Genetic Heterogeneity Profiling by Single Cell RNA Sequencing

Zilu Zhou, Graduate Group in Genomics and Computational Biology, University of Pennsylvania, Philadelphia, PA, USA. zhouzilu@pennmedicine.upenn.edu

Nancy R Zhang*, Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA, USA. nzh@wharton.upenn.edu

* Corresponding author

Abstract

Detection of genetically distinct subclones and profiling the transcriptomic differences between subclones is important for studying the evolutionary dynamics of tumors, as well as for accurate prognosis and effective treatment of cancer in the clinic. Single cell RNA-sequencing (scRNA-seq) is now ubiquitously adopted in ongoing and planned cancer studies for the profiling of intra-tumor transcriptional heterogeneity. Inference of clonal membership from scRNA-seq is currently unreliable. We propose DENDRO, an analysis method for scRNA-seq data that detects genetically distinct subclones, assigns each single cell to a subclone, and reconstructs the evolutionary tree. DENDRO utilizes information from single nucleotide mutations in transcribed regions and accounts for single-cell level expression stochasticity and technical noise. The accuracy of DENDRO is benchmarked on spike-in datasets and on scRNA-seq data with known subpopulation structure. We applied DENDRO to two scRNA-seq data sets, where the new approach allowed us to elucidate the interplay between genetic and transcriptomic intratumor variation.

Keywords

single-cell RNA sequencing, intratumor heterogeneity, clonal reconstruction, statistical modeling

Background

Tumor initiation, progression, metastasis and relapse are mediated by DNA alterations[1]. Single nucleotide alteration (SNA) is a basic type of DNA alteration that can affect the transcriptome and

37 contribute to intratumor heterogeneity [2]. Intratumor heterogeneity is a key mediator of patients'
38 response to treatment, as clonal selection can lead to the emergence of subclones that are drug
39 resistant [3]. Thus, identifying subclonal DNA alterations and assessing their impact on subclonal
40 transcriptional dynamics can elucidate the mechanisms of tumor evolution and determine potential
41 gene targets for therapy. To characterize intratumor genetic heterogeneity, most prior studies have
42 used bulk tumor DNA sequencing [4-11], but these approaches have limited resolution and power [12].

43 Breakthroughs in single-cell genomics promise to reshape cancer research by allowing comprehensive
44 cell type classification and rare subclone identification. For example, in breast cancer, single-cell DNA
45 sequencing (scDNA-seq) was used to distinguish normal cells from malignant cells, the latter of which
46 were further classified into subclones [13-15]. However, scDNA-seq still suffers from low coverage,
47 intensive labor and low consistency, hindering its widespread adoption in the cancer field [16]. Single-
48 cell RNA sequencing (scRNA-seq), on the other hand, is comparatively more mature and cost effective.
49 ScRNA-seq technologies such as Smart-seq2 [17], Drop-seq [18], and 10X Genomics Chromium™, which
50 can simultaneously profile thousands of cells with reasonable cost, have advanced tumor transcriptome
51 profiling to single cell resolution [19]. Previous single-cell studies revealed novel insights into the
52 developmental and immunological programs of gene expression in glioblastoma and breast cancer [20,
53 21]. Kim et al. and Tirosh et al. used scRNA-seq to elucidate drug target pathways and drug response
54 variability across samples and tumor metastasis [22, 23]. More recently, studies have focused on tumor
55 cellular architecture and tumor cell lineage in gliomas, leading to new expression signatures that are
56 predictive of prognosis [24-26]. However, despite the recognized importance of clonal DNA mutations in
57 tumor pathogenesis, this crucial aspect has not been explored in detail by these scRNA-seq studies, due
58 to the lack of appropriate computational methods for analyzing DNA-level mutation heterogeneity using
59 scRNA-seq data.

60 Only a small portion of the SNAs of each cell are expected to be seen in the read output of scRNA-seq.
61 To be sequenced, a SNA needs to fall in a transcribed region of the genome, at a location within the
62 transcript that will eventually be read. Gene transcription in single cells is typically bursty [27-29], and
63 thus an SNA residing in a gene that is expressed at the bulk tissue level may not be observed in a
64 particular cell, simply because the mutated allele, by chance, is not expressed in the given cell. We refer
65 to alleles that are not captured due to expression stochasticity as *biological dropouts*. Even if a mutated
66 allele is expressed, it has to be successfully converted to cDNA and then sequenced to be represented in
67 the final read output; we refer to alleles lost at this step as *technical dropouts*. In addition to dropout
68 events, post-transcriptional modification, such as RNA editing, and sequencing errors impede both
69 sensitivity and specificity of SNA discovery. As a result, methods developed for single cell SNA detection
70 using scDNA-seq, such as Monovar [30], as well as methods designed for SNA detection in bulk DNA or
71 RNA sequencing data do not yield accurate results in the scRNA-seq setting [31-36].

72 Intratumor genetic heterogeneity gives rise to intratumor transcriptomic heterogeneity, and cell-to-cell
73 transcriptional variation is driven by inter-clonal competition. Thus, in the analysis of scRNA-seq data of
74 tumor samples, the accurate classification of cells into clones and the characterization of each clone's
75 genetic alterations should be an integral step. The potential of identifying subclonal DNA mutations
76 using scRNA-seq data and the prospect of using these alterations for clonal inference has not been
77 extensively explored. This analysis gap is compounded by the lack of technology for profiling DNA
78 mutations and RNA expression in the *same* cell with acceptable accuracy, throughput, and cost.
79 Although one can apply scDNA-seq and scRNA-seq to the same tumor, the mutation analysis and RNA
80 quantification could not be conducted in the same set of cells.

81 Here we develop a new statistical and computational framework – DNA based EvolutionNary tree
82 preDiction by scRNA-seq technOlogy (DENDRO) - that reconstructs the phylogenetic tree for cells

83 sequenced by scRNA-seq based on genetic divergence calculated from DNA-level mutations. DENDRO
84 assigns each cell to a clone in the tree, and, for each clone, infers its mutation profile. DENDRO achieves
85 high power to detect genetically divergent subclones by addressing challenges unique to scRNA-seq,
86 including transcriptional variation and technical noise. A DENDRO clustering of scRNA-seq data allows
87 joint genetic and transcriptomic analysis on the same set of cells. We evaluate the accuracy of DENDRO
88 through spike-in benchmark data sets, and then illustrate its application through two case studies. In the
89 first case study on a metastasized renal cell carcinoma dataset, we validate DENDRO's results using
90 known subpopulation labels, and further elucidate the genetic-transcriptomic changes in a PDX tumor
91 model. In the second case study on a primary breast tumor dataset, DENDRO detects unbiased
92 subclones and jointly characterizes the transcriptomic and genetic changes between cells in lymph
93 nodes and primary resections, illustrating the clinical essence of studying intratumor heterogeneity. The
94 DENDRO package, implemented in R, is available at https://github.com/zhouzilu/DENDRO, where we
95 also provide a power calculation toolkit, DENDROplan, to aid in the design of scRNA-seq experiments for
96 subclonal mutation analysis using DENDRO.

97

98 ## Results

99 *Overview of DENDRO model and pipeline*

100 Figure 1A shows an overview of DENDRO's analysis pipeline. Input data can be in fastq or BAM format.
101 For fastq files, we recommend STAR 2-pass method for alignment, which accounts for splicing junctions
102 and achieve higher mapping quality [37]. In the next step, GATK tools are used for SNA detection [38];
103 parameters for this step are chosen to allow high sensitivity but low specificity, since the SNA loci will be
104 further refined in downstream analysis (see Methods and Figure S1 for details). Per cell counts of total

105 read coverage (N matrix) and mutation allele read coverage (X matrix) at SNA locations are extracted.
106 Based on these matrices, DENDRO then computes a cell-to-cell genetic divergence matrix, where entry
107 (c, c') of the matrix is a measure of the genetic divergence between cells c and c'. Details of this genetic
108 divergence evaluation are given in the next section. DENDRO next clusters the cells into genetically
109 distinct clones based on this pairwise divergence matrix, and selects the number of clones based on
110 inspection of the intra-cluster divergence curve. Reads from the same clone are then pooled together,
111 and the SNA profile for each clone is re-estimated based on the pooled reads, significantly improving
112 upon the previous SNA profiles computed at the single cell level. Finally, DENDRO generates a parsimony
113 tree using the clone-level mutation profiles, to more accurately reconstruct their evolutionary
114 relationship.

115

116 *Genetic divergence evaluation*

117 Due to the high rate of biological and technical dropout, SNA detection by GATK within each cell lacks
118 sensitivity. We also expect a high false positive rate due to the known high base error rate in scRNA-seq
119 protocols. Given this, simple approaches such as the Hamming or Euclidean distance computed using
120 the raw SNA genotype matrix or the allele frequency matrix do not accurately measure the genetic
121 divergence between cells.

