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] & Cole Trapnell[1,2]

Single-cell gene expression studies promise to reveal rare cell types and cryptic states, but the high variability of single-cell RNA-seq measurements frustrates efforts to assay transcriptional differences between cells. We introduce the Census algorithm to convert relative RNA-seq expression levels into relative transcript counts without the need for experimental spike-in controls. Analyzing changes in relative transcript counts led to dramatic improvements in accuracy compared to normalized read counts and enabled new statistical tests for identifying developmentally regulated genes. Census counts can be analyzed with widely used regression techniques to reveal changes in cell-fate-dependent gene expression, splicing patterns and allelic imbalances. We reanalyzed single-cell data from several developmental and disease studies, and demonstrate that Census enabled robust analysis at multiple layers of gene regulation. Census is freely available through our updated single-cell analysis toolkit, Monocle 2.

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, such changes are difficult to model[3,4] because of the high biological and technical noise in single-cell measurements. Many studies report high rates of 'dropout', wherein some cells of a nominally homogeneous population express high levels of a gene and others none at all. Drop-outs have spurred the deployment of hurdle models[5] that overcome the limitations of 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 at the level of transcript counts rather than read counts. Previous work[4] had suggested that comparing UMIs, rather than read counts, between cells would improve regression analysis. However, because UMI protocols are based on counting 3′-end tags, they do not report expression at allele or isoform resolution. Spike-in-based protocols use a linear regression between the normalized read counts and known molecular concentrations of spiked transcripts to convert relative cellular RNA read counts to transcript counts. However, exogenous standards must be carefully calibrated to avoid dominating the libraries, and may be subject to different rates of degradation or reverse transcription than endogenous RNA. Many published studies 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, branch expression analysis modeling (BEAM), for detecting genes that change after fate decisions in development. We also demonstrate that Census counts can be used to detect developmental regulation robustly at splice isoform and allele resolution. Census and BEAM are implemented in Monocle 2, the second major release of our open-source single-cell analysis toolkit (**Supplementary Software** and https://github.com/cole-trapnell-lab/monocle-release).

RESULTS

Estimating relative transcript counts in spike-in-free experiments

Census exploits two properties of single-cell RNA-seq data sets produced with current protocols (**Fig. 1a**). First, mRNA degradation after 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 poly(A) tail of mRNAs and thus generate full-length cDNAs[9]. In these libraries, genes are detected most frequently as a single cDNA molecule (**Fig. 1b** and **Supplementary Fig. 1**). Thus, all detectably expressed genes measured at or below the mode of the (log-transformed) relative

[1]Department of Genome Sciences, University of Washington, Seattle, Washington, USA. [2]Molecular and Cellular Biology Program, University of Washington, Seattle, Washington, USA. [3]Department of Applied Mathematics, University of Washington, Seattle, Washington, USA. Correspondence should be addressed to C.T. (coletrap@uw.edu).

