blood® Letters to Blood

TO THE EDITOR:

Concerning data inconsistencies in Burkitt lymphoma genome study

Christopher K. Rushton,[1] Kostiantyn Dreval,[1,2] and Ryan D. Morin[1-3]

[1]Department of Molecular Biology and Biochemistry, Simon Fraser University, Burnaby, BC, Canada; and [2]Genome Sciences Centre and [3]Lymphoid Cancer Research, BC Cancer, Vancouver, BC, Canada

In 2019, a pair of studies described the whole-genome sequencing (WGS)–based characterization of over 100 Burkitt lymphomas (BLs).[1,2] Panea et al[2] analyzed 101 patients representing sporadic, endemic, and HIV-associated clinical variants. Some of their conclusions were consistent with Grande et al[1] and others were contradictory and/or pointed to potential novel features of this disease. Panea et al nominated some new BL-related genes and highlighted the novelty of clustered noncoding mutations. To understand the factors leading to the divergent conclusions, we reanalyzed the sequencing data from Panea et al, which revealed several irregularities in their results that suggest serious errors in data processing.

On reviewing the mutations reported in supplemental Table 2 of Panea et al (Panea_S2), we noted that 936 (40.1%) variants were attributed to at least 2 of 101 patients. It is exceedingly improbable for cancers from different individuals to share many identical mutations, with the exception of hotspot mutations, because of strong selective pressure. Although 264 variants were attributed to loci that are affected by aberrant somatic hypermutation, a process that increases the local mutation rate, most variants reported in multiple patients (672, 72%) were outside these regions and have no known biological explanation.

In a recent erratum,[2] the authors clarified that exome data from these same patients, which were merged with the WGS data before depositing, were not used for their analyses. Accordingly, we first prepared separate binary alignment map (BAM) files from their deposited data that contained only the genome or exome reads based on read group information (supplemental Table 1; supplemental Figure 1A, available on the *Blood* website). For each variant reported in Panea_S2, we computationally evaluated the genome data for any supporting evidence. We considered any variant "corroborated" if at least 1 uniquely mapped read from that sample supported the mutant allele (supplemental Figure 1B). This is more lenient than the criteria described by the authors and should theoretically corroborate every variant reported, but the existence of 1388 (~30%) variants were not corroborated by this approach using either their original BAM files or the genome BAM files (supplemental Figure 1B; supplemental Tables 2-4).

Surprisingly, many of the duplicated variants could be corroborated in 1 patient but not in another, with a consistent pattern restricted to 27 samples (Figure 1A). Using clustering, we identified the cases with a pattern of mutations that could not be corroborated in a given sample tended to be corroborated in only 1 other sample, a pattern we refer to as directional data cross-contamination. Under the assumption that this was caused by the accidental combining of sequencing data from unrelated patients during the original analysis, we combined reads from the genome of the suspected contaminant (per directional pair) in silico and repeated our analysis (supplemental Figure 1C). This enabled the resolution of 643 of these uncorroborated variants, leaving 745 unresolved. The directional contamination pattern was almost exclusively observed among the African cohort (affecting 25 of 32 cases).

Given the large disparity between the actual data and the reported mutations, we considered the possibility that the observed data contamination could have been restricted to Panea_S2 rather than having influenced any of the main results. Upon scrutiny, the mutations presented in Figure 2 of Panea et al and the counts in supplemental Table 5 (Panea_S5) appear consistent with Panea_S2, indicating that this issue affected most, if not all, of their downstream analyses.

Although data quality cannot explain the reporting of variants with 0 read support, we were interested in how comparable the data were among the genomes sequenced because batch effects such as coverage can influence estimates of global mutation burden. We used Picard (CollectWGSMetrics) to calculate the average coverage of the genomes, which showed a broad range (~1× to 27.8×, mean = 14.1×) (Figure 1B). This was surprising because the authors claimed they were "targeting a mean genome coverage of 75×." This also revealed a large disparity in sequencing depth among samples, with cases in the African cohort generally having high coverage relative to the others. We found that the variants that could be corroborated in the raw data commonly had minimal read support (Figure 1C). These discrepancies between the reported results and the raw data have implications on some conclusions. Repeating the gene-wise comparison of mutation frequency between patients with EBV^+ and EBV^- using the corroborated mutations caused a loss of significance for 1 gene and a gain of 4 genes with a significant association (Figure 1D). Panea et al reported a difference in mutation load between patients with EBV type 1 and EBV type 2 as a novel finding, a difference which was not reproduced when we used only the corroborated variants (Figure 1E).

