CellTag Indexing: a genetic barcode-based multiplexing tool for single-cell technologies

Chuner Guo[1,2,3], Brent A. Biddy[1,2,3], Kenji Kamimoto[1,2,3], Wenjun Kong[1,2,3], Samantha A. Morris[1,2,3,*]

[1]Department of Developmental Biology;
[2]Department of Genetics;
[3]Center of Regenerative Medicine.
Washington University School of Medicine in St. Louis. 660 S. Euclid Avenue, Campus Box 8103, St. Louis, MO 63110, USA

*Correspondence: s.morris@wustl.edu

ABSTRACT

Single-cell technologies have seen rapid advancements in recent years, along with new analytical challenges and opportunities. These high-throughput assays increasingly require special consideration in experimental design, sample multiplexing, batch effect removal, and data interpretation. Here, we describe a lentiviral barcode-based multiplexing approach, 'CellTag Indexing', where we transduce and label samples that can then be pooled together for downstream application and analysis. By introducing predefined genetic barcodes that are transcribed and readily detected, we can reliably read out sample identity via genomic or transcriptomic profiling, permitting the simultaneous assessment of cell grouping and transcriptional state. We validate and demonstrate the utility of CellTag Indexing by sequencing transcriptomes at single-cell resolution using a variety of cell types including mouse pre-B cells, primary mouse embryonic fibroblasts, human HEK293T cells, and mouse induced endoderm progenitors. Furthermore, we establish CellTag Indexing as a valuable tool for multiplexing direct lineage reprogramming perturbation experiments. We present CellTag Indexing as a broadly applicable genetic multiplexing tool that is complementary with existing single-cell RNA-sequencing and multiplexing strategies.

INTRODUCTION

Many advances have been made in developing single-cell technologies in recent years, providing unique opportunities to investigate biological entities and processes with unprecedented resolution. As single-cell platforms are increasingly adopted for a variety of assays, this has presented new challenges in experimental design and data analysis. Combining data obtained from different experimental batches into one integrated dataset is one such challenge that complicates data analysis with batch effects.

In many typical single-cell RNA-sequencing (scRNA-seq) configurations, multiple biological samples are assayed separately then combined in a single dataset. For example, samples are loaded onto different "wells" of a scRNA-seq platform as separate runs, prepared as individual libraries, and later pooled together computationally. In this scenario, distinguishing true biological differences from potential batch effects arising from technical variation is an analytical and computational challenge that may produce systematic errors if not properly addressed. It was recently demonstrated that batch effects can drive aberrant clustering of the same biological sample processed by two different instruments[1], demonstrating how single-cell data analysis can be complicated by measurement errors. A scRNA-seq experiment should ideally be approached with care to ensure technical variations are minimized experimentally, or properly corrected computationally.

Several algorithms exist to enable computational correction of batch effects[2–4]. These methods aim to minimize technical artifacts in existing datasets by regressing out known factors of variation. However, these approaches are limited in that they require prior knowledge of the factors that need to be regressed out, which may not always be the case. Alternatively, samples may be pooled together and subsequently demultiplexed, based on their natural genetic variation[5], given that the samples are genetically distinct. This is a powerful approach that allows for the multiplexing of ~20 samples, as long as their genotypes are known. However, if such information is not readily available or if the samples are closely biologically related, demultiplexing by genetic variation may not be optimal.

Recently, several “label-and-pool” approaches have been developed to mark a sample with a distinct identity prior to pooling and loading onto a single scRNA-seq run[6–8]. For example, cells can be tagged with barcoded antibodies[7], or chemically labeled with DNA oligonucleotides[8], such that cell identities can be read out in parallel with their transcriptomes. This type of approach minimizes technical variation experimentally and offers additional advantages of streamlined library preparation and reduced sequencing costs. Here, we introduce a methodology to multiplex biological samples via genetic labeling with lentiviral ‘CellTag Indexing’. In this approach, lentivirally-delivered barcodes are transcribed and detected via scRNA-seq, allowing the labelling and subsequent identification of different cell populations. This method facilitates the demultiplexing of sample identity from pooled single-cell transcriptomes. We validate CellTag Indexing via species mixing of genetically distinct populations, demonstrating the accurate demultiplexing of cell identity using this method. Furthermore, we establish CellTag Indexing as a valuable demultiplexing tool through its application to the analysis of direct lineage reprogramming, where we overexpress a candidate gene to investigate its effects on reprogramming fibroblasts to induced endoderm progenitors. We present CellTag Indexing as a broadly applicable tool, easily deployed in cell culture- and transplantation-based assays, that will be compatible with many single-cell analysis modalities.

RESULTS

CellTag Indexing is a lentiviral barcode-based tool for genetically labeling biological samples

Here, we present a lentiviral CellTag toolbox to label cells with transcribed DNA barcodes, acting as cell/sample identifiers that can be read out from single cell transcriptomes[9]. Briefly, the CellTag method utilizes lentiviruses encoding GFP and a SV40 polyadenylation signal, where an 8-base pair (bp) index sequence is located within the GFP UTR (Fig. S1A). Using a defined barcode as an index, the CellTag virus is used to transduce and genetically label a sample. Once transduced, the viral GFP sequence produces polyadenylated and indexed transcripts that are efficiently captured as part of standard scRNA-seq library preparation pipelines, supporting the demultiplexing of original identity in downstream analysis. A set of predefined barcodes can be used for

indexing, and conversely, a complex library of multiple random barcodes can be used to label each cell with a unique combination of indexes for tracking clonal dynamics[9].

Species mixing of genetically distinct cells validates CellTag Indexing to label and subsequently demultiplex independent samples

We previously demonstrated that predefined CellTags can be used to index fibroblasts that are spiked into a reprogramming cell population for benchmarking[9]. To demonstrate the efficacy of CellTag Indexing for the simultaneous labeling of multiple biological samples, we applied it to a species mixing experiment consisting of human HEK293T cells and mouse embryonic fibroblasts (MEFs), inspired by previous 'barnyard' mixing experiments to assess cell co-encapsulation rates in droplet-based scRNA-seq[10].

