In-depth characterization of intratumoral heterogeneity in refractory B-cell non-Hodgkin lymphoma through the lens of a Research Autopsy Program

Keren Isaev,[1] Ting Liu,[1] Mehran Bakhtiari,[1] Kit Tong,[1] Rashmi Goswami,[2] Bernard Lam,[3] Ilinca Lungu,[3] Paul M. Krzyzanowski,[3] Amit Oza,[1] Neesha Dhani,[1] Anca Prica,[1] Michael Crump[1] and Robert Kridel[1]

[1]Princess Margaret Cancer Center - University Health Network; [2]Sunnybrook Health Sciences Center and [3]Ontario Institute for Cancer Research, Toronto, ON, Canada

Supplemental data and methods

Somatic structural variant and copy number alterations

Allele-specific and total copy number was inferred by Sequenza (3.0.0)[1] using matched germline control samples. Sequenza was also used to infer tumor ploidy and purity for all samples. The solution resulting in the highest purity estimate was chosen. Structural variants and large indels were called using Manta (1.6.0). The annotables package was used to obtain hg19 gene symbols and coordinates (https://github.com/stephenturner/annotables). Reported copy number events include those across chromosomes 1-22. Gene level analysis was reported based on 52,955 annotated genes on those chromosomes. To evaluate genes impacted by somatic copy number events, Sequenza output was merged with gene coordinates to obtain a list of genes impacted by amplifications or deletions for each sample.

Visualization of putative ancestral, shared and private mutations

Based on mutation, CNA and SV calls, putative ancestral somatic events were described as those that were successfully identified across all samples in a given patient. For SNVs and indels, these included events that were found in the same position with the same alternative allele. For CNAs, we visualized these events in multiple ways. First, we visualized the percentage of the genome altered across each patient using genome wide Sequenza CNA calls. Next, we merged copy number amplification and deletion call coordinates across samples in each patient to identify regions altered in all samples. Lastly, we also looked at the overlap of genes altered by copy number amplification and deletions across samples in each patient. For SVs, we first also summarized the number of Manta calls for each sample. Putative ancestral translocations found across all samples in a patient were identified as those affected the same chromosomes and genes.

Circos plots were generated using the R package BioCircos (https://github.com/lvulliard/BioCircos.R). All visualizations were created in R version 4.0.0.

ctDNA sample collection and sequencing

Plasma samples were collected ante-mortem (Patient 1) or at autopsy (Patients 2 and 3). ctDNA was extracted using the QIAamp Circulating Nucleic Acid Kit (Cat# 55114, Qiagen) and 150 ng were used as input for library preparation and capture using IDT LockDown probes targeting the entire coding or hotspot sequences of 80 genes recurrently mutated in lymphoma (158 kB). Unique molecular identifiers (UMIs, 4 mer DNA sequence followed by a T) were incorporated during library construction. Resulting libraries were then deep-sequenced using 2 x 100 bp paired-end reads.

ctDNA data processing and variant calling

FASTQ files were processed through ConsensusCruncher.[2] UMIs were used to obtain consensus DNA sequences. ConsensusCruncher was run with default parameters and the distribution of read depth across families was evaluated. Duplex and single strand consensus and singleton corrections were generated in the form of mapped and corrected bam files in each ConsensusCruncher category (dcs_SC, sscs_SC, dcs_sc_uniq). Read depth in panel regions was evaluated using CollectTargetedPcrMetrics (GATK4.1.8.1). The corrected bam files were adopted as reads input with matched normal controls and gnomad population control to call the somatic SNVs and indels by Mutect2 (GATK4.1.8.1) in each ConsensusCruncher category, and then followed with a filtering step suggested by GATK Best Practices Workflows (GetPileupSummaries;

CalculateContamination; FilterMutectCalls). Only those variants with the 'PASS' flag were kept in the final call set.

ctDNA and bulk tumor comparison

Mutations identified in ctDNA samples were compared to mutations identified from bulk tumor sequences across WGS samples in each patient. To minimize the bias between these two variant call sets, the WGS somatic variants were re-called by Mutect2 restricted in the same interval bed regions as ctDNA panel used, the parameters and filtering steps were the same as ctDNA dataset. We calculated the number of overlapped variants between ctDNA and WGS call sets and also the variants presented only in one type of samples but not another by the tool vcf-compare (vcftools v0.1.15).[3] Furthermore, we generated a merged VCF file to report the per-site genotypes among the samples with distinct backgrounds for each patient. The single sample VCF files were first created for three ctDNA classes and all WGS tumor samples in each patient, then merged into a multi-sample VCF file by bcftools (samtools v1.10).[4]

Phylogeny reconstruction using mutations patterns from spatially distinct samples

We used Treeomics to reconstruct the phylogeny of inter-regional tumor samples in each patient.[5] Treeomics infers robust phylogenetic trees while accounting for artifacts in sequencing and mutation calls, aiming to infer ancestral subclones that gave rise to tumors at various anatomical regions. Importantly, Treeomics assumes monoclonal seeding where polyclonal seeding and reseeding of metastases occur rarely. The maximum number of mutation patterns was set to five and whole exome filtering was applied to reduce run time for patients 2 and 3.

