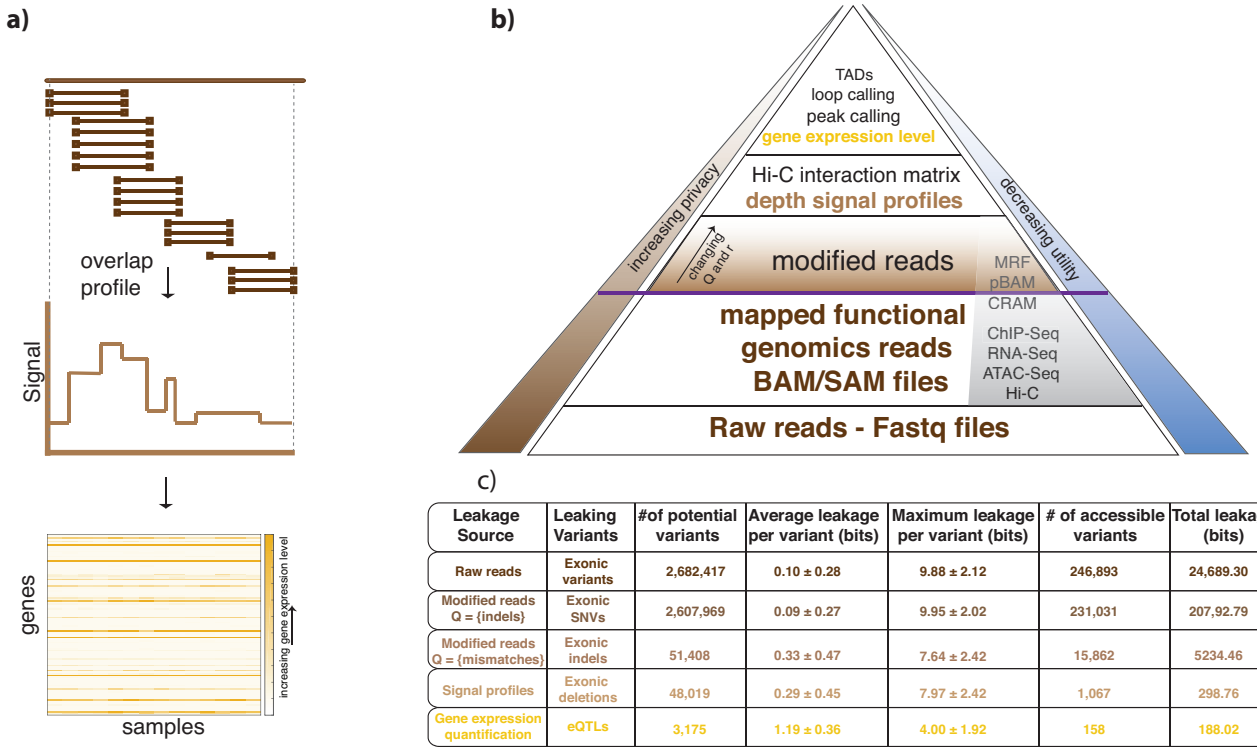


c)

| Leakage Source | Leaking Variants | #of potential variants | Average leakage per variant (bits) | Maximum leakage per variant (bits) | # of accessible variants | Total leakage (bits) |
|---|---|---|---|---|---|---|
| Raw reads | Exonic variants | 2,682,417 | 0.10 ± 0.28 | 9.88 ± 2.12 | 246,893 | 24,689.30 |
| Modified reads Q = {indels} | Exonic SNVs | 2,607,969 | 0.09 ± 0.27 | 9.95 ± 2.02 | 231,031 | 207,92.79 |
| Modified reads Q = {mismatches} | Exonic indels | 51,408 | 0.33 ± 0.47 | 7.64 ± 2.42 | 15,862 | 5234.46 |
| Signal profiles | Exonic deletions | 48,019 | 0.29 ± 0.45 | 7.97 ± 2.42 | 1,067 | 298.76 |
| Gene expression quantification | eQTLs | 3,175 | 1.19 ± 0.36 | 4.00 ± 1.92 | 158 | 188.02 |

