Genomic Characterization of HIV-Associated Plasmablastic Lymphoma Identifies Pervasive Mutations in the JAK–STAT Pathway

Zhaoqi Liu1,2, Ioan Filip1,2, Karen Gomez1,2, Dewaldt Engelbrecht3, Shabnum Meer4, Pooja N. Lalloo3, Pareen Patel3, Yvonne Perner5, Junfei Zhao1,2, Jiguang Wang6, Laura Pasqualucci7,9, Raul Rabadán1,2, and Pascale Willem3
Plasmablastic lymphoma (PBL) is an aggressive B-cell non-Hodgkin lymphoma associated with immunodeficiency in the context of human immunodeficiency virus (HIV) infection or iatrogenic immunosuppression. While a rare disease in general, the incidence is dramatically increased in regions of the world with high HIV prevalence. The molecular pathogenesis of this disease is poorly characterized. Here, we defined the genomic features of PBL in a cohort of 110 patients from South Africa (15 by whole-exome sequencing and 95 by deep targeted sequencing). We identified recurrent mutations in genes of the JAK–STAT signaling pathway, including STAT3 (42%), JAK1 (14%), and SOCS1 (10%), leading to its constitutive activation. Moreover, 24% of cases harbored gain-of-function mutations in RAS family members (NRAS and KRAS). Comparative analysis with other B-cell malignancies uncovered PBL-specific somatic mutations and transcriptional programs. We also found recurrent copy number gains encompassing the CD44 gene (37%), which encodes for a cell surface receptor involved in lymphocyte activation and homing, and was found expressed at high levels in all tested cases, independent of genetic alterations. These findings have implications for the understanding of the pathogenesis of this disease and the development of personalized medicine approaches.

SIGNIFICANCE: Plasmablastic lymphoma is a poorly studied and extremely aggressive tumor. Here we define the genomic landscape of this lymphoma in HIV-positive individuals from South Africa and identify pervasive mutations in JAK–STAT3 and RAS–MAPK signaling pathways. These data offer a genomic framework for the design of improved treatment strategies targeting these circuits.

See related commentary by Kükpers, p. 23.
profiles of 15 PBL cases to DLBCL and extranodal plasma-cytoma (11), and showed that B-cell receptor signaling genes including CD79A/B, BLK, LYN, and SWAP70 among others, were significantly downregulated in PBL compared with DLBCL; in contrast, targets of MYB and the oncogene MYC, as well as genes reflecting the known plasmacytic immunophenotype of PBL, were overexpressed. More recently, MYC and PRDM1 were investigated for mutations, structural rearrangements, and copy number gains in 36 PBL cases (12). PRDM1 mutations were found in 8 of 16 cases, frequently in association with MYC overexpression, suggesting a coordinate role in the pathogenesis of the disease. While these studies have shed light on this rare disorder, a systematic characterization of protein-changing alterations in PBL has not been performed.

The elucidation of genes and pathways that drive the initiation and maintenance of PBL is essential to better understand the biology of this cancer and, critically, to implement improved biomarkers and more effective treatment options. Here, we combined whole-exome and transcriptome sequencing followed by targeted resequencing of 110 HIV-associated PBL cases to elucidate the mutational, transcriptional, and copy number landscape of this disease. We show that mutations in various components of the JAK–STAT and MAPK–ERK pathway pervade this lymphoma, revealing this signaling cascade as a central oncogenic driver of the disease and a candidate for targeted therapy.

RESULTS

The Landscape of Somatic Mutations in PBL

To determine the mutational landscape of PBL, we performed whole-exome sequencing (WES) in a discovery panel of paired tumor and normal DNA collected from 15 HIV-positive patients, followed by deep targeted sequencing of the top 34 candidate genes in 95 additional cases (Supplementary Table S1 and S2; Methods). Somatic mutations were identified from WES data using SAVI2 (13), an empirical Bayesian method. Overall, 2,149 nonsynonymous somatic variants were found in the 15 discovery cases, with a median of 45 per case and a total of 1,528 affected genes, of which >95% of the mutations resulted in the amino acid changes Y640F (n = 11), D661V (n = 9), S614R (n = 5), and E616G/K (n = 4; Fig. 1B), which have been categorized as gain-of-function or likely gain-of-function mutations based on experimental validation (https://oncokb.org/gene/STAT3), and overlap with the STAT3 mutation pattern described in other aggressive B-cell lymphomas and T-cell and natural killer (NK) cell malignancies (15, 16). Three additional cases showed a >75% variant allele frequency in the absence of copy number changes, consistent with a copy neutral loss of heterozygosity. Sanger-based resequencing of the involved STAT3 exons, performed in 23 cases, validated all computationally identified mutations on these samples (Fig. 1F; Supplementary Table S6).