122 To more accurately estimate the cell-to-cell genetic divergence, we developed a statistical model that
123 accounts for technical dropout, sequencing error and expression stochasticity. Consider two cells, c
124 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'
125 come from the same subclone and thus share the same SNA profile. Let $X_c = (X_{c1}, \dots, X_{cm})$ be the

126 mutation allele read counts for this cell at the m SNA sites profiled, and let $N_c = (N_{c1}, \dots, N_{cm})$ be the

127 total read counts at these sites. We define the genetic divergence between the two cells as

128 $$d_{cc'} = \log P(X_c, X_{c'} | N_c, N_{c'}, I_c = I_{c'}) = \sum_{g=1}^{m} d_{cc'}^{g}$$

129 $$\text{where } d_{cc'}^{g} = -\log P\left(X_{cg}, X_{c'g} \middle| N_{cg}, N_{c'g}, I_c = I_{c'}\right)$$

130 In other words, $d_{cc'}$ is the negative log likelihood of the mutation allele counts of cells c and c', given

131 the total read counts and the event that the two cells belong to the same clone. If c and c' have

132 mutations in mismatched positions, the likelihood for X_c, $X_{c'}$ conditioned on $I_c = I_{c'}$ would be small,

133 giving a large value for $d_{cc'}$. By the assumption of independence between sites, $d_{cc'}$ is the sum of $d_{cc'}^{g}$,

134 where $d_{cc'}^{g}$ is the contribution of mutation site g to the divergence measure. In characterizing the

135 conditional distribution for X_{cg} and $X_{c'g}$, we use a Beta-Binomial distribution to model expression

136 stochasticity and a Binomial model to capture sequencing errors and rare RNA-editing events, please see

137 Methods for details.

138 To illustrate how transcriptional stochasticity affects the divergence calculation, consider the toy

139 example shown in Figure 1B, where we compare the contributions of two mutation sites, g and g', to

140 the pairwise divergence between cell c_1 and cell c_2. First, note that site g is identical to site g' in

141 mutation allele frequency and total read count in both cells. However, they contribute different values

142 of $d_{cc'}$ due to their different expression stochasticity: Gene g is bursty, that is, expression of both alleles

143 is sporadic leading to a U-shaped mutation allele frequency distribution across cells. In contrast, gene g'

144 is constitutive, that is, expression of both alleles is stable, leading to a bell shaped allele-frequency

145 distribution across cells. Thus, biological dropout is more likely for gene g than for gene g', which means

146 that, despite the identical mutation allele frequency and total coverage, $d_{c_1c_2}^{g} < d_{c_1c_2}^{g'}$.

147 *Accuracy assessment using in silico spike-ins*

148 We designed a spike-in procedure to assess the accuracy of DENDRO and to make realistic power
149 projections for subclone detection (Figure 2a). The procedure, in its basic mode, starts with an assumed
150 evolutionary tree, where the leaves are clades and mutations can be placed on the branches. In the
151 absence of prior information, a simple tree structure is used, such as the one shown in Figure 2A.
152 Parameters of simulation are (1) total number of mutations, (2) total number of cells, (3) the proportion
153 of cells in each clade, (4) the proportion of mutations along each branch, and (5) mean read coverage
154 across loci. Some of these parameters can be determined using bulk DNA-seq and/or bulk RNA-seq data,
155 if available, see our description of DENDROplan in next section. Parameters (1-4) determine the
156 mutation matrix Z. To get the matrix of alternative and total read counts for each mutation loci in each
157 cell (X_{cg} and N_{cg}, respectively), we overlay the mutation matrix onto the scRNA-seq data, see details in
158 Methods. This allows the spike-in framework to retain the expression stochasticity and sequencing error
159 of real scRNA-seq data. DENDRO is applied to the spike-in data to obtain the clonal clusters, which is
160 then compared with the known labels. Subclone detection is evaluated by three metrics: adjusted Rand
161 index, capture rate and purity. Details on these metrics are given in the Supplementary Material.

162 Using the above framework, we conducted a systematic evaluation of DENDRO's subclone detection
163 accuracy on an example scRNA-seq dataset (Figure 2). The results show: (1) As mutation burden of a
164 clade increases, the adjusted Rand index and clade-specific capture rate decrease, but cluster purity of
165 interested clade increases (Figure 2B), because there are few mutations remaining to distinguish the
166 other two subclones; (2) similarly, as cell proportion of a clade increases, its adjusted Rand index and
167 capture rate decrease, but cluster purity increases (Figure 2C); (3) more mutation site increases DENDRO
168 performance; (4) lower sequencing depth decreases accuracy; (5) As false positive rate of mutation calls
169 increasing, DENDRO performs worse; (6) lower sequencing error rate improves DENDRO accuracy; and

170 (7) variation in total expression of genes does not influence DENDRO's divergence measure (Figure S3).
171 The last point is due to DENDRO's reliance only on the distribution of the mutation allele frequency
172 conditioned on the total read coverage, and further illustrated by mutation spike-ins in a mixer cell
173 population with different expression pattern. This indicates that the divergence evaluation reflects
174 solely genetic distance not transcriptomic difference. We further illustrate this property by a simulation
175 study. See Supplementary Material and Figure S2 for details.

176 As a result, DENDRO's subclone identification is based solely on genetic divergence, and not on
177 expression profile. Previous scRNA-seq studies identify subclones based on total gene expression,
178 making it difficult to separate genetic and transcriptomic variations. Clustering based on total expression
179 counts and then performing differential gene expression analysis is prone to self-fulfilling prophecy, as
180 there will indeed be differential expressed genes because this was the clustering criteria. DENDRO's
181 approach makes the downstream differential gene expression analysis between subclones more
182 interpretable.

183 *Power analysis toolkit and experimental design*

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

190 As shown in Figure 2a, if bulk DNA sequencing and/or RNA sequencing data are available for the tumor
191 being studied, these data can be harnessed to make more realistic power calculations. For example, if

192 SNAs have been profiled using bulk DNA sequencing data, the set of mutations that lie in the exons of
193 annotated genes can be retrieved and used directly in constructing the spike-in data. Furthermore,
194 phylogeny construction algorithms for bulk DNA-seq data can be used to infer a putative tree structure
195 that can be used as input to DENDROplan [4, 39]. If bulk RNA-seq data is available, the bulk expression
196 level of the mutation-carrying genes can be used to predict the expression level of the mutation in the
197 single cell data. For details, see Methods. The power analysis tool is also available at
198 https://github.com/zhouzilu/DENDRO.

199 *Evaluation of DENDRO on a renal cell carcinoma and its metastasis*

200 We applied DENDRO to the analysis of the renal cell carcinoma dataset from Kim et al [22] (Figure 3).
201 This dataset is composed of 116 cells from three different tumors derived from one single patient: a
202 patient-derived xenograft (PDX) from the primary renal cell carcinoma (PDX_pRCC), a biopsy of the
203 metastasis to the lung 1 year after treatment of primary site (Pt_mRCC), and a PDX of the lung
204 metastasis renal cell carcinoma (PDX_mRCC) (Figure S4A). Transcriptomic profiles of each individual cell
205 were generated using the Smart-seq protocol [17]. The cells should share common early driver
206 mutations due to their shared origin from the same patient, but the metastasis and the cultivation of
207 each tumor in separate medium (human or mouse) should have allowed for the accumulation of new
208 mutations. Thus, we expect the three tumors to be clonally distinct. Given this prior expectation, we use
209 this dataset to examine the accuracy of DENDRO and to illustrate how DENDRO enables joint analysis of
210 the genetic and transcriptomic heterogeneity within the same group of cells.

211 GATK detected 2,867,029 mutation sites across all cells [38]. Mutations that are detected in less than 5%
212 (too rare) or more than 95% (too common) of the cells were removed, which leaves 72,206 mutations.
213 On average, 10801 mutations are detected in each cell and 17.35 cells possess the same mutation for

214 each loci (Figure S4B, C). For majority sites, only few cells have nonzero read coverage, highlighting the
215 fact that many mutations are missed due to technical and biological dropout (Figure S4D) [32-36].