In this experimental validation, HEK293T and MEFs were transduced with CellTag Indexing viruses with two predefined barcodes, CellTag 'A' and CellTag 'B', respectively, for 24-48 hours. The transduced cells were cultured for an additional 72 hours, to permit CellTag expression, prior to collection and methanol fixation as previously described[11], with a portion plated separately to visualize lentiviral transduction efficiency (determined to be ~90% for both HEK293T and MEFs; Fig. S1B). After rehydration, an equal-proportion of CellTagged HEK293T and MEFs were pooled and loaded onto one single lane of the 10x Genomics Chromium Single Cell 3' v2 system. The cell pool was 'super-loaded' in order to promote multiplets formation, to permit the assessment of our approach in identification of mixed samples[5]. Following library preparation, a total of 18,159 cells were sequenced, with an average of 10,263 reads per cell, and an inferred doublet rate of ~15%. Using the 10x Cell Ranger pipeline, single transcriptomes classify into 9,357 single human cells (hg19), 7,456 single mouse cells (mm10), and 1,346 multiplets by alignment to a modified hg19-mm10 reference genome, where transcriptomes with total unique molecular identifiers (UMI, corresponding to transcripts) counts exceeding the 1st percentile of the distribution for both genomes classify as multiplets (Fig. 1B)[12].

We used the output from the 10x pipeline as a benchmarking standard to validate our demultiplexing approach (Fig. 1A), where CellTag expression is used to infer sample identity. CellTag sequences were extracted from raw reads and collapsed using a sequence clustering approach[13], generating a digital gene expression (DGE) matrix of UMI counts for each CellTag sequence. A filtering step was applied to the DGE matrix to remove additional noise arising from PCR and sequencing errors, followed by normalization and log transformation. Overall, CellTag expression is detected in 82.3% of all cells. We then demultiplexed the transcriptomes by using a simple hierarchical classification system, where a cell is classified as a multiplet if its expression is positive for both CellTags, as 'non-determined' if its expression is negative for both, and otherwise as either 'human' or 'mouse' when the appropriate CellTag is detected. Using a stringent threshold for detecting robust expression (Fig. S1C), we can classify the single-cell transcriptomes into 5,679 human cells, 5,080 mouse cells, 571 multiplets, and 6,829 non-determined cells (Fig. 1C).

Comparison of the 10x- and CellTag-based classification after removing non-determined cells shows excellent agreement (Figs. 1D-F), with a Cohen's kappa of 0.8668 (95% CI [0.8582, 0.8754]). Furthermore, cells designated as multiplets by both 10x and CellTag demonstrate a clear upward shift in the mean numbers of transcripts per cell (Figs. 1G&H), suggesting that they likely represent true multiplets.

CellTagging does not alter cell physiology

To test the effects of lentiviral CellTag transduction on normal cell physiology, we cultured HAFTL pre-B cells as previously described[14]. We transduced one sample with CellTag lentivirus, while another sample was not transduced. To investigate the potential effects of CellTagging on gene expression, we sequenced single cells on 10x Genomics Chromium platform, from which we obtained 3,939 CellTagged transcriptomes and 2,064 control transcriptomes after library preparation, sequencing, alignment, and filtering.

We observe that CellTagged and control transcriptomes share similar quality metrics, with comparable numbers of genes detected, numbers of transcripts detected, and percent of

mitochondrial transcripts per cell (Fig. 2A). After filtering, dimension reduction, and clustering via Seurat[15], we find that CellTagged transcriptomes and control transcriptomes are evenly interspersed together (Fig. 2B) with minimal independent clustering (Fig. 2C) and comparable cluster compositions (Fig. 2D), suggesting that they possess very similar expression profiles. Assessment of B cell-specific markers curated from the Mouse Cell Atlas dataset[16] reveals that the two samples have indistinguishable levels of expression both on a single-cell level (Figs. S2A&B) and when averaged across the subpopulations (Fig. 2E). Genome-wide comparison of gene expression of the two samples shows a strong linear association with an R^2 value of 0.9998 (Fig. 2F), confirming that CellTagging does not alter cell identity or physiology.

CellTag Indexing enables multiplexing of perturbation experiments

As our CellTag Indexing strategy utilizes genetic lentiviral labeling, it is particularly suited for multiplexing otherwise genetically identical samples subjected to different experimental treatments. We therefore applied CellTag Indexing to a perturbation experiment in an in house biological system of direct lineage reprogramming. Briefly, MEFs can be reprogrammed into induced endoderm progenitors (iEPs) by overexpressing transcription factors Foxa1 and Hnf4α[17,18], a process marked by an early upregulation of Insulin-like growth factor-binding protein 3 (Igfbp3) in reprogrammed cells by our previous scRNA-seq analysis[9]. Igfbp3 has been implicated in stem cell function, migration, and homeostasis, particularly in the liver[19] and the large intestine[20]. As iEPs have been previously shown to engraft both the liver[17] and the colon (Fig. 3A)[18,21], we sought to investigate the effect of its overexpression in the generation of iEPs.

We designed a perturbation experiment where iEPs were generated, as previously described[17,18], with or without Igfbp3 overexpression(Fig. 3A). Each sample was then allowed to reprogram and form iEP colonies (Fig. S3A), and then transduced with a predefined CellTag Index for 48 hours at week 4. iEPs were collected and methanol fixed[11] on day 53. We have previously demonstrated that CellTags are not silenced over a 4-week period of reprogramming[9]. After rehydration, an equal-proportion pool was made and loaded onto the 10x Genomics Chromium platform, resulting in 6,080 single

transcriptomes with an average of 28,091 reads per cell. Using a stringent classification threshold (Fig. S3B), we classify all transcriptomes into 502 control cells, 1,221 Igfbp3 cells, 17 multiplets, and 4,940 non-determined cells using our established demultiplexing pipeline. Overall, we detect CellTag expression in 35% of all cells. The purpose of this relatively sparse labeling here also serves to demonstrate that cluster identity can be inferred (Fig. S3C), following low-level transduction and analysis. For the remainder of this study, we excluded multiplets and non-determined cells in the following analysis.

