Zhou et al. Genome Biology (2020) 21:10
https://doi.org/10.1186/s13059-019-1922-x

Genome Biology

METHOD Open Access

DENDRO: genetic heterogeneity profiling and subclone detection by single-cell RNA sequencing

Zilu Zhou[1,4], Bihui Xu[2], Andy Minn[3] and Nancy R. Zhang[4*]

Abstract

Although scRNA-seq is now ubiquitously adopted in studies of intratumor heterogeneity, detection of somatic mutations and inference of clonal membership from scRNA-seq is currently unreliable. We propose DENDRO, an analysis method for scRNA-seq data that clusters single cells into genetically distinct subclones and reconstructs the phylogenetic tree relating the subclones. DENDRO utilizes transcribed point mutations and accounts for technical noise and expression stochasticity. We benchmark DENDRO and demonstrate its application on simulation data and real data from three cancer types. In particular, on a mouse melanoma model in response to immunotherapy, DENDRO delineates the role of neoantigens in treatment response.

Keywords: Single-cell RNA sequencing, Intratumor heterogeneity, Cancer genomics, Phylogeny inference, Multi-omics analysis

Background

DNA alterations, especially single nucleotide alteration (SNA), and epigenetic modulation both contribute to intratumor heterogeneity [1], which mediates tumor initiation, progression, metastasis, and relapse [2, 3]. Intratumor genetic and transcriptomic variation underlie patients' response to treatment, as natural selection can lead to the emergence of subclones that are drug resistant [4]. Thus, identifying subclonal DNA alterations and assessing their impact on intratumor transcriptional dynamics can elucidate the mechanisms of tumor evolution and, further, uncover potential targets for therapy. To characterize intratumor genetic heterogeneity, most prior studies have used bulk tumor DNA sequencing [5–12], but these approaches have limited resolution and power [13].

Breakthroughs in single-cell genomics promise to reshape cancer research by allowing comprehensive cell type classification and rare subclone identification. For example, in breast cancer, single-cell DNA sequencing (scDNA-seq) was used to distinguish normal cells from malignant cells, the latter of which were further classified into subclones [14–16]. For the profiling of intratumor transcriptional heterogeneity, single-cell RNA sequencing (scRNA-seq), such as Smart-seq2 [17], Drop-seq [18], and 10X Genomics Chromium™, is now ubiquitously adopted in ongoing and planned cancer studies. ScRNA-seq studies have already led to novel insights into cancer progression and metastasis, as well as into tumor prognosis and treatment response, especially response variability in immune checkpoint blockade (ICB) [19–26]. Characterization of intratumor genetic heterogeneity and identification of subclones using scRNA-seq is challenging, as SNAs derived from scRNA-seq reads are extremely noisy and most studies have relied on the detection of chromosome-level copy number aberrations through smoothed gene expression profiles. Yet, as intratumor transcriptomic variation is partially driven by intratumor genetic variation, the classification of cells into subclones and the characterization of each subclone's genetic alterations should ideally be an integral step in any scRNA-seq analysis.

The appeal of subclone identification in scRNA-seq data is compounded by the shortage of technology for sequencing the DNA and RNA molecules in the *same* cell with acceptable accuracy, throughput, and cost [27–30]. Although one can apply both scDNA-seq and scRNA-seq

* Correspondence: nzh@wharton.upenn.edu
[4]Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA, USA
Full list of author information is available at the end of the article

to a given cell population, the mutation analysis and RNA quantification cannot be conducted in the same set of cells. Although there are now technologies for deep targeted sequencing of select transcripts matched with same-cell whole transcriptome sequencing [31, 32], these methods are still, in effect, profiling DNA-level variation by sequencing expressed transcripts, and are thus subject to the technical issues, especially dropout due to transcriptional stochasticity.

Subclone detection using scRNA-seq is difficult mainly because only a small portion of the SNAs of each cell is expected to be seen in the read output of scRNA-seq. This is because to be sequenced, an SNA needs to fall in a transcribed region of the genome, at a location within the transcript that will eventually be read by the chosen sequencing protocol. Even for SNAs that satisfy these requirements, the mutated allele is often missing in the read output due to *dropout*, especially in the heterozygous case. This is due, in part, to the bursty nature of gene transcription in single cells [33–35], where in any given cell, a substantial fraction of the genes are only expressed from one of the alleles. Thus, an SNA residing in a gene that is expressed at the bulk tissue level may not be observed in a particular cell, simply because the mutated allele, by chance, is not expressed in the given cell. We refer to alleles that are not captured due to expression stochasticity as *biological dropouts*. Even for a mutated allele that is expressed, it has to be successfully converted to cDNA and then sequenced to be represented in the final read output; we refer to alleles lost due to technical reasons as *technical dropouts*. In addition to dropout events, post-transcriptional modification, such as RNA editing, and sequencing errors impede both the sensitivity and the specificity of SNA discovery. As a result, methods developed for single-cell SNA detection using scDNA-seq, such as Monovar [36], as well as methods designed for SNA detection in bulk DNA or RNA sequencing data do not yield accurate results in the scRNA-seq setting [37–42].

Here we present a new statistical and computational framework—*D*NA based *E*volution*N*ary tree pre*D*iction by sc*R*NA-seq techn*O*logy (DENDRO)—that reconstructs the phylogenetic tree for cells sequenced by scRNA-seq based on genetic divergence calculated from DNA-level mutations. DENDRO assigns each cell to a leaf in the tree representing a subclone and, for each subclone, infers its mutation profile. DENDRO can detect genetically divergent subclones by addressing challenges unique to scRNA-seq, including transcriptional variation and technical noise. A DENDRO clustering of scRNA-seq data allows joint genetic and transcriptomic analysis on the same set of cells.

We evaluate DENDRO against existing approaches, through simulation data sets and a metastasized renal cell carcinoma dataset with known subpopulation labels, and show that DENDRO improved the accuracy of subclone detection. We then demonstrate the DENDRO to biological discovery through two applications. The first application profiles the treatment response in a melanoma model to immune checkpoint blockade therapy. DENDRO identified a subclone that contracted consistently in response to ICB therapy, and revealed that the contraction was driven by the high mutation burden and increased availability of predicted neoantigens. Transcriptional divergence between the subclones in this model was very weak, and thus, the neoantigen-driven subclonal dynamics would not have been detected without extracting DNA-level information. In the second application to a breast tumor dataset, DENDRO detected subclones and allowed for the joint characterization of transcriptomic and genetic divergence between cells in lymph node metastasis and cells in primary resections.

The DENDRO package, implemented in R, is available at https://github.com/zhouzilu/DENDRO, where we also provide a power calculation toolkit, DENDROplan, to aid in the design of scRNA-seq experiments for subclonal mutation analysis using DENDRO.

Results

Method overview

Overview of the DENDRO model and pipeline

Figure 1a shows an overview of DENDRO's analysis pipeline. Per cell counts of total read coverage (N matrix) and mutation allele read coverage (X matrix) at SNA locations are extracted after read alignment and SNA detection (details in the "Methods" section, Additional file 1: Figure S1). Based on these matrices, DENDRO then computes a cell-to-cell genetic divergence matrix, where entry (c, c') of the matrix is a measure of the genetic divergence between cells c and c'. Details of this genetic divergence evaluation will be given in the next section. DENDRO then clusters the cells into genetically distinct subclones based on this pairwise divergence matrix and selects the number of subclones based on inspection of the intra-cluster divergence curve. Reads from the same subclone are then pooled together, and the SNA profile for each subclone is re-estimated based on the pooled reads, which improves upon the previous SNA profiles computed at the single-cell level. Finally, DENDRO generates a parsimony tree using the subclone-level mutation profiles to more accurately reflect the evolutionary relationship between the subclones.

Genetic divergence evaluation

Due to the high rates of biological and technical dropout, SNA detection within each individual cell lacks sensitivity. We also expect low specificity due to the high