RECEIVED 26 SEPTEMBER 2016; ACCEPTED 22 DECEMBER 2016; PUBLISHED ONLINE 23 JANUARY 2017; DOI:10.1038/NMETH.4150

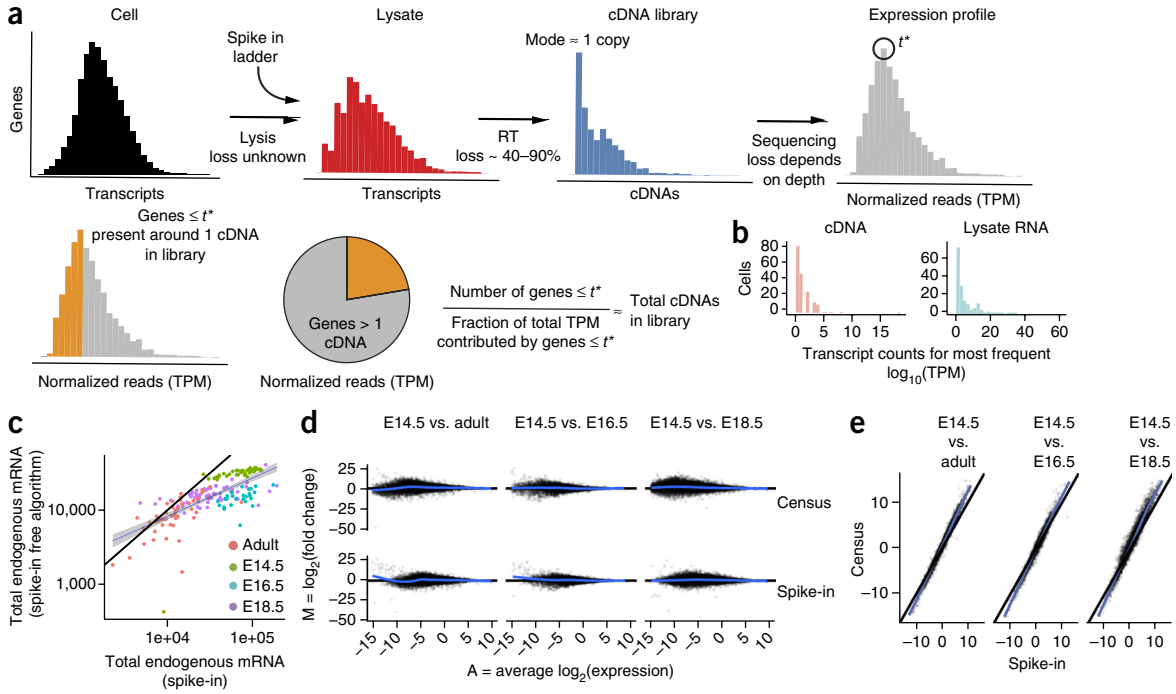

Figure 1 | Census approximation of relative transcript counts in single cells without external RNA standards. (**a**) Typical single-cell RNA-seq procedure for estimating mRNA abundances via spike-in standards. Losses alter the distribution of relative gene expression levels in a single cell. RT, reverse transcription. (**b**) Distribution of transcript counts corresponding to each cell's most frequently observed relative abundance (i.e., TPM) in cDNA or lysate RNA from lung epithelial data[25]. (**c**) Total transcripts per lung epithelial cell estimated using Census counts versus using spike-in controls. Blue line indicates linear regression. The shading around the blue line indicates the 95% confidence interval of the regression. Black line indicates perfect concordance. (**d**) MA plot for expressed genes based on contrasts between cells from embryonic day (E)14.5 and cells from all other time points. Census transcript counts (top); transcript counts derived by spike-in regression (bottom). (**e**) Fold changes in gene expression based on Census counts or spike-in regression of spike-ins, contrasting cells from E14.5 and all other time points.

abundance distribution in each cell should be present at around one cDNA copy (Online Methods).

We assessed Census' accuracy by reanalyzing several experiments that included spike-in controls[4,10–15] (accession codes for all data we analyzed are available in Online Methods). 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 (**Fig. 1c**). This is likely because of Census' inability to control for nonlinear 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 (**Fig. 1d**,**e**). Census produced accurate changes in relative transcript counts for seven additional data sets, including two based on UMIs[4,10,13], demonstrating that the algorithm can work well with different single-cell RNA-seq protocols (**Supplementary Figs. 1** and **2**). Downsampling and simulation experiments revealed that Census counts faithfully captured 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 Figs. 3** and **4**). Taken together, these benchmarking experiments show that Census recovered an accurate measure of changes in relative transcript counts between single cells without the need for spike-in controls.

Census counts improved 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 single-cell differential expression (SCDE)[19] (**Fig. 2a** and **Supplementary Fig. 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 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 (**Fig. 2b** and Online Methods). Tools designed for bulk RNA-seq analysis, such as DESeq2 (ref. 17), produce false discovery rates as high as 61%. SCDE, which includes explicit modeling of dropouts, 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 (**Fig. 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 in ref. 4 (**Supplementary Fig. 5**). For example, when targeting a false discovery rate of 10%, DESeq2's empirical false discovery rate dropped from 61% to 22%, with little to no drop in sensitivity, which remained as high as 82%. Monocle's false discovery rate dropped from 53% to 11%. The use of Census counts dramatically improved agreement between the tools, which agreed on 2,437 differentially expressed genes among a total of

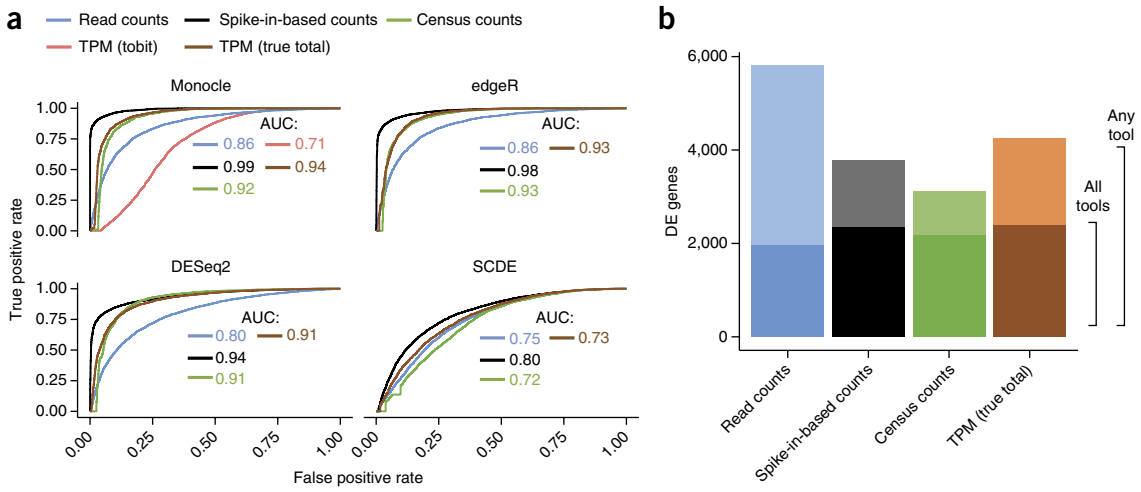

Figure 2 | Census counts improved the accuracy of differential expression analysis. (**a**) Receiver-operating characteristic (ROC) curves showing the accuracy of differential expression (DE) analysis between E14.5 and E18.5 lung epithelial cells[25]. Tools were provided with relative expression levels, normalized read counts, and transcript counts estimated with spike-ins or Census. A permutation-based test was applied to the spike-in-based expression levels to determine a ground truth set of DE genes. TPM (true total), counts derived by scaling TPM values by the correct per-cell total RNA. AUC, area under the curve. (**b**) Consensus between Monocle, DESeq2, edgeR and permutation tests using different measures of expression. Lighter bar colors, size of the union of DE genes reported by any of the four tests. Darker bar colors, number of DE genes identified by all tests.

