HHS Public Access

Author manuscript

Nat Methods. Author manuscript; available in PMC 2017 July 23.

Published in final edited form as:

Nat Methods. 2017 March ; 14(3): 309–315. doi:10.1038/nmeth.4150.

Single-cell mRNA quantification and differential analysis with Census

Xiaojie Qiu[1,2], **Andrew Hill**[1], **Jonathan Packer**[1], **Dejun Lin**[1], **Yi-An Ma**[3], and **Cole Trapnell**[1,2,*]

[1]Department of Genome Sciences, University of Washington, Seattle, WA, 98195, USA

[2]Molecular & Cellular Biology Program, University of Washington, Seattle, WA, 98195, USA

[3]Department of Applied Mathematics, University of Washington, Seattle, WA 98195, USA

Abstract

Single-cell gene expression studies promise to unveil rare cell types and cryptic states in development and disease through a stunningly high-resolution view of gene regulation. However, measurements from single-cell RNA-Seq are highly variable, frustrating efforts to assay how expression differs between cells. We introduce Census, an algorithm available through our single-cell analysis toolkit Monocle 2, which converts relative RNA-Seq expression levels into relative transcript counts without the need for experimental spike-in controls. We show that analyzing changes in relative transcript counts leads to dramatic improvements in accuracy compared to normalized read counts and enables new statistical tests for identifying developmentally regulated genes. We explore the power of Census through reanalysis of single-cell studies in several developmental and disease contexts. Census counts can be analyzed with widely used regression techniques to reveal changes in cell fate-dependent gene expression, splicing patterns, and allelic imbalances, demonstrating that Census enables robust single-cell analysis at multiple layers of gene regulation.

Introduction

Differential gene expression analysis, typically powered by statistical regression, is central to nearly all single-cell transcriptomic studies. As experiments now capture tens of thousands of cells[1,2], such regressions could in principle be used to detect gene regulatory changes across individual cells as a function of developmental progression, position in an embryo, or genetic sequence. However, they report measurements with high variability, frustrating efforts to build models that can detect such changes[3,4]. Numerous studies have reported high rates of “drop-out”, wherein some cells of a nominally homogeneous

*Corresponding author: coletrap@uw.edu.

Author Contributions

X.Q. and C.T. designed Census and the regression methods. X.Q. implemented the methods. X.Q. and A.H performed the analysis. J.P, D.L, and Y.M contributed to technical design. C.T. conceived the project. All authors wrote the manuscript.

Competing Financial Interest Statement

The authors declare no relevant financial interests.

population express high levels of a gene and others none at all. Drop-outs have spurred the deployment of hurdle models[5] that overcome limitations over simpler regression approaches, typically at a cost in speed, numerical stability, or design flexibility for the user.

Single-cell protocols that use exogenous RNA "spike-in" standards[6] or unique molecular identifiers[7,8] (UMIs) enable analysis to be performed at the level of transcript counts rather than read counts. Previous work by Grun *et al.* suggested that comparing UMIs, rather than read counts, between cells would improve regression analysis. However, because UMI protocols work by counting 3' end tags, they are limited to measuring gene expression and do not report expression at allele- or isoform-resolution. Spike-in-based protocols, which convert a cell's relative abundances to transcript counts through a linear regression between the spikes' normalized read counts and their known molecular concentrations, can report measurements at this resolution. However, exogenous standards must be carefully calibrated for single-cell experiments lest they dominate the libraries, and may be subject to different rates of degradation or reverse transcription than endogenous RNA. Many published studies have chosen to forgo the use of spike-in controls, restricting subsequent reanalysis.

Here, we introduce Census, an algorithm that converts conventional measures of relative expression such as transcript per million (TPM) in single cells to relative transcript counts without the need for spike-in standards or UMIs. "Census counts" eliminate much of the apparent technical variability in single-cell experiments and are thus easier to model with standard regression techniques than normalized read counts. We demonstrate the power of transcript count analysis with a new regression model, BEAM (Branch Expression Analysis Modeling), for detecting genes that change following fate decisions in development. We also analyze Census counts at the splice isoform and allele level, demonstrating that our approach robustly detects developmental regulation at those resolutions. Census and BEAM are implemented in Monocle 2, the second major release of our open-source single-cell analysis toolkit.

Results

Estimating relative transcript counts in spike-in-free experiments

Census exploits two properties of single-cell RNA-Seq datasets produced with current protocols (Figure 1a). First, mRNA degradation following cell lysis and inefficiencies in the reverse transcription reaction result in the capture of as few as 10% of the transcripts in a cell as cDNA. Second, most protocols rely on template-switching reverse transcriptases primed at the polyA tail of mRNAs and thus generate full-length cDNAs[9]. Such protocols typically generate libraries in which genes are detected most frequently as a single cDNA molecule (Figure 1b, Supplementary Figure 1). Thus, all detectably expressed genes measured at or below the mode of the (log-transformed) relative abundance distribution in each cell should be present at around 1 cDNA copy (see **Methods**).

We assessed Census' accuracy by re-analyzing several experiments that included spike-in controls[4,10,11,12,13,14,15]. Reanalysis of developing lung epithelial cells with Census recovered estimates of total per-cell transcript counts that were correlated with but not equal to those derived by linear regression against spike-in controls (Figure 1c), likely because of

Census' inability to control for non-linear cDNA amplification during library construction. However, changes in Census counts between groups of cells collected at the same time points were highly similar to changes measured via spike-in controls (Figure 1d,e). Census produced accurate changes in relative transcript counts for seven additional datasets, including two based on UMIs[4,10,13], demonstrating that the algorithm can work well with different single-cell RNA-Seq protocols. (Supplementary Figure 1,2). Downsampling and simulation experiments determined that Census counts faithfully capture changes in expression between groups of cells with as few as 100,000 reads per cell and over a wide range of mRNA capture rates (Supplementary Figure 3, 4). Taken together, these benchmarking experiments show that Census recovers an accurate measure of changes in relative transcript counts between single cells without the need for spike-in controls.

Census counts improve differential analysis accuracy

We next assessed whether using Census counts improved downstream differential analysis. We tested several popular tools[16,17] for differential expression with both read counts and relative transcript counts, including two tools specifically developed for single-cell data, Monocle[18], and SCDE[19] (Figure 2a, Supplementary Figure 5). When provided with read counts as a measure of expression, consensus between the tools was poor, with only 1,971 of 5,805 (34%) differentially expressed genes (DEGs) reported by all tools (except SCDE, which has very high precision but low recall), and few agreed with those reported by a nonparametric, permutation-based test between spike-in derived expression levels (Figure 2b, Methods). Tools designed for bulk RNA-Seq analysis, such as DESeq2[17], produce false discovery rates as high as 61%. SCDE, which includes explicit modeling of drop-outs returned few false positives but also captured a smaller fraction of the true positive set.

Repeating these tests using Census counts showed marked improvements in differential expression accuracy compared to read counts and TPM (Figure 2a). We attribute the improvements to the fact that the negative binomial distribution, which underlies most commonly used RNA-Seq analysis software[16,18,19], fits relative transcript count data much better than read count data, as noted by Grun *et al.*[4] (Supplementary Figure 5). For example, when targeting a false discovery rate of 10%, DESeq2's empirical false discovery rate dropped dramatically from 61% to 22% with little to no drop in sensitivity, which remains as high as 82%. Monocle's false discovery rate dropped from 53% to 11%. Importantly, using Census counts dramatically improved agreement between the tools, which all agreed on 2,437 DEGs among a total of 4,220 (70%), similar to the 62% (2,367 / 3,793) consensus genes obtained with spike-in derived levels (Figure 2b). Census also improved DE accuracy relative to gold standards derived from bulk RNA-Seq[18] measurements (Supplementary Figure 6). Taken together, our benchmarks demonstrate that single-cell relative transcript counts produced by Census can be more accurately compared with commonly used differential analysis methods than normalized read counts, and are thus preferable when spike-in standards or UMIs are unavailable.

Differential analysis of branch points in developmental trajectories reveals regulators of cell fate

Many single-cell gene expression studies aim to identify gene regulatory circuits that control cell-fate decisions made during development[20,21]. We recently developed Monocle, an algorithm that organizes single cells along trajectories and can describe the gene expression changes executed during cell differentiation. Monocle introduced the concept of "pseudotime", which quantifies each cell's progress through development. Pseudotime resolves cascades of gene regulatory changes that accompany differentiation and other dynamic cellular programs[18]. Monocle produces more reliable tests for differential expression along a trajectory when provided with Census counts than with relative expression values (Supplementary Figure 7).

Single-cell trajectories can have multiple outcomes, such as during the generation of alternative developmental lineages[22]. Analyzing cells at branch points where cells are diverted along two or more mutually exclusive paths could reveal the mechanisms by which such decisions are made. For example, scrutinizing genes upregulated in common myeloid progenitors but downregulated in common lymphoid progenitors has shed light on the molecular regulation of cell fate in hematopoiesis[23,24].

