Title

Multifunctional barcoding with ClonMapper enables high-resolution study of clonal dynamics during tumor evolution and treatment

Journal

Nature Cancer, 2(7)

ISSN

2662-1347

Authors

Gutierrez, Catherine
Al'Khafaji, Aziz M
Brenner, Eric
et al.

Publication Date

2021-07-01

DOI

10.1038/s43018-021-00222-8

Peer reviewed

Published in final edited form as:

Nat Cancer. 2021 July ; 2(7): 758–772. doi:10.1038/s43018-021-00222-8.

Multifunctional barcoding with ClonMapper enables high-resolution study of clonal dynamics during tumor evolution and treatment

Catherine Gutierrez[1,2,15], **Aziz M. Al'Khafaji**[3,4,5,15], **Eric Brenner**[3,4,15], **Kaitlyn E. Johnson**[4], **Satyen H. Gohil**[1,2,5,6,7], **Ziao Lin**[1,5,8], **Binyamin A. Knisbacher**[5], **Russell E. Durrett**[3,4], **Shuqiang Li**[2,5,9], **Salma Parvin**[2], **Anat Biran**[2,5], **Wandi Zhang**[2], **Laura Rassenti**[10], **Thomas J. Kipps**[10], **Kenneth J. Livak**[2,9], **Donna Neuberg**[11], **Anthony Letai**[1,2], **Gad Getz**[1,5,12,13], **Catherine J. Wu**[1,2,5,14,✉], **Amy Brock**[3,4,✉]

[1]Harvard Medical School, Boston, MA, USA.

[2]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA, USA.

[3]Institute of Cellular and Molecular Biology, University of Texas at Austin, Austin, TX, USA.

[4]Department of Biomedical Engineering, University of Texas at Austin, Austin, TX, USA.

[5]Broad Institute of MIT and Harvard, Cambridge, MA, USA.

[6]Department of Academic Haematology, University College London, London, UK.

[7]Department of Clinical Haematology, University College London Hospitals NHS Foundation Trust, London, UK.

[8]Harvard University, Cambridge, MA, USA.

[9]Translational Immunogenomics Laboratory, Dana-Farber Cancer Institute, Boston, MA, USA.

[10]Department of Medicine, University of California at San Diego Moores Cancer Center, La Jolla, CA, USA.

[11]Department of Data Sciences, Dana Farber Cancer Institute, Boston, MA, USA.

[12]Department of Pathology, Massachusetts General Hospital, Boston, MA, USA.

✉**Correspondence and requests for materials** should be addressed to C.J.W. or A.B. cwu@partners.org; amy.brock@utexas.edu.

Author contributions

C.G. and A.M.A. conceptualized and designed the research. C.G., A.M.A. and E.B. performed and analyzed the bulk of the experiments, under supervision by A. Brock. and C.J.W. K.E.J. performed all mathematical modeling and growth and death rate analyses. Z.L. performed all WGS analyses, supervised by G.G. R.D. generated the computational pipelines for barcode analysis and processed the RNA-seq and ATAC-seq data. S.L. and K.J.L. prepared the 10X Genomics scRNA-seq libraries. S.P. and A.L. performed and analyzed the BH3 profiling analyses. S.G. performed single-cell experiments and provided and facilitated transfer of data from patient samples. B.A.K. collected, processed and analyzed bulk RNA-seq from patients with CLL and compared to HG3. W.Z. and A. Biran processed patient samples for analysis. L.R. and T.J.K. provided primary CLL samples. E.B., K.E.J. and D.N. contributed to statistical analyses. All authors participated in data analysis. C.G., A.M.A. and E.B. wrote the manuscript. All authors reviewed the manuscript before submission. C.J.W. and A. Brock. equally and jointly supervised this work.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s43018-021-00222-8.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43018-021-00222-8.

[13]Center for Cancer Research, Massachusetts General Hospital, Boston, MA, USA.

[14]Department of Medicine, Brigham and Women's Hospital, Boston, MA, USA.

[15]These authors contributed equally: Catherine Gutierrez, Aziz M. Al'Khafaji, Eric Brenner.

Abstract

Lineage-tracing methods have enabled characterization of clonal dynamics in complex populations, but generally lack the ability to integrate genomic, epigenomic and transcriptomic measurements with live-cell manipulation of specific clones of interest. We developed a functionalized lineage-tracing system, ClonMapper, which integrates DNA barcoding with single-cell RNA sequencing and clonal isolation to comprehensively characterize thousands of clones within heterogeneous populations. Using ClonMapper, we identified subpopulations of a chronic lymphocytic leukemia cell line with distinct clonal compositions, transcriptional signatures and chemotherapy survivorship trajectories; patterns that were also observed in primary human chronic lymphocytic leukemia. The ability to retrieve specific clones before, during and after treatment enabled direct measurements of clonal diversification and durable subpopulation transcriptional signatures. ClonMapper is a powerful multifunctional approach to dissect the complex clonal dynamics of tumor progression and therapeutic response.

Tumors are composed of heterogeneous populations of cells with variable genetic and epigenetic characteristics that together contribute to their ability to adapt and further diversify in response to their environment. This intrinsic heterogeneity, revealed through recent high-throughput sequencing efforts, fuels cancer's powerful evolutionary capacity and underlies the short-lived nature of many therapeutic successes[1,2]. Certainly, the rational development of more effective treatments will require a more complete understanding of the processes by which individual cells arrive at each of their successful evolutionary outcomes.

Until now, studies of clonal evolution have primarily tracked variant allele frequencies (VAF) over time using bulk tumor DNA sequencing[3–6], but this approach is currently limited by the lower bounds of sequencing detection[2,7]. Recent efforts to monitor clonal dynamics at higher scale and resolution have leveraged DNA barcoding, which introduces a unique heritable marker into the genome of individual cells such that they and their descendants can be tracked over time[8–12]. While DNA barcoding has evolved in the past years to link clonal identity with single-cell transcriptomic measurements[13–18], these methods are destructive and thus do not allow subsequent molecular and functional analyses of specific clones of interest. Although approaches that enable live-cell clonal isolation exist, many require extensive manipulation, are limited in their sensitivity or lack integration with single-cell RNA sequencing (scRNA-seq)[19–21]. The ability to systematically identify, isolate and perform integrative characterizations (genetics, epigenetics, proteomics, live-cell assays) on live clones of interest represents the next major advance in analyzing the complex interactions underlying dynamic biological systems. To date, the lack of sensitive approaches to clonally dissect complex biological systems has constrained our capability to go beyond merely identifying patterns of clonal dynamics.

To address this challenge, we developed a lineage-tracing platform, ClonMapper, which couples the advantages of clonal characterization with clonal isolation. This is achieved by combining DNA barcoding with CROP-seq, a single-guide RNA (sgRNA) expression vector that can express and capture sgRNA barcodes in scRNA-seq protocols and COLBERT, a system that uses barcoded sgRNAs to drive the selective expression of a fluorescent reporter in clones of interest to enable their physical isolation (Fig. 1a–c)[19,22]. This platform thus allows for the comprehensive and systematic characterization of clones within complex evolving populations by enabling the efficient linking of clonal identity with single-cell transcriptomics as well as genetic, epigenetic and live-cell characteristics. We applied ClonMapper to study chronic lymphocytic leukemia (CLL), a low-grade B-cell malignancy whose typically indolent kinetics has provided an opportune setting in which to study clonal evolution. We demonstrate that this integrative platform allows us to identify differential clonal responses to therapy of a CLL model cell line, validate these signatures in primary CLL samples and measure the intrinsic stability of clonal transcriptomic identity.

Patterns of clonal selection in B-cell malignancies

To systematically uncover the clonal and transcriptomic heterogeneity of cancer cell populations at single-cell resolution, we constructed an expressed DNA barcode library using a variant of the CROP-seq vector containing a blue fluorescent protein (BFP) reporter and fully degenerate N(20) barcodes integrated into the sgRNA expression cassette (Fig. 1a). We transduced a representative CLL-like cell line, HG3, which we confirmed to closely reflect the transcriptomic profile of unmutated *IGHV* CLLs (Extended Data Fig. 1a,b), as well as the mantle cell lymphoma (MCL) cell line, REC1, with this highly diverse (~7.6×10^7 unique barcodes; Extended Data Fig. 1c,d) and well-distributed ClonMapper library at a multiplicity of infection (MOI) of 0.1 to minimize multiple barcode integrations per cell. BFP-tagged HG3 and REC1 cells were sorted and expanded to establish the desired medium and high-diversity barcoded populations of tens of thousands of unique clones, respectively (Methods). We exposed these barcoded cancer populations (baseline or TP0) to fludarabine/mafosfamide (FM) (the active in vitro analog of cyclophosphamide) and mafosfamide/doxorubicin/vincristine to reflect the frontline combination chemotherapy regimens for CLL and MCL, respectively. Eight parallel replicates of 1×10^7 cells each were treated with either combination chemotherapy for 72 h at a dose sufficient to generate a deep population bottleneck (95% population cell death; Extended Data Fig. 1e,f), before the drugs were removed and cells returned to standard growth medium.

We observed a massive decrease in viability following treatment across all eight replicates, with an eventual population recovery (TP1) at 21 and 17 d following treatment and with an 80% and 81% decrease in the number of detectable HG3 and REC1 clones, respectively (Fig. 1d). A majority of clones identified in one TP1 replicate were also present in at least one other replicate (~86.6% for HG3 clones; ~89.2% for REC1 clones; Fig. 1e) and a sizable proportion were shared across all replicates (~30% for HG3 clones; ~20% for REC1 clones; Fig. 1e). Given that the majority of clones with increased abundance ($\log_2$(fold change) from the pre-treatment population, TP0 > 0) were present in all eight replicates (Extended Data Fig. 1g) and totaled ~94% and ~62% of the composite HG3 and REC1 populations (Fig.

1e insets and Extended Data Fig. 1h), these findings collectively indicate that pre-existing mechanisms account for the underlying survival in both cell lines.

To characterize the gene expression profiles underlying the clonal consistency observed in the surviving population, we focused our studies on the HG3 cell line, as their diversity (thousands of clones, as opposed to hundreds of thousands in the MCL line) was more amenable to the throughput available for scRNA-seq analysis. We performed lineage-resolved scRNA-seq on two of eight TP1 replicates and a pre-treatment sample (TP0) using the 10X Chromium single-cell platform. We analyzed 8,975 total cells (median unique molecular identifier (UMI) of 10,776; median 2,592 genes per cell, Methods) and identified 1,050 unique expressed barcodes across seven transcriptionally defined Leiden clusters that segregated by treatment status (Fig. 1f–i, Methods).

We traced the trajectories of the ten most abundant clones at TP1 (54.6% of total clones; Supplementary Table 1) by annotating these clones across all transcriptionally defined clusters. All ten clones occupied five of six post-treatment TP1 clusters (0, 1, 3, 4 and 7) and only one of the pre-treatment TP0 clusters (5) (Fig. 1h). Clones present in clusters 0, 1, 3, 4 and 7 could be traced back uniquely to TP0 cluster 5, whereas clones present in cluster 6 could be tracked only to cluster 2, with nominal overlap of clones (7 of 1,050 detected clones) (Fig. 1i and Extended Data Fig. 1i). These data support the presence of two discrete pre-existing gene expression states at TP0; one composed of a unique subset of clones with ‘high survivorship’ (HS) potential that following treatment expands and accounts for the majority of relapse TP1 clones and another composed of a subset of clones with ‘low survivorship’ (LS) potential that decreases following the therapeutic bottleneck.

Phenotypes of surviving clones

We performed marker gene analysis between HS and LS subpopulations at TP0 using MetaCore (>0.5 $\log_2$(fold change) and q-value <0.05, Methods). The HS subpopulation at TP0 exhibited upregulated Wnt signaling (*WNT10A*, *TCF7*), Notch signaling (*JAG1*, *RBPJ*), inflammation (*CXCR4*, *CCL20*, *JAK1*) and chromatin modification (*PCGF5*, *AUTS2*) (Supplementary Table 2), whereas the LS subpopulations exhibited upregulated focal adhesion (*ITGB2* (CD18), *FGFR1* and *VCL*) and class II antigen presentation (*HLA-DMA*, *HLA-DMB*, *HLA-DOA)*. Both HS and LS gene expression profiles were largely maintained following treatment, indicating that these are relatively stable, potentially targetable phenotypes unique to specific clones and are not a result of the interconverting (and thus replenishing) transcriptomic states that have been noted in recent clonal dynamics studies in other disease models (Fig. 2a and Extended Data Fig. 2a,b)[23].

Two cell-surface markers, CXCR4 and CD18, distinguished the HS and LS subpopulations, respectively (Fig. 2b), with cells at TP0 being predominantly $CD18^+$ and at TP1, $CXCR4^+$ (Fig. 2c). Barcode sequencing of cells flow cytometrically isolated from TP0 confirmed that the identities of $CXCR4^+$ and $CD18^+$ cells reflected that of HS and LS clones identified by scRNA-seq (94.4% and 97.4% overlap, respectively; Extended Data Fig. 3a,b). Bulk transcriptome sequencing of these two subpopulations also demonstrated strong correlations with single-cell gene expression (Extended Data Fig. 3c) and marked upregulation of Wnt,

CXCR4 and Notch signaling (Fig. 2c–i, Supplementary Table 1 and Extended Data Fig. 3d) was evident at the protein level (Fig. 2d). This signature was preserved across populations at TP1, including in the $CD18^+$ population, indicating a likely convergence on this signature as a result of treatment.

To determine the overlap between chromatin accessibility and transcriptional signatures, we performed ATAC-seq on HS and LS subpopulations at TP0 (acquired through collection of $CXCR4^+$ and $CD18^+$ subpopulations). Of 1,275 gene promoter peaks detected as differentially accessible in HS versus LS subpopulations, 742 corresponded to differentially expressed genes identified by bulk RNA-seq (Pearson's correlation $r = 0.63$; Extended Data Fig. 3e). We observed a strong enrichment for accessible transcription factor binding sites involved in Wnt, Notch and CXCR4 signaling (*TCF7*, *HEY1*, *RBPJ*) that were also upregulated in HS clones (Extended Data Fig. 3f–g). We confirmed the cytoprotective contributions of these upregulated signaling pathways in the HS subpopulation; exposure of HG3 cells to FM in combination with targeted small molecule inhibitors of the CXCR4, Wnt or Notch pathways resulted in decreased cell viability compared to FM alone (Fig. 2e and Extended Data Fig. 3h).

The consistency of the HS signature at the mRNA and protein levels across post-treatment populations, together with the drug treatment findings, support a likely role of its constituent signaling pathways in chemotherapy survival (including CXCR4, Notch and Wnt, as noted in Supplementary Table 2). Indeed, analysis of matched transcriptomes collected before and after relapse to fludarabine-based chemotherapy in a cohort of 23 patients with *IGHV*-unmutated CLL revealed a marked enrichment of the HS ($P = 0.0024$) but not the LS signature (Fig. 2f,g; Methods; Supplementary Tables 3 and 4)[24,25]. Among the top three HS constituent pathways, CXCR4 and Notch signaling was enriched at relapse ($P = 0.003$ and 0.033, respectively). While Wnt signaling was not enriched at the bulk RNA-seq level (Extended Data Fig. 4a), analysis at single-cell resolution performed for 4 of 23 patients revealed that two of four samples (Patients 2 and 4) had cell subpopulations harboring a detectable Wnt signature that also exhibited a HS signature (Fig. 2h and Extended Data Fig. 4b,c), consistent with the subclonal dynamics observed in our model cell line.

We also evaluated whether these discovered signatures were relevant in the context of treatment with new targeted inhibitors, such as the BTK-targeting agent ibrutinib. Bulk CLL transcriptomes from 13 patients with *IGHV*-unmutated disease collected 6–12 months after initiating therapy were not enriched for Wnt, Notch or CXCR4 pathway signatures (Extended Data Fig. 4d). However, single-cell analysis from one of two ibrutinib-treated patients with CLL demonstrated that subpopulations expressing HS and LS signatures can exist simultaneously in patients and that treatment was associated with enrichment of cells expressing the HS and CXCR4 signatures (Fig. 2i,j and Extended Data Fig. 4e). Of note, previous studies have shown CXCR4 signaling to be transcriptionally upregulated in the setting of BCR/P13K blockade and support the activity of this pathway as a resistance mechanism in lymphoid malignancies[26,27]. Altogether, these data support these three signaling cascades as likely promoting survival to chemotherapy and even to ibrutinib therapy (Extended Data Fig. 4f).

Clonal growth dynamics and interactions

To determine whether basal growth kinetics differ between HS and LS clones, we used barcode and population measurements during expansion to TP0 to calculate the respective growth rate of individual clones (n = 5,058) (Extended Data Fig. 5a, Methods). Within this bulk population, no differences in growth rate were detected between HS and LS clones (Fig. 3a) and their proportions remained stable over time (Fig. 3b and Extended Data Fig. 5b). However, upon expansion of $CXCR4^+$ HS and $CD18^+$ LS subpopulations independently, the HS subpopulation exhibited a markedly faster baseline growth rate (doubling time (dT) = 18.84 h) than the bulk population or LS subpopulation (dT = 24.17 h and 31.21 h, respectively) (Fig. 3c). We further observed an increased growth rate among the LS subpopulation following treatment exposure (dT = 17.18 h), whereas that of the HS subpopulation remained unchanged (dT = 18.87 h) (Extended Data Fig. 5c).

These findings suggested that the LS clones impose antiproliferative effects on HS clones. We therefore co-cultured the two subpopulations, each marked by a unique fluorophore, in a range of proportions over 14 d (Methods). Populations consisting of only 10–20% HS clones displayed stability over time, reproducing the equilibrium observed in the bulk population. However, as the starting abundance of HS clones increased beyond 30%, their prevalence rose over time (P = 0.02; Fig. 3d and Extended Data Fig. 5d), though not at the rate of HS clones cultured in isolation.

