

1 **Fine-scale Population Structure and Demographic History of Han Chinese** 2 **Inferred from Haplotype Network of 111,000 Genomes**

3 Ao Lan^{1,†}, Kang Kang^{2,1,†}, Senwei Tang^{1,2,†}, Xiaoli Wu^{1,†}, Lizhong Wang¹, Teng Li¹, Haoyi
4 Weng^{2,1}, Junjie Deng¹, WeGene Research Team^{1,2}, Qiang Zheng^{1,2}, Xiaotian Yao^{1,*} & Gang
5 Chen^{1,2,3,*}

6 ¹ WeGene, Shenzhen Zaozhidao Technology Co., Ltd., Shenzhen 518042, China

7 ² Shenzhen WeGene Clinical Laboratory, Shenzhen 518118, China

8 ³ Hunan Provincial Key Lab on Bioinformatics, School of Computer Science and
9 Engineering, Central South University, Changsha 410083, China

10 † These authors contributed equally to this work.

11 * Correspondence: Xiaotian Yao: yaoxt@wegene.com & Dr. Gang Chen: cg@wegene.com

12

13 **ABSTRACT**

14 Han Chinese is the most populated ethnic group across the globe with a comprehensive
15 substructure that resembles its cultural diversification. Studies have constructed the genetic
16 polymorphism spectrum of Han Chinese, whereas high-resolution investigations are still
17 missing to unveil its fine-scale substructure and trace the genetic imprints for its demographic
18 history. Here we construct a haplotype network consisted of 111,000 genome-wide
19 genotyped Han Chinese individuals from direct-to-consumer genetic testing and over 1.3
20 billion identity-by-descent (IBD) links. We observed a clear separation of the northern and
21 southern Han Chinese and captured 5 subclusters and 17 sub-subclusters in haplotype
22 network hierarchical clustering, corresponding to geography (especially mountain ranges),
23 immigration waves, and clans with cultural-linguistic segregation. We inferred differentiated
24 split histories and founder effects for population clans Cantonese, Hakka, and Minnan-
25 Chaoshanese in southern China, and also unveiled more recent demographic events within
26 the past few centuries, such as *Zou Xikou* and *Chuang Guandong*. The composition shifts of
27 the native and current residents of four major metropolitans (Beijing, Shanghai, Guangzhou,
28 and Shenzhen) imply a rapidly vanished genetic barrier between subpopulations. Our study

29 yields a fine-scale population structure of Han Chinese and provides profound insights into
30 the nation's genetic and cultural-linguistic multiformity.

31

32 **INTRODUCTION**

33 Population genomics has provided magnificent insights into the evolutionary pathway and the
34 genetic composition of human beings. The prior large-scale studies, such as the 1000
35 Genomes Project (1KGP) (1000 Genomes Project Consortium et al., 2015), have
36 predominantly centered on the variation spectrum in human genomes, which empowered the
37 recognition of the genetic divergence of various populations across the globe. Comparing
38 with the variation-scale profile, the haplotype sharing network within a population may
39 administer a finer resolution for discriminating the substructures elicited by recent
40 demographic events such as migration, admixture, segregation, and natural selection
41 (Palamara et al., 2012; Powell et al., 2010; Speed and Balding, 2015). As two pilot studies,
42 the geographical subpopulation structures of the British and Finnish populations have been
43 well-demonstrated (Leslie et al., 2015; Martin et al., 2018). AncestryDNA, a direct-to-
44 consumer genetic testing (DTC-GT) service provider, also published the fine-scale
45 population structure in North America from their *in-house* biobank (Han et al., 2017).

46 As one of the most ancient nations, China is populated with the world's largest ethnic
47 group, Han Chinese. It is of great concern to conduct comprehensive genomics research to
48 testify the nation's historical records and legends, mine undocumented demographic events,
49 and map its cultural diversification with the genetic imprints. Former microarray-based
50 studies have identified an evident north-south genetic differentiation of Han Chinese (Chen et
51 al., 2009; Xu et al., 2009). The low-coverage sequencing of over 11,000 Han Chinese
52 uncovered a population structure along the east-west axis (Chiang et al., 2018). The deep
53 sequencing of over 10,000 Chinese has provided extensive genetic markers of high quality
54 (Cao et al., 2020). However, the limited sample volume of these studies remains insufficient
55 for a highly modularized nationwide haplotype network, and the hospital-based cohort may
56 also skew toward region-specific subpopulations. The largest published population study of
57 the Chinese people has utilized the ultra-low depth sequencing data from the non-invasive
58 prenatal testing to establish the nation-wide SNP spectrum (Liu et al., 2018), but lacks the
59 resolution on an individual scale. Nevertheless, these datasets cannot simultaneously afford

60 sufficient sample size, dense genetic markers to assemble shared haplotypes, and a well-
61 proportioned participant distribution across the country to unscrew the subpopulation
62 structure. A whole-genome genotyping dataset from a country-wide DTC-GT service
63 provider is still an ideal solution to balance the cost of effect of haplotype network
64 construction on a national scale.

65 In the present work, we create the haplotype network from the identity-by-descent
66 (IBD) segments shared by 110,955 consented DTC-GT users from WeGene, China. We
67 identify and annotate the subpopulation partitions using a hierarchical clustering approach
68 and map the genetic separations with linguistic and cultural differentiation or historic
69 demographic events.

70

71 **RESULTS**

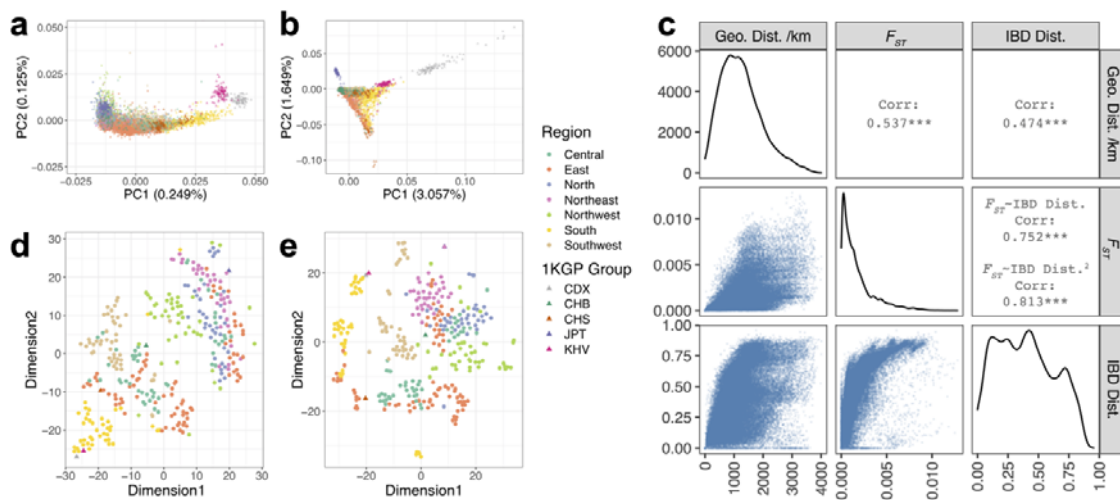
72 **Study Participants and the IBD Network Features**

73 The 110,955 consented participants with self-reported ethnicity, birthplace (in prefecture-
74 level), and current residence were recruited from the WeGene Biobank (**Figure S1**). All
75 participants were genotyped with one of two custom arrays: Affymetrix WeGene V1 Array or
76 Illumina WeGene V2 Array. After quality control, we utilized 350,140 autosomal single
77 nucleotide polymorphisms (SNPs) to identify IBD segments (**Figure S2**). We then yielded a
78 haplotype network composed of 102,822 vertices and 1.3 billion edges (total IBDs with a
79 minimal length of 2 centiMorgan between a pair of individuals).

