群体结构与系统发生分析

G05A：李桉

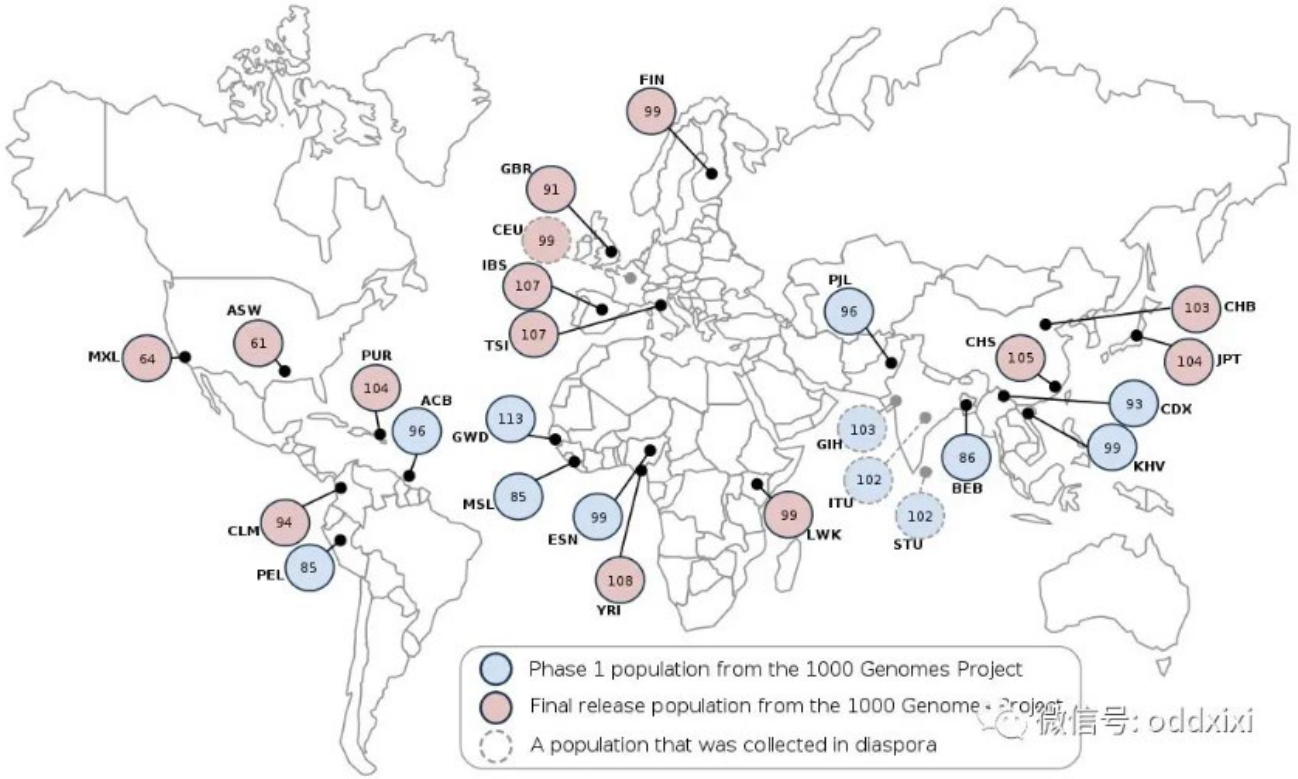
G05C：王抒扬

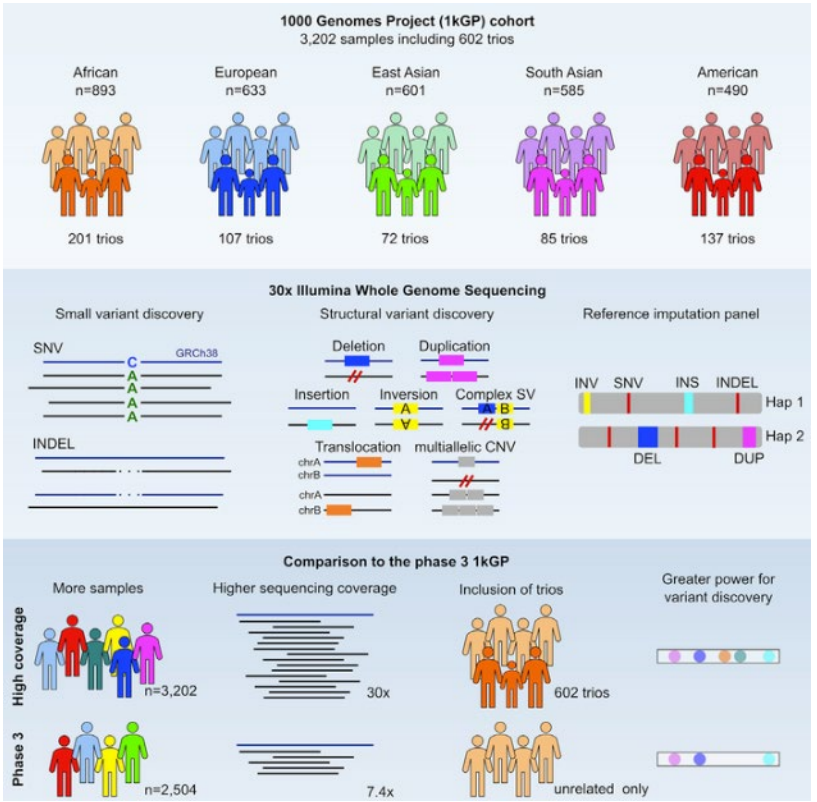
G05D：张丹

数据来源

- 千人基因组计划（1000 Genomes）第三期结构变异数据
- VCF格式

下载网址：https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/integrated_sv_map/supporting/GRCh38_positions/





VCF格式解读

- 前9列：CHROM POS ID REF ALT QUAL FILTER INFO FORMAT
- 之后每一列表示一个人的样本，包含2504个个体

```
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  HG00096 HG00097 HG00099 HG00100 HG00101 HG00102 HG00103 HG
00105 HG00106 HG00107 HG00108 HG00109 HG00110 HG00111 HG00112 HG00113 HG00114 HG00115 HG00116 HG00117 HG00118 HG00119 HG00120 HG00
121 HG00122 HG00123 HG00125 HG00126 HG00127 HG00128 HG00129 HG00130 HG00131 HG00132 HG00133 HG00136 HG00137 HG00138 HG00139 HG0014
0 HG00141 HG00142 HG00143 HG00145 HG00146 HG00148 HG00149 HG00150 HG00151 HG00154 HG00155 HG00157 HG00158 HG00159 HG00160 HG00171
HG00173 HG00174 HG00176 HG00177 HG00178 HG00179 HG00180 HG00181 HG00182 HG00183 HG00185 HG00186 HG00187 HG00188 HG00189 HG00190 HG
00231 HG00232 HG00233 HG00234 HG00235 HG00236 HG00237 HG00238 HG00239 HG00240 HG00242 HG00243 HG00244 HG00245 HG00246 HG00250 HG00
251 HG00252 HG00253 HG00254 HG00255 HG00256 HG00257 HG00258 HG00259 HG00260 HG00261 HG00262 HG00263 HG00264 HG00265 HG00266 HG0026
7 HG00268 HG00269 HG00271 HG00272 HG00273 HG00274 HG00275 HG00276 HG00277 HG00278 HG00280 HG00281 HG00282 HG00284 HG00285 HG00288
HG00290 HG00304 HG00306 HG00308 HG00309 HG00310 HG00311 HG00313 HG00315 HG00318 HG00319 HG00320 HG00321 HG00323 HG00324 HG00325 HG