In addition to STAT3, 15/110 PBL cases harbored heterozygous missense mutations of JAK1, encoding for a tyrosine kinase that phosphorylates STAT proteins. These events did not involve the canonical hotspot seen in many other cancers (loss-of-function K860Nfs), but occurred at a highly conserved amino acid position in the JH1 kinase domain, G1097D/V (Fig. 1C). Mutations at the JAK1 G1097 codon were previously reported in intestinal T-cell lymphomas (17) and anaplastic large cell lymphoma (15), and the G1097D substitution has been documented as a gain-of-function event that triggers aberrant phosphorylation of STAT3, leading to constitutive activation of the JAK–STAT signaling pathway (15). Of note, missense mutations in STAT3 and JAK1 were found to be mutually exclusive (P = 0.04, Fisher exact test), suggesting a convergent role in activating this cascade. In contrast to STAT3 and JAK1, mutations in SOCS1 (n = 11/110 cases) were more widely distributed, consistent with its previously established role as a target of AID-mediated aberrant somatic hypermutation (Fig. 1D). Although the consequences of most SOCS1 amino acid changes will require functional validation, the presence of a start-loss mutation in one case and a nonsense mutation in a second sample is expected to result in loss-of-function and inactivation of this negative JAK/STAT regulator, consistent with a tumor suppressor role.

The second most commonly mutated program was the MAPK–ERK pathway, acting in 28% of cases by mutations in the RAS gene family members NRAS (14%) and KRAS (9%), as well as in BRAF (5.5%) and MAP2K1 (3%). Nearly all (92.5%) RAS mutations were found at known functional hotspots that have been shown to affect the intrinsic RAS GTPase activity, namely G12, G13, and Q61 (ref. 18; Fig. 1E). BRAF, an integral component of the MAPK
Figure 1. The landscape of putative driver gene mutations in PBL. A, Sample information, MYC translocation, and somatic mutation information are shown for 110 cases of PBL samples. The heatmap represents individual mutations in each sample, color-coded by type of mutation. B–E, Individual gene mutation maps for frequently mutated genes, showing mutation subtype, position, and evidence of mutational hotspots, based on COSMIC database information. Y-axis counts at the bottom of the maps reflect the number of identified mutations in the COSMIC database. F, Sanger validation of single nucleotide variants (SNV) in STAT3 mutants.

signaling cascade, was found mutated in 6 cases, including 3 harboring mutations at or around the well-characterized V600 residue (namely: V600E, K601N, and T599TT) and 3 showing mutations at other common hotspots within the protein kinase domain (G464E, V471F, and M689V). The Valine at position 600 normally stabilizes the interaction between the BRAF glycine-rich loop and the activation segments, and its glutamine substitution confers an over 500-fold increase in activity, leading to constitutive activation of the MEK/ERK signaling cascade in the absence of extracellular stimuli. In line with their predicted gain of function, the mutant alleles in both RAS family members and BRAF were observed in heterozygosis (Supplementary Tables S3 and S5) and were actively expressed (median
Somatic Copy Number Changes

To identify recurrent copy number alterations (CNA) associated with PBL, we applied the SNP-FASST2 algorithm to WES data from the 15 discovery cases (Supplementary Fig. S4A), followed by validation in an independent cohort of 31 additional PBL samples that had been processed with Affymetrix OncoScan microarrays (ref. 21; Supplementary Table S7). Comparison of CNA calls obtained in parallel from the WES and OncoScan approach in a subset of 9 samples confirmed the data were highly consistent with each other (Supplementary Fig. S4B), supporting the robustness of the analysis.

Frequent copy number gains involved large chromosomal regions (>10 Mb) including chromosome 1q (20/46 cases, 43%) and the whole or most of chromosome 7 (13/46 cases, 28%). When applying the GISTIC 2.0 algorithm to the combined cohort of 46 cases, we uncovered regions of highly recurrent amplification including 6p22.2 (the most significant), 6p22.1 and 1q21.3, all of which encompass histone gene clusters (Fig. 2A). In particular, the chromosome 6p gain covered 36 genes that encode canonical histones and has been previously reported as a common alteration in a variety of cancers, where histone gains have been linked to genetic instability (22). The significant region on chromosome 1q21.3 also included the Ile6r gene and the antiapoptotic MCL1 gene, which showed increased gene expression (Fig. 2C) analogous to what has been described in 26% of activated B-cell like (ABC) DLBCLs (23). Although further studies will be needed to determine the functional impact of these alterations in PBL, chromosome 1q gains have been associated with unfavorable prognosis in multiple myeloma, suggesting a role in the pathobiology of the disease.