Estimating mutation cellular prevalence and clustering

We evaluated all mutations called in at least one sample in Patient 1 and a subset of mutations (non-synonymous) from Patient 2 and Patient 3 based on several criteria to improve efficiency of the analysis. Only mutations with major copy number greater than 0 and total copy number greater or equal to 2 were included for all patients. Bam-readcount was used to identify the reference and alternative alleles for each mutation across each sample in a given individual as well as their associated read counts.[6] SNVs and allele-specific copy number and purity were used to estimate cancer cell fractions for each mutation and these were clustered to infer tumor clones that had similar cellular prevalence. This analysis allowed for estimating clusters of similarly prevalent mutations while also highlighting likely ancestral mutations that were missed during the merging of mutations called by Strelka and Mutect2. The final numbers of unique mutations evaluated in this analysis for each patient were 3,146, 22,402 and 12,024 for patients 1, 2 and 3, respectively. Pyclone-VI[7] was then used to infer the clonal population structure in each patient. 'Pyclone-vi fit' was run with a maximum number of clusters set to 20 with beta-binomial probability density and with 50 restarts. Cellular prevalence values were estimated for each mutation across predicted clusters.

Clonal evolution reconstruction

Once clusters of mutations with similar cellular prevalence were determined by Pyclone-VI, Pairtree[8] was implemented to reconstruct each patient's clonal evolutionary history. Clusters with at least 50 (patient 1) or 100 mutations (patient 2 and 3) were considered and Pairtree was run with 4,000 trees per chain on Pyclone-VI clusters. Tree solutions were investigated and if the top scoring tree was polytumor, the next best solution with a single tumor initiating ancestor was

chosen for further analysis. The clone diversity index and the clone and mutation diversity index were also estimated where the former measures how many distinct cell subpopulations are present in each sample while the latter extends this measure to include information about tumor mutation burden.

Defining disease genes and mutations

We obtained a list of genes found to be previously mutated in MCL,[9] PMBCL[10] and DLBCL[11] from relevant publications. Further, when describing mutations in these genes as likely driver events, we focused only on those in coding regions. We also considered coding mutations previously annotated in COSMIC as potential driver events.

Mutation signature analysis

We used the MutationalPatterns package[12] to compare mutations observed across predicted cell populations (Subclones) to known COSMIC signatures.[13] We used strict signature refitting to assess the contribution of merged sets of single-base substitution (SBS) signatures. SBS signatures were first merged based on a cosine similarity cutoff of 0.5.

Supplemental references

1. Favero F, Joshi T, Marquard AM, et al. Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. Ann Oncol 2015;26(1):64–70.

2. Wang TT, Abelson S, Zou J, et al. High efficiency error suppression for accurate detection of low-frequency variants. Nucleic Acids Res 2019;47(15):e87.

3. Danecek P, Auton A, Abecasis G, et al. The variant call format and VCFtools. Bioinformatics 2011;27(15):2156–2158.

4. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009;25(16):2078–2079.

5. Reiter JG, Makohon-Moore AP, Gerold JM, et al. Reconstructing metastatic seeding patterns of human cancers. Nat Commun 2017;814114.

6. genome. GitHub - genome/bam-readcount: count DNA sequence reads in BAM files. https://github.com/genome/bam-readcount (accessed November 27, 2021).

7. Gillis S, Roth A. PyClone-VI: scalable inference of clonal population structures using whole genome data. BMC Bioinformatics 2020;21(1):571.

8. Wintersinger JA, Dobson SM, Stein LD, Dick JE, Morris QD. Reconstructing complex cancer evolutionary histories from multiple bulk DNA samples using Pairtree. bioRxiv 2021;2020.11.06.372219.

9. Beà S, Valdés-Mas R, Navarro A, et al. Landscape of somatic mutations and clonal evolution in mantle cell lymphoma. Proc Natl Acad Sci U S A 2013;110(45):18250–18255.

10. Mottok A, Hung SS, Chavez EA, et al. Integrative genomic analysis identifies key pathogenic mechanisms in primary mediastinal large B-cell lymphoma. Blood 2019;134(10):802–813.

11. Reddy A, Zhang J, Davis NS, et al. Genetic and Functional Drivers of Diffuse Large B Cell Lymphoma. Cell 2017;171(2):481–494.e15.

12. Blokzijl F, Janssen R, van Boxtel R, Cuppen E. MutationalPatterns: comprehensive genome-wide analysis of mutational processes. Genome Med 2018;10(1):33.