216 We compared 4 different clustering methods: DENDRO, hierarchical clustering based on the primary
217 genotype matrix Z generated by GATK ($Z_{cg} = 1$ when a mutation g is detected for cell c, $Z_{cg} =$
218 0 otherwise), hierarchical clustering based on the X/N matrix that preserve the variant allele
219 information and hierarchical clustering based on gene expression ($logTPM$). DENDRO gives the cleanest
220 separation between the three populations with adjusted Rand Index of 0.932 (1.0 indicates perfect
221 clustering, Figure 3A), as compared to 0.754 for Z matrix (Figure 3B), 0.519 for X/N (Figure 3C) and 0.489
222 for expression (Figure 3D). Inspection of the graph shows that, as expected, divergence between
223 primary tumor and metastasis exceeds divergence between patient sample and PDX sample, as
224 PDX_mRCC clusters with Pt_mRCC rather than PDX_pRCC. All of the other three methods successfully
225 separated the primary sample from the metastatic samples, but cannot differentiate between the two
226 metastasis samples. For DENDRO, the intra-cluster divergence curve kinks sharply at 3, and thus we
227 stopped at 3 clusters (Figure 3E and Methods). We annotated the clusters as PDX_mRCC, PDX_pRCC and
228 Pt_mRCC by their cell compositions (Table S1A). The fact that DENDRO found little mixing of cells
229 between the three tumors, and low genetic heterogeneity within each tumor, is unsurprising given how
230 these tumors were derived: Since relapsed metastasis consists of cells that have already underwent a
231 selection bottleneck, and since the PDX tumor is seeded by a small subsample of cells from the original
232 tumor, it is not surprising that no subclones are detected in each of the three tumors [40-42]. By
233 maximum likelihood, we called a total of 13,454 mutations after pooling the cells within each cluster
234 (Figure 3F and Methods). The unrooted phylogenetic tree computed with the revised mutation profiles
235 confirms that genetic divergence between metastasis and primary is greater than divergence between
236 PDX and primary sample.

237 DENDRO enables simultaneous clonal assignment and transcriptomic profiling of the same set of cells.
238 Plot of smoothed expression ordered by DENDRO shows unique expression patterns within each
239 subclone (Figure S6). We focused on the comparison of the two metastasized cell populations
240 (metastasis to lung and patient derived mouse xenograph). Even though PDX_mRCC was derived from
241 Pt_mRCC, the DENDRO analysis found substantial genetic divergence between the two cell populations.
242 To investigate further, we performed a differential expression analysis between PDX_mRCC and
243 Pt_mRCC with scDD and MAST, detecting 74 significant differentially expressed genes (Methods, Figure
244 S7E, Table S2A) [43-45]. Gene ontology analysis classified these 74 genes into two subgroups: immune-
245 related genes and cancer-related genes (Table 1, Table S2F) [46]. Immune-related differentially
246 expressed genes are enriched for the terms TNF-α signaling, complement system and allograft rejection.
247 On the other hand, cancer related differentially expressed genes overlap with the pathways including
248 hypoxia, KRAS signaling, mTORC1 signaling and epithelial mesenchymal transition.

249 Simultaneously, we compare the mutation profiles of these two clones. 9521 locus have different
250 mutated allele counts between these two populations and were further annotated by ANNOVAR [47].
251 After filtering, the preserved variants associated with 24 out of 74 differential expressed genes (Table
252 S2A). Next, we performed a similar GSEA on variants associated genes to identify mutation-related
253 pathway [46]. Interestingly, variant annotated genes are enriched in cancer-related pathways, including
254 mitotic spindle, mTORC1 signaling, EMT and hypoxia, overlapping substantially with the cancer-related
255 pathways identified by differential expression analysis; in comparison, none of the differentially
256 expressed genes from immune pathways showed up in this mutated gene analysis (Table 2, Table S2F).
257 In another word, cancer-related transcriptomic divergence between PDX_mRCC and Pt_mRCC is driven
258 directly by genetic alterations in the same genes, but immune-related differential expression is
259 influenced by non-DNA factors. This makes sense, since implantation of tumor cells from human to mice
260 alters their immune microenvironment [48-50], and thus is expected to alter immune-related signaling

261 within the implanted tumor cells. This illustrates how DENDRO extricates DNA variation from RNAs

262 allowing heir joint analysis. Differential expression and differential mutation analysis for the other

263 subclone pairs can be found in Table S2 and Figure S7.

264 Table 1. GO analysis on Differential Expressed Genes between Pt_mRCC and PDX_mRCC

Gene Set Name [# Genes (K)]				**p-value**	**FDR q-value**
Color key:	Cancer-related pathway	Immune-related pathway	Other		
HALLMARK_TNFA_SIGNALING_VIA_NFKB [200]				1.86 e-8	9.28 e-7
HALLMARK_HYPOXIA [200]*				5.04 e-7	1.26 e-5
HALLMARK_MTORC1_SIGNALING [200]*				1.15 e-5	1.92 e-4
HALLMARK_COMPLEMENT [200]				2.15 e-4	1.79 e-3
HALLMARK_GLYCOLYSIS [200]*				2.15 e-4	1.79 e-3
HALLMARK_KRAS_SIGNALING_UP [200]				2.15 e-4	1.79 e-3
HALLMARK_ALLOGRAFT_REJECTION [200]				3.17 e-3	1.58 e-2
HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSINSITION [200]*				3.17 e-3	1.58 e-2
HALLMARK_ESTROGEN_RESPONSE_EARLY [200]				3.17 e-3	1.58 e-2
HALLMARK_INFLAMMATORY_RESPONSE [200]				3.17 e-3	1.58 e-2

265

266 Table 2. GO analysis on Differential Mutated Genes between Pt_mRCC and PDX_mRCC

Gene Set Name [# Genes (K)]				**p-value**	**FDR q-value**
Color key:	Cancer-related pathway	Immune-related pathway	Other		
HALLMARK_UV_RESPONSE_DN [144]				1.39 e-29	6.93 e-28
HALLMARK_MYC_TARGETS_V1 [200]				2.75 e-27	6.87 e-26
HALLMARK_MITOTIC_SPINDLE [200]				7.97 e-24	1.33 e-22
HALLMARK_MTORC1_SIGNALING [200]*				9.49 e-20	1.19 e-18
HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSINSITION [200]*				1.91 e-17	1.91 e-16
HALLMARK_OXIDATIVE_PHOSPHORYLATION [200]				1.06 e-16	8.81 e-16
HALLMARK_HYPOXIA [200]*				5.69 e-16	4.06 e-15
HALLMARK_GLYCOLYSIS [200]*				2.97 e-15	1.86 e-14
HALLMARK_ANDROGEN_RESPONSE [101]				1.07 e-14	5.96 e-14
HALLMARK_HEME_METABOLISM [200]				7.4 e-14	3.7 e-13

267

268 *Analysis of scRNA-seq dataset of primary breast cancer*

269 We next demonstrate the application of DENDRO to a primary breast cancer dataset [21]. We focused
270 on two tumors (BC03 and BC09) that have the most cell sequenced (Table S3 and Figure S8). BC03
271 includes cells from primary tumor (here after BC03P) and from regional metastatic lymph nodes (here
272 after BC03LN). 132 single cell transcriptomes were profiled by Smart-seq protocol [17]. For this analysis,
273 we will first examine whether DENDRO can separate BC03 and BC09 based on their genetic differences,
274 and then, jointly characterize the transcriptomic and genetic changes between cells in lymph nodes and
275 primary resections.

276 GATK [46] detected a total of 2,364,823 mutation sites across the 132 cells, 353,647 of which have
277 mutation frequency greater than 5% and smaller than 95%, and were retained for downstream analysis
278 (Figure S8A, B, C). Figure 4 shows the clustering given by DENDRO. DENDRO separates BC09 cells from
279 BC03 cells with 100% accuracy (Figure 4A). The intra-cluster divergence curve kinks at 5 clones between
280 BC03 and BC09, yielding 3 subclones for BC03 and 2 for BC09 (Figure 4A, Figure S8D and Table S1B).
281 Within BC03, clones Mix_1 and Mix_2 contain a mixture of cells from the primary tumor and lymph
282 nodes, and clone LN_1 contains mostly cells from the lymph nodes. This suggests that tumor cells that
283 have transitioned to the lymph nodes are genetically heterogeneous, with some cells remaining
284 genetically similar to the primary population and others acquiring new genetic mutations. In
285 comparison, hierarchical clustering based on expression (using log transcripts-per-million values) does
286 not separate BC03 from BC09, and give a negative adjusted Rand index within BC03, indicating worse
287 than random classification (Figure 4B).

288 DENDRO pooled cells within each of the 5 clusters and re-estimated their mutation profiles. We define a
289 variant as subclonal if it is not present in all of the subclones within a tumor. Based on marginal
290 likelihood, we pick the top 10,000 variants to construct a phylogenetic tree (Figure 4C). As expected, the
291 two BC09 clusters are far separate from the BC03 clusters. Within BC03, the length of the branches

292 shows that the subclone containing mostly cells from lymph nodes (labeled BC03LN_1) is genetically
293 similar to the Mix_2 clone, and distant from the Mix_1 clone (Figure 4C).

