

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

Shuhua Xu

xushua@fudan.edu.cn

Fudan University <https://orcid.org/0000-0002-1975-1002>

Tang Xia

State Key Laboratory of Genetics and Development of Complex Phenotypes, Center for Evolutionary Biology, School of Life Sciences, Fudan University

Hongpu Chen

State Key Laboratory of Genetics and Development of Complex Phenotypes, Center for Evolutionary Biology

Xiaoyuan Zhang

Fudan University

Yingbing Shi

Fudan University <https://orcid.org/0000-0003-4495-1986>

Yang Gao

Fudan University <https://orcid.org/0000-0001-7057-1413>

Yumeng Sun

State Key Laboratory of Genetic Engineering, Human Phenome Institute, Zhangjiang Fudan International Innovation Center, Center for Evolutionary Biology, School of Life Sciences, Fudan University

Baonan Wang

Fudan University

Qi Liu

Fudan University

Yuhang Cui

Fudan University <https://orcid.org/0009-0008-9090-990X>

Mochi Wei

Fudan University

Shuangshuang Cheng

Fudan University

Lian Deng

Fudan University <https://orcid.org/0009-0008-7302-4122>

Zeshan Lin

Fudan University

Zhimin Feng

National Clinical Research Centre for Endocrine and Metabolic Diseases, Rui-Jin Hospital, Shanghai Jiao-Tong University School of Medicine

Yaoxi He

Kunming Institute of Zoology <https://orcid.org/0000-0003-3324-3239>

Jiucun Wang

State Key Laboratory of Genetic Engineering, Collaborative Innovation Center for Genetics and Development, School of Life Sciences and Human Phenome Institute <https://orcid.org/0000-0003-2765-0620>

Bing Su

Kunming Institute of Zoology, Chinese Academy of Sciences <https://orcid.org/0000-0002-4379-9014>

Sijia Wang

Chinese Academy of sciences <https://orcid.org/0000-0001-6961-7867>

Jingchun Luo

Fudan University <https://orcid.org/0000-0001-5138-6458>

Yan Zheng

Fudan University <https://orcid.org/0000-0003-1129-3147>

Xin Jin

BGI Research <https://orcid.org/0000-0001-7554-4975>

Yanan Cao

Ruijin Yangtze River Delta Health Institute, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine <https://orcid.org/0000-0002-3848-7040>

Yan Lu

Fudan University <https://orcid.org/0000-0001-6907-0443>

Li Jin

Fudan University <https://orcid.org/0000-0001-9201-2321>

Biological Sciences - Article

Keywords:

Posted Date: April 17th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9041912/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Additional Declarations: There is **NO** Competing Interest.

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

Running title: Population Architecture of CHIP in Chinese Cohorts

Xia Tang^{1,§}, Hongpu Chen^{1,§}, Xiaoyuan Zhang¹, Yingbing Shi¹, Yang Gao¹, Yumeng Sun^{1,2}, Baonan Wang¹, Qi Liu¹, Yuhang Cui^{1,2}, Mochi Wei¹, Shuangshuang Cheng^{1,2}, Lian Deng^{1,3}, Zeshan Lin¹, Zhimin Feng⁴, Yaoxi He⁵, Jiucun Wang^{1,2}, Bing Su^{5,6}, Sijia Wang^{6,7}, Jingchun Luo², Yan Zheng¹, The Han100K Initiative^{##}, Xin Jin^{8,9,*}, Yanan Cao^{4*}, Yan Lu^{1,2,*}, Li Jin^{1,2,*}, Shuhua Xu^{1,2,10*}

¹State Key Laboratory of Genetics and Development of Complex Phenotypes, Center for Evolutionary Biology, School of Life Sciences, Fudan University, Shanghai 200438, China;

²Human Phenome Institute, Zhangjiang Fudan International Innovation Center, Fudan University, Shanghai 201203, China

³Ministry of Education Key Laboratory of Contemporary Anthropology, Fudan University, Shanghai 201203, China;

⁴Ruijin Yangtze River Delta Health Institute, Department of Endocrine and Metabolic Diseases, Shanghai Institute of Endocrine and Metabolic Diseases, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China.

⁵State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming 650223, China;

⁶Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming 650223, China;

⁷Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai, China;

⁸BGI-Shenzhen, Shenzhen, Guangdong 518083, China;

⁹School of Medicine, South China University of Technology, Guangzhou, Guangdong 510006, China;

¹⁰Inner Mongolia Key Laboratory of Life Health and Bioinformatics, School of Life Science and Technology, Inner Mongolia University of Science and Technology, Baotou 014010, China.

[§]These authors contributed equally to this work;

*Correspondence and requests for materials should be addressed to X.J. (jinxin@genomics.cn), Y.C. (caoyanan@vip.sina.com), Y.L. (lueyan@fudan.edu.cn), L.J. (lijin@fudan.edu.cn), or S.X. (xushua@fudan.edu.cn).

##Full list of participants (collaborators) of the Han100K Initiative can be found online via <https://pog.fudan.edu.cn/han100k/#/participants>.

Summary

Clonal hematopoiesis of indeterminate potential (CHIP) an age-related expansion of hematopoietic clones with somatic mutations, associated with hematologic malignancies and cardiovascular disease, but remains poorly characterized beyond European-ancestry populations. Here, we analyzed high-depth sequencing data from 29,596 individuals, including a discovery cohort of 13,445 and an independent replication cohort of 16,151, representing the largest systematic investigation of CHIP across diverse Chinese ethnic groups to date. We estimated an overall CHIP prevalence of ~4.5%, slightly lower than reported in Europeans (~5.7%). While *DNMT3A* and *TET2* were the most frequently mutated genes, the relative contributions of other genes, including *KMT2D* and *ASXL1*, varied across populations. Interestingly, population differentiation at *AICDA* were consistent with a potential contribution to the higher frequency of *KMT2D*-associated CHIP in Chinese populations. Our findings suggest that current CHIP whitelists, being largely derived from European samples, should be applied cautiously and refined through broader population representation and ancestry-informed validation. Together, these results delineate the population-scale mutational landscape of CHIP in Chinese cohorts and provide a reference framework for studies of somatic hematopoiesis in East Asian populations.

Introduction

In the hematopoietic system, somatic mutations that confer a fitness advantage to hematopoietic stem cells (HSCs) and their progeny drive the age-related expansion of mutant clones, a phenomenon known as clonal hematopoiesis (CH)¹. Clonal hematopoiesis of indeterminate potential (CHIP) refers to the presence of blood cells that clones carrying somatic mutations in the absence of overt hematological malignancy or other diagnostic criteria for hematologic disease^{2,3}. CHIP is increasingly recognized as a common feature of aging, with a prevalence of > 10% in those older than 70 years³. The most commonly mutated genes in aging-related CHIP are *DNMT3A* and *TET2* followed by *ASXL1* (i.e., *DTA* genes) as a distant third, and a smaller frequency of mutations in the splicing genes *SF3B1*, *SRSF2*, *PRPF8*, and *U2AF1*. CHIP portends a strong underlying risk for the development of myelodysplastic syndrome (MDS) and other hematologic neoplasms as well as all-cause mortality², cardiovascular^{2,3}, infection⁴, stroke⁵, pulmonary disease⁶, liver disease⁷, and renal diseases⁸. Previous genome-wide association analyses (GWAS) in European-ancestry cohorts have identified both common and rare genetic variants predisposing to CHIP, including loci on *TERT* (telomere maintenance), *KPNA4*, *TET2*, *ATM*, *CHEK2* (DNA damage response), *CD164* and *PARP1* (DNA repair), by shaping the HSC function and/or leukocyte telomere length^{9–11}.

Despite these advances, the genetic architecture of CHIP in non-European populations remains poorly characterized, limiting insights into ancestry-specific risk factors and therapeutic targets. Populations can differ greatly in their rates and profiles of CHIP^{9–11}. To date, large-scale studies of CHIP have focused on individuals of European or European-dominant descent, leaving the impact of genetic ancestry on CHIP development incompletely understood. Carrier screening panels are mostly derived from European-descent populations and may miss CHIP subtypes common in non-Europeans. For instance, the UK Biobank, the largest whole genome dataset in the world, contains only 1,659 individuals of Chinese descent (~0.3%)¹². Similarly, the TOPMed included only 1,184 East Asians, representing approximately 1.2% of the cohort¹³. The only two published studies on CHIP in Chinese populations were limited to the single-center samples, with limited coverage of ancestry diversity^{14,15}. Given China's population of 1.44 billion and substantial genetic heterogeneity, as demonstrated by prior studies revealing pronounced population structure and complex ancestry composition among Chinese subgroups^{16–21}, a large-scale, ethnically

diverse investigation offers a unique opportunity to refine the mutation landscape of CHIP and enable both novel discovery and robust cross–population validation.

To characterize the landscape of clonal hematopoiesis and its ethnic diversity in Chinese populations, we conducted deep sequencing in a discovery cohort of 13,445 individuals spanning 23 ethnic groups and 32 administrative regions in China, along with an independent replication cohort of 16,151 individuals across 23 ethnic groups and 33 administrative regions, representing the largest CHIP analysis in Chinese populations to date. This study addresses critical gaps in CHIP characterization by: (i) systematically comparing CHIP prevalence estimates between Chinese and European cohorts; ii) examining mutation spectra to identify ancestry–related differences in gene–level contributions (e.g., *KMT2D* enrichment in East Asians); and iii) exploring the germline–somatic context of population–differentiated *AICDA* in relation to *KMT2D* variants enrichment in Chinese populations.

Results

Prevalence of CHIP in Chinese populations

We analyzed CHIP profiles in a discovery cohort of 13,445 individuals aged 18–107 years (mean 46.0, median 44, mean sequencing depth 74.4×), comprising Han and 22 ethnic minority groups from seven language families and covering 32 first–level administrative regions across China (see Methods, Fig. 1a, Tables S1, and Supplementary Figs. 1–2). Somatic variants were identified from whole–exome and whole–genome sequencing data using MuTect2²² with stringent filtration. For comparability with prior studies, we focused on a predefined whitelist of CHIP variants across 74 genes² (see Methods, Supplementary Fig. 1, and Table S2) to define the CHIP events and carriers.

In total, we detected 746 CHIP events in 600 individuals, corresponding to 596 variants across 58 genes, with each variant manually confirmed by Integrative Genomics Viewer (IGV) inspection (Additional File 1). The overall CHIP prevalence in the discovery cohort was 4.5% (Supplementary Figs. 1 and 3). The most frequently mutated genes were *DNMT3A*, *TET2*, *KMT2D*, *ASXL1*, and *PPM1D* (Supplementary Fig. 3).

In the independent replication cohort (16,151 individuals aged 18–98 years, mean age of 53.9 years, median 54.4 years, mean sequencing depth of 42.2×), the prevalence was 2.8% (95% CI: 2.5%–3.1%; 451/16,151), with 476 CHIP events identified using the same pipeline (Additional File 2, Supplementary Figs. 1, 2, 4, and 5). The lower prevalence was largely attributable to reduced sequencing coverage relative to the discovery cohort (Supplementary Fig. 1), with 342 whitelist variants detected across 36 genes (Supplementary Fig. 5). The most frequently mutated genes detected in the replication cohort were *DNMT3A*, *TET2*, *ASXL1*, *KMT2D*, *JAK2* and *PPM1D* (Supplementary Fig. 5).

We further applied the same pipeline to unrelated individuals from the 1000 Genomes Project (mean depth ~34×) to assess consistency with previous reports^{2,9,11}. This analysis identified 93 whitelist variants in 88 individuals, corresponding to a prevalence of 3.5% (95%CI: 2.9%–4.3%; Additional File 3, and Supplementary Fig. 6a), consistent with prior estimates in mixed ancestry populations²³. Notably, CHIP prevalence in East Asians from the 1000 Genomes Project (3.0%) was comparable to that observed in replication cohort (2.8%) at a similar sequencing depth.

