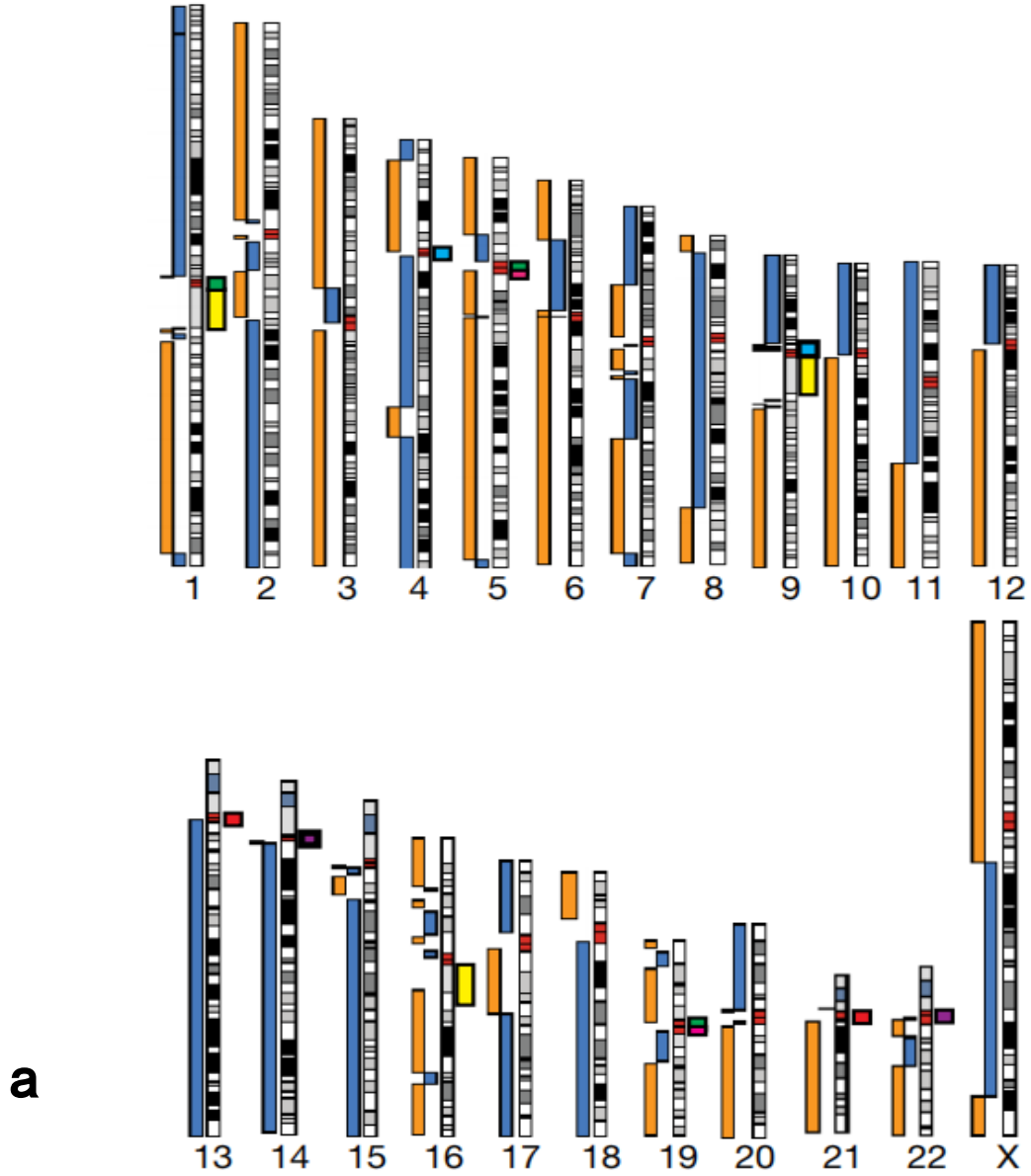
BGI华大

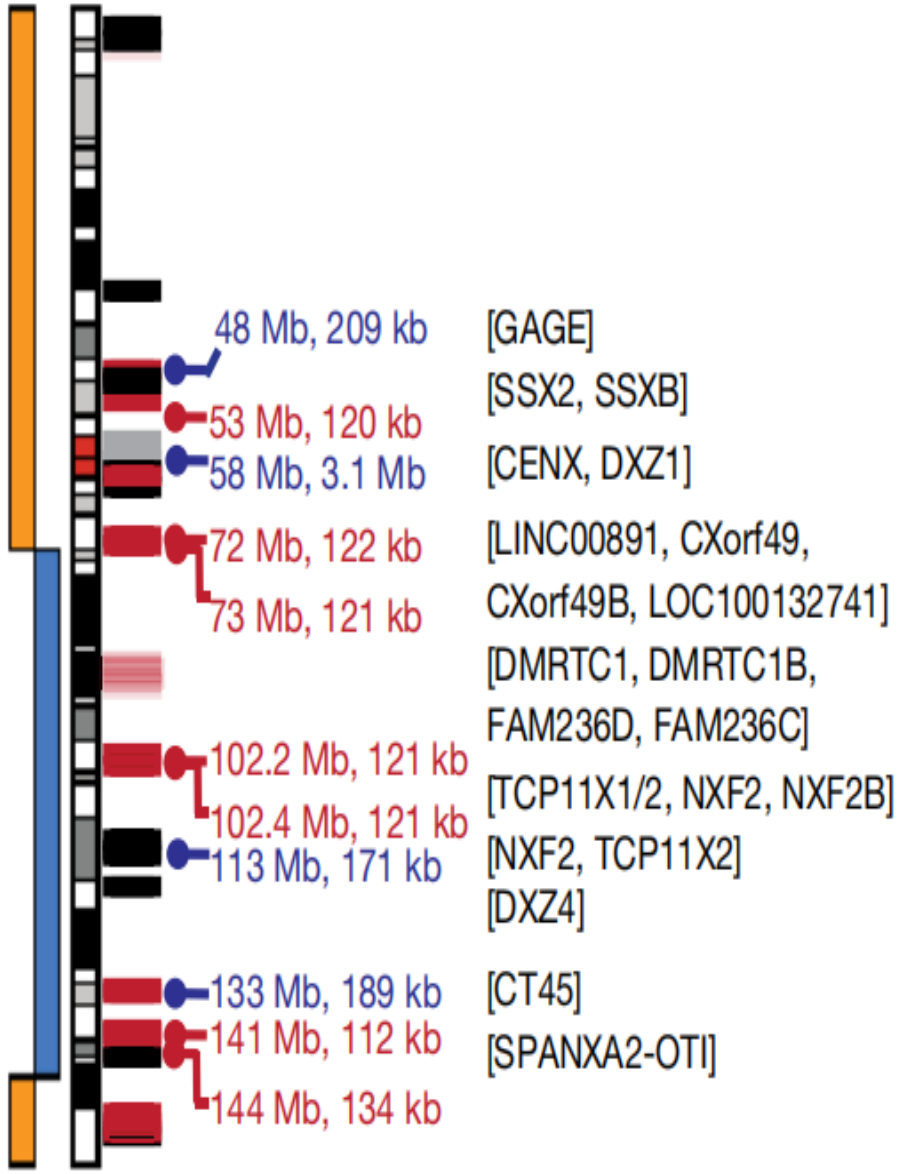
Telomere-to-telomere assembly of a complete human X chromosome

