Genomic Landscape and Molecular Subtypes of Primary Central Nervous System 1

Lymphoma 2

- Shengjie Li^{1,2,3,4,5,6*#}, Danhui Li^{7#}, Zuguang Xia^{8,9#}, Jianing Wu⁶, Jun Ren⁶, Yingzhu Li⁶, 3
- Jiazhen Cao^{8,10}, Ying Sun¹¹, Liyang Zhang¹², Hongwei Ye¹², Xingtao Zhou^{5*}, Chengxun Li^{13*}, 4
- Wenjun Cao^{6*}, Ying Mao^{1,2,3,4*} 5
- 1. Department of Neurosurgery, Huashan Hospital, Department of Clinical Laboratory, Eye & 6
- ENT Hospital, Shanghai Medical College, Fudan University, Shanghai, China 7
- 2. National Center for Neurological Disorders, Shanghai, China 8
- 9 3. Shanghai Key Laboratory of Brain Function Restoration and Neural Regeneration,
- Shanghai, China 10
- 4. Neurosurgical Institute of Fudan University, Shanghai, China 11
- 5. Eye Institute and Department of Ophthalmology, Eye & ENT Hospital, Fudan University, 12
- 13 Shanghai, China
- 6. Department of Clinical Laboratory, Eye & ENT Hospital, Shanghai Medical College, Fudan 14
- University, Shanghai, China 15
- 7. Department of Pathology, RenJi Hospital, School of Medicine, Shanghai JiaoTong 16
- 17 University, Shanghai, 200127, China
- 8. Department of Medical Oncology, Fudan University Shanghai Cancer Center, Fudan 18
- University, Shanghai, China 19
- 9. Department of Oncology, Shanghai Medical College, Fudan University, Shanghai, China 20
- 10. Department of Clinical Laboratory, Fudan University Shanghai Cancer Center, Fudan 21
- 22 University, Shanghai, China
- 11. GenomiCare Biotechnology (Shanghai) Co. Ltd., Shanghai, China 23
- 12. Guangzhou KingMed Center for Clinical Laboratory Co. Ltd., Guangzhou, China 24
- 25 13.Institutes of Biomedical Sciences, Fudan University, Shanghai, 200032, China

26

- #Shengjie Li, Danhui Li, and Zuguang Xia contributed equally to this work. 27
- 28

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

29 **Corresponding Authors:**

- 30 Wenjun Cao, Department of Clinical Laboratory, Eye & ENT Hospital, Fudan University, No.
- 83 Fenyang Road, Shanghai, China; Electronic address: wgkjyk@aliyun.com.
- 32 Chengxun Li, Institutes of Biomedical Sciences, Fudan University, Shanghai, 200032, China
- 33 Electronic address: licx21@m.fudan.edu.cn
- 34 Xingtao, Zhou, Eye Institute and Department of Ophthalmology, Eye & ENT Hospital, Fudan
- University, Shanghai, 200031, China. Electronic address: doctzhouxingtao@163.com.
- 36 Ying Mao, Department of Neurosurgery, Huashan Hospital, Fudan University, No. 12
- 37 Wulumuqi Road, Shanghai, 200040, China; Electronic address: maoying@fudan.edu.cn

All rights reserved. No reuse allowed without permission.



40 Highlights

- 1. In this study, the genomic landscape of 140 Chinese patients with primary central nervous
- 42 system lymphomas (PCNSLs) was evaluated.
- 43 2. Chinese PCNSL patients have a defining genetic signature that differs from that of both
- 44 PCNSL patients in other racial groups and DLBCL patients.
- 3. Three robust molecular subtypes of PCNSL related to clinical and molecular features were
 identified and validated.
- 47 4. The rate of EP300 mutation in PCNSLs was approximately three times higher among
- Asians than among Western patients, resulting in unfavorable outcomes independent of the
- 49 specific mutation site.
- 50

51 Abstract

Primary central nervous system lymphoma (PCNSL) is a rare and aggressive brain tumor 52 with a poor prognosis and almost exclusively comprises diffuse large B-cell lymphoma 53 (DLBCL). Its genetic characteristics and molecular subtypes in Chinese patients remain 54 poorly understood, which in turn makes developing effective new therapies challenging. In 55 our study, 140 Chinese patients with PCNSL that was newly diagnosed at one of three tertiary 56 care centers and who underwent extensive follow-up were included. With this sample, we 57 performed a genomic study aimed at expanding the genomic landscape and identifying new 58 59 molecular subtypes. We first confirmed that the molecular subtype categories of DLBCL, as previously published, are not applicable to PCNSLs in Chinese patients. We then identified 60 (n = 58) and validated (n = 82) three prominent genetic subtypes related to different clinical 61 and molecular features of PCNSL and further confirmed them in an independent external 62 Chinese PCNSL cohort (n = 36). We called these BMIs (from the co-occurrence of mutations 63 in two genes among BTG1, MYD88, and IRF4), which are associated with favorable 64 outcomes; E3s (so-called EP300 mutations), which are associated with unfavorable 65 outcomes; and UCs (unclassified, without characteristic mutations). Importantly, EP300 was 66 mutated in more PCNSLs from Asian patients (16.88%) than from Western patients (< 5.26%), 67 resulting in unfavorable outcomes independent of the specific mutation site. Our analysis 68 comprehensively reveals the genomic landscape of PCNSL in Chinese patients and 69 emphasizes the clinical value of molecular classification for improving precision medicine 70 71 strategies.

73 Introduction

Primary central nervous system lymphoma (PCNSL), a rare subtype of extranodal non-74 Hodgkin lymphoma, accounts for >60% of PCNSL cases and has an unfavorable prognosis 75 even when the patient receives standard treatment, which is based on a high-dose 76 methotrexate (HD-MTX) regimen^{1,2}. PCNSL is a unique lymphoma that differs from other 77 tumors and is characterized as a genetically heterogeneous disorder marked by a variety of 78 low-frequency mutations, somatic copy number alterations, and structural variants^{3–6}. Thus, 79 combining standard immunochemotherapy with promising novel agents that target specific 80 pathways via different molecular clusters may improve patient prognosis^{7,8}. 81

Pathologically, nearly 95% of PCNSLs are diffuse large B-cell lymphomas (DLBCLs). 82 Recently, DLBCL has been categorized into various molecular clusters on the basis of 83 genomic sequencing and RNA sequencing, including guartet categorization by L.M. Staudt's 84 team⁹, quintet categorization by *M.A. Shipp*'s team¹⁰, *L.M. Staudt*'s team's refined septet 85 method¹¹, and the LymphPlex classification proposed by W.L. Zhao's team¹². However, 86 PCNSL has been proven to be a biological entity that is molecularly distinct from DLBCL¹³, 87 so these molecular clusters of DLBCL may not be applicable to PCNSL. Notably, the 88 integration of genome-wide data from multiomic studies by *A. Alentorn* et al.¹⁴ showed four 89 molecular patterns of PCNSL using only gene expression data. These patterns have a 90 distinctive prognostic impact, providing a basis for future clinical stratification and subtype-91 based targeted interventions. Owing to the molecular heterogeneity between various ethnic 92 groups of PCNSL patients^{3,15}, the applicability of the proposed classifications for PCNSL to 93 other populations requires further confirmation. No effective biomarker or molecular 94 classification scheme exists to tailor therapies for individual PCNSL patients. Furthermore, 95 the issue of PCNSL molecular heterogeneity in Chinese patients has not been adequately 96 addressed, primarily because the data have come from small, single-center studies of 97 Chinese PCNSL patients^{16,17}. 98

⁹⁹ To address these issues, we conducted whole-exome sequencing (WES) on specimens ¹⁰⁰ obtained from 140 patients with newly diagnosed PCNSL who were seen at one of three ¹⁰¹ tertiary care cancer centers. These patients were divided into a discovery cohort and a

validation cohort. Among the 140 patients, 94.29% received treatment with the HD-MTX-102 based chemotherapy regimen, and all 140 patients underwent long-term follow-up. This 103 representative and clinically annotated PCNSL cohort was utilized to comprehensively detect 104 mutations. Additionally, to facilitate routine clinical implementation, we identified three 105 molecular subtypes composed only of single-nucleotide variations (SNVs); these SNVs were 106 associated with different outcomes, and this was confirmed in the validation cohort and in an 107 independent external Chinese PCNSL cohort¹⁷. Our results comprehensively reveal the 108 109 genomic landscape of PCNSLs in Chinese patients and provide a clinically actionable PCNSL classification system, thereby improving precision medicine strategies. 110

111

112 **Results**

113 **PCNSL genomic landscape**

Using WES, we detected mutations in 140 patients who were newly diagnosed with PCNSL 114 at one of three cancer centers; in 59% of cases, blood samples were lacking. Because 115 Epstein–Barr virus (EBV)-positive PCNSL is recognized as an immunodeficiency-associated 116 PCNSL in the current WHO CNS5 classification^{18,19}, all participants in the study who tested 117 negative for EBV were included. Additionally, DLBCL patients with hepatitis B virus (HBV) 118 and HIV infection may have different mutational profiles^{20,21}; thus, all participants in the study 119 who tested negative for HIV and HBV were included. The key demographic and clinical 120 121 characteristics of the patients are summarized in Figure S1 and Table S1. The study design and flow chart are displayed in Figure 1A. After filtering, we identified 115.938 genetic 122 variation events in the 140 PCNSL samples analyzed (median = 19.06 mutations/Mb, range 123 = 0.24–91.2 mutations/Mb; median number of variants = 949.5/sample) (Figure S2A, Figure 124 S3A). 125

We applied a filter to the 140 PCNSLs on the basis of the genes curated in the CIViC and
OncoKB databases to identify candidate cancer genes with hallmark mutations in PCNSL.
Some of the mutated genes were PIM1 (in 62.1% of PCNSLs), MYD88 (55.0%), KMT2D
(48.6%), CD79B (45.7%), PDE4DIP (45.7%), PCLO (42.9%), BTG2 (38.6%), LRP1B (37.9%),
FAT4 (37.1%), RNF213 (35%), FAT1 (34.3%), ZFHX3 (33.6%), DTX1 (31.4%), KMT2C

(31.4%), SETD1B (31.4%), MPEG1 (29.3%), CREBBP (27.9%), ANKRD11 (27.1%), KAT6B 131 (27.1%), BTG1 (25%), and PTEN (25%) (Figure 1B), which are involved in chromatin histone 132 modification, BCR-TLR-mediated NF-kB signaling, immune signaling, the cell cycle, PI3 133 kinase signaling, MAPK signaling, and Wnt/ β -catenin signaling (**Figure S2B**)^{3,5,9–11,22–29}. In 134 this study, we also identified several other pathways in PCNSL, such as pathways related to 135 genome integrity, RTK signaling, RNA abundance, TGFB signaling, TOR signaling, and 136 apoptosis (Figure S2B), many of which have defined roles in other cancers³⁰. The mutually 137 138 exclusive or cooccurring relationships of these genes are shown in Figure S3D. Survival associated with candidate cancer genes with mutation frequencies greater than 15% plus the 139 NOTCH2 and EZH2 genes⁹ is shown in **Figure S4**. 140