4,220 (70%), similar to the 62% (2,367/3,793 genes) consensus obtained with spike-in-derived levels (**Fig. 2b** and **Supplementary Data**). Census also improved differential expression accuracy relative to gold standards derived from bulk RNA-seq[18] measurements (**Supplementary Fig. 6**). Taken together, our benchmarks demonstrated 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.

Branch point analysis revealed regulators of cell fate

Many single-cell gene expression studies aim to identify gene regulatory circuits that control cell-fate decisions[20,21]. We recently developed Monocle, an algorithm that organizes single cells along trajectories and can describe gene-expression changes during 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 produced more reliable tests for differential expression along a trajectory when provided with Census counts than with relative expression values (**Supplementary Fig. 7**).

Analyzing cells at branch points between two or more mutually exclusive developmental trajectories[22] 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].

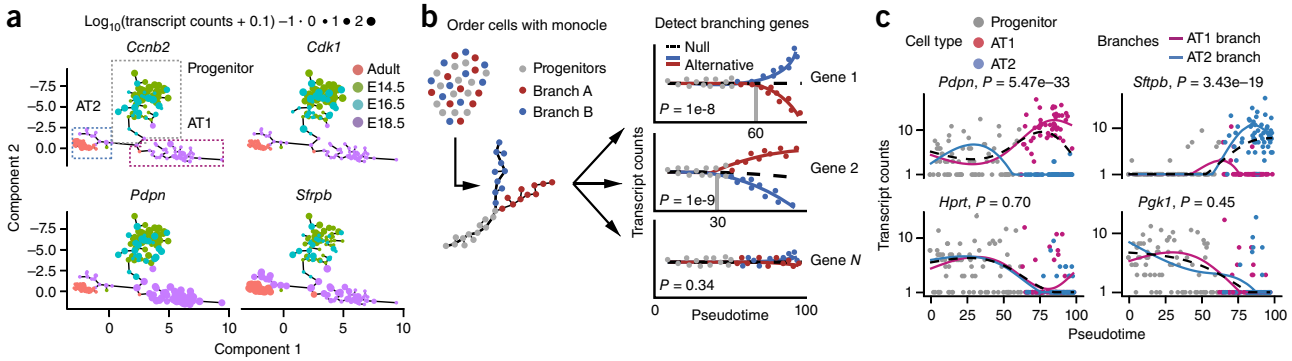

Figure 3 | BEAM identification of branch-dependent gene expression and potential drivers of lung epithelial fate specification. (**a**) Monocle recovered a branched single-cell trajectory beginning with bronchoalveolar progenitors and terminating at type I (AT1) and type II (AT2) pneumocytes. High expression of proliferation markers (*Ccnb2* and *Cdk1*) was restricted to progenitor cells, whereas high expression of AT1 (*Pdpn*) and AT2 (*Sftpb*) markers was restricted to their corresponding lineages. Size of circles denotes level of expression. (**b**) BEAM uses generalized linear models with natural splines to perform a regression on the data in which branch assignments are known (alternative model), fitting a separate curve for each branch. It also performs a regression in which branch assignments are not known (null model) by 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 AT1 and AT2 markers (*Ager* and *Sftpb*, respectively) and housekeeping genes (*Hprt* and *Pgk1*). Solid lines, smoothed expression curves for each branch in the alternative model. Dashed lines, fitted curve in null model used in the BEAM test.