CHIP prevalence and the proportion of larger clones (variant allele fraction, VAF > 10%) increased markedly with age ($P < 2.2 \times 10^{-16}$, Fig. 1b, and Supplementary Fig. 4b). The age-related increase varied by driver gene. For example, *CREBBP* mutations were detected at a younger age than *TET2* mutations (median difference ~6 years (Wilcoxon rank-sum test $P = 2.25 \times 10^{-6}$, Supplementary Fig. 7a), and *TET2* showed a clear preference for older ages in the replication cohort (Supplementary Fig. 7b). Overall CHIP prevalence was comparable between females and males (4.0% vs. 4.3%; χ^2 test $P = 0.43$; Fig. 1c and Supplementary Fig. 4c), although gene-level sex-biases were observed. *CREBBP* showed a lower female-to-male ratio ($P = 0.002$), whereas *CBLB* showed higher ratio ($P = 0.0001$). In contrast, *ASXL1* exhibited inconsistent sex patterns between the discovery and replication cohorts (Supplementary Figs. 7c–d). Most CHIP carriers harbored a single driver mutation (88.7% in discovery; 94.9% in replication (Fig. 1d and Supplementary Fig. 4d), with a mean of 1.24 (1.05 in replication) mutations per carrier. Detected variants were predominantly frameshift mutations ($n = 283$), missense ($n = 229$), stop-gained ($n = 182$), and splicing mutations ($n = 51$), with ~69.4% (520/749) predicted to be strongly deleterious (Supplementary Fig. 3).

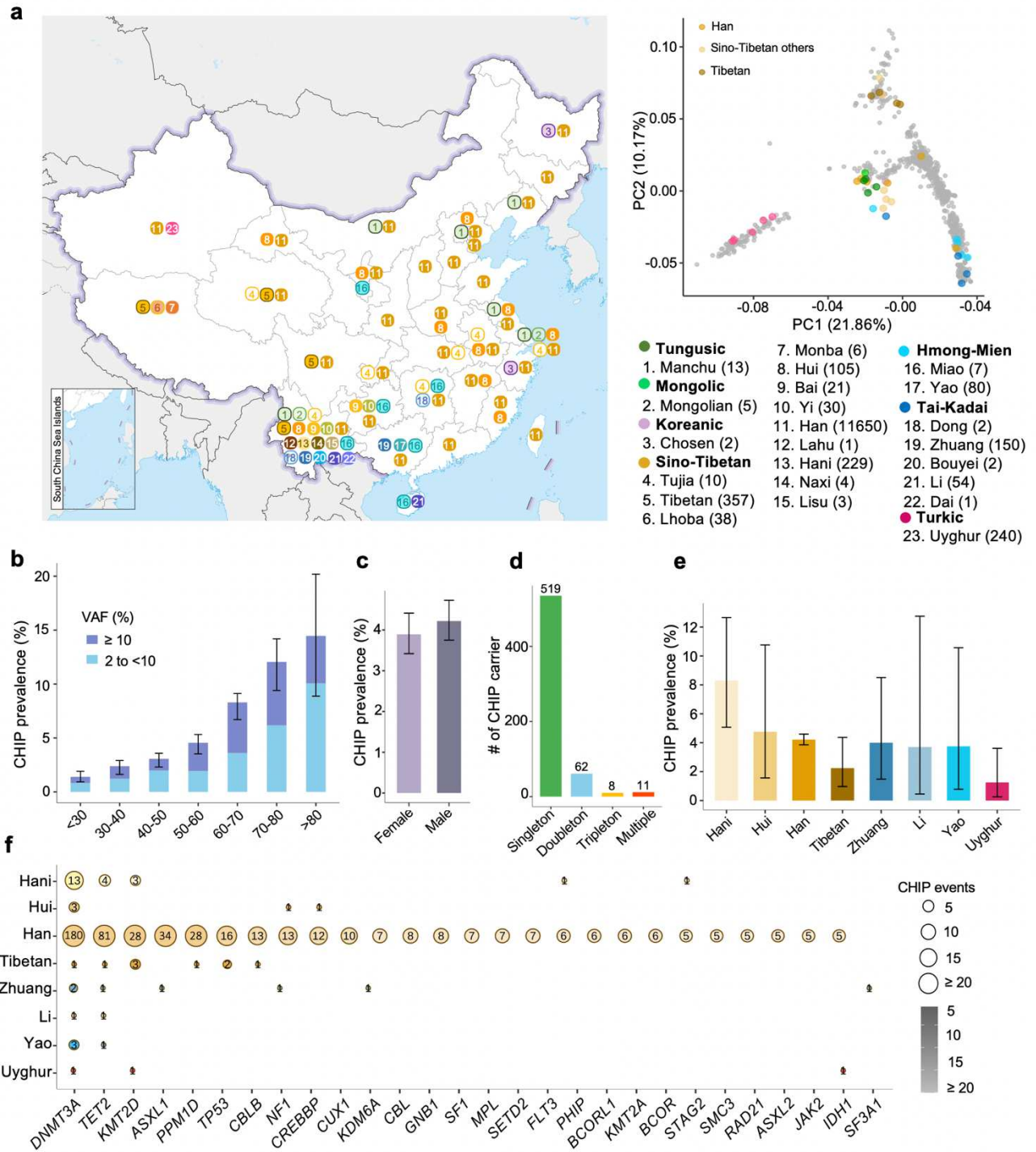


Figure 1 | Identifying CHIP in 13,445 Chinese populations including whole-genome or whole-exome samples (discovery cohort).

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 variant allele fraction (VAF). **c.** The distribution of overall CHIP prevalence in Chinese populations by gender. **d.** More than 84% of individuals with CHIP had only one somatic CHIP driver mutation variant identified (singleton). **e.** The observed CHIP prevalence varied greatly among different ethnic minorities, which only groups with detected at least one CHIP carriers were shown. However, the sample size varied greatly between ethnic subgroups, which was detailed in Notes of Supplementary Materials. **f.** Top frequent CHIP genes mutated in each subgroup. All detected CHIP genes were 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.

We identified 26 recurrent CHIP loci across 9 genes in the discovery cohort, including 15 (57.7%) predicted to be highly deleterious (Supplementary Fig. 8a). Hotspots varied markedly across geographical groups, with no clear age- or sex-specific patterns. Apparent differences between Han and ethnic minority groups were largely influenced by sample size disparities. Sixteen *DNMT3A* hotspots were detected, with the canonical p.R882 mutation observed in 30 individuals, consistent with reports in Western cohorts²⁴. Additional hotspots, including *KMT2D* p.AGTEP4826–4830AX and *GNBI* p.K57E, were found in 10 and 7 individuals, respectively, and replicated in 14 carriers (Supplementary Fig. 8b). In total, 22 recurrent loci were identified in the replication cohort, of which 10 overlapped with the discovery set, including 5 *DNMT3A* loci (e.g., p.R882, p.R320, p.R326, etc.), *JAK2* p.V617F, *GNBI* p.K57E, and *KMT2D* p.AGTEP4826–4830AX. Notably, no recurrent *TET2* loci were detected in the replication cohort, highlighting substantial heterogeneity even among closely related populations and suggesting convergent effects of diverse variants on CHIP.

CHIP prevalence varied widely across ethnic subgroups, ranging from 1.3% to 8.3%, 6.4–fold change (Fig. 1e), with a 3.8–fold range in the replication cohort (Supplementary Fig. 4e). Compared with Han individuals (4.4%, 95% CI: 3.9%–4.6%), prevalence was higher in Hani (8.3%, 95% CI: 5.4%–12.6%) and Hui (4.8%, 95% CI: 2.1%–10.7%), and lower in Li (3.7%, 95% CI: 1.0%–12.5%), Tibetan (2.2%, 95% CI: 1.1%–4.4%), and Uyghur (1.3%, 95% CI: 0.4%–3.6%). As Han samples comprised 84.4% of our cohort, their estimates closely matched the overall prevalence. *DNMT3A* mutations were observed in nearly all groups (Fig. 1f), except Mongolian samples (N = 241, aged 18–90 years, mean age of 42.48 years, median 41.8 years) in the replication cohort (Supplementary Fig. 4f). *TET2* mutations were broadly present but less frequent than

DNMT3A, with limited detection in Mongolian, Hui, and Uyghur groups (Fig. 1f and Supplementary Fig. 4f). *ASXL1* mutations were enriched Han, Hui, and Zhuang but were rare in other minorities. *KMT2D* mutations showed consistent enrichment across multiple groups, including Hani, Tibetan, Yi, and Uyghur (Fig. 1f and Supplementary Fig. 4f, and see Notes in Supplementary Materials for more details). Together, these results indicate substantial ethnic heterogeneity in CHIP architecture in Chinese populations compared with patterns described in European-ancestry populations.

CHIP heterogeneity between Chinese and European populations

Notably, the overall CHIP prevalence in our Chinese cohorts (4.5%) is comparable to that reported in Western populations (e.g., ~5% in the UK Biobank²³, 5.3% in 355,183 individuals of predominantly European ancestry²⁵, χ^2 test $P > 0.05$), contrasting with earlier reports suggesting a lower CHIP incidence in East Asians^{9,10}. However, these comparisons may be confounded by differences in age structure, sex distribution, and sequencing depth across studies. To minimize such biases, we conducted an independent comparison using East Asian (EAS, $N = 504$) and European (EUR, $N = 503$) samples from the 1000 Genomes Project (1KGP), which provides more uniform sequencing and processing conditions despite the smaller sample size. Using the same pipeline and criteria, we identified 29 CHIP carriers among EURs (5.77%) and 15 in EAS (2.98%), indicating a significantly lower prevalence in EAS than in EUR (χ^2 test $P = 0.04$). A detailed variant-level comparison further revealed substantial differences in the mutational spectrum between EAS and EUR. For example, *SETD2* (alias *KMT3A*) mutations were more frequent in Europeans, whereas *KMT2D* mutations were relatively enriched in East Asians (Additional File 3, and Table S6). Moreover, when restricting the analysis to the top 15 most frequently reported CHIP genes from eight large studies of European or European-dominant populations, the estimated prevalence in our Chinese cohorts decreased to 3.1% (see Methods), underscoring the influence of gene selection on cross-population comparisons. When the top 15 mutated genes in Western or Eastern populations were compared, only eight genes overlapped (see Methods; Fig. 2 and Supplementary Fig. 9a), and their mutation frequencies differ markedly between Eastern and Western populations. Thus, even when analyses were constrained to the same predefined whitelist in 74 CHIP genes and identical filtering criteria², substantial and non-negligible heterogeneity in CHIP architecture between European- and East Asian-ancestry populations remained evident.

Consistent discrepancies were also observed between our replication cohort and Western datasets (Notes in Supplementary Materials, Supplementary Figs. 9b and 10).

We next compare variants within the predefined whitelist of 74 CHIP genes across studies. The combined Western reference set comprised 16,854 CHIP events (5,820 unique variants) aggregated from three large cohorts: 17,182 WES samples (219 events, 171 unique variants)² (Table S8.1), 97,691 WGS samples (4,938 events, 2,531 unique variants)⁹ (Table S8.2), and 200,453 WES samples (11,697 events, 4,013 unique variants)¹¹ (Table S8.3). In contrast, our discovery cohort contain 596 unique CHIP variants (Table S4). Only 149 variants were shared between Western and Chinese datasets; when relaxed to genomic loci, 240 positions overlapped, highlighting a distinct population-specific CHIP spectrum. The proportion of nonsense variants in our cohorts (24.83%) was similar to that in Western populations (23.77%). Interestingly, CHIP variants on chromosome X were more frequently observed in Chinese individuals (6.01%) than in Western cohorts (3.14%), consistent with previous observations of higher rates of mosaic chromosome alterations in East Asian biobanks (Japan BBJ) compared with UK Biobank²⁶.

We next focused on the most frequent driver genes (Fig. 2). The top recurrent CHIP drivers in Chinese participants were *DNMT3A* (225 events in 217 participants), *TET2* (104 events in 99 participants), *KMT2D* (42 events in 38 participants), *ASXL1* (41 events in 39 participants), and *PPM1D* (31 events in 31 participants) (Fig. 2a), with a median VAF of 10.5%, 9.4%, 9.0%, 12.4%, and 11.7% respectively. Notably, *KMT2D* showed a relatively high prevalence, comparable with *ASXL1*, a pattern consistently observed in the replication cohort (Supplementary Fig. 10a) and supported by the 1000 Genomes data (Supplementary Fig. 6a). To systematically evaluate East–West differences, we combined the top 15 most frequent CHIP genes reported in Western and Eastern populations into a unified list of 22 genes (Supplementary Fig. 9a) and compared their overall prevalence (Fig. 2b). Mutations in canonical *DTA* genes (*DNMT3A*, *TET2*, and *ASXL1*), and DNA damage response genes (e.g., *TP53* and *PPM1D*) showed broadly similar prevalence across populations (Figs. 2b–c; Supplementary Figs. 10b–c). In contrast, mutations in *SF3B1* and *ATM* were more frequent in Western cohorts and have been associated with higher risk of progression to hematological malignancies³.