In the search for focal copy number alterations (CNAs) in the 140 PCNSLs, we detected 141 significant recurrent amplifications at chromosomal locations 1g21.1, 4g16.3, 11p15.5, 142 11q12.2, 17q12, 19p13.3, 22q11.1, Yq11.221, and Yq12. We also detected deletions at 11q11, 143 11q12.1, 14q32.33, 17q11.2, 18p11.21, 19p12, 19q13.42, and 22q11.21 (Figure 1C). The 144 arm-level CNAs among all 140 PCNSL samples are displayed in Figure S2C. The pattern of 145 146 somatic mutations caused by the different mutational processes in the genome, termed *mutational signatures*, was calculated and compared to the well-established signatures in 147 COSMIC³¹. The mutational signature contributions are shown in **Figure S2D**. The COSMIC 148 5, COSMIC 45, and COSMIC 1 signatures contributed the most to the PCNSL genome 149 (Figure S2E). These factors are associated with aging, tobacco smoking, NER deficiency, 150 the regulation of oxidative DNA damage repair, and the spontaneous deamination of 5-151 methylcytosine. Taken together, these results establish a comprehensive genomic landscape 152 of PCNSL in Chinese individuals. 153

154 Comparison of the genomic landscapes of the discovery and validation cohorts

Owing to differences in sample type between the discovery and validation cohorts (fresh tumor [discovery cohort] vs. formalin-fixed paraffin-embedded [FFPE] [validation cohort]) and the presence or absence of matched samples (paired [discovery cohort] vs. tumor-only [validation cohort]), we applied further filtering and recalibration steps to reduce false-positive calls in the validation cohort (for details, see the Methods subsection on somatic mutation

calling) and compared the genomic landscapes of the discovery and validation cohorts. The 160 key demographic and clinical characteristics of the patients in the discovery and validation 161 cohorts are summarized in **Table S2**. We identified 12,222 genetic variation events (median 162 = 3.84 mutations/Mb, median variants = 159/sample, Figure S3B, Figure S5A) in the 163 analyzed discovery cohort of PCNSL samples and 103,716 genetic variation events (median 164 = 6.85 mutations/Mb, median variants = 1597/sample, Figure S3C, Figure S5B) in the 165 analyzed validation cohort of PCNSL samples. The genetic variation profiles of the discovery 166 167 and validation cohort of PCNSL samples were similar, as revealed by principal component analysis (PCA) (Figure 2A), but the mutation rate was higher in the validation cohort of 168 PCNSL samples (Figure 2B). 169

PIM1, MYD88, CD79B, and KMT2D were the most frequently mutated candidate cancer 170 genes of PCNSL. PIM1 (70.7%), MYD88 (63.8%), CD79B (51.7%), and KMT2D (39.7%) 171 (Figure S5C) were identified in the discovery cohort, and PIM1 (56.1%), MYD88 (48.8%), 172 CD79B (41.5%), and KMT2D (54.6%) were also observed at similar mutation frequencies in 173 the validation cohort (Figure S5D). However, several of the less frequently mutated candidate 174 175 cancer genes in PCNSL whose mutation frequencies were significantly different when the discovery and validation cohorts were compared included ZFHX3, TRRAP, SZT2, RNF213, 176 RANBP2, PRKDC, PDE4DIP, PCLO, NUMA1, NOTCH1, NOTCH2, KMT2C, KMT2B, KAT6B, 177 FLT3, FAT4, FAT1, and ANKRD11 (all *P* < 0.05) (**Figure 2C**). The arm-level CNAs between 178 the discovery and validation cohort of PCNSL samples were mostly similar and are displayed 179 in Figure 2D. Significant differences in both recurrent amplifications (1g, 8g, and 12p) and 180 deletions (6p, 8p, 19p, and 19q) were observed. The results of significant recurrent focal 181 amplifications and deletions were visualized via GISTIC 2.0 in the discovery (Figure S5E) 182 183 and validation (Figure S5F) cohorts.

The tumor mutational burden (TMB) describes the number of mutations in one tumor sample and is often related to the prognosis of patients receiving clinical treatments³². The mutant-allele tumor heterogeneity (MATH) score reflects intratumor heterogeneity³³. Microsatellite instability (MSI) arises from the impaired function of DNA mismatch repair proteins, and its molecular characteristics are regarded as significant genetic markers³⁴.

Although the TMB (Figure 2E), MATH score (Figure 2F), and MSI (Figure 2G) were greater 189 in the validation cohort than in the discovery cohort (P < 0.001), the mean values were 190 relatively close. Furthermore, the variant allele frequency (VAF) values were greater in the 191 discovery cohort than in the validation cohort (P < 0.001, Figure 2H). The 3 signatures 192 extracted from the discovery samples presented cosine similarities of 86.7%, 93.1% and 94.4% 193 to COSMIC signatures 6, 45, and 10a, respectively (Figure S5G), whereas those extracted 194 from the validation samples presented similarities of 80.4%, 70.9%, and 77.9% to COSMIC 195 196 signatures 5, 1, and 23, respectively (Figure S5H). Taken together, these results suggest that the genetic variation profiles of the discovery and validation PCNSL samples were similar. 197

198 The unique genomic landscape of PCNSLs in Chinese patients

To date, only single-center studies with small sample sizes have reported heterogeneity in 199 the genomic landscape among different populations of Chinese PCNSL patients¹⁵. Thus, in 200 this large, multicenter study, we further compared the genetic landscape of our cohort with 201 previously published PCNSL/DLBCL data. The frequencies of recurrent genetic alterations in 202 PCNSL patients were compared with those in DLBCL patients²⁷. Mirror bar plots (**Figure 3A**) 203 revealed that a set of genes, such as PIM1 (62.1% vs. 16.88%), MYD88 (55% vs. 18.01%), 204 IRF4 (15% vs. 3.98%), and EP300 (19.29% vs. 5.97%), had significantly higher mutation 205 206 rates in Chinese PCNSL patients than in DLBCL patients, suggesting genetic diversity of the 207 cancer genome between diseases. Compared with those in PCNSLs in French patients ¹⁴ (Figure 3B), the mutation frequencies of EP300 (19.29% vs. 3.48%) and KMT2D (48.6% vs. 208 22.61%) were significantly higher in the PCNSLs of Chinese patients. Compared with those 209 in the PCNSLs³ of Japanese patients (**Figure 3C**), the mutation frequencies of MYD88 (55%) 210 vs. 85.6%) and BTG2 (38.6% vs. 87.8%) were significantly lower in the PCNSLs of Chinese 211 patients. These data revealed a unique mutational landscape of Chinese PCNSL patients in 212 terms of the frequency of SNVs. Next, the frequencies of recurrent SNAs with the GISTIC2.0 213 program in Chinese PCNSL patients were compared with those in DLBCL patients. At the 214 focal level of SNAs, Chinese PCNSL samples had significant levels of 11g12.4, 18p11.21, 215 19q13.42, 22q11.1, and 22q11.21 deletions along with amplifications in arms 1q21.1, 4p16.3, 216 11q12.2, and 17q12 (Figure 3D). Compared with PCNSLs from Japanese patients, those 217

from Chinese patients also had significant levels of deletions in arms 11q12.1, 18p11.21, 19q13.42, 22q11.1, and 22q11.21 along with amplifications in arms 1q21.1, 4p16.3, 11q12.2, and 17q12 (**Figure 3E**). These results suggest genetic diversity in the PCNSL genome between races.

Owing to the genetic diversity of the PCNSL genome across races and the genetic diversity 222 between the PCNSL and DLBCL cancer genomes, the existing molecular subtyping methods 223 for PCNSL¹⁴ and DLBCL^{9-12,35,36} may not be applicable to Chinese PCNSL patients. We 224 assessed the prognostic value of these published molecular subtypes for overall survival (OS) 225 and progression-free survival (PFS) in our Chinese PCNSL cohort. Actionable PCNSL 226 classification can perfectly differentiate prognoses, thereby improving precision medicine 227 strategies. According to cell-of-origin (COO) molecular subtyping^{35,36}, there were no 228 significant differences in OS (P = 0.42) or PFS (P = 0.67) between the germinal center type 229 (GCB) and non-GCB subtypes (Figure S6A). Similarly, according to LymphPlex molecular 230 subtyping¹², no significant differences in OS (P = 0.68) or PFS (P = 0.71) were observed 231 between the BN2-like, MCD-like, EZB-like, N1-like, TP53, and other subtypes (Figure S6B). 232 233 We further categorized our data on the basis of the DLBCL subtyping (C0-C5) proposed by *M. A. Shipp* et al.¹⁰ Within our cohort, the C4 subtype was associated with shorter OS (*P* = 234 0.009, P = 0.013) and PFS (P = 0.096, P = 0.047) than the C0 and C5 subtypes, respectively 235 (Figure S6C). However, the overall efficacy of this DLBCL subtyping system in predicting 236 outcomes was less than satisfactory (OS, P = 0.074). The data were also analyzed via the 237 five subtypes (BN2, EZB, MCD, N1, and others)⁹ and seven subtypes (A53, BN2, EZB, MCD, 238 N1, ST2, and others)¹¹ of DLBCL proposed by Louis M Staudt et al. There were no significant 239 differences in OS (P > 0.05) or PFS (P > 0.05) between BN2, N1, and other categories within 240 the 5-subtype model (Figure S6D). Overall, the seven-subtype model (Figure S6E) had 241 inferior performance, with no significant differences observed in OS (P = 0.61) or PFS (P = 242 0.47). With respect to the four subtypes (CS1-4) of PCNSL proposed by A Alentorn ¹⁴, our 243 data revealed that the OS (P = 0.12) and PFS (P = 0.54) of patients with the CS3 subtype 244 were shorter than those of patients with the other three subtypes (Figure S6F), but the 245 differences were not significant. Furthermore, the double-hit gene expression signature 246

defines a distinct subgroup of germinal center B-cell-like DLBCL, which is associated with a poor prognosis¹³. However, in this study, there were no significant differences in OS (P > 0.05) or PFS (P > 0.05) between the double- or triple-hit PCNSL patients and the other PCNSL patients, as shown in **Figure S7** (A: all; B: discovery cohort; C: validation cohort). Taken together, these results suggest that the molecular subtyping of PCNSL and DLBCL, as previously published, is not applicable to Chinese PCNSL patients.