华大生命科学研究院
BGI · research

whole-genome assembly and validation

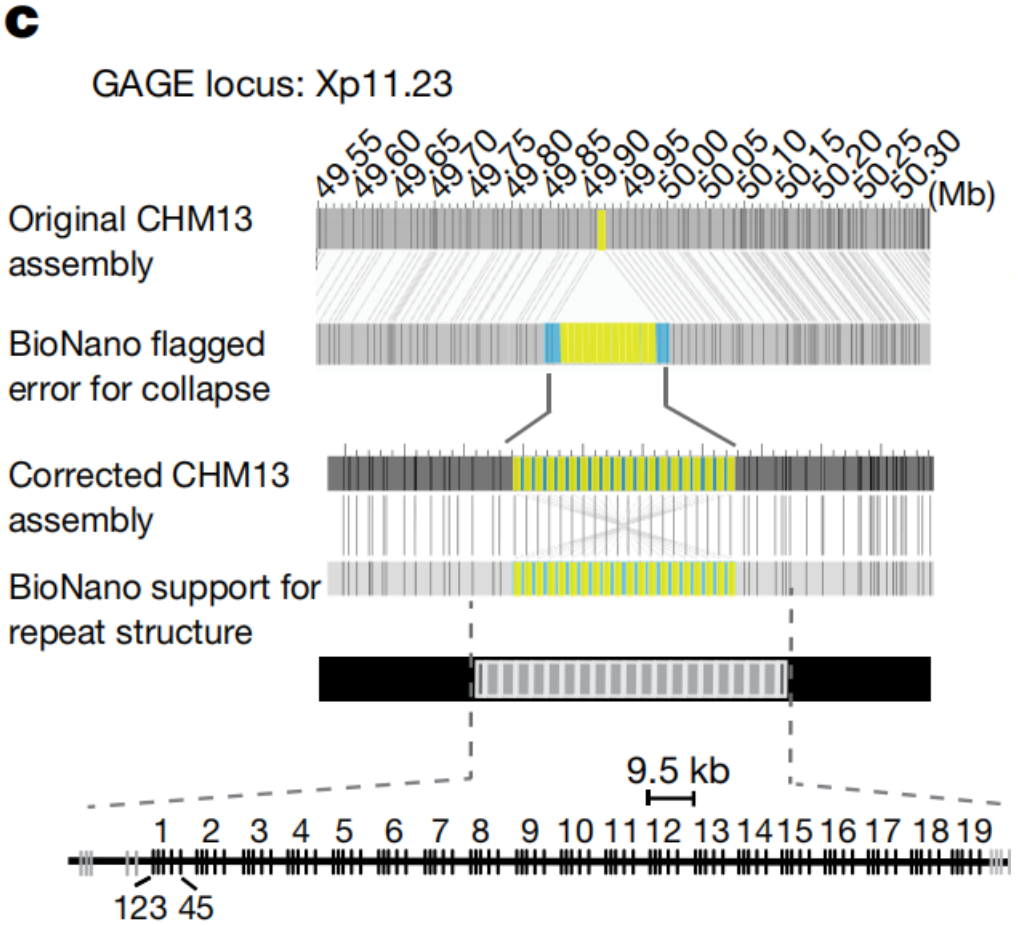


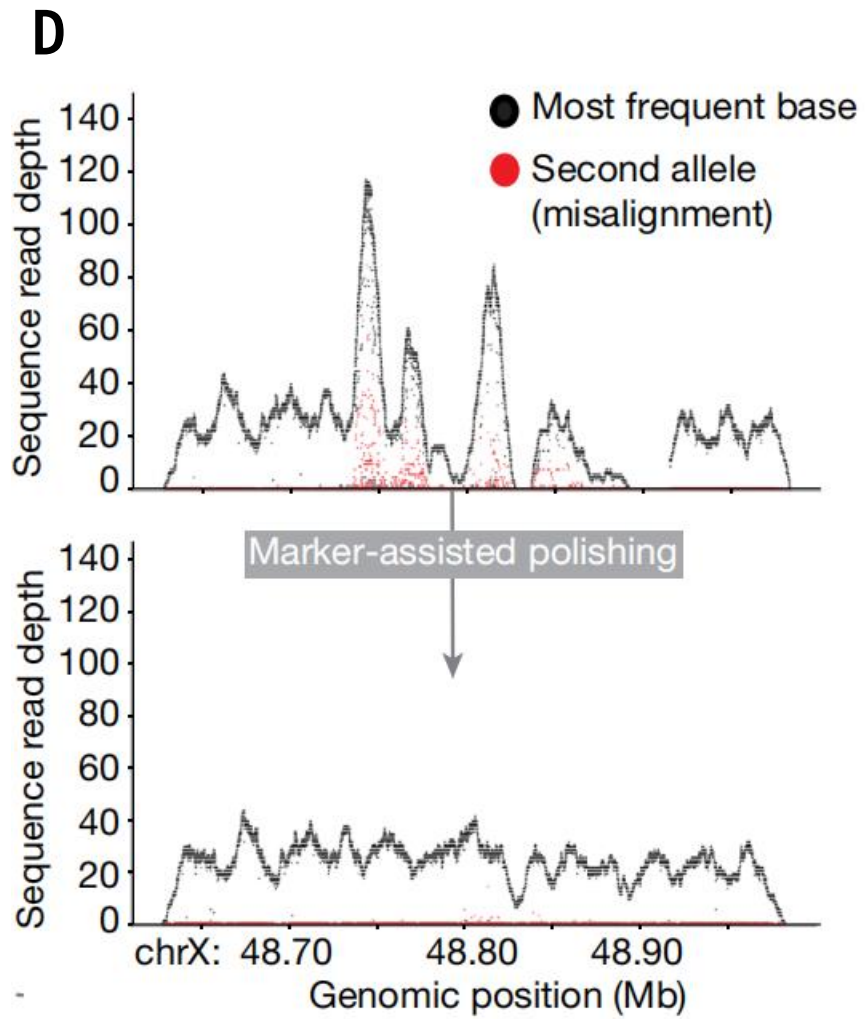
a



b

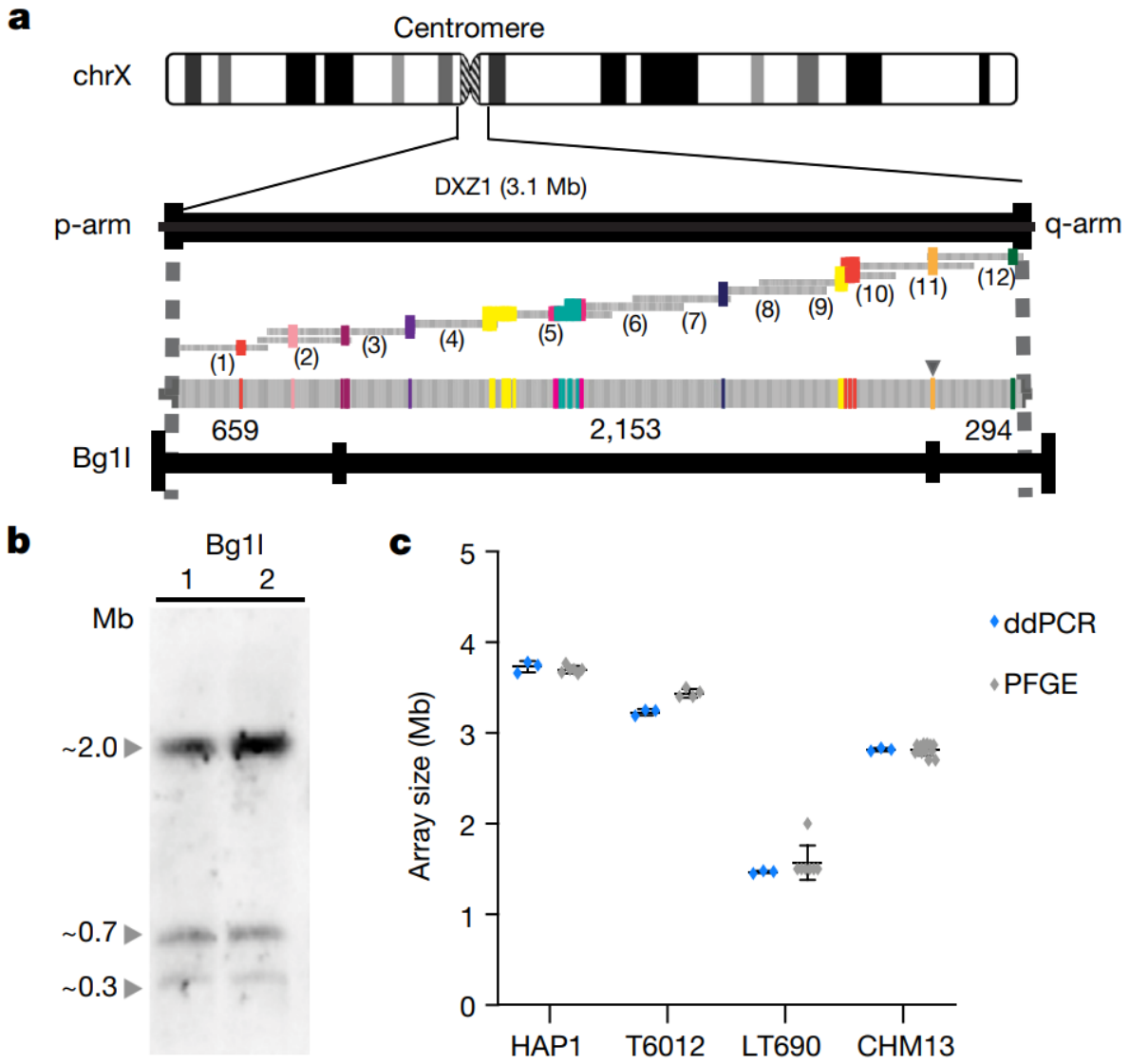
whole-genome assembly and validation

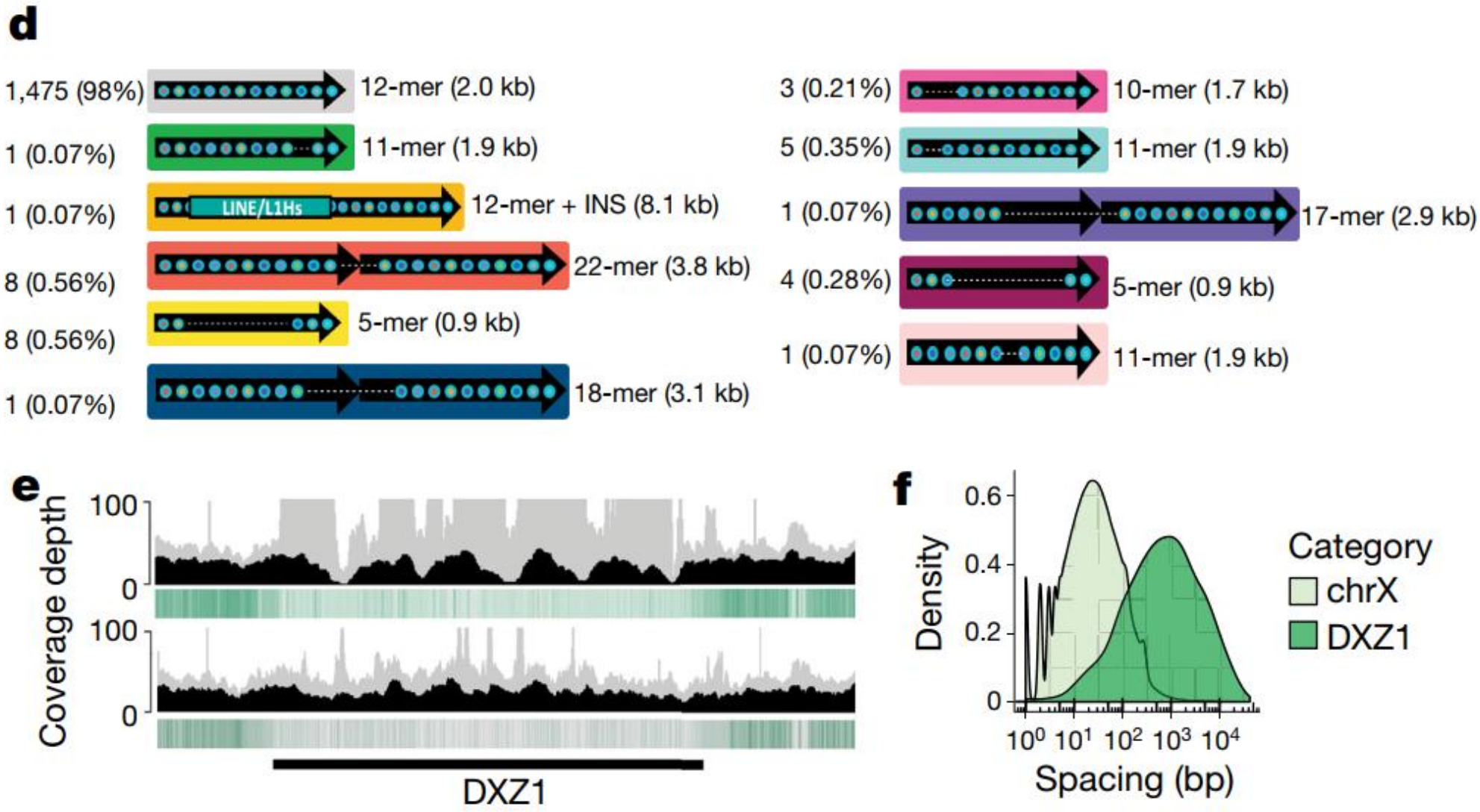




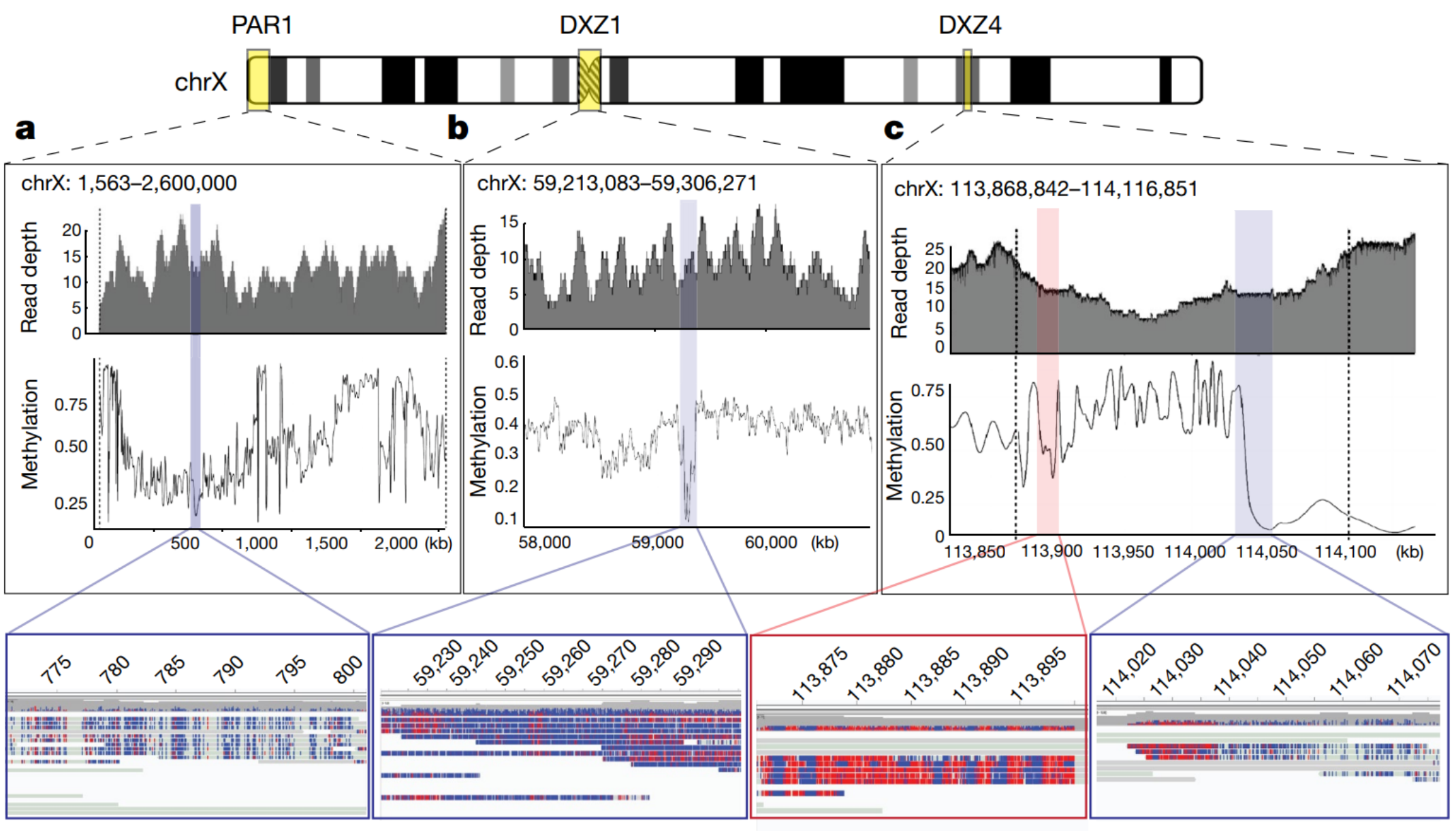
BGI华大

A finished human X chromosome




d
1,475 (98%)
12-mer (2.0 kb)
1 (0.07%)
11-mer (1.9 kb)
1 (0.07%)
LINE/L1Hs
12-mer + INS (8.1 kb)
8 (0.56%)
22-mer (3.8 kb)
8 (0.56%)
5-mer (0.9 kb)
1 (0.07%)
18-mer (3.1 kb)
3 (0.21%)
10-mer (1.7 kb)
5 (0.35%)
11-mer (1.9 kb)
1 (0.07%)
17-mer (2.9 kb)
4 (0.28%)
5-mer (0.9 kb)
1 (0.07%)
11-mer (1.9 kb)
e
Coverage depth
100
0
100
0
DXZ1
f
Density
0.6
0.4
0.2
0
10^0 10^1 10^2 10^3 10^4
Spacing (bp)
Category
chrX
DXZ1

BGI华大 Chromosome-wide DNA methylation maps



现有成果

1.同时重建了其他几条染色体
2.提供了使用现有技术完成整个人类基因组的技术方案

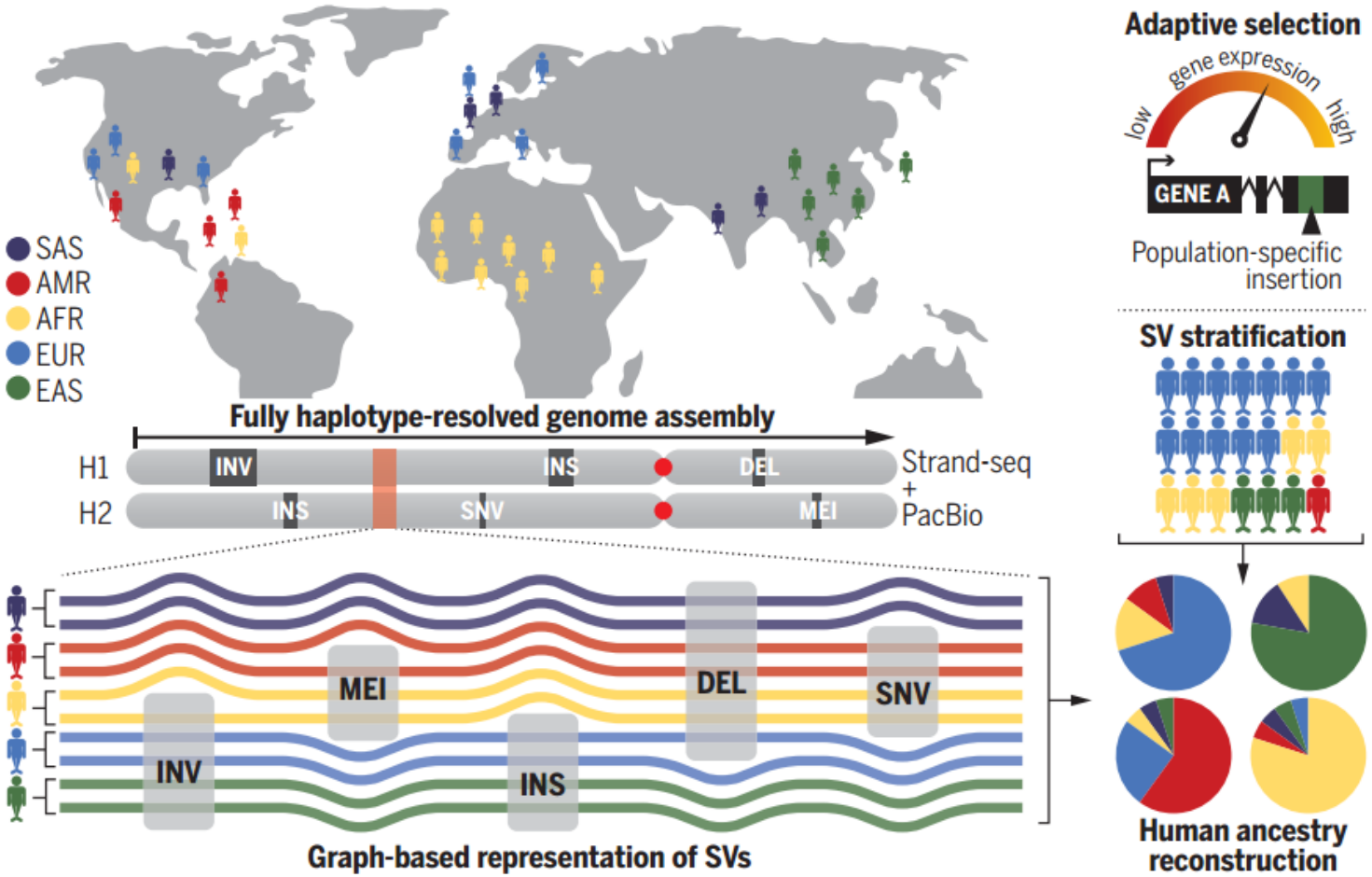
仍需克服的挑战

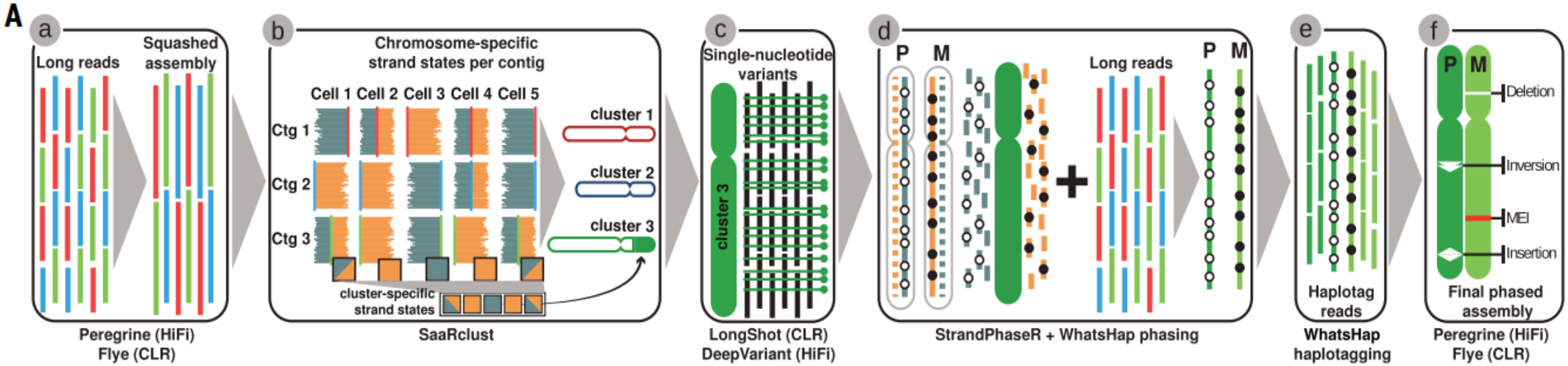
1.应用于二倍体样本将需要对潜在的单倍型进行定相，以避免混合复杂结构变异的区域
2.重复区域比X染色体更大的着丝粒卫星DNA将需要开发其他方法

