The genomic and transcriptional landscape of primary central nervous system lymphoma

Josefine Radke, Naveed Ishaque, Randi Koll, Zuguang Gu, Elisa Schumann, Lina Sieverling, Sebastian Uhrig, Daniel Hübschmann, Umut H. Toprak, Cristina López, Xavier Pastor Hostench, Simone Borgoni, Dilafruz Juraeva, Fabienne Pritsch, Nagarajan Paramasivam, Gnana Prakash Balasubramanian, Matthias Schlesner, Shashwat Sahay, Marc Weniger, Debora Pehl, Helena Radbruch, Anja Osterloh, Agnieszka Korfel, Martin Misch, Julia Onken, Katharina Faust, Peter Vajkoczy, Dag Moskopp, Yawen Wang, Andreas Jödicke, Lorenz Trümper, Ioannis Anagnostopoulos, Dido Lenze, Ralf Küppers, Michael Hummel, Clemens A. Schmitt, Otmar D. Wiestler, Stephan Wolf, Andreas Unterberg, Roland Eils, Christel Herold-Mende, Benedikt Brors, Reiner Siebert, Stefan Wiemann, Frank L. Heppner

ARTICLE

OPEN

Check for updates

The genomic and transcriptional landscape of primary central nervous system lymphoma

Josefine Radke [1,2,3,68✉], Naveed Ishaque [4,68✉], Randi Koll[1,2], Zuguang Gu[5], Elisa Schumann[1,2], Lina Sieverling [6,7], Sebastian Uhrig[5,8], Daniel Hübschmann [5,8,9,10,11], Umut H. Toprak[12], Cristina López[13,14], Xavier Pastor Hostench [5], Simone Borgoni[15], Dilafruz Juraeva [5], Fabienne Pritsch[1,2], Nagarajan Paramasivam[5], Gnana Prakash Balasubramanian[16,17], Matthias Schlesner [18,19], Shashwat Sahay[3,4], Marc Weniger[20,21], Debora Pehl[1], Helena Radbruch [1], Anja Osterloh[1], Agnieszka Korfel[22], Martin Misch[23], Julia Onken[23], Katharina Faust[23], Peter Vajkoczy[23], Dag Moskopp [24], Yawen Wang[24], Andreas Jödicke[25], Lorenz Trümper[26], Ioannis Anagnostopoulos[27], Dido Lenze[27], Ralf Küppers [20,21], Michael Hummel[27], Clemens A. Schmitt [28,29], Otmar D. Wiestler[30], Stephan Wolf[31], Andreas Unterberg[32], Roland Eils [4], Christel Herold-Mende[32], Benedikt Brors [5], ICGC MMML-Seq Consortium*, Reiner Siebert[13,69], Stefan Wiemann [11,15,69] & Frank L. Heppner [1,2,33,34,69]

Primary lymphomas of the central nervous system (PCNSL) are mainly diffuse large B-cell lymphomas (DLBCLs) confined to the central nervous system (CNS). Molecular drivers of PCNSL have not been fully elucidated. Here, we profile and compare the whole-genome and transcriptome landscape of 51 CNS lymphomas (CNSL) to 39 follicular lymphoma and 36 DLBCL cases outside the CNS. We find recurrent mutations in JAK-STAT, NFkB, and B-cell receptor signaling pathways, including hallmark mutations in *MYD88* L265P (67%) and *CD79B* (63%), and *CDKN2A* deletions (83%). PCNSLs exhibit significantly more focal deletions of HLA-D (6p21) locus as a potential mechanism of immune evasion. Mutational signatures correlating with DNA replication and mitosis are significantly enriched in PCNSL. *TERT* gene expression is significantly higher in PCNSL compared to activated B-cell (ABC)-DLBCL. Transcriptome analysis clearly distinguishes PCNSL and systemic DLBCL into distinct molecular subtypes. Epstein-Barr virus (EBV)+ CNSL cases lack recurrent mutational hot-spots apart from IG and *HLA-DRB* loci. We show that PCNSL can be clearly distinguished from DLBCL, having distinct expression profiles, *IG* expression and translocation patterns, as well as specific combinations of genetic alterations.

Central nervous system (CNS) lymphomas are predominantly aggressive neoplasms involving brain, meninges, spinal cord, and eyes[1,2]. Two clinical subtypes of CNSL can be distinguished: primary central nervous system lymphoma (PCNSL), which is confined to the CNS; and secondary central nervous system lymphoma (SCNSL) presenting initially with systemic, non-CNS or synchronous systemic and CNS involvement. The term SCNSL comprises all systemic lymphomas that spread to the CNS and its presentation, tropism, outcome and therapeutic options differ from PCNSL[3,4]. Typically, SCNSL are classified as diffuse large B-cell lymphoma (DLBCL), while other types such as follicular lymphoma (FL), T-cell lymphoma or Hodgkin lymphoma are extremely rare[5,6].

PCNSL incidence is increased in immunocompromised patients, in which the tumor cells are typically Epstein-Barr virus (EBV)-positive[1,7,8]. In contrast, PCNSL in immunocompetent patients is typically EBV-negative. The mechanisms leading to the observed exclusive topographical restriction of PCNSL to the CNS are not fully elucidated[9]. PCNSL is classified as DLBCL in the vast majority of cases (approx. 90%) which immunohistochemically most often show a non-germinal center B-cell-like (non-GCB) immunophenotype[1,10,11] according to the Hans classification[12]. The tumor cells express pan B-cell markers (CD19, CD20, and CD79a), the germinal center (GC)-associated molecule BCL6[13], and the post-GC-associated marker MUM1/IRF4[14]. By gene expression profiling, the tumor cells are most closely related to late germinal center (exit) B-cells[15]. Pathomechanistic genomic alterations involving Toll-like- and B-cell receptor (TLR, BCR) signaling pathways are identified in previous studies revealing a very high frequency of somatic nonsynonymous mutations in genes such as *MYD88, CARD11*, and *CD79B*[16–20]. Additionally, often homozygous HLA class II[21,22] and *CDKN2A* loss, recurrent *BCL6* translocations[23,24] and structural variants at chromosome band 9p24.1 (affecting *CD274/PD-L1 and PDCD1LG2/PD-L2*)[25] as well as *TBL1XR1* variants[26] are repeatedly described in PCNSL[27,28]. These mutational patterns suggest PCNSL to be genetically similar to recently described "MCD", "C5" or "MYD88-like" subtypes for which a derivation from long-lived memory B-cells is proposed[29–35].

The outcome of PCNSL, even in immunocompetent hosts, is poor compared to most primary systemic DLBCL[36], though probably not worse than that of DLBCL of the MCD/C5 group in general[30]. High-dose methotrexate (MTX) remains the commonly administered therapy but the use of rituximab (monoclonal anti-CD20 antibody) is shown to be effective[37,38]. However, reports on rituximab efficiency in PCNSL are conflicting[39–42]. Genomic studies suggest that lymphoma cell proliferation and survival are driven at least in part, by deregulated TLR, BCR, JAK-STAT, and NFκB signaling pathways inducing constitutive NFκB activation[43–45]. Therefore, inhibitors up- and downstream of NFκB such as ibrutinib, known to inhibit Bruton's tyrosine kinase (BTK) as critical mediator of BCR signaling, and lenalidomide which is shown to have indirect effects on tumor immunity are applied and seem to be effective therapeutic alternatives in PCNSL[46–51]. PD-L1/2 blockade is discussed as another therapeutic option[52].

Despite all progress in the molecular characterization of PCNSL in the last decades, our understanding of the genetic and transcriptional alterations of PCNSL is by far not comprehensive. The few previous next generation sequencing (NGS) studies of PCNSL are limited to target enrichment only of exons, or whole-genome analysis of very few samples[44,53–58]. Therefore, pathogenic mechanisms other than coding variants, such as non-coding and regulatory changes, structural variants or mutational mechanisms related to the genome-wide distribution of somatic hypermutation (SHM) are not fully elucidated in PCNSL.

Unbiased omics profiling, such as whole-genome sequencing (WGS) studies integrated with transcriptome sequencing, are currently the methods of choice to illuminate the role of non-coding mutations[59,60]. In addition, these approaches can unravel various molecular mechanisms deregulating driver genes in PCNSL, which are necessary for diagnosis, risk stratification, and treatment in the era of precision and targeted therapies.

In this work, we perform whole-genome and transcriptome sequencing in 51 B-cell lymphomas presenting in the CNS, including 42 PCNSL samples from immunocompetent patients, to comprehensively describe the mutational and transcriptional landscape of PCNSL.

Results

Study cohort. We enrolled CNSL samples from 51 adults diagnosed with PCNSL or SCNSL. According to the site of manifestation, the following subgroups were defined: PCNSL within the brain parenchyma (PCNSL; $n = 39$), PCNSL with meningeal manifestation (PCNSL-M; $n = 3$), SCNSL within the brain parenchyma (SCNSL; $n = 3$), SCNSL with meningeal manifestation (SCNSL-M; $n = 3$) and EBV-positive lymphomas (EBV+; $n = 3$). Median age was 69, mean age was 66.5 years at diagnosis (range 40–82 years). The female:male ratio was 1.3:1. Follow up data were available for 44 patients. The follow up time ranged from 1 to 104 months with a median survival of 15.0 months (Supplementary Fig. 1a). The detailed study cohort information and subgroup-specific demographics are given in Fig. 1a, b and Supplementary Data 1. Patient samples were histologically and immunohistochemically classified according to the WHO criteria[2,11,61,62], and further stratified according to the Hans classification[12] into non-GCB ($n = 37$) and GCB subgroup ($n = 5$, Fig. 1c, Supplementary Data 1). For nine samples, the tissue was not sufficient for non-GCB/GCB characterization. Furthermore, we integrated data from the ICGC MMML-Seq cohort (www.icgc.org) for comparison of WGS and transcriptome data from systemic DLBCL, FL, naïve B-cells, and GC B-cells[34,59,60].

Mutational landscape of central nervous system lymphoma (CNSL). WGS data of 38 CNSL (30 PCNSL, 1 PCNSL-M, 2 SCNSL, 3 SCNSL-M, and 2 EBV+ samples, Fig. 1b) was obtained with a median coverage of 77 (range 31–100) for tumors and 45 (range 27–85) for matched germline controls. We identified a median of 18584 (range: 1987–48280; median of the 30 PCNSL: 20263 (range: 9185–48280)) total SNVs, of which a median of 5759 (range: 686–16731; PCNSL: 6274 (range: 2850–16731)) were intronic, a median of 10218 (range: 983–24033; PCNSL: 10790 (range: 5063–24033) were intergenic, and a median of 194 (range: 47–436), PCNSL: 200 (range: 100–436) were nonsynonymous exonic variants (1% of all SNVs). Furthermore, we identified a median of 2406 (range: 711–9430; PCNSL: 2485 (range: 941–9430)) indels per CNSL sample, of which the majority was intergenic (1333 (range: 403–5218), PCNSL: 1375 (range: 517–5218). The median number of variants (SNVs and indels) in non-coding RNA genes was 2744 (range: 551–6913), PCNSL: 2901 (range: 1220–6913). Selected variants were verified using Sanger sequencing (see "Methods" section).

The CNSL cohort presented a median of 152 (PCNSL: 147) SVs (range: 24–517 (PCNSL: 47–517), inversions: 21 (PCNSL: 20), deletions: 76 (PCNSL: 81), duplications: 20 (PCNSL: 19), translocations: 14 (PCNSL: 16)). We also investigated chromosome level CNVs (based on 30% or more of a chromosome being amplified or deleted) and found a median of 8 CNVs (median 1 cnLOH (PCNSL: 2), median 4 gains (PCNSL: 4), and median 2 losses (PCNSL: 3)). The detailed mutational statistics (CNVs,
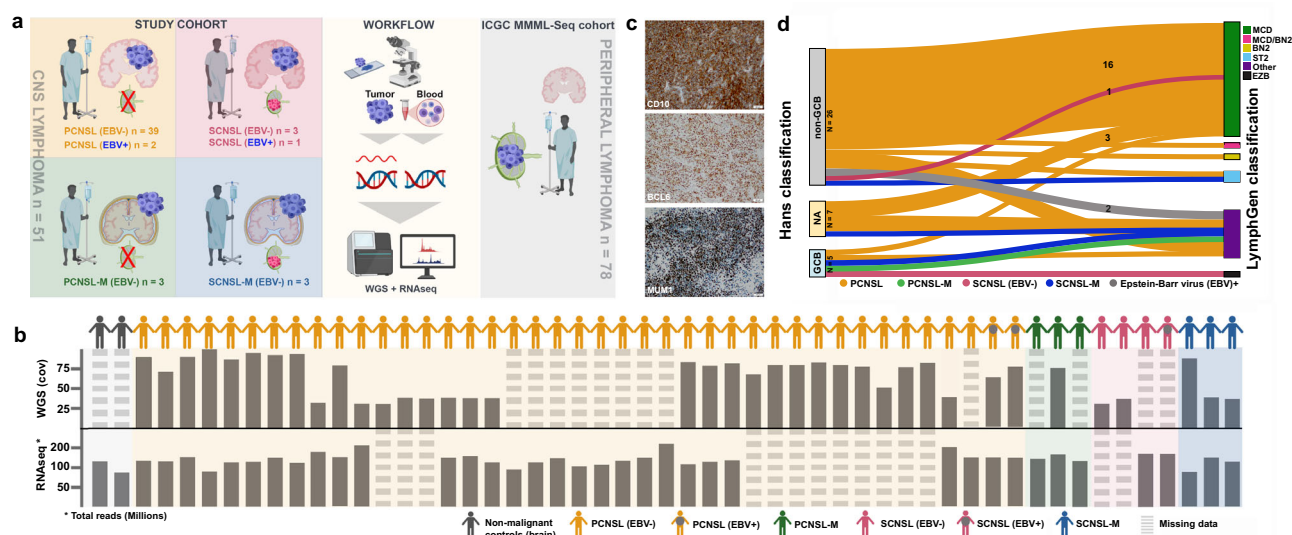
Fig. 1 Study design and multi-omic analysis of the CNSL cohort. Panel **a** demonstrates the study design, study cohort along with sample size, and sequencing approach. The central nervous system lymphoma (CNSL) cohort consists of 51 primary and secondary lymphoma (PCNSL and SCNSL) patients. Whole-genome sequencing was performed using tumor tissue and matched peripheral blood samples. For RNA sequencing, we additionally analyzed normal controls (non-malignant brain tissue (frontal lobe)) and included data from peripheral lymphomas without CNS manifestation for validation (ICGC MMML-seq cohort). The lower panel **b** lists all CNS lymphoma derived from PCNSL or SCNSL patients as well as controls. The depth of coverage (cov) for whole-genome sequencing and total number of reads in RNA sequencing are given for each tumor sample. Whole-genome sequencing data were obtained from $n = 38$ PCNSL/SCNSL patients. RNA sequencing data was generated from $n = 37$ PCNSL/SCNSL patients and $n = 2$ normal controls. In 24 PCNSL/SCNSL cases, we obtained patient-matched whole-genome and RNA sequencing data. Striped bars indicate data sets missing for individual samples. The CNSL specimen were examined according to the Hans classification (CD10, BCL6, and MUM1 (**c**)). Additionally, we classified all $n = 38$ whole-genome sequencing CNSL samples according to the genetic DLBCL subtypes using the LymphGen algorithm as described by Wright et al., 2020 (**d**). The results are displayed in a Sankey plot. Images in panels **a** and **b** were partially created with BioRender.com.

indels, SNVs, and SVs) of the CNSL, DLBCL, and FL samples are displayed in Supplementary Data 2.

PCNSL represents MCD genetic subtype of DLBCLs. Recent exome studies described the existence of different genetic subtypes of DLBCL, which show activation of distinct signaling pathways and different clinical outcomes[25,30,31]. We used the LymphGen algorithm described by Wright et al.[31] to classify our samples according to these genetic subtypes based on the obtained WGS data. The results of the CNSL cohort are displayed in Fig. 1d. In line with previous results[17,20,30,31], the majority of PCNSL samples were classified as MCD (based on the co-occurrence of *MYD88* L265P and *CD79B* mutations, 67%, 20/30). One sample was each assigned to BN2 (*BCL6* fusions and *NOTCH2* mutations, 3%, 1/30) and ST2 (*SGK1* and *TET2* mutated, 3%, 1/30), seven samples were non-subtyped cases ("Other", 23%, 7/30), and one sample was equally assigned to both groups BN2/MCD (3%, 1/30; Supplementary Data 1).

PCNSL samples classified as "Other" exhibited different CNV profiles affecting chromosome arms 1q, 2p, 2q, 3q, 4p, and 11p, as well as significantly more deletions of *CREBBP* compared to PCNSL samples classified as MCD by the LymphGen algorithm (Supplementary Fig. 2a). *CREBBP* gene inactivation is considered an early event in FLs and a subset of systemic DLBCL, mostly of GCB origin[63–67]. *CREBBP* inactivation is also described as a hallmark of the EZB class, but LymphGen's classification model is restricted to *CREBBP* point mutations and not focal deletions. The finding of a significantly increase number of *CREBBP* alterations ($p = 0.046$, Mann–Whitney U test) in PCNSLs classified as "Other" compared to MCD might, thus, imply a small subset of PCNSL to more resemble GCB-like DLBCL or, alternatively, the existence of a group of occult systemic

GCB-lymphomas with first clinical presentation in the CNS. Additionally, PCNSL-Other demonstrated significantly fewer mutations in *GRHPR, ETV6*, and *PIM1* (Supplementary Fig. 2b, c).

Driver mutations in CNSL. We first identified the genes recurrently mutated in CNSL (Fig. 2a) and used Metascape[68] for further pathway and process enrichment analysis. The top three level enriched terms were 'Regulation of hemopoiesis', 'Chromatin organization involved in negative regulation of transcription', and 'Cytokine signaling in immune system' ((hypergeometric test, FDR 8.91×10^{-9}, 1.04×10^{-4}, 1.17×10^{-4}, respectively; Fig. 2b). The enrichment analysis in TRRUST revealed 'Regulated by: STAT3' as the most significant term (hypergeometric test, FDR 3.98×10^{-7}; Fig. 2c). *STAT3* has been associated with intracranial spreading and poor survival in PCNSL[69,70], and reports of STAT3 inhibition via small molecules achieve complete tumor regression in vivo for lymphoma cell lines[71]. As *STAT3* is not highly mutated or hit by SVs or CNVs, its activation seems—in line with previous reports—induced by extrinsic factors such as infiltrating macrophages/microglial cells[72], or intrinsic factors such as activation downstream of MYD88[73].