253 **PCNSL molecular subtypes with clinical outcome implications**

To establish a molecular classification system for PCNSL, we initially carried out consensus 254 255 clustering on 58 patients (discovery cohort) who received immunochemotherapy (HD-MTXbased chemotherapy regimen) via WES data to generate cluster assignments (refer to 256 Methods - Classification model to identify molecular types). The molecular classification 257 scheme, presented in Figure 4A, Table S3 and Figure S8, was determined independently 258 of clinical information and finalized before the analysis of clinical data, allowing us to analyze 259 the relationships between genetic subtypes and survival in this entire cohort. Finally, EP300, 260 BTG1, MYD88, and IRF4 were included in the molecular subtyping of Chinese PCNSLs. 261

All nonsynonymous mutations in EP300, BTG1, MYD88, and IRF4 were visualized within 262 the functional domains of the encoded proteins (Figure S9). As previously reported^{10,37,38}, 263 264 MYD88 (L265P) was the most prevalent mutation, while EP300, BTG1, and IRF4 all harbored scattered mutations. We evaluated the prognostic significance of these four identified genes 265 for OS and PFS. The survival outcomes of patients with EP300 mutations were significantly 266 poorer than those of patients with wild-type EP300, and patients harboring mutations in either 267 BTG1 (OS, P = 0.017; PFS, P = 0.12), MYD88 (OS, P = 0.016; PFS, P = 0.069), or IRF4 (OS, 268 P = 0.045; PFS, P = 0.018) had favorable OS (Figure S10A) and PFS (Figure S10B) relative 269 to individuals with their wild-type counterparts. 270

In the discovery cohort, we identified three subtypes of PCNSL: BMI (n = 16, 27.59%), E3 (n = 9, 15.52%), and UC (n = 33, 56.90%). The E3 subtype was defined as those with mutations in the EP300 gene. The PCNSLs that carried mutations in at least two of the three genes, BTG1, MYD88, and IRF4, were characterized as the BMI subtype. The other PCNSLs that did not fit into either of these categories were classified under the UC subtype

(Unclassified). The three subtypes differed significantly in OS (P < 0.001) and PFS (P =276 0.043), the BMI subtype (P < 0.001, P = 0.023) had much more favorable outcomes than the 277 E3 and UC subtypes did, and the E3 subtype (P < 0.001, P = 0.037) had far worse outcomes 278 than the BMI and UC subtypes did (Figure 4B). These differences between the three 279 subtypes of PCNSL were detected in the validation cohort (OS, P = 0.020; PFS, P = 0.048; 280 n = 82, tumor-only sample), and patients with E3 had significantly shorter survival times than 281 those with BMI (OS, P = 0.007) or UC (OS, P = 0.043) (Figure 4C). Similar results were 282 observed in the entire cohort (n = 140, OS, P < 0.0001; PFS, P = 0.0004; Figure S10C): 26 283 patients were classified as BMI (18.57%), 27 as E3 (19.29%), and 87 as UC (62.14%). 284 Furthermore, the raw WES or WGS data, along with follow-up data from an independent 285 cohort of 36 Chinese patients with PCNSL¹⁷, were obtained. The three identified subtypes 286 significantly differed in OS (P = 0.0004; Figure 4D). Specifically, the BMI subgroup (P = 287 0.0004, P = 0.005) had notably more favorable outcomes than the E3 and UC subgroups did. 288 Conversely, the E3 subtype (P = 0.0004, P = 0.0231) had considerably worse outcomes than 289 either the BMI or UC subgroup did. These results further validate the robustness of our 290 291 molecular classification.

BTK inhibitors have been shown to improve the prognosis of PCNSL patients by blocking 292 B-cell receptor signaling pathways, which influence the growth and survival of B cells. 293 Consequently, we further analyzed whether the use of BTK inhibitors affects the robustness 294 of our classification. There were no significant differences in OS (P > 0.05, Figure S11A) or 295 PFS (P > 0.05, Figure S11B) between patients who received BTK inhibitors (n = 8) and those 296 who did not receive BTK inhibitors (n = 132). Additionally, after initial treatment options (HD-297 MTX combined with idarubicin (IDA), HD-MTX combined with rituximab (R), HD-MTX 298 299 combined with IDA and R, or a combination of BTK inhibitors) and consolidated therapy (WBRT, stem cell transplant) were adjusted, multivariable Cox regression analysis revealed 300 that the use of different treatment options did not affect the robustness of our classification 301 (Table S4). 302

For the E3 subtype, the mean mutation prevalence in the PCNSLs of Asian patients 303 (16.88%) was significantly higher than that in the PCNSLs of Western patients (5.26%) 304

(Figure 5A). Patients with the E3 subtype had significantly poorer survival, which was not 305 affected by the mutation site (OS, P = 0.82; PFS, P = 0.69; Figure 5B). Univariate Cox 306 regression analysis revealed that PCNSL molecular subtypes are associated with clinical 307 outcomes (Table S5), which was also confirmed by multivariate Cox regression hazard ratio 308 analysis after adjusting for important confounders (age, IELSG, MSKCC, LDH levels, deep 309 brain location, and CSF protein levels) (Figure 4E). These findings further confirmed that 310 PCNSL molecular subtypes are associated with clinical outcomes. The key demographic and 311 312 clinical characteristics of the three MS patients are summarized in Table S6.

We also constructed a Sankey diagram to visualize the correspondence between our samples and the previously described genetic subtypes of PCNSL and DLBCL (**Figure 4F**). The Chinese PCNSL subtypes were significantly distinct from those previously described. Taken together, these results indicate that three unique molecular subtypes of Chinese PCNSL are associated with clinical outcomes.

318 Genomic landscape and tumor microenvironment across PCNSL subtypes

We further aimed to determine whether these three molecular subtypes of Chinese PCNSL 319 patients could affect the mutational landscape, oncogenic pathways and tumor 320 microenvironment. Intergroup mutation distribution analysis was performed to identify unique 321 322 and shared mutations. A total of 6,102 (median variants per sample = 196.5), 41,548 (median variants per sample = 1719), and 73,440 mutated genes (median variants per sample = 918) 323 were identified in the BMI, E3, and UC subtypes, respectively, with 518 mutated genes shared 324 among the three subtypes (Figures 6A and S12A). Key candidate cancer genes in the 140 325 PCNSL patients, grouped by these three subtypes, are presented in **Figure 6B**. The mutation 326 frequencies of EP300, MYD88, IRF4, PCLO, FAT4, KMT2C, KMT2A, MPEG1, PRKDC, 327 SPEN, CARD11, RANBP2, NUMA1, NSD1, PASK, USP6, PTPN13, SLX4, ADGRA2, TET1, 328 and NOTCH2 were significantly different (P< 0.05) among these three subtypes. Arm-level 329 CNAs among the samples of the three subtypes are displayed in Figure 6C, showing 330 significant differences in recurrent deletions (in 6g and 12p) but not in amplifications (Figure 331 **6D**). The results of significant recurrent focal amplifications and deletions were visualized by 332 GISTIC 2.0 in the BMI, E3, and UC subtypes (Figure 12B). There was no significant 333

difference (P > 0.05) in the MATH (**Figure 6E**) or MSI (**Figure 6F**) scores among the three subtypes, but the TMB was greater in the E3 subgroup than in the BMI (P < 0.0001) or UC subgroup (P < 0.001) (**Figure 6G**). Furthermore, the VAF was lower in the E3 subgroup than in the BMI (P < 0.0001) or UC subgroup (P < 0.001) (**Figure 6H**).

The contributions of mutational signatures across the three subtypes are illustrated in 338 **Figure S12C**. The three signatures extracted from the E3 subtype samples displayed cosine 339 similarities of 82.0%, 93.8%, and 95.2% to the COSMIC signatures 6, 10a, and 45, 340 respectively (Figure S12D). In contrast, those extracted from BMI subtype samples 341 presented similarities of 77.3%, 81.7%, and 58.3% to COSMIC signatures 5, 84, and 87, 342 respectively (Figure S12D). The signatures from the UC subtype samples were 80.8%. 343 70.3%, and 91.7% and similar to the COSMIC signatures 5, 1, and 45 (Figure S12D), 344 respectively. 345

Hematoxylin–eosin (H&E) and immunohistochemical (IHC) staining of CD2, CD5, CD10, 346 CD19, CD20, CD79a, Ki-67, BCL2, BCL6, MUM1, c-Myc, and p53 from samples of different 347 PCNSL subtypes revealed significant malignant progression in patients with the E3 subtype 348 349 of PCNSL, as shown in Figure S13A and Table S7. This is evidenced by the formation of poorly differentiated and aggressively growing tumors (Ki-67), which contrasts with the other 350 two subtypes. Immunostaining revealed a greater presence of CD2+, CD5+, CD10+, CD19+, 351 CD20+, and CD79a+ cells in the E3 subtype. Additionally, a greater presence of BCL2+, 352 BCL6+, MUM1+, and c-Myc+ cells, which are markers associated with the diagnosis of B-cell 353 lymphomas³⁹, was observed in the E3 subtype. 354

To explore potential therapeutic strategies for the genetic subtypes of PCNSL, we 355 examined groups of genetic aberrations that target oncogenic signaling pathways (Figure 356 357 **S13B**). Genetic events affecting NF- κ B regulators that negatively regulate the stability of NF- κ B-dependent mRNAs were detected in 85.2% and 84.6% of the E3 and BMI subtype 358 samples but not in the UC subtype samples (P <0.05). These findings suggest that the E3 359 and BMI subtypes may be more responsive to BTK inhibitors⁴⁰. The PI3 kinase pathway, a 360 pathway that can indirectly activate NF- κ B, is genetically altered in 66.7% of E3 patients⁴¹. 361 We also noted a higher frequency of genetic alterations in chromatin histone modifiers (100%), 362

RTK signaling (77.8%), protein homeostasis/ubiquitination (66.7%), cell cycle pathways (77.8%), genome integrity (82.5%), Wnt/B-catenin signaling (59.3%), the chromatin SWI/SNF complex (70.4%), and splicing (37%) in E3 patients. Therefore, a combination of cyclin D-Cdk4,6 and PI3 kinase inhibitors might be beneficial for patients with the E3 subtype.

Taken together, these results suggest that the three molecular subtypes of PCNSL in Chinese patients each have a unique genomic landscape, tumor microenvironment, and oncogenic pathways.

370 Mutational landscape and oncogenic pathways in PCNSLs with different sites of onset

Next, we aimed to determine whether the site of PCNSL onset in Chinese patients influences the mutational landscape and oncogenic pathways. **Figure 7A** shows that the areas most often affected by PCNSL were the basal ganglia, brainstem, cerebellum, corpus callosum, frontal lobe, occipital lobe, parietal lobe, temporal lobe, thalamus, and ventricles, as well as combinations of multiple sites. The most frequently involved sites of PCNSL were the frontal lobe (22.14%), temporal lobe (15.71%), and basal ganglia (8.57%).

