Genomic Landscape and Molecular Subtypes of Primary Central Nervous System Lymphoma

Shengjie Li[1,2,3,4,5,6*#], Danhui Li[7#], Zuguang Xia[8,9#], Jianing Wu[6], Jun Ren[6], Yingzhu Li[6], Jiazhen Cao[8,10], Ying Sun[11], Liyang Zhang[12], Hongwei Ye[12], Xingtao Zhou[5*], Chengxun Li[13*], Wenjun Cao[6*], Ying Mao[1,2,3,4*]

1. Department of Neurosurgery, Huashan Hospital, Department of Clinical Laboratory, Eye & ENT Hospital, Shanghai Medical College, Fudan University, Shanghai, China

2. National Center for Neurological Disorders, Shanghai, China

3. Shanghai Key Laboratory of Brain Function Restoration and Neural Regeneration, Shanghai, China

4. Neurosurgical Institute of Fudan University, Shanghai, China

5. Eye Institute and Department of Ophthalmology, Eye & ENT Hospital, Fudan University, Shanghai, China

6. Department of Clinical Laboratory, Eye & ENT Hospital, Shanghai Medical College, Fudan University, Shanghai, China

7. Department of Pathology, RenJi Hospital, School of Medicine, Shanghai JiaoTong University, Shanghai, 200127, China

8. Department of Medical Oncology, Fudan University Shanghai Cancer Center, Fudan University, Shanghai, China

9. Department of Oncology, Shanghai Medical College, Fudan University, Shanghai, China

10. Department of Clinical Laboratory, Fudan University Shanghai Cancer Center, Fudan University, Shanghai, China

11. GenomiCare Biotechnology (Shanghai) Co. Ltd., Shanghai, China

12. Guangzhou KingMed Center for Clinical Laboratory Co. Ltd., Guangzhou, China

13.Institutes of Biomedical Sciences, Fudan University, Shanghai, 200032, China

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

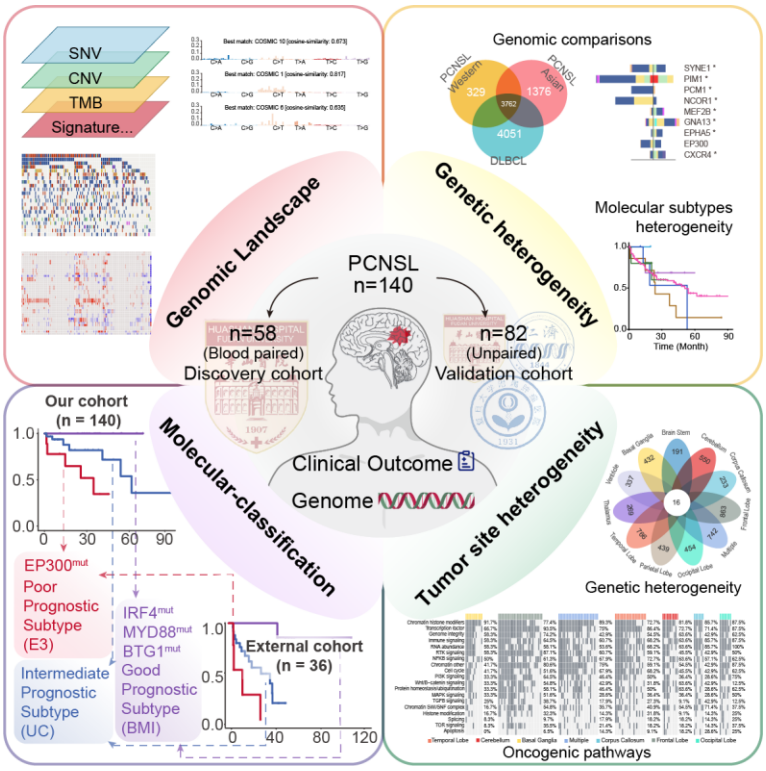
Corresponding Authors:

Wenjun Cao, Department of Clinical Laboratory, Eye & ENT Hospital, Fudan University, No. 83 Fenyang Road, Shanghai, China; Electronic address: wgkjyk@aliyun.com.

Chengxun Li, Institutes of Biomedical Sciences, Fudan University, Shanghai, 200032, China Electronic address: licx21@m.fudan.edu.cn

Xingtao, Zhou, Eye Institute and Department of Ophthalmology, Eye & ENT Hospital, Fudan University, Shanghai, 200031, China. Electronic address: doctzhouxingtao@163.com.

Ying Mao, Department of Neurosurgery, Huashan Hospital, Fudan University, No. 12 Wulumuqi Road, Shanghai, 200040, China; Electronic address: maoying@fudan.edu.cn



Highlights

1. In this study, the genomic landscape of 140 Chinese patients with primary central nervous system lymphomas (PCNSLs) was evaluated.
2. Chinese PCNSL patients have a defining genetic signature that differs from that of both 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.
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 specific mutation site.

Abstract

Primary central nervous system lymphoma (PCNSL) is a rare and aggressive brain tumor with a poor prognosis and almost exclusively comprises diffuse large B-cell lymphoma (DLBCL). Its genetic characteristics and molecular subtypes in Chinese patients remain poorly understood, which in turn makes developing effective new therapies challenging. In our study, 140 Chinese patients with PCNSL that was newly diagnosed at one of three tertiary care centers and who underwent extensive follow-up were included. With this sample, we performed a genomic study aimed at expanding the genomic landscape and identifying new 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 (n = 58) and validated (n = 82) three prominent genetic subtypes related to different clinical and molecular features of PCNSL and further confirmed them in an independent external Chinese PCNSL cohort (n = 36). We called these BMIs (from the co-occurrence of mutations in two genes among BTG1, MYD88, and IRF4), which are associated with favorable outcomes; E3s (so-called EP300 mutations), which are associated with unfavorable outcomes; and UCs (unclassified, without characteristic mutations). Importantly, EP300 was mutated in more PCNSLs from Asian patients (16.88%) than from Western patients (< 5.26%), resulting in unfavorable outcomes independent of the specific mutation site. Our analysis comprehensively reveals the genomic landscape of PCNSL in Chinese patients and emphasizes the clinical value of molecular classification for improving precision medicine strategies.

Introduction

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

Pathologically, nearly 95% of PCNSLs are diffuse large B-cell lymphomas (DLBCLs). Recently, DLBCL has been categorized into various molecular clusters on the basis of genomic sequencing and RNA sequencing, including quartet categorization by *L.M. Staudt*'s team[9], quintet categorization by *M.A. Shipp*'s team[10], *L.M. Staudt*'s team's refined septet method[11], and the LymphPlex classification proposed by *W.L. Zhao*'s team[12]. However, PCNSL has been proven to be a biological entity that is molecularly distinct from DLBCL[13], so these molecular clusters of DLBCL may not be applicable to PCNSL. Notably, the integration of genome-wide data from multiomic studies by *A. Alentorn* et al.[14] showed four molecular patterns of PCNSL using only gene expression data. These patterns have a distinctive prognostic impact, providing a basis for future clinical stratification and subtype-based targeted interventions. Owing to the molecular heterogeneity between various ethnic groups of PCNSL patients[3,15], the applicability of the proposed classifications for PCNSL to other populations requires further confirmation. No effective biomarker or molecular classification scheme exists to tailor therapies for individual PCNSL patients. Furthermore, the issue of PCNSL molecular heterogeneity in Chinese patients has not been adequately addressed, primarily because the data have come from small, single-center studies of Chinese PCNSL patients[16,17].

To address these issues, we conducted whole-exome sequencing (WES) on specimens obtained from 140 patients with newly diagnosed PCNSL who were seen at one of three 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-based chemotherapy regimen, and all 140 patients underwent long-term follow-up. This representative and clinically annotated PCNSL cohort was utilized to comprehensively detect mutations. Additionally, to facilitate routine clinical implementation, we identified three molecular subtypes composed only of single-nucleotide variations (SNVs); these SNVs were associated with different outcomes, and this was confirmed in the validation cohort and in an independent external Chinese PCNSL cohort[17]. Our results comprehensively reveal the genomic landscape of PCNSLs in Chinese patients and provide a clinically actionable PCNSL classification system, thereby improving precision medicine strategies.

Results

PCNSL genomic landscape

Using WES, we detected mutations in 140 patients who were newly diagnosed with PCNSL at one of three cancer centers; in 59% of cases, blood samples were lacking. Because Epstein–Barr virus (EBV)-positive PCNSL is recognized as an immunodeficiency-associated PCNSL in the current WHO CNS5 classification[18,19], all participants in the study who tested negative for EBV were included. Additionally, DLBCL patients with hepatitis B virus (HBV) and HIV infection may have different mutational profiles[20,21]; thus, all participants in the study who tested negative for HIV and HBV were included. The key demographic and clinical 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 variation events in the 140 PCNSL samples analyzed (median = 19.06 mutations/Mb, range = 0.24–91.2 mutations/Mb; median number of variants = 949.5/sample) (**Figure S2A, Figure S3A**).

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

131 (31.4%), SETD1B (31.4%), MPEG1 (29.3%), CREBBP (27.9%), ANKRD11 (27.1%), KAT6B
132 (27.1%), BTG1 (25%), and PTEN (25%) (**Figure 1B**), which are involved in chromatin histone
133 modification, BCR-TLR-mediated NF-κB signaling, immune signaling, the cell cycle, PI3
134 kinase signaling, MAPK signaling, and Wnt/β-catenin signaling (**Figure S2B**)[3,5,9–11,22–29]. In
135 this study, we also identified several other pathways in PCNSL, such as pathways related to
136 genome integrity, RTK signaling, RNA abundance, TGFB signaling, TOR signaling, and
137 apoptosis (**Figure S2B**), many of which have defined roles in other cancers[30]. The mutually
138 exclusive or cooccurring relationships of these genes are shown in **Figure S3D**. Survival
139 associated with candidate cancer genes with mutation frequencies greater than 15% plus the
140 NOTCH2 and EZH2 genes[9] is shown in **Figure S4**.