80 The principal component analysis (PCA) of the SNP profiles of the Han Chinese
81 individuals resembles previous population studies (Cao et al., 2020; Liu et al., 2018), with
82 similar proportions of variance explained by the first two PCs (0.25% and 0.13%) (**Figure**
83 **1a**). The PCA analysis of the IBD profiles exhibits a better separation among individuals
84 from different geographical regions (**Figure 1b**). Also, higher proportions of variance were
85 explained by the first two PCs of IBD (3.06% and 1.65%). IBD sharing indices were
86 calculated between pairs of prefectures. The IBD-based genetic distance (IBD distance,
87 calculated as $1 - \text{IBD sharing index}$), SNP-based genetic distance (fixation index, F_{ST}), and
88 the geographical distance between cities highly correlate with each other (Pearson's

89 correlation, $p < 0.01$) (**Figure 1c**). As the F_{ST} distribution was heavily right-skewed and the
 90 IBD distance emerges while F_{ST} remains low (Pearson's correlation between squared IBD
 91 distance and F_{ST} : 0.81), the IBD dissimilarity has the potential to achieve a higher resolution
 92 among the communities with similar genetic backgrounds. Both SNP- and IBD-based genetic
 93 dissimilarity projections (**Figure 1d-e**) are associated with the cities' spatial distribution
 94 (**Figure S3**), while the IBD analysis has presented better modularity for the prefectures from
 95 the same region (**Figure 1e**): for instance, the southern prefectures (yellow nodes) are
 96 immensely placed in specified modules in the IBD distance projection (ANOSIM test among
 97 the three southern provinces, $R = 0.55$, $p = 0.001$), while such partitions were less perceptible
 98 in the F_{ST} projection (**Figure 1d**), though also statistically significant (ANOSIM test, $R =$
 99 0.28 , $p = 0.001$). These city modules may preferably pronounce the genetic segregation
 100 among distinguished clans (Canton, Hakka, Min-Chaoshan, and Guangxi). Greater
 101 differentiation was also captured by the IBD distance between northern and northeastern
 102 China (ANOSIM test, $R = 0.29$ for IBD distance and 0.12 for F_{ST} , $p = 0.001$).

103



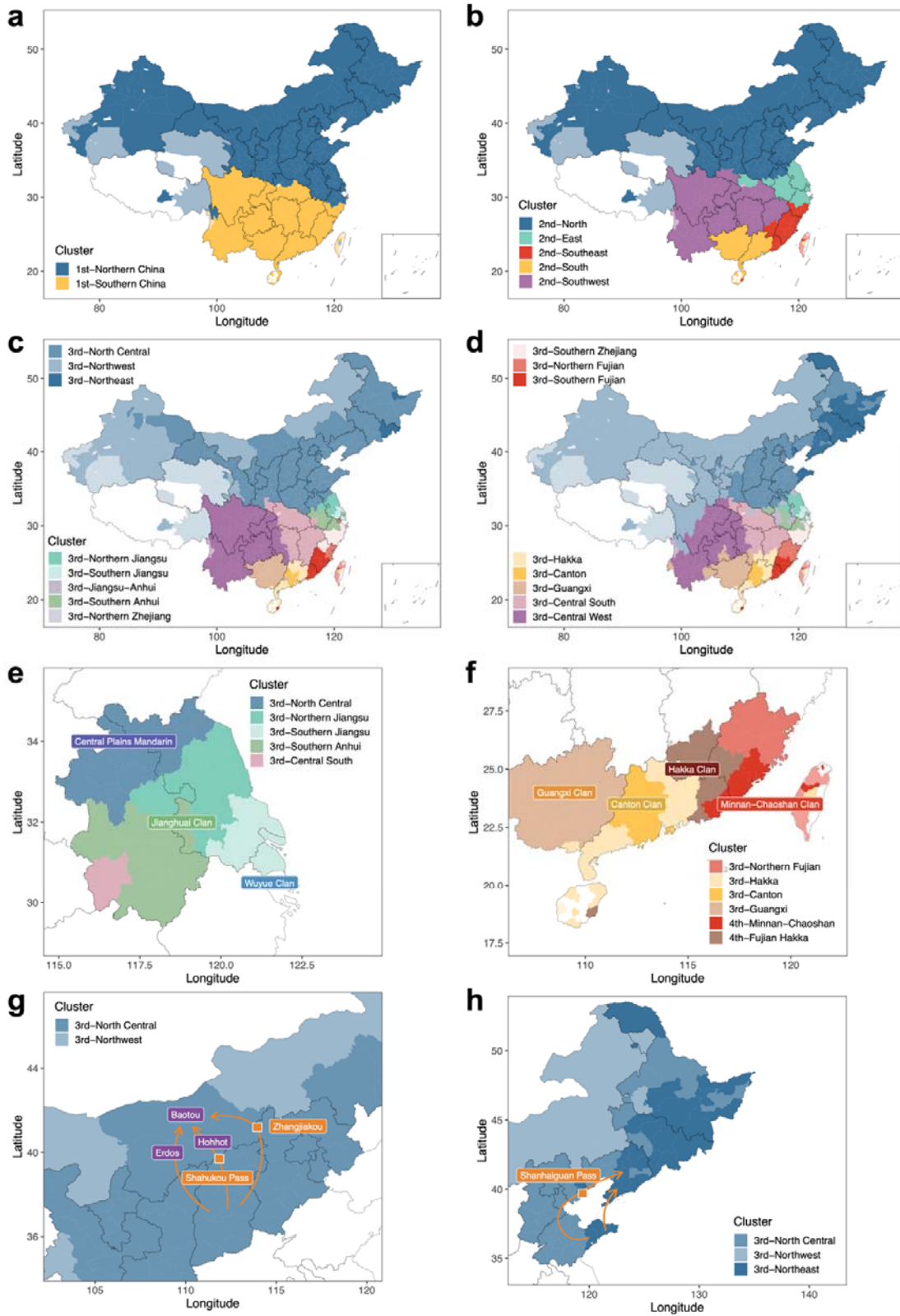
104

105 **Figure 1. The genetic dissimilarities among individuals and among different cities in**
 106 **China. a.** The PCA analysis of the SNP profiles of 5,000 randomly subsampled Han Chinese
 107 and 502 East Asian (EAS) samples from the 1000 Genomes Project (1KGP). **b.** The PCA
 108 analysis of the IBD profiles of the 5,502 individuals used in panel (a). **c.** The correlation between
 109 the inter-prefecture F_{ST} , IBD-based genetic distance, and geographic distance. **d.** The t-
 110 distributed Stochastic Neighbor Embedding (t-SNE) projection of the SNP-based genetic
 111 distances (F_{ST}) between prefecture pairs. **e.** The t-SNE projection of the IBD sharing indices
 112 between prefecture pairs. Panels a, b, d, and e share the same legend.

113

114 **Population Structure and Demographic Events**

115 Hierarchical clustering was applied to the haplotype network to obtain a fine-scale population
116 substructure recursively. The haplotype network clustering yielded two major clusters
117 harboring 61.8% and 36.6% of the vertices in the entire network, successfully divided the
118 population into the northern (1st-Northern China) and southern Chinese (2nd-Southern China).
119 The most abundant cluster in each prefecture was colored distinctly in **Figure 2a**. The second
120 stage clustering divided the southern population into three subclusters: 2nd-Southeast, 2nd-
121 South, and 2nd-Southwest, and separated the Yangtze River Delta region (2nd-East) from the
122 other northern population (2nd-North) (**Figure 2b**). In the third stage, more detailed partitions
123 could be identified (**Figures 2c-f, S4, and S5**), where the imprints from ethnic fusion, recent
124 movement, and linguistic-cultural division were able to be detected.