The second most significant focal CNA was a chromosome 11p13 regional gain targeting genes CD44 and PDHX, present in 17 of 46 cases (37%; Fig. 2A–C). CD44 is a nonkinase transmembrane glycoprotein that is induced in B cells upon antigen-mediated activation and is critically involved in multiple lymphocyte functions, including migration, homing, and the transmission of signals that regulate apoptosis. This CD44 protein is also thought to increase cancer cells’ adaptive plasticity in response to the microenvironment, thus giving them a survival and growth advantage (24). Analysis of 20 samples with available RNA-seq data revealed that CD44 was consistently and highly expressed in cases harboring copy number gains (n = 2), whereas PDHX levels were undetectable, indicating that CD44 is the specific target of the 11p13 amplicon. However, all samples displayed high CD44 mRNA expression, independent of the presence of genetic aberrations (Fig. 2C); consistently, IHC analysis with a specific CD44 antibody showed very strong membranous staining in all cases tested (Fig. 2D) and confirmed this finding in an independent panel of 38 cases. Of note, although CD44 expression can be detected in normal plasma cells at both RNA and protein levels, the signal was markedly lower than in plasmablastic lymphoma cells, suggesting that elevated expression of CD44 does not simply reflect the cellular ontology of these tumors, and that alternative regulatory mechanisms may lead to CD44 upregulation in cases lacking CNAs (Supplementary Fig. S5A–S5C; Supplementary Table S8).

Plasmablastic Lymphoma Displays a Distinct Genetic and Transcriptional Program

To explore the genomic features of PBL in relation to other lymphoid neoplasms arising from the mature B-cell
Figure 2. Recurrent copy number changes in PBL. A, GISTIC 2.0 results showing recurrent copy number changes in PBL samples. The green line indicates \( q \)-value = 1.0 \( \times \) 10\(^{-6}\). B, A zoomed-in view of 11p13 on 17 cases of PBL, which shows consistent focal copy number gains of CD44. The figure was generated by the IGV browser using CNV segment files from SNP-FASST2 algorithm. C, Scatter plot representations of genes located in regions with recurrent copy number gains in PBL (\( q \)-value < 1.0 \( \times \) 10\(^{-6}\), GISTIC 2.0). The horizontal axis indicates the \(-\log_{10}(q\text{-value})\) from the GISTIC report, and the vertical axis is the median gene expression level (normalized RPKM value) from PBL RNA-seq data (\( n = 20 \)). D, IHC showing strong CD44 protein expression on tumor cells membrane in a representative sample (A10), which has a focal CD44 copy number gain.
lineage, we performed unsupervised clustering analysis based on mutation frequency of the top mutated genes from three cancer types obtained from public repositories (Fig. 3A). The analysis included chronic lymphocytic leukemia (CLL; ref. 25), diffuse large B-cell lymphoma (DLBCL) transcriptionally defined as ABC and germinal center B-cell-like (GC-B) subtypes (26), and multiple myeloma (27). As expected on the basis of their presumed derivation from B cells committed to plasma cell differentiation, the mutational landscape of PBL was overall closer to multiple myeloma than to other mature B-cell malignancies, with mutations in RAS family members being detected in as many as 20% of cases in both diseases, while rare in DLBCL (Fig. 3A). Conversely, the mutational landscape of PBL was highly distinct from that of both GC-B and ABC-DLBCL (Fig. 3A). In particular, mutations affecting the methyltransferase KMT2D and acetyltransferase CREBBP, two among the most commonly mutated genes in DLBCL (28), were absent in plasmablastic lymphoma. ABC-DLBCL–specific mutations such as CD79A/B and MYD88 were also lacking in PBL. Of note, STAT3 was the top mutated gene in our cohort at significantly different frequencies compared with other B-cell malignancies, making it a hallmark of PBL (Fig. 3A). The differential genetic landscape of PBL is consistent with its status as a distinct entity among mature B-cell neoplasms.

To define the transcriptional profile of HIV-positive PBL and to identify unique signatures that may distinguish it from other lymphoma types, we performed RNA-seq analysis of 20 PBL samples (including 12 of the 15 discovery cases), and compared their transcriptome to that of normal B-cell subsets and other B-cell lymphoma types previously characterized in our laboratories and/or obtained from public repositories, including germinal center centroblasts (CB), naïve B cells (NB), and memory B cells (MB; ref. 29), as well as CLL (30), DLBCL (26), and multiple myeloma cell lines (31). As expected, hierarchical clustering of the top 1,000 most aberrantly expressed genes revealed that PBL and multiple myeloma were closer to each other compared with other B-cell malignancies and to normal B cells, reflecting their presumed cell of origin (Fig. 3B). Consistently, plasmablastic lymphoma and multiple myeloma lacked expression of common B-cell markers (CD19, CD20, CD40, and PAX5) and transcription factors involved in the germinal center reaction (BCL6, BCL7A, BCL11A, and SPIB), whereas expression of the master regulator of plasma cell differentiation PRDM1 and other plasma cell marker genes (CD138, XBP1, and IRF4) were increased (Supplementary Fig. S6). MYC expression was also higher in PBL and multiple myeloma. Other notable differences included the upregulation of IL6R, a known STAT3 target, and the downregulation of SWAP70, previously suggested as a potential biomarker of PBL (Supplementary Fig. S6; ref. 11). Finally, recent studies have suggested that EBV-positive PBLs evade immune recognition by expressing the programmed cell death protein 1 (PD-1) and its PD-L1 ligand (32). We confirmed high expression of PD-L1 in our PBL cohort, although PD-1 did not follow this pattern (Supplementary Fig. S6).