First, we found that there was no significant difference in the proportion of different disease 377 sites among the three subtypes (chi-square test, corpus callosum, P = 0.887; basal ganglia, 378 P=0.645; thalamus, P= 0.317; brainstem, P= 1.000; ventricle, P= 0.631; front lobe, P= 0.576; 379 380 multiple, P= 0.949; temporal lobe, P= 0.214; cerebellum, P= 0.831; occipital lobe, P= 0.778; parietal lobe, P= 0.887). BMI-related tumors were slightly more prevalent in the temporal lobe 381 (7 out of 26 patients vs. 15 out of 114 patients, chi-square test P = 0.082) than other subtypes 382 were. Conversely, E3 events were likely to be more frequent in the thalamus (2 out of 27 383 cases vs. 2 out of 113 cases, chi-square test P = 0.113). Larger studies are needed to confirm 384 this observation. 385

Next, we explored whether the tumor sites differed between the discovery and validation cohorts. As shown in **Figure S14A**, there was no significant difference in the proportions of different disease sites between the discovery and validation cohorts (chi-square test P =0.610). Furthermore, the proportions of deep brain lesions in the discovery and validation cohorts were not significantly different (P = 0.478).

391 Sixteen mutated genes were represented among all the sites (Figure 7B). Key candidate

cancer genes in the 140 PCNSL patients, grouped by site of disease onset, are presented in 392 Figure 7C. Mutations in the PIM1, CD79B, and MYD88 genes were relatively common across 393 all PCNSL onset sites. The frontal lobe had the highest mutation frequencies of five genes: 394 PIM1 (51.61%), MYD88 (51.61%), PDE4DIP (48.39%), PCLO (48.39%), and ZFHX3 395 (48.39%). In the temporal lobe, the most prevalent mutations were in PIM1 (68.18%), 396 PDE4DIP (63.64%), PCLO (59.09%), RNF213 (59.09%), and MYD88 (54.55%), with no 397 mutations observed in MET, PIK3CA, or TSC1. In the basal ganglia, the genes with the 398 399 highest mutation frequencies were PIM1 (66.67%), PDE4DIP (58.33%), MYD88 (50%), ZFHX3 (41.67%), and DTX1 (41.67%). The mutation frequencies of LRP1B and KMT2D were 400 significantly different (P< 0.05) among these sites. Arm-level CNAs among the samples of 401 these sites are displayed in Figure 7D, and significant differences in both recurrent 402 amplifications (2g, 10g, and 10p) and deletions (9p) were observed. The difference in the 403 MATH score was significant (P < 0.05) (Figure 7E) across these sites, but there was no 404 significant difference (*P* > 0.05) in the TMB (Figure 7F) or MSI (Figure 7G) score between 405 several sites. 406

407 Next, we further divided the tumor sites into deep brain and shallow brain tumor subgroups. A total of 854 (81%) mutated genes were represented between the deep brain and shallow 408 brain tumor subgroups (Figure S14B). The mutational landscape between deep brain and 409 shallow brain tumors is shown in Figure S14C. The mutation frequencies of LRP1B, FAT1, 410 KMT2B, CIC, PTPRB, MAP3KI, ARID1B, and CPS1 in the deep brain and shallow brain 411 tumors were significantly different (P< 0.05). Arm-level CNAs between deep brain and 412 shallow brain tumors are displayed in Figure S14D, and no significant differences (P > 0.05) 413 in either recurrent amplification or deletion were observed. There was no significant difference 414 415 (P > 0.05) in the MATH (Figure S14E), TMB (Figure S14F), or MSI (Figure S14G) score between deep brain and shallow brain tumors. 416

We further conducted pathway enrichment analysis (**Figure 7H**) and revealed significant involvement of NF- κ B signaling in the tumors of patients with onset in the parietal lobe and brainstem, each with an involvement rate of 100%. We also noted a significant difference (P <0.05) in the frequency of genetic alterations in transcription factors, other chromatin, immune

signaling, NF-κB signaling, chromatin histone modifiers, RTK signaling, PI3K signaling, RNA
abundance, the chromatin SWI/SNF complex, protein homeostasis/ubiquitination, and the
cell cycle among these sites. Furthermore, we noted a significant difference (P <0.05) in the
frequency of genetic alterations in the chromatin SWI/SNF complex but not in the other
pathways (Figure S14H). Taken together, these results indicate that the mutational
landscape in Chinese patients with PCNSL varies between sites of onset.

427

428 Discussion

Owing to the genetic, phenotypic, and tumor microenvironment heterogeneity of PCNSL. 429 identifying classification and prognostic biomarkers for patients is extremely challenging, 430 which in turn makes developing effective new therapies challenging. Here, we performed a 431 study of 140 Chinese PCNSL patients that was adequately powered to expand the genomic 432 landscape and explore its clinical significance utilizing an unprecedented sample size and a 433 multicenter approach. We defined recurrent mutations, CNAs, and associated cancer 434 435 candidate genes in Chinese PCNSL patients and compared these comprehensive genetic signatures with those of PCNSL patients and systemic DLBCL patients of other races, 436 revealing that genetic heterogeneity is a defining feature of PCNSL. Our results highlight the 437 complexity of PCNSLs, which have a median of 19.06 mutations/Mb and a median of 949.5 438 variants per sample. 439

Although the genetic variation profiles of the PCNSL discovery and validation samples were similar, notable differences were evident in the mutation frequencies of certain genes (e.g., NOTCH1, NOTCH2, and KMT2C), the TMB, the MSI, and the MATH score. These discrepancies might stem from variations in specimen types (fresh tissue versus FFPE) and the absence of paired samples in the validation cohort⁴². Furthermore, genetic heterogeneity between patient populations¹⁵ may be the main reason why the published classifications for DLBCL and PCNSL are not applicable to Chinese PCNSL patients.

More importantly, in our study, we identified three robust molecular subtypes of PCNSL in Chinese patients: BMI, E3, and UC. Distinct clinical outcomes, activated cellular pathways, and alterations in the genetic landscape distinguish the three subtypes and could be utilized

to personalize therapy for patients with different subtypes. For the E3 subtype, the mean 450 mutation rate was significantly higher in the PCNSLs of Chinese and Japanese patients 451 (16.88%) than in those of Western PCNSL patients (5.26%). This difference may be attributed 452 to variations in age, lifestyle, and Epstein–Barr virus infection rates^{17,19,43}. Patients with the 453 E3 subtype had significantly poorer survival outcomes, a finding that is consistent with 454 reported findings in DLBCL⁴⁴. EP300 and CREBBP are two closely related members of the 455 KAT3 family of histone acetyltransferases that function similarly⁴⁵. However, in this study, the 456 457 survival patients with EP300 mutations was significantly poorer, whereas CREBBP mutations did not contribute to an inferior prognosis (Figure S4). This discrepancy might stem from the 458 distinct roles of EP300 and CREBBP in PCNSL. Specifically, in a recent study, EP300, but 459 not CREBBP, was reported to play an essential role in supporting the viability of classical 460 Hodgkin's lymphoma by directly modulating the expression of the oncogenic MYC/IRF4 461 network, surface receptor CD30, immunoregulatory cytokine interleukin 10, and immune 462 checkpoint protein PD-L1⁴⁶. The molecular mechanisms through which EP300 mutations 463 facilitate the progression of PCNSL are still not understood. We plan to explore these 464 mechanisms in a forthcoming study. 465

PCNSLs that carried mutations in at least two of three genes, namely, BTG1, MYD88, and 466 IRF4, were characterized as the BMI subtype with favorable survival outcomes. The 467 prognostic significance of MYD88 mutations in PCNSL patients remains inconclusive. While 468 several studies have reported no effect on OS^{47–50}, only two have reported an unfavorable 469 outcome^{51,52}. In line with our findings, Olimpia et al.⁵³ and Claudio et al.⁵⁴ have indicated that 470 the survival time is significantly longer in patients with PCNSLs harboring MYD88 mutations. 471 This discrepancy may be attributed to heterogeneity in genetics, phenotype, and the tumor 472 473 microenvironment across various patient populations; variances in sample size; and the adjustment of confounding variables, including age and treatment methods. In this large study, 474 we found that Chinese PCNSL patients harboring MYD88 mutations have longer survival 475 times. However, further research is needed to verify these findings in diverse populations. 476 The associations of IRF4 and BTG1 mutations with clinical outcomes in PCNSL patients have 477 not been previously reported. This results of this study revealed that mutations in both IRF4 478

and BTG1 were associated with favorable outcomes. IRF4 has been identified as a driver 479 oncogene that transcriptionally regulates downstream target genes, such as MYC, and 480 coordinates the transcriptional program with NF-κB references^{55–58}. Numerous in vitro and 481 clinical studies have highlighted the abnormal overexpression and oncogenic roles of IRF4 482 in various mature lymphoid neoplasms^{58–61}. Therefore, we hypothesized that the reduced 483 expression and/or decreased transcriptional activity of IRF4 resulting from such mutations 484 may explain the association of IRF4 mutations with favorable survival outcomes in PCNSL 485 patients. BTG1 mutation has been associated with extranodal dissemination and unfavorable 486 outcomes in B-cell lymphoma patients^{62–64}. However, in this study, BTG1 mutation was linked 487 to favorable outcomes in PCNSL patients. In a substantial discovery cohort consisting of 533 488 patients with B-cell precursor acute lymphoblastic leukemia, *Blanca* and colleagues⁶⁵ 489 reported that deletions of BTG1 alone did not affect prognosis. The complete remission rate 490 among acute myeloid leukemia patients with low BTG1 expression was significantly greater 491 than that among patients with high BTG1 expression⁶⁶. The molecular mechanisms through 492 which mutations in BTG1 decelerate the progression of PCNSL warrant further research. 493

494 This study has several limitations. First, the proportion of unclassified PCNSLs (UC subtypes) is high, possibly due to the lack of multiomics data needed for classification. 495 However, the three-class classification based on WES proposed in this study has strong 496 clinical practicality and translational value, as it relies solely on the four genes identified 497 through WES. In other words, clinicians can predict patient outcomes and further perform 498 individualized management of PCNSLs by detecting mutations in these four genes without 499 the need for additional multiomics testing. Second, the three identified molecular subtypes 500 can be used to predict patient outcomes. However, it remains challenging to determine 501 502 whether these three molecular subtypes can still be used to accurately distinguish biological subtypes without additional RNA-seq data layers. 503

In summary, by performing WES on tumor specimens from 140 Chinese PCNSL patients, we identified three molecular subtypes that can be used to predict patient outcomes. These genetic signatures associated with PCNSL outcomes illuminate the path toward individualized management of PCNSL patients and the discovery of novel treatment

508 strategies for this challenging disease.