BGI华大

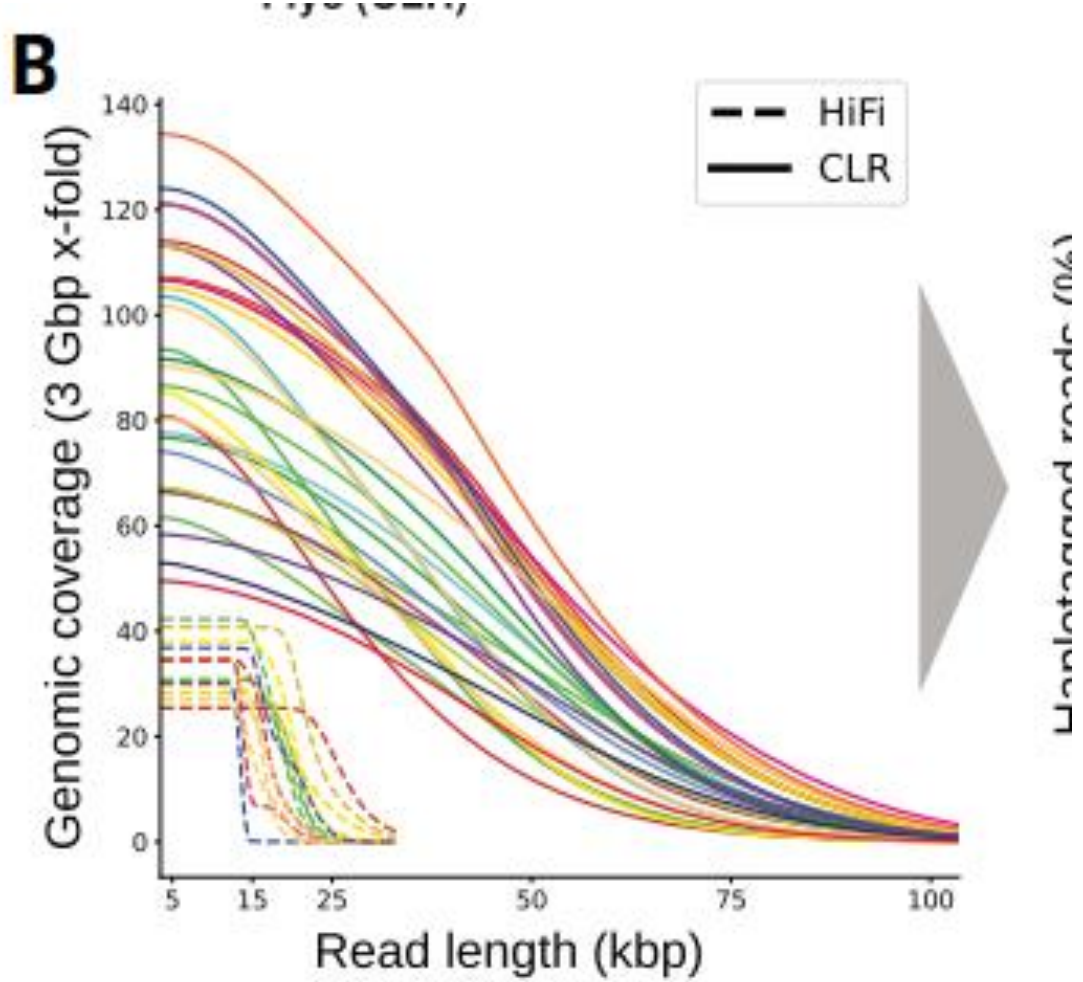
Haplotype-resolved diverse human genomes and integrated analysis of structural variation

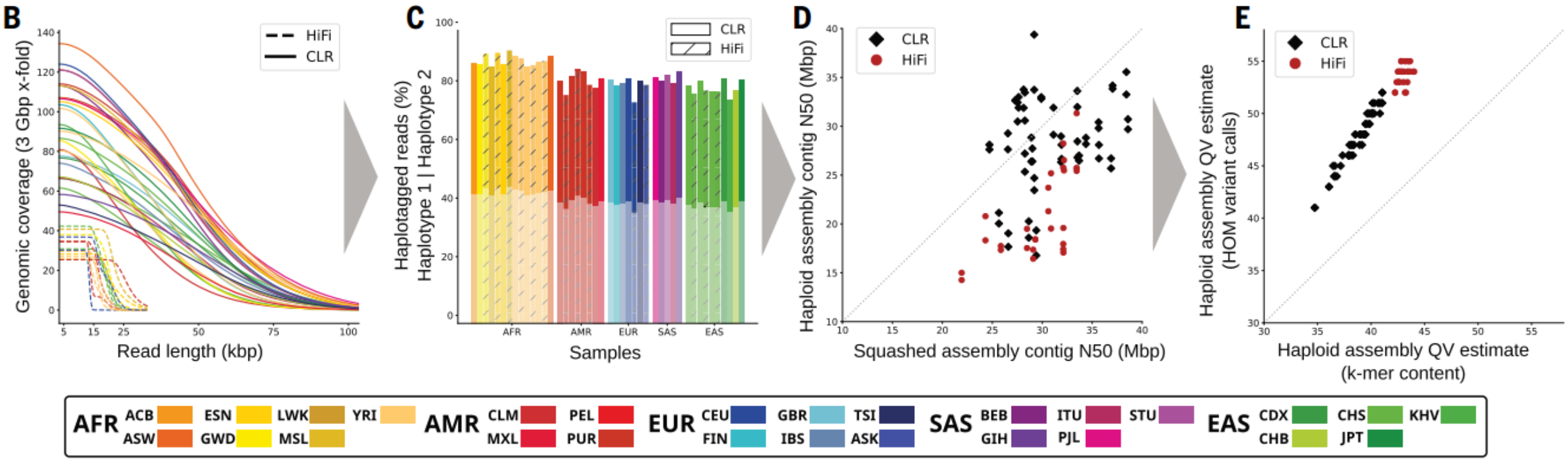
华大生命科学研究院
BGI · research


SAS
AMR
AFR
EUR
EAS
Fully haplotype-resolved genome assembly
H1
H2
INV
INS
DEL
INS
SNV
MEI
Strand-seq
+
PacBio
MEI
DEL
SNV
INV
INS
Graph-based representation of SVs
Adaptive selection
low
gene expression
high
GENE A
Population-specific
insertion
SV stratification
Human ancestry
reconstruction

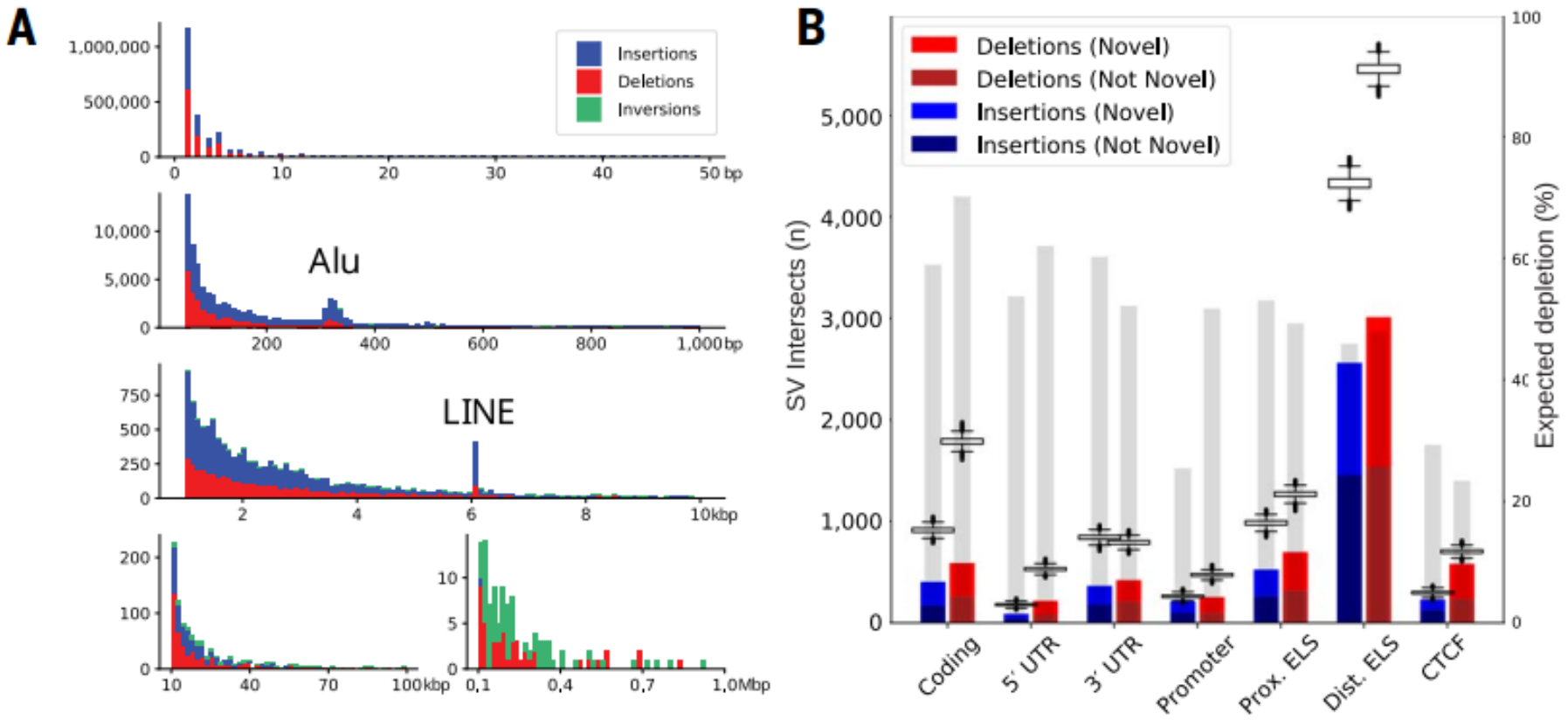


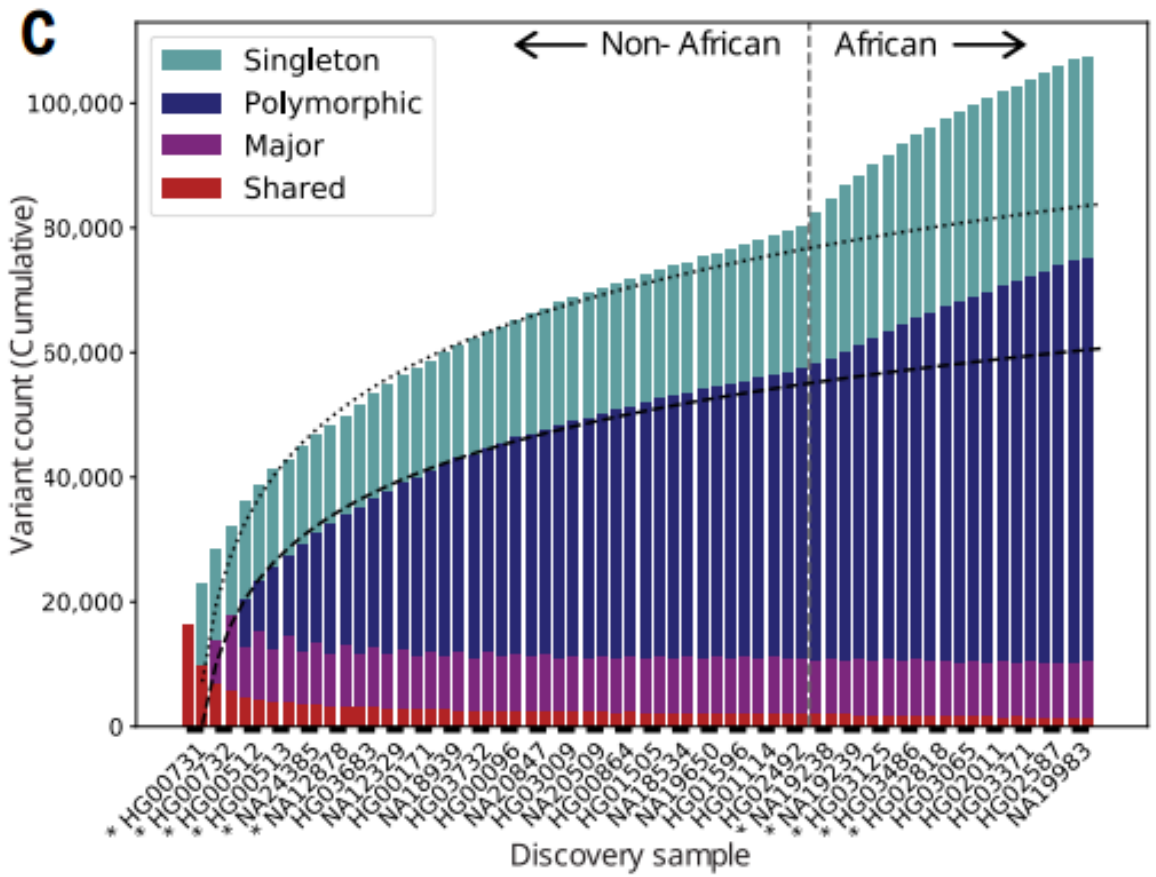
The Human Genome Structural Variation Consortium (HGSVC) recently developed a method for phased genome assembly that combines long–read PacBio whole–genome sequencing (WGS) and Strand–seq data to produce fully phased diploid genome assemblies without dependency on parent–child trio data

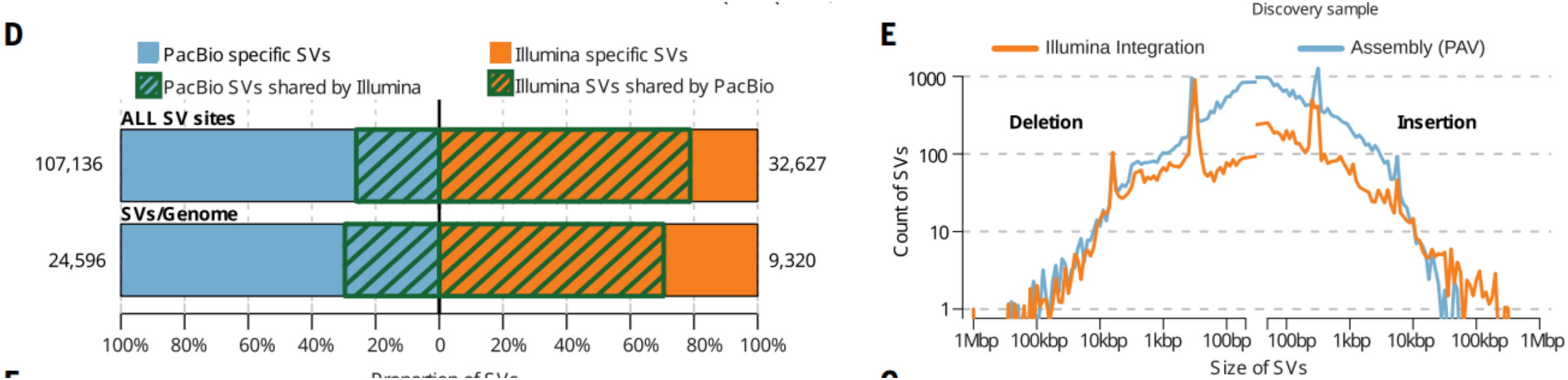

B
HiFi
CLR
Genomic coverage (3 Gbp x-fold)
0
20
40
60
80
100
120
140
5
15
25
50
75
100
Read length (kbp)



(B) Genomic coverage (y axis) as a function of the long-read length (x axis). (C) Fraction of reads that can be assigned ("haplotagged") to either haplotype 1 (semitransparent) or haplotype 2 for HiFi (hatched) and CLR (solid) datasets. (D) Contig-level N50 values for squashed (x axis) and haploid assemblies (y axis) for CLR (black diamonds) and HiFi (red circles) samples. (E) Haploid assembly QV estimates computed from distinct and shared k-mers (x axis) based on homozygous Illumina variant calls (y axis). Samples are colored according to the 1000GP population color scheme (15), with the exception of the added Ashkenazim individual NA24385/HG002 (Coriell family ID 3140) (ASK; dark blue).