We then performed functional enrichment analysis (gprofiler) to identify biological pathways that are preferentially enriched in PBL as compared with other B-cell malignancies. Whereas DLBCL and CLL were enriched for B-cell related biological programs (e.g., B-cell activation and proliferation, lymphocyte activation and differentiation, and adaptive immune response), multiple myeloma was enriched for mitotic cell-cycle processes, with highly expressed genes including MCM10, BIRC5, CENPE, BUB1, and AURKA (Fig. 3B). The most significantly enriched pathways in PBL were related to chromatin/nucleosome assembly (Fig. 3B); particularly, histone genes encoding basic nuclear proteins were highly expressed in PBL, possibly related to the recurrent copy number gains/amplification encompassing this gene cluster on chromosomes 6p22.2 and 1q21.3 (Supplementary Fig. S7A–S7D). However, the biological significance of this observation in relation to tumorigenesis remains unknown.

**Activation of the JAK–STAT Pathway in Plasmablastic Lymphoma**

Given the elevated frequency of mutations targeting the JAK–STAT pathway, we used computational and IHC approaches to assess the extent of activation of this signaling cascade in PBL (Methods). To this end, we first interrogated the genome-wide transcriptional profile of PBL for enrichment in the JAK–STAT signaling pathway using previously identified signatures available in the MSigDB database. This analysis revealed a significant positive enrichment for genes implicated in this pathway, consistent with constitutive activation (Fig. 4A). This signature was expressed at higher levels in PBL when compared with multiple myeloma and CLL (Fig. 4B), with the most significant difference being observed between PBL and multiple myeloma (t test, P = 4.7 × 10−11), which is congruent with the fact that the latter had few mutations in this pathway (Fig. 3A). We then performed IHC staining with anti-STAT3 and anti-phospho-STAT3 antibodies to assess the STAT3 subcellular localization and phosphorylation, used as a read-out for signaling activation. We observed strong positive nuclear signal in 61% of tested cases (19/31), which also stained positive for the phospho-STAT3 antibody. Of note, evidence of constitutively active STAT3 was detected even in the absence of genetic alterations affecting this pathway, suggesting alternative (epigenetic) mechanisms of activation and indicating a prominent role for this cascade in the pathogenesis of the disease (Fig. 4C and D; Supplementary Fig. S8; Supplementary Table S9).
**Figure 4.** Activation of JAK–STAT pathway in PBL. **A,** Preranked gene set enrichment analysis indicating the significant positive enrichment of KEGG JAK–STAT signaling pathway in PBL. The preranked gene list was generated on the basis of the median expression level on PBL samples. **B,** Enrichment comparisons of JAK–STAT pathway between PBL and other B-cell malignancies by performing single-sample GSEA (ssGSEA). Pairwise P values were derived from t test. **C** and **D,** IHC plots showing pSTAT3 protein expression in >75% of tumor cells, confirming STAT3 activation in STAT3 mutated cases. Results using anti-pSTAT3 antibody are photographed at ×100 and ×400 magnification.

**Virus Detection in Plasmablastic Lymphoma**

To analyze the viral and bacterial make-up of PBL tumors, we used the Pandora pipeline, which extracts and aligns nonhost genetic material from tumor RNA-seq data (Methods). Potentially pathogenic species were then identified by applying the BLAST algorithm against the NCBI database of viruses and bacteria reference genomes. In our PBL cohort, 12 of 20 samples contained HIV-1 transcripts, with only three samples exceeding 100 mapped reads (Supplementary Fig. S9). In addition, EBV transcripts were detected in 18 of 20 samples, HCMV (human cytomegalovirus, or human betaherpesvirus 5) in 3 of 20 samples, and KSHV (Kaposi sarcoma-associated herpes virus, human herpes virus 8) in one sample (Supplementary Fig. S9). These three herpes viruses have a broad tropism, naturally infect B cells, and are known to be associated with many tumor types.

