was extended to RNA-seq data, using ERCC read-counts for the noise term (Risso *et al.*, 2014). We extend that approach here to normalize single-cell data for dimensionality reduction and clustering. While RUV utilizes ordinary least squares regression to produce normalized counts, we implement weighted least squares to accommodate the heteroskedasticity found in single-cell data. To control for the effect of outliers in weight estimation, introduced via gene-level dropouts, we implement a robust variant: iteratively reweighted least-squares with a bisquare weight function. This approach allows SCell to produce counts normalized by any combination of: 1. Cyclins and cyclin-dependent kinases (a model of cell cycle state), and 2. a user supplied count matrix (enabling an arbitrary set of controls). To avoid over-fitting, SCell uses the first few singular vectors of the ERCC/Cyclin/user-defined count matrices to represent their span. The number singular vectors is chosen by the stick-breaking criterion (Cangelosi and Goriely, 2007).

As a complementary analysis, SCell utilizes canonical correlation analysis (CCA) to measure the effect of cell-cycle on gene expression. CCA measures correlations between two sets of multivariate measurements on the same set of samples. Given two standardized matrices $X = [x_{ij}]$ and $Y = [y_{ik}]$, $i = 1 \dots M$, $j = 1 \dots p$ and $k = 1 \dots q$, CCA solves the optimization problem $\max_{\|a_h\|=1, \|b_h\|=1} corr(Xa_h, Yb_h)$. Here $a_h = [a_{ih}]$ and $b_h = [b_{ih}]$, for $h = 1 \dots H$, are column vectors (Hotelling, 1936). We implement optional robust estimators: Spearman-rank correlation, percentage-bend correlation (Wilcox, 1994) and skipped-correlations (Wilcox, 2004). Intuitively the goal is to find a subspace of the intersection of the column-spans of X and Y, on which they are maximally correlated. Given M cells, we substitute for X, the $M \times p$ matrix of read-counts for cyclins and CDKs. For Y, we use the $M \times q$ matrix of non-cyclin/CDK gene expression values. Cyclin/CDK contribution to variance in gene expression can be estimated, by averaging Pearson correlations between the columns of the cyclin/CDK expression matrix X and the canonical factors of non-cyclin/CDK gene expression: Yb_h (Tenenhaus, 1998). Similarly, genes correlated to cyclin/CDK expression can be obtained by Pearson correlations of the columns of Y with the factors Xa_h (Fig. S2).

Lastly, SCell provides statistics for choosing a gene panel for downstream dimensionality reduction. If a gene's read-count is zero across a large number of cells then we may view that gene as being under-sampled, and we may want to discard that gene from inclusion in the gene panel. However, the observed number of zeroes may be expected if the gene has a low average expression and/or a high variance. Thus, we would like a statistic that can distinguish under-sampled genes from genes which are simply not expressed in a large number of cells. For that purpose, we use a score statistic derived from a generalized Poisson (GP) model, to test for zero-inflation (Yang *et al.*, 2010). The null hypothesis is that the read-counts for a given gene were drawn from a GP distribution, and the alternate hypothesis is that they were drawn from a zero-inflated GP distribution. Since the GP distribution itself accommodates over-dispersion (Consul and Jain, 1973), this approach allows us to delineate the contribution of zero counts to the observed variance. By placing a threshold on the Bengamini-Hodgeberg corrected q-values from this test, we can eliminate under-sampled genes from consideration. Secondly, we would like to identify genes with a high variance across cells, to adequately separate disparate samples in PCA space. For this we use the index of dispersion, $D = \frac{\sigma^2}{\mu}$. The index of dispersion can be used to test the null hypothesis that the gene's read-counts, across cells, were drawn from a common Poisson distribution; and, this test has a closed-form expression for its power function (Selby, 1965). SCell implements an interactive viewer to visualize gene variance vs. sampling (Fig. S3), and to select genes for downstream analysis based on thresholds on these statistics, and their false discovery rates (FDRs).

