Supplementary materials

Supplementary methods

NGS

cfDNA was isolated from plasma (median, 2.5 mL; range, 0.6-5.3 mL) using the AVENIO cfDNA isolation kit (Roche Molecular Systems, Inc., Branchburg, USA). After cfDNA isolation, double-stranded DNA (dsDNA) was quantified by the Qubit High Sensitivity dsDNA kit (Fischer Scientific, New Hampshire, USA), and the proportion of cfDNA samples (as opposed to contaminating high molecular weight DNA) was determined using a quantitative PCR-based cfDNA quality assessment. Up to 50 ng of cfDNA (average, 30.1 ng) was used as input for sequencing.

Library preparation and NGS were completed using a modified version of the AVENIO ctDNA analysis workflow (for research use only; not for use in diagnostic procedures), based on previously described CAPP-Seq technology [1, 2]. DNA was prepared for ligation and unique molecular identifier (UMI)-containing adapters were ligated onto the DNA fragments; the library was then amplified with universal PCR primers targeting the adapters which also contain unique dual index sample indices. Half the PCR product for each sample was captured with a ~314 kb panel designed to cover regions relevant for COO determination and minimal residual disease in DLBCL [2]. Each library was amplified to get a final sequencing library for each sample. Libraries were then pooled and sequenced on the HiSeq4000 (illumina, California, USA) with 2×150+2×8 reads, to an average of 60 million clusters per sample (range, 27-108 million clusters).

Variant calling

Variant callers were based on previously described algorithms for ctDNA variant calling [1, 3]. Variants were annotated using SnpEff (version 4.2). In total, 164637 unfiltered SNVs and 15396 indels were detected in plasma samples. As matched normal samples were unavailable for these samples, a combination of filtering schemes was used to exclude likely non-tumor-specific variants.

Variant filtering

Common variants in the Single Nucleotide Polymorphism Database (dbSNP) or variants with an AF > 0.1% in any of 1000 Genomes project (1KG) or Exome Aggregation Consortium (ExaC) populations were filtered out unless they were reported by the Catalogue of Somatic Mutations in Cancer (COSMIC) or the Cancer Genome Atlas Program (TCGA). Variants that were detected in low complexity regions of the genome, as well as regions known to contain immunoglobulin genes, were filtered out. Variants that were commonly detected using our targeted panel on healthy normal plasma samples were also discarded. For this blacklist filter, 22 healthy donor samples were sequenced with our workflow, and any variants occurring in more than two healthy donors with more than five supporting reads were excluded from consideration in patients with DLBCL. Finally, variants with low coverage (a depth of < 25% of median sample depth) were filtered out. After filtering, 58961 SNVs and 2954 indels were detected in plasma samples.

Tumor burden estimation

Mutant molecule per mL (MMPM) incorporates the allele fractions of the variant calls and the cell-free DNA (cfDNA) mass of the sample, specifically MMPM = (mean allele frequency [AF, 0-1 range] × extracted mass [ng] × 330 [copies/ng]) / plasma volume (mL).

Variant call concordance between tissue and plasma

As identical indel calls can be reported differently due to differences in methods and nomenclatures, the indel calls from tissue that were not detected in plasma were checked manually against possible indel call candidates detected in plasma. In total, 33 indels were detected that were called differently due to differences in alignment strategies and reporting (Supplementary Table 2). The concordance of short variant calls between tissue and plasma for patients by disease stage and plasma sample input mass was then analyzed and evaluated using a logistic regression model and the Cochran-Armitage test.

Prognostic associations

COO labels and corresponding LPS were predetermined from tissue samples using the NanoString Lymph2Cx assay. We analyzed the prognostic value of known molecular markers detected at baseline (variants in BCL2, BCL6, CARD11, CD79B, MYC, MYD88, and TP53)[4, 5, 6] using Kaplan–Meier plots to assess PFS, and estimated hazard ratios (HRs) using multivariate Cox regression (stratified by: immunochemotherapy regimen [G-CHOP vs R-CHOP], number of planned chemotherapy cycles, geographic region, and IPI score). A multivariate Cox regression analysis containing only MMPM and sum of the product of the greatest diameters (SPD) was performed to determine the prognostic value of MMPM on PFS after accounting for SPD.

Prognostic modeling using a combination of ctDNA features was also evaluated. For simplicity, all biomarkers were simplified to high or low risk (or high, intermediate and low for IPI); high-risk MMPM was defined as MMPM above the cohort median (315.85), high-risk COO was defined as non-GCB, and high-risk Chapuy was defined as S2, S3, or S5. The total number of high-risk biomarkers was then tabulated for each patient.