509

510 **METHODS**

511 Sample collection and clinicopathological features of PCNSL patients

In this study, a total of 140 newly diagnosed PCNSL patients were retrospectively enrolled 512 from January 2010 to December 2020 from three independent medical centers, namely, 513 514 Huashan Hospital of Fudan University, Renji Hospital of Shanghai Jiao Tong University, and Shanghai Cancer Center of Fudan University. The study comprised two cohorts: the 515 discovery cohort, which included 58 patients from Huashan Hospital of Fudan University, 516 Shanghai; and the validation cohort (n = 82), which consisted of 25 patients from Huashan 517 Hospital of Fudan University, 51 patients from Renji Hospital of Shanghai Jiao Tong University, 518 and 6 patients from Shanghai Cancer Center of Fudan University. In the discovery cohort, 519 paired blood samples were also obtained for analysis in conjunction with each tumor sample. 520 In the validation cohort, only unpaired FFPE tumor tissue samples were obtained for analysis. 521

None of the patients had received steroid treatment for PCNSL, and PCNSL tumors were 522 diagnosed on the basis of the World Health Organization criteria¹³. All patients underwent a 523 2-deoxy-2[F-18] fluoro-D-glucose positron emission tomography/computed tomography 524 (FDG PET/CT) scan and bone marrow aspiration to exclude systemic tumor manifestation 525 on the basis of the guidelines of the International PCNSL Collaborative Group⁶⁷ and the 526 European Association of Neuro-Oncology⁶⁸. PCNSL patients for whom ocular involvement 527 was present were excluded. Each tumor tissue sample was independently reviewed by at 528 least 2 pathologists to confirm that the tumor sample was histologically consistent with the 529 530 PCNSL. All patients underwent long-term follow-up, which ran to December 30, 2022. The detailed clinical parameters of the 140 PCNSL patients are presented in Table S1 and Figure 531 S1A. 532

533 Ethics statement

This study was conducted with the approval of the Institutional Review Boards at Huashan Hospital of Fudan University (Approval No. 2022-529), Shanghai Cancer Center of Fudan University (Approval No. 1612167-18), and Renji Hospital of Shanghai Jiao Tong University

(Approval No. LY2024-112-C). All participants provided written informed consent before
 participating in the clinical study and before the collection of tumor tissues. The research was
 carried out in strict adherence to the ethical principles of the Declaration of Helsinki and
 conformed to international norms of good clinical practice.

541 **Treatment Regimens**

All PCNSL patients were treated with HD-MTX-based combination immunochemotherapy. The following chemotherapy protocols were used for induction: HD-MTX combined with IDA, HD-MTX combined with R, HD-MTX combined with IDA and R, or a combination of BTK inhibitors. Whole-brain radiation therapy or stem cell transplantation is a form of consolidated therapy. The detailed therapeutic schedule was the same as that previously described⁶⁹.

547 Sample processing

The collection and processing of samples for this study were carried out by the Department 548 of Pathology or Department of Neurosurgery at Huashan Hospital of Fudan University, Renji 549 Hospital of Shanghai Jiao Tong University, and Shanghai Cancer Center of Fudan University. 550 551 In the discovery cohort, surplus tumor tissues and paired blood samples obtained during surgical procedures were collected. To guarantee the guality of these samples, each 552 specimen was promptly and accurately labeled with relevant patient information within 30 553 minutes of collection and then immediately snap-frozen in liquid nitrogen. These samples 554 were stored at -80°C until being shipped to GenomiCare Biotechnology (Shanghai, China). 555 In the validation cohort, the residual tumor tissue specimens were subjected to FFPE. These 556 FFPE samples were transported in a dry ice container to Sinotech Genomics Technologies 557 in Shanghai, China, accompanied by a time and temperature tracker to ensure proper 558 559 monitoring during shipment. Upon arrival, the samples were promptly stored in liquid nitrogen to maintain their integrity until further processing. 560

561 **DNA extraction**

562 For frozen fresh tissue and matched peripheral blood samples, total genomic DNA (gDNA) 563 was extracted via the Maxwell RSC Blood DNA Kit (AS1400, Promega) on a Maxwell RSC 564 system (AS4500, Promega) following the manufacturer's instructions. For FFPE tissue, total 565 gDNA was extracted via the QIAamp DNA FFPE Tissue Kit (56404, Qiagen) following the

566 manufacturer's instructions. The integrity and concentration of the gDNA were determined via

the Qsep100 System (BIOptic, China) and a Qubit 3.0 fluorometer (Thermo Fisher Scientific).

568 The OD₂₆₀ was measured with a NanoDrop One (Thermo Fisher Scientific).

569 WES library preparation and sequencing

For frozen tissue and matched peripheral blood samples, exome DNA was extracted via 570 the SureSelect Human All Exon V7 Kit (5991-9039EN, Agilent). The library was subsequently 571 prepared via the SureSelectXT Low Input Target Enrichment and Library Preparation system 572 (G9703-90000, Agilent). For the FFPE sample, the exome DNA was captured via the 573 574 SureSelect Human All Exon V8 Kit (5191-6874, Agilent) and prepared into a library via the SureSelect XT HS2 DNA Reagent Kit (G9983A, Agilent). The library was validated via the 575 use of an Agilent 2100 Bioanalyzer and a Qubit 3.0 fluorometer. Paired-end 150 bp read 576 sequencing was performed on an Illumina NovaSeg 6000. Image analysis and base calling 577 578 were performed via onboard RTA3 software (Illumina).

579 **Somatic mutation calling**

580 Quality control was conducted via FastQC software. The cleaned and trimmed FASTQ files were aligned to the UCSC human reference genome (hg19) via the Burrows–Wheeler Aligner 581 (BWA) with default parameters. For each paired blood sample, SNVs and short 582 insertions/deletions (INDELs) were identified via Sentieon TNseq⁷⁰ with default parameters. 583 The identified somatic mutations were removed if they did not satisfy any one of the following 584 criteria: a variant allele frequency (VAF) of at least 0.05, support from a minimum of three 585 reads, annotation by the Variant Effect Predictor (VEP) package⁷¹, and transformation into a 586 mutation annotation format (MAF) file for further analysis with maftools⁷². 587

When a paired normal sample was not available, a panel of normal (PON) files was used as a paired control to call the SNVs and InDels. A panel of normal files was created using the reads from clinical blood samples that showed no evidence of tumor contamination from 100 different individuals collected by Genomicare Biotechnology (Shanghai). The reads were retained when they passed the standard WES QC, as described in the WES subsection. The reads for reference alleles and alternative alleles were gathered for each candidate site in all 100 samples, except for false positives, which were those exhibiting a candidate allele

595 frequency exceeding 0.05 in more than 5 samples.

We applied an additional PON mask for all the candidate somatic SNV and INDEL sites to 596 exclude germline mutations, as previously reported^{10,73}. In general, in addition to the standard 597 retention rules, mutations were retained if they were known driver alterations highlighted as 598 being biologically significant in the COSMIC database. Mutations with a tumor mutation 599 frequency \geq 20% were also retained. Other germline mutations were retained if their existence 600 ratio in the Exome Aggregation Consortium (ExAC) exac all was < 0.00005, their existence 601 602 ratio in the Asian population exac eas was < 0.0003, and their existence ratio in previous GenomiCare samples germline gc was < 0.0006. Retained mutations were considered 603 604 somatic mutations.

Thus, the PON file and PON mask limit possible recurrent artifacts of sequencing and the presence of germline mutations in somatic mutations that are falsely selected owing to a lack of paired blood data.

608 Somatic copy number alteration calling and profiling

Using the method described in the ExomeCNV package⁷⁴, we implemented a normalized depth–coverage ratio method to identify CNVs in paired samples. To correct for five potential biases (the size of exonic regions, batch effects, both the quantity and quality of sequencing data, local GC content, and genomic mappability) that affect the raw read counts, we utilized a standard normal distribution model. Genes with a haploid copy number \geq 3 or \leq 1.2 were defined as amplified or deleted, respectively, and a minimum tumor content (purity) of 20% was needed.

616 Significance analysis of recurrent somatic copy number alterations

Sequenza⁷⁵ was used to generate the segment files. The segment files were used as the input for GISTIC2.0⁷⁶ to detect recurrent arm-level and focal peaks in copy number alterations via the parameters -js 40 -conf 0.99 -ta 0.2 -td 0.2 -brlen 0.8, with other parameters set to the defaults.

621 Cancer-related gene filtering

622 After the aforementioned steps of calling SNVs and CNVs, we further filtered the resulting 623 mutated genes by intersecting them with a designated group of genes recognized as cancer-

related genes. This group of genes was sourced from two prominent public cancer gene 624 databases. Part 1 of the group was obtained from the OncoKB curated cancer gene list 625 (https://www.oncokb.org/cancerGenes)⁷⁷. These genes are classified as cancer genes by 626 OncoKB on the basis of their inclusion in various sequencing panels, the Sanger Cancer 627 Gene Census, or the criteria established by *Vogelstein* et al.⁷⁸. Part 2 was derived from the 628 (https://github.com/griffithlab/civic-629 ranked CIViC gene candidates table server/blob/master/public/downloads/RankedCivicGeneCandidates.tsv)⁷⁹. The final list of 630 cancer-related genes consisted of all genes in Part 1 and genes in Part 2 only when its 631 'panel count' value (in the CIViC tsv file listed above) was ≥ 2 . 632

633 **Bioinformatic analysis**

The mutational signature classification was based on all single-base substitution (SBS) 634 signatures of the COSMIC mutational signatures⁸⁰. TMB was defined as the aggregate count 635 of somatic nonsynonymous mutations, including SNVs or INDELs, within the tumor exome 636 for each patient⁸¹. This count was then divided by the overall size of the targeted regions (35) 637 for WES), resulting in the TMB, expressed in counts per megabase (counts/Mb). The MATH 638 score³³ was calculated on the basis of the width of the variant allele frequency (VAF) 639 distribution, utilizing mattools for the analysis⁷². The pathway map was generated via 640 maftools as previously described³⁰. 641

642 Classification model for identifying molecular types

643 CoxNet survival analyses and least absolute shrinkage and selection operator (LASSO) 644 regressions were performed via the scikit-learn package in Python (version 3.9.7). K–M 645 survival curves (log-rank test) were used for OS analysis via the survival and survminer 646 packages in R (version 4.1.2). The molecular classification scheme is detailed as follows and 647 is presented in **Figure 4A**, **Table S3 and Figure S7**.

648 (1) Selection of 24 initial features in the discovery cohort

649 Molecular features derived from SNVs were selected to develop subsequent classification 650 models.

First, to identify genes with high variability for use in clustering, those with low mutation frequencies were filtered out. This process led to the selection of genes with a mutation

653 frequency greater than 5% for clustering purposes. A total of 150 genes were screened for 654 further analysis.

In the second phase, OS was analyzed as the dependent variable, with the aforementioned 655 150 genes serving as independent variables. The feature selection steps were as follows: (1) 656 a comprehensive approach involving CoxNet survival analyses and LASSO regressions was 657 employed to identify characteristic genes; (2) the importance score of each feature in the 658 model was calculated and ranked in descending order; (3) features that ranked in the top 50% 659 660 were recorded as potential features once; (4) processes 1-3 were repeated 500 times, and all potential features were recorded; and (5) among the feature sets obtained in the above 661 process, those recorded more than 250 times were selected. 662

Ultimately, 24 feature genes, namely, EP300, KMT2D, MPEG1, TUSC3, MYD88, PTCH1,
DST, ZNF521, BTG1, NOTCH2, PRDM1, ADGRL3, CDH11, CREBBP, ATRX, TCL1A,
GNA13, ETV6, IRF4, HIST2H3D, CD79B, GRM8, MYH11, and SPTA1, were selected.