Next, we used IntOGen and MutSigCV to discover putative driver mutations in the PCNSL WGS sub-cohort (Fig. 2d and Supplementary Data 3). We identified a total of 50 mutated driver genes, of which only 21 were previously known drivers. Many of the predicted drivers were associated with MCD enriched genes, including *MYD88* (67%), *CD79B* (63%), *OSBPL10* (83%), *HLA-A/B/C* (40%/63%/53%), *PRDM1* (40%), *TOX* (50%), *TBL1XR1* (40%), *CD58* (37%), *PIM1* (70%), *ETV6* (50%), *BCL11A* (30%), *CDKN2A* (83%), *GRHPR* (60%), *FOXC1* (20%), *and DAZAP1* (20%). These driver genes were significantly enriched for genes containing the *BCL6* binding motif
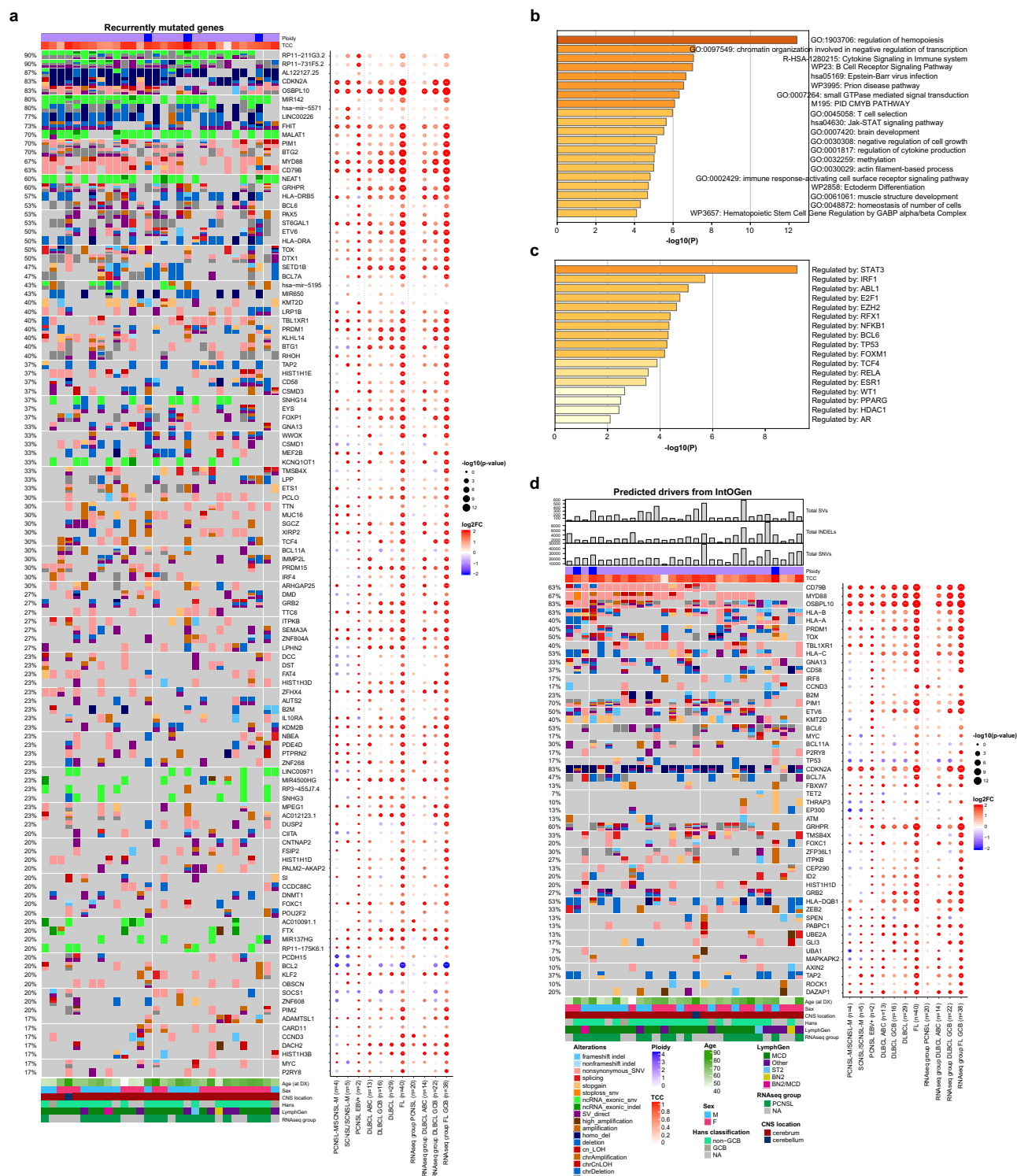
Fig. 2 Recurrent coding mutations and genomic drivers in PCNSL. Oncoprints of recurrently mutated genes, excluding IG and chromosome Y genes (**a**). Mutated genes are listed from top to bottom depending on their alteration frequency. The corresponding dot plot reflects the log2 fold change and significance of alteration frequencies in the other subcohorts and RNAseq subgroups compared to PCNSL. The recurrently mutated genes (listed in the oncoprint in a) were analyzed by Metascape[68] to identify pathway and process enrichment (**b**) and transcriptional regulatory networks (TRRUST) (**c**). Metascape adopts the hypergeometric test and employs the Benjamini-Hochberg correction for multiple testing. Oncoprints of driver genes in PCNSL (**d**). The top panel of the oncoprint shows the total numbers of structural variants (SVs), small insertions/deletions (INDELs), single nucleotide variants (SNVs), estimated ploidy, and genomic tumor cell content (TCC). Mutated genes are ranked by IntOGen. The corresponding dot plot reflects the log2 fold change and significance of alteration frequencies in the other subcohorts and RNAseq subgroups compared to PCNSL. In panels (**a**) and (**d**) the size of the dots demonstrate the significance according to a two-tailed Fisher's exact test not corrected for multiple testing.

(TRANSFAC and JASPAR PWMs, Enrichr enrichment test, adjusted $p = 0.03193$).

OSBPL10 was previously reported as a target of aSHM in PCNSL[53]. Consistent with observations in DLBCL[74], most of the identified mutations in PCNSL were confined to the exon 1 coding region (Supplementary Fig. 3).

Concerning *MYD88*, we only detected the classical pathogenic hotspot L265P mutation, which was validated by Sanger sequencing in all PCNSL samples investigated (100%, $n = 26$, Supplementary Data 4 and 5). Notably, *MYD88* mutation rates in the extension FFPE cohort were 78.9% in PCNSL and 55.6% in SCNSL. None of the five EBV-positive cases investigated harbored oncogenic *MYD88* L265P mutations, which is in line with previous findings[25,75]. Mutations in *TBL1XR1* also modulating TLR/MYD88 signaling[25] were identified in 40% of PCNSL (Fig. 2a, b). We investigated mutual exclusivity and co-occurrence patterns for *MYD88* among the driver genes that affect at least five patients using Fisher and CoMET test. We observed mutual exclusivity between alterations in *MYD88* and the NOTCH signaling inhibitor *SPEN*[30] (Fisher test, $p = 0.0009$, FDR $= 0.033$). In line with previous reports on ABC-DLBCLs[76,77], we found coexisting alterations in *MYD88* and *CD79B*. Nevertheless, this co-occurrence was not significant (Fisher test, $p = 0.16$, FDR $= 1.0$). However, *MYD88* was most significantly co-occurring with *TBL1XR1* (Fisher test, $p = 0.04$), both activating the NFκB signaling pathway[26]. Although this was not significant after correction for multiple testing (FDR $= 1$, Supplementary Data 6).

Compared to the MCD driver genes identified in the series presented by Wright et al.[31], our PCNSL series exhibited a higher proportion of samples with mutations in *PABPC1* (10% vs 0%), *P2RY8* (13% vs 1.2%), *ITPKB* (23% vs 2.5%), *GNA13* (20% vs 5.1%), and *B2M* (13.3% vs 2.8%). Furthermore, predicted driver genes in our PCNSL series included genes enriched in all other LymphGen classes: BN2 (*CCND3, BCL6, HIST1H1D, SPEN, PABPC1,* and *UBE2A*), EZB (*GNA13, IRF8, BCL7A, KM2TD,* and *EP300*), ST2 (*P2RY8, TET2, ZFP36L1,* and *ITPKB*) and A53 (*B2M* and *TP53*).

While the majority of identified drivers were reported by Wright et al., a number were not, including *FBXW7, ATM, TMSB4X,* THRAP3, *ID2, GRB2, ZEB2, GLI3, UBA1, MAPKAPK2, AXIN2, TAP2, ROCK1, CEP290,* and *HLA-DQB1*. These were previously recognized as general DLBCL drivers by Reddy et al.[78] and/or Chapuy et al.[29]. *ZEB2* was additionally identified as a genetic alteration associated either with the ABC subgroup[78] or the DLBCL C1 cluster[29] (Supplementary Data 3).

A remarkable finding was the identification of *MYC* mutations in 17% of PCNSL in the absence of *MYC* translocations. *MYC* alteration does not belong to the defining feature of the LymphGen algorithm nor has it been described as a driver in DLBCL by Chapuy et al.[29], though its functional relevance as oncogene in DLBCL has been shown by Reddy et al.[78]. Mutation of *MYC* in lymphomas is frequently linked to IGH translocations, which nevertheless are rare in the PCNSL as shown in the present as well as previous studies[23,79]. Whereas previous studies showing a high frequency of *MYC* mutations in PCNSL focused on the region underlying SHM in PCNSL[80], we here show that these mutations scatter across the gene (Supplementary Fig. 4). The function of the changes remains elusive but it is intriguing to speculate that at least part of them might contribute to the "double expression" of BCL2 and MYC in the absence of *MYC* translocation in PCNSL which has been associated with unfavorable outcome in systemic DLBCL[81].

Recurrent somatic alterations in non-protein-coding genes. The landscape of mutations affecting ncRNA in PCNSL was comparable to ABC-DLBCL, apart from significantly more mutations in *AL122127.1* and *AL122127.4* (Fig. 3a), situated in the *IGH* locus, and in *RP11-211G3.2*, situated in the first intron of *BCL6*. While the implications of these mutations are unclear, it is possible that these mutations are accumulated as part of the SHM/aSHM process affecting *IGH* and *BCL6*. Additionally, we identified recurrent aberrations in the aSHM target *MIR142* (80%; Fig. 3a, b) as well as *MALAT1* (70%) and *NEAT1* (60%), both located 53 kb apart on 11q13.1. The mechanistic roles of many ncRNAs are poorly understood because their exact function is difficult to assess. However, the lncRNAs *NEAT1* (nuclear enriched abundant transcript 1) and *MALAT1* (metastasis-associated lung adenocarcinoma transcript 1) are well known to play essential roles in the development and progression of various cancers by influencing gene expression by alternative splicing and epigenetic modification of regulatory elements[82–84]. Both, *MALAT1* and *NEAT1*, which have not been linked to PCNSL before are known to be mutated and highly expressed in DLBCL[34] and predict poor prognosis[85,86]. Further aberrations in lncRNAs affected *KCNQ1OT1* (33%) and *SNHG3* (23%), both reported to have oncogenic functions in multiple cancers[87,88] as well as *SNHG14* (37%), promoting immune evasion in DLBCL[89].

Kataegis shapes the mutational repertoire of PCNSL. Kataegis is a pattern of mutational hotspots that has been associated with a number of cancers[90], and is a frequent consequence of AID activity in lymphomas[91]. Many of the recurrently mutated genes in PCNSL were dominated by alterations that are located in these highly mutated hotspots[11], of which several have previously been described as targets of aSHM, such as *OSBPL10, PIM1, BTG2,* and *PAX5* (Fig. 3b)[25,53,80]. Of the 50 identified protein-coding driver genes and the top 50 mutated ncRNA in PCNSL, 15 and 21 were targeted by kataegis, respectively (Figs. 2a, b, 3a, b, additional supplement [https://doi.org/10.5281/zenodo.6054242][92]. Consistent with previous reports[93], expression of miRNA, lncRNA, antisense RNA, and protein coding genes with kataegis loci were expressed significantly higher than those without (Wilcoxon rank sum test, $p < 0.05$; Fig. 3c). This implicates that either aSHM preferentially targets highly expressed genes, or that aSHM may cause hyperactivation of these genes. Interestingly, the largest difference in RNA expression was observed in miRNA genes, again highlighting the importance of the non-coding alterations in PCNSL. This observation was consistent for subgroups, including systemic DLBCL (Supplementary Fig. 5).

Physiologically, SHM is the process of introducing mutations in the antibody genes to alter the antigen-binding site, increasing the immunoglobulin (IG) diversity[94]. Kataegis events were at *IGH* (100%), *IGL* (100%) and *IGK* (70%) loci but were also found outside IG loci, targeting *BTG2* (63%), *GRHPR* (50%), *PIM1* (43%), *DTX1* (40%), *OSBPL10* (37%), *ZNF860* (37%), *BCL6* (33%), *RHOH* (33%), *CXCR4* (30%), *BACH2* (27%), and *PAX5* (27%; Fig. 3b). The recurrently targeted genes in PCNSL mostly overlapped with those targeted in ABC-DLBCLs. However, samples with mutational hotspots in *BTG2, GRHPR, OSBPL10* and *ZNF860* were significantly more frequent in PCNSL (in 18, 15, 11 and 11 of 30 samples, respectively) compared to ABC-like DLBCL (in 2, 0, 0 and 0 of 13 samples, $p = 0.009, 0.001, 0.019$ and 0.019, Fisher's exact test, respectively). Taking all non-IG genes that overlapped a mutational hotspot in at least one PCNSL sample (242 genes, Supplementary Data 7), we found the BCR signaling pathway to be most significantly enriched (Enrichr enrichment test, adjusted $p = 0.0046$). Taken together, kataegis and aSHM play a decisive role in shaping the mutational repertoire of PCNSL and are associated with functional pathways in PCNSL pathogenesis.
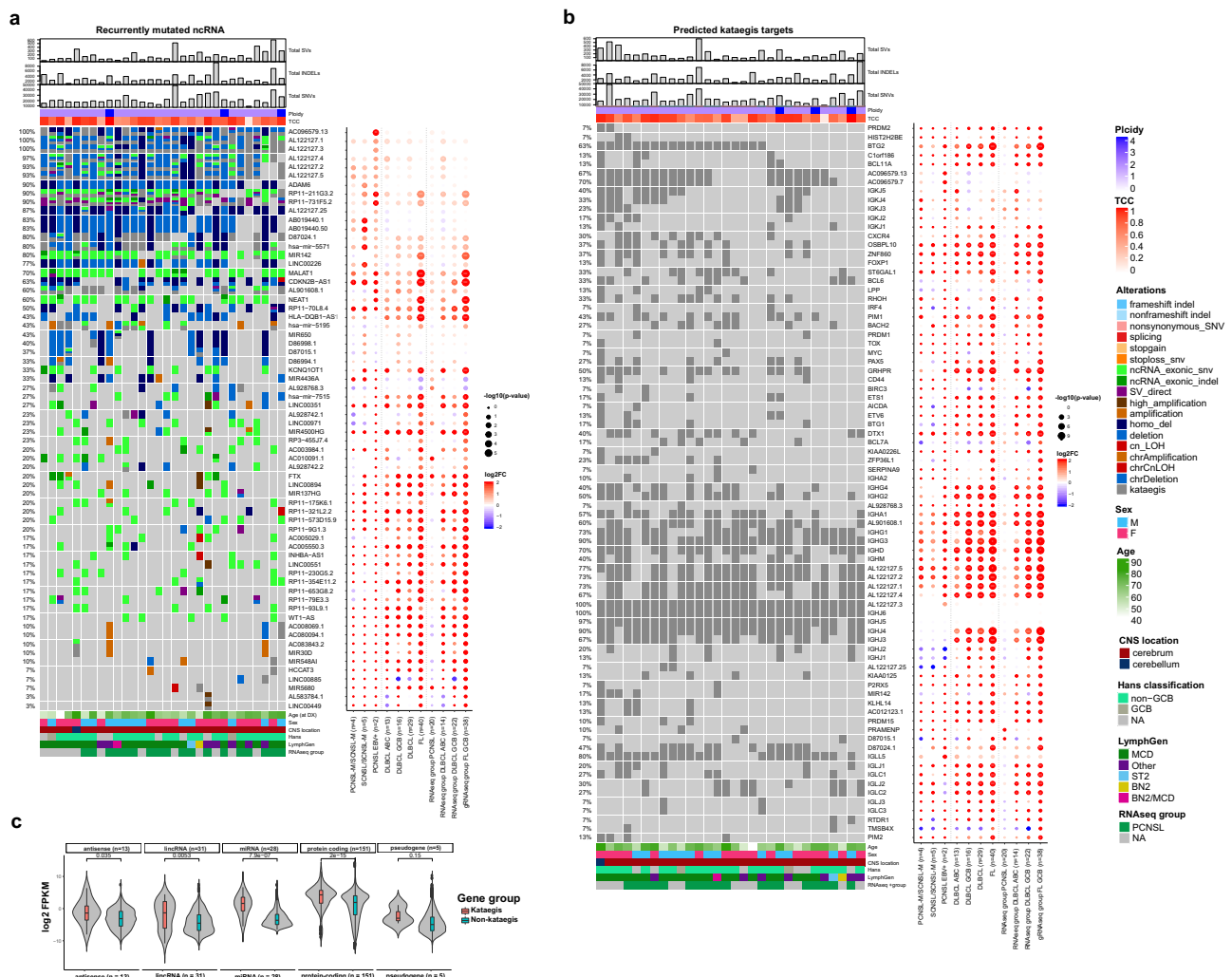
Fig. 3 Recurrent non-coding RNA mutations and Kataegis events in PCNSL. Oncoprint of recurrent non-protein-coding genes in PCNSL (**a**). Shown are the total numbers of structural variants (SVs), small insertions/deletions (INDELs), single nucleotide variants (SNVs), estimated ploidy, and tumor cell content (TCC) (top panel). Mutated genes are listed from top to bottom depending on their alteration frequency. Oncoprint of recurrent kataegis events in PCNSL (**b**). Mutated genes are listed from top to bottom depending on chromosome location. The corresponding dot plot reflects the log2 fold change and significance of alteration frequencies in the other subcohorts and RNAseq subgroups compared to PCNSL. In panels **a** and **b** the size of the dots demonstrate the significance according to a two-tailed Fisher's exact test not corrected for multiple testing. The violin plot **c** shows that the RNA expression of genes with kataegis loci compared to those without were significantly higher for antisense, long non-coding RNA, miRNA and protein coding genes (one-sided Wilcoxon rank sum test, $p < 0.05$). Box and whisker plots are embedded in the violin plots, showing median (center line), the upper and lower quartiles (the box), and the range of the data (the whiskers), excluding outliers.

While patterns of aSHM and kataegis were similar between CNSL and systemic DLBCL subtypes, we identified that EBV+ CNSL cases did not share many of the recurrent mutational hotspots apart from IGH and the *HLA-DRB* locus. (Fig. 3b, Supplementary Fig. 6a, b, Supplementary Data 5).