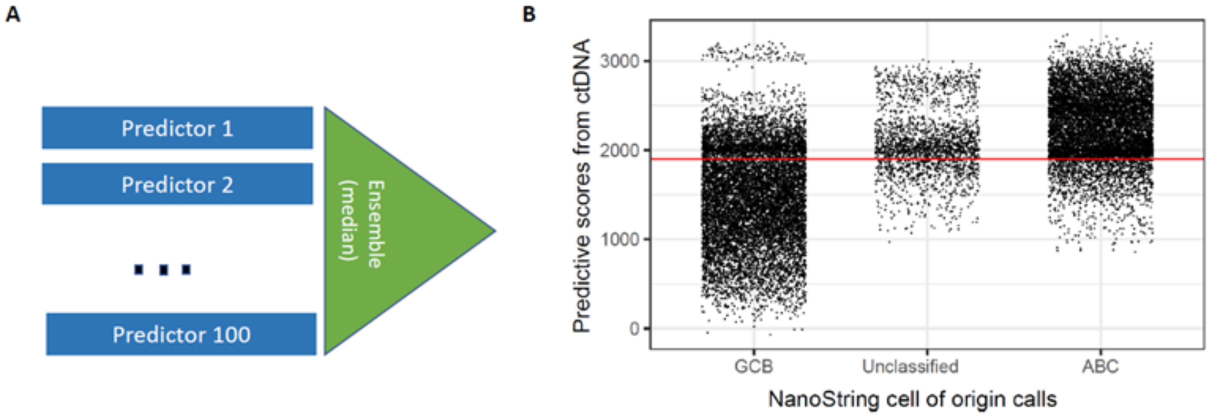
Supplementary tables

Supplementary Table 1. Sample information.

Supplementary Table 2. Variant calls for (A) somatic SNVs, (B) indels and, (C) translocations.

Supplementary figures

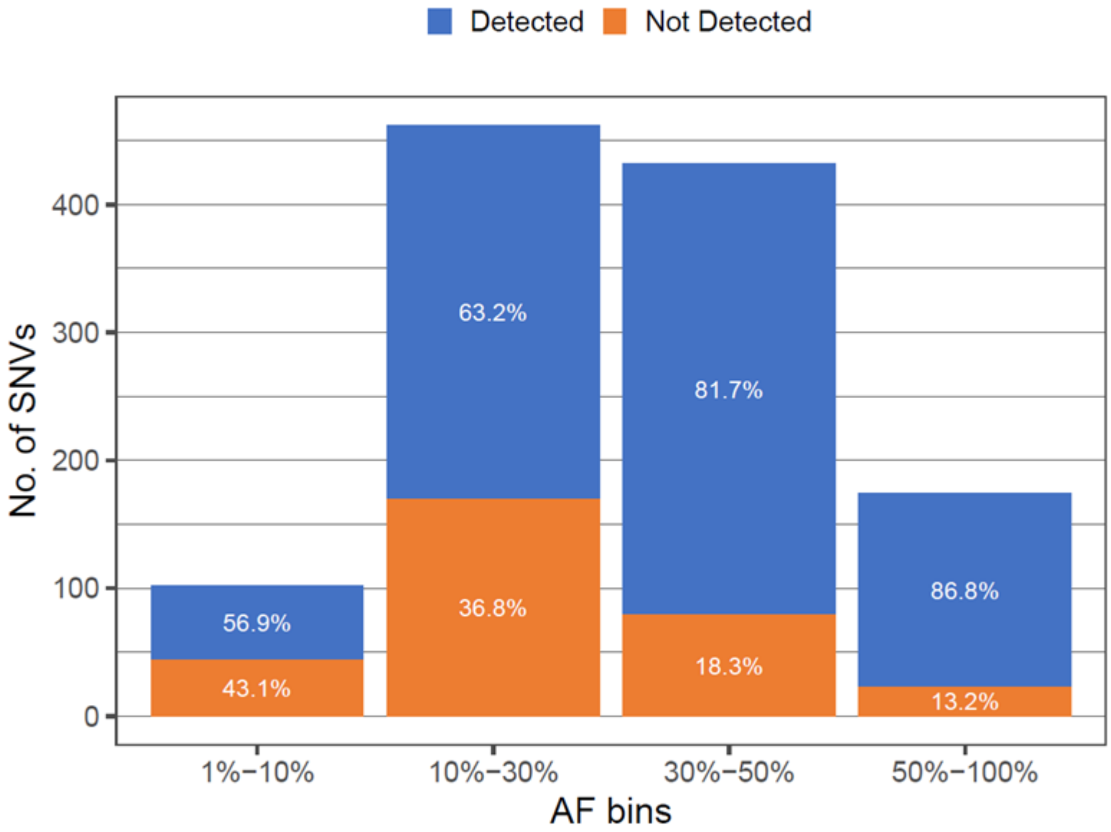
Supplementary Figure 1. Machine learning model for COO determination from plasma variant calls. (A) High-level diagram of the machine learning approach. Twenty-two features were fed into an ensemble of 100 XGBoost predictors to give predicted scores, with higher scores corresponding to ABC and lower scores corresponding to GCB. (B) Range of scores estimated from each of these predictors for samples in the training set.