13. Alexandrov LB, Kim J, Haradhvala NJ, et al. The repertoire of mutational signatures in human cancer. Nature 2020;578(7793):94–101.

Supplementary Figure legends

Supplementary Figure 1: Additional analysis of patient 1 (mantle cell lymphoma, MCL) (**A**) One likely reconstruction of phylogeny using Treeomics and patterns of single nucleotide variants (SNVs). Driver genes are labeled along relevant edges. The tree is not to scale and the number of mutations gained in each sample is labeled. (**B**) Tree constrained subclonal frequencies inferred by pairtree across three samples. (**C**) Summary of coding driver gene and COSMIC annotated mutations across each subclone for patient 1. (**D)** Relative contribution of mutation types across pairtree inferred clones. (**E**) Relative contribution of correlated COSMIC annotated mutation signatures across each subclone.

Supplementary Figure 2: Additional analysis of patient 2 (primary mediastinal B-cell lymphoma, PMBCL). (**A**) One likely reconstruction of phylogeny using Treeomics and patterns of single nucleotide variants (SNVs). Driver genes are labeled along relevant edges. The tree is not to scale and the number of mutations gained in each sample is labeled. (**B**) Tree constrained subclonal frequencies inferred by pairtree across three samples. (**C**) Summary of coding drive gene and COSMIC annotated mutations across each subclone for patient 1. (**D)** Relative contribution of mutation types across pairtree inferred clones. (**E**) Relative contribution of correlated COSMIC annotated mutation signatures across each subclone.

Supplementary Figure 3: Additional analysis of patient 1 (high-grade B-cell lymphoma, HGBCL). (**A**) One likely reconstruction of phylogeny using Treeomics and patterns of single nucleotide variants (SNVs). Driver genes are labeled along relevant edges. The tree is not to scale and the number of mutations gained in each sample is labeled. (**B**) Tree constrained subclonal

frequencies inferred by pairtree across three samples. (**C**) Summary of coding driver gene and

COSMIC annotated mutations across each subclone for patient 1. **(D)** Relative contribution of

mutation types across pairtree inferred clones. (**E**) Relative contribution of correlated COSMIC

annotated mutation signatures across each subclone.

Supplementary Figure 1
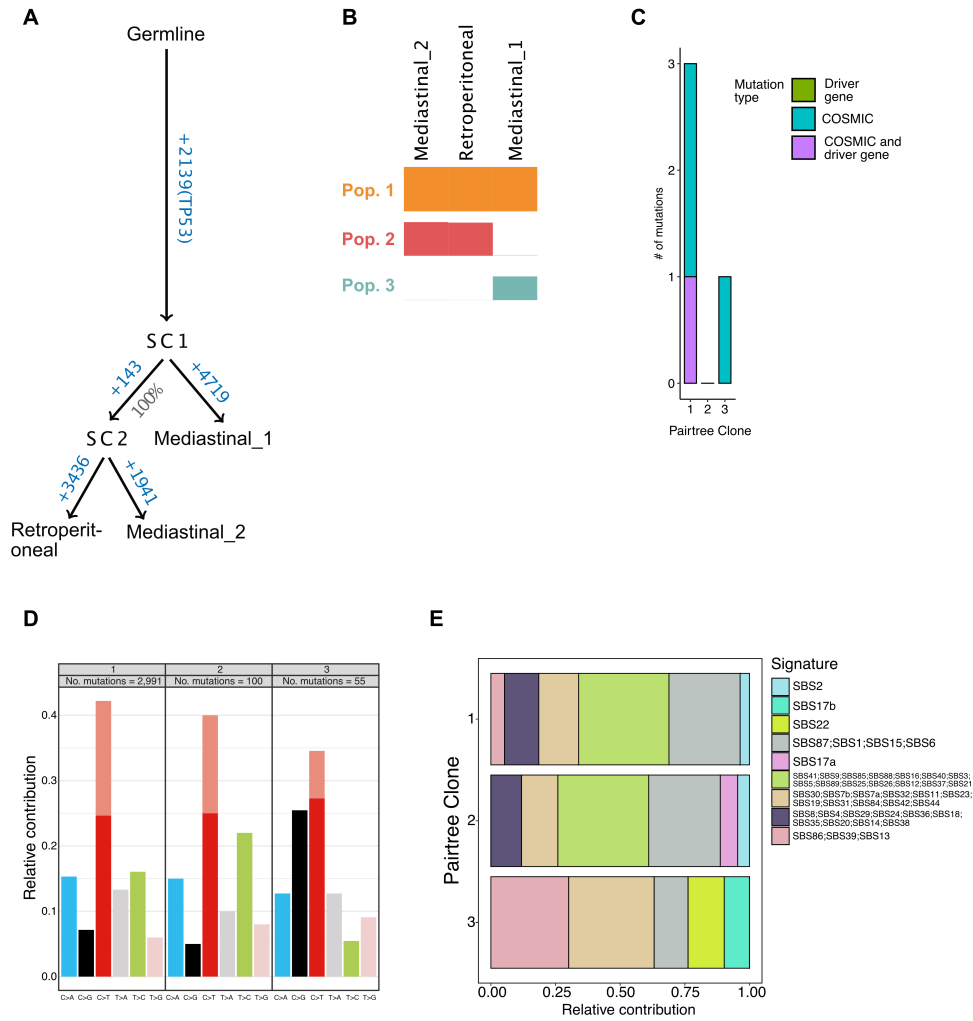
Supplementary Figure 2