00326 HG00327 HG00328 HG00329 HG00330 HG00331 HG00332 HG00334 HG00335 HG00336 HG00337 HG00338 HG00339 HG00341 HG00342 HG00343 HG00
```

- 每一行表示一个结构变异（SVTYPE表示结构变异），显示每个个体两个等位基因的种类
- 例：0|0表示这个人在这一位点两条同源染色体均为0类的等位基因

```
1       710330  ALU_umary_ALU_2 A       <INS:ME:ALU>    .       .       TSD=null;SVTYPE=ALU;MEINFO=AluYa4_5,1,223,-;SVLEN=222;CS=A
LU_umary;AC=35;AF=0.00698882;NS=2504;AN=5008;EAS_AF=0.0069;EUR_AF=0.0189;AFR_AF=0.0;AMR_AF=0.0072;SAS_AF=0.0041;SITEPOST=0.9998
        GT      0|0     0|0     0|0     0|0     0|0     0|0     0|0     0|0     0|0     0|0     0|0     0|0     0|0     0|0     0|0     0|0
     0|0     0|0     0|0     0|0     0|0     0|0     0|0     0|0     0|0     0|0     0|0     0|1     0|0     0|0     0|0     0|0
     0|0     0|1     0|0     0|0     0|0     0|0     0|0     0|0     0|0     0|0     0|0     0|0     0|0     0|0     0|0     0|0
     0|0     0|0     0|1     0|0     0|0     0|0     0|0     0|0     0|0     0|0     0|0     0|0     0|0     0|0     0|0     0|0
     1|0     0|0     0|1     0|0     0|0     0|0     0|0     0|0     0|0     0|0     0|0     0|0     0|0     0|0     0|0     0|0
     0|0     0|0     0|0     0|0     0|0     0|0     0|0     0|0     0|0     0|0     0|0     0|0     0|0     0|0     0|0     0|0
     0|0     0|0     0|1     0|0     0|0     0|0     0|0     0|0     0|0     0|0     0|0     0|0     0|0     1|0     0|0     0|0
     0|0     0|0     0|0     0|0     0|0     0|0     1|0     0|0     0|0     0|0     0|0     0|0     0|0     0|1     0|0     0|0
```