Figure 1

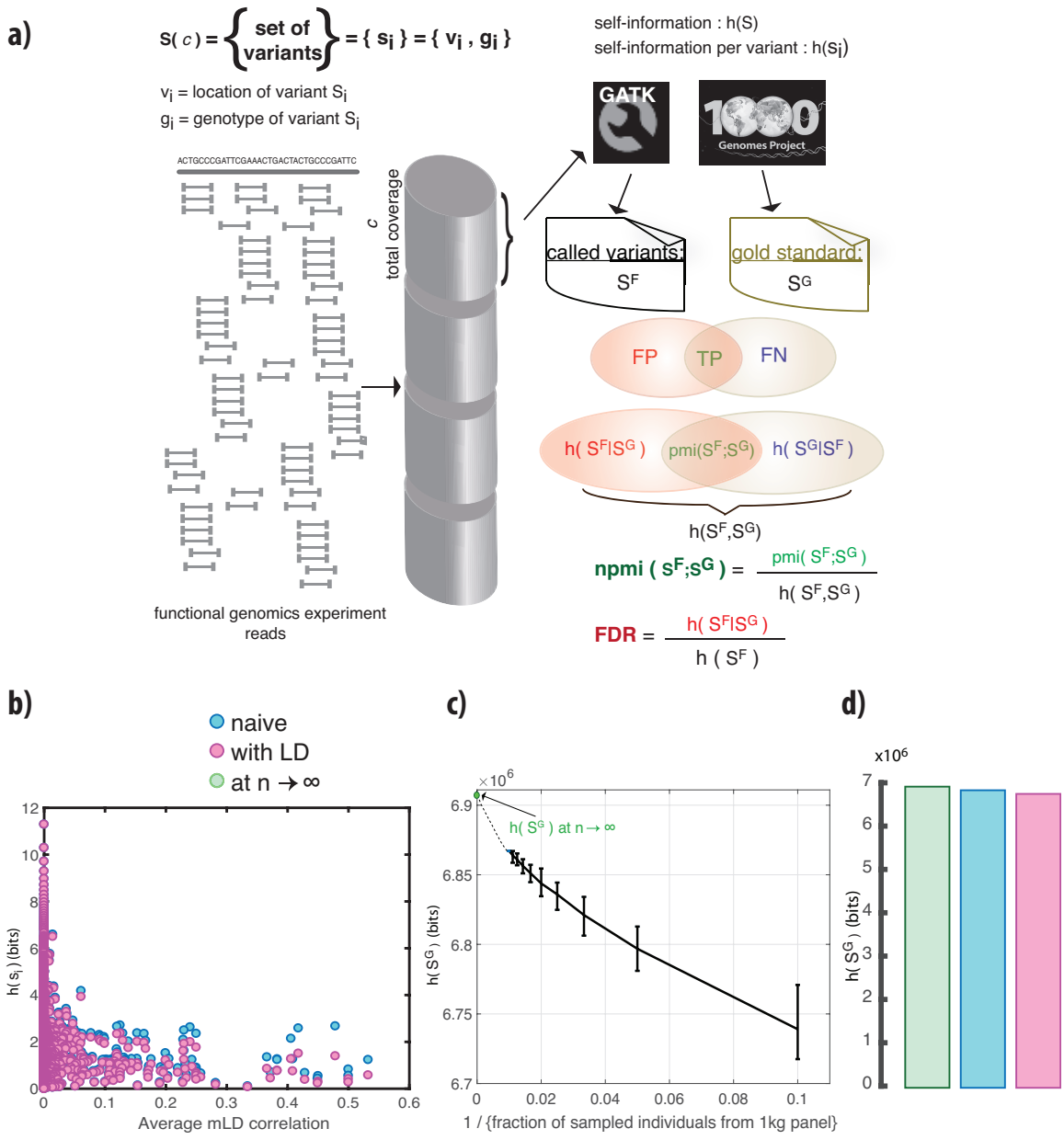
a)
S( c ) = { set of variants } = { s_i } = { v_i , g_i }
v_i = location of variant S_i
g_i = genotype of variant S_i
self-information : h(S)
self-information per variant : h(s_i)
GATK
1000 Genomes Project
ACTGCCCGATTCGAAACTGACTACTGCCCGATTC
c
total coverage
called variants: S^F
gold standard: S^G
FP
TP
FN
h( S^F|S^G )
pmi(S^F;S^G)
h( S^G|S^F )
h(S^F,S^G)
npmi ( s^F;s^G ) = pmi( S^F;S^G ) / h( S^F,S^G )
FDR = h( S^F|S^G ) / h ( S^F )
functional genomics experiment reads
b)
naive
with LD
at n → ∞
h( s_i ) (bits)
Average mLD correlation
c)
×10^6
h( S^G ) at n → ∞
h( S^G ) (bits)
1 / {fraction of sampled individuals from 1kg panel}
d)
x10^6
h( S^G ) (bits)