In aggregate, these findings support the notion that LS clones exert a suppressive growth effect on HS clones that is mitigated as LS cells decrease in abundance. To explore potential mechanisms underlying such a suppression, we applied the single-cell interactome analysis tool CellPhoneDB to identify candidate ligand–receptor pair interactions between the HS and LS subpopulations (Fig. 3e)[28]. Of seven ligand–receptor pairs identified in this manner, four were of ligands differentially expressed by HS cells with cognate receptors upregulated in LS cells, whereas three were of ligands differentially expressed by LS cells with cognate receptors upregulated in HS cells (P < 0.05). Notable among the latter was expression of *MDK*, which encodes a ligand regulator of proliferation in CLL, *LGALS9*, which encodes a tandem-repeat-type galectin demonstrated to exert antiproliferative effects on diverse cancer cell types and to directly suppress BCR signaling and *SORL1*, which encodes a low-density lipoprotein receptor with pleiotropic functions including cytokine signaling and chemoattraction[29–38]. We observed similar patterns of ligand–receptor expression in a number of patient single-cell gene expression samples, particularly in Patient 2 (Fig. 3e–i and Extended Data Fig. 5e), where expression of *LGALS9* and *SORL1* was associated with HS and LS-like subpopulations, respectively. Collectively, our data illustrate how ClonMapper can resolve the influence of direct cellular interactions on clonal dynamics and equilibrium.

High survivors exhibit drug tolerance

We subjected bulk, HS and LS subpopulations to an LD_{95} dose (dose lethal to 95% of population tested) of FM for 72 h and monitored these populations longitudinally using cell counts. Following an initial period of cell death and a prolonged period of stasis, the HS and

bulk population re-populated at days 10 and 13, respectively, whereas the LS subpopulation did not regrow within the observed window (Fig. 3f). Higher levels of apoptosis, measured longitudinally via live-cell imaging, were also observed in the LS versus HS subpopulation at baseline and over the course of treatment (Fig. 3g). Broad higher apoptotic priming (Bim and PUMA; Fig. 3h) and priming through targeted apoptotic cascades (Bad, MS1, Hrk, FS1) was detected at baseline in the LS versus HS subpopulations (Extended Data Fig. 5f). Thus, HS cells are less primed for apoptosis and less sensitive to FM treatment.

To identify the specific mechanisms employed by surviving cells during the early time points following treatment (when <5% of cells remain viable and much of the population is in stasis), we treated eight replicates of 1×10^7 cells with an LD_{95} dose of FM for 72 h, washed and performed barcode analysis on four replicates and lineage-resolved scRNA-seq analysis on the remaining four replicates (pooled) 12 d after treatment ('TP0.5'). Again, we observed a strong and consistent enrichment for HS clones (Fig. 3i,j and Extended Data Fig. 5g–j) and differential gene expression analysis revealed both HS and LS clusters at TP0.5 to exhibit apoptotic stress responses and upregulated drug efflux (*TAP1*, *CYCS*, *SLC25A4*; Fig. 3k). Gene set enrichment analysis (GSEA) of HS versus LS clusters at TP0.5 revealed that HS clones relied heavily on oxidative stress and DNA repair pathways for survival, while LS clones exhibited cellular senescence, inflammation and translational control mechanisms often employed to alleviate cellular damage (Extended Data Fig. 5k). At TP1, upon resolution of stasis, both populations displayed strong P53 and DNA damage responses (*PMAIP1*, *BBC3*, *GADD45A*) and upregulation of genes involved in response to reactive oxygen species (*ADA* and *MTHFD2*) (Supplementary Table 2), indicating a likely convergence on pathways enabling survival and resolution of cellular stress. Many genes involved in these DNA damage-response pathways were already expressed at higher levels in the HS versus LS subpopulations at TP0 (*FDXR*, *NINJ1*, *DDB2*), indicating potential priming of DNA repair pathways before therapy.

Clonal heterogeneity and transcriptomic identity

Genetic diversification is a hallmark of cancer which drives population heterogeneity and enables its evolution to more aggressive, malignant and resistant cancers. To understand the genetic differences defining the HG3 subpopulations, we performed whole-genome sequencing (WGS) on HS and LS subpopulations at TP0 and TP1. At TP0, we identified >5,000 unique mutations in each subpopulation. Of these, 27 LS and 36 HS mutations were nonsynonymous coding somatic single-nucleotide variants (sSNVs), with 14 LS and 19 HS mutations predicted to have functional impact, given that they are missense with elevated PolyPhen2 scores (>0.8 or nonsense, nonstop or splice site mutations (Fig. 4a and Supplementary Table 5). Of note, mutations in these genes, also with high PolyPhen2 scores, were found in 76 of 984 (7.7%) primary CLL samples previously characterized by whole-exome sequencing (Fig. 4a)[39]. One mutated gene, *TRIML2*, a p53 regulator previously reported to be associated with apoptotic response to oxidative stress and upregulated in oral cancer, was also found to rise in cancer cell fractions upon chemotherapy relapse (Extended Data Fig. 6a)[40–42].

Although HS clones exhibited a core transcriptional identity (that includes CXCR4/Wnt/Notch coexpression), we nonetheless observed substantial genetic heterogeneity in this subpopulation at TP0, with coding sSNVs present across a range of VAFs as clonal and subclonal events (Fig. 4b and Supplementary Table 6) and with varying degrees of clonal shift (>10% change in VAFs) following treatment. To evaluate this subpopulation heterogeneity at greater resolution, we performed clonal isolation of two lineages (direct descendants of their respective parental barcoded clones) that were equivalently highly abundant in the HS subpopulation at TP1 (lineage 1, 8.2%; lineage 2, 9.9%) (Fig. 4c). Using ClonMapper, we drove lineage-specific green fluorescent protein (GFP) expression of these lineages in the TP0 and TP1 HG3 cells and flow cytometrically isolated single GFP^+ cells. Three subclones from each lineage were expanded and characterized by WGS (Fig. 4d) and sSNVs were identified using the LS bulk population as a reference (Methods).

At baseline (TP0), we observed only 68.5% of sSNVs in the HS bulk population were shared by the two lineages at TP0, consistent with a picture that multiple additional lineages comprise the bulk population. Between the two lineages, ~41% of mutations were shared, indicating a likely common ancestor, but most mutations were unique to one lineage (Fig. 4e and Supplementary Table 7). Indeed, principal-component analysis on sSNVs found in the subclones and bulk populations showed that each lineage distinctly segregated from one another with proximal intra-lineage separation between TP0 and TP1 (Fig. 4f). Following exposure to therapy we observed increased diversification of the isolated lineages (~twofold increase in total and coding sSNVs) and bulk population (~fourfold increase), resulting in numerous unique, likely damaging subclonal mutations (Fig. 4g and Supplementary Table 8). This finding is reflected in primary tumors, as we observed elevated mutation burden in the exomes of primary CLL samples following relapse from chemotherapy ($P = 0.008$) (Extended Data Fig. 6b). sSNVs identified in the HS bulk population at TP1 were largely subclonal (VAF < 30%) as compared to TP0, consistent with the acquisition of de novo mutations (Fig. 4h). We further observed divergent genetic profiles within each lineage, with subclones of lineage 1, for example, sharing ~72.8% of sSNVs, whereas ~27.2% were unique and thus acquired during expansion from their initial barcoded ancestor (Fig. 4i). As with the bulk population, we observed a marked increase in the acquisition of new sSNVs in subclones following treatment (~twofold per clone). In sum, these findings reinforce the notion that the HS subpopulation consists of numerous genetically distinct lineages that are constantly diversifying during expansion and over the course of treatment.

The impact of mutation accumulation on the transcriptome has not been fully elucidated. While many studies center around the effects of individual or a small number of mutations on transcriptomic signatures, we sought to understand the impact of these numerous accumulating mutations on the overall transcriptional profile of each clone. To this end, we performed scRNA-seq of lineage 1 and 2 subclones from TP0 and TP1 (Fig. 5a; Methods), detecting 5,841 cells among a total of 12 subclones (mean 487 cells per subclone) with cells largely segregated on the basis of lineage identity and treatment status, recapitulating their clonal relatedness (Fig. 5b). To discern whether characteristic gene expression patterns existed among subclones, we utilized a multivariate classifier to learn the transcriptomic signatures of the six subclones within each time point. Data were divided into training and test sets and the latter was used to assess the clonal predicting accuracy of the models. The

models proved to be highly accurate, with weighted precision (positive predictive value) scores of 0.974 and 0.977 for TP0 and TP1, respectively (Fig. 5c,d). These results suggest that acquisition of genetic diversification contributes to measurable transcriptomic variation within these subclonal populations.

To discern whether these characteristic clonal transcriptomic signatures were stable over treatment course, we trained a new model to learn the transcriptomic signatures of the ten most abundant lineages from the TP1 scRNA-seq dataset and then subsequently assessed our ability to classify cells at TP0, TP0.5 and TP1 by their respective lineage identity (Fig. 3j). This ten-class model demonstrated high accuracy across all time points (weighted precision scores of 0.73 at TP0, 0.612 at TP0.5 and 0.827 at TP1) (Fig. 5e,f), supporting the idea that lineages have stable characteristic transcriptomic signatures that are conserved over treatment course. By ranking genes by their weighted coefficients from the trained model and visualizing the expression of the top 20 genes for each lineage across time points (Fig. 5g), we observed that most lineage-specific signatures consist of a constellation of gene expression values rather than expression of any single gene. Altogether, these observations link the accumulation of genetic alterations to transcriptomic variation and the development of characteristic subclonal signatures that are maintained over the course of expansion and treatment.

While great efforts have been made in measuring and identifying the impact of specific mutations in cancer, we have a poor sense of how mutations in aggregate affect the stability of the cell's gene regulatory network (GRN), the systems-level orchestration of gene expression that coordinates higher-level cellular function. Through this work, we demonstrate that substantial mutation accumulation can occur without catastrophic effects to the GRN, such that newly established subclonal signatures are maintained over time and even in the face of major stress such as chemotherapy exposure. Collectively, these findings support a model of clonal identity that is shaped by the balance of countervailing forces, namely genetic diversification and GRN stabilization of transcriptomic states.

Discussion

Technological advances over the last decade, from bulk genome to single-cell sequencing, have granted us insights into tumor heterogeneity at ever increasing resolution[7]. The next frontier of discovery naturally aims to uncover the adaptive processes that give rise to this heterogeneity. Herein we introduce ClonMapper, which can integrate the fate of individual cells with their genetic, epigenetic and transcriptomic characteristics to provide an unprecedented granular view of the process of clonal diversification and adaptation to chemotherapy, such as in this CLL model. In particular, we identify not only clonally distinct subpopulations and respective key signatures associated with differential responses to cytotoxic therapy, but also detect these same signatures in primary CLL samples following relapse from fludarabine-based chemotherapy and exposure to ibrutinib targeted therapy.

A key advantage of our platform is the ability to link clonal identity with respective fitness and transcriptomic profiles, allowing us to identify a 'HS' subset of clones, exhibiting

upregulated CXCR4, Wnt and Notch signaling, compared to a ‘LS’ subset. Rather than outright sensitivity or resistance, both HS and LS subpopulations exhibited immediate signs of cell toxicity upon treatment, followed by a prolonged period of stasis. HS clones were more viable than LS clones, however, and re-entered a proliferative state more rapidly. This range of drug response and recovery is distinct from the binary concept of resistance and we propose that gradations of ‘tolerance’ (low, intermediate, high) are more representative of the variable responses of heterogenous clones to chemotherapy.

The ability to isolate and co-culture HS and LS subpopulations enabled us to identify likely intercellular interactions that establish clonal equilibrium within the bulk population. We identified several ligand–receptor pairs between both populations that served as potential mediators of cell communication, including several driven by LGALS9, an important regulator of proliferation in CLL[29,35,36]. We have further detected paired LGALS9 and SORL1 expression in primary CLL cells that associate with LS and HS signatures. Our data support the notion that as abundance of LS cells diminished, their secretion of such ligands decreased below a threshold at which HS cell growth was no longer suppressed, resulting in accelerated expansion. The concept that competing cancer clones can exist in equilibrium, and that outside perturbation can disrupt this balance, is supported by observations in more complex patient settings, including in patients with CLL undergoing ‘watch and wait’ wherein clonal equilibrium can be maintained for decades before the need for treatment perturbs this balance and results in the rapid outgrowth of a minority of clones[25]. We further appreciate that such cell–cell interactions may be important in mediating therapeutic resistance in other diseases with aberrant CXCR4 expression, including Waldenstrom macroglobulinemia[43]. Thus, direct interactions between clones, rather than simply clonal competition for resources, may mediate this equilibrium and perturbing these interactions may provide new potential therapeutic opportunities.

A separate key insight of our study is the substantial genetic and transcriptomic heterogeneity present within each of the HS and LS subpopulations despite their distinct phenotypic profiles. Because ClonMapper enables us to map clonal identity to each clone’s respective genomic and transcriptomic profiles, we were able to: (1) clearly reveal the presence of diversification before and upon chemotherapeutic treatment at the bulk population and individual lineage level, with ~fourfold and ~twofold increase in unique sSNVs, respectively and (2) observe that lineages within the HS subpopulation, as well as subclones within a lineage, exhibit unique transcriptomic signatures that distinguish them over time and exposure to treatment, the likely result of ongoing genetic diversification that created thousands of unique subclones from a founding ancestral HS clone. Future functional studies will be necessary to establish causality of specific genetic and/or epigenetic events to characteristic changes in clonal transcription. We further note that a similar trend of mutation accumulation is evident in patients following relapse to therapy, with likely effects on resistance to subsequent therapy. While we are unable to directly conclude if this increase in mutation burden is specifically driven by mutagenic activity of the therapy, future studies using ClonMapper can be designed to discern such causality. Ultimately, our findings bring quantitative insights to the balance of two forces, accumulating genomic variation and transcriptomic cell-state stability. Together, they underscore the extent of genetic and transcriptomic diversification occurring in

heterogeneous tumor populations and serve as a glimpse of the source, rather than the results, of cancer's immense evolutionary potential.

While this first application was performed in the highly controlled setting of model cell line systems in vitro, future applications of ClonMapper can leverage the same workflows in primary human and murine tumor models, such as patient-derived xenografts or tumor allografts, to characterize the impact of tumor microenvironment and immune interactions on tumor evolution. Already, our findings highlight the variegating nature of cancer populations, a trait that drives immense evolutionary potential and thus makes these complex populations very difficult to treat. Until now, therapeutic approaches have been largely reactive in their efforts to target heterogeneous populations, a Sisyphean effort of repeatedly trying to outpace the vast evolutionary potential of cancer. Our insights draw us to the promise of new, proactive strategies like 'homogenization' which instead aim to purposefully address heterogeneity by funneling a diverse cancer population into a homogeneous, more targetable genotype or phenotype. Implementation of such an approach, however, requires tools to enable the systematic learning of the adaptive and evolutionary trajectories of cancer populations[44–50]. Integrative tools like ClonMapper are essential for the design and evaluation of homogenization strategies and as we have shown in the course of the current work, the ability to identify a highly drug-tolerant subpopulation and its corresponding gene expression state can inform rational and effective targeting strategies. Thus, by applying ClonMapper to the study of clonal dynamics in the setting of chemotherapy, we have gained deep insights into clonal dynamics and evolution, which can serve as a template for future discovery and development of new therapeutic strategies.

Methods

Further information on the research design is available in the Reporting Summary linked to this article. We describe patient cohorts in Supplementary Table 4f. Written informed consent was obtained from all patients. Samples were collected via protocols approved by institutional review boards or ethics and policy committees from the German CLL Study Group, Dana-Farber Cancer Institute and the CLL Research Consortium. All clinical trials were conducted in accordance with the Declaration of Helsinki and International Conference on Harmonization Guidelines for Good Clinical Practice.

Cell culture.

Cell lines utilized in this study were cultured as follows:

HG3 (DSMZ, ACC 765): RPMI 1640 (Gibco, 11875–093), 15% fetal bovine serum (FBS) (Sigma-Aldrich), 1% GlutaMAX (Gibco, 35050–061), 1% penicillin–streptomycin (PS) (Life Technologies).

REC1 (ATCC CRL-3004): RPMI 1640, 10% FBS, 1% PS.

HEK293T (ATCC CRL-3216): DMEM, 10% FBS, 1% PS.

Cells were incubated at 37 °C, 5% CO_2, passaged every 48 h and routinely tested for mycoplasma (VenorGeM Mycoplasma Detection kit; Sigma-Aldrich, MP0025).

ClonMapper library construction.

A modified version of CROP-seq-Guide-Puro[22] was assembled by replacing the puromycin resistance marker with a BFP marker to facilitate MOI measurement. PspLI (Thermo Fisher Scientific, FERFD0854) and MluI (Thermo Fisher Scientific, FERFD0564) were used to excise the puromycin resistance marker and a gBlock of the BFP fluorescent marker was cloned into the digested vector using the same restriction site overhangs. To increase transcript capture during RNA-seq workflows, the template switch sequence (AAGCAGTGGTATCAACGCAGAGTACATGGG) was cloned upstream of the hU6 promoter using restriction cloning with KflI (Thermo Fisher Scientific, FERFD2164) resulting in CROP-seq-BFP-TS-Guide (available up request). To assemble the high-complexity ClonMapper library, a 60-bp oligonucleotide containing a 20-bp random sequence (the barcode-sgRNA, BgL-BsmBI) and a reverse extension primer (RevExt-BgL-BsmBI) were ordered from Integrated DNA Technologies:

BgL-BsmBI:

GAGCCTCGTCTCCCACCGNNNNNNNNNNNNNNNNNNNNGTTTTGA GACGCATGCTGCA.