A second intriguing pattern was found in uncorroborated mutations in the sporadic and HIV cases, affecting 17 cases.

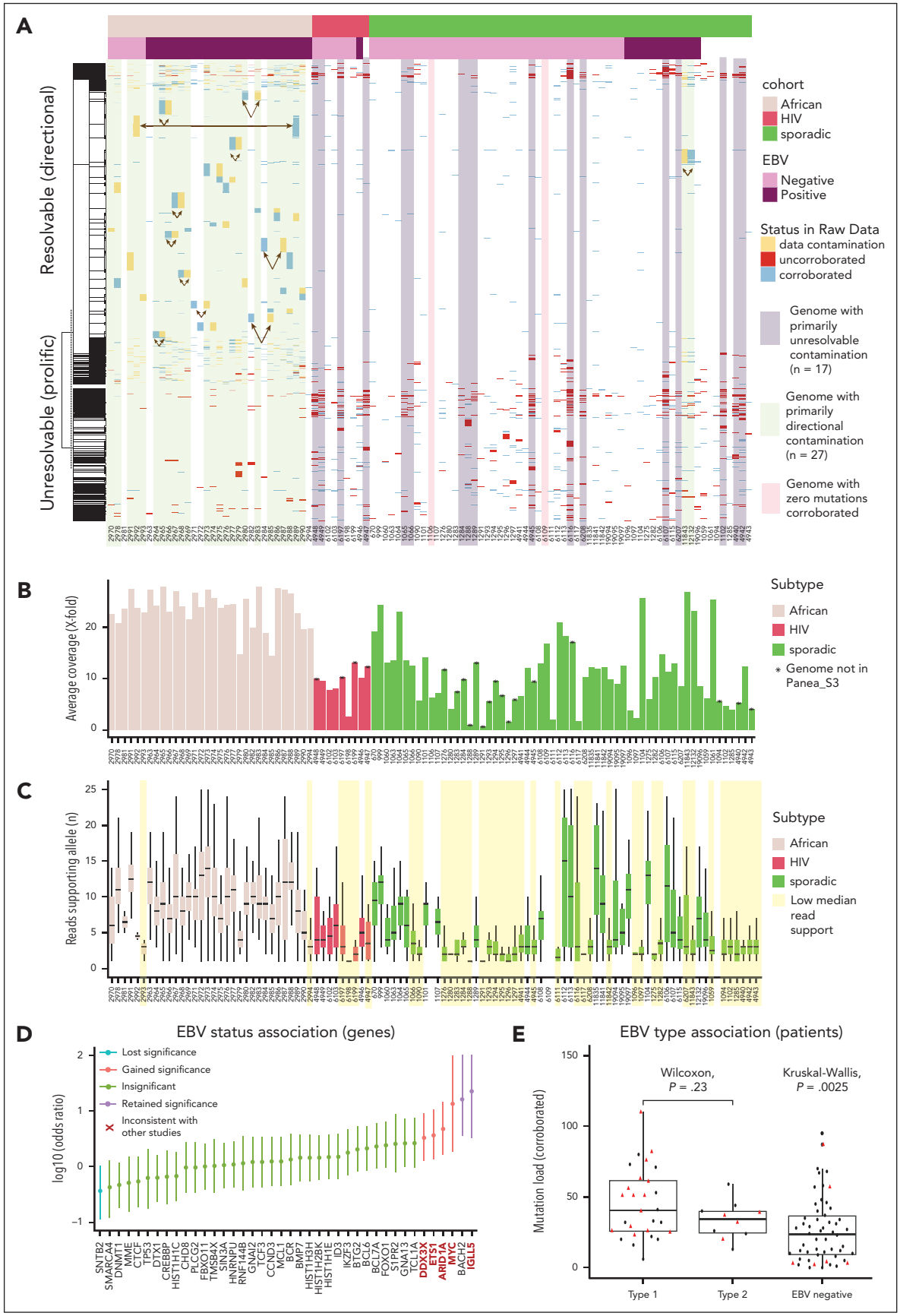

Figure 1. Patterns of uncorroborated mutations and directiona data contamination. (A) A heatmap showing a subset of the mutations reported in the 101 patients coloring each based on read support in the sequencing data from the sample specified ("corroborated," blue), uncorroborated in data from this patient but present in data from a sample identified as a potential contaminant ("data contamination," yellow), or with no read support in either ("uncorroborated," red). The patients are arranged according to their subtype

