River valleys shaped the maternal genetic landscape of Han Chinese

Yu-Chun Li[1,3,4,#], Wei-Jian Ye[2,#], Chuan-Gui Jiang[2,#], Zhen Zeng[2], Jiao-Yang Tian[1,3,4], Li-Qin Yang[1,3,4], Kai-Jun Liu[2], Qing-Peng Kong[1,3,4,*]

[1] State Key Laboratory of Genetic Resources and Evolution/Key Laboratory of Healthy Aging Research of Yunnan Province, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming 650223, China;

[2] Chengdu 23 Mofang Biotechnology Co., Ltd., Chengdu, 610000, China;

[3] CAS Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming 650223, China;

[4] Kunming Key Laboratory of Healthy Aging Molecular Mechanism Study, Kunming 650223, China.

[#] These authors contributed equally to this work.

*** Corresponding author:**

Qing-Peng Kong, State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming 650223, China;

Telephone: +86-871-68125403; Fax: +86-871-68125403

E-mail: kongqp@mail.kiz.ac.cn or qpkong@163.com

Supplementary Figures

Figure S1. PC map of Han populations based on haplogroup frequency matrix with samples assigned to terminal branches according to the PhyloTree (Table S4). Populations with a sample size of <20 individuals were excluded. (a) Han populations are classified into southern and northern Han groups; (b) Han populations are colored according to different river valleys; a and b are the same with Figure 2a and b, respectively; (c) Plot of haplogroup contribution of the first and second PC. Contribution of each haplogroup was calculated as factor scores for PC1 and PC2 with regression (REGR) using SPSS (v16.0) software.

Figure S2. PC map of Han populations based on haplogroup frequency matrix of major haplogroups (Table S5). Populations with a sample size of <20 individuals were excluded. (a) Han populations are classified into southern and northern Han groups; (b) Han populations are colored according to different river valleys; (c) Plot of haplogroup contribution of the first and second PC. Contribution of each haplogroup was calculated as the factor scores for PC1 and PC2 with regression (REGR) using SPSS (v16.0) software.

Figure S3. Correlations between (a) latitude and PC1, (b) latitude and PC2, (c) longitude and PC1, (d) longitude and PC2.

Figure S4. Unrooted neighbor-joining (NJ) tree of Han populations based on pair-wise Fst values (Table S6).

Figure S5. Testing the West-East divergence by removing populations with N<100. (a) PC map of populations with samples sizes more than 100 individuals. (b) Correlations between longitude and PC1.

Figure S6. Boxplot of Fst values (Table S9) showing significant higher Fst values between river valleys than those between geographic regions. [1]Fst values when populations located across more than one river valleys are excluded.

Figure S7. Hot zone distribution of (a) B4*, (b) A*, (c) D4*, (d) M7b1a1* and (e) F1* shown on a map of China. Blue-yellow-red represents low-middle-high frequencies, respectively.

Supplementary Tables

Table S1. Information on the 33 Han Chinese populations investigated in this study.

Table S2. Variant and haplogroup information of the 21,668 Han Chinese samples.

Table S3. Frequencies of mtDNA haplogroups in Han Chinese in different groups.

Table S4. Frequency matrix of haplogroups assigned to terminal lineages.

Table S5. Frequency matrix of major haplogroups.

Table S6. Pairwise Fst values of Han Chinese populations in this study.

Table S7. Fischer's exact test for frequency difference of each haplogroup between southern and northern Han Chinese groups.

Table S8. Classification of populations into different groups in AMOVA.

Table S9. Fst values between different Han groups.

Table S10. Coalescent ages of D4, B4 and M7, as well as their sub-branches.

Table S11. Proportions of lineages within different age periods in D4, M7 and B4.

Table S12. Variant and haplogroup information of the 218 newly sequenced Han samples.

Table S13. List of the 4,004 MT loci.

