China National GeneBank 国家基因库

文 献 阅 读

SCIENTIFIC RESEARCH LITERATURE ANALYSIS

报告人 ：张统庆、潘云瑞、张赛智

11/4/2021

【主题及背景】

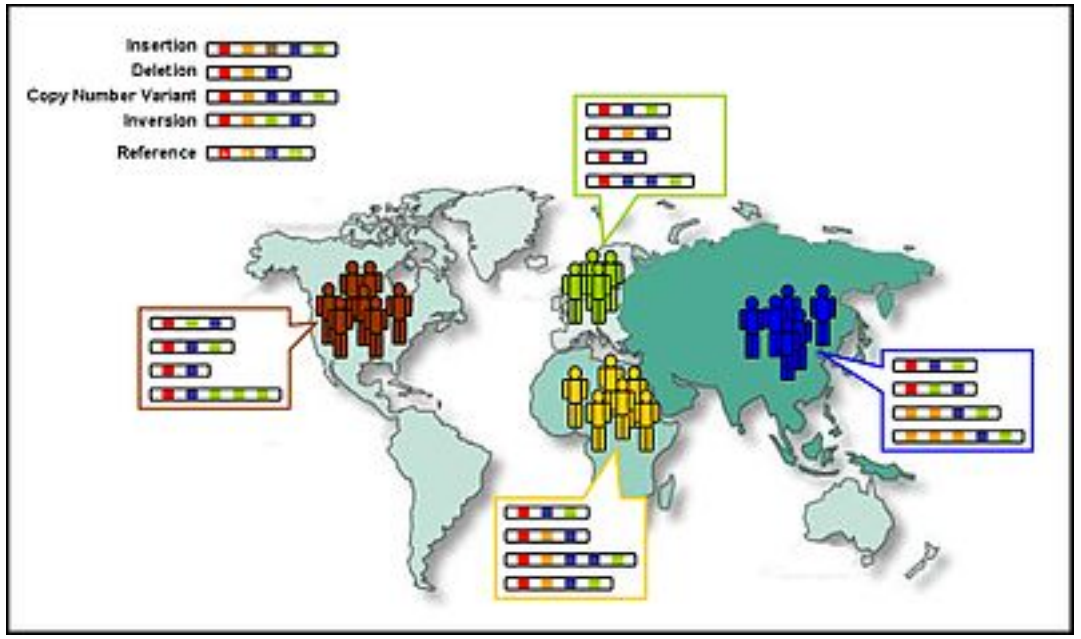
千人基因组计划（1000 Genomes Project）

2008.1 启动

2010（nature) A map of human genome variation from population-scale sequencing

2012（nature) An integrated map of genetic variation from 1092 human genomes

2015（nature) A global reference for human genetic variation



https://en.wikipedia.org/wiki/1000_Genomes_Project

A map of human genome variation from population-scale sequencing

The 1000 Genomes Project Consortium*

【 内容摘要 】

项目：1.对四个种群的179个个体进行低覆盖率全基因组测序

2.对两个母-父-子三人组进行高覆盖率测序

3.对七个种群的697个个体进行外显子靶向测序

结果：1500万SNPs, 100万indels，20,000个结构变异

变异的分布和分析

2010 Nature doi:10.1038/nature09534

【 研究背景和目的 】

现状：

已有的人类DNA序列变异推动了连锁不平衡（LD）和全基因组关联研究（GWAS）中致病基因的发现

当前研究的不足：

低频变异和罕见变异（此处分别定义为0.5%至5%MAF和低于0.5%MAF）的数量远远超过普通变异，并且对疾病的遗传结构也有显著影响，但尚未能够对其进行系统研究；充分理解常见和低频变异在人类表型变异中的作用需要一个更完整的人类DNA变异目录