RevExt-BgL-BsmBI:

TGCAGCATGCGTCTCAAAAC.

The following extension reaction was performed to generate double-stranded barcode-sgRNA oligonucleotides: 10 µl Q5 PCR Reaction Buffer, 1 µl 10 mM dNTP mix, 1 µl of 100 µM BgL-BsmBI barcoded template, 2 µl of 100 µM RevExt-BgL-BsmBI reverse extension primer, 0.5 µl Q5 polymerase and 35.5 µl water and incubated as follows: 98 °C for 2 min, 10× (65 °C for 30 s, 72 °C for 10 s), 72 °C for 2 min, hold at 4 °C. The double-stranded barcode-sgRNA oligonucleotide was purified using a QIAquick PCR Purification kit (QIAGEN, 28104). The double-stranded product contains two BsmBI sites that, upon digestion, generate complimentary overhangs for ligation into CROP-seq-BFP-TS-Guide. Then, 1.5 µg of BsmBI digested (Thermo Fisher Scientific, FERFD0454) CROP-seq-BFP-TS-Guide vector was entered into a Golden Gate assembly reaction with the double-stranded barcode-sgRNA insert at a molar ratio of 1:5 and cycled 100× (42 °C for 2 min and 16 °C for 5 min). This reaction was then purified and concentrated in 12 µl water using the DNA Clean & Concentrator kit (Zymo, D4033) and transformed into electrocompetent SURE 2 cells (Agilent, 200152). Transformants were inoculated into 500 ml of 2 × YT medium containing 100 µg ml^{-1} carbenicillin and incubated overnight at 37 °C. Bacterial cells were pelleted by centrifugation at 6,000 RCF at 4 °C for 15 min and plasmid DNA was extracted using a QIAGEN Plasmid Plus Midi kit (QIAGEN, 12943).

Next-generation barcode sequencing of cell populations.

ClonMapper plasmid sequencing.—Barcode sequences were amplified from 50 ng of plasmid DNA by PCR using primers containing flanking barcode annealing regions and Illumina adaptor/6-bp index sequences and underwent single-end sequencing for 75 cycles using an Illumina NextSeq platform. To estimate the total number of unique barcodes in the

plasmid pool, we used a fourth-order polynomial regression to fit the sampled data. This function plateaus at ~7.6×10^7 barcodes, corresponding to the projected number of unique barcodes present in the plasmid pool.

Sequencing of barcoded cell populations.—Genomic DNA (gDNA) was extracted from pre- and post-treated cell populations using the DNeasy Blood & Tissue kit (QIAGEN, 69504). To sequence the barcoded populations, 2 μg of gDNA was used per PCR reaction in eight parallel PCR reactions per sample (total of 36 μg gDNA for TP0 and 16 μg for TP1 replicates). All reactions were purified using a 1.8× AMPure XP bead cleanup (Beckman Coulter, A63880) and sequenced on an Illumina NextSeq platform by single-end sequencing for 75 cycles.

Barcode sequencing analysis.

Raw barcode sequencing data were processed with cutadapt (v.1.18) to identify reads containing barcodes and trim the 25-bp flanking adaptor sequences (5′: ATCTTGTGGAAAGGACGAAACACCG, 3′: GTTTTAGAGCTAGAAATAGCAAGTT). To accommodate barcode reads with sequencing errors, flanking sequence identification was permitted a maximum expected error of 0.1. Positively identified and trimmed barcode sequences of 20 bp in length with no ambiguous (N) bases were filtered for read quality (requiring all bases to have a minimum PHRED of 20). To compensate for amplification artifacts and sequencing error, trimmed barcode sequences were tightly clustered across samples using Starcode (v.1.3), implementing a message-passing algorithm with a maximum Levenshtein distance of 1 and minimum radius size of 3. Remaining singleton barcodes that were not clustered were removed, resulting in a median of 1.13×10^7 barcode sequences per post-treatment sample (min 9.7×10^6, max 1.35×10^7) and 2.06×10^8 sequences for the TP0 sample. Code to perform this process can be found at https://github.com/brocklab/cashier.

Lentivirus production of barcode-sgRNA libraries.

HEK293T cells were plated in a six-well plate at 600,000 cells per well in 3 ml antibiotic-free DMEM with 10% FBS. When cells reached 70–80% confluency, they were transfected with 9 μl of Lipofectamine 2000 (Thermo Fisher Scientific, 11668027), 1.5 μg per well PsPax2 (Addgene, 12260), 0.4 μg per well VSV-G (Addgene, 8454) and 2.5 μg of the ClonMapper library and medium was exchanged for 3 ml of DMEM + 20% FBS at 24 h. Medium containing viral particles was collected at 48 and 72 h post-transfection, pooled, filtered through a 0.45-μm Nalgene syringe filter (Thermo Fisher Scientific, 723–2945) coupled to a 10-ml syringe and concentrated using a 50 ml size-exclusion column (Vivaspin20 30,000 MWCO PES, Sartorius, VS2022) by centrifugation at 2,200 RCF at 4 °C for 2 h. Concentrated virus was removed from the upper chamber and stored at −80 °C in 50-μl aliquots for later use.

Lentiviral transduction with barcode-sgRNA libraries.

Cells were seeded in a six-well plate at 2×10^6 cells per well in 3 ml culture medium and transduced with the ClonMapper lentivirus pool using 0.8 μg ml^{-1} polybrene and centrifuged at 500 RCF for 1.5 h (HG3) or 0.5 μg ml^{-1} polybrene and centrifuged at 1,000

RCF (REC1), then incubated at 37 °C for 16 h before exchanging for fresh culture medium. To reduce the likelihood that two viral particles enter a single cell, the MOI was kept below 0.1. After 48 h, 1.2×10^6 BFP^+ HG3 cells were isolated by FACS (analyzed using FlowJo v.10.6.2), cultured for 27 d to 2.4×10^8 cells with a final diversity of 13,648 unique barcodes and cryopreserved in 12 vials of 2×10^7 cells each. Then, 1.2×10^6 REC1 cells were isolated by FACS after 48 h, cultured for 21 d to 1.6×10^8 cells with a final diversity of 643,541 unique barcodes and cryopreserved in 12 vials of 1.3×10^7 cells each.

Chemotherapy-resistance studies.

Dose–response curves were established by plating 200,000 HG3 or REC1 cells in 200 μl culture medium and treating with either fludarabine (Selleckem, S1491), mafosfamide (Santa Cruz, SC-211761), doxorubicin (Selleckem, S1208) or vincristine (Selleckem, S1241) at a range of 50 to 0.02 μM (1:2 dilutions). An LD_{50} was determined for each drug and a second set of dose–response curves were performed using an FM or mafosfamide/doxorubicin/vincristine combination at fixed concentration ratios (as determined by the respective LD_{50} drug doses) ranging from 50 to 0.02 μM. After 72 h, viability was determined using CellTiter-Glo (Promega, G7570). An LD_{95} dose at 72 h was selected to ensure stringent selection of the HG3 population. Given that LD_{95} and LD_{75} doses at 72 h of the mafosfamide/doxorubicin/vincristine combination resulted in substantial cell death and nonviable REC1 cultures beyond 72 h, we selected an LD_{50} dose resulting in 95% cell death at 6 d post-treatment. Before treatment of the barcoded populations, ~1.5×10^7 viable cells were thawed and expanded to a final cell count of 1.2×10^8 and 2.4×10^8 (for REC1 and HG3, respectively). Then, 1×10^7 barcoded cells were plated in eight parallel T25 flasks in 10 ml drug medium for 72 h, washed twice with PBS and re-plated in 10 ml drug-free medium. Upon outgrowth, cells were either pelleted for gDNA extraction or cryopreserved in FBS + 10% dimethylsulfoxide for future analysis.

Chemotherapy and inhibitor combination treatment studies.

Cells from the HG3 HS subpopulation ($CXCR4^+$ FACS sorted) were exposed to subcytotoxic concentrations of inhibitors: AMD3100 (Selleckem, S8030), LGK-974 (Selleckem, S7143) and DAPT (Selleckem, S2215). Viability was determined by performing 96-h dose–response cell viability assays with inhibitor concentrations ranging from 250 μM to 100 nM. Combination treatments were performed using 24-h pre-treatment with the inhibitor of interest, followed by a 72-h incubation with 3 μM FM. Cell viability was determined using CellTiter-Glo. These experiments were performed in biological triplicates.

Sample and library preparation for scRNA-seq.

HG3 cell line.—Cryopreserved samples were thawed, pelleted and resuspended in cold PBS–0.04% BSA. 100,000 viable BFP^+ cells were sorted into cold PBS–0.04% BSA, resuspended at 1×10^6 cells per ml and barcoded with a 10X Chromium Controller (10X Genomics) by reverse transcription of RNA from individual cells. Sequencing libraries were constructed using a Chromium Single Cell 3′ v2 reagent kit according to the manufacturer's instructions (10X Genomics, PN-120267). Sequencing was performed on an Illumina NovaSeq platform (paired-end read 1, 26 cycles; read 2, 98 cycles; index 1, 8 cycles)

according to the manufacturer's recommendations (10X Genomics). For assessment of subclonal transcriptional identity parental and daughter TP0 and TP1 clones were expanded and 0.5×10^6 cells from each were washed and incubated with 5 µl Trustain FcX for 10 min on ice (BioLegend, 422302). To minimize batch effects, all 12 samples were prepared in parallel and subjected to cell hashing by incubating with a unique combination of two TotalSeq-C antibodies for 3 min on ice (BioLegend, TotaISeq-C0251 anti-human Hashtag 1 (394661), TotaISeq-C0252 anti-human Hashtag 2 (394663), TotaISeq-C0253 anti-human Hashtag 3 (394665), TotaISeq-C0254 anti-human Hashtag 4 (394667), TotaISeq-C0255 anti-human Hashtag 5 (394669), TotaISeq-C0256 anti-human Hashtag 6 (394671), TotaISeq-C0257 anti-human Hashtag 7 (394673), TotaISeq-C0258 anti-human Hashtag 8 (394675), TotaISeq-C0259 anti-human Hashtag 9 (394677) and TotaISeq-C0260 anti-human Hashtag 10 (394679)). Cells were washed twice in cold PBS–0.04% BSA, normalized, mixed in equimolar amounts and resuspended to 1×10^6 ml^{-1}. Then, 20,000 cells were loaded onto a single 10X channel using the Single Cell 5′ reagent kit. Primary CLL cells from four fludarabine and two ibrutinib-treated patients were obtained from the CLL Research Consortium tissue core, housed at the University of California San Diego. All patients provided informed consent and study protocols were institutional review board approved (University of California San Diego Human Research Protections Program, Project 080918). Viable CD19$^+$ cells were purified by FACS or B-cell selection (BioLegend, 480082). Cells were resuspended into PBS–0.04% BSA at 1×10^6 cells ml^{-1} and processed using the single cell 3′ v2 (fludarabine) or 5′ (ibrutinib) kit according to manufacturer's instructions.

scRNA-seq analysis.

Sequencing data were aligned to the GRCh38 genome using Cell Ranger (v.3.1.0). For hashed samples, the data were demultiplexed using the HTODemux function from the Seurat package (v.3.2.0). Filtered matrices produced by Cell Ranger were loaded into scanpy (v.1.4.5). Cells were annotated by sample and clonal membership, where applicable. Cutoffs for further filtering were determined separately for each dataset, as is recommended practice[51].

Filtering the HG3 dataset of all lineages.—Only cells meeting the following requirements were retained for further analysis: (1) a minimum of 4,000 transcript counts, (2) a maximum of 8% of counts attributed to mitochondrial genes and (3) a minimum of 750 genes detected.

Filtering the hashed recalled subclones dataset.—Only cells meeting the following requirements were retained for further analysis: (1) a minimum of 5,000 transcript counts, (2) a maximum of 8% of counts attributed to mitochondrial genes and (3) a minimum of 2,500 genes detected.

Filtering the fludarabine-treated patient dataset.—Only cells meeting the following requirements were retained for further analysis: (1) a minimum of 600 transcript counts, (2) a maximum of 8% of counts attributed to mitochondrial genes and (3) a minimum of 200 genes detected.

Filtering the ibrutinib-treated patient dataset.—Only cells meeting the following requirements were retained for further analysis: (1) a minimum of 500 transcript counts, (2) a maximum of 10% of counts attributed to mitochondrial genes and (3) a minimum of 200 genes detected.

Normalization.—Normalization was conducted as follows (based on recommendations from studies comparing several normalization techniques)[51–53]: preliminary clustering of cells by constructing a nearest network graph and using scanpy's implementation of Leiden community detection (v.0.7.0)[54], calculating size factors using the R package scran (v.1.10.2)[55] and dividing counts by the respective size factor assigned to each cell. Normalized counts were transformed by adding a pseudocount of 1 and taking the natural log.

For regressing out cell cycle expression signatures, cells were assigned S phase and G2M-phase scores[56] and run through scanpy's regress_out function with both score categories as inputs.

For assigning cells to transcriptomically distinct clusters, every gene was centered to a mean of zero and principal-component analysis was performed. The top 40 principal components were used to construct a nearest-neighbor graph. Cells were then assigned to clusters using the Leiden algorithm.

Removal of non-CLL cells from primary samples.—A first round of clustering was performed. Cells in $CD19^{+}CD5^{+}$ clusters were retained and cells belonging to the remaining clusters were removed for subsequent analyses and plotting. A second round of clustering was performed on the new filtered CLL dataset.

Differential expression analysis.—Differential gene expression was assessed using scanpy's rank_genes_groups function with Wilcoxon tests to generate *P* values. A truncated normal (TN) test was also performed on all genes, yielding a second set of *P* values. Each set of *P* values was adjusted using Benjamini–Hochberg false discovery rate correction, producing *q*-values. A 'consensus' *q*-value was calculated by selecting the larger of the two *q*-values per gene. A gene was considered statistically significant if its consensus *q*-value was <0.05. Pathway analysis was performed using the MetaCore software (Clarivate Analytics v.2020) or GSEA[57].

Ligand–receptor analysis.—CellPhoneDB (v.2.1.1) was used for detecting receptor–ligand interactions between populations of cells[28]. The iterations parameter was set to 100, the threshold for proportion of cells expressing the specific ligand–receptor was set to 0.1 and the *P* value threshold was set to <0.05.

Clonal assignment in scRNA-seq data.

Cell- and UMI-labeled sequencing data in the unaligned BAM file reads from the 10X Genomics Cell Ranger pipeline were used to identify clone barcodes for each cell. Cutadapt was used to identify adaptors flanking the barcode sequences of interest (5′: ATCTTGTGGAAAGGACGAAACACCG, 3′: GTTTTAGAGCTAGAAATAGCAAGTT).

Owing to the extreme 3′ nature of the 10X data, only the 3′ adaptor and a minimum of 10 bp of the barcode portion of the read were required for barcode identification. To correct for processed barcode reads <20 bp, all barcode reads within a particular cell were compared. If an exact match to the 3′ end of a longer barcode read was found, the read was extended with the matching sequence. Length-corrected clone barcode reads were omitted if not found in the whitelist of known clone barcodes from the original TP0 population. For each cell, UMI-replicate reads were resolved by taking the mode barcode read for each UMI. To resolve instances where multiple clone reads in a particular cell were observed, we annotated cells by their maximum clone barcode given no other clone barcode surpassing 50% of the read count of the maximum clone. Clonal reads from cells with many major clone barcodes that could not be resolved were removed. Of 1,982 total cells analyzed at TP0, we identified 823 unique barcodes and of 6,993 total cells at TP1, we identified 413 unique barcodes.

Western blot analysis.

For all immunoblot assays, 5×10^6 cells were collected and lysed in 100 μl RIPA buffer (Sigma, R0278) supplemented with protease and phosphatase inhibitors (Sigma, 11836153001 and 4906845001). Protein concentration was determined using the Pierce bicinchoninic acid assay (Life Technologies, 23225) and 40 μg protein were loaded per lane of a 4–12% Bis-Tris gel (Life Technologies, NP0323). Proteins were transferred onto polyvinylidene fluoride membranes (Life Technologies, IB24002) using an iBlot 2 at 20 V for 8 min and subsequently incubated in 5% BSA, 1× TBST (Teknova, T9511) before adding one of the following primary antibodies (all from Cell Signaling Technology) at a 1:1,000 dilution for 1 h at room temperature: β-catenin (9562), TCF1/TCF7 (2203), Frizzled6 (5158), Notch1 (3608), Jagged1 (70109), RBPSUH (5313), p44/42 ERK1/2 (4370), p217/221 MEK1/2 (9121), KDM6A (33510) and GAPDH (2118). Membranes were washed in TBST before incubation with a 1:2,000 dilution of goat anti-rabbit HRP-conjugated IgG antibody (Thermo Fisher Scientific, 12348MI) for 1 h at room temperature. Membranes were visualized using SuperSignal West Pico Chemiluminescent Substrate (Life Technologies, 34580) and quantified using a BioRad ChemiDoc. All blots were performed in triplicate.

Bulk RNA-seq.

