Supplementary Materials

Mutational Architecture of Clonal Hematopoiesis of Indeterminate Potential in 29,596 Chinese Individuals

Xia Tang, Hongpu Chen, Xiaoyuan Zhang, Yingbing Shi, Yang Gao, Yumeng Sun, Baonan Wang, Qi Liu, Yuhang Cui, Mochi Wei, Shuangshuang Cheng, Lian Deng, Zeshan Lin, Zhimin Feng, Yaoxi He, Jiucun Wang, Bing Su, Sijia Wang, Jingchun Luo, Yan Zheng, The Han100K Initiative, Xin Jin, Yanan Cao, Yan Lu, Li Jin, Shuhua Xu*

*Contact: xushua@fudan.edu.cn

Supplementary Notes

CHIP profile across Chinese administrative regions

CHIP prevalence varied markedly across administrative regions in the discovery cohort, showing an almost 4.9–fold difference (Supplementary Fig. 21a). Higher prevalence was observed in Shanxi (11.3%, 95% CI: 9.0%–14.0%, 631 individuals aged 18–86 years, mean 62.4 years, median 68 years), Guangxi (6.1%, 95% CI: 5.0%–7.4%, 1,556 individuals aged 19–83 years, mean 51.93 years, median 54 years), Yunnan (5.8%, 95% CI: 4.0%–8.4%, 432 individuals aged 18–86 years, mean 45.56 years, median 44 years), and Shanghai (5.7%, 95% CI: 4.9%–6.6%, 2,949 individuals aged 18–89 years, mean 42.83 years, median 40 years) compared with the national estimate of 4.5% (regions with N > 400 only). In contrast, lower rates were observed in Henan (2.3%, N = 2,161), Hubei (2.3%, N = 1,462), and Jiangsu (3.5%, N = 1,497). A similar 5.4–fold regional variation was detected in the replication cohort (Supplementary Fig. 21b). However, CHIP prevalence showed no significant correlation with longitude or latitude (Pearson correlation, $P > 0.05$). These regional estimates are likely confounded by differences in in sample size, sequencing depth, sex structure, and substantial population mobility in metropolitan areas. Future regional comparisons will therefore require larger, well–balanced cohorts with regions control of these cofactors.

CHIP profile in Chinese ethnic subgroups

Ethnic minority groups accounted for 10.1% and 10.7% of the discovery and replication cohorts, respectively. Given known genetic differentiation between ethnic minorities and Han Chinese[1–7], we examined CHIP prevalence across ethnic subgroups, acknowledging uneven sample sizes and limited power for a low–incidence trait. Apparent heterogeneity was observed in both discovery (Fig. 1e and Supplementary Fig. 22a) and replicate (Supplementary Figs. 4e and 23a) cohorts, respectively. After adjusting for age and sex and applying matched resampling within Han subgroups (N = 173 in discovery; N = 240 in replication), within–subgroup variability was comparable to that observed between different ethnic subgroups (Supplementary Figs. 22b and 22d; Kolmogorov–Smirnov (K–S) test P = 0.98), whereas ethnic effects remained detectable (Supplementary Figs. 2c and 22e; K–S test P = 0.03), with replication analyses showing consistent trends. In contrast, ethnic effects might influence CHIP carriage differences. Results from the

replication cohort also support the ethnic effects on the CHIP variants among subgroups (Supplementary Figs. 23b and 23d, K–S test P = 0.5; Supplementary Figs. 23c and 23e, K–S test P = 0.01). Moreover, the absolute difference in CHIP prevalence between each minority group and Han Chinese positively correlated with genetic distance (F_{ST}), suggesting an ancestry–related component to CHIP variation (Supplementary Fig. 24). Across ethnic subgroups, we observed a positive association: greater genetic differentiation from Han Chinese (larger F_{ST}) tended to be accompanied by larger disparities in CHIP prevalence (Supplementary Fig. 24), supporting an ancestry–based foundation for CHIP carriage.

The CHIP prevalence by genders varied between different ethnic subgroups in the Chinese population. The Zhuang (3/101 females vs. 3/49 males, $P < 2.2 \times 10^{-16}$) and Tibetan exhibited a significant male preference in the discovery cohort (Supplementary Fig. 25a), which the Zhuang was replicated with a certain male preference in the replication cohort (3/138 females vs. 3/61 males, $P < 2.2 \times 10^{-16}$, Supplementary Fig. 25b). The Tibetan presented different gender–preferences in the discovery and replication cohorts (Supplementary Figs. 25a–b), indicating the great CHIP heterogeneity even within the population of similar ethnic background. Notably, only three female carriers were identified in 80 individuals of the Yao population (48 females and 32 males), in which no male carrier was identified, with an overall prevalence of 3.8% (3/80) (Fig. 1e and Supplementary Fig. 25a). In the replication cohort, only two female carriers were identified in 252 individuals of the Yi population (136 females and 116 males, without difference in the age distribution, Wilcoxon P = 0.68) (Supplementary Figs. 4e and 25b). The incidence of CHIP between the genders of Han, Man, Hui, and Zhuang was relatively comparable (Fig. 1e and Supplementary Fig. 4e), consistent with their closer genetic distance. These results supported the earlier knowledge or consensus that the observed disparities in CHIP prevalence between ethnic subgroups and genders were largely caused by their diverse genetic backgrounds[8], as well as their demographic histories[9].

Specifically, we detected the 2.2% CHIP prevalence in Tibetans in both discovery (N = 357, aged 18–80 years, mean age 39.87 years, median 39 years) and replication cohorts (N = 225, aged 19–78 years, mean age 35.7 years, median 32 years) according to the whitelist variant. We observed that the most frequent CHIP genes in Tibetan populations were *KMT2D* and *TP53*, while *DNMT3A* and *TET2* were detected in only one individual each, which are the most frequently

mutated genes observed in lowlanders (Fig. 1f, and Table S4 and S5). It is noteworthy that *TP53* is a classic oncogene, and its high frequency is supported by the high incidence of liver cancer in Tibetan populations[10]. This suggests a substantial difference between the CHIP mutation profiles of the plateau–adapted Tibetan population and the low–altitude Han population. Thus, we expanded our analysis to encompass all functional variants (see Methods) within the CHIP 74 genes, beyond the whitelist variant. We detected 78 events in 30 individuals (Table S7 and Supplementary Fig. 26), yielding a prevalence of 8.4% (30/357). No additional *DNMT3A* variants were identified, however, nine missense variants in *KMT2A*, five in *KMT2D*, five in *TET2* were added, which still deviated from our knowledge of the *DTA* model. Interestingly, Tibetans are highlanders who have a long–term inhabitation history on the Qinghai–Tibet plateau and are exposed to the unavoidable environmental stress of severe high–altitude hypoxia, which directly affects the oxygenation of the blood, resulting in insufficient oxygen supply[1,2,11]. Therefore, the unique mechanism of high–altitude adaptation might exhibit a potential association with the emergence of CHIP in highlanders. In addition, a significant lower CHIP prevalence of 1.3% is found among Uyghurs, a typical representative of the admixture of Eastern and Western ancestries, which may violate our discovery that the higher the proportion of EUR ancestors, the closer the CHIP prevalence is to EUR. We checked the age distribution of the Uyghur (aged 18–93 years, mean age of 34.25 years, median of 20 years) (Supplementary Fig. 10a), which may account for the lower prevalence of CHIP in this population. Therefore, this also reminds us that environmental factors such as age remain critical influences in CHIP analyses and that more fine–grained analyses are needed to clarify the extent and interpretability of ancestry effects.

Genomic differentiation of CHIP–related genes between Western and Eastern populations

To explore the genetic basis of East–West differences in CHIP, we analyzed 333 CHIP–related genes (Table S3) and GWAS salient loci obtained for CHIP as a query list (Table S15) using KGP East Asian (EAS) and European (EUR) populations from the gene (Supplementary Fig. 27) and SNP (Supplementary Fig. 28) levels, respectively. Among the 333 CHIP–related genes, 56 (~17%) had significant (Fixation index, F_{ST}, ranked top 0.1%) East–West differentiation, including nine whitelist genes, i.e., *CUX1*, *PTPN11*, *ASXL2*, *JAK1*, *ETNK1*, *CSF1R*, *IKZF3*, *KMT2A*, and *PDSS2*, with the F_{ST} of 0.76, 0.77, 0.75, 0.66, 0.63, 0.61, 0.6, 0.6, and 0.59, respectively (Supplementary Fig. 27 and Table S16). Specifically, the leading differentiated SNP,

rs60656667 (C > T) in *CUX1*, encoding DNA binding protein, reached a relatively low frequency in the West but high in the East (0.15 vs. 0.85, F_{ST} = 0.76, Supplementary Figs. 28 and 29a), consistent with the higher *CUX1*–CHIP prevalence in the Chinese populations (ranked top 12 frequent, Supplementary Fig. 2a). Moreover, the most differentiated gene was *FOXP1* (with an F_{ST} of 0.86) (Table S16), an expanded CHIP gene that acts as a tumor suppressor functioning on DNA–binding transcription factor activity and sequence–specific binding[12]. *FOXP1* exhibited the strongest gene–level differentiation (F_{ST} = 0.86), with derived allele (*FOXP1* rs56091574 T > C, with a CADD score of 17.6) enriched in East Asians compared to Western populations (0.81 vs. 0.19, Supplementary Figs. 28 and 29b, and Table S16–S17), suggesting potential population–specific regulatory influences.

Conversely, variants in *PTPN11* were more frequent in Europeans, concordant with the rarity of *PTPN11*–related CHIP in our Chinese cohorts. *PTPN11* encodes the protein tyrosine phosphatase (PTP) family, which is a signaling molecule that regulate a variety of cellular processes including oncogenic transformation[13]. Activating *PTPN11* mutants could promote hematopoietic progenitor cell–cycle progression and survival[13]. The derived allele frequency (rs2301756 A > G) was high in Western populations but low in Eastern populations (0.85 vs. 0.15, Supplementary Figs 27 and 29c, and Table S17). Additionally, the prevalence of *PTPN11*–CHIP in the Chinese population was correspondingly extremely rare, which we did not detected whitelist variants in *PTPN11* (Supplementary Fig. 3), and only eight expanded variants were detected in eight individuals among merged cohorts (N = 29,596 individuals) (Supplementary Fig. 13). Furthermore, significant differentiation was observed in *DNMT3B*, an important paralog of *DNMT3A*, encodes a DNA methyltransferase functioning in de novo methylation rather than maintenance methylation. Although the genomic region encompassing the *DNMT3B* did not meet the top 1% genome–wide threshold for statistical significance (empirical P = 0.03), it exhibited the highest signal of genetic differentiation per kb (N = 1.59 extremely differentiated loci per kb; Table S16). All the SNPs within this genomic region exhibited a pattern of higher frequency in European populations, while displaying a lower frequency in East Asian populations (Supplementary Fig. 28 and 29d, and Table S18).

Together, these results indicate that ancestry and gene–environment interactions likely underlie global heterogeneity in CHIP, and that population genetic differentiation in CHIP–related genes provide a genomic framework for interpreting ancestry–specific mutation spectra.

Supplementary Figures

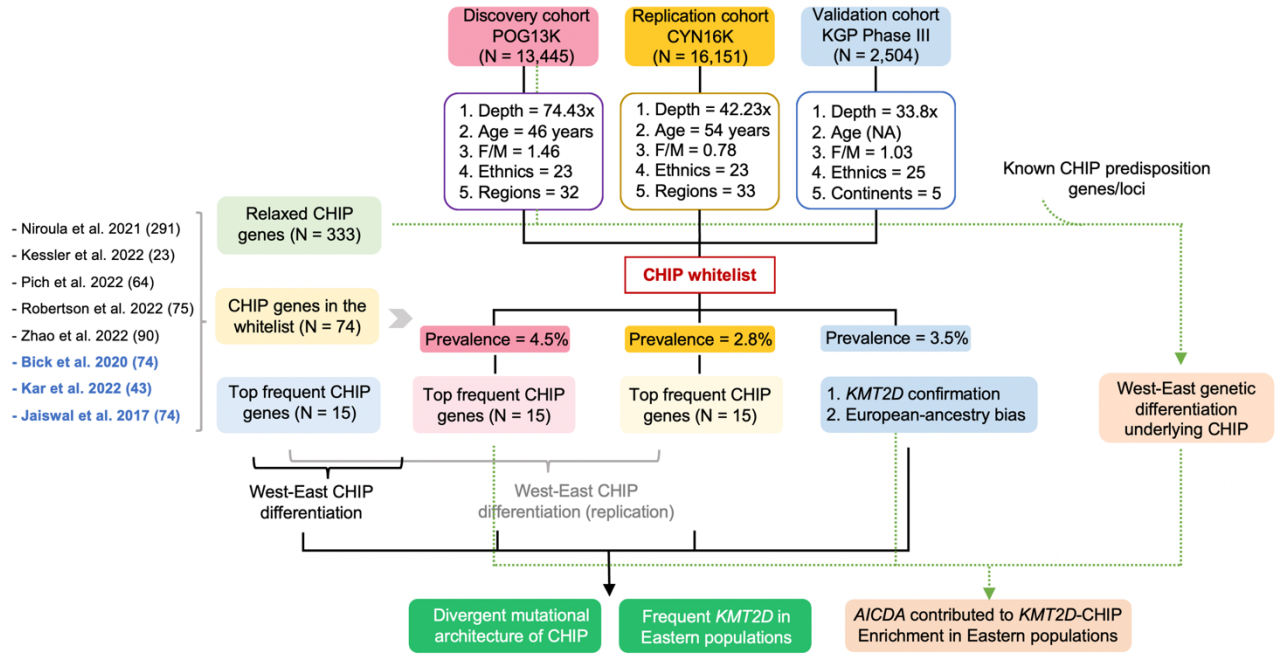


Figure S1. Schematic overview of CHIP gene variation in Chinese populations. Three cohorts used in this study: discovery (POG13K, N = 13,445), replication (CYN16K, N = 16,151), and validation (1KGP 30x WGS, N = 2,500). CHIP–related genes (N = 333) were collected from the union of eight previous CHIP studies [8,9,14–19], 74 whitelist CHIP genes were sourced from Jaiswal et al. in 2017[14], and top 15 frequent CHIP genes were the intersection of three studies (mainly based on European populations)[8,14,17]. We identified CHIP carriers defined by CHIP whitelist variants in the discovery cohort, and then progressed the replication and validation cohorts, respectively. Depth and age were the mean value of the cohort, respectively.

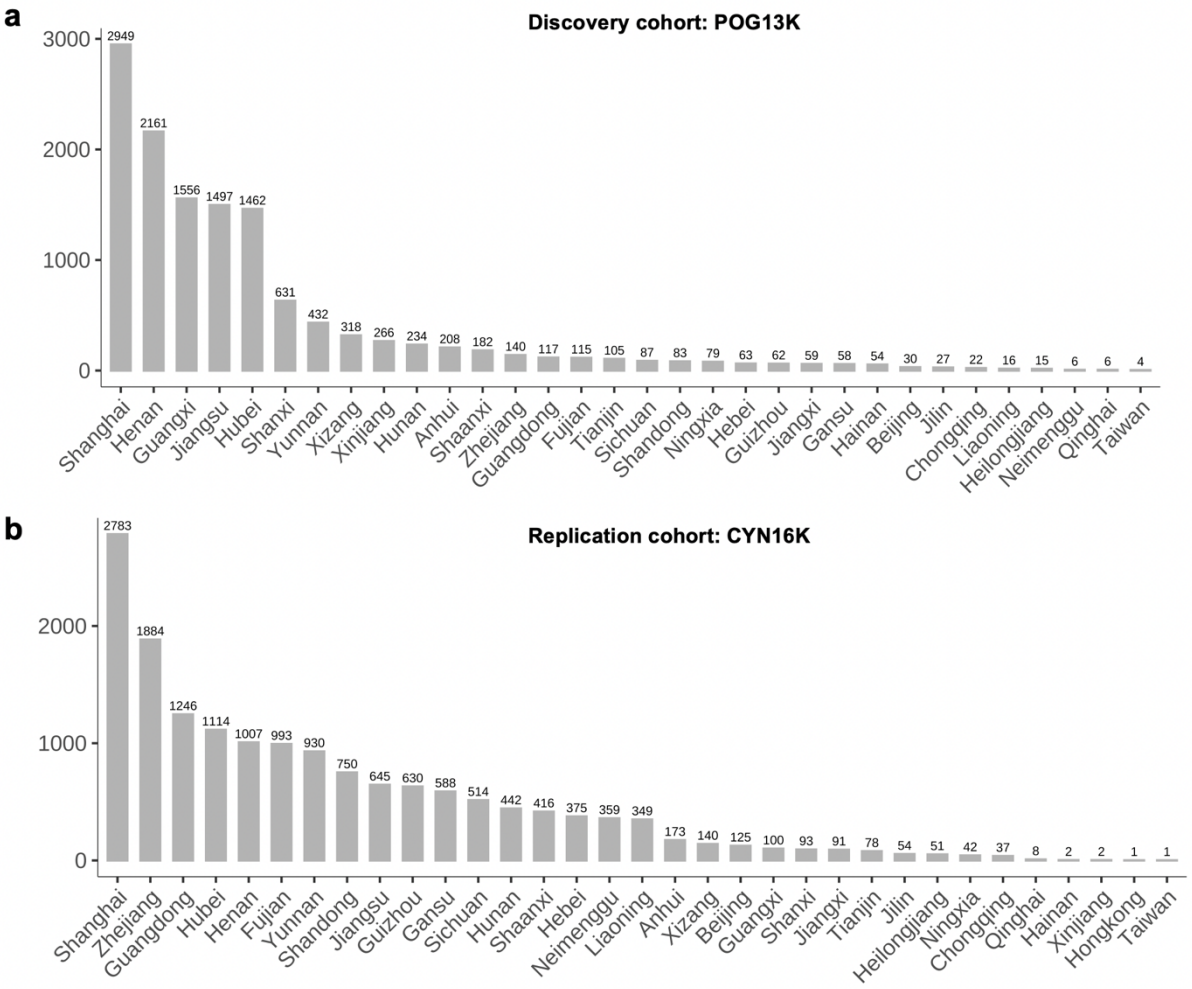


Figure S2. The distribution of sample size in different administrative regions within discovery cohort (**a**) and replication cohort (**b**). Only regions with a sample size greater than 400 were indicated.