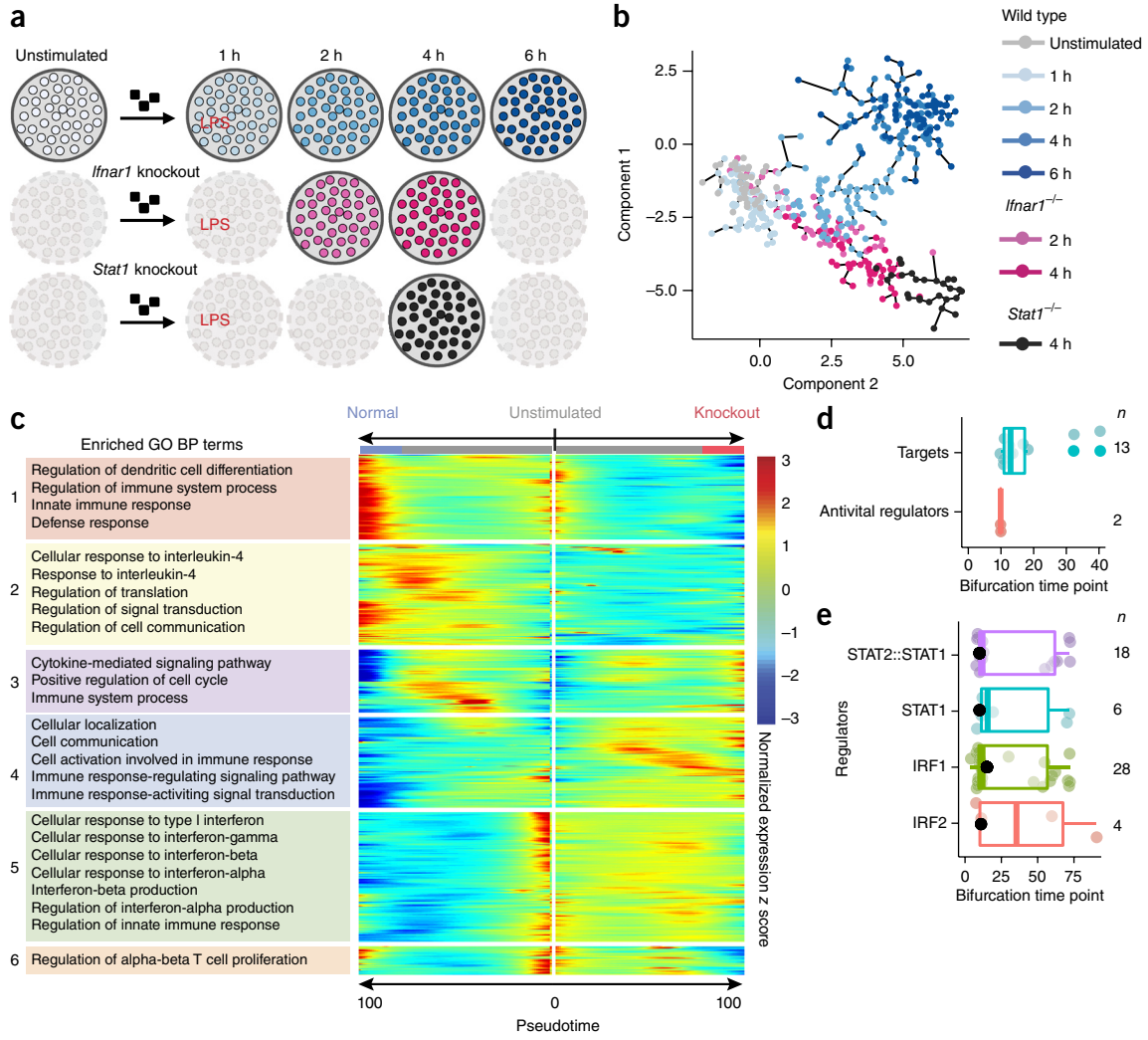

Figure 4 | Loss of interferon signaling generated a branch in the trajectory followed by immune-stimulated dendritic cells. (**a**) Experimental design used in ref. 36 to compare BMDCs from *Ifnar1*$^{-/-}$ and *Stat1*$^{-/-}$ knockout mice against the wild type as they respond to LPS. (**b**) Single-cell trajectory recovered by Monocle 2. (**c**) Six 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 (via the BEAM test) branching antiviral regulators and their significant branching targets collected from ref. 48 figure 4. (**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 in **c**. For all boxplots, upper and lower 'hinges' correspond to the first and third quartiles (the 25th and 75th percentiles), whiskers extend to the highest (or lowest) value that is within 1.5 × inter-quartile range of the hinge, or distance between the first and third quartiles. Points beyond the whiskers are the remaining data. The center line corresponds to the median.

To explore a developmental fate decision, we reanalyzed data on distal lung epithelium specification based on sequencing epithelial intermediates that give rise to type I (AT1) and type II (AT2) pneumocytes[25]. Monocle reconstructed a trajectory with a single branch point leading from progenitors to two outcomes corresponding to the AT1 and AT2 fates. Cells at the beginning of the trajectory exhibited high levels of proliferation markers[26] *Ccnb2* and *Cdk1*, whereas cells after the branch point exhibited much lower levels (**Fig. 3a**). High expression of the AT1 cell marker[27] *Pdpn* was restricted to cells on one branch, whereas cells expressing the AT2 marker[28] *Sftpb* at high levels were located on the other branch. Cells classified as AT1 and AT2 according to known markers[25] fell exclusively along the branches, with what the authors termed "bipotent progenitors" at or near the branch point. (**Supplementary Fig. 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 (**Fig. 3b**, **Supplementary Fig. 9** and Online Methods). BEAM identified 1,219 genes (false discovery rate (FDR) < 5%) as either AT1- or AT2-fate-dependent, including canonical markers[27] such as *Pdpn* and *Sftpb* (**Fig. 3c**). AT1-restricted genes were strongly enriched for ontology terms related to tube development, cytoskeletal remodeling and cell morphogenesis (**Supplementary Fig. 10** and **Supplementary Table 1**), whereas 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 Fig. 11**). These included several factors such as Tcf7l2 that are known to regulate lung development[30–35].

Disruption of interferon signaling induced a branch in the dendritic cell lipopolysaccharide-stimulation trajectory

Branch points in single-cell trajectories represent decision points that are enacted by regulated changes in the transcriptional program; they can arise during development, but also in response to mutations, treatment with drugs or other cellular perturbations. We re-analyzed single-cell data on the transcriptional response of mouse bone-marrow-derived dendritic cells (BMDCs) to lipopolysaccharide (LPS)[36] (**Fig. 4a**). In BMDCs, LPS triggers a paracrine feedback loop of type I interferon signaling mediated in part by *Stat1* (refs. 37–39). The authors compared BMDCs from wild-type mice to those from mice that lack the receptor for interferon alpha (*Ifnar1*$^{-/-}$) or *Stat1* (*Stat1*$^{-/-}$)[36]. 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 wild-type mice (**Fig. 4b**).

BEAM identified 870 genes (FDR < 5%) dependent on this branch, many associated with interferon signaling (**Fig. 4c**, **Supplementary Fig. 12** and **Supplementary Table 1**). Peaks corresponding to open chromatin collected in ref. 40 proximal to branch-dependent genes were enriched for Stat1/2 and Irf1/2 binding motifs (**Supplementary Fig. 13**). *Stat1/2* and *Irf1/2* 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 (**Fig. 4d**,**e**). Monocle 2 and BEAM demonstrated that loss of a key paracrine loop generated an 'alternative trajectory', suggesting that single-cell trajectory analysis can be useful for defining how a signaling pathway regulates a larger process.

