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 of BL that incorporates the combined effects of *MYC* and *ID3*.

They began the analysis of our publicly available data from the aligned sequencing data (binary alignment and map files). They separated sequencing reads into exonic and nonexonic reads by using the read group identifiers generated during alignment. This assumption is erroneous. Read groups are flagged during successive alignments and are not intended to identify exonic and nonexonic genomic reads. What they call the exome can also contain reads from the genome that were included in that alignment. Thus, we cannot comment further on their analysis except to point out that splitting the data into read groups does not recapitulate our analysis or accurately quantify the sequencing reads mapping to genes. This has been addressed in a published erratum.

They further examined the overlap of potential driver variants in supplemental Table 2. To be clear, most variants in our study affect only a single patient, thus precluding overlap. They perform elegant analyses that indicate that there is a significant overlap between the variants among the endemic and HIV-associated cases. We agree that this overlap is present. We disagree on what that indicates.

In addition to errors, there are methodological and biological explanations for the observed overlap, which they fail to consider. In our analysis, any variant that was somatic in one tumor was annotated as such for all others, even if it was not flagged as somatic in those other cases, as long as the population frequency of those variants is low. This is informed by the knowledge that some driver events that are somatic in some patients can also occur as germ line events in other patients. This is true of, for instance, the well-known *MYD88* L265P variant that is both a somatically mutated driver in many cancers and a rare germ line event in other patients. Our approach was intended to flag potential driver events across all cases by casting a wider net. However, this approach is also likely to flag germ line events and polymorphisms that were lacking in our control population frequency datasets at the time. This is not an error but rather an informed decision based on our knowledge of driver events.

By necessity, our set of patients who were HIV positive and with endemic BL were each narrowly drawn from a small, separate geographic region and disproportionately from population groups that are minorities in the United States. These minority groups, particularly those of African descent, are well-known to be underrepresented in population databases. We used the population frequencies available to us at the time. It is likely that some of our identified variants repeated across patients may turn out to be germ line variants prevalent in these groups. The underrepresentation of minority patient genotypes in population databases remains a major gap in the field that we, with many institutions around the world, are working to correct.

Finally, there are also possible biological reasons for the overlap of variants between cases. For instance, Gouveia et al[3] identified clusters of familial susceptibility for BL with highly overlapping, potential driver, variants among their patients. Information regarding the relatedness of our deidentified cases is not available to us but is conceivable, especially in endemic BLs that were drawn from a small geographic region in Africa.

The authors comment that our study likely undercounts genetic variants such as hotspot events in *TP53* and *EZH2* genes. That is partly true. Our study likely undercounts many other genetic variants. This is a direct consequence of our study design. The genomic part of our study is a discovery study intended to elucidate the most common drivers of BL. We cannot claim to have found or reported all the variants or drivers present in our data, but we have identified most frequent drivers that standard methods and population databases enabled at the time. Those drivers are highly concordant with other studies, including theirs. Our discovery methods and variant calling engender tradeoffs between sensitivity and specificity. The results from the Sanger sequencing in supplemental Table 4 indicates that our results have high specificity. These findings undoubtedly include false positives, false negatives, and exploratory results, as do nearly all genomic discovery studies.

Our discovery approach is in contrast with genomics in the clinical setting, with which we have considerable experience. For patients undergoing sequencing in the clinic, we usually reject cases achieving less than 200× coverage using a platform validated against clinical gold standards. Thus most, if not all, of the samples in our study and theirs[4] would be excluded. To establish whether a specific patient has a specific driver event requires that we carefully establish limits of detection, sensitivity, specificity, and accuracy of the assay along with informatics for variant detection, Epstein-Barr virus status, copy number alterations, and translocations. Each of these parameters can be affected by a statistical variation, limit of detection, tumor purity quantification, guanine-cytosine content and mappability of the region, and the DNA-sequencing platform and would need significant validation against clinical standards for those results to be reliable at patient level. Instead, our study was designed to go deeper using biological validation.

It is not uncommon to re-examine published genomic results with new data and tools and come to somewhat different conclusions. For instance, 3 previous contemporaneous, high-impact publications identified ID3 as a common, novel driver in BL with strikingly different frequencies of mutations: 34%,[5] 58%,[6] and 68%.[7] Our follow-up studies[2,4] reveal that the frequency is closer to 40%. Similarly, many putative drivers featured prominently in "Figure 1" in those papers[5,6] (including our own) have not held up in our own follow-up studies.[2,4] We point this out, not as criticism of the past work, but to note that it represents the nature of this science. Genomics is a fast-moving field, and almost none of the methods in our publication are still in use within our group. It is inevitable that new data, new patient cohorts, and new tools will enable a continued better understanding of the disease. Still, a preponderance of our drivers is directly corroborated by the other study. Our data remain a rich resource of BL genotypes, which we have shared transparently both as raw data and supplemental tables to enable the next round of discoveries.

We cannot solve all the issues with false discovery in genomic studies, but our study was designed to ameliorate them through a process of progressive validation. Although supplemental Table 2 has more than 200 000 elements generated by a purely computational analysis, supplemental Table 4 contains the subset of variants that we specifically validated with Sanger sequencing. Figure 5 describes proteomic characterization and a novel mouse model that validates the biological function of a single driver gene even more deeply. We believe that such an approach is essential to fully understand the genetic contributions to BL and other cancers.

Our paper includes many contributions relevant to the understanding of BL, including drivers, their expression, and functional roles in BL as well as the proteomic and in vivo characterization of the role of ID3. These results continue to provide a rich starting point for a more complete clinical and functional delineation of BL.

Authorship

Contribution: S.S.D. wrote the manuscript.

Conflict-of-interest disclosure: The author declares no competing financial interests.

Correspondence: Sandeep S. Dave, Center for Genomic and Computational Biology and Department of Medicine, Duke University, Durham, NC 27705; email: sandeep.dave@duke.edu.

Footnotes

Submitted 25 October 2022; accepted 25 October 2022; prepublished online on *Blood* First Edition 27 October 2022.

Publisher's note: The Letter to *Blood* by Rushton et al and the Response by Dave highlight the need for completeness and clarity in the reporting of methodology. The erratum to Panea et al[2] in this issue expands the methods for that article via an updated supplement and notes additional data made available through the European Genome Archive entry associated with the study. Full details are included in the erratum.

REFERENCES

1. Rushton CK, Dreval K, Morin RD. Concerning data inconsistencies in Burkitt lymphoma genome study. *Blood*. 2023;142(10):933-936.

2. Panea RI, Love CL, Shingleton JR, et al. The whole-genome landscape of Burkitt lymphoma subtypes. *Blood*. 2019;134(19):1598-1607.

3. Gouveia MH, Otim I, Ogwang MD, et al. Endemic Burkitt lymphoma in second-degree relatives in Northern Uganda: in-depth genome-wide analysis suggests clues about genetic susceptibility. *Leukemia*. 2021;35(4):1209-1213.

4. 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.

5. Love C, Sun Z, Jima D, et al. The genetic landscape of mutations in Burkitt lymphoma. *Nat Genet*. 2012;44(12):1321-1325.

6. Schmitz R, Young RM, Ceribelli M, et al. Burkitt lymphoma pathogenesis and therapeutic targets from structural and functional genomics. *Nature*. 2012;490(7418):116-120.

7. Richter J, Schlesner M, Hoffmann S, et al. Recurrent mutation of the ID3 gene in Burkitt lymphoma identified by integrated genome, exome and transcriptome sequencing. *Nat Genet*. 2012;44(12):1316-1320.

https://doi.org/10.1182/blood.2022018865