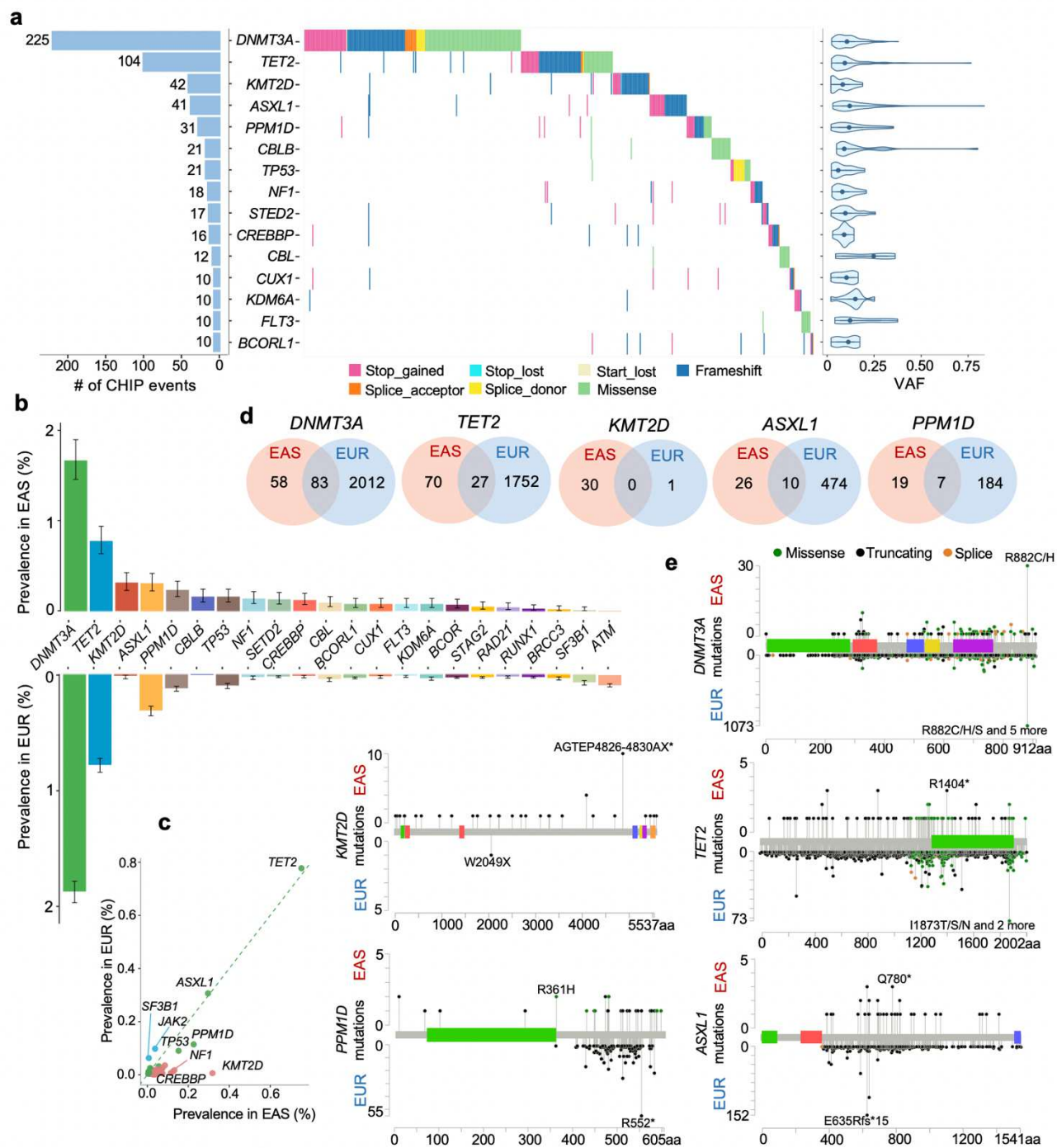


Figure 2 | Characterization of CHIP in the Chinese populations (discovery cohort).

a. Composite plot summarizing mutations in the 15 most common driver genes mutated in Chinese CHIP carriers. Each column in the waterfall plot represented a single individual, with mutation types color-coded. Bars on the left quantify mutations per gene as a mutation counts detected in corresponding genes. Violin plots on the right showed the distribution of variant allele fraction (VAF), with dots representing the mean value. **b.** The comparison of the prevalence of 22 frequent CHIP

genes between the Western and Eastern populations. CHIP prevalence in the Western populations was referred to the merged cohort of Jaiswal et al² and Bick et al⁹. The color saturation referred to the preference of the population. *DNMT3A* was the top frequent CHIP gene in both East and West, and the high prevalence disturbed the indication for other genes, thus we did not plot *DNMT3A*. **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.

At the variant level, 141 *DNMT3A* variants were identified in our cohort, 83 (58.9%) of which have been previously reported in Western population datasets (Fig. 2d and Table S4). Similarly, 97 *TET2* and 36 in *ASXL1* variants were detected with 27 (27.8%) and 10 (27.7%), respectively, overlapping with Western reports (Figs. 2d–e; and Table S4). In contrast, *KMT2D*–CHIP was markedly enriched in Chinese populations, with a 29–fold higher prevalence than in Western populations (Figs. 2b–e). Elevated frequencies were also observed for *CBLB*, *CREBBP* and *PPM1D* (Fig. 2b and Supplementary Fig. 11). Genes involved in transcriptional regulation and signaling pathways, including *CREBBP*, *CUX1*, and *CBL*, showed modestly higher prevalence in East Asians (0.09%) than in Western datasets (0.07%) (Fig. 2b and Supplementary Fig. 11). These patterns of driver gene heterogeneity and variant distribution were consistently replicated in the independent cohort (Supplementary Figs. 10 and 12), supporting the robustness of our pipeline and highlighting pronounced differentiation in CHIP architecture between Eastern and Western populations.

***KMT2D* is relatively common in East Asians**

A notable finding is the consistently higher prevalence of *KMT2D* mutations in Chinese populations, with detection rates of 0.29% in the discovery cohort (n = 13,445) and 0.14% in the replication cohort (n = 16,151), occasionally approaching or exceeding those of *ASXL1* when restricted to predefined CHIP whitelist variants (Fig. 2 and Supplementary Fig. 10). When the analysis was expanded to include all the functional variants within the 74 CHIP genes (see Methods), *KMT2D* remained more predominant in Chinese populations than expected based on canonical *DTA* genes, and its paralog *KMT2A* also showed an elevated burden (Supplementary Figs. 13a–b; Table S9–S10). Although missense variants in *KMT2D* are not included in the

whitelist definition, they were detected at a higher frequency in the Chinese cohort (N = 161, Table S11) and were supported by read-level validation (Additional File 4).

We identified a recurrent *KMT2D* hotspot (p.AGTEP4826–4830AX) and evaluated whether its frequency could reflect technical or batch-related artifact. No enrichment was observed across age, sex, ethnicity, or geographic region in either the discovery and replication cohorts (Table S4 and Table S5), and all calls were supported by sequencing reads (Additional Files 1–2). This hotspot mutation was not detected in the 1KGP dataset, likely due to limited sample size in East Asian samples. To further assess robustness, we merged the two cohorts and compared sequencing depth, VAF, and VAF–age distributions across the five most frequent CHIP genes. No significant differences were observed between *KMT2D* and *DNMT3A*, *TET2*, or *ASXL1* under either whitelist-defined (Supplementary Figs. 14a–c) or expanded CHIP variant sets (Supplementary Figs. 14d–f), arguing against data bias as an explanation for the elevated *KMT2D* frequency.

We next examined the distribution of the top five driver genes across ethnic groups, geographic regions, age strata, and sex using both whitelist-defined (Fig. 3) and expanded CHIP variants (Supplementary Fig. 15). Among ethnic groups with sample size > 200, *ASXL1* showed enrichment in Han individuals, whereas *KMT2D* was more diffusely distributed across multiple ethnic groups (Fig. 1f and 3a, Supplementary Fig. 4f), with a pattern more similar to *DNMT3A* and *TET2* (two-sided Wilcoxon tests $P = 0.09$ and $P = 0.47$, respectively). In contrast, the distribution for *ASXL1* differed significantly from *DNMT3A* and *TET2* ($P = 0.002$ and $P = 0.01$, respectively) (Fig. 3a). The differences became more pronounced when expanded variants were considered (Supplementary Fig. 15a).

Geographically, *DNMT3A*, *TET2*, *KMT2D* and *ASXL1* were detected in 18, 17, 14, and 12 regions, respectively (regions with $N > 400$; Fig. 3b). Notably, in highly diverse regions such as Yunnan and the Tibet Autonomous Region, *KMT2D* detection rates exceeded those of *ASXL1*, consistent with a stronger association of *ASXL1* with Han-ancestry. Across populations with greater minority ancestry, *KMT2D* appeared more frequently (Figs. 1f and 3b, and Supplementary Fig. 4f). Although both *KMT2D* and *ASXL1* differed from *DNMT3A* and *TET2*, *KMT2D* showed distributional patterns more similar to the canonical *DTA* genes (Kruskal–Wallis and K–S tests),

whereas ASXL1 exhibited greater divergence, a trend that was further accentuated in analyses using expanded variant set (Supplementary Fig. 15b).

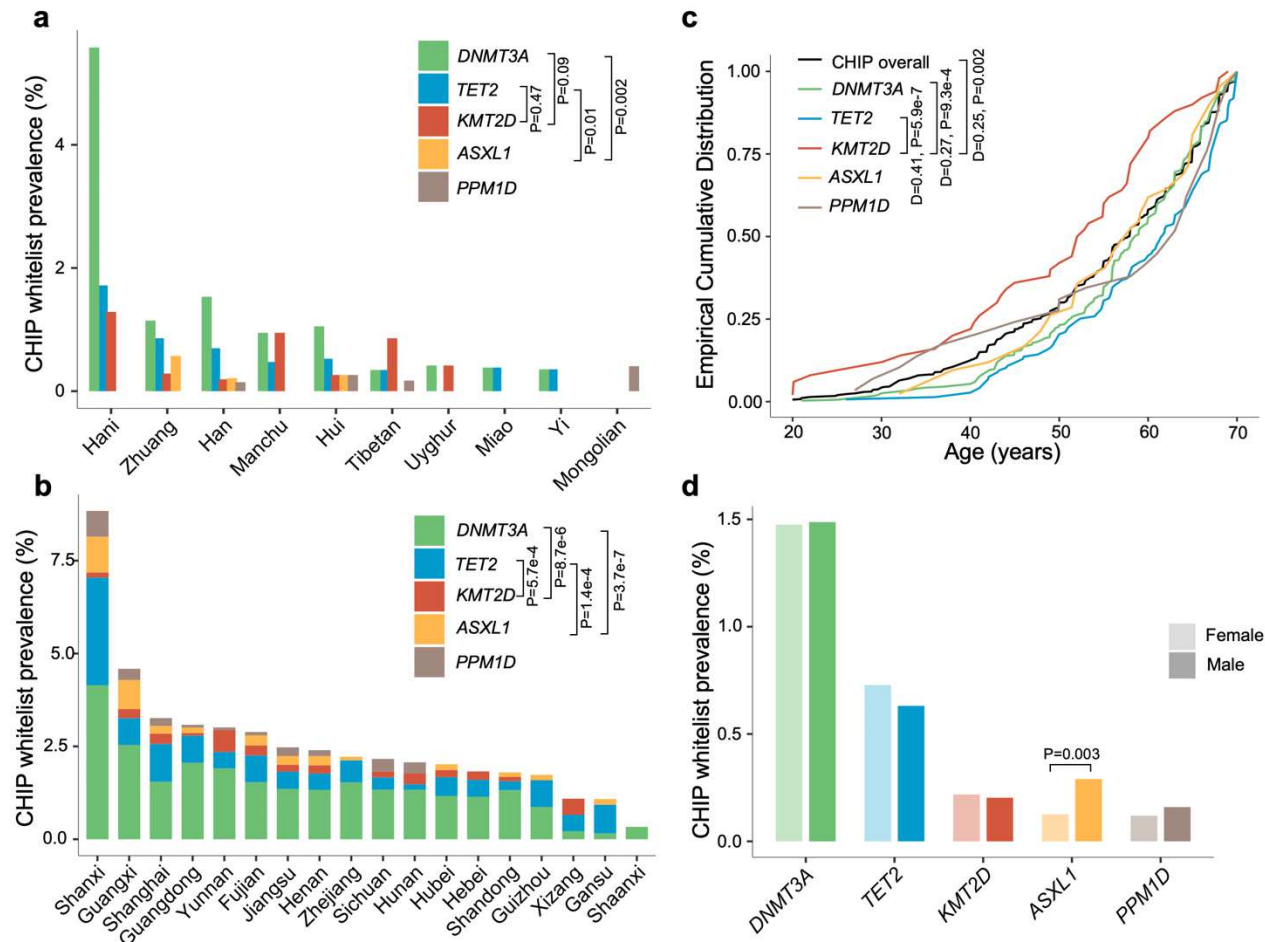


Figure 3 | Comparable ethnic, geographic, age, and sex distributions for *KMT2D* and other top frequent CHIP genes in Chinese populations (whitelist-defined CHIP in merged cohorts of 29,596 individuals).

a. The whitelist 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 whitelist 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 whitelist CHIP prevalence of top five frequent genes detected in females and males. A χ^2 test was used to evaluate differential carrier frequencies between sexes for each gene.