Recurrent copy number alterations (CNAs). Compared to systemic ABC-DLBCL and GCB-DLBCL of which the copy number profiles reflected previously published results[34,60,95], PCNSL demonstrated significantly more CN losses in 6p21 (*HLA-D* locus, Fig. 4a–c, Supplementary Data 8) as well as recurrent losses in 9p21 (*MTAP*, *CDKN2A/B*) and 19p13 (*CDKN2D*). The loss of the *HLA-D* locus that encode for MHC class II molecules lead to reduced immune surveillance and poor survival in DLBCL[96]. *CDKN2A* is an established tumor suppressor gene with roles in angiogenesis, cell death, invasiveness, and growth suppression[97–99]. Additionally,

we found deletions on chromosomes 1p13 and 3q13, affecting genes such as *CD58* and *CD80*, both candidates reported to lead to immune evasion[100]. Further CN losses were detected on chromosomes 8q12 (*TOX*), 12p13 (*ETV6*), and 15q21 (*B2M*) as well as 3p14, affecting the fragile site tumor suppressor gene, fragile histidine triad (*FHIT*). *TOX* deletions have been previously described by array-based imbalance profiling[101]. *TOX* is required for the development of various T-cell subsets and was described as putative tumor suppressor in MCD DLBCL[30]. *TOX* downregulation has been associated with poor prognosis in different cancers[102] and is a predictor for anti-PD1 response[103]. Significant CN gains in PCNSL mapped to 2q37 and 18q21 affecting *DIS3L2* and *MALT*. *DIS3L2* encodes for an exoribonuclease that is responsible for Perlman syndrome[104] and was recently described to promote HCC tumor progression by upregulating production of the oncogenic isoform of *RAC1*, *RAC1B*[105]. *MALT* is a regulator of NFκB signaling and potential therapeutic target in B-cell lymphoma[106].
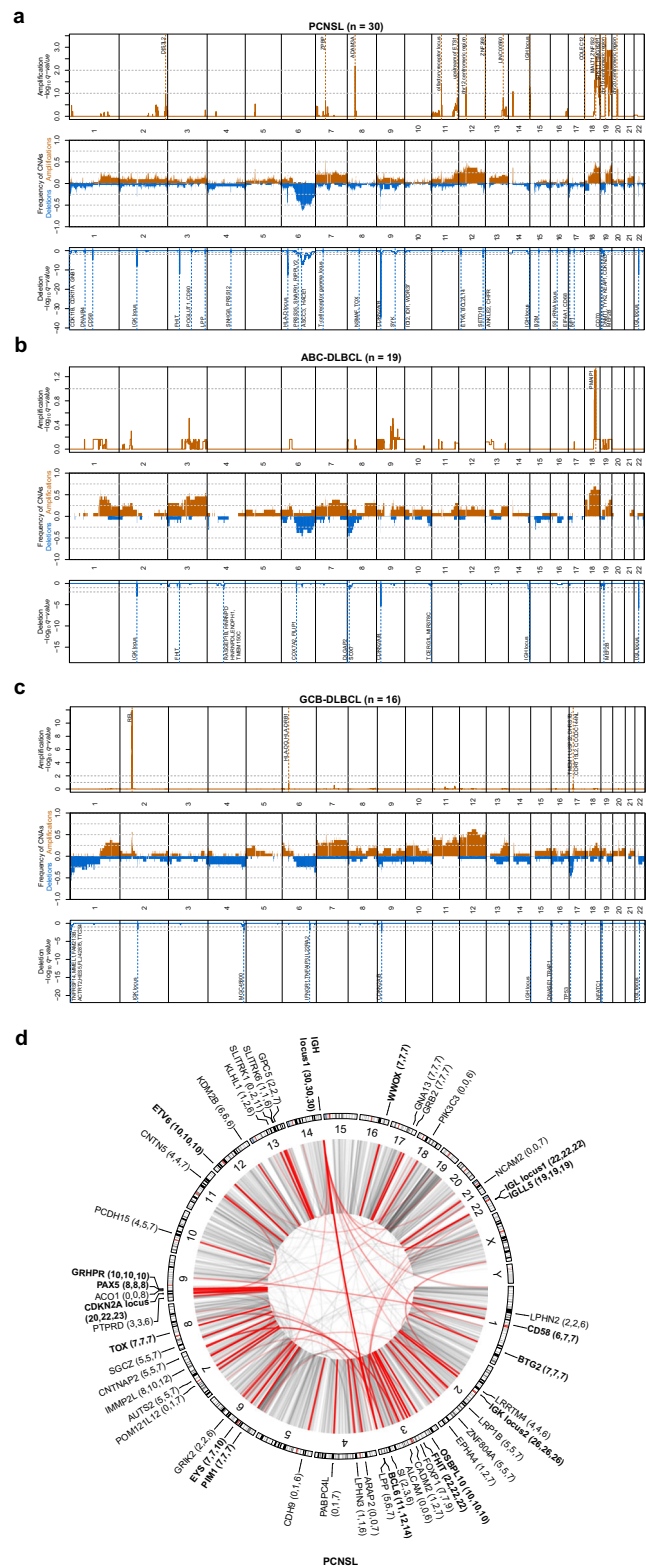
Fig. 4 Genomic structural variation in PCNSL. Recurrent somatic CNAs in PCNSL (**a**), ABC-DLBCL (**b**), and GCB-DLBCL (**c**). Relative prevalence of somatic copy number aberrations in tumor samples (middle panel), showing the presence of at least one copy number gain (orange bars), copy number loss (blue bars), as a proportion of analyzed samples. Significantly (q-value < 0.25) amplified and deleted regions and candidate genes are shown (upper and lower panel). For GCB-DLBCL, we added TP53 as this was detected in a significant broad deletion (Gistic2 p-value 0.0311), and the focal peak falls on the region including TP53. Circular visualization of genome rearrangements in PCNSL (**d**). The panels (from outside going inwards) represent recurrence per gene, chromosome ideograms, and chromosome numbering. Next to gene names, the number of SV breakpoints that lie direct on the gene, within 100 kbp of the genes, and closest to that gene are reported. Inter-chromosomal translocations are rendered with black (all) and red (highlighted if affected in 20% of samples) arcs in the center of the display. The significance of CNVs was calculated using the GISTIC 2.0 permutation test with Benjamini–Hochberg correction for multiple testing.

Recurrent structural variations (SVs). We defined SVs as genomic breakpoints, which can correspond the borders of amplifications and deletion, but also balanced translocations and inversions. PCNSL showed a median of 147 SVs (range: 24–517, Supplementary Data 2). IG gene rearrangements were found in all PCNSL, ABC-DLBCL, and GCB-DBCL cases and affected the *IGH* (100, 100, and 100%), *IGL* (73, 46, and 31%) and *IGK*

(87, 54, and 63%) loci. Furthermore, direct SVs affected *FHIT* (73, 23, and 38%), *CDKN2A* (67, 38, and 25%), *BCL6* (37, 21, and 19%), *OSBPL10* (33, 8, and 13%), *ETV6* (33%, 15%, 6%), *PAX5* (27, 0, and 13%), *PIM1* (23, 0, and 6%), *TOX* (23, 8, and 19%), *BTG2* (23, 8, and 0%), *WWOX* (23, 8, and 25%), as well as *CD58* (20, 8, and 19%; Fig. 4d, Supplementary Data 9). *WWOX* and *FHIT* represent common fragile site (CFS) and have been classified as tumor suppressor genes in DLBCL[107,108].

Recent studies have shown that translocation can act as enhancer hijacking even when the events is several hundred thousand base-pairs away from target genes[109]. To investigate this, we also annotated SV breakpoints to genes within 100 kbp and also to the closest genes. We found a number of genes involved in G protein-coupled receptor signaling (*ARAP2, LPHN2, LPHN3, EPHA4, ADGRL2,* and *GPC5*) consistent with observations in pan-cancer studies[110]. A number of other genes exhibited at least three times as many distal translocations (while still being the closest gene) than directly on the gene, including *PIK3C3, EPHA4, SI, ALCAM, NCAM2, CADM2, CDH9, PABPC4L, GRIK2, POM121L12, ACO1, KLHL1, SLITRK1,* and *SLITRK6*. Hyperactivation of PI3K signaling is one of the most common events in human cancers, and PIK3C3 has been shown to promote cell proliferation[111] and autophagy[112], and its inhibition has shown therapeutic benefit in bladder, hepatocellular (HCC), and colon cancer[113–115]. *EPHA4* has been described to promote cell proliferation and migration[116,117] and was associated with tumour aggressiveness and poor patient survival in human breast and rectal cancer[118,119]. Inhibition of EphA4 has been shown to overcome intrinsic resistance to chemotherapy[120]. Many of these other potential enhancer-hijacking targets do not have well-established roles in cancer pathogenesis, however, we did notice a number of genes involved in cell adhesion (*ALCAM, NCAM2, CADM2,* and *CDH9*) and 2 SLIT and NTRK like family members (*SLITRK1, SLITRK6*).

Immunoglobulin translocations implicate distinct CNSL subtypes. IG translocations are established oncogenic drivers of many lymphatic neoplasms[121–123]. *IGH-BCL6* fusions are recurrent in PCNSL[24], which mirrors observations of ABC-DLBCL[124]. *IGH-BCL2* fusions are more prominent in GCB-DLBCL[125]. We investigated the recurrent translocations (≥2 patients) in our cohort and identified five CNSL samples with *IGH-BCL6* translocations (Fig. 5a and Supplementary Fig. 7a–d). We also identified three cases with *IGH-BCL2* translocations (Fig. 5b and Supplementary Fig. 7e, f) one in each of SCNSL, SCNSL-M, and PCNSL-M,
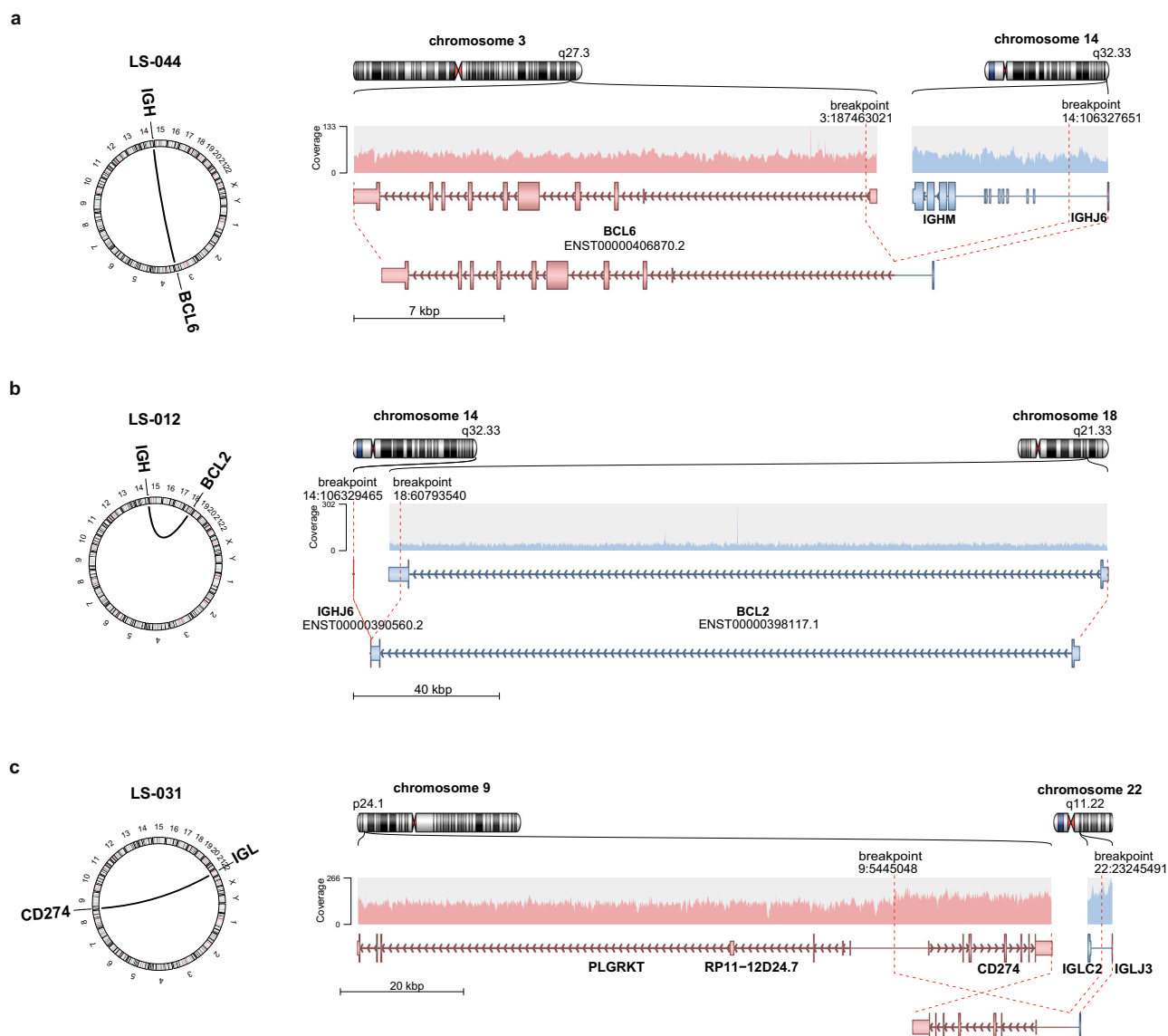
Fig. 5 Immunoglobulin translocations in PCNSL. Schematic representation of the translocation breakpoints involving *BCL6* (PCNSL patient LS-044; **a**), *BCL2* (SCNSL-M patient LS-012; **b**), and *CD274* (PD-L1; PCNSL patient LS-031; **c**). The left panels show a circular plot displaying the location of the translocation partners. The right panel shows the original and reconstructed translocation events. Shown are (from top to bottom) chromosome ideograms, read overage of the sites, the gene models, and the reconstructed translocation partners.

further implicating that meningeal and secondary CNSL are distinct from intraparenchymal PCNSLs. Two PCNSL cases showed *IGL* and *IGH* translocations with breakpoints close to *CD274* (PD-L1; Fig. 5c, Supplementary Fig. 7g), which resulted in strong PD-L1 protein expression (Supplementary Fig. 7h) and therefore implicates a potential target for immunotherapy. All other, non-recurrent translocations are listed in Supplementary Data 10.

The analyses of the IG breakpoints provided in all informative junctions evidence that these occurred due to illegitimate CSR or aberrant SHM, with the notable exception of the *IGH-BCL2* junctions, which were the consequence of an aberrant VDJ rearrangement (Supplementary Data 11). Thus, all IG translocations in PCNSL are supposed to occur in the GC process rather than in a pre B-cell.

Mutational signatures in PCNSL. Mutational signatures were analyzed with regard to SNVs (single base substitutions, SBS) and indels (ID) of all tumor samples as defined by

Alexandrov et al.[126] (Fig. 6). For single base substitution signatures (SBS) we found mutational patterns that have been associated with spontaneous deamination of 5-methylcytosine (SBS1), defective activity of the AID/APOBEC family (SBS2), failure of double-strand DNA break repair by homologous recombination (SBS3), SHM (SBS9), and damage by reactive oxygen species (SBS18). Additionally, the samples frequently revealed mutations caused by mutational signatures SBS5, SBS17b, and SBS40, which are of unknown etiology (Fig. 6a). The presence of SBS3, hallmark of defective DNA break repair by homologous recombination, and SBS40 may be therapeutically relevant as these indicate potential effectiveness of combination therapy with PARP inhibitors (e.g., Olaparib) alongside cytotoxic chemotherapy[127,128]. The three most prominent signatures in DLBCL, FL, and CNSL were SBS9, SBS5, and SBS40 (Fig. 6b). Direct comparison of PCNSL and DLBCL revealed that signature SBS1, which correlates with DNA replication at mitosis (mitotic clock)[126], was significantly enriched in PCNSL ($p = 0.0027$; Fig. 6c, Supplementary Fig. 8a–g).
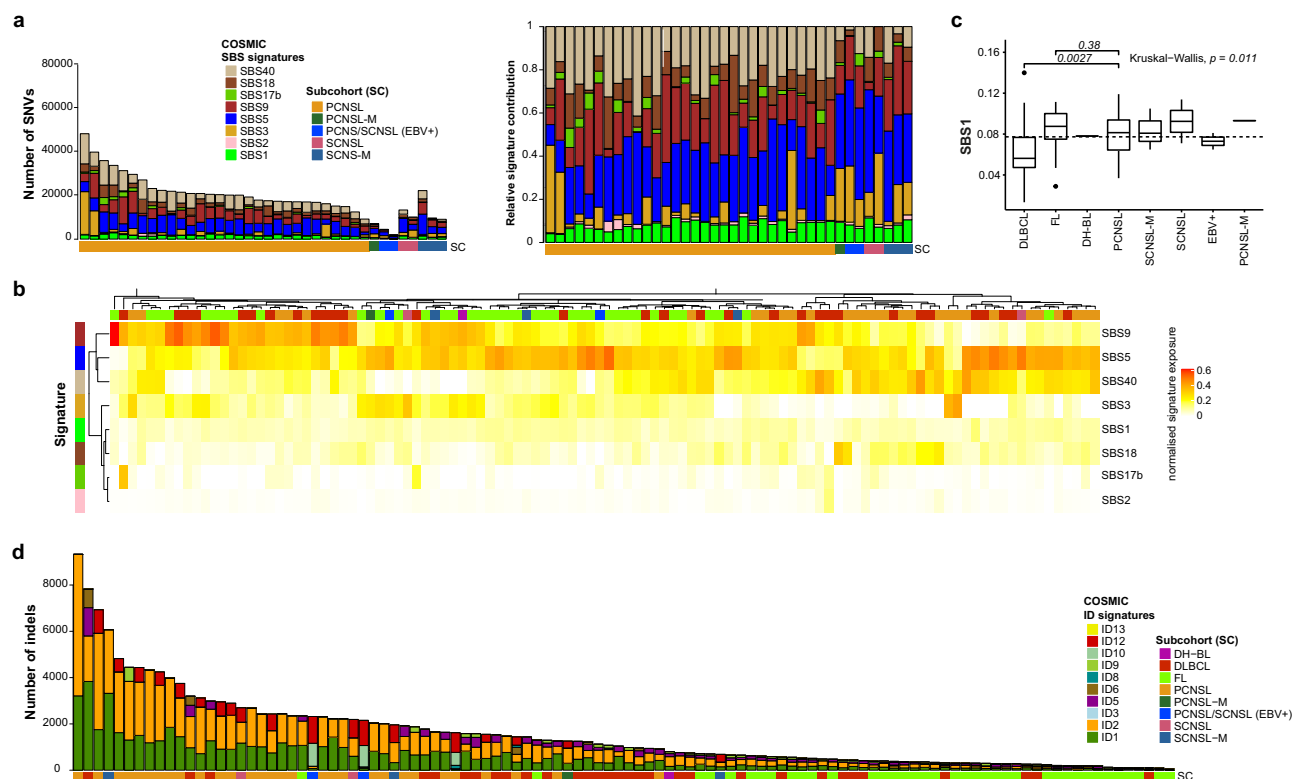
Fig. 6 Mutational signatures in PCNSL. Single base substitution (SBS) signature contribution in PCNSL and SCNSL. Stacked bar plots are ordered by subgroup and then decreasing mutations load (**a**). The left bar plot shows the number of SNVs and the right shows the normalized signature exposures. Each color corresponds to a mutations signature and the proportion of the color reflects the number or proportion of SNVs explained by a certain mutational signature. The heatmap shows the clustering pattern of the SBS mutational signatures in all CNSL and peripheral lymphomas (**b**), which revealed groupings mutational signatures SBS5, SBS9, and SBS40. Pairwise comparison of PCNSL with DLBCL or FL (Mann–Whitney U test) revealed that signature SBS1 was significantly enriched in PCNSL compared to DLBCL ($p = 0.027$; **c**). Box and whisker plots show median (center line), the upper and lower quartiles (the box), and the range of the data (the whiskers), excluding outliers. Small insertion and deletion (ID) signatures in CNSL and peripheral lymphoma (**d**) demonstrate elevated numbers of indels and mutational patterns associated with slippage during DNA replication of the replicated DNA strand (ID1) and template DNA strand (ID2) in PCNSL.

Analysis of small insertion and deletion signatures (ID) revealed mutational patterns associated with slippage during DNA replication of the replicated DNA strand (ID1) and template DNA strand (ID2); both of these signatures appeared significantly ($p < 1 \times 10^{-4}$, Wilcoxon) more prominent in PCNSL compared to DLBCL and FL (Fig. 6d), though different read-depths may have influence this analysis.