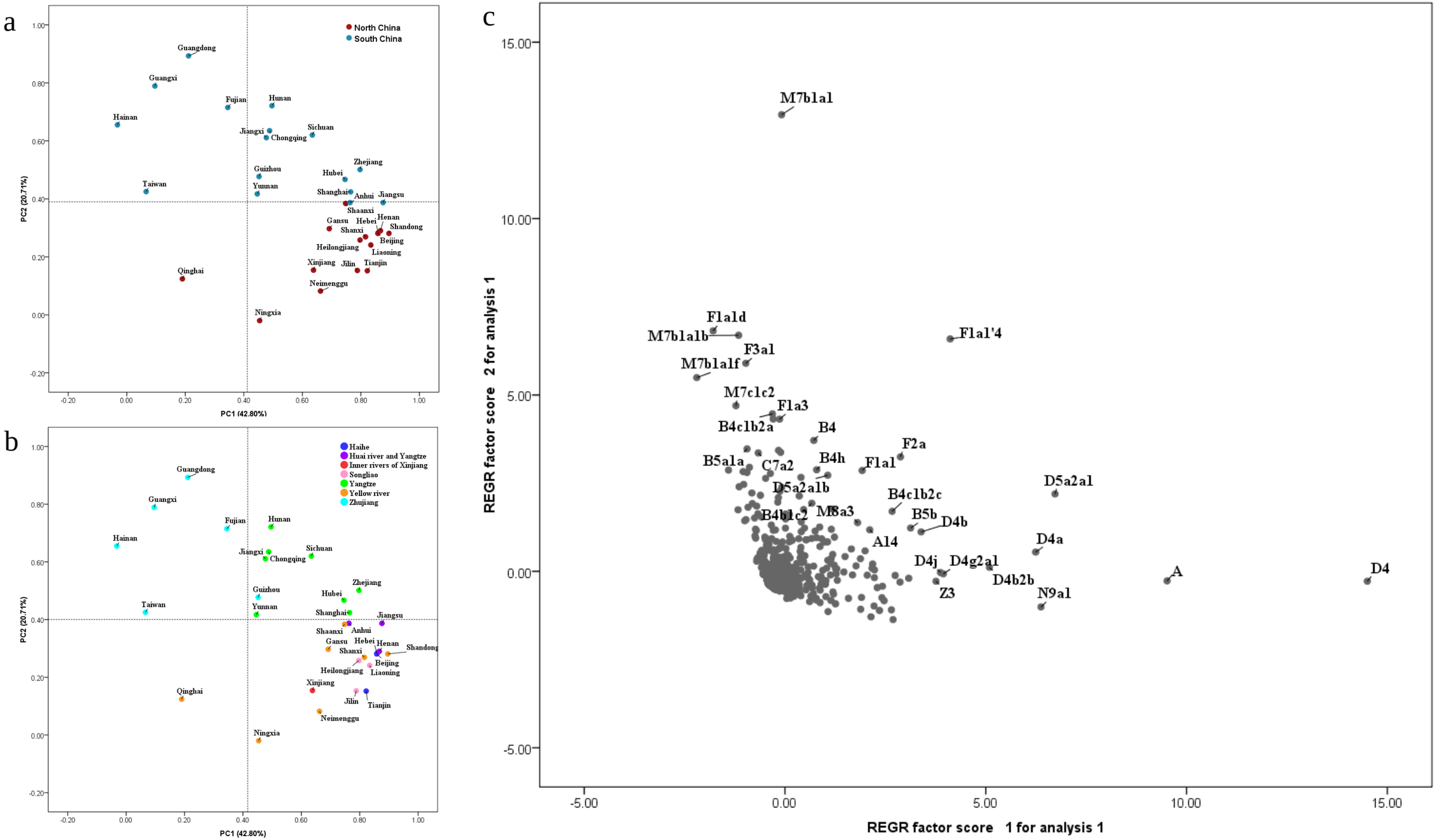

Figure S1. PC map of Han populations based on haplogroup frequency matrix with samples assigned to terminal branches according to the PhyloTree (Table S4). Populations with a sample size of <20 individuals were excluded. (a) Han populations are classified into southern and northern Han groups; (b) Han populations are colored according to different river valleys; a and b are the same with Figure 2a and b, respectively; c. Plot of haplogroup contribution of the first and second PC. Contribution of each haplogroup was calculated as factor scores for PC1 and PC2 with regression (REGR) using SPSS (v16.0) software.