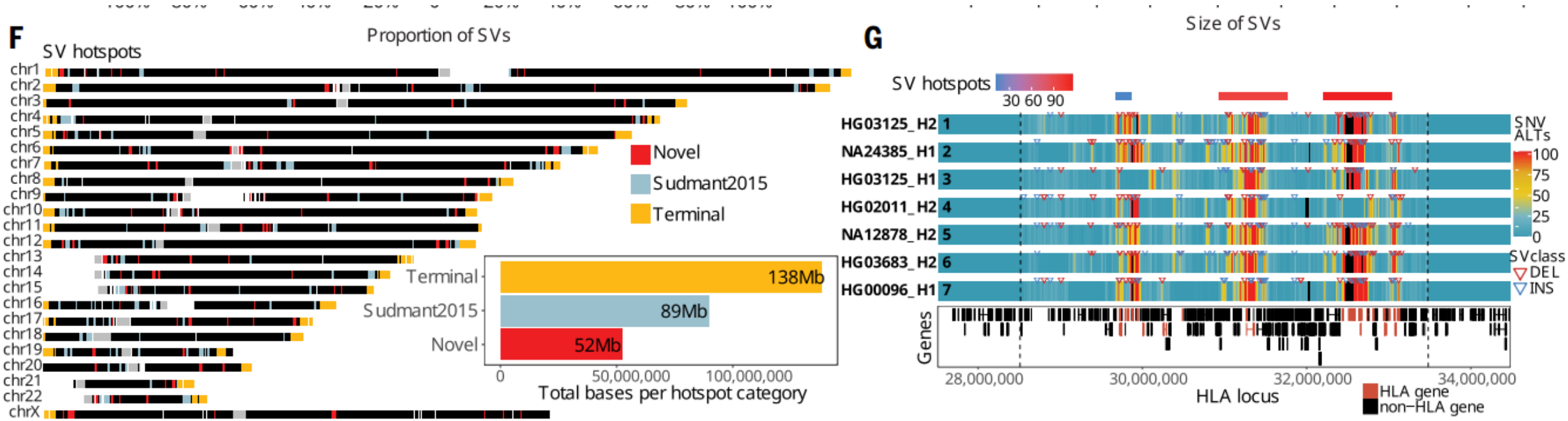

A
Insertions
Deletions
Inversions
Alu
LINE
50 bp
1,000bp
10kbp
100kbp
1,0Mbp
B
Deletions (Novel)
Deletions (Not Novel)
Insertions (Novel)
Insertions (Not Novel)
SV Intersects (n)
Expected depletion (%)
Coding
5' UTR
3' UTR
Promoter
Prox. ELS
Dist. ELS
CTCF


c
Singleton
Polymorphic
Major
Shared
← Non- African
African →
Variant count (Cumulative)
100,000
80,000
60,000
40,000
20,000
0
* HG00731
* HG00732
* HG00512
* HG00513
* NA24385
* NA12878
HG03683
NA12329
HG00171
NA18939
HG03732
HG00096
NA20847
HG03009
NA20509
HG00864
HG01505
NA18534
NA19650
HG01596
HG01114
* HG02492
* NA19238
* NA19239
* HG03125
* HG03486
HG02818
HG03065
HG02011
HG03371
HG02587
NA19983
Discovery sample



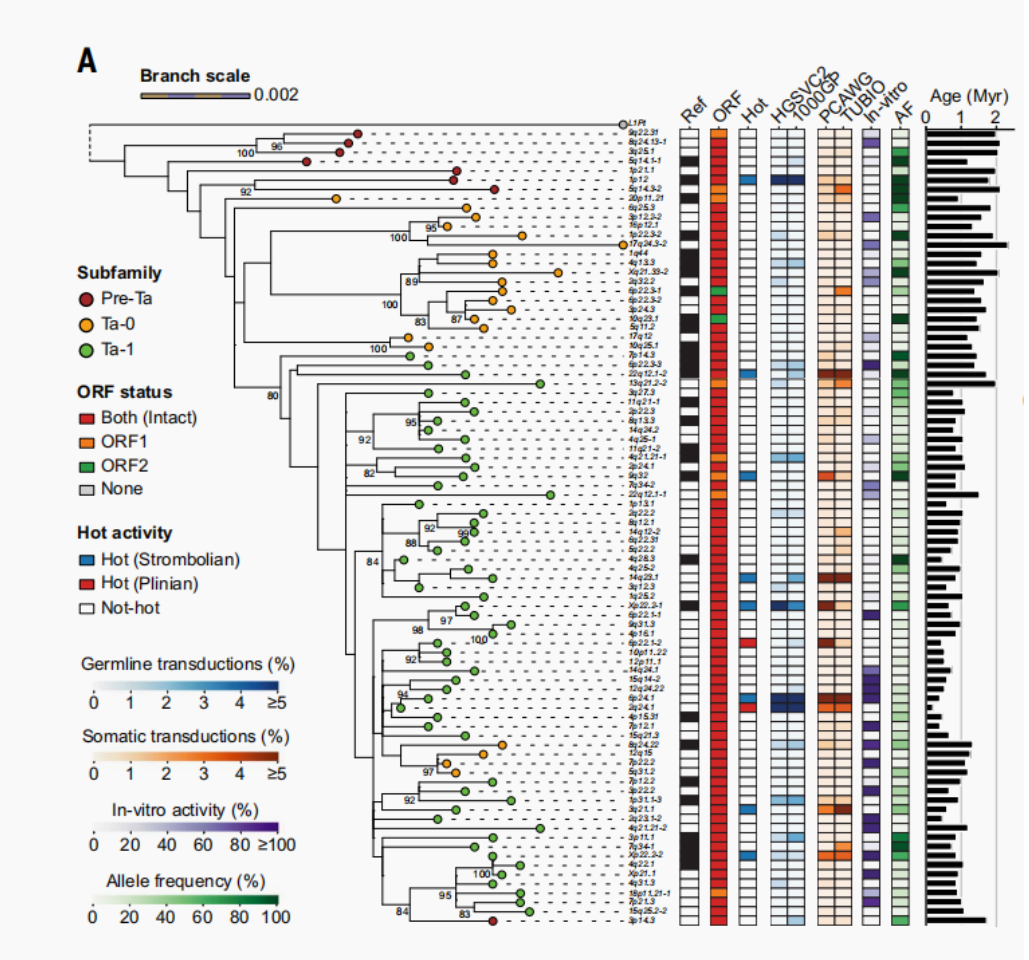
(D) Overlap between SVs detected by PacBio long-read assemblies and Illumina short-read alignments on 31 matched samples (NA24835, HG00514, HG00733, and NA19240 excluded). Top bar shows overall SV sites across 31 samples, and the bottom bar displays the average count of SVs per sample, with green stripes representing concordant SV calls between technologies.

(E) Length distribution of SVs detected with PacBio long-read assemblies and Illumina short-read alignments across all 31 matched samples.

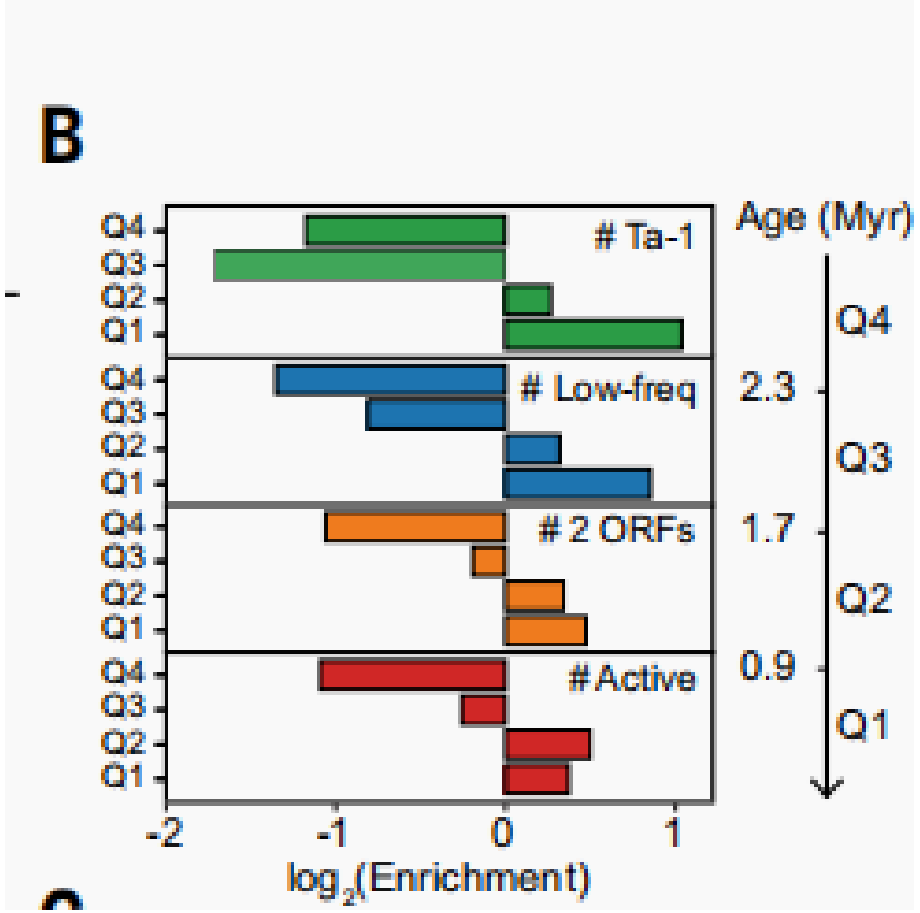


(F) Genome-wide distribution of SV hotspots divided in three categories: last 5 Mbp of chromosomes (yellow), overlapping (light blue), and previously unidentified (red) when compared with short-read SV analysis of 1000GP (23). (Inset) The total sequence length is represented by each hotspot category.
(G) Heatmap of seven selected SV haplotypes for 4-Mbp MHC region (chr6: 28,510,120 to 33,480,577; dashed lines) comparing regions of high-SNV (red) and low-diversity (blue) regions based on the number of alternate SNVs compared with the reference (GRCh38; alignment bin size 10 kbp, step 1 kbp). Phased SV insertions (blue open arrowheads) and deletions (red open arrowheads) are mapped above each haplotype. The most diverse regions correspond to SV hotspots (red and blue bars top row) and cluster with HLA genes (red bottom track)

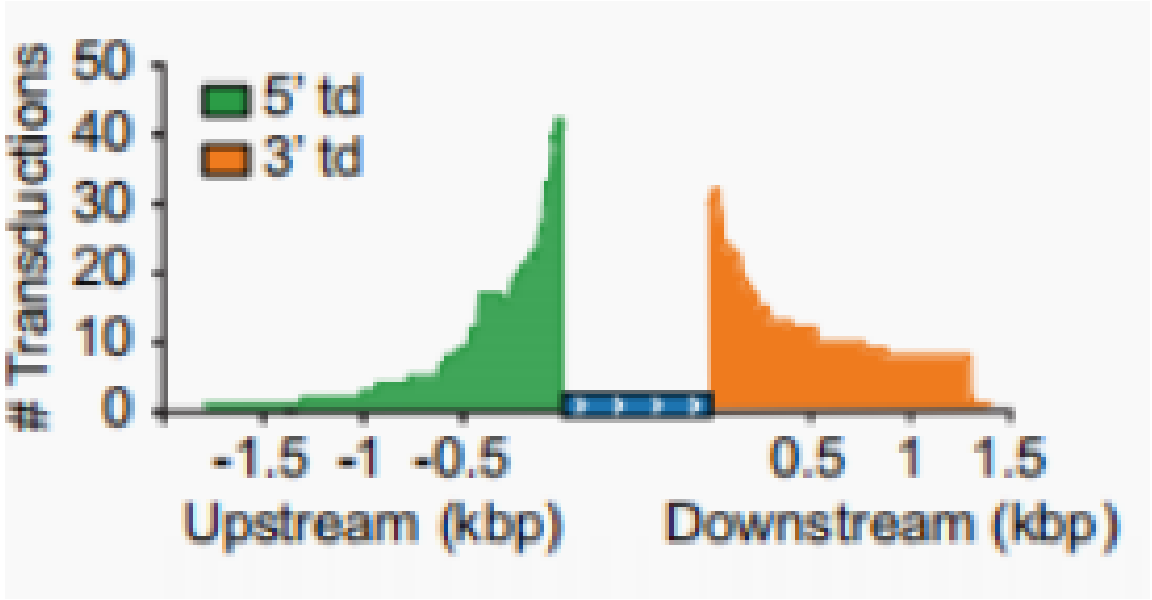
MEIs


A
Branch scale
0.002
Ref
ORF
Hot
HGSVC2
1000GP
PCAWG
TUBIO
In-vitro
AF
Age (Myr)
0 1 2
Subfamily
Pre-Ta
Ta-0
Ta-1
ORF status
Both (Intact)
ORF1
ORF2
None
Hot activity
Hot (Strombolian)
Hot (Plinian)
Not-hot
Germline transductions (%)
0 1 2 3 4 ≥5
Somatic transductions (%)
0 1 2 3 4 ≥5
In-vitro activity (%)
0 20 40 60 80 ≥100
Allele frequency (%)
0 20 40 60 80 100