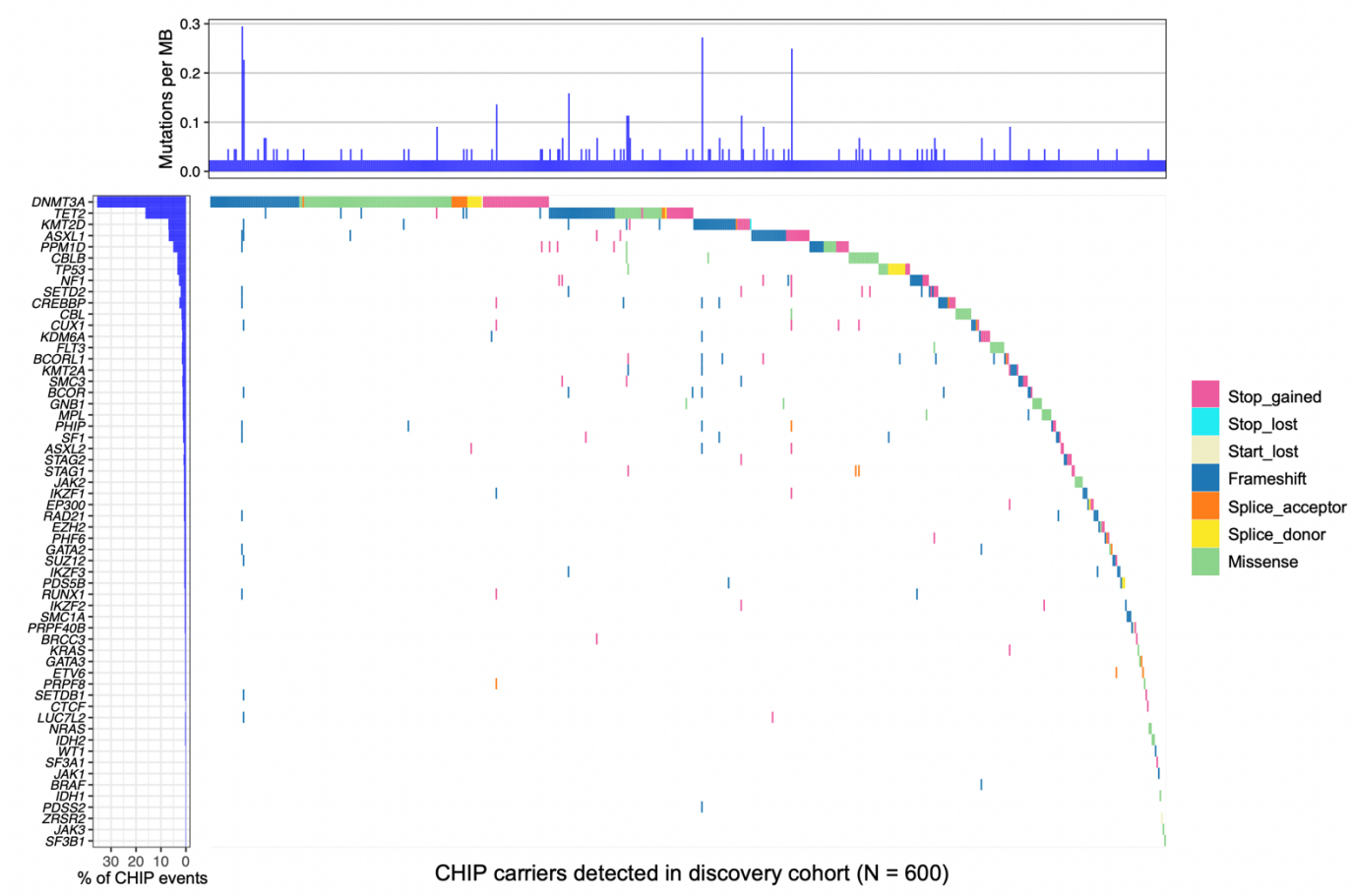


Figure S3. The landscape of CHIP in Chinese populations (whitelist variants detected in discovery cohort). The top histogram indicates the CHIP mutation burden in each carrier (the mutation burden is equal to the number of mutations per coverage space, multiplied by one million. The left histogram represents the proportion of mutations identified for each gene of all detected mutations. In the middle, x–axis indicated CHIP carriers, which every column represents the detected CHIP variants in each CHIP carrier, including stop_gained, stop_lost, start_lost, frameshift, splicing, and missense variants.

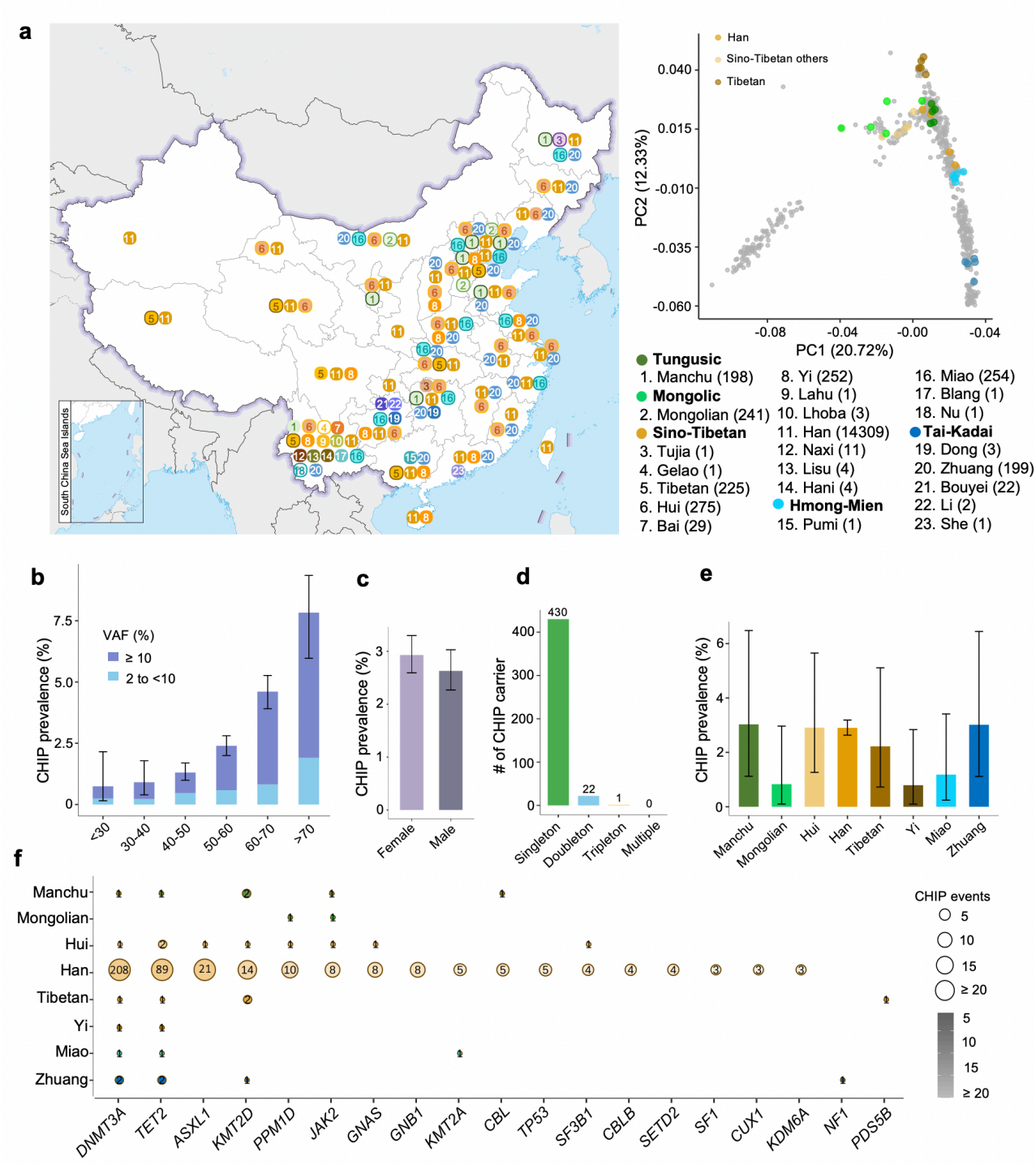


Figure S4. Identifying CHIP in 16,151 Chinese populations including whole–genome samples (replication cohort). **a.** Left: the geographical locations and ethnic, linguistic and genetic affiliations of the Chinese populations used in this study (see Table S1 for details). Top right: the results of principal component (PC) analysis based on whole–genome data

of the used Chinese populations (colored dots, every ethnic minority was highlighted by five individuals) in the context of East Asian populations (grey dots). **b.** Increased CHIP prevalence with the age by VAF. **c.** The distribution of overall CHIP prevalence in Chinese populations by gender, with 2.9% in females (95% CI: 2.6%–3.3%) and 2.6% in males (95% CI: 2.3%–3.0%). **d.** More than 84% of individuals with CHIP had only one somatic CHIP driver mutation variant identified (singleton). **e.** The CHIP prevalence varies greatly among different ethnic minorities, which only groups with detected at least one CHIP carriers are shown. **f.** Top frequent CHIP genes mutated in each subgroup. All detected CHIP genes are indicated if less than 10 in total were identified in that subgroup, except the Han subgroup, which only the genes detected more than five variants were indicated.

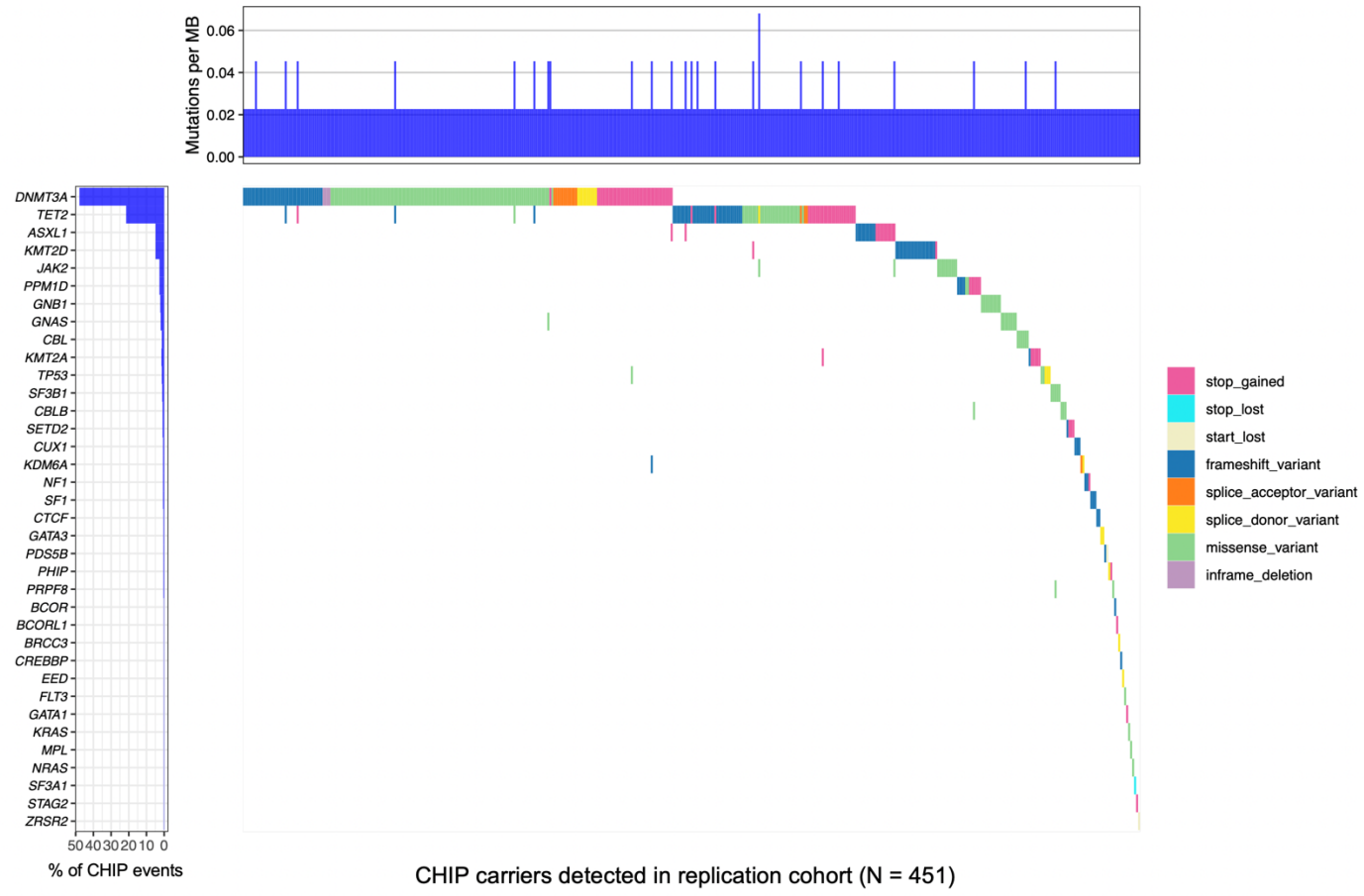


Figure S5. The landscape of CHIP in Chinese populations (whitelist variants detected in replication cohort). The legend is consistent with Figure S3.