We performed filtering by library size, number of genes expressed, and percentage of mitochondrial genes (Figs. S3D), followed by normalization and scaling as previously described[15]. After dimension reduction and clustering, analysis reveals four clusters (Fig. 3C). Interestingly, Igfbp3 cells predominantly cluster into three clusters (0, 2, and 3), while control cells consist of the majority of the remaining cluster, cluster 1 (Fig. 3B-D). Inspection of several known markers of iEP reprogramming shows that most cells have upregulated Apoa1, previously shown to mark reprogrammed iEPs[9], with very few fibroblasts remaining (Fig. 3E). Differential gene expression and gene list enrichment analysis[22,23] reveals that while control cells express markers associated with both small and large intestines, liver, and stomach, Igfbp3 cells separate into clusters that contain predominantly liver, bone marrow, and fibroblast signatures, respectively (Fig. 3E&F). The observed liver bias of Igfbp3 overexpression cells raises the possibility that its expression may drive reprogramming iEPs to adopt a more restricted lineage for their endoderm potential, as opposed to their control counterparts which become poised for differentiation toward several endodermal cell types. Findings such as these will assist in the design of more precise engineering of cell identity. Together, these results demonstrate the utility of CellTag Indexing for the accurate assessment of complex cell perturbation experiments.

DISCUSSION

Here, we introduce a broadly applicable and novel approach, CellTag Indexing, to multiplex biological samples for scRNA-seq, where each sample is genetically labeled with a predefined lentiviral GFP barcode to mark its biological identity. We demonstrate

that CellTagging does not interfere with cell identity, and validate the utility of CellTag Indexing via species mixing, showing that CellTag Indexing can be used to multiplex biological samples for scRNA-seq. We further demonstrate that CellTag Indexing can be used to multiplex biologically similar samples subjected to different experimental treatments, supporting experimental designs with minimal batch effects.

Our CellTag Indexing method provides the advantages of minimized technical variation by experimental design, broad compatibility with various single cell technologies, streamlined workflow and library preparation, reduced sequencing cost, and straightforward demultiplexing strategy. Furthermore, CellTag Indexing is designed for broad applications; its use of lentivirus as a labeling method represents a commonly used and very accessible biological tool with minimal setup costs and reagent requirements. CellTag Indexing conveniently utilizes GFP as a barcode carrier, which can act as a visual readout for estimating labeling efficiency. The GFP CellTag Index transcripts are abundantly expressed, and can be specifically amplified during library preparation to further increase detection efficiency.

Although we only demonstrate the multiplexing of two samples in our study here, it is possible to use CellTag Indexing for the simultaneous labeling of more than two samples, especially with the 'super-loading' strategy that was recently reported[5]. CellTag Indexing is also potentially compatible with single nucleus sequencing, as its transcripts are highly expressed and readily captured. Furthermore, cells labeled with CellTag Indexing can be cultured and subjected to additional assays prior to sequencing, for example in a completive transplant assay. We present CellTag Indexing here as a broadly-applicable tool complementary to existing methods for multiplexing and demultiplexing, providing a diverse panel of experimental and analytical strategies. As single-cell biology advances with increasing resolution and scale, future development in technologies will help guide studies to reveal deeper biological insights.

METHODS

Mice

Mouse embryonic fibroblasts were derived from the C57BL/6J strain (The Jackson Laboratory 000664). All animal procedures were based on animal care guidelines approved by the Institutional Animal Care and Use Committee.

Cell culture

HEK293T and mouse embryonic fibroblasts were cultured in Dulbecco's Modified Eagle Medium (Gibco) supplemented with 10% Fetal Bovine Serum (Gibco), 1% penicillin/streptomycin (Gibco), and 55 µM 2-mercaptoethanol (Gibco). HAFTL pre-B cells were cultured in RPMI1640 without phenol red (Lonza) supplemented with 10% charcoal/dextran-treated FBS (Hyclone) and 55 µM 2-mercaptoethanol (Gibco)[14].

Immunostaining and quantification

Transduced HEK293T and MEFs were cultured on a 4-chamber culture slide (Falcon) for 24 hr prior to fixation in 4% paraformaldehyde and staining in 300 nM DAPI in PBS. The slide was then mounted in ProLong Gold Antifade Mountant (Invitrogen). Images were acquired on a Nikon eclipse Ts2 inverted microscope.

For automatic quantification, images of CellTagged HEK 293T and MEF were processed with a custom python script to count GFP positive/negative cells. The proportion of GFP positive cells was calculated from DAPI and GFP images. First, DAPI images were transformed into binary images by thresholding fluorescent signal. The threshold values were determined by the Otsu method. The binary nucleus image was processed by watershed segmentation to separate individual cell areas completely. Inappropriately sized objects were filtered to remove noise and doublet cells. The intensity of the GFP signal per individual cell area was then quantified to distinguish between GFP positive cells and negative cells. These processes were run with Python 3.6.1 and its libraries: scikit-image 0.13.0, numpy 1.12.1, matplotlib 2.0.2, seaborn 0.8.1, jupyter 1.0.0.