Forty million HG3 cells from the bulk TP0 and TP1 populations were co-stained using human CXCR4-APC (1:10 dilution, Thermo Fisher Scientific, FAB173A) and CD18-PE (1:100 dilution, BioLegend, 302108) conjugated antibodies. CXCR4 and CD18 single-positive cells (representative of HS and LS subpopulations) were isolated using a BD FACSAria II. Three biological replicates were collected over a 2-week period. RNA was isolated using a RNeasy Micro kit (QIAGEN, 74004) and RNA quality was verified using an Agilent Bioanalyzer Pico Eukaryote chip. Then, 400 ng of total RNA was used as input for library preparation using the KAPA stranded mRNA HyperPrep kit (KAPA Biosystems, KK8580) per the manufacturer's instructions. Library yields and sizes were confirmed using an Agilent High Sensitivity DNA kit (Agilent Technologies, 5067–4626) on an Agilent 2100 bioanalyzer. Libraries were pooled in equimolar ratios and sequenced using a 75-cycle high output kit on an Illumina NextSeq 500 system. Bulk RNA sequencing data

for 23 fludarabine and 13 ibrutinib-treated patients were from dbGaP (accession numbers phs000922.v1.p1 and phs001431.v1.p, respectively).

Bulk RNA-seq analysis of HG3 samples.

Bulk RNA-seq data were processed by trimming Illumina sequencing adaptors with cutadapt and aligned to human genome build GRCh38 with bowtie2 (v.2.3.4.1). Featurecounts was used to count exon-mapping reads in gene transcripts in the Ensemble human reference annotation v.77 (ref. [58]). A sample-by-gene count matrix was derived from 'expected count' values in the RSEM outputs. Genes with <10 total counts across all samples were filtered out. Differential expression between the $CXCR4^{hi}$ and $CD18^{hi}$ populations was performed using DESeq2 with time point as a covariate.

Assigning high survivorship, low survivorship, Wnt, Notch and CXCR4 signaling scores to transcriptomes.

HS and LS genes were defined as those having a q-value < 0.05 and $\log_2(FC) > 0.5$ or $\log_2(FC) < 0.5$, respectively, from the HG3 bulk RNA-seq differential expression results. These two gene sets as well as the Wnt, Notch and CXCR4 signaling gene sets from the HG3 bulk RNA-seq results were utilized for assigning scores to transcriptomes (of bulk primary samples or single cells). Scores were assigned using scanpy's tl.score_genes() function.

Bulk RNA-seq analysis of patient samples.

Patient RNA-seq data were processed using the gtex-pipeline (v.9) (https://github.com/broadinstitute/gtex-pipeline) and aligned to a GENCODE (v.19)-annotated hg19 reference using STAR (v.2.6.1). RNA-SeQC (v.2.3.6; https://github.com/getzlab/rnaseqc) was used to produce TPM expression values. To measure the significance of change in scores from pre- to post-treatment, a one-sided t-test was performed. Change in tumor mutation burden in 71 patients with relapsed CLL (treated with fludarabine/cyclophosphamide or fludarabine/cyclophosphamide/rituximab) was calculated using a two-tailed paired t-test ($t = -2.72$, $P = 0.008$).

Fast-ATAC sequencing.

We employed the Fast-ATAC sequencing protocol as previously described[59]. Then, 50,000 HG3 cells were sorted into 1.5-ml microcentrifuge tubes and pelleted by centrifugation at 500 RCF for 5 min at 4 °C in a pre-cooled fixed-angle centrifuge. All supernatant was removed and 50 µl of transposase mixture (25 µl TD Buffer, 2.5 µl TDE1, 0.5 µl digitonin and 22 µl nuclease-free H20) (Illumina, 15027866 and 15027864; Promega G9441) was added to the cells and mixed gently by pipetting. Transposition reactions were incubated at 37 °C for 30 min and purified using a QIAGEN MinElute Reaction Cleanup kit (QIAGEN, 28204). Transposed fragments were amplified for 6–8 cycles as needed using Nextera i7 and i5 primers. Libraries were purified using double-sided bead purification (0.5× followed by 1.3× volume of AMPure XP beads) and eluted in nuclease-free H20. Library size and concentrations were confirmed on an Agilent High Sensitivity DNA Bioanalysis chip and by qPCR using a KAPA Library Quantification kit (Roche, KK4824). All libraries were

sequenced using paired-end, dual-index sequencing on a 75-cycle high output kit on an Illumina NextSeq 500 system.

ATAC-seq data analysis.

ATAC-seq data were processed uniformly with the Cromwell-based pipeline employed by the ENCODE (https://github.com/ENCODE-DCC/atac-seq-pipeline) project to reproducibly align and call irreproducible discovery rate (IDR)-resolved peaks in our samples. Relevant pipeline settings were as follows: Paired-end: true, auto_detect_adaptor enabled, idr and xcor enabled with idr_threshold of 0.05. Blacklist-filtered.narrowPeaks files from MACS2 were processed based on an available pipeline (https://rockefelleruniversity.github.io/RU_ATAC_Workshop.html) to generate a set of nonredundant peaks present in at least two of the samples. Counts of fragments in nucleosome-free regions for these peaks were calculated by running the featureCounts function from the Rsubread (v.2.0.1) package with the max fragment length set to 100 (ref. [60]). Differential accessibility between HS and LS subpopulations was performed using DESeq2 with time point as a covariate. The esATAC (v.1.8.0) package was used to annotate peaks to types of regions in the hg38 genome. Peaks that were significantly more accessible (q-value < 0.05) in the HS subpopulation were compared against peaks that were significantly more accessible in the LS subpopulation and analyzed for differences in transcription factor–binding motif enrichment. Input data were uploaded to the Analysis of Motif Enrichment (5.0.5) web portal (http://meme-suite.org/meme_5.0.5/tools/ame) and the JASPAR nonredundant vertebrates database was selected as the source for the position weight matrix motif.

Mathematical modeling of growth dynamics.

Growth rates of individual lineages were estimated by combining barcode sampling data with cell count measurements of the bulk population taken at 48, 96 and 144 h after thawing a vial of barcoded HG3 cells. Only barcodes observed in all three time points were considered for analysis to increase confidence in the growth rate estimate (thus removing 0.44%, 0.69% and 2% of unique barcodes from the 48, 96 and 144-h time points, respectively; the corresponding number of cells were also subtracted). The 48-h time point was considered as $t = 0$ as it was the first observed measurement. Barcode abundance (the number of reads for the ith barcode divided by the number of total reads) was calculated for each time point. To estimate the number of cells within each lineage, total cell number per time point was multiplied by the barcode abundance for the consensus set of barcodes. To generate a distribution of fitted growth rates, the estimated number of cells per barcode over time were fitted to the following exponential growth equation:

$$N_i(t) = N_{0i}e^{g_i t}$$

where i describes the ith barcode, N_{0i} is the calculated estimate of the number of cells in the ith barcoded population at the first measured time point and g_i is the growth constant assigned to the ith barcode based on the observed cell number in time data ($N(t)$) estimates. Each barcode was designated as an HS or LS clone based on whether the majority of reads correspond to a barcode present in the $CXCR4^+$ (HS) versus $CD18^+$ (LS)

flow-sorted populations. All other growth dynamic comparisons were performed by using the exponential growth equation above and fitting to observed cell counts from flow-sorted populations. All analyses were performed using MATLAB 2019a. Code is available at https://github.com/kej1993johnson/Lineage_transitions.

Co-culture experiments.

HS and LS subpopulations were fluorescently labeled using the Sleeping Beauty transposon system by nucleofecting 1.5×10^7 cells each with 15 μg of pCMV(CAT)T7-SB100 (transposase, Addgene, 34879) and 15 μg of either the pSBbi-GN (Addgene, 60517) or pSBbi-RN plasmids (Addgene, 60519) (transposons), respectively, to generate constitutively expressing GFP^+ and RFP^+ populations. Successfully transposed cells were purified twice by FACS 13 and 16 d post-nucleofection. Then, 500,000 fluorescently labeled cells were plated in 3 ml medium per well in a six-well plate, with each well containing a specified ratio of LS versus HS cells (100% LS cells, 90% LS versus 10% HS, 80% LS versus 20% HS, 70% LS versus 30% HS). Flow cytometric analysis was performed at 24-h intervals for 14 d (with continuous and even passaging at every measured time point) to determine the percentage of HS (GFP^+) versus LS (RFP^+) cells present within each sample over time. The initial and final proportion of HS and LS cells in each sample were used to compute percentage change in proportion over time. Samples were segregated into two groups; one with starting proportions reflecting that of the bulk population (<20% HS) and another with decreasing representation of the LS cells (>30% HS). A two-sample *t*-test comparing the percentage increase in proportion of HS cells between groups revealed a significance of $P = 0.023$. All experiments were performed in triplicate.

DNA damage and apoptosis assays.

Annexin V apoptosis assay.—HG3 cells were treated with an LD_{50} dose of FM (2 μM fludarabine + 2 μM mafosfamide) and plated on 0.01% polyornithine-coated plates (Sigma-Aldrich, P4957) in the presence of IncuCyte annexin V red reagent (Essen Bioscience, 4641; 1:200 dilution). Cells were imaged every 2 h for the length of time indicated using an IncuCyte S3 Live-Cell Analysis System (Essen Bioscience). Apoptosis index was determined by dividing red fluorescence area by total phase cell area and assay was conducted in duplicate.

BH3 profiling.—BH3 profiling was performed as described previously[61]. Cells were suspended in MEB2 buffer (150 mM mannitol, 10 mM HEPES-KOH (pH 7.5), 150 mM KCI, 1 mM EGTA, 1 mM EDTA, 0.1% BSA and 5 mM succinate), added to a 384-well plate and incubated at 25 °C for 60 min combined with different BH3-only peptides in 0.001% digitonin for permeabilization. Cell were fixed with 4% formaldehyde for 10 min at room temperature followed by neutralization by adding N2 buffer (1.7 M Tris base, 1.25 M glycine, pH 9.1). Cells were stained overnight at 4 °C with Hoechst 33342 (H3570, Invitrogen) and anti-cytochrome *c*-Alexa Fluor 488 (6H2.B4; 612308) and analyzed using an Intellicyt iQue flow cytometer to determine the rate of loss of cytochrome c in response to each BH3 peptide. Assays were conducted in triplicates and *P* values were calculated using the Holm–Sidak method for multiple comparison *t*-tests, with $\alpha = 0.05$. Dimethylsulfoxide

and alamethicin were used as negative and positive controls for cytochrome c release. All assays were performed in triplicate.

WGS.

gDNA was extracted from HG3 cells using a QIAGEN DNeasy Blood & Tissue kit and 350 ng were used as input into DNA fragmentation using a Covaris focused-ultrasonicator targeting 385-bp fragments. Additional size selection was performed using a solid-phase reversible immobilization cleanup. Library preparation was performed using a KAPA HyperPrep kit without amplification module (KAPA, KK8505) and quantified using qPCR (Kapa Library Quantification kit). Libraries were pooled and sequenced at 60× coverage on a NovaSeq 6000 S4 flowcell using 151-bp paired-end reads. Standard quality control metrics (error rates, percentage of reads passing filter, total Gb produced) were analyzed before downstream analysis. Output from Illumina software was processed by the Picard data-processing pipeline to yield BAM files containing demultiplexed, aggregated, aligned reads with well-calibrated quality scores.

WGS analysis.

High-confidence somatic mutation calls were made by applying MuTect[62], MuTect2[63] and Strelka[64] to sequencing data from HS and LS subpopulations, first using the HS sample as tumor and LS as matched normal and subsequently using the LS sample as tumor and HS sample as matched normal, such that we could call unique somatic mutations in both samples. Somatic mutation calls for all isolated lineages were made by using LS bulk as matched normal. High-confidence structural variant calls were similarly made, first using Manta[65], dRanger and SvABA[66] independently, then combining all results and running Breakpointer[67]. Commonly occurring germline variants and sequencing artifacts were filtered out using a panel of normal human samples. A realignment filter was applied to remove remaining false-positive calls. Estimation of allelic frequency of each SNV was determined by dividing the variant allele read counts by the summation of the variant allele and total reference allele read counts.

Clonal isolation.

Two lineages were isolated from the HS bulk populations at TP0 and TP1 with the following barcode identities:

Barcode 1: TAACGTTAGTGACAGACCTT

Barcode 2: TTTCACTTCTGGTTAGCCGT

Recall plasmids were constructed by using type IIS restriction cloning to introduce barcode-specific landing pad sequences into the Recall Plasmid upstream of a GFP reporter (available upon request)[19]. To isolate each lineage, a total of 1.8×10^7 cells from either the TP0 or TP1 bulk population were divided into six nucleofection reactions of 3×10^6 cells and co-nucleofected with 5 μg of recall plasmid and 4 μg of a plasmid containing the transcriptional activator variant dCas9-VPR (using an Amaxa 4D nucleofector, protocol

DN-100). GFP$^+$ cells were single-cell sorted 72 h later into 96-well plates and expanded for 3 weeks before validation (barcode amplification and sequencing).

Supervised machine learning of lineage and subclone transcriptomic identity.

Cells belonging to the ten most abundant lineages at TP1 were randomly assigned to separate training (75% of cells) and testing (25% of cells) datasets, stratified by lineage. The following two classifiers from sklearn (v.0.22.1)[68] were fitted to the training data: RidgeClassifierCV (α = (100,1e3,1e4), class_weight = 'balanced') and DummyClassifier (strategy = 'stratified'). The fitted classifiers were then individually tested on the TP1 testing set and on all cells belonging to the same ten lineages at TP0 and TP0.5. The precision_score (average = 'weighted') function from sklearn was used to assess performance of the classifiers. The top 20 genes contributing to lineage assignment were determined by extracting the ridge classifier coefficients assigned to each gene and ranking them. For the dataset of the hashed recalled subclones, the training and testing was done separately for each time point as subclones are not shared across time points.

Statistics and reproducibility.

Statistical analysis was performed using GraphPad Prism 6. For comparison of two groups, *P* values were calculated using a two-sample Student's *t*-test (after confirming homogeneity of variance) or Mann–Whitney *U*-test on independent variables or fitted growth rates. For scRNA-seq analysis, comparison of differential gene expression between clusters was performed as described above using the truncated normal test. For comparisons between scRNA-seq samples (for example all TP0 cells versus all TP1 cells), *P* values were calculated using a two-sided Wilcoxon rank-sum test. For DESeq2 analyses, *P* values were derived from a two-sided Wald test. Analysis of motif enrichment *P* values were derived using a one-tailed Fisher's exact text and adjusted using a Bonferroni correction. *E*-values were derived from the Bonferroni-adjusted *P* values by multiplying by the number of motifs in the motifs reference. All *q*-values described in this study were derived using the Benjamini–Hochberg procedure unless stated otherwise. Eight parallel replicates of 1 $\times$ 10^7 million barcoded cells each were chosen to enable observations of clonal patterns with ~800-fold (HG3) and ~15-fold (REC1) representation per barcode. The use of eight replicates enabled us to robustly identify constantly surviving clones (at our clonal survival rate ~20%, the odds of a clone surviving in all eight replicates by chance is 0.2 $\times$ 10^8 (0.000256%)). For patient data, no sample size calculation was performed; sample sizes were determined by availability and were sufficient to detect transcriptomic changes from pre-fludarabine and relapse bulk RNA-seq data. Patient data analysis was observational; randomization was not used. Blinding is not relevant to this study because group allocation was not performed.

Reporting Summary.

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

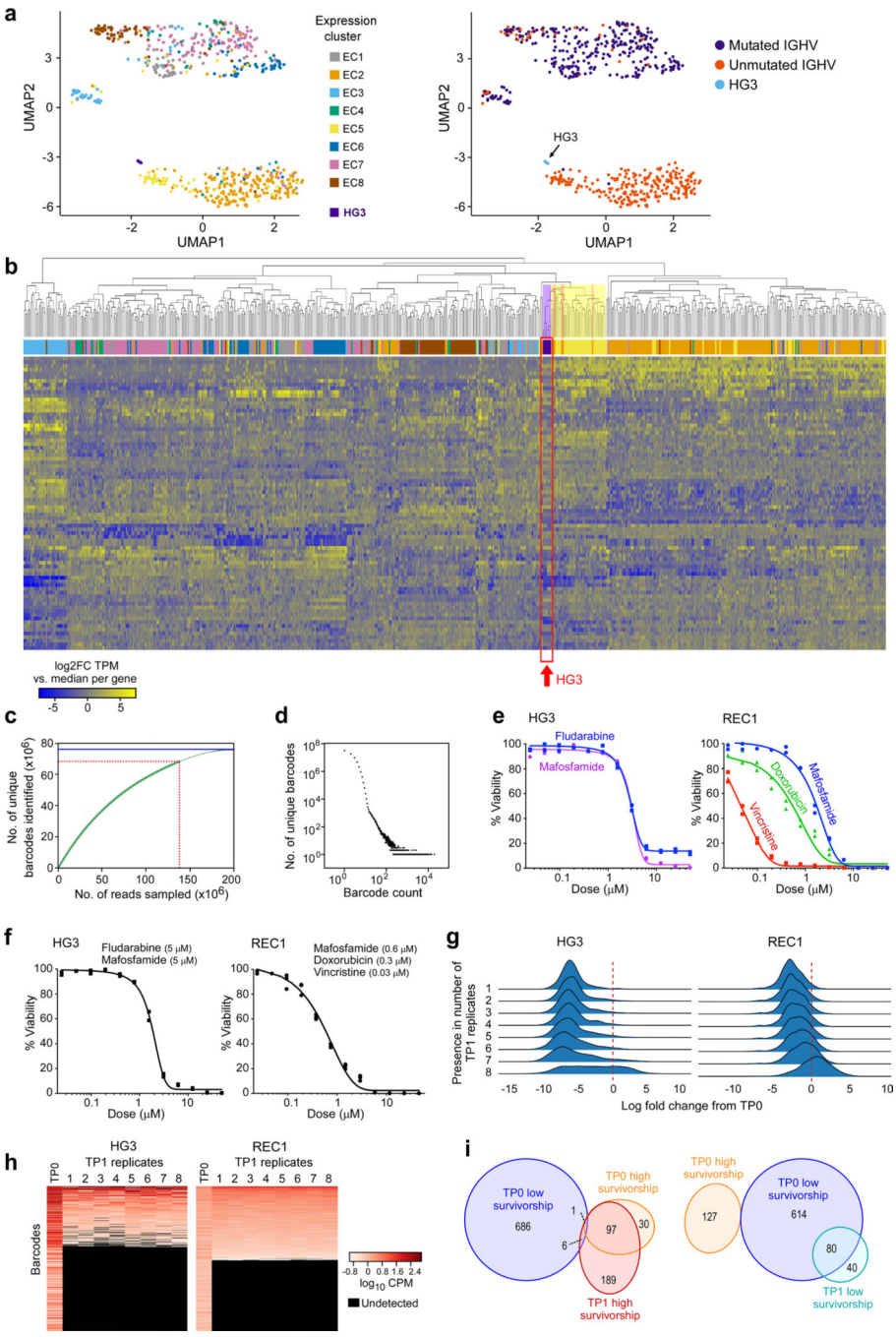
Data availability

All HG3 epigenomic sequencing data that support the findings of this study have been deposited in the Gene Expression Omnibus under accession code GSE151431. Previously published whole-exome sequencing data for fludarabine and ibrutinib-treated patients were from dbGaP with accession numbers phs000922. v1.p1 and phs001431.v1.p, respectively. Previously published scRNA-seq data for ibrutinib-treated patients are available from dbGaP with accession number phs2335.v1. All other data supporting the findings of this study are available from the corresponding authors on reasonable request. Source data are provided with this paper.