数据来源

- 群体注释数据（info.csv)

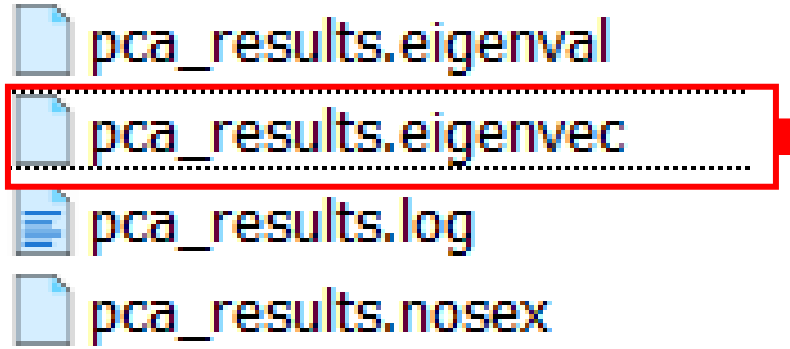
下载网址：https://www.internationalgenome.org/data-portal/sample

name	Sex	Biosample	Population_code	Population name	Superpopulation_code	Superpopulation_name	Populatio	Data collections
HG00459	male	SAME125	CHS	Southern Han Chinese	EAS	East Asian Ancestry	CHS	1000 Genomes 30x on GRCh38,1000 Geno
HG00473	female	SAME123	CHS	Southern Han Chinese	EAS	East Asian Ancestry	CHS	1000 Genomes on GRCh38,1000 Genomes
HG00478	male	SAME123	CHS	Southern Han Chinese	EAS	East Asian Ancestry	CHS	1000 Genomes on GRCh38,1000 Genomes
HG00480	male	SAME124	CHS	Southern Han Chinese	EAS	East Asian Ancestry	CHS	1000 Genomes 30x on GRCh38,1000 Geno
HG00500	male	SAME123	CHS	Southern Han Chinese	EAS	East Asian Ancestry	CHS	1000 Genomes on GRCh38,1000 Genomes
HG00512	male	SAME123	CHS	Southern Han Chinese	EAS	East Asian Ancestry	CHS	1000 Genomes on GRCh38,1000 Genomes
HG00524	male	SAME123	CHS	Southern Han Chinese	EAS	East Asian Ancestry	CHS	1000 Genomes on GRCh38,1000 Genomes
HG00531	female	SAME125	CHS	Southern Han Chinese	EAS	East Asian Ancestry	CHS	1000 Genomes on GRCh38,1000 Genomes
HG00315	female	SAME124	FIN	Finnish	EUR	European Ancestry	FIN	1000 Genomes on GRCh38,1000 Genomes
HG00327	female	SAME123	FIN	Finnish	EUR	European Ancestry	FIN	1000 Genomes on GRCh38,1000 Genomes
HG00334	female	SAME123	FIN	Finnish	EUR	European Ancestry	FIN	1000 Genomes on GRCh38,1000 Genomes
HG00339	female	SAME123	FIN	Finnish	EUR	European Ancestry	FIN	1000 Genomes on GRCh38,1000 Genomes
HG00542	male	SAME124	CHS	Southern Han Chinese	EAS	East Asian Ancestry	CHS	1000 Genomes on GRCh38,1000 Genomes
HG00341	male	SAME124	FIN	Finnish	EUR	European Ancestry	FIN	1000 Genomes on GRCh38,1000 Genomes
HG00554	female	SAME123	PUR	Puerto Rican	AMR	American Ancestry	PUR	1000 Genomes on GRCh38,1000 Genomes
HG00346	female	SAME125	FIN	Finnish	EUR	European Ancestry	FIN	1000 Genomes on GRCh38,1000 Genomes