EBV reactivation is considered a major driver of PBL, predominantly with a latency I infection program although low levels of LMP1 gene expression can be detected (33).
We thus performed a detailed RNA profiling of the EBV genome. Recapitulating previous reports, all PBL cases tested by in situ hybridization were positive for EBV-encoded RNA (EBER), while only 4 of 13 samples showed expression of the LMP1 protein on IHC (Supplementary Table S1). When using the levels of BLF1 and EBNA2, two genes that are invariably not expressed in PBL, as threshold for positive calls, we noted that nearly the entire viral genome was transcribed at background levels in the majority of samples (Fig. 5). This is analogous to what has been observed in nonreplicating infected B cells, where a background of transcripts can be detected in the absence of protein expression, due to regulation at the ribosomal level (34). However, several genes in the BamHI-A region of the virus were abundantly transcribed in at least 50% of cases across the cohort, including those encoding for components of the viral replication machinery (namely, the DNA polymerase catalytic subunit BALF5, the single-stranded DNA-binding protein BALF2, and the lytic origin of DNA replication oriLyt), the BART encoded protein RPMS1, the viral envelope glycoprotein BALF4, and the G-protein–coupled receptor BLF1, which is expressed predominantly during the immediate early phases of infection in vitro (35, 36). LF1, LF2, and LF3 were also highly expressed in PBL, but their functional roles are less understood. Expression of the LMP-1 mRNA was detected in 17 of 20 samples, but at much lower levels (Fig. 5), consistent with the known EBV latency programs of PBL (33).

**DISCUSSION**

Our study provides a comprehensive snapshot of the genetic landscape of HIV-associated PBL and reveals highly recurrent somatic mutations affecting the JAK–STAT3 and RAS–MAPK signaling pathway as a genetic hallmark of this disease, with STAT3 representing the most prominent target. These findings underscore a central role for this transcription factor in the pathogenesis of PBL and have implications for the diagnosis and treatment of these diseases. The STAT3 protein is an important player in multiple immune cells where it modulates a variety of physiological processes. Within the B-cell lineage, a selective role has been recognized for this factor in the differentiation of B cells into plasma cells upon antigen stimulation, as documented by both in vitro and in vivo studies (14). Briefly, in response to CD40L- and IL21-mediated signaling by T follicular helper cells, STAT3 is phosphorylated by activated JAK kinases, translocates into the nucleus as homo- or hetero-dimers, and activates the transcription of multiple targets, including the plasma cell master regulator BLIMP1. In turn, BLIMP1 downregulates BCL6 expression, an absolute prerequisite to GC exit and plasma cell differentiation (14). The STAT3 amino acid changes identified in our study include experimentally documented gain-of-function events that are predicted to have oncogenic effects by enhancing its phosphorylation and transactivation potential (37). In addition, other well-documented genetic mechanisms were found that can activate the JAK–STAT signaling pathway, including mutually exclusive upstream mutations of JAK1 or JAK2 (18/110 cases) and the loss of the STAT3 negative regulator SOCS1 (mutated in 11/110 cases). Notably, a targeted sequencing study published while this manuscript was under revision identified recurrent STAT3 mutations in 5 of 42 cases, all of which were EBV positive (38). Thus, STAT3 dysregulation may contribute to the pathogenesis of PBL by rendering these cells signaling-independent while providing proliferation and survival signals.

Constitutive activation of the JAK–STAT signaling pathway has been reported in a number of solid and hematologic malignancies and plays a central role in two lymphomas that are immunophenotypically closely related to plasmablastic lymphoma: primary effusion lymphoma (PEL) and
ALK-positive large B-cell lymphoma (LBCL). PEL is a rare and aggressive AIDS-defining disease, which is associated with infection by HHV8 and is clinically distinguishable from PBL by the presence of lymphomatous effusion in body cavities. In these cells, constitutive STAT3 activity is achieved via expression of the HHV8 viral protein IL6, which contributes to the disease in an autocrine fashion by promoting proliferation and survival (39). In ALK-positive LBCL, STAT3 activation is sustained by the ALK kinase mediated by chromosomal translocations with the CLTC gene or the NPM1 gene (40, 41). However, direct genetic alterations of STAT3 are rare in mature B-cell lymphomas. In particular, while multiple genomic hits leading to potentiation of the JAK-STAT oncogenic pathway have been detected in 87% of Hodgkin lymphomas as well as in primary mediastinal B-cell lymphomas, the most commonly affected STAT member in these tumors is STAT6 (42–45). The high incidence of STAT3 mutations in HIV-associated PBL points to STAT or JAK inhibitors as promising treatment options in this lymphoma type. While anti-STAT3 therapeutic attempts are still in development, JAK inhibitor therapy, currently used in the clinical setting, was shown to be an effective antagonist to STAT3 activation, inducing apoptosis in both anaplastic large T-cell lymphomas and ovarian cancer (46).