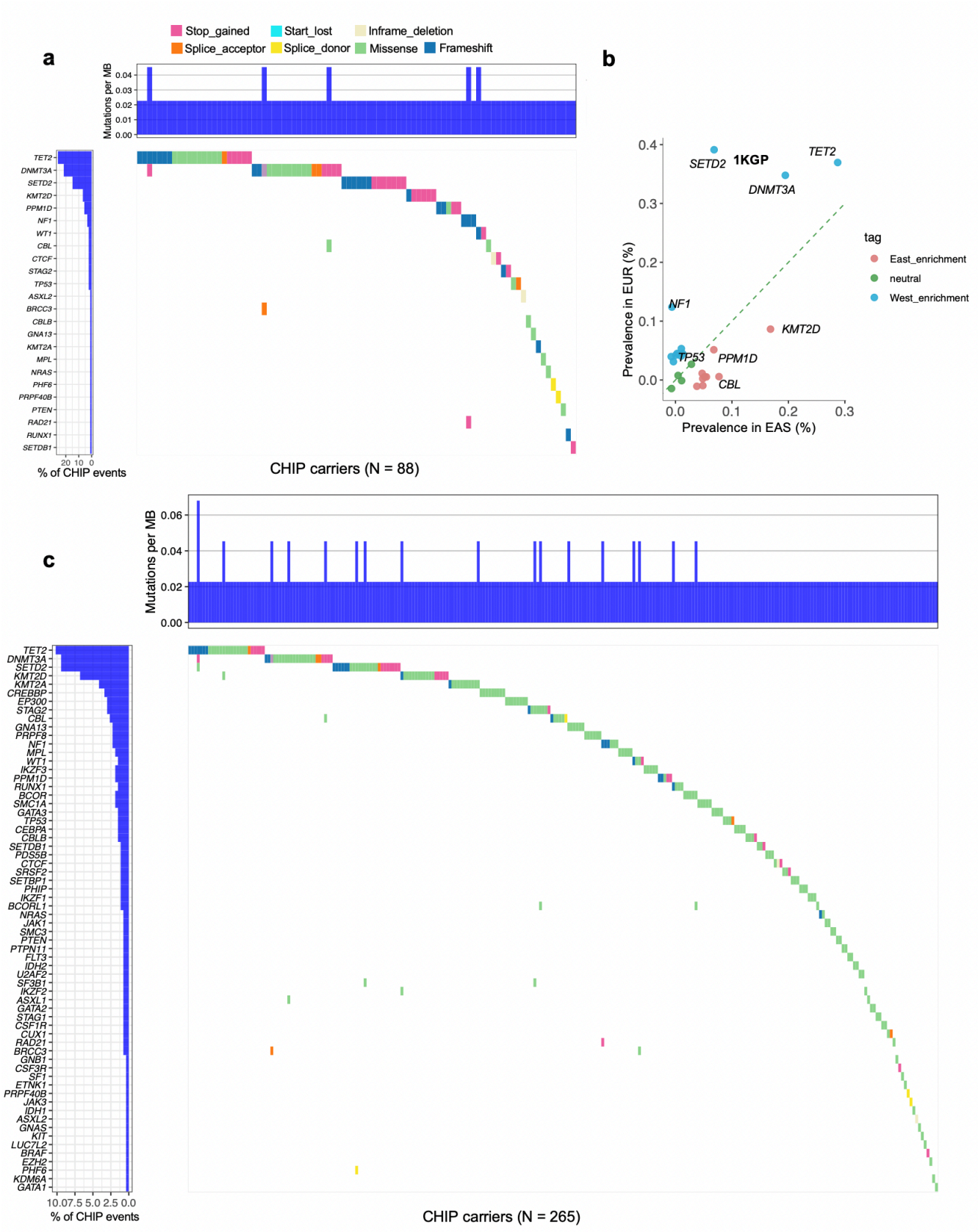

a
Stop_gained
Start_lost
Inframe_deletion
Splice_acceptor
Splice_donor
Missense
Frameshift
Mutations per MB
0.04
0.03
0.02
0.01
0.00
TET2
DNMT3A
SETD2
KMT2D
PPM1D
NF1
WT1
CBL
CTCF
STAG2
TP53
ASXL2
BRCC3
CBLB
GNA13
KMT2A
MPL
NRAS
PHF6
PRPF40B
PTEN
RAD21
RUNX1
SETDB1
20 10 0
% of CHIP events
CHIP carriers (N = 88)
b
SETD2
1KGP
TET2
DNMT3A
NF1
KMT2D
TP53
PPM1D
CBL
0.4
0.3
0.2
0.1
0.0
Prevalence in EUR (%)
0.0
0.1
0.2
0.3
Prevalence in EAS (%)
tag
East_enrichment
neutral
West_enrichment
c
Mutations per MB
0.06
0.04
0.02
0.00
TET2
DNMT3A
SETD2
KMT2D
KMT2A
CREBBP
EP300
STAG2
CBL
GNA13
PRPF8
NF1
MPL
WT1
IKZF3
PPM1D
RUNX1
BCOR
SMC1A
GATA3
TP53
CEBPA
CBLB
SETDB1
PDS5B
CTCF
SRSF2
SETBP1
PHIP
IKZF1
BCORL1
NRAS
JAK1
SMC3
PTEN
PTPN11
FLT3
IDH2
U2AF2
SF3B1
IKZF2
ASXL1
GATA2
STAG1
CSF1R
CUX1
RAD21
BRCC3
GNB1
CSF3R
SF1
ETNK1
PRPF40B
JAK3
IDH1
ASXL2
GNAS
KIT
LUC7L2
BRAF
EZH2
PHF6
KDM6A
GATA1
10.0 7.5 5.0 2.5 0.0
% of CHIP events
CHIP carriers (N = 265)

Figure S6. The landscape of CHIP in global populations (whitelist variants identified in 1KGP 30x WGS 2,416 unrelated individuals). **a.** The landscape of CHIP in the validation cohort. The legend is consistent with Figure S3. **b.** The prevalence of CHIP genes (detected by whitelist variants) varied greatly between EUR and EAS populations. The top differentiated mutated CHIP genes were labeled with blue (EUR) and red (EAS) respectively. **c.** The landscape of expanded CHIP (i.e., all the functional variants in the canonical 74 CHIP genes, including “stop_gained”, “stop_lost”, “start_lost”, “frameshift_variant”, “splice_acceptor_variant”, “splice_donor_variant”, and “missense_variant”) in the validation cohort. The legend is consistent with Figure S3.

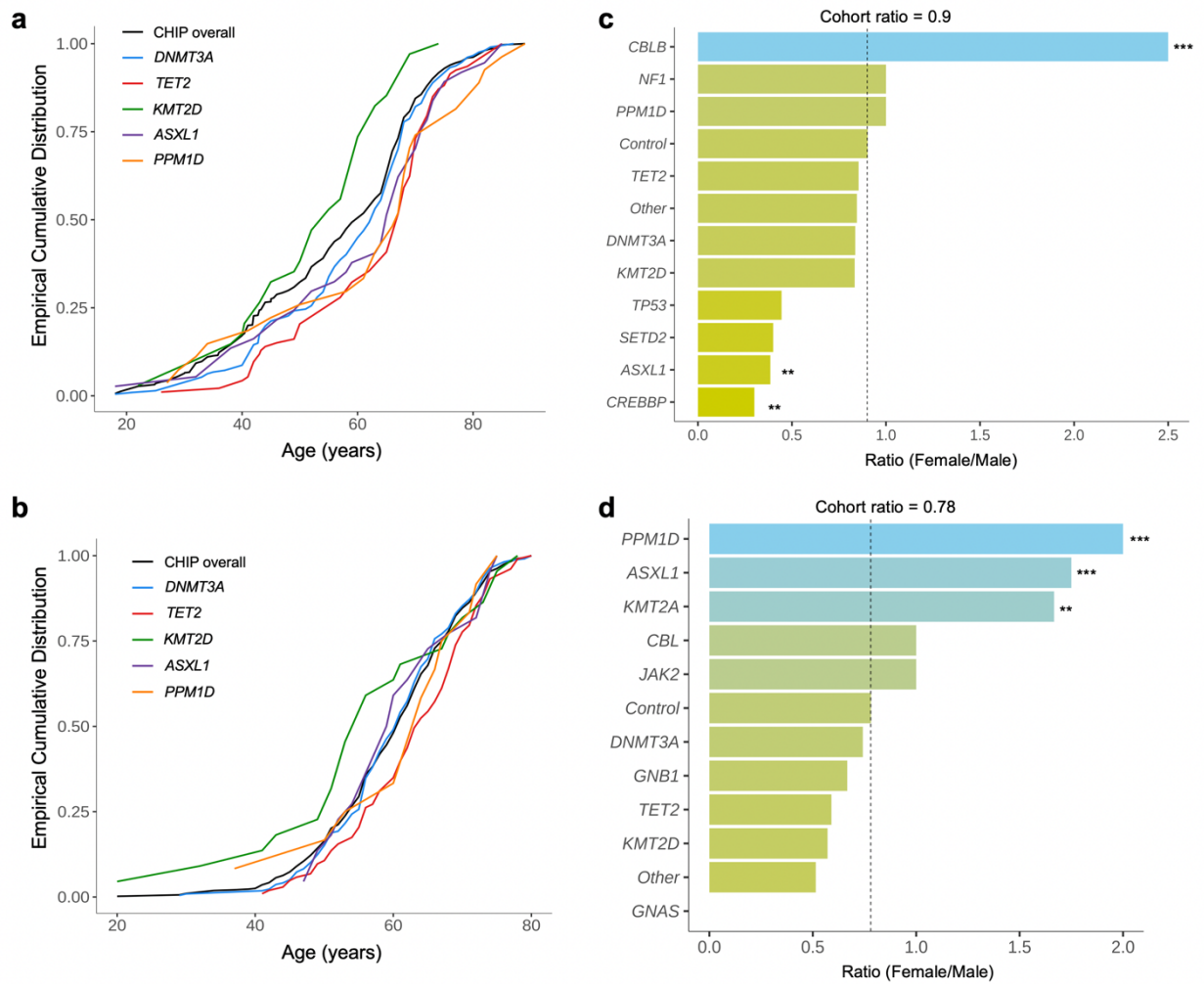


Figure S7. The heterogeneity of age– and sex–bias of CHIP in Chinese populations. Empirical cumulative distribution (ECD) of the age of individuals (discovery cohort in **a**; replication cohort in **b**) with CHIP overall (black) and stratified by the 5 most common driver genes. Bar plot showing the female to male (F:M) ratio of CHIP carriers with mutations in the ten most common driver genes (discovery cohort in **c**; replication cohort in **d**). 'Other' represents the remaining driver genes grouped together and 'Control' the ratio for individuals without CHIP. Dotted vertical line shows the F:M ratio observed in the full cohort, respectively. ***, P < 0.001; **, P < 0.01; *, P < 0.05. P–values are from a Chi–squared test comparing the distribution for each gene with 'Control.

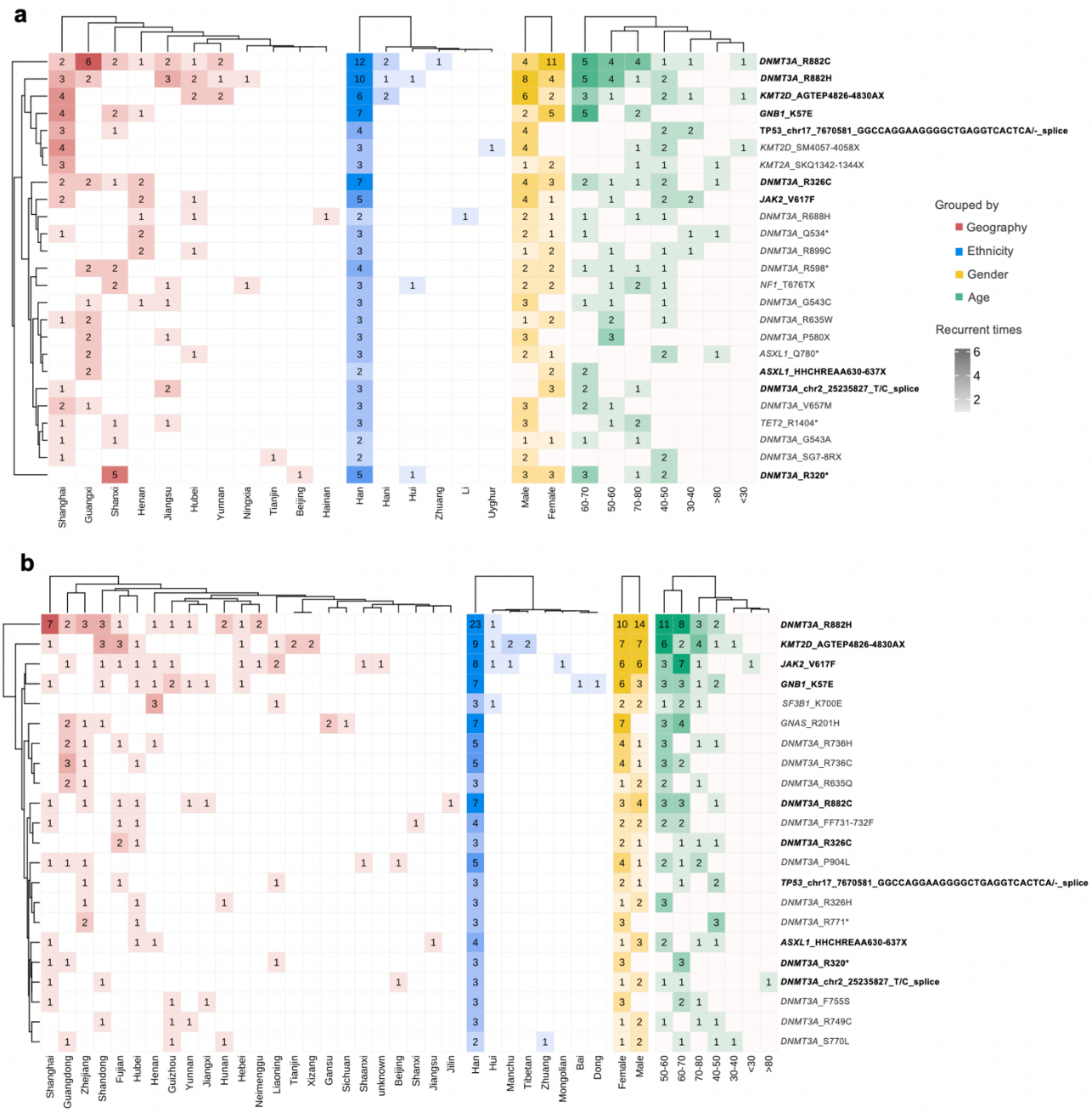


Figure S8. CHIP hotspots in Chinese populations in both discovery cohort (**a**) and replication cohort (**b**). The landscape of hotspots was consistently viewed from four perspectives (subplots): geographical region, ethnicity, gender, and age decades. A hotspot was defined as a recurrently mutated region in at least three individuals in each hotspot, respectively. The amino acid change was noted if the variant resulted in a protein alteration, and the chromosome location with nucleotide change was recorded if the

variant was annotated without a protein alteration. The Hotspots identified by both the discovery and duplicate cohorts is marked in bold.

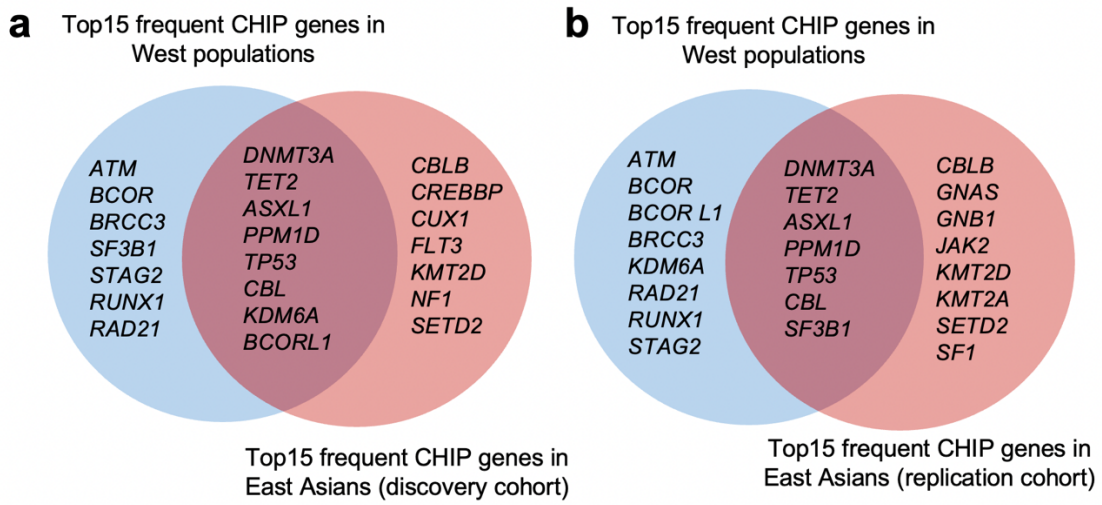


Figure S9. The Venn plot between top 15 frequent CHIP genes detected in Western (related to Figure S1) and Eastern populations of discovery (**a**) and replication cohort (**b**), in which only eight and seven genes were intersected respectively. Therefore, we took the union of all these 22 or 23 genes as the final frequent CHIP genes to character the CHIP differences between the West and East (related to Figure 2 and Figure S16), respectively.

