Leveraging single cell RNA sequencing experiments to model intra-tumor heterogeneity

Meghan C. Ferrall-Fairbanks (1), Markus Ball (2,3),
Eric Padron (2), and Philipp M. Altrock* (1)

1: Department of Integrated Mathematical Oncology, Moffitt Cancer Center and Research Institute

2: Department of Malignant Hematology, Moffitt Cancer Center and Research Institute, Tampa, FL 33612, USA

3: present address: Institute of Pathology, University Hospital Cologne, 50937 Köln, Germany

*Corresponding author: Philipp M. Altrock, philipp.altrock@moffitt.org
Moffitt Cancer Center, 12902 USF Magnolia Drive, SRB 24007, Tampa, FL 33612, USA
Phone: +1 813 745-5897 Fax: +1 813 745- 8357

Running Title: General intra-tumor heterogeneity can be calculated from scRNA-seq data.

Key Words: Acute Myeloid Leukemia. Tumor Heterogeneity. Mathematical Oncology. Diversity. Single Cell RNA sequencing.

The authors declare no competing financial interests.

ABSTRACT

PURPOSE: Many cancers can be treated with targeted. Almost inevitably, tumors develop resistance to targeted therapy, either from preexistence or by evolving new genotypes and traits. Intra-tumor heterogeneity serves as a reservoir for resistance, which often occurs due to selection of minor cellular sub-clones. On the level of gene expression, the 'clonal' heterogeneity can only be revealed by high-dimensional single cell methods. We propose to use a general diversity index (GDI) to quantify heterogeneity on multiple scales and relate it to disease evolution.

METHODS: We focused on individual patient samples probed with single cell RNA sequencing to describe heterogeneity. We developed a pipeline to analyze single cell data, via sample normalization, clustering and mathematical interpretation using a generalized diversity measure, and exemplify the utility of this platform using single cell data.

RESULTS: We focused on three sources of RNA sequencing data: two healthy bone marrow (BM) samples, two acute myeloid leukemia (AML) patients, each sampled before and after BM transplant (BMT), four samples of pre-sorted lineages, and six lung carcinoma patients with multi-region sampling. While healthy/normal samples scored low in diversity overall, GDI further quantified in which respect these samples differed. While a widely used Shannon diversity index sometimes reveals less differences, GDI exhibits differences in the number of potential key drivers or clonal richness. Comparing pre and post BMT AML samples did not reveal differences in heterogeneity, although they can be very different biologically.

CONCLUSION: GDI can quantify cellular heterogeneity changes across a wide spectrum, even when standard measures, such as the Shannon index, do not. Our approach offers wide applications to quantify heterogeneity across samples and conditions.

INTRODUCTION

In many cancers, there still exists a critical need to understand the mechanisms of therapy resistance evolution. For example, Acute Myeloid Leukemia (AML) is an aggressive hematologic malignancy that is hallmarked by proliferation of immature myeloid cells in the bone marrow and life-threatening ineffective hematopoiesis[1]. AML is the most common adult leukemia, with an incidence of about 20,000 cases yearly and a 5-year survival of only 26%[2,3]. The diagnosis of AML requires greater than 20% of myeloid immature cells (myeloblasts) in the peripheral blood or bone marrow. The median survival of untreated AML is measured in weeks[4]. Several AML targeted therapies have been recently approved, e.g. midostaurin for FLT3 mutated patients and enasidenib for those with mutations in IDH2[5,6]. These mutations occur at rates of 25% (FLT3) and 5% (IDH2) of all AML patients and their targeted therapies are generally well tolerated relative to chemotherapeutic counterparts[7]. However, midostaurin (and even more potent FLT3 inhibitors in clinical trial[8]) does not fully eradicated the disease, leading to refractory or relapsing AML in most patients[9]. The complete response rate for enasidenib in relapse/refractory IDH2 mutated AML is less than 20%. Further refinements in patient selection are required to realize mutationally-directed therapy[5]. Little is known regarding the emerging resistance mechanism and whether targeted therapies (single or combination) against AML alone can ever be successful.

The conventional dogma postulates that therapeutic resistance occurs via the acquisition of mutations that result in clonal evolution. Emerging data suggests that these mutations are either subclonally present or present at frequencies detectable using digital PCR or ultra-deep sequencing technologies at diagnosis or prior to progression. Very low level somatic mutations are also detected in pre-leukemic states[10-13]. Somatic mutations are often present years before

69 the diagnosis of therapy related myeloid neoplasms[14,15]. Interestingly, these mutations are
70 commonly associated with disease progression and transformation[16]. The presence of such low-
71 frequency genetic markers suggests that high levels of intra-tumor heterogeneity (ITH) persist
72 over long periods of time, and that pre-existing ITH is a primary driver of future therapy
73 resistance, while variation in transcription over time shapes the disease phenotype. A clinically
74 relevant summary metric to describe ITH on the transcriptional level has not been developed.

75 scRNA-sequencing technologies can present a cost-effective method to identify
76 transcriptomic heterogeneity and directly measure ITH. Proof of concept studies have been
77 performed in AML, using DROP-seq that yields potentially cost-effective single cell annotations
78 of thousands of transcripts per cell. In triple negative breast cancer, intercellular heterogeneity
79 of gene expression programs within tumors is variable and correlates with genomic clonality[17].
80 A study in Chronic Myeloid Leukemia (CML) demonstrated that scRNA-seq was capable of
81 segregating patients with discordant responses to targeted tyrosine kinase inhibitor therapy[18].
82 These data provide rationale to explore ITH in scRNA-seq data, and to determine whether
83 defined measures of ITH can be predictive of progression, and eventually leveraged to mitigate
84 progression and relapse.

85 Our goal is to quantify ITH in cancer such that it has maximal predictive value, in particular
86 in hematologic malignancies. To this end we here present a platform that uses a generalized
87 diversity index that characterizes cell population heterogeneity across a spectrum of scales
88 (orders of diversity)[19]. These scales range from clonal richness (low order of diversity reveals
89 the number of distinct subpopulations), to more classical measures such as Shannon or
90 Simpson indices (intermediate order of diversity), to the number of most abundant cell types that

91 possibly act as the key drivers of heterogeneity before transformation or perturbation by therapy
92 (high order of diversity).
93
94 **MATERIALS AND METHODS**
95 We created a computational and modeling approach to develop a robust statistical picture
96 of the persistent and emerging variability in scRNA-seq data, chiefly based on drop-seq
97 technologies; the 10x Genomics platform offered a variety of datasets that were linked to disease
98 and treatment dynamics[20]. We specifically used the datasets of two healthy/control bone marrow
99 mononuclear cell samples (BMMCs), two individuals with AML BMMCs sampled pre and post
100 bone marrow transplant (BMT), to develop and test our ITH pipeline. pipelines.
101 First, we ran publicly available FASTQ-format files (a typical output from a drop -seq
102 experiment) through the cellranger count pipeline and then through the cellranger aggr pipeline,
103 to pool the samples together for comparisons during cluster analysis, interrogated through the
104 10x Genomics Loupe™ Cell Browser (Fig. S1). To test the robustness and valid our diversity
105 metrics and the ITH pipeline, we extended our analysis to include additional publicly available
106 datasets for other hematopoietic cell types (CD34+, CD14+, CD19+, and CD4+)[20], as well as six
107 patient normal-tumor matched lung cancer samples[21], for which we used the same approaches
108 and pipelines. To calculate summary metrics (outlined in **Fig. 1**), first the transcript expression
109 data was clustered into groups of cells with similar transcript expressions (cellranger aggr). Next,
110 we quantified the distance between each of the clusters to determine if clusters separated based
111 on healthy or disease status (healthy vs AML). A Euclidean distance was calculated between
112 the mean expression values for each gene, of each cluster, to establish a distance metric (**Fig.
113 2**).