666 (2) Differential feature selection

To determine the most suitable features among the initial 24 for the classification model, Kaplan–Meier analyses were conducted. Features qualifying for further analysis met the following criteria: a log-rank test P value of less than 0.2 and any group with more than 5 patients. Ultimately, 6 DEGs, namely, EP300, CREBBP, DST, MYD88, BTG1, and IRF4, were identified.

(3) Model training utilizing random combinations of 6 differential features

To adhere to clinical practice guidelines for predicting patient outcomes and implementing 673 individualized management strategies for PCNSL patients, patients with PCNSL were 674 categorized into three subgroups (those with good, poor, and intermediate prognoses) in the 675 present study. Six differential features were ultimately selected for building a classification 676 model to classify patients into the three subtypes. These features were randomly combined 677 to form three types, yielding a total of 288 combinations. To identify the most appropriate 678 679 classification model among these combinations, Kaplan-Meier analyses were performed. Models were chosen on the basis of a log-rank test P value of less than 0.05 and subgroup 680 sample sizes greater than 8. Ultimately, seven candidate molecular subtype combinations 681

met these criteria: (molecular markers shown in the format of subtype 1 vs. subtype 2 vs.
subtype 3): ① EP300 (predominant) vs. MYD88/BTG1/IRF4≥2 vs. others; ② EP300 vs.
MYD88/BTG1/IRF4≥2 (predominant) vs. others; ③ BTG1 (predominant) vs. CREBP/DST≥1
vs. others; ④ BTG1 vs. CREBP/DST≥1 (predominant) vs. others; ⑤ BTG1/IRF4≥1 vs.
EP300/CREBP/DST≥1 vs. others; ⑥ BTG1/IRF4≥1 vs. EP300/DST≥1 vs. others; and ⑦
CREBP/DST≥1 vs. MYD88/BTG1/IRF4≥2 vs. others.

688 (4) Validation of candidate molecular subtypes

The WES data and OS data of 82 patients in the validation cohort were included to validate 689 690 the candidate molecular subtypes. The mutation frequencies of EP300, CREBBP, DST, MYD88, BTG1, and IRF4 in the validation cohort were greater than 5%. Kaplan-Meier 691 survival analysis was used to validate the seven candidate molecular subtype combinations. 692 As shown in Figure S7E. the performance of [EP300 (predominant) vs. 693 MYD88/BTG1/IRF4≥2 vs. others] (3-way P = 0. 0202) and [EP300 vs. MYD88/BTG1/IRF4≥2 694 vs. others (predominant)] (3-way P = 0. 0346) was better than those of the other five 695 molecular subtypes (3-way P > 0.05 for both). Additionally, compared with the [EP300 vs. 696 MYD88/BTG1/IRF4≥2 vs. others (predominant)] molecular subtype, the Kaplan-Meier 697 survival curve of [EP300 (predominant) vs. MYD88/BTG1/IRF4≥2 vs. others] showed no 698 699 crossover. Finally, the performance of [EP300 (predominant) vs. MYD88/BTG1/IRF4≥2 vs. others] was the best, so we selected it as the molecular subtype of Chinese PCNSL patients. 700 In other words, PCNSL patients with both EP300 and BMI (BTG1, IRF4, and MYD88) variants 701 702 were categorized into the E3 subtype.

703 Hematoxylin & eosin and immunohistochemical staining

PCNSL tumor tissue was fixed with 10% paraformaldehyde and embedded in paraffin. The paraffin-embedded tissues were cut into 4-µm-thick sections. These sections were then subjected to staining with hematoxylin and eosin (H&E) or for immunohistochemistry. For H&E staining, the sections were stained with hematoxylin and eosin following standard protocols (ab245880, Abcam). Immunohistochemistry was performed for ki-67 (ab15580, Abcam), CD2 (ab314761, Abcam), CD5 (ab75877, Abcam), CD10 (ab256494, Abcam), CD19 (ab134114, Abcam), CD20 (ab78237, Abcam), CD79a (ab79414, Abcam), BCL2 (MA5-11757,

Thermo Scientific), BCL6 (PA5-14259, Thermo Scientific), MUM1 (ab247079, Abcam), c-Myc
(MA1-980, Thermo Scientific), and p53 (MA5-14516, Thermo Scientific) via an automated
Leica BOND-III staining system. H&E and immunohistochemistry images were obtained with
a KFBIO scanner (KF-PRO-005-EX, Zhejiang, China) and visualized via KFBIO SlideViewer
software. Each staining was independently reviewed by at least 2 pathologists.

716 In situ hybridization of fluorescence

PCNSL tumor tissue was fixed with 10% paraformaldehyde and embedded in paraffin. The paraffin-embedded tissues were cut into 4-µm-thick sections. Fluorescence in situ hybridization was performed in accordance with standard protocols via the following commercial probes: MYC break-apart, BCL-2 break-apart, and BCL-6 break-apart probes (LBP Medicine Science & Technology Co., Ltd., Guangzhou, China). Staining was independently evaluated by two hematopathologists, and discrepancies were resolved by another hematopathologist.

724 HIV, HBV, and EBV detection

Antibody and antigen levels of HIV were measured via an Elecscy HIV combi PT assay kit (Roche, Germany) with a Cobas e 601 analyzer (Roche). EBV-specific PCR was performed via an EBV virus nucleic acid test kit (Sansure Biotech, China) with an Applied Biosystems 7500 Real-Time PCR System (Thermo Scientific). HBV-specific PCR was performed via an HBV virus nucleic acid test kit (Sansure Biotech, China) with an Applied Biosystems 730 Real-Time PCR System (Thermo Scientific).

731

732 Statistical analysis

All the statistical analyses were performed via R version 4.1.2 and GraphPad Prism version
9. The specific statistical analyses employed for each figure are detailed in the respective
legends.

We used PASS software to calculate the sample size for our study, with a significance level of 0.05 and a power of 0.80. The observed group proportions were 0.6, 0.2, and 0.2, whereas the expected proportions were 0.33 for each group. The sample size required to meet these criteria was 25 participants. Consequently, the sample size utilized in this study is deemed adequate.

Whole-exome sequencing data were processed via nf-core/sarek v3.4.2⁸² of the nf-core collection of workflows⁸³, which uses reproducible software environments from the Bioconda⁸⁴ and Biocontainers⁸⁵ projects. The pipeline was executed with Nextflow v24.04.2⁸⁶. The software tools and their versions are displayed in **Table S8**.

745

746 Data and code availability

The raw sequence data have been deposited in the Genome Sequence Archive⁸⁷ in the National Genomics Data Center⁸⁸, China National Center for Bioinformation/Beijing Institute of Genomics, Chinese Academy of Sciences (GSA-Human HRA006122), which is publicly accessible at <u>https://ngdc.cncb.ac.cn/gsa-human</u>. Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

The raw sequencing data of the DLBCL WES data are accessible from the TCGA database 752 and were reanalyzed via pipelines and filter settings that were identical to those used in the 753 present study. The raw WES data of the Japanese PCNSL patients were obtained from the 754 Japanese Genotype–Phenotype Archive (JGA, http://trace.ddbj.nig.ac.jp/jga) and reanalyzed 755 via pipelines and filtering settings that were identical to those used in the present study, which 756 is hosted by DDBJ under the accession number JGAS0000000021³. The raw sequencing 757 WES/WGS data of an external independent Chinese PCNSL cohort (n = 36)¹⁷ were obtained 758 from the China National Center for Bioinformation/Beijing Institute of Genomics, Chinese 759 760 Academy of Sciences and reanalyzed via pipelines and filtering settings that were identical to those used in the present study, under the accession number GSA-Human HRA002475. 761

The frequencies of recurrent genetic alterations in DLBCLs²⁷ and French PCNSLs¹⁴ are accessible from the manuscript or supplemental information and were not reanalyzed via identical pipelines and filtering settings as those used in the present study.

765

Acknowledgment: This study was funded by the National Natural Science Foundation of China (82302582), the Shanghai Municipal Health Commission Project (20224Y0317), the Shanghai Science and Technology Commission (17430750200), the Join Breakthrough Project for New Frontier Technologies of Shanghai Hospital Development Center

- (SHDC12016120), and the Youth Medical Talents–Clinical Laboratory Practitioner Program(2022-65).
- 772
- 773 **Competing Interests:** The authors declare no potential conflicts of interest.
- 774

Author contributions: SJ. L, CX. L, DH. L, WJ. C, ZG. X, XT. Z, and Y. M conceived and designed the project. SJ. L, DH. L, WJ. C, JZ. C, J. R, and ZG. X collected the clinical samples. SJ. L, JN. W, JZ. C, LY. Z, HW. Y, and Y. S analyzed the WES data and performed bioinformatic analyses. SJ. L, DH. L, J. R, LY. Z, HW. Y, and JN. W integrated the sequencing data, drew the display items. SJ. L, DH. L, and CX. L wrote the manuscript. XT. Z and Y. M oversaw the ethical guidelines and data regulation. WJ. C and Y. M supervised the project. All of the authors contributed to the final version of the paper.

782

783 **References**

- Chen, T. *et al.* Evidence-based expert consensus on the management of primary central nervous system
 Iymphoma in China. *J Hematol Oncol* **15**, 136 (2022).
- 786 2. Ferreri, A. J. M. *et al.* Primary central nervous system lymphoma. *Nat. Rev. Dis. Primers* 9, 29 (2023).
- 787 3. Fukumura, K. *et al.* Genomic characterization of primary central nervous system lymphoma. *Acta* 788 *Neuropathol. (Berl.)* 131, 865–875 (2016).
- Braggio, E. *et al.* Genome-Wide Analysis Uncovers Novel Recurrent Alterations in Primary Central Nervous
 System Lymphomas. *Clin. Cancer Res.* 21, 3986–3994 (2015).
- 791 5. Radke, J. *et al.* The genomic and transcriptional landscape of primary central nervous system lymphoma.
 792 *Nat. Commun.* 13, 2558 (2022).
- 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).
- 795 7. Rosenwald, A. et al. The use of molecular profiling to predict survival after chemotherapy for diffuse large-