KMT2D carriers ranged in age from 20 to 78 years, spanning all age groups, with prevalence increasing with age (Fig. 3c). This pattern was broadly consistent with those observed for *DNMT3A*, *TET2*, *ASXL1*, and overall CHIP carriers under both whitelist-defined and expanded variant sets (Fig. 3c and Supplementary Fig. 15c). Notably, *KMT2D* prevalence appeared to rise earlier in life than that of canonical *DTA* genes, consistent with gene-specific differences in age-related CHIP dynamics¹¹. When expanded variants were considered, the age distribution of *KMT2D* did not differ significantly from that of the overall expanded CHIP (Supplementary Fig. 15c).

No significant sex bias was observed for *KMT2D*, *DNMT3A*, *TET2*, or *PPM1D*, under either whitelist or expanded definitions. In contrast, *ASXL1* exhibited a clear male predominance (χ^2 test $P = 0.003$ for whitelist-defined, and $P = 1.1 \times 10^{-11}$ for expanded variants) (Fig. 3d and Supplementary Fig. 15c). Consistently, *KMT2D* variants were also detected at a higher rate in East Asian samples from the 1KGP (Supplementary Fig. 6b), and functional *KMT2D* variants were present across global populations (Supplementary Fig. 6c). Notably, associations between *KMT2D* and leukemia became more predominant when analyses include multi-ancestry datasets rather than European-only cohorts (Supplementary Figs. 16a–b), based on the Human Disease Knowledge Portal²⁷, suggesting a potentially greater contribution of *KMT2D* in non-European populations²⁷. Collectively, these results support that *KMT2D*-associated CHIP is relatively common in East Asian populations.

***AICDA* may contribute to the higher mutation burden of *KMT2D* in East Asians**

We did not observe strong population differentiation in shared SNPs within *KMT2D* between East Asians and Europeans (F_{ST} not within the top 1% or 5% genome-wide), despite a significantly higher germline variant burden in East Asians. This suggests that the elevated somatic mutation burden of *KMT2D* in East Asians is more likely related to an intrinsically higher germline variant load (33.9 vs. 16.6 variants per individual, Supplementary Fig. 17a), rather than strong positive selection. Consistently, the rare variant burden in *KMT2D* was ~7.9-fold higher in East Asians than in Europeans (Table S13).

Thus, we assumed the flanking regions covering the genetic determinants of the accumulated variants in *KMT2D* by ‘cis’, which was previously reported as the mainly associations

with somatic mutations²⁸. We first examined cis-regulatory effects by analyzing ± 1 Mb flanking regions of *KMT2D* (Supplementary Fig. 18), and identified several highly differentiated loci, including a downstream region (~ 262 kb) linked to *KMT2D* (LD $r^2 = 0.89$) and overlapping *ADCY6*, *RP11-579D7.8* and *LINC00935*, which was underlying significant natural selection in East Asians (Supplementary Fig. 18). However, as the latter two are non-coding RNA genes, these signals are more likely to influence transcriptional regulation rather than directly explain elevated mutation rates in *KMT2D*.

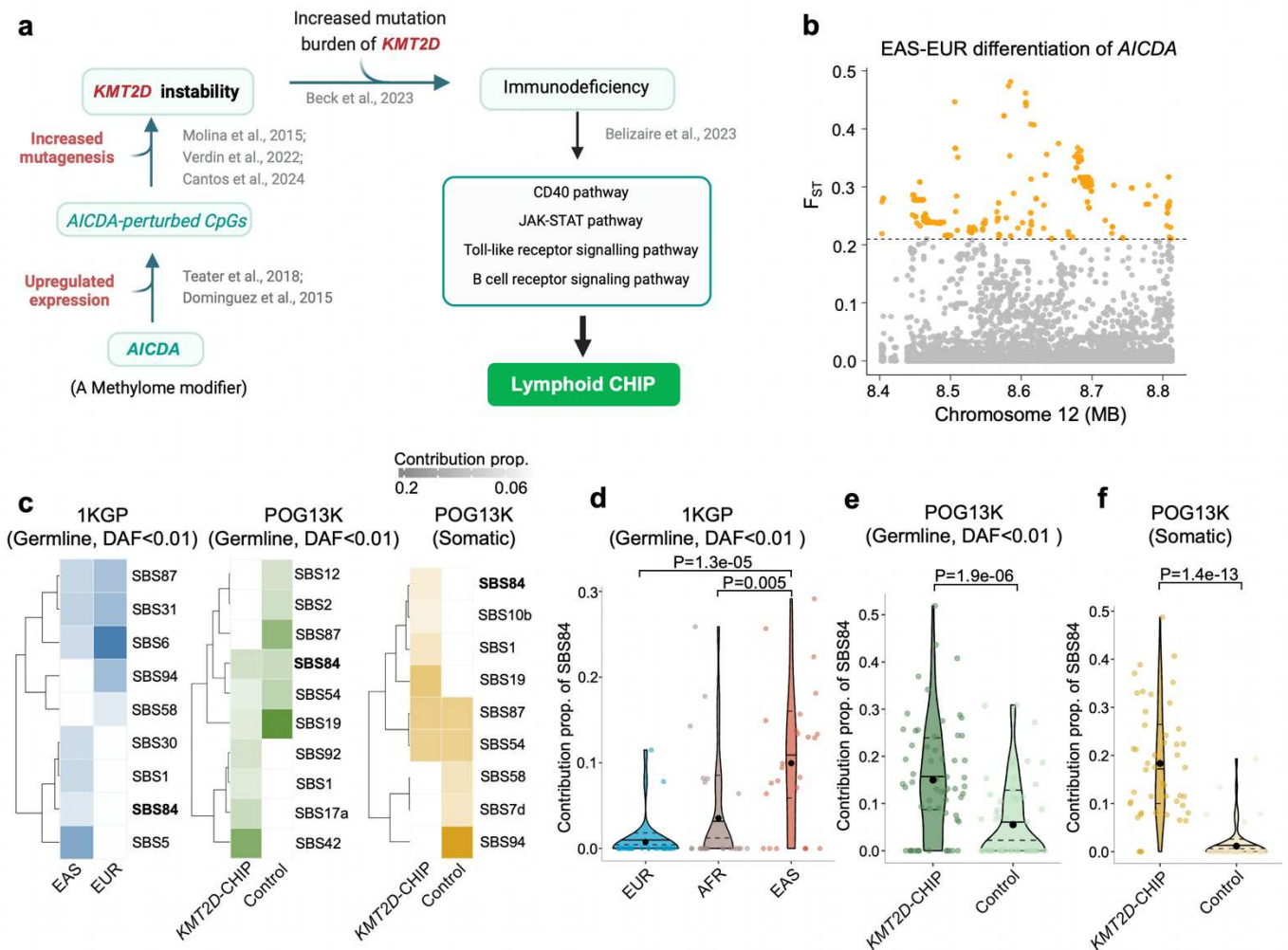


Figure 4 | The genetic basis underlying the higher mutation burden of *KMT2D* in EAS.

a. The pathway that *AICDA* activated the occurrence of *KMT2D* variants leading to the lymphoid CHIP. **b.** F_{ST} of *AICDA* between EAS and EUR. **c.** The mutation signature (single base substitution, SBS) of *KMT2D* variants between EAS and EUR. **d.** The relative proportions of *AICDA*-induced mutation signature SBS84 (according to COSMIC: <https://cancer.sanger.ac.uk/signatures/sbs/>) in *KMT2D*

variants among EAS, EUR, and AFR from 1KGP. EAS, East Asians; EUR, Europeans; AFR, Africans. **e.** The relative proportions of *AICDA*-induced mutation signature SBS84 in germline variants in *KMT2D* between *KMT2D*-CHIP cases and non-carriers within our cohorts. **f.** The relative proportions of *AICDA*-induced mutation signature SBS84 in somatic variants in *KMT2D* between *KMT2D*-CHIP cases and non-carriers within our cohorts.

We therefore focused on pathway-level mechanisms (Fig. 4a). *KMT2D* functions upstream of multiple tumor suppressor pathways, particularly those involved in B cell-activation and the occurrence of lymphoid tumors²⁹, and is known to be susceptible to genotoxic activity from activation-induced cytidine deaminase (*AICDA/AID*), which can increase genomic instability, and further synergize the lymphomas development³⁰ (Fig. 4a). Analysis of *AICDA* and its ± 200 kb flanking regions revealed 392 highly differentiated loci (top 5% genome-wide) between EAS and EUR populations, with lead SNPs (e.g., rs2580874-A, rs10841924-A, and rs11046277-T) showing high F_{ST} (0.46–0.48) (Fig. 4b), strong linkage in EAS ($LD\ r^2 = 0.98$), and markedly higher allele frequencies in EAS (0.53 vs. 0.07 in EUR), consistent with population-specific selection signals ($|iHS| > 2$) (Supplementary Fig. 17b).

Stratified analyses showed that East Asian individuals carrying differentiated *AICDA* loci exhibited a ~ 1.56 -fold higher germline *KMT2D* variant burden than Europeans (25.9 vs. 16.6 variants per individual; $p < 0.001$, Supplementary Fig. 17d), exceeding the average baseline difference (1.03-fold; Supplementary Fig. 17c). Replication in our Chinese cohort further confirmed that *KMT2D*-CHIP cases harbored an even higher germline and rare variant burden (Supplementary Figs. 17d–e). Moreover, somatic mutations in CHIP cases were significantly enriched for canonical *AID*-related mutations compared with controls ($P < 7.84 \times 10^{-5}$, Fisher's exact test), supporting a role for *AICDA*-associated mutagenesis.

Mutation signature decomposition of the population-specific rare variants (derived allele frequency $< 1\%$) revealed a distinct *KMT2D* mutation spectrum between East Asian and European populations (Fig. 4c left), consistent across germline (Fig. 4c middle) and somatic (Fig. 4c right) levels. Notably, East Asians showed a markedly elevated *AID*-related mutational signature (Single Base Substitution 84, SBS84), which was minimal in European and intermediate in African populations ($10.69\% \pm 0.85\%$ in East Asians vs. $0.77\% \pm 0.55\%$ in Europeans and $3.54\% \pm 1.36\%$ in Africans; Fig. 4d and Supplementary Figs. 19–20). This pattern was replicated in our cohorts,

where *KMT2D*–CHIP cases exhibited higher *AICDA*–associated signatures than matched controls at both germline and somatic level ($15.0\% \pm 1.55\%$ vs. $5.52\% \pm 1.06\%$) (Figs. 4d–e). Consistently, mutations in an *AICDA* context (RCY C > T) were significantly more frequent in East Asians, intermediate in Africans, and lowest in Europeans ($P < 2.2 \times 10^{-16}$, Wilcoxon signed–rank test). We also observed a notable difference of *AICDA*–induced somatic mutations between our *KMT2D*–CHIP case and control groups (number of mutations = 14.4 in cases vs 8.2 in controls, Figs. 4f).