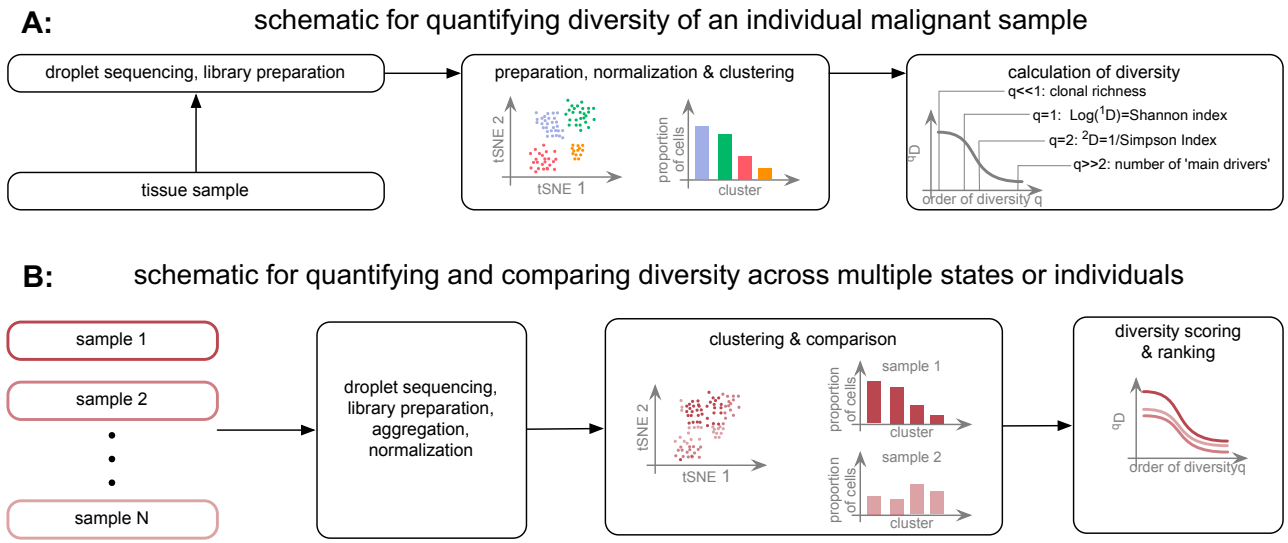


Figure 1. Schematic of our single cell RNA sequencing-based approach. A: Workflow of calculating a generalized diversity index (GDI) for a single sample. After sequencing and library preparation, normalization to reduce the number of false negatives or false positives is applied, for example using 10x-genomics platform, or. Clustering can then be applied (Loupe^{TM} Cell browser, or other platforms, see Supplement), from which diversity can be calculated. **B**: A similar approach can be used when multiple samples are compared. Data normalization and clustering now have to be implemented considering all samples (see Supplement), from which diversity scoring can inform a ranking of intratumor-heterogeneity across samples. Single dots in the tSNE plots represent single cells, which might either be marked according to their cluster classification or according to their sample of origin.

Second, we sought to characterize across-sample differences by calculating the Kolmogorov-Smirnov (KS) distance[22] of the cell count distributions in each cluster, in order to compare samples or pooled samples of the same condition (e.g. disease v. healthy) in terms of the cellular distribution over the identified clusters (**Fig. 3 A-G**).

Third, we calculated an ecological diversity index[23] using the cellular frequencies over clusters, across a range of order of diversity (**Fig. 3H, I**). To assess the robustness of our diversity metric, we performed down sampling of the original datasets and found the relative change in diversity index, across a range of order of diversity, to determine the sensitivity of our diversity metric (**Fig. 4**).

Last, we applied our ITH pipeline and diversity metric to two additional datasets, (1) a hematopoietic cell type datasets comparing CD34+ cells with CD4+, CD14+, and CD19+ cell populations[20], and (2) a lung cancer dataset with tumor-normal matched tissue sites taken from

135 six different lung cancer patients[21] (**Fig. 5**, Supplement). Further specific details of our methods,
136 such as cells per sample, are described in the Supplement and is available online (including all
137 code used to generate our results): https://github.com/mcfefa/scRNAseq.
138

139 **RESULTS AND DISCUSSION**

140 We established proof of concept that we can generate clinically relevant summary metrics
141 of ITH by analyzing publicly available scRNA-seq data[20,21]. Within BMMC samples from
142 diagnosed AML and healthy control groups, we sought to establish how to summarize both inter-
143 and intra-sample ITH. First, we clustered transcript expression of two healthy individuals, two
144 AML patients, each sampled twice—once before and once after allogenic bone marrow
145 transplant.

146 As a verification we sought to distinguish between healthy and AML samples based on
147 the mean expression values across cells, across clusters (**Fig. 2A)**. With the 23 clusters
148 identified (**Fig. 2B, C**), a network of clusters emerged, displayed as an undirected graph where
149 the distance between mean UMI counts determines the thickness and length of the edges (**Fig.**
150 **2D**). The size of the node was chosen to indicate of the total number of cells in that cluster. We
151 colored each node according to the condition (health, pre BMT AML, post BMT AML) that was
152 in the majority (breakdown of actual proportions per cluster shown in **Fig. 2E**). This showed that
153 indeed the large clusters with mostly healthy cells are most similar in average gene expression,
154 while the large clusters with mostly AML cells cluster separately (to the right). The post BMT
155 cells clustered more closely to the healthy than pre BMT AML samples. This supports the idea
156 that these patients were potentially still transitioning but closer to a healthy phenotype. However,
157 some AML-dominant clusters still grouped near the healthy/post BMT super cluster. Based on

this "bulk" measure alone, one may not be able to easily distinguish between healthy and diseased cells. Therefore, other quantifications and metrics to describe the gene expression differences may better discriminate clinical and stages.

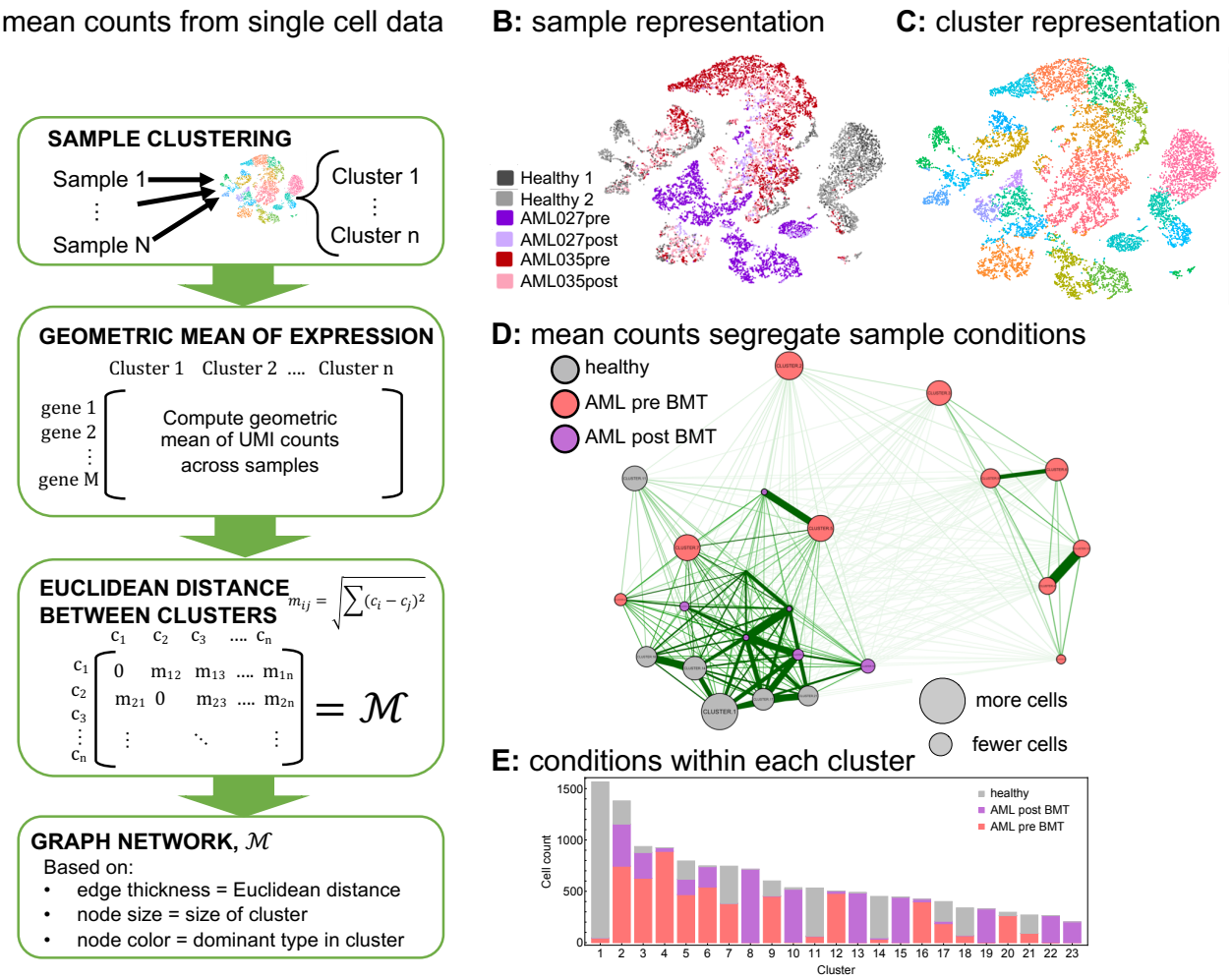


Figure 2. Mean cellular gene expression across clusters within patients can separate disease conditions to some degree. Here we built a network based on mean differences in overall expression. **A**: The geometric mean of UMI counts across samples and genes was calculated for each cluster. Then a Euclidean distance was calculated between clusters. Here we used publicly available scRNA-seq data[20]: two healthy donor BMMCs, two AML patient BMMCs pre bone marrow transplant (preBMT), and post bone marrow transplant (postBMT). These six samples were then clustered (using 10x-genomics Loupe^{TM} Cell browser; for alternative clustering methods see Supplement), for which we show the sample-based (**B**) and cluster-based (**C**) tSNE-plots out of Loupe^{TM} browser. Each dot represents a single cell, which is colored either according to sample of origin or its assigned cluster. The cluster-based differences in mean gene expression over unique molecular identifier (UMI) counts the gave rise to a "clustering of the clusters" (**D**). The nodes in the resulting graph were colored based on the dominant cell type from each condition present in each cluster; gray for healthy, red for AML pre BMT and purple for AML post BMT, the distance between nodes was chosen inversely proportional to the difference in mean gene expression level, the individual distributions of cells from a specific condition in each cluster are shown in **E**.