Census counts enabled single-cell differential splicing analysis

Methods for detecting splicing changes in single-cell data are beginning to appear, but have suffered from isoform-level measurement variability. For example, SingleSplice[41] uses a hurdle model to compare variation in isoform frequencies against expected technical variation. However, its contrasts are limited to tests for excess variability in 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 human 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] (**Fig. 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 Fig. 14a**,**b**), coinciding with shifts in the actin family from widely expressed members (*ACTB* and *ACGT*) to partial replacement with muscle-specific ones (*ACTA1* and *ACTA2*) (**Supplementary Fig. 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.

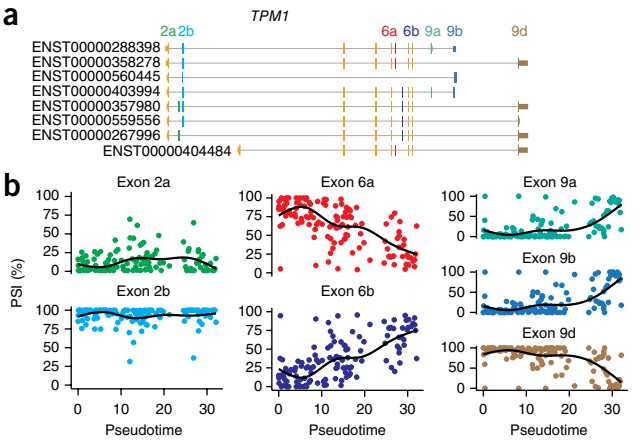

Figure 5 | Census enabled robust analysis of differential splicing during human myoblast differentiation. (**a**) Splicing structure of human gene *TPM1*, with the three alternatively spliced sets of exons highlighted. (**b**) Percent-spliced-in (PSI) values for *TPM1* alternative exons. PSI values were 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.

Census counts enabled 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. A recent study tracked single-cell gene expression from preimplantation 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 that, 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 embryos, consistent with little to no transcription from the embryonic genome (**Fig. 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 (**Fig. 6a**). Paternal X-chromosome inactivation manifested in female embryos by the 16-cell stage, with progressively fewer genes exhibiting contributions from the paternal X chromosome (**Fig. 6b**,**c**), except for genes known to escape inactivation (**Supplementary Fig. 15**).

The original study reported widespread stochastic monoallelic gene expression in individual cells[45]. This claim was challenged by an allele-specific expression analysis in embryonic stem cells that used a statistical model to attribute much of the apparent stochastic monoallelic expression to technical sources[47]. We tested whether using Census to estimate allelic transcript counts instead of allelic read counts would reduce observed stochastic monoallelic expression to expected levels. Consistent with the generative model used in ref. 47, the expected rate of monoallelic expression was near 100% for genes expressed at a single copy, and decreased with increasing expression (**Fig. 6d**). Of 6,608 'allele-informative' genes in the genome, 95.0% produced monoallelic transcript counts within the expected range. In contrast, only 77% of genes fell within the range obtained by fitting similar models to normalized

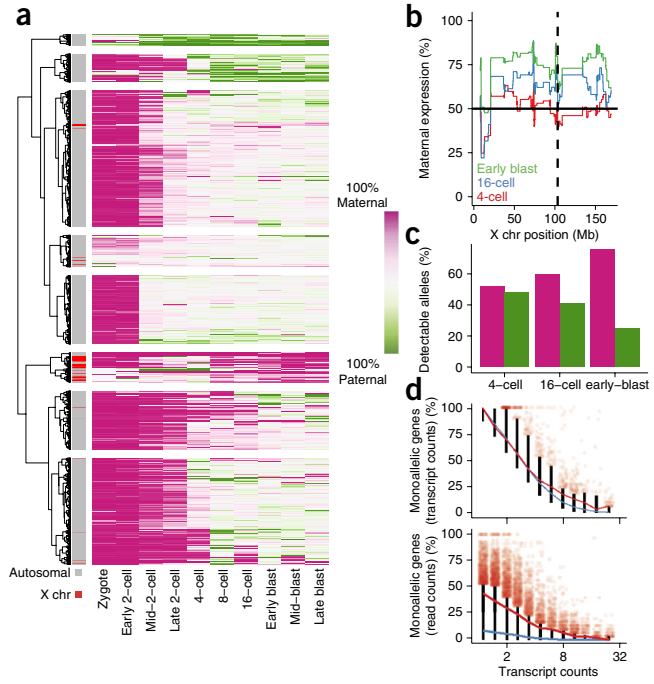

Figure 6 | Census detected shifts in allelic balance in single cells during embryogenesis. (**a**) A quasibinomial regression model detected changes in allelic balance in single cells as a function of embryo stage. (**b**) Spread of X-chromosome inactivation as measured by Census for female embryos at different stages (compare with ref. 45 figure 2b). (**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.

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 (owing 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 dropouts and other nuisance variation. Here we showed 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 one cDNA molecule in each cell's library. Through reanalysis of several data sets, we showed that this is the case with most current protocols, owing to mRNA capture rates lower than 50% and the 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 showed through extensive benchmarking that differential analysis results with Census counts were highly concordant with those from spike-ins. 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 Census' power, we developed three regression-based methods to detect gene regulatory changes. 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. To our surprise, branched cell trajectories arose 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 that encoding TPM1 showing a sequence of pseudotime-dependent shifts. The third method captures changes in allelic transcript counts derived with Census. By reanalyzing data from preimplantation embryos, we confirmed the authors' timing of transcriptional activation of the embryonic genome and X-chromosome inactivation[45]. In contrast to the original study, we did 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.