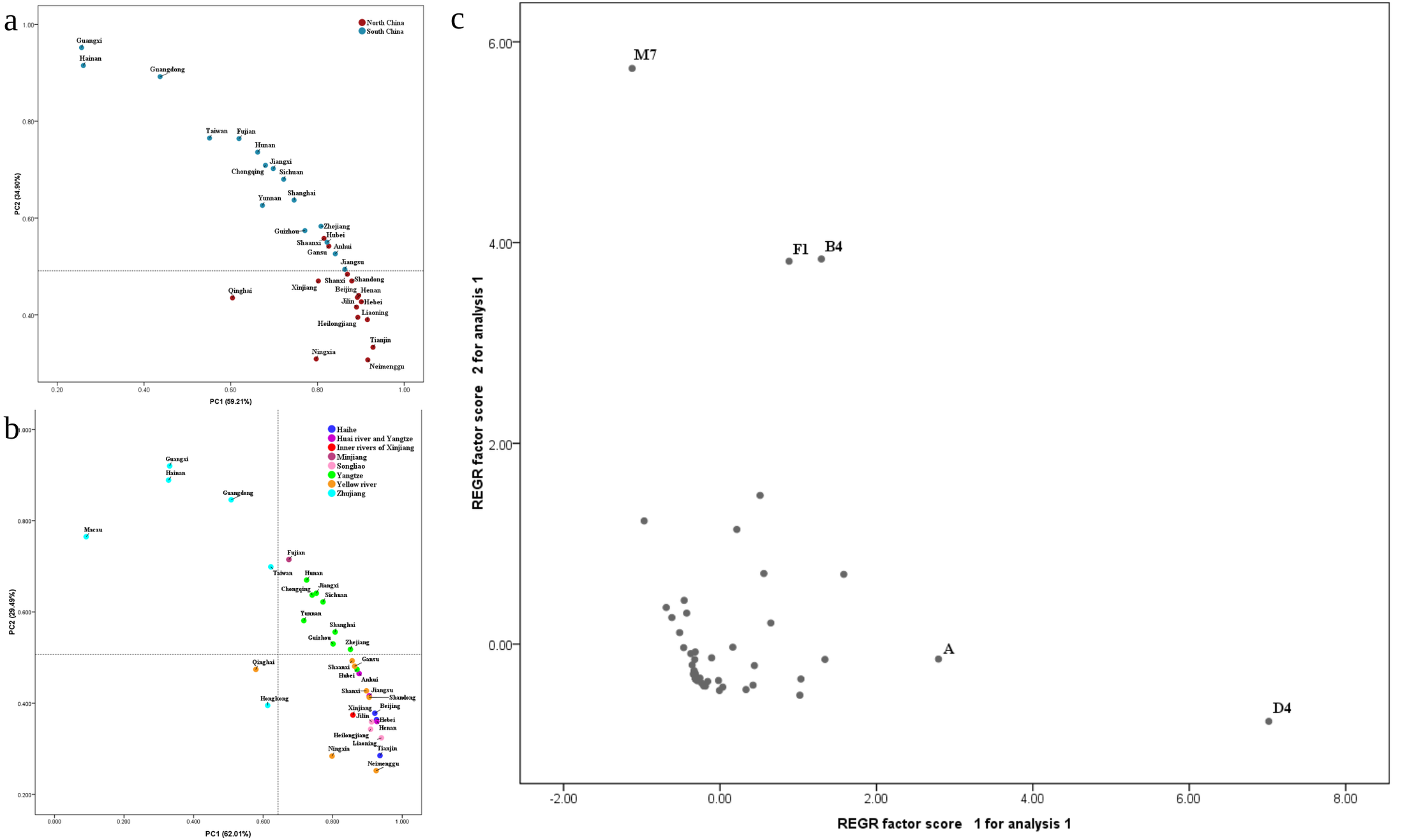

Figure S2. PC map of Han populations based on haplogroup frequency matrix of major haplogroups (Table S5). Populations with a sample size of <20 individuals were excluded. a. Han populations are classified into southern and northern Han groups; b. Han populations are colored according to different river valleys; c; Plot of haplogroup contribution of the first and second PC. Contribution of each haplogroup was calculated as the factor scores for PC1 and PC2 with regression (REGR) using SPSS (v16.0) software.