294 To profile clone-specific gene expression patterns, we first plot window-smoothed expression with cells
295 grouped by DENDRO clustering [23]. Unsurprisingly, there are broad chromosome-level differences in
296 expression patterns between groups (Figure S9), most likely due to copy number aberrations and/or
297 epigenetic remodeling. Next, we focused on the expression of the PAM50 genes (Figure 4D), and found
298 substantial differences in their expression between subclones. DENDRO detected one rare subclone,
299 BC09_2, with only six cells (<5% of the total number of cells) which has strong basal-like signature.
300 Interestingly, in BC03, the subclone LN_1 has the TNBC/basal-like subtype with invasive gene signature,
301 while the subclone Mix_2 has the *ESR1*$^{+}$ subtype. Thus, the genetic divergence of LN_1 from Mix_2 is
302 accompanied by its acquisition of an invasive metastatic expression signature. If the comparison were
303 made directly between cells from the primary site and cells from the lymph node without DENDRO's
304 subclone assignment, these expression patterns would be much weaker since the clones do not cleanly
305 separate by site. Comparing with original paper assigning each tumor to one specific breast cancer
306 subtype, with the aid of DEDNRO, subtype identification in sub-tumor resolution allows better therapy
307 design to potentially target all subclones and reduce tumor relapse risk.

308 We next conducted a transcriptome-wide search for pathways that have differential expression
309 between clones (Methods and Table S4), and assessed their overlap with pathways that are
310 differentially mutated between clones. Focusing on tumor BC03, pathways for G2M checkpoint and
311 *KRAS* signaling are up-regulated in lymph node metastasis subclone BC03LN_1, while pathways for
312 estrogen response and apoptosis are down-regulated, indicating more invasive phenotype (Table S4E).
313 In addition, *GAPDH* is up-regulated in the metastatic clone (BC03LN_1) and down-regulated in the two
314 mix-cell clones, consistent with previous findings [51, 52] (Figure S10D). Differentially expressed genes

315 between other subclone pairs in BC03 are also enriched in estrogen response, apoptosis, and DNA
316 repair. (Table S4C, D). In parallel, subclone-specific mutated genes are highly enriched in cancer-related
317 pathways including MYC target, G2M checkpoints and mitotic spindle, and immune related pathways
318 such as, interferon response, TNF-a signaling and inflammatory response (Table S4). Interestingly, few of
319 the differentially mutated genes are associated with estrogen and androgen responses, suggesting that
320 differential expression of hormone related genes are not mediated directly by genetic mutations in
321 these pathways. This is consistent with the recent studies that found epigenetic alteration, such as
322 histone acetylation and methylation, to underlie hormones receptor signaling in breast cancer [53-56].
323 DNA-RNA joint analysis between other subclones are included in the Table S4 and Figure S10. Overall,
324 this example illustrates how DENDRO enables the joint assessment of genetic and transcriptomic
325 contributions to clonal diversity at single-cell resolution.

326

327 Discussion

328 We have described DENDRO, a statistical framework to study intratumor DNA-level heterogeneity using
329 scRNA-seq data. DENDRO starts with mutations detected directly from the scRNA-seq reads, which are
330 very noisy due to a combination of factors: Errors are introduced in the reverse-transcription,
331 sequencing and mapping steps in the experiment, low sequencing depth and molecule conversion
332 efficiency lead to technical dropouts, and expression burstiness at the single cell level lead to biological
333 dropouts. DENDRO overcomes these obstacles by utilizing the fact that cells from the same subclone
334 possess almost identical mutation profile. Given noisy mutation profiles and allele specific read counts,
335 DENDRO first reconstructs clonal substructure by computing a distance between each pair of cells that
336 quantifies their genetic divergence after accounting for transcriptional bursting, dropout and sequencing
337 error. Then, DENDRO re-estimates subclone-specific mutation profiles by pooling reads across cells

338 within the same subclone. These re-estimated mutations profiles are then passed to downstream
339 mutation analysis and phylogenetic tree reconstruction.

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

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

356 We have demonstrated that proper computational modeling can excavate the DNA-level heterogeneity
357 in scRNA-seq data. Yet, there is always limitation in working with RNA. While rare RNA editing events are
358 absorbed by the parameter ϵ, DENDRO cannot distinguish subcluster-specific constituent RNA editing
359 events from subcluster-specific DNA mutations. In the extreme and unlikely scenario where RNA editing
360 events are common and pervasive, DENDRO's cluster would reflect RNA editing. In such cases, we

361 recommend using matched bulk DNA-seq of the same tumor to filter the loci detected in the first step of
362 DENDRO, keeping only those that are supported by at least on read in the bulk DNA-seq data.

363

364 Conclusions

365 We have developed DENDRO, a statistical method for tumor phylogeny inference and clonal
366 classification using scRNA-seq data. DENDRO accurately infers the phylogeny relating the cells and
367 assigns each single cell from the scRNA-seq data set to subclone. DENDRO allows us to (1) cluster cells
368 based on genetic divergence while accounting for transcriptional bursting, technical dropout and
369 sequencing error, as benchmarked by in silico mixture and a spike-in analysis, (2) characterize the
370 transcribed mutations for each subclone, and (3) perform DNA-RNA joint analysis with the same set of
371 cells and examine the relationship between transcriptomic variation and mutation profile. We illustrate
372 DENDRO through two case studies. In the first case study on a metastasized renal cell carcinoma
373 dataset, we validated DENDRO's results using known subpopulation labels and elucidated the genetic-
374 transcriptomic changes in the PDX tumor model. In the second case study on a primary breast tumor
375 dataset, DENDRO brought forth new insights on the interplay between intratumor transcriptomic
376 variation and subclonal divergence.

377

378 Methods

379 *SNA calling pipeline*

380 Figure S1 illustrates the SNA calling pipeline. Raw scRNA-seq data is aligned by STAR 2-pass method [37].
381 Raw variants calling was made using the Haplotype Caller (GATK tool) on the BAM files after sorting,
382 joining read groups, removing duplicated reads, removing overhangs into intronic regions, realigning
383 and recalibration [38]. Conventionally, there are two methods from GATK tools for mutation detection:
384 haplotype caller and mutect2. Haplotype caller has a RNA-seq setting which handle splice junctions
385 correctly, but assumes VAF around 50%, while mutect2 can detect mutations with low VAF but does
386 account for splice junction. The reason we select haplotype caller instead of mutect2 is that we extract
387 allele read counts for all cells as long as one of the cells is listed as carrying the mutation. Thus, as long
388 as one cell has VAF reaching 50%, this mutation would be detected. Calls with stand_call_conf greater
389 than 20 and population frequency greater than 5% but less than 95% were preserved for further
390 analysis. Admittedly, such lenient filtering typically introduces false positive sites. However, our priority
391 at this step is to minimize false negative rate, while the genetic divergence matrix in the following step
392 robustly estimates cell population substructure. Both the coverage of the alternative allele and the total
393 read coverage are extracted for each site for further analysis.

394

395 *Genetic Divergence and Beta-Binomial framework*

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

399 $$d_{cc'}^{g} = \log \frac{1}{P(X_{cg}, X_{c'g} | N_{cg}, N_{c'g}, I_c = I_{c'})}$$

400 $$= \log \frac{\mathrm{P}\left(X_{cg}, X_{c'g} \middle| N_{cg}, N_{c'g}, I_c = I_{c'}\right) + \mathrm{P}(X_{cg}, X_{c'g} | N_{cg}, N_{c'g}, I_c \neq I_{c'})}{P(X_{cg}, X_{c'g} | N_{cg}, N_{c'g}, I_c = I_{c'})})$$

401 where $X_c = (X_{c1}, X_{c2}, \dots X_{cg}, \dots X_{cm})$ are the mutation allele read counts for cell c and $N_c =$

402 $(N_{c1}, N_{c2}, \dots N_{cg}, \dots N_{cm})$ are the total read counts at these sites. More intuitively, if cells c and c' are not

403 from the same clonal group, the numerator has larger value compared to denominator. Thus $d_{cc'}^{g}$ is

404 large, indicating bigger divergence between the two cells. With further derivation (see Methods), $d_{cc'}^{g}$ is

405 a function of the five following probabilities:

406 $d_{cc'}^{g}$

407 $= f\left(P_g; P\left(X_{cg} \middle| N_{cg}, Z_{cg} = 0\right); P\left(X_{cg} \middle| N_{cg}, Z_{cg} = 1\right); P\left(X_{c'g} \middle| N_{c'g}, Z_{c'g} = 0\right); P\left(X_{c'g} \middle| N_{c'g}, Z_{c'g} = 1\right)\right)$

408 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

409 the cells estimated by GATK calls.

410 In the above formula for $d_{cc'}^{g}$, $P\left(X_{cg} \middle| N_{cg}, Z_{cg} = 0\right)$ and $P\left(X_{c'g} \middle| N_{cg}, Z_{cg} = 0\right)$ reflect reverse-

411 transcription/sequencing/mapping errors and rare RNA editing events, because when there is no

412 mutation (i.e. $Z_{cg} = 0$), all mutation reads reflect such technical errors or RNA editing. Let ϵ denote the

413 combined rate of technical error and RNA editing, we have

414 $$P\left(X_{cg} \middle| N_{cg}, Z_{cg} = 0\right) \sim \text{Binomial}(X_{cg} | N_{cg}, \epsilon)$$

415 For cases where there are mutations (i.e. $Z_{cg} = 1$), the distribution of mutated read counts given total

416 read counts is modeled with a Beta Binomial distribution, which is capable of modeling technical

417 dropout and transcriptional bursting, and is supported by previous allele specific expression studies [28,

418 60] .

