The genomic formation of Tanka people, an isolated “Gypsies in water” in the coastal region of Southeast China

Guanglin He[1,2,#,*], Yunhe Zhang[3,#], Lan-Hai Wei[1,9, #,*], Mengge Wang[4,5], Xiaomin Yang[1], Jianxin Guo[1], Jin Sun[6], Rong Hu[1], Chuan-Chao Wang[1,7,8,*], Xian-Qing Zhang[1,*]

[1]Department of Anthropology and Ethnology, Institute of Anthropology, National Institute for Data Science in Health and Medicine, State Key Laboratory of Cellular Stress Biology, School of Life Sciences, State Key Laboratory of Marine Environmental Science, Xiamen University, Xiamen, 361005, China
[2]School of Humanities, Nanyang Technological University, Nanyang Avenue, 639798, Singapore
[3]School of Public Administration, Zhejiang Gongshang University, Hangzhou, 310018, China
[4]Guangzhou Forensic Science Institute, Guangzhou, 510080, China
[5]Faculty of Forensic Medicine, Zhongshan School of Medicine, Sun Yat-sen University, No. 74 Zhongshan Road II, Guangzhou, 510080, China
[6]Xingyi Normal University for Nationalities, Xingyi 562400, China
[7]School of Basic Medical Sciences, Zhejiang University School of Medicine, Hangzhou, 310000, China
[8]Institute of Asian Civilizations, Zhejiang University, Hangzhou, 310000, China
[9]B&R International Joint Laboratory for Eurasian Anthropology, Fudan University, Shanghai 200438, China
[#]These authors contributed equally to this work.
[*]Correspondence: Guanglinhe@163.com (GLH); Ryan.lh.wei@xmu.edu.cn (LHW); wang@xmu.edu.cn (CCW); and xqz@xmu.edu.cn (XQZ)