Interestingly, only CNSL samples but not DLBCL or FL revealed mutations caused by mutational signature ID12 that is of unknown etiology and has been observed in prostate adenocarcinoma and soft tissue liposarcoma[126].

PCNSL RNA expression signatures are distinct from systemic DLBCL. The relative rarity of PCNSL and limited availability of fresh frozen tissue have thus far complicated the implementation of larger molecular studies needed for patient stratification. To unravel the molecular signature of PCNSL, we employed an unsupervised consensus clustering approach (using the cola tool[129]) to identify expression groupings between PCNSL samples and samples from the ICGC MMML-seq project (mainly consisting of non-GCB and GCB type DLBCLs, and FLs). This yielded the following major clusters: FL, PCNSL, GCB-type DLBCL, ABC-type DLBCL, non-tumorous GC B-cells, and naïve B-cells (Fig. 7a). For each cluster, we identified signature gene sets that significantly correlated with the groupings. Interestingly, all meningeal PCNSL (PCNSL-M) and SCNSL-M grouped together

with either GCB- or ABC-DLBCL, clearly indicating that these subtypes are molecularly and pathomechanistically distinct from intraparenchymal CSNL, which formed one separate cluster suggesting a distinct signature of CNS tropism. The ABC-type DLBCL cluster was enriched for *MYD88* mutant samples, which were still distinct from *MYD88* mutant PCNSL at the gene expression level (Fig. 7a).

To further exclude an impact of potentially contaminating surrounding CNS tissue on gene expression signatures, we analyzed total RNA from normal brain controls ($n = 2$) and compared this to PCNSL. To investigate the gradient of various tumor cell contents of samples, we spiked increasing concentrations of RNA from non-diseased brain tissue into a PCNSL sample with very high tumor cell content (0, 20, 40, 60, and 80%). Then, we further stratified the PCNSL group by another round of consensus clustering using two different classification methods, which both revealed two groups (Fig. 7b and Supplementary Fig. 9a–c). The first PCNSL expression group (PCNSL subcluster 1) consisted of samples with high tumor cell content (determined by WGS and histopathological analysis). Expression of its signature gene set did not show similarity to normal brain tissue expression. However, the second PCNSL expression group (PCNSL subcluster 2) contained mainly samples with lower tumor cell content, and expression of its signature gene set was indeed similar to normal brain tissue expression (Fig. 7b). We identified the PCNSL signature gene sets relative to ABC and GCB type DLBCLs and FLs, and removed potential background signatures from contaminating brain tissue.
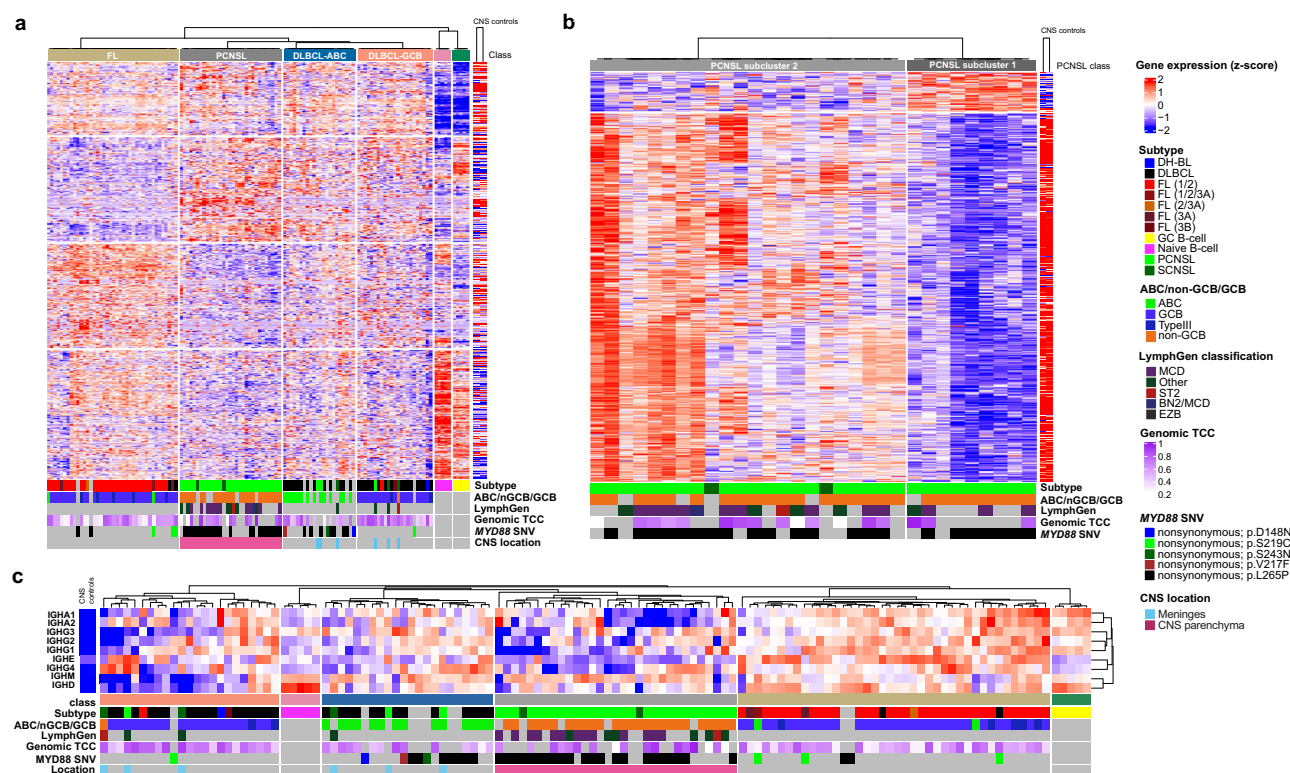
Fig. 7 Transcriptomic signatures distinguish PCNSL from DLBCL. Global gene expression patterns clearly separate PCNSL from other subtypes (ABC-DLBCL, GCB-DLBCL, and FL; **a**). The heatmap shows unsupervised consensus clustering of gene expression data. All SCNSL-M and PCNSL-M cluster with non-CNS-DLBCL, distinct from intraparenchymal PCNSL. The ABC-DLBCL cluster is enriched for *MYD88* mutant samples, which are distinct from *MYD88* L265P mutant PCNSL. Samples of normal germinal center (GC) and naïve B-cells were included as controls. Furthermore, normal brain tissue (CNS controls) samples were added to analyze the impact of normal brain tissue contamination at the RNA level. Consensus clustering (skmeans) with intraparenchymal PCNSL samples and CNS controls revealed two groups (**b**). The first PCNSL expression group (PCNSL subcluster 1, on the right) consisted of samples with high tumor cell content, the second (PCNSL subcluster 2, on the left) contained mainly samples with a lower tumor cell content, which signatures correlated well with normal brain tissue expression. PCNSL can be distinguished from ABC-DLBCL, GCB-DLBCL, and FL based on the expression of *IG* constant genes (**c**).

The marker genes in each group were identified based on differential gene expression analysis (Supplementary Data 12). Among the marker genes for PCNSL were e.g., *LAPTM5*, a CD40-related gene expressed in malignant B-cell lymphoma[130] and *ITGAE*, mediating cell adhesion, migration, and lymphocyte homing through interaction with E-cadherin[131]. We used Metascape[68] for functional analysis of marker genes. Further pathway and process enrichment analysis revealed that pathways such as 'ribonucleoprotein complex biogenesis', 'mRNA processing', 'cell cycle', 'RNA modification', 'DNA conformation change', and 'DNA-templated transcription, initiation' were enriched (Supplementary Fig. 9d). The top three level Gene Ontology biological processes included 'cellular component organization or biogenesis', 'metabolic process', and 'localization' (Supplementary Fig. 9e).

Expression of IGHM is characteristic for PCNSL. Additionally, we analyzed the expression of *IG* constant genes, which again revealed the same clusters as the unsupervised consensus clustering approach, demonstrating that PCNSL can be differentiated from DLBCL based on only the expression of *IG* constant genes. In contrast to DLBCL and FL, PCNSL show generally low expression of *IG* constant genes, but higher expression of *IGHM* (Fig. 7c).

TERT expression but not telomere content upregulated in PCNSL. Telomerase activity and telomerase reverse transcriptase (*TERT*) gene expression have been reported as prognostic factors

in PCNSL patients[132]. We used TelomereHunter, a software for detailed characterization of telomere maintenance mechanisms[133] to estimate the telomere content in a representative cohort of PCNSL, SCNSL, peripheral lymphoma, as well as non-tumorous naïve and GC B-cells as control[59]. In approximately 1/3 of the samples, the TCC-corrected telomere content was higher in the tumor than in the matched control (whole blood) (Fig. 8a and Supplementary Fig. 10a). Nevertheless, telomere content was not significantly different between the different histological, clinical and molecular subgroups irrespective of whether the results were corrected for TCC (Fig. 8b and Supplementary Fig. 10b (tumor/control log2 ratio)) or not (Supplementary Fig. 10c (uncorrected for the control sample)). As expected, we found a negative correlation between age and telomere content in the control (Supplementary Fig. 10d). However, expression of the *TERT* gene, the main activity of the encoded protein is the elongation of telomeres, was significantly higher in GC B-cells[59] and in PCNSL compared to ABC-DLBCL (Fig. 8c and Supplementary Fig. 10e). This was consistent with observations when stratifying samples by RNA subgroups, where *TERT* expression was significantly higher in PCNSL compared to ABC-DLBCL, GCB-DLBCL, and FL (Fig. 8d).

Interestingly, the higher *TERT* expression in PCNSL significantly correlated with normalized telomere content (uncorrected for the control sample: Pearson's $R = 0.67$, $p = 0.003$; (Fig. 8e) and telomere content T/C log2 ratio (Supplementary Fig. 10f)). However, SCNSL-M, DLBCL and FL did not show such trends
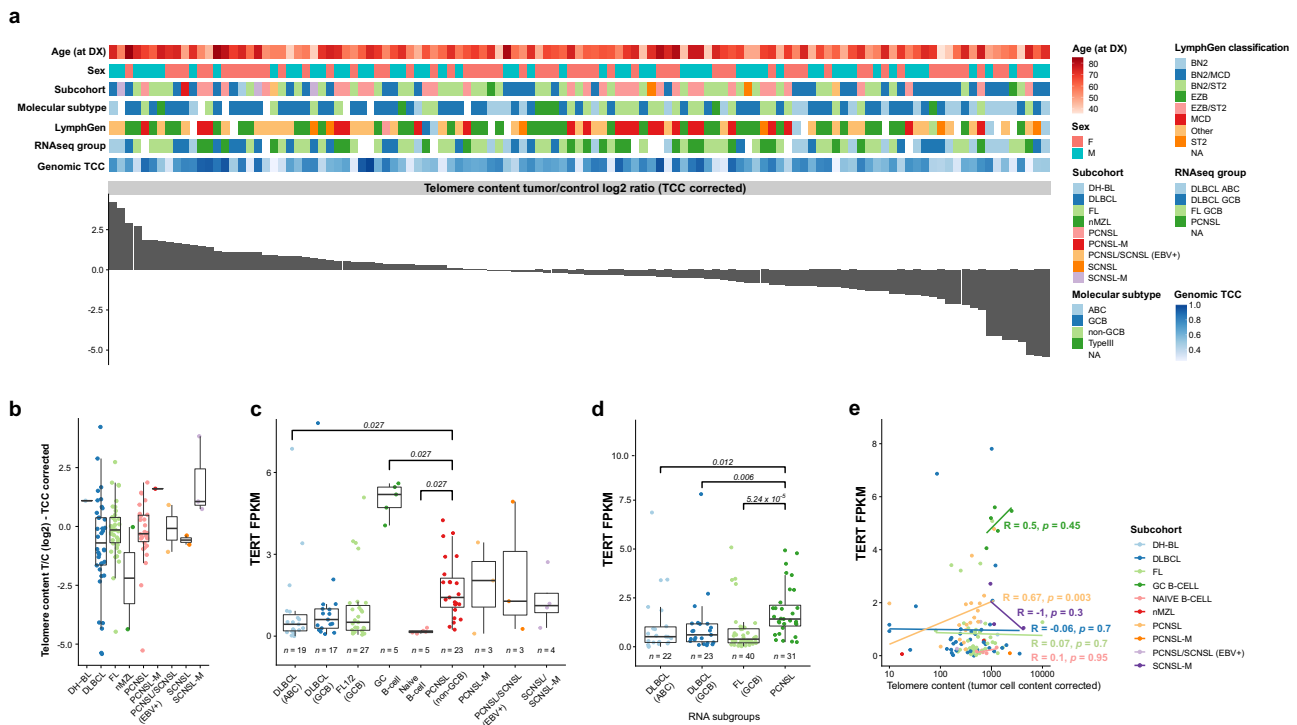
Fig. 8 TERT expression correlates with telomere content in PCNSL. Telomere content in DLBCL, FL, PCNSL, and SCNSL was estimated from WGS data using TelomereHunter. In about 1/3 of the samples, the telomere content was higher in the tumor than in the control sample (**a**). There was no statistical difference between PCNSL and other subcohorts in telomere content (**b**). TERT expression was significantly higher in PCNSL compared to subcohort ABC-DLBCL (**c**) and compared to RNA subgroups ABC-DLBCL, GCB-DLBCL, and FL (**d**). *P*-values in panels **c** and **d** are calculated using the one-sided Wilcox test adjusted for multiple testing using the Holm-Bonferroni method. Box and whisker plots show median (center line), the upper and lower quartiles (the box), and the range of the data (the whiskers), excluding outliers. A significant positive correlation between TERT expression and telomere content was only observed in PCNSL (**e**), using the Pearson correlation-coefficient test without correction for multiple testing.

($R = -1$, $p = 0.3$, $R = -0.06$, $p = 0.7$, and $R = 0.07$, $p = 0.7$, respectively (Fig. 8e)). This suggests that *TERT* has an active role in combatting telomere degradation in PCNSL. Two well-known promoter hotspot mutations (-124C>T (C228T) and -146C>T (C250T)) have been described to increase *TERT* expression and cell-cycle progression[134,135]. These mutations have been found in several solid and hematological malignancies including different brain tumors and PCNSL[136–138]. Therefore, we next investigated the *TERT* promoter mutation status in our WGS ($n = 38$) and FFPE extension cohort ($n = 31$). Sanger sequencing of the *TERT* promoter region was performed (i) for all WGS samples having only low coverage in the promoter sequence (below 40×, $n = 6$, Supplementary Fig. 10g), (ii) the FFPE extension cohort, and (iii) three oligodendrogliomas, known to carry high frequency *TERT* promoter mutation[139]. We detected no *TERT* promoter mutations in 67 samples of PCNSL and SCNSL (Supplementary Data 5, Supplementary Fig. 10h), while the well-known TERT rs2853669 polymorphism, which has been associated with increased cancer risk[140], was identified in 40% (14/35 (8 PCNSL, 6 SCNSL)) of the patients in the extension FFPE cohort. The Sanger sequencing results of two samples were not conclusive.

Discussion
Here we have performed a comprehensive analysis of recurrent protein coding and non-coding mutations, CNVs, SVs, and driver mutations in a large cohort of PCNSL and compared the genetic features to systemic DLBCL and FL. The vast majority of PCNSLs are of non-GCB-DLBCL subtype[141] and share many genetic alterations with non-CNS ABC-DLBCL in the same signaling pathways. Previous studies made use of whole-exome sequencing[25,53] which (i) limits the investigation to protein-

coding regions and (ii) may not be ideal for understanding the patterns of mutational hotspots—e.g., attributed to AID induced SHM in B-cell non-Hodgkin lymphomas[142]—as well as the structural variation in genomes[143]. PCNSL showed significantly more SNVs and indels compared to systemic DLBCL, even in intronic and intergenic regions, also underlining the importance of non-protein coding aberrations in PCNSL pathogenesis. Many of the recurrent mutations in non-protein coding genes affected non-coding RNAs (ncRNAs), which are among other functions involved in epigenetic regulation of gene expression, cell differentiation, and development[82,144]. The molecular profile of SCNSL, on the other hand, corresponded to that of systemic DLBCL.

In line with previous results[25,43,145], we here demonstrate that PCNSL are defined by recurrent and often biallelic *CDKN2A* deletions, *MYD88* L265P mutations, and mutations that activate BCR signaling, genetic hallmarks of the DLBCL subtype MCD/C5[30,31]. Furthermore, we found high frequencies of SVs affecting the *IGH*, *IGL*, and *IGK* loci as well as losses of chromosome 6p affecting the *HLA* gene complex as a mechanism to escape recognition by cytotoxic T-cells[146]. *MYD88* L265P mutation and *CDKN2A* loss have been described as early mutational events in PCNSL[45] and we confirmed both to be major drivers in PCNSL. While *TP53* alterations seem to play a minor role in PCNSL, the *CDKN2A/B* genes encode several proteins that regulate either the p53 (p19 ARF) or the RB1 (p16 INK4a) pathway[147,148], underlining the relevance of the TP53 pathway in the context of PCNSL and cell cycle control.

The frequencies of *MYD88* mutations had varied between 38 and 94% in previous PCNSL studies[26,31,53,54,149], which might reflect a selection bias among small study populations, given the

rarity of PCNSL. This huge range could alternatively result also from an imprecise definition of PCNSL, which includes all malignant NHL within the brain, eyes, spinal cord, or leptomeninges without systemic involvement. In contrast, we here defined PCNSL as only intraparenchymal CNS-DLBCL and found a high prevalence of the *MYD88* L265P variant in this cohort (WGS cohort, extension FFPE cohort; mean: 73%). This is further supported by our robust classification of PCNSL by the RNA sequencing results, which demonstrated that the expression profiles of PCNSLs were distinct from PCNSL-M, SCNSL-M, and peripheral DLBCL without CNS manifestation, the latter three entities sharing similar profiles.

Moreover, SHM has previously been described as having a pathogenic role in PCNSL development and that its extent was greater there than in systemic DLBCL[80]. In agreement with previous reports, we identified several aSHM targets including the proto-oncogenes*PIM1, PAX5, BTG2*, and *OSBPL10*[25,53,80]. Exploiting a WGS approach, we observe additional mutational hotspots indicative of aSHM also in other genes including *MIR142, FHIT, ETV6, BTG1, GRHPR*, and *CD79B*. Our data suggest that katagis loci are reasonable indications of aSHM. We observed significantly higher RNA expression of genes with putative aSHM loci compared to those without. In addition, these putative aSHM loci were significantly enriched in genes involved in BCR signaling. Together this implicates that BCR signaling genes are both upregulated and targeted by putative aSHM, raising the question of cause and effect—is aSHM upregulating these genes, or is the high expression levels of these genes priming them for aSHM? This becomes even more complex when considering that highly expressed genes should have lower mutational rates due to transcriptional coupled repair[150].

The landscape of CNAs and SVs revealed potentially clinically exploitable deletion of *TOX* as a predictor for anti-PD1 response[103], amplification of *MALT1*, whose inhibition has been shown to be selectively toxic for ABC-DLBCL[151], and potential enhancer-hijacking events involving *PIK3C3* and *EPHA4*, whose inhibition has shown therapeutic advantage in a number of cancer models[113–115,120].

While the genetic landscape of PCNSL was described in some detail before[16–18,20,24,25,45,53,55], studies investigating the global gene expression profile of PCNSL have been scarce so far. Therefore, we performed RNA sequencing of 37 CNSL samples and 2 normal brain controls. Global gene expression profiles demonstrates that PCNSL are indeed distinct and can be distinguished from systemic ABC-DLBCL. This was perfectly mirrored based on the expression repertoire of *IG* constant genes, implicating the role of B-cell maturation in classification of PCNSL and other lymphomas, as employed in leukemia and multiple myeloma[152,153].