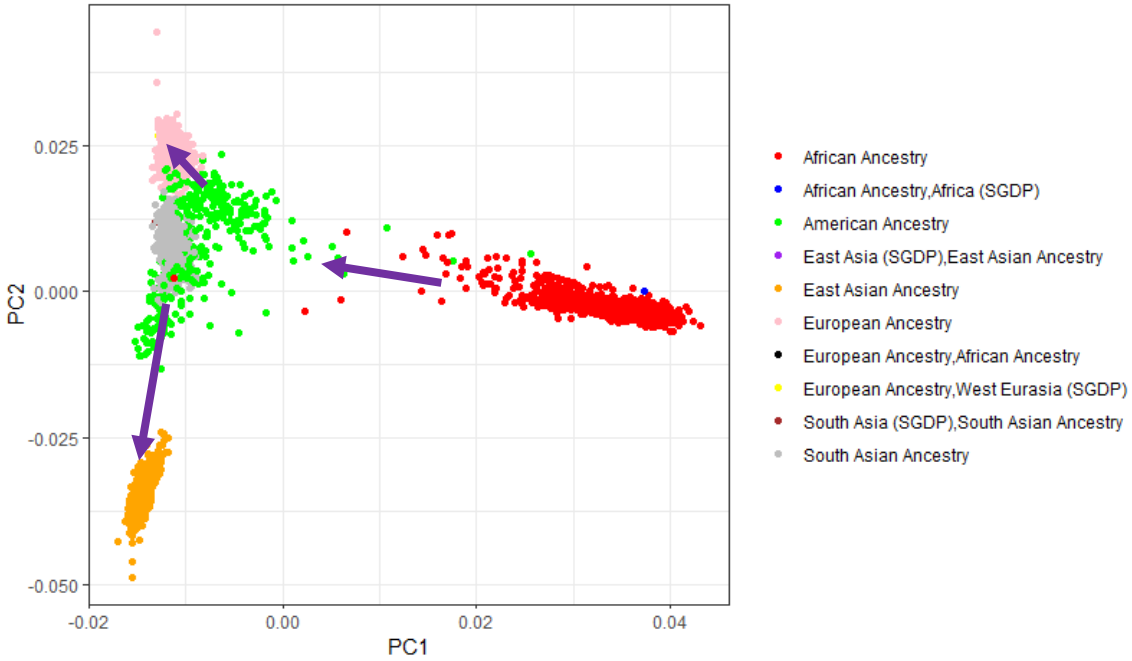
- 第一列为个体编号， Population name，Superpopulation_name为所属的群体和超群体（例如HG00459个体属于东亚，汉族）；
- Population_code和Superpopulation_code为对应群体的缩写

主成分分析PCA

- 代码：./plink --vcf g.vcf --pca --out pca_results
- 输出



特征向量文件用于可视化



- 非洲群体在PC1与其它群体分离，
- 美洲、南亚、欧洲群体在空间上接近，
- 东亚群体在PC2上与其它群体分离。

群体结构分析（Linux）

- vcf格式转换为plink格式

```
./plink --vcf g.vcf -recode12 --out admixture/plink_result
```

- 根据-hwe 0.0001进行质控，去除稀有单倍型

```
./plink --noweb --file ./admixture/plink_result -hwe 0.0001 --make-bed --out admixture/QC
```

- 用多个K值进行admixture群体结构分析(K表示假设的祖先单倍型个数)

```
for K in 2 3 4 5 6
do
/rd1/home/leb27/final/dist/admixture_linux-1.3.0/admixture --cv ./admixture/QC.bed $K -j16 > admixture_k${K}.out
done
```