141 In the search for focal copy number alterations (CNAs) in the 140 PCNSLs, we detected
142 significant recurrent amplifications at chromosomal locations 1q21.1, 4q16.3, 11p15.5,
143 11q12.2, 17q12, 19p13.3, 22q11.1, Yq11.221, and Yq12. We also detected deletions at 11q11,
144 11q12.1, 14q32.33, 17q11.2, 18p11.21, 19p12, 19q13.42, and 22q11.21 (**Figure 1C**). The
145 arm-level CNAs among all 140 PCNSL samples are displayed in **Figure S2C**. The pattern of
146 somatic mutations caused by the different mutational processes in the genome, termed
147 *mutational signatures*, was calculated and compared to the well-established signatures in
148 COSMIC[31]. The mutational signature contributions are shown in **Figure S2D**. The COSMIC
149 5, COSMIC 45, and COSMIC 1 signatures contributed the most to the PCNSL genome
150 (**Figure S2E**). These factors are associated with aging, tobacco smoking, NER deficiency,
151 the regulation of oxidative DNA damage repair, and the spontaneous deamination of 5-
152 methylcytosine. Taken together, these results establish a comprehensive genomic landscape
153 of PCNSL in Chinese individuals.

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

155 Owing to differences in sample type between the discovery and validation cohorts (fresh
156 tumor [discovery cohort] vs. formalin-fixed paraffin-embedded [FFPE] [validation cohort]) and
157 the presence or absence of matched samples (paired [discovery cohort] vs. tumor-only
158 [validation cohort]), we applied further filtering and recalibration steps to reduce false-positive
159 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 key demographic and clinical characteristics of the patients in the discovery and validation cohorts are summarized in **Table S2**. We identified 12,222 genetic variation events (median = 3.84 mutations/Mb, median variants = 159/sample, **Figure S3B, Figure S5A**) in the analyzed discovery cohort of PCNSL samples and 103,716 genetic variation events (median = 6.85 mutations/Mb, median variants = 1597/sample, **Figure S3C, Figure S5B**) in the analyzed validation cohort of PCNSL samples. The genetic variation profiles of the discovery 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 PCNSL samples (**Figure 2B**).

PIM1, MYD88, CD79B, and KMT2D were the most frequently mutated candidate cancer genes of PCNSL. PIM1 (70.7%), MYD88 (63.8%), CD79B (51.7%), and KMT2D (39.7%) (**Figure S5C**) were identified in the discovery cohort, and PIM1 (56.1%), MYD88 (48.8%), CD79B (41.5%), and KMT2D (54.6%) were also observed at similar mutation frequencies in the validation cohort (**Figure S5D**). However, several of the less frequently mutated candidate cancer genes in PCNSL whose mutation frequencies were significantly different when the discovery and validation cohorts were compared included ZFHX3, TRRAP, SZT2, RNF213, RANBP2, PRKDC, PDE4DIP, PCLO, NUMA1, NOTCH1, NOTCH2, KMT2C, KMT2B, KAT6B, FLT3, FAT4, FAT1, and ANKRD11 (all $P < 0.05$) (**Figure 2C**). The arm-level CNAs between the discovery and validation cohort of PCNSL samples were mostly similar and are displayed in **Figure 2D**. Significant differences in both recurrent amplifications (1q, 8q, and 12p) and deletions (6p, 8p, 19p, and 19q) were observed. The results of significant recurrent focal amplifications and deletions were visualized via GISTIC 2.0 in the discovery (**Figure S5E**) 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 treatments[32]. The mutant-allele tumor heterogeneity (MATH) score reflects intratumor heterogeneity[33]. Microsatellite instability (MSI) arises from the impaired function of DNA mismatch repair proteins, and its molecular characteristics are regarded as significant genetic markers[34].

Although the TMB (**Figure 2E**), MATH score (**Figure 2F**), and MSI (**Figure 2G**) were greater in the validation cohort than in the discovery cohort ($P < 0.001$), the mean values were relatively close. Furthermore, the variant allele frequency (VAF) values were greater in the discovery cohort than in the validation cohort ($P < 0.001$, **Figure 2H**). The 3 signatures extracted from the discovery samples presented cosine similarities of 86.7%, 93.1% and 94.4% to COSMIC signatures 6, 45, and 10a, respectively (**Figure S5G**), whereas those extracted from the validation samples presented similarities of 80.4%, 70.9%, and 77.9% to COSMIC 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.

The unique genomic landscape of PCNSLs in Chinese patients

To date, only single-center studies with small sample sizes have reported heterogeneity in the genomic landscape among different populations of Chinese PCNSL patients[15]. Thus, in this large, multicenter study, we further compared the genetic landscape of our cohort with previously published PCNSL/DLBCL data. The frequencies of recurrent genetic alterations in PCNSL patients were compared with those in DLBCL patients[27]. Mirror bar plots (**Figure 3A**) revealed that a set of genes, such as PIM1 (62.1% vs. 16.88%), MYD88 (55% vs. 18.01%), IRF4 (15% vs. 3.98%), and EP300 (19.29% vs. 5.97%), had significantly higher mutation rates in Chinese PCNSL patients than in DLBCL patients, suggesting genetic diversity of the 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. 22.61%) were significantly higher in the PCNSLs of Chinese patients. Compared with those in the PCNSLs[3] of Japanese patients (**Figure 3C**), the mutation frequencies of MYD88 (55% vs. 85.6%) and BTG2 (38.6% vs. 87.8%) were significantly lower in the PCNSLs of Chinese patients. These data revealed a unique mutational landscape of Chinese PCNSL patients in terms of the frequency of SNVs. Next, the frequencies of recurrent SNAs with the GISTIC2.0 program in Chinese PCNSL patients were compared with those in DLBCL patients. At the focal level of SNAs, Chinese PCNSL samples had significant levels of 11q12.4, 18p11.21, 19q13.42, 22q11.1, and 22q11.21 deletions along with amplifications in arms 1q21.1, 4p16.3, 11q12.2, and 17q12 (**Figure 3D**). Compared with PCNSLs from Japanese patients, those

218 from Chinese patients also had significant levels of deletions in arms 11q12.1, 18p11.21,
219 19q13.42, 22q11.1, and 22q11.21 along with amplifications in arms 1q21.1, 4p16.3, 11q12.2,
220 and 17q12 (**Figure 3E**). These results suggest genetic diversity in the PCNSL genome
221 between races.

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

247 defines a distinct subgroup of germinal center B-cell-like DLBCL, which is associated with a
248 poor prognosis[13]. However, in this study, there were no significant differences in OS (P > 0.05)
249 or PFS (P > 0.05) between the double- or triple-hit PCNSL patients and the other PCNSL
250 patients, as shown in **Figure S7** (A: all; B: discovery cohort; C: validation cohort). Taken
251 together, these results suggest that the molecular subtyping of PCNSL and DLBCL, as
252 previously published, is not applicable to Chinese PCNSL patients.

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

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

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

271 In the discovery cohort, we identified three subtypes of PCNSL: BMI (n = 16, 27.59%), E3
272 (n = 9, 15.52%), and UC (n = 33, 56.90%). The E3 subtype was defined as those with
273 mutations in the EP300 gene. The PCNSLs that carried mutations in at least two of the three
274 genes, BTG1, MYD88, and IRF4, were characterized as the BMI subtype. The other PCNSLs
275 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 ($P = 0.043$), the BMI subtype ($P < 0.001$, $P = 0.023$) had much more favorable outcomes than the E3 and UC subtypes did, and the E3 subtype ($P < 0.001$, $P = 0.037$) had far worse outcomes than the BMI and UC subtypes did (**Figure 4B**). These differences between the three subtypes of PCNSL were detected in the validation cohort (OS, $P = 0.020$; PFS, $P = 0.048$; n = 82, tumor-only sample), and patients with E3 had significantly shorter survival times than those with BMI (OS, $P = 0.007$) or UC (OS, $P = 0.043$) (**Figure 4C**). Similar results were observed in the entire cohort (n = 140, OS, $P < 0.0001$; PFS, $P = 0.0004$; **Figure S10C**): 26 patients were classified as BMI (18.57%), 27 as E3 (19.29%), and 87 as UC (62.14%). Furthermore, the raw WES or WGS data, along with follow-up data from an independent cohort of 36 Chinese patients with PCNSL[17], were obtained. The three identified subtypes significantly differed in OS (P = 0.0004; **Figure 4D**). Specifically, the BMI subgroup (P = 0.0004, P = 0.005) had notably more favorable outcomes than the E3 and UC subgroups did. Conversely, the E3 subtype (P = 0.0004, P = 0.0231) had considerably worse outcomes than either the BMI or UC subgroup did. These results further validate the robustness of our molecular classification.

BTK inhibitors have been shown to improve the prognosis of PCNSL patients by blocking B-cell receptor signaling pathways, which influence the growth and survival of B cells. Consequently, we further analyzed whether the use of BTK inhibitors affects the robustness of our classification. There were no significant differences in OS (P > 0.05, **Figure S11A**) or PFS (P > 0.05, **Figure S11B**) between patients who received BTK inhibitors (n = 8) and those who did not receive BTK inhibitors (n = 132). Additionally, after initial treatment options (HD-MTX combined with idarubicin (IDA), HD-MTX combined with rituximab (R), HD-MTX 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 that the use of different treatment options did not affect the robustness of our classification (**Table S4**).

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

305 (**Figure 5A**). Patients with the E3 subtype had significantly poorer survival, which was not
306 affected by the mutation site (OS, P = 0.82; PFS, P = 0.69; **Figure 5B**). Univariate Cox
307 regression analysis revealed that PCNSL molecular subtypes are associated with clinical
308 outcomes (**Table S5**), which was also confirmed by multivariate Cox regression hazard ratio
309 analysis after adjusting for important confounders (age, IELSG, MSKCC, LDH levels, deep
310 brain location, and CSF protein levels) (**Figure 4E)**. These findings further confirmed that
311 PCNSL molecular subtypes are associated with clinical outcomes. The key demographic and
312 clinical characteristics of the three MS patients are summarized in **Table S6**.
313 We also constructed a Sankey diagram to visualize the correspondence between our
314 samples and the previously described genetic subtypes of PCNSL and DLBCL (**Figure 4F**).
315 The Chinese PCNSL subtypes were significantly distinct from those previously described.
316 Taken together, these results indicate that three unique molecular subtypes of Chinese
317 PCNSL are associated with clinical outcomes.

318 **Genomic landscape and tumor microenvironment across PCNSL subtypes**

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

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 **Figure S12C**. The three signatures extracted from the E3 subtype samples displayed cosine similarities of 82.0%, 93.8%, and 95.2% to the COSMIC signatures 6, 10a, and 45, respectively (**Figure S12D**). In contrast, those extracted from BMI subtype samples presented similarities of 77.3%, 81.7%, and 58.3% to COSMIC signatures 5, 84, and 87, respectively (**Figure S12D**). The signatures from the UC subtype samples were 80.8%, 70.3%, and 91.7% and similar to the COSMIC signatures 5, 1, and 45 (**Figure S12D**), respectively.