M3 Dimensionality reduction, clustering and expression kinetics

SCell implements PCA for dimensionality reduction. Two interactive windows allow the user to explore samples in PCA space, with gene-level and sample-level metadata displayed in an interactive window upon mouse-over. Genes and samples can be selected to be added to user-defined gene and sample lists. PCA can be recomputed at any time from the user's working sample list. This permits "iterative" PCA analysis, where a cluster identifying a particular cell type can then be analyzed separately via PCA to learn its sub-populations. SCell implements several methods for clustering, including: k-means, Minkowski-weighted k-means (de Amorim, 2012), clustering via a Gaussian mixture model (Friedman and Rubin, 1967), the clustering "with scatter" algorithm DBSCAN (Ester *et al.*, 1996) and user-defined clusters. Additionally, rotations of the principle components (or more generally rotations in factor analysis) can be used to improve interpretability of feature loadings; we implement Vari-max rotation, that a user may optionally apply post-PCA (Kaiser, 1958).

Recently, the Magwene lineage tracing algorithm, designed for microarray data (Magwene *et al.*, 2003), was re-implemented on single-cell RNA-seq data (Trapnell *et al.*, 2014), to infer lineage trajectories. This method is based on Menger's Theorem: the Traveling Salesman Path (TSP), *of points sampled* *from a smooth curve*, converges to that curve as the number of samples increases (Giesen, 1999). Importantly, in this limit the TSP preserves the order of samples along the original curve (Giesen, 1999). The diameter of a Minimum Spanning Tree (MST) also enjoys this property (Figueiredo and Gomes, 1994). A limitation of this approach, however, is that in practice samples are not derived from a smooth curve, but rather are distributed stochastically about some unknown trajectory. A recent application of a MST approach, to k-nearest neighbor graphs of mass cytometry data, acknowledged it was prone to “short circuits” (see Supplementary Information of Bendall *et al.*, 2014). The Magwene algorithm itself uses P-Q trees to cluster cells, post MST construction, to summarize noisy branches. However, cell-cycle state and technical noise are significant contributors to variation in the data. Fluctuations due to cell cycle may dominate the TSP/MST, if the sample of cells is not sufficiently large or complex. Moreover, forcing a MST to pass through each individual cell may not be realistic when the cells considered are derived from different samples and not the daughters of one another. None the less, we would like to reconstruct a “typical” trajectory, representing gene expression of an average cell and its daughters as it progresses through lineage commitment. For this purpose we construct paths, not on individual cells but on cluster centroids. SCell gives the user the option of fitting either a MST to the full set of cluster centroids, or fitting a minimum-distance path (MDP) between two user-defined centroids within the Gabriel Proximity Graph (GPG) of the data. The GPG is defined so that each cell’s PCA coordinates define a vertex. There is an edge between cell i and cell j, if $d_{ij}^2 < d_{ik}^2 + d_{kj}^2$, for all k not equal to i or j; where d_{ij} is the Euclidean distance from cell i to cell j, in PCA space. GPGs are frequently used to model the backbone of ad-hoc networks (Chiwewe and Hancke, 2012), and for other machine-learning applications (Torres *et al.*, 2012; Fragkiadaki, 2012; Choo *et al.*, 2007). Notably, the GPG contains the Euclidean MST as a subgraph.

Having constructed a path of interest in PCA space, we are still left with the challenge of summarizing gene expression along that trajectory. Rather than summarizing gene expression by projecting cells onto the MST or MDP, we take the alternate approach of directly regressing gene expression on PCA coordinates. This allows us to estimate gene expression along an arbitrary trajectory. SCell provides options for automatically fitting loess/lowess regressions, as well as several interpolation algorithms (linear, cubic spline, biharmonic and thin-plate spline) to construct estimates of gene expression in PCA space and evaluate gene expression along a MST or MDP (Fig. S4).

M4 Experimental Procedures

Tissue samples

De-identified fetal cortical tissue was collected with prior patient consent in strict observance of the legal and institutional ethical regulations from elective pregnancy termination specimens at San Francisco General Hospital. Protocols were approved by the Human Gamete, Embryo and Stem Cell Research Committee (institutional review board) at the University of California, San Francisco. GW16/18 samples were embedded in 3.5% low melting point agarose (Fisher) and sectioned using a Leica VT1200S vibrating blade microtome in artificial cerebrospinal fluid containing 125 mM NaCl, 2.5 mM KCl, 1 mM MgCl2, 1 mM CaCl2, 1.25 mM NaH2PO4. Tissue was dissociated in a pre-warmed solution of Papain and 2000 units/mL of DNase freshly diluted in Earl’s Balanced Salt Solution according to manufacturer’s instructions (Worthington Biochem. Corp.). The resulting single cell suspensions were incubated at 37° C for 20-30 minutes and centrifuged for 5 minutes at 300g. After removing the Papain/DNaseI supernatant, tissue was re-suspended in 0.5 mL of sterile Dulbecco`s Phosphate Buffered Saline (DPBS) containing 3% FBS (Sigma) and 1000 units of DNAse and manually triturated by pipetting up and down approximately 10 times. The suspension was passed through a 40 μm strainer cap (BD Falcon) to yield a uniform single cell suspension.