PCNSL are highly proliferative[154]. *TERT* activation confers unlimited proliferation, and activating *TERT* promoter mutations are frequent in different types of human cancers[155]. Mutations at two hotspots positions ($-124G>A$ and $-146G>A$) are causal for enhanced *TERT* promoter activity. Bruno et al. have previously reported these *TERT* promoter mutations to be present in PCNSL located in the splenium[136]. Therefore, we investigated 69 CNSL (including 49 PCNSL), but could not identify any *TERT* promoter mutations suggesting that this mechanisms of *TERT* activation is likely not relevant in PCNSL. Yet, we observed significantly more *TERT* expression in PCNSL compared to non-CNS ABC-DLBCL and this was consistent when stratifying the cohort based on our RNAseq subgrouping. However, we were not able to identify increased telomere content in PCNSL (or MCD) compared to the other groups, suggesting a role of telomere maintenance to overcome telomere shortening, which is imposed by the high levels of proliferation. This concept was supported by a within-group correlation of *TERT* expression and normalized telomere content (Pearson's $R = 0.67$, $p = 0.0023$), implicating a role of *TERT* in overcoming telomere degradation in PCNSL. Supporting the high proliferation in these tumors, PCNSL showed significantly elevated presence of the mutational signatures SBS1, which correlates with DNA replication, as well as of ID1 and ID2 which are associated with slippage during DNA replication.

With our study, we have substantiated the genomic and transcriptomic alterations characterizing PCNSL. We show that PCNSL can be clearly distinguished from systemic DLBCL, having distinct expression profiles, *IG* expression, and translocation patterns, as well as specific combinations of genetic alterations that are characterized by genomic instability, BCR activation, and most importantly, oncogenic TLR and NFκB signaling, which should be in the focus of future drug development.

Methods

CNS lymphoma (CNSL) study cohort. All procedures performed in this study were in accordance with the ethical standards of the respective institutional research committees and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. The ethics committee of the Medical Faculty Heidelberg and the Charité ethics committee (Charitéplatz 1, 10117 Berlin, Germany, approval number: EA1/245/13) approved the study. Informed consent was obtained from all participants in the study. Fresh frozen and paraffin-embedded PCNSL and SCNSL tumor tissue and matching blood samples (germline control) were acquired from the Department of Neuropathology, Charité, Berlin (Germany), and the Department of Neurosurgery, Heidelberg (Germany) from chemotherapy-naïve patients. Age at diagnosis, tumor localization, peripheral manifestation (bone marrow biopsy result, CT/MRI scan), first line therapy, as well as overall survival (OS) in months were evaluated. Additional control samples from age-matched, postmortem, and non-neoplastic brain ($n = 2$) were analyzed. The diagnosis was confirmed by at least two experienced (neuro)pathologists. The morphologic characteristics were assessed by using the fresh-frozen (FF) as well as the formalin-fixed and paraffin-embedded (FFPE) tissue sections of the respective tumor specimen. The tumor cell content in the cryopreserved sample material was estimated to be at least 60% based on histomorphological evaluation. Immuno-phenotypic characterization was performed on FFPE tissue sections (Supplementary Fig. 1b) of each tumor biopsy using an immunohistochemical panel including antibodies directed against CD20, CD10, BCL6, CD3, Ki67, MUM1/IRF4, and EBV (LMP1). To further exclude an EBV association, all cases with unclear EBV immunohistochemistry ($n = 9$) were investigated by an EBV-specific PCR as previously described[156] (Supplementary Fig. 1c, see "Methods"). For classifying GCB or non-GCB types, all samples were stratified according to the Hans classification[12] (CD10, BCL6, MUM1). We enrolled CNSL from a total of 51 patients for whole-genome (WGS, $n = 38$) and RNA sequencing (RNAseq, $n = 37$), including $n = 24$ samples subjected to both workflows. The study cohort and sample size as well as the experimental design, analysis workflow, diagnosis, and quality metrics of WGS and RNAseq are displayed in Fig. 1 and Supplementary Data 1. We included DLBCL confined to the CNS as PCNSL according to the recent WHO classification of tumors of hematopoietic and lymphoid organs and one of the tumors of the central nervous system. DLBCL, which presented initially with systemic, non-CNS, or synchronous systemic and CNS involvement were included as SCNSL[1,2,11,61,62]. In our SCNSL cohort, three patients presented with initial lymph node manifestation, one patient with testicular involvement, and three patients with involvement of parotid gland, liver, or urinary tract, respectively.

ICGC MMML-Seq Consortium samples. For comparison, we used and reanalyzed an early release of meanwhile published whole-genome and RNA sequencing data obtained by the ICGC MMML-Seq Consortium from systemic diffuse large B-cell lymphoma (DLBCL, total: $n = 36$, WGS: $n = 29$, RNAseq: $n = 36$, both workflows: $n = 29$), follicular lymphoma (FL, total: $n = 39$, WGS: $n = 39$, RNAseq: $n = 38$, both workflows: $n = 38$), and one "double hit" (DH)-lymphoma with a molecular BL signature[34]. In addition, we included WGS and RNAseq data from a single EBV-PCNSL case as well as RNAseq data from two nodal marginal zone lymphomas (nMZL) as well as naïve ($n = 5$) and GC B-cells ($n = 5$) as normal controls[157]. These data were obtained by the ICGC MMML-Seq consortium in accordance to protocols previously published[59,60].

DNA and RNA isolation. DNA and RNA were obtained from fresh frozen CNSL tumor samples. RNA and genomic DNA were isolated from 15 to 30 10 μm cryosection slices (depending on the tissue size). DNA from tumor samples and their matched blood controls was isolated according to standard procedures. Total RNA from tumor samples was extracted using the RNeasy® Plus Mini Kit (Qiagen, Hilden, Germany) according to the manufacturer's instructions. The RNA integrity

number (RIN) was determined using an Agilent 4200 TapeStation system (Agilent Technologies, Santa Clara, CA).

Whole-genome sequencing and data processing

Whole-genome sequencing. The DNA libraries of the tumor and matched control samples were prepared according to the Illumina TruSeq Nano DNA Library protocol using the TruSeq Nano DNA library Preparation Kit (Illumina, Hayward, CA; estimated insert size of 350 bp). Paired-end sequencing was performed on Illumina HiSeq X (2 × 150 bp) instruments using the TruSeq SBS kit, Version 3.

Alignment of sequencing reads. Sequencing reads were aligned using the DKFZ alignment workflow from ICGC Pan-Cancer Analysis of Whole Genome projects (DKFZ AlignmentAndQCWorkflows v1.2.73, https://github.com/DKFZ-ODCF/AlignmentAndQCWorkflows). Briefly, read pairs were mapped to the human reference genome (build 37, version hs37d5) using bwa mem (version 0.7.8) with minimum base quality threshold set to zero [-T 0] and remaining settings left at default values[158], followed by coordinate sorting with biobambam bamsort (version 0.0.148) with compression option set to fast (1) and marking duplicate read pairs with biobambam bammarkduplicates with compression option set to best (9)[159]. To allow the required and meaningful comparability to previous whole genome sequencing studies in lymphomas[34,95,160,161], the human reference genome version GRCh37/hg19 was used.

Small mutation calling and annotation. Somatic small variants (SNVs and indels) in matched tumor normal pairs were called using the DKFZ in-house pipelines (SNVCallingWorkflow v1.2.166-1, https://github.com/DKFZ-ODCF/SNVCallingWorkflow; IndelCallingWorkflow v1.2.177, https://github.com/DKFZ-ODCF/IndelCallingWorkflow) as previously described[162]. Briefly, the SNVs were identified using samtools and bcftools version 0.1.1957[163] and then classified as somatic or germline by comparing the tumor sample to the control, and later assigned a confidence which is initially set to 10, and subsequently reduced based on overlaps with repeats, DAC blacklisted regions, DUKE excluded regions, self-chain regions, segmental duplication records as introduced by the ENCODE project[164] and additionally if the SNV exhibited PCR or sequencing strand bias. Only SNVs with confidence 8 or above were considered for further analysis. Tumor and matched blood samples were analyzed by Platypus[165] to identify indels. Indel calls were filtered based on Platypus internal confidence calls, and only indels with confidence 8 or greater were used for subsequent analysis. In order to remove recurrent artifacts and misclassified germline events, somatic indels that were identified as germline in at least two patients in the CNS lymphoma cohort were excluded.

The protein coding effect of somatic SNVs and indels from all samples were annotated using ANNOVAR[166] according to GENCODE gene annotation (version 19) and overlapped with variants from dbSNP10 (build 141) and the 1000 Genomes Project database. Mutations of interest were defined as somatic SNV and indels that were predicted to cause protein coding changes (non-synonymous SNVs, gain or loss of stop codons, splice site mutations, and both frameshift and non-frameshift indels), and also synonymous exonic mutations on non-coding genes.

Tumor in normal contamination detection. We applied the TiNDA (tumor in normal detection algorithm) workflow to account for potential tumor in normal contamination leading to false negative calls as previously described[162]. The TiNDA algorithm is implemented in the DKFZ indel calling workflow (v1.2.177, https://github.com/DKFZ-ODCF/IndelCallingWorkflow). Briefly, the B-allele frequency (BAF) was calculated from the tumor and control samples. Positions overlapping with common variants were filtered out. Then, the clustering algorithm from Canopy[167] was applied to the BAF values for the positions in tumor vs control using a single pass run, assuming 9 clusters. The clusters that were determined to be tumor-in-normal had to have 75% of positions above the identity line (where the VAF in the tumor sample is the same as the VAF in the control sample). These identified mutations were then reclassified as somatic instead of the original germline annotation. All but 4 CNSL WGS samples exhibited evidence for tumor in normal. On average 31 SNVs (range 0–136) were "rescued" in PCNSL, 6 in PCNSL-M (6-6, single sample), 27 (19–34) in SCNSL, 22 (9–43) in SCNSL-M, and 0 in the EBV-positive sub-cohorts (0-0, 2 samples). In total, only 6 SNVs with protein coding effects were rescued, including the MYD88 p.L265P mutation in sample LS-0102, which had 3 of 47 read support in the control, and 86 of 170 reads supporting the variant in the tumor sample (Supplementary Data 13). In our series, we only found a very low level of tumor in normal. The rescued mutations followed a genomic distribution similar to the overall mutational landscape of PCNSL. We observed that rescued mutations were 1% exonic (compared to 1% in the mutational landscape), 32% intronic (c.f. 33%), 53% intergenic (c.f. 55%), and 13% on ncRNA (c.f. 12%).

Genomic structural rearrangements. Genomic structural rearrangements (SVs) were detected using SOPHIA v.34.0[168] as implemented in the DKFZ structural variation calling workflow (SophiaWorkflow:1.2.16, https://github.com/DKFZ-ODCF/SophiaWorkflow). Briefly, SOPHIA uses supplementary alignments as produced by bwa-mem as indicators of a possible underlying SV. SV candidates are filtered by comparing them to a background control set of sequencing data obtained using normal blood samples from a background population database of 3261 WGS samples from patients from published ICGC-PedBrain, ICGC-MMMLseq and ICGC-Prostate studies and DKFZ-HIPO studies, sequenced using Illumina HiSeq 2000, 2500 (100 bp) and HiSeq X (151 bp) platforms and aligned uniformly using the same workflow as in this study. All studies have been approved by appropriate ethics committees. Gencode V19 was used for the gene annotations. We used the script draw_fusions.R from the Arriba package[169] to visualize SVs generated by SOPHIA.

Copy number alterations and allelic imbalances. Allele-specific copy-number aberrations were detected using ACEseq (allele-specific copy-number estimation from WGS)[170] as implemented in the DKFZ CNV calling workflow (ACEseqWorkflow:1.2.8-4, https://github.com/DKFZ-ODCF/ACEseqWorkflow). ACEseq determines absolute allele-specific copy numbers as well as tumor ploidy and tumor cell content (TCC) based on coverage ratios of tumor and control as well as the B-allele frequency (BAF) of heterozygous SNPs. SVs called by SOPHIA were incorporated to improve genome segmentation.

Final copy number segments were further smoothed to calculate the total number of gains and losses. Neighboring segments were merged if they rounded to the same copy number and deviated by less than 0.5 copies in case of segments <20 kb or deviated by less than 0.3 copies otherwise. Remaining segments <500 kb were merged with their closer neighbor based on allele-specific and total copy number and once again segments smaller than 2 Mb deviating by less than 0.4 copies were merged. Based on the resulting segments the number of gains and losses was estimated.

Furthermore, the fraction of aberrant genome was calculated as the fraction of the genome that is classified either as duplication or deletion (>0.7 deviation from the ploidy) or was identified as a loss of heterozygosity.

Classification of mutational hotspots (kataegis events). Mutational hotspots indicating putative kataegis events (likely due to SHM or aberrant SHM (aSHM)) were defined as regions with at least 6 somatic SNVs within an average intermutational distance of 1000 bp or less, as previously used by Alexandrov and colleagues[90]. A gene was described to be targeted by kataegis if its definition (from Gencode version 19 gene models) overlapped with at least 1 kataegis region in at least 1 sample. While many of these kataegis loci are indeed SHM/aSHM targets, located 2.5 kb from the transcription start site (TSS), we cannot completely control for all PCNSL-specific TSSs due to the normal brain background tissue.

Mutational signatures. Supervised mutational signature analysis was performed using YAPSA development version 3.13[171] using R 4.0.0. Briefly, the linear combination decomposition (LCD) of the mutational catalog with known and pre-defined PCAWG COSMIC signatures[126] was computed by non-negative least squares (NNLS). The mutational signature analysis was applied to the mutational catalogs for SNVs (or single base substitutions, SBS) and indels of all tumor samples. Signature-specific cutoffs were applied and cohort level analysis was used for detecting signatures as recommended by Huebschmann et al.[34]. The cutoff used corresponds to "cost factors" of 10 for SNVs and 3 for indels in the modified ROC analysis.

Integration of different variant types. SNVs, indels, SVs, and CNAs were integrated in order to account for all variant types in the recurrence analysis. All genes with SNVs or indels in coding regions (nonsynonymous, stop gain, stop loss, splicing, frameshift, and non-frameshift events) and ncRNA (exonic) were included. Any SV with breakpoints directly lying on a gene (SV direct) were considered for oncoprints, however, SVs were also annotated to a gene when they were either within 100 kb of a gene (SV near), or to the closest gene (SV close) for SV recurrence analysis to account for regulatory mutations such as enhancer hijacking events. Genes were annotated with CNAs if they were completely or partially affected. Chromosome level CNVs events were determined once >30% of a chromosome arm was altered. Only focal CNA events were taken into account for variant integration, as these are more likely to target specific genes within the affected region than large events such as whole chromosome arm events. To capture the precise target of CNVs, we employed results from GISTIC. Finally, genes affected by SNVs, indels, directly hit by SVs, or genes with focal CNAs were considered for the recurrence analysis and added to the oncoprints. The mutations were integrated and plotted as oncoprint plots using using R v3.4.0 (library yapsa v3.13), perl v5.26.2 (libraries perl-getopt-long v2.50) and bedtools v2.16.2. SV cohort plots were generated using perl v5.20.0, bedtools v2.24.0, R v3.3.1 (libraries circlize v0.4.5 and dplyr v0.7.8), using the gencode v19 gene models for annotation.

Mutual exclusivity and inclusivity analysis. Mutual exclusivity analysis was performed to investigate the relationship between *MYD88* mutations with other implicated drivers from the IntOGen analysis including SNVs, indels, SVs, CNAs.

The minimal recurrence threshold was set to 5. We applied the commonly used Fisher's exact test and the CoMET test[172] (v0.1.5) for both co-occurrence and mutual exclusivity using R (v3.4.0). Fisher's right tailed test was used to support co-occurrence when the number of samples with alterations in both genes is significantly higher than expected by chance. Additionally, Fisher's left tailed test was used to suggest mutual exclusivity when the number of samples with alterations in both genes is significantly lower than expected. Resultant p-values were corrected for multiple testing by FDR.

Mutational significance analysis. The IntOGen pipeline[173] algorithm was applied to identify significant cancer drivers in the core set of PCNSL samples (n = 30) based on the hg19 genome assembly. IntOGen v 3.0.8 was installed via conda from the bbglab anaconda channel. The relevant conda environment setup included explicit definitions of python v3.5.0 (with libraries scipy v0.16.0, pycurl v7.43.0.3, numpy v1.10.0, pandas 0.17.0). In addition, a local installation of perl v5.20.3.1 (with libraries perl-digest-perl-md5 v1.9, perl-threaded v5.26.0) was used, with installation of perl libraries Digest-MD5 v2.52 via cpan and perl-DBI v1.626 via yum package managers. The background intogen database (bgdata) was automatically downloaded using the command 'intogen -setup' which downloaded the 20150729 background databases. The IntOGen run specific parameters included running on 4 cores, Matlab Compiler Runtime v8.1 (2013a) and MutSigCV v1.4. Significance thresholds of 10% FDR were used for oncodrivefm, oncodriveclust and mutsig. Sample thresholds of 2 and 5 were used for oncodrivefm and oncodriveclust respectively. IntOGen reported 50 genes to be significant drivers.

Significant CNV were identified using GISTIC v2.0.23 using MCR v83, using the following parameters: "-broad 1 –genegistic 1 -savegene=1 -brlen=0.8 -conf=0.2 -maxseg=2500".

Telomere content estimation. The telomere content was determined from WGS data using the software tool TelomereHunter[59] (v1.1.0) which uses python v3.5.6 (using libraries pyyaml v3.13, pysam v0.9.1, pynacl v1.2.1), samtools v1.3.1, bcftools v1.3.1 and htslib v1.3.2. Telomere hunter was run using default settings (filtering of telomere reads: at least 6 telomere repeats per 100 bp read length)[133]. Briefly, unmapped reads or reads with a very low alignment confidence (mapping quality lower than 8) containing six non-consecutive instances of the four most common telomeric repeat types (TTAGGG, TCAGGG, TGAGGG, and TTGGGG) were extracted. The telomere content was determined by normalizing the telomere read count to all reads in the sample with a GC-content of 48–52%. In the case of tumor samples, the telomere content was further corrected for the tumor cell content (TCC, as estimated by ACEseq) using the following formula as previously described[59], which corrects for inter-patient differences in telomere content assuming that the non-malignant cells in the tumor sample have a similar telomere content as in the control sample, as shown in Eq. 1:

$$T_{TCCcorrected} = (T - C(1 - TCC))/TCC \qquad (1)$$

Here, T and C are the telomere contents of the tumor and control sample, and $T_{TCCcorrected}$ is the TCC-corrected telomere content of the tumor sample.

LymphGen classification. All WGS samples were classified according to the LymphGen v2.0 algorithm described by Wright et al.[31] which categorizes DLBCL samples into the different genetic subtypes MCD, N1, A53, BN2 ST2, EZB (MYC + and MYC−), based on genetic aberrations in subtype predictor genes. The algorithm requires information on mutations, copy-number alterations, and fusions. The results of the DKFZ SNV and indels calling workflows were used to define the small mutations. BCL2 and BCL6 translocations were determined by Sophia calls, and copy number changes were derived from the DKFZ CNV workflow (ACEseq) results. The outputs from all workflows were filtered for somatic regions with all different variations occurring in exons and 5′UTR region of the gene. The files were created using Python and Perl scripts based on the description provided on the LymphGen website [https://llmpp.nih.gov/lymphgen/LymphGenInstructions.pdf?v=1600863825]. The individual sample inputs are further merged together to form the input dataset for the LymphGen algorithm and uploaded to the website [https://llmpp.nih.gov/lymphgen/lymphgendataportal.php] for classification the samples.