Code availability

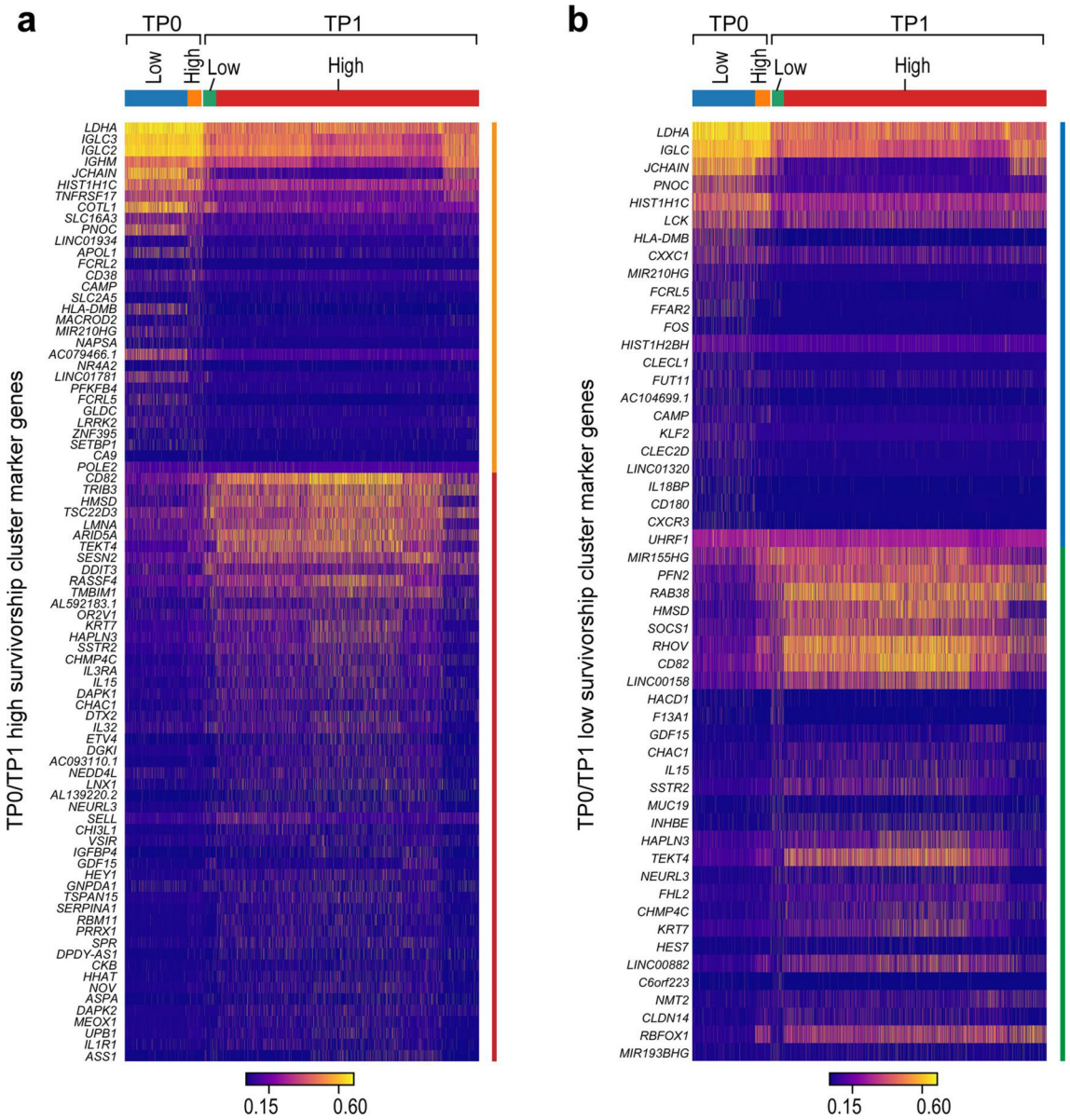
The single-cell RNA data were processed using Cell Ranger v.3.1 (https://www.10xgenomics.com/) and analyzed with the Python package scanpy v.1.4.5 (https://scanpy.readthedocs.io/en/stable/). Barcode sequence processing was conducted using our cashier algorithm (v.0.1) (https://github.com/brocklab/cashier). Code for modeling growth dynamics is also available on GitHub (https://github.com/kej1993johnson/Lineage_transitions). Any additional code is available from the corresponding authors upon request.

Extended Data

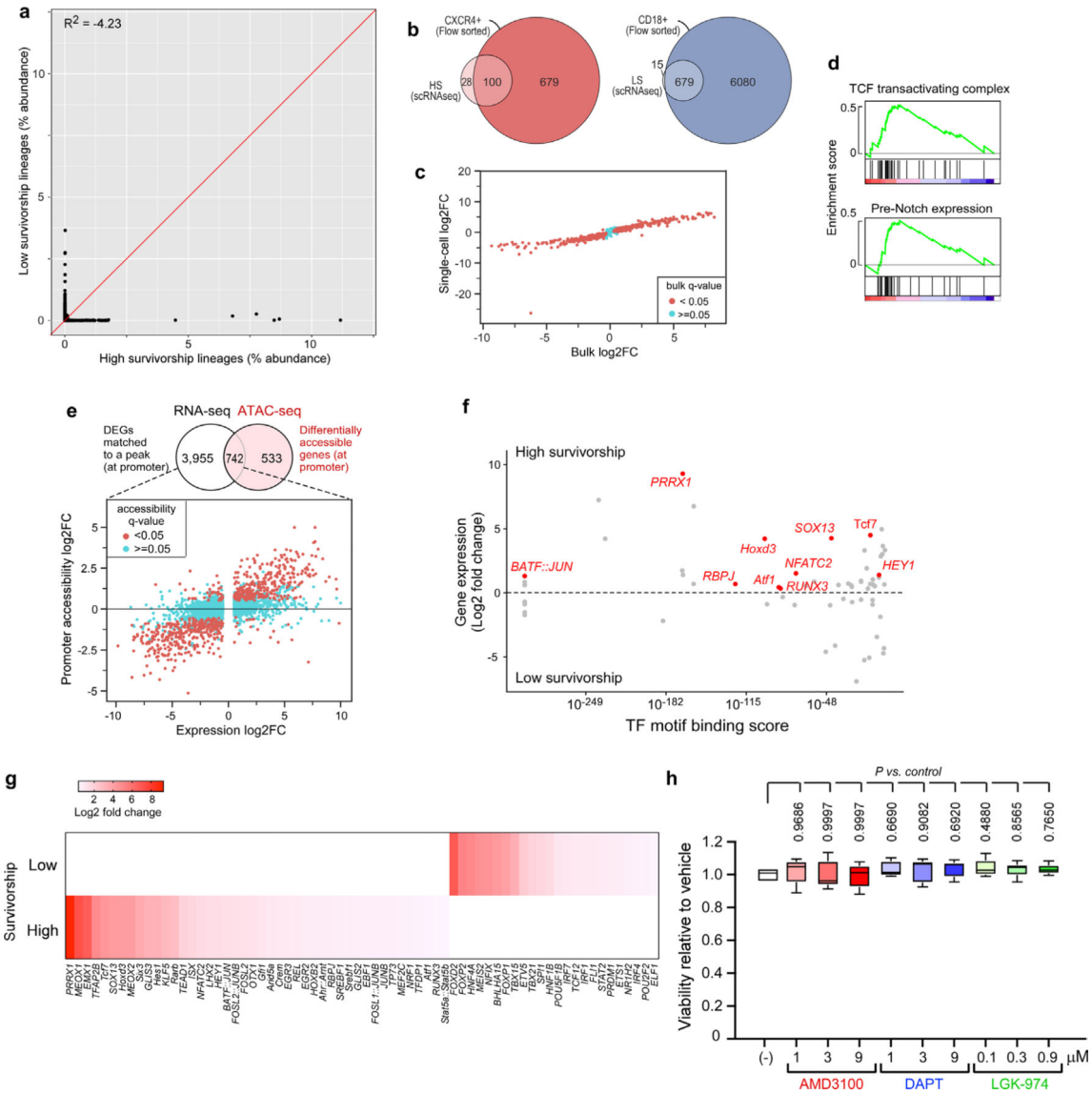


Extended Data Fig. 1 |. Generation and analysis of high-diversity barcode populations.
a, UMAP of expression profiles derived from bulk RNA-seq of 6 untreated HG3 samples (TP0) and 610 treatment-naïve primary CLLs annotated by expression cluster (left) and *IGHV* mutation status (right)[39]. All 616 samples were pre-processed together by using PEER factors to remove technical effects[69]. **b**, Hierarchical clustering of the expression profiles used in (**a**) further supports the similarity between HG3 and *IGHV*-unmutated

primary CLLs as noted by the dendrogram sub-tree highlights (red, HG3; yellow, *IGHV*-unmutated). **c**, Diversity of the ClonMapper library as estimated by deep sequencing. Green dots correspond to total unique read counts at corresponding read depths. Approximately 68 million unique barcodes were identified at a read depth of 138 million reads (red line), and fitting of the sampled data (blue line) yields unique barcode estimates of ~7.6×10^7. **d**, Barcode frequency distribution of the ClonMapper library as quantified by deep sequencing (depth of 138 million reads with no base below Q30). **e**, CellTiter-Glo analysis of 72-hour dose curves using each chemotherapeutic alone ($n = 2$ cell culture replicates from one independent experiment, respectively) or (**f**) in combination at molar ratios equivalent to each drug's respective LD_{50} concentrations (HG3 $n = 2$, REC1 $n = 3$ cell culture replicates as in **d** above). **g**, Ridgeline plots representing the log2 fold change of barcode proportions from TP0 to TP1, where barcodes are grouped by their presence across 8 barcoded cell culture replicates. Only barcodes at or above 0.005% (HG3) and 0.0005% (REC1) abundance at TP0 are shown. **h**, Unique barcode read counts of TP0 and TP1 barcoded cell culture replicates 1–8 normalized to $\log_{10}$ counts per million. Barcodes are sorted in descending order by the sum of their counts across TP1 columns. **i**, Venn diagrams of clones occupying HS versus LS clusters at TP0 and TP1.

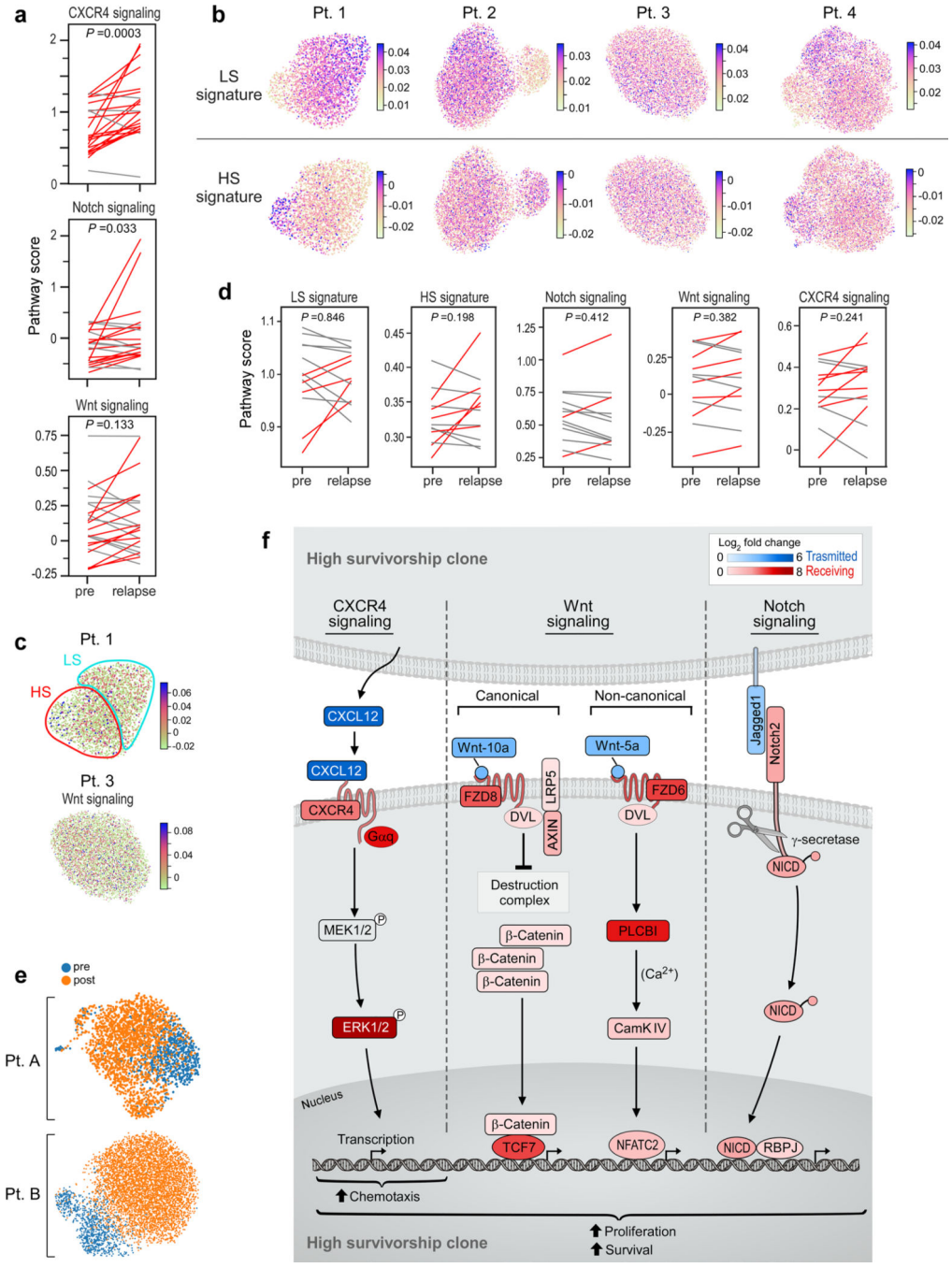


Extended Data Fig. 2 |. Marker gene expression over time across survivorship clusters.
a, Heatmap of marker gene expression upregulated in TP0 vs TP1 HS (top; rows annotated by vertical orange bar), and TP1 vs TP0 HS (bottom; rows annotated by vertical red bar). **b**, Heatmap of marker gene expression upregulated in TP0 vs TP1 LS (top; rows annotated by vertical blue bar), and TP1 vs TP0 LS (top; rows annotated by vertical green bar). All with log2(FC) > 2, $q < 0.05$.