125

126

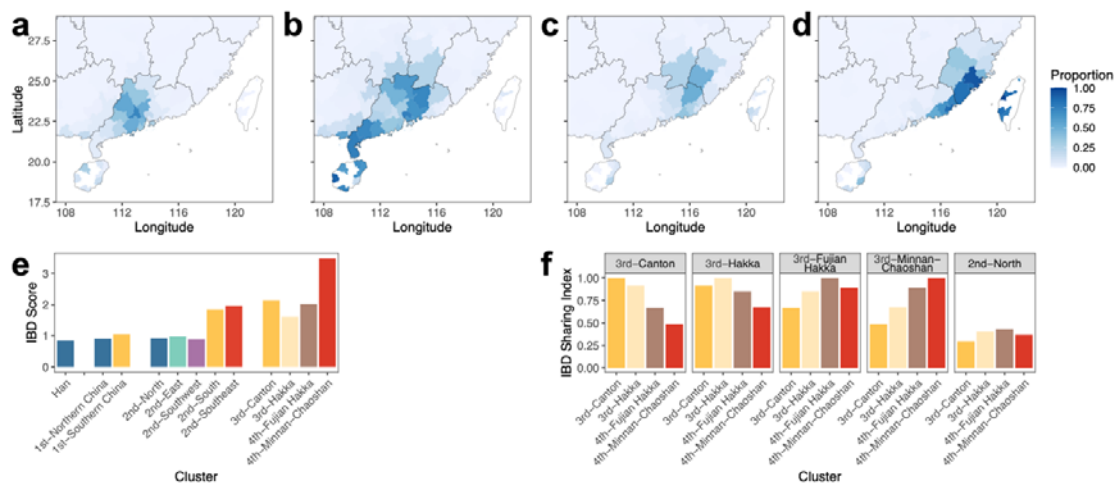
127 **Figure 2. The hierarchical clustering of the haplotype network. a-c.** The most populated 1st-
128 level **(a)**, 2nd-level **(b)**, and 3rd-level **(c)** cluster in each prefecture. **d.** The 3rd-level cluster with
129 the largest odds ratio in each prefecture. **e.** The spatial distribution of the 3rd-level subclusters in
130 Jiangsu province is accompanying the linguistic-cultural division. The most populated clusters
131 were shown. **f.** Three major clans in Guangdong, Canton, Hakka (falling into two clusters), and
132 Min-Chaoshan, can be distinguished from the haplotype network clusters. The clusters with the
133 largest odds ratio were shown. **g-h.** The paths of the *Zou Xikou* **(g)** and *Chuang Guandong* **(h)**
134 migration waves. In panels **a-d**, 50% transparency was applied to the prefectures with small
135 sample sizes ($n < 10$), and prefectures with no valid samples were left blank.

136 The subpopulation partitioning may attribute to an interplay of multiple factors
137 including geography, politics, cultural, ethnic fusion, and natural selection. The southern
138 boundary of the 2nd-North cluster is generally consistent with the Qinling-Huaihe Line
139 **(Figure 2b)**, the geographical dividing line for northern and southern China, as the two parts
140 differ from each other in climate, staple crop and culture. Such separation was clearly
141 pronounced by the haplotype cluster distribution in Jiangsu and Anhui provinces that locates
142 in the Huai River basin, where the Wuyue clan, Jianghuai clan, and the central plain
143 mandarin speaking regions could be distinguished **(Figure 2e)**. In Guangdong province, the
144 spatial division of the three major clans (Canton, Hakka, and Minnan-Chaoshan) could also
145 be linked with distinct haplotype subclusters **(Figure 3f)**. In the north, the pattern of the 3rd-
146 Northwest cluster is substantially following the geographic placement of the Mongolic and
147 Altaic ethnic minorities. The outlier in of the Hetao Plain in the central of Inner Mongolia,
148 where the leading cluster assembles the Central Plains, may imply the historic migration
149 wave *Zou Xikou* (go beyond the western pass) during the Qing dynasty **(Figure 2g)**. Similarly,
150 the Shandong Peninsula and most northeastern cities partook a common subcluster by the
151 largest odds ratio, 3rd-Northeast **(Figure 2h)**, which also implies the *Chuang Guandong* (rush
152 beyond the Shanhaiguan Pass) immigration wave. In the PCA analysis for SNP of the
153 individuals from the 3rd-North Central and 3rd-Northeast, no detachment could be discerned
154 **(Figure S6)**.

155 More subclusters could be classified in the south of the Qinling-Huaihe line (3rd-level
156 subclusters, north: 3, south: 14). Guangdong and Fujian residents have formed various clans
157 with specified languages, cultures, and habitations, and the differentiation is also portrayed
158 by separate haplotype subclusters in this study **(Figure 3a-d)**. Much higher IBD scores were
159 observed in the 2nd-South (1.40) and 2nd-Southeast (1.57) populations **(Figure 3e)**,
160 particularly for the 4th-Minnan-Chaoshan subcluster (3.11), compared with the other clusters

161 (for 2nd-North: 0.57, 2nd-East: 0.62, and 2nd-Southwest: 0.55, respectively). High IBD scores
 162 imply strong founder effects for these Han subpopulations, in line with the historic records
 163 for their southward migrations. In the meantime, the IBD sharing index between 3rd-Canton
 164 and 4th-Minnan-Chaoshan (0.41) was lower than the median IBD sharing index between two
 165 random clusters (0.61, one-sample Wilcoxon signed-rank test, $p < 2 \times 10^{-16}$) (**Figure 3f**),
 166 suggesting a high genetic disparity between these clans, though residing in adjacent regions
 167 for over a thousand year. The two Hakka subclusters exhibit the highest IBD sharing with the
 168 2nd-North cluster (0.40 and 0.43), while 3rd-Canton shared the least (0.29).

169



170

171 **Figure 3. The distribution of the subclusters of the major clans in Guangdong and their**
 172 **population dynamics. a-d.** Each subcluster's population fraction in Guangdong province and
 173 adjacent regions. **e.** The IBD score of each subcluster. **f.** The IBD sharing indices between
 174 subclusters.

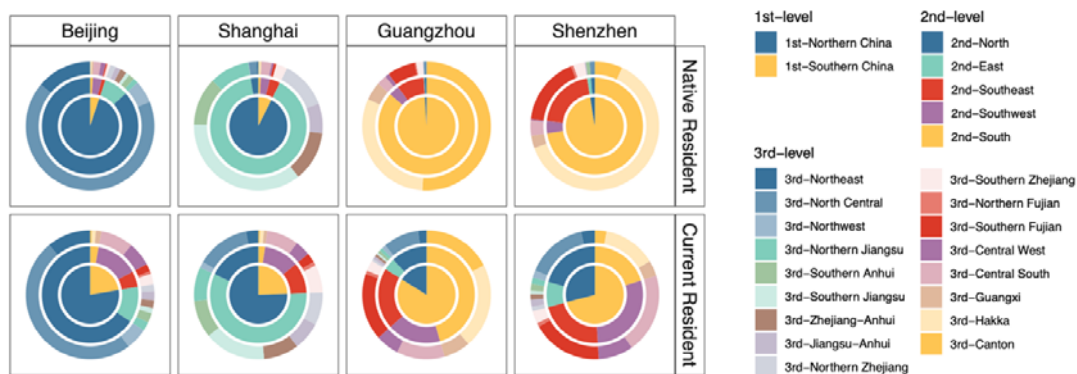
175

176 Modern Population Flows

177 In the contemporary era, economics is also shaping the new population substructures at an
 178 exceptionally rapid pace. We analyzed the modern population flows by comparing the
 179 participants' birthplaces and current residences. In the four major metropolises in China,
 180 most of the native residents (classified by participants' birthplace) belong to the local cluster
 181 and subclusters (**Figure 4**): for instance, 86.8% of the Beijing native residents belong to the
 182 2nd-North cluster; 51.5% of the Guangzhou native residents were members of 3rd-Canton and