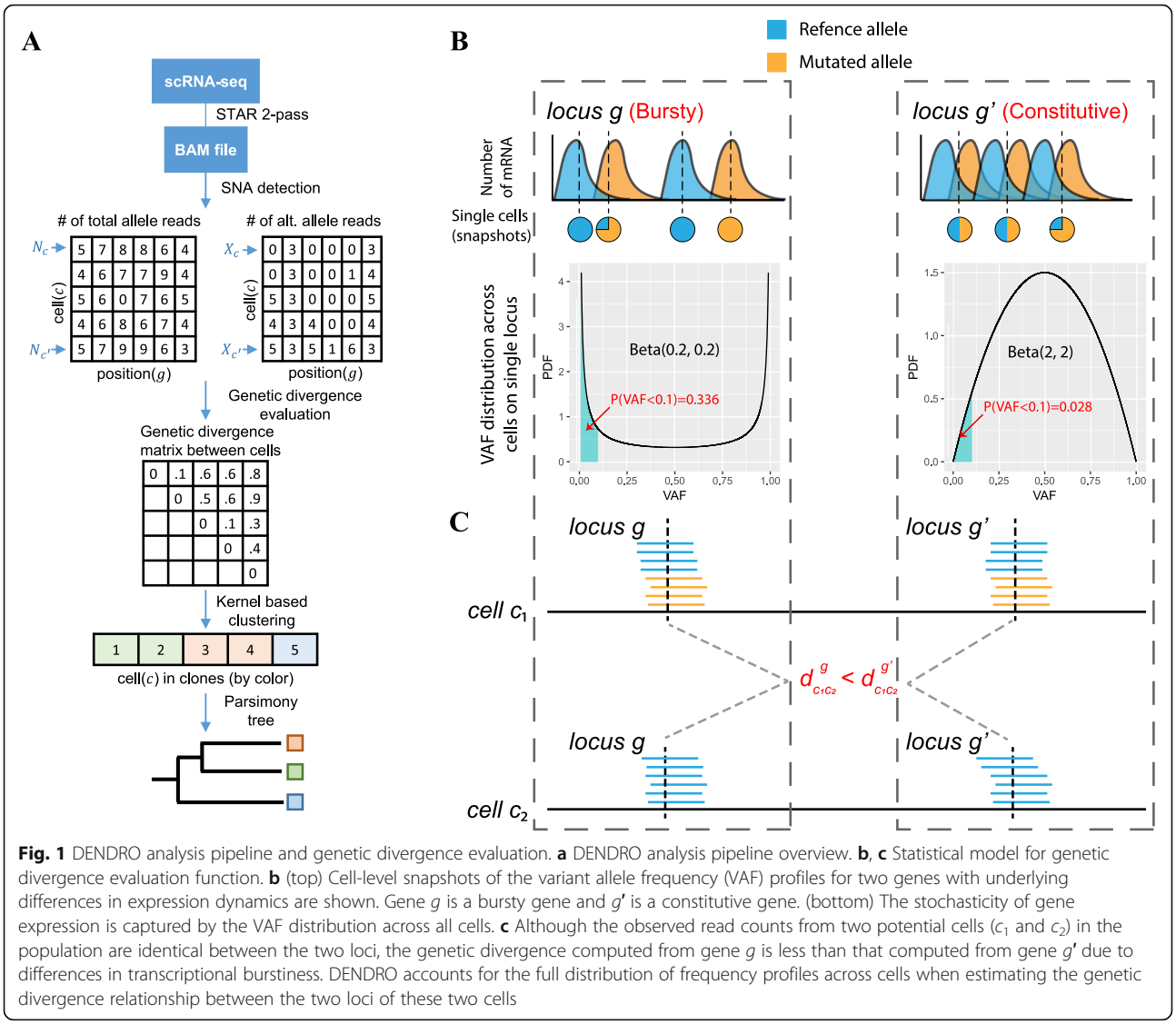

Fig. 1 DENDRO analysis pipeline and genetic divergence evaluation. **a** DENDRO analysis pipeline overview. **b**, **c** Statistical model for genetic divergence evaluation function. **b** (top) Cell-level snapshots of the variant allele frequency (VAF) profiles for two genes with underlying differences in expression dynamics are shown. Gene *g* is a bursty gene and *g*′ is a constitutive gene. (bottom) The stochasticity of gene expression is captured by the VAF distribution across all cells. **c** Although the observed read counts from two potential cells (c_1 and c_2) in the population are identical between the two loci, the genetic divergence computed from gene *g* is less than that computed from gene *g*′ due to differences in transcriptional burstiness. DENDRO accounts for the full distribution of frequency profiles across cells when estimating the genetic divergence relationship between the two loci of these two cells

base error rate in scRNA-seq protocols. Thus, simple distance measures such as the Hamming or Euclidean distances evaluated on the raw SNA genotype matrix or the raw allele frequency matrix do not accurately reflect the genetic divergence between cells.

To more accurately estimate the cell-to-cell genetic divergence, we have developed a statistical model that accounts for technical dropout, sequencing error, and expression stochasticity. Consider two cells, *c* and *c*', and let I_c and $I_{c'}$ index the clonal group to which the cells belong. That is, $I_c = I_{c'}$ if cells *c* and *c*' come from the same subclone and thus share the same SNA profile. Let $X_c = (X_{c1}, \ldots, X_{cm})$ be the mutation allele read counts for this cell at the *m* SNA sites profiled, and $N_c = (N_{c1}, \ldots, N_{cm})$ be the total read counts at these sites. We define the genetic divergence between the two cells as

$$d_{cc'} = -\log P\left(X_c, X_{c'} | N_c, N_{c'}, I_c = I_{c'}\right) = \sum_{g=1}^{m} d^g_{cc'}$$

where $d^g_{cc'} = -\log P(X_{cg}, X_{c'g} | N_{cg}, N_{c'g}, I_c = I_{c'})$.

In other words, $d_{cc'}$ is the negative log likelihood of the mutation allele counts of cells *c* and *c*′, given the total read counts and the event that the two cells belong to the same subclone. If *c* and *c*′ have mutations in mismatched positions, this likelihood for X_c, $X_{c'}$ conditioned on $I_c = I_{c'}$ would be small, giving a large value for $d_{cc'}$. By the assumption of independence between sites, $d_{cc'}$ is the sum of $d^g_{cc'}$, where $d^g_{cc'}$ is the contribution of mutation site *g* to the divergence measure. In characterizing the conditional distribution for X_{cg} and $X_{c'g}$, we use a beta-binomial distribution to model expression stochasticity and a

binomial model to capture sequencing errors and rare RNA editing events. Referring to Fig. 1b, mutations residing in bursty genes, such as gene g, would tend to have U-shaped allele frequency distributions and are more likely to be "dropped" due to low or zero expression. In contrary, mutations residing in constitutive (non-bursty) genes, such as gene g' in Fig. 1b, would have bell-shaped allele frequency distributions and can be genotyped more reliably. Thus, even if the read counts for the mutation loci residing in genes g and g' are identical across two cells (c_1 and c_2 in Fig. 1c), the locus in g' would contribute a higher value, compared to the locus in g, to the divergence between cells c_1 and c_2. Please see the "Methods" section for details.

Accuracy assessment

Accuracy assessment by simulation experiment

First, we designed a simulation procedure to assess the accuracy of DENDRO versus existing approaches and to make realistic power projections for subclone detection (Fig. 2a). Since DENDRO is currently the only method for SNA-based subclone detection using scRNA-seq data alone, we benchmarked against more straightforward approaches such as hierarchical clustering based on mutation allele frequencies and genotypes respectively. The simulation procedure starts with an assumed evolutionary tree, where the leaves are subclones and mutations can be placed on the branches. In the absence of prior information, a simple tree structure is used, such as the one shown in Fig. 2a. Parameters of simulation are (1) total number of mutations, (2) total number of cells, (3) the proportion of cells in each clade, (4) the proportion of mutations along each branch, and (5) mean read coverage across loci. Some of these parameters can be determined using bulk DNA-seq and/or bulk RNA-seq data if available (the "Methods" section). Parameters (1–4) determine the mutation profile matrix (Fig. 2a). To get the matrix of alternative allele (X_{cg}) and total read counts (N_{cg}) for each mutation loci in each cell, we overlay a reference scRNA-seq data with allele-specific read counts onto a designed mutation matrix, which is generated from the simulated tree (see the "Methods" section for details). This allows the simulated datasets to retain the expression stochasticity and sequencing error of real scRNA-seq data. DENDRO is then applied to the read count matrices to obtain the subclone clusters, which is then compared with the known labels. Accuracy is evaluated by three metrics: adjusted Rand index, capture rate, and purity (Additional file 2: Supplementary Materials). Such simulation procedure can also facilitate experiment design, as it predicts the expected clustering accuracy by DENDRO given sequencing parameters and available bulk data for the tumor (see DENDROplan in the "Methods" section).

Using the above framework, we conducted a systematic evaluation of DENDRO's subclone detection accuracy on an example scRNA-seq dataset with allelic information [43]. The results, compiled in Fig. 2b, show that DENDRO has better performance than simply clustering on mutation allele frequencies or the directly estimated mutation profiles from scRNA-seq data. Due to high burstness of the scRNA-seq dataset and limited sequencing depth, we found that Z-matrix, on average, underperformed in all scenario, indicating the necessity of the DENDRO framework. We also quantified how accuracy depends on the mutation burden, mutation read depth, mutation distribution, subclone cell proportion, and cell populations (Additional file 1: Figure S3 and Additional file 2: Supplementary Materials). Even when there are only 100 mutations with relatively low average coverage (read depth equals to 1), DENDRO can still extract meaningful clustering results (average ARI≈ 0.8). More importantly, variation in total expression of genes does not influence DENDRO's divergence measure. DENDRO shows consistent results in simulation analysis between populations of single cell type and multiple cell types (Additional file 1: Figure S3). This is due to DENDRO's reliance only on the distribution of the mutation allele frequency conditioned on the total read coverage, as illustrated by the simulation study (Additional file 1: Figure S2 and Additional file 2: Supplementary Materials). The divergence evaluation reflects solely genetic distance not transcriptomic difference, allowing for easy interpretation. A more extensive simulation analysis can be found in the Additional file 2: Supplementary Materials.

Accuracy assessment on a renal cell carcinoma and its metastasis

We also benchmarked DENDRO against existing methods on the renal cell carcinoma dataset from Kim et al. [21] (Fig. 2c). This dataset contained 116 cells sequenced using the Smart-seq technology [17], obtained from three tumors derived from one patient: a patient-derived xenograft (PDX) from the primary renal cell carcinoma (PDX_pRCC), a biopsy of the metastasis to the lung 1 year after treatment of primary site (Pt_mRCC), and a PDX of the lung metastasis renal cell carcinoma (PDX_mRCC) (Additional file 1: Figure S4a). The cells should share common early driver mutations due to their shared origin from the same patient, but the metastasis and the cultivation of each tumor in separate medium (human or mouse) should have allowed for the accumulation of new mutations. Thus, we expect the three tumors to be clonally distinct. This knowledge allows us to use this dataset to benchmark accuracy and to illustrate how DENDRO enables joint analysis of the genetic and transcriptomic heterogeneity at single-cell resolution.