The second important finding of this study is the identification of frequent hotspot mutations in RAS–MAPK family members. Functional RAS activation is a common molecular feature of multiple myeloma, particularly in the relapsed/refractory setting, while it is rarely observed in de novo DLBCL, suggesting a specific role in the pathogenesis of plasma cell dyscrasias. Interestingly, and different from multiple myeloma, NRAS, and KRAS mutations in PBL were never concurrently observed in the same case and were often well represented in the dominant tumor clone, consistent with early events. These data have direct implications for the clinical exploration of treatments inhibiting this pathway.

Our study also showed overexpression of the transmembrane glycoprotein CD44, frequently associated with copy number gains/amplifications at this locus. CD44 is an adhesion molecule that mediates cellular interaction with the microenvironment and participates in the trafficking of neoplastic cells in multiple myeloma, CLL, and ALL (47); moreover, CD44 was shown to increase cell resistance to apoptosis and to enhance cancer cell invasiveness. Whereas the role of CD44 in PBL requires functional dissection, its high expression is likely to compound the aggressiveness of the disease, as previously described in DLBCL of the ABC subtype (48). In light of this, successful anti-CD44–targeted therapy in a mice xenograft model of human multiple myeloma may in future represent an attractive therapeutic option for PBL (49).

The finding of increased histones mRNA abundance in PBL, together with recurrent copy number gains encompassing this gene cluster on chromosome 6, is of interest because it emerged as a distinctive feature of this disease compared with other lymphoid malignancies. Histones represent a basic component of the chromatin structure and their involvement may suggest a selective role for nucleosomal plasticity in the pathogenesis of PBL, which warrants further investigations.

Due to the small number of EBV-negative cases in our cohort (n = 2) and the overall rarity of this subset, we could not assess whether EBV status is associated with differing genetic features, as recently reported for Burkitt lymphoma (50) and suggested by the evidence of distinct transcriptomic profiles between EBV-positive and -negative PBL (32). Moreover, it remains to be determined whether regional differences exist with PBL occurring outside the context of HIV immunodeficiency, or within HIV-associated PBL, given the high homogeneity of our cohort from the Gauteng region in South Africa. Thus, additional work will be required to specifically address these questions.

We found that most of the EBV genome was transcribed at very low levels in the majority of PBL cases, while prominent expression was detected for several genes in the BamHI-A region of the virus, including some early lytic genes. However, transcription of the key early lytic gene BZLF1 was noticeably absent, suggesting an incomplete lytic program. Supporting this notion, a recent study showed that increased STAT3 expression, as observed in the majority of PBL cases, decreases the susceptibility of latently infected cells to EBV lytic activation signals via an RNA-binding protein PCBP2 (51). Thus, further protein expression studies, in particular for BRLF1 and for the BLLFI-encoded viral envelope glycoprotein gp350, are required to assess whether the observed transcriptome program translates in lytic infection. Indeed, recent translation ribosome profiling studies have clearly demonstrated a marked heterogeneity of lytic genes translation and complex levels of intracellular translation repression mechanisms at work in infected B cells (34).

In conclusion, the results of our study characterize HIV-associated PBL as a distinct subset of aggressive B-cell lymphoma and significantly contribute to our knowledge about the molecular pathogenesis of this disease through the identification of recurrently mutated genes, uncovering a major role for the dysregulation of JAK–STAT3 and RAS–MAPK signaling pathways. These results reveal new points of potential therapeutic intervention in these patients.

**METHODS**

For complete experimental details and computational analyses, see also Supplementary Methods.

**Patient Cohorts**

Prospective samples (n = 15 with paired tumor-normal tissue; discovery cohort) were collected from patients with suspected PBL, upon informed written consent in line with the Declaration of Helsinki, and diagnosis was further confirmed by two independent pathologists. Matched normal DNA for these 15 patients was extracted from peripheral blood samples that were documented to be tumor-free. For targeted sequencing (n = 95 samples; extension cohort), formalin-fixed paraffin-embedded (FFPE) material was retrieved from the archives of the Department of Oral Pathology and Anatomical Pathology, National Health Laboratory Service and the University of the Witwatersrand. This panel included 36 monoral PBL samples that were part of a published cohort (52). The study protocol was approved by the local Human Research Ethics Committee (IRB Reference M150390).
summary of demographics and phenotypic markers of the discovery cohort are displayed in Supplementary Table S1, whereas demographic information for the extension cohort is included in Supplementary Table S3. For downstream nucleic acid extraction, the tumor area (>70% tumor cells) was rinsed for microdissection in all samples. A third cohort of 31 well-characterized PBL samples obtained as archived FFPE material (local Human Research Ethics Committee IRB reference M101171 and 96/2011; Supplementary Table S7) was used to validate and refine copy number aberration results observed in the WES data by using Microarray OncoScan (21).