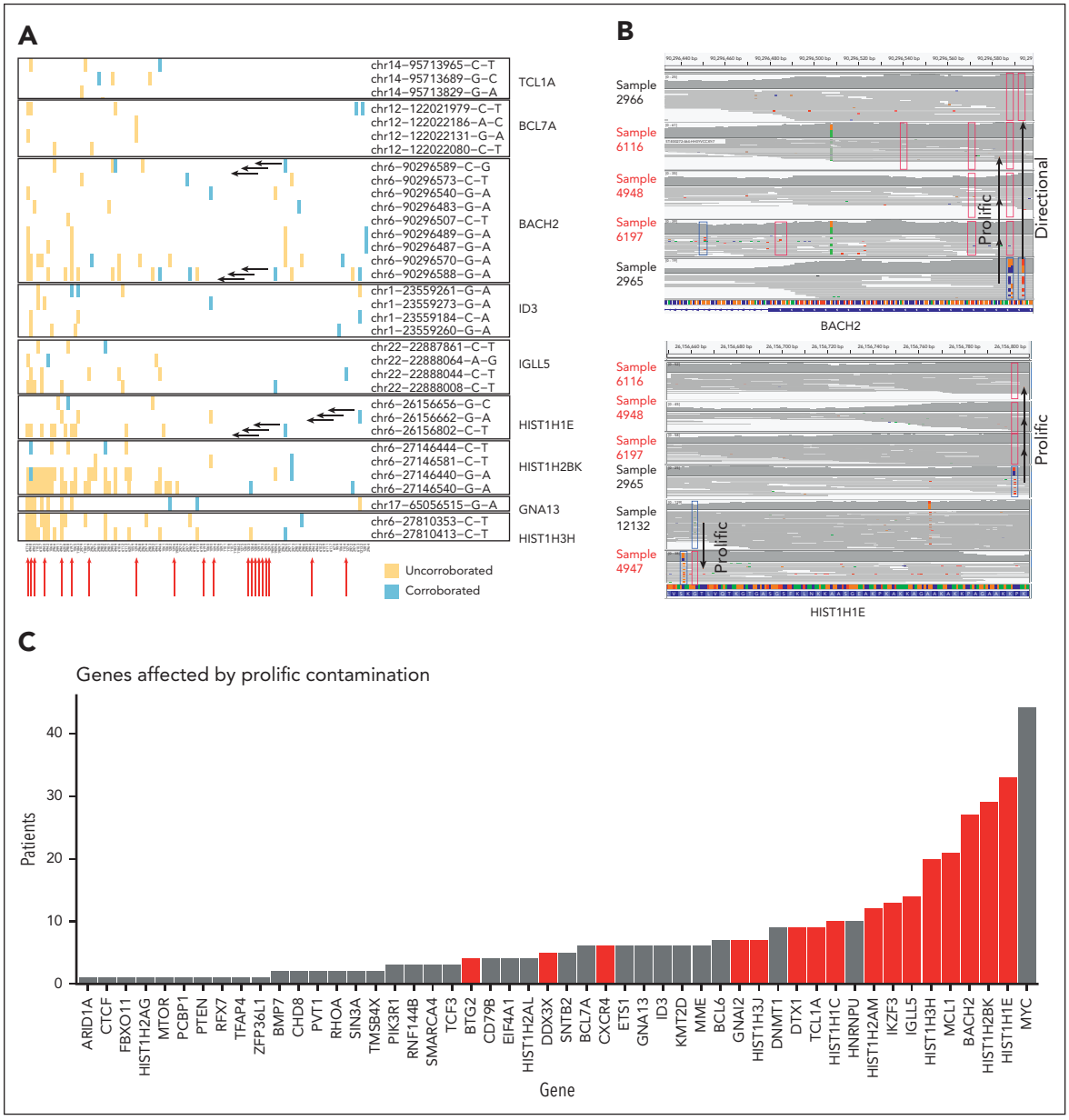

Figure 2. Patterns of uncorroborated mutations and prolific data contamination. (A) Corroboration status for variants reported among 9 representative regions affected by clustered mutations according to Panea_S3. For each region, mutations reported in multiple patients are shown on a separate row, with blue boxes indicating the patient(s) in which the variant could be corroborated and yellow boxes indicating uncorroborated mutations. The rows annotated with sets of arrows correspond to the variants labeled as "prolific" in panel B. The reported existence each of these prolific variants was uncorroborated in multiple patients and typically only corroborated in 1 to 2 patients. (B) Representative examples of prolific mutations that could not be corroborated. Integrative Genomics Viewer visualizations of the sequencing data for regions of the *BACH2* (top) and *HIST1H1E* (bottom) genes include boxes outlining the locations of corroborated (dark blue) or uncorroborated (red) variants in individual genomes. Sample 2965 is the source of both directional contamination (of sample 2966) and the other samples shown. Samples indicated with red arrows (A) or red labels (B) were indicated as mutated in Panea_S2 but were absent from supplemental Table 3. (C) Effect of prolific contamination on the reported rate of mutations in affected genes. The red bars indicate genes with significantly lower frequency in the reanalysis when compared to the mutations reported in Panea_S2. bp, base pair.