According to Wright et al.[31], the results are displayed in Supplementary Data 1. Samples, where only RNA was available, were listed as "NA" in Supplementary Data 1.

RNA sequencing and data processing
RNA library preparation and sequencing. RNA libraries of the tumor samples and normal brain samples were prepared using the TruSeq RNA library preparation Kit Set A and B, following the manufacturer's instructions at an insert size of ~300 bp. Two barcoded libraries were pooled per lane and sequenced on Illumina HiSeq2000 or HiSeq4000 platforms.

RNAseq alignment and expression quantification. RNAseq reads were aligned and gene expression quantified using the DKFZ RNAseq workflow (v1.2.22-6, https://github.com/DKFZ-ODCF/RNAseqWorkflow) as previously described[174]. Briefly, the RNAseq read pairs were aligned to the STAR index generated reference genome (build 37, version hs37d5) using STAR in 2 pass mode (version

2.5.2b)[174,175]. Duplicate reads were marked using sambamba (version 0.6.5) and BAM files were coordinate sorted using SAMtools (version 0.1.19). featureCounts (version 1.5.1)[176] was used to perform non-strand-specific read counting for genes over exon features based on the Gencode V19 gene model (without excluding read duplicates). When both read pairs aligned uniquely (indicated by a STAR alignment quality score of 255) they were used towards gene read counts. For total library abundance calculations, during TPM and FPKM expression values estimation, genes on chromosomes X, Y, MT, and rRNA and tRNA were omitted as they can introduce library size estimation biases.

Hierarchical consensus clustering was applied using the cola package (version 1.5.6) with "MAD" as top-value method and "kmeans" as partitioning method. Classification on CNS samples was applied using cola with "ATC" as top-value method and "skmeans" as partitioning method. All other parameters took default values[129].

RNA dilution experiment. To further investigate the impact of brain tissue contamination in unsupervised clustering analysis of gene expression data on PCNSL, we performed a serial dilution experiment with total RNA from a PCNSL sample considered "pure" (LS-027, estimated tumor cell content >80%) and a normal brain tissue control (CTRL). Total RNA from LS-027 was mixed with CTRL RNA with increasing concentrations (0, 20, 40, 60, and 80%) and sequenced. The z-score transformed TPM expression levels for PCNSL subcluster 1 and subcluster 2 signature genes for the serially diluted H050-0027 sample was compared against the cohort and individually using clustering analysis.

Differential expression (DE) analysis to identify signature genes. DE of genes was analyzed using DESeq2 (version 1.14.1) with default settings using raw read counts from featureCounts. Genes without any count in all samples were excluded from the analysis.

Validation of whole-genome and RNA sequencing results
Sanger sequencing. Bidirectional Sanger sequencing (bSS) was performed (i) to validate WGS results in the PCNSL/SCNSL study cohort if sufficient DNA quantity was available (total: n = 35; PCNSL: n = 26; PCNSL-M: n = 1; SCNSL: n = 2; SCNSL: n = 3; EBV+ : n = 2), and (ii) to identify mutations of recurrently mutated candidates in a larger set of additional PCNSL/SCNSL FFPE samples (FFPE extension cohort). The following genes were analyzed: MYD88, KMT2D (MLL2), HLA-B, SETD1B, HIST1H1E, CD79B, BTG1, MYC, TP53, TERT, GRHPR, TBL1XR1, DST, PRDM15, OBSCN, FAT4, GRP98, and OSBPL10. Briefly, the PCR conditions were: 94 °C for 4 min (1 cycle), followed by 3 cycles of 94 °C for 30 s, 61 °C for 45 s, 72 °C for 60 s, 3 cycles of 94 °C for 30 s, 59 °C for 45 s, 72 °C for 60 s, 3 cycles of 94 °C for 30 s, 57 °C for 45 s, 72 °C for 60 s, 31 cycles of 94 °C for 30 s, 55 °C for 45 s, 72 °C for 60 s, and finally extension at 72 °C for 10 min with AmpliTaq™ 360 DNA Polymerase (Applied Biosystems, Waltham, USA). The PCR primers for the genomic regions of interest are displayed in Supplementary Data 14. Sequencing was performed at Eurofins Genomics, Ebersberg, Germany. 180/189 (95%) of the selected variants (allele frequency above 10%) identified by WGS were confirmed. The results are displayed in Supplementary Data 4 and 5.

Formaldehyde-fixed paraffin-embedded (FFPE) CNSL extension cohort. Candidate genes were validated in 31 additional FFPE specimens of PCNSL (n = 19), SCNSL (n = 9), and n = 3 EBV positive cases. The DNA was extracted using the QIAamp DNA FFPE Tissue Kit (Qiagen, Hilden, Germany) according to the manufacturer's instructions. The following, recurrently mutated genes (exons) were investigated: PIM1 (exons 1–4), MYD88 (exon 5), GPHPR (exons 2, 4, and 5), TBL1XR1 (exons 7, 8, 10, 12, and 14), KMT2D (exons 32, 34, 38, 48, and 50), KLHL14 (exon 2), HLA-B (exons 2, 3), PRDM15 (exons 9–12), GPR98 (exons 65, 70, and 81), DST (exons 13, 14, 23, 24, and 36), OBSCN (exons 32, 63, 64, and 85), FAT4 (exons 1, 9, and 17), HIST1H3D (exon 2), HIST1H1E (exon 1), TERT (promoter region). The results are displayed in Supplementary Data 5.

Fluorescence in situ hybridization (FISH). FISH analysis was performed as previously described[177]. Briefly, 4 μm FFPE sections were deparaffinized, dehydrated, and incubated in pre-treatment solution (Dako, Denmark) for 10 min at 95–99 °C. Samples were treated with pepsin solution for 3–6 min at 37 °C, washed, dehydrated, air dried, and incubated with the respective DNA probe: CDKN2A (9p21.3): Orange, Biocare Medical, USA; Vysis CEP 9 SpectrumGreen Probe (Abbott), The Netherlands). The sections were sealed, denatured in humidified atmosphere at 82 °C for 5 min, and then incubated overnight at 45 °C to achieve hybridization. After post-hybridization washing, slides were counterstained with 4′6-diamidino-2-phenylindole (DAPI) and analyzed using an automated scanning system (Duet, BioView Ltd. Rehovot, Israel; Supplementary Fig. 1d).

Real-time quantitative PCR (RT-qPCR). We performed SYBR Green quantitative real-time PCR (qPCR) measuring six amplicons covering the CDKN2A/B gene (Supplementary Fig. 1e) as well as five amplicons covering the FHIT, NOTCH4, SPIB, and MIR650 gene. The primer sequences are annotated in Supplementary

Data 14. qPCR analysis was performed on ABI Prism 7900HT Sequence Detection System (Applied Biosystems, Foster City, CA, USA).

Immunohistochemical (IHC) procedures. Immunohistochemical stainings were performed on a Benchmark XT autostainer (Ventana Medical Systems, Tuscon, AZ, USA) with standard antigen retrieval methods (CC1 buffer, pH8.0, Ventana Medical Systems, Tuscon, AZ, USA) using 7 µm-thick frozen or 4 µm-thick FFPE tissue sections. The following primary antibodies were used: monoclonal mouse anti-BCL6 (DAKO, M7211, 1:10), monoclonal mouse anti-CD10 (Novocastra, NCL-CD10-270, 1:10), monoclonal mouse anti-CD20 (DAKO, M0755, 1:400), polyclonal rabbit anti-CD3 (DAKO, A0452, 1:100), monoclonal mouse anti-CD45 (DAKO, M0701, 1:400), monoclonal mouse anti-CD79a (DAKO, M7051, 1:100), monoclonal mouse anti-EBV-LMP1 (DAKO, M0897, 1:1000), monoclonal mouse anti-Ki-67 (DAKO, M7240, 1:100), monoclonal mouse anti-MUM1 (DAKO, M7259, 1:50), monoclonal mouse anti-PD-L1 (Cell Signaling, 13684, 1:200). The iVIEW DAB Detection Kit (Ventana Medical Systems, Tuscon, AZ, USA) was used according to the manufacturer's instructions. Sections were counterstained with hematoxylin, dehydrated in graded alcohol and xylene, mounted, and coverslipped. IHC stained sections were evaluated by two skilled neuropathologists with concurrence. The DLBCL subtypes of GCB and non-GCB were categorized using CD10, BCL6, and MUM1 according to the Hans classification[12].

Epstein-Barr virus PCR. EBV-specific PCR was performed as previously described[156]. Briefly, a highly conserved region of the EBNA-1 (BKRF1) gene specific for EBV was amplified by endpoint PCR using the following primers: 5′-GAG GGT GGT TTG GAA AGC-3′ and 5′-AAC AGA CAA TGG ACT CCC TTA G-3′, 0.1 µM each. The PCR conditions were: 95 °C for 5 min (1 cycle), followed by 40 cycles of 94 °C for 1 min, 55 °C for 2 min, 72 °C for 3 min, and finally extension at 72 °C for 7 min with ThermoPrime™ Taq DNA Polymerase (Thermo Fisher Scientific, Waltham, USA). Subsequently, amplification products were analyzed by ELISA (Roche, Basel, Switzerland).

Statistics and reproducibility. No statistical methods were used to predetermine sample sizes. We included all individuals with DLBCL of the CNS where sufficient material was available as specified in the description of study design. No data were excluded from the analyses. Statistical details for each analysis are mentioned in each figure legend or in the respective part of the text. WGS, RNA-sequencing, Sanger sequencing, quantitative real-time PCR, immunohistochemical stainings, and FISH were performed in a blinded fashion. Evaluation of histological and immunohistochemical stainings, as well as FISH images, was performed separately by at least two independent (neuro-)pathologists in Berlin and Heidelberg. Histological staining, immunohistochemistry, and FISH analyses were replicated at least once. CDKN2A/B FISH was performed exemplarily for n = 4 CNSL patients. The representative images shown were adjusted in brightness and contrast to different degrees (depending on the need resulting from the range of brightness and contrast of the raw images) in Adobe Photoshop, and for these cases, raw image files are publicly available [https://doi.org/10.5281/zenodo.6054242][92]. The experiments were not randomized.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The whole genome sequencing (WGS) and RNA sequencing data of the 51 CNSL samples generated in this study as well as the raw data of Sanger sequencing and quantitative real-time PCR data have been deposited in the European Genome-Phenome Archive under the accession number EGAS00001005339. The data are available under controlled access due to the sensitive nature of genome sequencing data, and access can be obtained by contacting the appropriate Data Access Committee listed for each dataset in the study. Access will be granted to commercial and non-commercial parties according to patient consent forms and data transfer agreements. We have an institutional process in place to deal with requests for data transfer. A response to requests for data access can be expected within 14 days. After access has been granted, the data is available for two years. Access to the ICGC MMML-Seq raw sequencing data is available via the EGA under the accession number EGAS00001002199 and EGAS00001001692. Access to the ICGC MMML-Seq data is available via the data access committee of the ICGC (www.ICGC.org). Raw image files of histological stainings, immunohistochemistry, and FISH images generated in this study, as well as all somatic mutation calls, integrated mutations tables, and RNAseq counts on which the analysis was performed have been deposited publicly at Zenodo [https://doi.org/10.5281/zenodo.6054242][92]. The uncropped PCR gel images as well as the processed real-time PCR data and Kaplan-Meier survival data shown in Supplementary Fig. 1 are provided in the Source Data file with this paper. The remaining data are available within the article, Supplementary Information or Source Data file.

References

1. Louis, D. N., Ohgaki, H., Wiestler, O. D. & Cavenee, W. K. *WHO Classification of Tumours of the Central Nervous System* 4 edn, 1 (International Agency for Research on Cancer Lyon, 2016).
2. Swerdlow, S. H. et al. *WHO Classification of Tumours of Haematopoietic and Lymphoid Tissues* 2 (International Agency for Research on Cancer Lyon, 2008).
3. Maciocia, P. et al. Treatment of diffuse large B-cell lymphoma with secondary central nervous system involvement: encouraging efficacy using CNS-penetrating R-IDARAM chemotherapy. *Br. J. Haematol.* **172**, 545–553 (2016).
4. Ferreri, A. J. M. Secondary CNS lymphoma: The poisoned needle in the haystack. *Ann. Oncol.* **28**, 2335–2337 (2017).
5. Ferreri, A. J. Risk of CNS dissemination in extranodal lymphomas. *Lancet Oncol.* **15**, e159–e169 (2014).
6. Malikova, H. et al. Secondary central nervous system lymphoma: Spectrum of morphological MRI appearances. *Neuropsychiatr. Dis. Treat.* **14**, 733–740 (2018).
7. Bashir, R., McManus, B., Cunningham, C., Weisenburger, D. & Hochberg, F. Detection of Eber-1 RNA in primary brain lymphomas in immunocompetent and immunocompromised patients. *J. Neurooncol.* **20**, 47–53 (1994).
8. Kleinschmidt-DeMasters, B. K., Damek, D. M., Lillehei, K. O., Dogan, A. & Giannini, C. Epstein Barr virus-associated primary CNS lymphomas in elderly patients on immunosuppressive medications. *J. Neuropathol. Exp. Neurol.* **67**, 1103–1111 (2008).
9. Montesinos-Rongen, M., Siebert, R. & Deckert, M. Primary lymphoma of the central nervous system: Just DLBCL or not? *Blood* **113**, 7–10 (2009).
10. Grommes, C. & DeAngelis, L. M. Primary CNS lymphoma. *J. Clin. Oncol.: Off. J. Am. Soc. Clin. Oncol.* **35**, 2410–2418 (2017).
11. Louis, D. N. et al. The 2016 World Health Organization Classification of Tumors of the Central Nervous System: A summary. *Acta Neuropathol.* **131**, 803–820 (2016).
12. Hans, C. P. et al. Confirmation of the molecular classification of diffuse large B-cell lymphoma by immunohistochemistry using a tissue microarray. *Blood* **103**, 275–282 (2004).
13. Klein, U. et al. Transcriptional analysis of the B cell germinal center reaction. *Proc. Natl Acad. Sci. USA* **100**, 2639–2644 (2003).
14. Klein, U. et al. Transcription factor IRF4 controls plasma cell differentiation and class-switch recombination. *Nat. Immunol.* **7**, 773–782 (2006).
15. Montesinos-Rongen, M. et al. Gene expression profiling suggests primary central nervous system lymphomas to be derived from a late germinal center B cell. *Leukemia* **22**, 400–405 (2008).
16. Deckert, M., Montesinos-Rongen, M., Brunn, A. & Siebert, R. Systems biology of primary CNS lymphoma: from genetic aberrations to modeling in mice. *Acta Neuropathol.* **127**, 175–188 (2014).
17. Montesinos-Rongen, M. et al. Activating L265P mutations of the MYD88 gene are common in primary central nervous system lymphoma. *Acta Neuropathol.* **122**, 791–792 (2011).
18. Montesinos-Rongen, M. et al. Mutations of CARD11 but not TNFAIP3 may activate the NF-kappaB pathway in primary CNS lymphoma. *Acta Neuropathol.* **120**, 529–535 (2010).
19. Bodor, C. et al. Molecular subtypes and genomic profile of primary central nervous system lymphoma. *J. Neuropathol. Exp. Neurol.* **79**, 176–183 (2020).
20. Montesinos-Rongen, M., Schafer, E., Siebert, R. & Deckert, M. Genes regulating the B cell receptor pathway are recurrently mutated in primary central nervous system lymphoma. *Acta Neuropathol.* **124**, 905–906 (2012).
21. Jordanova, E. S. et al. Hemizygous deletions in the HLA region account for loss of heterozygosity in the majority of diffuse large B-cell lymphomas of the testis and the central nervous system. *Genes Chromosomes Cancer* **35**, 38–48 (2002).
22. Riemersma, S. A. et al. Extensive genetic alterations of the HLA region, including homozygous deletions of HLA class II genes in B-cell lymphomas arising in immune-privileged sites. *Blood* **96**, 3569–3577 (2000).
23. Montesinos-Rongen, M. et al. Interphase cytogenetic analysis of lymphoma-associated chromosomal breakpoints in primary diffuse large B-cell lymphomas of the central nervous system. *J. Neuropathol. Exp. Neurol.* **61**, 926–933 (2002).
24. Schwindt, H. et al. Chromosomal translocations fusing the BCL6 gene to different partner loci are recurrent in primary central nervous system lymphoma and may be associated with aberrant somatic hypermutation or defective class switch recombination. *J. Neuropathol. Exp. Neurol.* **65**, 776–782 (2006).

25. Chapuy, B. et al. Targetable genetic features of primary testicular and primary central nervous system lymphomas. *Blood* **127**, 869–881 (2016).

26. Gonzalez-Aguilar, A. et al. Recurrent mutations of MYD88 and TBL1XR1 in primary central nervous system lymphomas. *Clin. Cancer Res.: Off. J. Am. Assoc. Cancer Res.* **18**, 5203–5211 (2012).

27. Kreher, S. et al. Prognostic impact of B-cell lymphoma 6 in primary CNS lymphoma. *Neuro-Oncol.* **17**, 1016–1021 (2015).

28. Riemersma, S. A. et al. High numbers of tumour-infiltrating activated cytotoxic T lymphocytes, and frequent loss of HLA class I and II expression, are features of aggressive B cell lymphomas of the brain and testis. *J. Pathol.* **206**, 328–336 (2005).

29. Chapuy, B. et al. Molecular subtypes of diffuse large B cell lymphoma are associated with distinct pathogenic mechanisms and outcomes. *Nat. Med.* **24**, 679–690 (2018).

30. Schmitz, R. et al. Genetics and pathogenesis of diffuse large B-cell lymphoma. *N. Engl. J. Med.* **378**, 1396–1407 (2018).

31. Wright, G. W. et al. A probabilistic classification tool for genetic subtypes of diffuse large B cell lymphoma with therapeutic implications. *Cancer Cell* **37**, 551–568.e514 (2020).

32. Venturutti, L. & Melnick, A. The dangers of deja vu: Memory B-cells as the cell-of-origin of ABC-DLBCLs. *Blood* https://doi.org/10.1182/blood.2020005857 (2020).

33. Venturutti, L. et al. TBL1XR1 mutations drive extranodal lymphoma by inducing a pro-tumorigenic memory fate. *Cell* **182**, 297–316.e227 (2020).

34. Hubschmann, D. et al. Mutational mechanisms shaping the coding and noncoding genome of germinal center derived B-cell lymphomas. *Leukemia* https://doi.org/10.1038/s41375-021-01251-z (2021).

35. Lacy, S. E. et al. Targeted sequencing in DLBCL, molecular subtypes, and outcomes: A Haematological Malignancy Research Network report. *Blood* **135**, 1759–1771 (2020).

36. Niparuck, P. et al. Treatment outcome and prognostic factors in PCNSL. *Diagn. Pathol.* **14**, 56 (2019).

37. Coiffier, B. et al. CHOP chemotherapy plus rituximab compared with CHOP alone in elderly patients with diffuse large-B-cell lymphoma. *N. Engl. J. Med.* **346**, 235–242 (2002).