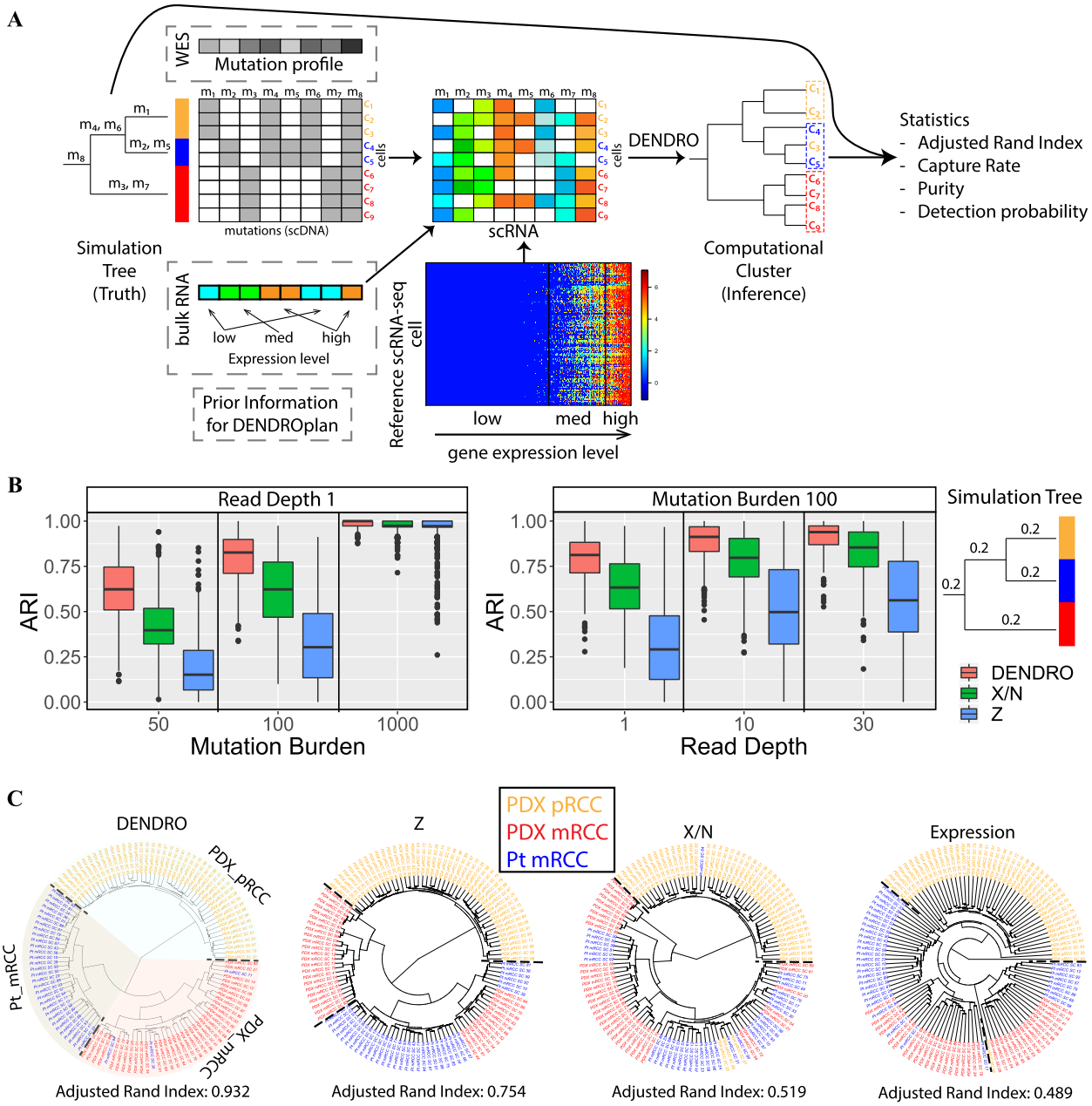

Fig. 2 DENDRO accuracy assessment. **a** The overall simulation analysis pipeline. Mutation matrix (cell-by-loci) is generated according to a simulated evolutionary tree, where the leaves are subclones and mutations can be placed on the branches. Matrices of alternative allele (X_{cg}) and total read counts (N_{cg}) are sampled from a scRNA-seq dataset with known transcriptomic allele-specific read counts. DENDRO cluster is further applied, and its performance is assessed by adjusted Rand index (global accuracy), capture rate (subclone-specific sensitivity), and purity (subclone-specific precision). See Additional file 2: Supplementary Materials for detailed definition. Gray dashed line indicates optional input for DENDROplan, where bulk DNA-seq and bulk RNA-seq can guide the tree simulation and read count sampling procedure. **b** Cluster accuracy via simulation studies. Various parameters show effects on cluster accuracy (measured by adjusted Rand index) based on tree structure on the most right. Left panel: effect of mutation burden on fixed read depth. Right panel: effect of read depth on fixed mutation burden. **c** Evaluation of DENDRO on a renal cell carcinoma and its metastasis. (Left to right) (1) DENDRO clustering result from primary and metastatic renal cell carcinoma dataset. Background colors represent DEDRO clustering result. (2) Clustering of the same dataset using Z matrix (indicator matrix, $Z_{ij} = 1$ when detected a mutation for cell i at locus j by GATK tool). (3) Clustering of the same dataset using $\frac{X}{N}$ matrix (mutation allele frequency matrix). (4) Clustering of the same dataset using expression ($\log(TPM + 1)$)

We compared four different clustering methods: (1) DENDRO, (2) hierarchical clustering based on the primary genotype matrix Z generated by GATK ($Z_{cg} = 1$ when a mutation g is detected for cell c, $Z_{cg} = 0$ otherwise), (3) hierarchical clustering based on the $\frac{X}{N}$ matrix that preserve the variant allele frequency information, and (4) hierarchical clustering based on gene expression ($\log TPM$). DENDRO gives the cleanest separation between the three populations with adjusted Rand Index of 0.932 (1.0 indicates perfect clustering, Fig. 2c panel 1), as compared to 0.754 for Z matrix (Fig. 2c panel 2), 0.519 for $\frac{X}{N}$ matrix (Fig. 2c panel 3), and 0.489 for expression (Fig. 2c panel 4). Inspection of the tree shows that, as expected, divergence between primary tumor and metastasis exceeds divergence between patient sample and PDX sample, as PDX_mRCC clusters with Pt_mRCC rather than PDX_pRCC. All of the other three methods successfully separated the primary sample from the metastatic samples, but could not differentiate between the two metastasis samples.

For DENDRO, the intra-cluster divergence curve flattened at 3, and thus, we stopped splitting at 3 clusters (Additional file 1: Figure S4e and the "Methods" section). We annotated the clusters as PDX_mRCC, PDX_pRCC and Pt_mRCC by their cell compositions (Additional file 3: Table S3a). DENDRO found minimal sharing of subclones among the tumors derived from three sources and low genetic heterogeneity within each tumor. This is unsurprising since relapsed metastasis consists of cells that have already undergone selection, and since the PDX tumors are each seeded by a small subsample of cells from the original tumor, each tumor consists of unique subclones not detected in other sites [44–46]. Additional joint analysis of transcriptome and DNA mutations can be found in Additional file 2: Supplementary Materials and Addtional file 4: Table S4.

DENDRO analysis of the melanoma model in response to immune checkpoint blockade highlights the role of neoantigens

Immune checkpoint blockade (ICB) of the inhibitory receptors CTLA4 and PD1 can result in durable responses in multiple cancer types [47]. Features intrinsic to cancer cells that can impact ICB treatment outcome include their repertoire of neoantigens [48], tumor mutational burden (TMB) [49], and expression of PDL1 [50]. DENDRO analysis of scRNA-seq data allows joint DNA-RNA analysis of single cells, thus enabling the simultaneous quantification of tumor mutational burden, the prediction of neoantigen repertoire, and the characterization of gene expression profile at subclonal resolution. Thus, to demonstrate the power of DENDRO and to better understand the relationship between ICB response and intratumor heterogeneity, we profiled the single-cell transcriptomes across three conditions derived from two melanoma cell lines (Fig. 3a): B16 melanoma cell line, which has shown modest initial response to ICB treatment but eventually grows out, and Res 499 melanoma cell line (R499), which was derived from a relapsed B16 tumor after combined treatment of radiation and anti-CTLA4 and is fully resistant to ICB [51]. B16 was evaluated with and without anti-PD1 treatment, as we wanted a tumor model that captures a transient ICB response. A total of 600 tumor cells were sequenced with Smart-seq technology from six mice across three conditions: two mice with B16 without treatment (B16), two mice with B16 after anti-PD1 treatment (B16PD1), and two mice with R499 without treatment (R499) (Fig. 3a and the "Methods" section). The existence of multiple subclones in B16 and R499 was suggested by bulk WES analysis [51, 52]. Our goal here is to determine whether the subclones differ in anti-PD1 response, and if so, what are the subclonal differences.

A DENDRO analysis of 4059 putative mutation sites across 460 cells retained after QC (see the "Methods" section and Additional file 1: Figure S9a, b, c) yields the clustering displayed in Fig. 3b, with four subclones suggested by the intra-cluster divergence curve (Additional file 1: Figure S9d). All subclones are shared among the three conditions, which is not unexpected given that all tumor cells were derived from the same parental cell line. However, the subclonal proportions vary significantly between conditions (Fig. 3b). The subclonal proportions of B16PD1 are approximately intermediate between that of B16 and R499 (Fig. 3c). This is expected as R499 had gone through immune editing whereas B16PD1, at the time of harvest, was still undergoing immune editing and was at the transient response state. Furthermore, the selective pressure of radiation plus anti-CTLA4 is likely more than that of anti-PD1 treatment, as the former but not the latter results in complete responses in our B16 model [51]. The frequency of Clone 2 is lower in B16PD1 and R499, indicating sensitivity to anti-PD1 treatment, while the frequencies of Clone 3 and Clone 4 increase after treatment and are the highest in R499, indicating resistance to therapy (Fig. 3c, S10a).

To explore why subclones vary in sensitivity to anti-PD1 treatment, we compared the mutation profile of Clone 2 to the other subclones. We pooled cells in each of the four subclones and re-estimated their mutation profiles, which were then used to construct a phylogenetic tree (Fig. 3d). The phylogeny suggests that Clone 3 and Clone 4 are genetically closer to each other than to Clone 2, and thus, their similarity in treatment response may be in part due to similarity in their mutation profiles. The re-estimated mutation profiles show that Clone 2 has the highest tumor mutation burden, which