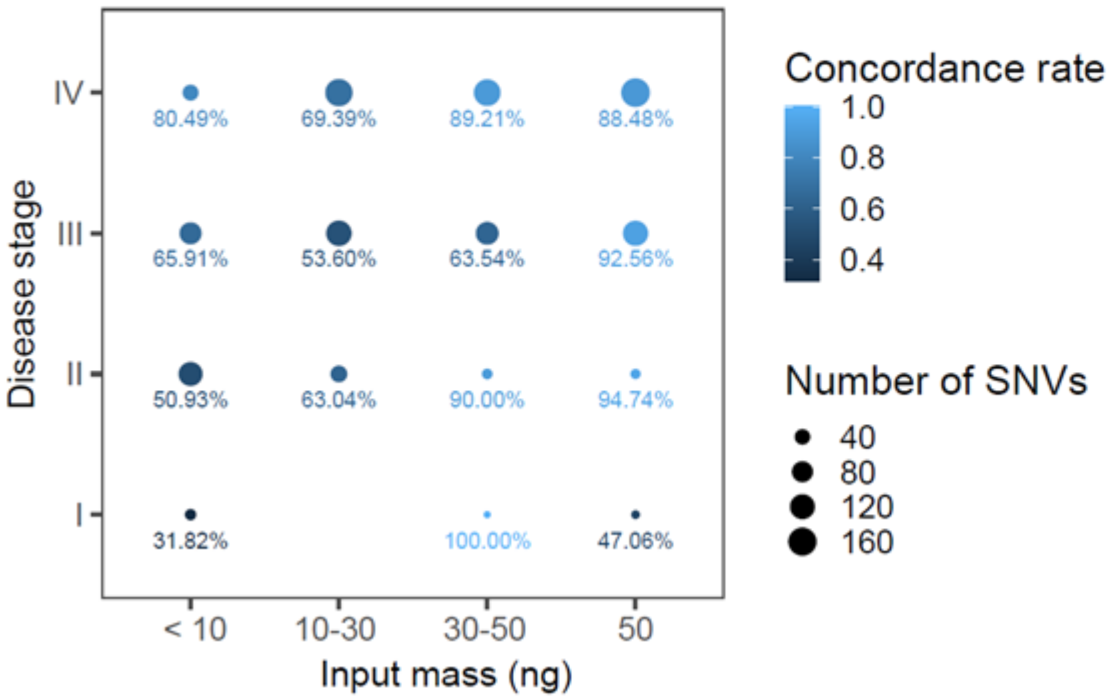
ABC, activated B-cell like; COO, cell of origin; ctDNA, circulating tumor DNA; GCB, germinal center B-cell like.

Supplementary Figure 2. PPA of SNV calls between tissue and plasma by allele frequency

AF, allele frequency; PPA, positive percentage agreement; SNV, single nucleotide variant