群体结构分析（Linux）

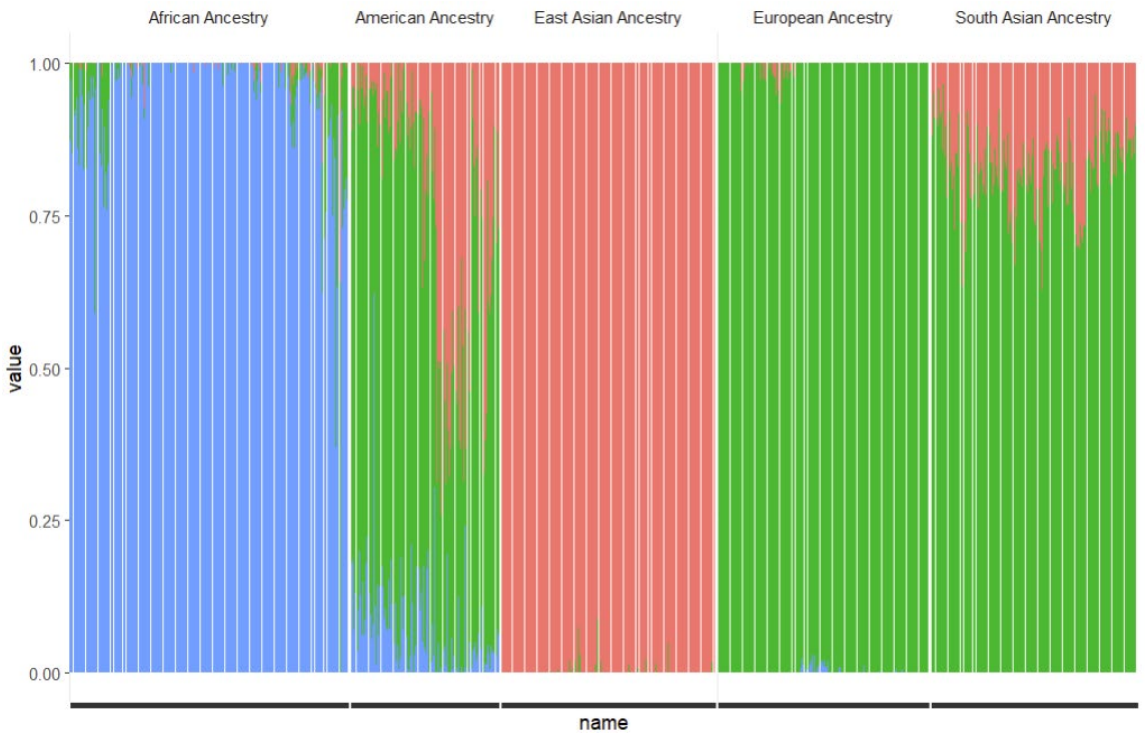
- 对每一个K，会生成三个文件，后缀分别为.out、.P和.Q

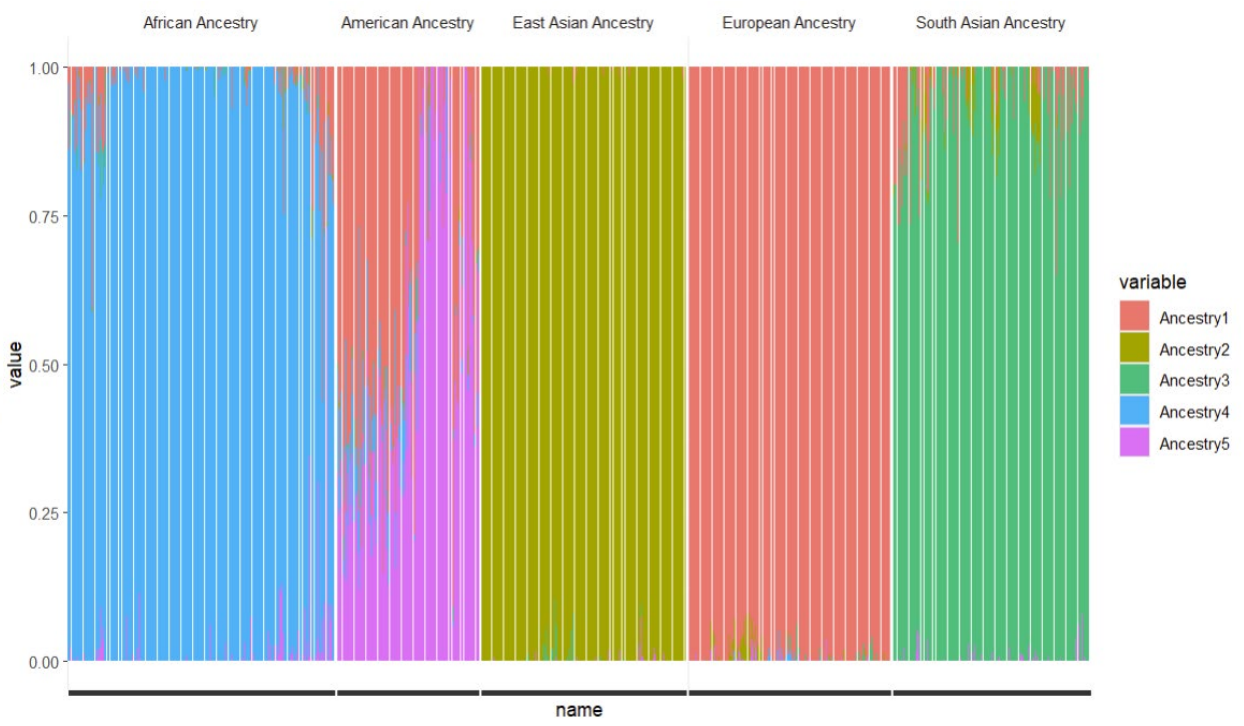
- .Q文件用于可视化，有k列，表示k个祖先单倍型，每一行为一个个体，顺序与g.vcf一致。.Q文件整理为admixture_k5.csv

- K=2、3、4、5、6的CV error（交叉验证错误率）分别为0.04509、0.04365、0.04314、0.04293、0.04327（均较低）

name	Ancestry1	Ancestry2	Ancestry3	Ancestry4	Ancestry5
HG00096	0.981228	0.00001	0.00001	0.00001	0.018742
HG00097	0.99996	0.00001	0.00001	0.00001	0.00001
HG00099	0.99996	0.00001	0.00001	0.00001	0.00001
HG00100	0.99996	0.00001	0.00001	0.00001	0.00001
HG00101	0.99996	0.00001	0.00001	0.00001	0.00001
HG00102	0.99996	0.00001	0.00001	0.00001	0.00001
HG00103	0.99996	0.00001	0.00001	0.00001	0.00001
HG00105	0.99996	0.00001	0.00001	0.00001	0.00001
HG00106	0.993064	0.000013	0.00001	0.00001	0.006903
HG00107	0.99996	0.00001	0.00001	0.00001	0.00001
HG00108	0.99996	0.00001	0.00001	0.00001	0.00001
HG00109	0.99996	0.00001	0.00001	0.00001	0.00001
HG00110	0.99996	0.00001	0.00001	0.00001	0.00001
HG00111	0.99996	0.00001	0.00001	0.00001	0.00001
HG00112	0.99996	0.00001	0.00001	0.00001	0.00001
HG00113	0.99996	0.00001	0.00001	0.00001	0.00001