174 To determine metrics better at discriminating between healthy and disease AML, we
175 analyzed and summarized inter- and intra-heterogeneity in two different ways. First, we
176 considered each samples' grouping of cells into clusters of similar gene expression. To this end,
177 we used the Kolmogorov-Smirnov (KS) distance, which compares two discrete probability mass
178 functions (the fraction of cells per cluster, **Fig. 3A**). We identified rather similar distributions within
179 the same condition, and rather different distributions between conditions, with post BMT being
180 a notable exception (**Fig. 3B-D**). The KS distance between the two healthy samples was 0.139,
181 between the two pre BMT AML samples it was 0.174, but between the two post BMT samples it
182 was 0.551. Since we had clustered all six samples together, we could also compare them pooled
183 by condition, which revealed that indeed conditions distribute differently across the identified
184 cellular subpopulations in high-dimensional gene expression space (**Fig. 3E-G**).

185 Second, we calculated a general diversity index (GDI) for each condition (healthy, pre
186 BMT AML, post BMT). The mathematical definition of GDI, ^{q}D, is shown in **Fig. 3H**. We
187 established segregation of the different clinical conditions according to this ecology-based
188 diversity index[23]. The pre BMT AML samples had consistently higher diversity index compared
189 to the healthy sample and this held true across the entire order of diversity range, q (**Fig. 3I**).
190 Interestingly, on this level the post BMT samples also scored unanimously higher in GDI. This
191 could indicate that post BMT settings may require a certain amount of time after transplant to
192 evolve toward a healthy spectrum of intra-leukemic diversity. Also, in a comparison of the
193 individual samples within each condition (**Fig. 4, A-C**), the post BMT samples were most different
194 from each other.

195 To interrogate the robustness of GDI further and to establish confidence in the metric, we
196 down sampled the dataset and then re-clustered and calculated the ^{q}D spectrum (**Fig. 4D-I**).

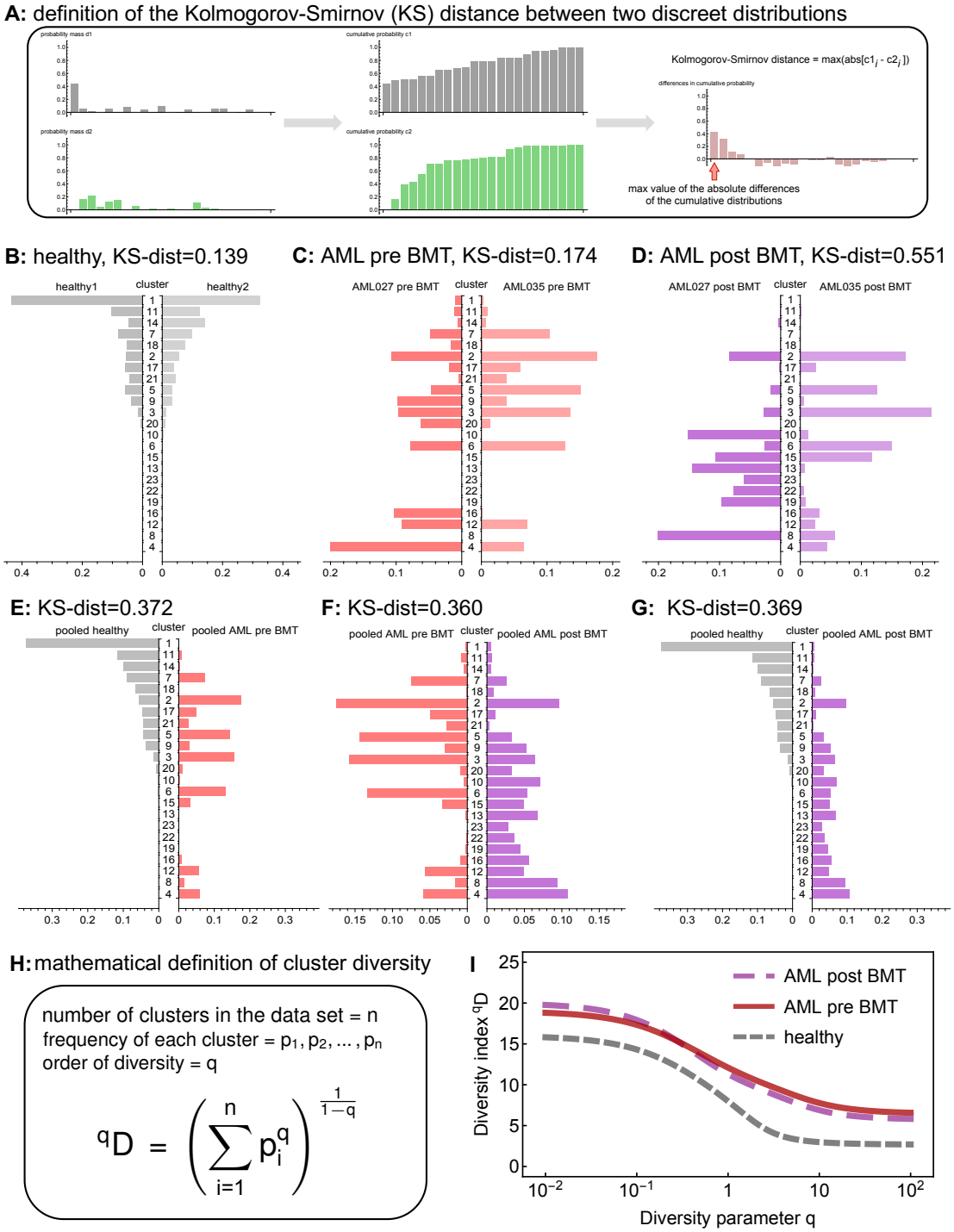


Figure 3. Cluster-based diversity scoring reveals strong differences between healthy individuals and cancer patients. In our analysis, using data from Zheng, et. al.[20], we evaluated our ability to score significant differences in "cluster diversity" across healthy and AML samples pre and post bone marrow transplant (BMT). As a first indicator of between-sample or between-condition differences we used the Kolmogorov-Smirnov (KS) distance for discrete probability mass functions (**A**). Little difference was found in the KS distance within condition-differences (**B**, **C**), except in post BMT samples (**D**). Between condition-differences were larger when comparing pooled samples across conditions (**E**, **F**, **G**). We calculated a general diversity index (GDI) qD (**H**) to quantify diversity across "orders of diversity" q. For all orders of diversity measure, AML patients (pre and post BMT) had a higher diversity index compared to healthy individuals (**I**, 2 samples per condition), suggesting GDI can be used as a metric for stratification.

207 During down sampling, we analyzed each sample individual by randomly removing 50% of the
208 cells, then calculating the number of clusters identified for that individual's transcript expression,
209 and finally calculating the diversity index for specific q values of interest, including $q=10^{-2}$, $q=10^{-1}$,
210 q=1 (which relates to the Shannon index), q=2 (which defines the inverse of the Simpson
211 index), q=10, and $q=10^{2}$. The distributions shown were obtained from 1000 runs of independent
212 down sampling. Intriguingly, these distributions showed that with removing half of the cells, the
213 diversity scores did not change more than a unit or two in either direction. When comparing to
214 the diversity spectrum shown in **Fig. 3I**, this suggests that if healthy diversity spectrum were
215 shifted up by two units (10% of the maximum) and the AML samples diversity spectrum were
216 shifted down by two units, there would still be visible separation between the healthy and AML
217 conditions.

218 To further validate our metric, we implemented on our approach with two other datasets.
219 One dataset described different hematopoietic cellular subtypes, CD34+, CD4+, CD14+, and
220 CD19+[20]. CD34+ is a hematopoietic progenitor cell marker and represent a polyclonal population
221 that includes many different subtypes (hematopoietic stem cells, multipotent progenitor cells,
222 common myeloid progenitor cells, common lymphoid progenitor cells, megakaryocytes erythroid
223 progenitor cells and granulocytes macrophage progenitor cells) all of which express CD34[24-26].
224 The CD34+ polyclonal population contrasted the CD4+, CD14+, and CD19+ populations, which
225 represent more homogenous cellular populations (helper T-cells, monocytes, and B-cells,
226 respectively). This clonality pattern was recovered by GDI (**Fig. 5A**), where the CD34+
227 population had a considerable higher diversity score across the spectrum. Interestingly, lower
228 values of q seem to separate differentiated cells or robustly.
229

