nature methods

BRIEF COMMUNICATION

https://doi.org/10.1038/s41592-019-0392-0

Multiplexed detection of proteins, transcriptomes, clonotypes and CRISPR perturbations in single cells

Eleni P. Mimitou[1], Anthony Cheng[2,3], Antonino Montalbano[4,5], Stephanie Hao[1], Marlon Stoeckius[1], Mateusz Legut[4,5], Timothy Roush[4,5], Alberto Herrera[6], Efthymia Papalexi[4,5], Zhengqing Ouyang[2,3,7], Rahul Satija[4,5], Neville E. Sanjana[4,5], Sergei B. Koralov[6] and Peter Smibert[1*]

Multimodal single-cell assays provide high-resolution snapshots of complex cell populations, but are mostly limited to transcriptome plus an additional modality. Here, we describe expanded CRISPR-compatible cellular indexing of transcriptomes and epitopes by sequencing (ECCITE-seq) for the high-throughput characterization of at least five modalities of information from each single cell. We demonstrate application of ECCITE-seq to multimodal CRISPR screens with robust direct single-guide RNA capture and to clonotype-aware multimodal phenotyping of cancer samples.

High-throughput single-cell RNA sequencing (scRNA-seq) has rapidly progressed from a tremendous technical achievement to a standard tool for phenotypic interpretation of complex biological systems. Recently, substantial progress has been made in combining readouts of other modalities with scRNA-seq in high-throughput assays, including genome sequence, chromatin accessibility, methylation, immunophenotype (reviewed in ref. [1]) and synthetic markers of cell lineage (reviewed in ref. [2]). Additionally, several approaches have recently been reported that allow detection of CRISPR-mediated perturbations along with the transcriptome of single cells using specialized vectors that link the expression of single-guide RNAs (sgRNAs) to separate transcripts that can be captured by standard scRNA-seq methods[3–6]. Collectively, these methods enable the use of scRNA-seq as an unbiased readout of pooled CRISPR-based genetic screens, but all current methods suffer from limitations related to the need to determine the identity of the guide by a proxy polyadenylated transcript[7].

Previously, we and others have layered detection of proteins on top of scRNA-seq to enable integration of robust and well-characterized protein markers with unbiased transcriptomes of single cells[8,9]. Our method, cellular indexing of transcriptomes and epitopes by sequencing (CITE-seq) is compatible with oligo-dT based scRNA-seq approaches and enables simultaneous protein detection using DNA oligo-labeled antibodies against cell surface markers. Given that protein levels are typically much higher than corresponding messenger RNAs (mRNAs), detection of proteins via antibody-derived tags (hereafter called protein tags) is a more robust measure of gene expression. In a series of experiments, we demonstrated the value of multimodal analysis to reveal phenotypes that could not be discovered using scRNA-seq alone[8].

Here, we extend the use of CITE-seq and the related Cell Hashing method for multiplexing and doublet detection[10], to 5′ capture-based scRNA-seq methods, exemplified by the 10x Genomics 5P/V(D)J system, allowing the detection of surface proteins together with the scRNA-seq and clonotype features. Oligos partially complementary to the gel bead-associated template switch oligos (TSO) in the 10x Genomics 5P/V(D)J kit were covalently conjugated to antibodies as described[10] and used to label cells. Annealing and extension during the reverse-transcription reaction associates the cell barcode and unique molecular identifier (UMI) from the gel bead oligo with the antibody tag in parallel with the mRNAs in the same droplet (Fig. 1a) (see Methods). Separate detection of expanded CRISPR-compatible cellular indexing of transcriptomes and epitopes by sequencing (ECCITE-seq) protein tags and cell hashtags is achieved using different amplification handles[10].

We further adapt the system to enable direct and robust capture of sgRNAs from existing guide libraries and commonly used vectors compatible with pooled cloning. In contrast to commonly used 3′ tag scRNA-seq methods, the 10x Genomics 5P workflow appends the barcode via TSO, using a soluble poly(dT) oligo to prime reverse transcription, opening up the possibility of adding custom reverse-transcription primers to sequences of interest. sgRNAs have a structure that lends themselves to direct capture: a variable region at the 5′ end and an invariant scaffold at the 3′ end[11,12]. We leveraged the scaffold as an annealing site for an additional reverse-transcription primer, which after copying the variable guide sequence and template switching with the bead-derived TSO during reverse transcription, acquires a cell barcode and UMI in parallel with other modalities (mRNA, protein tags, hashtags) (Fig. 1a). A mixture of human and mouse cells transduced with different sets of non-targeting sgRNAs was well-resolved by transcriptome, surface protein and sgRNA content, demonstrating the specificity of this approach (Fig. 1b, Supplementary Fig. 1a and Supplementary Tables 1 and 2).