To explore a developmental fate decision at single-cell resolution, we reanalyzed RNA-seq data from a recent study investigating the specification of the distal lung epithelium[25]. Treutlein *et al.* sequenced developing epithelial cells to define the cellular intermediates giving rise to type I (AT1) and type II (AT2) pneumocytes. Monocle reconstructed a trajectory with a single branch point leading from progenitors to two outcomes corresponding to the AT1 and AT2 fates. The beginning of the trajectory contained cells with high levels of markers of active proliferation[26] (*Ccnb2, Cdk1*), whereas these genes were expressed at much lower levels after the branch point (Figure 3a). High expression of a known marker of AT1 cells[27] (*Pdpn*) was restricted to cells on one branch of the tree, whereas cells expressing an AT2 marker[28] (*Sftpb*) at high levels were located on the other branch. Cells classified as AT1 and AT2 according to known markers by Treutlein *et al.* fell exclusively along the branches, with what the authors termed "bipotent progenitors (BP)" at or near the branch point. (Supplementary Figure 8).

To detect cell fate-dependent genes in a statistically robust manner, we developed BEAM, a generalized linear modeling (GLM) [29] strategy for analyzing branched single-cell trajectories (Figure 3b; Supplementary Figure 9, see Methods). BEAM identified 1,219 genes (FDR < 5%) as either AT1- or AT2- fate dependent, including canonical markers[27] such as as *Pdpn* and *Sftpb* (Figure 3c). AT1-restricted genes were strongly enriched for ontological terms related to tube development, cytoskeletal remodeling, and cell morphogenesis (Supplementary Figure 10, Supplementary Table 1), while AT2-restricted genes were enriched for terms related to lipid processing, consistent with the production of lipid-rich surfactant by AT2 cells in the mature lung. Regulatory DNA elements proximal to these genes were enriched for binding sites of 74 transcription factors, eleven of which exhibited significant branch-dependent expression. Supplementary Figure 11) These factors included several such as Tcf7l2 that are well known to regulate lung development[30-35].

Disruption of interferon signaling induces a branch in the dendritic cell LPS stimulation trajectory

Branch points in single-cell trajectories represent steps in a program of transcriptional change in which cells must choose between one of several mutually exclusive gene expression programs. Branches could arise not only during development, but also in response to mutations, treatment with drugs, or other cellular perturbations. We reanalyzed a recent study[36] from Shalek and colleagues, which dissected the transcriptional response of murine bone marrow-derived dendritic cells (BMDCs) to lipopolysaccharide (LPS) (Figure 4a). In BMDCs, LPS triggers a paracrine feedback loop of type I interferon signaling mediated in part by *Stat1*[37-39]. The authors compared BMDCs from wild-type (WT) mice to those from mice that lack the receptor for *Interferon alpha* (*Ifnar1*-/-) or *Stat1* (*Stat1*-/-). Monocle recovered a trajectory with a single branch point, with cells from *Infar*-/- or Stat1-/- mice distributed on an alternative trajectory in response to LPS stimulation compared with those from WT mice (Figure 4b).

BEAM identified 870 genes (FDR < 5%), many associated with interferon signaling, dependent on this branch (Figure 4c, Supplementary Figure 12). Peaks corresponding to open chromatin collected by Lavin et al[40] proximal to branch-dependent genes are enriched for *Stat1/2* and *Irf1*|*2* binding motifs (Supplementary Figure 13). These factors were themselves significantly branch-dependent, with branching pseudotimes substantially earlier than their putative targets, confirming that BEAM can distinguish the regulatory factors that drive branching in single-cell trajectories from genes downstream (Figure 4e, f). Monocle 2 and BEAM demonstrated that loss of a key paracrine loop generates an “alternative trajectory”, suggesting that single-cell trajectory analysis can be useful for defining how a signaling pathway regulates a larger process.

Census counts enable single-cell differential splicing analysis

Methods for detecting splicing changes in single-cell RNA-Seq experiments are beginning to appear, but have grappled with isoform-level measurement variability. For example, Welch *et al.* described SingleSplice[41], which uses a hurdle model to compare observed variation in isoform frequencies against expected technical variation, but its contrasts are limited to tests for excess variability within groups of cells, rather than as a function of arbitrary variables in a regression, and it requires calibration with spike-in standards.

We used Census to estimate isoform-level transcript counts in differentiating myoblasts, a classic model system for vertebrate splicing. Modeling isoform counts from each gene as a Dirichlet-multinomial distribution captured pseudotime-dependent shifts in splicing in 74 genes (FDR < 0.1), including well-characterized components of the molecular machinery required for muscle contraction such as tropomyosin *TPM1,* which has been intensely studied in myoblasts as a model of alternative splicing[42,43] (Figure 5). *TPM1* has three well-characterized sets of alternatively spliced exons, with exons 6b and 9b excluded in myoblasts but included in myotubes[44]. These exons became progressively more frequent in *TPM1* mRNAs, with inclusion of exon 6b preceding inclusion of exon 9b. Each isoform of the 74 differentially spliced genes showed one of seven distinct pseudotemporal expression patterns, (Supplementary Figure 14a, b) coinciding with shifts in the actin family from

widely expressed members (*ACTB, ACGT*) partly replaced with muscle specific ones (*ACTA1, ACTA2*). (Supplementary Figure 14c). Our analysis supports the view that cytoskeletal reorganization during myoblast differentiation is globally coordinated not only at the level of genes but across individual splice variants.

Census counts enable allelic balance analysis in single cells

Single-cell analysis could in principle shed light on the degree to which the two alleles of each gene are regulated in a coordinated manner. Recently, Deng *et al.* tracked gene expression genome wide in single-cells from pre-implantation mouse embryos of mixed genetic background (CAST/EiJ × C57BL/6J) [45]. Coupling allele-level relative abundances from Kallisto[46] with Census produced relative allele transcript counts which, when modeled similarly to isoform counts, recapitulated many of the key observations made in the initial study. As expected, nearly all RNAs matched the maternal allele in zygotes and early 2-cell embyros, consistent with little to no transcription from the embryonic genome (Figure 6a). Allelic balance for most genes equilibrated to 50% as transcription from the embryonic genome began in mid- to late- 2-cell embryos, with the X chromosome notably excepted (Figure 6a). Inactivation of the paternal X chromosome in female embryos was manifest by the 16-cell stage, with progressively fewer genes exhibiting contributions from the paternal X (Figure 6b, c), although genes known to escape inactivation were notably excepted (Supplementary Figure 15).

In addition to pre-implantation allelic dynamics, Deng *et al.* reported widespread stochastic monoallelic gene expression in individual cells. This claim has been challenged by Kim *et al.* [47], who analyzed allele-specific expression in embryonic stem cells using a statistical model that attributed much of the apparent stochastic monoallelic expression to technical sources. We tested whether using Census to estimate allelic transcript counts instead of allelic read counts would reduce the apparent stochastic monoallelic expression to expected levels. Consistent with the generative model used by Kim *et al.*, the expected rate of monoallelic expression was near 100% for genes expressed at a single copy, and decreased with increasing expression (Figure 6e). Of 6,608 “allele-informative” genes in the genome, 95.0% produced observed monoallelic transcript counts within the expected range. In contrast, only 77% of genes fell within the range obtained by fitting similar models to normalized read counts for each allele. We interpret this to mean that a substantial portion of apparent monoallelic expression arose because the sequenced libraries correspond to a small proportion of the true RNA molecules in each cell (due to dropout), a technical artifact that is accounted for when allelic gene expression is modeled using Census-estimated relative transcript counts but not when it is modeled using normalized read counts.

Discussion

Efforts to detect changes in gene regulation in development have grappled with high technical and biological variability, demanding specialized statistical methods that explicitly model drop-outs and other nuisance variation. Here, we show that analyzing changes in relative transcript counts leads to dramatic reductions in apparent technical variability compared to normalized read counts, making single-cell RNA-Seq compatible with widely

used regression techniques. We have developed Census, a normalization algorithm that can convert relative expression levels from read counts into per-cell transcript counts without the need for spike-in standards or UMIs. The algorithm requires only that genes are most frequently present at 1 cDNA molecule in each cell's library. We show through reanalysis of several datasets that this is the case with most current protocols, owing to mRNA capture rates lower than 50% and their generation of full-length cDNAs during reverse transcription. Census cannot control for amplification biases, and thus does not produce estimates of lysate mRNA abundances that perfectly match those derived with spike-ins or UMIs. When spike-ins or UMIs are available, transcript counts should be recovered using them rather than Census. However, we show through extensive benchmarking that differential analysis results with Census counts are highly concordant with those from spike-ins. Importantly, tools widely used for bulk RNA-Seq analysis that perform poorly when provided with read counts work vastly better with Census counts, alleviating the need for software tailored for single-cell RNA-Seq.