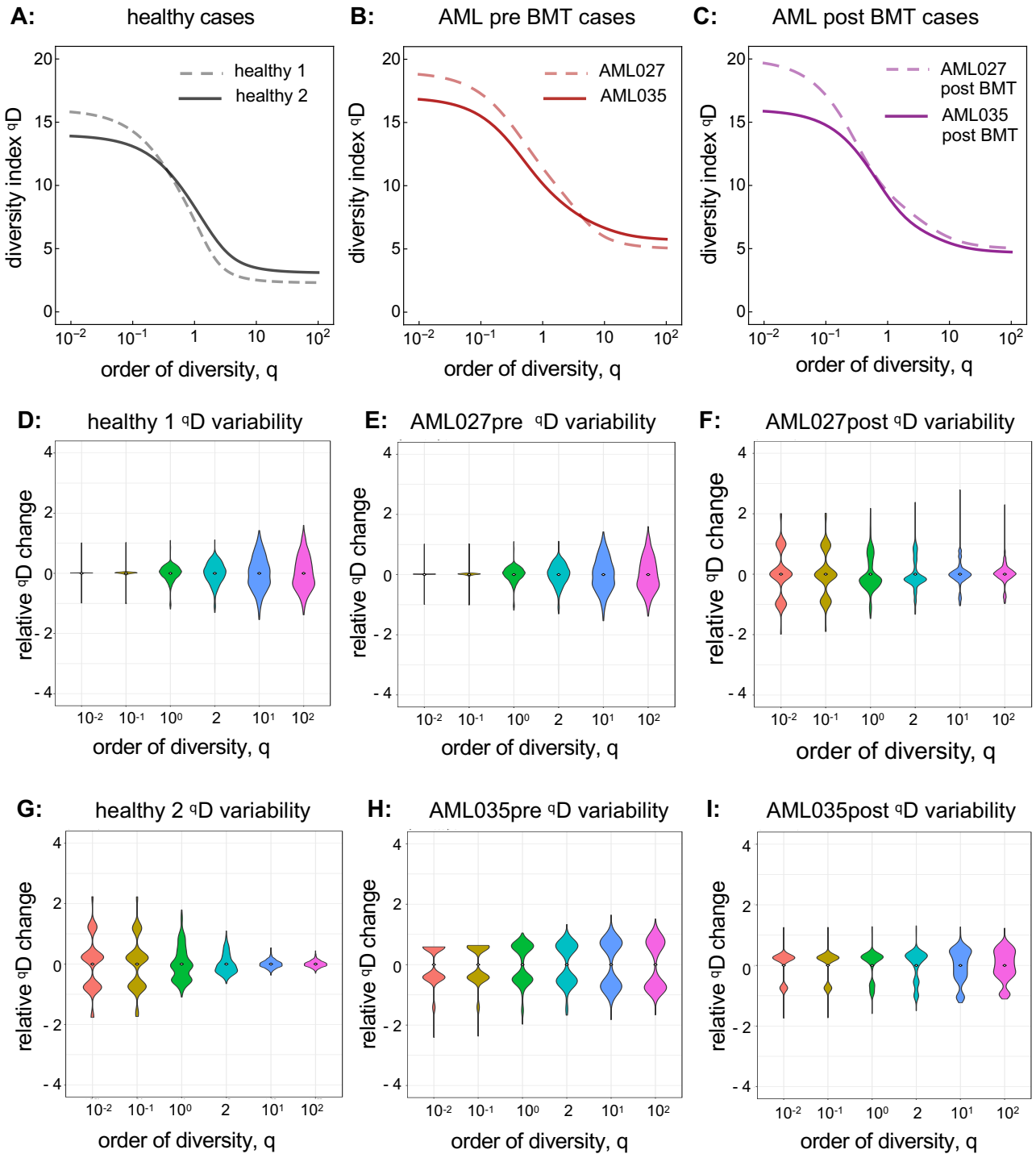
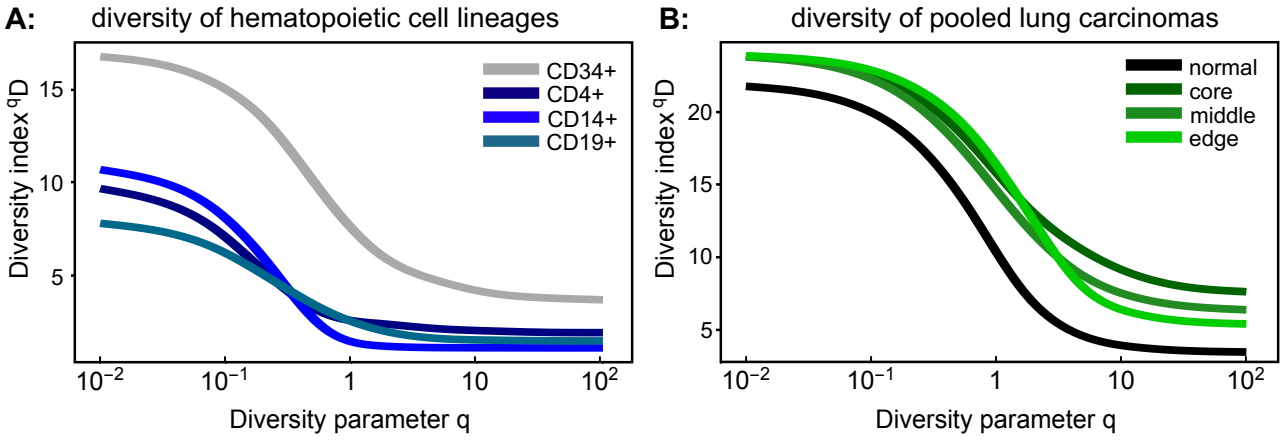


Figure 4. AML patients have consistently higher diversity compared to healthy individuals. Individual diversity spectrums were reported for healthy (**A**), AML pre BMT (**B**), and AML post BMT (**C**) samples (each line is one from one sample). Cell-gene matrices were down sampled to 50% of the cells 1000 times, and qD scores were calculated (using the full pipeline, see Supplement) for specific values of q=0.01, 0.1, 1, 2, 10, 100. The distributions of relative qD changes for healthy (**D**, **G**), AML (**E**, **H**), and postBMT AML (**F**, **I**) samples showed that generally, lower q values lead to less change in measured diversity. Across all cases the diversity score did not change by more than two units (relative change is measured by dividing the entire distribution by the distribution mean). For sample sizes, see Supplement. BMT: Bone marrow transplant.

238 Finally, we quantified ITH using lung cancer scRNA-seq samples[21]. We analyzed six
239 different patients, each with up to three different tumor sites (core, middle, and edge) and a
240 patient-matched adjacent normal lung tissue sample. Using our GDI metric, we see that the
241 diversity spectrum of the normal lung tissue was much lower than any of the tumor site diversity
242 scores (pooled conditions, **Fig. 5B**). Interestingly here, more clear separation of conditions was
243 achieved at high orders of diversity q, indicating differences in the number of driver clones at
244 different sites within the tumors. These additional datasets further support the ability of quantified
245 diversity metric to discriminated between healthy and diseases states, which can be applied in
246 a clinical setting.



247

248 **Figure 5. Higher diversity indicates higher clonality in in normal tissues and solid tumors.** (**A**) Additional available data
249 from Zheng, et. al.[20] for CD34+ cells, CD4+ helper T-cells, CD14+ monocytes and CD19+ B-cells were run through pipeline S
250 (see Supplement), and the continuum of diversity was calculated for each population. The naturally polyclonal population
251 (CD34+) shows the highest diversity score. Each of the other differentiated immune cell compartments are more homogeneous
252 across orders of diversity. (**B**) In solid tumors, location matters: normal-tumor matched lung carcinoma samples were obtained
253 from publicly available data for six lung cancer patients[21] (individual patients: see Supplement). The diversity metric across q
254 shows an increase in diversity within tumors across different tumor locations.

255

256 **SUMMARY AND CONCLUSIONS**

257 Single-cell RNA sequencing efforts have greatly helped to uncover population structures
258 and mapping to specific cellular population patterns[27]. Although these methods can also
259 elucidate tumorigenesis[28], immune-profiles[29], and detect and track genomic profiles of

260 clones[30,31], the overall utility of scRNA-seq for cancer progression survival metrics has been
261 elusive[32]. Here, we demonstrated the potential utility of two scRNA-seq based scores of cellular
262 heterogeneities, using a generalized diversity index that may be elevated in disease.
263 Remarkably, using previously published data without further processing, this quantification of
264 intra-tumor heterogeneity was able to accurately distinguish AML from healthy individuals, as
265 well as from post-transplant conditions. These data suggest that ITH can be estimated by
266 diversity-based summary statistics, and that these summary statistics can be leveraged to
267 predict clinical outcomes.

268 Our work here aimed at optimizing and identifying a clinically relevant summary index for
269 ITH in the context of AML, for which targeted single-cell genome sequencing was also able to
270 sensitively uncovered complex clonal evolution[33]. We anticipate that our ILH metric will be
271 prognostic for leukemia-free survival (LFS) and potentially overall survival (OS), even after
272 correction for known clinical prognostic variables. We have also shown how this metric can also
273 be used to effectively describe heterogeneity in other malignancies, including solid tumors such
274 as lung carcinomas.

275 From a clinical perspective in terms of tumor heterogeneity and emergence of resistance
276 clones during targeted therapy[34,35], we expect our metric can discriminate patients clinically. We
277 hypothesize that these heterogeneity metrics would be elevated *independently* (at least *a priori*)
278 in potentially highly resistant patients. The advantage of a more general metric used here is that
279 it allows us to look across many orders of diversity, and potentially pick a desired range of
280 heterogeneity quantification. For example, one might be interested especially in lower values of
281 q, where a higher diversity score may indicate an individual (sample) more at risk of resistance
282 evolution, as it shows high standing variation. On the other hand, differences at high values of q

283 point to key differences in the number of important driver clones, which might uncover distinct
284 vulnerabilities that can be targeted in combination or adaptively.