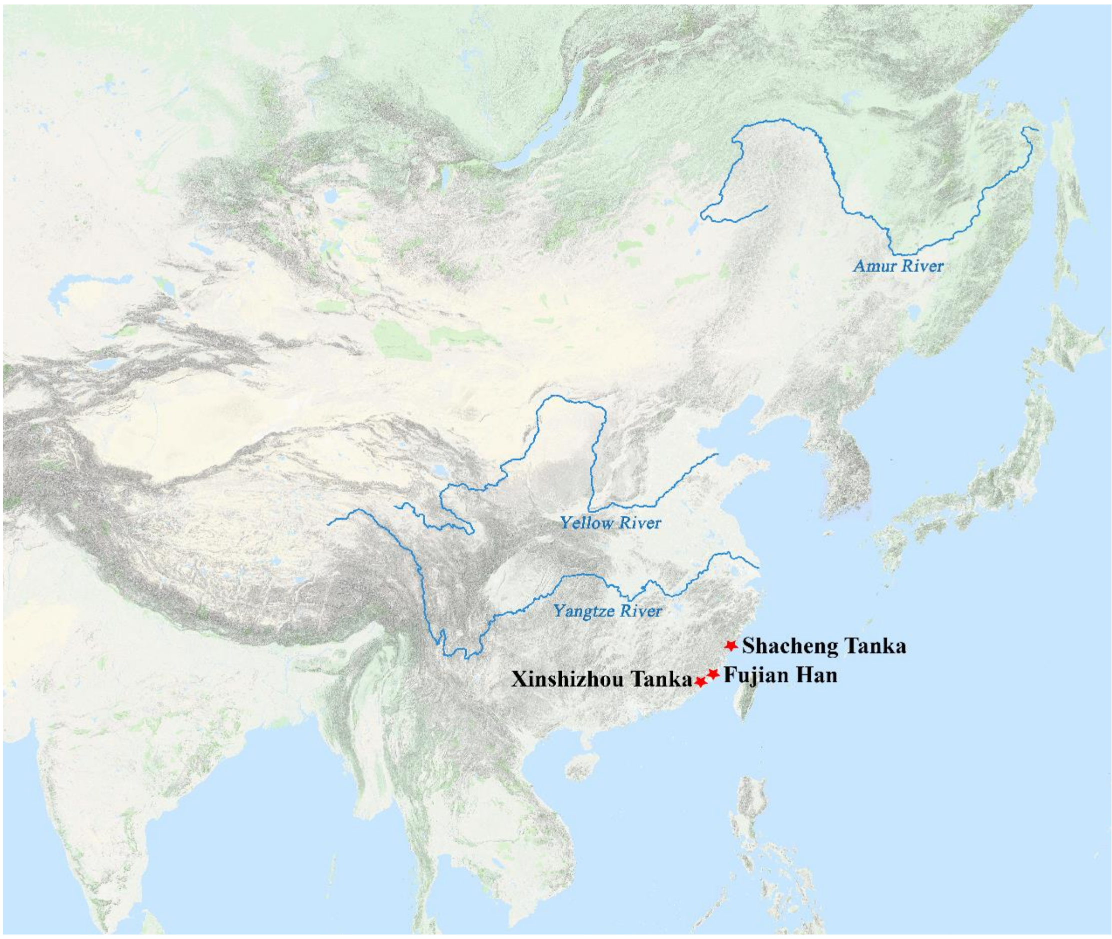

Figure 1. Geographical positions of two Tanka populations and one Han Chinese population collected from Fujian province in southeastern China.

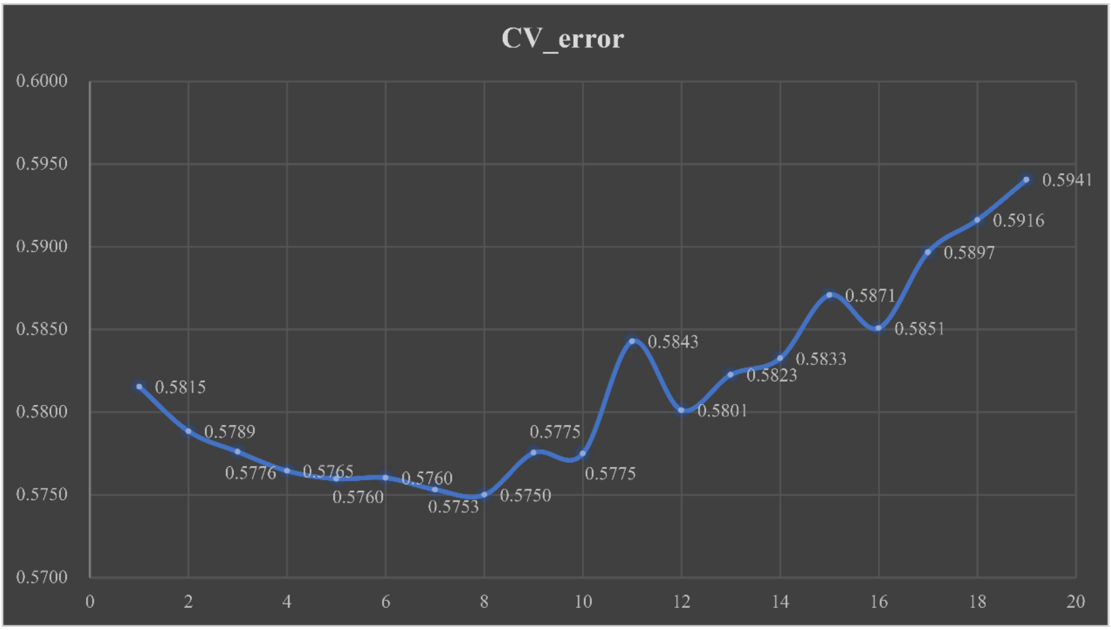

Figure S2. Cross-validation error in the model-based ADMIXTURE analyses. The best model is the eight-source-based mixed model with the smallest cross-validation error (0.5750).