To illustrate the detection of six modalities (transcriptome, T-cell receptor (TCR) α/β and γ/δ, surface protein, sample identity by hashtags and sgRNA) in a single experiment (Supplementary Fig. 1b), we generated a cell mixture comprising human peripheral blood mononuclear cells (PBMCs), two human T-cell lymphoma lines (MyLa and Sez4) and mouse NIH-3T3 cells that had been transduced with a library of non-targeting sgRNA-generating constructs (Fig. 1c and Supplementary Table 2). Cell hashtags specific to human cells were used to distinguish the three human samples, and the hashtag distribution was consistent with transcriptome-based clustering (Fig. 1c(i)). ECCITE-seq antibodies directed against human or mouse CD29 label cells according to their species of origin (Fig. 1c(ii)), illustrating the ability of ECCITE-seq to

[1]Technology Innovation Laboratory, New York Genome Center, New York, NY, USA. [2]The Jackson Laboratory for Genomic Medicine, Farmington, CT, USA. [3]Department of Genetics and Genome Sciences, University of Connecticut Health Center, Farmington, CT, USA. [4]New York Genome Center, New York, NY, USA. [5]Department of Biology, New York University, New York, NY, USA. [6]Department of Pathology, New York University School of Medicine, New York, NY, USA. [7]Institute for Systems Genomics and Department of Biomedical Engineering, University of Connecticut, Storrs, CT, USA. *e-mail: psmibert@nygenome.org

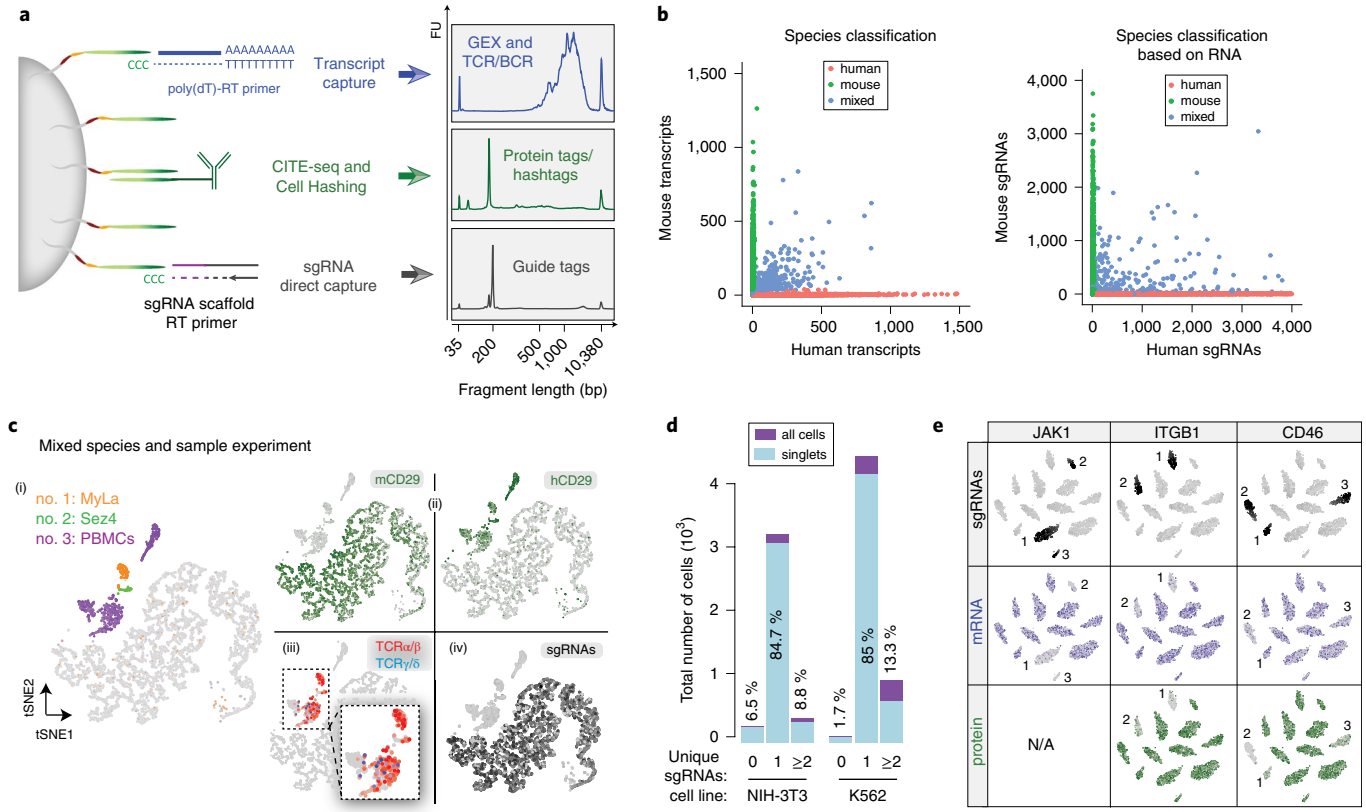

Fig. 1 | ECCITE-seq allows simultaneous detection of transcriptome, proteins, clonotypes and CRISPR perturbations. a, Schematic overview of the multiple cellular modalities captured by ECCITE-seq. RT, reverse transcription; bp, base pair. **b**, Species-mixing experiment. Left: number of transcripts associated with each cell barcode (red, >90% human reads; green, >90% mouse reads; blue, >10% human and mouse (multiplet). Right: sgRNA reads associated with each cell barcode. Points are colored on the basis of species classifications using transcripts. **c**, Transcriptome-based clustering of single-cell expression profiles of the mixed human and mouse sample (total cells = 5,935), illustrating the five modalities of ECCITE-seq: Transcriptome, Cell Hashing (i), Protein (ii), T-cell antigen receptors (α/β: red, γ/δ: blue) (iii) and sgRNAs (iv). **d**, Number of cell barcodes before (purple) or after (blue) removal of cell doublets assigned zero, one or two unique guides per cell in two independent experiments and cell lines. **e**, K562 cells were clustered on the basis of normalized and scaled sgRNA counts. Highlighted are counts for sgRNAs (black) targeting the indicated gene and counts for respective mRNA (blue) and protein tags (green), for all cells assigned to single guides ($n = 4{,}120$). FU, fluorescence units.