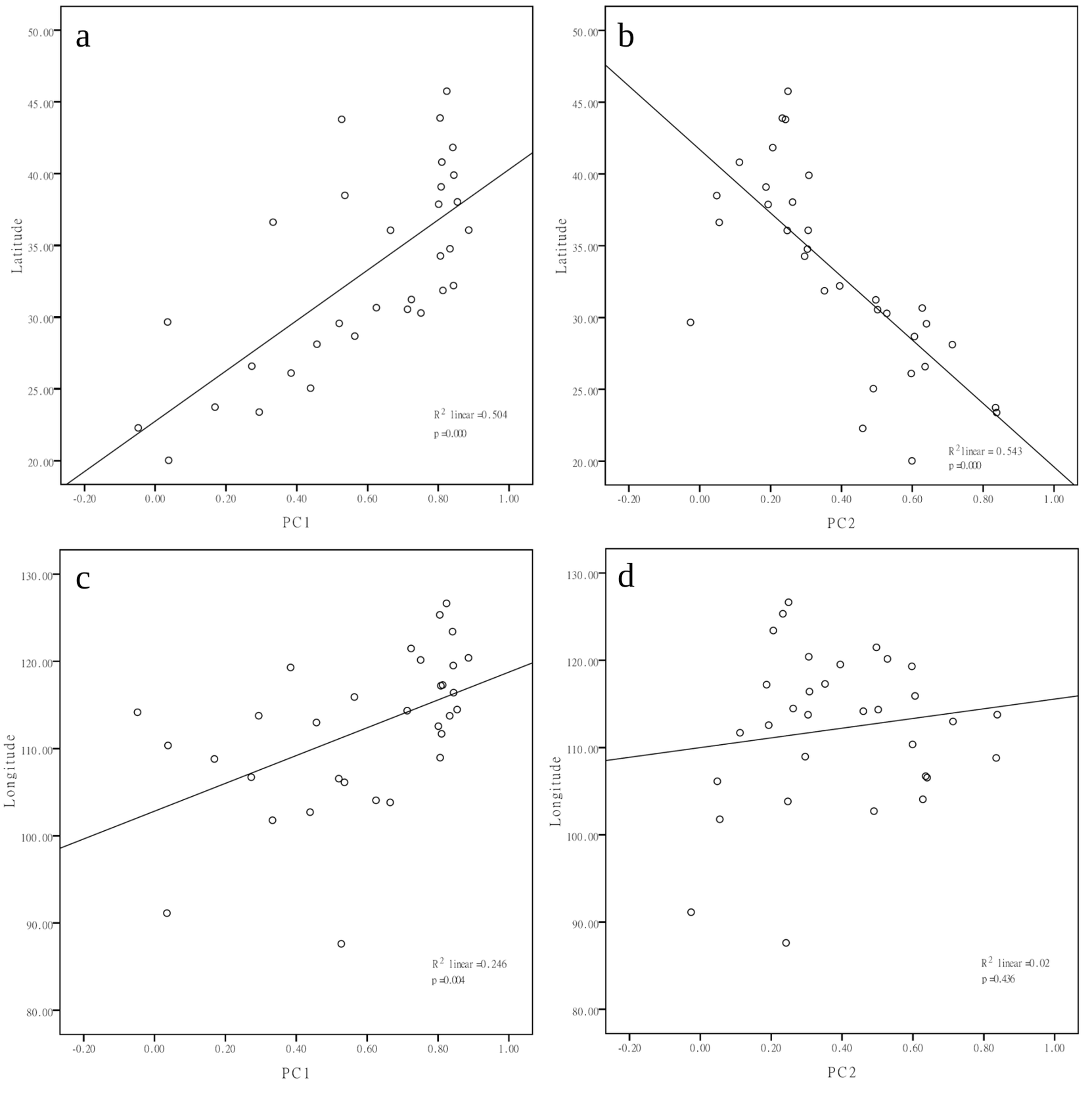

Figure S3. Correlations between (a) latitude and PC1, (b) latitude and PC2, (c) longitude and PC1, (d) longitude and PC2.