183 30.8% were from 3rd-Hakka. However, the compositions of all these cities soon become an
184 admixture of immigrants cross the country (**Figure 4**): only 17.5% and 21.3% of the current
185 Guangzhou residents remain members of 3rd-Canton and 3rd-Hakka; the youngest one,
186 Shenzhen, whose *de jure* population emerged from 0.3 million to over 20 million in the past
187 40 years, the fraction of the former dominant subcluster 3rd-Hakka reduced from 62.4% to
188 only 13.3%. Accordingly, the 3rd-level cluster alpha-diversity (Shannon index) of these four
189 cities increased from 1.47 ± 0.34 to 2.25 ± 0.25 (one-tailed, paired sample *t*-test, $p = 0.003$).

190



191

192 **Figure 4. The resident composition by IBD network clusters in four major metropolises,**
193 **in comparison to the native residents (according to the birthplace) and the current**
194 **residents.** The inner to outer circles represent the compositions of the 1st, 2nd, and 3rd-level
195 clusters respectively.

196

197 DISCUSSION

198 As a biobank-scale population study of Chinese, we revealed the fine-scale subpopulation
199 structure of Han Chinese by constructing a haplotype network of 110,000 genomes. The
200 haplotype network shows a marked dependency between genetic distance and geography, but
201 also process a step further to disclose the population substructures derived from recent
202 demographic events or cultural and linguistic separation.

203 Previous large-scale studies on the Chinese population did not reach a fine-scale
204 resolution for the population substructure, due to the limitation of sample size or genetic
205 marker quantity and quality (Cao et al., 2020; Liu et al., 2018; Xu et al., 2009). The current

206 biobanks in China also lack essential volume or nationwide representativeness of participants:
207 only 6.3% of the 510,000 China Kadoorie Biobank participants were genome-wide
208 genotyped (Chen et al., 2011), while the Taiwan Biobank project only sampled Han Chinese
209 residing in Taiwan (Chen et al., 2016; Fan et al., 2008). Hence, the whole-genome
210 genotyping dataset from DTC-GT services becomes a preferred solution to reveal the
211 subpopulation structure that balanced the issues of participant distribution, sample size,
212 genotyping cost, and marker density.

213 Unlike the North American haplotype network constructed from close relatives to
214 reveal the post-Columbus population expansion (Han et al., 2017), we employed full-
215 spectrum IBD pairs to trace the demographic events over a longer timescale. This enables
216 founder effect estimation and cross-community dissimilarity analysis, which successfully
217 revealed the genetic disparity among the clans in south China.

218 **Discrepant Application Scenarios between SNPs and IBDs**

219 In the present study, the haplotype network and the SNP spectrum have provided related but
220 independent information. In some scenarios, the SNP-based analysis lacks the essential
221 resolution to subdivide population substructures with similar genetic makeup: for instance,
222 the 3rd-Northeast clustering harboring the *Chuang Guandong* offsprings could not be
223 distinguished from the other northern Chinese by SNPs.

224 **Geographical Impacts: Mountain Range > Climate > River**

225 Mountain ranges have predominantly shaped the partition of the population substructure.
226 Different subclusters with a considerable genetic distance reside on both sides of the major
227 mountain ranges, such as the Qinling Mountains, Five Ridges, Wuyi Mountains, and Xuefeng
228 Mountains. Different climate zones, the temperate zone, and the subtropical zone, also harbor
229 different subpopulations, as revealed by the population composition of Anhui and Jiangsu
230 provinces. There is no significant geographical isolation in this region, while different clans
231 with disparate languages or dialects have formed, which also correlates with the rice farming
232 and wheat (or millet before the Bronze Age) farming regions: wheat was cultivated in the
233 Central Plains Mandarin speaking region, Wuyue relies on rice, while Jianghuai formed a
234 cline. On the contrary, the isolation effect of great rivers, for instance, the Yangtze River, was

235 not observed: the two sides of the Yangtze river always resemble each other in the
236 subpopulation compositions, no matter in its upper-middle reaches, or in the delta region.

237 **War, Migration, and Politics: Keys to Population Split and National Fusion**

238 War is a critical factor for ancient immigration, population split, and fusion. The southern
239 Han Chinese clans are purported to be offerings of diverse southward movements from Qin
240 to Song dynasties (Meacham, 1999; Wen et al., 2004). Cantonese was purported to be
241 originated between Qin (221 to 206 BC) to Tang (618 to 907 AD) dynasties; Minnan-
242 Chaoshan formed between Jin (266 to 420 AD) and Tang dynasties; the Hakka clan was
243 composed of various southward movements between Tang and Qing (1612 to 1912 AD)
244 dynasties, with a relatively short history and manifold origins. These histories were supported
245 by the IBD network analysis, where Hakka has the lowest IBD score, but the highest IBD
246 sharing index with northern clusters, suggesting a relatively late split with the Central Plains
247 population. Cantonese and Minnan-Chaoshanese, though reside in adjacent regions, exhibited
248 notable disparity, supporting the different origins. The 3rd-Canton cluster's low IBD sharing
249 index with the northern communities may also suggest its oldest split time, which is in line
250 with historic records.

251 The haplotype network also successfully unveiled more recent demographic events
252 driven by politics. *Zou Xikou* and *Chuang Guandong* were the largest recent migration waves
253 of Han Chinese majorly happening within the past centuries, driven by politics. The
254 population increase in the Central Plains imposed much pressure on the authorities. As a
255 consequence, the Qing regime released the immigration ban for the Han people to reside
256 beyond the Great Wall, the former reserved land of the ruling ethnic groups, Man and
257 Mongol. As a result of the demic diffusion of Han Chinese, most of the northeastern Han
258 people are offsprings of the *Chuang Guandong* wave. In our study, the genetic relationship
259 between the Shandong Peninsula, the major origin of *Chuang Guandong*, and the
260 northeastern Chinese was disclosed. As the most populated cluster (3rd-North Central) differs
261 from the cluster with the largest OR (3rd-Northeast) in northeast China, the two clusters may
262 imply the offsprings of migrants from different migration waves or choosing different routes:
263 the inland residents using the land route via the Shanhaiguan Pass, or the coastal migrants
264 using the sea route and landed on the Liaodong Peninsula (**Figure 2h**). Similarly, the *Zou*
265 *Xikou* migrants from Shanxi province settled down to the traditional Mongolic regions
266 including Baotou and Hohhot and became the largest local population now (**Figure 2g**).

267 **The Rapidly Vanished Population Boundaries**

268 Though the Chinese populations have comprehensive substructures involving its long history
269 and cultural pluralism, the genetic divergence between subpopulations may vanish over the
270 coming decades, which may resemble the national fusion process that happened in Hispanic
271 Latin America. Our analysis of the shifts of the metropolitans' residents has confirmed the
272 irreversible trend. Admittedly, the user distribution of a DTC-GT service could heavily skew
273 toward youngsters and the current residents of the most developed regions and cities (**Figure**
274 **S1**), particularly new economic migrants, which may result in an overestimation of the
275 level of population mixing. The rapidly growing economy, coped with the emerging
276 transportation capacity, has been speedily eliminating the genetic barriers between
277 subpopulations. As the admixture increases, it might become more difficult to trace the
278 demographic histories of a nation from either SNPs or IBDs. In this golden time for human
279 population genomics, biobanking and biobank-scale studies are essential to mining the
280 memories coded in our DNA.