detect differentially expressed proteins within a sample. Clonotypes for TCR α/β (following 10x protocol) and TCR γ/δ (custom adaptation, see Methods) were detected in the PBMC and lymphoma cell clusters (Fig. 1c(iii)). Finally, guide tags, derived directly from sgRNA molecules were specifically and robustly detected only in mouse cells (Fig. 1c(iv)). The use of Cell Hashing together with sgRNA detection allowed us to distinguish between apparent 'doublets' where cells have been infected with two viruses ($n = 325/390$), from doublets resulting from co-encapsulation of two cells in the same droplet ($n = 65/390$) (Fig. 1d). sgRNA capture was highly efficient, with sgRNAs detected in 93.5% of mouse cells (Fig. 1d), in proportions consistent with genomic DNA-based detection from bulk cells (Supplementary Fig. 1c).

ECCITE-seq is designed to enable interrogation of single-cell transcriptomes together with surface protein markers in the context of CRISPR screens. To illustrate this, we infected K562 cells with a CRISPR library comprising guides targeting genes encoding cell surface markers (CD29 and CD46), intracellular signaling molecules (JAK1 and p53), as well as two non-targeting controls (Supplementary Table 1). We leveraged the Cell Hashing feature to remove cell doublets and observed very high rates of guide capture (confident detection of guide sequences in 98.3% of cells), in proportions consistent with genomic DNA-based detection (Fig. 1d and Supplementary Fig. 1d). Clustering on the basis of sgRNA counts of cells assigned to one guide revealed 13 distinct clusters, corresponding to the 13 guides in the experiment. Loss of expression of target genes at the level of mRNA and protein was readily apparent for *ITGB1* (the gene encoding CD29 protein) and *CD46* (Fig. 1e) and similarly apparent at the mRNA level for *JAK1*. To demonstrate that the capture of additional modalities has no detrimental effect on transcript capture, we performed scRNA-seq alone on the same aliquot of cells and confirmed no reduction in transcripts per cell (Supplementary Fig. 1e).

Cellular perturbations measured at transcript and protein level by ECCITE-seq reveal important features to consider, exemplified by CD46: most cells have detectable levels of protein, which collapse in cells with targeting sgRNAs (Supplementary Fig. 1f). mRNA reduction is also apparent in cells with targeting sgRNAs, albeit less notably. Many cells have undetectable levels of *CD46* mRNA even in the absence of targeting guides, probably reflecting the high drop-out rates of scRNA-seq and the increased sensitivity that comes with protein detection.

The low drop-out of protein detection[8,9] suggests that ECCITE-seq could be more sensitive in detecting expression phenotypes than scRNA-seq alone. To test this for single genes, we compared clusters assigned to each given guide to the two non-targeting clusters and determined the *P* value of detecting the expected gene-expression change in randomly-sampled cells ranging from 10–100 per group. (Supplementary Fig. 1g). The number of cells needed to detect the direct consequence of a given perturbation is