Together, these findings suggest that the elevated *KMT2D* mutation burden in East Asians is unlikely to result from long–term positive selection on *KMT2D* itself, but is more plausibly driven by a higher underlying mutation rate, potentially influenced by population differentiation in *AICDA* and related mutational processes.

Potential ancestry bias in current CHIP whitelist variants

The observed difference in CHIP between Western and Eastern populations promoted us to re–examine global CHIP distribution using 1KGP 30× data, which minimized data generation and sample–size biases. When restricted to the CHIP whitelist (Table S2), prevalence was highest in EUR (5.96%), nearly double that in EAS (2.98%) and South Asian (SAS, 2.86%), while Africans (AFR, 3.34%) and American (AMR, 3.46%) showed intermediate levels (Fig. 5a and Table S6). Consistently, previous reports also show higher CHIP prevalence in European–ancestry populations under comparable conditions (Table S14). Although *DNMT3A* and *TET2* were detected across populations, their frequencies varied by ancestry. In addition, certain genes showed ancestry–specific enrichment, such as *SETD2* in EUR, *KMT2D* in EAS, *WT1* in AFR, and *PPM1D* in SAS, whereas *ASXL1* was not markedly prevalent after accounting for ancestry (Fig. 5b).

When the query was expanded to include all functional variants within the canonical 74–gene panel (expanded CHIP; see Methods), prevalence converged across populations, with no significant ancestry differences. Specifically, AFR showed 12.16%, followed by EUR (11.73%), SAS (11.66%), and AMR (11.52%) (Fig. 5c and Table S11). EAS remained slightly lower (9.32%), but the difference from EUR was not significant (χ^2 test, $P = 0.25$). Gene–level disparities (e.g., *TET2*, *DNMT3A*, and *SETD2*) were also attenuated (Fig. 5d), while *KMT2D* exhibited a notable prevalence in AFR. Nonetheless, ancestry–associated enrichment for specific CHIP genes

persisted. All reported CHIP variants from both whitelist and expanded analyses were manually verified by IGV with reads support (Additional File 3).

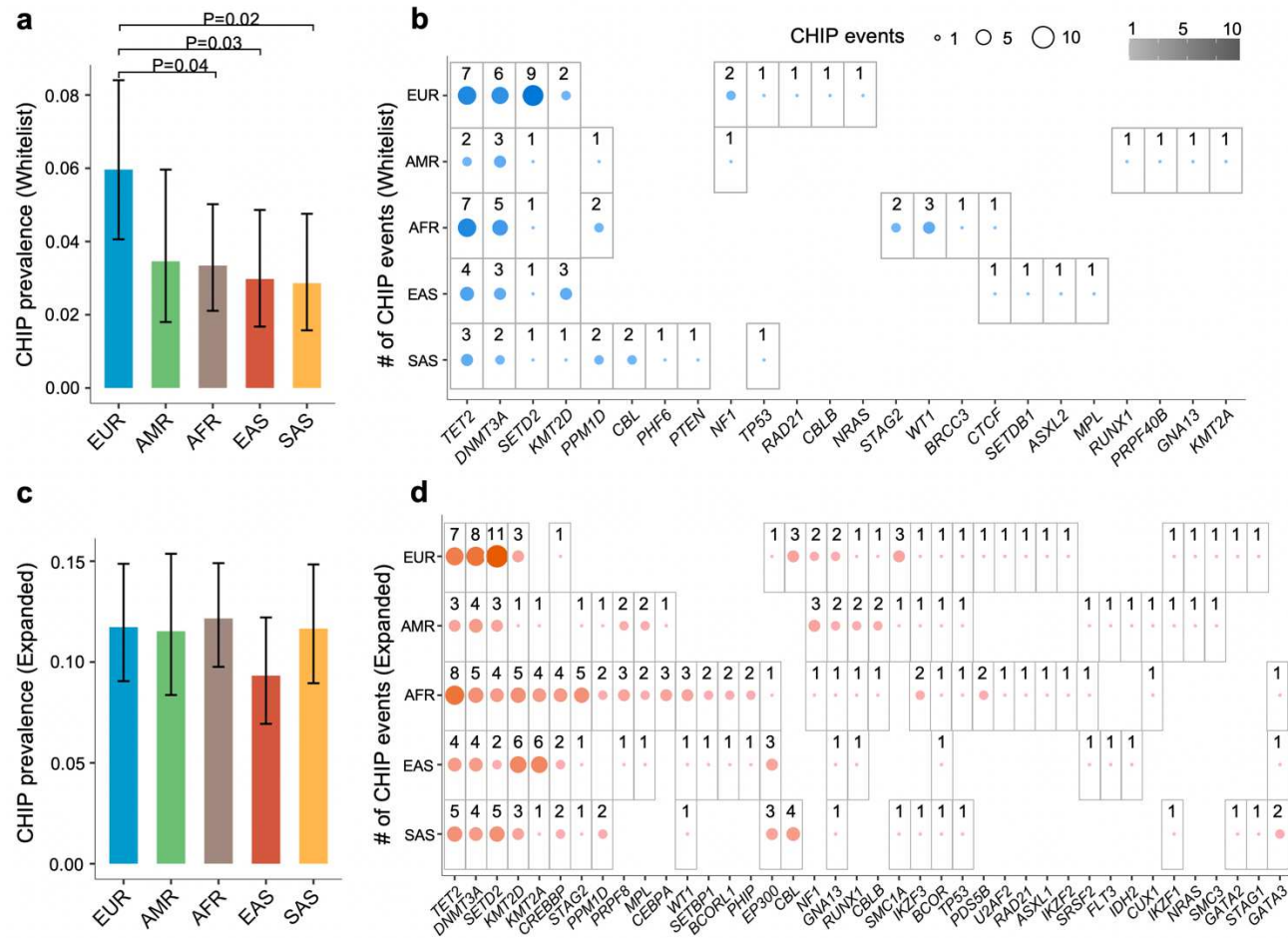


Figure 5 | Potential ancestry bias for the current CHIP whitelist variants mainly curated from European ancestry–dominant populations. **a.** The CHIP prevalence according to the whitelist variants among the 1KGP unrelated individuals from five continents. The χ^2 test was applied to the contingency table of ancestry (EUR vs non–EUR) by presence/absence of whitelist variants, respectively. All reported variants were confirmed by Integrative Genomics Viewer with reads support (Additional File 3). EUR, Europeans; AMR, Americans; AFR, Africans; EAS, East Asians; SAS, South Asians. **b.** Counts of whitelist CHIP events observed within each ancestry group were displayed (related to Table S6). The size and color saturation of the dots represent the number of events. **c.** The CHIP prevalence according to the expanded variant sets among the 1KGP unrelated individuals from five continents, using the same legend as in **a**. **d.** Counts of expanded CHIP events observed within each ancestry group were displayed (related to Table S12), using the same legend as in **b**.

Together, our results indicate that the current CHIP whitelist, largely curated from European-ancestry datasets, likely contains ancestry bias. This may lead to systematic underestimation of CHIP in non-European populations, with the bias increasing with genetic distance from Europeans. A similar pattern has been reported in a large Mexican cohort of ~130,000 admixed individuals³¹, and other previous reports (Table S14). In contrast, the overall functional variant burden across CHIP genes did not differ significantly among populations. Accurate CHIP classification therefore requires population-specific clinical validation and longitudinal follow-up linking genotype to phenotype, similar to the framework used to establish the original whitelist by Jaiswal et al.². Importantly, non-European populations were underrepresented in whitelist construction. Thus, applying the current CHIP whitelist across diverse ancestries should be done cautiously, and future efforts should prioritize broader population representation and ancestry-informed longitudinal validation.

Discussion

In this study, we present the largest and most comprehensive analysis of clonal hematopoiesis of indeterminate potential (CHIP) in Chinese populations to date, encompassing 29,596 individuals from 28 ethnic groups across China, including 13,445 individuals from 23 ethnic groups and 32 administrative regions in the discovery cohort, and 16,151 individuals from 23 ethnic groups and 33 administrative regions in the replication cohort. By leveraging high-depth sequencing and broad multi-ethnic representation, this work addresses critical gaps in CHIP research arising from the historical underrepresentation of non-European populations, which has limited our understanding of ancestral diversity in clonal hematopoiesis. The combination of large sample size, deep sequencing, independent replication, and rigorous quality-control pipelines with read-level confirmation of each reported CHIP variant establishes this dataset as a valuable resource for large-scale CHIP studies in East Asian populations. Our key findings are threefold: (i) CHIP prevalence in Chinese populations is lower than that reported in Europeans, reaching 4.5% in the discovery cohort and 2.8% in the replication cohort; (ii) substantial East-West heterogeneity exists in CHIP mutational spectra, with significant enrichment of *KMT2D* in Asian-ancestry-dominant populations, suggesting that drivers beyond canonical *DTA* genes may play important roles when ancestral diversity is fully considered; and (iii) *AICDA*-associated

mutagenesis, together with pronounced East–West differentiation at *AICDA*, may contribute to the elevated burden of *KMT2D*–CHIP in East Asians.

A direct East–West comparison reveals substantial divergence in CHIP spectra between Chinese and European cohorts. Although *DNMT3A* and *TET2* were consistently common across populations, *KMT2D* was strikingly enriched in Chinese populations and, in some subgroups, even approached or exceeded the prevalence of *ASXL1*. Notably, *KMT2D* emerged as the third or fourth most frequent CHIP drivers across multiple language–family groups (e.g., Sino–Tibetan, Hmong–Mien, and Tai–Kadai) and across 14 geographic regions in China, without evident ethnic, geographic, sex, or age bias. Its distribution pattern more closely resembled that of *DNMT3A* and *TET2* than that of *ASXL1* and remained consistent across both whitelist–defined and expanded variants sets.

Biologically, *KMT2D* encode histone methyltransferase and transcription 10^{-8} 10^{-12} 10^{-5} coactivators and is among the most frequently mutated epigenetic regulators in human cancers³². It is recurrently mutated or downregulated in acute myeloid leukemia (AML), where its deficiency promotes leukemogenesis through dysregulated ribosome biogenesis³³and activation of mTOR pathway³⁴. Loss of *Kmt2d* in hematopoietic stem and progenitor cells and AML cells is associate with enhanced ribosome biogenesis, enlarged nucleoli, and increased rRNA and protein synthesis³⁴. Depending on cancer context, *KMT2D* can function as either a tumor suppressor gene or an oncogene, and in leukemia it has been characterized as a haploinsufficient tumor suppressor¹, which were evident to be recurrently mutated in myeloid cancers. Indeed, *KMT2D* functions as either a tumor suppressor gene or an oncogene depending on the cancer type, while it has been demonstrated as a haploinsufficient tumor suppressor in leukemia³⁵ that regulates ribosome biogenesis³³. Consistently, pan–cancers data indicate that *KMT2D* mutations occur across a wide range of malignancies, including hematopoietic and lymphoid cancers³⁶.

Despite its established role in hematopoietic carcinogenesis, *KMT2D* has been reported at a very low prevalence in European or European–dominant cohorts^{2,9–11,37–39}. Consequently, some large–scale studies (e.g., in both UK Biobank and All of US⁴⁰) have proposed excluding *KMT2D* from CHIP gene lists, and many subsequent CHIP analyses did not include this gene in their queried panels^{9–11}, including earlier studies of East Asian populations¹⁴. In contrast, our results demonstrated that *KMT2D* driver variants are not rare in East Asian CHIP and show appreciable

prevalence in Chinese populations, as well as elevated incidence in Chinese patients with myeloid and lymphocytic leukemia^{41,42}. Furthermore, supporting evidence also comes from smaller East Asian cohorts, such as a Korean CHIP study based on 112 Koreans⁴³ reported a notably higher prevalence of *KMT2D* compared with near absence in European-dominated datasets. We further show that *AICDA*-associated mutagenesis may partially explain the elevated *KMT2D* burden, consistent with higher germline mutation rates and somatic mutational accumulation linked to population-specific genetic backgrounds. Integration of somatic mutation profiles with population-genomic metrics (e.g., F_{ST} , $iHS/XP-EHH$, and LD patterns) suggests that population-specific evolutionary forces, together with germline-somatic interactions, may shape the architecture of CHIP risk across ancestries.