Figure 1 (continued) and Epstein-Barr virus (EBV) status. The rows are ordered based on hierarchical clustering with the dendrogram (left) showing the clustering based on mutations reported in each patient. We noted that a minority of uncorroborated variants could not be resolved by any directional contamination, and these "unresolvable" variants appeared to be more common among sporadic and HIV-associated BL cohorts. Genomes were classified based on the predominant pattern where this could be inferred. (B) The average coverage depth for the genomes. (C) Box plot showing the distribution of reads supporting the nonreference allele for corroborated variants. Genomes with their variants supported by a minimal number of reads (mean supporting reads, <4) are highlighted in yellow. (D) A forest plot showing the log-transformed odds ratio estimate from Fisher exact tests comparing the mutation frequency of corroborated variants in EBV^+ and EBV^- cases. Genes with points above y = 0 had more mutations in EBV^+ cases. *SNTB2*, the only gene reported as enriched for mutations in EBV^- cases, is no longer significant ($q > 0.1$, false discovery rate). Bold red type indicates a gene that is significantly associated with EBV status in this analysis but not in other studies that have compared mutation frequency between EBV^+ and EBV^- BL.[1,3] (E) The mutation burden of each patient based on the corroborated variants is shown as a box-whisker plot with patients stratified on the reported EBV type. Cases that benefited from directional contamination are indicated in red triangles and the rest are black points. Although a significant global difference is observed (Kruskal-Wallis test), post hoc pairwise tests show an insignificant difference between cases with type 1 and type 2 EBV (Wilcoxon rank-sum test).