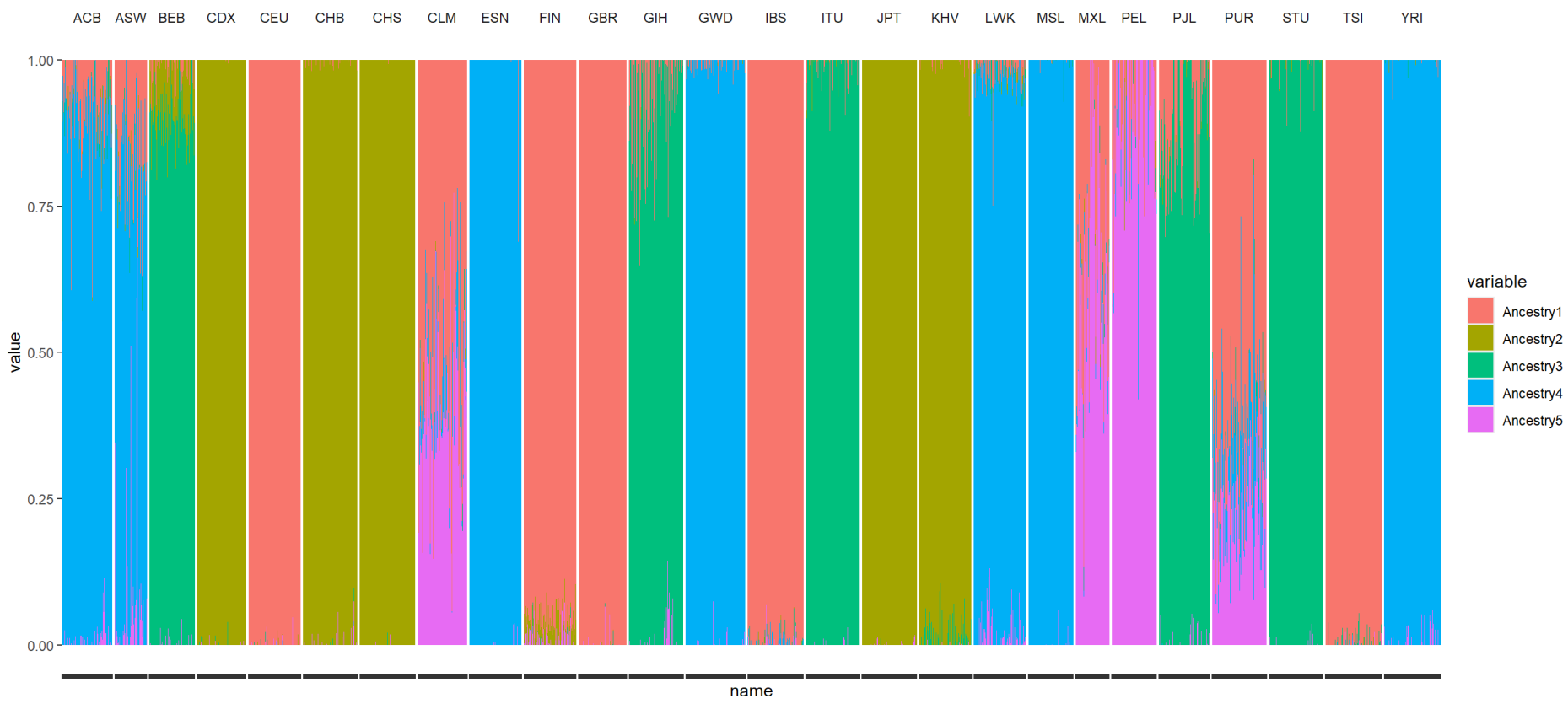
群体结构分析可视化（R）





- 非洲群体为主要由祖先单倍型3组成，同时包含部分祖先单倍型1和2的多态性，美洲群体多态性最高
- 东亚群体和欧洲群体多态性较单一，分别仅包含祖先单倍型1和2，暗示其起源较晚
- K=5时可以更清晰的看出东亚、南亚、欧洲单倍型存在显著差异，不同的K值有时可以提供不同的理解

群体结构分析可视化（R）



按population code进行划分

构建系统发生树

- 系统发生树构建方法主要包括2大类，基于距离的方法（如UPGMA、NJ法，需要获得距离矩阵）、基于性状的方法（如最大似然法、贝叶斯法，需要序列文件，如fasta格式）。