To illustrate their power, we have developed three regression-based methods for detecting changes in Census counts. The first, BEAM, builds on our previous work tracking gene expression changes in single-cell trajectories, helping pinpoint the moment at which cell-fate decisions occur in a complex biological process. BEAM identified hundreds of genes differentially regulated during specification of the type I and type II pneumocytes in the alveolar epithelium. Surprisingly, branched cell trajectories arise not only in development, but also in response to genetic perturbations, suggesting that branch analysis may be useful in many biological contexts. The second method uses Census counts to find genes undergoing pseudotime-dependent changes in splicing. Reanalysis of differentiating myoblasts showed widespread alteration in isoform ratios in genes involved in muscle contraction and cytoskeletal structure, with some genes such as TPM1 showing a sequence of pseudotime-dependent shifts. The third method captures changes in allelic transcript counts derived with Census. By reanalyzing data from pre-implantation embryos, we confirmed the authors' timing of transcriptional activation of the embryonic genome and X chromosome inactivation. In contrast to the original study, we do not see substantial evidence of random, monoallelic expression on the autosomes, and attribute this observation to inadequate modeling of dropouts in normalized read counts. Monoallelic expression at the transcript count level was in line with expectations under a simple overdispersed binomial regression model.

Together, our analyses show that single-cell differential expression analyses conducted at the level of normalized transcript counts are more robust and accurate than analyses of normalized read counts. We provide a new algorithm, Census, that makes relative transcript count analysis widely accessible, as well as examples of regression models, in particular BEAM, that leverage them for high-resolution dissection of gene regulation. We expect that such techniques will continue to unveil new mechanisms of gene regulation, including at the allele and isoform level, in development and disease.

Data Availability Statement

Code availability

A version of monocle 2 (version: 1.99) used to produce all the figures, supplementary data is provided in Supplementary Software. The newest Monocle 2 is available through Bioconductor as well as GitHub (https://github.com/cole-trapnell-lab/monocle-release). Supplementary Software also includes a helper package including helper functions as well as all analysis code which can reproduce all figures in this study.

Data availability

Eleven public sc RNA-seq datasets are used in this study, of which 8 datasets used ERCC spike-in. Here is a summary list of all the data:

Datasets with spike-in:

Lung: GSE52583[25]

Noise model: GSE54695 [4]

Neuron reprogramming: GSE67310 [15]

Human Preimplantation Embryos: E-MTAB-3929 [10]

Pancreas: E-MTAB-5061[11]

Cortex: http://linnarssonlab.org/cortex/ [12]

Marker-free: GSE54006 [13]

Quantitative assessment data: GSE51254 [14]

Datasets without spike-in:

HSMM: GSE52529 [18]

Dendritic cell knockout: GSE41265 [36]

Allele-specific gene expression: GSE45719 [45]

Online Methods

A generative model for single-cell RNA-seq experiments with a spike-in ladder

Census is motivated by a generative model of single-cell (sc) RNA-Seq similar to the one developed by Kim *et al.*[47]. When performing sc-RNA-seq, each individual cell is lysed to recover its endogenous RNA molecules, some fraction of which may be degraded or lost. Lysis thus involves an RNA recovery rate a. Spike-in transcripts are then added into the cell lysate. Note that spike-in transcripts are added to the lysate as naked RNA, and thus may be

degraded at different rates from the endogenous RNA. We denote the ladder recovery rate as β. The RNA counts in the lysate can be written:

$$Cell\ lysate: \begin{cases} Y^l_{ij} & \approx & \alpha_i Y^c_{ij} \\ S^l_{ij} & \approx & \beta_i S_{\cdot j} \end{cases},$$

where Y^l, S^l, S, are the transcript counts of endogenous RNA in cell lysate, spike-in transcript counts in cell lysate and the spike-in transcript counts added into the cell lysate. The first subscript in all variables (here and below) corresponds to cell while the second subscript corresponds to gene index. Note that we are not able to directly observe Y^c_{ij}, the true transcript counts for gene j in cell i and thus α is an unknown variable.

The RNA molecules and spike-in transcripts will then be subjected to reverse transcription and amplified to make a cDNA library. The expected number of cDNA molecules generated from each RNA molecules is denoted by θ. The cDNA counts can be written:

$$cDNA: \begin{cases} Y^d_{ij} & = & Y^l_{ij} \cdot \theta_i \\ S^d_{ij} & = & S^l_{ij} \cdot \theta_i \end{cases},$$

where Y^d, S^d, are the cDNA counts of endogenous RNA, spike-in cDNA counts successfully converted from the corresponding transcript counts Y^l, S^l in cell lysate under a uniform capture rate θ, which for current protocols is less than 1.

Our model generates sequencing reads from the cDNA. The relative cDNA abundances are calculated as $\frac{Y^d_{ij}}{\sum_{j=1}^{n}\left(Y^d_{ij}+S^d_{ij}\right)}$ for endogenous RNA, or $\frac{S^d_{ij}}{\sum_{j=1}^{n}\left(Y^d_{ij}+S^d_{ij}\right)}$ for spike-in RNA.

The model then generates γ reads per cDNA molecule on average; with sufficient sequencing, γ will be larger than 1; we expect each cDNA molecule to generate at least one sequencing read. This process can be regarded as a multinomial sampling of R reads $\left(R_i{=}\gamma\sum_{j=1}^{n}\left(Y^d_{ij}+S^d_{ij}\right)\right)$ from the distribution of relative cDNA abundances mentioned above which can be represented as:

$$Read\ counts: \begin{cases} Y^r_{ij} \sim multinomial\left(\frac{Y^d_{ij}}{\sum_{j=1}^{n}\left(Y^d_{ij}+S^d_{ij}\right)}, R^e_i\right) \\ S^r_{\cdot j} \sim multinomial\left(\frac{S^d_{ij}}{\sum_{j=1}^{n}\left(Y^d_{ij}+S^d_{ij}\right)}, R^s_i\right) \end{cases},$$

where R^e_i, R^s_i, denotes the reads sampled for cDNA from the endogenous RNA or spike- in RNA in cell i, $Y^r_{ij}, S^r_{\cdot j}$ denotes the reads sampled for cDNA from the endogenous RNA j or spike-in RNA j in cell i.

The model described here is essentially a special case of the model in Kim *et al.*, and differs mainly in that their model describes transcript-level capture rates and sequencing rates with beta and gamma distributions, respectively. In contrast, we simply use global constants for these rates. As Census does not make use of variance estimates from the generative model, this simpler model is sufficient for calculating key statistics (e.g. mode of the transcript counts) needed to convert relative to absolute abundances.

A simulator for the sc-RNA-seq process

To generate an *in silico* library for a single cell, we built a simulator that first selects G genes at random from a relative expression profile (P_{bulk}) derived from a bulk RNA-Seq experiment to represent the hypothetical relative abundance of a single-cell in cell lysate. These values are rescaled to proportions (i.e. summing to 1), or ρ_{scaled}.

$$\rho_{scaled} \sim scale\left(uniform\left(P_{bulk}\left(1,2,\ldots\right),\mathrm{G}\right)\right)$$

These proportions are then used to parameterize a multinomial distribution from which T transcripts are drawn to obtain the transcripts in the library space where we also consider there is α_i percentage of the RNA is degraded. Therefore, we have:

$$library{:}Y_{ij}^{l} \sim multinomial\left(\rho_{scaled},\left(1-\alpha_i\right)T_i\right)$$

To this pool of transcripts, a fixed number of spike-in transcripts are added, forming a mixture of simulated “endogenous” and “spike-in” mRNAs where the degradation of spike-in transcripts is represented by β_i. Of these, θ_i percent are selected uniformly at random to simulate incomplete mRNA capture by the reverse transcription process. Finally, the abundances of these cDNAs relative to one another were used to parameterize another multinomial, from which R_i reads are sampled. The read counts are then used to calculate the relative abundance for the spike-in and the endogenous RNA.

In this study, we systematically simulated the sc RNA-seq process obtained from bulk RNA-Seq measurements made in Trapnell and Cacchiarelli *et al*[18] by varying the gene number G, capture rate θ, endogenous RNA degradation α, spike-in degradation β, total endogenous transcript count T and total number of reads R. Results based on simulation are shown in Supplemental Figure 4.

Estimating the capture rate based on spike-in ladder

Similar to Kim *et al.*[47], spike-in transcripts can be used to infer the rate at which lysate RNAs are converted to cDNA. The probability of observing a particular spike-in transcript in the sequenced read counts can be used to estimate the capture rate θ. For a given spike-in transcript i with transcript counts s calculated using the above procedure, the probability to observe at least one copy of this transcript is $p = 1 – (1 – \theta)^s$. We assume the capture rate, θ, is the same for all spike transcripts and thus can use the following objective function to estimate the capture rate using all spike transcripts:

$$\min_{\theta} \sum_{i} (1 - (1 - \theta)^s - o_i^s)^2$$

where o_i^s is the probability for all transcripts with *s* copies have non-zero TPM values. In order to robustly estimate θ, we assume a constant capture rate for cells collected in each time point (lung or neuron experiment) or the whole dataset (other experiments) and pool them for estimating θ.