Hematoxylin–eosin (H&E) and immunohistochemical (IHC) staining of CD2, CD5, CD10, CD19, CD20, CD79a, Ki-67, BCL2, BCL6, MUM1, c-Myc, and p53 from samples of different PCNSL subtypes revealed significant malignant progression in patients with the E3 subtype 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 two subtypes. Immunostaining revealed a greater presence of CD2+, CD5+, CD10+, CD19+, CD20+, and CD79a+ cells in the E3 subtype. Additionally, a greater presence of BCL2+, BCL6+, MUM1+, and c-Myc+ cells, which are markers associated with the diagnosis of B-cell lymphomas[39], was observed in the E3 subtype.

To explore potential therapeutic strategies for the genetic subtypes of PCNSL, we examined groups of genetic aberrations that target oncogenic signaling pathways (**Figure S13B**). Genetic events affecting NF-κB regulators that negatively regulate the stability of NF-κB-dependent mRNAs were detected in 85.2% and 84.6% of the E3 and BMI subtype samples but not in the UC subtype samples ($P < 0.05$). These findings suggest that the E3 and BMI subtypes may be more responsive to BTK inhibitors[40]. The PI3 kinase pathway, a pathway that can indirectly activate NF-κB, is genetically altered in 66.7% of E3 patients[41]. We also noted a higher frequency of genetic alterations in chromatin histone modifiers (100%),

363 RTK signaling (77.8%), protein homeostasis/ubiquitination (66.7%), cell cycle pathways
364 (77.8%), genome integrity (82.5%), Wnt/B-catenin signaling (59.3%), the chromatin SWI/SNF
365 complex (70.4%), and splicing (37%) in E3 patients. Therefore, a combination of cyclin D-
366 Cdk4,6 and PI3 kinase inhibitors might be beneficial for patients with the E3 subtype.

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

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

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

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

386 Next, we explored whether the tumor sites differed between the discovery and validation
387 cohorts. As shown in **Figure S14A**, there was no significant difference in the proportions of
388 different disease sites between the discovery and validation cohorts (chi-square test P =
389 0.610). Furthermore, the proportions of deep brain lesions in the discovery and validation
390 cohorts were not significantly different (P = 0.478).

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

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

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

417 We further conducted pathway enrichment analysis (**Figure 7H**) and revealed significant
418 involvement of NF-κB signaling in the tumors of patients with onset in the parietal lobe and
419 brainstem, each with an involvement rate of 100%. We also noted a significant difference (P
420 <0.05) in the frequency of genetic alterations in transcription factors, other chromatin, immune

421 signaling, NF-κB signaling, chromatin histone modifiers, RTK signaling, PI3K signaling, RNA
422 abundance, the chromatin SWI/SNF complex, protein homeostasis/ubiquitination, and the
423 cell cycle among these sites. Furthermore, we noted a significant difference ($P < 0.05$) in the
424 frequency of genetic alterations in the chromatin SWI/SNF complex but not in the other
425 pathways (**Figure S14H**). Taken together, these results indicate that the mutational
426 landscape in Chinese patients with PCNSL varies between sites of onset.
427

428 **Discussion**

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

440 Although the genetic variation profiles of the PCNSL discovery and validation samples were
441 similar, notable differences were evident in the mutation frequencies of certain genes (e.g.,
442 NOTCH1, NOTCH2, and KMT2C), the TMB, the MSI, and the MATH score. These
443 discrepancies might stem from variations in specimen types (fresh tissue versus FFPE) and
444 the absence of paired samples in the validation cohort[42]. Furthermore, genetic heterogeneity
445 between patient populations[15] may be the main reason why the published classifications for
446 DLBCL and PCNSL are not applicable to Chinese PCNSL patients.

447 More importantly, in our study, we identified three robust molecular subtypes of PCNSL in
448 Chinese patients: BMI, E3, and UC. Distinct clinical outcomes, activated cellular pathways,
449 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 mutation rate was significantly higher in the PCNSLs of Chinese and Japanese patients (16.88%) than in those of Western PCNSL patients (5.26%). This difference may be attributed to variations in age, lifestyle, and Epstein–Barr virus infection rates[17,19,43]. Patients with the E3 subtype had significantly poorer survival outcomes, a finding that is consistent with reported findings in DLBCL[44]. EP300 and CREBBP are two closely related members of the KAT3 family of histone acetyltransferases that function similarly[45]. However, in this study, the 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 distinct roles of EP300 and CREBBP in PCNSL. Specifically, in a recent study, EP300, but not CREBBP, was reported to play an essential role in supporting the viability of classical Hodgkin's lymphoma by directly modulating the expression of the oncogenic MYC/IRF4 network, surface receptor CD30, immunoregulatory cytokine interleukin 10, and immune checkpoint protein PD-L1[46]. The molecular mechanisms through which EP300 mutations facilitate the progression of PCNSL are still not understood. We plan to explore these mechanisms in a forthcoming study.

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

479 and BTG1 were associated with favorable outcomes. IRF4 has been identified as a driver
480 oncogene that transcriptionally regulates downstream target genes, such as MYC, and
481 coordinates the transcriptional program with NF-κB references[55–58]. Numerous in vitro and
482 clinical studies have highlighted the abnormal overexpression and oncogenic roles of IRF4
483 in various mature lymphoid neoplasms[58–61]. Therefore, we hypothesized that the reduced
484 expression and/or decreased transcriptional activity of IRF4 resulting from such mutations
485 may explain the association of IRF4 mutations with favorable survival outcomes in PCNSL
486 patients. BTG1 mutation has been associated with extranodal dissemination and unfavorable
487 outcomes in B-cell lymphoma patients[62–64]. However, in this study, BTG1 mutation was linked
488 to favorable outcomes in PCNSL patients. In a substantial discovery cohort consisting of 533
489 patients with B-cell precursor acute lymphoblastic leukemia, *Blanca* and colleagues[65]
490 reported that deletions of BTG1 alone did not affect prognosis. The complete remission rate
491 among acute myeloid leukemia patients with low BTG1 expression was significantly greater
492 than that among patients with high BTG1 expression[66]. The molecular mechanisms through
493 which mutations in BTG1 decelerate the progression of PCNSL warrant further research.

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

504 In summary, by performing WES on tumor specimens from 140 Chinese PCNSL patients,
505 we identified three molecular subtypes that can be used to predict patient outcomes. These
506 genetic signatures associated with PCNSL outcomes illuminate the path toward
507 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**

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

522 None of the patients had received steroid treatment for PCNSL, and PCNSL tumors were
523 diagnosed on the basis of the World Health Organization criteria[13]. All patients underwent a
524 2-deoxy-2[F-18] fluoro-D-glucose positron emission tomography/computed tomography
525 (FDG PET/CT) scan and bone marrow aspiration to exclude systemic tumor manifestation
526 on the basis of the guidelines of the International PCNSL Collaborative Group[67] and the
527 European Association of Neuro-Oncology[68]. PCNSL patients for whom ocular involvement
528 was present were excluded. Each tumor tissue sample was independently reviewed by at
529 least 2 pathologists to confirm that the tumor sample was histologically consistent with the
530 PCNSL. All patients underwent long-term follow-up, which ran to December 30, 2022. The
531 detailed clinical parameters of the 140 PCNSL patients are presented in **Table S1 and Figure**
532 **S1A**.

533 **Ethics statement**

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

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

541 **Treatment Regimens**

542 All PCNSL patients were treated with HD-MTX-based combination immunochemotherapy.
543 The following chemotherapy protocols were used for induction: HD-MTX combined with IDA,
544 HD-MTX combined with R, HD-MTX combined with IDA and R, or a combination of BTK
545 inhibitors. Whole-brain radiation therapy or stem cell transplantation is a form of consolidated
546 therapy. The detailed therapeutic schedule was the same as that previously described[69].

547 **Sample processing**

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

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
567 the Qsep100 System (BIOptic, China) and a Qubit 3.0 fluorometer (Thermo Fisher Scientific).
568 The OD_{260} was measured with a NanoDrop One (Thermo Fisher Scientific).

569 **WES library preparation and sequencing**

570 For frozen tissue and matched peripheral blood samples, exome DNA was extracted via
571 the SureSelect Human All Exon V7 Kit (5991-9039EN, Agilent). The library was subsequently
572 prepared via the SureSelectXT Low Input Target Enrichment and Library Preparation system
573 (G9703-90000, Agilent). For the FFPE sample, the exome DNA was captured via the
574 SureSelect Human All Exon V8 Kit (5191-6874, Agilent) and prepared into a library via the
575 SureSelect XT HS2 DNA Reagent Kit (G9983A, Agilent). The library was validated via the
576 use of an Agilent 2100 Bioanalyzer and a Qubit 3.0 fluorometer. Paired-end 150 bp read
577 sequencing was performed on an Illumina NovaSeq 6000. Image analysis and base calling
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
581 were aligned to the UCSC human reference genome (hg19) via the Burrows–Wheeler Aligner
582 (BWA) with default parameters. For each paired blood sample, SNVs and short
583 insertions/deletions (INDELs) were identified via Sentieon TNseq[70] with default parameters.
584 The identified somatic mutations were removed if they did not satisfy any one of the following
585 criteria: a variant allele frequency (VAF) of at least 0.05, support from a minimum of three
586 reads, annotation by the Variant Effect Predictor (VEP) package[71], and transformation into a
587 mutation annotation format (MAF) file for further analysis with maftools[72].

588 When a paired normal sample was not available, a panel of normal (PON) files was used
589 as a paired control to call the SNVs and InDels. A panel of normal files was created using the
590 reads from clinical blood samples that showed no evidence of tumor contamination from 100
591 different individuals collected by Genomicare Biotechnology (Shanghai). The reads were
592 retained when they passed the standard WES QC, as described in the WES subsection. The
593 reads for reference alleles and alternative alleles were gathered for each candidate site in all
594 100 samples, except for false positives, which were those exhibiting a candidate allele

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 exclude germline mutations, as previously reported[10,73]. In general, in addition to the standard retention rules, mutations were retained if they were known driver alterations highlighted as being biologically significant in the COSMIC database. Mutations with a tumor mutation frequency ≥20% were also retained. Other germline mutations were retained if their existence ratio in the Exome Aggregation Consortium (ExAC) exac_all was < 0.00005, their existence 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 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.