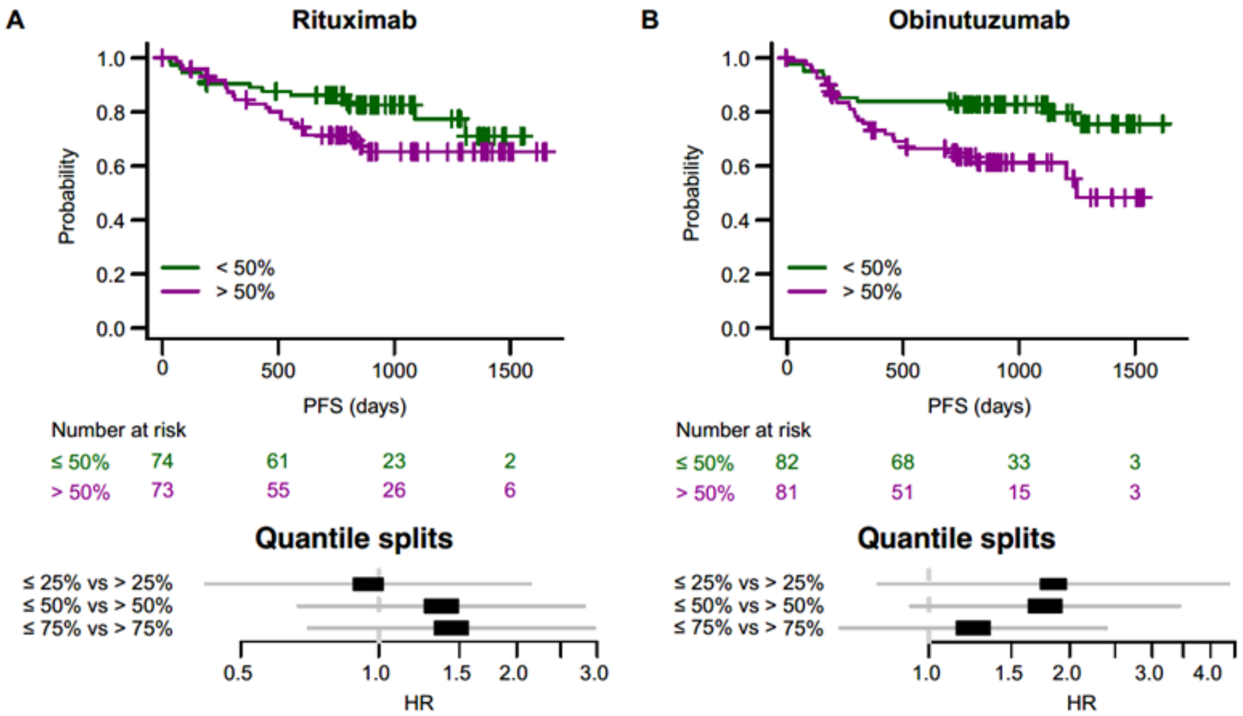


Supplementary Figure 3. Impact of cfDNA input mass and Ann Arbor stage on PPA between tissue-based and plasma-based SNV calls.



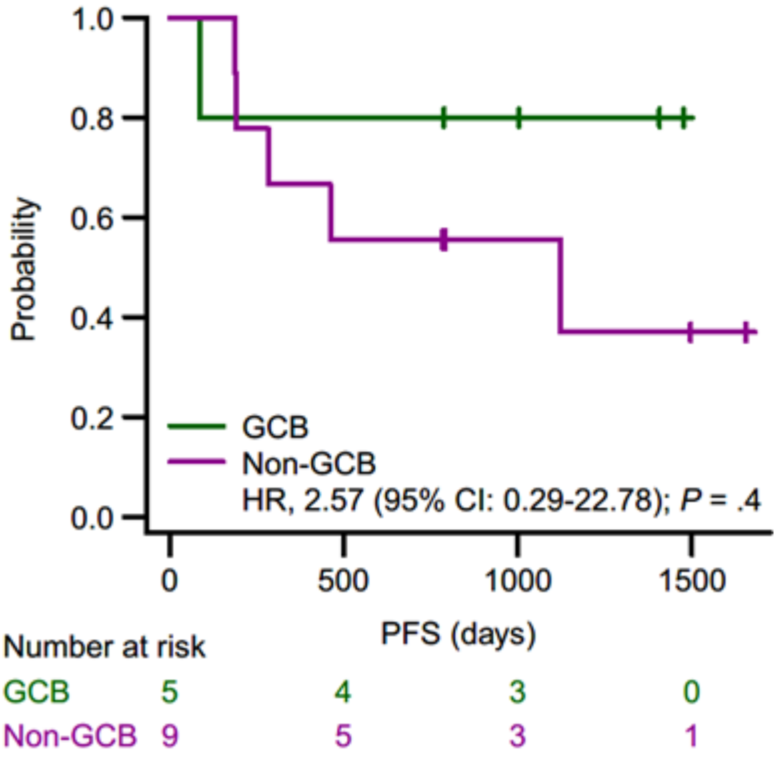
cfDNA, cell-free DNA; PPA, positive percent agreement; SNV, single nucleotide variant.

Supplementary Figure 4. Association between tumor burden assessment from plasma (MMPM) and prognosis for each treatment regimen. PFS according to median MMPM values for patients treated with (A) rituximab or (B) obinutuzumab. HRs and 95% CIs for quantile splits are shown below each plot for MMPM split points of 25%, 50%, and 75%.