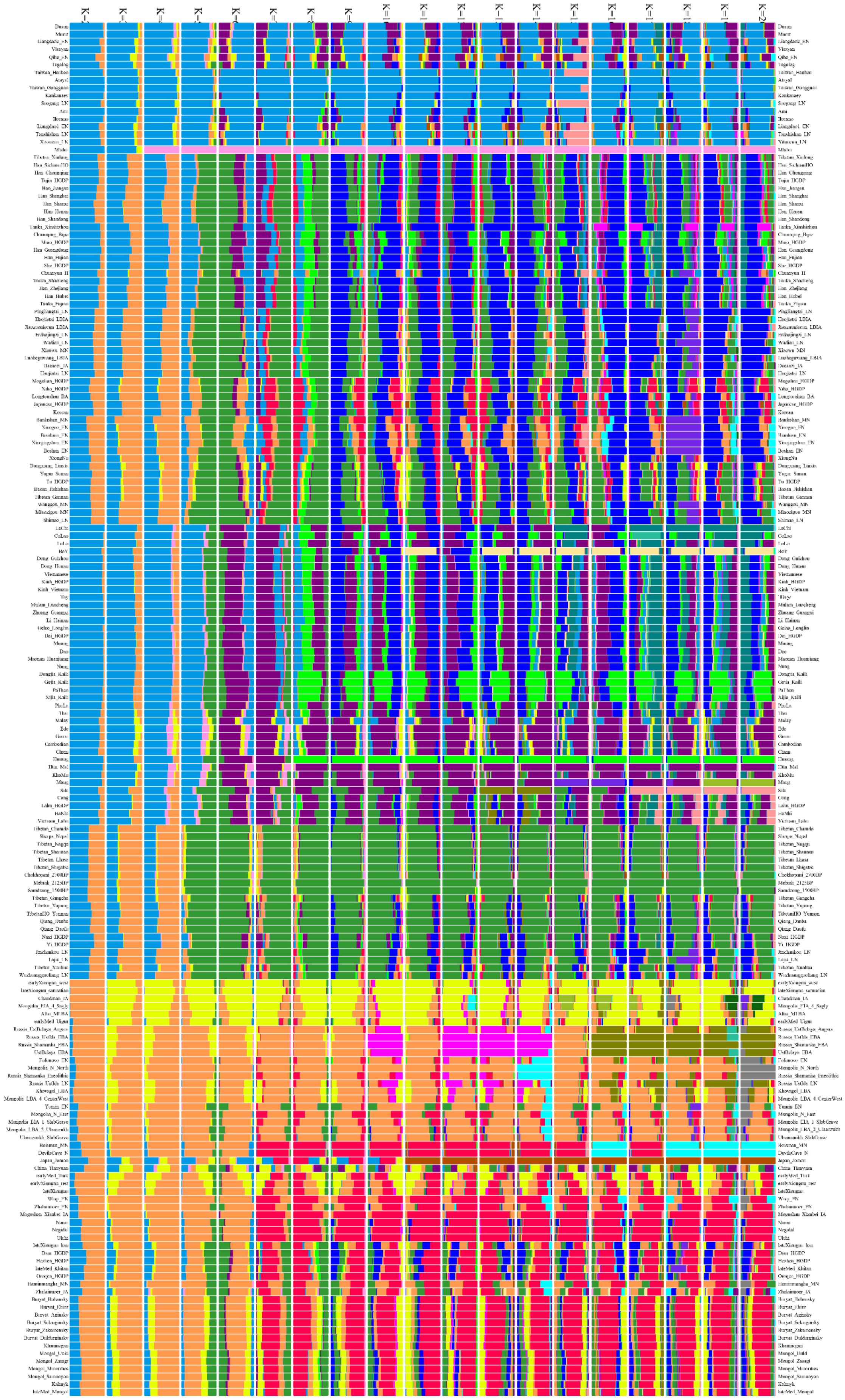
K=2
K=3
K=4
K=5
K=6
K=7
K=8
K=9
K=10
K=11
K=12
K=13
K=14
K=15
K=16
K=17
K=18
K=19
K=20
Dusun
Murut
Liangdao2_EN
Visayan
Qihe_EN
Tagalog
Taiwan_Hanben
Atayal
Taiwan_Gongguan
Kankanaey
Suogang_LN
Ami
Ilocano
Liangdao1_EN
Tanshishan_LN
Xitoucun_LN
Mlabri
Tibetan_Xinlong
Han_SichuanHO
Han_Chongqing
Tujia_HGDP
Han_Jiangsu
Han_Shanghai
Han_Shanxi
Han_Henan
Han_Shandong
Tanka_Xinshizhou
Chuanqing_Bijie
Miao_HGDP
Han_Guangdong
Han_Fujian
She_HGDP
Chuanyun_H
Tanka_Shacheng
Han_Zhejiang
Han_Hubei
Tanka_Fujian
Pingliangtai_LN
Haojiatai_LDIA
Jiaozuoniecun_LDIA
Erdaojingzi_LN
Wadian_LN
Xiaowu_MN
Luoheguxiang_LBIA
Dacaozi_IA
Haojiatai_LN
Mongolian_HGDP
Xibo_HGDP
Longtoushan_BA
Japanese_HGDP
Korean
Banlashan_MN
Xiaogao_EN
Bianbian_EN
Xiaojingshan_EN
Boshan_EN
XiongNu
Dongxiang_Linxia
Yugur_Sunan
Tu_HGDP
Baoan_Jishishan
Tibetan_Gannan
Wanggou_MN
Miaozigou_MN
Shimao_LN
LaChi
CoLao
LoLo
HoY
Dong_Guizhou
Dong_Hunan
Vietnamese
Kinh_HGDP
Kinh_Vietnam
Tay
Mulam_Luocheng
Zhuang_Guangxi
Li_Hainan
Gelao_Longlin
Dai_HGDP
Muong
Dao
Maonan_Huanjiang
Nung
Dongjia_Kaili
Gejia_Kaili
PaThen
Xijia_Kaili
PhuLa
Thai
Malay
Ede
Giarai
Cambodian
Cham
Hmong