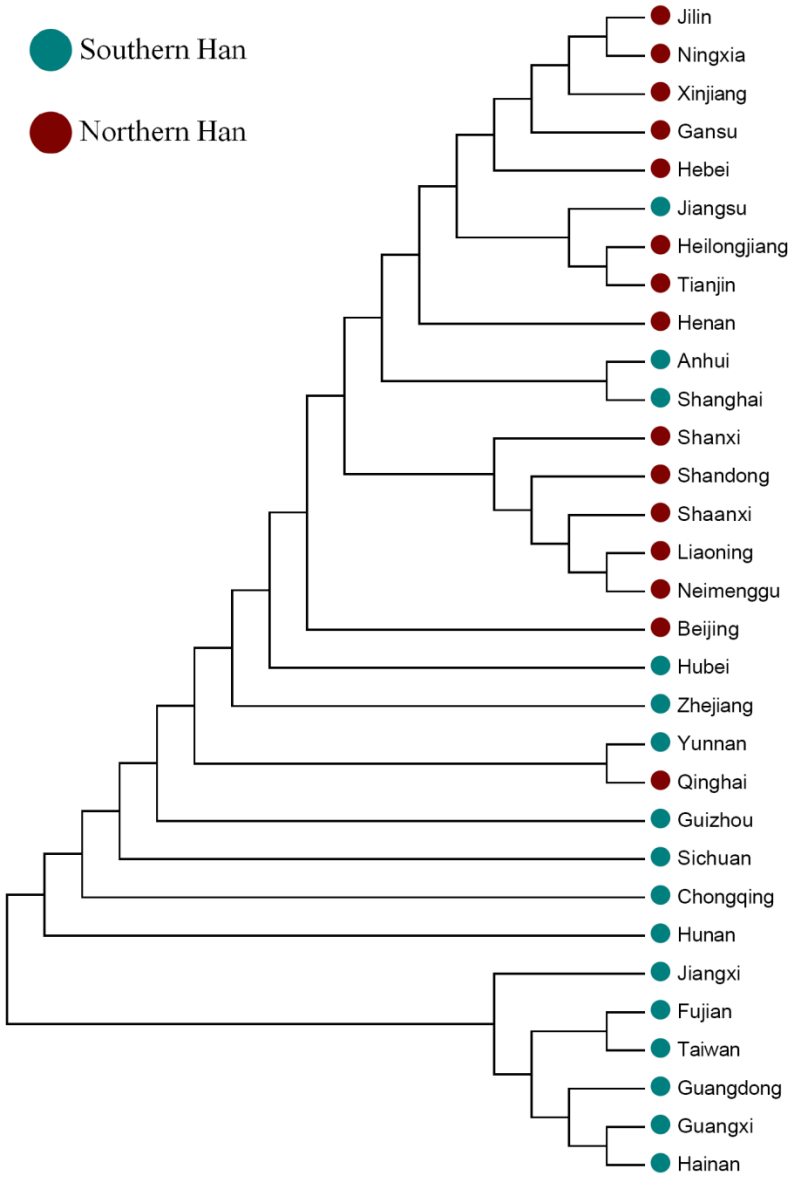

Figure S4. Unrooted Neighbor-Joining (NJ) tree of Han populations based on pair-wise Fst values (Table S6).