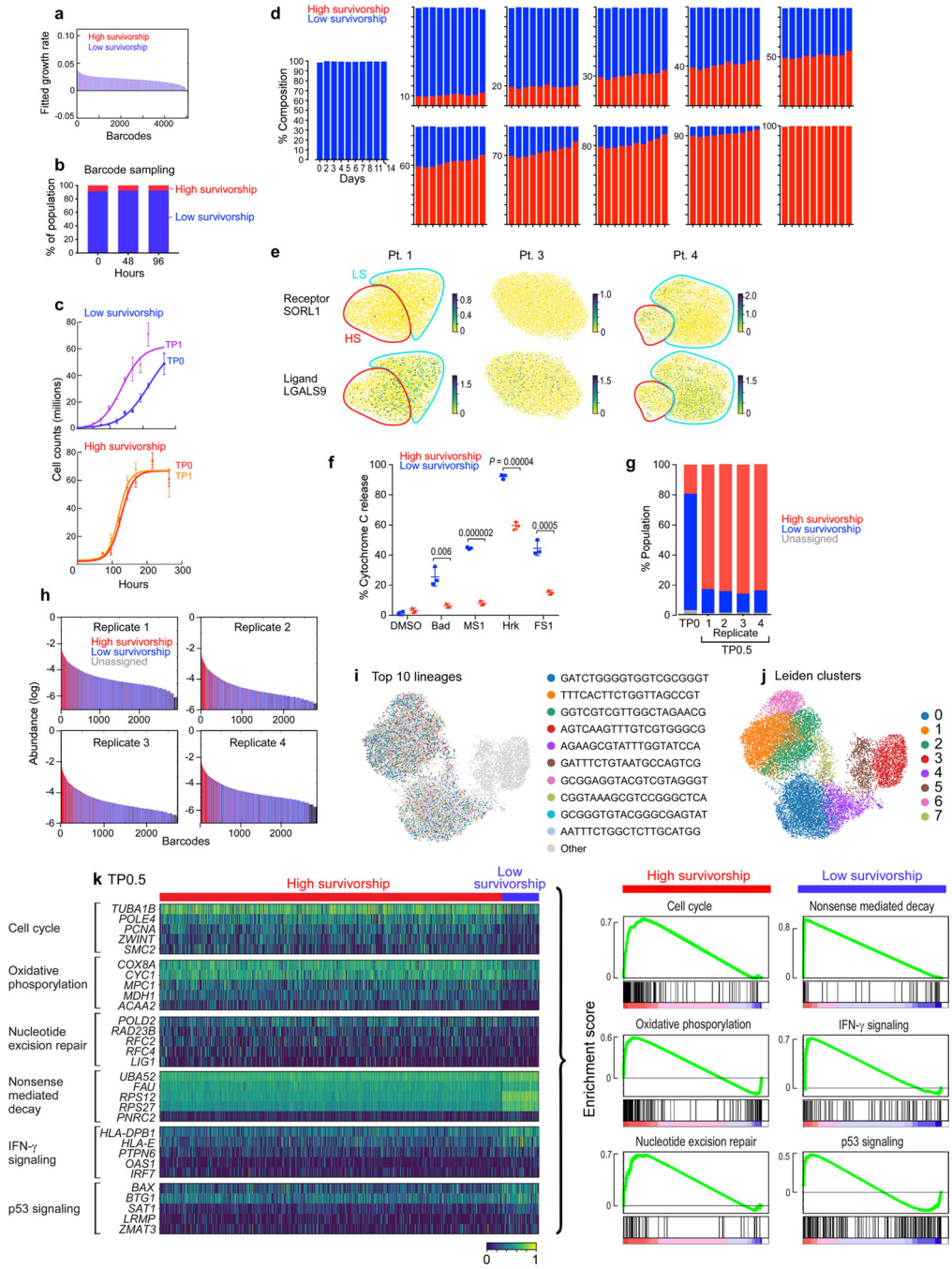


Extended Data Fig. 3 |. Molecular characteristics of high and low survivorship populatios. **a**, Correlation plots of unique barcode abundance as determined by barcode sampling of CD18+ and CXCR4+ isolated populations (n = 1 biologically independent sample). **b**, Venn diagram of overlap between unique barcodes identified by barcode sampling of FACS-sorted populations and unique barcodes identified by scRNA-seq. **c**, Correlation plot of differentially expressed genes between HS and LS subpopulations as measured by scRNA-seq and bulk RNA-seq (n = 1,921 genes). **d**, Two top pathways (FDR < 0.25, nominal P < 0.05) identified through gene set enrichment analysis of differentially expressed genes identified through bulk RNA-sequencing. **e**, Correlation plot of genes differentially expressed by RNA-seq that also have differentially accessible peaks by ATAC-seq. Spearman correlation = 0.63; n = 4010 genes. **f**, Transcription factor motif enrichment within differentially accessible ATAC-seq peaks (DESeq2 two-sided Wald test q < 0.05, n = 6 paired samples) in HS versus LS subpopulations. Motif scores are E-values produced by AME using Fisher's exact test. Red; transcription factors involved in Wnt and

Notch signaling. **g**, Log2(FC) of transcription factor motif enrichment within differentially accessible ATAC-seq peaks (DESeq2 2-sided Wald test $q < 0.05$, $n = 6$ paired samples) in HS versus LS subpopulations, colored by gene expression. **h**, Viability assay of the HS subpopulation treated with small molecule inhibitors of the Wnt, Notch, or CXCR4 pathways. ANOVA with Dunnett's test was used to adjust the p values for comparison of each experimental condition to the single control ($n = 3$ biologically independent samples). Whiskers, minimum and maximum values, box, first and third quartile, line, median.



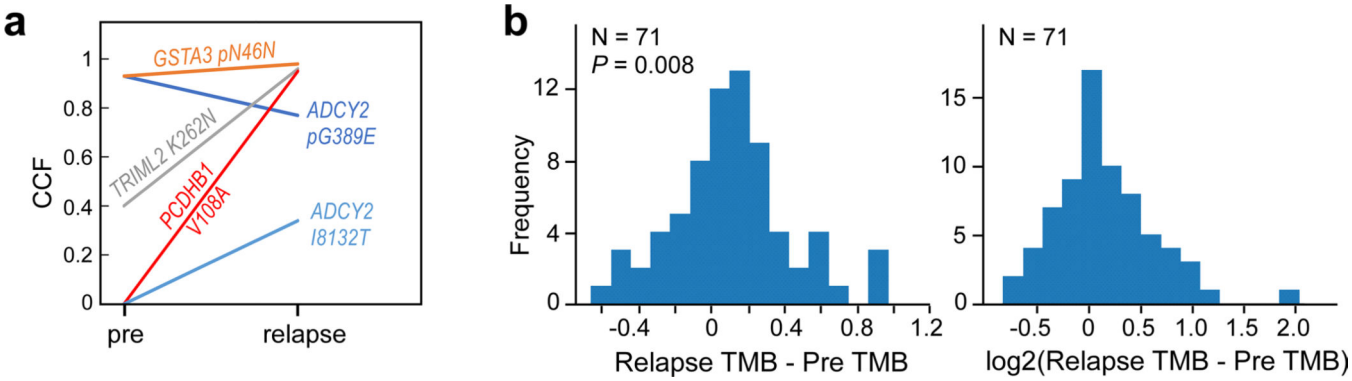
Extended Data Fig. 4 |. High and low survivorship signatures in primary CLL samples.
a, Transcriptomic enrichment analysis of the CXCR4, Wnt, and Notch signaling scores (n = 23 patients; one-sided t-test; adjustments were not made for multiple comparisons). **b**, UMAP of scRNA-seq from patient samples (n = 4 patients) post-FCR therapy. Samples are annotated by HS and LS signatures. **c**, UMAP of scRNA-seq from patients 1 and 3 annotated by Wnt signature. **d**, Signaling enrichment in matched transcriptomes collected from a 13-patient cohort with *IGHV*-unmutated CLL before and 6–12 months after initiating ibrutinib therapy (one-sided t-test; adjustments were not made for multiple comparisons). **e**, UMAP of scRNA-seq from patient samples before and during ibrutinib therapy (n = 2 patients). **f**, Schematic of CXCR4, Wnt and Notch signaling in HS clones. Transmitting ligands and receiving receptors are both upregulated in HS as compared to LS clones. Intensity of color (blue if transmitting, red if receiving) indicates the log2(FC) in expression identified through bulk RNA-sequencing.



Extended Data Fig. 5 |. Differential growth dynamics of high and low survivorship subpopulations.

a, Fitted growth rate of unique clones at TP0 across 4 parallel barcoded cell culture replicates labeled by HS (red) vs LS (blue) classification. **b**, Percentage of population belonging to HS or LS subpopulations over 96-hour outgrowth. **c**, Cell counts of bulk, HS and LS subpopulations at TP0 and TP1. LS TP1 cell counts were statistically significantly higher ($P = 0.00029$) than TP0 cell counts. There was no statistically significant difference between HS TP0 and TP1 cell counts ($P = 0.639$). P-values calculated using

a two-sided two-sample Student's t-test (n = 1 of 2 biologically independent experiments with 3 technical replicates each, error bars denote mean±SD; adjustments not made for multiple comparison). **d**, Stacked bar graphs depict HS and LS subpopulations (marked by GFP or RFP, respectively) co-cultured at different proportions and analyzed by flow cytometry over time. **e**, UMAP of scRNA-seq data from patient samples post-FCR therapy annotated by LGALS9 and SORL1 expression (red outline, high survivorship expression; blue outline, low survivorship expression - as demonstrated in Extended Data Fig. 4c). **f**, BH3 profiling of LS versus HS subpopulations. Statistical significance determined using the Holm-Sidak method for multiple comparison two-sided t-tests, alpha=0.05 (n = 1 biologically independent experiment with 3 technical replicates, error bars denote mean ± SD). **g**, Percent abundance of barcodes at TP0 and TP0.5 across 4 parallel barcoded cell culture replicates labeled by HS (red), LS (blue), and unassigned (grey) classification. **h**, Ordered distribution of barcode abundance at each timepoint, where each bar represents a lineage. Red, HS. Blue, LS. Gray, unassigned. **i-j**, UMAP of scRNA-seq data from TP0, TP0.5 and TP1 with top 10 lineages and Leiden clusters annotated (n = 14,505 cells). **k**, Scaled expression of genes corresponding to pathways from GSEA analysis (input from 2-sided Wilcoxon rank-sum comparing TP0.5 HS versus LS clusters).



Extended Data Fig. 6 |. Clonal diversification.

a, Cancer cell fractions (CCF) of representative HS mutations enriched upon chemotherapy relapse in cohort of 71 patients with CLL treated with fludarabine, cyclophosphamide and rituximab[24,25]. **b**, Change in tumor mutation burden (TMB) in 71 patients with relapsed CLL treated with fludarabine and cyclophosphamide or fludarabine, cyclophosphamide, and rituximab. Raw difference (left; two-tailed paired t-test, t = −2.72, P = 0.008). Right, log2(FC) fold change.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We acknowledge S. Pollock, L. Nguyen, H. Lyon and C. Patterson for expert project management. We thank C. Hahn, I. Leschiner and B. Persaud for excellent input on human sequencing analysis. We acknowledge support from the National Institutes of Health (5R21CA212928, 5R01CA226258 to A.B.; and 3P01CA206978-03S1, 1U10CA180861-01, 1P01CA206978-01 to C.J.W.). C.J.W. is a Scholar of the Leukemia and Lymphoma Society and C.G. is a Scholar through the American Society of Hematology MMSAP Program and the F31 Diversity

Individual Predoctoral Fellowship program through the NCI. E.B. is a recipient of the University of Texas at Austin Provost's Graduate Excellence Fellowship and the F.M. Jones & H.L. Bruce Endowed Graduate Fellowship. K.J. is grateful for support through a National Science Foundation Graduate Research Fellowship (1610403). B.A.K. was supported by a long-term EMBO fellowship (ALTF 14-2018). We thank the Dana-Farber Flow Cytometry Core, the Broad Institute Walk-Up Sequencing Core and the Genome Sequencing and Analysis Facility at the University of Texas at Austin for their services. All scRNA-seq workflows were conducted in the Translational Immunogenomics Laboratory at the Dana-Farber Cancer Institute.

Competing interests

C.J.W. holds equity in BioNTech Inc. and receives research funding from Pharmacyclics. C.J.W. and D.N. have been consultants for H3 Biomedicine and received research funding from Celgene. G.G. receives research funds from IBM and Pharmacyclics and is an inventor of several bioinformatics-related patents, including related to MuTect and ABSOLUTE. All other authors declare no competing interests.

References

1. Marusyk A, Janiszewska M & Polyak K Intratumor heterogeneity: the rosetta stone of therapy resistance. Cancer Cell 37, 471–484 (2020). [PubMed: 32289271]
2. Turajlic S, Sottoriva A, Graham T & Swanton C Resolving genetic heterogeneity in cancer. Nat. Rev. Genet. 20, 404–416 (2019). [PubMed: 30918367]
3. Egan JB et al. Whole-genome sequencing of multiple myeloma from diagnosis to plasma cell leukemia reveals genomic initiating events, evolution, and clonal tides. Blood 120, 1060–1066 (2012). [PubMed: 22529291]
4. Kasar S et al. Whole-genome sequencing reveals activation-induced cytidine deaminase signatures during indolent chronic lymphocytic leukaemia evolution. Nat. Commun. 6, 8866 (2015). [PubMed: 26638776]
5. Landau DA et al. Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. Cell 152, 714–726 (2013). [PubMed: 23415222]
6. Roth A et al. PyClone: statistical inference of clonal population structure in cancer. Nat. Methods 11, 396–398 (2014). [PubMed: 24633410]
7. Williams MJ, Sottoriva A & Graham TA Measuring clonal evolution in cancer with genomics. Annu. Rev. Genomics Hum. Genet. 20, 309–329 (2019). [PubMed: 31059289]
8. Bhang HE et al. Studying clonal dynamics in response to cancer therapy using high-complexity barcoding. Nat. Med. 21, 440–448 (2015). [PubMed: 25849130]
9. Hata AN et al. Tumor cells can follow distinct evolutionary paths to become resistant to epidermal growth factor receptor inhibition. Nat. Med. 22, 262–269 (2016). [PubMed: 26828195]
10. Kalhor R, Mali P & Church GM Rapidly evolving homing CRISPR barcodes. Nat. Methods 14, 195–200 (2017). [PubMed: 27918539]
11. Pei W et al. Polylox barcoding reveals haematopoietic stem cell fates realized in vivo. Nature 548, 456–460 (2017). [PubMed: 28813413]
12. Kebschull JM & Zador AM Cellular barcoding: lineage tracing, screening and beyond. Nat. Methods 15, 871–879 (2018). [PubMed: 30377352]
13. Kong W et al. CellTagging: combinatorial indexing to simultaneously map lineage and identity at single-cell resolution. Nat. Protoc. 15, 750–772 (2020). [PubMed: 32051617]
14. Weinreb C, Rodriguez-Fraticelli A, Camargo FD & Klein AM Lineage tracing on transcriptional landscapes links state to fate during differentiation. Science 10.1126/science.aaw3381 (2020).
15. Wagner DE et al. Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. Science 360, 981–987 (2018). [PubMed: 29700229]
16. Wagner DE & Klein AM Lineage tracing meets single-cell omics: opportunities and challenges. Nat. Rev. Genet. 10.1038/s41576-020-0223-2 (2020).
17. Raj B, Gagnon JA & Schier AF Large-scale reconstruction of cell lineages using single-cell readout of transcriptomes and CRISPR-Cas9 barcodes by scGESTALT. Nat. Protoc. 13, 2685–2713 (2018). [PubMed: 30353175]
18. Alemany A, Florescu M, Baron CS, Peterson-Maduro J & van Oudenaarden A Whole-organism clone tracing using single-cell sequencing. Nature 556, 108–112 (2018). [PubMed: 29590089]

19. Al'Khafaji AM, Deatherage D & Brock A Control of lineage-specific gene expression by functionalized gRNA barcodes. ACS Synth Biol. 7, 2468–2474 (2018). [PubMed: 30169961]
20. Feldman D et al. CloneSifter: enrichment of rare clones from heterogeneous cell populations. BMC Biol. 18, 177 (2020). [PubMed: 33234154]
21. Umkehrer C et al. Isolating live cell clones from barcoded populations using CRISPRa-inducible reporters. Nat. Biotechnol. 10.1038/s41587-020-0614-0 (2020).
22. Datlinger P et al. Pooled CRISPR screening with single-cell transcriptome readout. Nat. Methods 14, 297–301 (2017). [PubMed: 28099430]
23. Neftel C et al. An integrative model of cellular states, plasticity, and genetics for glioblastoma. Cell 178, 835–849 (2019). [PubMed: 31327527]
24. Landau DA et al. Mutations driving CLL and their evolution in progression and relapse. Nature 526, 525–530 (2015). [PubMed: 26466571]
25. Gruber M et al. Growth dynamics in naturally progressing chronic lymphocytic leukaemia. Nature 570, 474–479 (2019). [PubMed: 31142838]
26. Chen L et al. CXCR4 upregulation is an indicator of sensitivity to B-cell receptor/PI3K blockade and a potential resistance mechanism in B-cell receptor-dependent diffuse large B-cell lymphomas. Haematologica 105, 1361–1368 (2020). [PubMed: 31471373]
27. Li L, Chai Y, Wu C & Zhao L Chemokine receptor CXCR4: an important player affecting the molecular-targeted drugs commonly used in hematological malignancies. Expert Rev. Hematol 13, 1387–1396 (2020). [PubMed: 33170753]
28. Efremova M, Vento-Tormo M, Teichmann SA & Vento-Tormo R CellPhoneDB: inferring cell–cell communication from combined expression of multi-subunit ligand–receptor complexes. Nat. Protoc. 15, 1484–1506 (2020). [PubMed: 32103204]
29. Giovannone N et al. Galectin-9 suppresses B cell receptor signaling and is regulated by I-branching of N-glycans. Nat. Commun. 9, 3287 (2018). [PubMed: 30120234]
30. Giovannone N, Smith LK, Treanor B & Dimitroff CJ Galectin–glycan interactions as regulators of B cell immunity. Front. Immunol. 9, 2839 (2018). [PubMed: 30564237]
31. Kobayashi T et al. Galectin-9 exhibits anti-myeloma activity through JNK and p38 MAP kinase pathways. Leukemia 24, 843–850 (2010). [PubMed: 20200560]
32. Kuroda J et al. Targeting activating transcription factor 3 by galectin-9 induces apoptosis and overcomes various types of treatment resistance in chronic myelogenous leukemia. Mol. Cancer Res. 8, 994–1001 (2010). [PubMed: 20571063]
33. Kageshita T et al. Possible role of galectin-9 in cell aggregation and apoptosis of human melanoma cell lines and its clinical significance. Int. J. Cancer 99, 809–816 (2002). [PubMed: 12115481]
34. Wiersma VR et al. The glycan-binding protein galectin-9 has direct apoptotic activity toward melanoma cells. J. Invest. Dermatol. 132, 2302–2305 (2012). [PubMed: 22572821]
35. Cohen S et al. The cytokine midkine and its receptor RPTPzeta regulate B cell survival in a pathway induced by CD74. J. Immunol. 188, 259–CD269 (2012). [PubMed: 22140262]
36. Cohen S & Shachar I Midkine as a regulator of B cell survival in health and disease. Br. J. Pharmacol. 171, 888–895 (2014). [PubMed: 24111754]
37. Pietila M et al. SORLA regulates endosomal trafficking and oncogenic fitness of HER2. Nat. Commun. 10, 2340 (2019). [PubMed: 31138794]
38. Campion O et al. Contribution of the low-density lipoprotein receptor family to breast cancer progression. Front. Oncol. 10, 882 (2020). [PubMed: 32850302]
39. Knisbacher B et al. The CLL-1100 project: towards complete genomic characterization and improved prognostics for CLL. In 62nd ASH Annual Meeting and Exposition, 5–8 December (2020).
40. Feng YC et al. c-Myc inactivation of p53 through the pan-cancer lncRNA MILIP drives cancer pathogenesis. Nat. Commun. 11, 4980 (2020). [PubMed: 33020477]
41. Hayashi F et al. Increased expression of tripartite motif (TRIM) like 2 promotes tumoral growth in human oral cancer. Biochem. Biophys. Res. Commun. 508, 1133–1138 (2019). [PubMed: 30554657]