Htin_Mal
KhoMu
Mang
Sila
Cong
Lahu_HGDP
HaNhi
Vietnam_Lahu
Tibetan_Chamdo
Sherpa_Nepal
Tibetan_Nagqu
Tibetan_Shannan
Tibetan_Lhasa
Tibetan_Shigatse
Chokhopani_2700BP
Mebrak_2125BP
Samdzong_1500BP
Tibetan_Gangcha
Tibetan_Yajiang
TibetanHO_Yunnan
Qiang_Danba
Qiang_Daofu
Naxi_HGDP
Yi_HGDP
Jinchankou_LN
Lajia_LN
Tibetan_Xunhua
Wuzhuangguoliang_LN
earlyXiongnu_west
lateXiongnu_sarmatian
Chandman_IA
Mongolia_EIA_4_Sagly
Altai_MLBA
earlyMed_Uigur
Russia_UstBelaya_Angara
Russia_UstIda_EBA
Russia_Shamanka_EBA
UstBelaya_EBA
Fofonovo_EN
Mongolia_N_North
Russia_Shamanka_Eneolithic
Russia_UstIda_LN
Khovsgol_LBA
Mongolia_LBA_4_CenterWest
Yumin_EN
Mongolia_N_East
Mongolia_EIA_1_SlabGrave
Mongolia_LBA_2_Ulaanzuukh
Ulaanzuukh_SlabGrave
Boisman_MN
DevilsCave_N
Japan_Jomon
China_Tianyuan
earlyMed_Turk
earlyXiongnu_rest
lateXiongnu
Wuqi_EN
Zhalainuoer_EN
Mogushan_Xianbei_IA
Nanai
Negidal
Ulchi
lateXiongnu_han
Daur_HGDP
Hezhen_HGDP
lateMed_Khitan
Oroqen_HGDP
Haminmangha_MN
Zhalainuoer_IA
Buryat_Bohansky
Buryat_Ehirit
Buryat_Aginsky
Buryat_Selenginsky
Buryat_Zakamensky
Buryat_Duldurginsky
Khamnigan
Mongol_Huld
Mongol_Zasagt
Mongol_Monenthus
Mongol_Sumnyan
Kalmyk
lateMed_Mongol