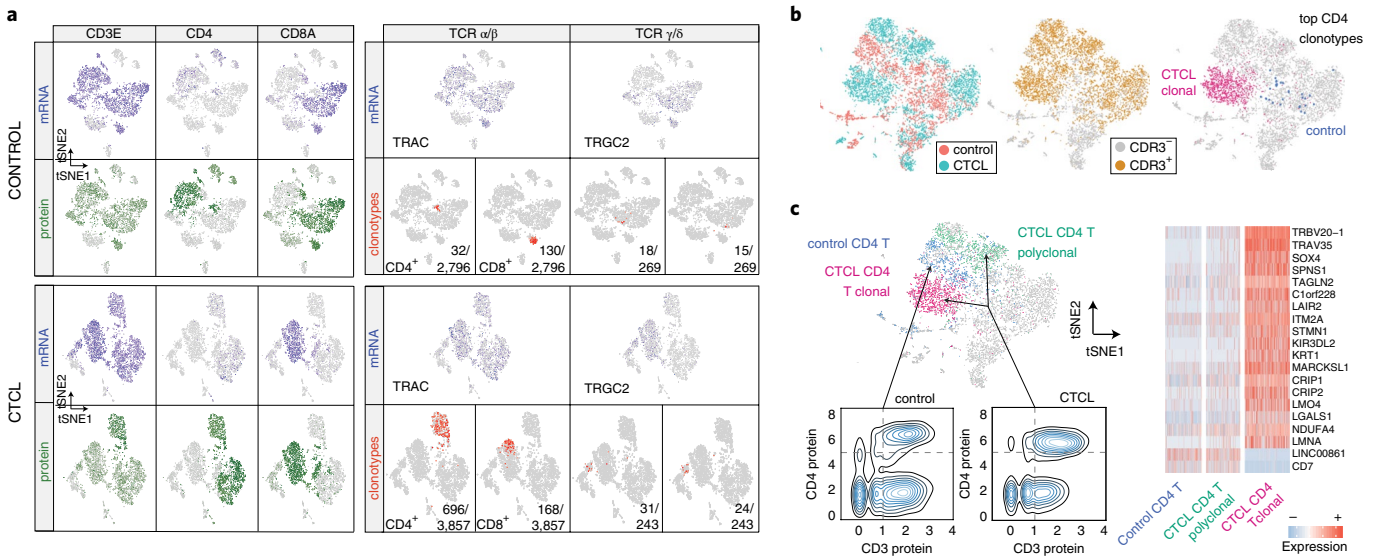

Fig. 2 | ECCITE-seq couples clonotype determination with immunophenotyping. a, Transcriptome-based clustering of PBMCs from healthy donor (top) and a patient with CTCL (bottom) after removing cell doublets (3,733 and 3,866 cells respectively). Projected is the mRNA (blue) and protein (green) signal for CD3, CD4 and CD8, as well as the most abundant CD4+ or CD8+ TCR α/β or two most abundant TCR γ/δ clonotypes (red). **b**, Transcriptome-based clustering of the combined dataset after merging, depth normalization and cell alignment. Highlighted from left to right is sample identity, productive TCR rearrangement and most abundant CD4+ TCRα/β clonotype. **c**, Differentially expressed genes between CD3+ CD4+ T clusters of control and CTCL PBMCs, as defined by in silico gating based on CD3 and CD4 protein counts coupled with clonotype determination.

markedly reduced when using protein detection as a readout compared to mRNA, increasing the numbers of perturbations that can be assessed for a given number of cells. Additionally, as exemplified by CD46, the gene-expression change triggered by two out of three sgRNAs (CD46.1 and CD46.3) was confidently detected only at the level of protein, even when considering all cells assigned to these sgRNAs. In practical terms, future applications of this technology will rely on detection of changes in gene-expression signatures and it stands to reason that these signatures will be more robust with protein components.

We next constructed a 49-marker panel of ECCITE-seq antibodies to deeply profile PBMCs from a healthy donor and a patient with cutaneous T-cell lymphoma (CTCL) (Fig. 2, Supplementary Fig. 2 and Supplementary Table 3) and prepared libraries for hashtags, protein tags, TCR α/β, TCR γ/δ and transcriptome. After hashtag demultiplexing to remove doublets, cells were clustered on the basis of transcriptome (Fig. 2a and Supplementary Fig. 2). The majority of markers showed enrichment at the level of both protein and RNA (not shown) in expected clusters, consistent with our previous 3′ CITE-seq results[8]. We additionally recovered TCR α/β and γ/δ clonotype information for both the control and CTCL samples. Select markers and clonotypes are shown in Fig. 2a and Supplementary Fig. 2.

For further comparative analysis, cells from both samples were computationally merged[13] and clustering based on either RNA or protein showed agreement in detecting most cell sub-populations and their gene-expression signatures (Supplementary Fig. 3). in silico gating based on CD3 and CD4 protein levels coupled with clonotypic information enabled differential gene-expression analysis comparing monoclonal T cells with polyclonal T cells from both the patient and the healthy donor sample (Fig. 2b,c). This analysis reveals a distinct gene-expression signature of the malignant CTCL cells, consistent with previous studies[14], and illustrates the power of ECCITE-seq to combine immunophenotype, clonotype and transcriptome information.

The enhancements to the CITE-seq toolkit enable detailed phenotypic and functional characterization of single cells. The recovery of clonotype information together with surface protein marker expression allowed fine separation of specific cell populations of interest, enabling careful determination of molecular phenotypes. Analogous to the use of TCR clonotype information in this study, we have recently used expressed mutations to define and further characterize clonal populations in scRNA-seq datasets (genotyping of transcriptomes[15]), an approach that could readily be combined with ECCITE-seq. The method we describe is inherently customizable and we envisage additional oligo-tagged ligands, such as peptide-loaded major histocompatibility complexes for detecting specific TCRs, labeled antigens for detection of antigen specific B cells or antibodies directed against intracellular proteins being added to future iterations of this system. The combination of Cell Hashing together with direct sgRNA capture will enhance perturbation screens with single-cell readouts by allowing the analysis of greater numbers of cells for a given budget. The 'super-loading' afforded by this knowledge will additionally drive down the per-cell cost of single-cell CRISPR screens, which will also require fewer cells per guide to detect expression phenotypes that feature both protein and mRNA. The modular nature of ECCITE-seq allows the tailoring of readouts of such screens, potentially enabling the investigator to interrogate panels of transcripts and proteins of interest in response to their perturbations in addition to, or instead of, the transcriptome. This is in line with the high-dimensional phenotyping of multiple proteins in CRISPR-based pooled screens using Pro-Codes and CyTOF as readout[16]. While this method can more economically achieve precise quantification of intracellular and extracellular protein levels in millions of single cells, it cannot interrogate the single-cell transcriptome simultaneously, it lacks the scalability of DNA barcoding and requires sgRNA cloning in special constructs. ECCITE-seq is readily applicable with minor modifications to any sgRNA library with a 3′ invariant scaffold sequence and, by allowing direct capture of sgRNA molecules, overcomes documented problems of barcode swapping events observed with Perturb-seq[7]. While this work was under review, a conceptually similar method to capture sgRNAs in the context of scRNA-seq was described in ref. [17]. Direct guide capture allows compatibility with applications

using multiple different guides per cell; for example, combinatorial screens targeting more than one gene per cell[18,19] or lineage tracing using multiple homing sgRNAs[20]. Our approach additionally provides a roadmap for targeted capture of specific RNA molecules including non-polyadenylated transcripts.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data availability and code availability are available at https://doi.org/10.1038/s41592-019-0392-0.