- 796 B-cell lymphoma. *N. Engl. J. Med.* **346**, 1937–1947 (2002).
- 8. Wilson, W. H. *et al.* Targeting B cell receptor signaling with ibrutinib in diffuse large B cell lymphoma. *Nat.*
- 798 *Med.* **21**, 922–926 (2015).
- 9. Schmitz, R. *et al.* Genetics and Pathogenesis of Diffuse Large B-Cell Lymphoma. *N. Engl. J. Med.* **378**,
- 800 1396–1407 (2018).
- 10. 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).
- 11. 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.e14 (2020).
- 805 12. Shen, R. *et al.* Simplified algorithm for genetic subtyping in diffuse large B-cell lymphoma. *Signal*806 *Transduct. Target. Ther.* 8, 145 (2023).
- 807 13. Swerdlow, S. H. *et al.* The 2016 revision of the World Health Organization classification of lymphoid
 808 neoplasms. *Blood* 127, 2375–2390 (2016).
- 14. Hernández-Verdin, I. *et al.* Molecular and clinical diversity in primary central nervous system lymphoma.
- 810 Ann. Oncol. **34**, 186–199 (2023).
- 15. Yuan, X. et al. Analysis of the genomic landscape of primary central nervous system lymphoma using
- whole-genome sequencing in Chinese patients. *Front. Med.* **17**, 889–906 (2023).
- 813 16. He, X. *et al.* Analysis of genomic alterations in primary central nervous system lymphoma. *Medicine*814 (*Baltimore*) 102, e34931 (2023).
- 17. Zhu, Q. et al. Whole-Genome/Exome Sequencing Uncovers Mutations and Copy Number Variations in
- Primary Diffuse Large B-Cell Lymphoma of the Central Nervous System. *Front Genet* **13**, 878618 (2022).
- 18. Gandhi, M. K. et al. EBV-associated primary CNS lymphoma occurring after immunosuppression is a

- 818 distinct immunobiological entity. *Blood* **137**, 1468–1477 (2021).
- 19. Kaulen, L. D. *et al.* Integrated genetic analyses of immunodeficiency-associated Epstein-Barr virus- (EBV)
- positive primary CNS lymphomas. *Acta Neuropathol* **146**, 499–514 (2023).
- 821 20. Ren, W. et al. Genetic landscape of hepatitis B virus-associated diffuse large B-cell lymphoma. Blood 131,
- 822 2670–2681 (2018).
- 21. Berhan, A., Bayleyegn, B. & Getaneh, Z. HIV/AIDS Associated Lymphoma: Review. *Blood Lymphat Cancer*
- 824 **12**, 31–45 (2022).
- 22. Pasqualucci, L. et al. Analysis of the coding genome of diffuse large B-cell lymphoma. Nat Genet 43, 830–
- 826 837 (2011).
- 827 23. Morin, R. D. *et al.* Frequent mutation of histone-modifying genes in non-Hodgkin lymphoma. *Nature* **476**,
- 828 298–303 (2011).
- 829 24. Lohr, J. G. *et al.* Discovery and prioritization of somatic mutations in diffuse large B-cell lymphoma (DLBCL)
- by whole-exome sequencing. *Proc Natl Acad Sci U S A* **109**, 3879–3884 (2012).
- 831 25. Morin, R. D. et al. Mutational and structural analysis of diffuse large B-cell lymphoma using whole-
- 832 genome sequencing. *Blood* **122**, 1256–1265 (2013).
- 833 26. de Miranda, N. F. C. C. et al. Exome sequencing reveals novel mutation targets in diffuse large B-cell
- 834 Iymphomas derived from Chinese patients. *Blood* **124**, 2544–2553 (2014).
- 27. Reddy, A. *et al.* Genetic and Functional Drivers of Diffuse Large B Cell Lymphoma. *Cell* **171**, 481-494.e15
- 836 (2017).
- 837 28. Bruno, A. *et al.* Mutational analysis of primary central nervous system lymphoma. *Oncotarget* 5, 5065–
 838 5075 (2014).
- 839 29. Chapuy, B. et al. Targetable genetic features of primary testicular and primary central nervous system

- 840 lymphomas. *Blood* **127**, 869–881 (2016).
- 30. Sanchez-Vega, F. *et al.* Oncogenic Signaling Pathways in The Cancer Genome Atlas. *Cell* **173**, 321337.e10 (2018).
- 843 31. Alexandrov, L. B. et al. Signatures of mutational processes in human cancer. Nature 500, 415–421 (2013).
- 32. Rizvi, N. A. et al. Cancer immunology. Mutational landscape determines sensitivity to PD-1 blockade in
- 845 non-small cell lung cancer. *Science* **348**, 124–128 (2015).
- 33. Mroz, E. A. & Rocco, J. W. MATH, a novel measure of intratumor genetic heterogeneity, is high in poor-
- 847 outcome classes of head and neck squamous cell carcinoma. *Oral Oncol* **49**, 211–215 (2013).
- 848 34. Yamamoto, H. *et al.* Microsatellite instability: A 2024 update. *Cancer Sci* **115**, 1738–1748 (2024).
- 35. Alizadeh, A. A. et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling.
- 850 *Nature* **403**, 503–511 (2000).
- 36. 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).
- 853 37. Phelan, J. D. *et al.* A multiprotein supercomplex controlling oncogenic signalling in lymphoma. *Nature*854 560, 387–391 (2018).
- 38. Ngo, V. N. *et al.* Oncogenically active MYD88 mutations in human lymphoma. *Nature* **470**, 115–119 (2011).
- 39. van Imhoff, G. W. et al. Prognostic impact of germinal center-associated proteins and chromosomal
- breakpoints in poor-risk diffuse large B-cell lymphoma. J Clin Oncol 24, 4135–4142 (2006).
- 40. Iwasaki, H. *et al.* The IkB kinase complex regulates the stability of cytokine-encoding mRNA induced by
- TLR-IL-1R by controlling degradation of regnase-1. *Nat Immunol* **12**, 1167–1175 (2011).
- 41. Kloo, B. et al. Critical role of PI3K signaling for NF-kappaB-dependent survival in a subset of activated B-
- 861 cell-like diffuse large B-cell lymphoma cells. *Proc Natl Acad Sci U S A* **108**, 272–277 (2011).

- 42. Spencer, D. H. et al. Comparison of clinical targeted next-generation sequence data from formalin-fixed
- and fresh-frozen tissue specimens. *J Mol Diagn* **15**, 623–633 (2013).
- 43. Yokoyama, A. *et al.* Age-related remodelling of oesophageal epithelia by mutated cancer drivers. *Nature*
- **565**, 312–317 (2019).
- 44. Huang, Y.-H. et al. CREBBP/EP300 mutations promoted tumor progression in diffuse large B-cell
- 867 lymphoma through altering tumor-associated macrophage polarization via FBXW7-NOTCH-CCL2/CSF1
- axis. *Signal Transduct Target Ther* **6**, 10 (2021).
- 45. Meyer, S. N. et al. Unique and Shared Epigenetic Programs of the CREBBP and EP300 Acetyltransferases
- 870 in Germinal Center B Cells Reveal Targetable Dependencies in Lymphoma. *Immunity* **51**, 535-547.e9
- 871 (2019).
- 46. Wei, W. *et al.* Analysis and therapeutic targeting of the EP300 and CREBBP acetyltransferases in anaplastic
- large cell lymphoma and Hodgkin lymphoma. *Leukemia* **37**, 396–407 (2023).
- 47. Nayyar, N. et al. MYD88 L265P mutation and CDKN2A loss are early mutational events in primary central
- 875 nervous system diffuse large B-cell lymphomas. *Blood Adv* **3**, 375–383 (2019).
- 48. Zhou, Y. et al. Analysis of Genomic Alteration in Primary Central Nervous System Lymphoma and the
- 877 Expression of Some Related Genes. *Neoplasia* **20**, 1059–1069 (2018).
- 49. Zheng, M. et al. Frequency of MYD88 and CD79B mutations, and MGMT methylation in primary central
- 879 nervous system diffuse large B-cell lymphoma. *Neuropathology* **37**, 509–516 (2017).
- 50. Yamada, S., Ishida, Y., Matsuno, A. & Yamazaki, K. Primary diffuse large B-cell lymphomas of central
- 881 nervous system exhibit remarkably high prevalence of oncogenic MYD88 and CD79B mutations. *Leuk*
- 882 *Lymphoma* **56**, 2141–2145 (2015).
- 51. Takano, S. *et al.* MyD88 Mutation in Elderly Predicts Poor Prognosis in Primary Central Nervous System

- Lymphoma: Multi-Institutional Analysis. *World Neurosurg* **112**, e69–e73 (2018).
- 885 52. Hattori, K. et al. MYD88 (L265P) mutation is associated with an unfavourable outcome of primary central
- 886 nervous system lymphoma. *Br J Haematol* **177**, 492–494 (2017).
- 53. Curran, O. E. et al. MYD88 L265P mutation in primary central nervous system lymphoma is associated
- 888 with better survival: A single-center experience. *Neurooncol Adv* **3**, vdab090 (2021).
- 54. Agostinelli, C. *et al.* Genomic Profiling of Primary Diffuse Large B-Cell Lymphoma of the Central Nervous
- 890 System Suggests Novel Potential Therapeutic Targets. *Mod Pathol* **36**, 100323 (2023).
- 891 55. Wong, R. W. J. et al. Feed-forward regulatory loop driven by IRF4 and NF-κB in adult T-cell
- 892 leukemia/lymphoma. *Blood* **135**, 934–947 (2020).
- 893 56. Tsuboi, K. et al. MUM1/IRF4 expression as a frequent event in mature lymphoid malignancies. Leukemia
- **14**, 449–456 (2000).
- 57. Nakagawa, M. *et al.* Targeting the HTLV-I-Regulated BATF3/IRF4 Transcriptional Network in Adult T Cell
 Leukemia/Lymphoma. *Cancer Cell* 34, 286-297.e10 (2018).
- 897 58. Weilemann, A. *et al.* Essential role of IRF4 and MYC signaling for survival of anaplastic large cell lymphoma.
 898 *Blood* 125, 124–132 (2015).
- 59. Shaffer, A. L. *et al.* IRF4 addiction in multiple myeloma. *Nature* **454**, 226–231 (2008).

900 60. Amanda, S. *et al.* IRF4 drives clonal evolution and lineage choice in a zebrafish model of T-cell lymphoma.

- 901 *Nat Commun* **13**, 2420 (2022).
- 902 61. Yang, Y. *et al.* Exploiting synthetic lethality for the therapy of ABC diffuse large B cell lymphoma. *Cancer*903 *Cell* **21**, 723–737 (2012).
- 904 62. Delage, L. et al. BTG1 inactivation drives lymphomagenesis and promotes lymphoma dissemination
- 905 through activation of BCAR1. *Blood* **141**, 1209–1220 (2023).