Figure S3. Results of model-based ADMIXTURE analyses results with the predefined ancestral sources ranging from two to twenty.

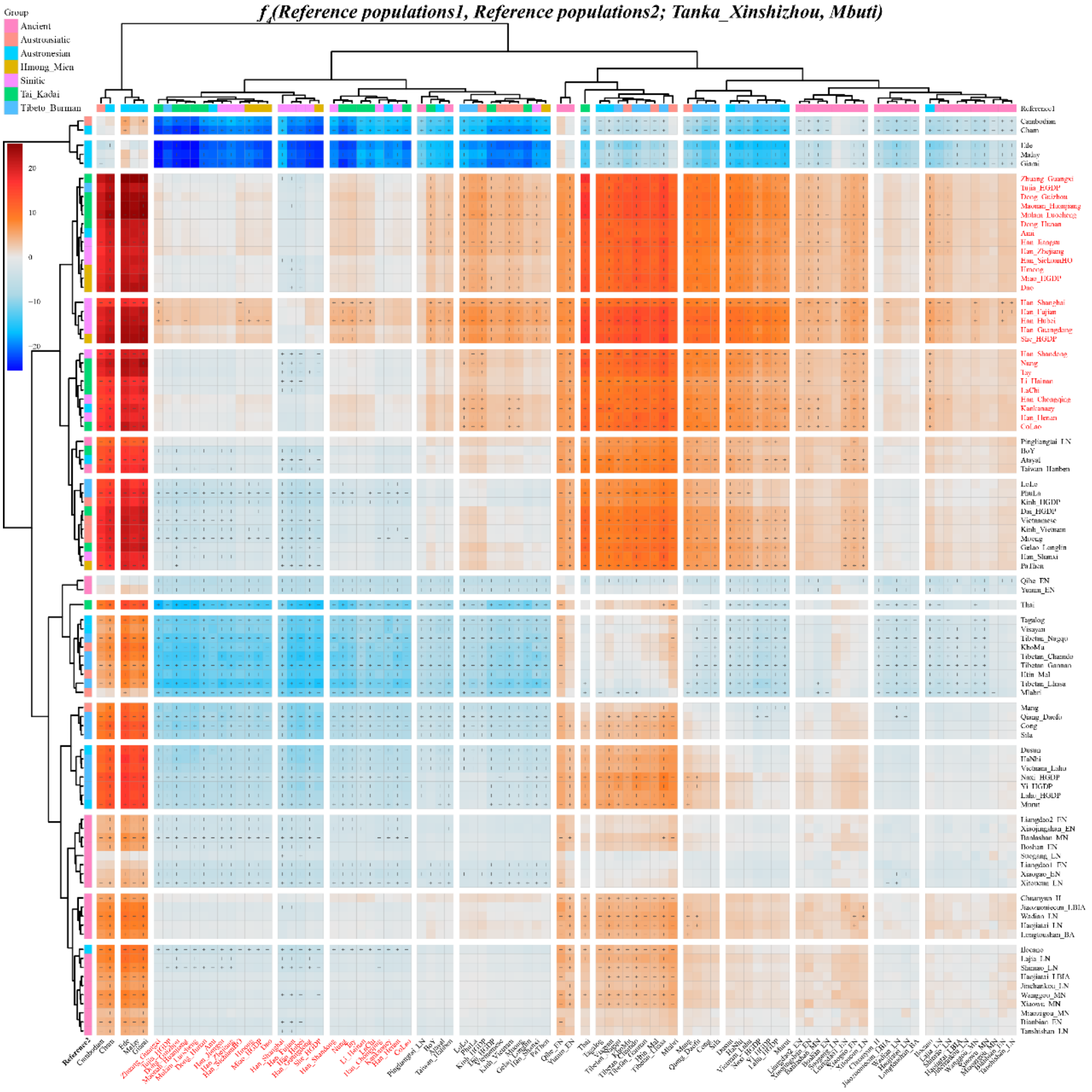

Figure S4. Formal test of genomic affinity in Xinshizhou Tanka people inferred from the two-population comparison f_4-statistics in the form f_4*(Reference population1, Reference population2; Tanka_Xinshizhou, Mbuti).* Red color denoted the positive f_4-values, which suggested Xinshizhou Tanka people shared more derived mutations with reference population1 (left population lists), and blue color showed the negative f_4-values, which suggested Xinshizhou Tanka people shared more alleles with reference population2 (bottom population lists). Statistically significant results were marked with the '+'.