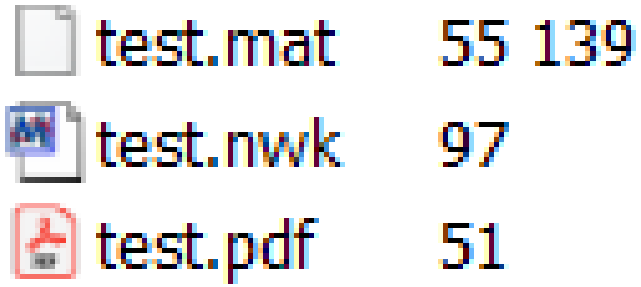
近邻相接法（NJ法）通过最小化进化树的总分支长度构建拓扑结构，其核心步骤如下：

- 根据DNA序列差异计算两两物种间的遗传距离
- 通过迭代合并最近邻节点形成树状结构
- 每次合并后重新计算剩余节点间的校正距离

该方法采用枝长优化策略，通过星状分解法逐步完善树形结构，尤其适用于分枝长度差异显著的进化关系分析。

用VCF2Dis生成距离矩阵

- 这里由于使用结构变异数据，无法生成序列文件，故使用VCF2Dis生成距离矩阵
- 软件包：VCF2Dis-1.54.zip
- 服务器中位置：/rd1/home/leb27/final/VCF2Dis-1.54/bin/VCF2Dis -i g.vcf -o test.mat