Generation of iEPs

Mouse embryonic fibroblasts were converted to iHeps/iEPs as previously described[17,18]. Briefly, fibroblasts were prepared from E13.5 embryos, cultured on gelatin, and serially transduced with Hnf4α-t2a-Foxa1 retrovirus over the course of 5 days, followed by culture on collagen in hepato-medium, which is DMEM:F-12 (Gibco) supplemented with 10% FBS, 1% penicillin/streptomycin, 55 μM 2-mercaptoethanol, 10 mM nicotinamide (Sigma-Aldrich), 100 nM dexamethasone (Sigma-Aldrich), 1 ug/mL insulin (Sigma-Aldrich), and 20 ng/ml epidermal growth factor (Sigma-Aldrich). For the perturbation experiment, reprogramming 3-factor iEPs were serially transduced with Igfbp3 retrovirus on day 8 and day 10.

Lenti- and retrovirus production

Lentiviruses were produced by transfecting HEK293T cells with lentiviral pSMAL vector and packing plasmids pCMV-dR8.2 dvpr (Addgene plasmid 8455) and pCMV-VSV-G (Addgene plasmid 8454) using X-tremeGENE 9 (Sigma-Aldrich). Viruses were collected 48 and 72 hr after transfection. Retroviruses were similarly produced, with retroviral pGCDNsam vector and packaging plasmid pCL-Eco (Imgenex).

CellTag Indexing

CellTag lentiviral constructs were generated by introducing an 8bp variable region into the 3' UTR of GFP in the pSmal plasmid[24] using a gBlock gene fragment (Integrated DNA Technologies) and megaprimer insertion. Individual clones were picked and Sanger sequenced to generate predefined barcodes.

scRNA-seq procedure

10x Genomics Chromium Single Cell 3' Library & Gel Bead Kit v2, Chromium Single Cell 3' Chip kit v2, and Chromium i7 Multiplex kit were used according to the manufacturer's protocols. cDNA libraries were quantified on the Agilent 2200 TapeStation and sequenced on Illumina HiSeq 2500.

CellTag demultiplexing

Reads containing the CellTag sequence were extract from the processed, filtered, and unmapped reads BAM files produced in intermediate steps of the 10x pipeline. Reads that contained the CellTag "motif" were identified: “CCGGTNNNNNNNNGAATTC”. Following extraction of reads from the BAM file, a custom gawk script was utilized to parse the output, capturing the Read ID, Sequence, Cell Barcode, UMI, CellTag Sequence, and Aligned Genes of each read. CellTags and surrounding motifs aligning to genes were filtered out. This parsed output was then used to construct a Cell Barcode x CellTag UMI matrix. CellTags were grouped by Cell Barcodes and then the number of unique UMIs for each Cell Barcode, CellTag pair was counted. The matrix was then filtered to remove any cell barcodes not found in the filtered Cell Ranger file. Finally, the CellTags were filtered to remove any that were represented by ≤1 UMI. The construction and filtering of the CellTag UMI matrix accomplished using a custom R script. CellTag sequences were collapsed using Starcode with the sphere clustering algorithm[13], after which the DGE was correspondingly collapsed in R. Then, a filtering of predefined CellTag sequences was applied to the DGE, followed by normalization and log transformation. Each cell barcode was then assigned a classification by a simple hierarchical algorithm, where a cell is classified as a multiplet if its expression is positive for both CellTags, as ‘non-determined’ if its expression is negative for both, and otherwise as either ‘human’ or ‘mouse’ when either CellTag is detected, using a threshold of 1 for detecting robust expression.

scRNA-seq analysis

The R package Seurat[15] was used for data processing and visualization. For the iEP dataset, we removed cells with a low number of genes detected (<200), cells with a high number of UMI detected (>100000), and cells with a high proportion of UMI counts attributed to mitochondrial genes (>0.2), resulting in a filtered matrix of 1,652 cells and 16,153 genes. The filtered expression matrix was then subjected to log normalization with the default scale factor of 10,000, variable gene detection resulting in 940 variable genes (average expression between 0.2 and 3, dispersion greater than 0.5), and scaling to remove unwanted sources of variation driven by number of detected UMIs and mitochondrial gene expression. Linear dimension reduction was performed, followed by

clustering using the first 15 principal components as the input and a resolution of 0.6. Non-linear dimension reduction and visualization by tSNE, and differential gene expression analysis were performed using the default parameters. Individual lists of gene markers for each cluster was then uploaded into the interactive Enrichr gene list enrichment analysis tool to assess enrichment of upregulated genes in mouse tissues from BioGPS, using the Mouse Gene Atlas gene list under the Cell Types category. The results were visualized using the bar graph option, sorted by combined score of p-value and z-score.

ACKNOWLEDGEMENTS

We would like to thank members of the Morris lab for critical discussion, John Dick for the kind gift of the pSMAL-GFP construct[24], and Genome Technology Access Center for sequencing. . This work was funded by National Institute of General Medical Sciences R01 GM126112; National Human Genome Research Institute R21 HG009750 (to S.A.M and R.D.M); Chan Zuckerberg Initiative Grants HCA-A-1704-01646 and HCA2-A-1708-02799; The Children's Discovery Institute of Washington University in St. Louis and St. Louis Children's Hospital MI-II-2016-544; and Washington University Digestive Diseases Research Core Center, National Institute of Diabetes and Digestive and Kidney Diseases P30 DK052574, all to S.A.M. S.A.M is supported by a Vallee Scholar Award, B.A.B is supported by a NIH T32 Training Grant, and K.K is supported by a Japan Society for the Promotion of Science Postdoctoral Fellowship for Research Abroad.