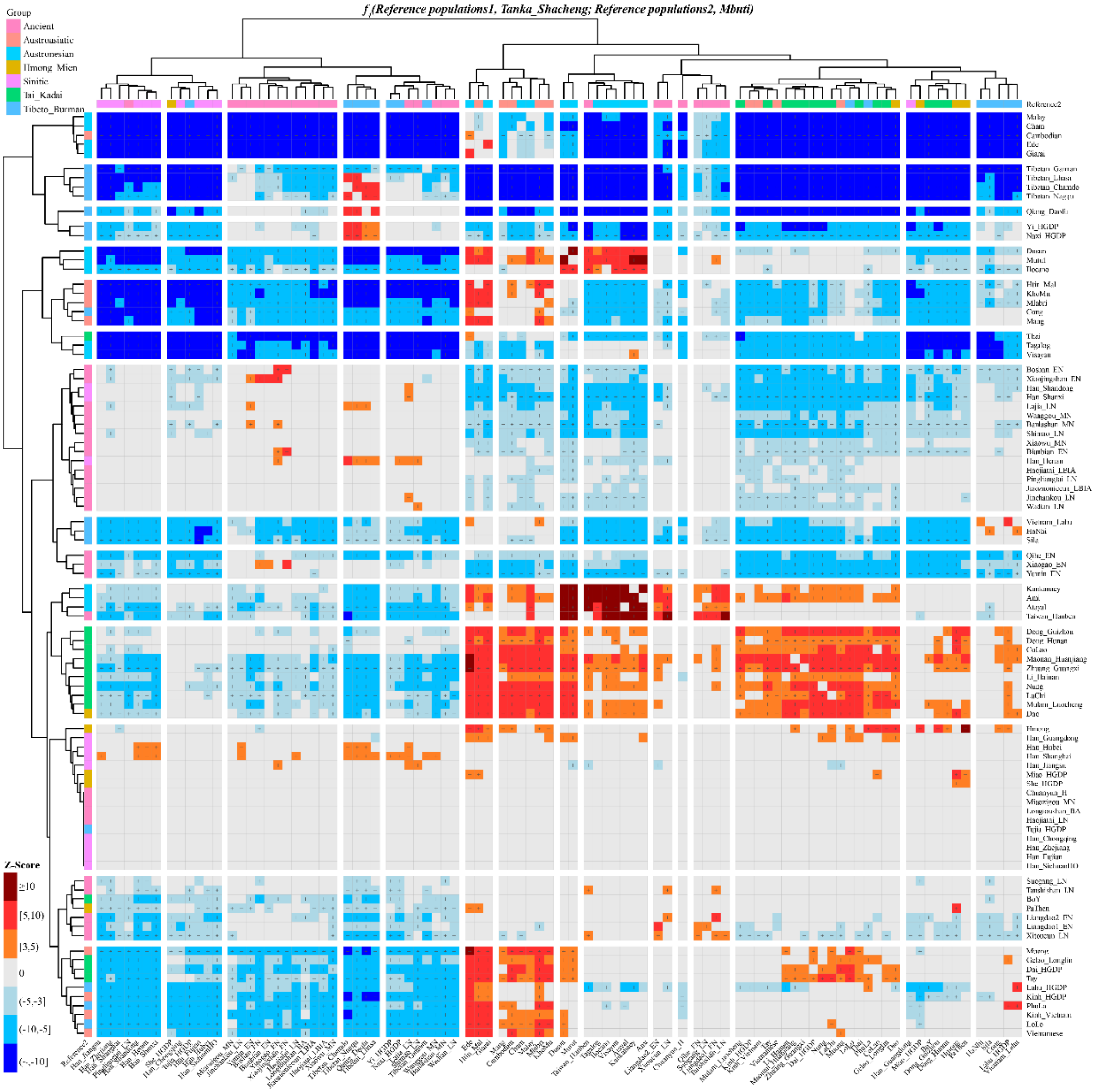

Figure S5. A formal test of genomic continuity and admixture in Shacheng Tanka people inferred from the two-population comparison f_4-statistics in the form f_4*(Reference population1, Reference population2; Tanka_Shacheng, Mbuti).* Red color denoted the positive f_4-values, which suggested reference population2 (bottom population lists) shared more derived mutations with reference population1 (left population lists), and blue color showed the negative f_4-values, which suggested reference population2 shared more alleles with Shacheng population and gray color showed no statistically significant results were observed. Statistically significant results were marked with the '+'.