38. Schmitt, A. M. et al. Rituximab in primary central nervous system lymphoma—A systematic review and meta-analysis. *Hematol. Oncol.* **37**, 548–557 (2019).

39. Seidel, S. & Schlegel, U. Have treatment protocols for primary CNS lymphoma advanced in the past 10 years. *Expert Rev. Anticancer Ther.* **19**, 909–915 (2019).

40. van der Meulen, M. et al. Primary therapy and survival in patients aged over 70-years-old with primary central nervous system lymphoma: A contemporary, nationwide, population-based study in the Netherlands. *Haematologica* **106**, 597–600 (2021).

41. Van Dijck, R., Doorduijn, J. K. & Bromberg, J. E. C. The role of rituximab in the treatment of primary central nervous system lymphoma. *Cancers* https://doi.org/10.3390/cancers13081920 (2021).

42. Bromberg, J. E. C. et al. Rituximab in patients with primary CNS lymphoma (HOVON 105/ALLG NHL 24): A randomised, open-label, phase 3 intergroup study. *Lancet Oncol.* **20**, 216–228 (2019).

43. Deckert, M. et al. Modern concepts in the biology, diagnosis, differential diagnosis and treatment of primary central nervous system lymphoma. *Leukemia* **25**, 1797–1807 (2011).

44. Braggio, E. et al. Genome-wide analysis uncovers novel recurrent alterations in primary central nervous system lymphomas. *Clin. Cancer Res.: Off. J. Am. Assoc. Cancer Res.* **21**, 3986–3994 (2015).

45. Nayyar, N. et al. MYD88 L265P mutation and CDKN2A loss are early mutational events in primary central nervous system diffuse large B-cell lymphomas. *Blood Adv.* **3**, 375–383 (2019).

46. Grommes, C., Nayak, L., Tun, H. W. & Batchelor, T. T. Introduction of novel agents in the treatment of primary CNS lymphoma. *Neuro-Oncol.* **21**, 306–313, https://doi.org/10.1093/neuonc/noy193 (2019).

47. Ghesquieres, H. et al. Lenalidomide in combination with intravenous rituximab (REVRI) in relapsed/refractory primary CNS lymphoma or primary intraocular lymphoma: a multicenter prospective 'proof of concept' phase II study of the French Oculo-Cerebral lymphoma (LOC) Network and the Lymphoma Study Association (LYSA)dagger. *Ann. Oncol.* **30**, 621–628 (2019).

48. Soussain, C. et al. Ibrutinib monotherapy for relapse or refractory primary CNS lymphoma and primary vitreoretinal lymphoma: Final analysis of the phase II 'proof-of-concept' iLOC study by the Lymphoma study association (LYSA) and the French oculo-cerebral lymphoma (LOC) network. *Eur. J. Cancer* **117**, 121–130 (2019).

49. Kotla, V. et al. Mechanism of action of lenalidomide in hematological malignancies. *J. Hematol. Oncol.* **2**, 36 (2009).

50. Lionakis, M. S. et al. Inhibition of B cell receptor signaling by Ibrutinib in primary CNS lymphoma. *Cancer Cell* **31**, 833–843 e835 (2017).

51. Grommes, C. et al. Ibrutinib unmasks critical role of Bruton tyrosine kinase in primary CNS lymphoma. *Cancer Discov.* **7**, 1018–1029 (2017).

52. Nayak, L. et al. PD-1 blockade with nivolumab in relapsed/refractory primary central nervous system and testicular lymphoma. *Blood* **129**, 3071–3073 (2017).

53. Vater, I. et al. The mutational pattern of primary lymphoma of the central nervous system determined by whole-exome sequencing. *Leukemia* **29**, 677–685 (2015).

54. Bruno, A. et al. Mutational analysis of primary central nervous system lymphoma. *Oncotarget* **5**, 5065–5075 (2014).

55. Zhou, Y. et al. Analysis of genomic alteration in primary central nervous system lymphoma and the expression of some related genes. *Neoplasia* **20**, 1059–1069 (2018).

56. Takashima, Y. et al. Target amplicon exome-sequencing identifies promising diagnosis and prognostic markers involved in RTK-RAS and PI3K-AKT signaling as central oncopathways in primary central nervous system lymphoma. *Oncotarget* **9**, 27471–27486 (2018).

57. Kaulen, L. D. et al. Whole exome sequencing identifies novel SLIT2 mutations in primary CNS lymphoma (3962). *Neurology* **94**, 3962 (2020).

58. Gandhi, M. K. et al. EBV-tissue positive primary CNS lymphoma occurring after immunosuppression is a distinct immunobiological entity. *Blood* https://doi.org/10.1182/blood.2020008520 (2020).

59. Lopez, C. et al. Genomic and transcriptomic changes complement each other in the pathogenesis of sporadic Burkitt lymphoma. *Nat. Commun.* **10**, 1459 (2019).

60. Consortium, I. T. P.-C. Ao. W. G. Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020).

61. Arber, D. A. et al. The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia. *Blood* **127**, 2391–2405 (2016).

62. Leonard, J. P., Martin, P. & Roboz, G. J. Practical implications of the 2016 revision of the World Health Organization classification of lymphoid and myeloid neoplasms and acute leukemia. *J. Clin. Oncol.: Off. J. Am. Soc. Clin. Oncol.* **35**, 2708–2715 (2017).

63. Zhang, J. et al. The CREBBP acetyltransferase is a haploinsufficient tumor suppressor in B-cell lymphoma. *Cancer Discov.* **7**, 322–337 (2017).

64. Pasqualucci, L. et al. Inactivating mutations of acetyltransferase genes in B-cell lymphoma. *Nature* **471**, 189–195 (2011).

65. Meyer, S. N. et al. Unique and shared epigenetic programs of the CREBBP and EP300 acetyltransferases in germinal center B cells reveal targetable dependencies in lymphoma. *Immunity* **51**, 535–547.e539 (2019).

66. Schmidt, J. et al. CREBBP gene mutations are frequently detected in in situ follicular neoplasia. *Blood* **132**, 2687–2690 (2018).

67. Loeffler, M. et al. Genomic and epigenomic co-evolution in follicular lymphomas. *Leukemia* **29**, 456–463 (2015).

68. Zhou, Y. et al. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat. Commun.* **10**, 1523 (2019).

69. Yang, X. et al. STAT3 activation is associated with interleukin-10 expression and survival in primary central nervous system lymphoma. *World Neurosurg.* **134**, e1077–e1084 (2020).

70. Tang, D. et al. Clinicopathologic significance of MYD88 L265P mutation and expression of TLR4 and P-STAT3 in primary central nervous system diffuse large B-cell lymphomas. *Brain Tumor Pathol.* **38**, 50–58 (2021).

71. Bai, L. et al. A potent and selective small-molecule degrader of STAT3 achieves complete tumor regression in vivo. *Cancer Cell* **36**, 498–511 e417 (2019).

72. Komohara, Y. et al. M2 macrophage/microglial cells induce activation of Stat3 in primary central nervous system lymphoma. *J. Clin. Exp. Hematop.* **51**, 93–99 (2011).

73. Ngo, V. N. et al. Oncogenically active MYD88 mutations in human lymphoma. *Nature* **470**, 115–119 (2011).

74. Dobashi, A. et al. TP53 and OSBPL10 alterations in diffuse large B-cell lymphoma: Prognostic markers identified via exome analysis of cases with extreme prognosis. *Oncotarget* **9**, 19555–19568 (2018).

75. Gandhi, M. K. et al. EBV-associated primary CNS lymphoma occurring after immunosuppression is a distinct immunobiological entity. *Blood* **137**, 1468–1477 (2021).

76. Wang, J. Q. et al. Synergistic cooperation and crosstalk between MYD88(L265P) and mutations that dysregulate CD79B and surface IgM. *J. Exp. Med.* **214**, 2759–2776 (2017).

77. Visco, C. et al. Oncogenic mutations of MYD88 and CD79B in diffuse large B-cell lymphoma and implications for clinical practice. *Cancers* https://doi.org/10.3390/cancers12102913 (2020).

78. Reddy, A. et al. Genetic and functional drivers of diffuse large B. *Cell Lymphoma Cell* **171**, 481–494.e415 (2017).

79. Nosrati, A. et al. MYC, BCL2, and BCL6 rearrangements in primary central nervous system lymphoma of large B cell type. *Ann. Hematol.* **98**, 169–173 (2019).

80. Montesinos-Rongen, M., Van Roost, D., Schaller, C., Wiestler, O. D. & Deckert, M. Primary diffuse large B-cell lymphomas of the central nervous

system are targeted by aberrant somatic hypermutation. *Blood* **103**, 1869–1875 (2004).

81. Brunn, A. et al. Frequent triple-hit expression of MYC, BCL2, and BCL6 in primary lymphoma of the central nervous system and absence of a favorable MYC(low)BCL2 (low) subgroup may underlie the inferior prognosis as compared to systemic diffuse large B cell lymphomas. *Acta Neuropathol.* **126**, 603–605 (2013).

82. Amodio, N. et al. MALAT1: A druggable long non-coding RNA for targeted anti-cancer approaches. *J. Hematol. Oncol.* **11**, 63 (2018).

83. Li, S., Li, J., Chen, C., Zhang, R. & Wang, K. Pan-cancer analysis of long non-coding RNA NEAT1 in various cancers. *Genes Dis.* **5**, 27–35 (2018).

84. Fujimoto, A. et al. Whole-genome mutational landscape and characterization of noncoding and structural mutations in liver cancer. *Nat. Genet.* **48**, 500–509 (2016).

85. Deng, L. et al. Aberrant NEAT1_1 expression may be a predictive marker of poor prognosis in diffuse large B cell lymphoma. *Cancer Biomark.* **23**, 157–164 (2018).

86. Wang, Q. M., Lian, G. Y., Song, Y., Huang, Y. F. & Gong, Y. LncRNA MALAT1 promotes tumorigenesis and immune escape of diffuse large B cell lymphoma by sponging miR-195. *Life Sci.* **231**, 116335 (2019).

87. Zheng, Z. H., You, H. Y., Feng, Y. J. & Zhang, Z. T. LncRNA KCNQ1OT1 is a key factor in the reversal effect of curcumin on cisplatin resistance in the colorectal cancer cells. *Mol. Cell Biochem.* https://doi.org/10.1007/s11010-020-03856-x (2020).

88. Xu, B. et al. LncRNA SNHG3, a potential oncogene in human cancers. *Cancer Cell Int.* **20**, 536 (2020).

89. Tian, Y. et al. lncRNA SNHG14 promotes oncogenesis and immune evasion in diffuse large-B-cell lymphoma by sequestering miR-152-3p. *Leuk Lymphoma* https://doi.org/10.1080/10428194.2021.1876866 (2021).

90. Alexandrov, L. B. et al. Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).

91. Casellas, R. et al. Mutations, kataegis and translocations in B cells: Understanding AID promiscuous activity. *Nat. Rev. Immunol.* **16**, 164–176 (2016).

92. Radke, J. et al. The genomic and transcriptional landscape of primary central nervous system lymphoma (1.1.0) [Data set]. *Zenodo* https://doi.org/10.5281/zenodo.6054242 (2021).

93. Khodabakhshi, A. H. et al. Recurrent targets of aberrant somatic hypermutation in lymphoma. *Oncotarget* **3**, 1308–1319 (2012).

94. Papavasiliou, F. N. & Schatz, D. G. Somatic hypermutation of immunoglobulin genes: Merging mechanisms for genetic diversity. *Cell* **109**, S35–S44 (2002).

95. Morin, R. D. et al. Mutational and structural analysis of diffuse large B-cell lymphoma using whole-genome sequencing. *Blood* **122**, 1256–1265 (2013).

96. Rimsza, L. M. et al. Loss of MHC class II gene and protein expression in diffuse large-B-cell lymphoma is related to decreased tumor immunosurveillance and poor patient survival regardless of other prognostic factors: A follow-up study from the Leukemia and Lymphoma Molecular Profiling Project. *Blood* **103**, 4251–4258 (2004).

97. Baruah, P. et al. Impact of p16 status on pro- and anti-angiogenesis factors in head and neck cancers. *Br. J. Cancer* **113**, 653–659 (2015).

98. Eymin, B., Leduc, C., Coll, J. L., Brambilla, E. & Gazzeri, S. p14ARF induces G2 arrest and apoptosis independently of p53 leading to regression of tumours established in nude mice. *Oncogene* **22**, 1822–1835 (2003).

99. England, N. L. et al. Identification of human tumour suppressor genes by monochromosome transfer: rapid growth-arrest response mapped to 9p21 is mediated solely by the cyclin-D-dependent kinase inhibitor gene, CDKN2A (p16INK4A). *Carcinogenesis* **17**, 1567–1575 (1996).

100. Challa-Malladi, M. et al. Combined genetic inactivation of beta2-Microglobulin and CD58 reveals frequent escape from immune recognition in diffuse large B cell lymphoma. *Cancer Cell* **20**, 728–740 (2011).

101. Schwindt, H. et al. Chromosomal imbalances and partial uniparental disomies in primary central nervous system lymphoma. *Leukemia* **23**, 1875–1884 (2009).

102. Yu, X. & Li, Z. TOX gene: a novel target for human cancer gene therapy. *Am. J. Cancer Res.* **5**, 3516–3524 (2015).

103. Kim, K. et al. Single-cell transcriptome analysis reveals TOX as a promoting factor for T cell exhaustion and a predictor for anti-PD-1 responses in human cancer. *Genome Med* **12**, 22 (2020). https://doi.org/10.1186/s13073-020-00722-9.

104. Astuti, D. et al. Germline mutations in DIS3L2 cause the Perlman syndrome of overgrowth and Wilms tumor susceptibility. *Nat. Genet.* **44**, 277–284, https://doi.org/10.1038/ng.1071 (2012).

105. Xing, S. et al. DIS3L2 promotes progression of hepatocellular carcinoma via hnRNP U-mediated alternative splicing. *Cancer Res.* **79**, 4923–4936 (2019).

106. McAllister-Lucas, L. M., Baens, M. & Lucas, P. C. MALT1 protease: A new therapeutic target in B lymphoma and beyond? *Clin. Cancer Res.: Off. J. Am. Assoc. Cancer Res.* **17**, 6623–6631 (2011).

107. Kameoka, Y. et al. Contig array CGH at 3p14.2 points to the FRA3B/FHIT common fragile region as the target gene in diffuse large B-cell lymphoma. *Oncogene* **23**, 9148–9154 (2004).

108. Roy, D., Sin, S. H., Damania, B. & Dittmer, D. P. Tumor suppressor genes FHIT and WWOX are deleted in primary effusion lymphoma (PEL) cell lines. *Blood* **118**, e32–e39 (2011).

109. Haller, F. et al. Enhancer hijacking activates oncogenic transcription factor NR4A3 in acinic cell carcinomas of the salivary glands. *Nat. Commun.* **10**, 368 (2019).

110. Zhang, Y. et al. High-coverage whole-genome analysis of 1220 cancers reveals hundreds of genes deregulated by rearrangement-mediated cis-regulatory alterations. *Nat. Commun.* **11**, 736 (2020).

111. Zhou, X., Takatoh, J. & Wang, F. The mammalian class 3 PI3K (PIK3C3) is required for early embryogenesis and cell proliferation. *PLoS One* **6**, e16358 (2011).

112. Munson, M. J. & Ganley, I. G. MTOR, PIK3C3, and autophagy: Signaling the beginning from the end. *Autophagy* **11**, 2375–2376 (2015).

113. Chen, C. H. et al. Dual inhibition of PIK3C3 and FGFR as a new therapeutic approach to treat bladder cancer. *Clin. Cancer Res.: Off. J. Am. Assoc. Cancer Res.* **24**, 1176–1189 (2018).

114. Kumar, B. et al. PIK3C3 inhibition promotes sensitivity to colon cancer therapy by inhibiting cancer stem cells. *Cancers* https://doi.org/10.3390/cancers13092168 (2021).

115. Liu, F. et al. PIK3C3 regulates the expansion of liver CSCs and PIK3C3 inhibition counteracts liver cancer stem cell activity induced by PI3K inhibitor. *Cell Death Dis.* **11**, 427 (2020).

116. Fukai, J. et al. EphA4 promotes cell proliferation and migration through a novel EphA4-FGFR1 signaling pathway in the human glioma U251 cell line. *Mol. Cancer Ther.* **7**, 2768–2778 (2008).

117. Iiizumi, M. et al. EphA4 receptor, overexpressed in pancreatic ductal adenocarcinoma, promotes cancer cell growth. *Cancer Sci.* **97**, 1211–1216 (2006).

118. Hachim, I. Y. et al. Transforming growth factor-beta regulation of ephrin type-A receptor 4 signaling in breast cancer cellular migration. *Sci. Rep.* **7**, 14976 (2017).

119. Lin, C. Y. et al. High expression of EphA4 predicted lesser degree of tumor regression after neoadjuvant chemoradiotherapy in rectal cancer. *J. Cancer* **8**, 1089–1096 (2017).

120. Kina, S. et al. Targeting EphA4 abrogates intrinsic resistance to chemotherapy in well-differentiated cervical cancer cell line. *Eur. J. Pharm.* **840**, 70–78 (2018).

121. Schmitz, R., Ceribelli, M., Pittaluga, S., Wright, G. & Staudt, L. M. Oncogenic mechanisms in Burkitt lymphoma. *Cold Spring Harb. Perspect. Med.* https://doi.org/10.1101/cshperspect.a014282 (2014).

122. Kuppers, R. & Dalla-Favera, R. Mechanisms of chromosomal translocations in B cell lymphomas. *Oncogene* **20**, 5580–5594 (2001).

123. Seifert, M., Scholtysik, R. & Kuppers, R. Origin and pathogenesis of B cell lymphomas. *Methods Mol. Biol.* **1956**, 1–33 (2019).

124. Krull, J. E. et al. Somatic copy number gains in MYC, BCL2, and BCL6 identifies a subset of aggressive alternative-DH/TH DLBCL patients. *Blood Cancer J.* **10**, 117 (2020).

125. Scott, D. W. et al. High-grade B-cell lymphoma with MYC and BCL2 and/or BCL6 rearrangements with diffuse large B-cell lymphoma morphology. *Blood* **131**, 2060–2064 (2018).

126. Alexandrov, L. B. et al. The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).

127. Fong, P. C. et al. Inhibition of poly(ADP-ribose) polymerase in tumors from BRCA mutation carriers. *N. Engl. J. Med.* **361**, 123–134 (2009).

128. Poti, A. et al. Correlation of homologous recombination deficiency induced mutational signatures with sensitivity to PARP inhibitors and cytotoxic agents. *Genome Biol.* **20**, 240 (2019).

129. Gu, Z., Schlesner, M. & Hubschmann, D. cola: An R/Bioconductor package for consensus partitioning through a general framework. *Nucleic Acids Res.* **49**, e15 (2021).

130. Seimiya, M. et al. Stage-specific expression of Clast6/E3/LAPTM5 during B cell differentiation: elevated expression in human B lymphomas. *Int. J. Oncol.* **22**, 301–304 (2003).

131. Kim, Y., Shin, Y. & Kang, G. H. Prognostic significance of CD103+ immune cells in solid tumor: A systemic review and meta-analysis. *Sci. Rep.* **9**, 3808 (2019).

132. Harada, K. et al. Telomerase activity in central nervous system malignant lymphoma. *Cancer* **86**, 1050–1055 (1999).

133. Feuerbach, L. et al. TelomereHunter—in silico estimation of telomere content and composition from cancer genomes. *BMC Bioinform.* **20**, 272 (2019).