本实验结果和目的：

开发和比较利用高通量平台进行全基因组测序的不同策略；三个项目

【 工作流程】

三个项目的共同流程：

（1）发现

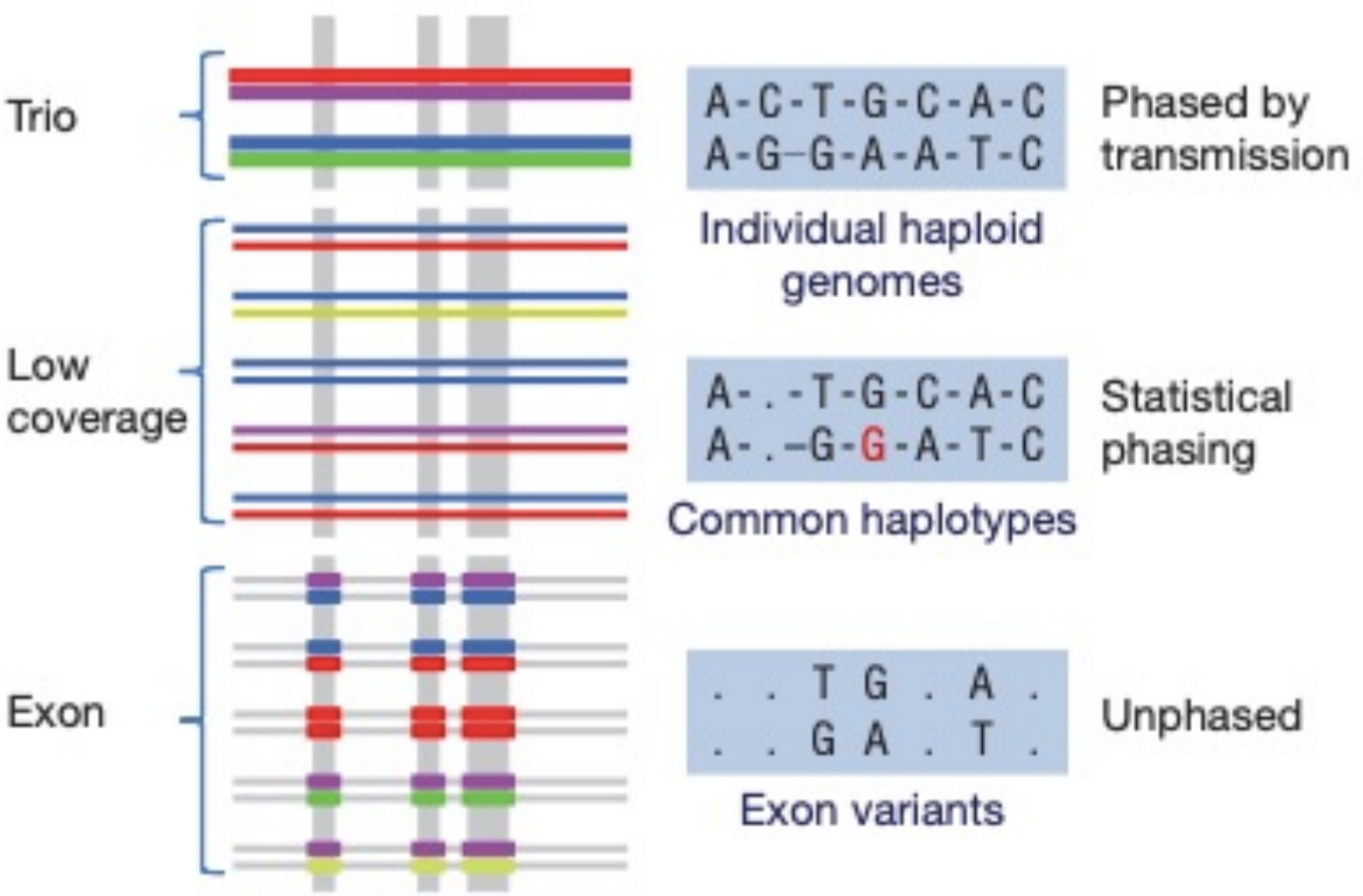
（2）过滤

（3）基因分型

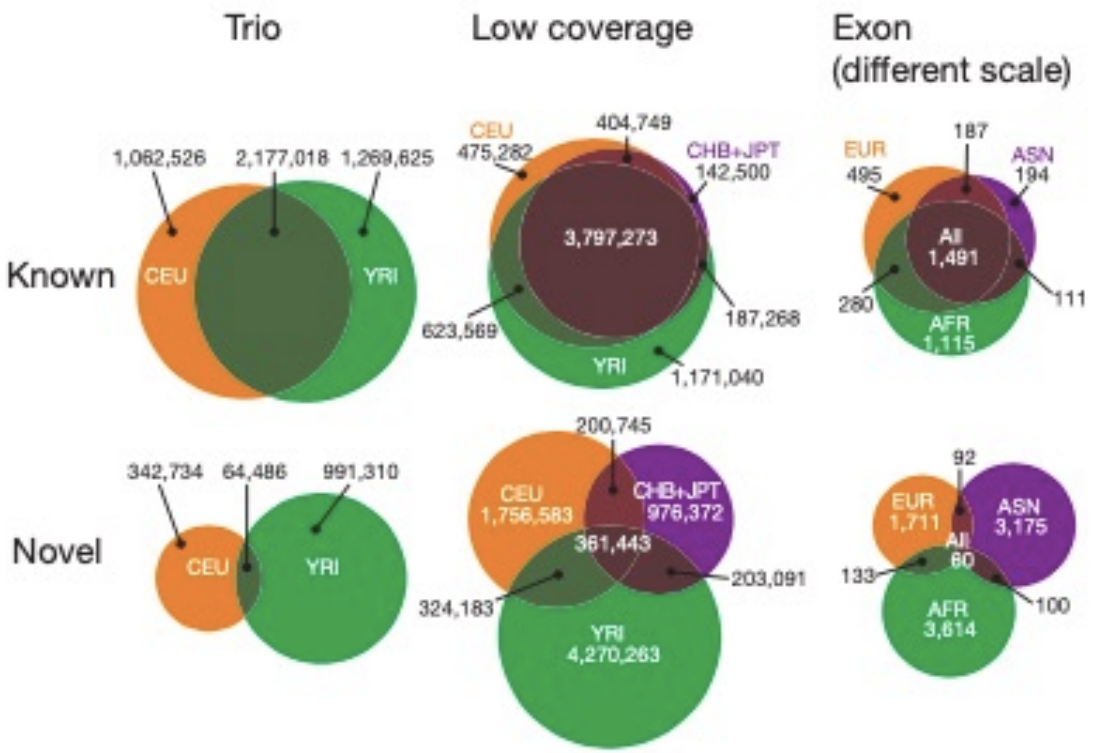
（4）验证

三个项目在设计上的不同点：

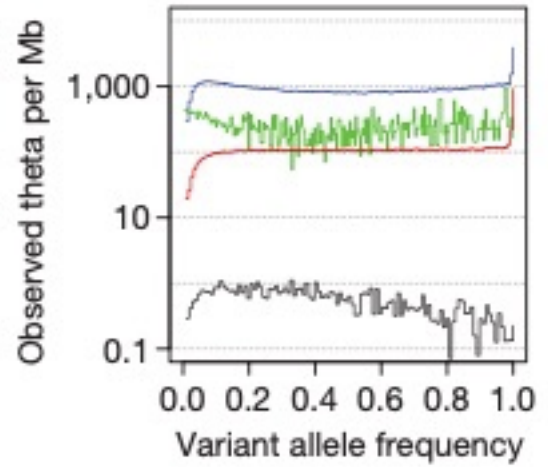
见右图



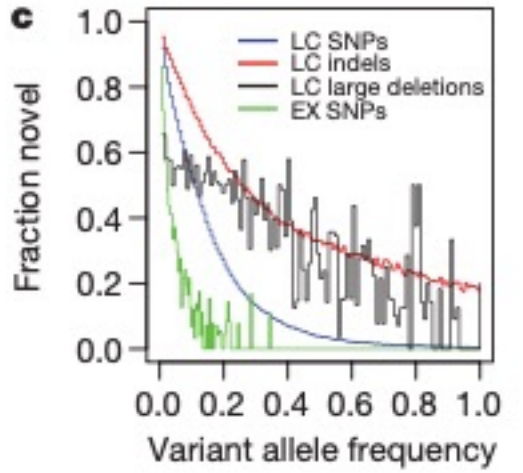
【 结果和发现】



新发现的SNPs有一种强烈的趋势，即只在一个分析小组中发现



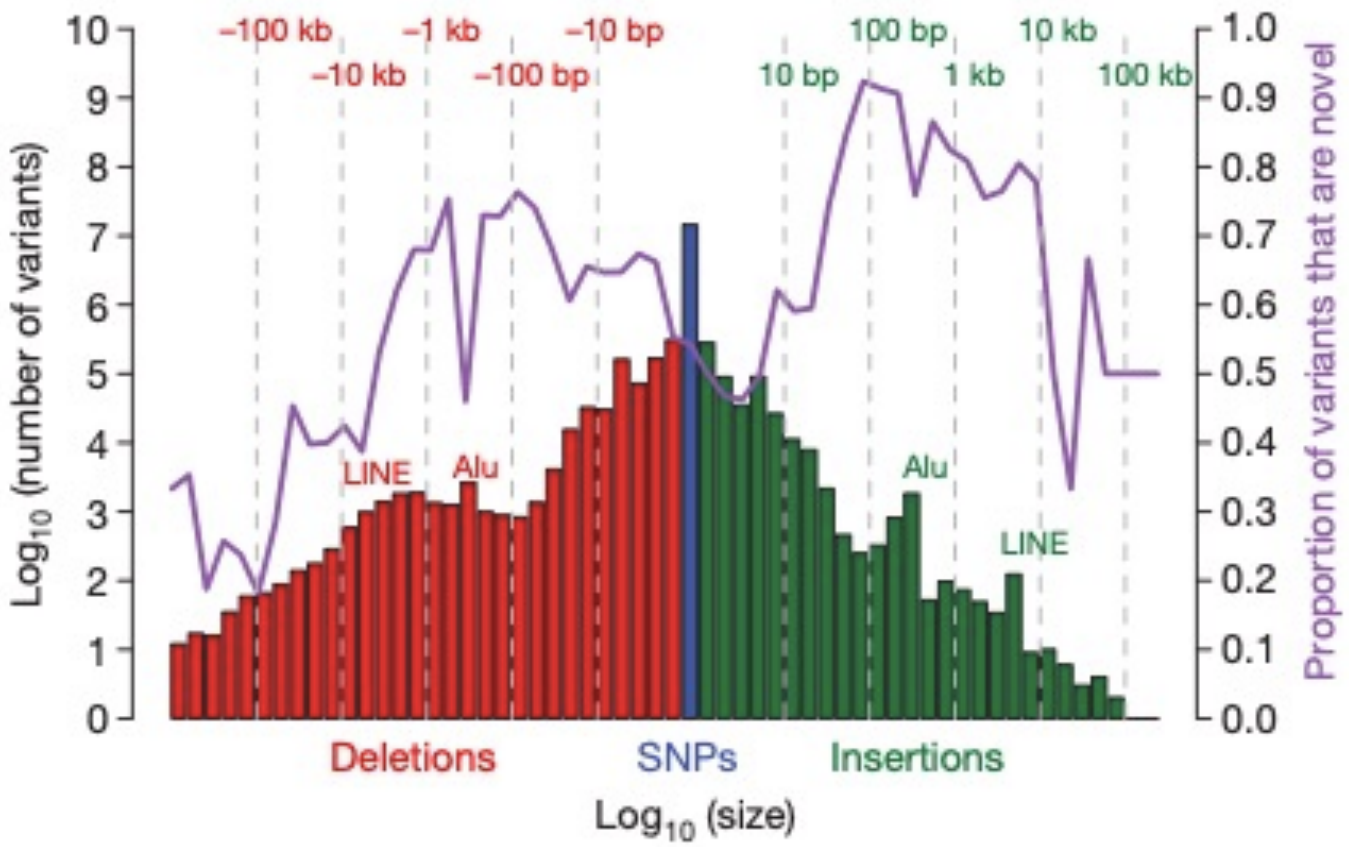
外显子和低覆盖率项目中较大的样本量使我们能够检测到大量的低频变异



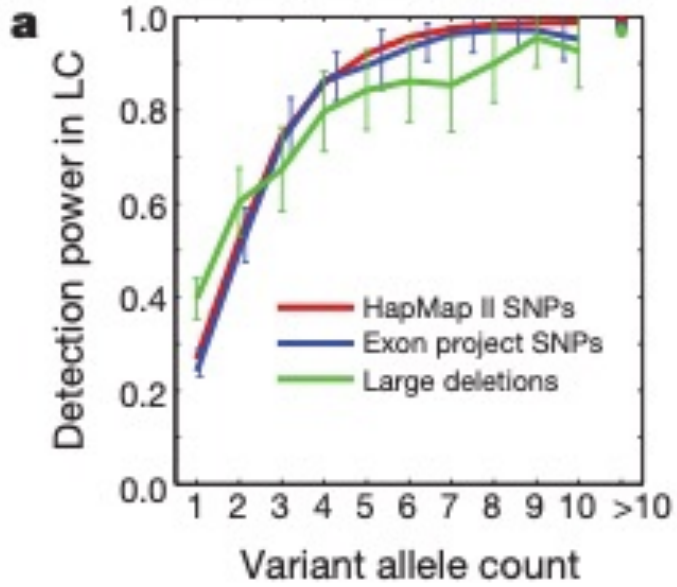
发现的高频变异也大多数收录在了原有的数据库（dbSNP）中

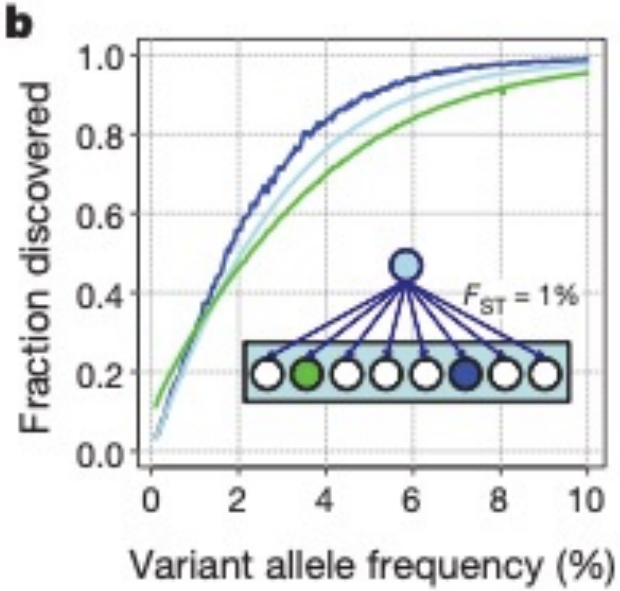
【 结果和发现】

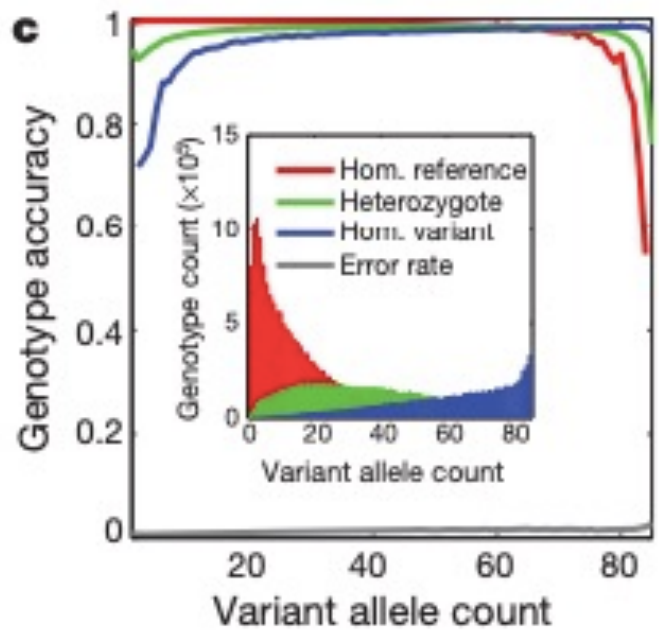
对于低频的SNPs、indels和结构性变异，公共数据库的完整性很差

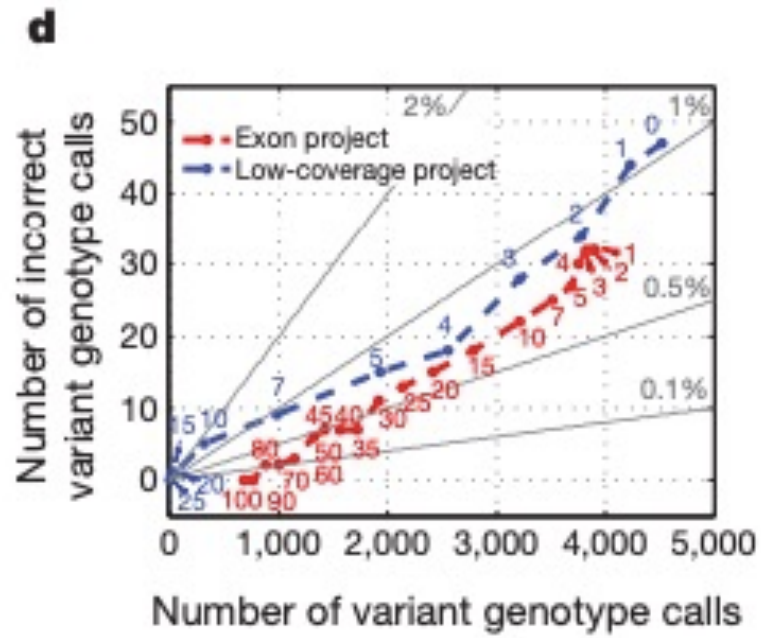


【 结果和发现】


a
Detection power in LC
1.0
0.8
0.6
0.4
0.2
0.0
1 2 3 4 5 6 7 8 9 10 >10
Variant allele count
HapMap II SNPs
Exon project SNPs
Large deletions


b
Fraction discovered
1.0
0.8
0.6
0.4
0.2
0.0
0 2 4 6 8 10
Variant allele frequency (%)
F_{ST} = 1%

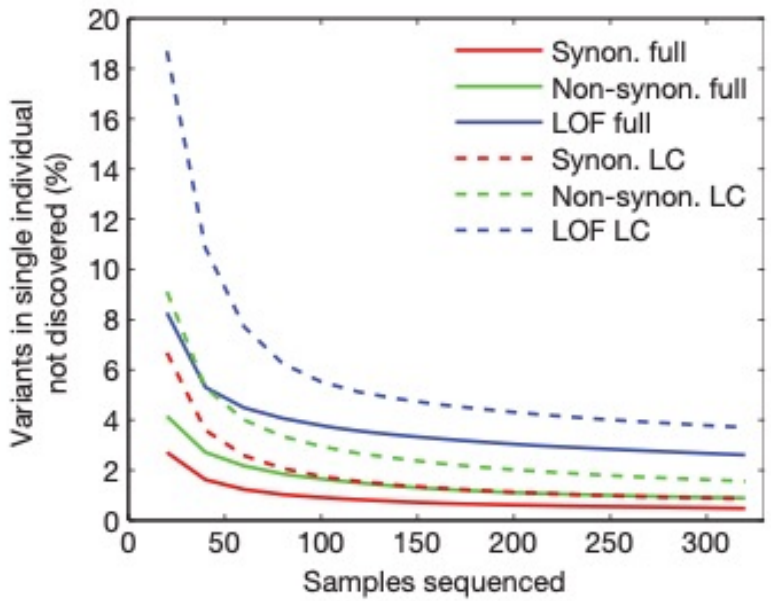

c
Genotype accuracy
1
0.8
0.6
0.4
0.2
0
20 40 60 80
Variant allele count
Genotype count (×10⁶)
15
10
5
0
0 20 40 60 80
Variant allele count
Hom. reference
Heterozygote
Hom. variant
Error rate

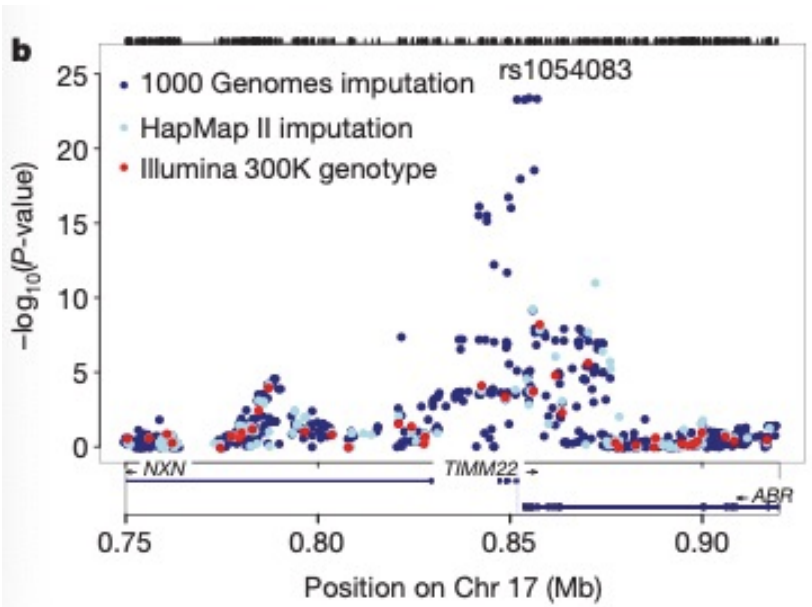

d