We expect that the use of normalized transcript counts, available through Census, will continue to unveil new mechanisms of gene regulation, including at the allele and isoform level, in development and disease.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

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

ACKNOWLEDGMENTS

We thank J. Shi and S. Xu for technical discussions, M. Kircher for cluster computation support, and J. Shendure, R. Hause, D. Cusanovich, B. Trapnell, J. Whitsett and members of the Trapnell laboratory for comments on the manuscript. This work was supported by US National Institutes of Health (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 a National Science Foundation (NSF) Graduate Research Fellowship.

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.-A.M. contributed to technical design. C.T. conceived the project. All authors wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Macosko, E.Z., Basu, A., Satija, R., Nemesh, J. & Shekhar, K. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).
2. Klein, A.M. *et al.* Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187–1201 (2015).
3. Shalek, A.K. *et al.* Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature* **498**, 236–240 (2013).
4. Grün, D., Kester, L. & van Oudenaarden, A. Validation of noise models for single-cell transcriptomics. *Nat. Methods* **11**, 637–640 (2014).
5. Finak, G. *et al.* MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* **16**, 278 (2015).
6. Jiang, L. *et al.* Synthetic spike-in standards for RNA-seq experiments. *Genome Res.* **21**, 1543–1551 (2011).
7. Fu, G.K., Hu, J., Wang, P.-H. & Fodor, S.P.A. Counting individual DNA molecules by the stochastic attachment of diverse labels. *Proc. Natl. Acad. Sci. USA* **108**, 9026–9031 (2011).
8. Hug, H. & Schuler, R. Measurement of the number of molecules of a single mRNA species in a complex mRNA preparation. *J. Theor. Biol.* **221**, 615–624 (2003).
9. Picelli, S., Faridani, O.R., Björklund, A.K. & Winberg, G. Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* **9**, 171–181 (2014).
10. Petropoulos, S. *et al.* Single-cell RNA-seq reveals lineage and X chromosome dynamics in human preimplantation embryos. *Cell* **165**, 1012–1026 (2016).
11. Segerstolpe, Å. *et al.* Single-cell transcriptome profiling of human pancreatic islets in health and type 2 diabetes. *Cell Metab.* **24**, 593–607 (2016).
12. Zeisel, A. *et al.* Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* **347**, 1138–1142 (2015).
13. Jaitin, D.A. *et al.* Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* **343**, 776–779 (2014).
14. Wu, A.R. *et al.* Quantitative assessment of single-cell RNA-sequencing methods. *Nat. Methods* **11**, 41–46 (2014).
15. Treutlein, B. *et al.* Dissecting direct reprogramming from fibroblast to neuron using single-cell RNA-seq. *Nature* **534**, 391–395 (2016).
16. Robinson, M.D., McCarthy, D.J. & Smyth, G.K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
17. Love, M.I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
18. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**, 381–386 (2014).
19. Kharchenko, P.V., Silberstein, L. & Scadden, D.T. Bayesian approach to single-cell differential expression analysis. *Nat. Methods* **11**, 740–742 (2014).
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* **6**, 468–478 (2010).
21. Buganim, Y. *et al.* Single-cell expression analyses during cellular reprogramming reveal an early stochastic and a late hierarchic phase. *Cell* **150**, 1209–1222 (2012).
22. Zhou, J.X. & Huang, S. Understanding gene circuits at cell-fate branch points for rational cell reprogramming. *Trends Genet.* **27**, 55–62 (2011).
23. Moignard, V. *et al.* Decoding the regulatory network of early blood development from single-cell gene expression measurements. *Nat. Biotechnol.* **33**, 269–276 (2015).
24. Marco, E. *et al.* Bifurcation analysis of single-cell gene expression data reveals epigenetic landscape. *Proc. Natl. Acad. Sci. USA* **111**, E5643–E5650 (2014).
25. Treutlein, B. *et al.* Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature* **509**, 371–375 (2014).
26. Hochegger, H., Takeda, S. & Hunt, T. Cyclin-dependent kinases and cell-cycle transitions: does one fit all? *Nat. Rev. Mol. Cell Biol.* **9**, 910–916 (2008).
27. Desai, T.J., Brownfield, D.G. & Krasnow, M.A. Alveolar progenitor and stem cells in lung development, renewal and cancer. *Nature* **507**, 190–194 (2014).
28. Chi, X., Garnier, G., Hawgood, S. & Colten, H.R. Identification of a novel alternatively spliced mRNA of murine pulmonary surfactant protein B. *Am. J. Respir. Cell Mol. Biol.* **19**, 107–113 (1998).
29. McCullagh, P. & Nelder, J.A. *Generalized Linear Models* 2nd edn. (CRC Press, 1989).
30. Shu, W. *et al.* Foxp2 and Foxp1 cooperatively regulate lung and esophagus development. *Development* **134**, 1991–2000 (2007).
31. Yin, Y. *et al.* An FGF-WNT gene regulatory network controls lung mesenchyme development. *Dev. Biol.* **319**, 426–436 (2008).
32. Shu, W., Yang, H., Zhang, L., Lu, M.M. & Morrisey, E.E. 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.* **276**, 27488–27497 (2001).
33. Wan, H. *et al.* Kruppel-like factor 5 is required for perinatal lung morphogenesis and function. *Development* **135**, 2563–2572 (2008).
34. Xu, Y. *et al.* C/EBPα is required for pulmonary cytoprotection during hyperoxia. *Am. J. Physiol. Lung Cell. Mol. Physiol.* **297**, L286–L298 (2009).
35. Okubo, T. & Hogan, B.L.M. Hyperactive Wnt signaling changes the developmental potential of embryonic lung endoderm. *J. Biol.* **3**, 11 (2004).
36. Shalek, A.K. *et al.* Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature* **510**, 363–369 (2014).
37. Darnell, J.E. Jr., Kerr, I.M. & Stark, G.R. Jak-STAT pathways and transcriptional activation in response to IFNs and other extracellular signaling proteins. *Science* **264**, 1415–1421 (1994).
38. Honda, K. *et al.* IRF-7 is the master regulator of type-I interferon–dependent immune responses. *Nature* **434**, 772–777 (2005).
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.* **201**, 1435–1446 (2005).
40. Lavin, Y. *et al.* Tissue-resident macrophage enhancer landscapes are shaped by the local microenvironment. *Cell* **159**, 1312–1326 (2014).
41. Welch, J.D., Hu, Y. & Prins, J.F. Robust detection of alternative splicing in a population of single cells. *Nucleic Acids Res.* **44**, e73 (2016).
42. Perrin, B.J. & Ervasti, J.M. The actin gene family: function follows isoform. *Cytoskeleton* **67**, 630–634 (2010).
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* **66**, 798–815 (2009).
44. Gunning, P., O'Neill, G. & Hardeman, E. Tropomyosin-based regulation of the actin cytoskeleton in time and space. *Physiol. Rev.* **88**, 1–35 (2008).
45. Deng, Q., Ramsköld, D., Reinius, B. & Sandberg, R. Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science* **343**, 193–196 (2014).
46. Bray, N.L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).
47. Kim, J.K., Kolodziejczyk, A.A., Ilicic, T., Teichmann, S.A. & Marioni, J.C. Characterizing noise structure in single-cell RNA-seq distinguishes genuine from technical stochastic allelic expression. *Nat. Commun.* **6**, 8687 (2015).
48. Amit, I. *et al.* Unbiased reconstruction of a mammalian transcriptional network mediating pathogen responses. *Science* **326**, 257–263 (2009).