Specifically, a preponderance of variants that were attributed to many samples but could generally be corroborated in only 1, which we termed "prolific" data contamination (Figure 2). These variants affected a more limited set of loci, mostly corresponding to regions deemed to contain clustered mutations described in supplemental Table 3 of Panea et al (Panea_S3). We encountered numerous discrepancies between the genomes and mutations attributed to each region according to Panea_S2 and Panea_S3, with the latter only documenting mutations in 81 of the genomes. The 20 genomes absent from Panea_S3 were among those with the lowest average coverage achieved, which could explain their exclusion from that analysis. Despite having been (presumably) excluded from that analysis, these genomes had mutations attributed to them within these regions according to Panea_S2. Figure 2 shows some of these regions and highlights the uncorroborated variants with the prolific pattern. We found additional discrepancies between these tables with many additional samples having mutations reported only in Panea_S2 but not documented in Panea_S3. If only corroborated variants are considered, 16 of the BL driver genes described in this study are mutated in significantly fewer patients than reported (binomial exact test) (Figure 2C). All these genes were affected by examples of prolific contamination.

We then sought an explanation for the variants that could not be explained by data cross-contamination from other genomes. We checked for uncorroborated variants supported by exome but not genome sequencing data, which revealed 146 mutations across 39 patients that could only be corroborated by the exome data, with 17 patients having at least 3 apparent exome-derived mutations (supplemental Figures 1D and 2). This suggests that the analysis in Panea et al relied on another source of data to identify the variants reported in Panea_S2. These results are concerning because they cannot be consolidated with the analysis as described by Panea et al.

Although the effects on each conclusion from Panea et al has not been evaluated, we demonstrated that ~30% of the reported mutations are not supported by their WGS data, which caused a significant inflation of the mutation prevalence of at least 16 genes and the rate of coding mutations in 9 genes (supplemental Figure 3). These lead to different associations between mutation frequency of genes and EBV status and negated the reported association between mutation load and EBV type. We also noted that numerous mutations in each of *TP53* and *EZH2* were identified in our analysis but not reported in the study (supplemental Figures 4 and 5, respectively), drawing further questions about the analytical approaches used. Collectively, we feel these issues draw into question the validity of the remaining conclusions.

Authorship

Contribution: R.D.M. conceived the study, generated the figures and tables, and wrote the manuscript; and all authors performed the analyses.

Conflict-of-interest disclosure: The authors declare no competing financial interests.

ORCID profile: R.D.M., 0000-0003-2932-7800.

Correspondence: Ryan D. Morin, Department of Molecular Biology and Biochemistry, Simon Fraser University, 8888 University Dr, Burnaby, BC V5A 1S6, Canada; email: rdmorin@sfu.ca.

Footnotes

Submitted 30 March 2022; accepted 16 June 2022; prepublished online on *Blood* First Edition 27 October 2022.

The data reported in this article have been deposited in the European Genome-phenome Archive database at the European Bioinformatics Institute (accession number EGAS00001003778). The data were used in a form agreed by the user institution with the data access committee for the Dave laboratory, Duke University.

Data can be accessed through a request as detailed on the European Genome-phenome Archive website at https://ega-archive.org/.

The online version of this article contains a data supplement.

REFERENCES

1. Grande BM, Gerhard DS, Jiang A, et al. Genome-wide discovery of somatic coding and noncoding mutations in pediatric endemic and sporadic Burkitt lymphoma. *Blood*. 2019;133(12):1313-1324.

2. Panea RI, Love CL, Shingleton JR, et al. The whole-genome landscape of Burkitt lymphoma subtypes. [published correction appears in *Blood*. 2022;139(8):1256]. *Blood*. 2019;134(19):1598-1607.

3. Kaymaz Y, Oduor CI, Yu H, et al. Comprehensive transcriptome and mutational profiling of endemic Burkitt lymphoma reveals EBV type–specific differences. *Mol Cancer Res*. 2017;15(5):563-576.

https://doi.org/10.1182/blood.2022016505

RESPONSE

Burkitt lymphoma genomic discovery studies, drivers, and validation

Sandeep S. Dave

Center for Genomic and Computational Biology and Department of Medicine, Duke University, Durham, NC

Rushton et al[1] refer to our work on Burkitt lymphoma (BL)[2] that identified the genetic drivers of different subgroups of BL as well as functionally validated the drivers in BL using a CRISPR screen and created the first in vivo model