WES and RNA-Seq

Both exome and RNA-seq library preparations and sequencing were outsourced to Centrillion Genomics Services and BGI (Americas Corporation). Whole-exome libraries were prepared using the Agilent SureSelect Human All Exon V6 Kit (Agilent Technologies) and sequenced on HiSeq 2500 using TruSeq SBS v2 Reagent Kit (Illumina) at 2 × 100 bp paired-end reads with on-target coverage of 100X per sample. RNA-seq libraries were prepared using the Illumina TruSeq Stranded Total RNA Sample Prep Kit (Illumina). Flow cells with multiplexed samples were run on the HiSeq 2500 using an Illumina TruSeq SBS v2 Reagent Kit at 2 × 100 bp paired-end reads and a coverage of 50M reads per sample.

Mutation Calling

Fastq files were aligned to the human genome assembly (hg19) using the Burrows-Wheeler Aligner (version 0.6.2). Before further analysis, the initially aligned BAM files were subjected to preprocessing that sorted, indexed, and marked duplicated reads using SAMtools (version 1.7) and Picard (version 1). To identify somatic mutations from WES data for tumor samples with matched blood control, we applied the variant-calling software SAVI2, based on an empirical Bayesian WES data for tumor samples with matched blood control, we applied the variant-calling software SAVI2, based on an empirical Bayesian

Design of Targeted Sequencing Panel

The full coding exons of 34 genes found recurrently mutated in the discovery cohort and/or previously implicated in lymphoma were analyzed in an extension panel of 95 cases (Supplementary Table S3) by targeted capture and next generation sequencing. Genes were selected according to the following criteria: (i) allele frequency > 15%; (ii) mutated in at least 2 of 15 cases; (iii) expressed in normal or transformed B cells; and (iv) functionally annotated. In addition, we included 16 genes that were only mutated in one sample but have known roles in the pathogenesis of lymphoma and 4 genes that were not found mutated in the discovery cohort but have been previously implicated in PBL. Genes lacking clear functional annotation and/or known to represent common nonspecific mutational targets in sequencing studies (e.g., TTN, PCL0) were excluded. The complete gene list is reported in Supplementary Table S4.

Targeted Next-Generation Sequencing

The entire coding region of the 34 selected genes was isolated using the IDT xGen Predesigned Gene Capture Custom Target Enrichment Technology (Integrated DNA Technologies) and subjected to library preparation and next-generation sequencing on the Illumina HiSeq platform with 2 × 150 bp configuration. Targeted sequencing was performed at GENEWIZ. Read alignments and conventional preprocessing were conducted as described for the WES analysis. For samples lacking matched normal control, variants were filtered out if found in dbSNP database (dbSNP138) as well as any normal sample of the WES cohort.

Copy Number Analysis

For copy number analysis from WES data (n = 15), the Biodiscovery Multiscale BAM Reference Builder (53) was used to construct a multiscale reference (MSR) file from 14 paired normal samples alignments (BAM). The MSR file was used as reference for copy-number variation (CNV) calling of all tumor alignments with the SNP-FASST2 algorithm (54), using Nexus Copy Number, v10.0 (BioDiscovery, Inc.; ref. 54). Gains and losses were defined as at least +0.3 and -0.3 log, ratio changes, respectively, in the tumor alignment.

To validate CNV calling from WES data, 31 additional samples were processed on Oncoscan FFPE Express Arrays (Affymetrix, Thermo Fisher Scientific) (Supplementary Table S7) according to the manufacturer’s instructions, followed by scanning on a GeneChip Scanner 3000 7G, with the Affymetrix GeneChip Command Console (AGCC). Paired A+T and G+C CEL files were combined and analyzed with Chromosome Analysis Suite v3.3.0.139 (ChAS, Applied Biosystems, Thermo Fisher Scientific), using the OncoScan CNV workflow for FFPE, without manual recentering. We confirmed that copy number calls from Oncoscan and WES data were consistent with each other by performing Oncoscan on 7 cases that had been sequenced by WES and comparing the calls obtained from the two methods. Probe genomic coordinates were aligned to hg19 and the resulting OSCHP files were analyzed by Nexus Copy Number v10.0 using the SNP-FASST2 algorithm (54) with default parameters.

To detect recurrent copy number aberrations, we applied the GISTIC 2.0 algorithm using GenePattern (https://www.genepattern.org/) to the copy number segmentations of the combined cohort of 46 patients (15 cases analyzed by WES, 31 cases analyzed by OncoScan). Recurrent regions of copy number aberrations with q-value < 1.0 × 10−4 were considered significant. Furthermore, we assessed the expression level of each gene within the significant GISTIC peaks using the median value of quantile normalized RPKM from RNA-seq data (n = 20).