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 α. 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 as

$$\text{Cell lysate}: \begin{cases} Y_{ij}^{l} \approx \alpha_i Y_{ij}^{c} \\ S_{ij}^{l} \approx \beta_i S_{ij} \end{cases},$$

where Y^l, S^l and 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 and the second subscript corresponds to gene index. Note that we cannot directly observe Y_{ij}^{c}, 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:

$$\text{cDNA}: \begin{cases} Y_{ij}^{d} = Y_{ij}^{l} \cdot \theta_i \\ S_{ij}^{d} = S_{ij}^{l} \cdot \theta_i \end{cases},$$

where Y^d and S^d are the cDNA counts of endogenous RNA, spike-in cDNA counts successfully converted from the corresponding transcript counts Y^l, and 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

$$\frac{Y_{ij}^{d}}{\sum_{j=1}^{n}(Y_{ij}^{d} + S_{ij}^{d})}$$

for endogenous RNA or

$$\frac{S_{ij}^{d}}{\sum_{j=1}^{n}(Y_{ij}^{d} + S_{ij}^{d})}$$

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

$$R_i = \gamma \sum_{j=1}^{n}(Y_{ij}^{d} + S_{ij}^{d})$$

from the distribution of relative cDNA abundances mentioned above which can be represented as

$$\text{Read counts}: \begin{cases} Y_{ij}^{r} \sim \text{multinomial}\left(\dfrac{Y_{ij}^{d}}{\sum_{j=1}^{n}(Y_{ij}^{d} + S_{ij}^{d})}, R_i^{e}\right) \\ S_{ij}^{r} \sim \text{multinomial}\left(\dfrac{S_{ij}^{d}}{\sum_{j=1}^{n}(Y_{ij}^{d} + S_{ij}^{d})}, R_i^{s}\right) \end{cases},$$

where R_i^e, R_i^s denotes the reads sampled for cDNA from the endogenous RNA or spike-in RNA in cell i, Y_{ij}^{r}, $S_{.j}^{r}$ 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.*[47], 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 (for example, 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}

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

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 an RNA recovery rate α_i. Therefore, we have

$$\text{Library}: Y_{ij}^{l} \sim \text{multinomial}(\rho_{\text{scaled}}, (\alpha_i)T_i)$$

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 ladder recovery rate is represented as 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 distribution, 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 **Supplementary Figure 4**.

doi:10.1038/nmeth.4150

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 based on spike-in ladder (see **Supplementary Note 1**), the probability to observe at least one copy of this transcript is $\rho = 1 - (1 - \theta)^s$. We assumed 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 nonzero TPM values. In order to robustly estimate θ, we assumed a constant capture rate for cells collected in each time point (lung or neuron experiment[15,25]) or the whole data set (other experiments) and pooled 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.,

$$\mathrm{TPM}_{ij} \propto \frac{Y_{ij}}{\sum_{j=1} Y_{ij}}$$

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_i} 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 $(1/\theta)$ 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. Although 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 assumed a value of 0.25, which is close to the lung and neuron experiments of refs. 25 and 15.