Received: 29 October 2018; Accepted: 14 March 2019;
Published online: 22 April 2019

References

1. Hu, Y. et al. *Front. Cell. Dev. Biol.* **6**, 28 (2018).
2. Kester, L. & van Oudenaarden, A. *Cell Stem Cell* **23**, 166–179 (2018).
3. Jaitin, D. A. et al. *Cell* **167**, 1883–1896.e15 (2016).
4. Adamson, B. et al. *Cell* **167**, 1867–1882.e21 (2016).
5. Dixit, A. et al. *Cell* **167**, 1853–1866.e17 (2016).
6. Datlinger, P. et al. *Nat. Methods* **14**, 297–301 (2017).
7. Hill, A. J. et al. *Nat. Methods* **15**, 271–274 (2018).
8. Stoeckius, M. et al. *Nat. Methods* **14**, 865–868 (2017).
9. Peterson, V. M. et al. *Nat. Biotechnol.* **35**, 936–939 (2017).
10. Stoeckius, M. et al. *Genome Biol.* **19**, 224 (2018).
11. Jinek, M. et al. *Science* **337**, 816–821 (2012).
12. Cong, L. et al. *Science* **339**, 819–823 (2013).
13. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. *Nat. Biotechnol.* **36**, 411–420 (2018).
14. Fanok, M. H. et al. *J. Investig. Dermatol.* **138**, 1116–1125 (2018).
15. Nam, A. S. et al. High throughput droplet single-cell Genotyping of Transcriptomes (GoT) reveals the cell identity dependency of the impact of somatic mutations. Preprint at https://doi.org/10.1101/444687 (2018).
16. Wroblewska, A. et al. *Cell* **175**, 1141–1155.e16 (2018).
17. Replogle, J. M. et al. Direct capture of CRISPR guides enables scalable, multiplexed, and multi-omic Perturb-seq. Preprint at https://doi.org/10.1101/503367 (2018).
18. Shen, J. P. et al. *Nat. Methods* **14**, 573–576 (2017).
19. Najm, F. J. et al. *Nat. Biotechnol.* **36**, 179–189 (2018).
20. Kalhor, R. et al. *Science* **361**, 893 (2018).

Acknowledgements

We thank N. Ødum (University of Copenhagen) for the kind gift of cell lines. Samples from patients were obtained with the help of M.B. Natan Zommer and J.-A. Latkowski. Work in S.B.K.'s laboratory was supported by the NIH R01 grant no. HL-125816, the Colton Center for Autoimmunity, funding from the Hirschl/Weill-Coulier Trust and a grant from the Spatz Foundation. Work in the NYGC Technology Innovation lab was supported by the NIH R21 grant no. HG-009748 to P.S. and the Chan Zuckerberg Initiative grant no. HCA-A-1704-01895 to P.S. and R.S. N.E.S. is supported by NYU and NYGC startup funds, NIH/NHGRI (R00HG008171 and DP2HG010099), NIH/NCI (R01CA218668), DARPA (D18AP00053), the Sidney Kimmel Foundation, the Melanoma Research Alliance, and the Brain and Behavior Foundation. M.L. is supported by a Hope Funds for Cancer Research postdoctoral fellowship. Work in Z.O.'s laboratory was supported by NIH R35 grant GM124998. We thank L. Yang, W. Stephenson, S. Jaini and K. Pandit for helpful discussions. We thank B. Fritz from 10x Genomics for providing kits for development of 5P compatible CITE-seq reagents and B. Yeung and K. Nazor from BioLegend for providing some of the unconjugated antibodies used in this study.

Author contributions

E.P.M. and P.S. conceived and designed the study with input from A.M., S.H., M.S., E.P., R.S., N.E.S. and S.B.K. E.P.M. performed all experiments, aided by S.H. M.S. A.C., A.H., Z.O. and S.B.K. provided samples and performed analysis on the CTCL study. A.M., M.L., T.R. and N.E.S. worked with E.P.M. to design and generate CRISPR libraries and provide experimental and analytical guidance. E.P.M. and P.S. wrote the paper.

Competing interests

M.S. and P.S. are listed as co-inventors on a patent application related to this work (US provisional patent application 62/515–180).

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41592-019-0392-0.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to P.S.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

See Protocol Exchange (https://doi.org/10.1038/protex.2019.025) and Supplementary Protocol for a step-by-step protocol for ECCITE-seq.