134. Panebianco, F., Nikitski, A. V., Nikiforova, M. N. & Nikiforov, Y. E. Spectrum of TERT promoter mutations and mechanisms of activation in thyroid cancer. *Cancer Med* **8**, 5831–5839 (2019).

135. Bell, R. J. et al. Understanding TERT promoter mutations: A common path to immortality. *Mol. Cancer Res.* **14**, 315–323 (2016).

136. Bruno, A. et al. TERT promoter mutations in primary central nervous system lymphoma are associated with spatial distribution in the splenium. *Acta Neuropathol.* **130**, 439–440 (2015).

137. Stogbauer, L., Stummer, W., Senner, V. & Brokinkel, B. Telomerase activity, TERT expression, hTERT promoter alterations, and alternative lengthening of the telomeres (ALT) in meningiomas—a systematic review. *Neurosurg. Rev.* https://doi.org/10.1007/s10143-019-01087-3 (2019).

138. Ichimura, K. TERT promoter mutation as a diagnostic marker for diffuse gliomas. *Neuro-Oncol.* **21**, 417–418 (2019).

139. Lee, Y. et al. The frequency and prognostic effect of TERT promoter mutation in diffuse gliomas. *Acta Neuropathol. Commun.* **5**, 62 (2017).

140. Liu, Z. et al. Association between TERT rs2853669 polymorphism and cancer risk: A meta-analysis of 9,157 cases and 11,073 controls. *PLoS One* **13**, e0191560 (2018).

141. Grommes, C. & Younes, A. Ibrutinib in PCNSL: The curious cases of clinical responses and aspergillosis. *Cancer Cell* **31**, 731–733 (2017).

142. Rheinbay, E. et al. Analyses of non-coding somatic drivers in 2658 cancer whole genomes. *Nature* **578**, 102–111, https://doi.org/10.1038/s41586-020-1965-x (2020).

143. Biesecker, L. G., Shianna, K. V. & Mullikin, J. C. Exome sequencing: The expert view. *Genome Biol.* **12**, 128 (2011).

144. Fatica, A. & Bozzoni, I. Long non-coding RNAs: New players in cell differentiation and development. *Nat. Rev. Genet.* **15**, 7–21 (2014).

145. Jardin, F. et al. Diffuse large B-cell lymphomas with CDKN2A deletion have a distinct gene expression signature and a poor prognosis under R-CHOP treatment: A GELA study. *Blood* **116**, 1092–1104 (2010).

146. Fangazio, M. et al. Genetic mechanisms of HLA-I loss and immune escape in diffuse large B cell lymphoma. *Proc. Natl Acad. Sci. USA* https://doi.org/10.1073/pnas.2104504118 (2021).

147. Monti, S. 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* **22**, 359–372 (2012).

148. Tao, W. & Levine, A. J. P19(ARF) stabilizes p53 by blocking nucleo-cytoplasmic shuttling of Mdm2. *Proc. Natl Acad. Sci. USA* **96**, 6937–6941 (1999).

149. Yamada, S., Ishida, Y., Matsuno, A. & Yamazaki, K. Primary diffuse large B-cell lymphomas of central nervous system exhibit remarkably high prevalence of oncogenic MYD88 and CD79B mutations. *Leuk. Lymphoma* **56**, 2141–2145 (2015).

150. Georgakopoulos-Soares, I. et al. Transcription-coupled repair and mismatch repair contribute towards preserving genome integrity at mononucleotide repeat tracts. *Nat. Commun.* **11**, 1980 (2020).

151. Ferch, U. et al. Inhibition of MALT1 protease activity is selectively toxic for activated B cell-like diffuse large B cell lymphoma cells. *J. Exp. Med.* **206**, 2313–2320 (2009).

152. Oakes, C. C. et al. DNA methylation dynamics during B cell maturation underlie a continuum of disease phenotypes in chronic lymphocytic leukemia. *Nat. Genet.* **48**, 253–264 (2016).

153. Bodker, J. S. et al. A multiple myeloma classification system that associates normal B-cell subset phenotypes with prognosis. *Blood Adv.* **2**, 2400–2411 (2018).

154. Sugita, Y. et al. Primary central nervous system lymphomas and related diseases: Pathological characteristics and discussion of the differential diagnosis. *Neuropathology* **36**, 313–324 (2016).

155. Vinagre, J. et al. Frequency of TERT promoter mutations in human cancers. *Nat. Commun.* **4**, 2185 (2013).

156. Stocher, M. et al. Parallel detection of five human herpes virus DNAs by a set of real-time polymerase chain reactions in a single run. *J. Clin. Virol.* **26**, 85–93 (2003).

157. Kretzmer, H. et al. DNA methylome analysis in Burkitt and follicular lymphomas identifies differentially methylated regions linked to somatic mutation and transcriptional control. *Nat. Genet.* **47**, 1316–1325 (2015).

158. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

159. Tischler, G. & Leonard, S. biobambam: tools for read pair collation based algorithms on BAM files. *Source Code Biol. Med.* **9**, 13 (2014).

160. Arthur, S. E. et al. Genome-wide discovery of somatic regulatory variants in diffuse large B-cell lymphoma. *Nat. Commun.* **9**, 4001 (2018).

161. Zhang, J. et al. Genetic heterogeneity of diffuse large B-cell lymphoma. *Proc. Natl Acad. Sci. USA* **110**, 1398–1403 (2013).

162. Ishaque, N. et al. Whole genome sequencing puts forward hypotheses on metastasis evolution and therapy in colorectal cancer. *Nat. Commun.* **9**, 4782 (2018).

163. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

164. Consortium, E. P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).

165. Rimmer, A. et al. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat. Genet.* **46**, 912–918 (2014).

166. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).

167. Jiang, Y., Qiu, Y., Minn, A. J. & Zhang, N. R. Assessing intratumor heterogeneity and tracking longitudinal and spatial clonal evolutionary history by next-generation sequencing. *Proc. Natl Acad. Sci. USA* **113**, E5528–E5537 (2016).

168. Sahm, F. et al. Meningiomas induced by low-dose radiation carry structural variants of NF2 and a distinct mutational signature. *Acta Neuropathol.* **134**, 155–158 (2017).

169. Uhrig, S. et al. Accurate and efficient detection of gene fusions from RNA sequencing data. *Genome Res.* **31**, 448–460 (2021).

170. Kleinheinz, K. et al. ACEseq—allele specific copy number estimation from whole genome sequencing. Preprint at *bioRxiv* https://doi.org/10.1101/210807 (2017).

171. Hubschmann, D. et al. Analysis of mutational signatures with yet another package for signature analysis. *Genes Chromosomes Cancer* https://doi.org/10.1002/gcc.22918 (2020).

172. Leiserson, M. D., Wu, H. T., Vandin, F. & Raphael, B. J. CoMEt: A statistical approach to identify combinations of mutually exclusive alterations in cancer. *Genome Biol.* **16**, 160 (2015).

173. Gonzalez-Perez, A. et al. IntOGen-mutations identifies cancer drivers across tumor types. *Nat. Methods* **10**, 1081–1082 (2013).

174. Paramasivam, N. et al. Mutational patterns and regulatory networks in epigenetic subgroups of meningioma. *Acta Neuropathol.* **138**, 295–308 (2019).

175. Dobin, A. et al. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

176. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).

177. Jurmeister, P. et al. Parallel screening for ALK, MET and ROS1 alterations in non-small cell lung cancer with implications for daily routine testing. *Lung Cancer* **87**, 122–129 (2015).

Acknowledgements

Author contributions

J.R., R.K., E.S., F.P., S.B., D.L., and M.W. performed experiments and analyzed data. Z.G., L.S., D.H., U.H.T., X.P.H., N.P., M.S., C.L., D.J., G.P.B., S.S., S.U., and N.I. carried out the bioinformatics analyses and processed data. D.P., H.R., and A.O. performed sample acquisition and histological analyses. I.A. performed histological analyses. D.L. performed FISH analyses. A.K., M.M., J.O., K.F., P.V., D.M., Y.W., A.J., and A.U. clinically assessed or operated patients and contributed to sample acquisition. C.H.M., S.W.i., and J.R. coordinated sample acquisition. S.W.i. and S.W.o. oversaw sequencing of the samples. R.S., M.S., L.T., and R.K.ü. provided data from the MMML cohort. Experimental design and execution were overseen by F.L.H., S.W.i., O.W., C.H.M., B.B., C.A.S., R.S., M.H., R.K.ü., and R.E. J.R., N.I., S.W.i., F.L.H., and R.S. designed the study. RS coordinated the ICGC MMML-Seq network. J.R. and N.I. wrote the manuscript. J.R. prepared the final figures. F.L.H., S.W.i., O.W., C.H.M., B.B., C.A.S., R.S, and M.H. critically revised the manuscript. All authors critically read, reviewed, and approved the final manuscript.

Funding
Open Access funding enabled and organized by Projekt DEAL.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-022-30050-y.

Correspondence and requests for materials should be addressed to Josefine Radke or Naveed Ishaque.

Peer review information *Nature Communications* thanks Mark Roschewski and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at http://www.nature.com/reprints

[1]Department of Neuropathology, Charité - Universitätsmedizin Berlin, corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin, Berlin Institute of Health, Berlin, Germany. [2]German Cancer Consortium (DKTK), Partner Site Charité Berlin, Berlin, Germany. [3]Institute of Pathology, Universitätsmedizin Greifswald, Greifswald, Germany. [4]Digital Health Center, Berlin Institute of Health (BIH) at Charité - Universitätsmedizin Berlin, Berlin, Germany. [5]Molecular Precision Oncology Program, NCT Heidelberg, Heidelberg, Germany. [6]Division of Applied Bioinformatics, German Cancer Research Center (DKFZ), Heidelberg, Germany. [7]Division of Translational Medical Oncology, National Center for Tumor Diseases (NCT), Heidelberg, Germany. [8]Computational Oncology, Molecular Diagnostics Program, NCT Heidelberg, Heidelberg, Germany. [9]Heidelberg Institute for Stem Cell Technology and Experimental Medicine (HI-STEM gGmbH), Heidelberg, Germany. [10]Department of Pediatric Oncology, Hematology and Immunology, University Hospital, Heidelberg, Germany. [11]German Cancer Consortium (DKTK), Heidelberg, Germany. [12]Division Neuroblastoma Genomics, German Cancer Research Center (DKFZ), Heidelberg, Germany. [13]Institute of Human Genetics, Ulm University & Ulm University Medical Center, Ulm, Germany. [14]Institut d'Investigacions Biomèdiques August Pi Sunyer (IDIBAPS), CIBER de Cáncer, Barcelona, Spain. [15]Division of Molecular Genome Analysis, German Cancer Research Center (DKFZ), Heidelberg, Germany. [16]Hopp Children's Cancer Center, NCT Heidelberg (KiTZ), Heidelberg, Germany. [17]Division of Pediatric Neurooncology, German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ), Heidelberg, Germany. [18]Bioinformatics and Omics Data Analytics, German Cancer Research Center (DKFZ), Heidelberg, Germany. [19]Biomedical Informatics, Data Mining and Data Analytics, Augsburg University, Augsburg, Germany. [20]Institute of Cell Biology (Cancer Research), Medical Faculty, University of Duisburg-Essen, Essen, Germany. [21]German Cancer Consortium (DKTK), Partner Site Essen, Essen, Germany. [22]Department of Hematology, Oncology and Cancer Immunology, Charité - Universitätsmedizin Berlin, corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin, Berlin Institute of Health, Berlin, Germany. [23]Department of Neurosurgery, Charité - Universitätsmedizin Berlin, corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin, Berlin Institute of Health, Berlin, Germany. [24]Department for Neurosurgery, Vivantes Klinikum im Friedrichshain, Berlin, Germany. [25]Department for Neurosurgery, Vivantes Klinikum Neukölln, Berlin, Germany. [26]Department of Hematology and Oncology, Georg-Augusts-University of Göttingen, Göttingen 37075, Germany. [27]Department of Pathology, Charité - Universitätsmedizin Berlin, corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin, Berlin Institute of Health, Berlin, Germany. [28]Medical Department of Hematology and Oncology, Johannes Kepler University, Kepler Universitätsklinikum, Linz, Austria. [29]Department of Hematology, Oncology and Tumor Immunology and Molekulares Krebsforschungszentrum, Charité - Universitätsmedizin Berlin, corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin, Berlin Institute of Health, Berlin, Germany. [30]German Cancer Research Center (DKFZ), Heidelberg, Germany. [31]Genomics and Proteomics Core Facility, German Cancer Research Center (DKFZ), Heidelberg, Germany. [32]Department of Neurosurgery, University Hospital Heidelberg, Heidelberg, Germany. [33]Cluster of Excellence, NeuroCure, Charitéplatz 1, 10117 Berlin, Germany. [34]German Center for Neurodegenerative Diseases (DZNE) Berlin, 10117 Berlin, Germany.[68]These authors contributed equally: Josefine Radke, Naveed Ishaque. [69]These authors jointly supervised this work: Reiner Siebert, Stefan Wiemann, Frank L. Heppner. *A list of authors and their affiliations appears at the end of the paper. ✉email: josefine.radke@med.uni-greifswald.de; naveed.ishaque@bih-charite.de

ICGC MMML-Seq Consortium

Coordination (C1) Reiner Siebert[13,35], Susanne Wagner[35], Andrea Haake[35], Julia Richter[35,36] & Gesine Richter[35]

Data Center (C2) Roland Eils[4,37,38], Chris Lawerenz[37], Jürgen Eils[37], Jules Kerssemakers[37], Christina Jaeger-Schmidt[37] & Ingrid Scholz[37]

Clinical Centers (WP1) Anke K. Bergmann[35,39], Christoph Borst[40], Friederike Braulke[26], Birgit Burkhardt[41,42], Alexander Claviez[39], Martin Dreyling[43], Sonja Eberth[43], Hermann Einsele[44], Norbert Frickhofen[45], Siegfried Haas[40], Martin-Leo Hansmann[46], Dennis Karsch[47], Nicole Klepl[26], Michael Kneba[47], Jasmin Lisfeld[41],

Luisa Mantovani-Löffler[48], Marius Rohde[41], German Ott[49], Christina Stadler[26], Peter Staib[50], Stephan Stilgenbauer[51], Lorenz Trümper[26] & Thorsten Zenz[52]

Normal Cells (WPN) Martin-Leo Hansmann[46], Dieter Kube[26], Ralf Küppers [20,21] & Marc Weniger[20,21]

Pathology and Analyte Preparation (WP2-3) Siegfried Haas[40], Michael Hummel[27], Wolfram Klapper[36], Ulrike Kostezka[53], Dido Lenze[27], Peter Möller[54], Andreas Rosenwald[55], German Ott[49] & Monika Szczepanowski[36]

Sequencing and genomics (WP4-7) Ole Ammerpohl[13,35], Sietse M. Aukema[35,36], Vera Binder[56], Arndt Borkhardt[56], Andrea Haake[35], Jessica I. Hoell[56], Ellen Leich[55], Peter Lichter[57], Cristina López[13,35], Inga Nagel[35], Jordan Pischimariov[55], Bernhard Radlwimmer[57], Julia Richter[35,36], Philip Rosenstiel[58], Andreas Rosenwald[55], Markus Schilhabel[58], Stefan Schreiber[59], Inga Vater[35], Rabea Wagener[13,35] & Reiner Siebert[13,35]

Bioinformatics (WP8-9) Stephan H. Bernhart[60,61,62], Hans Binder[60,61], Benedikt Brors[5], Gero Doose[60,61,62], Roland Eils[4,37,38], Steve Hoffmann[60,61,62], Lydia Hopp[60], Daniel Hübschmann[5,8,9,10,11], Kortine Kleinheinz[37,38], Helene Kretzmer[60,61,62], Markus Kreuz[63], Jan Korbel[64], David Langenberger[60,61,62], Markus Loeffler[64], Maciej Rosolowski[64], Matthias Schlesner[18,19], Peter F. Stadler[60,61,62,65,66,67] & Stephanie Sungalee[64]

[35]Institute of Human Genetics, Christian-Albrechts-University, Kiel, Germany. [36]Institute of Pathology, Hematopathology Section, Christian-Albrechts-University, Kiel, Germany. [37]Division of Theoretical Bioinformatics (B080), German Cancer Research Center (DKFZ), Heidelberg, Germany. [38]Department for Bioinformatics and Functional Genomics, Institute of Pharmacy and Molecular Biotechnology and Bioquant, University of Heidelberg, Heidelberg, Germany. [39]Department of Pediatrics, University Hospital Schleswig-Holstein, Campus Kiel, Kiel, Germany. [40]Department of Internal Medicine/Hematology, Friedrich-Ebert-Hospital, Neumünster, Germany. [41]Pediatric Hematology and Oncology, University Hospital Muenster, Muenster, Germany. [42]Pediatric Hematology and Oncology, University Hospital Giessen, Giessen, Germany. [43]Department of Medicine III - Campus Grosshadern, University Hospital Munich, Munich, Germany. [44]Department of Medicine and Poliklinik II, University Hospital Würzburg, Würzburg, Germany. [45]Department of Medicine III, Hematology and Oncology, Dr. Horst-Schmidt-Kliniken of Wiesbaden, Wiesbaden, Germany. [46]Senckenberg Institute of Pathology, University of Frankfurt Medical School, Frankfurt am Main, Germany. [47]Department of Internal Medicine II: Hematology and Oncology, University Medical Center Schleswig-Holstein, Campus Kiel, Kiel, Germany. [48]Hematology and Oncology, Hospital of Internal Medicine II, St-Georg Hospital Leipzig, Leipzig, Germany. [49]Department of Pathology, Robert-Bosch-Hospital, Stuttgart, Germany. [50]Clinic for Hematology and Oncology, St.-Antonius-Hospital, Eschweiler, Germany. [51]Department for Internal Medicine III, University of Ulm and University Hospital of Ulm, Ulm, Germany. [52]National Centre for Tumor Disease, Heidelberg, Germany. [53]Comprehensive Cancer Center Ulm (CCCU), University Hospital Ulm, Ulm, Germany. [54]Institute of Pathology, University of Ulm and University Hospital of Ulm, Ulm, Germany. [55]Institute of Pathology, Comprehensive Cancer Center Mainfranken, University of Würzburg, Würzburg, Germany. [56]Department of Pediatric Oncology, Hematology and Clinical Immunology, Heinrich-Heine-University, Düsseldorf, Germany. [57]Division of Molecular Genetics, German Cancer Research Center (DKFZ), Heidelberg, Germany. [58]Institute of Clinical Molecular Biology, Christian-Albrechts-University, Kiel, Germany. [59]Department of General Internal Medicine, University Kiel, Kiel, Germany. [60]Interdisciplinary Center for Bioinformatics, University of Leipzig, Leipzig, Germany. [61]Department of Computer science, University of Leipzig, Leipzig, Germany. [62]Transcriptome Bioinformatics, LIFE Research Center for Civilization Diseases, University of Leipzig, Leipzig, Germany. [63]Institute for Medical Informatics Statistics and Epidemiology, University of Leipzig, Leipzig, Germany. [64]EMBL Heidelberg, Genome Biology, Heidelberg, Germany. [65]RNomics Group, Fraunhofer Institute for Cell Therapy and Immunology IZI, Leipzig, Germany. [66]Santa Fe Institute, Santa Fe, NM, USA. [67]Max-Planck-Institute for Mathematics in Sciences, Leipzig, Germany.