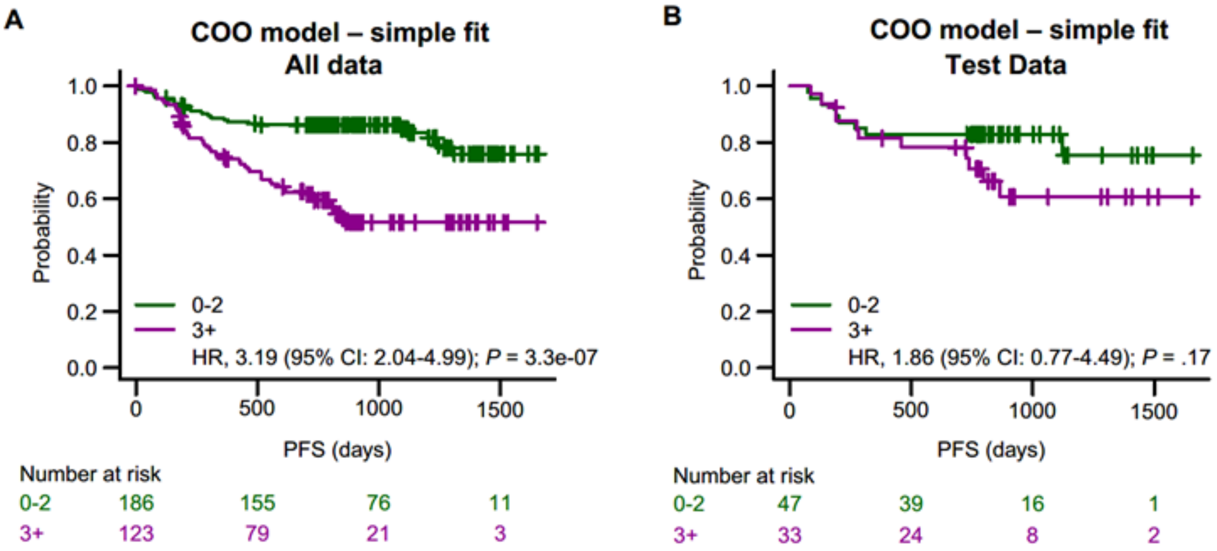
CI, confidence interval; HR, hazard ratio; MMPM, mutant molecule per mL; PFS, progression-free survival.

Supplementary Figure 5. KM curve showing PFS for the validation set split by COO classification as determined by the plasma-based machine learning method for patients with COO unclassified by tissue-based NanoString Lymph2Cx (n = 14).



CI, confidence interval; COO, cell of origin; GCB, germinal center B-cell like; HR, hazard ratio; KM, Kaplan–Meier; PFS, progression-free survival.

Supplementary Figure 6. Multi-modal model for prognostic prediction using MMPM, TP53 status, BCL2/MYC translocation status, COO calling from NanoString Lymph2CX on FFPE tissue samples, and IPI score. KM curves showing PFS for high- versus low-risk features are shown in (A) the full dataset (N = 309), and (B) the test dataset (n = 80).



CI, confidence interval; COO, cell of origin; FFPE, formalin-fixed paraffin-embedded; HR, hazard ratio; IPI, International Prognostic Index; KM, Kaplan–Meier; MMPM, mutant molecule per mL; PFS, progression-free survival.

References

1. Newman AM, Lovejoy AF, Klass DM, et al. Integrated digital error suppression for improved detection of circulating tumor DNA. Nat Biotechnology. 2016;34(5):547-555. doi: 10.1038/nbt.3520.

2. Kurtz DM, Scherer F, Jin MC, et al. Circulating tumor DNA measurements as early outcome predictors in diffuse large B-cell lymphoma. J Clin Oncol. 2018;36(28):2845-2853. doi: 10.1200/JCO.2018.78.5246.

3. Newman AM, Bratman SV, Stehr H, et al. FACTERA: a practical method for the discovery of genomic rearrangements at breakpoint resolution. Bioinformatics (Oxford, England). 2014;30(23):3390-3. doi: 10.1093/bioinformatics/btu549.

4. Monti S, Chapuy B, Takeyama K, et al. Integrative analysis reveals an outcome-associated and targetable pattern of p53 and cell cycle deregulation in diffuse large B cell lymphoma. Cancer Cell. 2012;22(3):359-72. doi: 10.1016/j.ccr.2012.07.014.

5. Xu-Monette ZY, Wu L, Visco C, et al. Mutational profile and prognostic significance of TP53 in diffuse large B-cell lymphoma patients treated with R-CHOP: report from an International DLBCL Rituximab-CHOP Consortium Program Study. Blood. 2012;120(19):3986-96. doi: 10.1182/blood-2012-05-433334.

6. Schuetz JM, Johnson NA, Morin RD, et al. BCL2 mutations in diffuse large B-cell lymphoma. Leukemia. 2012;26(6):1383-90. doi: 10.1038/leu.2011.378.