REFERENCES

1. Hicks, S. C., Townes, F. W., Teng, M. & Irizarry, R. A. Missing data and technical variability in single-cell RNA-sequencing experiments. *Biostatistics* (2017). doi:10.1093/biostatistics/kxx053
2. Ritchie, M. E. *et al.* Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* (2015). doi:10.1093/nar/gkv007
3. Shaham, U. *et al.* Removal of batch effects using distribution-matching residual networks. *Bioinformatics* (2017). doi:10.1093/bioinformatics/btx196
4. Haghverdi, L., Lun, A. T. L., Morgan, M. D. & Marioni, J. C. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.* (2018). doi:10.1038/nbt.4091
5. Kang, H. M. *et al.* Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat. Biotechnol.* (2018). doi:10.1038/nbt.4042
6. Cao, J. *et al.* Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science (80-.).* (2017). doi:10.1126/science.aam8940
7. Stoeckius, M. *et al.* Cell 'hashing' with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. *bioRxiv* (2017). doi:10.1101/237693
8. Gehring, J., Park, J. H., Chen, S., Thomson, M. & Pachter, L. Highly Multiplexed Single-Cell RNA-seq for Defining Cell Population and Transcriptional Spaces. *bioRxiv* 315333 (2018). doi:10.1101/315333
9. Biddy, B. A., Waye, S. E., Sun, T. & Morris, S. A. Single-cell analysis of clonal dynamics in direct lineage reprogramming: a combinatorial indexing method for lineage tracing. *bioRxiv* (2017). at <http://biorxiv.org/content/early/2017/04/28/127860>
10. Macosko, E. Z. *et al.* Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161,** 1202–1214 (2015).
11. Alles, J. *et al.* Cell fixation and preservation for droplet-based single-cell transcriptomics. *BMC Biol.* (2017). doi:10.1186/s12915-017-0383-5
12. Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* (2017). doi:10.1038/ncomms14049

13. Zorita, E., Cuscó, P. & Filion, G. J. Starcode: Sequence clustering based on all-pairs search. *Bioinformatics* (2015). doi:10.1093/bioinformatics/btv053
14. Bussmann, L. H. *et al.* A Robust and Highly Efficient Immune Cell Reprogramming System. *Cell Stem Cell* (2009). doi:10.1016/j.stem.2009.10.004
15. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* (2018). doi:10.1038/nbt.4096
16. Han, X. *et al.* Mapping the Mouse Cell Atlas by Microwell-Seq. *Cell* (2018). doi:10.1016/j.cell.2018.02.001
17. Sekiya, S. & Suzuki, A. Direct conversion of mouse fibroblasts to hepatocyte-like cells by defined factors. *Nature* **475,** 390–393 (2011).
18. Morris, S. A. *et al.* Dissecting engineered cell types and enhancing cell fate conversion via CellNet. *Cell* **158,** 889–902 (2014).
19. Steiger-Luther, N. C., Darwiche, H., Oh, S.-H., Williams, J. M. & Petersen, B. E. Insulin-like growth factor binding protein-3 is required for the regulation of rat oval cell proliferation and differentiation in the 2AAF/PHX model. *Hepat. Med.* **2010,** 13–32 (2010).
20. D'Addio, F. *et al.* Circulating IGF-I and IGFBP3 Levels Control Human Colonic Stem Cell Function and Are Disrupted in Diabetic Enteropathy. *Cell Stem Cell* (2015). doi:10.1016/j.stem.2015.07.010
21. Miura, S. & Suzuki, A. Generation of Mouse and Human Organoid-Forming Intestinal Progenitor Cells by Direct Lineage Reprogramming. *Cell Stem Cell* (2017). doi:10.1016/j.stem.2017.08.020
22. Chen, E. Y. *et al.* Enrichr: Interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* (2013). doi:10.1186/1471-2105-14-128
23. Kuleshov, M. V. *et al.* Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* (2016). doi:10.1093/nar/gkw377
24. Lu, R., Neff, N. F., Quake, S. R. & Weissman, I. L. Tracking single hematopoietic stem cells in vivo using high-throughput sequencing in conjunction with viral genetic barcoding. *Nat. Biotechnol.* (2011). doi:10.1038/nbt.1977