285 Diversity measures have long received attention in ecology and evolution[19,36]. We here
286 approached to measure diversity, and thus tumor heterogeneity, using a general definition of
287 non-spatial diversity[37], in form of the quantity ^{q}D (**Fig. 3H-I**). This approach considers all possible
288 orders of diversity q, but also allows to compare disease stages according to a specific diversity
289 index (fixed choice of q), which emerge as special cases of ^{q}D. The species (clonal) richness of
290 a sample is given by q=0. The Shannon index (log scale) can be found when q approaches 1.
291 The Simpson index, which approximated the probability that any two cells are identical emerges
292 from the case q=2. Both Shannon and Simpson index have been used in mathematical and
293 statistical models of cancer evolutionary dynamics to quantify tumor heterogeneity as it
294 potentially changes during tumor growth, with disease progression, or during treatment[38-40].
295 Shannon entropy-based statistics have also been used to quantify single cell heterogeneity, to
296 deliver insights into emerging or disappearing clones during transitions between clinical
297 conditions[41].

298 Single cell RNA sequencing experiments give a snapshot of the cell population state on
299 the level of gene expression, and it can characterize how individual cell's transcriptomes
300 compare to the bulk. In contrast to mass cytometry, drop-seq is fast and extremely high-
301 throughput. Other single cell technologies, like flow cytometry can be used to generate single
302 cell data for a relatively small subset of potential markers that distinguish between normal and
303 disease, and it requires that the researcher/clinician knows what these markers are in advance.
304 Among a variety of outcomes that may be distinguished by our metric, one can be the study for

305 segregating samples based on disease severity, for which additional follow-up knowledge will
306 be needed.

307 Further extending our results to the potential impact in a clinical setting in leukemias,
308 intraleukemic heterogeneity (ILH) is a known reservoir for tumor resistance and clinical
309 refractoriness to targeted therapies[9]. For current targeted therapies to treat AML, clinical
310 responses have been modest. Specific means to change ILH can be of particular appeal in such
311 cases, as they might help render tumors less aggressive and hinder their ability to rapidly evolve
312 resistance. In particular, hypomethylating agents (HMA, cytosine analogs that irreversibly bind
313 to DNA methyltransferase, an enzyme required for methylation of CpG-rich DNA) have the
314 potential to diminish ILH[42]. Transcriptome changes upon treatment with HMA therapy have not
315 been analyzed at single cell resolution. Our analyses here provide a quantitative basis to
316 understand and reliably track these changes.

317 Clear separation of diversity metrics by condition, as we show it, might not be expected
318 in general. A weakness of our approach is that it does not consider any meaning of the
319 associated phenotypes or genotypes. Therefore, as it stands, our method cannot be transferred
320 to improve the predictive power of existing bulk signatures. Hence, existing survival data is
321 unlikely to be useful to prove that GDI is predictive of survival, and novel databased that uniquely
322 connect high throughput single cell experiments with clinical outcomes are needed.

323 Once the appropriate cohorts are established, however, changes in an individual's
324 diversity score could indicate unique features of disease progression. In the context of adaptive
325 therapy[43,44], which aims at tumor burden control rather than difficult tumor eradication, it might
326 be critical to identify the appropriate scale of diversity that best predicts outcomes. One could
327 speculate that there is an optimal window of diversity that should be maintained—very low

diversity could indicate fast disease progression and very high diversity could mean that the tumor could adapt to the treatment schedule too quickly. The concept we introduced here is sufficiently flexible in its ability to quantify optimally predictive windows of diversity that should be maintained during adaptive therapy.

AUTHORSHIP CONTRIBUTIONS

P.M.A., E.P. conceived the study. M.C.F., M.B. and P.M.A. performed computational experiments and statistical analyses. M.C.F., and P.M.A. performed mathematical modeling. M.C.F., M.B., E.P. and P.M.A. analyzed the data. M.C.F., M.B., E.P. and P.M.A. wrote the manuscript. P.M.A. and E.P. supervised the project.

341 **REFERENCES**

342

343 1. Perl AE: The role of targeted therapy in the management of patients with AML.
344 Hematology Am Soc Hematol Educ Program 2017:54-65, 2017

345 2. Percival ME, Tao L, Medeiros BC, et al: Improvements in the early death rate
346 among 9380 patients with acute myeloid leukemia after initial therapy: A SEER database
347 analysis. Cancer 121:2004-12, 2015

348 3. Medeiros BC, Satram-Hoang S, Hurst D, et al: Big data analysis of treatment
349 patterns and outcomes among elderly acute myeloid leukemia patients in the United States.
350 Annals of Hematology 94:1127-1138, 2015

351 4. Arber DA, Orazi A, Hasserjian R, et al: The 2016 revision to the World Health
352 Organization classification of myeloid neoplasms and acute leukemia. Blood 127:2391-2405,
353 2016

354 5. Stein EM, DiNardo CD, Pollyea DA, et al: Enasidenib in mutant IDH2 relapsed or
355 refractory acute myeloid leukemia. Blood 130:722-731, 2017

356 6. Stone RM, Mandrekar SJ, Sanford BL, et al: Midostaurin plus Chemotherapy for
357 Acute Myeloid Leukemia with a FLT3 Mutation. N Engl J Med 377:454-464, 2017

358 7. Papaemmanuil E, Gerstung M, Bullinger L, et al: Genomic Classification and
359 Prognosis in Acute Myeloid Leukemia. N Engl J Med 374:2209-2221, 2016

360 8. Cooper TM, Cassar J, Eckroth E, et al: A Phase I Study of Quizartinib Combined
361 with Chemotherapy in Relapsed Childhood Leukemia: A Therapeutic Advances in Childhood
362 Leukemia & Lymphoma (TACL) Study. Clinical Cancer Research 22:4014-4022, 2016

363 9. Smith CC, Paguirigan A, Jeschke GR, et al: Heterogeneous resistance to
364 quizartinib in acute myeloid leukemia revealed by single-cell analysis. Blood 130:48-58, 2017

365 10. Paguirigan AL, Smith J, Meshinchi S, et al: Single-cell genotyping demonstrates
366 complex clonal diversity in acute myeloid leukemia. Sci Transl Med 7:281re2, 2015

367 11. Corces-Zimmerman MR, Hong WJ, Weissman IL, et al: Preleukemic mutations in
368 human acute myeloid leukemia affect epigenetic regulators and persist in remission. Proc Natl
369 Acad Sci U S A 111:2548-53, 2014

370 12. Ding L, Ley TJ, Larson DE, et al: Clonal evolution in relapsed acute myeloid
371 leukaemia revealed by whole-genome sequencing. Nature 481:506-10, 2012

372 13. Padron E, Yoder S, Kunigal S, et al: ETV6 and signaling gene mutations are
373 associated with secondary transformation of myelodysplastic syndromes to chronic
374 myelomonocytic leukemia. Blood 123:3675-7, 2014

375 14. Wong TN, Ramsingh G, Young AL, et al: Role of TP53 mutations in the origin and
376 evolution of therapy-related acute myeloid leukaemia. Nature 518:552-555, 2015

377 15. Gillis NK, Ball M, Zhang Q, et al: Clonal haemopoiesis and therapy-related myeloid
378 malignancies in elderly patients: a proof-of-concept, case-control study. Lancet Oncol 18:112-
379 121, 2017

380 16. Zuffa E, Franchini E, Papayannidis C, et al: Revealing very small FLT3 ITD
381 mutated clones by ultra-deep sequencing analysis has important clinical implications in AML
382 patients. Oncotarget 6:31284-94, 2015

383 17. Karaayvaz M, Cristea S, Gillespie SM, et al: Unravelling subclonal heterogeneity
384 and aggressive disease states in TNBC through single-cell RNA-seq. Nature Communications
385 9:3588, 2018

386 18. Giustacchini A, Thongjuea S, Barkas N, et al: Single-cell transcriptomics uncovers
387 distinct molecular signatures of stem cells in chronic myeloid leukemia. Nat Med 23:692-702,
388 2017

389 19. Hill MO: Diversity and evenness: a unifying notation and its consequences.
390 Ecology 54:427-432, 1973

391 20. Zheng GX, Terry JM, Belgrader P, et al: Massively parallel digital transcriptional
392 profiling of single cells. Nat Commun 8:14049, 2017

393 21. Lambrechts D, Wauters E, Boeckx B, et al: Phenotype molding of stromal cells in
394 the lung tumor microenvironment. Nat Med 24:1277-1289, 2018

395 22. Carruth J, Tygert M, Ward R: A comparison of the discrete Kolmogorov-Smirnov
396 statistic and the Euclidean distance. arxiv.org:arXiv:1206.6367, 2012

397 23. Lou J: Entropy and diversity. Oikos 113:363-375, 2006

398 24. Greaves MF, Brown J, Molgaard HV, et al: Molecular features of CD34: a
399 hemopoietic progenitor cell-associated molecule. Leukemia 6 Suppl 1:31-6, 1992

400 25. Krause DS, Fackler MJ, Civin CI, et al: CD34: structure, biology, and clinical utility.
401 Blood 87:1-13, 1996

402 26. Zhang Y, Gao S, Xia J, et al: Hematopoietic Hierarchy - An Updated Roadmap.
403 Trends Cell Biol, 2018

404 27. Trapnell C, Cacchiarelli D, Grimsby J, et al: The dynamics and regulators of cell
405 fate decisions are revealed by pseudotemporal ordering of single cells. Nat Biotechnol 32:381-
406 386, 2014

407 28. Ellsworth DL, Blackburn HL, Shriver CD, et al: Single-cell sequencing and
408 tumorigenesis: improved understanding of tumor evolution and metastasis. Clin Transl Med
409 6:15, 2017