42. Kung CP, Khaku S, Jennis M, Zhou Y & Murphy ME Identification of TRIML2, a novel p53 target, that enhances p53 SUMOylation and regulates the transactivation of proapoptotic genes. Mol. Cancer Res. 13, 250–262 (2015). [PubMed: 25256710]
43. Xu L et al. Acquired mutations associated with ibrutinib resistance in Waldenstrom macroglobulinemia. Blood 129, 2519–2525 (2017). [PubMed: 28235842]
44. Efferth T et al. Collateral sensitivity of natural products in drug-resistant cancer cells. Biotechnol. Adv. 38, 107342 (2020). [PubMed: 30708024]
45. Goldman A et al. Targeting tumor phenotypic plasticity and metabolic remodeling in adaptive cross-drug tolerance. Sci. Signal. 10.1126/scisignal.aas8779 (2019).
46. Imamovic L et al. Drug-driven phenotypic convergence supports rational treatment strategies of chronic infections. Cell 172, 121–134 (2018). [PubMed: 29307490]
47. Lin KH et al. Using antagonistic pleiotropy to design a chemotherapy-induced evolutionary trap to target drug resistance in cancer. Nat. Genet. 52, 408–417 (2020). [PubMed: 32203462]
48. Tong M et al. New insights from the widening homogeneity perspective to target intratumor heterogeneity. Cancer Commun. 38, 17 (2018).
49. Acar A et al. Exploiting evolutionary steering to induce collateral drug sensitivity in cancer. Nat. Commun. 11, 1923 (2020). [PubMed: 32317663]
50. Zhang J, Cunningham JJ, Brown JS & Gatenby RA Integrating evolutionary dynamics into treatment of metastatic castrate-resistant prostate cancer. Nat. Commun. 8, 1816 (2017). [PubMed: 29180633]
51. Luecken MD & Theis FJ Current best practices in single-cell RNA-seq analysis: a tutorial. Mol. Syst. Biol. 15, e8746 (2019). [PubMed: 31217225]
52. Buttner M, Miao Z, Wolf FA, Teichmann SA & Theis FJ A test metric for assessing single-cell RNA-seq batch correction. Nat. Methods 16, 43–49 (2019). [PubMed: 30573817]
53. Vieth B, Parekh S, Ziegenhain C, Enard W & Hellmann I A systematic evaluation of single cell RNA-seq analysis pipelines. Nat. Commun. 10, 4667 (2019). [PubMed: 31604912]
54. Traag VA, Waltman L & van Eck NJ From Louvain to Leiden: guaranteeing well-connected communities. Sci. Rep. 9, 5233 (2019). [PubMed: 30914743]
55. Lun AT, Bach K & Marioni JC Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. Genome Biol. 17, 75 (2016). [PubMed: 27122128]
56. Zhang JM, Kamath GM & Tse DN Valid post-clustering differential analysis for single-cell RNA-seq. Cell Syst. 9, 383–392 (2019). [PubMed: 31521605]
57. Mootha VK et al. PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat. Genet. 34, 267–273 (2003). [PubMed: 12808457]
58. Liao Y, Smyth GK & Shi W featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 30, 923–930 (2014). [PubMed: 24227677]
59. Corces MR et al. Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. Nat. Genet. 48, 1193–1203 (2016). [PubMed: 27526324]
60. Liao Y, Smyth GK & Shi W The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. Nucleic Acids Res. 47, e47 (2019). [PubMed: 30783653]
61. Ryan JA, Brunelle JK & Letai A Heightened mitochondrial priming is the basis for apoptotic hypersensitivity of $CD4^{+}CD8^{+}$ thymocytes. Proc. Natl Acad. Sci. USA 107, 12895–12900 (2010). [PubMed: 20615979]
62. Cibulskis K et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat. Biotechnol. 31, 213–219 (2013). [PubMed: 23396013]
63. Benjamin D et al. Calling somatic SNVs and indels with Mutect2. Preprint at bioRxiv 10.1101/861054 (2019).
64. Kim S et al. Strelka2: fast and accurate calling of germline and somatic variants. Nat. Methods 15, 591–594 (2018). [PubMed: 30013048]
65. Chen X et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics 32, 1220–1222 (2016). [PubMed: 26647377]

66. Wala JA et al. SvABA: genome-wide detection of structural variants and indels by local assembly. Genome Res. 28, 581–591 (2018). [PubMed: 29535149]

67. Sun R et al. Breakpointer: using local mapping artifacts to support sequence breakpoint discovery from single-end reads. Bioinformatics 28, 1024–1025 (2012). [PubMed: 22302574]

68. Pedregosa F et al. Scikit-learn: machine learning in Python. JMLR 12, 2825–2830 (2011).

69. Stegle O, Parts L, Durbin R & Winn J A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. PLoS Comput. Biol. 6, e1000770 (2010). [PubMed: 20463871]

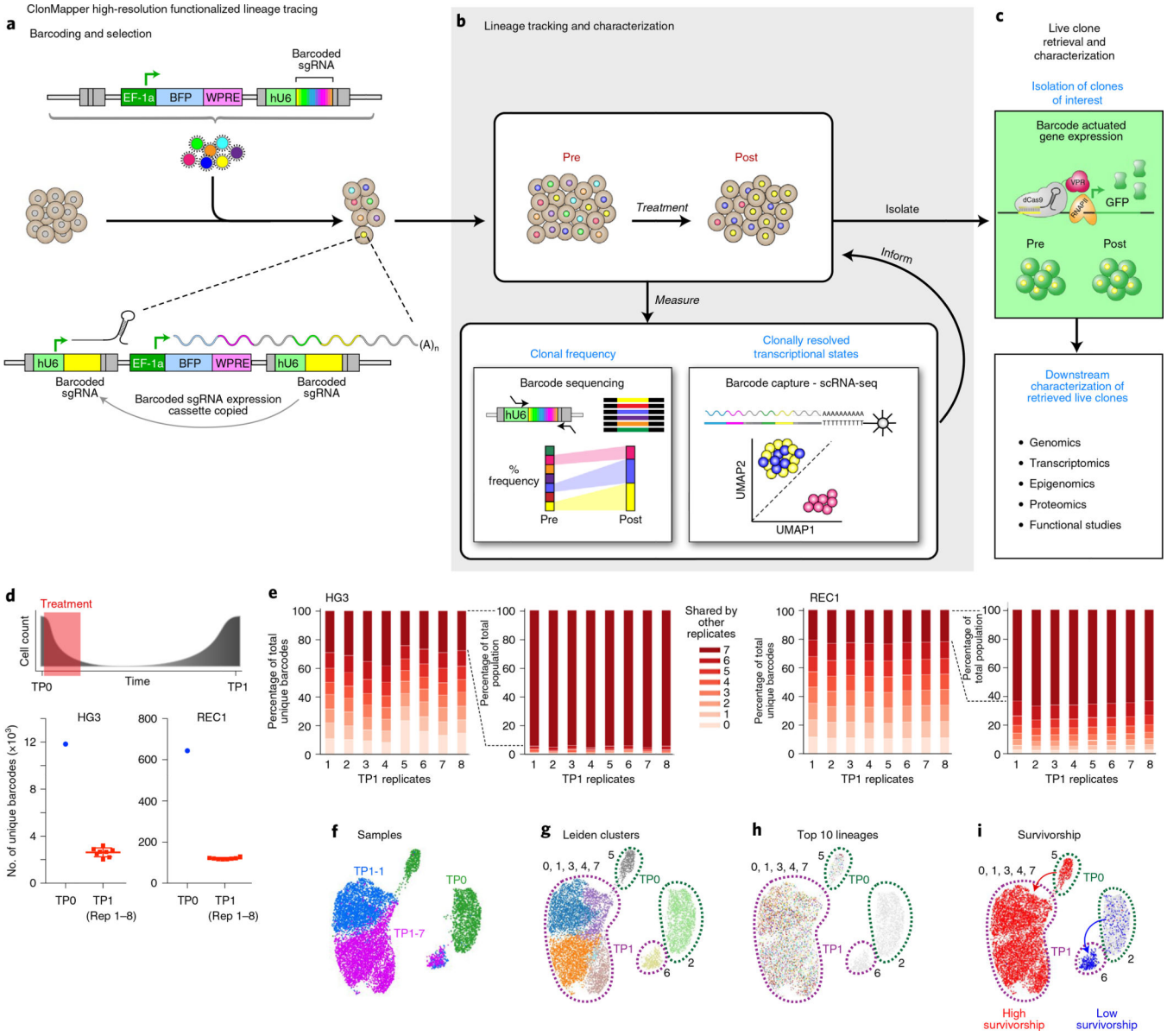


Fig. 1 |. Expressed barcodes reveal clonal dynamics of response to chemotherapeutic treatment in MCL and CLL cell lines.

a–c, Experimental workflow for the multifunctional ClonMapper system. Barcoding and selection of tagged cells (**a**). Characterization of clonal dynamics through targeted barcode sequencing and clone-resolved scRNA-seq (**b**). Live clone retrieval and functional analysis of clones of interest as determined in **b** (**c**). **d**, Parallel drug treatment replicates of eight ClonMapper-tagged HG3 and REC1 cell lines. TP, time point. Red indicates treatment window (top). Unique barcode counts at TP0 (n = 1 barcoded parental cell culture population) and in TP1 (n = 8 parallel barcoded cell culture replicates) replicates for the HG3 and REC1 cell lines (error bars denote mean ± s.d.) (bottom). **e**, Percentage of observed unique barcodes present in each TP1 replicate and shared across replicates (n = 8 parallel barcoded cell culture replicates) (left). Shared barcodes weighted by population abundance (right). **f–i**, Uniform manifold approximation and projection (UMAP) of 8,975

cells from scRNA-seq of the HG3 barcoded population at TP0 and TP1. Sample annotations are green, TP0; blue, TP1 replicate 1; and purple, TP1 replicate 7 (**f**). Leiden cluster annotations reveal substantial pre- and post-treatment heterogeneity (**g**). Annotation of the ten most abundant clones reveals their exclusive presence in TP1 clusters 0, 1, 3, 4, 7 and in TP0 cluster 5 (each color represents a unique clone) (**h**). Clonal survivorship classification based on clonal residence in either TP1 clusters 0, 1, 3, 4 and 7 (red, HS) or cluster 6 (blue, LS) (**i**). Clones not detected in TP1 are labeled in gray (unassigned).

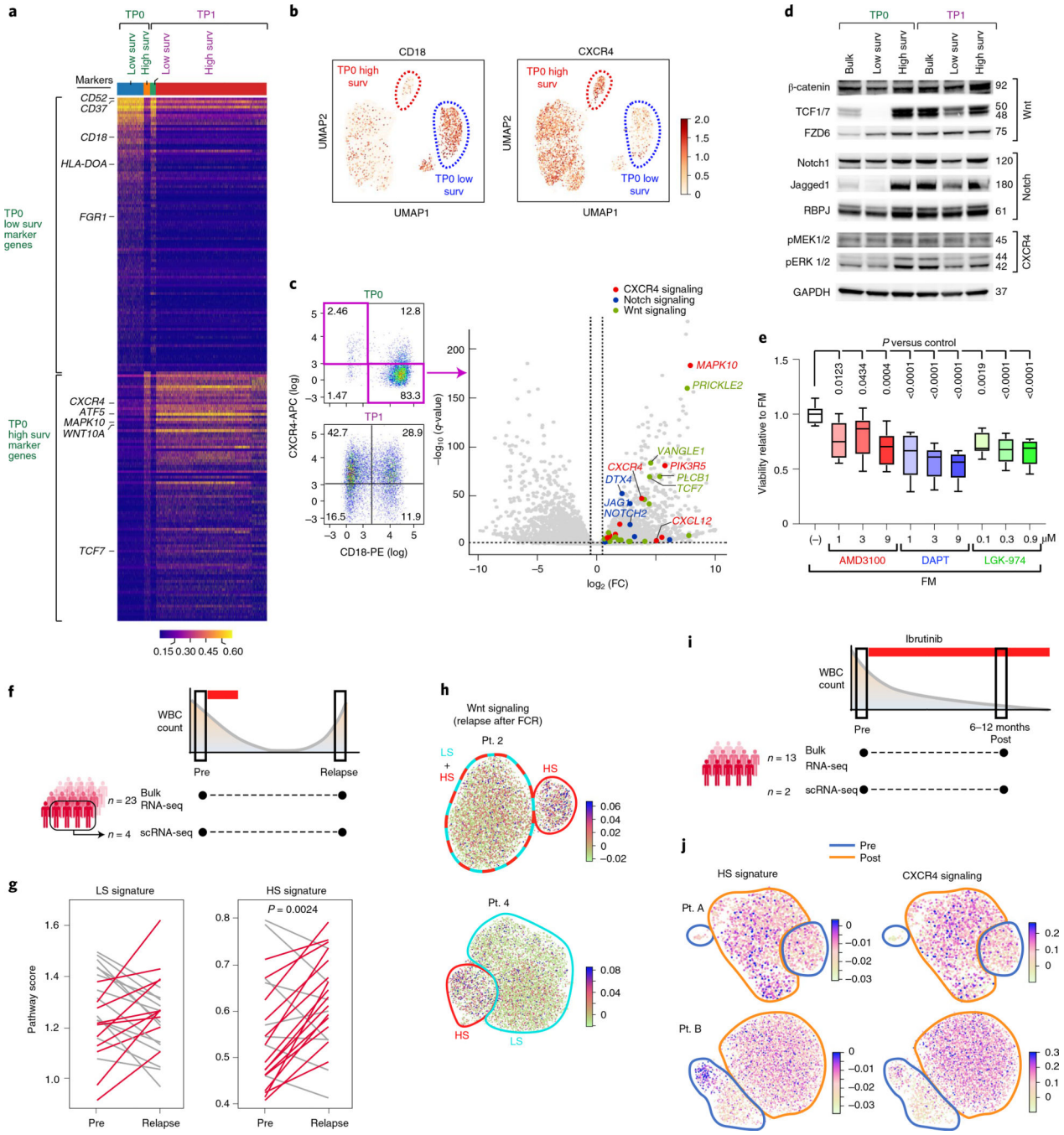


Fig. 2 |. Characterization and clinical associations of high and low survivorship subpopulations.
a, Scaled expression of marker genes distinguishing HS and LS (surv) clusters at TP0 (n = 393 HS cells, n = 1,589 LS cells) and TP1 (n = 6,656 HS cells, n = 337 LS cells) (two-sided Wilcoxon rank-sum test with $q < 0.05$ and $\log_2(\text{FC}) > 2$, truncated normal test with $q < 0.05$). FC, fold change. For each gene, expression was scaled from 0–1 to maintain a balanced colormap. Genes are ranked by $\log_2(\text{FC})$; genes of interest are highlighted. **b**, UMAPs of cells from **a** annotated by CXCR4 and CD18 expression. **c**, Flow cytometric analysis of TP0 and TP1 populations stained with CXCR4–APC and CD18–PE antibodies.

CXCR4+ and CD18+ cells were FACS isolated for subsequent characterization (inset). Volcano plot displays elevated Wnt, Notch and CXCR4 signaling in CXCR4+ versus CD18+ cells as identified by bulk RNA sequencing (DESeq2 two-sided Wald test $q < 0.05$, $\log_2(FC) > 0.5$, $n = 6$ paired samples). **d**, One representative western blot of CXCR4+ and CD18+ subpopulations at TP0 and TP1 (gel pieces were cropped as noted by white space; findings reproduced in $n = 3$ biologically independent experiments). **e**, Viability assay of the HS subpopulation treated with FM alone or co-treated with FM and small molecule inhibitors of the Wnt (LGK-974), Notch (DAPT) or CXCR4 (AMD3100) pathways. Analysis of variance with Dunnett's test was used to adjust the *P* values for comparison of each experimental condition to the single control ($n = 3$ biologically independent samples). All *P* values are two-sided; whiskers, min and max values; box, first and third quartile; line, median. **f**,**g**, Matched transcriptomes collected from a 23-patient cohort with *IGHV*-unmutated CLL before and after relapse to fludarabine-based chemotherapy. Four patient CLL samples additionally underwent scRNA-seq. Transcriptomic enrichment analysis of HS and LS signature scores (Methods; one-sided paired *t*-test) (**g**). WBC, white blood cell. **h**, UMAP of scRNA-seq from patient (Pt.) samples ($n = 2$ patients) following fludarabine, cyclophosphamide and rituximab (FCR) therapy. Samples are annotated by Wnt signaling score. **i**, Matched transcriptomes collected from a 13-patient cohort with *IGHV*-unmutated CLL before and 6–12 months after initiating ibrutinib therapy. **j**, UMAP of scRNA-seq from patient samples ($n = 2$ patients) before and during ibrutinib therapy. Samples are annotated by HS signature and CXCR4 signaling scores.