Number of incorrect variant genotype calls
50
40
30
20
10
0
0 1,000 2,000 3,000 4,000 5,000
Number of variant genotype calls
Exon project
Low-coverage project
2%
1%
0.5%
0.1%

【 结果和发现】

Table 2 | Estimated numbers of potentially functional variants in genes

Class	Combined total	Combined novel	Low coverage		High-coverage trio		Exon capture		
			Total	Interquartile*	Total	Individual range	Total	Interquartile*	GENCODE extrapolation
Synonymous SNPs	60,157	23,498	55,217	10,572–12,126	21,410	9,193–12,500	5,708	461–532	11,553–13,333
Non-synonymous SNPs	68,300	34,161	61,284	9,966–10,819	19,824	8,299–10,866	7,063	396–441	9,924–11,052
Small in-frame indels	714	383	666	198–205	289	130–178	59	1–3	~25–75
Stop losses	77	40	71	9–11	22	4–14	6	0–0	~0–0
Stop-introducing SNPs	1,057	755	951	88–101	192	67–100	82	2–3	~50–75
Splice-site-disrupting SNPs	517	399	500	41–49	82	28–45	3	1–1	~50
Small frameshift indels	954	551	890	227–242	433	192–280	37	0–1	~0–25
Genes disrupted by large deletions	147	71	143	28–36	82	33–49	ND	ND	ND
Total genes containing LOF variants	2,304	NA	1,795	272–297	483	240–345	77	3–4	~75–100
HGMD 'damaging mutation' SNPs	671	NA	578	57–80	161	48–82	99	2–4	~50–100