410 29. Schelker M, Feau S, Du J, et al: Estimation of immune cell content in tumour tissue
411 using single-cell RNA-seq data. Nat Commun 8:2032, 2017

412 30. Pellegrino M, Sciambi A, Treusch S, et al: High-throughput single-cell DNA
413 sequencing of acute myeloid leukemia tumors with droplet microfluidics. Genome Res 28:1345-
414 1352, 2018

415 31. Navin NE: The first five years of single-cell cancer genomics and beyond. Genome
416 Res 25:1499-507, 2015

417 32. Wang L, Livak KJ, Wu CJ: High-dimension single-cell analysis applied to cancer.
418 Mol Aspects Med 59:70-84, 2018

419 33. Pellegrino M, Sciambi A, Treusch S, et al: High-throughput single-cell DNA
420 sequencing of acute myeloid leukemia tumors with droplet microfluidics. Genome research 28:1-
421 8, 2018

422 34. Gatenby RA, Silva AS, Gillies RJ, et al: Adaptive therapy. Cancer Res 69:4894-
423 903, 2009

424 35. Staňková K, Brown JS, Dalton WS, et al: Optimizing cancer treatment using game
425 theory: A review. JAMA Oncology, 2018

426 36. MacArthur RH: Patterns of species diversity. . Biological Reviews 40:510-533,
427 1965

428 37. Tuomisto H: A consistent terminology for quantifying species diversity? Yes, it
429 does exist. Oecologia 164:853-60, 2010

430 38. Almendro V, Cheng YK, Randles A, et al: Inference of tumor evolution during
431 chemotherapy by computational modeling and in situ analysis of genetic and phenotypic cellular
432 diversity. Cell Rep 6:514-27, 2014

433 39. Marusyk A, Tabassum DP, Altrock PM, et al: Non-cell-autonomous driving of
434 tumour growth supports sub-clonal heterogeneity. Nature 514:54-8, 2014

435 40. Altrock PM, Liu LL, Michor F: The mathematics of cancer: integrating quantitative
436 models. Nat Rev Cancer 15:730-45, 2015

437 41. Li J, Smalley I, Schell MJ, et al: SinCHet: a MATLAB toolbox for single cell
438 heterogeneity analysis in cancer. Bioinformatics 33:2951-2953, 2017

439 42. Kaminskas E, Farrell AT, Wang YC, et al: FDA drug approval summary: azacitidine
440 (5-azacytidine, Vidaza) for injectable suspension. Oncologist 10:176-82, 2005

441 43. Zhang J, Cunningham JJ, Brown JS, et al: Integrating evolutionary dynamics into
442 treatment of metastatic castrate-resistant prostate cancer. Nat Commun 8:1816, 2017

443 44. Gallaher JA, Enriquez-Navas PM, Luddy KA, et al: Spatial Heterogeneity and
444 Evolutionary Dynamics Modulate Time to Recurrence in Continuous and Adaptive Cancer
445 Therapies. Cancer Res 78:2127-2139, 2018

446
447

SUPPLEMENTAL MATERIAL

Leveraging single cell RNA sequencing experiments to model intra-tumor heterogeneity

Meghan C. Ferrall-Fairbanks (1), Markus Ball (2,3),
Eric Padron (2), and Philipp M. Altrock* (1)

1: Department of Integrated Mathematical Oncology, Moffitt Cancer Center and Research Institute

2: Department of Malignant Hematology, Moffitt Cancer Center and Research Institute, Tampa, FL 33612, USA

3: present address: Institute of Pathology, University Hospital Cologne, 50937 Köln, Germany

*Corresponding author: Philipp M. Altrock, philipp.altrock@moffitt.org
Moffitt Cancer Center, 12902 USF Magnolia Drive, SRB 24007, Tampa, FL 33612, USA
Phone: +1 813 745-5897 Fax: +1 813 745- 8357

467 **SUPPLEMENTAL MATERIALS AND METHODS**

468 DROPLET-BASED $scRNA_{SEQ}$ SAMPLES

469 The 10X Genomics platform offered a variety of datasets[1] that were used to create our
470 pipeline for quantifying intraleukemic heterogeneity. The major focus of our analysis and pipeline
471 development was of Healthy/Control and AML patient bone marrow mononulear cells (BMMCs).
472 The Healthy Controls 1 and 2 BMMCs had been sequenced on Illumina Hiseq 2500 Rapid Run
473 V2 with 90-135 thousand reads per cell and 2000-2400 cells detected. The AML027 and
474 AML035 pre-transplant BMMCs had been sequenced on Illumnia NextSeq 500 High Output with
475 16.6-58 thousand reads per cell and 3500-3900 cells detected. The AML027 and AML035 post-
476 transplant BMMCs had also been sequenced on Illumina NextSeq 500 High Output with 41-51
477 thousand reads per cell and 900-3900 cells detected. Additional datasets[1] for CD34+, CD14+,
478 CD19+, and CD4+ (hematopoietic cell type set) cells as well as patient normal-tumor matched
479 lung cancer samples[2] available in ArrayExpress under accessions E-MTAB-6653 and E-MTAB-
480 6149 were used to test the robustness of the pipeline. The CD34+ dataset was CD34+ cells
481 enriched from peripheral blood mononuclear cells (PBMCs), sequenced on an Illumnia NextSeq
482 500 High Output with 24.7 thousand reads per cell with 9000 cells detected. The CD14+ dataset
483 was enriched from PBMCs, sequenced on Illumnia NextSeq 500 Output with 100 thousand
484 reads per cell with 2600 cells detected. The CD19+ dataset was enriched from PBMCs,
485 sequenced on Illumnia NextSeq 500 Output with 25 thousand reads per cell with 10000 cells
486 detected. The CD4+ dataset was enriched from PBMCs, sequenced on Illumnia NextSeq 500
487 High Output with 21 thousand reads per cell with 11000 cells detected. All lung tissue samples
488 were prepared as single-cell suspensions, sequenced on Illumina HiSeq4000, and aimed for an
489 estimated 4000 cells per library[2]. There were six patients analyzed, each patient had 4 distinct

490 tissues locations sampled: an adjuacent normal lung sample (normal), a sample from the tumor
491 core (core), a sample from the tumor margin (edge), and a sample between the tumor core and
492 margin/edge (middle).

493 INITIAL DATA PROCESSING

494 First, the all samples were run through the 10x Genomics cellranger count pipeline for
495 transcriptome alignment, calling individual cell barcodes, estimating multiplet rates, and finally
496 clustering of normalization corrected FPKM (fragments per kilobase million; expression level)
497 values (**Fig. S1**, pipeline X). In our analysis of the Healthy versus AML high-throughput scRNA-
498 seq, we used cellranger aggr to pool together the individual dataset results, so we could compare
499 intra-tumor heterogeneity (ITH) between samples with the same clustering. Using a graph-based
500 clustering algorithm, 23 different clusters were identified based on clustering cells by expression
501 similarity (shown in **Fig. 2B-C**). The clustering was performed by forming a graph with cells as
502 vertices and edges indicating pairs of cells that are sufficiently similar (implemented by
503 CellRanger). The CellRanger graph-based clustering with Louvain Modularity Optimization[3] was
504 used to partition this graph into clusters of similar cells. For analysis with the hematopoietic cell
505 set and the lung cancer samples the cell/gene matrices produced by cellranger count pipeline
506 were then loaded into R using the Seurat package[4] (designed for QC, analysis, and exploration
507 of single cell RNA-seq data; **Fig. S1**, pipelines M and S). With Seurat, cells were removed that
508 had either fewer than 200 UMIs and greater than 6000 UMIs and more than 10% mitochondrial
509 DNA[2], then scaled and normalized before clustering using graph-based clustering.

510 QUANTIFYING SUMMARY DIVERSITY METRICS

511 After the data was clustered, we then sought to quantify whether the clustering could
512 disentangle healthy and AML samples by average gene expression. To approach within-sample

513 differences in overall gene expression, we computed Euclidean distance matrices of the mean
514 expression values in each cluster, to establish a distance metric of cluster differences and a
515 subsequent diversity of distances across samples (standard deviation, ANOVA) (**Fig. 2A**). For
516 each cluster, the geometric mean of each unique molecular identifier (UMI) across all genes
517 were computed, then the Euclidean distance was computed between clusters. This was plotted
518 as a graph where each node represents each clusters identified in the leukemic dataset (**Fig.**
519 **2D**). The size of the node indicates the total number of cells in that cluster and the color identifies
520 the major species (AML, Healthy, or postBMT) present in that cluster. The distribution of cells
521 per cluster (**Fig. 2E**) was also included to show which other conditions were also present in each
522 cluster and how the number of cells compared to the major species used to dictate the color of
523 the nodes in graph describing the connectivity of the clusters in **Fig. 2D**.

524 Next, we sought to characterize across-sample differences by calculating the
525 Kolmogorov-Smirnov (KS) distance[5] of the cell count distributions in each cluster (**Fig. 3A**). The
526 KS distance is a non-parametric measure that is calculated as the supremum of paired
527 differences of two empirical probability mass functions[5]. Last, we calculated a continuum of the
528 ecological diversity index[6] based on the individual cell frequencies in each of the clusters
529 identified by the graph-based clustering (**Fig. 3H**), across all clusters, which can be written as:

530
$$^{q}D = \left(\sum_{I=1}^{n} p_i^q \right)^{\frac{1}{1-q}}$$

531 where n is the number of clusters in the data set, p_n is the frequency of each cluster, and q is
532 the order of diversity. q is a hyperparameter that would be optimized in a clinical setting. The two
533 distributions for AML were joined together and the two distributions of the healthy indiviuals were

534 joined together and plotted across orders of diversity (**Fig. 4A-C**). The same technique was used
535 to group the lung cancer samples based on location as reported in (**Fig. 5B**).