281

282 **METHODS**

283 **Study Design**

284 **Participants.** All participants involved in this study were drawn from consenting WeGene
285 customers. Participants with self-reported ethnicity, prefecture-level birthplace, and current
286 residence were included (n = 110,955), and the demographic data were collected in April
287 2020. The East Asian samples (EAS) from the 1000 Genomes Project (1KGP) (n = 504) were
288 integrated into the database. Duplicated genetic profiles from the same individual (n = 144)
289 and profiles with relatedness up to the second-degree kinship (n = 8,493) were identified with
290 *King* V2.2.1 (Manichaikul et al., 2010) with default parameters and excluded from analyses.
291 Finally, 102,822 genetic profiles were acquired for analyses.

292 **Ethnic approval and compliance.** Informed consent for online research was obtained from
293 all individual participants included in the study. The study was approved by the Ethical
294 Committee of Shenzhen WeGene Clinical Laboratory. The study was conducted following

295 the human and ethical research principles of The Ministry of Science and Technology of the
296 People's Republic of China (Regulation of the Administration of Human Genetic Resources,
297 July 1, 2019).

298 **DNA sampling and genotyping assay.** Saliva samples for DNA extraction were collected
299 processed following the previously published protocol (Kang et. al, in press). Samples were
300 genotyped on one of two custom arrays: Affymetrix WeGene V1 Array (596,744 SNPs) by
301 Affymetrix GeneTitan MC Instrument, and Illumina WeGene V2 Array (742,762 SNPs) by
302 Illumina iScan System. A minimal genotyping call of 98.5% was required for a valid sample.

303 **Data Processing**

304 **Genetic marker quality control.** Indels, heterosomal loci, and loci with more than two
305 allelic states were removed from the genotyping data. For both arrays, SNP markers were
306 filtered with *Plink* V1.9 (Purcell et al., 2007) with parameters "*--maf 0.001 --geno 0.05*"
307 respectively. Only the intersection of the two arrays with identical allelic states was retained.
308 To minimize the impact of the batch effect between the two arrays, for each biallelic SNP, a
309 Chi-square test was performed among the three genotypes, and the *p*-values were Bonferroni
310 corrected. SNPs with significant batch effect (*false discovery rate (FDR) < 0.01*) were
311 eliminated. PCA analyses for the SNP sets before and after batch effect removal were
312 illustrated in **Figure S7**. The density of the SNP markers used for IBD detection was shown
313 in **Figure S8**.

314 **1KGP sample integration.** The genotypes of the selected genetic markers of the 504 EAS
315 samples were extracted with *VCFTools* V0.1.15 (Danecek et al., 2011). The genotypes of
316 SNPs with inconsistent allelic states with the WeGene samples were set to a missing value.
317 Then the genetic profiles of the 504 EAS samples were concatenated with the WeGene
318 samples.

319 **Genotype phasing.** For the WeGene samples and 1KGP samples, we employed Eagle V2.3.5
320 (Loh et al., 2016) for a reference panel-free genotype phasing, using the default parameters.

321 **IBD detection and merging.** To minimize false-positive haplotype sharing, we identified the
322 IBD segments (with a minimal length of 1 cM) with *Refined-IBD* (Browning and Browning,
323 2013) with default parameters. We then merged adjacent IBD segments with a gap less than

324 0.6 cM and no more than one genotype discordance in the gap region as one consecutive IBD
325 segment, using the *merge-ibd-segments* function. In sum, 4,585 million IBD segments were
326 yielded.

327 **IBD segment quality control.** We exclude the IBD segments with overlaps with any of the
328 following regions annotated by the UCSC hg19 reference genome (<http://genome.ucsc.edu/>):
329 centromeres, telomeres, acrocentric short chromosomal arms, heterochromatic regions,
330 clones, and contigs identified in the "gaps" table.

331 For each SNP marker, the amount of IBD segments harboring it was summarized as the IBD
332 coverage. 25% and 75% quantiles (Q1 and Q3) and the interquartile range (IQR) were
333 calculated. The regions with an IBD coverage $\geq 75\% Q3 + 1.5 \times IQR$ were marked as IBD
334 hotspots (**Figure S9 and Table S2**). IBD segments fell in or overlapped with such IBD
335 hotspots were discarded.

336 **Hierarchical clustering.** The haplotype network was constructed with edges representing
337 and weighted by the total shared IBD length (≥ 2 cM) between each pair of individuals. For
338 the detection of population substructures recursively, we retained the edges corresponding to
339 a total IBD ≥ 3 cM and applied the Louvain method for the hierarchical clustering (Blondel et
340 al., 2008). The R package *igraph* was employed to apply. The clustering was performed for
341 five levels. If a cluster or subcluster contained less than 50 nodes or was composed with $< 1\%$
342 nodes of its parent cluster, or was the only subcluster of its parent cluster, its next-level
343 clustering stopped. In the 3rd to 5th levels, a cluster might be subdivided into fragmented and
344 meaningless subclusters. To avoid this, we summarized the node counts in a subcluster \times
345 prefecture matrix, and pairwise calculated the Spearman's correlation between subclusters.
346 The subclusters with pairwise correlation coefficients ≥ 0.8 were merged as one subcluster
347 and would not be subdivided during the next-level clustering.

348 In each prefecture, the proportions and odds ratios (OR) of each cluster were calculated. The
349 dominating clusters were named according to the cluster's geographical distribution. The
350 statistics of major clusters were summarized in **Table S1**. The geographical distributions of
351 the clusters were shown in **Figures S5 and S6**.

352 **Statistics**

353 **IBD score, IBD sharing index, and genetic distances.** IBD score was introduced to
354 represent the mean total IBD length among all individual pairs within a community,
355 following the previously published method (Consortium, 2019). IBD scores were calculated
356 for prefectures, clusters, ethnic groups, and community subsets. For community i with a size
357 of n_i , k and l are an individual pair belonging to community i , the IBD score of community i
358 was calculated as:

$$359 \quad IBD\ score_i = \frac{\sum_{k,l}^{n_i} (total\ IBD\ length_{k,l})}{n_i(n_i-1)/2} \quad Eq. 1$$

360 IBD sharing index was introduced to represent the mean total IBD length among all
361 individual pairs from two communities and normalized by the IBD scores of the two
362 communities to eliminate founder effects in different degrees. For community i and j with
363 sizes of n_i and n_j , respectively, k is a member of community i and l is a member of
364 community j , the IBD sharing index between i and j was calculated as:

$$365 \quad IBD\ sharing\ index_{i,j} = \frac{\sum_k^{n_i} \sum_l^{n_j} (total\ IBD\ length_{k,l})}{n_i \times n_j \times \sqrt{IBD\ score_i \times IBD\ score_j}} \quad Eq. 2$$

366 IBD distance between two communities was calculated as $1 - IBD$ sharing index. The
367 IBD distances < 0 were rescaled to 0.

368 **Data projection.** Principal component analysis (PCA) was applied to the SNP profiles and
369 the IBD profiles of 5,000 randomly subsampled Han Chinese individuals and the 502 EAS
370 samples. For all quality-controlled SNPs, the redundant markers sharing the same linkage
371 disequilibrium (LD) block were removed from the PCA analysis with *Plink* V1.9 (Purcell et
372 al., 2007) with parameters "--indep-pairwise 50 5 0.5". Finally, 138,725 SNP markers were
373 retained for the PCA analysis for SNPs. For the IBD profiles, the IBD sharing matrix among
374 the 5,502 individuals was used as the input. PCA analysis was performed with *GCTA* V1.9
375 (Yang et al., 2011) with the function *GCTA-PCA*.