Single Cell Capture and Library Preparation

The capture of single cells, generation of cDNA, and preparation of sequencing libraries was performed as previously described (Pollen *et al.*, 2014). Briefly, cells were captured using the C1 TM Single-Cell Auto Prep Integrated Fluidic Circuit (IFC) following the methods described in protocol PN 100-7168, http://www.fluidigm.com/. The PCR thermal protocol was adapted from (Fan *et al.*, 2012) using the SMARTer® Ultra Low RNA Kit (Cat. No. 63495, PT5163-1). cDNA products were quantified using high sensitivity DNA chips (Agilent) and then diluted to a final concentration of 0.15–0.30 ng/μL using C1 TM Harvest Reagent. The Nextera® XT DNA Sample Preparation Kit (Illumina) was then used to convert diluted cDNA reaction products into sequencing libraries following manufacturer's instructions, with minor modifications. Reactions were run at one quarter of the recommended volume, the tagmentation step was extended to 10 minutes, and the extension time during PCR was increased from 30 seconds to 60 seconds. After the PCR step, samples were pooled, cleaned twice with 0.9X Agencourt AMPure XP SPRI beads (Beckman Coulter), eluted in DNA suspension buffer (Teknova) or EB buffer (Qiagen) buffer and quantified using High Sensitivity DNA Chip (Agilent).

Alignment of RNA sequencing data

An average of 2.9 million 100bp, paired-end reads were generated per library. Using cutadapt under the Trim Galore! wrapper with the default settings, reads were trimmed for quality, and Nextera transposase sequences were removed. Reads shorter than 20bp were discarded. Read level quality control was then assessed using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Reads were aligned to the NCBI human reference sequence GRCh37, by Tophat2 (Kim et al., 2013) using the --prefilter-multihits option and a guided alignment via the NCBI RefSeq transcriptome reference. Expression for RefSeq genes was quantified by the featureCounts routine, in the subRead library (Liao *et al.*, 2013), using only uniquely mapping reads and discarding chimeric fragments and unpaired reads.

Supplementary Figure Legends

Fig. S1. (A) The Lorenz statistic estimates the fraction of genes expressed at background levels, across all samples. By comparing a reference sample to a Poisson noise simulation, samples with a high level of background can be triaged. (B) Cells triaged by a Lorenz statistic (QC pass is $q<0.05$, QC fail otherwise); top panel: the reciprocal of Simpson diversity, bottom panel: PRESEQ coverage estimate. (C) Heat-map illustrating the percentage of genes expressed above a given expression quantile. Quantiles of expression are computed across samples.

Fig. S2. (A) Average correlation between the cyclin/CDK given on the x-axis and the canonical factors of the gene expression matrix. (B) Average correlation between the gene given on the x-axis and the canonical factors of the cyclin/CDK expression matrix.

Fig. S3. (A) The gene panel selection tool. Genes are ordered by the percentage of cells that express them on the x-axis. Log index of dispersion percentile ranks genes by variability, on the y-axis. (B) A 2D lowess regression fit of PAX6 expression.

Fig. S4. (A) A k-means clustering in PCA space. Top panel: the purple cluster is enriched for PAX6, the magenta cluster for EOMES and the orange cluster for NEUROD6. This indicates that these cells likely come from the excitatory lineage. Middle panel: an iterative PCA of the purple, magenta and orange clusters from the top panel. A Gabriel graph minimum cost path is indicated in black. Bottom panel: an iterative PCA of the green cluster from the top panel, enriched for markers of inhibitory neurons. (B) Gene expression for markers of the excitatory neuronal lineage (PAX6, EOMES, NEUROD6), and the inhibitory neuronal lineage (DLX6, DLX6-AS1, GAD1), plotted over the PCA in the top panel of (A). (C) Genes characterizing varied classes of inhibitory neurons, plotted over the PCA in the bottom panel of (A).