功能分析与疾病关系研究

An integrated map of genetic variation from 1,092 human genomes

The 1000 Genomes Project Consortium*

【 内容摘要 】

描述了来自14个种群的1,092个个体的基因组结果

3800万个单核苷酸多态性、140万个短的插入和缺失，以及14000多个较大的缺失

表明了来自不同种群的个体携带不同的罕见和常见变异体，低频变异体显示出巨大的地理差异

在相关人群中捕获了高达98%的频率为1%的SNPs，能够分析来自不同人群（包括混血人群）的常见和低频率变异

【 研究背景 】

现状：

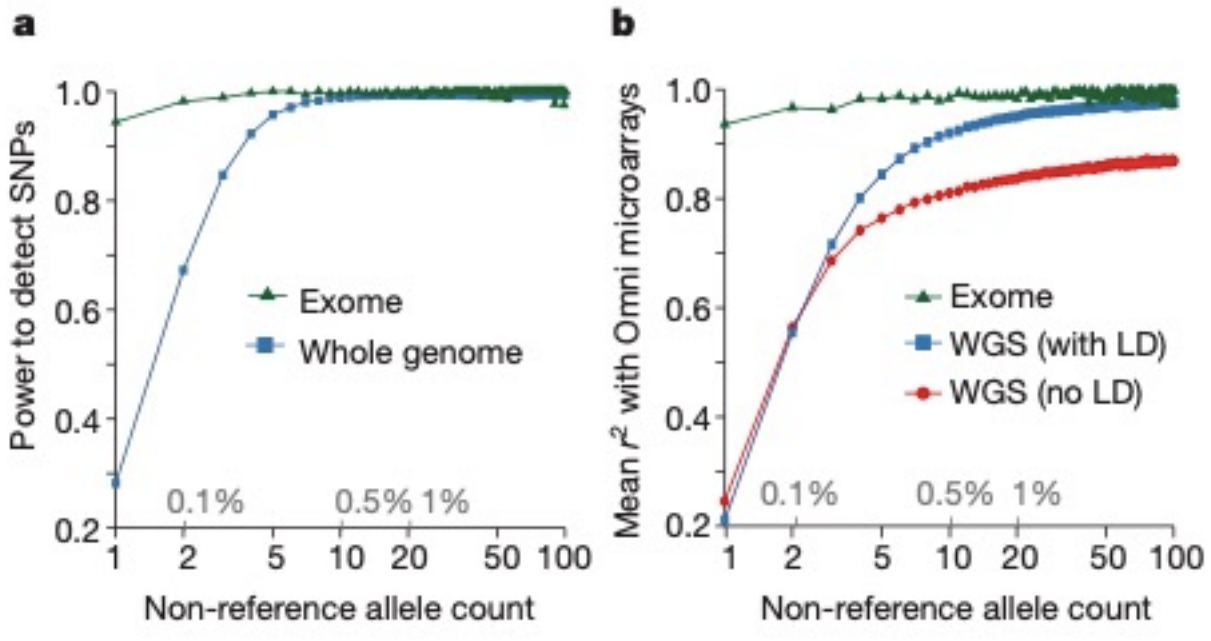
在千人基因组的试验阶段发现了了95%以上的常见（0.5%频率）变异

当前研究的不足：

低频率变体，特别是编码外显子组以外的变体，仍然没有得到很好的描述

低频变异有潜在的功能性突变，因此具有研究意义

由于低频变体往往是最近产生的，它们表现出更高的群体分化水平