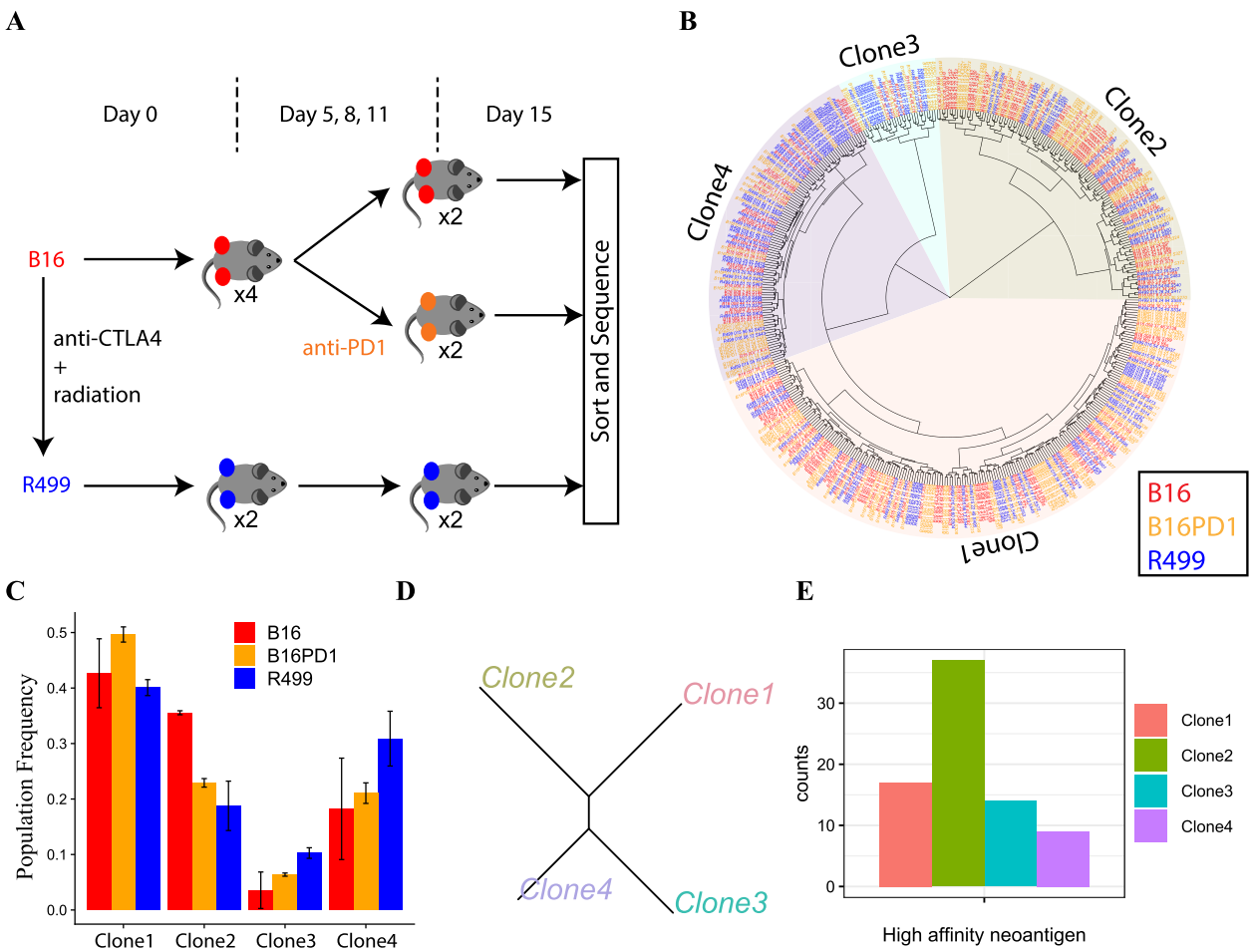

Fig. 3 Clonal composition alternations with anti-PD1 treatments and cell lines. **a** Experimental overview. For each condition at day 15, we have two biological replicates. There are total 600 cells from 6 tumors sequenced. **b** DENDRO cluster result. No clone is exclusively associated with any tumor condition. **c** Frequencies of the subclonal populations in B16, B16PD1, and R499. **d** Neighbor-joining phylogenetic tree given detected subclones. **e** Number of high-affinity neoantigens predicted for each clone. Clone 2 has the highest number of neoantigens

has been associated with increased likelihood of ICB response [53, 54]. We then predicted the quantity of high-affinity (≤ 100 nm) neoantigens in each subclone given its mutation profile [52]. As shown in Fig. 3e, Clone 2 has twice as many high-affinity neoantigens as the other three subclones. The high level of neoantigens can lead to better T cell recognition, resulting in increased efficacy of anti-PD1 treatment [55].

Analysis of gene expression, on the other hand, did not yield detectable known signatures associated with anti-PD1 treatment sensitivity. Projections based on the expression of highly variable genes, as shown in PCA and t-SNE plots (Additional file 1: Figure S8), did not yield meaningful clusters. Differential expression analysis between each subclone and the other subclones found few genes with adjusted P value < 0.05, indicating similar expression across subclones that is concordant with the lack of structure in the expression PCA and t-SNE plots. Expressions of *Pdl1* (aka. *Cd274*) showed no differences between subclones (KS-test: P value > 0.42, Additional file 1: Figure S10b). In addition, there were no detectable chromosome-level differences in smoothed gene expression, indicating that there are no large CNV events that distinguish the subclones (Additional file 1: Figure S11). DENDRO, detecting exonic mutations from scRNA-seq data, enabled the finding of subclones in this data, the prediction of neoantigen load of each subclone, and the analysis of subclonal dynamics due to treatment. Our analysis suggests that the genetic heterogeneity, rather than transcriptomic heterogeneity, contributes to treatment efficacy in this tumor model.

Simultaneous analysis of genetic and transcriptomic variation in single-cell breast cancer

We next applied DENDRO to the analysis of data from a study of primary and metastasized breast cancer [20].

We focused on tumors from two patients (BC03 and BC09) that had the most cells sequenced (Additional file 1: Figure S12 and Additional file 5: Table S5). Patient BC03 had cells sequenced from the primary tumor (here after BC03P) as well as cells from regional metastatic lymph nodes (here after BC03LN), whereas patient BC09 had cells sequenced only from the primary resection. One hundred thirty-two single-cell transcriptomes were profiled by Smart-seq protocol [17]. We first assess whether DENDRO separated BC03 cells from BC09 cells, since inter-individual genetic distances should far exceed intra-individual genetic distances owning to the randomness of passenger mutations [19, 22, 56]. Then, we examine the transcriptomic and genetic heterogeneity within each tumor.

GATK [57] detected a total of 2,364,823 mutation sites across the 132 cells; 353,647 passed QC (the "Methods" section) and were retained for downstream analysis (Additional file 1: Figure S12a, b, c). Figure 4 shows the clustering determined by DENDRO. DENDRO separates BC09 cells from BC03 cells with 100% accuracy (Fig. 4a). The intra-cluster divergence curve flattened at five subclones: three subclones for BC03 and two for BC09 (Fig. 4a, Additional file 1: Figure S12d and Additional file 3: Table S3b). Within BC03, Clone Mix_1 and Clone Mix_2 contained a mixture of cells from the primary tumor and lymph nodes, and Clone LN_1 contained mostly cells from the lymph nodes. This suggests that tumor cells that have metastasized to the lymph nodes belong to an intermediate stage and are genetically heterogeneous, with some cells remaining genetically similar to the primary population and others acquiring new genetic mutations, coherent with previous studies [58, 59]. In comparison, hierarchical clustering based on expression (using log transcripts-per-million values) did not separate BC03 from BC09 and gave a negative adjusted Rand index within BC03, indicating effectively random assignment of cells to the two patients (Fig. 4).

We then pooled cells within each of the 5 clusters and re-estimated their mutation profiles with DENDRO. We defined a variant as subclonal if it was not present in all of the subclones within a tumor. Based on detection marginal likelihood, we picked the top 10,000 most confident variants to construct a phylogenetic tree (Fig. 4c). As expected, the two BC09 clusters are far from the three BC03 clusters. Within BC03, the length of the branches shows that the subclone containing mostly cells from lymph nodes (labeled BC03LN_1) is genetically more similar to Clone Mix_2 compared to Clone Mix_1 (Fig. 4c). In addition, window-smoothed expression plot with cells grouped by DENDRO clustering shows broad chromosome-level shifts in expression patterns between subclones, most likely due to copy number aberrations that are consistent with SNAs (Additional file 1: Figure S13) [22].

A comparison of the transcriptomes of the subclones revealed substantial differences in the expression of PAM50 genes, which are prognostic markers for breast cancer (Fig. 4d) [60]. DENDRO detected one rare subclone, BC09_2, with only six cells (< 5% of the total number of cells) which had a strong basal-like signature. Interestingly, in BC03, Clone LN_1 has the TNBC/basal-like subtype with an invasive gene signature, while Clone Mix_2 has the *ESR1*$^{+}$ subtype. Thus, the genetic divergence of Clone LN_1 from Clone Mix_2 is accompanied by its acquisition of an invasive metastatic expression signature. In a direct comparison between cells from the primary site and cells from the lymph node without distinguishing subclones, these expression differences would be much weaker since the subclones do not cleanly separate by site. Compared with the original analysis that assigned each tumor to one specific breast cancer subtype, this analysis identifies subclones with different expression phenotypes, potentially allowing for better therapy design that targets all subclone phenotypes to reduce the risk of tumor relapse.

Existing scRNA-seq studies of cancer tissue cluster cells based on total gene expression or copy number profiles derived from smoothed total expression, making it difficult to separate the effects of subclonal copy number aberrations from transcriptomic variation [19, 22, 24]. Differential expression analysis based on clusters derived from total expression is prone to self-fulfilling prophecy, as there would indeed be differentially expressed genes because this is the clustering criteria. Because DENDRO's subclone identification is based solely on genetic divergence, and not on expression profile, the downstream differential gene expression analysis can be precisely attributed to transcriptional divergence between subclones.

Hence, we conducted a transcriptome-wide search for pathways that have differential expression between subclones (the "Methods" section and Additional file 6: Table S6), and assessed their overlap with pathways that are differentially mutated between subclones. Focusing on tumor BC03, pathways for G2M checkpoint and *KRAS* signaling are upregulated in lymph node metastasis Clone BC03LN_1, while pathways for estrogen response and apoptosis are downregulated, indicating a more invasive phenotype (Additional file 6: Table S6e). In addition, *GAPDH* is upregulated in the metastatic subclone (BC03LN_1) and downregulated in the two mix-cell subclones, consistent with previous findings [61, 62] (Additional file 1: Figure S14d). Differentially expressed genes between other subclone pairs in BC03 are also enriched in estrogen response, apoptosis, and DNA repair (Additional file 6: Table S6c, d). In parallel, subclone-specific mutated genes are highly enriched in