Somatic copy number alteration calling and profiling

Using the method described in the ExomeCNV package[74], 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 ≥3 or ≤1.2 were defined as amplified or deleted, respectively, and a minimum tumor content (purity) of 20% was needed.

Significance analysis of recurrent somatic copy number alterations

Sequenza[75] was used to generate the segment files. The segment files were used as the input for GISTIC2.0[76] 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.

Cancer-related gene filtering

After the aforementioned steps of calling SNVs and CNVs, we further filtered the resulting 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 databases. Part 1 of the group was obtained from the OncoKB curated cancer gene list (https://www.oncokb.org/cancerGenes)[77]. These genes are classified as cancer genes by OncoKB on the basis of their inclusion in various sequencing panels, the Sanger Cancer Gene Census, or the criteria established by *Vogelstein* et al.[78]. Part 2 was derived from the ranked CIViC gene candidates table (https://github.com/griffithlab/civic-server/blob/master/public/downloads/RankedCivicGeneCandidates.tsv)[79]. The final list of cancer-related genes consisted of all genes in Part 1 and genes in Part 2 only when its 'panel_count' value (in the CIViC tsv file listed above) was ≥ 2.

Bioinformatic analysis

The mutational signature classification was based on all single-base substitution (SBS) signatures of the COSMIC mutational signatures[80]. TMB was defined as the aggregate count of somatic nonsynonymous mutations, including SNVs or INDELs, within the tumor exome for each patient[81]. This count was then divided by the overall size of the targeted regions (35 for WES), resulting in the TMB, expressed in counts per megabase (counts/Mb). The MATH score[33] was calculated on the basis of the width of the variant allele frequency (VAF) distribution, utilizing maftools for the analysis[72]. The pathway map was generated via maftools as previously described[30].

Classification model for identifying molecular types

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

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

Molecular features derived from SNVs were selected to develop subsequent classification 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

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

In the second phase, OS was analyzed as the dependent variable, with the aforementioned 150 genes serving as independent variables. The feature selection steps were as follows: (1) a comprehensive approach involving CoxNet survival analyses and LASSO regressions was employed to identify characteristic genes; (2) the importance score of each feature in the model was calculated and ranked in descending order; (3) features that ranked in the top 50% 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 process, those recorded more than 250 times were selected.

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.

(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 individualized management strategies for PCNSL patients, patients with PCNSL were categorized into three subgroups (those with good, poor, and intermediate prognoses) in the present study. Six differential features were ultimately selected for building a classification model to classify patients into the three subtypes. These features were randomly combined to form three types, yielding a total of 288 combinations. To identify the most appropriate 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 sample sizes greater than 8. Ultimately, seven candidate molecular subtype combinations

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.

(4) Validation of candidate molecular subtypes

The WES data and OS data of 82 patients in the validation cohort were included to validate 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 survival analysis was used to validate the seven candidate molecular subtype combinations. As shown in **Figure S7E**, the performance of [EP300 (predominant) vs. MYD88/BTG1/IRF4≥2 vs. others] (3-way P = 0. 0202) and [EP300 vs. MYD88/BTG1/IRF4≥2 vs. others (predominant)] (3-way P = 0. 0346) was better than those of the other five molecular subtypes (3-way P > 0.05 for both). Additionally, compared with the [EP300 vs. MYD88/BTG1/IRF4≥2 vs. others (predominant)] molecular subtype, the Kaplan–Meier survival curve of [EP300 (predominant) vs. MYD88/BTG1/IRF4≥2 vs. others] showed no 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. In other words, PCNSL patients with both EP300 and BMI (BTG1, IRF4, and MYD88) variants were categorized into the E3 subtype.

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.

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.

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 7500 Real-Time PCR System (Thermo Scientific).

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 v3.4.2[82] of the nf-core collection of workflows[83], which uses reproducible software environments from the Bioconda[84] and Biocontainers[85] projects. The pipeline was executed with Nextflow v24.04.2[86]. The software tools and their versions are displayed in **Table S8**.

Data and code availability

The raw sequence data have been deposited in the Genome Sequence Archive[87] in the National Genomics Data Center[88], China National Center for Bioinformation/Beijing Institute of Genomics, Chinese Academy of Sciences (GSA-Human HRA006122), which is publicly accessible at https://ngdc.cncb.ac.cn/gsa-human. 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 and were reanalyzed via pipelines and filter settings that were identical to those used in the present study. The raw WES data of the Japanese PCNSL patients were obtained from the Japanese Genotype–Phenotype Archive (JGA, http://trace.ddbj.nig.ac.jp/jga) and reanalyzed via pipelines and filtering settings that were identical to those used in the present study, which is hosted by DDBJ under the accession number JGAS00000000021[3]. The raw sequencing WES/WGS data of an external independent Chinese PCNSL cohort (n = 36)[17] were obtained from the China National Center for Bioinformation/Beijing Institute of Genomics, Chinese 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.

The frequencies of recurrent genetic alterations in DLBCLs[27] and French PCNSLs[14] 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.

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).

Competing Interests: The authors declare no potential conflicts of interest.

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.

References

1. Chen, T. *et al.* Evidence-based expert consensus on the management of primary central nervous system lymphoma in China. *J Hematol Oncol* **15**, 136 (2022).
2. Ferreri, A. J. M. *et al.* Primary central nervous system lymphoma. *Nat. Rev. Dis. Primers* **9**, 29 (2023).
3. Fukumura, K. *et al.* Genomic characterization of primary central nervous system lymphoma. *Acta Neuropathol. (Berl.)* **131**, 865–875 (2016).
4. Braggio, E. *et al.* Genome-Wide Analysis Uncovers Novel Recurrent Alterations in Primary Central Nervous System Lymphomas. *Clin. Cancer Res.* **21**, 3986–3994 (2015).
5. Radke, J. *et al.* The genomic and transcriptional landscape of primary central nervous system lymphoma. *Nat. Commun.* **13**, 2558 (2022).
6. 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).
7. Rosenwald, A. *et al.* The use of molecular profiling to predict survival after chemotherapy for diffuse large-

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. Med.* **21**, 922–926 (2015).
9. Schmitz, R. *et al.* Genetics and Pathogenesis of Diffuse Large B-Cell Lymphoma. *N. Engl. J. Med.* **378**, 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).
12. Shen, R. *et al.* Simplified algorithm for genetic subtyping in diffuse large B-cell lymphoma. *Signal Transduct. Target. Ther.* **8**, 145 (2023).
13. Swerdlow, S. H. *et al.* The 2016 revision of the World Health Organization classification of lymphoid neoplasms. *Blood* **127**, 2375–2390 (2016).
14. Hernández-Verdin, I. *et al.* Molecular and clinical diversity in primary central nervous system lymphoma. *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).
16. He, X. *et al.* Analysis of genomic alterations in primary central nervous system lymphoma. *Medicine (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).

819 19. Kaulen, L. D. *et al.* Integrated genetic analyses of immunodeficiency-associated Epstein-Barr virus- (EBV)
820 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).

823 21. Berhan, A., Bayleyegn, B. & Getaneh, Z. HIV/AIDS Associated Lymphoma: Review. *Blood Lymphat Cancer*
824 **12**, 31–45 (2022).

825 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)
830 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 lymphomas derived from Chinese patients. *Blood* **124**, 2544–2553 (2014).

835 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).

841 30. Sanchez-Vega, F. *et al.* Oncogenic Signaling Pathways in The Cancer Genome Atlas. *Cell* **173**, 321-

842 337.e10 (2018).

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

844 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).

846 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).

849 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).

851 36. Hans, C. P. *et al.* Confirmation of the molecular classification of diffuse large B-cell lymphoma by

852 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).

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

856 39. van Imhoff, G. W. *et al.* Prognostic impact of germinal center-associated proteins and chromosomal

857 breakpoints in poor-risk diffuse large B-cell lymphoma. *J Clin Oncol* **24**, 4135–4142 (2006).

858 40. Iwasaki, H. *et al.* The IκB kinase complex regulates the stability of cytokine-encoding mRNA induced by

859 TLR-IL-1R by controlling degradation of regnase-1. *Nat Immunol* **12**, 1167–1175 (2011).

860 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).

862 42. Spencer, D. H. *et al.* Comparison of clinical targeted next-generation sequence data from formalin-fixed
863 and fresh-frozen tissue specimens. *J Mol Diagn* **15**, 623–633 (2013).
864 43. Yokoyama, A. *et al.* Age-related remodelling of oesophageal epithelia by mutated cancer drivers. *Nature*
865 **565**, 312–317 (2019).
866 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
868 axis. *Signal Transduct Target Ther* **6**, 10 (2021).
869 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).
872 46. Wei, W. *et al.* Analysis and therapeutic targeting of the EP300 and CREBBP acetyltransferases in anaplastic
873 large cell lymphoma and Hodgkin lymphoma. *Leukemia* **37**, 396–407 (2023).
874 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).
876 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).
878 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).
880 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).
883 51. Takano, S. *et al.* MyD88 Mutation in Elderly Predicts Poor Prognosis in Primary Central Nervous System

884 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).

887 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).

889 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*

894 **14**, 449–456 (2000).

895 57. Nakagawa, M. *et al.* Targeting the HTLV-I-Regulated BATF3/IRF4 Transcriptional Network in Adult T Cell

896 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).

899 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).

63. Mlynarczyk, C. *et al.* BTG1 mutation yields supercompetitive B cells primed for malignant transformation. *Science* **379**, eabj7412 (2023).

64. Tijchon, E. *et al.* Tumor suppressors BTG1 and BTG2 regulate early mouse B-cell development. *Haematologica* **101**, e272-276 (2016).

65. Scheijen, B. *et al.* Tumor suppressors BTG1 and IKZF1 cooperate during mouse leukemia development and increase relapse risk in B-cell precursor acute lymphoblastic leukemia patients. *Haematologica* **102**, 541–551 (2017).

66. Li, Y. *et al.* The predictive value of BTG1 for the response of newly diagnosed acute myeloid leukemia to decitabine. *Clin Epigenetics* **16**, 16 (2024).

67. Abrey, L. E. *et al.* Report of an international workshop to standardize baseline evaluation and response criteria for primary CNS lymphoma. *J Clin Oncol* **23**, 5034–5043 (2005).