Supplementary Figure 3
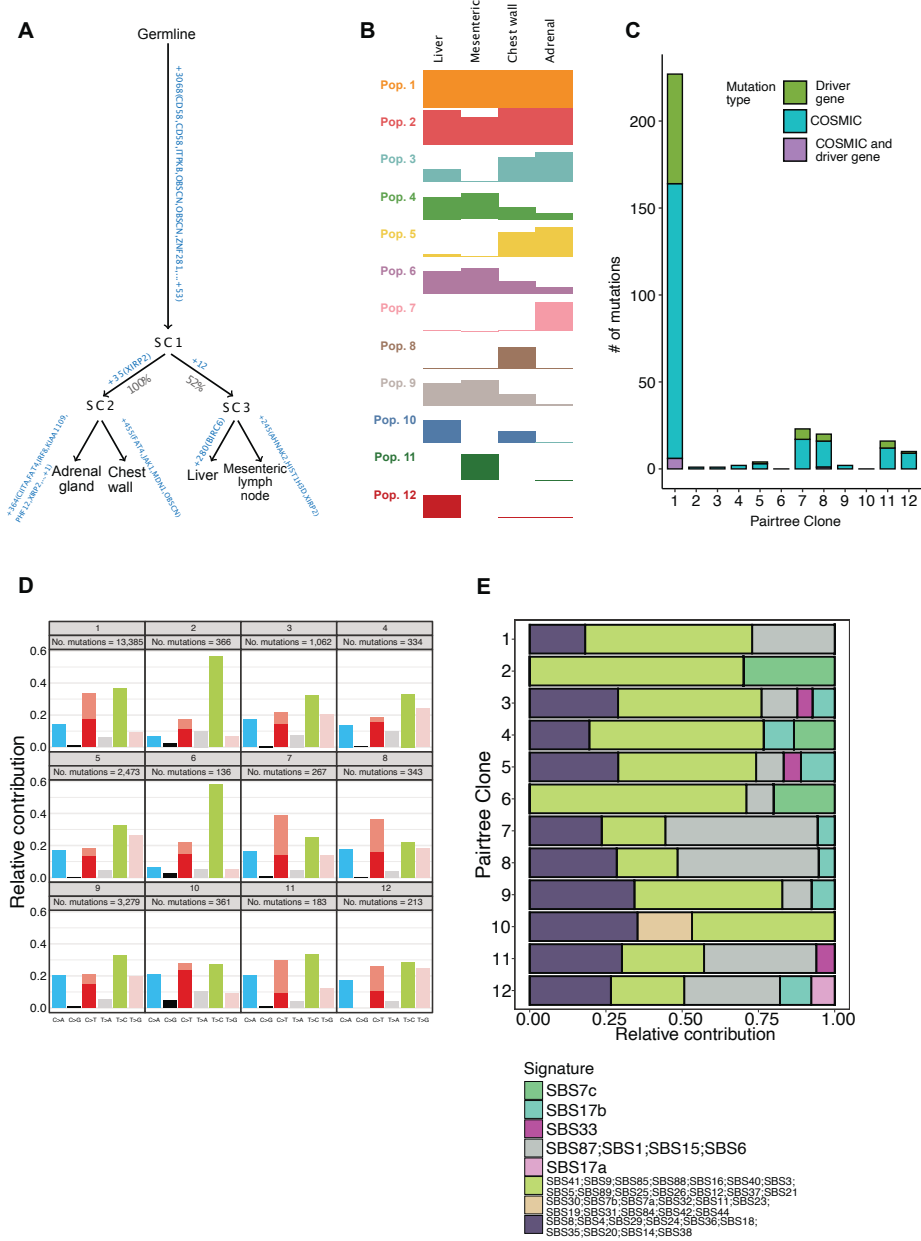
Supplementary Table list

Supplementary Table 1: Overview of samples.

Supplementary Table 2: Whole genome sequencing summary statistics for 27 tumor samples and three control samples.

Supplementary Table 3: Coding mutations identified using whole-genome sequencing across three patients.

Supplementary Table 4: Mutations across Pyclone-VI and Pairtree clones.

Tables available as excel files