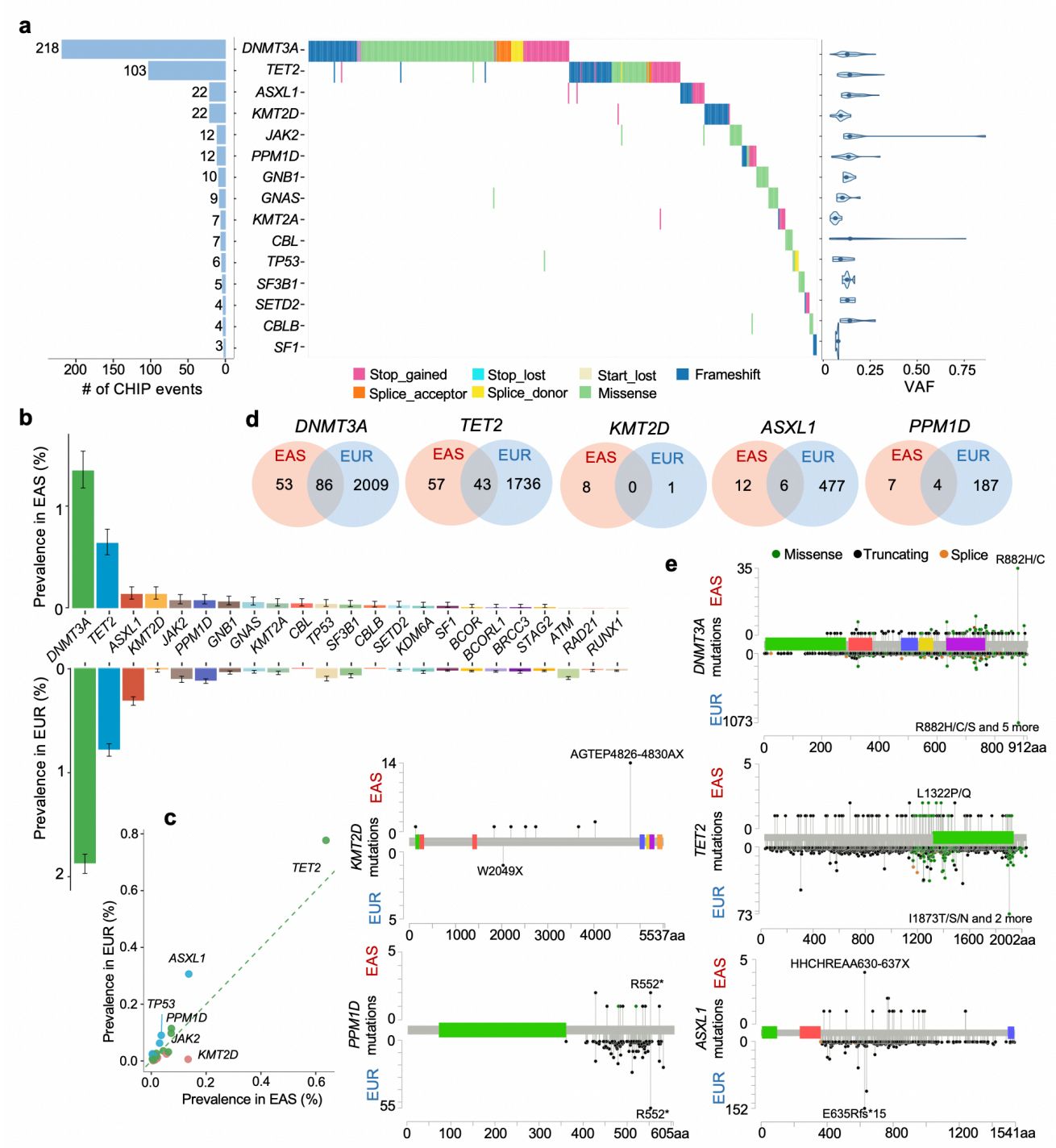


Figure S10. Characterization of CHIP in the Chinese populations (replication cohort). **a.** Composite plot summarizing mutations in the 15 most common driver genes mutated in Chinese CHIP carriers. Each column in the waterfall plot represents a single individual, with mutation types of color–coded. Bars on the left quantify mutations per gene as a mutation counts detected in corresponding genes. Violin plots on the right show the

distribution of variant allele frequency (VAF), with dots representing mean value. **b.** The comparison of the prevalence of 23 frequent CHIP genes between the Western and Eastern populations. CHIP prevalence in the Western populations was referred to the merged cohort of Jaiswal et al[14] and Bick et al[8]. The color saturation refers to the preference of the population. **c.** CHIP variants (SNV/Indels) varied greatly between Western and Eastern populations. The top differentiated mutated CHIP genes were labeled with blue (Western) and red (Eastern) respectively. **d.** Venn diagram of the distribution of specific CHIP variants detection between the Western and Eastern populations of the top five frequent genes. **e.** Lollipop plot of the detailed CHIP variants identified in the Western and Eastern populations.

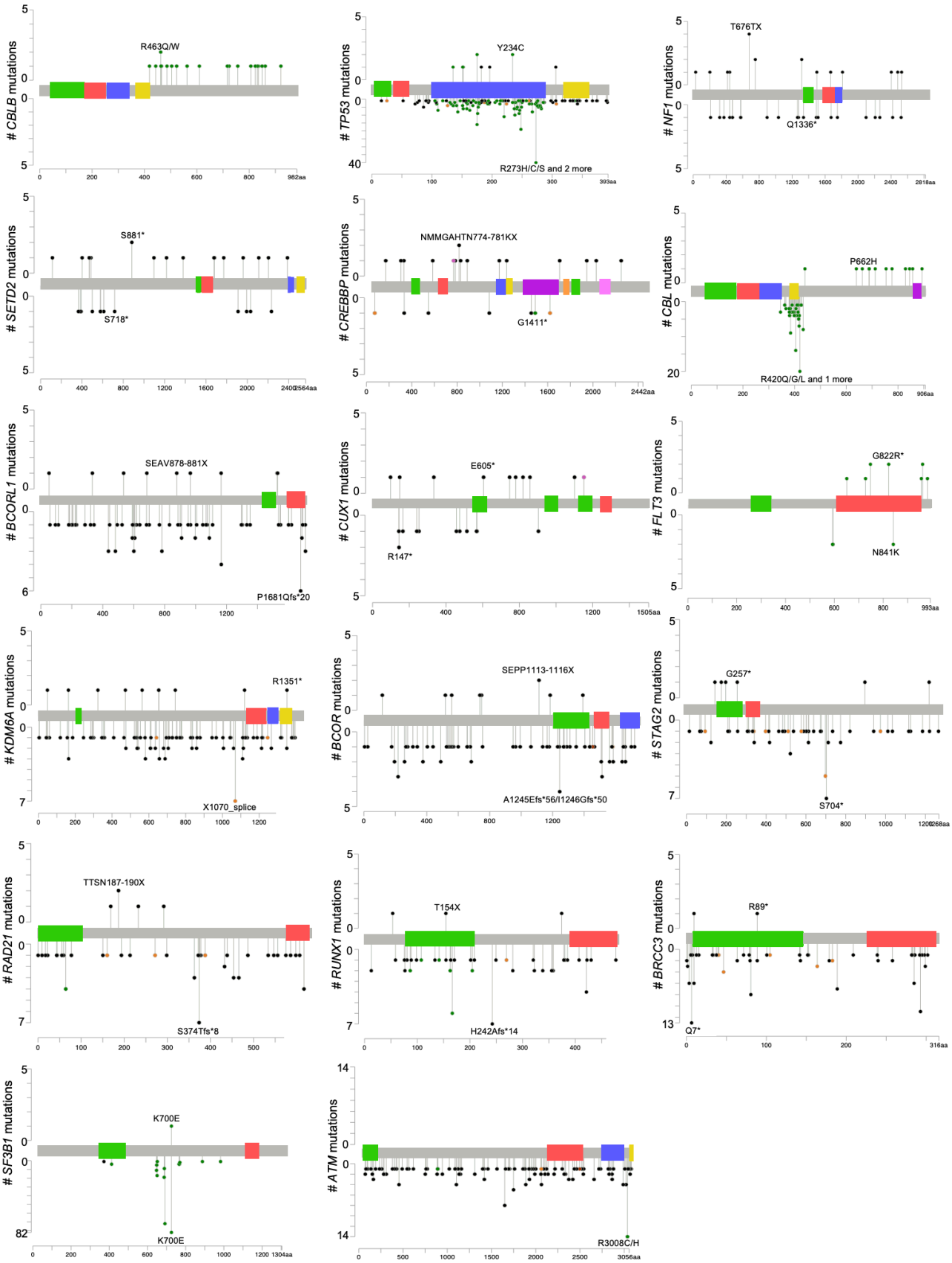

CBLB mutations
R463Q/W
982aa
TP53 mutations
Y234C
R273H/C/S and 2 more
393aa
NF1 mutations
T676TX
Q1336*
2818aa
SETD2 mutations
S881*
S718*
2564aa
CREBBP mutations
NMMGAHTN774-781KX
G1411*
2442aa
CBL mutations
P662H
R420Q/G/L and 1 more
906aa
BCORL1 mutations
SEAV878-881X
P1681Qfs*20
CUX1 mutations
E605*
R147*
1505aa
FLT3 mutations
G822R*
N841K
993aa
KDM6A mutations
R1351*
X1070_splice
BCOR mutations
SEPP1113-1116X
A1245Efs*56/I1246Gfs*50
STAG2 mutations
G257*
S704*
1268aa
RAD21 mutations
TTSN187-190X
S374Tfs*8
RUNX1 mutations
T154X
H242Afs*14
BRCC3 mutations
R89*
Q7*
316aa
SF3B1 mutations
K700E
K700E
1304aa
ATM mutations
R3008C/H
3056aa

Figure S11. Lollipop plot of the detailed CHIP variants identified in the Western and Eastern populations (discovery cohort). The Y–axis above represents the number of variants detected in our discovery cohort, and the below represents the number of variants detected in Western populations (N = 315,326).

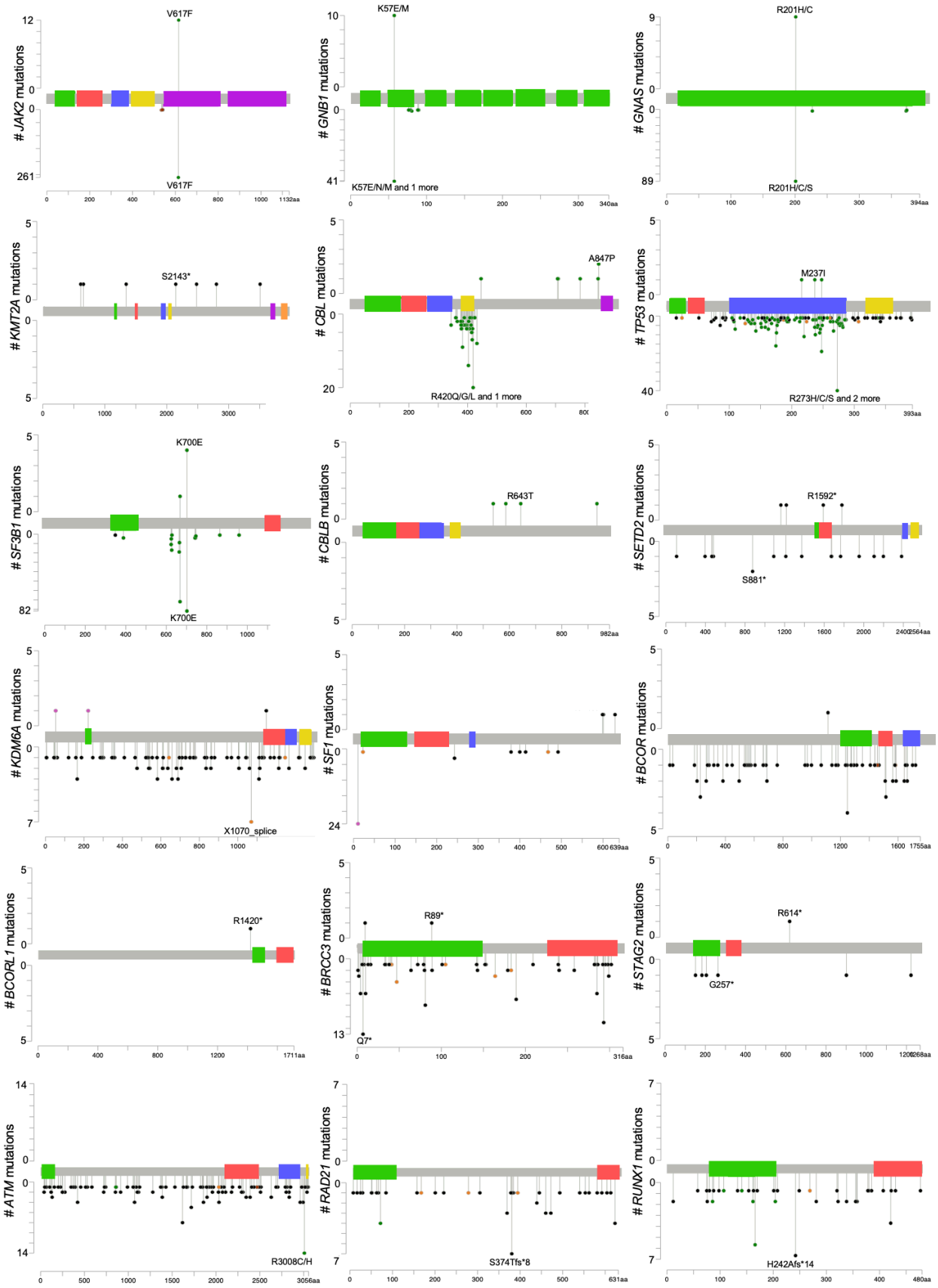
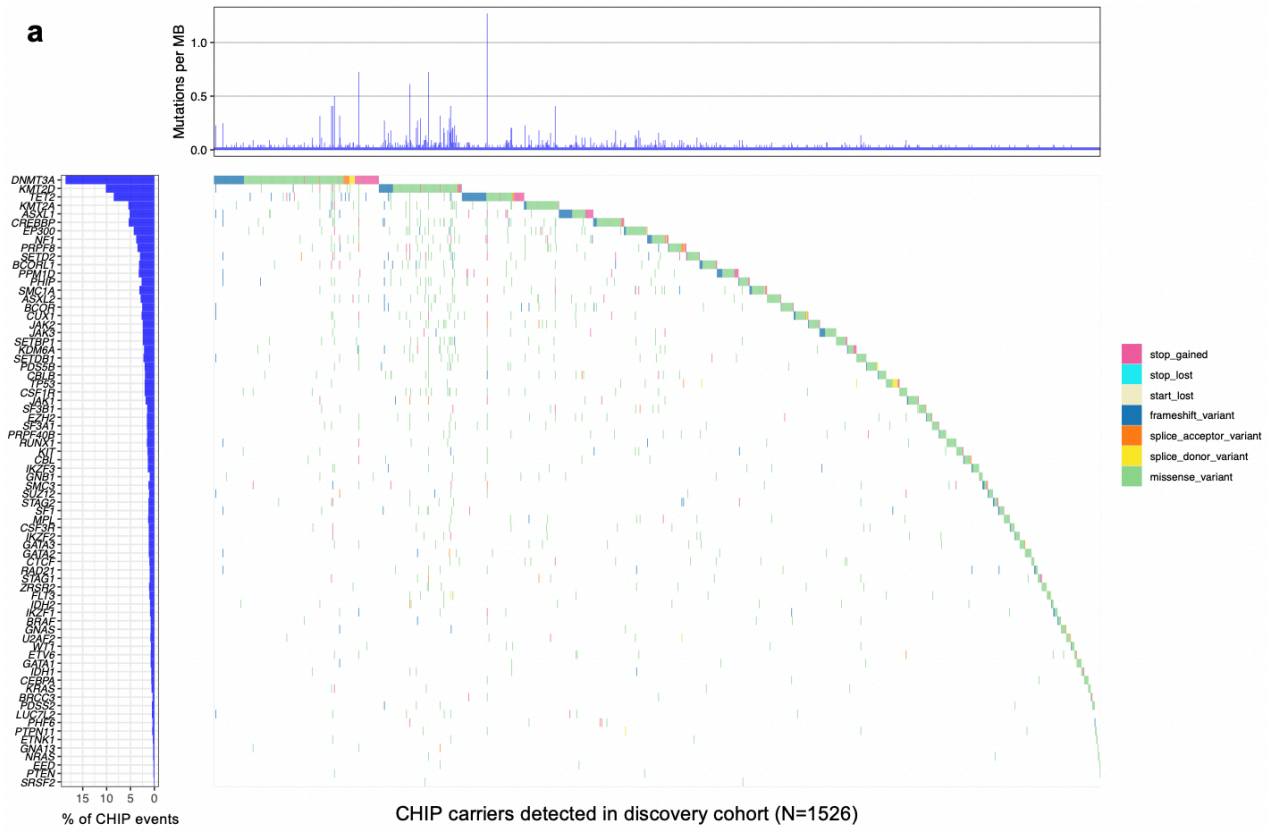

#JAK2 mutations
V617F
V617F
#GNB1 mutations
K57E/M
K57E/N/M and 1 more
#GNAS mutations
R201H/C
R201H/C/S
#KMT2A mutations
S2143*
#CBL mutations
A847P
R420Q/G/L and 1 more
#TP53 mutations
M237I
R273H/C/S and 2 more
#SF3B1 mutations
K700E
K700E
#CBLB mutations
R643T
#SETD2 mutations
R1592*
S881*
#KDM6A mutations
X1070_splice
#SF1 mutations
#BCOR mutations
#BCORL1 mutations
R1420*
#BRCC3 mutations
R89*
Q7*
#STAG2 mutations
R614*
G257*
#ATM mutations
R3008C/H
#RAD21 mutations
S374Tfs*8
#RUNX1 mutations
H242Afs*14

Figure S12. Lollipop plot of the detailed CHIP variants identified in the Western and Eastern populations (replication cohort). The legend is consistent with Figure S11.


a
Mutations per MB
1.0
0.5
0.0
DNMT3A
KMT2D
TET2
KMT2A
ASXL1
CREBBP
EP300
NF1
PRPF8
SETD2
BCORL1
PPM1D
PHIP
SMC1A
ASXL2
BCOR
CUX1
JAK2
JAK3
SETBP1
KDM6A
SETDB1
PDS5B
CBLB
CBL
CSF1R
JAK1
SF3B1
EZH2
SF3A1
PRPF40B
RUNX1
KIT
NF1
IKZF3
GNB1
SMC3
SUZ12
STAG2
U2AF2
SF1
MPL
CSF3R
IKZF2
IKZF1
GATA3
GATA2
CTCF
RAD21
STAG1
ZRSR2
FLT3
IDH2
IKZF1
BRAF
GNAS
U2AF1
WT1
ETV6
GATA1
IDH1
CEBPA
KRAS
BRCC3
PDS5A
LUC7L2
PHF6
PTPN11
ETNK1
GNA13
NRAS
EED
PTEN
SRSF2
15 10 5 0
% of CHIP events
stop_gained
stop_lost
start_lost
frameshift_variant
splice_acceptor_variant
splice_donor_variant
missense_variant
CHIP carriers detected in discovery cohort (N=1526)