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).

69. Li, Q. *et al.* Improvement of outcomes of an escalated high-dose methotrexate-based regimen for patients with newly diagnosed primary central nervous system lymphoma: a real-world cohort study. *Cancer Manag Res* **13**, 6115–6122 (2021).

70. Freed, D., Aldana, R., Weber, J. A. & Edwards, J. S. The Sentieon Genomics Tools–A fast and accurate solution to variant calling from next-generation sequence data. *BioRxiv* 115717 (2017).

71. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol* **17**, 122 (2016).

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).

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).

940 80. Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101

941 (2020).

942 81. Chalmers, Z. R. *et al.* Analysis of 100,000 human cancer genomes reveals the landscape of tumor

943 mutational burden. *Genome Med* **9**, 34 (2017).

944 82. Garcia, M. *et al.* Sarek: A portable workflow for whole-genome sequencing analysis of germline and

945 somatic variants. *F1000Res* **9**, 63 (2020).

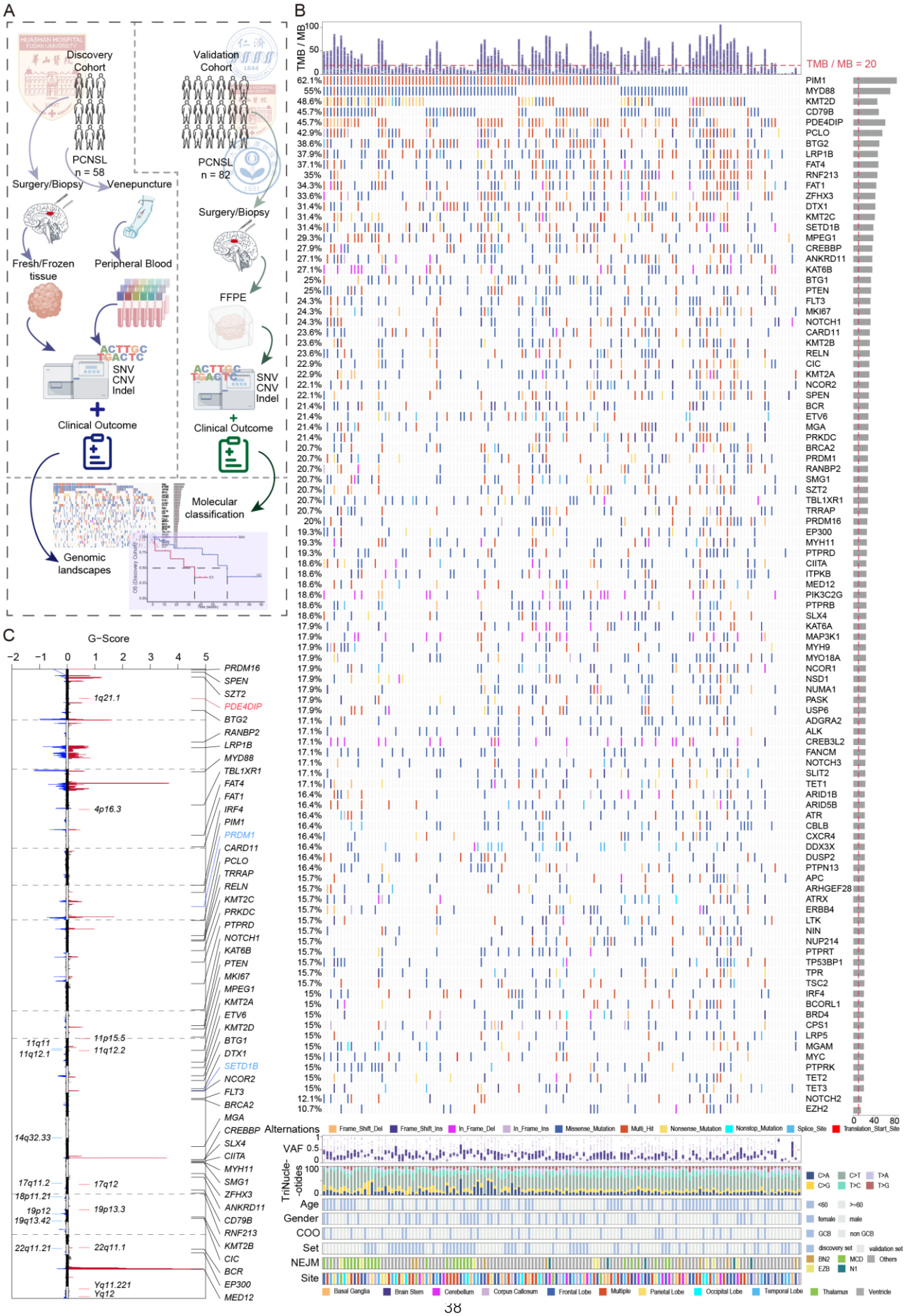
946 83. Ewels, P. A. *et al.* The nf-core framework for community-curated bioinformatics pipelines. *Nat Biotechnol*

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).
957


A
Discovery Cohort
PCNSL n = 58
Surgery/Biopsy
Venepuncture
Fresh/Frozen tissue
Peripheral Blood
SNV
CNV
Indel
Clinical Outcome
Validation Cohort
PCNSL n = 82
Surgery/Biopsy
FFPE
SNV
CNV
Indel
Clinical Outcome
Molecular classification
Genomic landscapes
B
TMB / MB
TMB / MB = 20
Alternations
VAF
TriNucle-otides
Age
Gender
COO
Set
NEJM
Site
C
G-Score
1q21.1
4p16.3
11q11
11q12.1
11p15.5
11q12.2
14q32.33
17q11.2
17q12
18p11.21
19p12
19q13.42
19p13.3
22q11.21
22q11.1
Yq11.221
Yq12

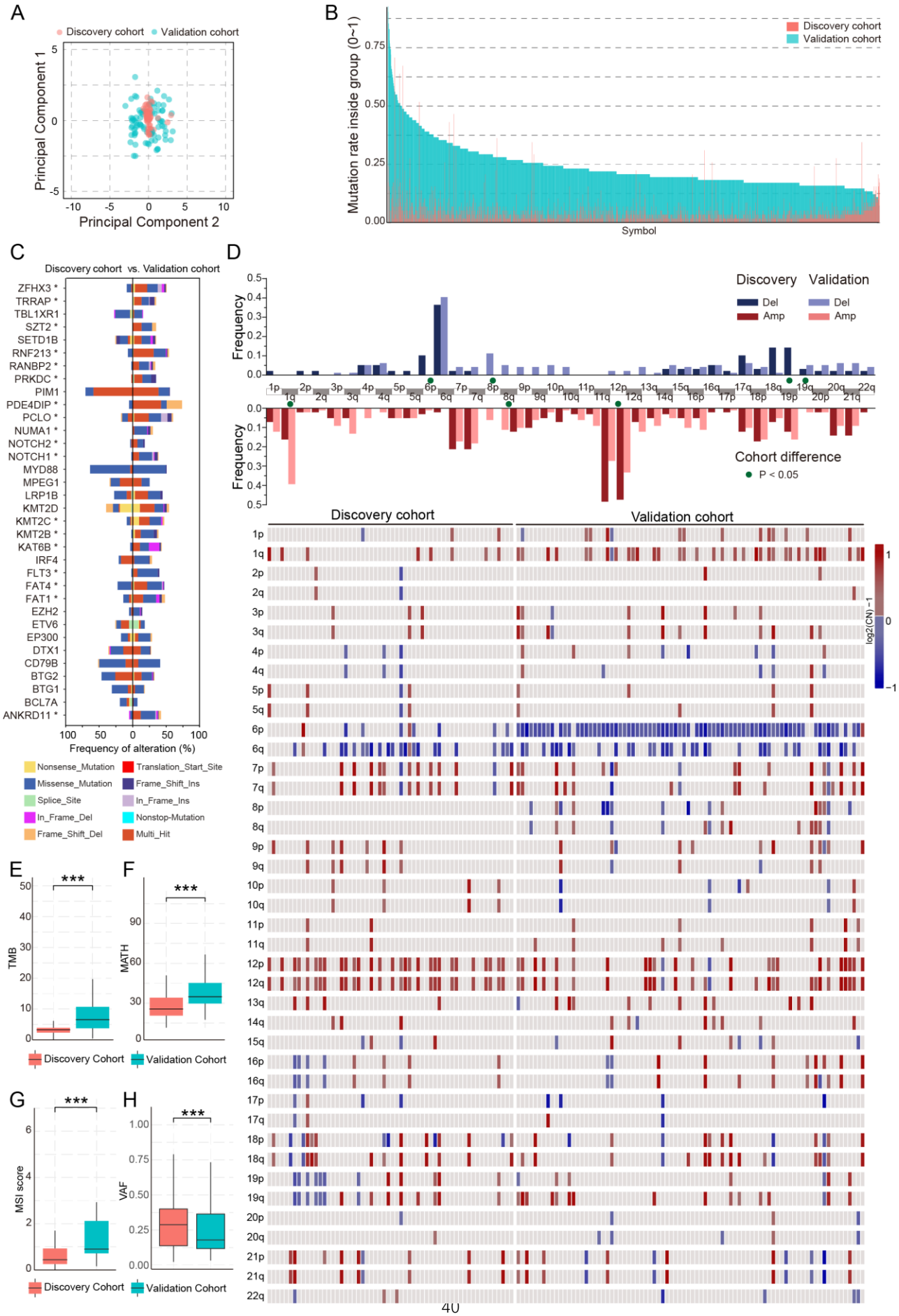
959 **Figure 1. Study design and the PCNSL mutational landscape in Chinese patients**

960 A: Study design, including workflow and data composition for the discovery cohort and
961 validation cohort.

962 B: Number and frequency of recurrent mutations. Gene–sample matrix of recurrently mutated
963 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.