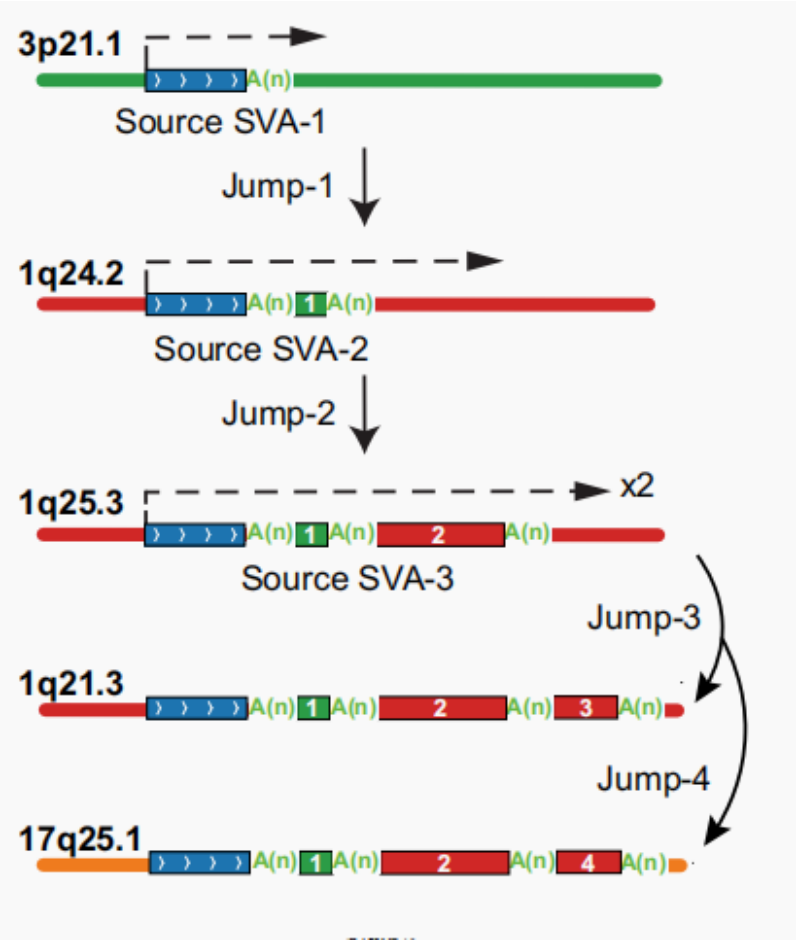
进化年龄与L1的亚家族、活性水平和等位基因频率等特征相关。

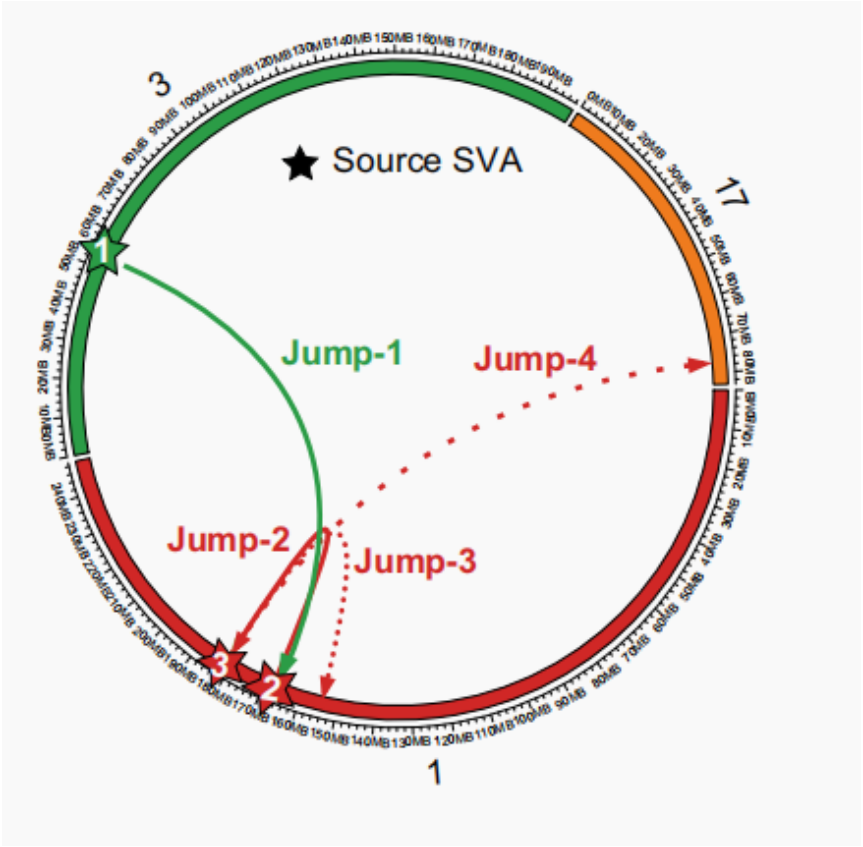


基于侧翼序列分析的5‘和3’ SVA介导的转导(td)的大小分布和数量

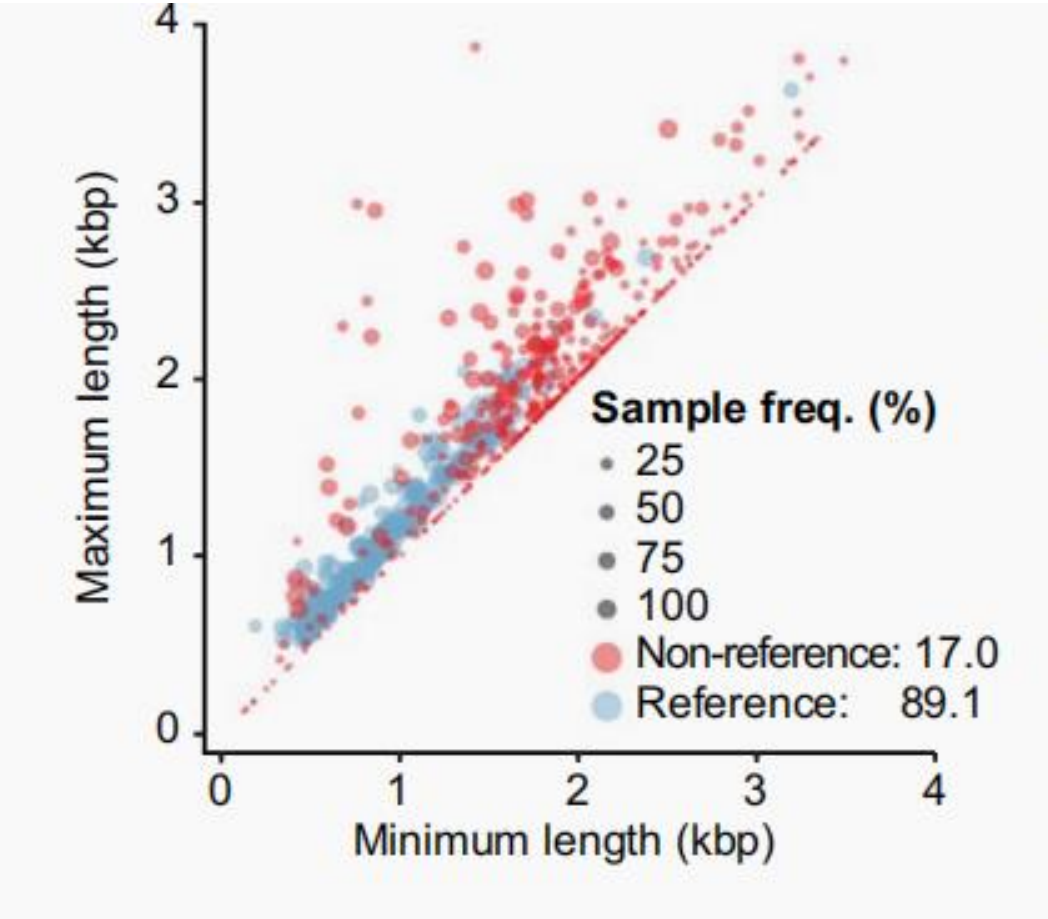


由SVA介导的一系列转导事件

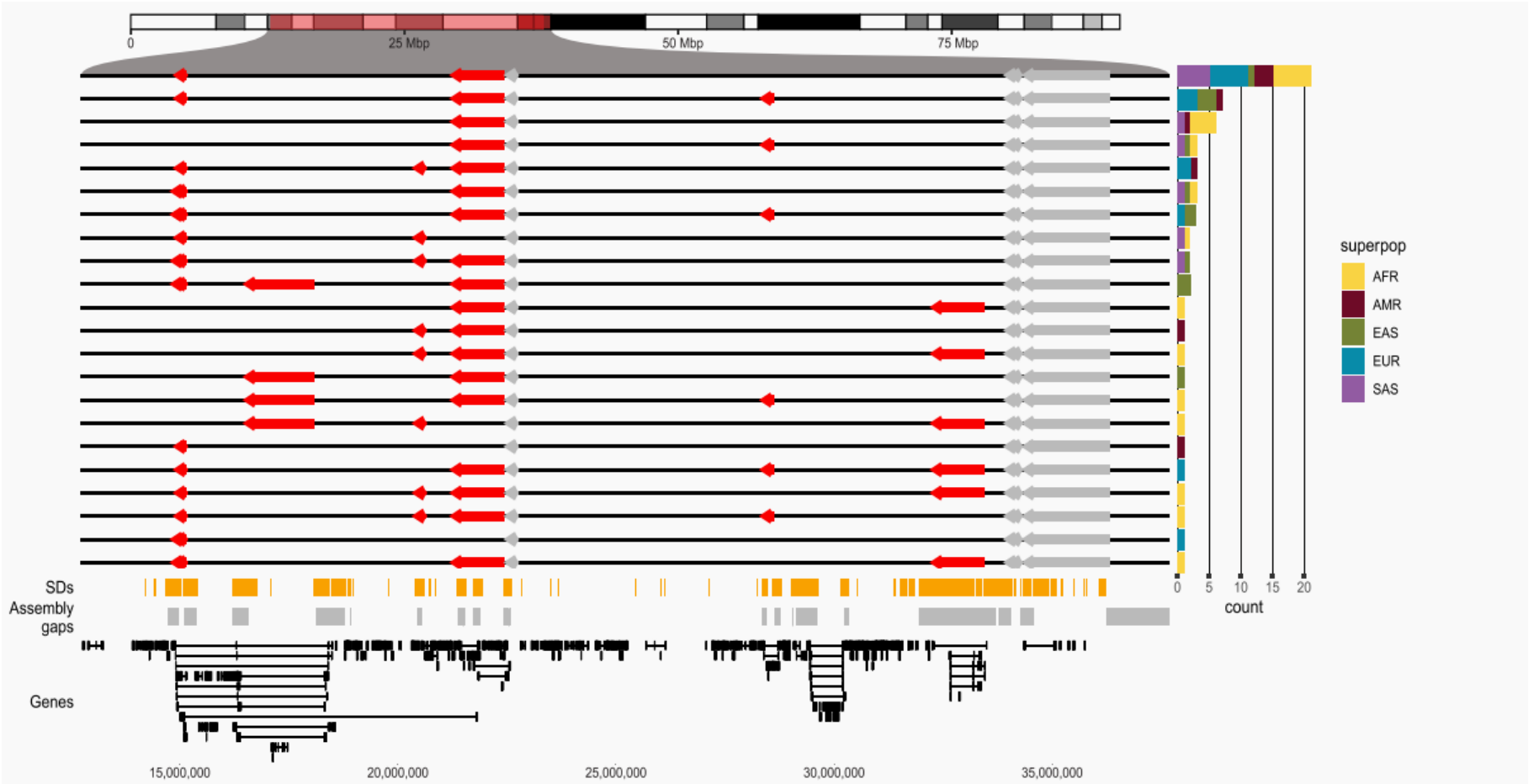




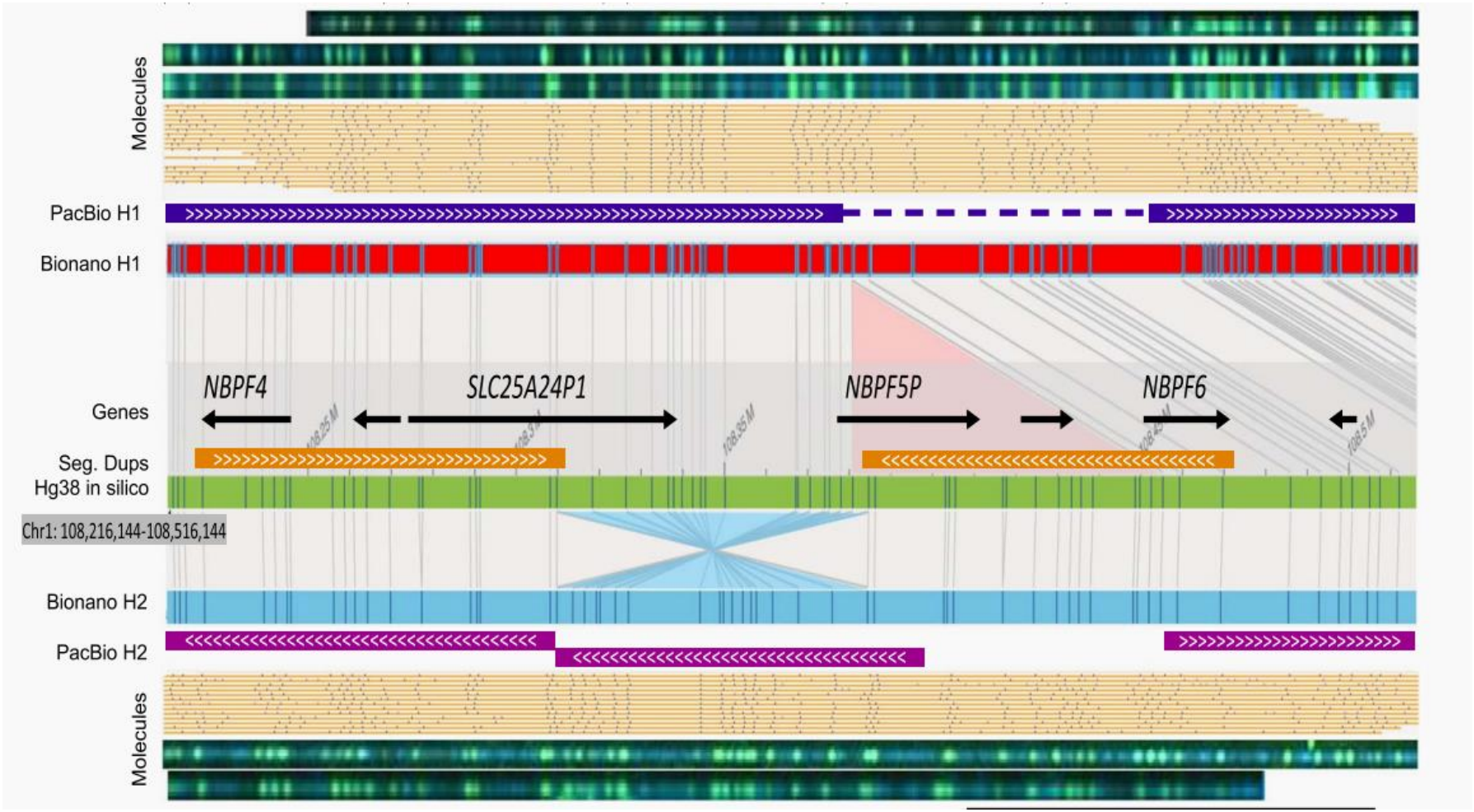
参考元素和非参考SVA元素的VNTR长度的分布



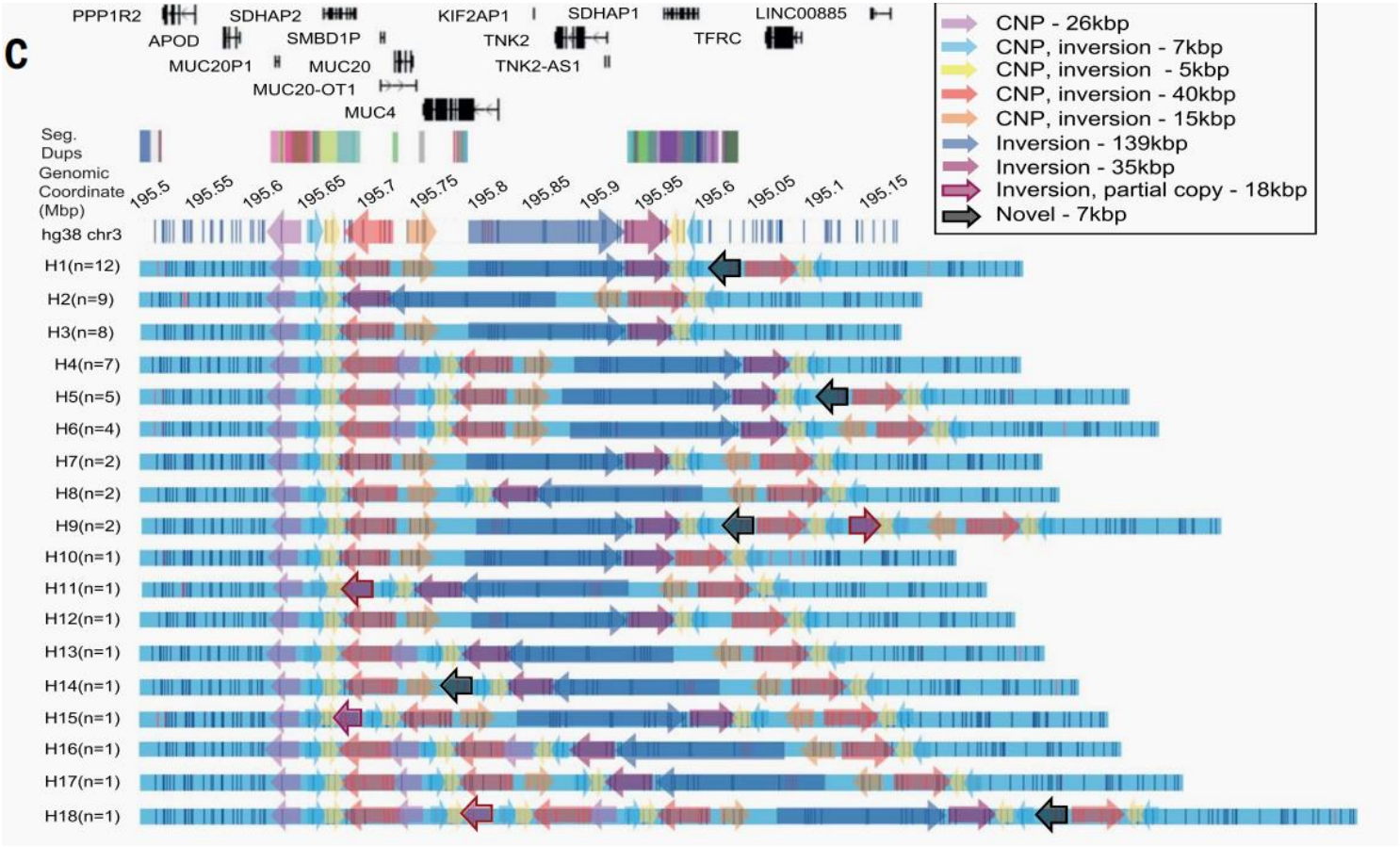
Inversions


0
25 Mbp
50 Mbp
75 Mbp
superpop
AFR
AMR
EAS
EUR
SAS
0
5
10
15
20
count
SDs
Assembly gaps
Genes
15,000,000
20,000,000
25,000,000
30,000,000
35,000,000