536 DOWNSAMPLING

537 We sought to test the robustness of our metrics by downsampling the number of cells and
538 re-calculating the number of clusters identified as well as changes in key diversity scores. This
539 downsampling was preformed removing increments of 10% of the data at a time and the results
540 were reported as the summary statistics of 1000 runs at each downsampling percentile (**Fig.**
541 **S2**). The downsampling results were also used to quantify confidence in the diversity index
542 spectrum and was calculated for the following orders of diversity: 10^{-2}, 10^{-1}, 10^{0}, 2, 10^{1}, 10^{2}. The
543 relative change in diversity score was reported for the downsampling results with 50% the initial
544 amount of cells in **Fig. 4D-I**. Relative change in the diversity score was calculated by subtracting
545 the mean of the diversity score from all 1000 and reported as the distributions of scores around
546 that mean score.

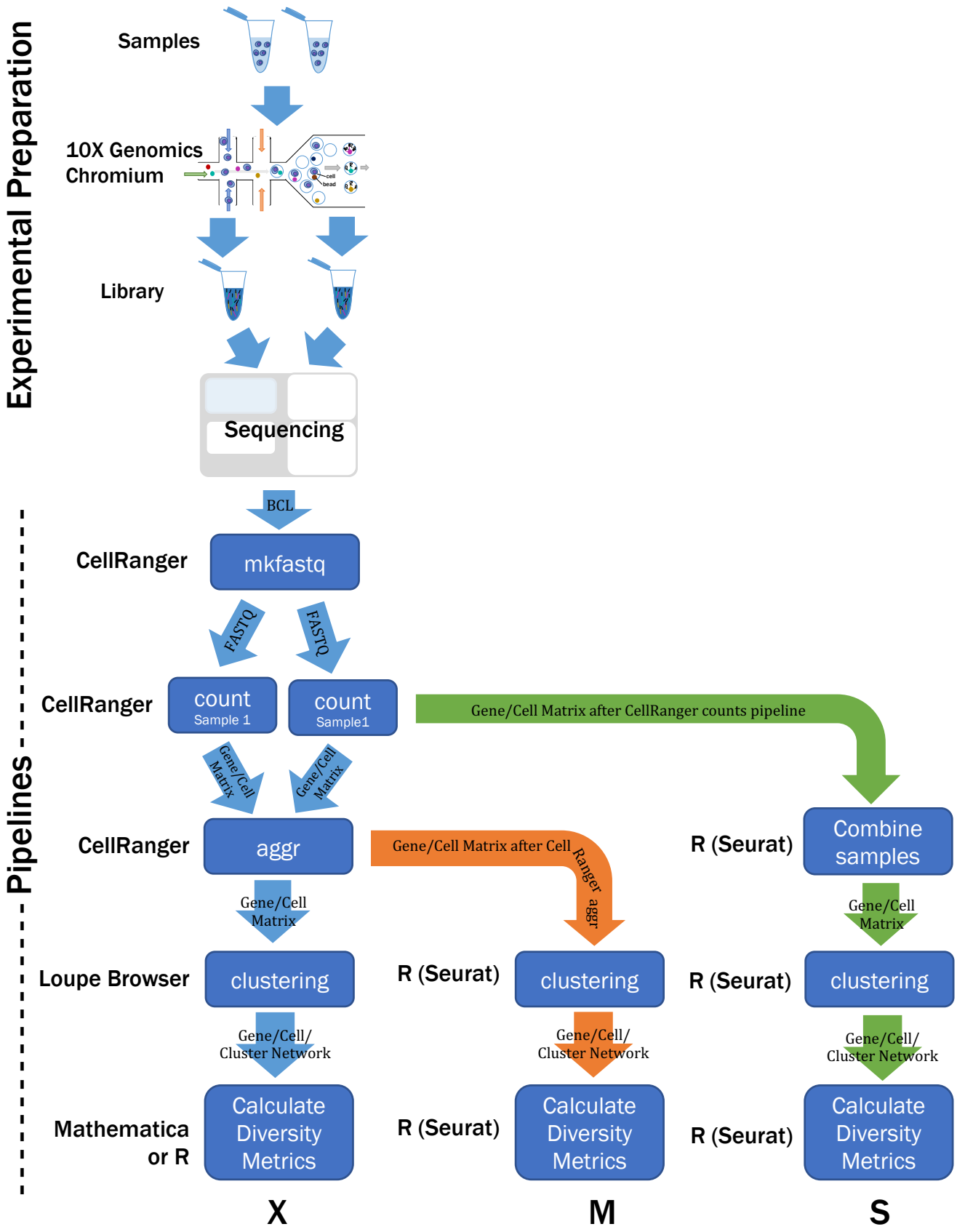
547 CODE AVAILABILITY

548 The code used in the pipelines described in **Fig. S1** were uploaded to a GitHub repository
549 (https://github.com/mcfefa/scRNAseq). These pipelines were implemented using data run
550 through at least cellranger counts, and then post-processed with CellRanger Loupe Browser,
551 Mathematica, and R. Code available includes:

552 - Mathematica code and R scripts and data described to implement all pipelines (described
553 in **Fig. S1**)
554 - Mathematica code for calculating the geometric mean of UMIs and Euclidean distances
555 between clusters (for **Fig. 2**)
556 - R code to draw the graph of the clusters (**Fig. 2D**)

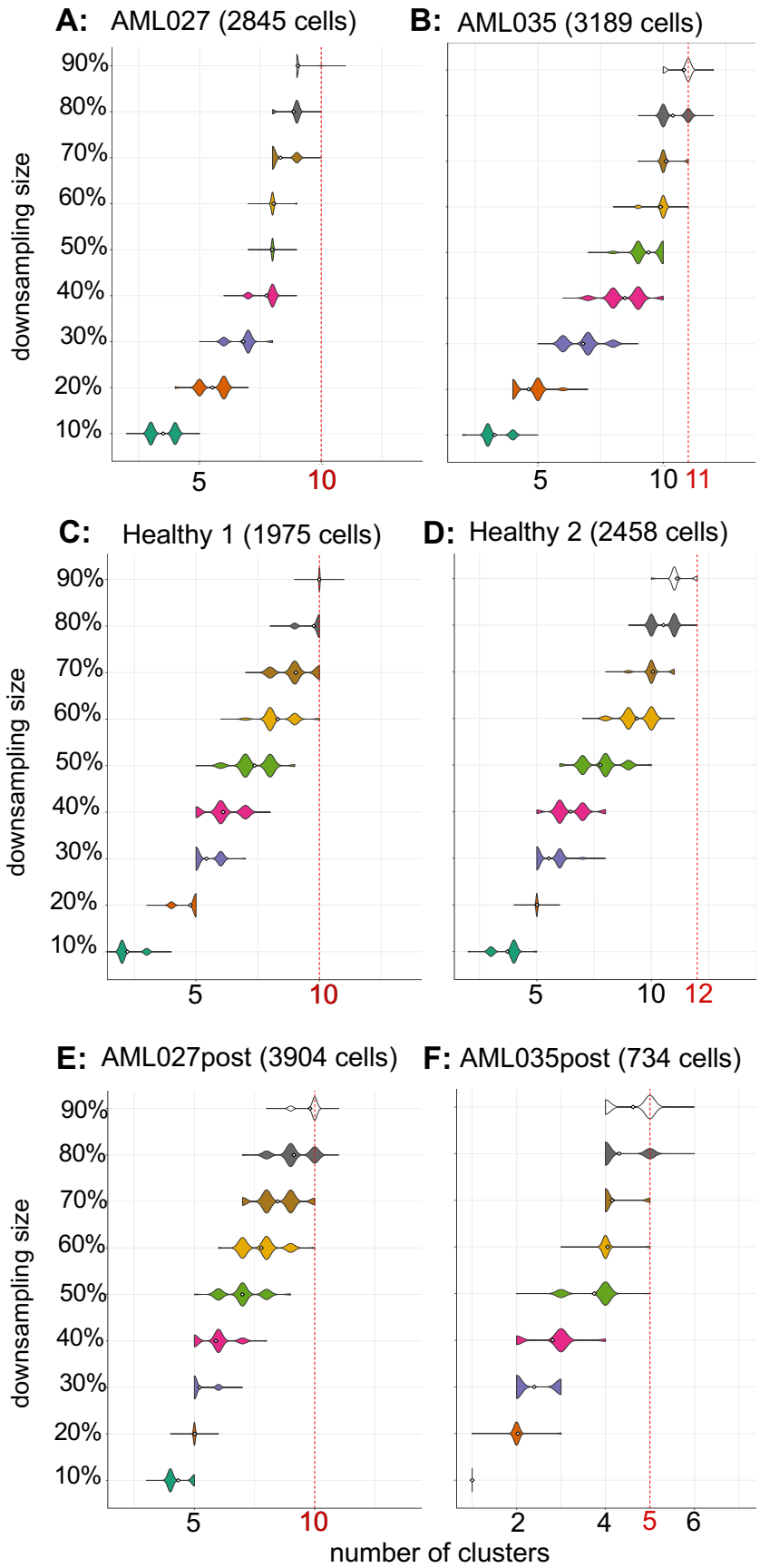
557 • Mathematica code for calculating the KS distance and diversity spectrum across all
558 clusters for the leukemic dataset (**Fig. 3**)
559 • R code used to downsample datasets, then cluster and calculate diversity indices (**Fig. 4**
560 and **S2**)
561 • R code for calculating the diversity spectrum for the hematopioteic subtype and matched
562 lung carcinoma datasets (**Fig. 5**)
563

