Supplemental Materials Accompanying Panea et al, Blood 2019

External validity:
Our paper was published in the same year as Grande et al (Blood 2019). That paper provides useful context for us to evaluate the external validity of our approach and results.

The results about the landscape of BLs from our paper are very similar to results described in other papers. The top driver genes in our study (e.g. MYC, BACH2, IGLL5, BACH2, SIN3A, BCL7A, DDX3X, BTG2, FBXO11, FOXO1, ID3) are also supported by the data in Grande et al (Blood 2019). Some of the genes that we describe as drivers (e.g. HNRNPU) were validated in later publication(s). Beyond discovery of drivers, we also pursued an in-depth functional approach to validation that included a CRISPR screen to validate drivers, and a detailed characterization of a novel mouse model that illuminates the role of ID3 as a tumor suppressor in vivo.

	Grande et al	Our paper
Number of Cases	91	101
All BL subtypes included?	No (HIV not included)	Yes, all three subtypes included
Additional sequencing done	Targeted	WES + RNAseq
Transcriptomics	No	Yes
Functional Validation	No	Yes
CRISPR screen	No	Yes
Mouse model	No	Yes

Thus, our approaches are similar to others used in the field and have generated results that are highly overlapping with other genomic studies of BL.

Discussion of Coverage
This discussion was previously included in response to the letter to the editor. We reproduce it as it provides the context for understanding the tradeoffs that our (and every study) entails. While we targeted a maximum of 75X in our sample preparation, the actual coverage achieved can vary widely based on the specific alterations in the sample, mappability of the reads, quality of the sample and sequencing run itself.

We should point out that there is no single standard for coverage of whole genomes. For instance, we used the widely cited 1000 genomes project as controls—most of those genomes were sequenced at 5X coverage at the time of our analysis. The specific question of coverage and variant discovery was addressed by Bentley et al (Nature 2008). As shown in the adjoining figure from that paper, even 5X and 10X coverage enables the identification of over 30% and 60% of all variants. Other whole genome studies have used different levels of coverage—e.g. Sudmant et al (Nature 2015) used a coverage of 7.4X. Whole genome sequencing does not refer to a particular coverage, but the actual technique employed.

Ideally, the coverage would be determined by study goals. If the goal is to discover all variants in a specific patient (e.g. clinical sequencing), 100X coverage is inadequate. FDA and NY Dept. of Health guidelines

recommend 500X coverage for clinical sequencing. 100X coverage is also inadequate for systematic discovery of all structural alterations. For instance, MYC translocations, which are a defining feature of Burkitt lymphoma, cannot be reliably detected (AUC>0.95) at 100X coverage in those cases. Finally, there are limitations imposed by the short read format of the Illumina sequencer itself that further limits the sensitivity and specificity of whole genome sequencing.

Given the rareness of Burkitt lymphoma, we believe that a case with lower than “ideal” levels of coverage still has the potential to be informative, either by confirming other events in our own study or other studies. To our knowledge, even 3 years after publication, our study remains one of the largest published sources of whole genome sequencing in HIV+ Burkitt lymphoma.

Our study was designed to identify repeated genetic events (at any genomic location) that give rise to Burkitt lymphoma. Our study was quite well-powered to detect recurrent events such as a MYD88 L265P hotspot or the described TERT promoter region mutation, and progressively less powered for other types of genetic events. Similar considerations apply to all genomic studies.