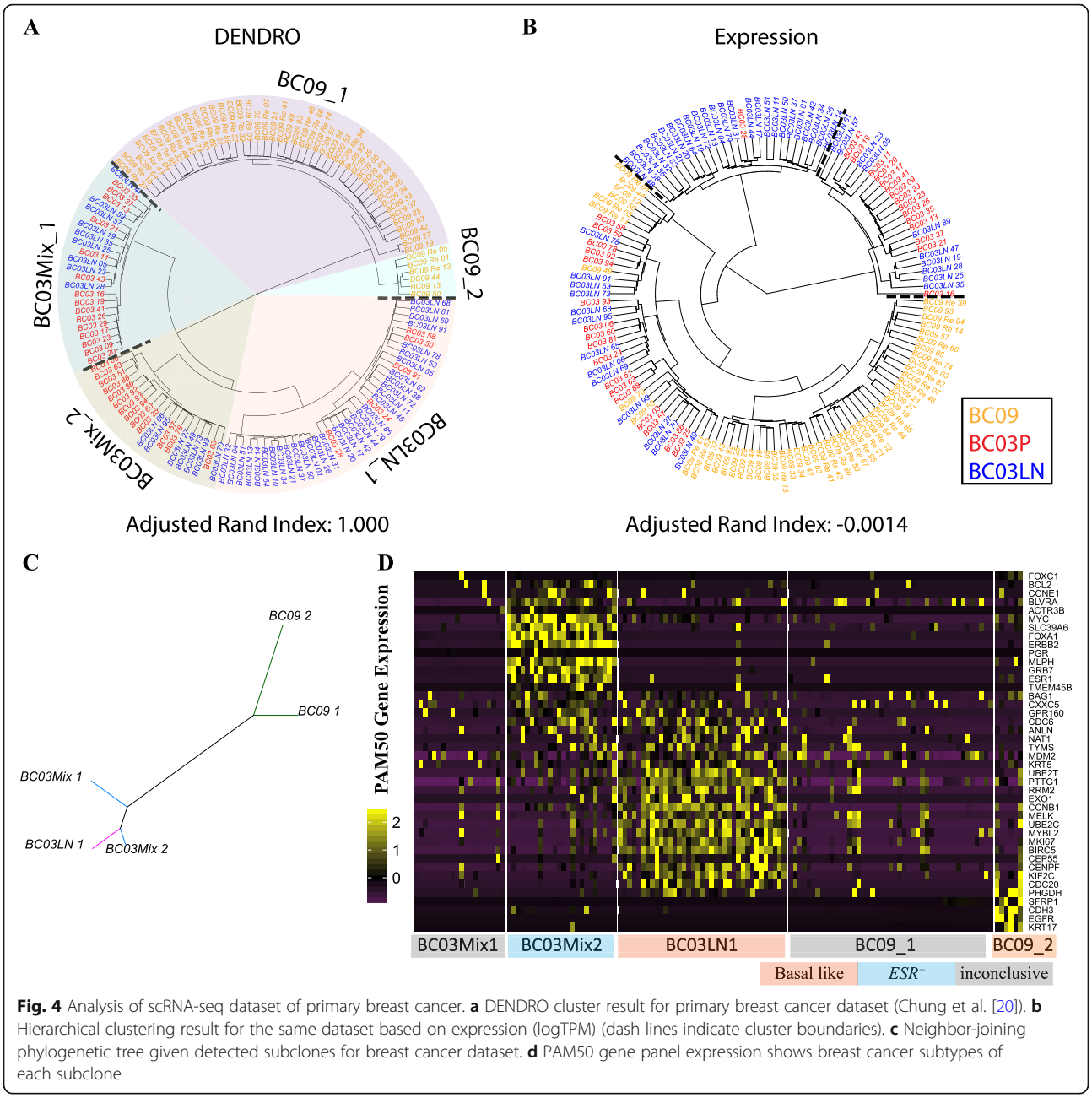

Fig. 4 Analysis of scRNA-seq dataset of primary breast cancer. **a** DENDRO cluster result for primary breast cancer dataset (Chung et al. [20]). **b** Hierarchical clustering result for the same dataset based on expression (logTPM) (dash lines indicate cluster boundaries). **c** Neighbor-joining phylogenetic tree given detected subclones for breast cancer dataset. **d** PAM50 gene panel expression shows breast cancer subtypes of each subclone

cancer-related pathways, including MYC target, G2M checkpoints, and mitotic spindle, and immune-related pathways, such as interferon response, TNF-a signaling, and inflammatory response (Additional file 6: Table S6). Interestingly, few of the differentially mutated genes are associated with estrogen and androgen responses, suggesting that the differential expression of hormone-related genes is not mediated directly by genetic mutations in these pathways. This is consistent with the recent studies that epigenetic alteration, such as histone acetylation and methylation, regulates hormone receptor signaling in breast cancer [63–66]. DNA-RNA joint analysis between other subclones is included in the Additional file 6: Table S6 and Additional file 1: Figure S14. Overall, this example illustrates how DENDRO enables the joint assessment of genetic and transcriptomic contributions to clonal diversity at single-cell resolution.

Discussion

We have described DENDRO, a statistical framework to reconstruct intratumor DNA-level heterogeneity using scRNA-seq data. DENDRO starts with mutations detected directly from the scRNA-seq reads, which are very noisy due to a combination of factors: (1) errors are

introduced in reverse-transcription, sequencing, and mapping; (2) low sequencing depth and low molecule conversion efficiency leading to technical dropouts; and (3) expression burstiness at the single-cell level leading to biological dropouts. DENDRO overcomes these obstacles through the statistical modeling of each component. Given noisy mutation profiles and allele-specific read counts, DENDRO computes a distance between each pair of cells that quantifies their genetic divergence after accounting for transcriptional bursting, dropout, and sequencing error. Then, DENDRO clusters the cells based on this distance as subclone and re-estimates a more robust subclone-specific mutation profile by pooling reads across cells within the same cluster. These re-estimated mutations profiles are then passed to downstream mutation analysis and phylogenetic tree reconstruction.

Importantly, the genetic divergence used by DENDRO for cell clustering is based solely on allelic expression ratios and do not reflect the difference in total expression between cells at mutation sites. Thus, DENDRO differs from, and complements, existing tools that cluster cells based on total expression. In fact, as shown by simulation analysis, DENDRO clusters the cells based on true underlining mutation profiles and is robust to changes in total gene expression. As expected, the numbers of cells, the depth of sequencing, the actual number of subclonal mutations, and the phylogenetic tree structure all influence the power of DENDRO. To aid researchers in experiment design, we developed DENDROplan, which predicts DENDRO's clustering accuracy given basic experimental parameters and the expected informative mutation count, which can be obtained from bulk DNA sequencing.

Ideally, joint sequencing of the DNA and RNA on the same cells would allow us to relate genomic profiles to transcriptomic variations. Currently, there is yet no scalable technology for doing this. Separately performing scDNA-seq and scRNA-seq on different batches of cells within the same tumor would meet the nontrivial challenge of matching the subclones between the two data sets. DENDRO takes advantage of the central dogma and utilizes computational methods to extract genetic divergence information from noisy mutation calls in coding regions. Through two case studies, we illustrate the insights gained from the subclonal mutation and expression joint analysis that DENDRO enables.

We have demonstrated that proper computational modeling can excavate the DNA-level heterogeneity in scRNA-seq data. Yet, there are always limitations in working with RNA. While rare RNA editing events are absorbed by the parameter ϵ, DENDRO cannot distinguish subclone-specific constituent RNA editing events from subclone-specific DNA mutations. In the extreme and unlikely scenario where RNA editing events are common and pervasive, DENDRO's cluster would reflect RNA editing. In such cases, we recommend using matched bulk DNA-seq of the same tumor to filter the loci detected in the first step of DENDRO, keeping only those that are supported by at least one read in the bulk DNA-seq data. In addition, DENDRO's analysis is restricted to transcribed regions, as variants are detected using transcriptomic data, and thus ignores non-coding mutations which can sometimes be informative for tumor evolution [67–70].

Tag-based scRNA-seq (10X, Drop-seq, etc.) is now commonly adopted for cancer sequencing, but we do not recommend applying DENDRO to this sequencing design because of two reasons: (1) limited number of variants can be detected with tag-based methods as they only profile a small fraction of the transcript (3-prime or 5-prime end); and (2) the sequencing depth of tag-based methods are critically low (< 0.1X), resulting in unreliable variant calling. However, we do anticipate that emerging technologies, such as long-read full-transcript scRNA-seq technologies [71] and transcriptome-based deep targeted sequencing [31, 32] will overcome these limitations of tag-based scRNA-seq. Given proper experimental design, we expect that these emerging technologies will be ideally suited for the joint analysis of exonic somatic mutations and gene expression.

Conclusions

We have developed DENDRO, a statistical method for tumor phylogeny inference and clonal classification using scRNA-seq data. DENDRO accurately infers the phylogeny relating the cells and assigns each single cell from the scRNA-seq data set to subclone. DENDRO allows us to (1) cluster cells based on genetic divergence while accounting for transcriptional bursting, technical dropout, and sequencing error, as benchmarked by in silico mixture and a simulation analysis; (2) characterize the transcribed mutations for each subclone; and (3) perform single-cell multi-omics analysis by examining the relationship between transcriptomic variation and mutation profile with the same set of cells. We evaluate the performance of DENDRO through a simulation analysis and a data set with known subclonal structure. We further illustrate DENDRO through two case studies. In the first case study of relationship between intratumor heterogeneity and ICB treatment response, DENDRO estimates tumor mutation burden and predicts repertoire of high-affinity neoantigens in each subclone from scRNA-seq. In the second case study on a primary breast tumor dataset, DENDRO brought forth new insights on the interplay between intratumor transcriptomic variation and subclonal divergence.

Methods

scRNA-seq alignment and SNA calling pipeline

Additional file 1: Figure S1 illustrates the SNA calling pipeline. Raw scRNA-seq data is aligned by STAR 2-pass method (default parameters), which accounts for splicing junctions and achieve higher mapping quality [72]. Transcripts per million (TPM) was quantified using RSEM (default parameters) [73]. In the next step, raw variant calling is made using the Haplotype Caller (GATK tool) on the BAM files after sorting, joining read groups, removing duplicated reads, removing overhangs into intronic regions, realigning, and recalibration [74]. Conventionally, there are two methods from GATK tools for mutation detection: haplotype caller and mutect2. Haplotype caller has a RNA-seq setting which handles splice junctions correctly, but assumes VAF around 50%, while mutect2 can detect mutations with low VAF but does not account for splice junction. The reason we select haplotype caller instead of mutect2 is that we extract allele read counts for all cells as long as one of the cells is listed as carrying the mutation. Thus, as long as one cell has VAF reaching 50%, this mutation would be detected. Calls with stand_call_conf greater than 20 and population frequency greater than 5% but less than 95% were preserved for further analysis. Admittedly, such lenient filtering typically introduces false-positive sites. However, our priority at this step is to minimize false-negative rate, while the genetic divergence matrix in the following step robustly estimates cell population substructure. Both the coverage of the alternative allele and the total read coverage are extracted for each site for further analysis.

Data preprocessing and quality control

To ensure robustness of downstream analysis, we filtered out low-quality cells, variants, and genes. We retained cells with (1) > 10,000 reads mapped, (2) < 10% mitochondria gene expression, and (3) > 1000 gene detected; genes with > 5 cells detected (TPM > 0 as detected); and variants with > 2 cells detected by GATK. Original TPM values as defined by RSEM were added a value of 1 (to avoid zeros) and then log-transformed for downstream transcriptomic analysis.