- 906 63. Mlynarczyk, C. *et al.* BTG1 mutation yields supercompetitive B cells primed for malignant transformation.
 907 *Science* **379**, eabj7412 (2023).
- 908 64. Tijchon, E. *et al.* Tumor suppressors BTG1 and BTG2 regulate early mouse B-cell development.
 909 *Haematologica* 101, e272-276 (2016).
- 910 65. Scheijen, B. et al. Tumor suppressors BTG1 and IKZF1 cooperate during mouse leukemia development
- 911 and increase relapse risk in B-cell precursor acute lymphoblastic leukemia patients. *Haematologica* 102,
 912 541–551 (2017).
- 913 66. Li, Y. et al. The predictive value of BTG1 for the response of newly diagnosed acute myeloid leukemia to
- 914 decitabine. *Clin Epigenetics* **16**, 16 (2024).
- 915 67. Abrey, L. E. *et al.* Report of an international workshop to standardize baseline evaluation and response
 916 criteria for primary CNS lymphoma. *J Clin Oncol* 23, 5034–5043 (2005).
- 917 68. Hoang-Xuan, K. *et al.* Diagnosis and treatment of primary CNS lymphoma in immunocompetent patients:
- guidelines from the European Association for Neuro-Oncology. *Lancet Oncol* **16**, e322-332 (2015).
- 919 69. Li, Q. et al. Improvement of outcomes of an escalated high-dose methotrexate-based regimen for
- 920 patients with newly diagnosed primary central nervous system lymphoma: a real-world cohort study.
- 921 *Cancer Manag Res* **13**, 6115–6122 (2021).
- 922 70. Freed, D., Aldana, R., Weber, J. A. & Edwards, J. S. The Sentieon Genomics Tools-A fast and accurate
- 923 solution to variant calling from next-generation sequence data. *BioRxiv* 115717 (2017).
- 924 71. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol* **17**, 122 (2016).
- 925 72. Mayakonda, A., Lin, D.-C., Assenov, Y., Plass, C. & Koeffler, H. P. Maftools: efficient and comprehensive
- analysis of somatic variants in cancer. *Genome Res* 28, 1747–1756 (2018).
- 927 73. Wang, Y. et al. Comprehensive identification of somatic nucleotide variants in human brain tissue.

- 928 *Genome Biol* **22**, 92 (2021).
- 929 74. Sathirapongsasuti, J. F. et al. Exome sequencing-based copy-number variation and loss of heterozygosity
- 930 detection: ExomeCNV. *Bioinformatics* **27**, 2648–2654 (2011).
- 931 75. Favero, F. *et al.* Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data.
- 932 Ann Oncol **26**, 64–70 (2015).
- 933 76. Mermel, C. H. *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic
- 934 copy-number alteration in human cancers. *Genome Biol* **12**, R41 (2011).
- 935 77. Chakravarty, D. et al. OncoKB: A Precision Oncology Knowledge Base. JCO Precis Oncol 2017,
- 936 PO.17.00011 (2017).
- 937 78. Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **339**, 1546–1558 (2013).
- 938 79. Griffith, M. et al. CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation
- 939 of variants in cancer. *Nat Genet* **49**, 170–174 (2017).
- 80. Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101
- 941 (2020).
- 81. Chalmers, Z. R. *et al.* Analysis of 100,000 human cancer genomes reveals the landscape of tumor
 mutational burden. *Genome Med* 9, 34 (2017).
- 82. Garcia, M. *et al.* Sarek: A portable workflow for whole-genome sequencing analysis of germline and
 somatic variants. *F1000Res* 9, 63 (2020).
- 83. Ewels, P. A. *et al.* The nf-core framework for community-curated bioinformatics pipelines. *Nat Biotechnol*947 38, 276–278 (2020).
- 948 84. Grüning, B. et al. Bioconda: sustainable and comprehensive software distribution for the life sciences. Nat
- 949 *Methods* **15**, 475–476 (2018).

- 950 85. da Veiga Leprevost, F. et al. BioContainers: an open-source and community-driven framework for
- 951 software standardization. *Bioinformatics* **33**, 2580–2582 (2017).
- 952 86. P, D. T. *et al.* Nextflow enables reproducible computational workflows. *Nature biotechnology* **35**, (2017).
- 953 87. Chen, T. et al. The Genome Sequence Archive Family: Toward Explosive Data Growth and Diverse Data
- 954 Types. *Genomics Proteomics Bioinformatics* **19**, 578–583 (2021).
- 955 88. CNCB-NGDC Members and Partners. Database Resources of the National Genomics Data Center, China
- 956 National Center for Bioinformation in 2022. *Nucleic Acids Res* **50**, D27–D38 (2022).

All rights reserved. No reuse allowed without permission.



38

959 Figure 1. Study design and the PCNSL mutational landscape in Chinese patients

- A: Study design, including workflow and data composition for the discovery cohort and
 validation cohort.
- 962 B: Number and frequency of recurrent mutations. Gene–sample matrix of recurrently mutated
- genes ranked by mutation frequency (n = 140). The total mutation density across the cohort
- 964 is displayed at the top, with the variant allele fraction and trinucleotides at the bottom.
- 965 C: GISTIC2.0 results of significant recurrent focal amplifications (red) and deletions (blue).
- 966 Genes affected by each focal event are annotated (n = 140). X-axis: plot of chromosomes;
- 967 **Y-axis: G score**.





970 Figure 2. Comparison of mutation profiles between the discovery cohort and validation

971 cohort of PCNSL patients

- 972 A: Principal component analysis of whole-exome sequencing (WES) data between the
- 973 discovery cohort and validation cohort of PCNSL patients
- 974 B: Mutation rate of WES data between the discovery cohort and validation cohort of PCNSL
- 975 patients
- 976 C: A mirror bar plot showing the frequencies of genetic alterations in the discovery cohort of
- 977 PCNSL patients compared with those in the validation cohort.
- 978 D: Arm-level copy number alterations of the discovery cohort and validation cohort samples
- are displayed. The frequencies of amplifications and deletions were compared.
- 980 E: Comparison of the tumor mutational burden (TMB) between the discovery cohort and the
- validation cohort of PCNSL patients
- 982 F: The levels of mutant-allele tumor heterogeneity (MATH) in the discovery cohort and the 983 validation cohort of PCNSL patients were compared.
- 984 G: The levels of microsatellite instability (MSI) in the discovery cohort and the validation 985 cohort of PCNSL patients were compared.
- 986 H: The variant allele frequency (VAF) in the discovery cohort and the validation cohort of 987 PCNSL patients were compared.
- 988 The discovery cohort included blood-paired fresh tumor tissue samples (n = 58), and the
- validation cohort included unpaired FFPE tumor tissue samples (n = 82). Independent-
- samples t tests and chi-squared tests were used. *P < 0.05; **P < 0.01; ***P < 0.001.

All rights reserved. No reuse allowed without permission.



992 Figure 3. Mutation profiles of Chinese PCNSLs, DLBCLs, and PCNSLs of other races

- 993 A: A mirror bar plot showing the frequencies of genetic alterations between Chinese PCNSL
- patients (n = 140) and DLBCL patients (n = 955).
- B: A mirror bar plot showing the frequencies of genetic alterations between Chinese PCNSL
- patients (n = 140) and French PCNSL patients (n = 115).
- 997 C: A mirror bar plot showing the frequencies of genetic alterations between Chinese PCNSL
- patients (n = 140) and Japanese PCNSL patients (n = 41).
- 999 D: GISTIC-defined recurrent copy number focal deletions (blue) and gains (red) as mirror
- plots in DLBCL (TCGA, n = 46) and Chinese PCNSLs (n = 140) are shown. X-axis: plot of chromosomes; Y-axis: G score.
- E: The GISTIC2.0-defined recurrent copy number focal deletions (blue) and gains (red) in Japanese PCNSL patients (n = 41) and Chinese PCNSL patients (n = 140) are shown as mirror plots. Y-axis: plot of chromosomes; X-axis: G score.
- 1005 The chi-squared test was used. *P < 0.05. In the figure legend of A-C, 'Inferred' means that 1006 the mutation types were not directly provided but were calculated from the supplemental 1007 tables of the corresponding paper; thus, the results are 'Inferred', and the *in facto* mutation 1008 number will not be less than the inferred value.

All rights reserved. No reuse allowed without permission.



1010 Figure 4. Integrated genetic drivers reveal PCNSL molecular subtypes with clinical

1011 outcome implications

- 1012 A: Schematic of the typing strategies used to identify PCNSL molecular subtypes in the
- discovery cohort (n = 58) and in the validation cohort (n = 82).
- 1014 B: Kaplan-Meier estimates of overall survival (OS) and progression-free survival (PFS)
- among patients belonging to each molecular subtype in the discovery cohort.
- 1016 C: Kaplan-Meier estimates of overall survival (OS) and progression-free survival (PFS)
- among patients belonging to each molecular subtype in the validation cohort.
- 1018 D: Kaplan-Meier estimates of overall survival (OS) among patients belonging to each
- 1019 molecular subtype in an independent external cohort (n=36).
- 1020 E: Hazard ratio estimates of overall survival in PCNSL patients (n = 140).
- 1021 F: Sankey diagram illustrating the relationships between our PCNSL molecular subtypes and
- 1022 the published DLBCL and PCNSL molecular subtypes.
- 1023
- 1024
- 1025





1026

Figure 5. Mutation frequency among races and prognosis of EP300 1027

- A: The prevalence of EP300 mutations in PCNSL patients of different races. 1028
- B: Kaplan-Meier analyses for comparisons of the overall survival (OS) and progression-free 1029
- survival (PFS) of patients with mutations within and those with mutations outside the EP300 1030
- domain. 1031
- 1032



1034 Figure 6. Mutation profiles across three molecular subtypes of PCNSL

- 1035 A: Venn diagram showing unique and shared mutations among the three identified PCNSL 1036 subtypes.
- 1037 B: Number and frequency of recurrent mutations and the gene–sample matrix of recurrently
- 1038 mutated genes among the three PCNSL subtypes. The relative abundance across the 1039 molecular subtypes is displayed on the right.
- 1040 C: Arm-level copy number alterations among the three molecular subtype samples are 1041 displayed.
- 1042 D: The frequencies of amplifications and deletions were compared among the three 1043 molecular subtype samples.
- 1044 E: The levels of mutant-allele tumor heterogeneity (MATH) were compared among the three 1045 molecular subtype samples.
- 1046 F: The levels of microsatellite instability (MSI) were compared among the three MS samples.
- 1047 G: The tumor mutational burden (TMB) was compared among the three MS samples.
- 1048 H: The levels of variant allele frequency (VAF) were compared among the three MS samples.
- Independent-samples t tests and chi-squared tests were used. *P < 0.05; **P < 0.01; ***P <
- 1050 **0.001; ns** *P* > **0.05**.



1053 Figure 7. Mutational profiles across different sites of onset in PCNSL

- 1054 A: The number and percentage of PCNSL patients with various molecular subtypes at 1055 different disease onset sites.
- 1056 B: Venn diagram of unique and shared mutations among PCNSLs with different sites of onset.
- 1057 C: The number and frequency of recurrent mutations, along with a gene–sample matrix of
- recurrently mutated genes in PCNSL patients at different sites of onset, are presented. The relative abundance across the molecular subtypes is displayed on the right.
- 1060 D: Arm-level copy number alterations across different sites of onset are displayed. The 1061 frequencies of amplifications and deletions were compared.
- 1062 E: The levels of mutant-allele tumor heterogeneity (MATH) were compared among the 1063 samples representing different sites of onset.
- 1064 F: Comparisons of the tumor mutational burden (TMB) among the different sites of onset 1065 samples were performed.
- 1066 G: The levels of microsatellite instability (MSI) were compared among the samples 1067 representing different sites of onset.
- 1068 H: Pathways affected by oncogenes in PCNSL patients with different sites of onset.
- 1069 The chi-squared test and one-way ANOVA were used. *P < 0.05.