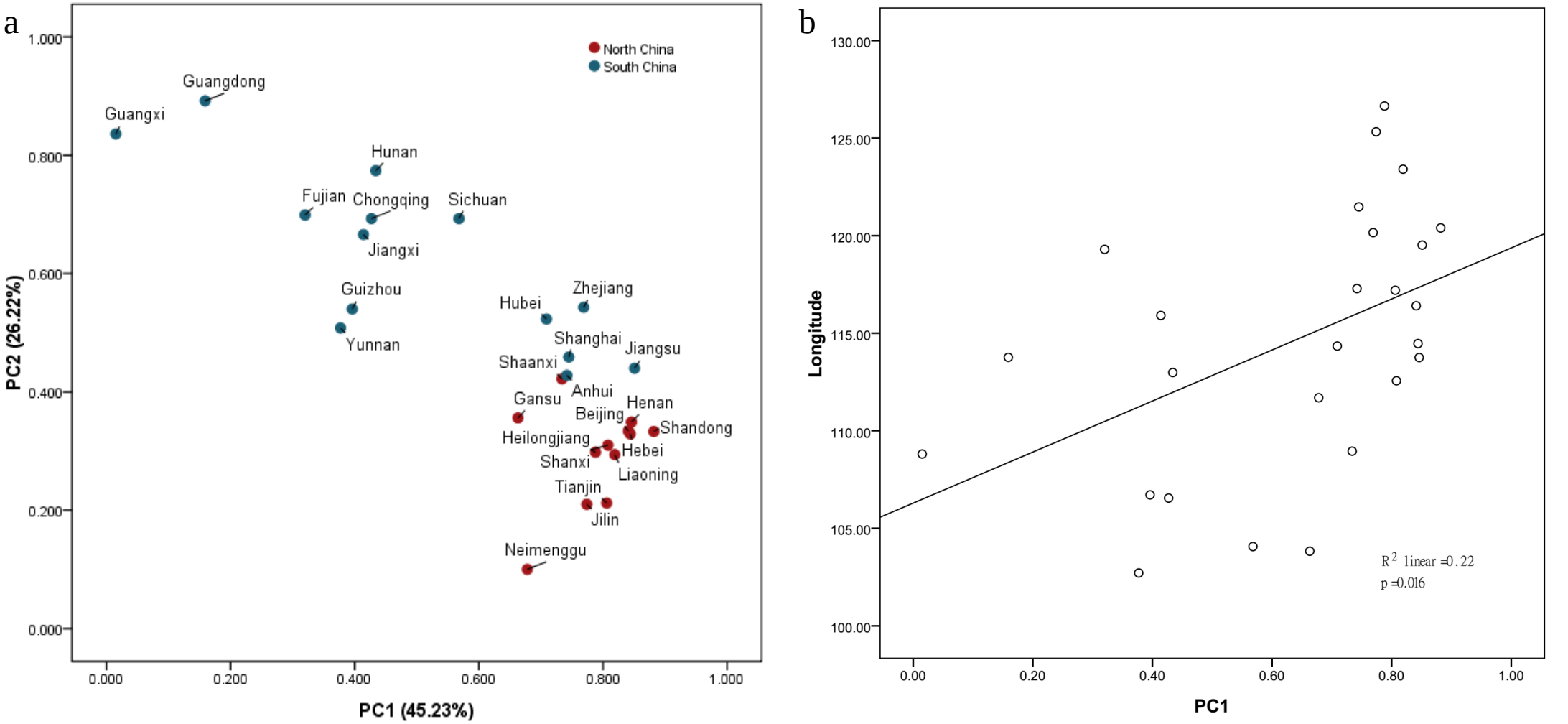

Figure S5. Testing the West-East divergence by removing populations with N<100. (a) PC map of populations with samples sizes more than 100 individuals; (b) Correlations between longitude and PC1.