Current knowledge of CHIP prevalence and clinical implications has largely been derived from European cohorts. Our findings, together with global re-analyses, indicate that this European-centric perspective may obscure ancestry-specific features of CHIP, partly due to biases in the current CHIP whitelist, which was predominantly curated from European datasets. The canonical 74-genes CHIP panel (often with predefined hotspot variants) may therefore be insufficient to capture the full spectrum of CHIP across global populations, particular when analyses focus on only on the most frequent genes identified in Western cohorts. This limitation likely explains why *KMT2D* was not detected in previous Chinese CHIP studies based on smaller, single-center datasets with limited ancestral diversity^{14,15}. Our results highlight the need for ancestry-aware screening strategies and support the inclusion of *KMT2D* as a key CHIP driver in East Asian populations. More broadly, CHIP gene panels should be iteratively updated as data from underrepresented populations, including African, Middle Eastern, Southeast Asian, and South Asian populations, become available.

Importantly, a substantial proportion of clonal expansions in blood remain unexplained by currently recognized CHIP drivers. In our dataset, the accumulation of high-confidence somatic variants (~50.3 per individual at ~40× depth) far exceeded the number of variants classified as CHIP. This observation is consistent with the concept of ‘driverless’ clonal hematopoiesis, in which clonal expansion occur without identifiable canonical drivers and may account for the majority of clonal expansions in older individuals. Such expansions have been associated with increased risks of infection, mortality, and hematological malignancies, underscoring the

importance of expanding the search for additional driver genes and mechanisms, particularly in non-European populations. The observed East–West differences in CHIP architecture therefore motivate future investigations to systematically identify ancestry–specific CHIP drivers in Eastern populations.

Several limitations warrant cautious interpretation. First, despite the large overall sample size, some ethnic minority groups remain underrepresented, limiting the power of fine-scale ancestry-specific analyses. Second, expanded CHIP analyses rely on broader gene sets with heterogeneous levels of functional evidence across populations. Third, cross-cohort comparisons are subject to potential batch effects, demographic differences, and substantial variation in sequencing depth, which may directly influence mutation detection sensitivity and reproducibility, particularly at lower variant allele fractions. Although we incorporated some of the most comprehensive high-depth sequencing data currently available for Chinese minority populations, the dataset still does not fully capture the extensive ancestral diversity within China, and non-Han subgroup sample sizes remain modest. Nevertheless, this study represents the first systematic exploration of CHIP across multiple and diverse Chinese ethnic minority groups, which have been largely overlooked in global research that has predominantly focused on Han populations. Given the underrepresentation of East Asian populations in global CHIP studies, especially ethnic minority subgroups, our dataset provides an important initial framework, and the subgroup-specific findings reported here should be re-evaluated as larger high-depth datasets become available.

Additionally, our analyses are primarily based on available datasets not originally designed for CHIP research, resulting in incomplete phenotypic information (e.g., smoking history and other environmental exposures). Further validation using orthogonal approaches, including long-read sequencing and functional assays, is needed to confirm *KMT2D*–CHIP and other CHIP variants and to evaluate their effects on hematopoietic lineage bias. Experimental studies are also required to directly test whether *AICDA*–mediated processes contribute to elevated somatic mutational burden in *KMT2D* and to refine predictions of CHIP onset and progression. It is also important to acknowledge that CHIP detection is highly dependent on sequencing depth; thus, studies with different sequencing depths and ancestry compositions may yield discrepant prevalence estimates^{9,10,31}. Nevertheless, across both discovery and replication cohorts, as well as in 1KGP

analyses, we consistently observed a robust enrichment of *KMT2D* in East Asian–ancestry–predominant populations, supporting the reliability and ancestry relevance of our findings.

In summary, our study demonstrates substantial differences in both overall prevalence and mutational landscapes of CHIP between Chinese and European–dominant populations. We propose that ancestry–specific CHIP architectures may exist, exemplified by the elevated prevalence of *KMT2D* mutations in East Asians, potentially influenced in part by population differentiation at *AICDA*. These results suggest that current CHIP gene panels, largely derived from European–ancestry cohorts, require revision to incorporate drivers that are more frequent in non–European populations. More broadly, our findings underscore the necessity of ancestry–aware screening frameworks and diversity–inclusive study designs in precision medicine, which are essential for accurately characterizing clonal hematopoiesis and for reducing global health disparities.

Methods

Study populations

In this study we included the Chinese populations consisting of 29,596 participants with blood DNA whole genome or –exome sequencing (WGS or WES) data available from the POG13K (discovery cohort, N = 13,445) and CYN16K (replication cohort, N = 16,151). All samples were sampled from mainland China crossing 33 first–level administrative regions, and covering 28 self–reported minority ethnic subgroups who represent underrepresented populations and minority groups, including Manchu, Mongolian, Chosen, Tujia, Tibetan, Lhoba, Monba, Hui, Bai, Yi, Han, Lahu, Hani, Naxi, Lisu, Dai, Gelao, Pumi, Miao, Yao, Blang, Nu, Dong, Zhuang, Bouyei, Li, She, and Uyghur populations. The study excluded individuals with a diagnosis of hematologic malignancy. All Chinese participants in the study (N = 29,596) are being used for the first time to investigate CHIP. We also utilized the 1KGP unrelated individuals (N = 2,504, 30× WGS) to make an independent validation and the genetic differentiation of CHIP related genes between EAS and EUR. Descriptions of the individual cohort characteristics and sequencing methods are briefly described below (see [Table S1](#) for additional information).

POG13K. The POG13K is a population–based cohort of 13,445 middle–aged individuals in the mainland China covering 32 first–level administrative regions. WGS was performed on

5,258 samples sequenced as previously described^{44–47}, and additional datasets of 8,187 WES were generated as previously reported^{48–50}, with the mean sequencing depth of 74.4× across the cohort. The age of participants at the time of blood draw was obtained for a subset comprising 12,213 of the samples, aged 18–107 years (mean age of 46.00 years, median 44 years). The age distribution varied across the constituent cohorts (Table S1). The POG13K was consisted of Han and 22 other ethnic minorities from seven language families, including Tungusic (Manchu), Mongolic (Mongolian), Koreanic (Chosen), Sino–Tibetan (Tujia, Tibetan, Lhoba, Monba, Hui, Bai, Yi, Lahu, Hani, Naxi, and Lisu), Hmong–Mien (Miao and Yao), Tai–Kadai (Dong, Zhuang, Bouyei, Li, and Dai), and Turkic (Uyghur) (Fig. 1a). The Han population accounted for 94.9%. The gender information was collected from a total of 12,744 individuals, comprising 6,708 males and 6,036 females. The samples were primarily collected from Shanghai, Henan Province, Guangxi Zhuang Autonomous Region, Jiangsu Province, Hubei Province, Shanxi Province, and Yunnan Province (sample sizes > 400), covering the major North–South and East–West axes of mainland China (Supplementary Fig. 2a).

CYN16K. The CYN16K is a comparable population–based cohort of 16,151 Chinese individuals covering 33 first–level administrative regions in China, which was utilized as an independent replication cohort in this study. The CYN16K was consisted of 23 ethnic subgroups from five language families, including Tungusic (Manchu), Mongolic (Mongolian), Sino–Tibetan (Tujia, Gelao, Tibetan, Hui, Bai, Yi, Lahu, Lhoba, Han, Hani, Naxi, and Lisu), Hmong–Mien (Pumi, Miao, Blang, and Nu), and Tai–Kadai (Dong, Zhuang, Bouyei, Li, and She) (Supplementary Fig. 4a). The Han population accounted for 89.3%. All data in the CYN16K cohort were derived from WGS, with an overall mean sequencing depth of 42.2× as previously reported^{51,52}, comprising 7,073 males and 9,078 females. The age of participants at the time of blood draw was obtained for a subset comprising 16,148 of the samples, aged 18–98 years (mean age of 53.87 years, median 54.4 years). All samples were clinically confirmed to be free of hematologic malignancy at the time of sampling. The samples were primarily collected from Shanghai, Zhejiang Province, Guangdong Province, Hubei Province, Henan Province, Fujian Province, Yunnan Province, Shandong Province, Jiangsu Province, Guizhou Province, Sichuan Province, Hunan Province, and Shaanxi Province (sample sizes > 400), mainly covering the Eastern, Central, Southern and Southwestern regions of China (Supplementary Fig. 2b). represent

the regional diversity of China. CYN16K and POG13K were both collected from different regions in China, and both represent the regional diversity of China.

1000 Genomes Project (1KGP) cohort. We utilized 1KGP unrelated individuals (N = 2,504) with high-coverage WGS data ($\sim 33.8\times$)⁵³ as an external validation set, which included 504 East Asian, 503 European, 489 South Asian, 347 American, and 657 African populations. This eliminated the sequencing batch between different sample collection and data-generation processes, while also controlling the ratio of samples from East Asian and European populations. As a result, the dataset is well suited for investigating East-West CHIP differences and their genetic differentiation. We downloaded the raw FASTQ data and processed germline and somatic mutation calling, as well as CHIP filtration, using the same pipeline to ensure consistency with our cohort processing.

Ethical statement

All the authors in this study are well educated and aware of the ethical issues, and are capable of explaining explicitly the results to minority communities. Written informed consent was obtained from all human participants in each of the studies that contributed to the Chinese population cohort utilized in this analysis with approval of study protocols by ethics committees at participating institutions, as summarized in [Table S1](#). The personal identifiers of all samples, if any existed, were stripped off before sequencing and analysis. Secondary analysis of the Chinese populations cohort data as described in this manuscript was approved by the Biomedical Research Ethics Committee of Fudan University (No. BE25149I; FE25478I). All procedures were carried out in accordance with the ethical standards of the Responsible Committee on Human Experimentation and the 1964 Helsinki Declaration, its later amendments (2000), or comparable ethical standards.

Whole-exome and -genome sequencing (WES & WGS), and pre-processing

In POG13K and 1KGP, whole-exome and genome sequencing data for genomic DNA was generated as previously described⁴⁴⁻⁵⁰. Raw data were cleaned using the following steps: both paired reads were removed if any one of the reads 1) contained a sequencing adapter, 2) the ratio of low-quality base (base quality ≤ 12) $> 50\%$ or 3) the ratio of unknown base ('N' base) $> 10\%$. Low-quality reads and bases were filtered out using FastQC (V.0.11.9), and these qualified data

were analyzed by downstream bioinformatics. Sequencing read alignment to the GRCh38 genome reference ‘GRCh38_full_analysis_set_plus_decoy_hla.fa’ using BWA with ‘alt-aware’ strategy, and to generate both germline and somatic variant calling ready .bam file after Base Quality Score Recalibration (BQSR) by GATK. In CYN16K, whole genome sequencing data from genomic DNA was generated as previously described^{51,52}, which raw sequence reads for both forward and reverse were filtered with $Q30 \geq 80\%$, and $Max_n_content < 2\%$. The mapping quality was filtered with mapping rate $\geq 95\%$, mismatch rate $< 1\%$, duplicate rate $< 10\%$, mean sequence depth $\geq 30\times$, $20\times$ coverage $\geq 80\%$, and contamination $< 5\%$ ⁵¹.

Germline variant calling

Within the POG13K and CYN16K cohorts, germline single-nucleotide variants (SNVs) and small insertions/deletions (indels) were called and jointly genotyped using GATK HaplotypeCaller, respectively, as previously described^{44–52}.

Somatic variant calling

Putative somatic SNPs and short indels were called with GATK Mutect2 v.4.2.1.0²² from .bam file after BQSR, and filtered with a stringent pipeline to remove artifacts and germline variants. In brief, Mutect2 searches for sites where there is evidence for variation and then performs local reassembly. This software aims to use mapping quality measures as well as allele frequency information to identify somatic mutations against a background of germline mutations and sequencing errors. We used data generated from gnomAD v4.1 as the reference source for germline allele frequency. A public panel of healthy controls ([gs://gatk-best-practices/somatic-hg38/1000g_pon.hg38.vcf.gz](https://gatk-best-practices/somatic-hg38/1000g_pon.hg38.vcf.gz)) was used to remove potential recurrent artifacts with GATK FilterMutectCalls. In addition, the ‘-orientation-bias-artifact-priors’ option was also specified to remove read orientation artifacts based on priors generated with LearnReadOrientationModel. All detected variants were annotated using Ensembl Variant Effect Predictor (VEP, V.107).