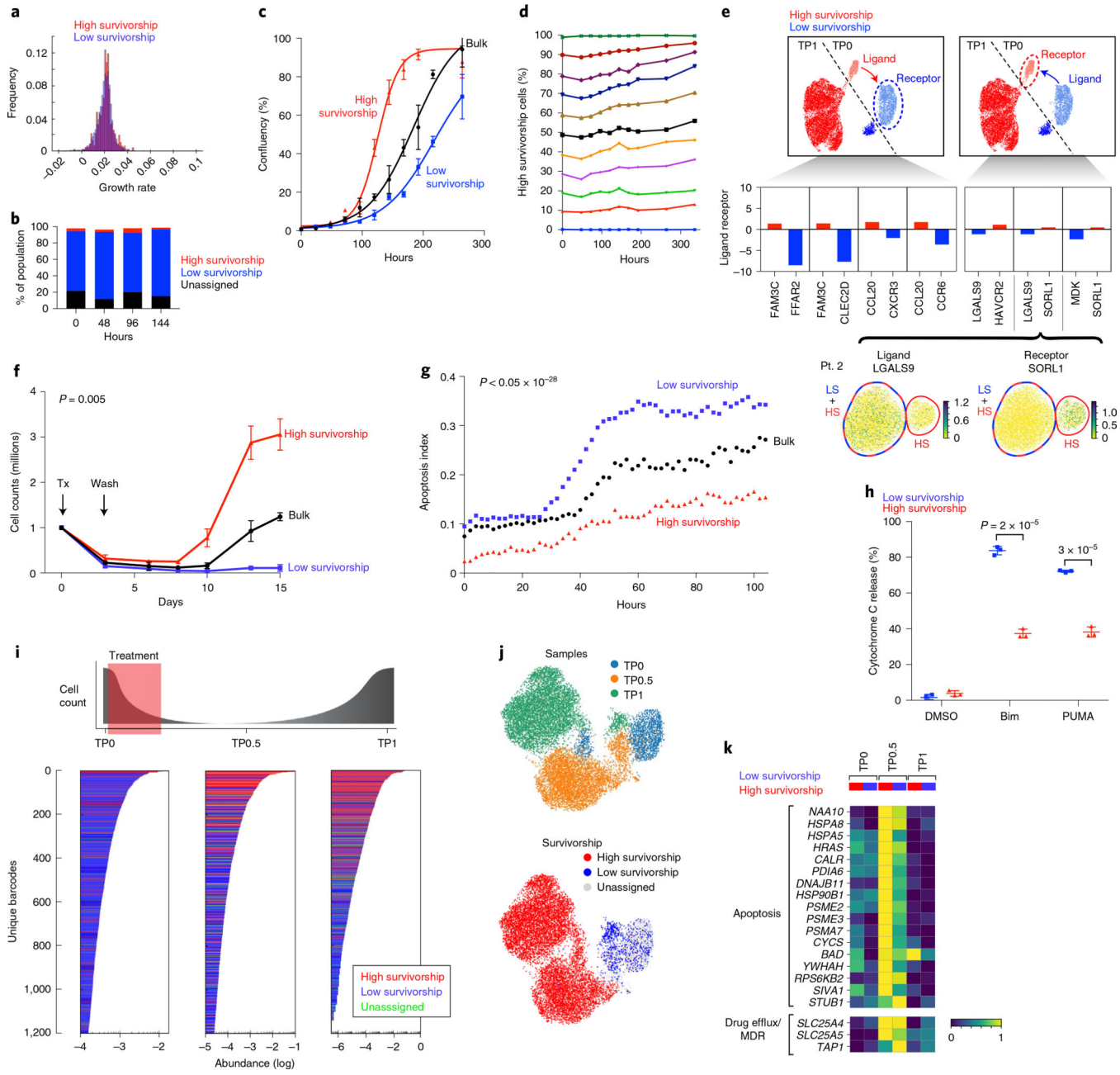


Fig. 3 |. High survivorship clones exhibit elevated drug tolerance.

a, Frequency distribution of fitted lineage growth rates of individual clones annotated by survivorship classification (n = 462 HS clones, n = 4,596 LS clones). **b**, Flow cytometric analysis of expanding TP0 population (n = 1 biologically independent sample, same sample as in **a**). **c**, Growth dynamics of bulk TP0, HS and LS subpopulations. P value calculated using a two-sided two-sample Student's t-test on growth rates fitted to individual replicates for HS and LS samples ($P < 0.005$, n = 1 (of two) representative experiment with three technical replicates, error bars denote mean ± s.d. **d**, Percentage of HS cells detected over time by flow cytometry (in samples with different initial starting proportions; n = 1 independent experiment). **e**, CellPhoneDB analysis of interacting ligand pairs in HS (red)

versus LS (blue) clones in scRNA-seq analysis (*y* axis = $\log_2$(FC), $P < 0.05$; $n = 393$ TP0 HS cells, $n = 1{,}589$ TP0 LS cells). Representative patient ($n = 1$ patient) with LGALS9 and SORL1 expression associated with LS and HS expression, respectively (inset). **f**, Cell counts over time following 72-h FM treatment of bulk TP0 (black), HS (red) and LS (blue) subpopulations. *P* value was calculated by calculating regrowth rates from each replicate starting at day 8 and comparing groups of growth rates using a two-sided two-sample *t*-test ($n = 1$ (of two) independent experiment with three biologically independent cell culture samples, error bars denote mean ± s.d.). **g**, Live-cell imaging quantification of annexin V in bulk TP0, HS and LS subpopulations treated with an LD_{50} FM dose. LD_{50}, dose lethal to 50% of population tested. Apoptosis index indicated by red fluorescence/total object count ($n = 1$ independent experiment with two technical replicates). **h**, Differential baseline apoptotic priming in HS and LS subpopulations as determined by BH3 profiling. Statistical significance determined using the Holm–Sidak method for multiple comparison two-sided *t*-tests, with $\alpha = 0.05$ ($n = 1$ biologically independent experiment with three technical replicates, error bars denote mean ± s.d.). DMSO, dimethylsulfoxide. **i**, Ordered distribution of barcode abundance at each time point, where each bar represents a lineage. Red, HS; blue, LS; green, unassigned. **j**, UMAP visualizations of scRNA-seq from TP0, TP0.5 and TP1. Similar to the HS and LS classification in Fig. 1i, TP0.5 cells from Leiden clusters 0 and 4 were designated as 'TP0.5 HS' ($n = 4{,}987$ cells) and from Leiden cluster 5 as 'TP0.5 LS' ($n = 543$ cells; Extended Data Fig. 5j). **k**, Scaled mean expression of genes relating to apoptosis and drug efflux across survivorship clusters. For each gene, values were scaled from 0–1 to maintain a balanced color map (two-sided Wilcoxon rank-sum test with $q < 0.05$ and $\log_2(\text{FC}) > 0.5$, truncated normal test with $q < 0.05$). MDR, multidrug resistance.

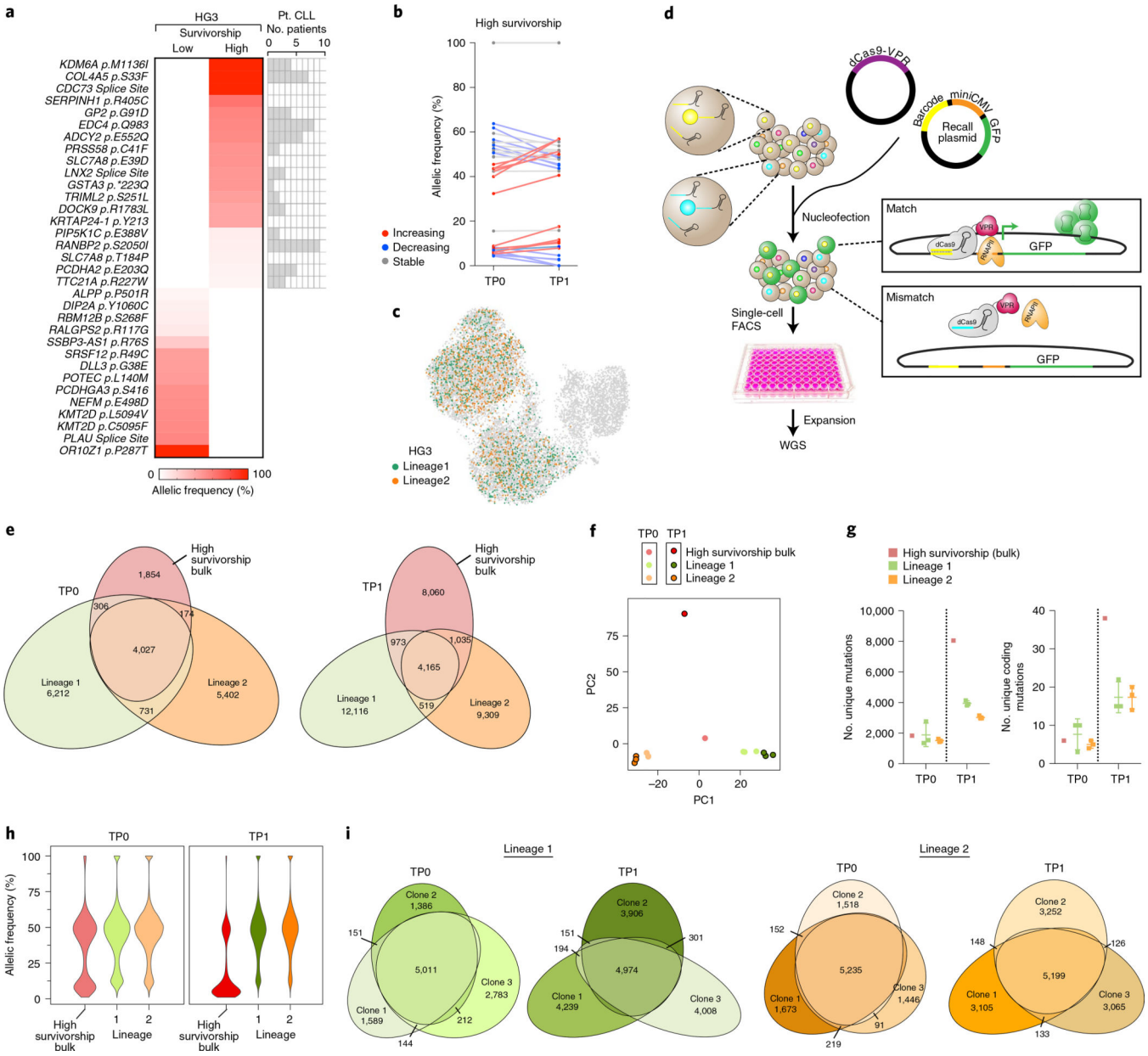


Fig. 4 |. ClonMapper-enabled whole-genome sequencing reveals clonal and subclonal genetic diversification.

a, VAFs of sSNVs (PolyPhen2 score > 0.8) identified in the HS and LS bulk subpopulation at TP0 (left). Prevalence of these mutations in primary CLL patient samples (n = 984 patients) (right). **b**, VAFs of sSNVs identified in the HS bulk subpopulation at TP0 and TP1. Gray, stable VAFs; red, increasing VAFs; blue, decreasing VAFs (change ≥ 10%). **c**, UMAP of scRNA-seq from TP0, TP0.5 and TP1 annotated by isolated lineages 1 (n = 1,017 cells) and 2 (n = 989 cells). **d**, COLBERT-mediated isolation and FACS single-cell sorting of two HS lineages of interest. **e**, Euler diagram of sSNVs identified in the bulk population and lineages 1 and 2 at TP0 and TP1. **f**, Principal-component analysis plot of principal components 1 and 2 segregate lineages and bulk populations by sSNV and treatment status

(*n* = 3 subclones per lineage per time point). **g**, Scatter-plot of total (left) and coding (right) sSNVs identified in the HS bulk (*n* = 1 biologically independent sample) and isolated lineages (*n* = 3 biologically independent samples, respectively) at TP0 and TP1. Error bars denote mean ± s.d. **h**, Violin plots of sSNV allelic frequency distributions across the HS bulk (red) and isolated lineage 1 (green) and 2 (orange) samples at TP0 and TP1. **i**, Euler diagrams of sSNVs identified in three subclones each of lineages 1 (green) and 2 (orange) at TP0 and TP1.

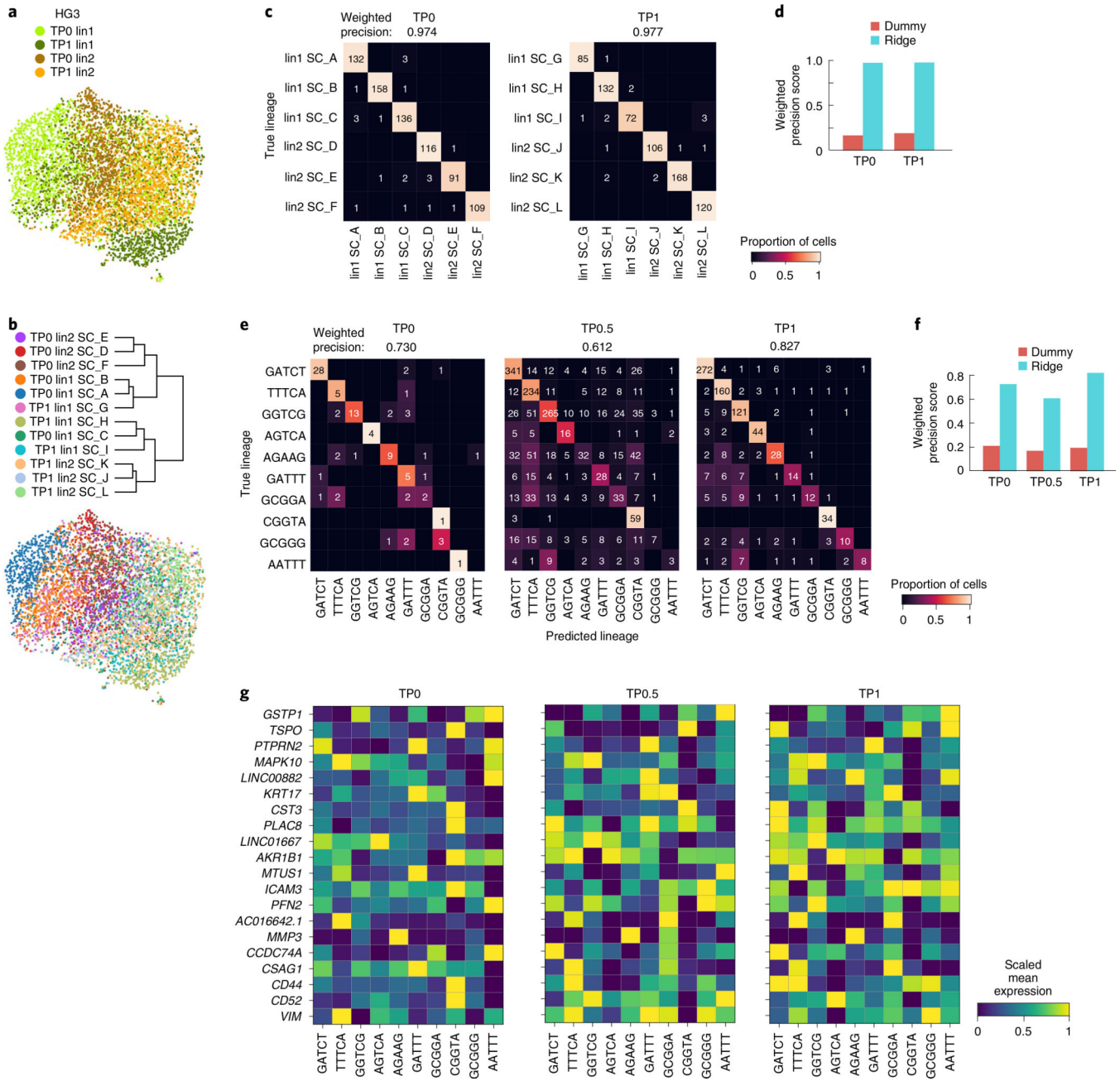


Fig. 5 |.

Clones retain characteristic transcriptomic identity over the course of treatment. **a**, UMAP of scRNA-seq of 12 isolated subclones annotated by the two parental HS lineages (lin1 and lin2) at TP0 and TP1. **b**, UMAP annotated by subclone, parental lineage and time point. Dendrogram showing subclonal similarities associated with parental lineage and time point. **c**, Confusion matrices depicting the ridge classifier predictions of subclonal identity of cells from **b**. Within each time point, data were split into a training set and a test set. Color intensity indicates proportion of cells assigned to a lineage. Raw values indicate number of cells assigned to a lineage. **d**, Bar graph depicting ability of a Ridge and dummy classifier

to predict subclone identity of cells from lineage 1 and lineage 2 from TP0 and TP1. **e**, Confusion matrices depicting the ridge classifier predictions of lineage identity of cells from Extended Data Fig. 5i. **f**, Bar graph depicting ability of a Ridge and dummy classifier to predict lineage identity of cells in the ten most abundant lineages at TP1. Classifiers were trained on a subset of cells from TP1 and applied across time points. **g**, Scaled mean expression of the ridge classifier's top 20 most important genes. For each gene, expression was scaled from 0 to 1.0 to maintain a balanced color map.