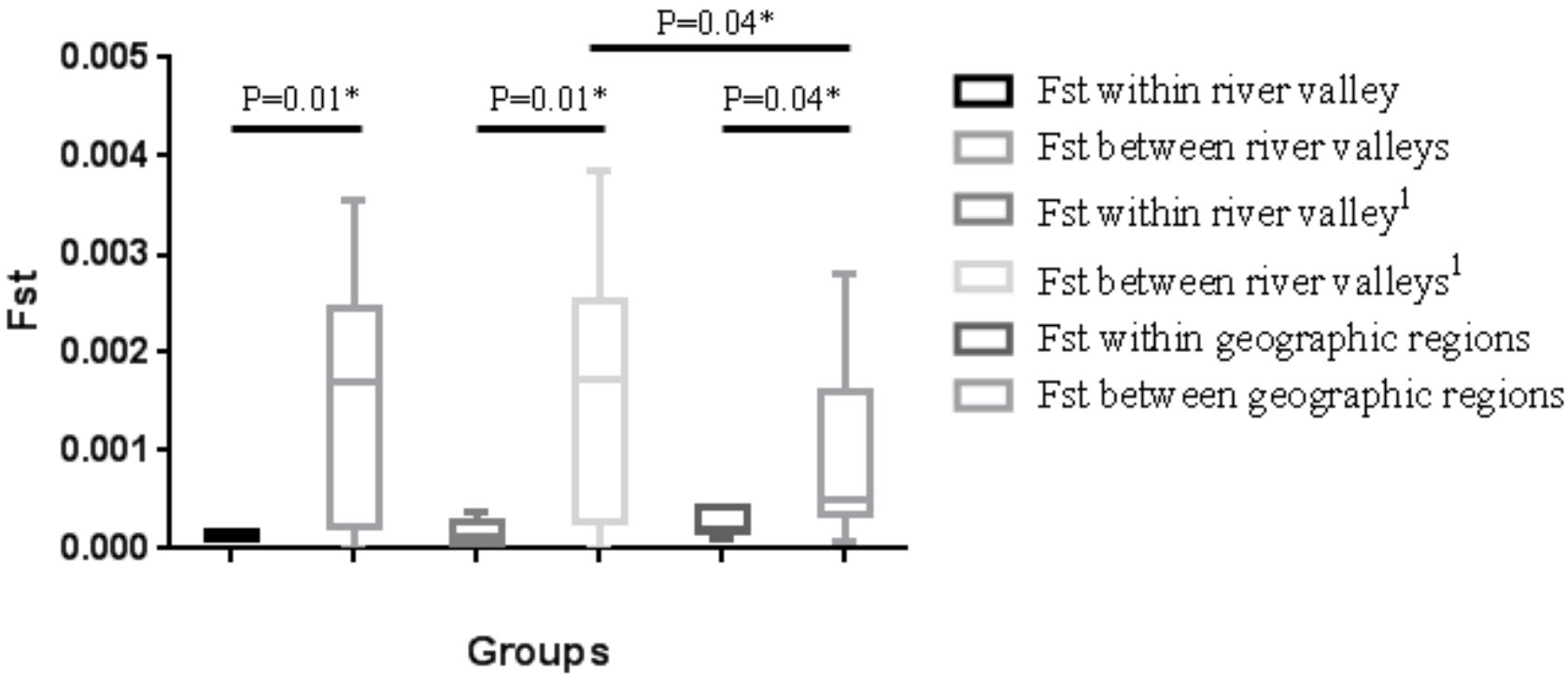

Figure S6. Boxplot of Fst values (Table S9) showing significant higher Fst values between river valleys than those between geographic regions. [1]Fst values when populations located across more than one river valleys are excluded.