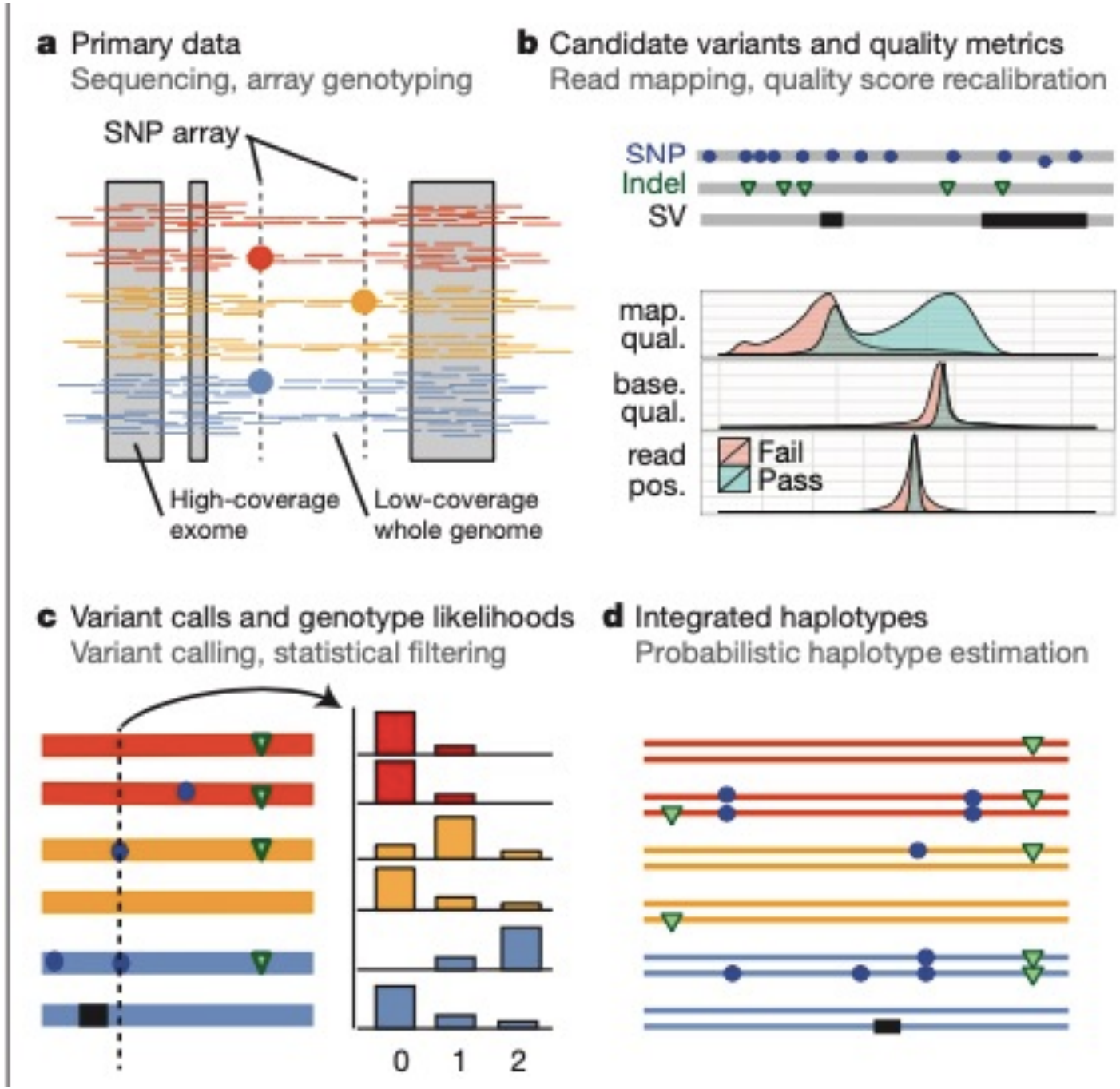
【 1092基因组map的构建 】

a. 获取数据

b. 筛选变异

c. 使用表型可能性过滤

d. 整合统计单倍型（使用LD等信息）

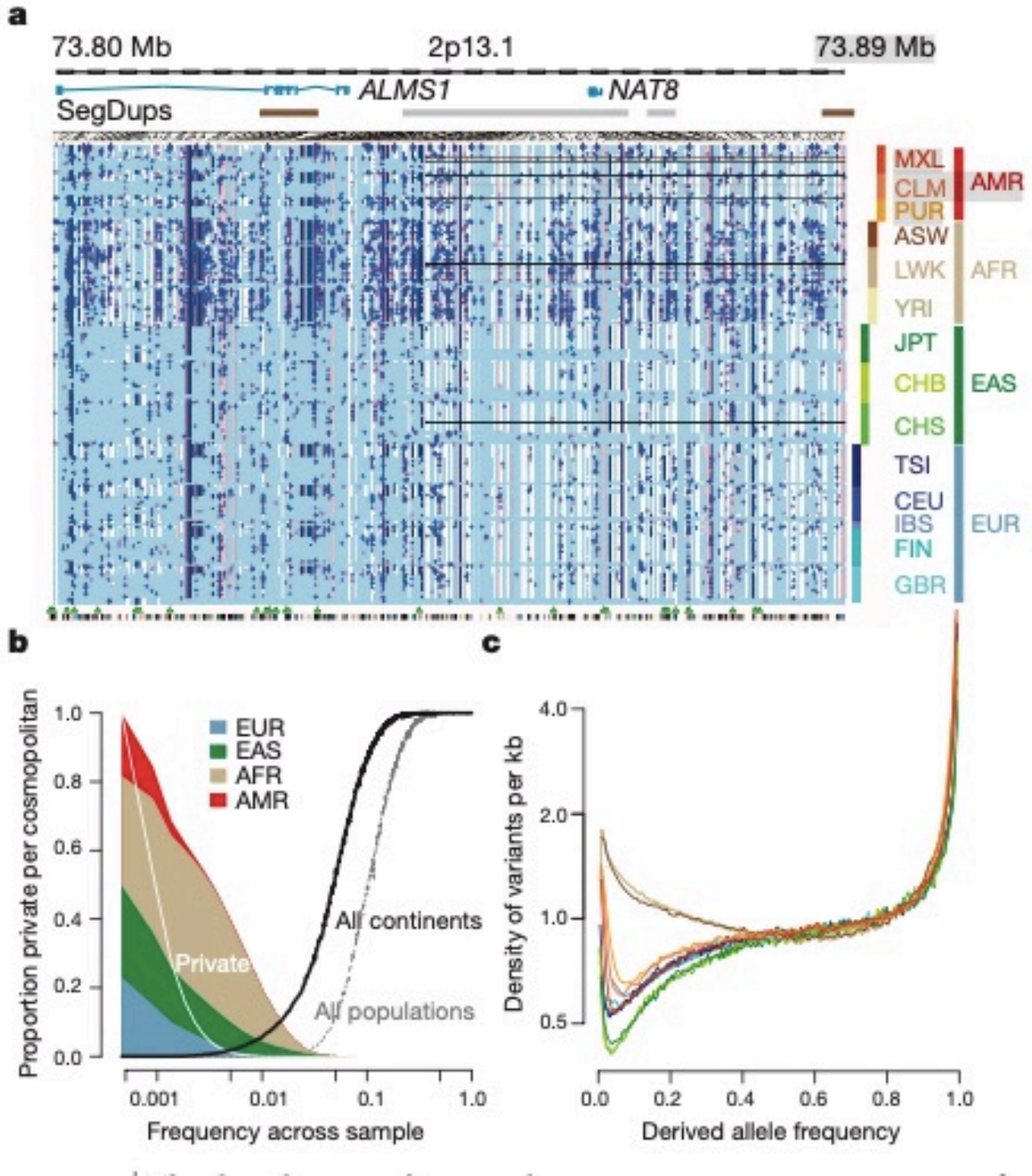


【种群内部和之间的遗传变异】

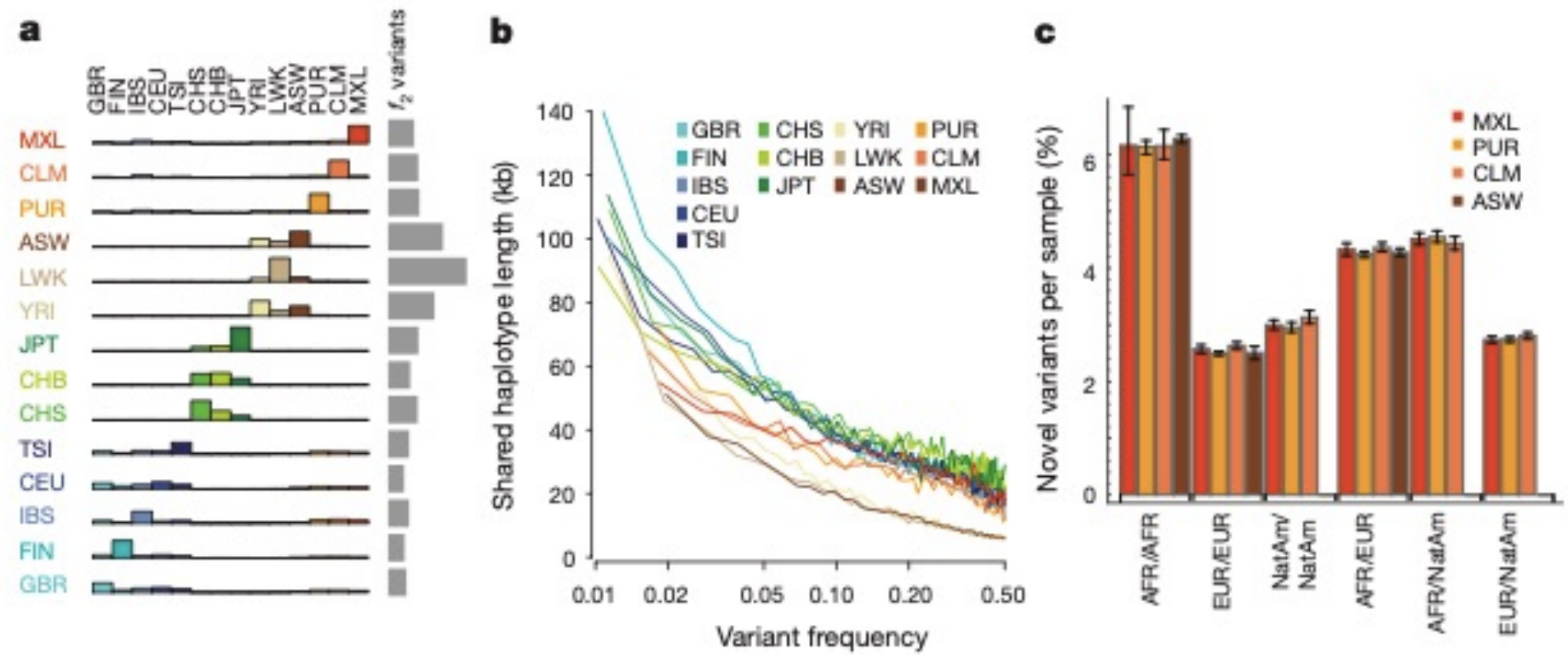
a. 2号染色体上横跨ALMS1和NAT8基因的100kb区域的单倍型概况

b. 各人群中不同频率变异的组成分布

c. 衍生等位基因频率的密度分布



【种群内部和之间的遗传变异】

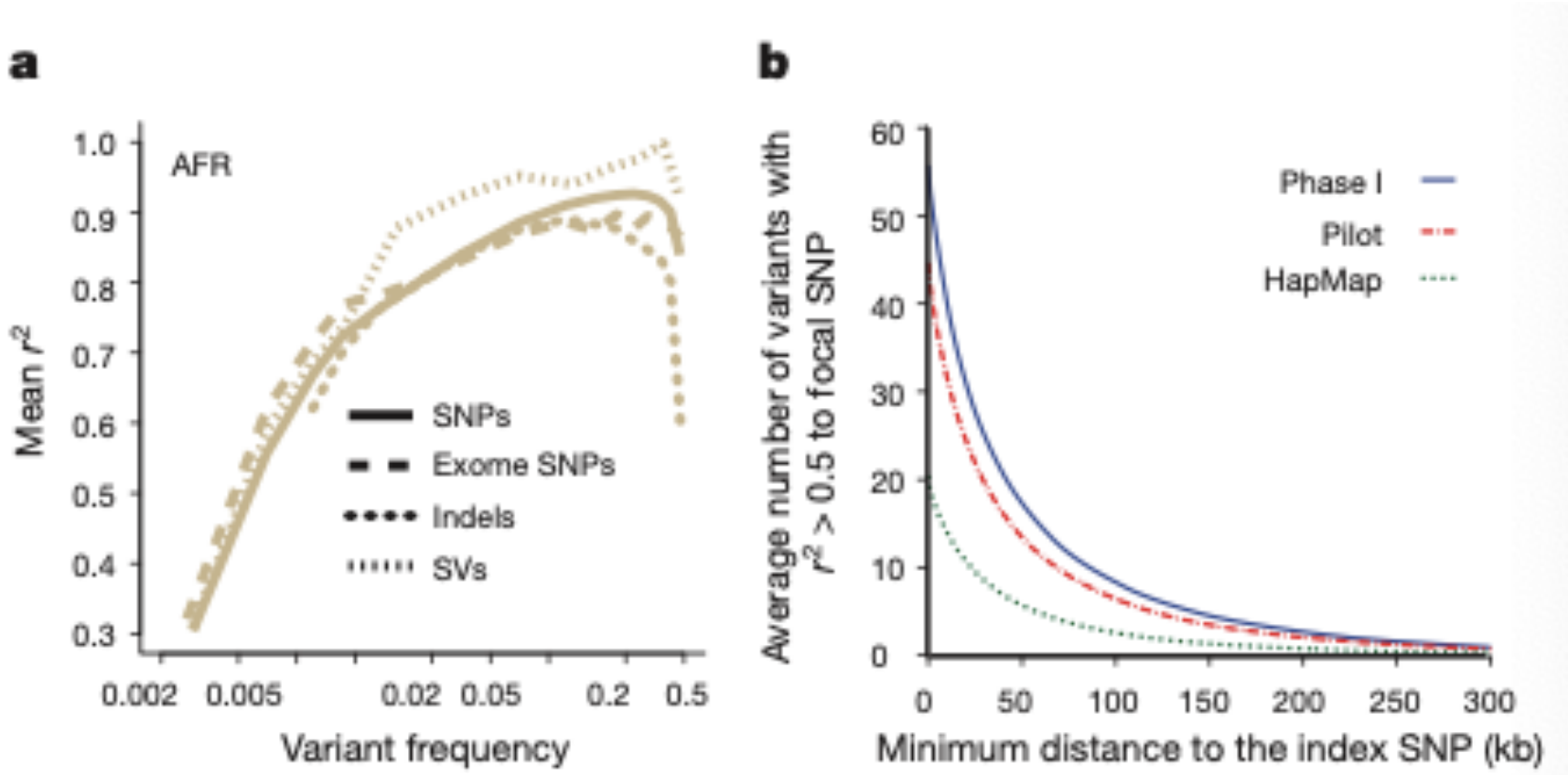


a. f2变异在各人群中的分布

b. 各频率的变体在不同人群中共享的长度

c. 各人群平均每个个体新变异的数量

【千人基因组计划数据在医学遗传学中的应用】



Imputation for GWAS