Genetic divergence and beta-binomial framework

Consider two cells: c and c'. Let I_c and I_c denote the clonal group to which the cells belong, i.e., $I_c = I_{c'}$ if and only if cells c and c' come from the same subclone. We define the genetic divergence at loci g, by $d^g_{cc'}$:

$$d^g_{cc'} = \log \frac{\mathrm{P}(X_{cg}, X_{c'g}|N_{cg}, N_{c'g})}{P(X_{cg}, X_{c'g}|N_{cg}, N_{c'g}, I_c = I_{c'})}.$$

$$= \log \frac{\mathrm{P}(X_{cg}, X_{c'g}|N_{cg}, N_{c'g}, I_c = I_{c'}) + \mathrm{P}(X_{cg}, X_{c'g}|N_{cg}, N_{c'g}, I_c \neq I_{c'})}{P(X_{cg}, X_{c'g}|N_{cg}, N_{c'g}, I_c = I_{c'})}$$

where $X_c = (X_{c1}, X_{c2}, \ldots X_{cg}, \ldots X_{cm})$ are the mutation allele read counts for cell c and $N_c = (N_{c1}, N_{c2}, \ldots N_{cg}, \ldots N_{cm})$ are the total read counts at these sites. More intuitively, if cells c and c' are not from the same clonal group, the probability of cells c and c' from the same cells given data (i.e., denominator) has smaller value. Thus, $d^g_{cc'}$ is large, indicating bigger divergence between the two cells. With further derivation (Additional file 2: Supplementary Materials), $d^g_{cc'}$ is a function of the five following probabilities:

$$d^g_{cc'} = f\big(P_g; P\big(X_{cg}|N_{cg}, Z_{cg} = 0\big); P\big(X_{cg}|N_{cg}, Z_{cg} = 1\big);$$

$$P\Big(X_{c'g}|N_{c'g}, Z_{c'g} = 0\Big); P\Big(X_{c'g}|N_{c'g}, Z_{c'g} = 1\Big)\Big)$$

where $Z_{cg} \in \{0, 1\}$ is SNA indicator for cell c at site g and $P_g = P(Z_g = 1)$ is mutation frequency across the cells estimated by GATK calls.

In the above formula for $d^g_{cc'}$, $P(X_{cg}|\, N_{cg}, Z_{cg} = 0)$ and $P(X_{c'g}|N_{c'g}, Z_{c'g} = 0)$ reflect reverse-transcription/sequencing/mapping errors and rare RNA editing events, because when there is no mutation (i.e., $Z_{cg} = 0, Z_{c'g} = 0$), all mutation reads reflect such technical errors or RNA editing. Let ϵ denote the combined rate of technical error and RNA editing, we have

$$P\big(X_{cg}|N_{cg}, Z_{cg} = 0\big) \sim \mathrm{Binomial}\big(X_{cg}|N_{cg}, \epsilon\big)$$

where ϵ is set to 0.001 based on prior knowledge [75].

For cases where there are mutations (i.e., $Z_{cg} = 1$), the distribution of mutated read counts given total read counts is modeled with a beta-binomial distribution, which is capable of modeling technical dropout and transcriptional bursting, and is supported by previous allele-specific expression studies [34, 76].

$$P(X_{cg}|N_{cg}, Z_{cg} = 1) \sim \int_0^1 \mathrm{Binomial}\Big(X_{cg}|N_{cg}, Q_{cg} = q\Big) dF(q),$$

$$q \sim \mathrm{Beta}\Big(\alpha_g, \beta_g\Big)$$

where Q_{cg} indicates proportion of mutated alleles expressed in cell c at site g, with beta distribution as prior. Respectively, α_g and β_g represent gene activation and deactivation rate, which are estimated empirically across cells based on the first and second moment estimators.

Through optimized vectorization, given a data set of 500 cells with 2500 variants, genetic divergence matrix can be computed under 2 min in a normal desktop with 16 GB of RAM (single thread). Analytically, the algorithm is of complexity $O(N^2 * G)$, where N is the number of cells and G is the number of variants.

Kernel-based clustering and optimal cluster assignment

We cluster the cells using a kernel-based algorithm, such as hierarchical clustering. Given that there are

multiple sorting schemes, we leave the user to choose it. For the default-sorting scheme, we recommend "ward. D" [77]. This is because $d_{cc'}$ behaves like a log likelihood ratio, which should follow a χ^2 distribution when the two cells share the same subclone. The "ward. D" method has been shown to work well in Euclidian space. Empirically, among different hierarchical clustering algorithms on the renal cell carcinoma dataset (Additional file 1: Figure S5), "ward. D"-based hierarchical clustering performs the best.

To determine the number of clusters, we use an intra-cluster divergence curve computed from the divergence matrix. Existing software rely on AIC, BIC, or another model selection metric [78, 79]. However, since we only have the "distance" matrix, these traditional methods cannot be applied. Let N_k be the number of cell pairs in cluster C_k and N be the total number of pairs between cells for all clusters. Let K be the number of clusters. The weighted sum of intra-cluster distance W_K is

$$W_K = \sum_{k=1}^{K} N_k \sum_{(i,j)\in C_k} \frac{d_{ij}}{N}$$

Note that small clusters are naturally down-weighted in the above metric. DENDRO relies on visual examination of the intra-cluster divergence curve (W_K plotted against K) to find the "elbow point," which can be taken as a reasonable clustering resolution.

Simulation analysis

In our simulation analysis, we adopt a scRNA-seq dataset from Deng et al. as the reference, which, by crossing two mouse strains, obtained transcriptomic allele-specific read counts for every SNPs in exonic regions in each cell [43]. In this case, the Deng et al. data maintained the expression stochasticity in scRNA-seq data. To overlay the read counts on simulated mutation profile, for every simulated locus, we sampled a SNP from this reference. For cells with mutation at this locus, we randomly assigned one allele of the sampled SNP as mutated allele. For cells without mutation, we set the mutated allele counts as 0 and the total read counts as sum of the two alleles from the reference. We further added binomial noise ($p_\epsilon = 0.001$, suggested by [75]) to mimic sequencing error. When analyzing DENDRO performance in terms of various number of mutation sites, number of cells, proportion of cells in each clade, and proportion of mutations along each branch, we only take a subset of cells (cells in early blastocyst, mid blastocyst, and late blastocyst stages) to ensure the expression homogeneity. On the other hand, we utilize a mixture cell population (cells in 16-cell stages and blastocyst stages) to test the robustness of DENDRO performance with regard to various expression profiles.

Power analysis toolkit and experimental design

Before conducting a single-cell RNA-seq experiment on a tumor sample, it is important to project how subclone detection power depends on the number of cells sequenced and the coverage per cell. To facilitate experiment design, we have developed a tool, DENDROplan (Fig. 2a), that predicts the expected clustering accuracy by DENDRO given sequencing parameters and available bulk data for the tumor. Given an assumed tree structure and a target accuracy, DENDROplan computes the necessary read depth and number of cells needed.

As shown in Fig. 2a, if bulk DNA sequencing and/or RNA sequencing data are available for the tumor being studied, these data can be harnessed to make more realistic power calculations. For example, if SNAs have been profiled using bulk DNA sequencing data, the set of mutations that lie in the exons of annotated genes can be retrieved and used directly in constructing the simulation data. Furthermore, phylogeny construction algorithms for bulk DNA-seq data can be used to infer a putative tree structure that can be used as input to DENDROplan [5, 79]. If bulk RNA-seq data is available, the bulk expression level of the mutation-carrying genes can be used to predict the expression level of the mutation in the single-cell data. In another word, variants in high-expressed genes in bulk will be sampled from high-expressed variant loci in scRNA reference and vice versa. The power analysis tool is also available at https://github.com/zhouzilu/DENDRO.

SNA inference in "bulk" and phylogenetic tree construction

As stated previously, DENDRO further inferred SNA after pooling the reads from all cells within each cluster. Because, with our choice of thresholds, we identify SNAs in single cells with high sensitivity, the "bulk" level SNAs should be a subset of the SNAs in single cells, and mutation allele counts and total allele counts should provide us with enough information for SNA detection using a maximum likelihood framework [80], which accounts for both sequencing error and rare RNA editing events. Suppose s is the genotype (number of reference allele) at a site and assume m, the ploidy, equals to 2. Then, the likelihood is:

$$\mathcal{L}(s) = \frac{1}{m^k} \prod_{j=1}^{l} [(m-s)\epsilon + s(1-\epsilon)] \prod_{j=l+1}^{k} [(m-s)(1-\epsilon) + s\epsilon]$$

where k is the number of reads at a site and the first l bases ($l \le k$) be the same to reference and the rests are same to alternative allele. ϵ is the sequencing error and rare RNA editing combined rate. s^* is the maximum likelihood estimator of the genotype:

$$s^* = \underset{s}{\operatorname{argmax}} -\mathcal{L}(s)$$

Given mutation profiles, DENDRO then constructs a phylogenetic tree with the neighbor-joining method, which can more accurately capture the evolutionary relationship between different subclones [81] than the initial tree given by hierarchical clustering.

Differential gene expression, mutation annotation, and gene ontology analysis

We use Seurat and scDD to identify differentially expressed genes between tumors and between tumor subclones [82–84]. For each comparison, we apply two different methods: MAST implemented by Seurat and scDD. Genes with adjusted p value < 0.05 count as significant differentially expressed gene for each method. We further intersect these two sets of differentially expressed genes to increase robustness. Subclonal mutations are annotated by ANNOVAR with default parameters, and variants associated with intergenic regions were discarded for downstream analysis [85]. For GO analysis, we apply Gene Set Enrichment Analysis tool [57]. Hallmark gene sets serve as the fundamental database with FDR q value < 0.05 as significant.

Single-cell RNA-seq of tumor model derived from B16

Six C57bl/6 mice were injected on both flanks with either B16 or R499: four with B16 and two with R499. Two of the mice implanted with B16 were treated with 200 μg of anti-PD1 per mouse on days 5, 8, and 11. On day 15, all tumors were harvested and made into single-cell suspension. One hundred thousand CD45-negative tumor cells were sorted on Aria to enrich for live tumor cells and loaded on SMARTer ICELL8 cx Single-Cell System prior to full-length single-cell RNA sequencing library preparation using Smart-seq following the manufacturer's recommendations. Four hundred sixty cells and 11,531 genes passed standard QC and were retained for downstream analysis.

Neoantigen prediction

Based on gene expression from RNA-seq data, variants from unexpressed transcripts are removed. The MHC-I binding affinities of variants are then predicted using NetMHC version 4.0 for H-2-Kb and H-2-Db using peptide lengths from 8 to 11 [86]. Given subclonal mutation profile, we further assign the neoantigens to each subclone.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s13059-019-1922-x.

Additional file 1. Supplementary figures

Additional file 2. Supplementary material and Table S1 and Table S2

Additional file 3. Table S3

Additional file 4. Table S4

Additional file 5. Table S5

Additional file 6. Table S6

Additional file 7. Review history.