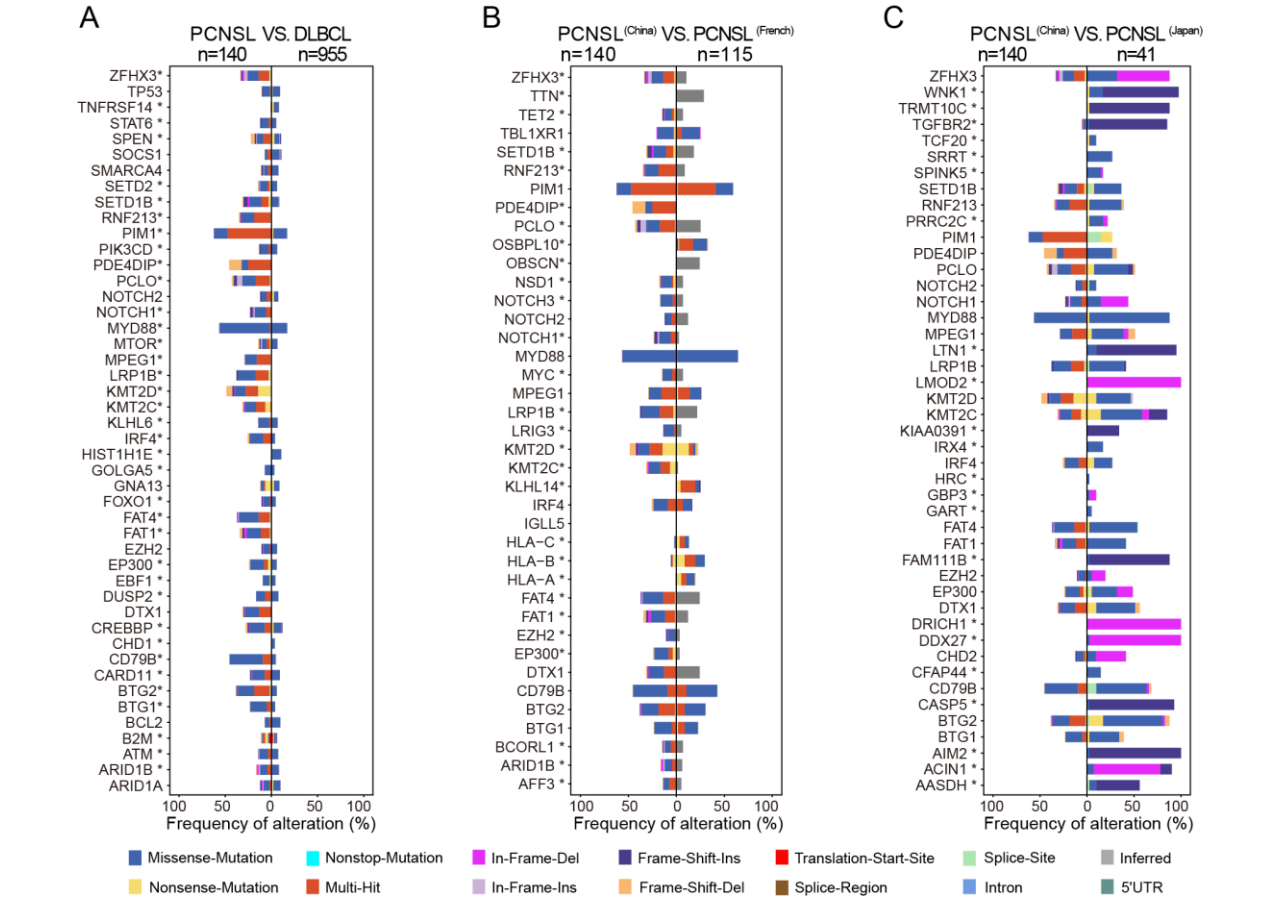
965 C: GISTIC2.0 results of significant recurrent focal amplifications (red) and deletions (blue).
966 Genes affected by each focal event are annotated (n = 140). X-axis: plot of chromosomes;
967 Y-axis: G score.

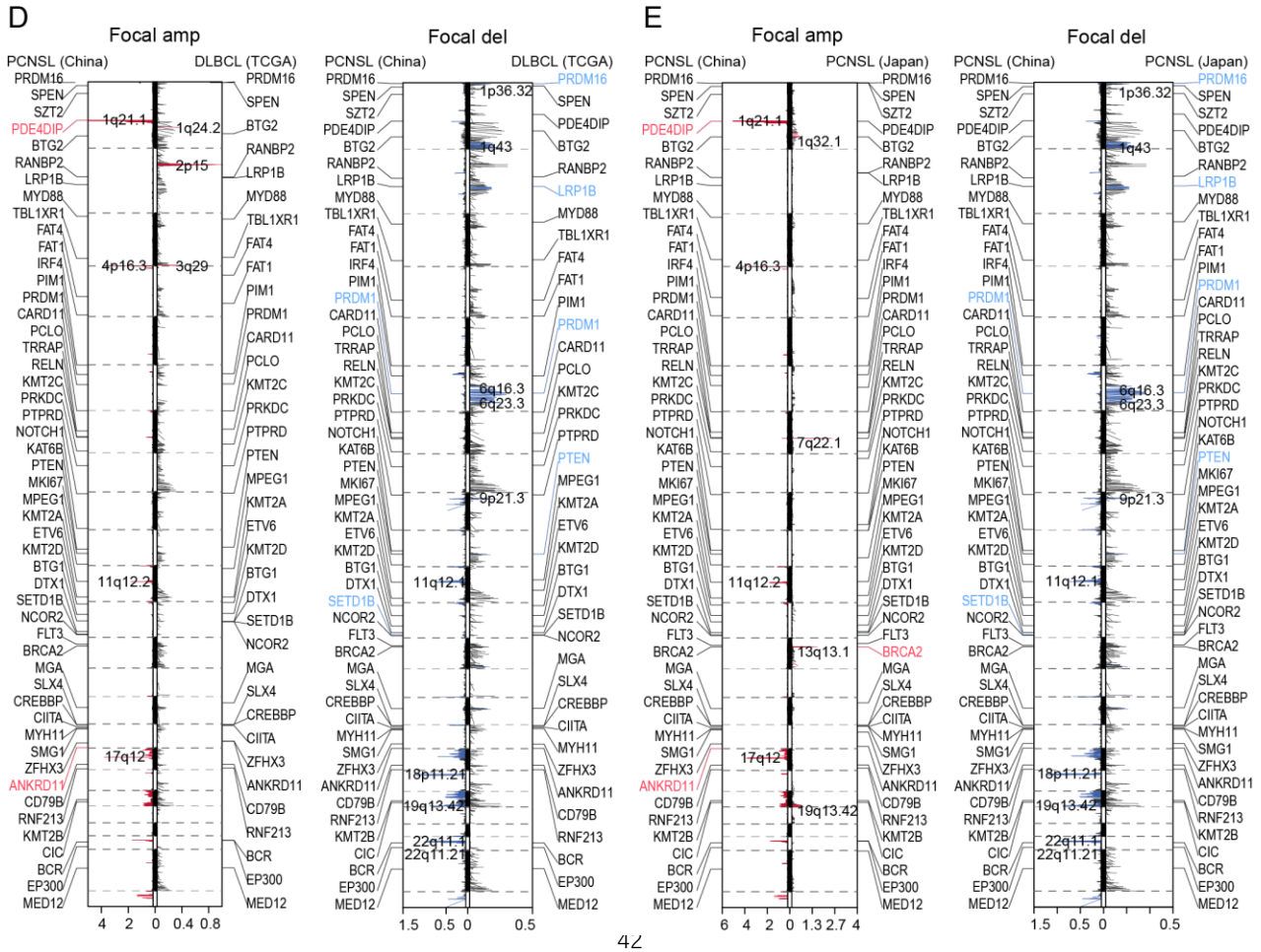
968


A
Discovery cohort
Validation cohort
Principal Component 1
Principal Component 2
B
Discovery cohort
Validation cohort
Mutation rate inside group (0~1)
Symbol
C
Discovery cohort vs. Validation cohort
Frequency of alteration (%)
Nonsense_Mutation
Missense_Mutation
Splice_Site
In_Frame_Del
Frame_Shift_Del
Translation_Start_Site
Frame_Shift_Ins
In_Frame_Ins
Nonstop-Mutation
Multi_Hit
D
Discovery
Validation
Del
Amp
Frequency
Cohort difference
P < 0.05
Discovery cohort
Validation cohort
log2(CN)
E
TMB
F
MATH
G
MSI score
H
VAF
Discovery Cohort
Validation Cohort

969

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
979 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
981 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
989 validation cohort included unpaired FFPE tumor tissue samples (n = 82). Independent-
990 samples t tests and chi-squared tests were used. *$P < 0.05$; **$P < 0.01$; ***$P < 0.001$.


A
PCNSL VS. DLBCL
n=140
n=955
B
PCNSL(China) VS. PCNSL(French)
n=140
n=115
C
PCNSL(China) VS. PCNSL(Japan)
n=140
n=41
Frequency of alteration (%)
Missense-Mutation
Nonsense-Mutation
Nonstop-Mutation
Multi-Hit
In-Frame-Del
In-Frame-Ins
Frame-Shift-Ins
Frame-Shift-Del
Translation-Start-Site
Splice-Region
Splice-Site
Intron
Inferred
5'UTR


D
Focal amp
Focal del
PCNSL (China)
DLBCL (TCGA)
E
PCNSL (Japan)

991

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
994 patients (n = 140) and DLBCL patients (n = 955).