Antibody-oligo conjugates. Antibodies used for CITE-seq and Cell Hashing were obtained as purified, unconjugated reagents from BioLegend and were covalently and irreversibly conjugated to barcode oligos by iEDDA-click chemistry as previously described[10,21]. See Supplementary Tables 3–5 for a list of antibodies, clones and barcodes used for ECCITE-seq.

Cell staining with barcoded antibodies. Cells were stained with barcoded antibodies as previously described for CITE-seq[8] and Cell Hashing[10]. Briefly, approximately 1.5–2 million cells per sample were resuspended in 1× CITE-seq staining buffer (2% BSA, 0.01% Tween in PBS) and incubated for 10 min with Fc receptor block (TruStain FcX, BioLegend) to block FC receptor-mediated binding. Subsequently, cells were incubated with mixtures of barcoded antibodies for 30 min at 4 °C. Antibody concentrations were 1 µg per test, as recommended by the manufacturer (BioLegend) for flow cytometry applications. For some highly expressed markers, tags can take up unacceptably high proportions of the protein-tag libraries. In these cases (determined empirically from previous experiments), we reduced the concentration of the oligo-tagged antibodies in the panel by diluting with un-tagged antibody. Oligo-labeled CD44 and CD45 were diluted 1:10 and therefore used at an effective concentration of 0.1 µg per stain. After staining, cells were washed three times by resuspension in PBS containing 2% BSA and 0.01% Tween, followed by centrifugation (300*g* 5 min at 4 °C) and supernatant exchange. After the final wash, cells were resuspended in PBS and filtered through 40 µm cell strainers.

ECCITE-seq on a 10x Genomics instrument. Stained and washed cells were loaded into a 10x Genomics single-cell V(D)J workflow and processed according to the manufacturer's instructions with the following modifications:

(1) 12 pmol of an reverse-transcription primer complementary to sgRNA scaffold sequences was spiked into the reverse-transcription reaction (only when sgRNA capture was desired). gd_RT_v4: AGCAAGTGAGAAGCATCGTGT-CAAAGCACCGACTCGGTGCCAC.
(2) During the complementary DNA (cDNA) amplification step, 1 pmol of hashtag additive (GTGACTGGAGTTCAGACGTGTGCTC), 1 pmol of guide-tag additive (AGCAAGTGAGAAGCATCGTGTC) (only when sgRNA capture was desired) and 2 pmol of protein-tag additive primers (CCTTG-GCACCCGAGAATTCC) were spiked into the cDNA amplification PCR.
(3) Following PCR, 0.6× solid phase reversible immobilization (SPRI) beads were used to separate the large cDNA fraction derived from cellular mRNAs (retained on beads) from the protein-tag-, hashtag- and guide-tag-containing fraction (in supernatant). The complementary DNA fraction was processed according to the 10x Genomics Single-Cell V(D)J protocol to generate the transcriptome library and the TCR α/β library. To amplify TCR γ/δ transcripts we implemented a strategy similar to TCR α/β approach from 10x Genomics with a two-step PCR: during target enrichment 1 we used SI-PCR (AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACAC-GACGCTC) and a mix of R1_hTRDC (AGCTTGACAGCATTGTACTTCC) and R1_hTRGC (TGTGTCGTTAGTCTTCATGGTGTTCC), followed by target enrichment 2 with a generic P5 oligo (AATGATACGGCGACCAC-CGAGATCTACAC) and a mix of R2_hTRDC (TCCTTCACCAGACAAGC-GAC) and R2_hTRGC (GATCCCAGAATCGTGTTGCTC). cDNA and TCR (α/β and γ/δ) enriched libraries were further processed according to the 10x Genomics Single-Cell V(D)J protocol.
(4) An additional 1.4× reaction volume of SPRI beads was added to the protein-tag/hashtag/guide-tag fraction from step 3, to bring the ratio up to 2.0×. Beads were washed with 80% ethanol, eluted in water and an additional round of 2.0× SPRI performed to remove excess single-stranded oligonucleotides carried over from the cDNA amplification reaction. After final elution, separate PCR reactions were set up to generate the protein-tag library (SI-PCR and RPI-x primers), the hashtag library (SI-PCR and D7xx_s) and the guide-tag library (SI-PCR and Next_nst_x). The protein-tag and hashtag libraries were prepared as previously described[10]. Following the cDNA amplification, the sgRNA sequences are converted to an Illumina library by amplification with smRNA_nst_x (v.3): CAA-GCAGAAGACGGCATACGAGATxxxxxxxxGTGACTGGAGTTCCTTG-GCACCCGAGAATTCCATTCTAGCTCTAAAAC or Next_nst_x (v4): CAAGCAGAAGACGGCATACGAGATxxxxxxxxGTCTCGTGGGCTCGGA-GATGTGTATAAGAGACAGTATTTCTAGCTCTAAAAC together with the SI-PCR primer. 'x' nucleotides indicate the sample index sequenced by the Illumina i7 index read. Before the final library PCR, sgRNA molecules can be further enriched by performing extra rounds of amplification with guide-tag additive and SI-PCR primers.