A global reference for human genetic variation

The 1000 Genomes Project Consortium*

【 内容摘要 】

利用低覆盖率全基因组测序、深度外显子组测序和密集的微阵列基因分型的组合，重建了来自26个种群的2504人的基因组

8800万个变体（8470万SNPs，360万个indels，以及6万个结构变异）

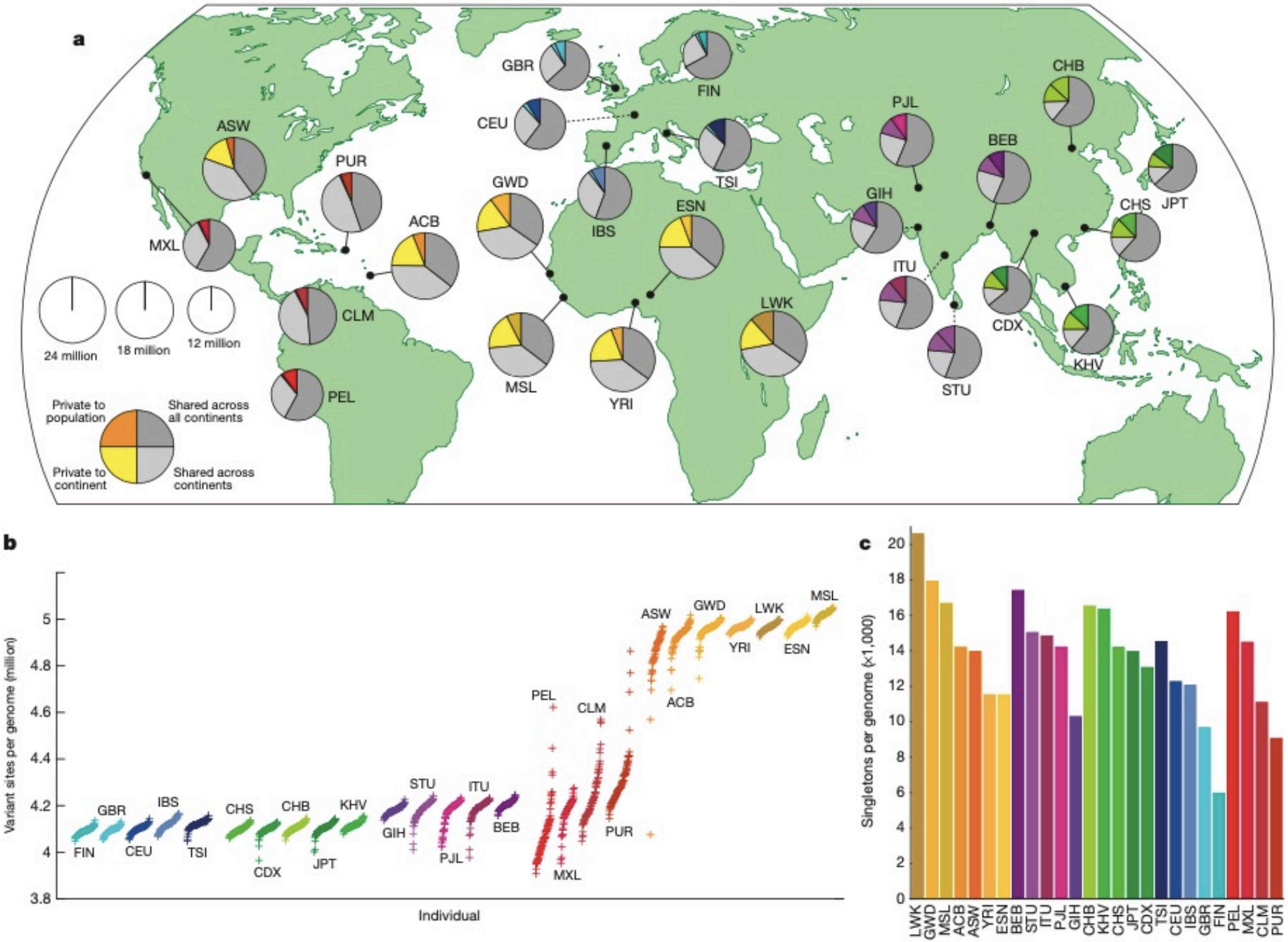
描述了全球样本的遗传变异的分布，并讨论了对常见疾病研究的影响。

【构建数据集】

a.样本来源

b.人群差异

c.每个基因组的平均单体数量

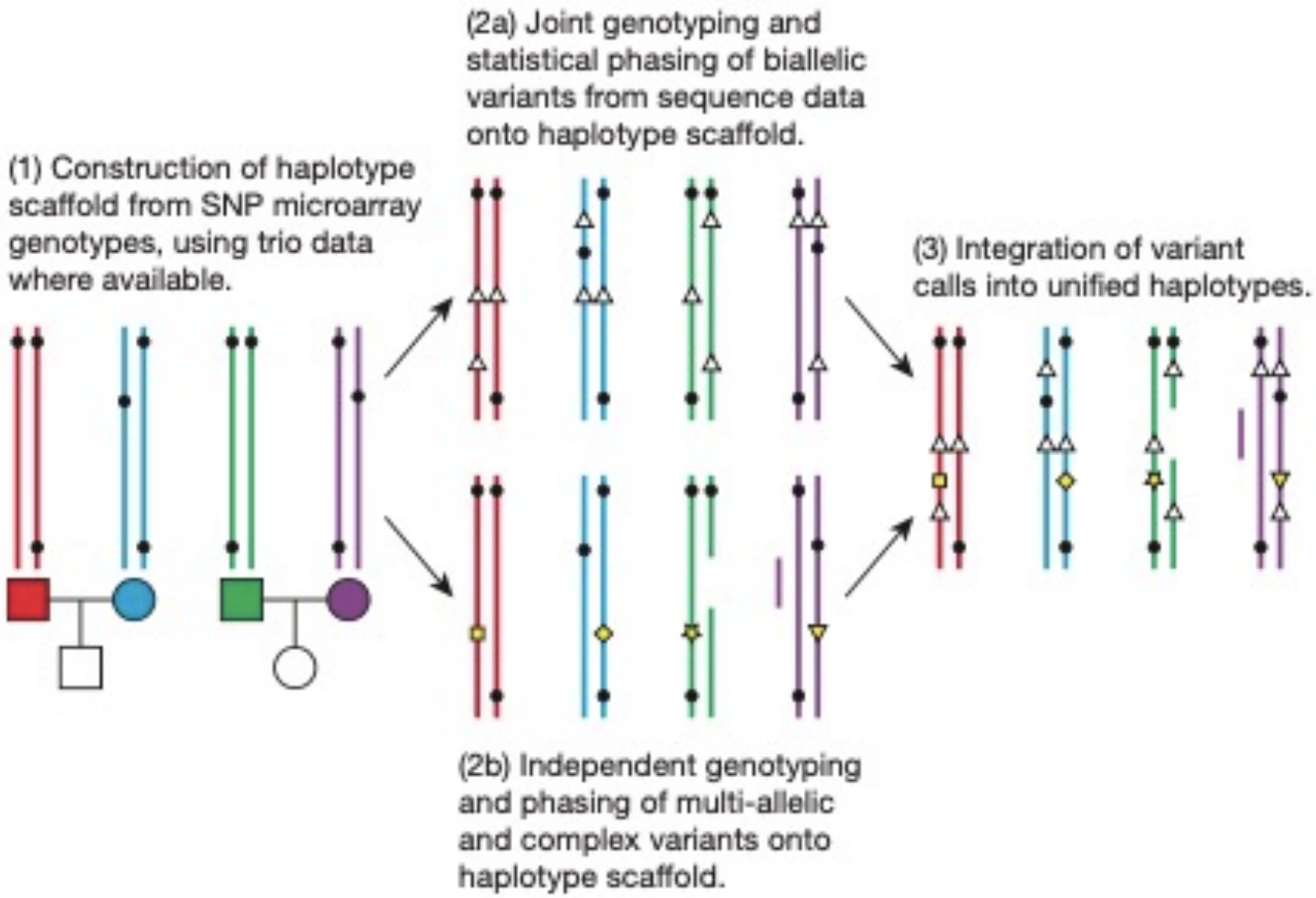


【构建单倍型框架】

1.从SNP微阵列基因型构建单倍型支架

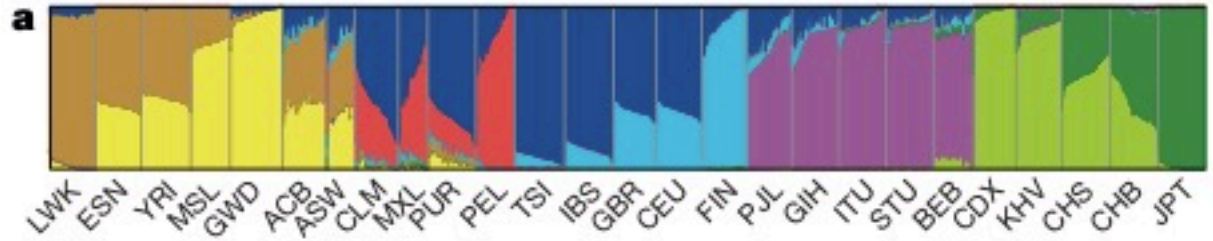
2.从序列数据到单倍型支架，对双亲变异体进行联合基因分型和统计分期；对多等位基因和复杂变异体进行独立的基因分型和分阶段到单倍型支架上