564 **Supplemental Figure 1. Workflow diagram for the scRNAseq quantifying intraleukemia**
565 **heterogeneity for an example comparing two samples.** Cells were processed through the
566 10X Genomics Chromium and ultimately sequenced, producing a raw base call (BCL) file, which
567 was demultiplexed for each flowcell directory and converted to a FASTQ file using cellranger
568 mkfastq. Then cellranger count was run separately for each library to align, filter, and count
569 barcodes and UMIs. These instances were then aggregated into a single instance using
570 cellranger aggr to normalize runs to the same sequencing depth and recompute analysis on
571 combined dataset. Following along pipeline X, from the aggregated analysis, the Loupe Browser
572 contains the clustered data, which were exported and run through a Mathematic script to
573 calculate the diversity metrics. This pipeline was used to generate Figs. 3 and 4. This pipeline
574 was expanded to give the user increased flexibility by using the gene-barcode matrices output
575 by cellranger and then processing the data in R using the Seurat package, which allows users
576 to set the filtering criteria (M) and allows data to be combined without running through the
577 cellranger aggr pipeline (S) as well as both new pipelines allow users to implement additional
578 clustering algorithms. Pipeline S was used to generate Fig. 5 from gene-barcode matrices
579 available from additional data from Zheng, et. al.[18].
580


Experimental Preparation
Samples
10X Genomics Chromium
Library
Sequencing
Pipelines
BCL
CellRanger
mkfastq
FASTQ
FASTQ
CellRanger
count Sample 1
count Sample1
Gene/Cell Matrix after CellRanger counts pipeline
Gene/Cell Matrix
Gene/Cell Matrix
CellRanger
aggr
Gene/Cell Matrix after Cell Ranger aggr
R (Seurat)
Combine samples
Gene/Cell Matrix
Gene/Cell Matrix
Loupe Browser
clustering
R (Seurat)
clustering
R (Seurat)
clustering
Gene/Cell/ Cluster Network
Gene/Cell/ Cluster Network
Gene/Cell/ Cluster Network
Mathematica or R
Calculate Diversity Metrics
R (Seurat)
Calculate Diversity Metrics
R (Seurat)
Calculate Diversity Metrics
X
M
S

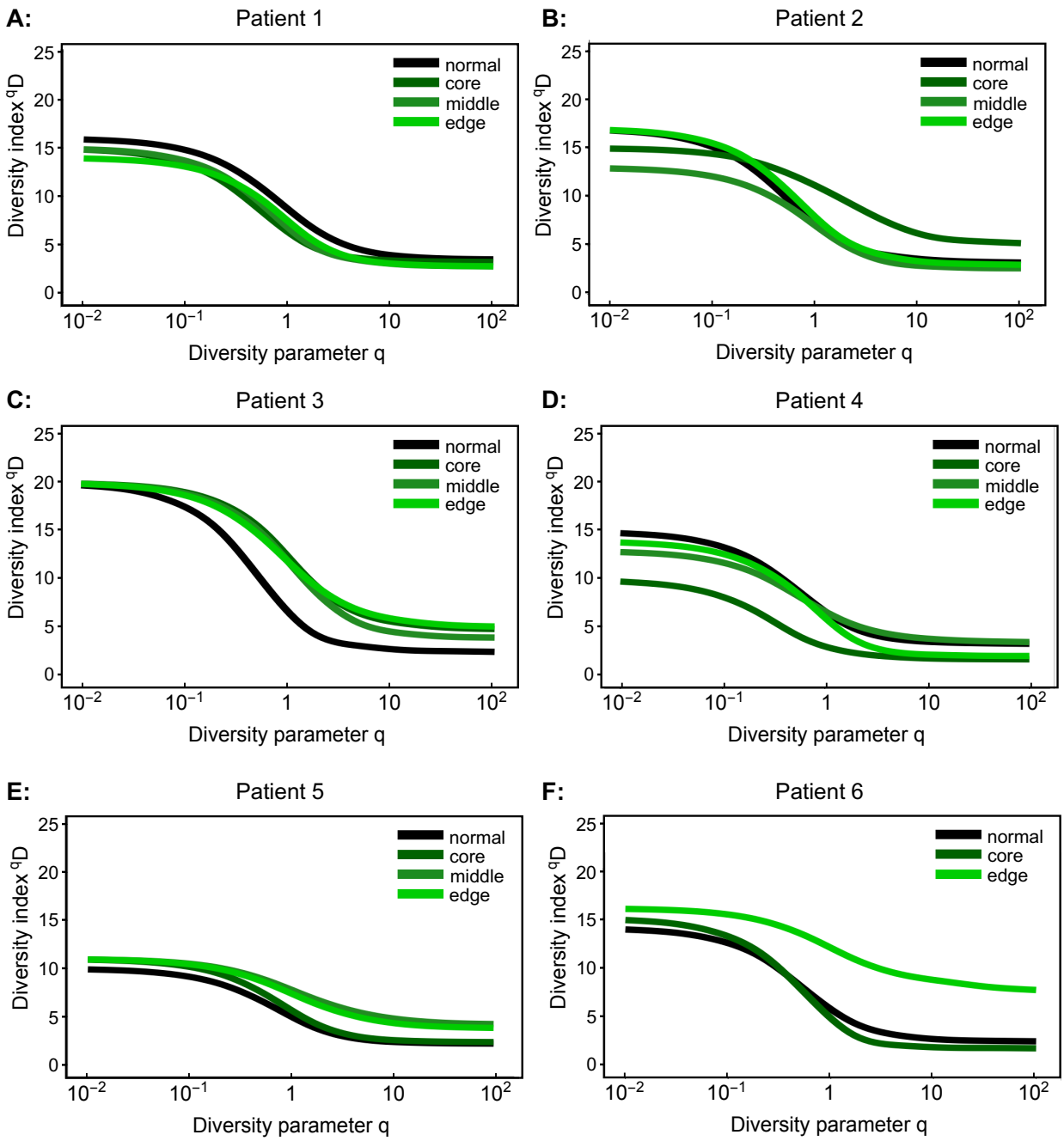
581

582 **Supplemental Figure 2. AML populations converge to consistent number of clusters that**
583 **describe the population sooner than healthy populations. graph-based clustering.** Each
584 original population's gene-barcode matrix obtain from publicly available data from Zheng, et.
585 al.[18] was then downsampled to contain 10% to 90% of the cells originally in the matrix. This data
586 was loaded into R using the Seurat package, downsampled, and then clustered to determine
587 how sensitive the cluster metrics was to the starting number of cells. Downsampling was
588 performed 1000 times per cutoff and the violin plots show the distribution of clusters identified.
589 The red-dashed lined indicates the number of clusters identified with no cells removed from the
590 dataset. Gray numbers indicate the number of cells present in each downsampling condition.
591 AML populations (**A**, **B**) showed quickest convergence to number of clusters after about 1500
592 cells were present in clustering. Healthy populations (**C**, **D**) showed that as more cells were
593 added, an additional cluster could be found. AML post-bone marrow transplant populations (**E**,
594 **F**) behaved more similar to healthy than AML populations.
595


A: AML027 (2845 cells)
B: AML035 (3189 cells)
C: Healthy 1 (1975 cells)
D: Healthy 2 (2458 cells)
E: AML027post (3904 cells)
F: AML035post (734 cells)
downsampling size
90%
80%
70%
60%
50%
40%
30%
20%
10%
5
10
11
12
2
4
6
number of clusters

596

Supplemental Figure 3: Individual lung cancer patient diversity scores show same trends as aggregated clustering results with tumor samples have greater diversity scores than their matched normal counterpart. Individual patients were run through pipeline S to aggregate normal, core, middle, and edge tissue samples.



603 **REFERENCES**

604

605 1. Zheng GX, Terry JM, Belgrader P, et al: Massively parallel digital transcriptional
606 profiling of single cells. Nat Commun 8:14049, 2017

607 2. Lambrechts D, Wauters E, Boeckx B, et al: Phenotype molding of stromal cells in
608 the lung tumor microenvironment. Nat Med 24:1277-1289, 2018

609 3. 10xGenomics: https://support.10xgenomics.com/single-cell-gene-
610 expression/software/pipelines/latest/algorithms/overview, 2018

611 4. Butler A, Hoffman P, Smibert P, et al: Integrating single-cell transcriptomic data
612 across different conditions, technologies, and species. Nat Biotechnol 36:411-420, 2018

613 5. Carruth J, Tygert M, Ward R: A comparison of the discrete Kolmogorov-Smirnov
614 statistic and the Euclidean distance. arxiv.org:arXiv:1206.6367, 2012

615 6. Lou J: Entropy and diversity. Oikos 113:363-375, 2006

616
617