Complex structural variation



在染色体3q29处的单倍型结构复杂性



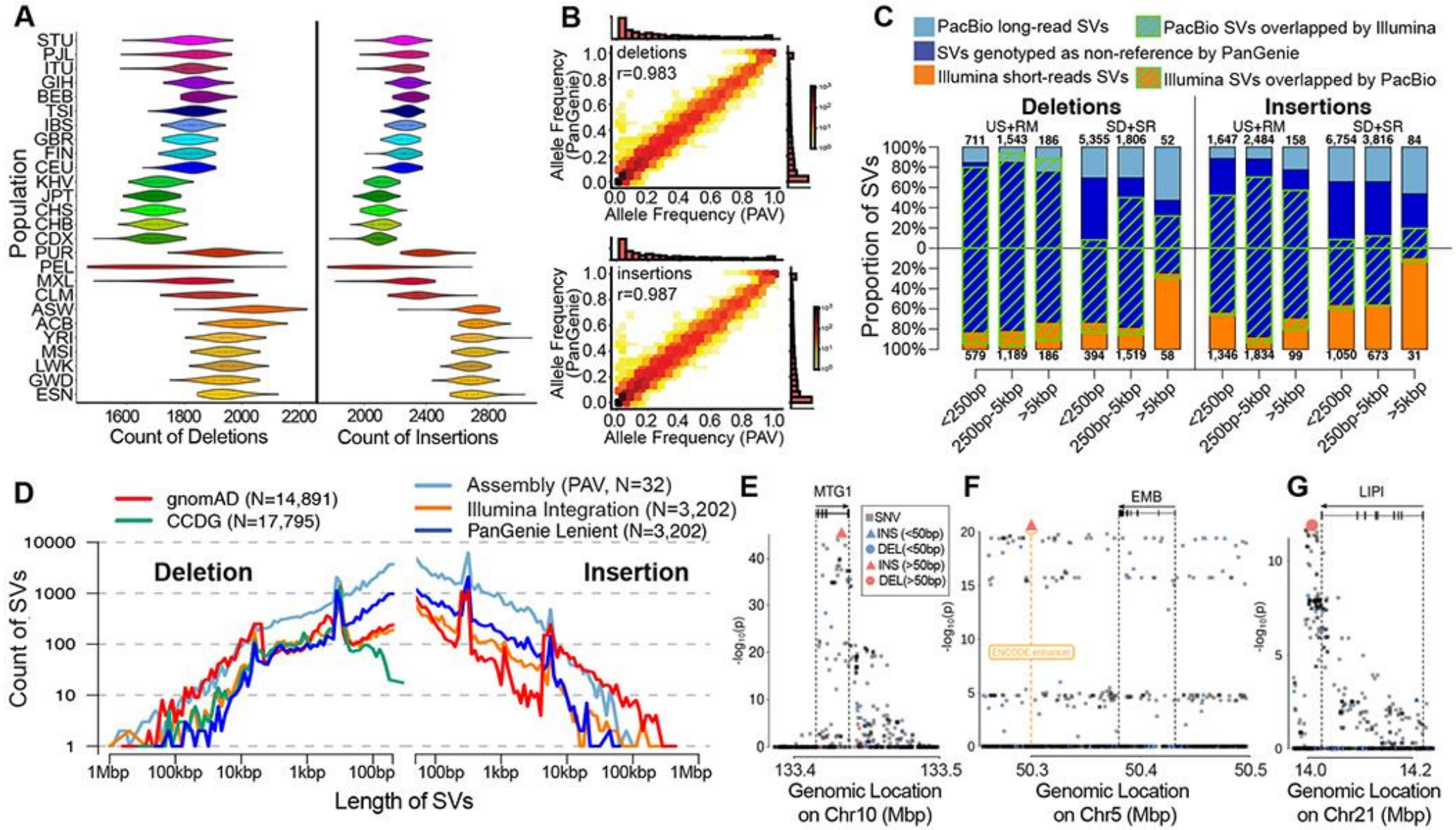
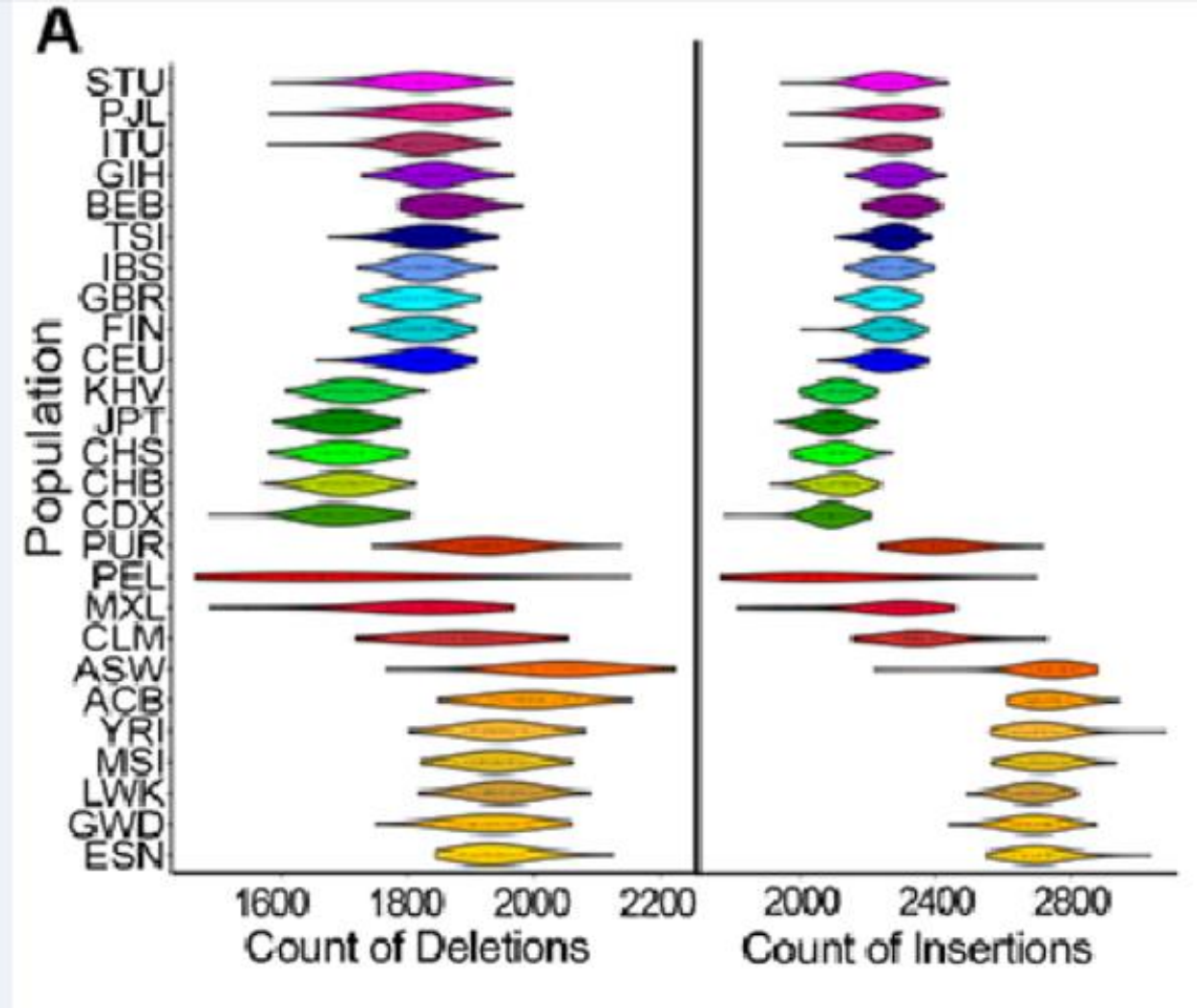

A
Population
STU
PJL
ITU
GIH
BEB
TSI
IBS
GBR
FIN
CEU
KHV
JPT
CHS
CHB
CDX
PUR
PEL
MXL
CLM
ASW
ACB
YRI
MSI
LWK
GWD
ESN
1600 1800 2000 2200
Count of Deletions
2000 2400 2800
Count of Insertions
B
deletions
r=0.983
Allele Frequency (PanGenie)
Allele Frequency (PAV)
insertions
r=0.987
C
PacBio long-read SVs
PacBio SVs overlapped by Illumina
SVs genotyped as non-reference by PanGenie
Illumina short-reads SVs
Illumina SVs overlapped by PacBio
Deletions
Insertions
US+RM
SD+SR
Proportion of SVs
711 1,543 186 5,355 1,806 52
1,647 2,484 158 6,754 3,816 84
579 1,189 186 394 1,519 58
1,346 1,834 99 1,050 673 31
<250bp
250bp-5kbp
>5kbp
D
gnomAD (N=14,891)
CCDG (N=17,795)
Assembly (PAV, N=32)
Illumina Integration (N=3,202)
PanGenie Lenient (N=3,202)
Deletion
Insertion
Count of SVs
1Mbp 100kbp 10kbp 1kbp 100bp
Length of SVs
E
MTG1
SNV
INS (<50bp)
DEL(<50bp)
INS (>50bp)
DEL(>50bp)
133.4 133.5
Genomic Location on Chr10 (Mbp)
F
EMB
ENCODE enhancer
50.3 50.4 50.5
Genomic Location on Chr5 (Mbp)
G
LIPI
14.0 14.2
Genomic Location on Chr21 (Mbp)

Figure 5: SV genotyping and eQTL analysis

Genotyping



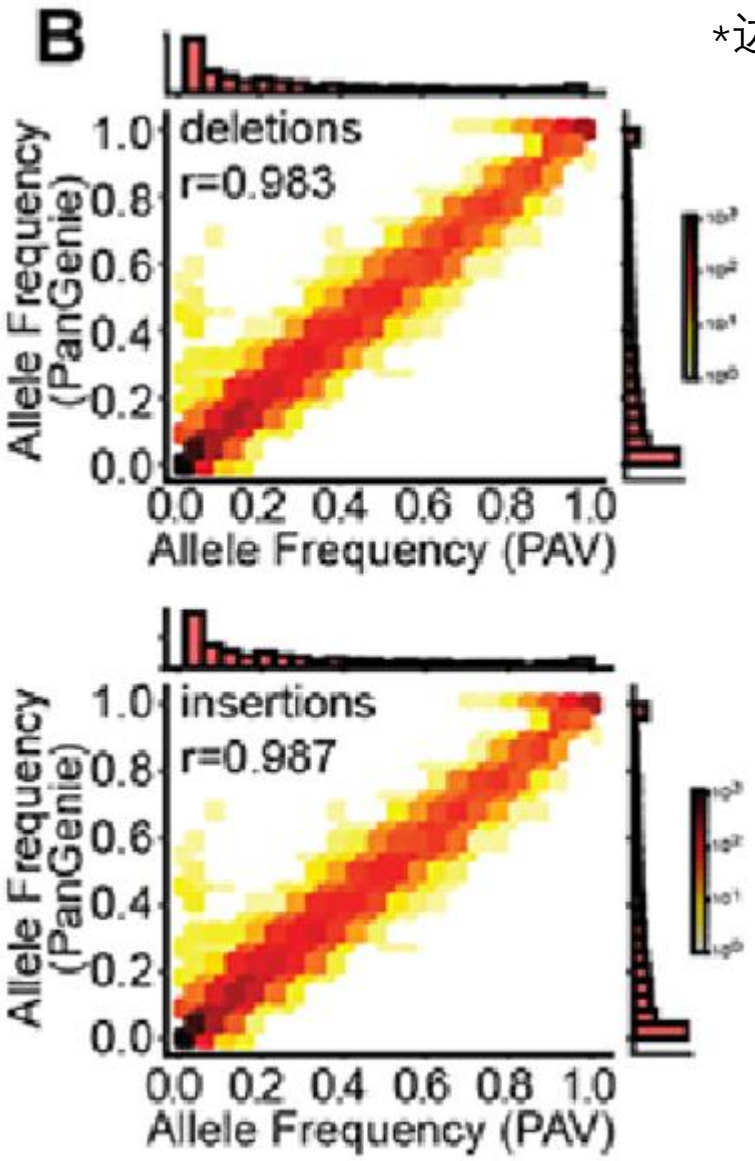
PanGenie:利用基于单倍型解析的基因组组装构建的泛基因组参考，结合来自原始短读长测序数据的 k-mer 计数信息，对广泛的遗传变异进行基因分型，并提供对短读无法访问的区域的访问。

*使用由15.5M SNVs、1.03M indels（1-49 bp）和96.1k SVs（其中等位基因脱落率<20%）组成的参考集进行基因分型步骤，并将这些变体的基因分型纳入1000GP WGS数据集，观察预期的多样性模式。

Figure 5: SV genotyping and eQTL analysis

*边际直方图为线性比例

Genotyping



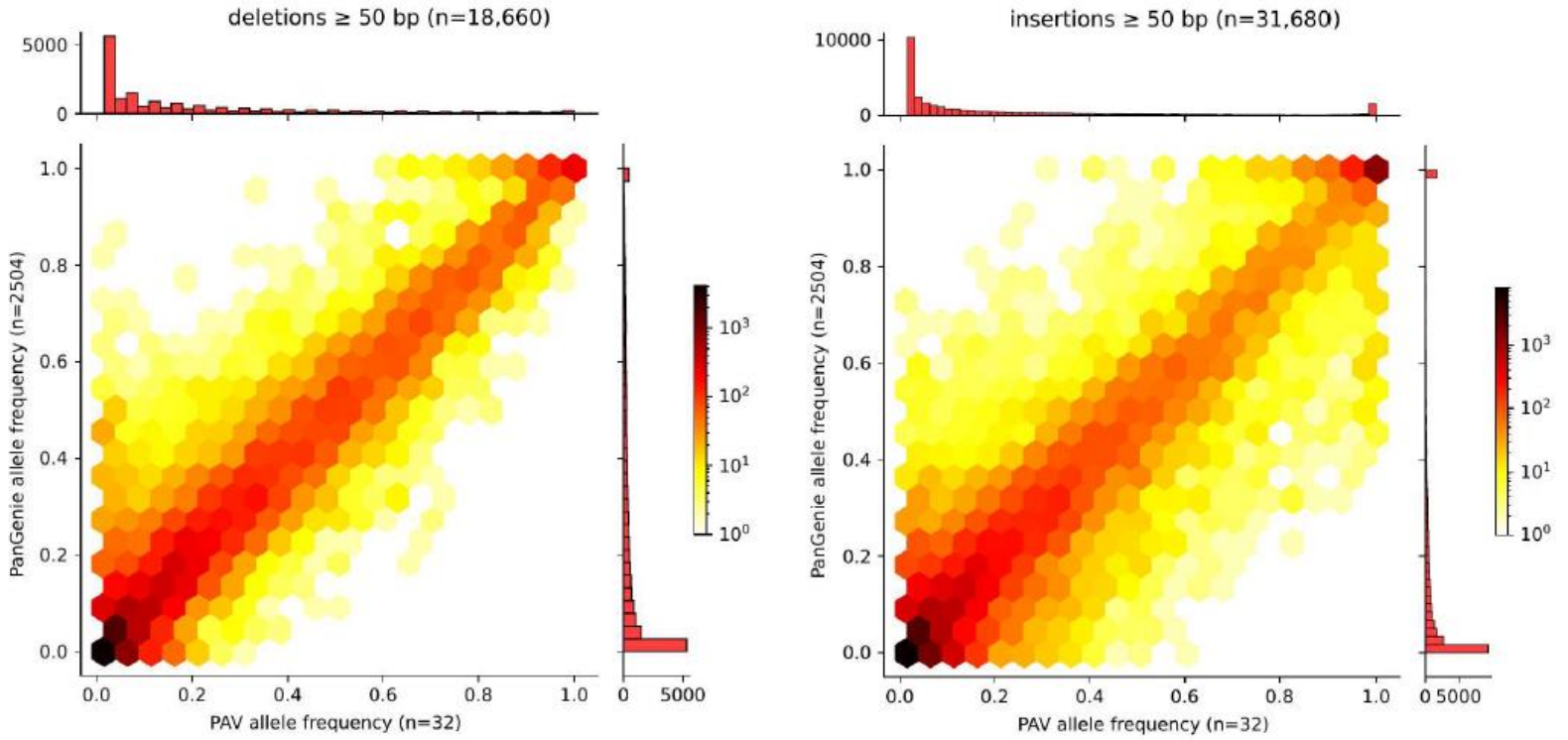


Fig. S29. Allele frequencies of SVs in the lenient callset

For PanGenie, allele frequencies were computed based on the genotypes of all 2,504 unrelated samples. The PAV allele frequencies were computed based on all 64 assemblies. Only SVs (≥50 bp) contained in our lenient callset (cutoff -0.5, n=50,340) were considered.