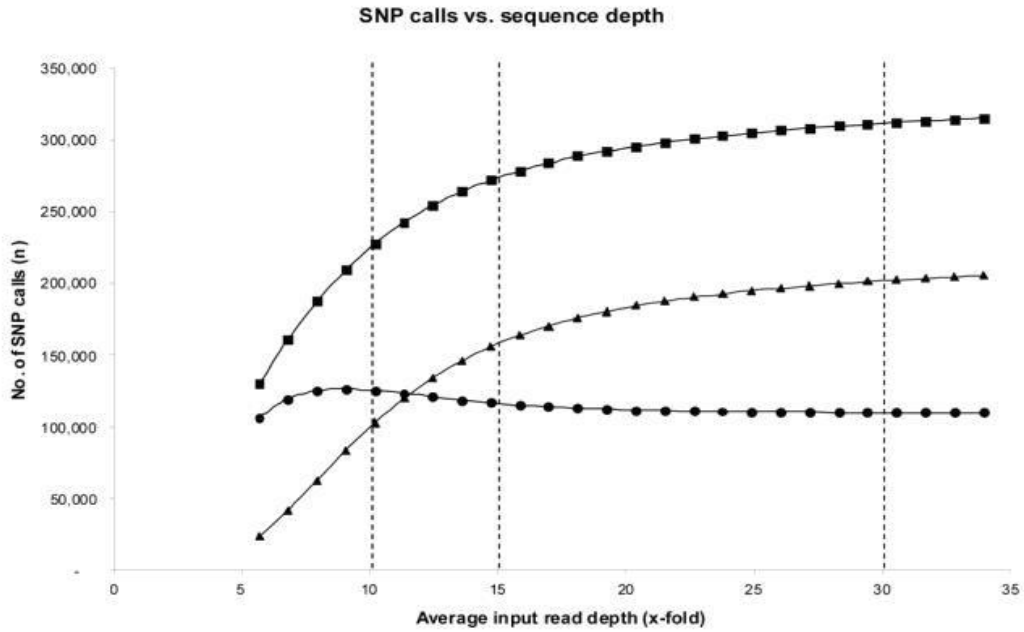

Figure 1 shows the effects of sequencing coverage on the detection of all variants (squares), heterozygous variants (triangles) and homozygous variants(circles). Even at 5X and 10X depth, a significant proportion of variants are identified. Variant identification plateaus around 30X. Source: Bentley et al, Nature 2008.

As shown in the adjoining figure, empirical evidence from Bentley et al indicates that variant identification plateaus around 30X coverage. Our study achieved an average of 28X coverage. Recently, the 1000 genomes study was updated to provide “high coverage” genomes (Byrska-Bishop et al, Cell 2022). The average coverage in that set is 30X.

Use of Whole Exome Sequencing and RNA seq data

The study was designed as a discovery study using whole genome sequencing. Indeed, whole genome sequencing was the primary discovery set for putative driver events described in Table S2 and the driver

genes in Table S5. In reviewing our results, we discovered an error where we failed to describe the use of exome and RNA sequencing data from the same samples.

All the individual variants in Table S2 were identified from whole genome sequencing data. In addition, whole exome sequencing and RNAseq data were used to identify patients in whom those variants were present, requiring a minimum of two reads supporting the variant already identified by whole genome sequencing.

RNAseq has been used previously as a primary discovery tool in lymphomas and a number of other cancers. In particular, RNAseq was the primary discovery approach used for identifying drivers in diffuse large B cell lymphoma (Morin et al, Nature 2011) and Burkitt lymphoma (Schmitz et al, Nature 2012). In our study, RNAseq was used in a much more limited fashion to identify variants that were already known to be somatically mutated in our Burkitt lymphoma cases by whole genome sequencing.

Thus, the study design resembles that used by Grande et al (Blood 2019) to identify Burkitt lymphoma drivers. Their study followed up whole genome sequencing with targeted sequencing to improve detection of variants in already discovered driver genes, whereas we followed up whole genome sequencing with whole exome and RNA sequencing. The vast majority of the results arise from whole genome sequencing.

Addressing Sample "Contamination"

We further examined all the individual sequencing files for evidence of contamination as suggested by Rushton et al (Blood 2022). We found no evidence of sample-level contamination in any of the cases.

However, we discovered an issue in the labeling of RNAseq data files (N=25, endemic BL cases) and WGS files from cases with multiple sequencing runs (N=2, sporadic BL cases). This issue led to a proportion of these samples appearing to have overlap as the counted variants originated from more than one source. These issues were resolved and the results were recalculated using the correct data. We removed variants that originated from the mismatched sample and were not supported by the mismatch-resolved data and provide this in the updated Table S2. Overall, this table is over 97% identical to the original table. We have also updated driver genes in updated Table S5. The overall set of drivers has not changed.

The genomic data have been relabeled appropriately and will be available in EGA under the same accession number as previously.

Verification of Somatic Events

In order to verify the findings of our paper, we used one of the tools we had previously used for the generation of the data that enabled verification of the results, without needing normal controls. We ran the widely used variant calling tool HaplotypeCaller (version 4.3.0) with default settings in single-sample and joint-genotype calling mode to evaluate the variants that we previously described in Table S2. We found that this tool identified over 85% of the unique variants in Table S2 from whole genome sequencing. This indicates that our identified variants are supported by whole genome sequencing data.

We further used genotyping to identify if a SNV found from discovery was also found in other samples in table S2. We identified SNVs that were found with at least two reads of support for the alternate allele and recovered over 99% of all SNVs that were in Table S2. This shows that our original approach for identifying variants is reproducible from our sequencing data. We also examined the insertions/deletions (indels) that were in table S2. As genotyping is not possible for indels, we manually reviewed each indel using Integrated Genomics Viewer and eliminated any that appeared to arise from poor alignment.