CHIP filtration

Predefined whitelist CHIP. To identify CHIP driver variants, we first retrieved Mutect2 PASS somatic variants occurring in a previously defined whitelist of 74 genes with leukemogenic

driver mutations² (Table S2). We applied a series of postprocessing filters to further remove false-positive variants caused by sequencing artifacts and putative germline polymorphisms. Only variants supported by 1) at least two alternate allele reads, 2) at least one from both forward and reverse reads, 3) a total read depth of large than 15×, 4) a variant allele fraction (VAF) of at least 2%, and 5) present at a high VAF of $\geq 35\%$, unless it was a clear cancer somatic hotspot according to criteria as previously described⁵⁴, were retained. Variant allele fraction (VAF) was calculated as the ratio of the depth at the variant allele to the total depth at that locus. Samples were defined as the CHIP carrier if detected with one or more whitelist CHIP variants.

Expanded CHIP. Additionally, using identical variant filtering criteria, we extended CHIP detection to functional variants in 74-gene panel beyond the whitelist. All the functional variants in the canonical 74 CHIP genes were retained as expanded CHIP variants, including “stop_gained”, “stop_lost”, “start_lost”, “frameshift_variant”, “splice_acceptor_variant”, “splice_donor_variant”, and “missense_variant” annotated by VEP V.107, as previously described^{9,11}. We applied same postprocessing filters as the whitelist CHIP, to the expanded CHIP variants.

Validation of calls

We performed manual review in Integrative Genomics Viewer (IGV) of CHIP whitelist mutations reported in our study, i.e., POG13K, CYN16K, and 1KGP, respectively (Additional Files 1–3), and the missense variant in *KMT2D* were also confirmed by IGV (Additional File 4).

CHIP gene panel to characterize the East–West CHIP difference

Considering that our dataset (e.g., the discovery cohort POG13K) include WGS and WES data, and that the current findings of CHIP studies are mainly based on UKB, we have selected the union of three Western studies (i.e., 17,182 WES cohorts², 97,691 WGS cohorts⁹, and 200,453 WES cohorts¹¹) as Western CHIP findings (Table S8.1–S8.3) to compare with our Chinese populations. We take the union among the top 15 frequently mutated CHIP genes in Western populations^{2,9,11} (Supplementary Fig. 1), and the top 15 in Chinese populations (POG13K and CYN16K, respectively), as the most frequently mutated CHIP genes to character the difference between Western and East Asian populations, i.e., 22 genes for Western vs. POG13K (Supplementary Fig. 9a), and 23 genes for Western vs. CYN16K (Supplementary Fig. 9b).

696 Additionally, we merged the previously defined CHIP genes (Table S2) and 259 genes recurrently
697 mutated in mature lymphoid neoplasms and absent in the canonical CHIP gene panel ($N = 74$), as
698 the CHIP-related genes ($N = 333$) (Supplementary Fig. 1 and Table S3) to investigate the genetic
699 differentiation between EAS and EUR. These genes were selected based on reported mutational
700 frequency at diagnosis and contribution to disease pathogenesis, molecular classification, and
701 clinical risk stratification within prevalent lymphoma subtypes, including chronic lymphocytic
702 leukemia, diffuse large B-cell lymphoma, follicular lymphoma, mantle cell lymphoma, marginal
703 zone lymphoma, Hodgkin's lymphoma (HL), and peripheral T-cell lymphoma⁵⁵.

704 **Estimating genetic differentiation and population-specific selection of CHIP between** 705 **populations**

706 Genetic differentiation between Western and Eastern populations (1KGP, EUR vs. EAS)
707 were measured using vcftools⁵⁶ with F_{ST} (Fixation Index) according to Weir and Cockerham⁵⁷
708 using SNPRelate v1.16.0⁵⁸ which allows the correction for different sample sizes of populations.
709 We investigated the population differentiation of CHIP from the variant and gene level,
710 respectively, i.e., the variant-level F_{ST} between EUR and EAS for each SNV and INDEL; the gene
711 level F_{ST} was obtained by partitioning the genome into sliding windows of 50 Kb in length shift
712 by 25 Kb and only the top 10 F_{ST} value within each window were retained⁵⁹. For each ethnic
713 minority in our Chinese cohorts, we also estimated their genetic differentiation between Han
714 Chinese using Weir and Cockerham's F_{ST} ⁵⁷ using the same pipeline.

715 We then explored the signatures of positive selection underlying the genetic basis of CHIP
716 among EUR and EAS using haplotype-homozygosity-based methods (i.e., integrated Haplotype
717 Score, iHS, and Cross-Population Extended Haplotype Homozygosity, XP-EHH). We first
718 inferred the ancestral and derived states for each variant using the ancestral alignments of Homo
719 sapiens from Ensembl release 109 (GRCh38 build)⁶⁰, which were based on six non-human
720 primates. Next, we calculated the iHS and XP-EHH values using selscan
721 (<https://github.com/szpiech/selscan>)⁶¹. Only variants exhibited $|iHS|$ or $XP-EHH > 2$ were
722 considered as candidate signals of selective sweep.

Rare variant burden for essential CHIP genes in Western populations and our Chinese cohort

Based on the variant frequencies obtained from gnomAD (v4.1) 76,156 WGS sequencing data (Number of East Asian individuals = 2,604, Number of non-Finnish European N = 39,435), we first analyzed the population-specific rare burden of ancestry-specific rare variants in CHIP-related genes at the germline level variants burden). We define SNPs with minor allelic frequencies (MAF) less than 0.001 in East Asian ancestral populations as rare variants specific to East Asian ancestry and SNPs with MAF less than 0.001 in European ancestral populations as rare variants specific to European ancestry. Finally, we used the Wilcoxon rank sum test to compare the population-specific mutation burden of rare variants on CHIP-related genes.

Mutation burden analyses

We extracted *KMT2D* germline variants from *KMT2D*-CHIP cases (N = 39) identified in POG13K. To equalize sample size with the *KMT2D*-CHIP cohort, EUR individuals from the 1KGP were randomly downsampled (39 individuals per draw) and repeated 100 times to derive a distribution of *KMT2D* germline variants; one batch of 39 individuals which was closest to the mean across the 100 resamples was selected as the EUR reference sample. We plotted the *KMT2D* germline variants of the *KMT2D*-CHIP and the downsampled EUR individuals on the same axis to directly show the difference between them (Supplementary Fig. 17a). Given the well-established genetic differences between EAS and EUR populations, genome-wide germline variants burden was compared directly between EAS and EUR individuals to establish a baseline reference for East-West differential ratio (as indicated in Supplementary Fig. 17c). We then investigated whether the East-West differentiation of *AICDA* contributed to the East-West differences of germline variants burden in *KMT2D*. Guided by *AICDA* differentiation site (i.e., with a higher frequency in EAS, but lower in EUR), together with *KMT2D*-CHIP individuals from POG13K, we extracted their germline *KMT2D* variants from these three groups (EAS, EUR, *KMT2D*-CHIP), and compared the *KMT2D* all-variants and rare-variants burden, respectively (Supplementary Figs. 17d-e). The *AICDA*-differentiation-aware ratio was then contrasted with the East-West baseline to indicate the deviation.

Mutation signature analyses

To compare the mutational spectrum of single-nucleotide variants in *KMT2D* across different populations, we performed three-fold mutation signatures (i.e., single base substitution, SBS) analysis, leveraging: 1) the population-specific rare germline mutations of *KMT2D* of East Asian and European populations in 1KGP dataset (30×, 2022 release) to simulate the mutational process of somatic mutations⁶²; 2) germline rare variants and 3) somatic mutations of *KMT2D* of CHIP case and control from our POG13K cohort. To offset the risk introduced by imbalance genomic coverage, we retained individuals with whole-genome sequencing^{63,64}. Ancestral statuses were determined using the ancestral sequence from Ensembl release 109 (http://ftp.ensembl.org/pub/release-109/fasta/ancestral_alleles/). We then fit the mutational spectrum using R package ‘deconstructSigs’⁶⁵ against the COSMIC v3.4 SBS signatures, which represent signatures of 96 mutation classes. Active mutational processes were determined with a signature contribution cutoff of 6% to reduce the false positive discovery rate, following an established workflow⁶⁶. Furthermore, we perturbed the original mutation spectrum of each group by 100 times of bootstrapping, and estimated the signature contributions in each iteration.

Statistical analysis and data visualization

All statistical analysis and data visualization for Figs. were generated using R v4.3.1 (<https://www.r-project.org/>), mainly utilizing libraries from ggplot2 v3.4.3, GenVisR v1.30.0 and ggpubr v0.6.0.

Acknowledgements

We thank all the participants of this project. This study was supported by the National Key Research and Development Program of China (No. 2023YFC2605400), the National Natural Science Foundation of China (NSFC) grants (32288101, 32470649, 323B2013, 32300499, 82270842, 82470830), the Shanghai Science and Technology Commission Program (25JS2810100, 23JS1410100, QNKJ2024023), the Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (JYB2025XDXM508), the Shanghai Municipal Science and Technology Major Project (2017SHZDZX01), the CAMS Innovation Fund for Medical Sciences (2019-I2M-5-066), and the Office of Global Partnerships (Key Projects Development Fund). The computational work in this study was supported by the CFFF Computing Platform of Fudan University. We thank LetPub (www.letpub.com.cn) for its linguistic assistance during the preparation of this manuscript. The funders had no role in the study design, data collection, analysis, decision to publish, or preparation of the manuscript.

Author Contributions

S.X. and X.T. conceived and designed the study. S.X., L.J., X.J., Y.C. and Y.L. supervised the project. Y.L., Y.G., X.T., Y.H., and Q.L., contributed to sample collection. S.X., YN.C., L.J., X.J., J.W., S.W., J.L., Y.Z., and B.S. contributed data and reagents. Y.L. managed laboratory work and discussed with constructive comments to the project. B.W. contributed to the High Performance Computing server management, ensuring the smooth operation of the computing environment. X.T. developed the pipeline for processing somatic variants calling and CHIP filtering, and Y.S. posted the workflow to GitHub by snakemake. Y.G. updated the pipeline for processing the germline mutation of NGS data, X.T., YN.C., Y.G., Y.S. and Y.S. performed variant-calling analysis. X.T., Y.C., H.C., X.Z., M.W. and Y.G. contributed to data analysis. X.T. and Y.S. phased the phenotype data. H.C. and X.T. performed West-East differentiation and population genetic analysis. YH.C. performed the genetic differentiation of ethnic minorities and Han Chinese. X.Z. and X.T. contributed to the IGV confirmation and prepared related Supplementary Materials. H.C, X.Z. and S.C. helped the analysis of germline-somatic interplay. L.D. took charge of the ethical issues and application for data archiving. X.T. wrote the main paper and prepared the additional files with the input from H.C. and X.Z.. S.X. and Y.L. substantially

revised the manuscript. All authors discussed the results and implications and commented on the manuscript.

Data availability

The use of genetic data in this study has been approved by the National Health Commission of the People's Republic of China (2025BAT00815). The genetic data have been deposited at the NGDC database (China National Center for Bioinformation) with the accession number of PRJCA036736, and are publicly available as of the date of publication. All data generated or analyzed during this study are included in this Article, its Supplementary Information, and publicly available repositories. Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

Code availability

All original code has been deposited at <https://github.com/Shuhua-Group/CHIP-mutation-calling-NGS-pipeline>. The analyses were performed using publicly available software and web-based applications are described in the Methods section.

814 **References**

- 815 1. Jaiswal, S. *et al.* Age-Related Clonal Hematopoiesis Associated with Adverse Outcomes.
816 *N Engl J Med* **371**, 2488–2498 (2014).
- 817 2. Jaiswal, S. *et al.* Clonal Hematopoiesis and Risk of Atherosclerotic Cardiovascular
818 Disease. *N Engl J Med* **377**, 111–121 (2017).
- 819 3. Jaiswal, S. & Ebert, B. L. Clonal hematopoiesis in human aging and disease. *Science* **366**,
820 ean4673 (2019).
- 821 4. Zhou, Y. *et al.* Clonal hematopoiesis is not significantly associated with COVID-19
822 disease severity. *Blood* **140**, 1650–1655 (2022).
- 823 5. Qiu, X. *et al.* Somatic mutation contributing to clonal haematopoiesis is a risk factor of
824 recurrent stroke in first-ever acute ischaemic stroke: A prospective cohort study. *Stroke Vasc*
825 *Neurol* **8**, 103–110 (2023).
- 826 6. Ahmad, H., Jahn, N. & Jaiswal, S. Clonal Hematopoiesis and Its Impact on Human
827 Health. *Annu. Rev. Med.* **74**, 249–260 (2023).
- 828 7. Wong, W. J. *et al.* Clonal haematopoiesis and risk of chronic liver disease. *Nature* **616**,
829 747–754 (2023).
- 830 8. Vlasschaert, C. *et al.* Clonal hematopoiesis of indeterminate potential is associated with
831 acute kidney injury. *Nat Med* **30**, 810–817 (2024).
- 832 9. Bick, A. G. *et al.* Inherited causes of clonal haematopoiesis in 97,691 whole genomes.
833 *Nature* **586**, 763–768 (2020).
- 834 10. Kessler, M. D. *et al.* Common and rare variant associations with clonal haematopoiesis
835 phenotypes. *Nature* **612**, 301–309 (2022).
- 836 11. Kar, S. P. *et al.* Genome-wide analyses of 200,453 individuals yield new insights into the
837 causes and consequences of clonal hematopoiesis. *Nat Genet* **54**, 1155–1166 (2022).