Acknowledgements

We thank Dr. Nan Lin for informing useful tools for biological insight analysis and Dr. Kai Tan and Dr. Mingyao Li for helpful comments and suggestions.

Review history

The review history is available as Additional file 7.

Peer review information

Barbara Cheifet was the primary editor of this article and handled its editorial process and peer review in collaboration with the rest of the editorial team.

Authors' contributions

ZZ and NRZ formulated the model. ZZ developed and implemented the algorithm and conducted all computational analyses. BX and AM designed the immunotherapy case study. BX performed the experiments of the immunotherapy case. ZZ, BX, and NRZ wrote the manuscript. All authors read and approved the final manuscript.

Funding

This work was supported by the National Institutes of Health (NIH) grant 1P01CA210944-01 to AM and BX, 5R01-HG006137-07 to ZZ and NRZ, and 1U2CCA233285-01 to NRZ.

Availability of data and materials

DENDRO is an open-source R package available at https://github.com/zhouzilu/DENDRO with license GPL-3.0 [87]. Original source script for this manuscript is stored with digital object identifier (DOI) at https://doi.org/10.5281/zenodo.3521087 [88]. Public datasets for simulation analysis, renal cell carcinoma validation, and breast cancer analysis can be found at the National Center for Biotechnology Information Gene Expression Omnibus (GEO) under accession numbers GSE45719, GSE73122, and GSE75688 respectively [89–91]. Sequencing data for anti-PD1 experiment in melanoma cell lines can be found at GEO with accession number GSE139248 [92].

Ethics approval and consent to participate

All animal experiments were performed according to protocols approved by the Institutional Animal Care and Use Committee of the University of Pennsylvania, (IACUC#804835).

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Graduate Group in Genomics and Computational Biology, University of Pennsylvania, Philadelphia, PA, USA. [2]Department of Radiation Oncology, Parker Institute for Cancer Immunotherapy, Abramson Family Cancer Research Institute, Graduate Group in Cell and Molecular Biology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. [3]Department of Radiation Oncology, Parker Institute for Cancer Immunotherapy, Abramson Family Cancer Research Institute, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. [4]Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA, USA.

Received: 16 September 2019 Accepted: 16 December 2019
Published online: 14 January 2020

References

1. Gamazon ER, Stranger BE. The impact of human copy number variation on gene expression. Brief Funct Genomic Proteomic. 2015;14:352–7.

2. Hanks S, Coleman K, Reid S, Plaja A, Firth H, Fitzpatrick D, Kidd A, Mehes K, Nash R, Robin N, et al. Constitutional aneuploidy and cancer predisposition caused by biallelic mutations in BUB1B. Nat Genet. 2004;36:1159–61.
3. Vicente-Duenas C, Hauer J, Cobaleda C, Borkhardt A, Sanchez-Garcia I. Epigenetic priming in cancer initiation. Trends Cancer. 2018;4:408–17.
4. Burrell RA, McGranahan N, Bartek J, Swanton C. The causes and consequences of genetic heterogeneity in cancer evolution. Nature. 2013; 501:338–45.
5. Jiang Y, Qiu Y, Minn AJ, Zhang NR. Assessing intratumor heterogeneity and tracking longitudinal and spatial clonal evolutionary history by next-generation sequencing. Proc Natl Acad Sci U S A. 2016;113:E5528–37.
6. Deshwar AG, Vembu S, Yung CK, Jang GH, Stein L, Morris Q. PhyloWGS: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. Genome Biol. 2015;16:35.
7. Zare H, Wang J, Hu A, Weber K, Smith J, Nickerson D, Song C, Witten D, Blau CA, Noble WS. Inferring clonal composition from multiple sections of a breast cancer. PLoS Comput Biol. 2014;10:e1003703.
8. Carter SL, Cibulskis K, Helman E, McKenna A, Shen H, Zack T, Laird PW, Onofrio RC, Winckler W, Weir BA, et al. Absolute quantification of somatic DNA alterations in human cancer. Nat Biotechnol. 2012;30:413–21.
9. Li B, Li JZ. A general framework for analyzing tumor subclonality using SNP array and DNA sequencing data. Genome Biol. 2014;15:473.
10. Oesper L, Mahmoody A, Raphael BJ. THetA: inferring intra-tumor heterogeneity from high-throughput DNA sequencing data. Genome Biol. 2013;14:R80.
11. Ha G, Roth A, Khattra J, Ho J, Yap D, Prentice LM, Melnyk N, McPherson A, Bashashati A, Laks E, et al. TITAN: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. Genome Res. 2014;24:1881–93.
12. Miller CA, White BS, Dees ND, Griffith M, Welch JS, Griffith OL, Vij R, Tomasson MH, Graubert TA, Walter MJ, et al. SciClone: inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution. PLoS Comput Biol. 2014;10:e1003665.
13. Navin NE. The first five years of single-cell cancer genomics and beyond. Genome Res. 2015;25:1499–507.
14. Navin N, Kendall J, Troge J, Andrews P, Rodgers L, McIndoo J, Cook K, Stepansky A, Levy D, Esposito D, et al. Tumour evolution inferred by single-cell sequencing. Nature. 2011;472:90–4.
15. Wang Y, Waters J, Leung ML, Unruh A, Roh W, Shi X, Chen K, Scheet P, Vattathil S, Liang H, et al. Clonal evolution in breast cancer revealed by single nucleus genome sequencing. Nature. 2014;512:155–60.
16. Gao R, Davis A, McDonald TO, Sei E, Shi X, Wang Y, Tsai PC, Casasent A, Waters J, Zhang H, et al. Punctuated copy number evolution and clonal stasis in triple-negative breast cancer. Nat Genet. 2016;48:1119–30.
17. Picelli S, Bjorklund AK, Faridani OR, Sagasser S, Winberg G, Sandberg R. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. Nat Methods. 2013;10:1096–8.
18. Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, Peshkin L, Weitz DA, Kirschner MW. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell. 2015;161:1187–201.
19. Patel AP, Tirosh I, Trombetta JJ, Shalek AK, Gillespie SM, Wakimoto H, Cahill DP, Nahed BV, Curry WT, Martuza RL, et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science. 2014;344:1396–401.
20. Chung W, Eum HH, Lee HO, Lee KM, Lee HB, Kim KT, Ryu HS, Kim S, Lee JE, Park YH, et al. Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. Nat Commun. 2017;8:15081.
21. Kim KT, Lee HW, Lee HO, Song HJ, Jeong da E, Shin S, Kim H, Shin Y, Nam DH, Jeong BC, et al: Application of single-cell RNA sequencing in optimizing a combinatorial therapeutic strategy in metastatic renal cell carcinoma. Genome Biol 2016, 17:80.
22. Tirosh I, Izar B, Prakadan SM, Wadsworth MH 2nd, Treacy D, Trombetta JJ, Rotem A, Rodman C, Lian C, Murphy G, et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science. 2016; 352:189–96.
23. Jerby-Arnon L, Shah P, Cuoco MS, Rodman C, Su MJ, Melms JC, Leeson R, Kanodia A, Mei S, Lin JR, et al. A cancer cell program promotes T cell exclusion and resistance to checkpoint blockade. Cell. 2018;175: 984–97 e924.
24. Tirosh I, Venteicher AS, Hebert C, Escalante LE, Patel AP, Yizhak K, Fisher JM, Rodman C, Mount C, Filbin MG, et al. Single-cell RNA-seq supports a developmental hierarchy in human oligodendroglioma. Nature. 2016;539: 309–13.
25. Venteicher AS, Tirosh I, Hebert C, Yizhak K, Neftel C, Filbin MG, Hovestadt V, Escalante LE, Shaw ML, Rodman C, et al. Decoupling genetics, lineages, and microenvironment in IDH-mutant gliomas by single-cell RNA-seq. Science. 2017;355.
26. Li H, Courtois ET, Sengupta D, Tan Y, Chen KH, Goh JJL, Kong SL, Chua C, Hon LK, Tan WS, et al. Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. Nat Genet. 2017;49:708–18.
27. Macaulay IC, Ponting CP, Voet T. Single-cell multiomics: multiple measurements from single cells. Trends Genet. 2017;33:155–68.
28. Dey SS, Kester L, Spanjaard B, Bienko M, van Oudenaarden A. Integrated genome and transcriptome sequencing of the same cell. Nat Biotechnol. 2015;33:285–9.
29. Macaulay IC, Haerty W, Kumar P, Li YI, Hu TX, Teng MJ, Goolam M, Saurat N, Coupland P, Shirley LM, et al. G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. Nat Methods. 2015;12:519–22.
30. Suva ML, Tirosh I. Single-cell RNA sequencing in cancer: lessons learned and emerging challenges. Mol Cell. 2019;75:7–12.
31. van Galen P, Hovestadt V, Wadsworth Ii MH, Hughes TK, Griffin GK, Battaglia S, Verga JA, Stephansky J, Pastika TJ, Lombardi Story J, et al. Single-cell RNA-Seq reveals AML hierarchies relevant to disease progression and immunity. Cell. 2019;176:1265–81 e1224.
32. Nam AS, Kim KT, Chaligne R, Izzo F, Ang C, Taylor J, Myers RM, Abu-Zeinah G, Brand R, Omans ND, et al. Somatic mutations and cell identity linked by genotyping of transcriptomes. Nature. 2019;571:355–60.
33. Raj A, Peskin CS, Tranchina D, Vargas DY, Tyagi S. Stochastic mRNA synthesis in mammalian cells. PLoS Biol. 2006;4:e309.
34. Jiang Y, Zhang NR, Li M. SCALE: modeling allele-specific gene expression by single-cell RNA sequencing. Genome Biol. 2017;18:74.
35. Padovan-Merhar O, Nair GP, Biaesch AG, Mayer A, Scarfone S, Foley SW, Wu AR, Churchman LS, Singh A, Raj A. Single mammalian cells compensate for differences in cellular volume and DNA copy number through independent global transcriptional mechanisms. Mol Cell. 2015;58:339–52.
36. Zafar H, Wang Y, Nakhleh L, Navin N, Chen K. Monovar: single-nucleotide variant detection in single cells. Nat Methods. 2016;13:505–7.
37. Piskol R, Ramaswami G, Li JB. Reliable identification of genomic variants from RNA-seq data. Am J Hum Genet. 2013;93:641–51.
38. Brennecke P, Anders S, Kim JK, Kolodziejczyk AA, Zhang X, Proserpio V, Baying B, Benes V, Teichmann SA, Marioni JC, Heisler MG. Accounting for technical noise in single-cell RNA-seq experiments. Nat Methods. 2013;10:1093–5.
39. Pierson E, Yau C. ZIFA: dimensionality reduction for zero-inflated single-cell gene expression analysis. Genome Biol. 2015;16:241.
40. Vallejos CA, Marioni JC, Richardson S. BASiCS: Bayesian analysis of single-cell sequencing data. PLoS Comput Biol. 2015;11:e1004333.
41. Ding B, Zheng L, Zhu Y, Li N, Jia H, Ai R, Wildberg A, Wang W. Normalization and noise reduction for single cell RNA-seq experiments. Bioinformatics. 2015;31:2225–7.
42. Qiu X, Hill A, Packer J, Lin D, Ma YA, Trapnell C. Single-cell mRNA quantification and differential analysis with Census. Nat Methods. 2017;14:309–15.
43. Deng Q, Ramskold D, Reinius B, Sandberg R. Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. Science. 2014;343:193–6.
44. Eirew P, Steif A, Khattra J, Ha G, Yap D, Farahani H, Gelmon K, Chia S, Mar C, Wan A, et al. Dynamics of genomic clones in breast cancer patient xenografts at single-cell resolution. Nature. 2015;518:422–6.
45. Gerlinger M, Rowan AJ, Horswell S, Math M, Larkin J, Endesfelder D, Gronroos E, Martinez P, Matthews N, Stewart A, et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. N Engl J Med. 2012;366:883–92.
46. Shi YJ, Tsang JY, Ni YB, Tse GM. Intratumoral heterogeneity in breast cancer: a comparison of primary and metastatic breast cancers. Oncologist. 2017;22:487–90.
47. Ribas A, Wolchok JD. Cancer immunotherapy using checkpoint blockade. Science. 2018;359:1350–5.
48. Schumacher TN, Schreiber RD. Neoantigens in cancer immunotherapy. Science. 2015;348:69–74.
49. Rizvi NA, Hellmann MD, Snyder A, Kvistborg P, Makarov V, Havel JJ, Lee W, Yuan J, Wong P, Ho TS, et al. Cancer immunology. Mutational landscape