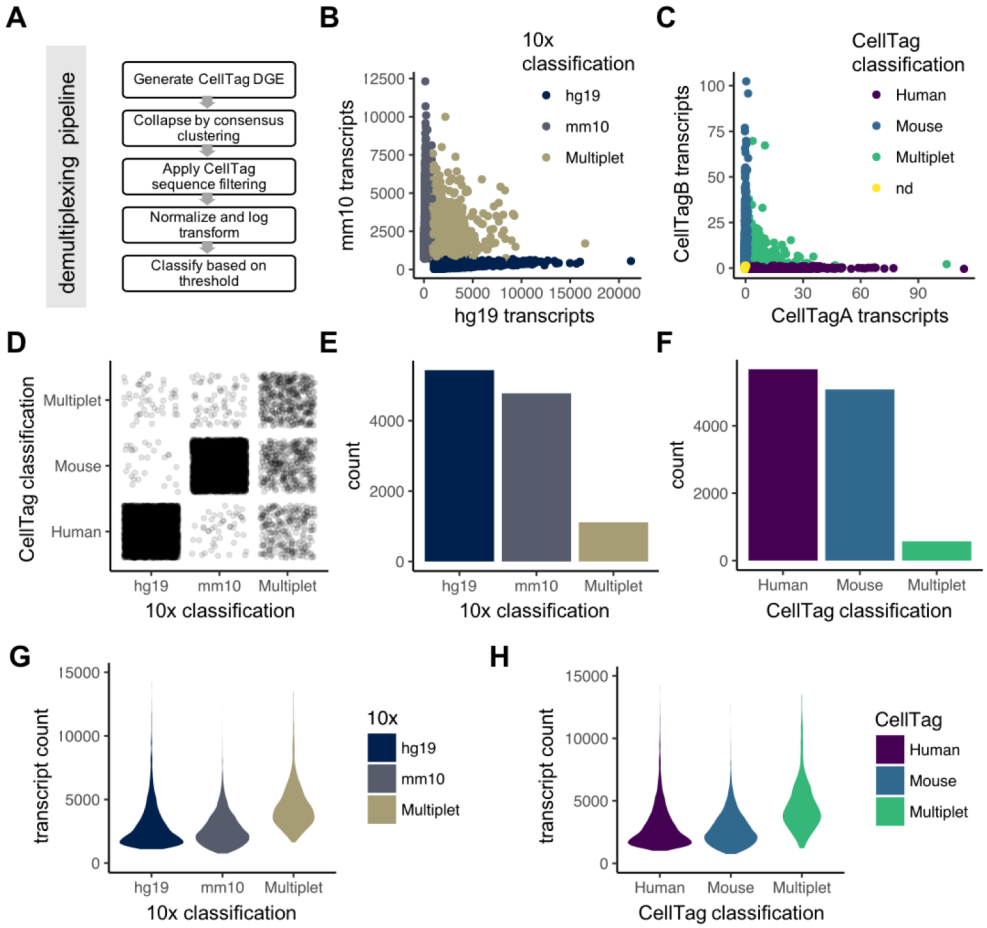

Figure 1. Species mixing experiment validates CellTag Indexing to label and subsequently demultiplex independent samples. (A) Overview of the CellTag demultiplexing pipeline. (B) Classification of the species mixed transcriptomes by 10x Cell Ranger into 9,357 single human cells (hg19), 7,456 single mouse cells (mm10), and 1,346 multiplets. (C) Classification of the species mixed transcriptomes by CellTag Index demultiplexing into 5,679 human cells, 5,080 mouse cells, 571 multiplets, and 6,829 non-determined cells. (D) Comparison of 10x and CellTag classifications after removing non-determined cells. (G) Distribution of total number of transcript detected in each single transcriptome, grouped by 10x classification. (H) Distribution of total number of transcript detected in each single transcriptome, grouped by CellTag classification.

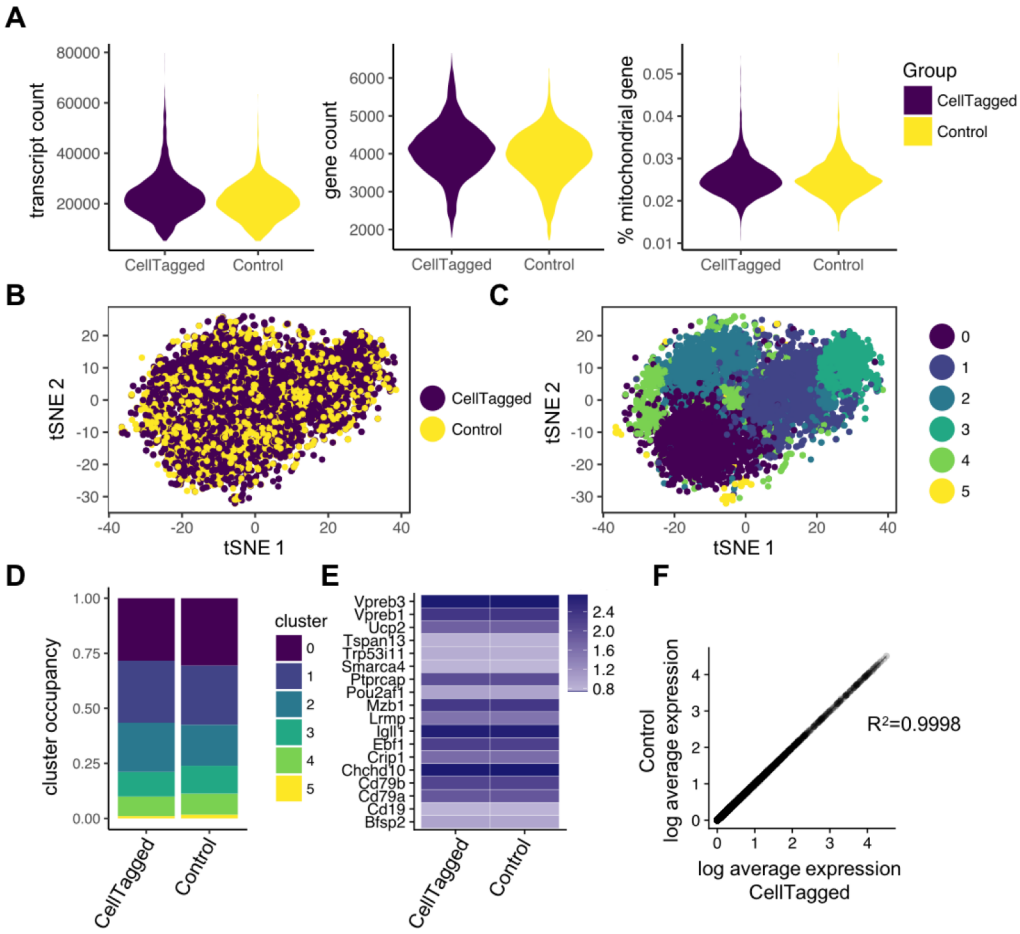

Figure 2. CellTagging does not alter cell physiology. (A) Distribution of total number of transcripts detected, total number of genes detected, and percent of detected transcripts attributed to mitochondrial genes in each single transcriptome, grouped by sample identity. (B) tSNE visualization of single transcriptomes, colored by sample identity. (C) tSNE visualization of single transcriptomes, colored by cluster identity. (D) Cluster composition of each sample, colored by cluster identity. (E) Log average gene expression for B cell markers, grouped by sample identity. (F) Log average gene expression for all detected genes in control cells and CellTagged cells, fitted with linear regression with a R^2 value of 0.9998.