- 838 12. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data.
839 *Nature* **562**, 203–209 (2018).
- 840 13. Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed
841 Program. *Nature* **590**, 290–299 (2021).
- 842 14. Zhao, K. *et al.* Somatic and Germline Variants and Coronary Heart Disease in a Chinese
843 Population. *JAMA Cardiol* **9**, 233 (2024).
- 844 15. Wang, K. *et al.* The impact of age and number of mutations on the size of clonal
845 hematopoiesis. *Proc. Natl. Acad. Sci. U.S.A.* **121**, e2319364121 (2024).
- 846 16. Gao, Y. *et al.* A pangenome reference of 36 Chinese populations. *Nature* **619**, 112–121
847 (2023).
- 848 17. The HUGO Pan-Asian SNP Consortium *et al.* Mapping Human Genetic Diversity in
849 Asia. *Science* **326**, 1541–1545 (2009).
- 850 18. Xu, S. *et al.* Genomic Dissection of Population Substructure of Han Chinese and Its
851 Implication in Association Studies. *The American Journal of Human Genetics* **85**, 762–774
852 (2009).
- 853 19. Feng, Q. *et al.* Genetic History of Xinjiang’s Uyghurs Suggests Bronze Age Multiple-
854 Way Contacts in Eurasia. *Molecular Biology and Evolution* **34**, 2572–2582 (2017).
- 855 20. Pan, Y. *et al.* Genomic diversity and post-admixture adaptation in the Uyghurs. *National*
856 *Science Review* **9**, nwab124 (2022).
- 857 21. Wen, J. *et al.* Ancestral origins and post-admixture adaptive evolution of highland Tajiks.
858 *National Science Review* **11**, nwae284 (2024).
- 859 22. Benjamin, D. *et al.* Calling Somatic SNVs and Indels with Mutect2. (2019)
860 doi:10.1101/861054.
- 861 23. Wen, S. *et al.* Comparative analysis of 136,401 Admixed Americans and 419,228
862 Europeans reveals ancestry-specific genetic determinants of clonal haematopoiesis. (2024)
863 doi:10.1101/2024.02.07.24302442.

- 864 24. Feusier, J. E. *et al.* Large-scale Identification of Clonal Hematopoiesis and Mutations
865 Recurrent in Blood Cancers. *Blood Cancer Discovery* **2**, 226–237 (2021).
- 866 25. Uddin, M. M. *et al.* Germline genomic and phenomic landscape of clonal hematopoiesis
867 in 323,112 individuals. (2022) doi:10.1101/2022.07.29.22278015.
- 868 26. Terao, C. *et al.* Chromosomal alterations among age-related haematopoietic clones in
869 Japan. *Nature* **584**, 130–135 (2020).
- 870 27. Jurgens, S. J. *et al.* Rare coding variant analysis for human diseases across biobanks and
871 ancestries. *Nat Genet* **56**, 1811–1820 (2024).
- 872 28. Weinstock, J. S. *et al.* The genetic determinants of recurrent somatic mutations in 43,693
873 blood genomes. *Sci. Adv.* **9**, eabm4945 (2023).
- 874 29. Von Beck, K., Von Beck, T., Ferrell, P. B., Bick, A. G. & Kishtagari, A. Lymphoid
875 clonal hematopoiesis: Implications for malignancy, immunity, and treatment. *Blood Cancer J.*
876 **13**, 5 (2023).
- 877 30. Ortega-Molina, A. *et al.* The histone lysine methyltransferase KMT2D sustains a gene
878 expression program that represses B cell lymphoma development. *Nat Med* **21**, 1199–1208
879 (2015).
- 880 31. Wen, S. *et al.* Comparative analysis of the Mexico City Prospective Study and the UK
881 Biobank identifies ancestry-specific effects on clonal hematopoiesis. *Nat Genet* **57**, 572–582
882 (2025).
- 883 32. Issa, G. C. *et al.* Therapeutic implications of menin inhibition in acute leukemias.
884 *Leukemia* **35**, 2482–2495 (2021).
- 885 33. Xu, J., Zhong, A., Chen, C., Liu, Y. & Niu, T. KMT2D Deficiency Promotes Myeloid
886 Leukemias through Regulating Ribosome Biogenesis. *Blood* **140**, 5855–5856 (2022).
- 887 34. Xu, J. *et al.* KMT2D Deficiency Promotes Myeloid Leukemias which Is Vulnerable to
888 Ribosome Biogenesis Inhibition. *Advanced Science* **10**, 2206098 (2023).

- 889 35. Xu, J., Liu, Y., Chen, C. & Niu, T. KMT2D Is a Haploinsufficient Tumor Suppressor in
890 Acute Leukemia. *Blood* **132**, 1511–1511 (2018).
- 891 36. Sondka, Z. *et al.* COSMIC: A curated database of somatic variants and clinical data for
892 cancer. *Nucleic Acids Research* **52**, D1210–D1217 (2024).
- 893 37. Steensma, D. P. Clinical consequences of clonal hematopoiesis of indeterminate
894 potential. *Blood Advances* **2**, (2018).
- 895 38. Uddin, M. M. *et al.* Longitudinal profiling of clonal hematopoiesis provides insight into
896 clonal dynamics. *Immun Ageing* **19**, 23 (2022).
- 897 39. Weinstock, J. S. *et al.* Aberrant activation of TCL1A promotes stem cell expansion in
898 clonal haematopoiesis. *Nature* **616**, 755–763 (2023).
- 899 40. Vlasschaert, C. *et al.* A practical approach to curate clonal hematopoiesis of
900 indeterminate potential in human genetic datasets. *Blood* blood.2022018825 (2023)
901 doi:10.1182/blood.2022018825.
- 902 41. Zhong, W. *et al.* Comparison of gene mutation spectra in younger and older Chinese
903 acute myeloid leukemia patients and its prognostic value. *Gene* **770**, 145344 (2021).
- 904 42. Yi, S. *et al.* High incidence of MYD88 and KMT2D mutations in Chinese with chronic
905 lymphocytic leukemia. *Leukemia* **35**, 2412–2415 (2021).
- 906 43. Bolton, K. L. *et al.* Clonal hematopoiesis is associated with risk of severe Covid-19. *Nat*
907 *Commun* **12**, 5975 (2021).
- 908 44. Pan, Y. *et al.* Lineage-specific positive selection on *ACE2* contributes to the genetic
909 susceptibility of COVID-19. *National Science Review* **9**, nwac118 (2022).
- 910 45. Zhu, H. *et al.* A Chinese host genetic study discovered IFNs and causality of laboratory
911 traits on COVID-19 severity. *iScience* **24**, 103186 (2021).
- 912 46. Wu, P. *et al.* The trans-omics landscape of COVID-19. *Nat Commun* **12**, 4543 (2021).

- 913 47. Wang, F. *et al.* Initial whole-genome sequencing and analysis of the host genetic
914 contribution to COVID-19 severity and susceptibility. *Cell Discov* **6**, 83 (2020).
- 915 48. Hao, M. *et al.* The HuaBiao project: Whole-exome sequencing of 5000 Han Chinese
916 individuals. *Journal of Genetics and Genomics* **48**, 1032–1035 (2021).
- 917 49. Gao, Q. *et al.* Integrated Proteogenomic Characterization of HBV-Related Hepatocellular
918 Carcinoma. *Cell* **179**, 561–577.e22 (2019).
- 919 50. Wu, R. *et al.* Exome Sequencing Identified Susceptible Genes for High Residual Risks in
920 Early-Onset Coronary Atherosclerotic Disease. *Clinical Cardiology* **47**, e70066 (2024).
- 921 51. The ChinaMAP Consortium *et al.* The ChinaMAP analytics of deep whole genome
922 sequences in 10,588 individuals. *Cell Res* **30**, 717–731 (2020).
- 923 52. Li, L. *et al.* The ChinaMAP reference panel for the accurate genotype imputation in
924 Chinese populations. *Cell Res* **31**, 1308–1310 (2021).
- 925 53. Byrska-Bishop, M. *et al.* High-coverage whole-genome sequencing of the expanded 1000
926 Genomes Project cohort including 602 trios. *Cell* **185**, 3426–3440.e19 (2022).
- 927 54. Bolton, K. L. *et al.* Cancer therapy shapes the fitness landscape of clonal hematopoiesis.
928 *Nat Genet* **52**, 1219–1226 (2020).
- 929 55. Niroula, A. *et al.* Distinction of lymphoid and myeloid clonal hematopoiesis. *Nat Med* **27**,
930 1921–1927 (2021).
- 931 56. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158
932 (2011).
- 933 57. Weir, B. S. & Cockerham, C. C. Estimating F-Statistics for the Analysis of Population
934 Structure. *Evolution* **38**, 1358 (1984).
- 935 58. Zheng, X. *et al.* A high-performance computing toolset for relatedness and principal
936 component analysis of SNP data. *Bioinformatics* **28**, 3326–3328 (2012).

937 59. Pickrell, J. K. *et al.* Signals of recent positive selection in a worldwide sample of human
938 populations. *Genome Res.* **19**, 826–837 (2009).

939 60. Martin, F. J. *et al.* Ensembl 2023. *Nucleic Acids Research* **51**, D933–D941 (2023).

940 61. Szpiech, Z. A. & Hernandez, R. D. Selscan: An Efficient Multithreaded Program to
941 Perform EHH-Based Scans for Positive Selection. *Molecular Biology and Evolution* **31**, 2824–
942 2827 (2014).

943 62. Mathieson, I. & Reich, D. Differences in the rare variant spectrum among human
944 populations. *PLoS Genet* **13**, e1006581 (2017).

945 63. Vali-Pour, M. *et al.* The impact of rare germline variants on human somatic mutation
946 processes. *Nat Commun* **13**, 3724 (2022).

947 64. Park, S., Supek, F. & Lehner, B. Systematic discovery of germline cancer predisposition
948 genes through the identification of somatic second hits. *Nat Commun* **9**, 2601 (2018).

949 65. Rosenthal, R., McGranahan, N., Herrero, J., Taylor, B. S. & Swanton, C.
950 deconstructSigs: Delineating mutational processes in single tumors distinguishes DNA repair
951 deficiencies and patterns of carcinoma evolution. *Genome Biol* **17**, 31 (2016).

952 66. Caswell, D. R. *et al.* The role of APOBEC3B in lung tumor evolution and targeted cancer
953 therapy resistance. *Nat Genet* **56**, 60–73 (2024).

954

955 **Supplemental information**

956 **Supplementary Materials** including Supplementary Notes, Figs., and Tables, which the
957 Supplementary Tables were a combined excel file containing supplementary tables.

958 **Additional file 1.** CHIP whitelist mutations identified and confirmed in the discovery cohort.

959 **Additional File 2.** CHIP whitelist mutations identified and confirmed in the replication cohort.

960 **Additional File 3.** CHIP whitelist mutations identified and confirmed in the validation cohort.

961 **Additional File 4.** *KMT2D* somatic missense mutations identified and confirmed in the discovery
962 and replication cohort.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplTablesCHIP20260304.xlsx](#)
- [SupplCHIP20260305.pdf](#)
- [AdditionalFile1POG13KIGVreducedSize.pdf](#)
- [AdditionalFile1POG13KIGV20260304.pdf](#)
- [AdditionalFile2CYN16KIGV20260304.pdf](#)
- [AdditionalFile3KGPIGV20260304.pdf](#)
- [AdditionalFile4KMT2DmissenseIGV20260304.pdf](#)