determines sensitivity to PD-1 blockade in non-small cell lung cancer. Science. 2015;348:124–8.
50. Tumeh PC, Harview CL, Yearley JH, Shintaku IP, Taylor EJ, Robert L, Chmielowski B, Spasic M, Henry G, Ciobanu V, et al. PD-1 blockade induces responses by inhibiting adaptive immune resistance. Nature. 2014;515:568–71.
51. Twyman-Saint Victor C, Rech AJ, Maity A, Rengan R, Pauken KE, Stelekati E, Benci JL, Xu B, Dada H, Odorizzi PM, et al. Radiation and dual checkpoint blockade activate non-redundant immune mechanisms in cancer. Nature. 2015;520:373–7.
52. Benci JL, Johnson LR, Choa R, Xu Y, Qiu J, Zhou Z, Xu B, Ye D, Nathanson KL, June CH, et al. Opposing functions of interferon coordinate adaptive and innate immune responses to cancer immune checkpoint blockade. Cell. 2019;178:933–48 e914.
53. Patel SA, Minn AJ. Combination cancer therapy with immune checkpoint blockade: mechanisms and strategies. Immunity. 2018;48:417–33.
54. Goodman AM, Kato S, Bazhenova L, Patel SP, Frampton GM, Miller V, Stephens PJ, Daniels GA, Kurzrock R. Tumor mutational burden as an independent predictor of response to immunotherapy in diverse cancers. Mol Cancer Ther. 2017;16:2598–608.
55. Rosenthal R, Cadieux EL, Salgado R, Bakir MA, Moore DA, Hiley CT, Lund T, Tanic M, Reading JL, Joshi K, et al. Neoantigen-directed immune escape in lung cancer evolution. Nature. 2019;567:479–85.
56. Navin NE. Cancer genomics: one cell at a time. Genome Biol. 2014;15:452.
57. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102:15545–50.
58. Naxerova K, Reiter JG, Brachtel E, Lennerz JK, van de Wetering M, Rowan A, Cai T, Clevers H, Swanton C, Nowak MA, et al. Origins of lymphatic and distant metastases in human colorectal cancer. Science. 2017;357:55–60.
59. Wong JS, Warren LE, Bellon JR. Management of the regional lymph nodes in early-stage breast cancer. Semin Radiat Oncol. 2016;26:37–44.
60. Parker JS, Mullins M, Cheang MC, Leung S, Voduc D, Vickery T, Davies S, Fauron C, He X, Hu Z, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. J Clin Oncol. 2009;27:1160–7.
61. Zhang JY, Zhang F, Hong CQ, Giuliano AE, Cui XJ, Zhou GJ, Zhang GJ, Cui YK. Critical protein GAPDH and its regulatory mechanisms in cancer cells. Cancer Biol Med. 2015;12:10–22.
62. Tarrado-Castellarnau M, Diaz-Moralli S, Polat IH, Sanz-Pamplona R, Alenda C, Moreno V, Castells A, Cascante M. Glyceraldehyde-3-phosphate dehydrogenase is overexpressed in colorectal cancer onset. Transl Med Commun. 2017;2:6.
63. Mann M, Cortez V, Vadlamudi RK. Epigenetics of estrogen receptor signaling: role in hormonal cancer progression and therapy. Cancers (Basel). 2011;3:1691–707.
64. Green KA, Carroll JS. Oestrogen-receptor-mediated transcription and the influence of co-factors and chromatin state. Nat Rev Cancer. 2007;7:713–22.
65. Dreijerink KM, Mulder KW, Winkler GS, Hoppener JW, Lips CJ, Timmers HT. Menin links estrogen receptor activation to histone H3K4 trimethylation. Cancer Res. 2006;66:4929–35.
66. Kim H, Heo K, Kim JH, Kim K, Choi J, An W. Requirement of histone methyltransferase SMYD3 for estrogen receptor-mediated transcription. J Biol Chem. 2009;284:19867–77.
67. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA Jr, Kinzler KW. Cancer genome landscapes. Science. 2013;339:1546–58.
68. Tokheim CJ, Papadopoulos N, Kinzler KW, Vogelstein B, Karchin R. Evaluating the evaluation of cancer driver genes. Proc Natl Acad Sci U S A. 2016;113:14330–5.
69. Zhang W, Bojorquez-Gomez A, Velez DO, Xu G, Sanchez KS, Shen JP, Chen K, Licon K, Melton C, Olson KM, et al. A global transcriptional network connecting noncoding mutations to changes in tumor gene expression. Nat Genet. 2018;50:613–20.
70. Cuykendall TN, Rubin MA, Khurana E. Non-coding genetic variation in cancer. Curr Opin Syst Biol. 2017;1:9–15.
71. Singh M, Al-Eryani G, Carswell S, Ferguson JM, Blackburn J, Barton K, Roden D, Luciani F, Giang Phan T, Junankar S, et al. High-throughput targeted long-read single cell sequencing reveals the clonal and transcriptional landscape of lymphocytes. Nat Commun. 2019;10:3120.
72. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29:15–21.
73. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011;12:323.
74. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20:1297–303.
75. Pfeiffer F, Grober C, Blank M, Handler K, Beyer M, Schultze JL, Mayer G. Systematic evaluation of error rates and causes in short samples in next-generation sequencing. Sci Rep. 2018;8:10950.
76. Skelly DA, Johansson M, Madeoy J, Wakefield J, Akey JM. A powerful and flexible statistical framework for testing hypotheses of allele-specific gene expression from RNA-seq data. Genome Res. 2011;21:1728–37.
77. Ward JH. Hierarchical grouping to optimize an objective function. J Am Stat Assoc. 1963;58:236.
78. Goutte C, Hansen LK, Liptrot MG, Rostrup E. Feature-space clustering for fMRI meta-analysis. Hum Brain Mapp. 2001;13:165–83.
79. Urrutia E, Chen H, Zhou Z, Zhang NR, Jiang Y. Integrative pipeline for profiling DNA copy number and inferring tumor phylogeny. Bioinformatics. 2018;34:2126–8.
80. Li B, Chen W, Zhan X, Busonero F, Sanna S, Sidore C, Cucca F, Kang HM, Abecasis GR. A likelihood-based framework for variant calling and de novo mutation detection in families. PLoS Genet. 2012;8:e1002944.
81. Schliep KP. phangorn: phylogenetic analysis in R. Bioinformatics. 2011; 27:592–3.
82. Finak G, McDavid A, Yajima M, Deng J, Gersuk V, Shalek AK, Slichter CK, Miller HW, McElrath MJ, Prlic M, et al. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. Genome Biol. 2015;16:278.
83. Korthauer KD, Chu LF, Newton MA, Li Y, Thomson J, Stewart R, Kendziorski C. A statistical approach for identifying differential distributions in single-cell RNA-seq experiments. Genome Biol. 2016;17:222.
84. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol. 2018;36:411–20.
85. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38:e164.
86. Karosiene E, Lundegaard C, Lund O, Nielsen M. NetMHCcons: a consensus method for the major histocompatibility complex class I predictions. Immunogenetics. 2012;64:177–86.
87. Zhou Z. Genetic heterogeneity profiling by single cell RNA sequencing. Github: https://github.com/zhouzilu/DENDRO; 2019. Accessed 28 Oct 2019.
88. Zhou Z: Genetic heterogeneity profiling by single cell RNA sequencing Zenodo: https://doi.org/10.5281/zenodo.3521087; 2019. Accessed 29 Oct 2019.
89. Deng Q, Ramsköld D, Reinius B, Sandberg R: Single-cell RNA-Seq reveals dynamic, random monoallelic gene expression in mammalian cells. GSE45719: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE45719; 2014. Accessed 10 Jan 2014.
90. Kim KT, Lee HW, Lee HO, Song HJ, Jeong da E, Shin S, Kim H, Shin Y, Nam DH, Jeong BC, et al: Single-cell transcriptome profiling for metastatic renal cell carcinoma patient-derived cells. GSE73122: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE73122; 2015. Accessed 18 Sept 2015.
91. Chung W, Eum HH, Lee HO, Lee KM, Lee HB, Kim KT, Ryu HS, Kim S, Lee JE, Park YH, et al: Single cell RNA sequencing of primary breast cancer. GSE75688: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE75688; 2016. Accessed 09 Dec 2016.
92. Zhou Z, Xu B: DENDRO: genetic heterogeneity profiling and subclone detection by single-cell RNA sequencing. GSE139248: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE139248; 2019. Accessed 10 Nov 2019.

Publisher's Note