We further plotted the updated mutations and their frequencies in Figure 2 below. These findings are essentially identical to those described in the original Figure 2A.

Figure 2 shows the updated mutations after the removal of artifacts described above. Overall, 100% of the drivers remain unchanged. The patterns of mutations remain essentially identical to the original Figure 2A.

Recapitulating key associations from the paper (Figures 2, 3)

1. <u>Associations between EBV subtypes and mutation load</u>

   We recalculated the associations EBV subtypes and the updated mutation load using an ANOVA test, Krushkal-Wallis test to check for differences in the subtypes (EBV-, EBV+ Type 1, EBV+ Type2). We see a significant association (ANOVA test p=0.04 and KW test p=0.017) and similar trends to what was observed in our paper in Figure 2B. The difference of mutational load between EBV type 1 and 2 shows similar patterns as before, with higher mutational load in Type 1. However the results are not significant (ANOVA test, P=0.16), which is partly a reflection of the small number of cases that were Type 2.

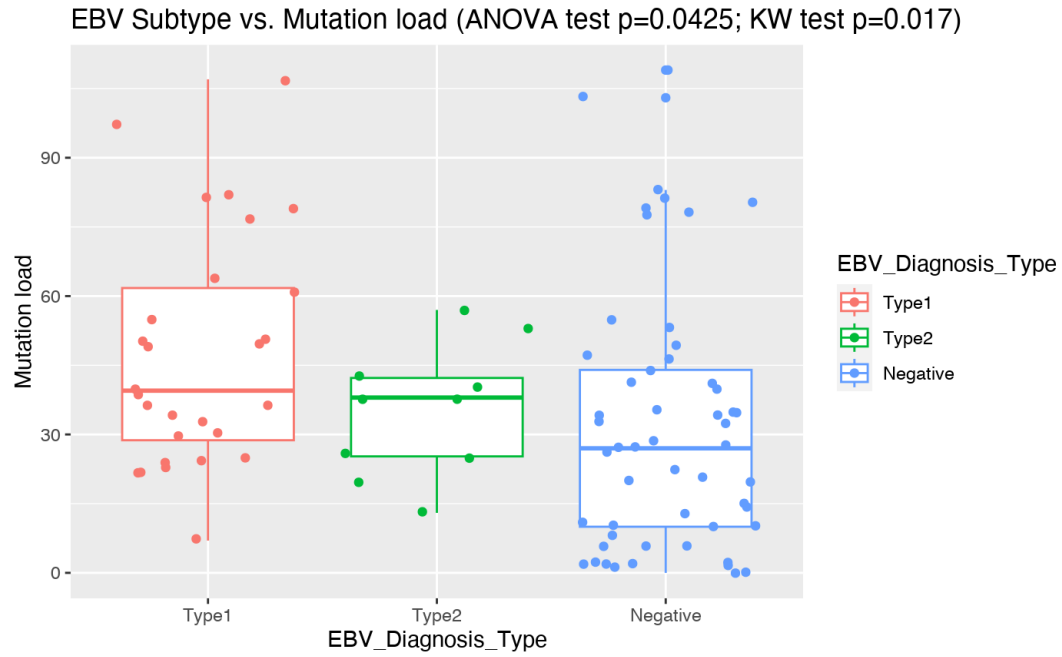

Figure 3 shows the relationship of mutational load and EBV status. EBV positive cases are associated with higher mutational load with type 1 cases showing higher mutational load compared to EBV type 2 (ANOVA test, P=0.043).

2. <u>Recapitulating Associations of Driver Genes and EBV status (Figure 2D and 2E).</u>

We tested the association with the updated gene list and found essentially identical results to Figure 2D and 2E. All the genes depicted in Figures 2D and 2E remain significant, with p-values listed in the table below.

Association of driver gene mutations with BL subtypes is shown below.

Genes	**%sporadic**	**%HIV**	**%African**	**Chisq p-value**
DNMT1	0.79	0.17	0.04	4.04E-08
SNTB2	0.76	0.24	0.00	3.16E-06
CTCF	0.70	0.20	0.10	0.00148228
BCL6	0.51	0.02	0.47	0.00783904
MME	0.77	0.19	0.04	0.000690325
CHD8	0.81	0.15	0.04	0.00113058
FBXO11	0.69	0.17	0.14	0.006918049
BCL7A	0.43	0.11	0.46	0.037754964
HNRNPU	0.76	0.15	0.09	0.001582532

Association of driver gene mutations with EBV subtypes is shown below.