376 The SNP-based inter-city genetic distance was calculated as the fixation index (F_{ST})
377 using *VCFtools* V0.1.15 (Danecek et al., 2011). The SNPs used for F_{ST} calculation were the
378 same SNP set for IBD detection. T-distributed stochastic neighbor embedding (t-SNE) was
379 used for the genetic distances among cities.

380 **Basic statistics and visualization.** Data process, statistics, and visualization were performed
381 using *R* and *R* packages including *igraph*, *vegan* (Oksanen et al., 2007), *reshape2* (Wickham,
382 2012), *tidyverse* (Wickham et al., 2019), *RCy3* (Gustavsen et al., 2019), *ggplot2* (Wickham,
383 2016), *ggally* (Schloerke et al., 2011), *ggtree* (Yu et al., 2017), *heatmap* (Kolde and Kolde,
384 2015), *patchwork* (Pedersen, 2017), and *ggnewscale* (Campitelli, 2019).

385 **Data availability.** In light of our commitment to customer privacy and regulations from the
386 Administration of Human Genetic Resource of China, we will not be publishing the raw data
387 from WeGene customers. For the purpose of reproducing the analyses, we can share the
388 haplotype network topology on request after a compliance review. For questions about the
389 analyses in this research, please contact the WeGene Research Team by email
390 (research@wegene.com).

391

392 **Acknowledgments**

393 We thank all WeGene users who consented to share their genotype and demographic
394 information for research purposes. We thank Prof. Dr. Chuan-Chao Wang from Xiamen
395 University for his valuable suggestions and comments on this study. We also thank the
396 employees of WeGene Inc. who contributed to the development of the infrastructure that
397 made this research possible.

398

399 **Conflict of Interest**

400 The authors AL, KK, ST, XW, LW, TL, HW, JD, QZ, XY, and GC work for WeGene
401 (Shenzhen Zaozhidao Technology Co. Ltd. or Shenzhen WeGene Clinical Laboratory).

402

403 **SUPPLEMENTARY INFORMATION**

404 This document includes 9 supplementary figures and 2 supplementary tables.

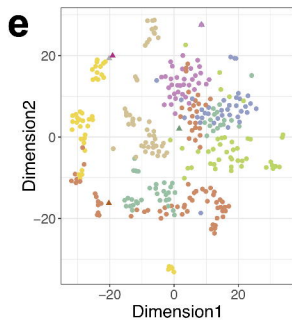
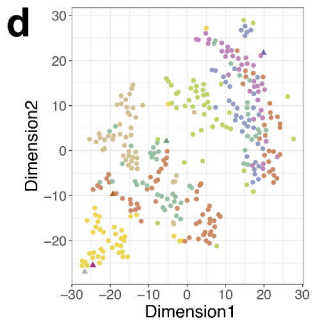
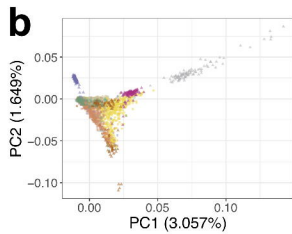
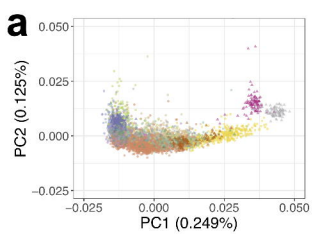
405

406 REFERENCES

407

- 408 1000 Genomes Project Consortium, Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P.,
409 Kang, H.M., Korbelt, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., *et al.* (2015). A global
410 reference for human genetic variation. *Nature* 526, 68-74.
- 411 Blondel, V.D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of
412 communities in large networks. *Journal of statistical mechanics: theory and experiment*
413 2008, P10008.
- 414 Browning, B.L., and Browning, S.R. (2013). Improving the accuracy and efficiency of identity-
415 by-descent detection in population data. *Genetics* 194, 459-471.
- 416 Campitelli, E. (2019). ggnewscale: Multiple Fill and Color Scales in 'ggplot2 '. R package
417 version 02 0 URL: [https://CRAN](https://CRAN.R-project.org/package=ggnewscale)
418 Cao, Y., Li, L., Xu, M., Feng, Z., Sun, X., Lu, J., Xu, Y., Du, P., Wang, T., Hu, R., *et al.* (2020). The
419 ChinaMAP analytics of deep whole genome sequences in 10,588 individuals. *Cell Res.*
- 420 Chen, C.H., Yang, J.H., Chiang, C.W.K., Hsiung, C.N., Wu, P.E., Chang, L.C., Chu, H.W., Chang,
421 J., Song, I.W., Yang, S.L., *et al.* (2016). Population structure of Han Chinese in the modern
422 Taiwanese population based on 10,000 participants in the Taiwan Biobank project. *Hum Mol*
423 *Genet* 25, 5321-5331.
- 424 Chen, J., Zheng, H., Bei, J.X., Sun, L., Jia, W.H., Li, T., Zhang, F., Seielstad, M., Zeng, Y.X.,
425 Zhang, X., *et al.* (2009). Genetic structure of the Han Chinese population revealed by
426 genome-wide SNP variation. *Am J Hum Genet* 85, 775-785.
- 427 Chen, Z., Chen, J., Collins, R., Guo, Y., Peto, R., Wu, F., Li, L., and China Kadoorie Biobank
428 collaborative, g. (2011). China Kadoorie Biobank of 0.5 million people: survey methods,
429 baseline characteristics and long-term follow-up. *Int J Epidemiol* 40, 1652-1666.
- 430 Chiang, C.W.K., Mangul, S., Robles, C., and Sankararaman, S. (2018). A Comprehensive Map
431 of Genetic Variation in the World's Largest Ethnic Group-Han Chinese. *Mol Biol Evol* 35,
432 2736-2750.
- 433 Consortium, G.K. (2019). The GenomeAsia 100K Project enables genetic discoveries across
434 Asia. *Nature* 576, 106.
- 435 Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E.,
436 Lunter, G., Marth, G.T., and Sherry, S.T. (2011). The variant call format and VCFtools.
437 *Bioinformatics* 27, 2156-2158.
- 438 Fan, C.T., Lin, J.C., and Lee, C.H. (2008). Taiwan Biobank: a project aiming to aid Taiwan's
439 transition into a biomedical island. *Pharmacogenomics* 9, 235-246.
- 440 Gustavsen, J.A., Pai, S., Isserlin, R., Demchak, B., and Pico, A.R. (2019). RCy3: network
441 biology using cytoscape from within R. *F1000Research* 8.
- 442 Han, E., Carbonetto, P., Curtis, R.E., Wang, Y., Granka, J.M., Byrnes, J., Noto, K., Kermany,
443 A.R., Myres, N.M., Barber, M.J., *et al.* (2017). Clustering of 770,000 genomes reveals post-
444 colonial population structure of North America. *Nat Commun* 8, 14238.
- 445 Kolde, R., and Kolde, M.R. (2015). Package 'pheatmap'. *R Package* 1, 790.
- 446 Leslie, S., Winney, B., Hellenthal, G., Davison, D., Boumertit, A., Day, T., Hutnik, K., Royrvik,
447 E.C., Cunliffe, B., Wellcome Trust Case Control, C., *et al.* (2015). The fine-scale genetic
448 structure of the British population. *Nature* 519, 309-314.