Libraries were pooled to desired quantities and sequenced on either an Illumina HiSeq 2500 (rapid run flowcell: recipe 26 cycles read 1, 8 cycles index, 39 cycles read 2) or on a NovaSeq 6000 (S2 flowcell: recipe 26 cycles read 1, 8 cycles index, 91 cycles read 2). Reads were trimmed as required for downstream processing. A detailed and regularly updated point-by-point protocol for CITE-seq, Cell Hashing, ECCITE-seq and future updates can be found at www.cite-seq.com and on the Nature Protocol Exchange.

Cells. The samples from the patient and the control were collected at New York University Langone Medical Center in accordance with protocols approved by the New York University School of Medicine Institutional Review Board and Bellevue Facility Research Review Committee (IRB no. i15–01162). Patients with CTCL were diagnosed according to the WHO classification criteria. After written informed consent was obtained, peripheral blood samples were harvested. PBMCs were isolated from the blood of patients and healthy controls by gradient centrifugation using Ficoll-Paque PLUS (GE Healthcare) and Sepmate-50 tubes (Stemcell). Buffy coat PBMCs were collected and washed twice with PBS 2% FBS and cryopreserved in freezing medium (40% Roswell Park Memorial Institute medium (RPMI) 1640, 50% FBS and 10% DMSO). Cryopreserved PBMCs were thawed for 1–2 min in a 37 °C water bath, washed twice in warm PBS 2% FBS and resuspended in complete medium (RPMI 1640 supplemented with 10% FBS and 2 mM L-Glut). Control and CTCL PBMCs were stained with a 49-antibody panel (Supplementary Table 3) and Cell Hashing antibodies (Supplementary Table 5), before loading into two separate 10x Genomics Chromium lanes.

The Sez4 cell line is derived from the blood of a patient with Sézary syndrome[22], and the MyLa 2059 line is derived from a plaque biopsy sample of a patient with mycosis fungoides[23]. Sez4 cells were cultured in RPMI 1640 medium with 2 mM L-glutamine, 1% Pen/Strep, 500 units per ml of rh IL-2 (Corning) and 10% human serum. MyLa 2059 cells were cultured in RPMI 1640 medium with 2 mM L-glutamine, 1% Pen/Strep and 10% fetal bovine serum.
All cells were incubated at 37 °C, 5% CO_2 in a humidified incubator. The cells were cryopreserved in 90% FBS 10% DMSO and aliquots of 1–1.5 million cells were thawed on the day of the experiment. PBMCs were obtained cryopreserved from AllCells and used immediately after thawing. NIH-3T3 and HEK293FT cells expressing non-targeting sgRNAs were maintained according to standard procedures in Dulbecco's Modified Eagle's Medium (Thermo Fisher) supplemented with 10% fetal bovine serum (Thermo Fisher) and 1 µg ml^{-1} puromycin, at 37 °C with 5% CO_2. K562 cells expressing targeting and non-targeting guides were maintained in RPMI supplemented with 10% fetal bovine serum and 1 µg ml^{-1} puromycin, at 37 °C with 5% CO_2.

Lentivirus production and transduction. DNA oligos encoding the sgRNAs were individually synthesized (Integrated DNA Technologies) and cloned into the lentiviral transfer vector LentiCRISPR v.2 (ref. [24]) (Addgene Plasmid: 52961). Equal amounts of each sgRNA vector were mixed and packaged into lentiviral particles through transfection with packaging plasmids in HEK293FT cells, as previously described[25].

For transduction of HEK293FT, the lentiviral guide pool consisted of ten non-targeting human guides in one experiment and 10 non-targeting and 11 gene-targeting human guides in another experiment (Supplementary Table 1). For transduction of K562, the pool consisted of 2 non-targeting and 11 targeting human guides (Supplementary Table 1). For transduction of NIH-3T3, the pool consisted of 10 non-targeting mouse guides (Supplementary Table 2). NIH-3T3, HEK293FT and K562 cells were infected at multiplicity of infection = 0.05 and selected and maintained in 1 µg ml^{-1} puromycin. NIH-3T3 cells used in the proof-of-principle experiment were maintained in culture for several weeks, allowing drift in the representation of guides. Following transduction, K562 cells were stored in liquid nitrogen and were allowed to grow for 2 days before the ECCITE-seq run.

Single-cell data processing. Fastq files from the 10x libraries with four distinct barcodes were pooled together and processed using the cellranger count pipeline, v.2.1.1. Reads were aligned to the GRCh38 (human healthy and CTCL PBMC datasets) or hg19-mm10 concatenated reference (human–mouse experiment). For protein-tag, hashtag and guide-tag quantification, we used a previously developed tag quantification pipeline (v.1.3.2), available at https://github.com/Hoohm/CITE-seq-Count, run with default parameters (maximum Hamming distance of 1). For the TCR libraries, fastq files from the 10x libraries with four distinct barcodes were pooled together, processed using the cellranger vdj pipeline and reads were aligned to the GRCh38 reference genome.