Census

Census aims to convert relative abundances X_{ij} into lysate transcript counts Y_{ij}. Without loss of generality, we consider relative abundances is on the TPM scale, and assume that a gene's TPM value is proportional to the relative frequencies of its mRNA within the total pool of mRNA in a given cell's lysate, i.e., $TPM_{ij} \propto \frac{Y_{ij}}{\sum_{j=1} Y_{ij}}$. The generative model discussed above predicts that when only a minority of the transcripts in a cell is captured in the library, signal from most detectably expressed genes will originate from a single mRNA. Because the number of sequencing reads per transcript is proportionate to molecular frequency after normalizing for length (i.e. TPM or FPKM), all such genes in a given cell should have similar TPM values.

Census works by first identifying the (log-transformed) TPM value in each cell *i*, written as x_i^*, that corresponds to genes from which signal originates from a single transcript. Because our generative model predicts that these most detectable genes should fall into this category, we simply estimate x_i^* as the mode of the log-transformed TPM distribution for cell *i*. This mode is obtained by log-transforming the TPM values, performing a Gaussian kernel density estimation and then identifying the peak of the distribution. Given the TPM value for a single transcript in cell *i*, it is straightforward to convert all relative abundances to their lysate transcript counts. We estimate the total number of mRNAs captured for cell *i:*

$$M_i = \frac{1}{\theta} \cdot \frac{n_i}{F_{x_i}(x_i^*) - F_{x_i}(\varepsilon)}$$

where F_x represents the cumulative distribution function for the TPM values for cell *i*, ε is a TPM value below which no mRNA is believed to be present (by default, $\varepsilon = 0.1$), and n_i is the number of genes with TPM values in the interval (ε, x_i^*). That is, we simply calculate the total number of single-mRNA genes and divide this number by the fraction of the library contributed by them to estimate the total number of captured mRNAs in the cell. This number is scaled by $\frac{1}{\theta}$ to yield an estimate for the number of mRNAs that were in the cell's lysate, including those that were not actually captured. This scaling step is performed mainly to facilitate comparison with spike-in derived estimates. While we do not know the capture rate θ *a priori*, it is a highly protocol-dependent quantity that appears to have little

dependence on cell type or state. Throughout our analysis, we assume a value of 0.25, which is close to the lung and neuron experiments of Truetlein et al.

With an estimate of the total lysate mRNAs M_i in cell i, we simply rescale its TPM values into mRNA counts for each gene:

$$\hat{Y}_{ij} = X_{ij} \cdot \frac{M_i}{10^6}$$

Limitations of Census

Census and our generative model of single-cell RNA-Seq assume that TPM is proportional to the true relative abundance in the cell lysis, i.e., $TPM_{ij} \propto \frac{Y_{ij}}{\sum_{j=1} Y_{ij}}$. However, non-linear amplification at any stage of the library construction protocol could distort this relationship. We can see this distortion when fitting the linear regression model, $\log(\text{TPM}_{ij}) = k * \log(Y_{ij}) + b$, to the spike-in data recovers a value of k that deviates from 1, which indicates that $\text{TPM}_{ij} \propto (Y_{ij})^k$. In practice, we find that k ranges from around 0.5 to near 1, depending on the protocol and the laboratory. We have not observed k much larger than 1.

The inability to estimate k without making strong assumptions surrounding the expected number of total RNAs in a given cell means that Census and indeed any measure of relative abundance not normalized by spike-in standards will be limited in its ability to recapitulate the transcript counts derive from spike-based conversion. We argue here that this limitation is not onerous in differential analysis because its impact on fold changes between cells is small.

Testing for branch-dependent expression

Monocle assigns each cell a pseudotime value and a “State” encoding the segment of the trajectory it resides upon based on the PQ-tree algorithm (see the supplemental material for Trapnell and Cacchiarelli et al for further information[18]). Transcript counts values were variance-stabilized[49] via the technique described by Anders and Huber prior to tree construction.

In Monocle 2, we extended the capability to test for branch-dependent gene expression by formulating the problem as a contrast between two negative binomial GLMs.

The null model

$$NB(Census\ counts) \sim sm.ns(Pseudotime)$$

for the test assumes the gene being tested is not a branch specific gene, whereas the alternative model:

$$NB(Census\ counts) \sim sm.ns(Pseudotime) + Branch + sm.ns(Pseudotime):Branch$$

assumes that the gene is a branch specific gene where : represents an interaction term between branch and transformed pseudotime, NB means negative binomial distribution. Each model includes a natural spline (here with three degrees of freedom) describing smooth changes in mean expression as a function of pseudotime. The null model fits only a single curve, whereas the alternative will fit a distinct curve for each branch. Our current implementation of Monocle 2 relies on VGAM's "smart" spline fitting functionality, hence the use of the *sm.ns()* function instead of the more widely used *ns()* function from the splines package in R[50]. Likelihood ratio testing was performed with the VGAM *lrtest()* function, similar to Monocle's other differential expression tests[50]. A significant branch-dependent genes means that the gene has distinct expression dynamics along each branch, with smoothed curves that have different shapes.

To fit the full model, each cell must be assigned to the appropriate branch, which is coded through the factor "Branch" in the above model formula. Monocle's function for testing branch dependence accepts an argument specifying which branches are to be compared. These arguments are specified using the 'State' attribute assigned by Monocle during trajectory reconstructions. For example, in our analysis of the Truetlein *et al* data [25], Monocle reconstructed a trajectory with two branches (L_{AT1}, L_{AT2} for AT1 and AT2 lineages, respectively), and three states (S_{BP}, S_{AT1}, S_{AT2} for progenitor, AT1, or AT2 cells). The user specifies that he or she wants to compare L_{AT1} and L_{AT2} by providing S_{AT1} and S_{AT2} as arguments to the function. Monocle then assigns all the cells with state S_{AT1} to branch L_{AT1} and similarly for the $AT2$ cells. However, the cells with S_{BP} must be members of both branches, because they are on the path from each branch back to the root of the tree. In order to ensure the independence of data points required for the LRT as well as the robustness and stability of our algorithm, we implemented a strategy to partition the progenitor cells into two groups, with each branch receiving a group. The groups are computed by simply ranking the progenitor cells by pseudotime and assigning the odd-numbered cells to one group and the even numbered cells to the other. We assign the first progenitor to both branches to ensure they start at the same time which is required for downstream spline fitting and clustering. The branch plots in Figure 3d visualize the branch specific spline curves fit by this method.

Branch time point detection

The branching time point for each gene can be quantified by fitting a separate spline curves for each branch from all the progenitor to each cell fate. To robustly detect the pseudotime point (t_β^i) when a gene i with a branching expression pattern starts to diverge between two cell fates L_1, L_2, we developed the branch time point detection algorithm. The algorithm starts from the end of stretched pseudotime (pseudotime $t = 100$, see below) to calculate the divergence ($D_i(t = 100) = x_{L_1}(t = 100) - x_{L_2}(t = 100)$) of gene i's expression ($x_{L_1}(t = 100)$, $x_{L_2}(t = 100)$) between two cell fates, L_1 L_2, (for a branching gene, the divergence at this moment should be large if not the largest across pseudotime). It then moves backwards to find the latest intersection point between two fitted spline curves, which corresponds to the time when the gene starts to diverge between two branches. To add further flexibility, the algorithm moves forward to find the time point when the gene expression diverges up to a

user controllable threshold (ε), or $D_1(t) \geq \varepsilon(t)$, and defines this time point as the branch time point, t_β^i, for that particular gene *i*.

Analysis of human skeletal muscle myoblasts

We used the HSMM data from our previous publication[18] to benchmark the performance of developmental tree reconstruction and pseudotime DEG test between relative abundance or census counts. Relative abundances are converted into transcript counts using Census with default parameters with parameter t^* estimated from the relative abundance data for each cell. Potential contaminating fibroblast cells with transcript counts of *Mef2c* less than 5 and *Myf5* less than 1 were removed which yields 142 cells for downstream analysis.

The union of genes which are differentially expressed between the four time points in relative abundance or recovered transcript counts scale are used to reduce dimension and order the cells. Transcript counts were variance stabilized. The ordering of developmental trajectories between these two approaches is compared using spearman correlation. Pseudotime tests are performed on both the relative abundance and transcript counts scale where the pseudotime dependent genes are collected as those with q values less than 0.05 (Benjamini-Hochberg correction). The benchmark set is obtained from the permutation test based on a modified algorithm from the glm.perm package as previously described (see ***section Benchmarking differential expression analysis*** in supplementary notes).

Differential splicing analysis was conducted by first converting isoform-level TPM values from Cufflinks to transcript counts using Census with default parameters. Each gene's isoform-level transcript counts $Z_1, \ldots, Z_k$ were then modeled using a generalized linear model with a Dirichlet-multinomial response using the VGAM package (version 1.0-1). The Dirichlet-multinomial distribution is a compound distribution, where the probabilities that parameterize a multinomial are themselves drawn from a Dirichlet distribution with an additional over dispersion parameter $\boldsymbol{\phi}$. That is, the Dirichlet encodes the frequencies of the isoforms $\boldsymbol{\pi}$ and the variation in this frequency vector, while the multinomial captures the sampling of actual transcripts according to these frequencies. The Dirichlet has proven effective in previous analyses of splicing changes in bulk RNA-Seq studies [51].