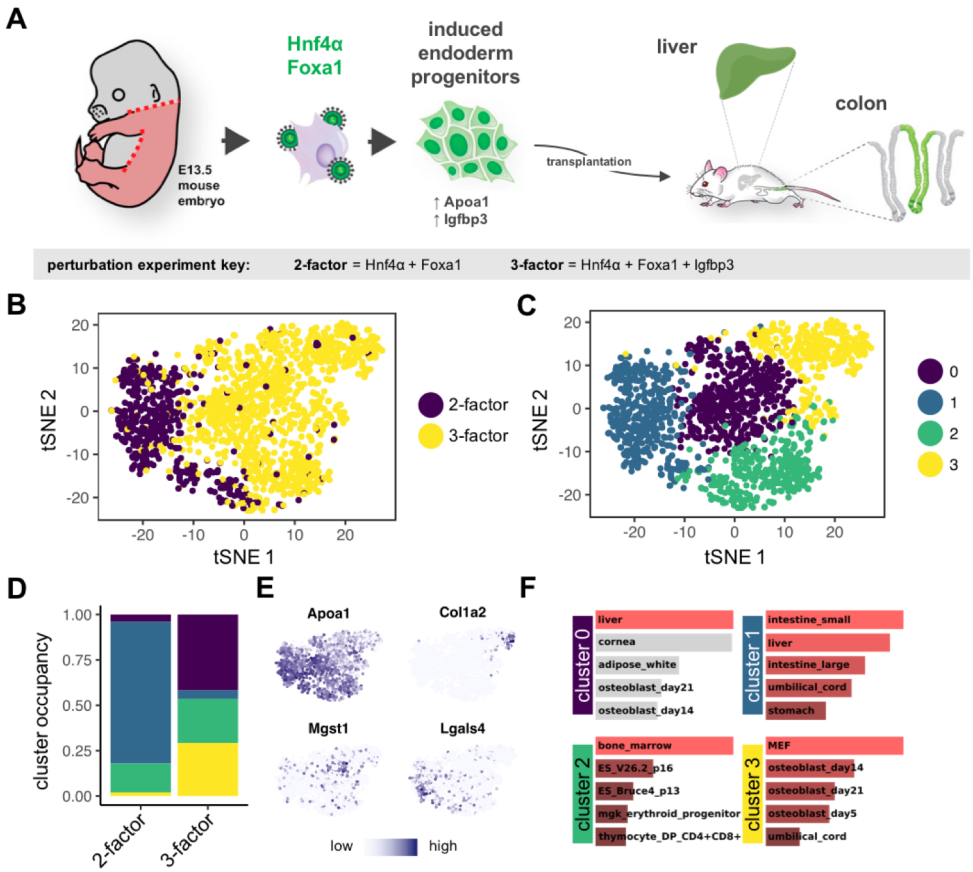

Figure 3. CellTag Indexing enables multiplexing of direct lineage reprogramming perturbation experiments. (A) Schematic of iEP generation and its potential to engraft both the liver and the large intestine; perturbation experiment setup is such that two-factor iEPs were generated by overexpressing Foxa1 and Hnf4α, and three-factor iEPs were generated by overexpression Foxa1, Hnf4α, and Igfbp3. (B) tSNE visualization of single transcriptomes showing two samples with minimal overlap, colored by sample identity. (C) tSNE visualization of single transcriptomes separated into four clusters, colored by cluster identity. (D) Cluster composition of each sample, with control cells largely located to cluster 1, and Igfbp3 cells consisting of clusters 0, 2, and 3. (E) Expression patterns of iEP marker Apoa1, fibroblast marker Col1a2, liver marker Mgst1, and intestinal marker Lgals4. (F) Enrichment of BioGPS mouse tissue upregulated gene lists by each cluster.

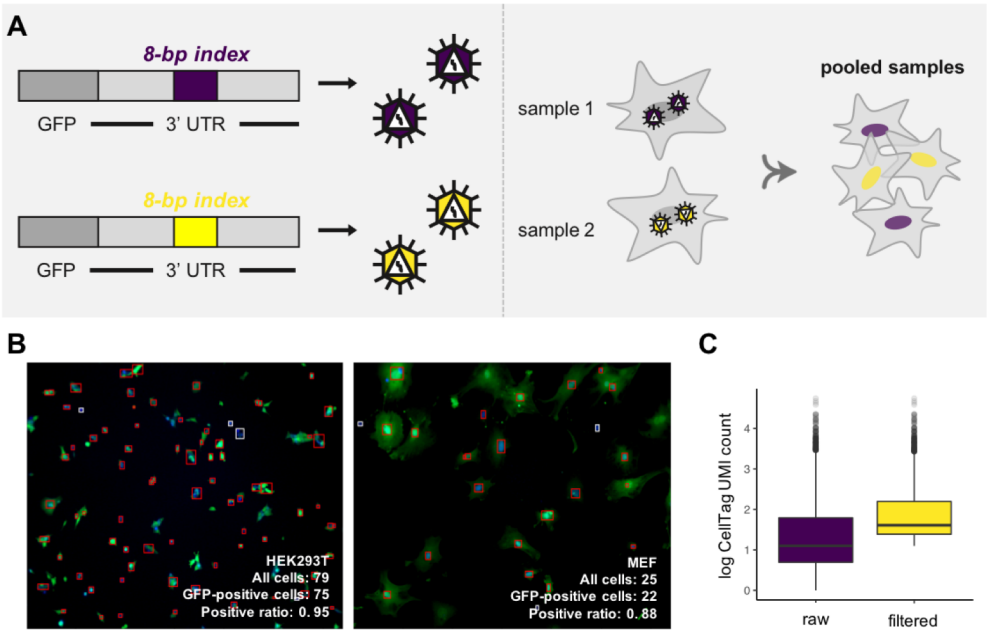

Figure S1. Species mixing experiment validates CellTag Indexing to label and subsequently demultiplex independent samples. (A) Schematic of CellTag design and sample multiplexing strategy. (B) Fluorescent microscopy images of CellTagged HEK293T (left panel) and MEFs (right panel) with automated quantification, where red boxes designate automatically-detected GFP-positive cells, and white boxes designate automatically-detected GFP-negative cells. (C) Log CellTag transcript count before and after filtering.