Figure 5: SV genotyping and eQTL analysis

Added value from graph-based genotyping into short read WGS data

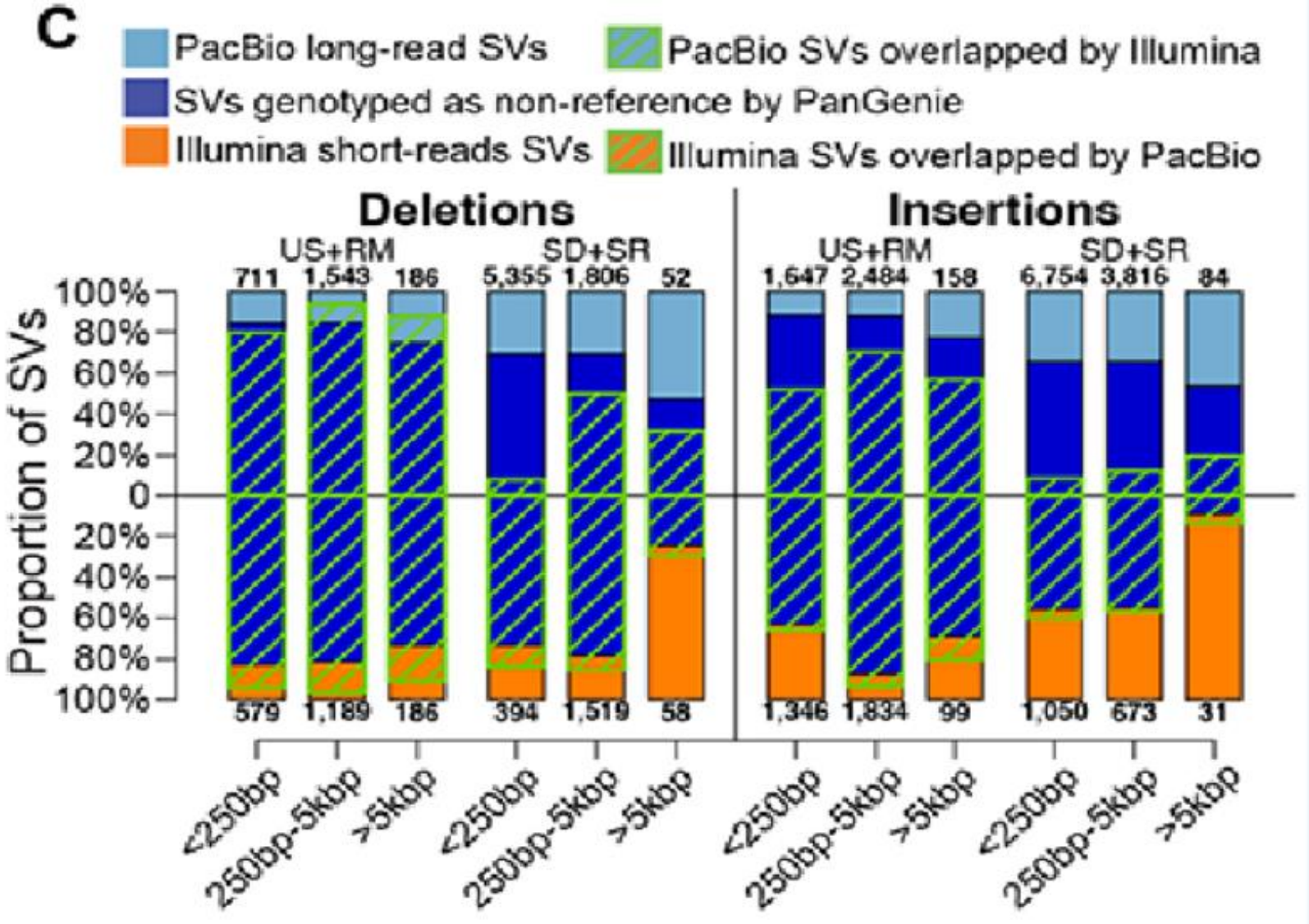
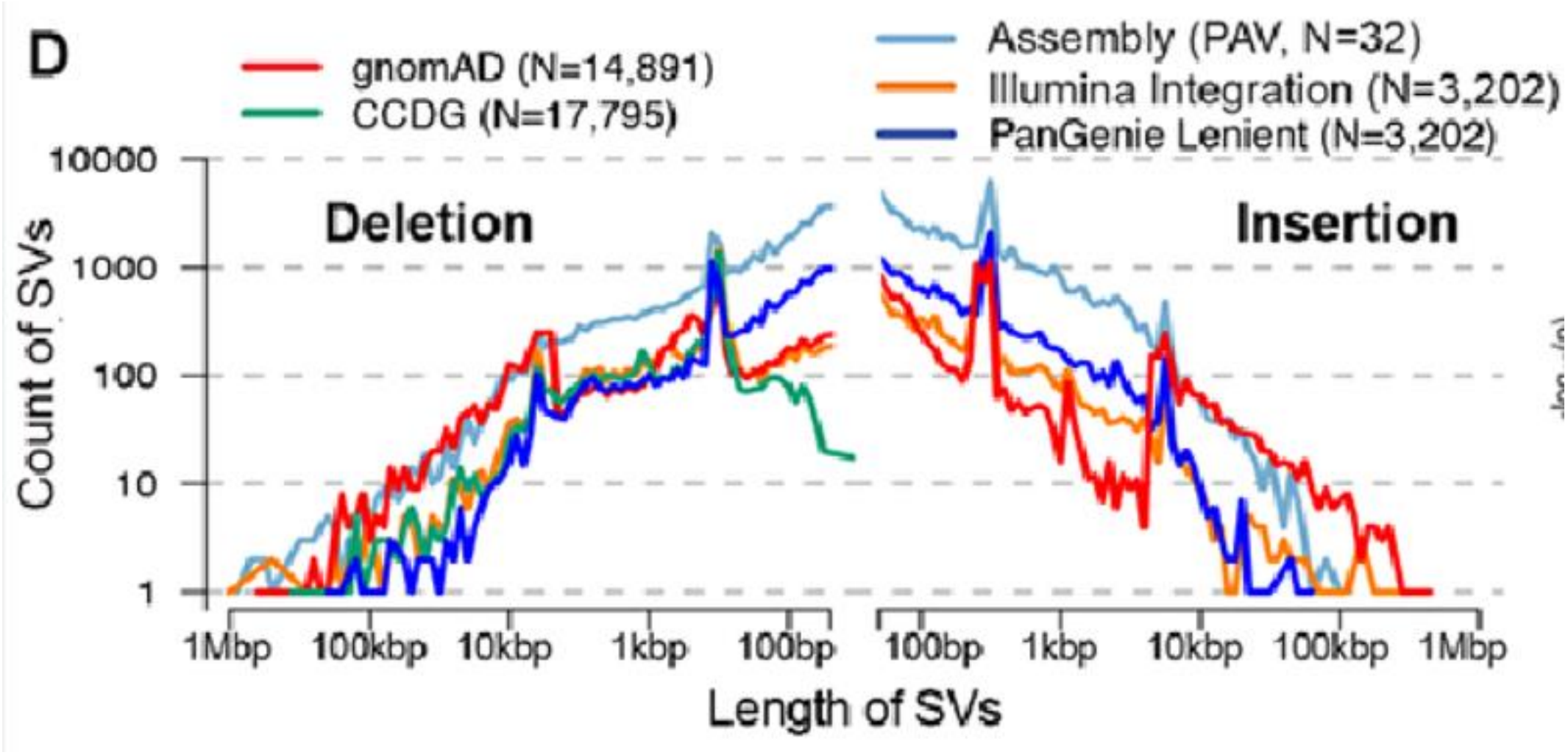


Figure 5: SV genotyping and eQTL analysis

Added value from graph-based genotyping into short read WGS data

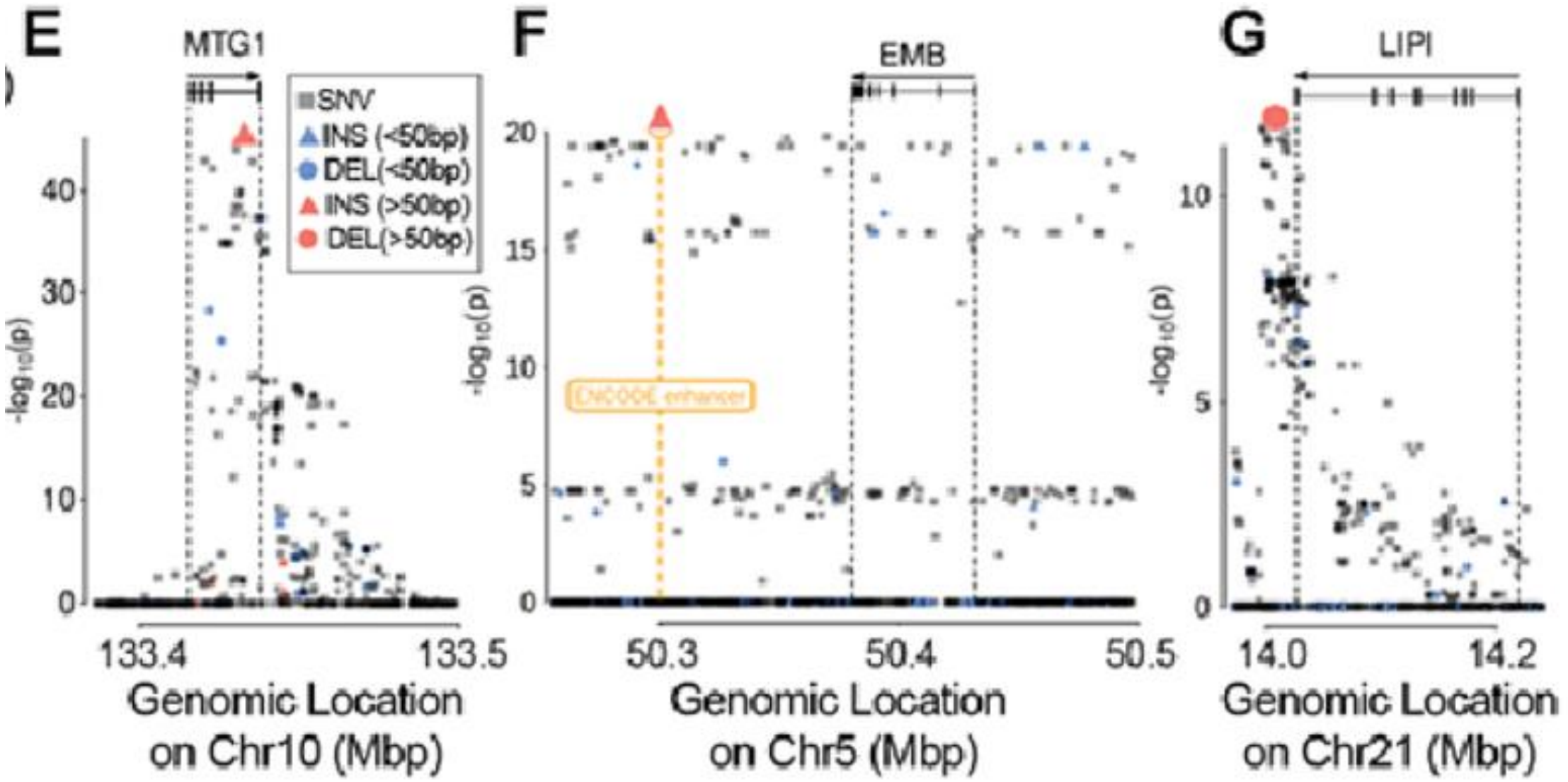


*突变基因数据来自常见疾病基因组学中心（CCDG）和基因组聚集数据库（gnomAD）

Figure 5: SV genotyping and eQTL analysis

*quantitative trait loci (eQTL)

QTL analyses



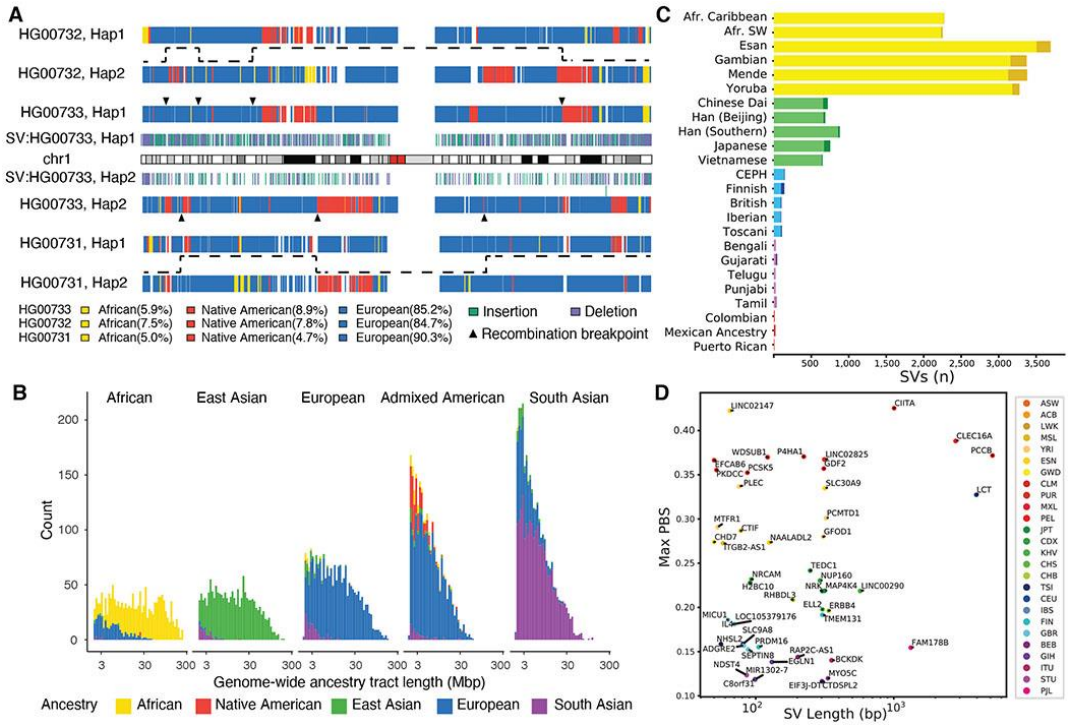

A
HG00732, Hap1
HG00732, Hap2
HG00733, Hap1
SV:HG00733, Hap1
chr1
SV:HG00733, Hap2
HG00733, Hap2
HG00731, Hap1
HG00731, Hap2
HG00733 African(5.9%) Native American(8.9%) European(85.2%) Insertion Deletion
HG00732 African(7.5%) Native American(7.8%) European(84.7%)
HG00731 African(5.0%) Native American(4.7%) European(90.3%) Recombination breakpoint
C
Afr. Caribbean
Afr. SW
Esan
Gambian
Mende
Yoruba
Chinese Dai
Han (Beijing)
Han (Southern)
Japanese
Vietnamese
CEPH
Finnish
British
Iberian
Toscani
Bengali
Gujarati
Telugu
Punjabi
Tamil
Colombian
Mexican Ancestry
Puerto Rican
500 1,000 1,500 2,000 2,500 3,000 3,500
SVs (n)
B
African
East Asian
European
Admixed American
South Asian
Count
200
150
100
50
0
3 30 300
Genome-wide ancestry tract length (Mbp)
Ancestry African Native American East Asian European South Asian
D
Max PBS
0.40
0.35
0.30
0.25
0.20
0.15
0.10
LINC02147
CIITA
CLEC16A
PCCB
WDSUB1
P4HA1
LINC02825
GDF2
FCAB6
PCSK5
PKDCC
PLEC
SLC30A9
CT
PCMTD1
MTFR1
CTIF
CHD7
NAALADL2
GFOD1
TTGB2-AS1
TEDC1
NRCAM
NUP160
H2BC10
NRK
MAP4K4
LINC00290
RHBDL3
ELL2
ERBB4
MICU1
LOC105379176
TMEM131
SLC9A8
NHSL2
PRDM16
FAM178B
ADGRE2
SEPTIN8
RAP2C-AS1
EGLN1
BCKDK
NDST4
MIR1302-7
MYO5C
C8orf31
EIF3J-DT
LCTDSPL2
10^2
10^3
SV Length (bp)
ASW
ACB
LWK
MSL
YRI
ESN
GWD
CLM
PUR
MXL
PEL
JPT
CDX
KHV
CHS
CHB
TSI
CEU
IBS
FIN
GBR
BEB
GIH
ITU
STU
PJL