With an estimate of the total lysate mRNAs M_i in cell i, we simply rescaled 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 sc-RNA-seq assume that TPM is proportional to the true relative abundance in the cell lysis, i.e.,

$$\mathrm{TPM}_{ij} \propto \frac{Y_{ij}}{\sum_{j=1} Y_{ij}}$$

However, nonlinear amplification at any stage of the library construction protocol could distort this relationship. We can see this distortion when fitting the linear regression model, $\log(\mathrm{TPM}_{ij}) = k \log(Y_{ij}) + b$, to the spike-in data that recovers a value of k that deviates from 1, which indicates that $\mathrm{TPM}_{ij} \propto (Y_{ij})^k$. In practice, we found that k ranged 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 supplementary information for Trapnell and Cacchiarelli *et al.*[18] for further information). Transcript counts values were variance-stabilized[49] via the technique described by in ref. 49 before 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

$$\mathrm{NB}(\text{Census counts}) \sim \mathrm{sm.ns}(\text{Pseudotime})$$

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

$$\mathrm{NB}(\text{Census counts}) \sim \mathrm{sm.ns}(\text{Pseudotime}) + \text{Branch} + \mathrm{sm.ns}(\text{Pseudotime}) : \text{Branch}$$

assumes that the gene is a branch specific gene where : represents an interaction term between branch and transformed pseudotime,

and 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 gene is one that 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.*[25] data, Monocle reconstructed a trajectory with two branches (L_{AT1}, L_{AT2} for AT1 and AT2 lineages, respectively) and three states (S_{BP}, S_{AT1} or 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 were 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 assigned 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 3c** show 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^i_β) 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 **Supplementary Note 1**) to calculate the divergence ($D_i(t = 100) = x_{L1}(t = 100) - x_{L2}(t = 100)$) of the expression for gene i ($x_{L1}(t = 100)$, $x_{L2}(t = 100)$ 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 $D_i(t) \geq \varepsilon(t)$, and defines this time point as the branch time point, t^i_β, for that particular gene i.

Analysis of human skeletal muscle myoblasts. We used the human skeletal muscle myoblast (HSMM) data from our previous publication[18] to benchmark the performance of developmental tree reconstruction and pseudotime differentially expressed gene 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 that were differentially expressed between the four time points in relative abundance or recovered transcript counts scale were used to reduce dimension and order the cells. Transcript counts were variance-stabilized. The ordering of developmental trajectories between these two approaches was compared using Spearman correlation. Pseudotime tests were performed on both the relative abundance and transcript counts scale where the pseudotime-dependent genes were collected as those with q values less than 0.05 (Benjamini-Hochberg correction). The benchmark set was obtained from the permutation test based on a modified algorithm from the glm.perm package as previously described (see **Supplementary Note 1**, benchmarking differential expression analysis).

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 ϕ. That is, the Dirichlet encodes the frequencies of the isoforms π 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:

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

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

$$\text{Dirmultinomial}(Z_1, \ldots, Z_k \mid \pi, \phi) \sim 1$$

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 preimplantation 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

doi:10.1038/nmeth.4150

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:

$$P(x = k) = \binom{n}{k} p(p + k\phi)^{k-1}(1 - p - k\phi)^{n-k}$$

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

$$\text{quasibinomial}(Z_m, Z_p) \sim \text{stage}$$

and a null

$$\text{quasibinomial}(Z_m, Z_p) \sim 1$$

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-cell, 16-cell 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-chromosome genes. Embryos comprised of these cells were annotated as male.

To quantify the number of genes escaping X-chromosome 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.

We provide a table that describes all variables used in Census (**Supplementary Table 2a**), BEAM (**Supplementary Table 2b**), isoform switch analysis (**Supplementary Table 2c**) and allele specific analysis (**Supplementary Table 2d**).

Code availability. A version of monocle 2 (version: 1.99) used in this study is provided as **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 that can be used to reproduce all figures and data in this study.

Data availability. Eleven public sc RNA-seq data sets are used in this study, of which eight data sets used ERCC spike-in. Data sets with spike-in were lung: GSE52583 (ref. 25); noise model: GSE54695 (ref. 4); neuron reprogramming: GSE67310 (ref. 15); human preimplantation embryos: E-MTAB-3929 (ref. 10); pancreas: E-MTAB-5061 (ref. 11); cortex: http://linnarssonlab.org/cortex/ (ref. 12); marker-free: GSE54006 (ref. 13); and quantitative assessment data: GSE51254 (ref. 14). Data sets without spike-in were HSMM: GSE52529 (ref. 18); dendritic cell knockout: GSE41265 (ref. 36); and allele-specific gene expression: GSE45719 (ref. 45).

49. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106 (2010).
50. Yee, T.W. *Vector Generalized Linear and Additive Models* (Springer, 2015).
51. Katz, Y., Wang, E.T., Airoldi, E.M. & Burge, C.B. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat. Methods* **7**, 1009–1015 (2010).
52. Keane, T.M. *et al.* Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature* **477**, 289–294 (2011).
53. Corbel, C., Diabangouaya, P., Gendrel, A.-V., Chow, J.C. & Heard, E. Unusual chromatin status and organization of the inactive X chromosome in murine trophoblast giant cells. *Development* **140**, 861–872 (2013).
54. Yang, F., Babak, T., Shendure, J. & Disteche, C.M. Global survey of escape from X inactivation by RNA-sequencing in mouse. *Genome Res.* **20**, 614–622 (2010).