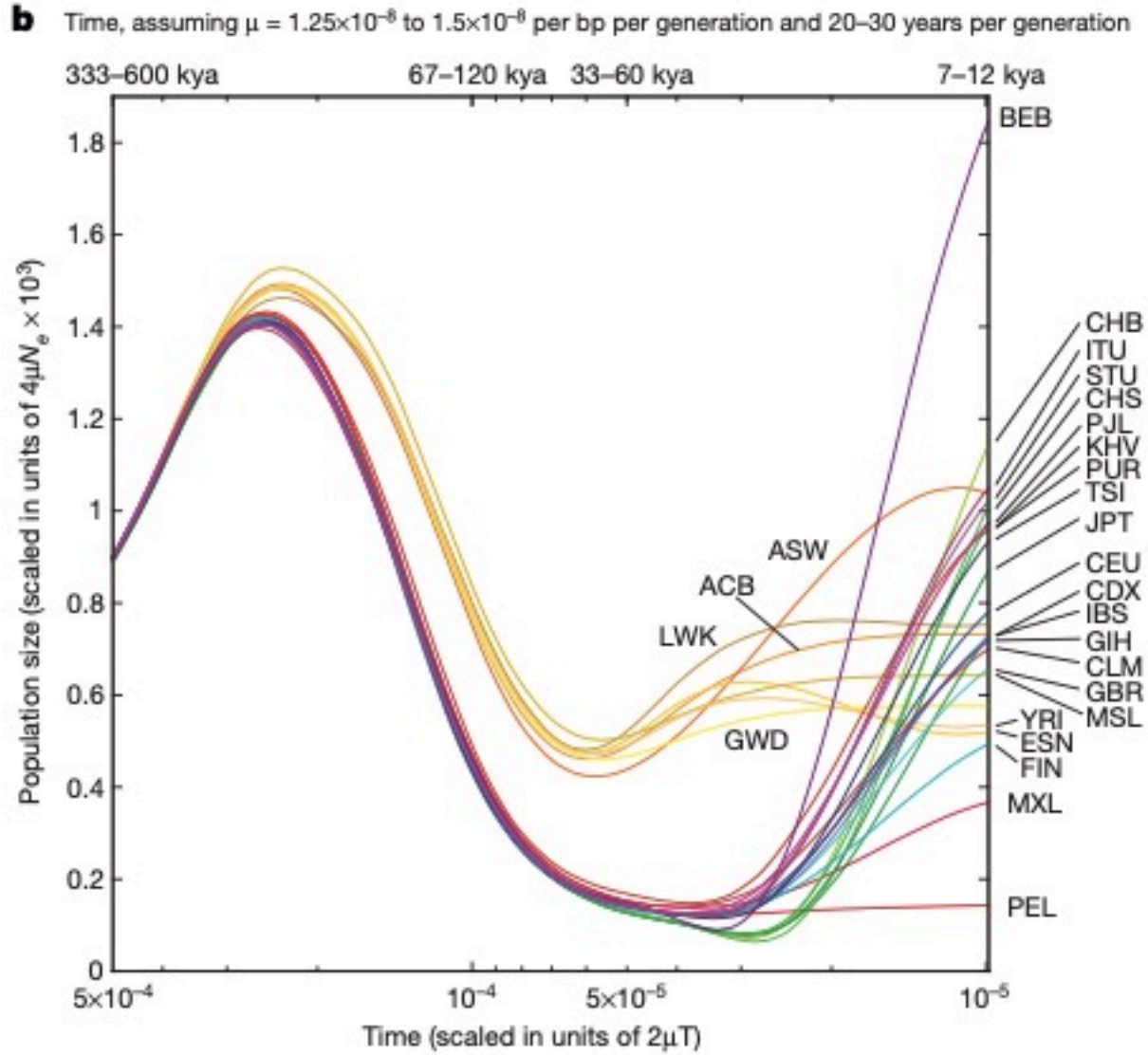
3.将变体调用整合到统一的单倍型中



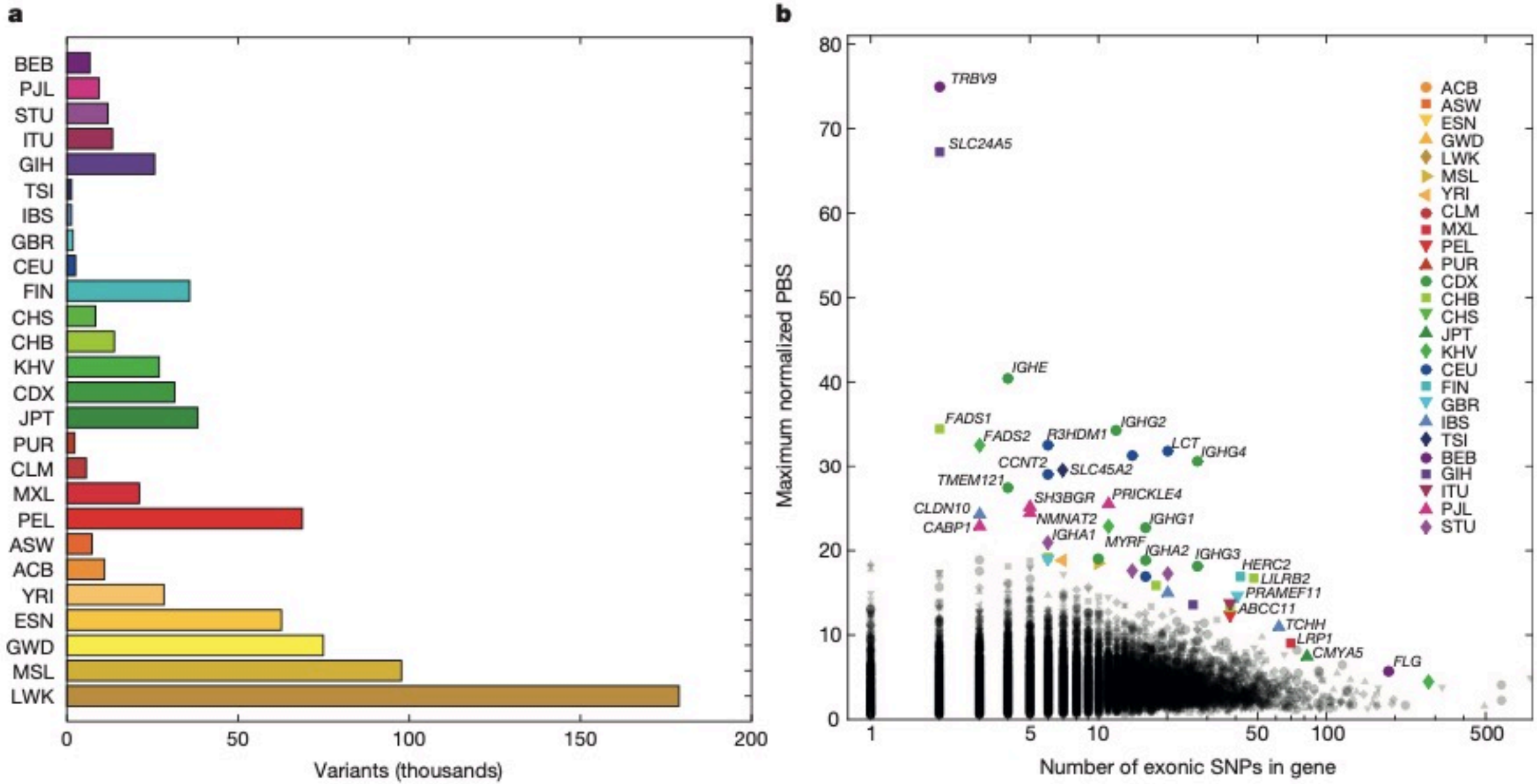
【种群间遗传变异的共享】

【人口统计学意义】

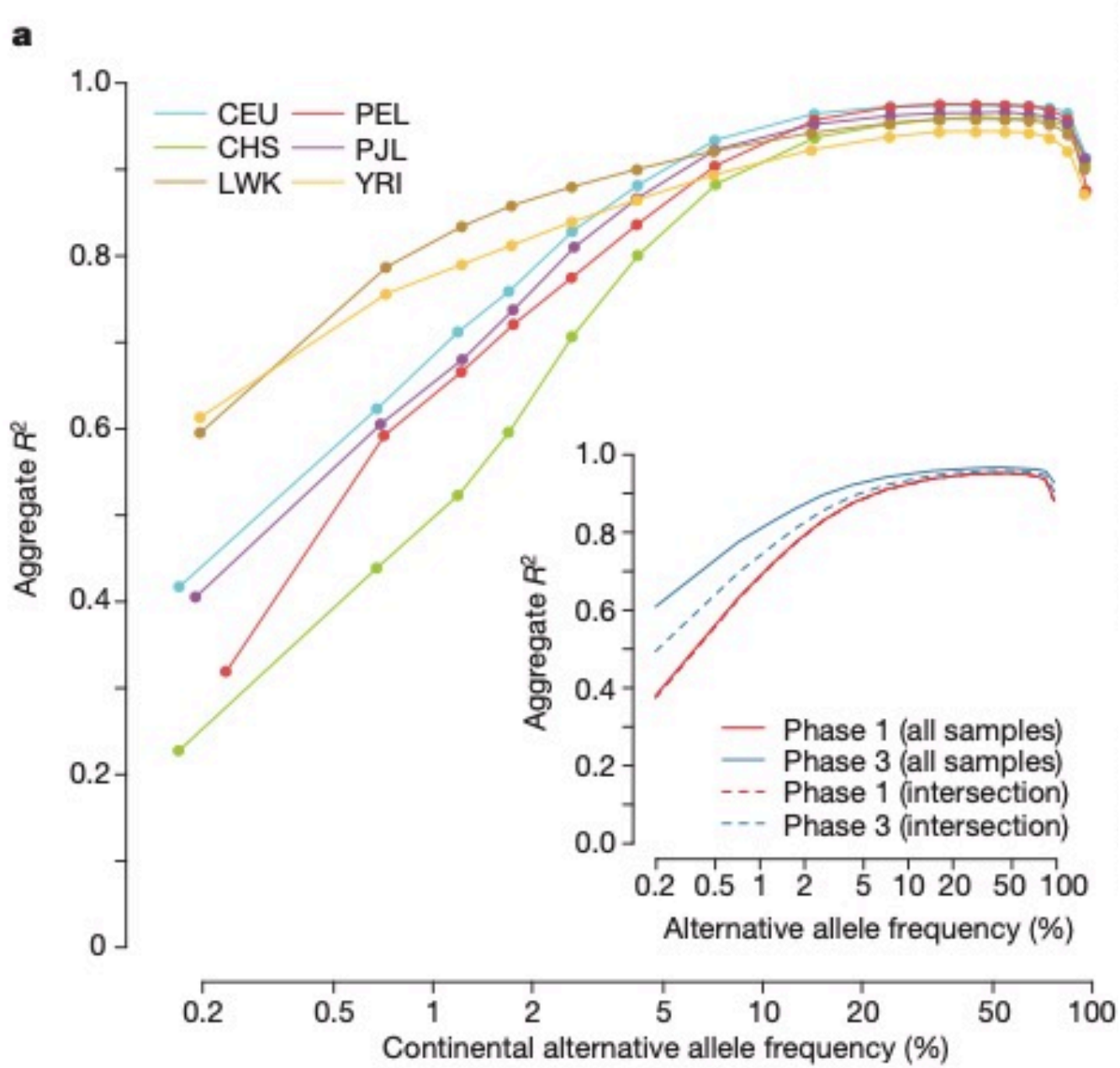




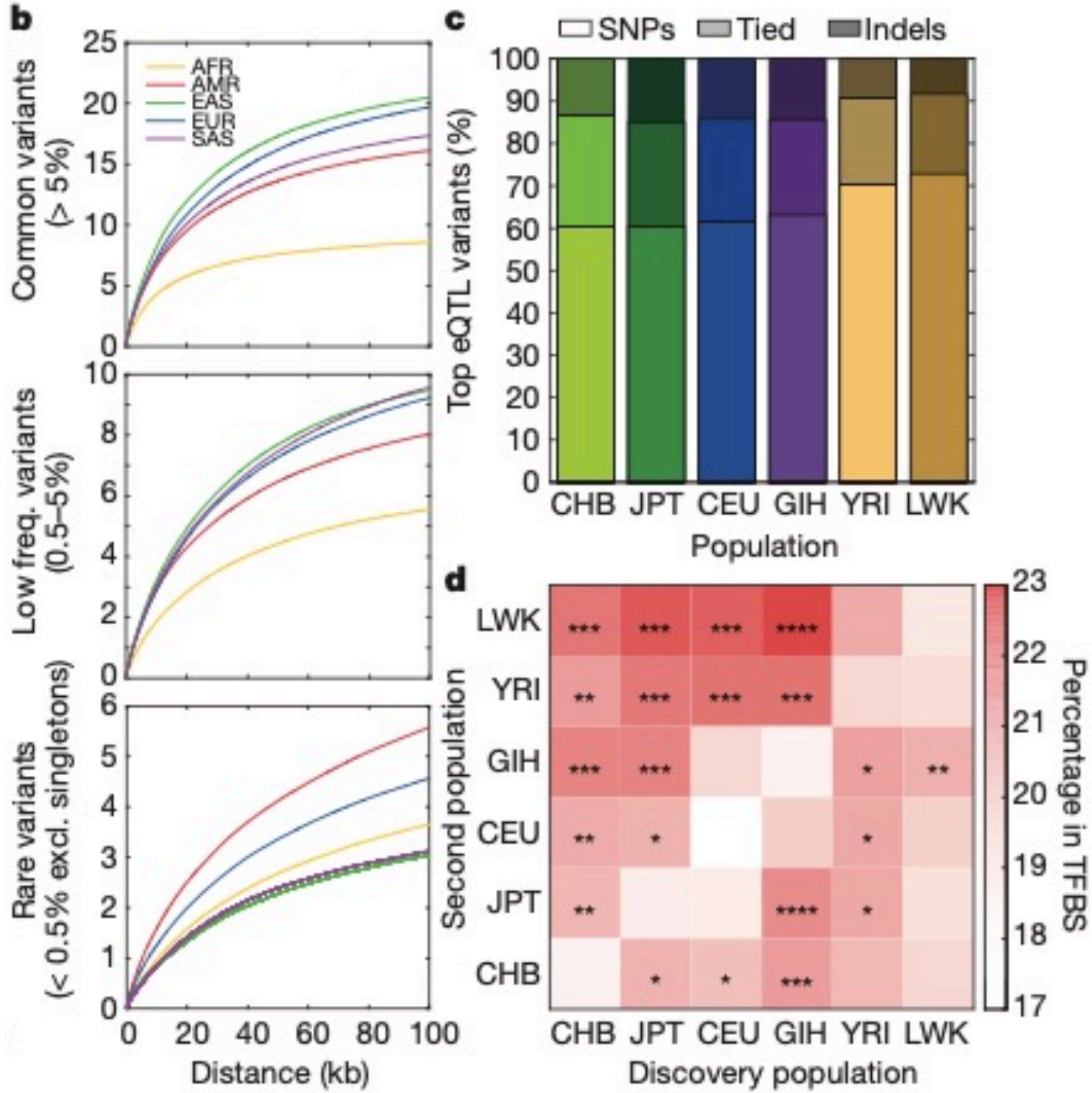
【人口统计学意义】


a
BEB
PJL
STU
ITU
GIH
TSI
IBS
GBR
CEU
FIN
CHS
CHB
KHV
CDX
JPT
PUR
CLM
MXL
PEL
ASW
ACB
YRI
ESN
GWD
MSL
LWK
0
50
100
150
200
Variants (thousands)
b
Maximum normalized PBS
Number of exonic SNPs in gene
TRBV9
SLC24A5
IGHE
FADS1
FADS2
R3HDM1
IGHG2
LCT
IGHG4
CCNT2
SLC45A2
TMEM121
SH3BGR
PRICKLE4
CLDN10
CABP1
NMNAT2
IGHG1
IGHA1
MYRF
IGHA2
IGHG3
HERC2
LILRB2
PRAMEF11
ABCC11
TCHH
LRP1
CMYA5
FLG
ACB