449 Liu, S., Huang, S., Chen, F., Zhao, L., Yuan, Y., Francis, S.S., Fang, L., Li, Z., Lin, L., Liu, R., *et al.*
450 (2018). Genomic Analyses from Non-invasive Prenatal Testing Reveal Genetic Associations,
451 Patterns of Viral Infections, and Chinese Population History. *Cell* *175*, 347-359 e314.
452 Loh, P.R., Danecek, P., Palamara, P.F., Fuchsberger, C., Y, A.R., H, K.F., Schoenherr, S., Forer,
453 L., McCarthy, S., Abecasis, G.R., *et al.* (2016). Reference-based phasing using the Haplotype
454 Reference Consortium panel. *Nature genetics* *48*, 1443-1448.
455 Manichaikul, A., Mychaleckyj, J.C., Rich, S.S., Daly, K., Sale, M., and Chen, W.M. (2010).
456 Robust relationship inference in genome-wide association studies. *Bioinformatics* *26*, 2867-
457 2873.
458 Martin, A.R., Karczewski, K.J., Kerminen, S., Kurki, M.I., Sarin, A.P., Artomov, M., Eriksson,
459 J.G., Esko, T., Genovese, G., Havulinna, A.S., *et al.* (2018). Haplotype Sharing Provides
460 Insights into Fine-Scale Population History and Disease in Finland. *Am J Hum Genet* *102*,
461 760-775.
462 Meacham, W. (1999). Neolithic to historic in the Hong Kong region. *Bulletin of the Indo-*
463 *Pacific Prehistory Association* *18*, 121-128.
464 Oksanen, J., Kindt, R., Legendre, P., O'Hara, B., Stevens, M.H.H., Oksanen, M.J., and Suggests,
465 M. (2007). The vegan package. *Community ecology package* *10*, 719.
466 Palamara, P.F., Lencz, T., Darvasi, A., and Pe'er, I. (2012). Length distributions of identity by
467 descent reveal fine-scale demographic history. *Am J Hum Genet* *91*, 809-822.
468 Pedersen, T. (2017). Patchwork: the composer of ggplots. R package version 0.0. 1.
469 Powell, J.E., Visscher, P.M., and Goddard, M.E. (2010). Reconciling the analysis of IBD and
470 IBS in complex trait studies. *Nat Rev Genet* *11*, 800-805.
471 Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar,
472 P., de Bakker, P.I., Daly, M.J., *et al.* (2007). PLINK: a tool set for whole-genome association
473 and population-based linkage analyses. *Am J Hum Genet* *81*, 559-575.
474 Schloerke, B., Crowley, J., Cook, D., Hofmann, H., Wickham, H., Briatte, F., Marbach, M.,
475 Thoen, E., Elberg, A., and Larmarange, J. (2011). Ggally: Extension to ggplot2.
476 Speed, D., and Balding, D.J. (2015). Relatedness in the post-genomic era: is it still useful? *Nat*
477 *Rev Genet* *16*, 33-44.
478 Wen, B., Li, H., Lu, D., Song, X., Zhang, F., He, Y., Li, F., Gao, Y., Mao, X., and Zhang, L. (2004).
479 Genetic evidence supports demic diffusion of Han culture. *Nature* *431*, 302-305.
480 Wickham, H. (2012). reshape2: Flexibly reshape data: a reboot of the reshape package. R
481 package version 1.
482 Wickham, H. (2016). ggplot2: elegant graphics for data analysis (springer).
483 Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L.D.A., François, R., Grolemund,
484 G., Hayes, A., Henry, L., and Hester, J. (2019). Welcome to the Tidyverse. *Journal of Open*
485 *Source Software* *4*, 1686.
486 Xu, S., Yin, X., Li, S., Jin, W., Lou, H., Yang, L., Gong, X., Wang, H., Shen, Y., and Pan, X. (2009).
487 Genomic dissection of population substructure of Han Chinese and its implication in
488 association studies. *The American Journal of Human Genetics* *85*, 762-774.
489 Yang, J., Lee, S.H., Goddard, M.E., and Visscher, P.M. (2011). GCTA: a tool for genome-wide
490 complex trait analysis. *Am J Hum Genet* *88*, 76-82.
491 Yu, G., Smith, D.K., Zhu, H., Guan, Y., and Lam, T.T.Y. (2017). ggtree: an R package for
492 visualization and annotation of phylogenetic trees with their covariates and other
493 associated data. *Methods in Ecology and Evolution* *8*, 28-36.
494