To test for pseudotime-dependent shifts in the frequencies of the isoforms produced by each gene, we fit the following model to the observed isoform-level Census RNA counts:

$$Dirmultinomial(Z_1, \ldots, Z_k | \boldsymbol{\pi}, \boldsymbol{\phi}) \sim sm.ns(Pseudotime)$$

Only isoforms with at least one copy detected in at least 15 cells were included in the model for each gene, in order to ensure numerical stability within VGAM. We then compared this full model to the null

$$Dirmultinomial(Z_1, \ldots, Z_k | \boldsymbol{\pi}, \boldsymbol{\phi}) \sim 1$$

by likelihood ratio test. Note that each gene's ϕ was estimated by maximum likelihood separately, as we did not wish to assume that these dispersion parameters are a smooth function of expression level, as is commonly done in RNA-Seq.

Analysis of pre-implantation embryos

Allele-specific relative gene expression values (transcripts per million) were estimated by applying Kallisto[46] to the raw reads of Deng *et al.*[45] using an allele-specific transcriptome index. This index consisted of cDNA sequences from GENCODE vM9, corresponding to the paternal (C57BL/6J) alleles, plus the same sequences with maternal (CAST/EiJ) SNP alleles overlaid (CAST genotypes from Keane *et al.* [52]; only homozygous variants relative to the C57BL/6J reference were used).

The TPM values for the two alleles for each gene were converted to allelic RNA counts using Census with default parameters. The number of RNA molecules from each allele of each gene were modeled using a quasibinomial GLM. The quasibinomial is a binomial that allows for over (or under) dispersion with respect to the binomial through a parameter ϕ. Its probability mass function is:

$$P(x{=}k){=}\left(\begin{array}{c} n \\ k \end{array}\right) p(p{+}k\phi)^{k-1}(1-p-k\phi)^{n-k}$$

where p encodes the probability that an RNA originated from the maternal allele (without loss of generality).

Quasibinomial GLMs were fit to each gene using VGAM, using the option "dispersion=0" to direct VGAM to estimate the dispersion parameter for each model from each gene's maternal and paternal RNA counts Z_m and Z_p, respectively. To test for embryo stage-dependent allelic balance shifts in each gene, we fit a full model

$$quasibinomial(Z_m, Z_p) \sim stage$$

And a null

$$quasibinomial(Z_m, Z_p) \sim 1$$

to these data, and compared them using an F-test [29]. As for isoform-level modeling, the dispersion parameter was fit separately for each gene. We note that the quasibinomial is similar to the beta-binomial, the two category case of the Dirichlet-Multinomial. We explored the use of the beta-binomial for this analysis, and while we reached qualitatively similar conclusions regarding escape from X inactivation and monoallelic expression, we felt that the quasibinomial provided a better fit for the data.

Analysis of X chromosome inactivation was performed on female embryos at the 4-, 16- and early blastocyst stages. Embryos were sexed by hierarchically clustering cells on the basis of variance stabilized transcript counts for genes on the Y chromosome. Cells fell into two

clearly defined clusters, only one of which expressed “informative” Y genes. Embryos comprised of these cells were annotated as male.

To quantify the number of genes escaping X inactivation at each stage, we used the quasibinomial GLMs to assess the probability that less than 10% of the RNA from a gene originated from the inactive chromosome. (10% is a widely accepted threshold for escape from X inactivation[53,54]). To do so, we constructed a 95% prediction interval on the allelic ratio for each gene by simulating random variates from its GLM via the VGAM package’s simulate.vlm(). That is, we calculated the number of simulated observations that were less than 10% percent maternal or paternal. Using this statistic, we calculated a significance score for contribution from the maternal and paternal alleles for each gene on the X chromosome, corrected these for multiple testing (via Benjamini-Hochberg), and reported the number of genes with significant maternal and paternal contributions.

We used a similar simulation-based procedure to construct prediction intervals for expected monoallelic expression. After fitting a quasibinomial GLM for each (autosomal) gene’s allele RNA counts, we simulated 100 random variates from each gene’s model and counted the number of times the model reported RNAs from only one of the two alleles. We then collected these counts into quantiles based on the gene’s expression level to generate 95% prediction intervals for monoallelic expression as a function of expression level. The exact same fitting, simulation, and prediction interval estimation procedure was used for both RNA counts and estimated allelic read counts from Kallisto.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Jianghong Shi, Song Xu for technical discussions and Martin Kircher for cluster computation support. We are grateful to Jay Shendure, Ron Hause, Darren Cusanovich, Bruce Trapnell, Jeff Whitsett and members of the Trapnell laboratory for comments on manuscript. This work was supported by NIH Grant DP2 HD088158. C.T. is partly supported by a Dale. F. Frey Award for Breakthrough Scientists and an Alfred P. Sloan Foundation Research Fellowship. A.H. is supported by an NSF Graduate Research Fellowship.

References

1. Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. 2015; doi: 10.1016/j.cell.2015.05.002
2. Klein AM, et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell. 2015; 161:1187–1201. [PubMed: 26000487]
3. Shalek AK, et al. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. Nature. 2013; doi: 10.1038/nature12172
4. Grün D, Kester L, van Oudenaarden A. Validation of noise models for single-cell transcriptomics. Nature methods. 2014; 11:637–640. [PubMed: 24747814]
5. Finak G, et al. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. Genome biology. 2015; 16:278. [PubMed: 26653891]
6. Jiang L, et al. Synthetic spike-in standards for RNA-seq experiments. Genome Research. 2011; 21:1543–1551. [PubMed: 21816910]

7. Fu GK, Hu J, Wang PH, Fodor SPA. Counting individual DNA molecules by the stochastic attachment of diverse labels. Proceedings of the National Academy of Sciences of the United States of America. 2011; 108:9026–9031. [PubMed: 21562209]
8. Hug H, Schuler R. Measurement of the number of molecules of a single mRNA species in a complex mRNA preparation. Journal of theoretical biology. 2003; 221:615–624. [PubMed: 12713944]
9. Picelli S, Faridani OR, Bj o rklund AK, Winberg G. Full-length RNA-seq from single cells using Smart-seq2. 2014; doi: 10.1038/nprot.2014.006
10. Petropoulos S, et al. Single-Cell RNA-Seq Reveals Lineage and X Chromosome Dynamics in Human Preimplantation Embryos. Cell. 2016; 165:1012–1026. [PubMed: 27062923]
11. Segerstolpe Å, et al. Single-Cell Transcriptome Profiling of Human Pancreatic Islets in Health and Type 2 Diabetes. Cell Metabolism. 2016; 24:593–607. [PubMed: 27667667]
12. Zeisel A, et al. Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science. 2015; 347:1138–1142. [PubMed: 25700174]
13. Jaitin DA, et al. Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. Science. 2014; 343:776–779. [PubMed: 24531970]
14. Wu AR, et al. Quantitative assessment of single-cell RNA-sequencing methods. Nature methods. 2013; 11:41–46. [PubMed: 24141493]
15. Treutlein B, et al. Dissecting direct reprogramming from fibroblast to neuron using single-cell RNA-seq. Nature. 2016; 534:391–395. [PubMed: 27281220]
16. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010; 26:139–140. [PubMed: 19910308]
17. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome biology. 2014; 15
18. Trapnell C, et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nature Biotechnology. 2014; 32:381–386.
19. Kharchenko PV, Silberstein L, Scadden DT. Bayesian approach to single-cell differential expression analysis. Nature methods. 2014; 11:740–742. [PubMed: 24836921]
20. Tang F, et al. Tracing the derivation of embryonic stem cells from the inner cell mass by single-cell RNA-Seq analysis. Cell stem cell. 2010; 6:468–478. [PubMed: 20452321]
21. Buganim Y, et al. Single-cell expression analyses during cellular reprogramming reveal an early stochastic and a late hierarchic phase. Cell. 2012; 150:1209–1222. [PubMed: 22980981]
22. Zhou JX, Huang S. Understanding gene circuits at cell-fate branch points for rational cell reprogramming. Trends in genetics : TIG. 2011; 27:55–62. [PubMed: 21146896]
23. Moignard V, et al. Decoding the regulatory network of early blood development from single-cell gene expression measurements. Nature Biotechnology. 2015; 33:269–276.
24. Marco E, et al. Bifurcation analysis of single-cell gene expression data reveals epigenetic landscape. Proceedings of the National Academy of Sciences of the United States of America. 2014; 111:E5643–50. [PubMed: 25512504]
25. Treutlein B, et al. Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. Nature. 2014; 509:371–375. [PubMed: 24739965]
26. Hochegger H, Takeda S, Hunt T. Cyclin-dependent kinases and cell-cycle transitions: does one fit all? Nature Reviews Molecular Cell Biology. 2008; 9:910–916. [PubMed: 18813291]
27. Desai TJ, Brownfield DG, Krasnow MA. Alveolar progenitor and stem cells in lung development, renewal and cancer. Nature. 2014; 507:190–194. [PubMed: 24499815]
28. Chi X, Garnier G, Hawgood S, Colten HR. Identification of a novel alternatively spliced mRNA of murine pulmonary surfactant protein B. Am J Respir Cell Mol Biol. 1998; 19:107–113. [PubMed: 9651186]
29. MCCULLAGH, P., Nelder, JA. Generalized Linear Models. Second Edition. CRC Press; 1989.
30. Shu W, et al. Foxp2 and Foxp1 cooperatively regulate lung and esophagus development. Development (Cambridge, England). 2007; 134:1991–2000.