419 $$P\left(X_{cg} \middle| N_{cg}, Z_{cg} = 1\right) \sim \int_0^1 \text{Binomial}\left(X_{cg} \middle| N_{cg}, Q_{cg} = q\right) dF(q),$$

420 $$q \sim \text{Beta}(\alpha_g, \beta_g)$$

421 where Q_{cg} indicates proportion of mutated alleles expressed in cell c at site g, with Beta distribution as
422 prior. Respectively, α_g and β_g represent gene activation and deactivation rate, which are estimated
423 empirically. See Supplementary Material for detailed formula derivation.

424 *Kernel based clustering and optimal cluster assignment*

425 We cluster the cells using a kernel-based algorithm, such as hierarchical clustering. Given that there are
426 multiple sorting schemes, we leave the user to choose it. For the default-sorting scheme, we
427 recommend "ward.D" [61]. This is because $d_{cc'}$ behaves like a log likelihood ratio , which should follow a
428 χ^2 distribution when the two cells share the same subclone. The "ward.D" method has been shown to
429 work well in Euclidian space,. Empirically, among different hierarchical clustering algorithms on the renal
430 cell carcinoma dataset (Figure S5) "ward.D" based hierarchical clustering performs the best.

431 To determine the number of clusters we use an intra-cluster divergence curve computed from the
432 divergence matrix. Existing software rely on AIC, BIC, or another model selection metric, [39, 62].
433 However, since we only have the "distance" matrix, these traditional methods cannot be applied. Let N_k
434 be the number of cell pairs in cluster C_k and N be the total number of pairs between cells for all
435 clusters. Let K be the number of clusters. The weighted sum of intra-cluster distance W_K is

436
$$W_K = \sum_{k=1}^{K} N_k \sum_{(i,j)\in C_k} \frac{d_{ij}}{N}$$

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

440

441 *Spike-in analysis*

442 In our spike-in analysis, we adopt a dataset from Deng et al, which has allele specific read counts [63].
443 We further added binomial noise to mimic sequencing error. When analyzing DENDRO performance in
444 terms of various number mutation site, number of cells, proportion of cells in each clade and proportion
445 of mutations along each branch, we only take a subset of cells with similar cell types. On the other hand,
446 we utilize a mixture cell population to test the robustness of DENDRO performance with regard to
447 various expression profiles.

448

449 *SNA inference in "bulk" and phylogenetic tree construction*

450 As stated previously, DENDRO further inferred SNA after pooling the reads from all cells within each
451 cluster. Because, with our choice of thresholds, we identify SNAs in single cells with high sensitivity, the
452 "bulk" level SNAs should be a subset of the SNAs in single cells, and mutation allele counts and total
453 allele counts should provide us with enough information for SNA detection using a maximum likelihood
454 framework [64], which accounts for both sequencing error and rare RNA-editing events. Suppose g is
455 the genotype (number of reference allele) at a site and assume m, the ploidy, equals to 2. Then the
456 likelihood is:

457 $$\mathcal{L}(g) = \frac{1}{m^k} \prod_{j=1}^{l} [(m-g)\epsilon + g(1-\epsilon)] \prod_{j=l+1}^{k} [(m-g)(1-\epsilon) + g\epsilon]$$

458 $$g^* = \underset{g}{\operatorname{argmax}} -\mathcal{L}(g)$$

459 where k is number of reads at a site and the first l bases ($l \leq k$) be the same to reference and the rests
460 are same to alternative allele. ϵ is the sequencing error and rare RNA-editing combined rate. Given

461 mutation profiles, DENDRO then constructs a phylogenetic tree with the neighbor-joining method,
462 which can more accurately capture the evolutionary relationship between different subclones [65] than
463 the initial tree given by hierarchical clustering.

464

465 *Differential gene expression, mutation annotation and gene ontology analysis*

466 We used Seurat and scDD to identify differentially expressed genes between tumors and between tumor
467 subclones [43-45]. For each comparison, we used two different methods: MAST implemented by Seurat
468 and scDD. Genes with adjusted p-value < 0.05 counts as significant differentially expressed gene for each
469 method. We further intersect these two sets of differentially expressed genes to increase robustness.
470 Subclonal mutations are annotated by ANNOVAR with default parameters and variants associated with
471 intergenic regions were discarded for downstream analysis [47]. For GO analysis, we applied Gene Set
472 Enrichment Analysis tool [46]. Hallmark gene sets were used as fundamental database with FDR q-value
473 < 0.05 as significant.

474

475 List of abbreviations

476 SNA: Single nucleotide alteration

477 scDNA-seq: Single-cell DNA sequencing

478 scRNA-seq: Single-cell RNA sequencing

479 PDX: Patient-derived xenograft

TPM: Transcripts per kilobase million

Declarations

Acknowledgements

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

Funding

This work was supported by National Institutes of Health (NIH) grant 5R01-HG006137-07 to ZZ and 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.

Authors' contributions

NRZ initiated and envisioned the study. ZZ and NRZ formulated the model. ZZ developed and implemented the algorithm and conducted all analyses. ZZ and NRZ and wrote the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

498 *Consent for publication*

499 Not applicable.

500 *Ethics approval and consent to participate*

501 Not applicable.

502 1. Hanks, S., et al., *Constitutional aneuploidy and cancer predisposition caused by biallelic*
503 *mutations in BUB1B.* Nat Genet, 2004. **36**(11): p. 1159-61.
504 2. Gamazon, E.R. and B.E. Stranger, *The impact of human copy number variation on gene*
505 *expression.* Briefings in Functional Genomics, 2015. **14**(5): p. 352-357.
506 3. Burrell, R.A., et al., *The causes and consequences of genetic heterogeneity in cancer*
507 *evolution.* Nature, 2013. **501**(7467): p. 338-45.
508 4. Jiang, Y., et al., *Assessing intratumor heterogeneity and tracking longitudinal and spatial*
509 *clonal evolutionary history by next-generation sequencing.* Proc Natl Acad Sci U S A,
510 2016. **113**(37): p. E5528-37.
511 5. Deshwar, A.G., et al., *PhyloWGS: reconstructing subclonal composition and evolution*
512 *from whole-genome sequencing of tumors.* Genome Biol, 2015. **16**: p. 35.
513 6. Zare, H., et al., *Inferring clonal composition from multiple sections of a breast cancer.*
514 PLoS Comput Biol, 2014. **10**(7): p. e1003703.
515 7. Carter, S.L., et al., *Absolute quantification of somatic DNA alterations in human cancer.*
516 Nat Biotechnol, 2012. **30**(5): p. 413-21.
517 8. Li, B. and J.Z. Li, *A general framework for analyzing tumor subclonality using SNP array*
518 *and DNA sequencing data.* Genome Biol, 2014. **15**(9): p. 473.
519 9. Oesper, L., A. Mahmoody, and B.J. Raphael, *THetA: inferring intra-tumor heterogeneity*
520 *from high-throughput DNA sequencing data.* Genome Biol, 2013. **14**(7): p. R80.
521 10. Ha, G., et al., *TITAN: inference of copy number architectures in clonal cell populations*
522 *from tumor whole-genome sequence data.* Genome Res, 2014. **24**(11): p. 1881-93.
523 11. Miller, C.A., et al., *SciClone: inferring clonal architecture and tracking the spatial and*
524 *temporal patterns of tumor evolution.* PLoS Comput Biol, 2014. **10**(8): p. e1003665.
525 12. Navin, N.E., *The first five years of single-cell cancer genomics and beyond.* Genome Res,
526 2015. **25**(10): p. 1499-507.
527 13. Navin, N., et al., *Tumour evolution inferred by single-cell sequencing.* Nature, 2011.
528 **472**(7341): p. 90-4.
529 14. Wang, Y., et al., *Clonal evolution in breast cancer revealed by single nucleus genome*
530 *sequencing.* Nature, 2014. **512**(7513): p. 155-60.
531 15. Gao, R., et al., *Punctuated copy number evolution and clonal stasis in triple-negative*
532 *breast cancer.* Nat Genet, 2016. **48**(10): p. 1119-30.
533 16. Gawad, C., W. Koh, and S.R. Quake, *Single-cell genome sequencing: current state of the*
534 *science.* Nat Rev Genet, 2016. **17**(3): p. 175-88.