References

de Amorim,R.C. (2012) Constrained clustering with Minkowski Weighted K-Means. *2012 IEEE 13th Int. Symp. Comput. Intell. Informatics*, 13–17.

Anders,S. and Huber,W. (2010) Differential expression analysis for sequence count data. *Genome Biol.*, **11**, R106.

Bendall,S.C. *et al.* (2014) Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. *Cell*, **157**, 714–25.

Buettner,F. *et al.* (2015) Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat. Biotechnol.*, **33**.

Cangelosi,R. and Goriely,A. (2007) Component retention in principal component analysis with application to cDNA microarray data. *Biol. Direct*, **2**, 2.

Chiwewe,T. and Hancke,G. (2012) A Distributed Topology Control Technique for Low Interference and Energy Efficiency in Wireless Sensor Networks. *Ind. Informatics, IEEE …*, **8**, 11–19.

Choo,J. *et al.* (2007) MOSAIC: A Proximity Graph Approach for Agglomerative Clustering. In, Song,I. *et al.* (eds), *Data Warehousing and Knowledge Discovery SE - 21*, Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 231–240.

Consul,P.C. and Jain,G.C. (1973) A Generalization of the Poisson Distribution. *Technometrics*, **15**, 791–799.

Daley,T. and Smith,A.D. (2014) Modeling genome coverage in single cell sequencing. *Bioinformatics*, **30**, 1–7.

Ester,M. *et al.* (1996) A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *Second Int. Conf. Knowl. Discov. Data Min.*, 226–231.

Fan,J.-B. *et al.* (2012) Highly Parallel Genome-Wide Expression Analysis of Single Mammalian Cells. *PLoS One*, **7**, e30794.

Figueiredo,L. De and Gomes,J.D.M. (1994) Computational morphology of curves. *Vis. Comput.*, 1–7.

Fragkiadaki,K. (2012) Video segmentation by tracing discontinuities in a trajectory embedding. *2012 IEEE Conf. Comput. Vis. Pattern Recognit.*, 1846–1853.

Friedman,H.P. and Rubin,J. (1967) On Some Invariant Criteria for Grouping Data. *J. Am. Stat. Assoc.*, **62**, 1159–1178.

Giesen,J. (1999) Curve Reconstruction, the Traveling Salesman problem and menger's Theorem on Length. *Proc. fifteenth Annu. Symp. …*, 207–216.

Good,I. (1953) The Population Frequencies of Species and the Estimation of Population Parameters. *Biometrika*, **40**, 237–264.

Hotelling,H. (1936) Relations between two sets of variates. *Biometrika*.

Kaiser,H.F. (1958) The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, **23**, 187–200.

Leek,J.T. and Storey,J.D. (2007) Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.*, **3**, 1724–35.

Liao,Y. *et al.* (2013) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 1–8.

Magwene,P.M. *et al.* (2003) Reconstructing the temporal ordering of biological samples using microarray data. *Bioinformatics*, **19**, 842–850.

Orlitsky,A. *et al.* (2003) Always Good Turing: asymptotically optimal probability estimation. *Science*, **302**, 427–431.

Pollen,A.A. *et al.* (2014) Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nat Biotech*, **advance on**.

Risso,D. *et al.* (2014) Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat. Biotechnol.*, **32**.

Selby,B. (1965) The index of dispersion as a test statistic. *Biometrika*, **52**, 627–629.

Simpson,E.H. (1949) Measurement of Diversity. *Nature*, **163**, 688–688.

Tenenhaus,M. (1998) La regression PLS: theorie et pratique.

Torres,L.B. *et al.* (2012) A Computational Geometry Approach for Pareto-Optimal Selection of Neural Networks. In, Villa,A.P. *et al.* (eds), *Artificial Neural Networks and Machine Learning – ICANN 2012 SE - 13*, Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 100–107.

Trapnell,C. *et al.* (2014) The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.*

Wilcox,R. (2004) Inferences Based on a Skipped Correlation Coefficient. *J. Appl. Stat.*, **31**, 131–143.

Wilcox,R.R. (1994) The percentage bend correlation coefficient. *Psychometrika*, **59**, 601–616.

Yang,S. (1977) General Distribution Theory of the Concomitants of Order Statistics Author. *Ann. Stat.*, **5**, 996–1002.

Yang,Z. *et al.* (2010) Score Tests for Zero-Inflation in Overdispersed Count Data. *Commun. Stat. - Theory Methods*, **39**, 2008–2030.