31. Yin Y, et al. An FGF-WNT gene regulatory network controls lung mesenchyme development. Developmental Biology. 2008; 319:426–436. [PubMed: 18533146]
32. Shu W, Yang H, ZHANG L, Lu MM, Morrisey EE. Characterization of a new subfamily of winged-helix/forkhead (Fox) genes that are expressed in the lung and act as transcriptional repressors. J Biol Chem. 2001; 276:27488–27497. [PubMed: 11358962]
33. Wan H, et al. Kruppel-like factor 5 is required for perinatal lung morphogenesis and function. Development (Cambridge, England). 2008; 135:2563–2572.
34. Xu Y, et al. C/EBP{alpha} is required for pulmonary cytoprotection during hyperoxia. Am J Physiol Lung Cell Mol Physiol. 2009; 297:L286–98. [PubMed: 19465518]
35. Okubo T, Hogan BLM. Hyperactive Wnt signaling changes the developmental potential of embryonic lung endoderm. Journal of Biology. 2004; 3:11. [PubMed: 15186480]
36. Shalek AK, et al. Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. Nature. 2014; doi: 10.1038/nature13437
37. Darnell JE, Kerr IM, Stark GR. Jak-STAT pathways and transcriptional activation in response to IFNs and other extracellular signaling proteins. Science. 1994; 264:1415–1421. [PubMed: 8197455]
38. Honda K, et al. IRF-7 is the master regulator of type-I interferon-dependent immune responses. Nature. 2005; 434:772–777. [PubMed: 15800576]
39. Gautier G, et al. A type I interferon autocrine-paracrine loop is involved in Toll-like receptor-induced interleukin-12p70 secretion by dendritic cells. J Exp Med. 2005; 201:1435–1446. [PubMed: 15851485]
40. Lavin Y, et al. Tissue-resident macrophage enhancer landscapes are shaped by the local microenvironment. Cell. 2014; 159:1312–1326. [PubMed: 25480296]
41. Welch JD, Hu Y, Prins JF. Robust detection of alternative splicing in a population of single cells. Nucleic acids research. 2016; 44:e73–e73. [PubMed: 26740580]
42. Perrin BJ, Ervasti JM. The actin gene family: function follows isoform. Cytoskeleton (Hoboken). 2010; 67:630–634. [PubMed: 20737541]
43. Tondeleir D, Vandamme D, Vandekerckhove J, Ampe C, Lambrechts A. Actin isoform expression patterns during mammalian development and in pathology: insights from mouse models. Cell Motil Cytoskeleton. 2009; 66:798–815. [PubMed: 19296487]
44. Gunning P, O'Neill G, Hardeman E. Tropomyosin-based regulation of the actin cytoskeleton in time and space. Physiol Rev. 2008; 88:1–35. [PubMed: 18195081]
45. Deng Q, Ramsköld D, Reinius B, Sandberg R. Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. 2014; doi: 10.1126/science.1245316
46. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nature Biotechnology. 2016; 34:525–527.
47. Kim JK, Kolodziejczyk AA, Illicic T, Teichmann SA, Marioni JC. Characterizing noise structure in single-cell RNA-seq distinguishes genuine from technical stochastic allelic expression. Nature Communications. 2015; 6:8687.
48. Amit I, et al. Unbiased reconstruction of a mammalian transcriptional network mediating pathogen responses. Science. 2009; 326:257–263. [PubMed: 19729616]
49. Anders S, Huber W. Differential expression analysis for sequence count data. Genome biology. 2010; 11
50. Yee, TW. Vector Generalized Linear and Additive Models. Springer; 2015.

Method-only Reference

51. Katz Y, Wang ET, Airoldi EM, Burge CB. Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nature methods. 2010; 7:1009–1015. [PubMed: 21057496]
52. Keane TM, et al. Mouse genomic variation and its effect on phenotypes and gene regulation. Nature. 2011; 477:289–294. [PubMed: 21921910]

53. Corbel C, Diabangouaya P, Gendrel AV, Chow JC, Heard E. Unusual chromatin status and organization of the inactive X chromosome in murine trophoblast giant cells. Development (Cambridge, England). 2013; 140:861–872.

54. Yang F, Babak T, Shendure J, Disteche CM. Global survey of escape from X inactivation by RNA-sequencing in mouse. Genome Research. 2010; 20:614–622. [PubMed: 20363980]

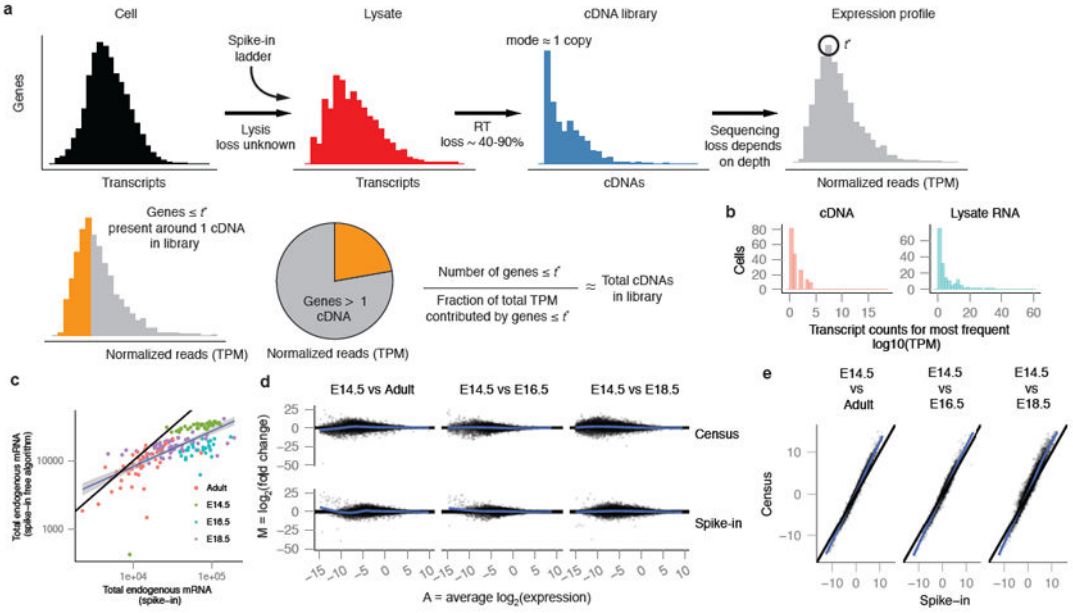

Figure 1. Census approximates relative transcript counts in single cells without external RNA standards

(a) The typical procedure for estimating lysate mRNA abundances via spike-in standards in single-cell RNA-Seq. Losses at various stages alter the distribution of relative gene expression levels within a single cell. (b) Distribution of the transcript counts corresponding to each cell's most frequently observed relative abundance (i.e. TPM) in cDNA or lysate RNA space in the lung epithelial data from Treutlein *et al.* Modes are obtained by log-transforming the data, performing a Gaussian kernel density estimation, and then exponentiating back to the original scale. (c) Total transcripts per lung epithelial cell estimated via spike-in controls versus counts from the spike-free algorithm in Census. Blue line indicates linear regression. Black line indicates perfect concordance. (d) MA plot for expressed genes based on contrasts between cells from E14.5 and cells from all other time points. The top panels show Census transcript counts while the bottom panels show transcript counts derived by spike-in regression. (e) Fold changes for expressed genes based on data from Census transcripts or transcripts with spike-in regression of contrasts between cells from E14.5 and cells from all other time points.