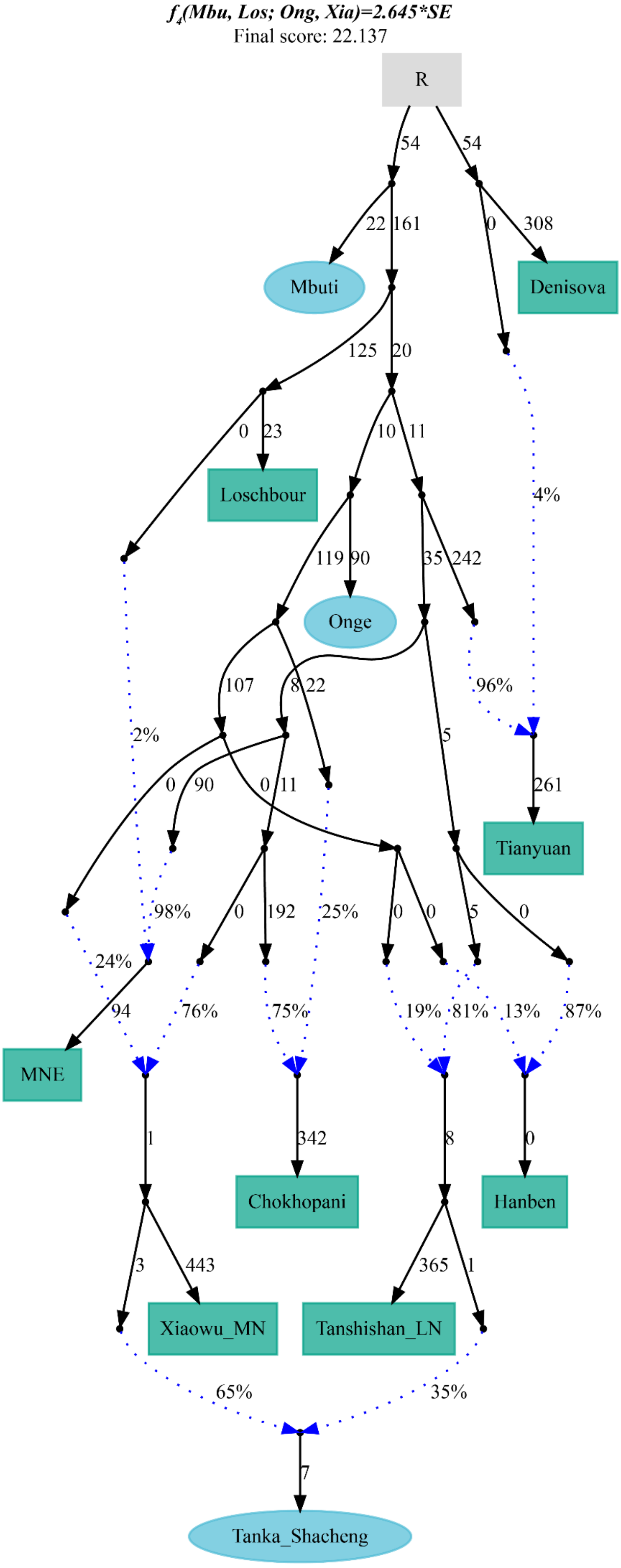
f_4(Mbu, Los; Ong, Xia)=2.645*SE
Final score: 22.137
R
54
54
22
161
0
308
Mbuti
Denisova
125
20
0
23
10
11
Loschbour
4%
119
90
35
242
Onge
107
8
22
96%
5
2%
0
90
0
11
261
Tianyuan
0
192
25%
0
0
5
0
98%
24%
94
76%
75%
19%
81%
13%
87%
MNE
1
342
8
0
Chokhopani
Hanben
3
443
365
1
Xiaowu_MN
Tanshishan_LN
65%
35%
7
Tanka_Shacheng

Figure S6. Genetic drift-based phylogenetic phylogeny showed population split and gene flow events for Shacheng Tanka. Tanka people were modeled as the admixture of two Neolithic East Asian lineages. Genetic drift was marked as 1000 times of f_2 values. Dot blue lines denoted the admixture events and corresponding admixture proportions were also marked.