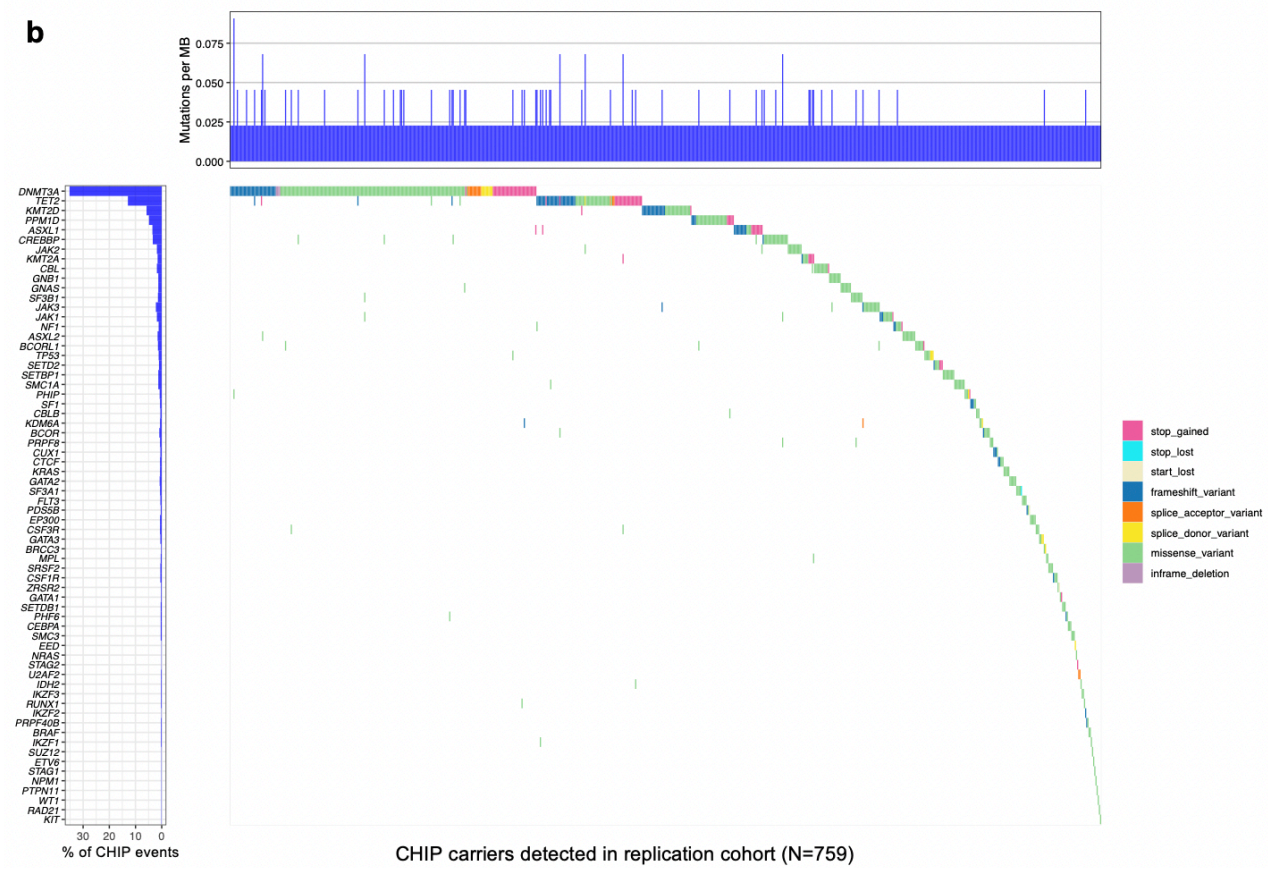

b
Mutations per MB
0.075
0.050
0.025
0.000
DNMT3A
TET2
KMT2D
PPM1D
ASXL1
CREBBP
JAK2
KMT2A
CBL
GNB1
GNAS
SF3B1
JAK3
JAK1
NF1
ASXL2
BCORL1
TP53
SETD2
SETBP1
SMC1A
PHIP
SF1
CBLB
KDM6A
BCOR
PRPF8
CUX1
CTCF
KRAS
GATA2
SF3A1
FLT3
PDS5B
EP300
CSF3R
GATA3
BRCC3
MPL
SRSF2
CSF1R
ZRSR2
GATA1
SETDB1
PHF6
CEBPA
SMC3
EED
NRAS
STAG2
U2AF2
IDH2
IKZF3
RUNX1
IKZF2
PRPF40B
BRAF
IKZF1
SUZ12
ETV6
STAG1
NPM1
PTPN11
WT1
RAD21
KIT
30 20 10 0
% of CHIP events
stop_gained
stop_lost
start_lost
frameshift_variant
splice_acceptor_variant
splice_donor_variant
missense_variant
inframe_deletion
CHIP carriers detected in replication cohort (N=759)

Figure S13. The landscape of CHIP of expanded variants detected in the discovery (**a**) and replication cohort (**b**). The legend is consistent with Figure S3.

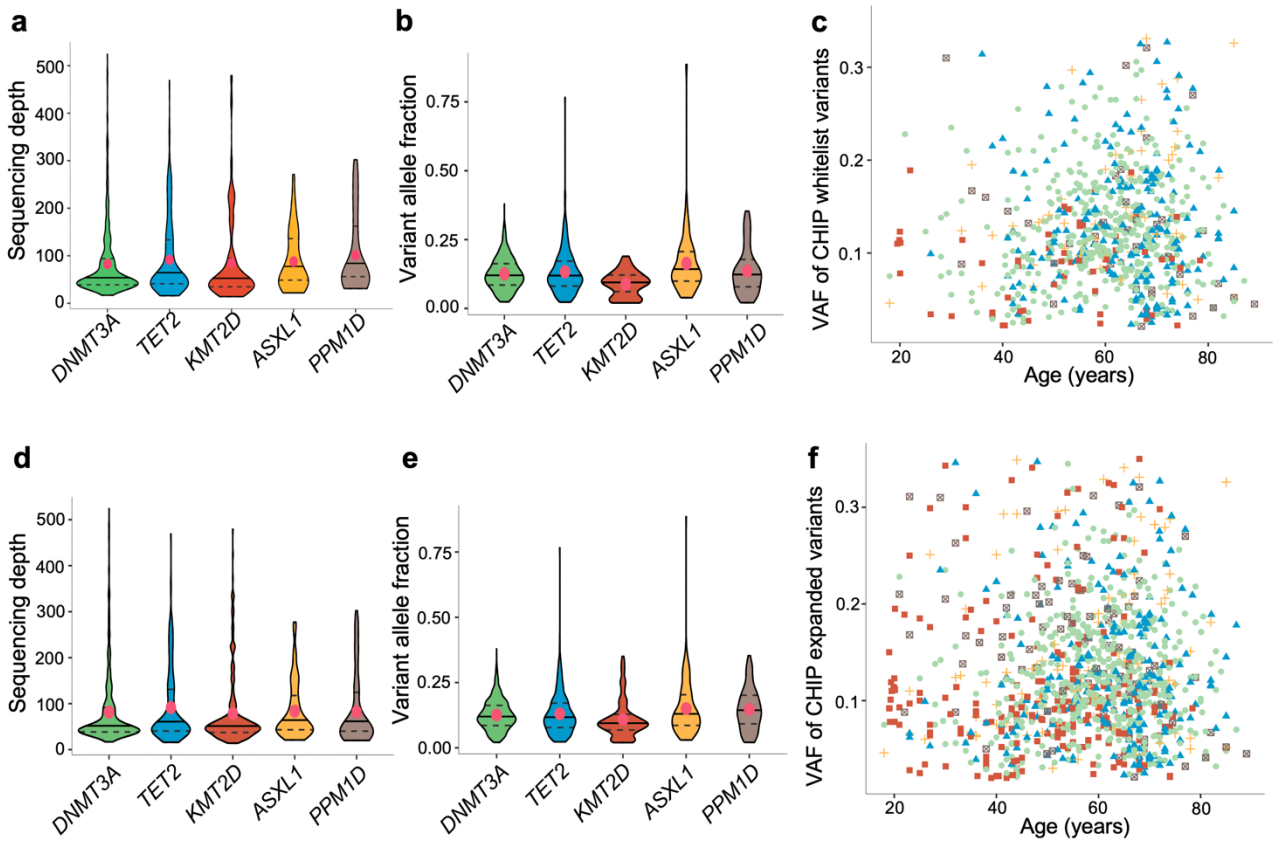


Figure S14. Sequencing depth, variant allele fraction (VAF), and VAF~age distributions for the top five CHIP genes detected with whitelist (**a**, **b**, **c**) and expanded variants (**d**, **e**, **f**) in merged cohorts (i.e., merged POG13K and CYN16K, N = 20,596).

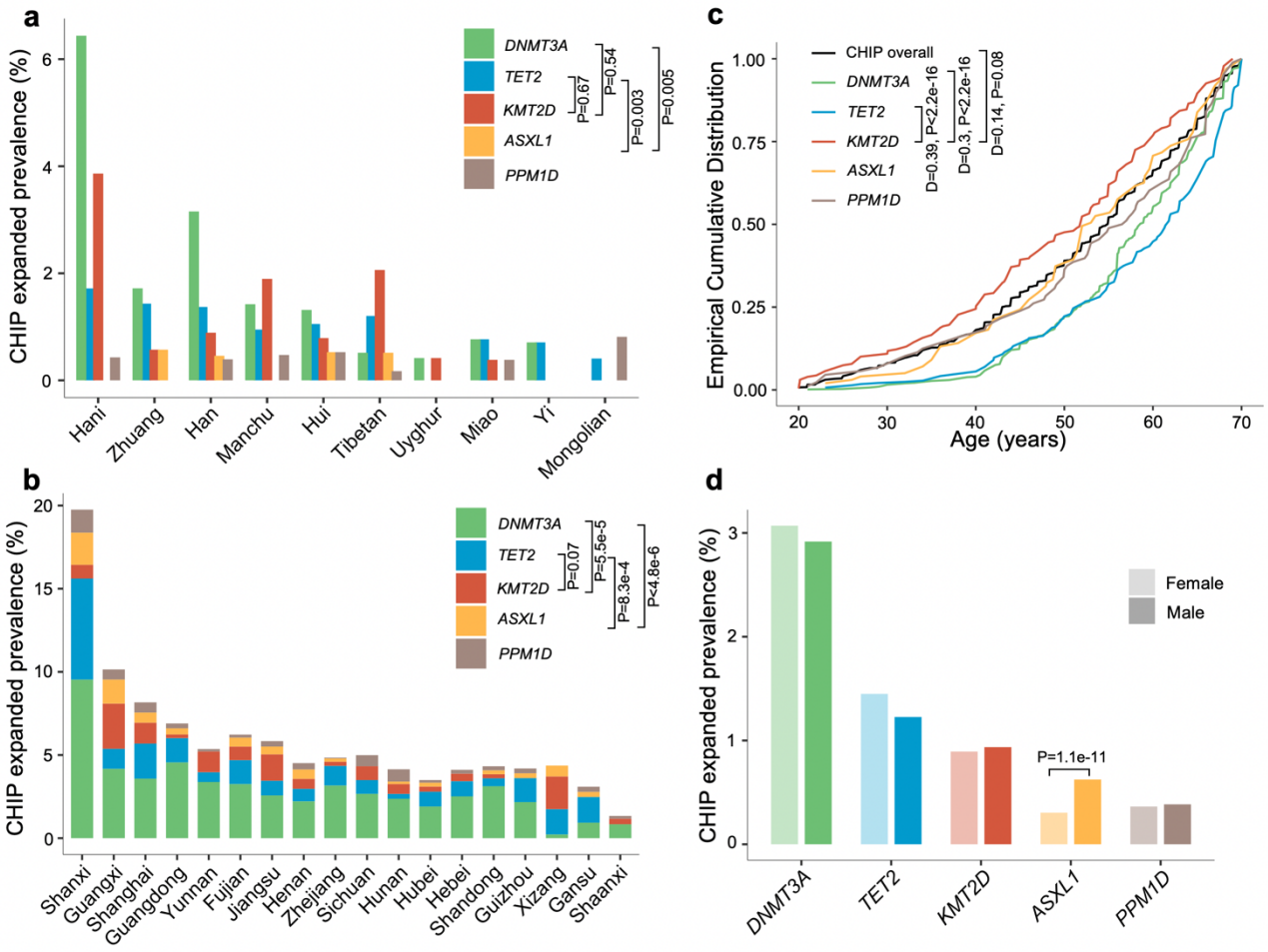


Figure S15 Comparable ethnic, geographic, age, and sex distributions for *KMT2D* and other top frequent CHIP genes in Chinese populations (expanded CHIP in merged cohorts of 29,596 individuals). **a**. The expanded CHIP prevalence of top five frequent genes detected in ethnic subgroups (only ethnic subgroups with a sample size > 200 were reported) . Two–sided Wilcoxon rank–sum test for comparing distributions between groups. **b**. The expanded CHIP prevalence of top five frequent genes detected in geographic regions (only regions with a sample size > 400 were reported). Differences in distributions across geographic regions were assessed with the two–sample Kolmogorov–Smirnov (K–S) test. **c.** The age distribution of CHIP carriers detected with top five frequent genes. Differences in distributions across ages were assessed with the two–sample K–S test. **d.** The expanded CHIP prevalence of top five frequent genes detected in females and males. A Chi–squared test was used to evaluate differential carrier frequencies between sexes for each gene.

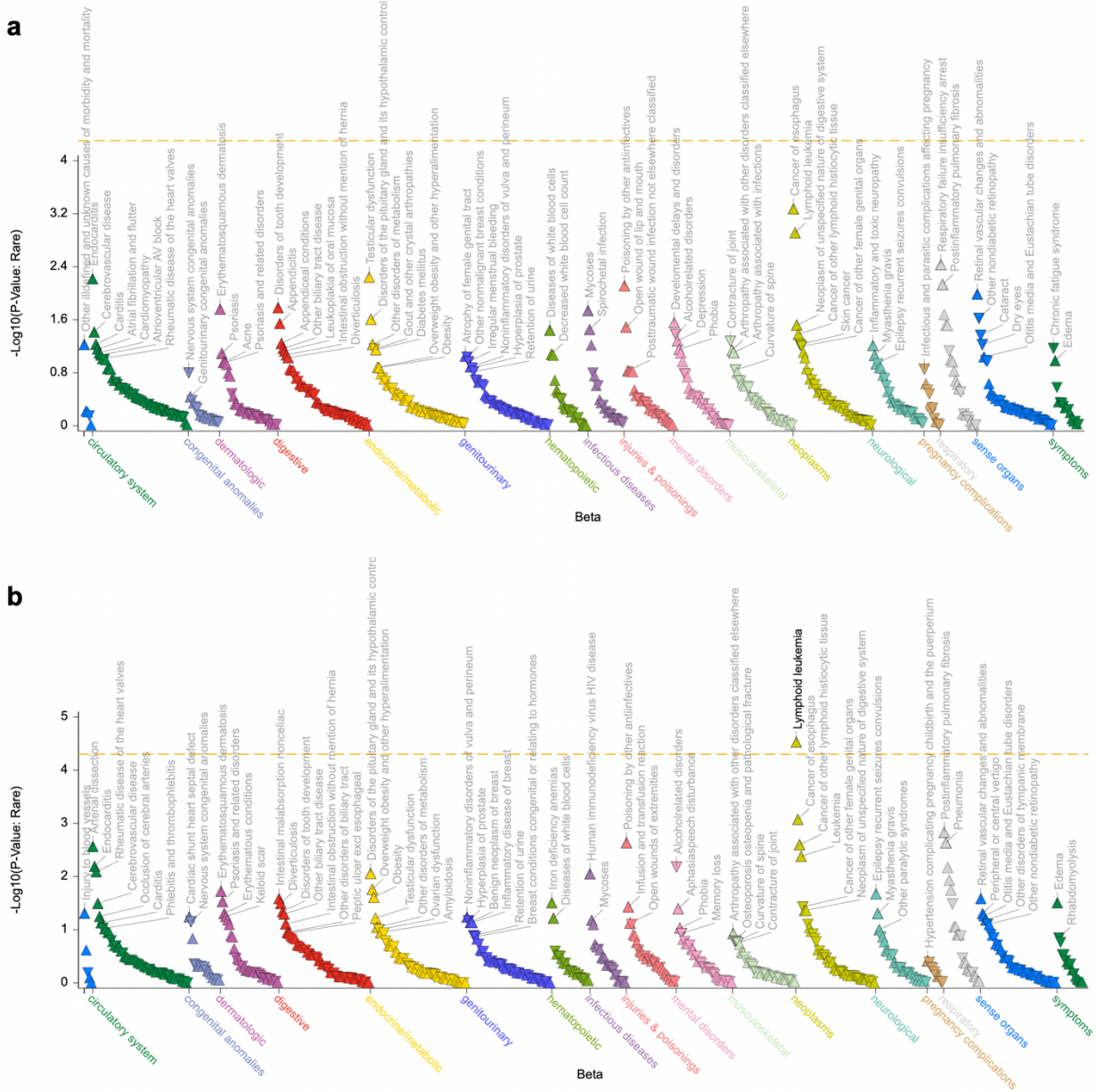


Figure S16. *KMT2D* was significantly associated with leukemia when the ancestry was changed from European only **a.** to the Mixed ancestry **b.** with African, American and Asian, from the Human Disease Knowledge Portal[20] (https://hugeamp.org:8000/research.html?pageid = 600_traits_app_home), suggesting an increased role of *KMT2D* in individuals of non–European ancestry. The cohorts included UK Biobank, All of US, and Mass General Brigham Biobank (MGB)[20].