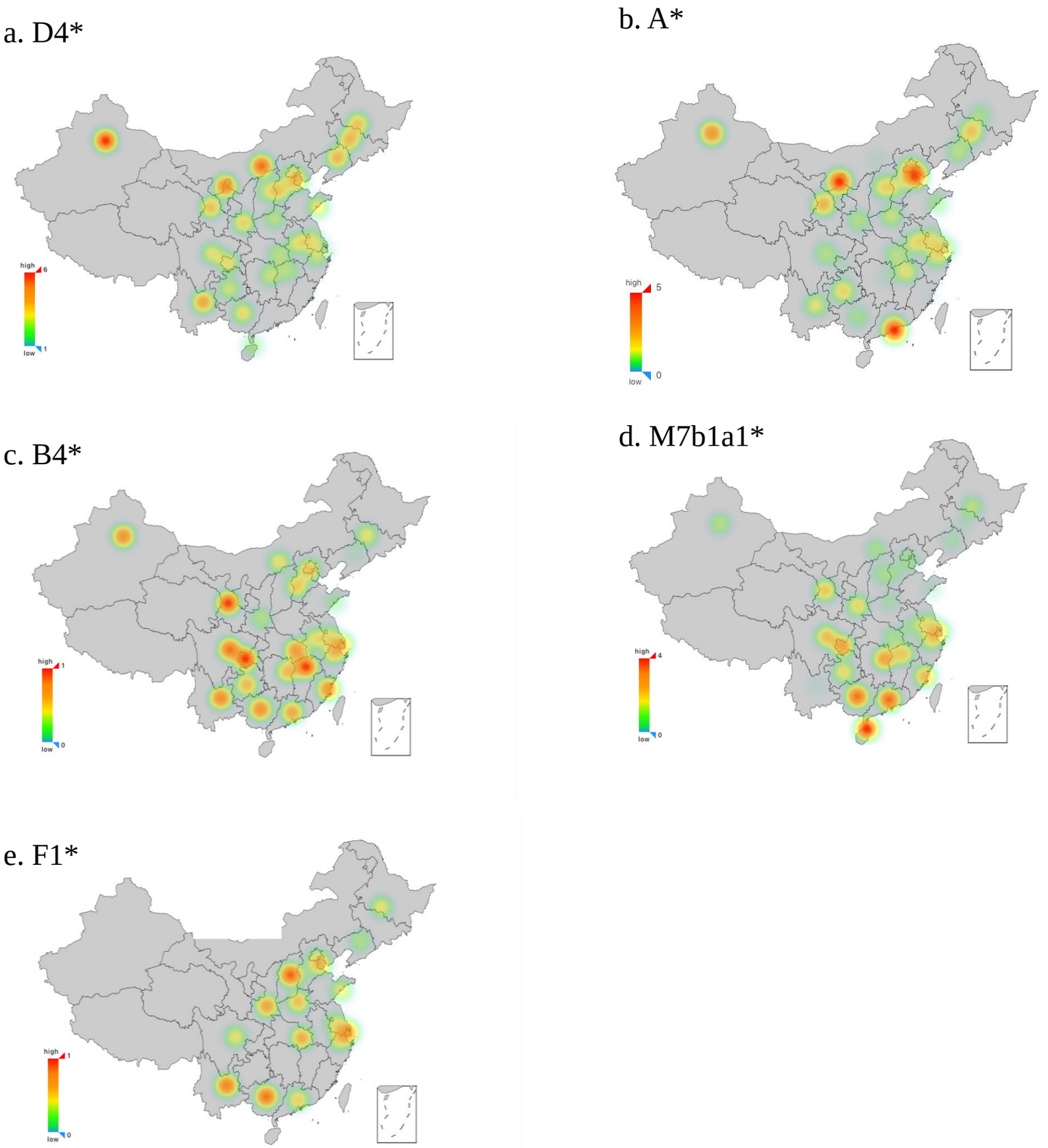

Figure S7. Hot zone distributions of (a) D4*, (b) A*, (c) B4*, (d) M7b1a1* and (e) F1* shown on a map of China. Blue-yellow-red represents low-middle-high frequencies, respectively.