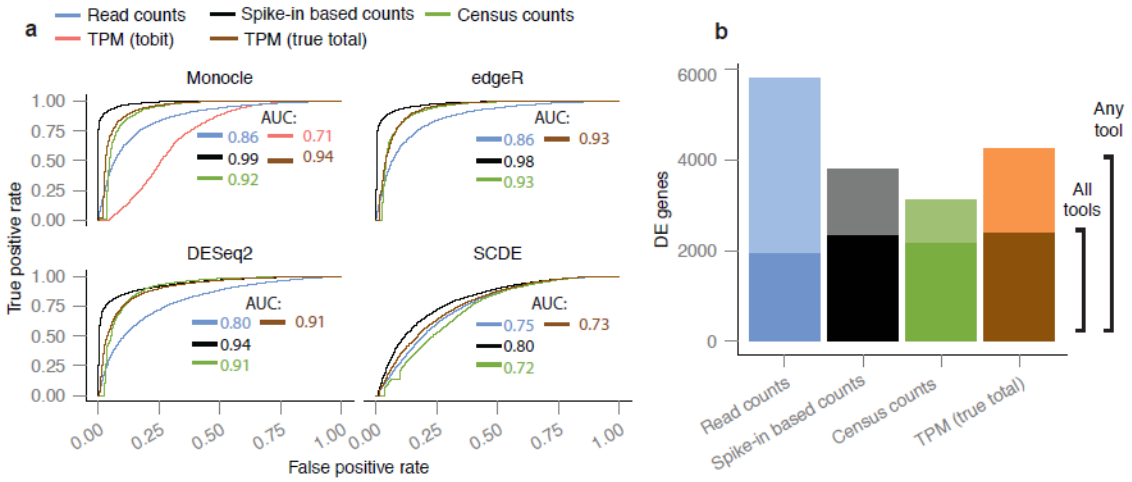

Figure 2. Use of Census counts improves accuracy of differential expression analysis and can be performed on libraries with or without spike controls

(a) Receiver-operating characteristic (ROC) curves showing differential expression (DE) analysis accuracy from various tools provided with relative expression levels, normalized read counts, and transcript counts estimated with spike-ins or Census. Cells from E14.5 and E18.5 from Treutlein *et al.* were provided to each tool. A permutation-based test was applied to the spike-in-based expression levels to determine a ground truth set of DE genes. In addition to Census and spike controls, we include transcript counts derived by scaling the TPM values by the correct per-cell total RNAs. This control shares Census' inability to control for amplification bias, but begins with the same total per-cell transcript counts available through spike-ins. Comparing this control to spike-based regression reveals the impact of amplification bias on differential analysis in single cells. Comparing it to Census assesses how error in estimating total transcript counts translates into error in differential analysis. (b) Consensus in differential analysis results between Monocle, DESeq2, edgeR, and permutation tests using different measures of expression. The total height of each bar reflects the size of the union of DE genes reported by any of the four tests. The smaller bar reports the number of DE genes identified by all tests.

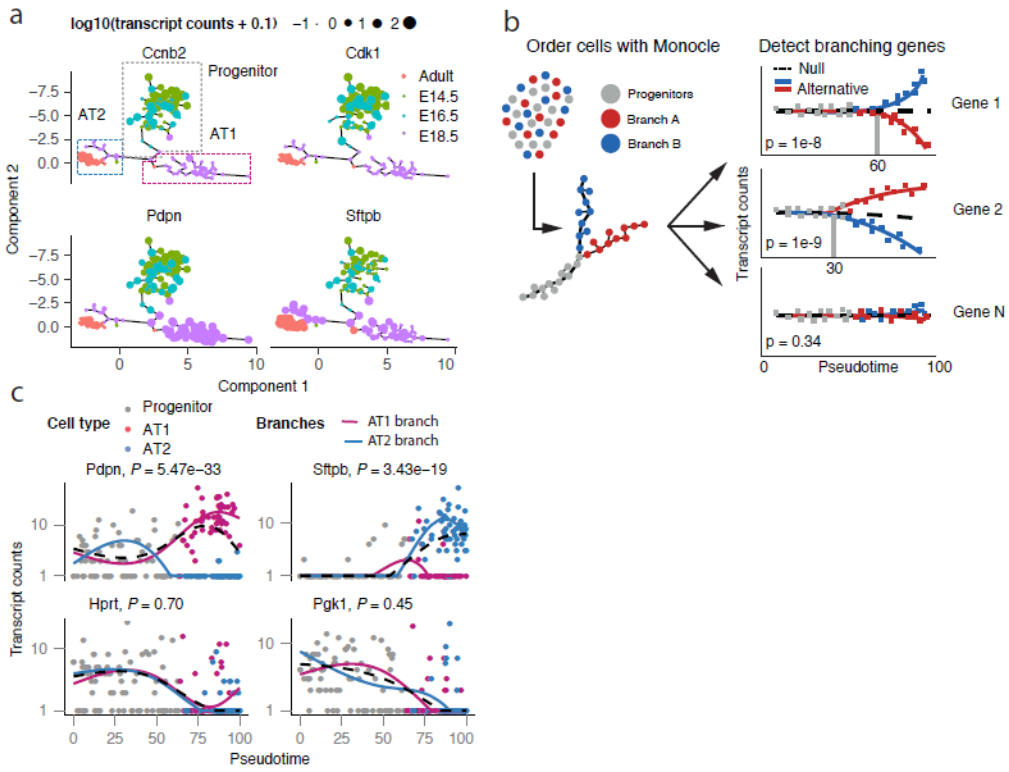

Figure 3. BEAM identifies genes with branch-dependent expression and potential drivers during lung epithelial fate specification

(a) Monocle 2 recovers a branched single-cell trajectory beginning with bronchoalveolar progenitors (BP) and terminating at type I (AT1) and type II (AT2) pneumocytes. High expression of known markers of proliferation (*Ccnb2, Cdk2*) is restricted to progenitor cells, whereas high expression of known AT1 (Pdpn) and AT2 (Sftpb) markers is restricted to their corresponding lineages. Size of circles denotes level of expression. (b) Branching Expression Analysis Modeling (BEAM) is a statistical framework for identifying genes with expression that changes over a single-cell trajectory in a branch-dependent manner. BEAM first uses generalized linear models with natural splines to perform a regression on the data in which the branch assignments of the cells are known (alternative model), fitting a separate curve for each branch. It also performs another regression in which the branch assignments are not known (null model), fitting a single curve for all the data, and then compares these models via a likelihood ratio test. (c) Null and alternative model fits for the AT1/2 markers (*Ager* / *Sftpb*) and housekeeping genes (*Hprt* and *Pgk1*). Solid lines indicate the smoothed expression curves for each branch in the alternative model while dashed line corresponding to the fitted curve in the null model used in the BEAM test.

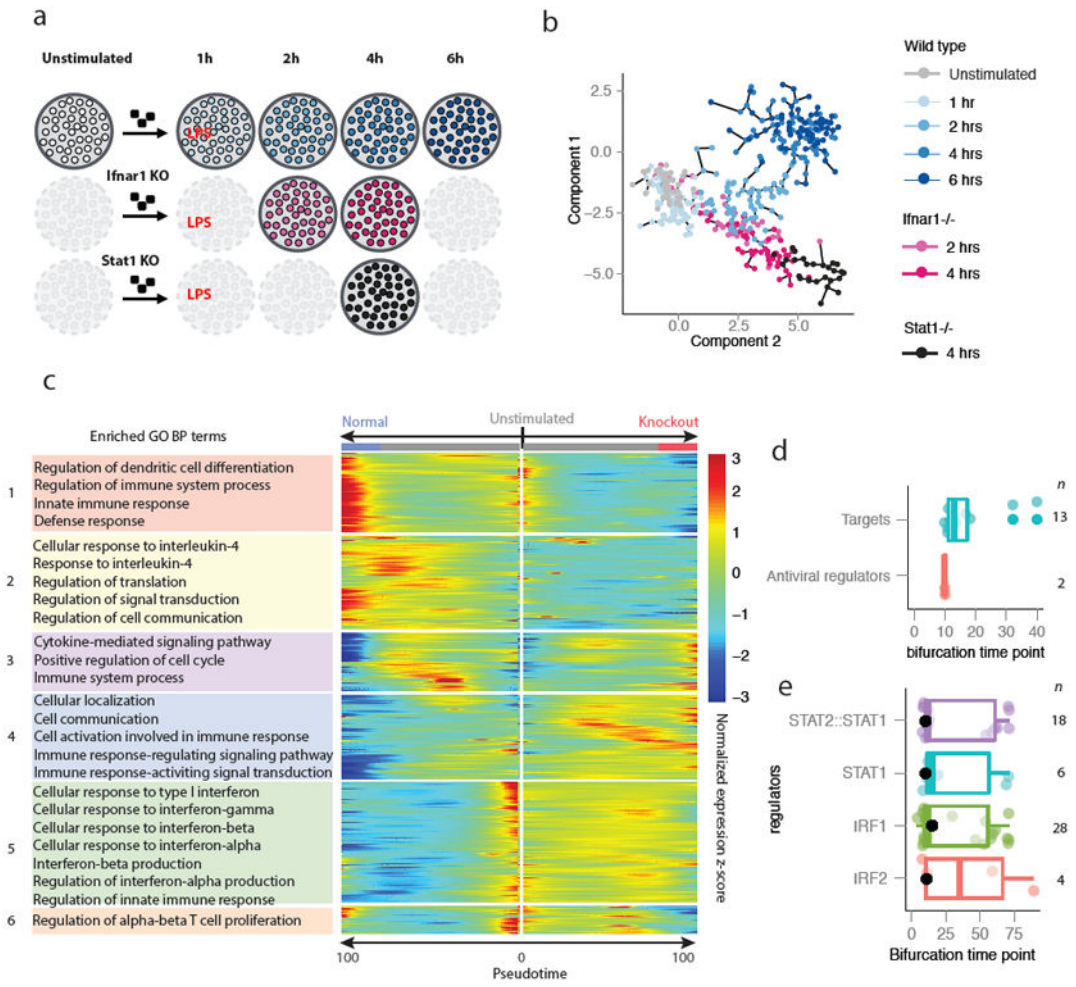

Figure 4. Loss of interferon signaling generates a branch in the trajectory followed by immune-stimulated dendritic cells