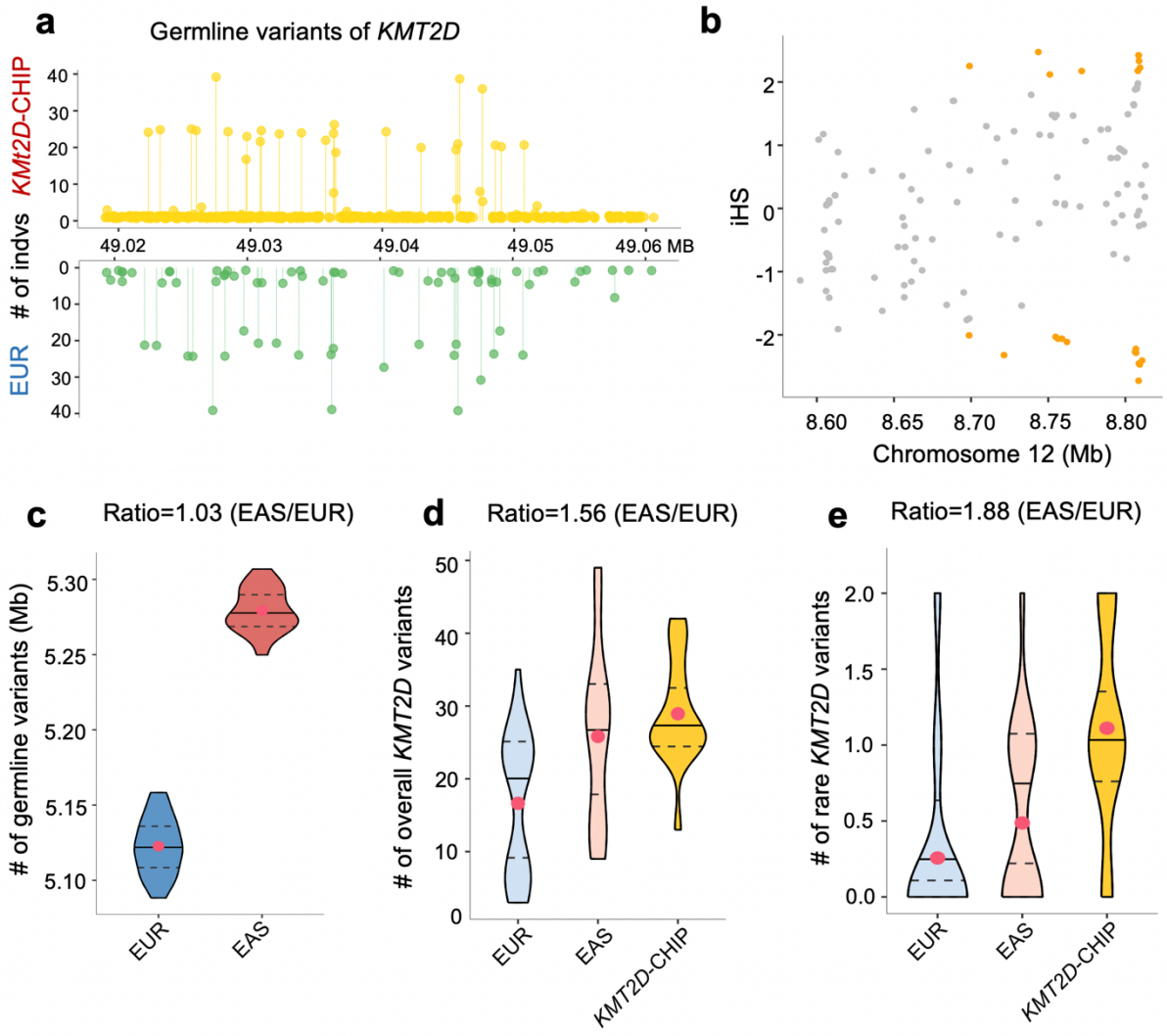


Figure S17. East–West differentiated *AICDA* may contribute to the higher variants burden in *KMT2D* in East Asians (related to Figure 4). **a.** Germline variants of *KMT2D* in East Asians (*KMT2D*–CHIP carriers in POG13K) and Europeans (1KGP, EUR). **b.** Genetic differentiation of *AICDA*. iHS (Integrated Haplotype Score) value of *ACIDA* and its flanking regions (± 200kb). The points of absolute value of iHS > 2 were labeled with color. **c.** The distribution of genome–wide germline mutation burdens of EAS and EUR (1KGP). **d**. The distribution of genome–wide germline mutation burdens of *KMT2D* (All variants were included) in EAS with *AICDA* highly differentiated loci and EUR, as well as that in *KMT2D*–CHIP cases in our Chinese cohort. **e.** The distribution of genome–wide germline rare variant burden of *KMT2D* in EAS with *AICDA* highly differentiated loci and EUR, as well as that in *KMT2D*–CHIP cases in our Chinese cohort.

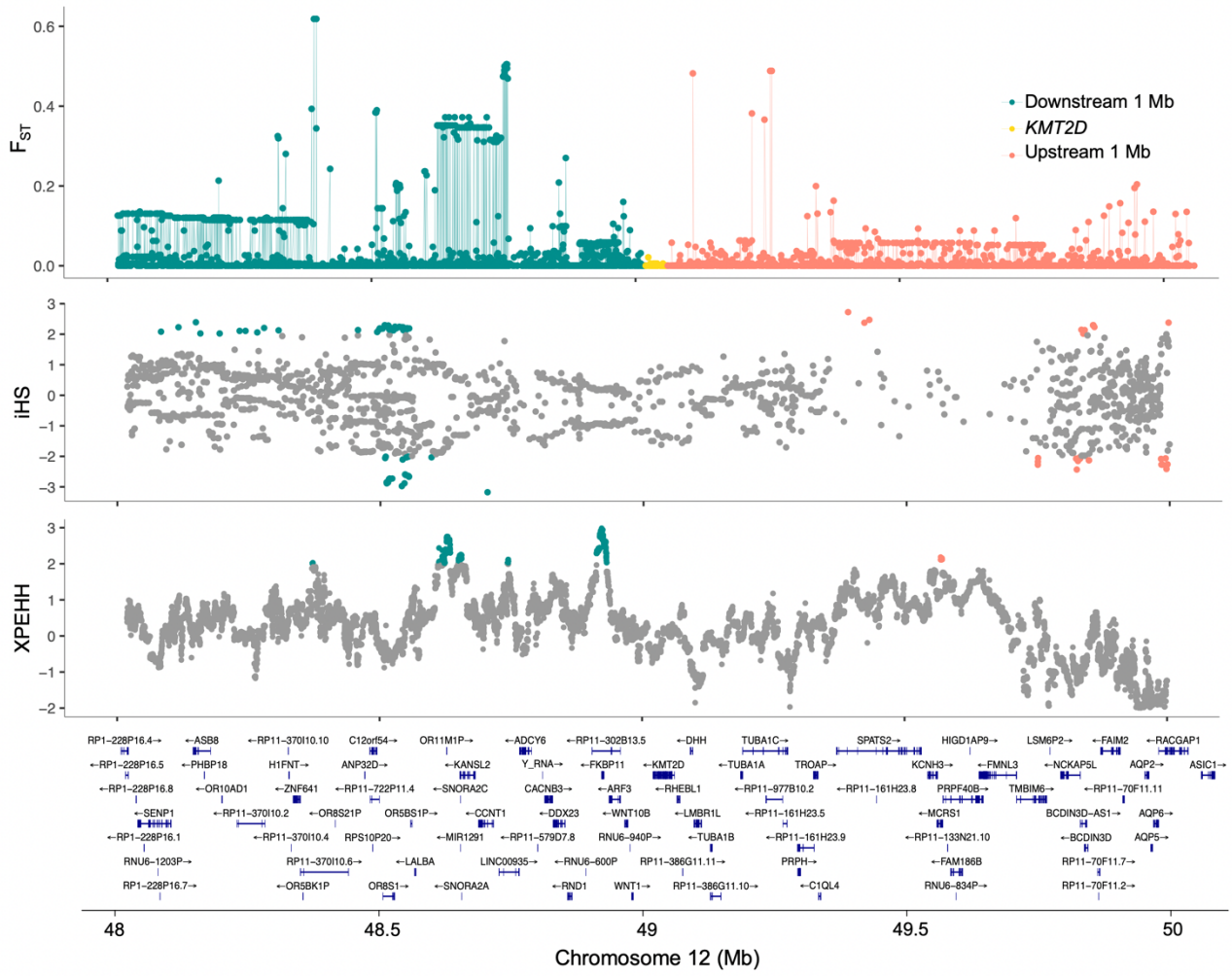


Figure S18. Genetic differentiation of *KMT2D* and its flanking regions. Fixation Index (F_{ST}), Integrated Haplotype Score (iHS), Cross Population Extended Haplotype Homozygosity (XPEHH), and the covered genes of *KMT2D* and its flanking regions (± 1Mb). The points of absolute value of iHS > 2, and the XPEHH value > 2 was labeled with color.

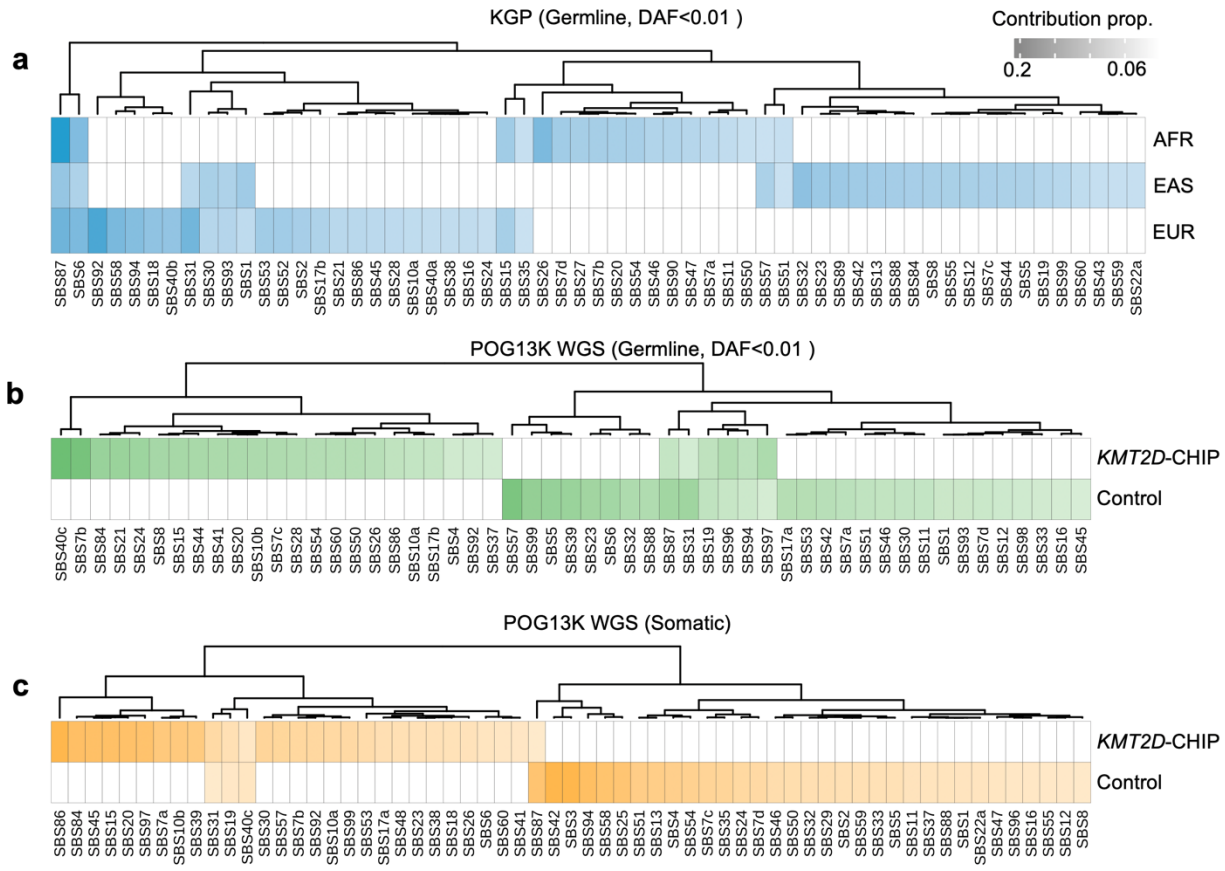


Figure S19. The mutation signature (single base substitution, SBS) of *KMT2D* variants among EAS, EUR, and AFR (**a**) from 1KGP, and between *KMT2D*–CHIP cases and non–carriers within our cohorts (germline in **b** and somatic in **c**), with bootstrapping 100 times, respectively. EAS, East Asians; EUR, Europeans; AFR, Africans.

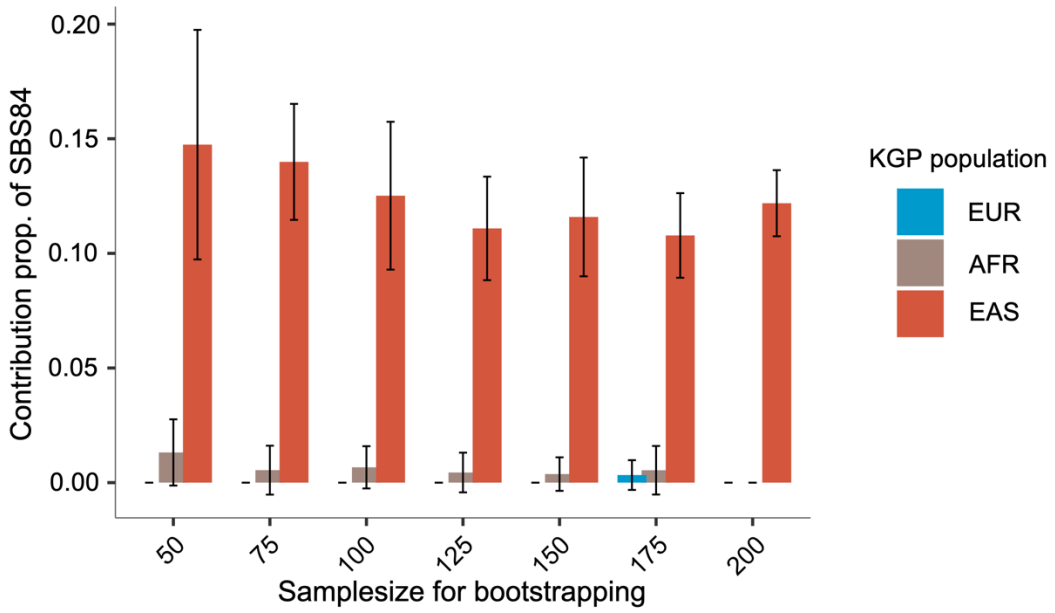


Figure S20. The relative contribution of SBS84 in *KMT2D* variants among EAS, EUR, and AFR. The sample size for bootstrapping was from 50–200, 100 times per gradient. EAS, East Asians; EUR, Europeans; AFR, Africans.

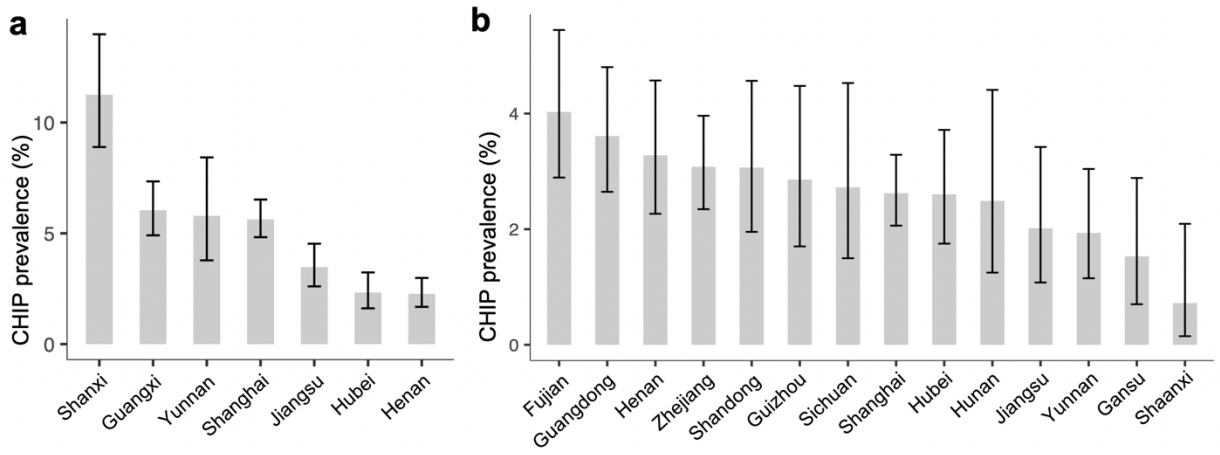


Figure S21. The distribution of population prevalence of CHIP among geographical regions in Chinese populations in discovery (**a**) and replication cohort (**b**). The distribution of CHIP prevalence in different regions of China varies greatly but without significant correlation with latitude and longitude. The bar plot indicated the CHIP prevalence, and which only regions with sample size at least 400 individuals are shown.