535 17. Picelli, S., et al., *Smart-seq2 for sensitive full-length transcriptome profiling in single cells.*
536 Nat Methods, 2013. **10**(11): p. 1096-8.
537 18. Klein, A.M., et al., *Droplet barcoding for single-cell transcriptomics applied to embryonic*
538 *stem cells.* Cell, 2015. **161**(5): p. 1187-1201.
539 19. Cloney, R., *Cancer genomics: Single-cell RNA-seq to decipher tumour architecture.* Nat
540 Rev Genet, 2017. **18**(1): p. 2-3.
541 20. Patel, A.P., et al., *Single-cell RNA-seq highlights intratumoral heterogeneity in primary*
542 *glioblastoma.* Science, 2014. **344**(6190): p. 1396-401.
543 21. Chung, W., et al., *Single-cell RNA-seq enables comprehensive tumour and immune cell*
544 *profiling in primary breast cancer.* Nat Commun, 2017. **8**: p. 15081.
545 22. Kim, K.T., et al., *Application of single-cell RNA sequencing in optimizing a combinatorial*
546 *therapeutic strategy in metastatic renal cell carcinoma.* Genome Biol, 2016. **17**: p. 80.
547 23. Tirosh, I., et al., *Dissecting the multicellular ecosystem of metastatic melanoma by*
548 *single-cell RNA-seq.* Science, 2016. **352**(6282): p. 189-96.
549 24. Tirosh, I., et al., *Single-cell RNA-seq supports a developmental hierarchy in human*
550 *oligodendroglioma.* Nature, 2016. **539**(7628): p. 309-313.
551 25. Venteicher, A.S., et al., *Decoupling genetics, lineages, and microenvironment in IDH-*
552 *mutant gliomas by single-cell RNA-seq.* Science, 2017. **355**(6332).
553 26. Li, H., et al., *Reference component analysis of single-cell transcriptomes elucidates*
554 *cellular heterogeneity in human colorectal tumors.* Nat Genet, 2017. **49**(5): p. 708-718.
555 27. Raj, A., et al., *Stochastic mRNA synthesis in mammalian cells.* PLoS Biol, 2006. **4**(10): p.
556 e309.
557 28. Jiang, Y., N.R. Zhang, and M. Li, *SCALE: modeling allele-specific gene expression by single-*
558 *cell RNA sequencing.* Genome Biol, 2017. **18**(1): p. 74.
559 29. Padovan-Merhar, O., et al., *Single mammalian cells compensate for differences in*
560 *cellular volume and DNA copy number through independent global transcriptional*
561 *mechanisms.* Mol Cell, 2015. **58**(2): p. 339-52.
562 30. Zafar, H., et al., *Monovar: single-nucleotide variant detection in single cells.* Nat
563 Methods, 2016. **13**(6): p. 505-7.
564 31. Piskol, R., G. Ramaswami, and J.B. Li, *Reliable identification of genomic variants from*
565 *RNA-seq data.* Am J Hum Genet, 2013. **93**(4): p. 641-51.
566 32. Brennecke, P., et al., *Accounting for technical noise in single-cell RNA-seq experiments.*
567 Nat Methods, 2013. **10**(11): p. 1093-5.
568 33. Pierson, E. and C. Yau, *ZIFA: Dimensionality reduction for zero-inflated single-cell gene*
569 *expression analysis.* Genome Biol, 2015. **16**: p. 241.
570 34. Vallejos, C.A., J.C. Marioni, and S. Richardson, *BASiCS: Bayesian Analysis of Single-Cell*
571 *Sequencing Data.* PLoS Comput Biol, 2015. **11**(6): p. e1004333.
572 35. Ding, B., et al., *Normalization and noise reduction for single cell RNA-seq experiments.*
573 Bioinformatics, 2015. **31**(13): p. 2225-7.
574 36. Qiu, X., et al., *Single-cell mRNA quantification and differential analysis with Census.* Nat
575 Methods, 2017. **14**(3): p. 309-315.
576 37. Dobin, A., et al., *STAR: ultrafast universal RNA-seq aligner.* Bioinformatics, 2013. **29**(1):
577 p. 15-21.

578 38. McKenna, A., et al., *The Genome Analysis Toolkit: a MapReduce framework for analyzing*
579 *next-generation DNA sequencing data.* Genome Res, 2010. **20**(9): p. 1297-303.
580 39. Urrutia, E., et al., *Integrative pipeline for profiling DNA copy number and inferring tumor*
581 *phylogeny.* Bioinformatics, 2018. **34**(12): p. 2126-2128.
582 40. Eirew, P., et al., *Dynamics of genomic clones in breast cancer patient xenografts at*
583 *single-cell resolution.* Nature, 2015. **518**(7539): p. 422-6.
584 41. Gerlinger, M., et al., *Intratumor heterogeneity and branched evolution revealed by*
585 *multiregion sequencing.* N Engl J Med, 2012. **366**(10): p. 883-892.
586 42. Shi, Y.J., et al., *Intratumoral Heterogeneity in Breast Cancer: A Comparison of Primary*
587 *and Metastatic Breast Cancers.* Oncologist, 2017. **22**(4): p. 487-490.
588 43. Finak, G., et al., *MAST: a flexible statistical framework for assessing transcriptional*
589 *changes and characterizing heterogeneity in single-cell RNA sequencing data.* Genome
590 Biol, 2015. **16**: p. 278.
591 44. Korthauer, K.D., et al., *A statistical approach for identifying differential distributions in*
592 *single-cell RNA-seq experiments.* Genome Biol, 2016. **17**(1): p. 222.
593 45. Butler, A., et al., *Integrating single-cell transcriptomic data across different conditions,*
594 *technologies, and species.* Nat Biotechnol, 2018. **36**(5): p. 411-420.
595 46. Subramanian, A., et al., *Gene set enrichment analysis: a knowledge-based approach for*
596 *interpreting genome-wide expression profiles.* Proc Natl Acad Sci U S A, 2005. **102**(43): p.
597 15545-50.
598 47. Wang, K., M. Li, and H. Hakonarson, *ANNOVAR: functional annotation of genetic*
599 *variants from high-throughput sequencing data.* Nucleic Acids Res, 2010. **38**(16): p.
600 e164.
601 48. Straussman, R., et al., *Tumour micro-environment elicits innate resistance to RAF*
602 *inhibitors through HGF secretion.* Nature, 2012. **487**(7408): p. 500-4.
603 49. Michor, F. and V.M. Weaver, *Understanding tissue context influences on intratumour*
604 *heterogeneity.* Nat Cell Biol, 2014. **16**(4): p. 301-2.
605 50. Augsten, M., *Cancer-associated fibroblasts as another polarized cell type of the tumor*
606 *microenvironment.* Front Oncol, 2014. **4**: p. 62.
607 51. Zhang, J.Y., et al., *Critical protein GAPDH and its regulatory mechanisms in cancer cells.*
608 Cancer Biol Med, 2015. **12**(1): p. 10-22.
609 52. Tarrado-Castellarnau, M., et al., *Glyceraldehyde-3-phosphate dehydrogenase is*
610 *overexpressed in colorectal cancer onset.* Translational Medicine Communications, 2017.
611 **2**(1): p. 6.
612 53. Mann, M., V. Cortez, and R.K. Vadlamudi, *Epigenetics of estrogen receptor signaling: role*
613 *in hormonal cancer progression and therapy.* Cancers (Basel), 2011. **3**(3): p. 1691-707.
614 54. Green, K.A. and J.S. Carroll, *Oestrogen-receptor-mediated transcription and the influence*
615 *of co-factors and chromatin state.* Nat Rev Cancer, 2007. **7**(9): p. 713-22.
616 55. Dreijerink, K.M., et al., *Menin links estrogen receptor activation to histone H3K4*
617 *trimethylation.* Cancer Res, 2006. **66**(9): p. 4929-35.
618 56. Kim, H., et al., *Requirement of histone methyltransferase SMYD3 for estrogen receptor-*
619 *mediated transcription.* J Biol Chem, 2009. **284**(30): p. 19867-77.
620 57. Han, L., et al., *The Genomic Landscape and Clinical Relevance of A-to-I RNA Editing in*
621 *Human Cancers.* Cancer Cell, 2015. **28**(4): p. 515-528.