ASW
ESN
GWD
LWK
MSL
YRI
CLM
MXL
PEL
PUR
CDX
CHB
CHS
JPT
KHV
CEU
FIN
GBR
IBS
TSI
BEB
GIH
ITU
PJL
STU

【共享单倍型和imputation】



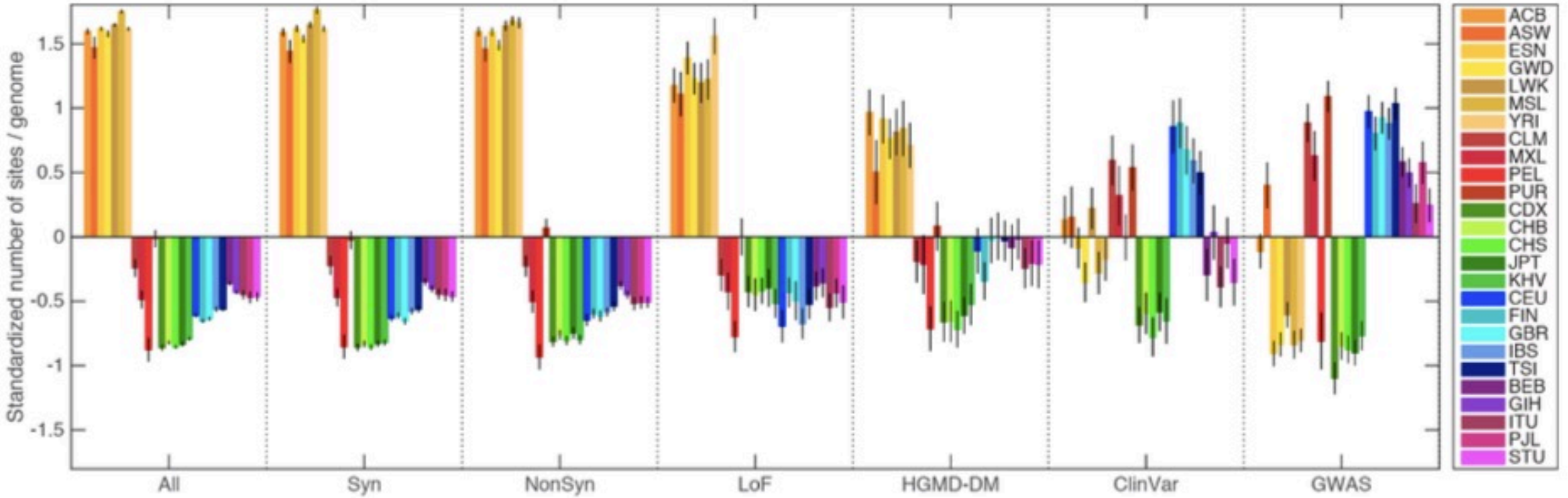
【遗传关联研究】


b
Common variants (> 5%)
Low freq. variants (0.5–5%)
Rare variants (< 0.5% excl. singletons)
AFR
AMR
EAS
EUR
SAS
Distance (kb)
c
SNPs
Tied
Indels
Top eQTL variants (%)
CHB JPT CEU GIH YRI LWK
Population
d
Second population
Discovery population
Percentage in TFBS

一 些 结 论

SOME CONCLUSIONS

1. 技术进步使全基因组测序能够应用于各种重要的医学样本
2. 千人基因组计划的样本提供了人类遗传变异的广泛的共性和特点
3. 项目样本和由此产生的数据可以广泛共享，使测序策略和分析方法可以在一组基准样本上轻松比较。



谢 谢 观 看

THANKS FOR WATCHING

2021/11/4