Pandora: A High-Performing Pipeline for Quantifying the Bacterial and Viral Microenvironment of Bulk RNA-Seq Samples

Pandora is an open-source pipeline (https://github.com/RabadanLab/Pandora) that takes as input the total RNA sequence reads from a single sample and outputs the spectrum of detected microbial transcripts, focusing on bacteria and viruses. The Pandora workflow is divided into the following four main modules: (i) mapping to the human host genome using STAR and Bowtie2 to filter out the host reads from downstream analysis; (ii) de novo assembly of host-subtracted short reads using Trinity (35) to create contiguous full-length transcripts (contigs), which help increase the accuracy of alignment to the correct species of origin in the next step; (iii) identification of the most likely species of origin for each assembled contig with BLAST; and (iv) filtering and parsing of the BLAST results into a final report on the detected microbial abundances. We also performed gene expression profiling of all the reads that mapped to the EBV genome (56) using FeatureCounts (57) to fully characterize lytic and latent EBV programs.

Data Availability

The data that support the findings of this study are available upon request. The sequencing data have been deposited in NCBI Sequence Read Archive (SRA) under accession number PRJNA598849.
Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Authors’ Contributions

Z. Liu: Conceptualization, investigation, visualization, methodology, writing-original draft, writing-review, and editing. J. Filip: Investigation, methodology, and writing-original draft. K. Gomez: Investigation, methodology, writing-review, and editing. D. Engelbrecht: Resources, validation, investigation, and visualization. S. Meer: Resources and investigation. P. Laloo: Resources, validation, and investigation. P. Patel: Resources, validation and investigation. Y. Perner: Resources and investigation. J. Zhao: Investigation, visualization, and methodology. J. Wang: Investigation, writing-review, and editing. L. Pasqualucci: Conceptualization, resources, supervision, funding acquisition, investigation, methodology, writing-original draft, project administration, writing-review, and editing. R. Rabadan: Conceptualization, supervision, funding acquisition, investigation, methodology, project administration, writing-review, and editing. P. Willem: Conceptualization, resources, supervision, funding acquisition, validation, investigation, visualization, writing-original draft, project administration, writing-review, and editing.

Acknowledgments

This work has been funded by NIH grants R21 CA192854 (to P. Willem, L. Pasqualucci, and R. Rabadan), R01GM117591 and U54 CA193313 (to R. Rabadan), and was initiated under the Columbia-South Africa Training Program for Research on HIV-associated Malignancies D43 CA153715, with the support of Judith Jacobson. We thank Stephen P. Goff and Henri-Jacques Delecluse for their helpful suggestions. We also thank Sonja Boy for providing additional samples for the independent copy number validation cohort, Nicole Crawford for assistance in samples collection and data cura-
tion, and Jacky Brown for help with Sanger sequencing. Whole exome capture and sequencing, and RNA sequencing were completed at CenterBio Bioscience, Inc and BGI Tech Solutions (Hong Kong). Targeted DNA sequencing was performed at Geneviz Inc. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked advertisement in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received April 2, 2020; revised May 11, 2020; accepted May 11, 2020; published first June 10, 2020.

REFERENCES

14. Hillmer EJ, Zhang H, Li HS, Watowich SS. STAT3 signaling in immu-
18. Prior IA, Lewis PD, Mattos C. A comprehensive survey of Ras muta-
20. Taddesse-Heath L, Meloni-Elong A, Scherre J, Kelly JC. Jaffe ES. Plas-
mblastic lymphoma with MYC translocation: evidence for a com-
24. Taddesse-Heath L, Meloni-Elong A, Scherre J, Kelly JC. Jaffe ES. Plas-
mblastic lymphoma with MYC translocation: evidence for a com-
Genomic Characterization of Plasmablastic Lymphoma


Genomic Characterization of HIV-Associated Plasmablastic Lymphoma Identifies Pervasive Mutations in the JAK–STAT Pathway

Zhaoqi Liu, Ioan Filip, Karen Gomez, et al.


**Updated version**
Access the most recent version of this article at:
doi: 10.1158/2643-3230.BCD-20-0051

**Supplementary Material**
Access the most recent supplemental material at:
http://bloodcancerdiscov.aacrjournals.org/content/suppl/2020/06/11/2643-3230.BCD-20-0051.DC1

**Cited articles**
This article cites 55 articles, 15 of which you can access for free at:
http://bloodcancerdiscov.aacrjournals.org/content/1/1/112.full#ref-list-1

**Citing articles**
This article has been cited by 2 HighWire-hosted articles. Access the articles at:
http://bloodcancerdiscov.aacrjournals.org/content/1/1/112.full#related-urls

**E-mail alerts**
Sign up to receive free email-alerts related to this article or journal.

**Reprints and Subscriptions**
To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org.

**Permissions**
To request permission to re-use all or part of this article, use this link
http://bloodcancerdiscov.aacrjournals.org/content/1/1/112.
Click on "Request Permissions" which will take you to the Copyright Clearance Center's (CCC) Rightslink site.