(a) Experimental design used by Shalek *et al.* to compare BMDCs from *Ifnar1*-/- and *Stat1*-/- knockout mice against the wild type as they respond to LPS. (b) Single-cell trajectory recovered by Monocle 2. (c) Kinetic clusters of branch-dependent genes identified by BEAM are functionally enriched for interferon signaling and other immune-related processes. (d) Branch time point for the significant branching antiviral regulators and their significant branching targets collected from Fig. 4 of. [48]) (e) Branch time points for the TFs with motifs enriched in nearby DHS site from significant branch genes from cluster 5 and their potential target genes in cluster 5 (panel c). For all boxplots in this study, the upper and lower “hinges” correspond to the first and third quartiles (the 25th and 75th percentiles). The whiskers extend from the upper (or lower) hinge to the highest (or lowest) value that is within 1.5 * IQR of the hinge, where IQR is the inter-quartile range, or distance between the first and third quartiles. Data beyond the end of the whiskers are outliers and plotted as points. The center line corresponds to the median.

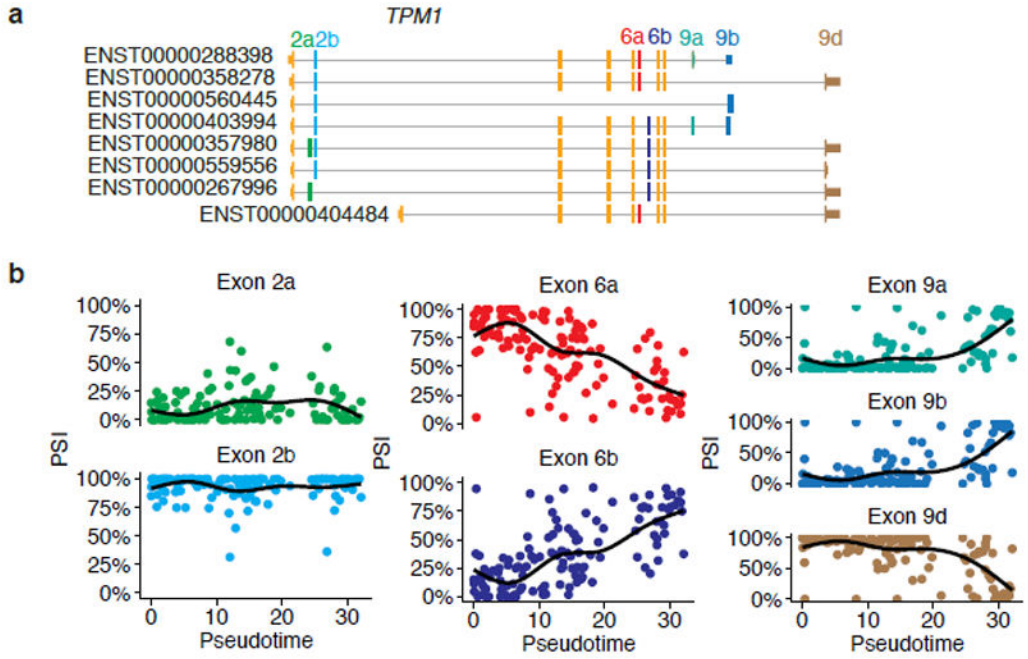

Figure 5. Census enables robust analysis of differential splicing during cell differentiation

(a) Splicing structure of TPM1, with the three alternatively spliced sets of exons highlighted. (b) Percent-spliced-in (PSI) values for TPM1 alternative exons. PSI values were computed by summing Census counts for isoforms including each exon and dividing by the total TPM1 transcript count in each cell. Black lines indicate loess smoothing of the PSI values as a function of pseudotime.

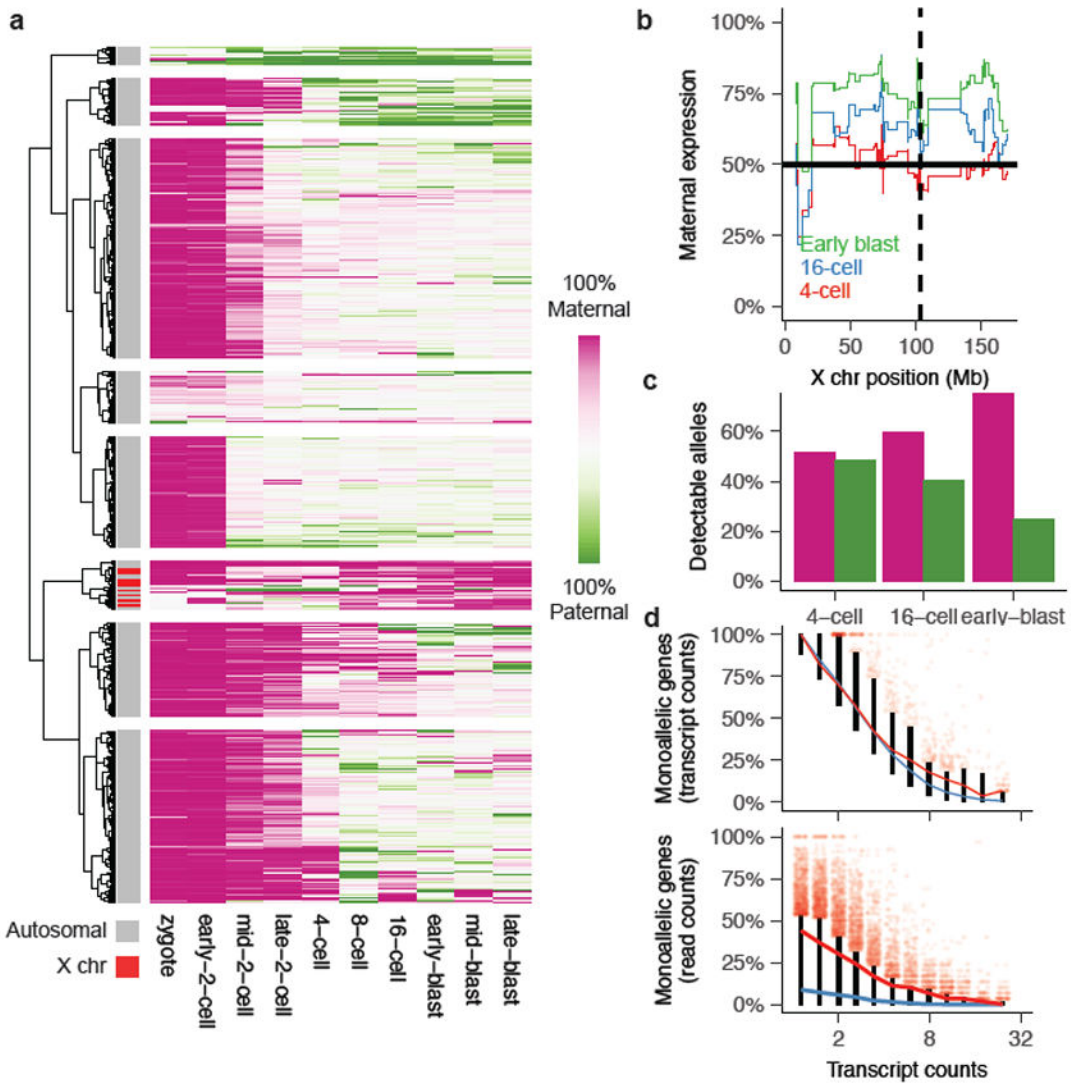

Figure 6. Census detects shifts in allelic balance in single cells during embryogenesis

(a) A quasibinomial regression model detects changes in allelic balance in single cells as a function of embryo stage. (b) Spread of X chromosome inactivation as measured by Census for female embryos at the 4-cell, 16-cell, and early blastocyst stage. Compare with Fig 2B from Deng *et a l*[45]. (c) Number of genes with at least 10% contribution from the maternal and paternal copies of X chromosome. (d) Observed monoallelic expression in single cells from late stage embryos as measured by Census transcript counts (top) or normalized read counts (bottom). Red line indicates median fraction of monoallelic calls as a function of average transcript count across cells. Only autosomal genes are shown. Black bars indicate 95% prediction interval generated by a quasibinomial regression model fit to each gene, with the median of the gene intervals indicated by the blue line. Light red points indicate individual genes that fall outside the prediction interval.