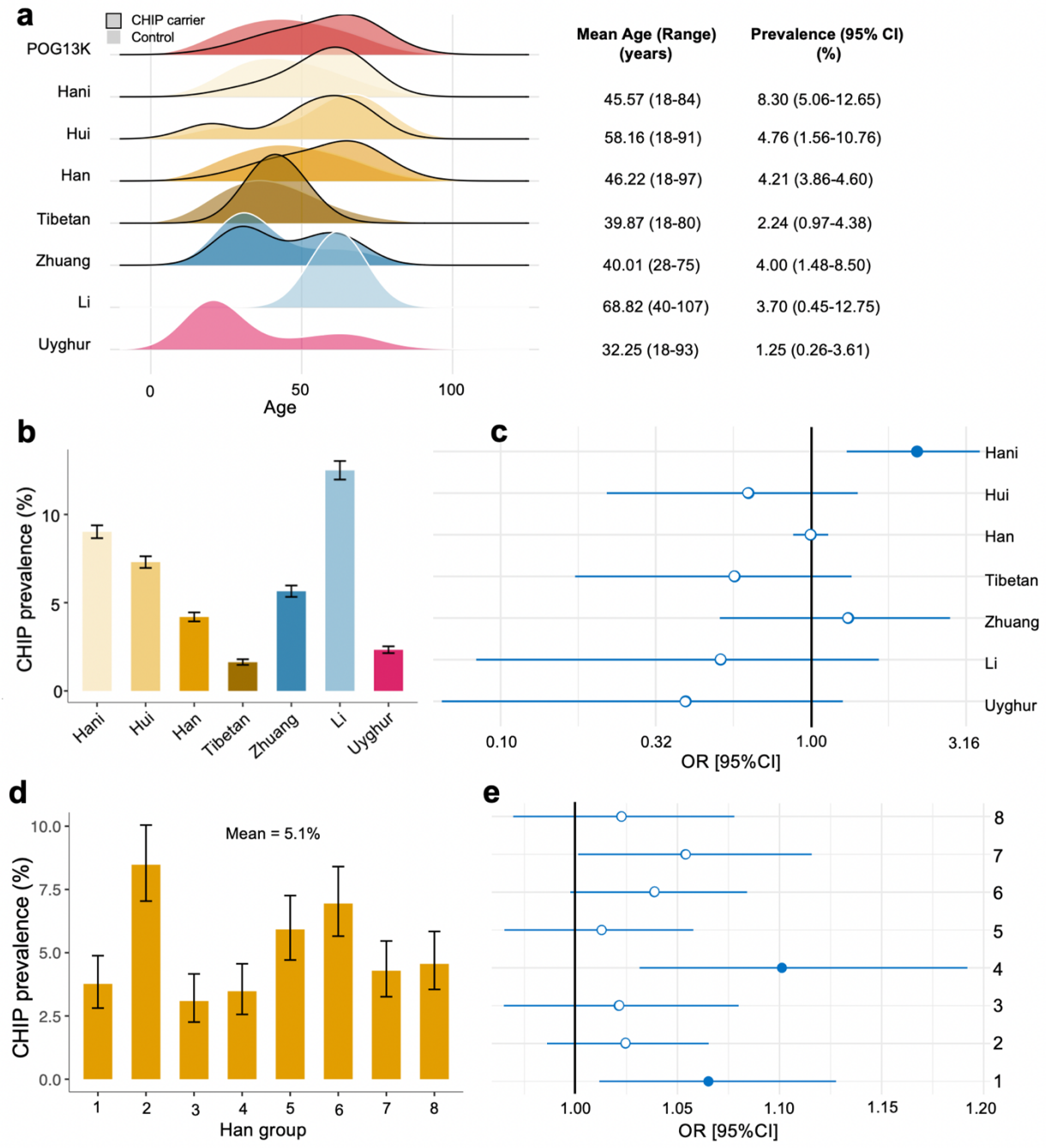


Figure S22. CHIP heterogeneity within ethnic subgroups in discovery cohort. **a**. The density of age distribution between all samples and CHIP carriers in the discovery cohort, and ethnic minorities, respectively. The mean age, age range, and CHIP prevalence were indicated in the right, correspondingly. **b.** The adjusted CHIP prevalence within ethnic subgroups in discovery cohort, adjusted for age and sex, related to Figure 1E. **c.** The odds ratios and unadjusted two–sided P values were derived from a logistic regression

model with all CHIP as the outcome and ethnic cohort as the predictor, adjusted for age, and sex. Solid circles represent significant associations ($P < 0.05$); hollow circles represent non–significant associations ($P \geq 0.05$). **d.** The adjusted CHIP prevalence within Han subgroup, which was randomly sampled into eight groups according to the mean sample size of minority groups and the prevalence was similarly adjusted. **e**. The legend is consistent with (**c**), while the ethnic group was Han. The two–sample Kolmogorov–Smirnov (K–S) was tested for (**b**) and (**d**) with P value of 0.98, and for (**c**) and (**e**) with P value of 0.03.

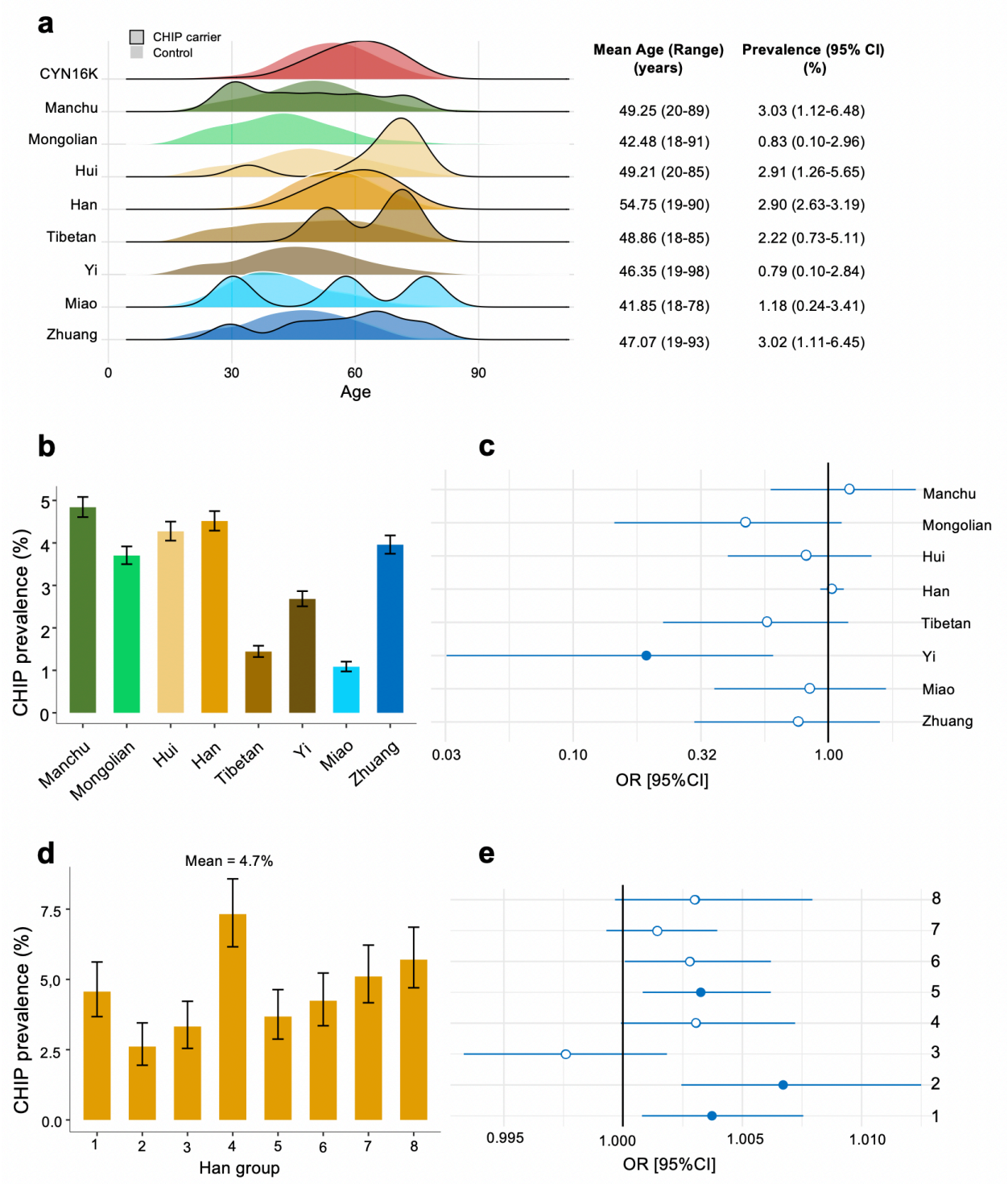


Figure S23. CHIP heterogeneity within ethnic subgroups in replication cohort, related to Figure S4E. The legend is consistent with Figure S10. The two–sample Kolmogorov–Smirnov (K–S) was tested for (**b**) and (**d**) with P value of 0.5, and for (**c**) and (**e**) with P value of 0.01.

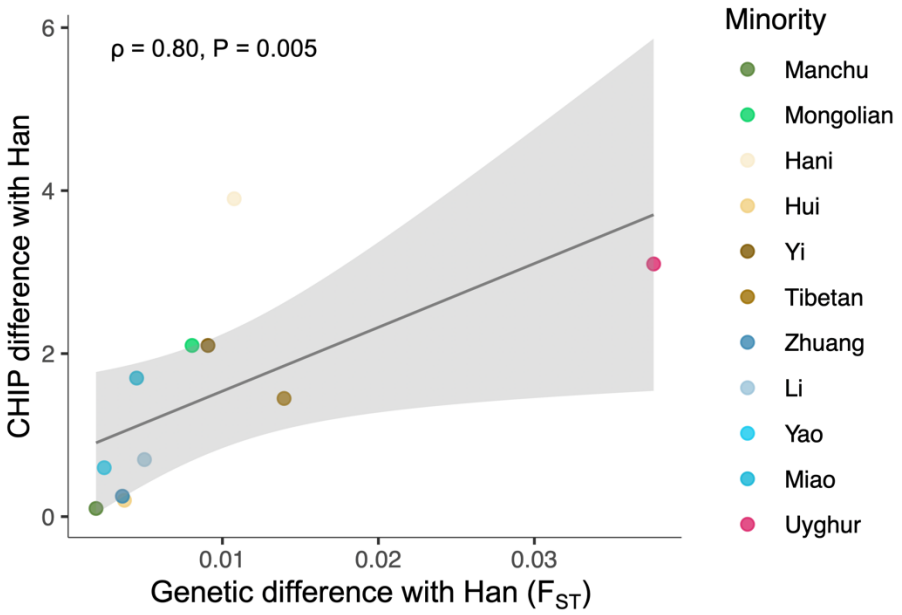


Figure S24. The correlation between the genetic difference and the CHIP difference between ethnic minorities and Han Chinese. For each minority, the CHIP difference was defined as the absolute difference in CHIP prevalence from Han Chinese. If a minority was reported in both the discovery and replication cohorts, its CHIP difference was averaged across the two cohorts. Genetic differentiation was quantified by F_{ST} (Fixation Index) between the minority and Han Chinese. Spearman's rank correlation was used to test the association (ρ = 0.80, P = 0.005).

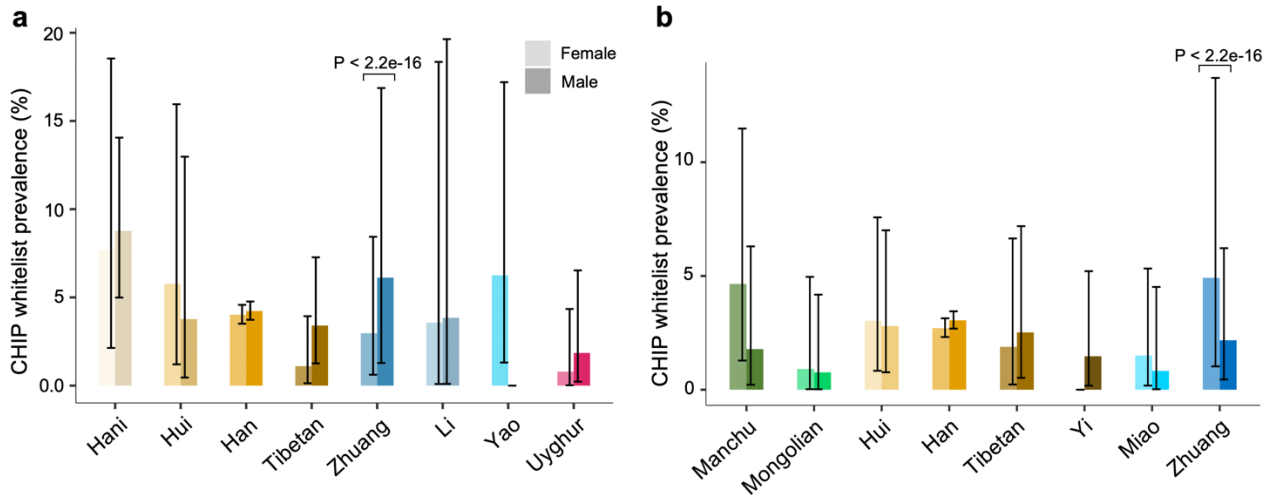


Figure S25. The whitelist CHIP prevalence detected in females and males in ethnic subgroups (discovery cohort in **a**; replication cohort in **b**). A Chi–squared test was used to evaluate differential carrier frequencies between sexes for each gene. Only female CHIP carriers were detected in Yao populations (N = 80).

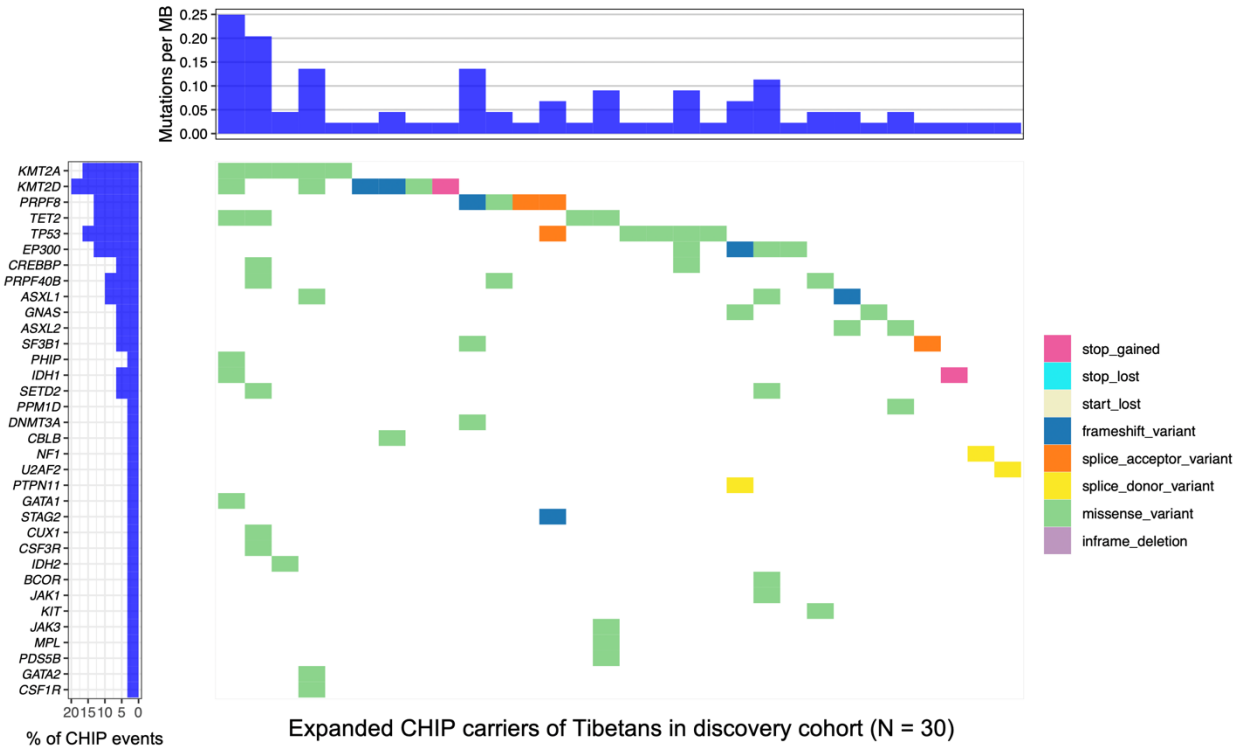


Figure S26. The landscape of expanded CHIP variants detected in Tibetans in discovery cohort. The legend is consistent with Figure S3.