Figure 6

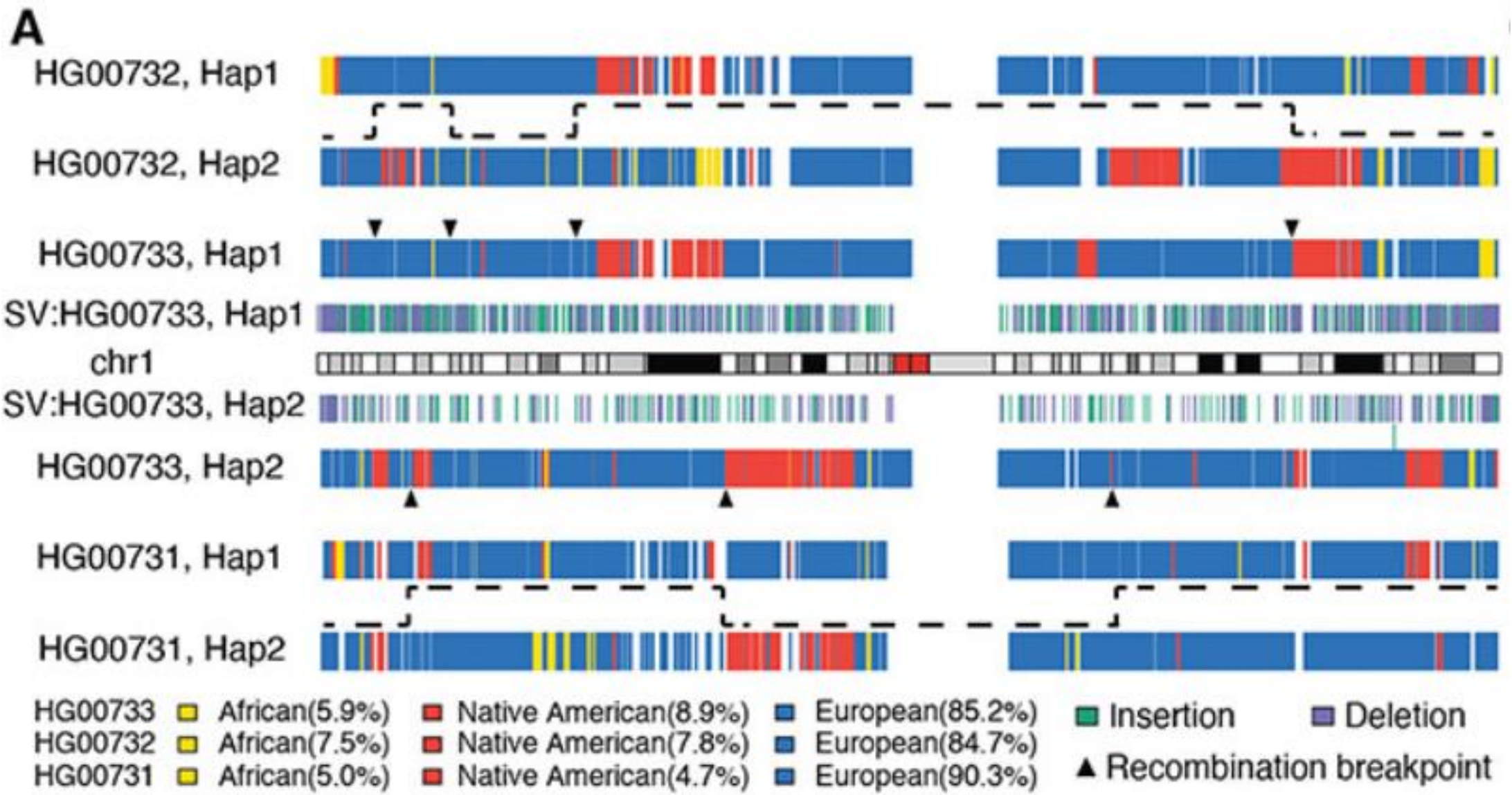

A
HG00732, Hap1
HG00732, Hap2
HG00733, Hap1
SV:HG00733, Hap1
chr1
SV:HG00733, Hap2
HG00733, Hap2
HG00731, Hap1
HG00731, Hap2
HG00733 African(5.9%) Native American(8.9%) European(85.2%)
HG00732 African(7.5%) Native American(7.8%) European(84.7%)
HG00731 African(5.0%) Native American(4.7%) European(90.3%)
Insertion
Deletion
Recombination breakpoint

Figure 6

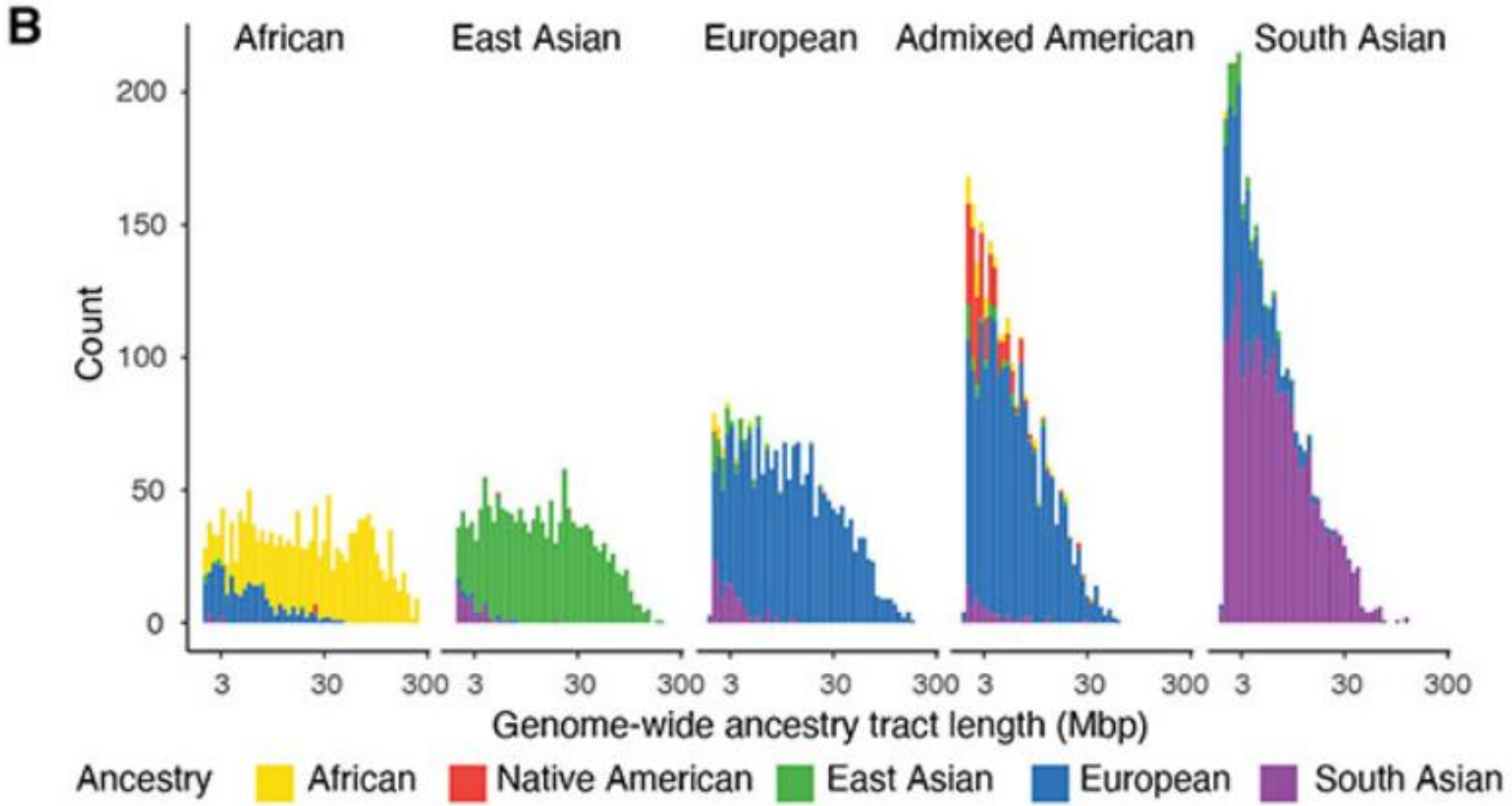

B
African
East Asian
European
Admixed American
South Asian
Count
200
150
100
50
0
3
30
300
Genome-wide ancestry tract length (Mbp)
Ancestry
African
Native American
East Asian
European
South Asian

Figure 6

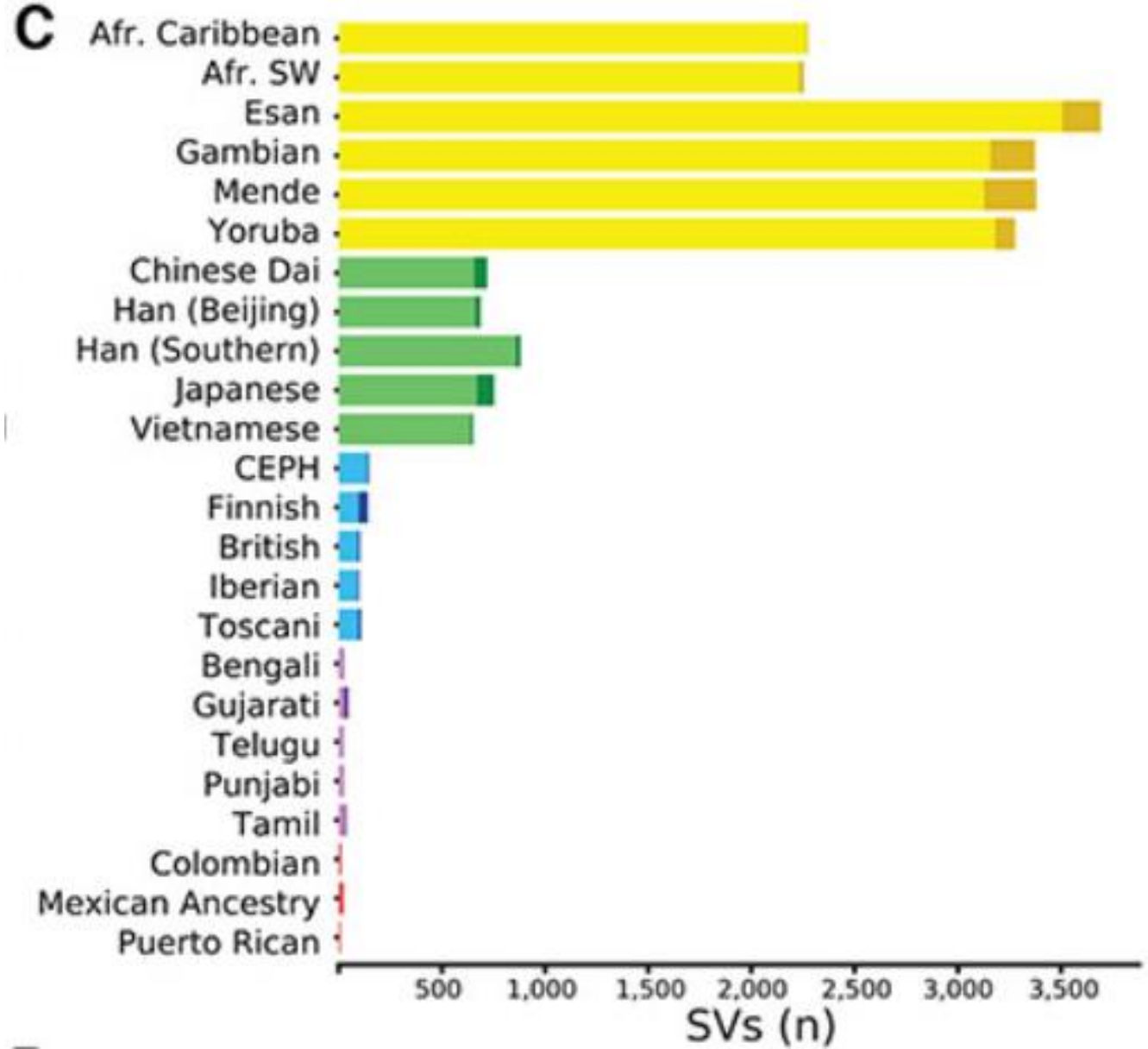
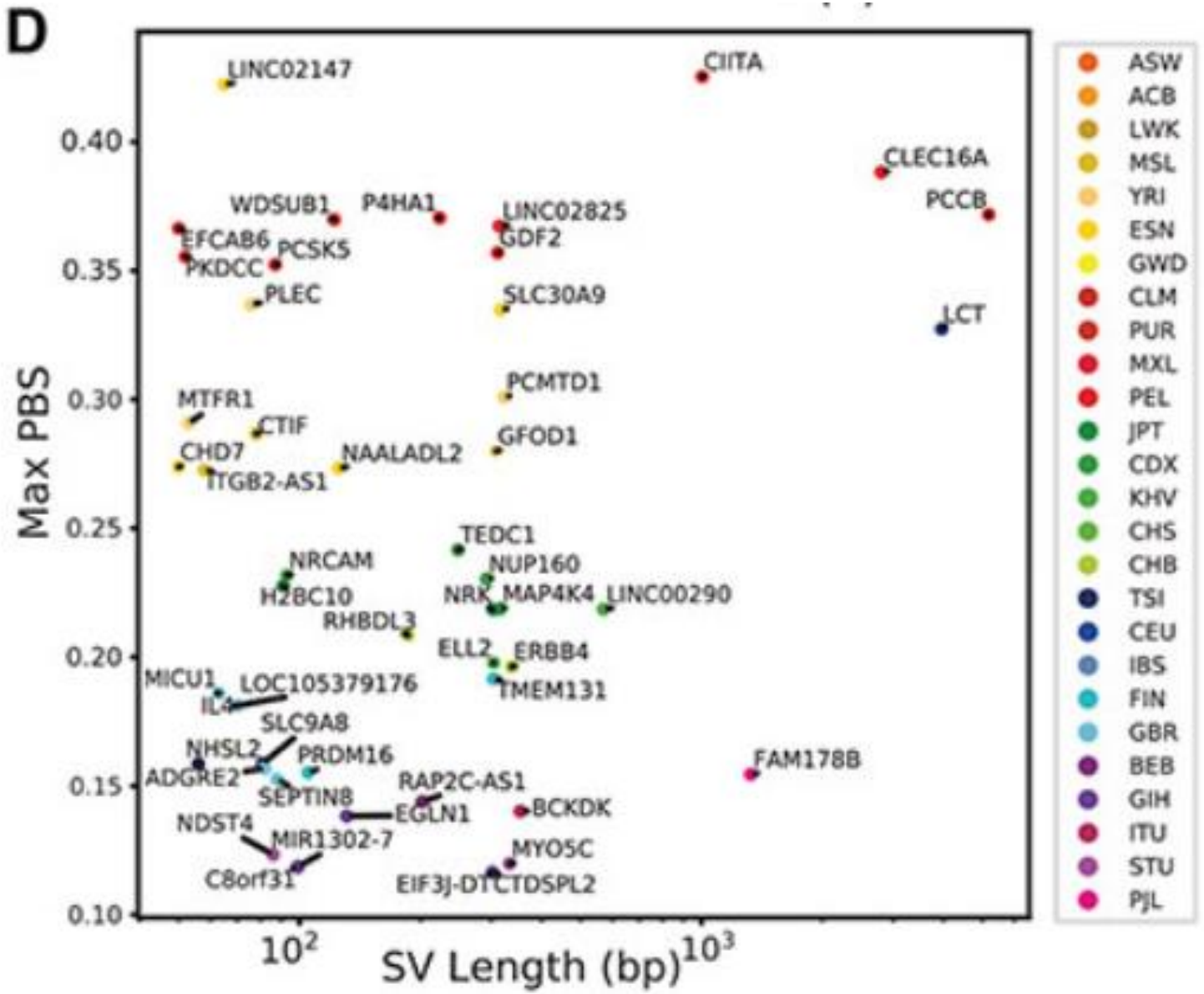

C
Afr. Caribbean
Afr. SW
Esan
Gambian
Mende
Yoruba
Chinese Dai
Han (Beijing)
Han (Southern)
Japanese
Vietnamese
CEPH
Finnish
British
Iberian
Toscani
Bengali
Gujarati
Telugu
Punjabi
Tamil
Colombian
Mexican Ancestry
Puerto Rican
500
1,000
1,500
2,000
2,500
3,000
3,500
SVs (n)

Figure 6


D
LINC02147
CIITA
CLEC16A
PCCB
WDSUB1
P4HA1
LINC02825
GDF2
EFCAB6
PCSK5
PKDCC
PLEC
SLC30A9
LCT
PCMTD1
MTFR1
CTIF
GFOD1
CHD7
NAALADL2
ITGB2-AS1
TEDC1
NRCAM
NUP160
H2BC10
NRK
MAP4K4
LINC00290
RHBDL3
ELL2
ERBB4
MICU1
LOC105379176
TMEM131
IL4
SLC9A8
NHSL2
PRDM16
FAM178B
ADGRE2
SEPTIN8
RAP2C-AS1
EGLN1
BCKDK
NDST4
MIR1302-7
MYO5C
C8orf31
EIF3J-DT
CTDSPL2
0.40
0.35
0.30
0.25
0.20
0.15
0.10
Max PBS
10^2
10^3
SV Length (bp)
ASW
ACB
LWK
MSL
YRI
ESN
GWD
CLM
PUR
MXL
PEL
JPT
CDX
KHV
CHS
CHB
TSI
CEU
IBS
FIN
GBR
BEB
GIH
ITU
STU
PJL

现有成果

1.通过比较组装的人类基因组直接发现所有形式的遗传变异（PAV）

2.有更多的SVs（63%）被指定为基于同源（>50 bp）的突变机制

3. 全基因组QTL扫描可以弥补分子和临床表型之间的差距，并作为由遗传变异类介导的功能效应的代理。

4.具有准确基因型的单倍体解决的SV也将促进SV的进化和种群遗传学研究，包括对反复突变率、种群分层和选择扫荡的估计。

仍需克服的挑战

1.长读组装仍未完全解决SD附近和内部更复杂的SV形式

2.需要新方法对多平行的VNTRs/STRs以及嵌入在较大重复序列（如SDs和中心粒）中的SVs进行基因分型