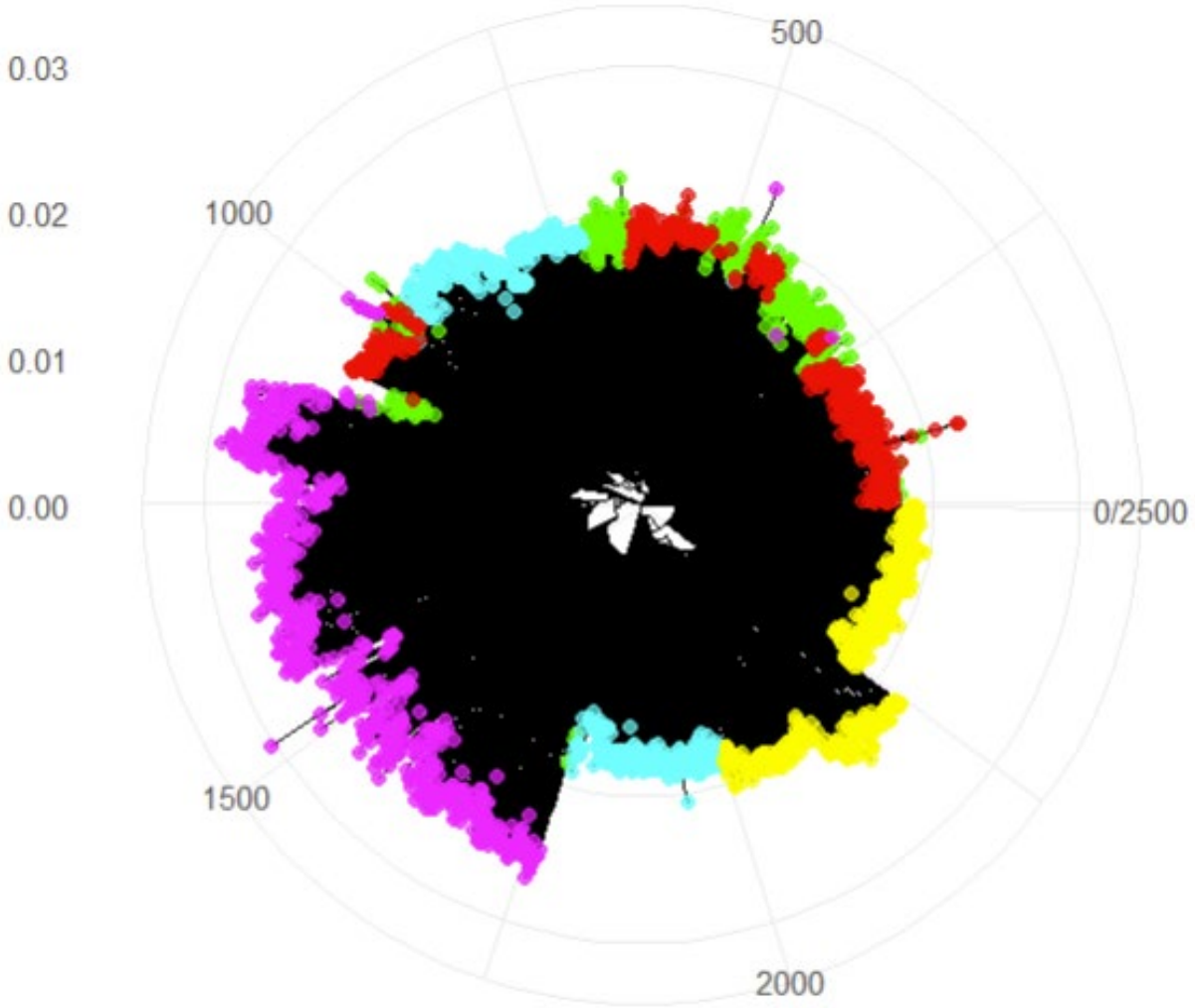


	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	HG00096	0	0.032454	0.032309	0.034157	0.031195	0.032818	0.030518	0.031835	0.033182	0.034077	0.033167	0.031639	0.032469	0.031894	0.031341	0.034296	0.032076
2	HG00097	0.032454	0	0.032076	0.033502	0.033575	0.032141	0.035504	0.034215	0.031952	0.037039	0.03693	0.035387	0.032047	0.031108	0.035227	0.036879	0.035664
3	HG00099	0.032309	0.032076	0	0.033073	0.033852	0.032207	0.034383	0.03383	0.03193	0.036268	0.037141	0.034958	0.031937	0.032309	0.035693	0.037069	0.035293
4	HG00100	0.034157	0.033502	0.033073	0	0.035154	0.032884	0.036355	0.035664	0.032629	0.038066	0.038131	0.036144	0.032527	0.032534	0.037323	0.038626	0.036828
5	HG00101	0.031195	0.033575	0.033852	0.035154	0	0.033662	0.033735	0.033102	0.033881	0.035009	0.035373	0.033269	0.034201	0.033247	0.033735	0.036312	0.034055
6	HG00102	0.032818	0.032141	0.032207	0.032884	0.033662	0	0.035067	0.034019	0.031872	0.036748	0.037127	0.034754	0.032265	0.031734	0.035657	0.036974	0.03586
7	HG00103	0.030518	0.035504	0.034383	0.036355	0.033735	0.035067	0	0.033873	0.034856	0.032178	0.033211	0.03431	0.034732	0.034128	0.031624	0.034012	0.031115
8	HG00105	0.031835	0.034215	0.03383	0.035664	0.033102	0.034019	0.033873	0	0.03383	0.035635	0.036071	0.034186	0.033917	0.033218	0.03463	0.036595	0.034659
9	HG00106	0.033182	0.031952	0.03193	0.032629	0.033881	0.031872	0.034856	0.03383	0	0.036479	0.036472	0.034368	0.032432	0.031071	0.035999	0.037447	0.035708
10	HG00107	0.034077	0.037039	0.036268	0.038066	0.035009	0.036748	0.032178	0.035635	0.036479	0	0.035329	0.035278	0.036734	0.036632	0.03348	0.035314	0.033531
11	HG00108	0.033167	0.03693	0.037141	0.038131	0.035373	0.037127	0.033211	0.036071	0.036472	0.035329	0	0.035649	0.036472	0.036173	0.034012	0.035169	0.034092
12	HG00109	0.031639	0.035387	0.034958	0.036144	0.033269	0.034754	0.03431	0.034186	0.034368	0.035278	0.035649	0	0.034427	0.034587	0.035205	0.037185	0.034725
13	HG00110	0.032469	0.032047	0.031937	0.032527	0.034201	0.032265	0.034732	0.033917	0.032432	0.036734	0.036472	0.034427	0	0.031049	0.035518	0.037265	0.035671
14	HG00111	0.031894	0.031108	0.032309	0.032534	0.033247	0.031734	0.034128	0.033218	0.031071	0.036632	0.036173	0.034587	0.031049	0	0.035118	0.036697	0.035074
15	HG00112	0.031341	0.035227	0.035693	0.037323	0.033735	0.035657	0.031624	0.03463	0.035999	0.03348	0.034012	0.035205	0.035518	0.035118	0	0.034412	0.031661

- 生成3个文件，test.mat为距离矩阵，可用于在R中构建邻接树，并整理为distance.csv，test.nwk为树文件，可在mega或iTOL（iTOL: Interactive Tree Of Life）进行美化，pdf为直接生成的树文件。

系统发生树可视化（R）

- ape、ggtree包



- 非洲群体、东亚群体构成单系群
- 欧洲群体、美洲群体构成多系群，可能由于群体近期存在基因交流，或存在多次群体的扩散事件。
- 南亚群体在系统树构成两个分支，暗示其可能由两个不同来源的亚群体组成。
- 此外，距离法在处理复杂的进化历史时效果较差，对于SNP数据，可以用vcf2phylip转换为phylip或fasta格式，用最大似然法（如IQtree）进行系统发生树构建。且仅根据结构变异数据对演化历史进行估计会存在偏差

谢谢！