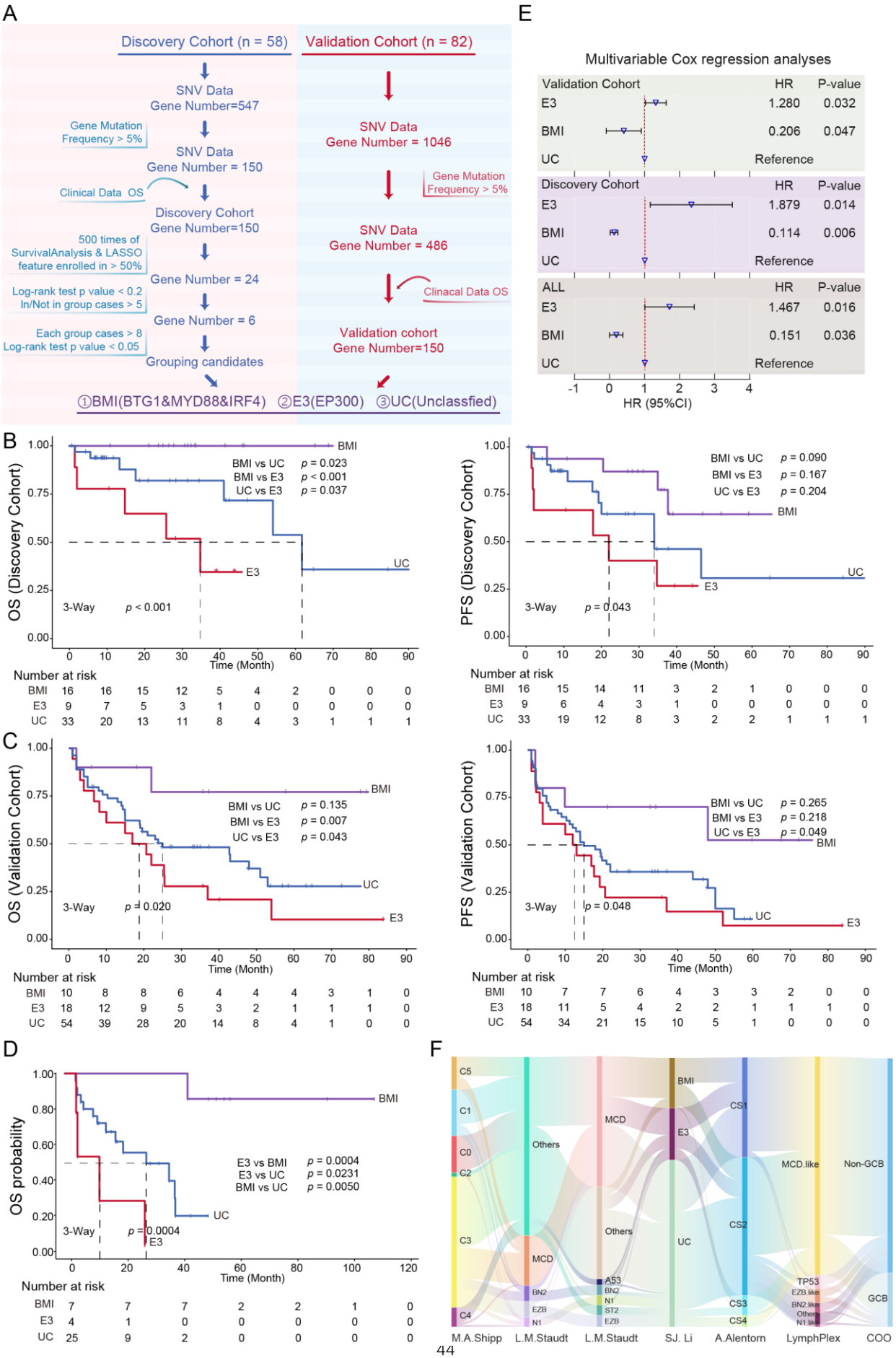
995 B: A mirror bar plot showing the frequencies of genetic alterations between Chinese PCNSL
996 patients (n = 140) and French PCNSL patients (n = 115).

997 C: A mirror bar plot showing the frequencies of genetic alterations between Chinese PCNSL
998 patients (n = 140) and Japanese PCNSL patients (n = 41).

999 D: GISTIC-defined recurrent copy number focal deletions (blue) and gains (red) as mirror
1000 plots in DLBCL (TCGA, n = 46) and Chinese PCNSLs (n = 140) are shown. X-axis: plot of
1001 chromosomes; Y-axis: G score.

1002 E: The GISTIC2.0-defined recurrent copy number focal deletions (blue) and gains (red) in
1003 Japanese PCNSL patients (n = 41) and Chinese PCNSL patients (n = 140) are shown as
1004 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.


A
Discovery Cohort (n = 58)
Validation Cohort (n = 82)
SNV Data
Gene Number=547
Gene Mutation Frequency > 5%
SNV Data
Gene Number = 150
Clinical Data OS
Discovery Cohort
Gene Number=150
500 times of SurvivalAnalysis & LASSO feature enrolled in > 50%
Gene Number = 24
Log-rank test p value < 0.2 In/Not in group cases > 5
Gene Number = 6
Each group cases > 8 Log-rank test p value < 0.05
Grouping candidates
SNV Data
Gene Number = 1046
Gene Mutation Frequency > 5%
SNV Data
Gene Number = 486
Clinacal Data OS
Validation cohort
Gene Number=150
①BMI(BTG1&MYD88&IRF4) ②E3(EP300) ③UC(Unclassfied)
E
Multivariable Cox regression analyses
Validation Cohort HR P-value
E3 1.280 0.032
BMI 0.206 0.047
UC Reference
Discovery Cohort HR P-value
E3 1.879 0.014
BMI 0.114 0.006
UC Reference
ALL HR P-value
E3 1.467 0.016
BMI 0.151 0.036
UC Reference
HR (95%CI)
B
OS (Discovery Cohort)
BMI vs UC p = 0.023
BMI vs E3 p < 0.001
UC vs E3 p = 0.037
3-Way p < 0.001
Time (Month)
Number at risk
BMI 16 16 15 12 5 4 2 0 0 0
E3 9 7 5 3 1 0 0 0 0 0
UC 33 20 13 11 8 4 3 1 1 1
PFS (Discovery Cohort)
BMI vs UC p = 0.090
BMI vs E3 p = 0.167
UC vs E3 p = 0.204
3-Way p = 0.043
Number at risk
BMI 16 15 14 11 3 2 1 0 0 0
E3 9 6 4 3 1 0 0 0 0 0
UC 33 19 12 8 3 2 2 1 1 1
C
OS (Validation Cohort)
BMI vs UC p = 0.135
BMI vs E3 p = 0.007
UC vs E3 p = 0.043
3-Way p = 0.020
Number at risk
BMI 10 8 8 6 4 4 4 3 1 0
E3 18 12 9 5 3 2 1 1 1 0
UC 54 39 28 20 14 8 4 1 0 0
PFS (Validation Cohort)
BMI vs UC p = 0.265
BMI vs E3 p = 0.218
UC vs E3 p = 0.049
3-Way p = 0.048
Number at risk
BMI 10 7 7 6 4 3 3 2 0 0
E3 18 11 5 4 2 2 1 1 1 0
UC 54 34 21 15 10 5 1 0 0 0
D
OS probability
E3 vs BMI p = 0.0004
E3 vs UC p = 0.0231
BMI vs UC p = 0.0050
3-Way p = 0.0004
Number at risk
BMI 7 7 7 2 2 1 0
E3 4 1 0 0 0 0 0
UC 25 9 2 0 0 0 0
F
C5
C1
C0
C2
C3
C4
Others
MCD
BN2
EZB
N1
MCD
Others
A53
BN2
N1
ST2
EZB
BMI
E3
UC
CS1
CS2
CS3
CS4
MCD.like
TP53
EZB.like
BN2.like
Others
N1.like
Non-GCB
GCB
M.A.Shipp
L.M.Staudt
L.M.Staudt
SJ. Li
A.Alentorn
LymphPlex
COO

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
1013 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)
1015 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)
1017 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