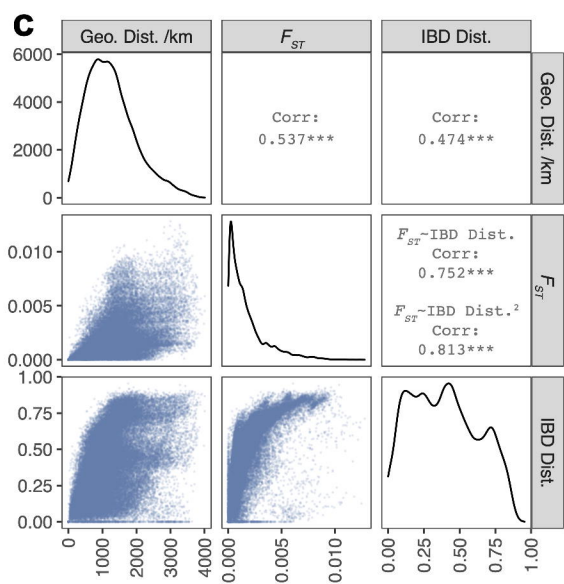


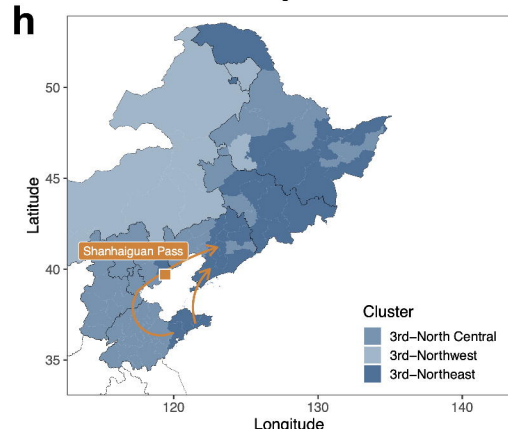
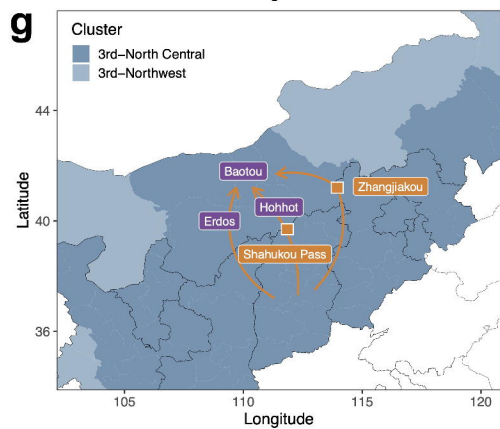
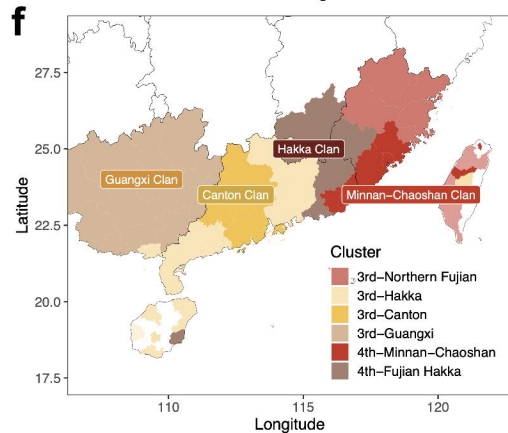
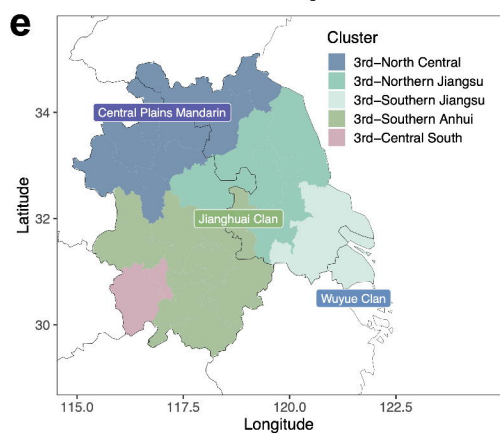
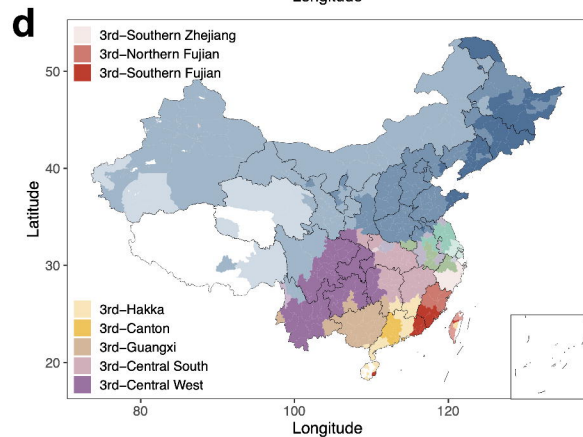
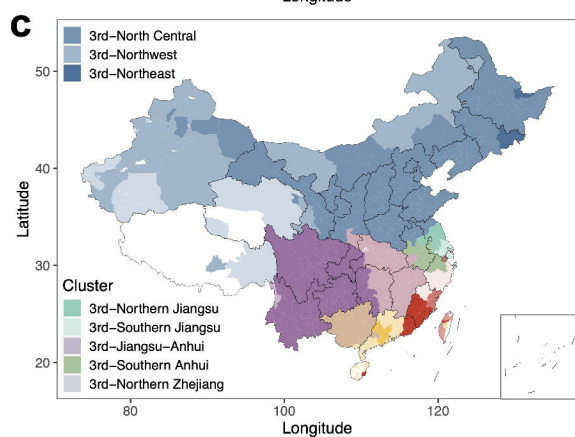
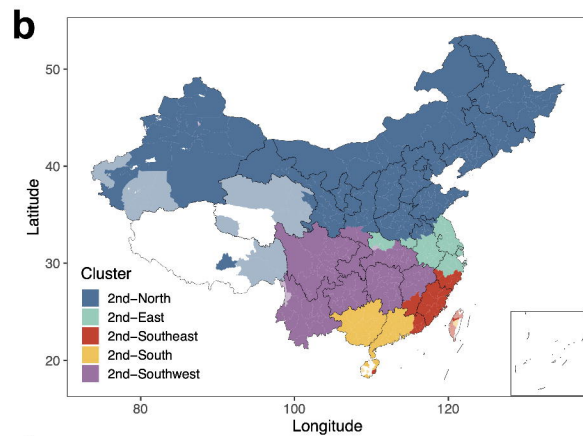
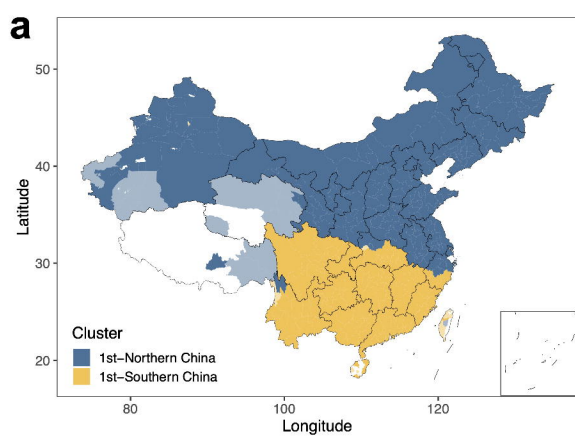
Region

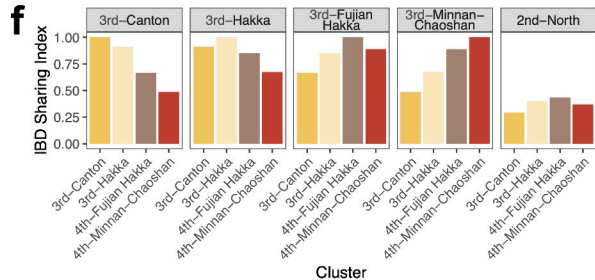
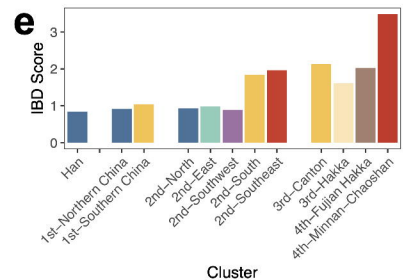
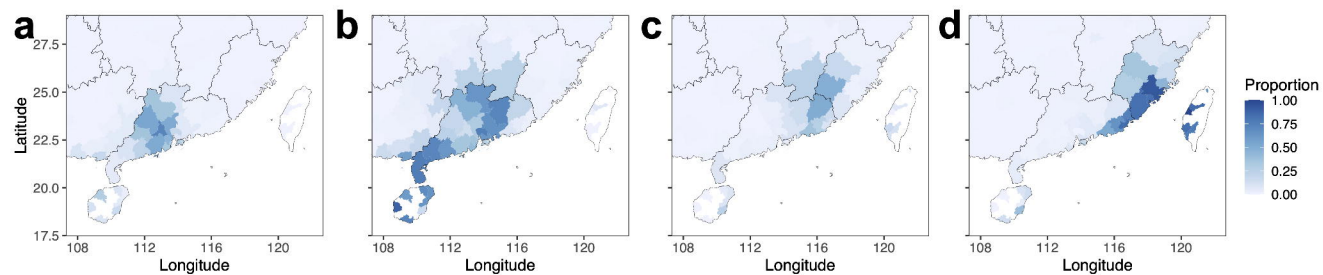
- Central
- East
- North
- Northeast
- Northwest
- South
- Southwest

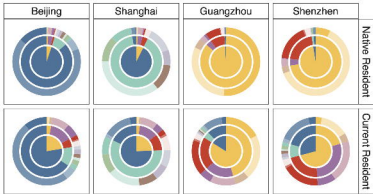
1KGP Group

- ▲ CDX
- ▲ CHB
- ▲ CHS
- ▲ JPT
- ▲ KHV









1st-level

- 1st-Northern China
- 1st-Southern China

2nd-level

- 2nd-North
- 2nd-East
- 2nd-Southeast
- 2nd-Southwest
- 2nd-South

3rd-level

- 3rd-Northeast
- 3rd-North Central
- 3rd-Northwest
- 3rd-Northern Jiangsu
- 3rd-Southern Anhui
- 3rd-Southern Jiangsu
- 3rd-Zhejiang-Anhui
- 3rd-Jiangsu-Anhui
- 3rd-Northern Zhejiang
- 3rd-Southern Zhejiang
- 3rd-Northern Fujian
- 3rd-Southern Fujian
- 3rd-Central West
- 3rd-Central South
- 3rd-Guangxi
- 3rd-Hakka
- 3rd-Canton