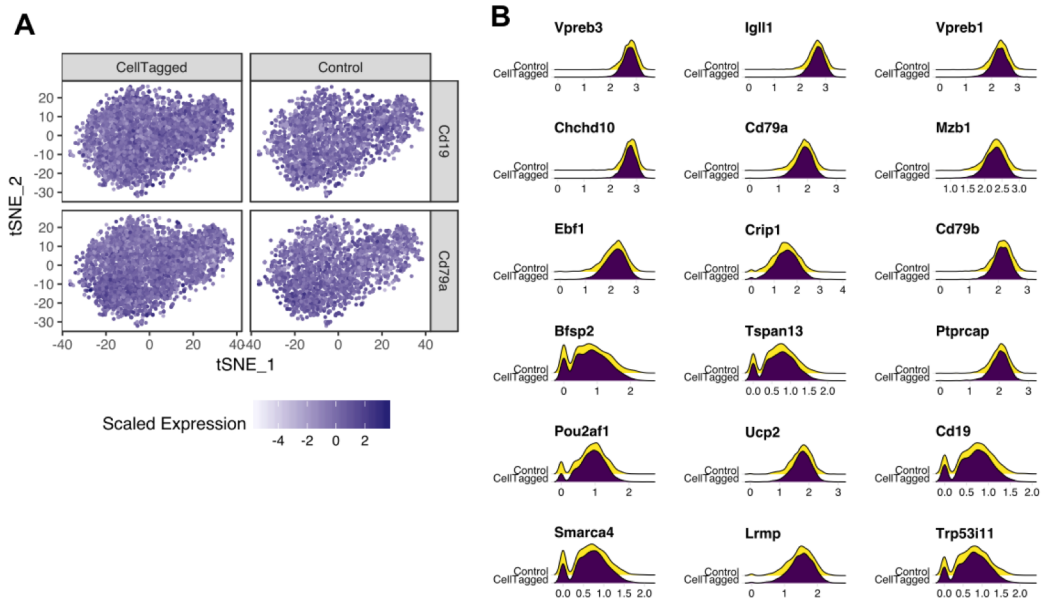

Figure S2. CellTagging does not alter cell physiology. (A) Gene expression heatmaps of B cell markers Cd19 and Cd79a in CellTagged and control single cells. (B) Single-cell gene expression distribution of eighteen B cell markers in CellTagged and control single cells.

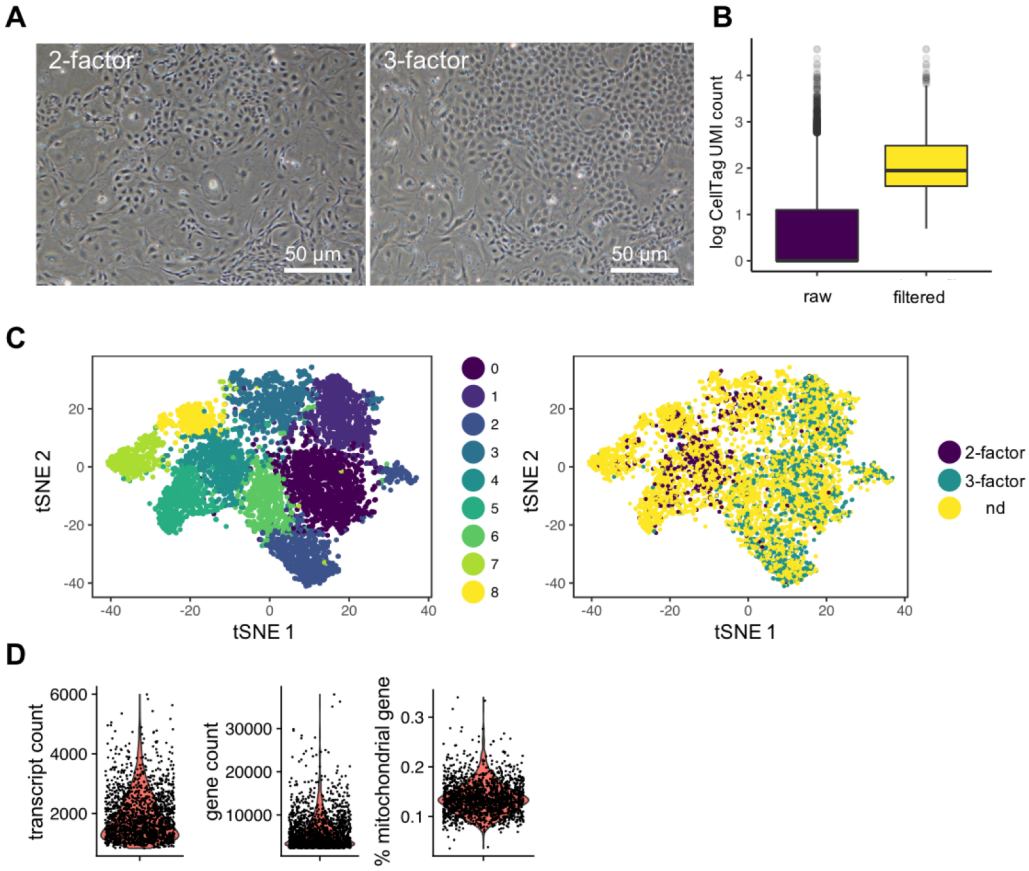

Figure S3. CellTag Indexing enables multiplexing of direct lineage reprogramming perturbation experiments. (A) Transmitted light microscopy images of day 21 two-factor and three-factor iEPs. Scale bar, 50 µm. (B) Log CellTag transcript count before and after filtering. (C) tSNE visualization of the perturbation experiment when including non-determined (nd) cells, colored by cluster identity (left panel) and sample identity (right panel). (D) Distribution of total number of transcript detected, total number of gene detected, and percent of detected transcripts attributed to mitochondrial genes in each single transcriptome.