Genes	%EBV+	%EBV-	Chisq p-value
IGLL5	0.51	0.49	0.07177579
BACH2	0.57	0.43	0.000353537
SNTB2	0.23	0.77	0.034120409

3. Recomputing results from integrative analysis from Figure 3B

The integrative analysis in Figure 3 provided an exploratory analysis of the relationship between different mutations and expression. The main finding was that there are two broad sets of gene sets, signaling pathways and DNA damage repair. We recalculated the integrative analysis performed in Figure 3 using updated mutation data. We recomputed the differential genes associated with driver mutations, and performed Gene Set Enrichment Analysis using Reactome gene sets and identified equivalent patterns to what was observed in Figure 3B. As before, we observed two broad clusters of overlapping gene sets - Signaling pathways (top cluster) and DNA damage (bottom cluster) and driver gene clusters (left vs. right) as seen in Figure 3B. As before, these results are meant to provide an illustration of the analyses and biological exploration that is feasible with our data. The original figure included manual curation of the pathways which was not performed here.

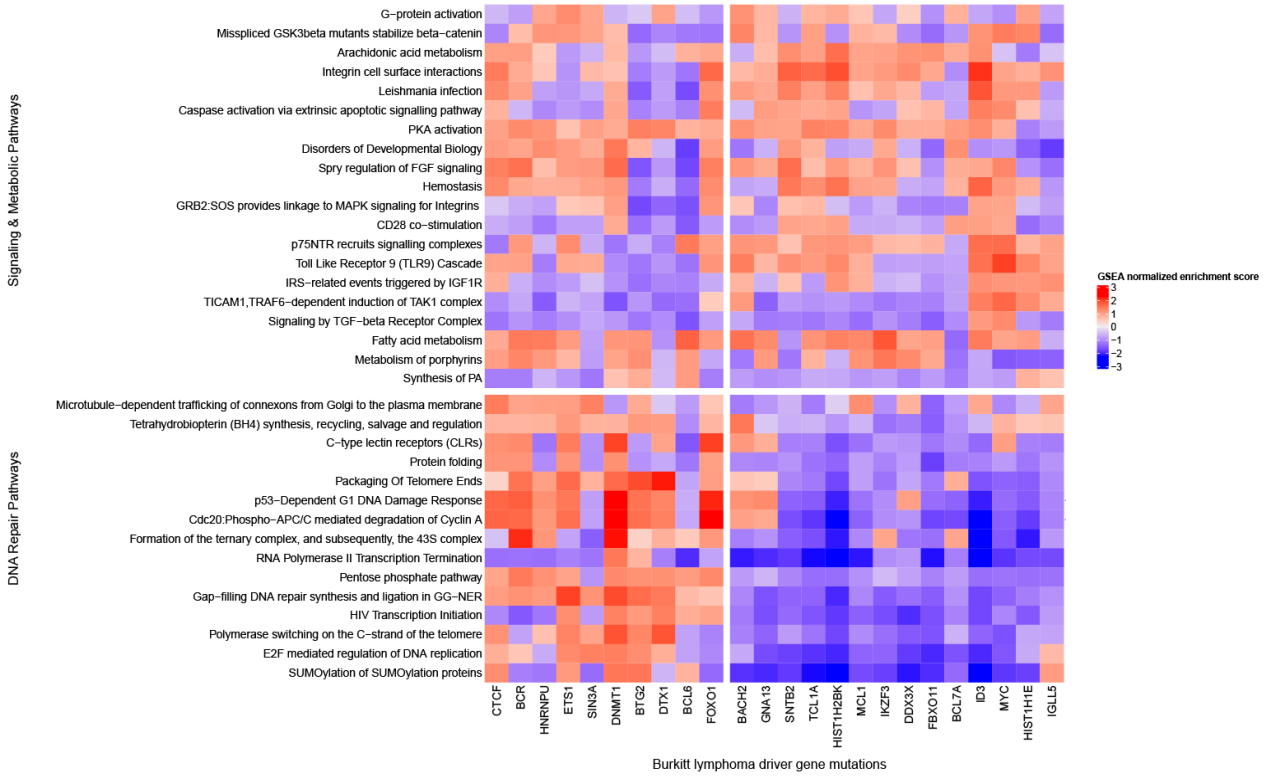

Figure 4 shows the relationship of individual mutations and expression. Gene sets associated with each mutation were averaged and clustered as described. We observed two main clusters that comprise signaling and DNA repair pathways. This is illustrative of the types of analyses enabled by our data.