Seurat. Normalization and downstream analysis of RNA data were performed using the Seurat R package (v.2.3.0)[13], which enables the integrated processing of multimodal single-cell datasets. Protein-tag, hashtag and guide-tag raw counts were normalized using centered log ratio transformation, where counts were divided by the geometric mean of the corresponding tag across cells and log-transformed[8]. For demultiplexing based on hashtag or guide-tag counts we used the HTODemux function within the Seurat package as described[10]. To calculate the significance in detecting the target gene-expression change between the targeting guide clusters and the non-targeting clusters we used FindAllMarkers with maximum cell number ranging from 10–100, in ten sampling iterations for each cell number. For the TCR libraries, productive clonotypes were filtered and

their raw counts were inserted into the Seurat object under a new assay slot. Raw counts were normalized using centered log ratio transformation and scaled. For comparison between the healthy donor and CTCL data, both Seurat objects were merged and depth-normalized when performing cell alignment (or batch normalization) using RunCCA with a default parameter of 30 canonical vectors[13]. The top ten aligned components were used for visualization with t-SNE (t-distributed stochastic neighbor embedding) as well as clustering with modularity optimization. The top 20 genes upregulated in each cluster (FindAllMarkers) was used to label the cluster. For protein-tag clustering, distance matrices of the combined object were computed before generating t-SNE plots.

Definition of CD4 T cells and Malignant clone. In an analogous strategy to what is used for data visualization in flow cytometry, biaxial KDE plots were made using log(protein-tag counts + 1) of CD3 and CD4. Cells in both samples were gated at a threshold ≥4.5 (log scale) for CD4 protein-tag counts and ≥1.0 (log scale) for CD3 protein-tag counts, defining $CD4^+$ T cells. CTCL Malignant cells were defined as CD4 T cells that possessed the most abundant TCRβ CDR3 amino acid sequence, CSARFLRGGYNEQFF, while CTCL CD4 polyclonal cells were CD4 T cells that did not possess this sequence.

Single-cell differential analysis. Comparisons were done using Wilcoxon rank sum test (FindMarkers) between 'CTCL Malignant' and 'CTCL CD4 polyclonal' as well as between 'CTCL Malignant' and 'control CD4 Normal'. Significant genes were defined using q value <0.05 and $|\text{avg\_log}_2\text{FC}| > 1.0$. All Ribosomal Protein (^RP[SL][:digit:]) genes as well as Y, X-escapee and X-variable genes were removed from the differentially expressed list. Heatmaps were made using the union of both sets of significant genes.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Data generated in this project have been deposited to the Gene Expression Omnibus with the accession code GSE126310.

References

21. Van Buggenum, J. A. G. L. et al. *Sci. Rep.* **6**, 22675 (2016).
22. Abrams, J. T. et al. *J. Investig. Dermatol.* **96**, 31–37 (1991).
23. Kaltoft, K. et al. *In Vitro Cell. Dev. Biol.* **28A**, 161–167 (1992).
24. Sanjana, N. E., Shalem, O. & Zhang, F. *Nat. Methods* **11**, 783–784 (2014).
25. Patel, S. J. et al. *Nature* **548**, 537–542 (2017).

nature research

Corresponding author(s): Peter Smibert

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐ A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☒	☐ Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	No software was used
Data analysis	To make the protein and guide oligo tag tables we used CITE-seq count, version 1.3.2: https://github.com/Hoohm/CITE-seq-Count. Gene expression tables were prepared using the cellranger count pipeline, and TCR libraries were processed using the cellranger vdj pipeline. For multimodal analysis we used the Seurat package, version 2.3.0.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Data generated in this project have been deposited to the Gene Expression Omnibus (GEO) with the accession code GSE126310. Raw human sequences are excluded to avoid disclosure of polymorphisms and other potential identifying information.

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Single cell expression profiling datasets with 2,000-10,000 cell input
Data exclusions	Data generated in this project have been deposited to the Gene Expression Omnibus (GEO) with the accession code GSE126310. Processed data filter out cells with low quality, indicative of ambient RNA
Replication	We have performed at least 10 ECCITE-seq experiments capturing various combinations of modalities during method development and application. Results in this study are representative
Randomization	Not relevant to this study
Blinding	Blinding not implemented, not feasible to do so within the context of this experimental design.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☒	☐ Unique biological materials
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	All antibodies used for this study are commercially available and listed in Supplementary tables 3,4,5
Validation	All are well-established clones, validated by the provider

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	The Sez4 cell line is derived from the blood of an SS patient (Abrams et al., 1991), and the MyLa 2059 line is derived from a plaque biopsy sample of an MF patient (Kaltoft et al., 1992). Drs. Odum and Kaltoft generously shared the lines with us. Other

	cell lines used in this study are obtained from ATCC.
Authentication	Karyotyping and phenotypic analysis (FACS, Western)
Mycoplasma contamination	Cell lines were not tested for mycoplasma contamination
Commonly misidentified lines (See ICLAC register)	No commonly misidentified lines were used

Human research participants

Policy information about studies involving human research participants

Population characteristics	The healthy donor was 29 years old. The 39-year old patient was diagnosed with MF stage IV/SS based on histological analysis of punch biopsy of lesions and significant presence of aberrant T cells in the blood as evaluated by flow-cytometric analysis. Sample was taken 1 month following cessation of UVB therapy.
Recruitment	Patient and control samples were collected at New York University Langone Medical Center in accordance with protocols approved by the New York University School of Medicine Institutional Review Board and Bellevue Facility Research Review Committee (IRB#i15-01162). Exclusion criteria include prior history of other hematologic malignancies, active disease resulting in immunodeficiency, anemia and current pregnancy. Sezary patient was diagnosed and staged according to the WHO classification criteria.