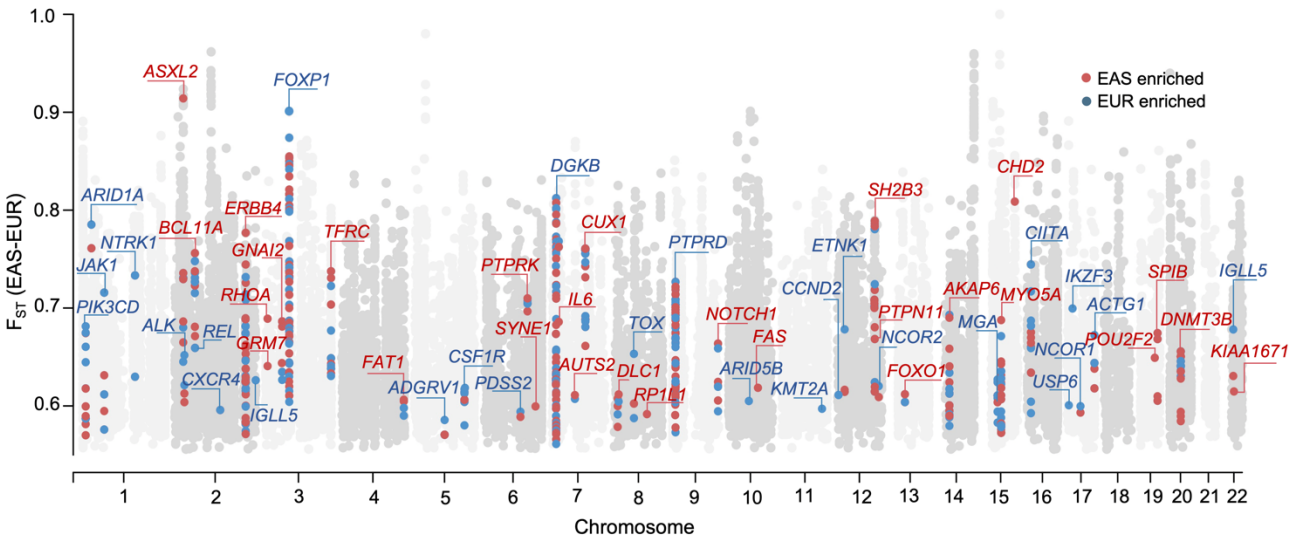


Figure S27. Genetic differentiation of CHIP in Chinese populations. Significant differentiation between East Asian (EAS) and European populations (EUR) (using 1KG, and CEU vs. CHB). The genome–wide SNPs with fixation index (F_{ST}) values ranked top 0.1% (i.e., $F_{ST} > 0.55$) are plotted as the gray background. The SNPs located on expanded CHIP genes are highlighted with colored dots, with blue dots representing a significantly higher frequency of this locus in EUR and red dots representing a significantly higher frequency in EAS.

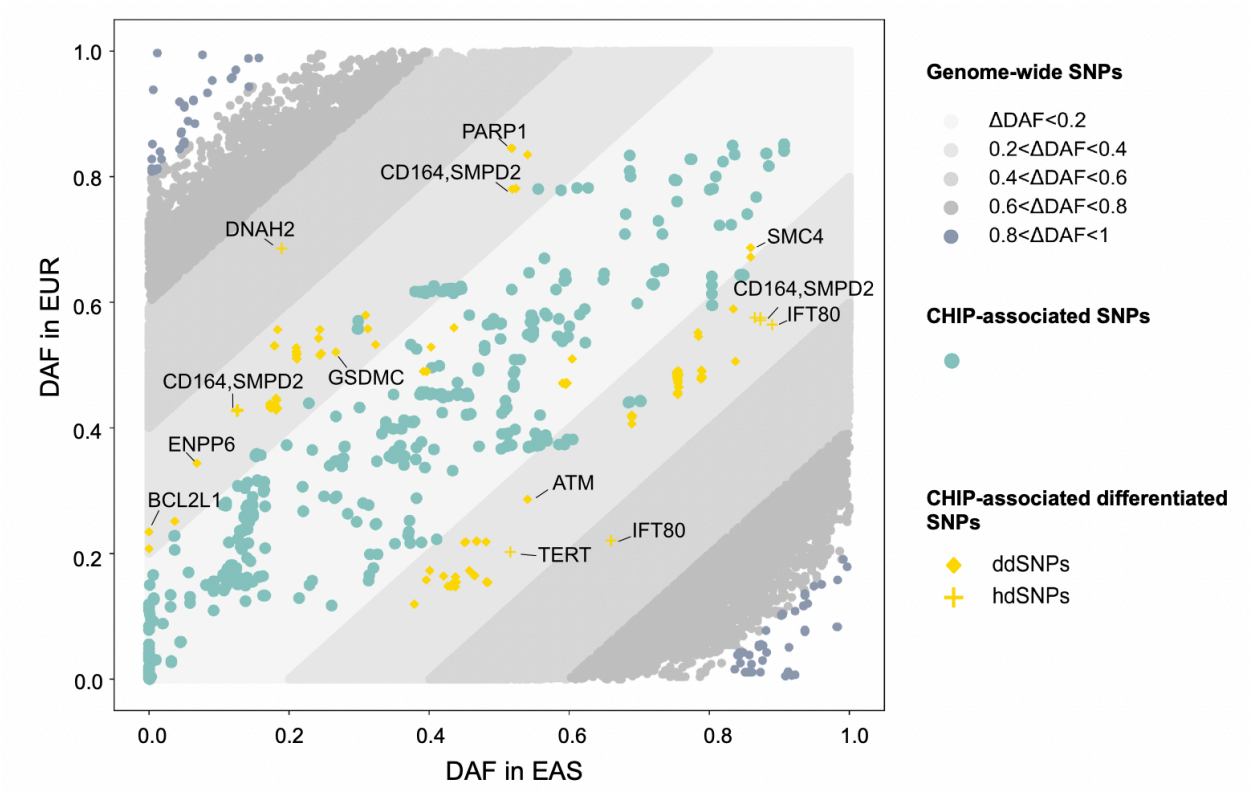
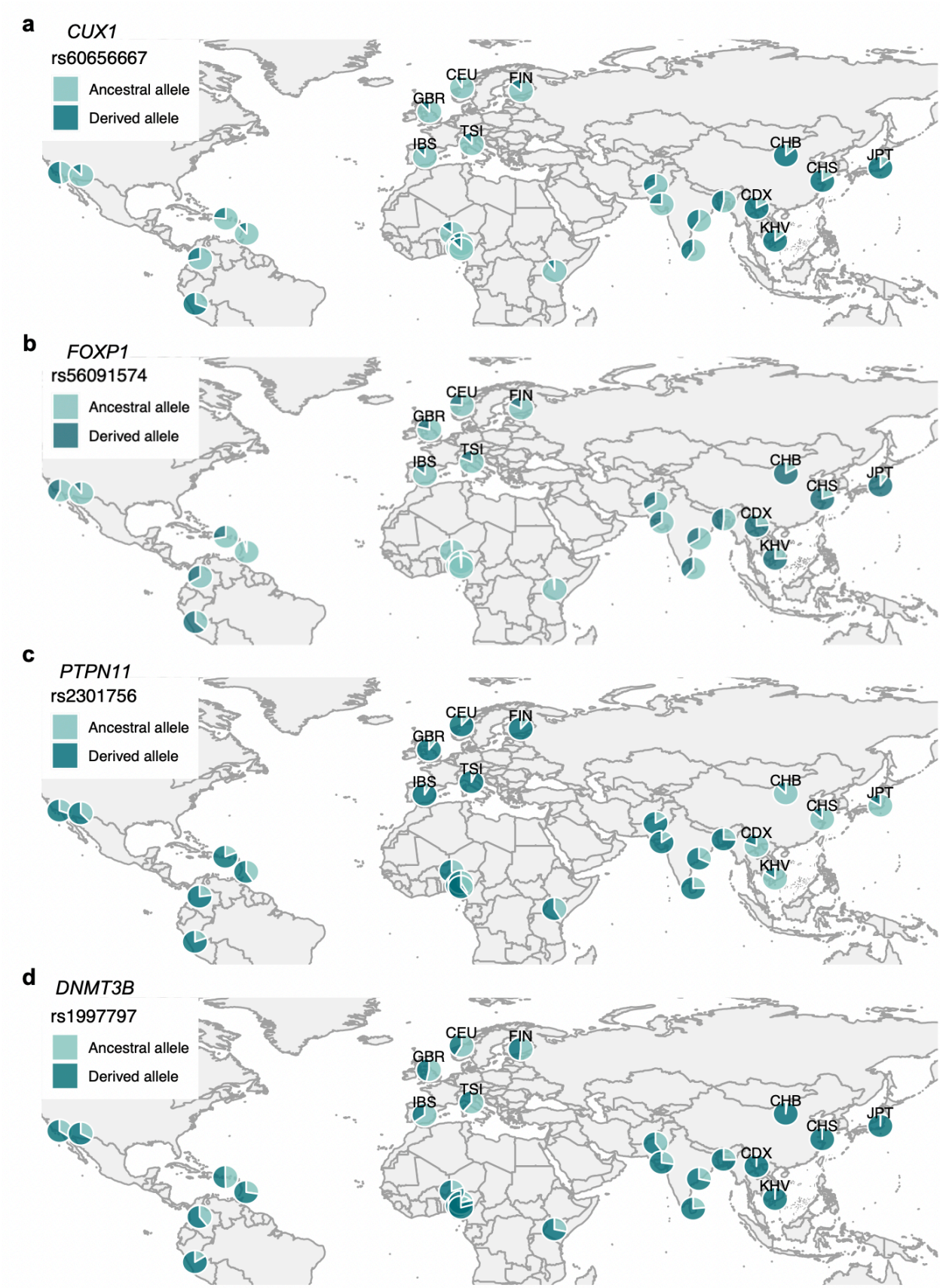


Figure S28. Genetic differentiation of CHIP in Chinese populations. The significant differentiation between East Asian and European populations (based on 1KGP, CEU vs. CHB). The genome–wide SNPs with fixation index (F_{ST}) values ranked top 0.1% (i.e., F_{ST} > 0.55) were plotted as the gray background. The SNPs located on expanded CHIP genes were highlighted with colored dots, with blue dots representing a significantly higher frequency of this locus in Europeans (EUR) and red dots representing a significantly higher frequency in East Asians (EAS). DAF, derived allele frequency. ddSNPs, differentiated SNPs, top 5%; hdSNPs, highly differentiated SNPs, top 1%.


a
CUX1
rs60656667
Ancestral allele
Derived allele
CEU
FIN
GBR
TSI
IBS
CHB
JPT
CHS
CDX
KHV
b
FOXP1
rs56091574
Ancestral allele
Derived allele
CEU
FIN
GBR
TSI
IBS
CHB
JPT
CHS
CDX
KHV
c
PTPN11
rs2301756
Ancestral allele
Derived allele
CEU
FIN
GBR
TSI
IBS
CHB
JPT
CHS
CDX
KHV
d
DNMT3B
rs1997797
Ancestral allele
Derived allele
CEU
FIN
GBR
TSI
IBS
CHB
JPT
CHS
CDX
KHV

1

Figure S29. Geographical distribution of CHIP mutations showing significant East–West differentiation (based on 1KGP, CEU vs. CHB). *CUX1* is an essential CHIP gene with an F_{ST} of 0.76 (**a**), *PTPN11* a core CHIP gene with an F_{ST} (Fixation Index) of 0.77 (**b**), *FOXP1* an expanded CHIP gene with an F_{ST} of 0.86 (**c**), and *DNMT3B* an expanded CHIP gene with a SNP–specific F_{ST} of 0.65 (**d**). Allele frequency is based on 1KGP.

Supplementary Tables (An independent Excel file)

Table S1. Cohorts included in the Chinese CHIP analyses.

Table S2. Previously validated CHIP mutations queried in this study (from Jaiswal et al. 2017).

Table S3. CHIP–related genes including myeloid and lymphoid leukemia genes (collected from previously eight CHIP studies).

Table S4. CHIP whitelist variants identified among 13,445 Chinese individuals (discovery cohort POG13K).

Table S5. CHIP whitelist variants identified among 16,151 Chinese individuals (replication cohort CYN16K).

Table S6. CHIP whitelist variants identified among 2,504 1KGP unrelated individuals (validation cohort 1KGP 30x data).

Table S7. All functional variants in CHIP 74 genes identified among 357 plateau–adapted Tibetans.

Table S8.1. CHIP previously reported in Western populations (Jaiswal et al. 2017).

Table S8.2. CHIP previously reported in Western populations (Bick et al. 2020).

Table S8.3. CHIP previously reported in Western populations (Kar et al. 2022).

Table S9. All functional variants identified in 13,445 individuals (discovery cohort POG13K).

Table S10. All functional variants identified in 16,151 individuals (replication cohort CYN16K).

Table S11. *KMT2D* somatic missense mutations identified and confirmed in the discovery and replication cohort.

Table S12. All functional variants identified in 1KGP unrelated individuals (N = 2,504).

Table S13. Rare variants burden of highly differentiated CHIP genes between the West (i.e., 39,345 Europeans from gnomAD) and East (i.e., 10,842 Chinese with WGS) populations.

Table S14. Previous reports of CHIP prevalence among populations.

Table S15. Previously reported CHIP predisposing genetics determinants (mainly from European descendants).

Table S16. Top 0.001 differentiated CHIP genes between the West and East populations (1KGP, CEU vs. CHB).

Table S17. Top 0.001 differentiated CHIP genetic variants between the West and East populations (1KGP, CEU and CHB).

Table S18. Significantly differentiated CHIP genetic variants (previously reported GWAS SNPs) between the West and East populations (1KGP, CEU vs. CHB).

References

1. Lu, D. *et al.* Ancestral Origins and Genetic History of Tibetan Highlanders. *The American Journal of Human Genetics* **99**, 580–594 (2016).

2. Deng, L. *et al.* Prioritizing natural-selection signals from the deep-sequencing genomic data suggests multi-variant adaptation in Tibetan highlanders. *National Science Review* **6**, 1201–1222 (2019).

3. Feng, Q. *et al.* Genetic History of Xinjiang's Uyghurs Suggests Bronze Age Multiple-Way Contacts in Eurasia. *Molecular Biology and Evolution* **34**, 2572–2582 (2017).

4. Gao, Y. *et al.* PGG.Han: The Han Chinese genome database and analysis platform. *Nucleic Acids Research* **48**, D971–D976 (2020).

5. Pan, Y. *et al.* Genomic diversity and post-admixture adaptation in the Uyghurs. *National Science Review* **9**, nwab124 (2022).

6. Gao, Y. *et al.* A pangenome reference of 36 Chinese populations. *Nature* **619**, 112–121 (2023).

7. Wen, J. *et al.* Ancestral origins and post-admixture adaptive evolution of highland Tajiks. *National Science Review* **11**, nwae284 (2024).

8. Bick, A. G. *et al.* Inherited causes of clonal haematopoiesis in 97,691 whole genomes. *Nature* **586**, 763–768 (2020).

9. Kessler, M. D. *et al.* Common and rare variant associations with clonal haematopoiesis phenotypes. *Nature* **612**, 301–309 (2022).

10. Zheng, R. S. Cancer statistics in China, 2016. *Zhonghua Zhong Liu Za Zhi* **45**, 212–220 (2023).

11. Beall, C. M. Two routes to functional adaptation: Tibetan and Andean high-altitude natives. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 8655–8660 (2007).

12. Kim, J.-H. *et al.* Molecular networks of FOXP family: Dual biologic functions, interplay with other molecules and clinical implications in cancer progression. *Mol Cancer* **18**, (2019).

13. Yang, Z., Li, Y., Yin, F. & Chan, R. J. Activating PTPN11 mutants promote hematopoietic progenitor cell-cycle progression and survival. *Experimental Hematology* **36**, 1285–1296 (2008).

14. Jaiswal, S. *et al.* Clonal Hematopoiesis and Risk of Atherosclerotic Cardiovascular Disease. *N Engl J Med* **377**, 111–121 (2017).

15. Robertson, N. A. *et al.* Longitudinal dynamics of clonal hematopoiesis identifies gene-specific fitness effects. *Nat Med* **28**, 1439–1446 (2022).

16. Zhao, K. *et al.* Somatic and Germline Variants and Coronary Heart Disease in a Chinese Population. *JAMA Cardiol* **9**, 233 (2024).

17. Kar, S. P. *et al.* Genome-wide analyses of 200,453 individuals yield new insights into the causes and consequences of clonal hematopoiesis. *Nat Genet* **54**, 1155–1166 (2022).

18. Pich, O., Reyes-Salazar, I., Gonzalez-Perez, A. & Lopez-Bigas, N. Discovering the drivers of clonal hematopoiesis. *Nat Commun* **13**, 4267 (2022).

19. Niroula, A. *et al.* Distinction of lymphoid and myeloid clonal hematopoiesis. *Nat Med* **27**, 1921–1927 (2021).

20. Jurgens, S. J. *et al.* Rare coding variant analysis for human diseases across biobanks and ancestries. *Nat Genet* **56**, 1811–1820 (2024).