622 58. Mayba, O., et al., *MBASED: allele-specific expression detection in cancer tissues and cell*
623 *lines.* Genome Biol, 2014. **15**(8): p. 405.
624 59. Rhee, J.K., et al., *Allelic imbalance of somatic mutations in cancer genomes and*
625 *transcriptomes.* Sci Rep, 2017. **7**(1): p. 1653.
626 60. Skelly, D.A., et al., *A powerful and flexible statistical framework for testing hypotheses of*
627 *allele-specific gene expression from RNA-seq data.* Genome Res, 2011. **21**(10): p. 1728-
628 37.
629 61. Ward, J.H., *Hierarchical Grouping to Optimize an Objective Function.* Journal of the
630 American Statistical Association, 1963. **58**(301): p. 236-&.
631 62. Goutte, C., et al., *Feature-space clustering for fMRI meta-analysis.* Hum Brain Mapp,
632 2001. **13**(3): p. 165-83.
633 63. Deng, Q., et al., *Single-cell RNA-seq reveals dynamic, random monoallelic gene*
634 *expression in mammalian cells.* Science, 2014. **343**(6167): p. 193-6.
635 64. Li, B., et al., *A likelihood-based framework for variant calling and de novo mutation*
636 *detection in families.* PLoS Genet, 2012. **8**(10): p. e1002944.
637 65. Schliep, K.P., *phangorn: phylogenetic analysis in R.* Bioinformatics, 2011. **27**(4): p. 592-3.
638

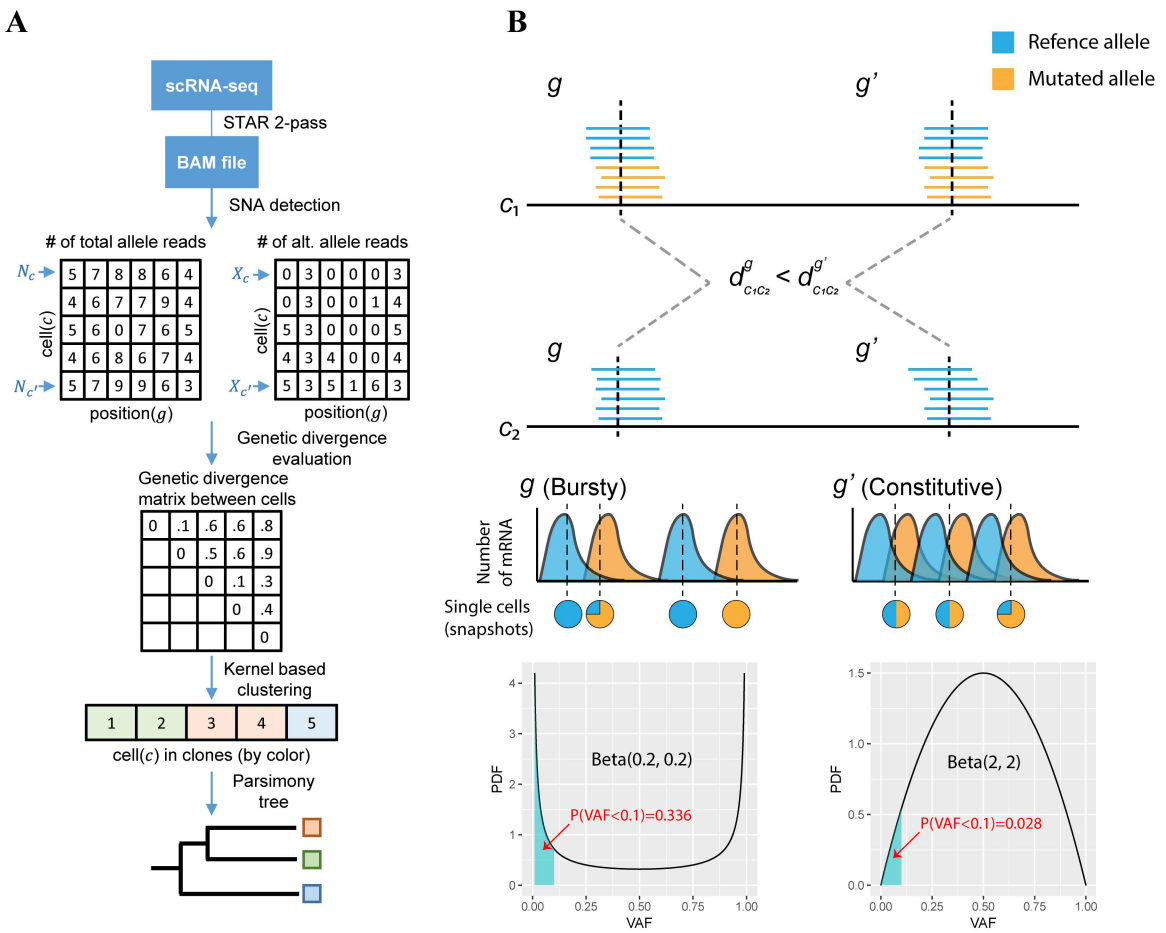

Figure 1. DENDRO analysis pipeline and genetic divergence evaluation. a DENDRO analysis pipeline overview. **b** Statistical model for genetic divergence evaluation function. In this example, we show relationship of cell c_1 and c_2 with gene g and g' in terms of their genetic divergence. Middle panels shows the stochasticity of cell expression and bottom panels illustrates the distribution of variants allele frequency across cells for these two genes, indicating that gene g is a bursty gene and g' is a constitutive gene.

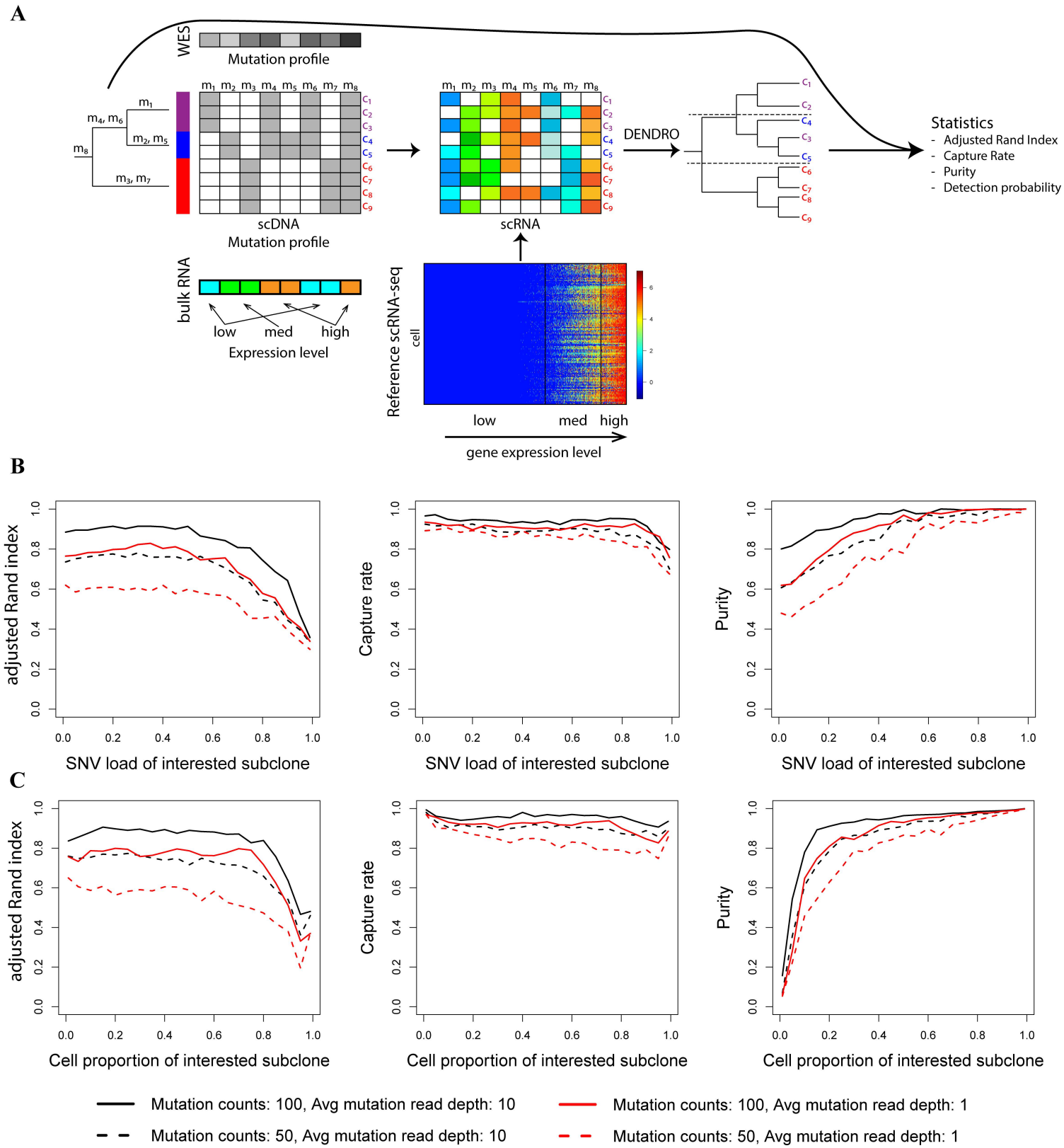

Figure 2. DENDRO accuracy assessment by spike-in analysis. a The overall spike-in analysis pipeline. Three statistics measure the accuracy of DENDRO - adjusted Rand index (global accuracy, 1 indicating perfect classification), capture rate (subclone specific accuracy, out of all cells from the clade, how many of them are correctly assigned) and purity (subclone specific accuracy, out of all cells from the detected cluster, how many are actually from the clade). **b** Statistics under different SNV load of a given clade of interest (i.e. as fraction of total mutation counts). **c** Statistics under different cell proportion of a given clade of interest. (Mutation counts: number of mutation identified; Avg mutation read depth: average read depth of all the mutation sites.)