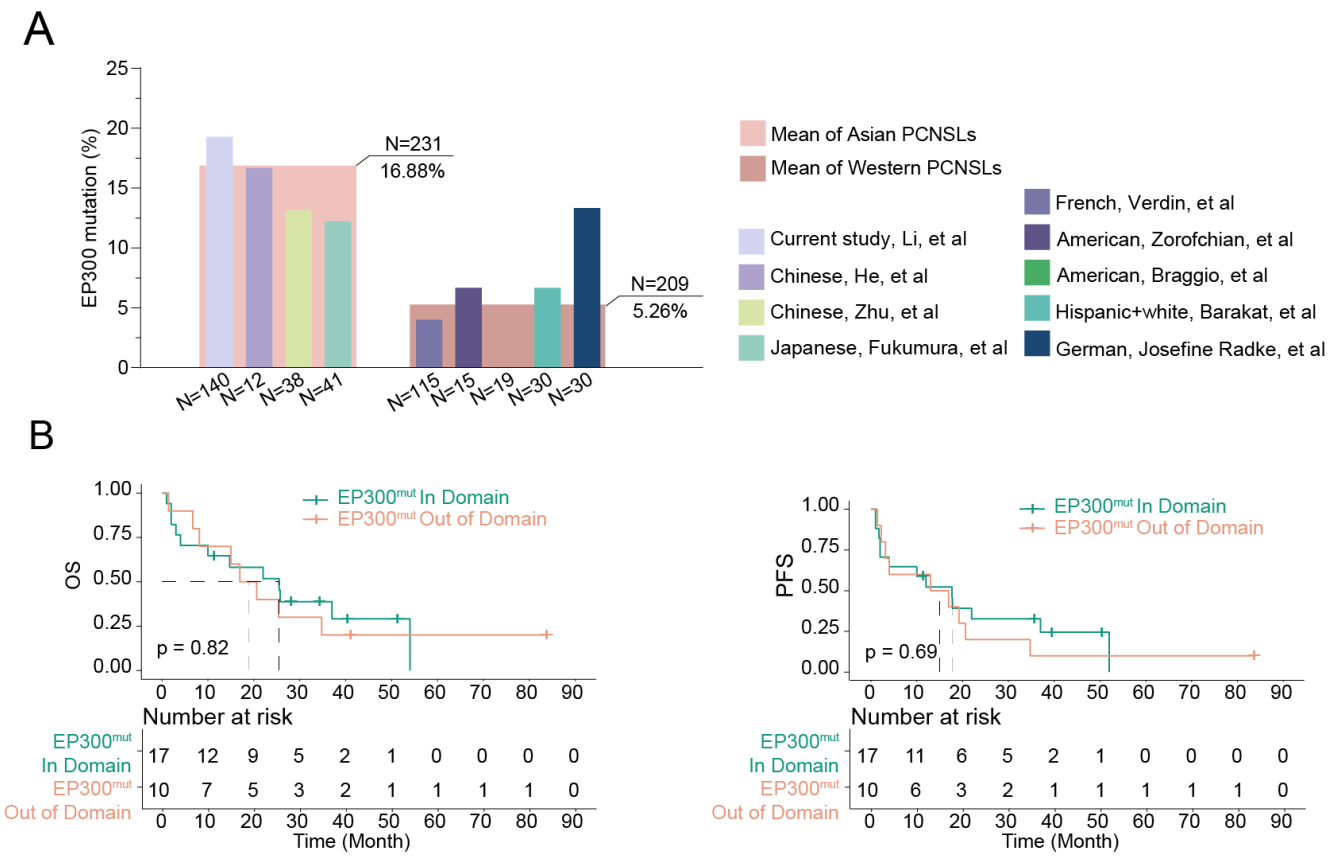
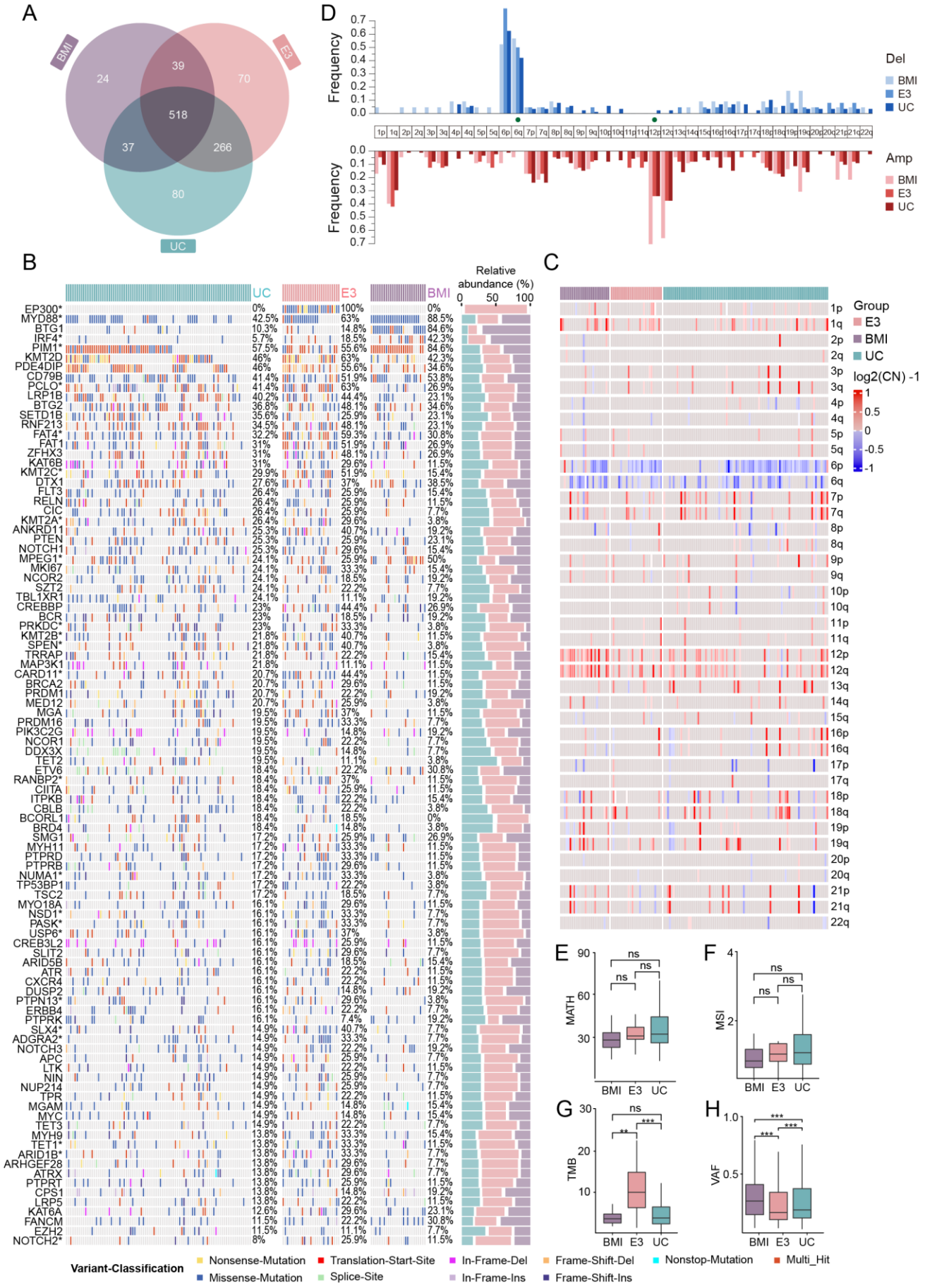


Figure 5. Mutation frequency among races and prognosis of EP300

A: The prevalence of EP300 mutations in PCNSL patients of different races.

B: Kaplan–Meier analyses for comparisons of the overall survival (OS) and progression-free survival (PFS) of patients with mutations within and those with mutations outside the EP300 domain.


A
BMI
E3
UC
24
39
70
518
37
266
80
D
Frequency
Del
BMI
E3
UC
Amp
B
UC
E3
BMI
Relative abundance (%)
Variant-Classification
Nonsense-Mutation
Missense-Mutation
Translation-Start-Site
Splice-Site
In-Frame-Del
In-Frame-Ins
Frame-Shift-Del
Frame-Shift-Ins
Nonstop-Mutation
Multi_Hit
C
Group
E3
BMI
UC
log2(CN)
E
MATH
F
MSI
G
TMB
H
VAF
ns

1033

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.

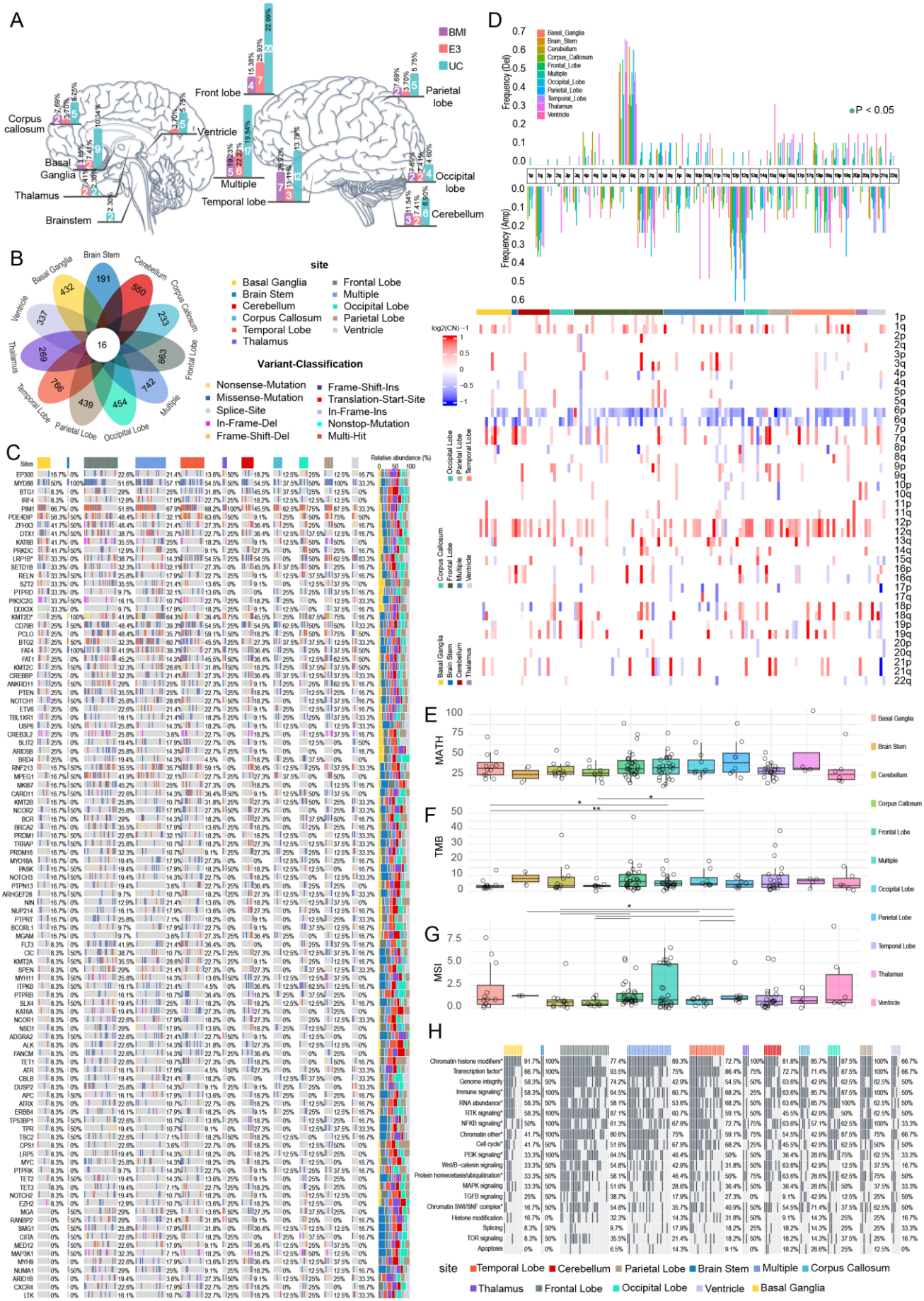
1046 F: The levels of microsatellite instability (MSI) were compared among the three MS samples.

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.

1049 Independent-samples t tests and chi-squared tests were used. *$P < 0.05$; **$P < 0.01$; ***$P <$
1050 0.001; ns $P > 0.05$.

1051


A
BMI
E3
UC
Front lobe
Parietal lobe
Corpus callosum
Ventricle
Basal Ganglia
Thalamus
Brainstem
Multiple
Temporal lobe
Occipital lobe
Cerebellum
B
Basal Ganglia
Brain Stem
Cerebellum
Corpus Callosum
Frontal Lobe
Multiple
Occipital Lobe
Parietal Lobe
Temporal Lobe
Thalamus
Ventricle
site
Variant-Classification
Nonsense-Mutation
Missense-Mutation
Splice-Site
In-Frame-Del
Frame-Shift-Del
Frame-Shift-Ins
Translation-Start-Site
In-Frame-Ins
Nonstop-Mutation
Multi-Hit
C
Sites
Relative abundance (%)
D
Frequency (Del)
Frequency (Amp)
P < 0.05
log2(CN)
E
MATH
F
TMB
G
MSI
H
Chromatin histone modifiers*
Transcription factor*
Genome integrity
Immune signaling*
RNA abundance*
RTK signaling*
NFKB signaling*
Chromatin other*
Cell cycle*
PI3K signaling*
Wnt/β-catenin signaling
Protein homeostasis/ubiquitination*
MAPK signaling
TGFB signaling
Chromatin SWI/SNF complex*
Histone modification
Splicing
TOR signaling
Apoptosis
site
Temporal Lobe
Cerebellum
Parietal Lobe
Brain Stem
Multiple
Corpus Callosum
Thalamus
Frontal Lobe
Occipital Lobe
Ventricle
Basal Ganglia

1052

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
1058 recurrently mutated genes in PCNSL patients at different sites of onset, are presented. The
1059 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. *$P < 0.05$.

1070