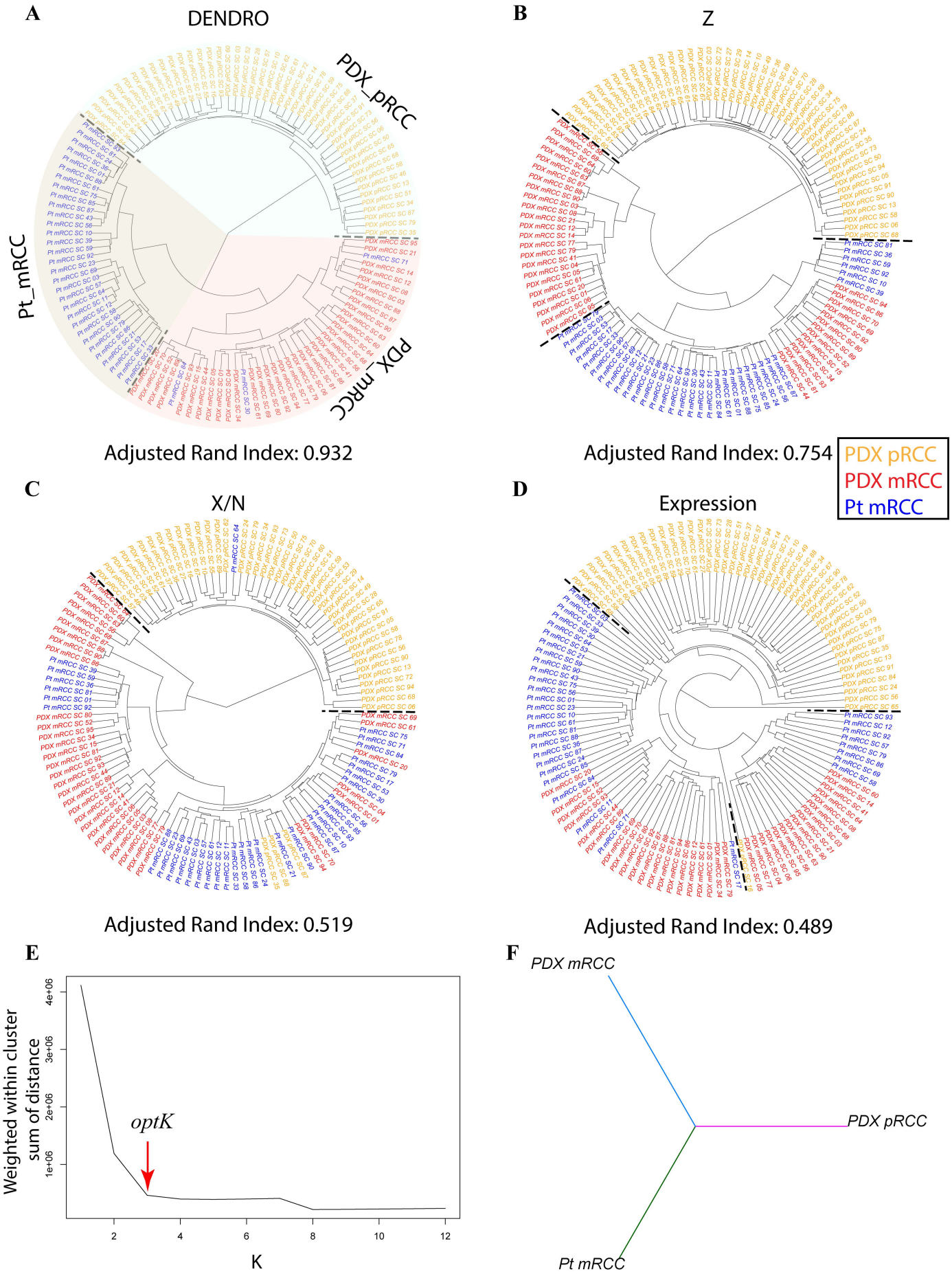

Figure 3. Evaluation of DENDRO on a renal cell carcinoma and its metastasis. a DENDRO clustering result from primary and metastatic renal cell carcinoma dataset. Background colors represent DEDRO clustering result. **b** Clustering of the same dataset using *Z* matrix (indicator matrix, $z_{ij}=1$ when detected a mutation for cell *i* at site *j* by GATK tool). **c** Clustering of the same dataset using *X/N* matrix (mutation allele frequency matrix) **d** Clustering of the same dataset using expression (*log*TPM). **e** Intra-cluster divergence curve to select optimal number of cluster by DENDRO. Here, *optK=3* **f** Phylogenetic tree of three cell populations identified by DEDNRO. (In **a** - **d**, dashlines indicate cluster boundaries)

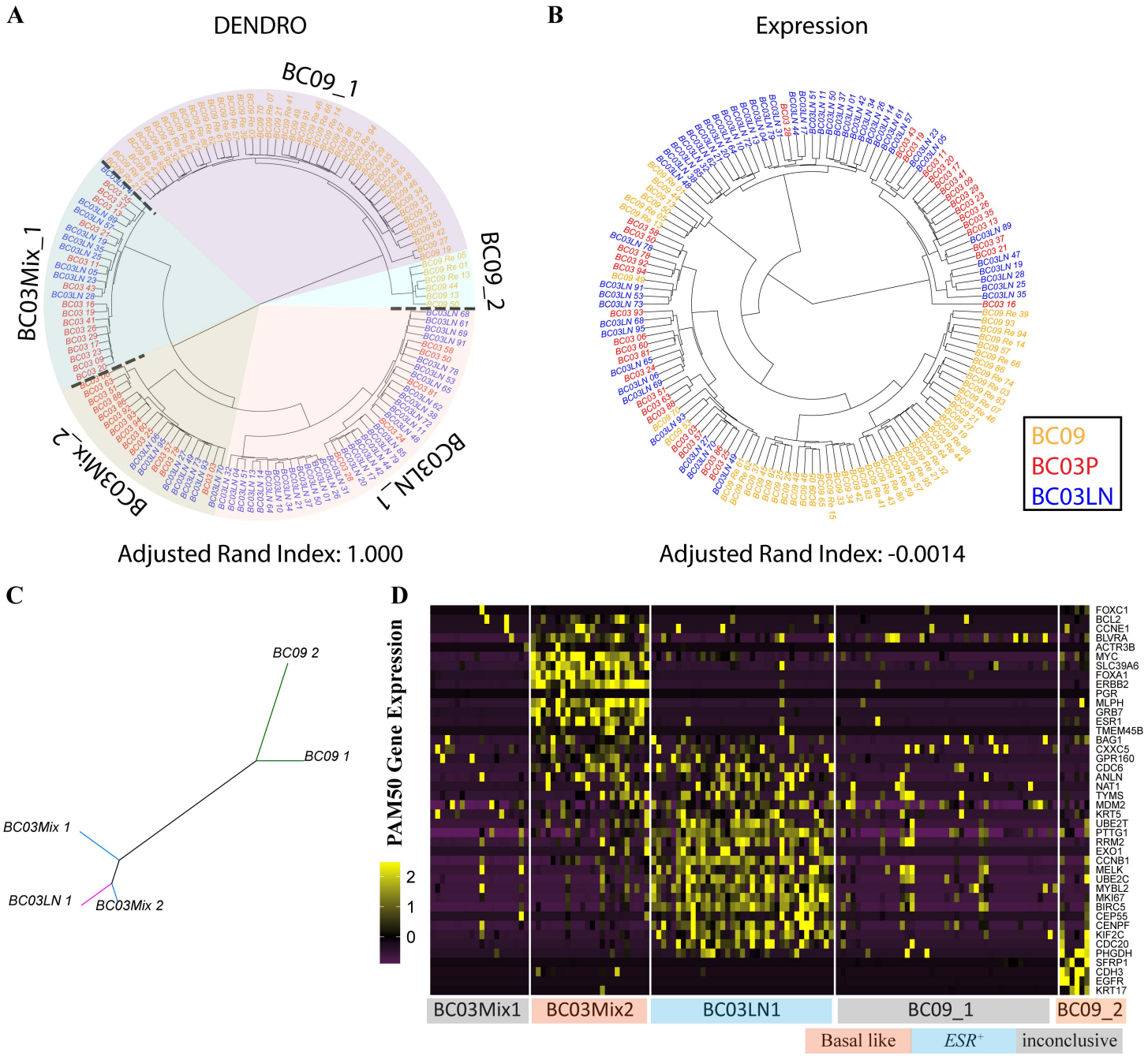

Figure 4. Analysis of scRNA-seq dataset of primary breast cancer. a DENDRO cluster result for primary breast cancer dataset (Chung et al., 2017). **b** Hierarchical clustering result for the same dataset based on expression (*log*TPM). (dashlines indicate cluster boundaries). **c** Neighbor joining phylogenetic tree given detected subclones for breast cancer dataset. **d** PAM50 gene panel expression shows breast cancer subtypes of each subclone.