Figure 2

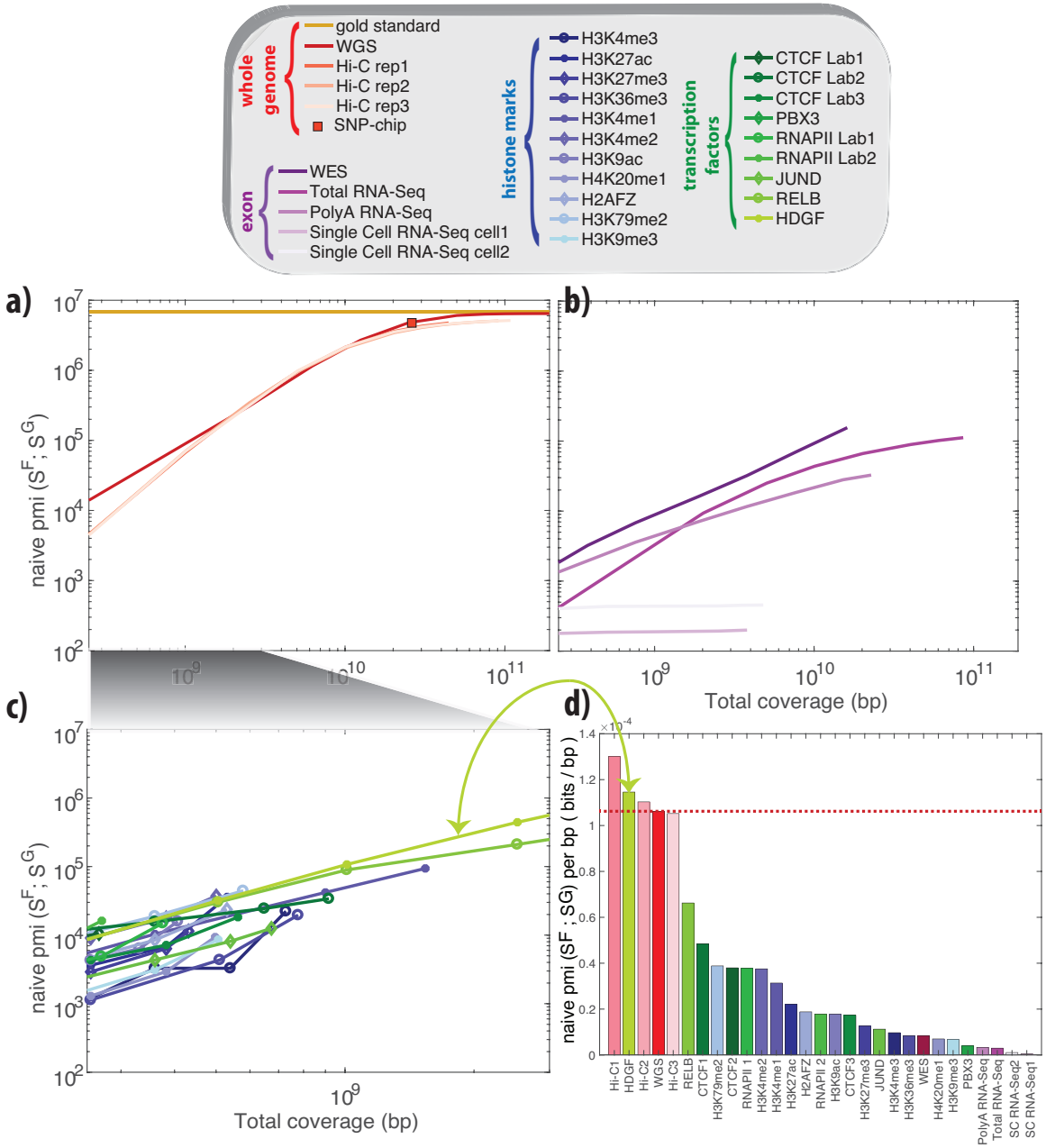
whole genome
gold standard
WGS
Hi-C rep1
Hi-C rep2
Hi-C rep3
SNP-chip
exon
WES
Total RNA-Seq
PolyA RNA-Seq
Single Cell RNA-Seq cell1
Single Cell RNA-Seq cell2
histone marks
H3K4me3
H3K27ac
H3K27me3
H3K36me3
H3K4me1
H3K4me2
H3K9ac
H4K20me1
H2AFZ
H3K79me2
H3K9me3
transcription factors
CTCF Lab1
CTCF Lab2
CTCF Lab3
PBX3
RNAPII Lab1
RNAPII Lab2
JUND
RELB
HDGF
a)
b)
c)
d)
naive pmi (S^F ; S^G)
Total coverage (bp)
naive pmi (SF ; SG) per bp ( bits / bp )
Hi-C1
HDGF
Hi-C2
WGS
Hi-C3
RELB
CTCF1
H3K79me2
CTCF2
RNAPII 1
H3K4me2
H3K4me1
H3K27ac
H2AFZ
RNAPII 2
H3K9ac
CTCF3
H3K27me3
JUND
H3K4me3
H3K36me3
WES
H4K20me1
H3K9me3
PBX3
PolyA RNA-Seq
Total RNA-Seq
SC RNA-Seq2
SC RNA-Seq1

Figure 3

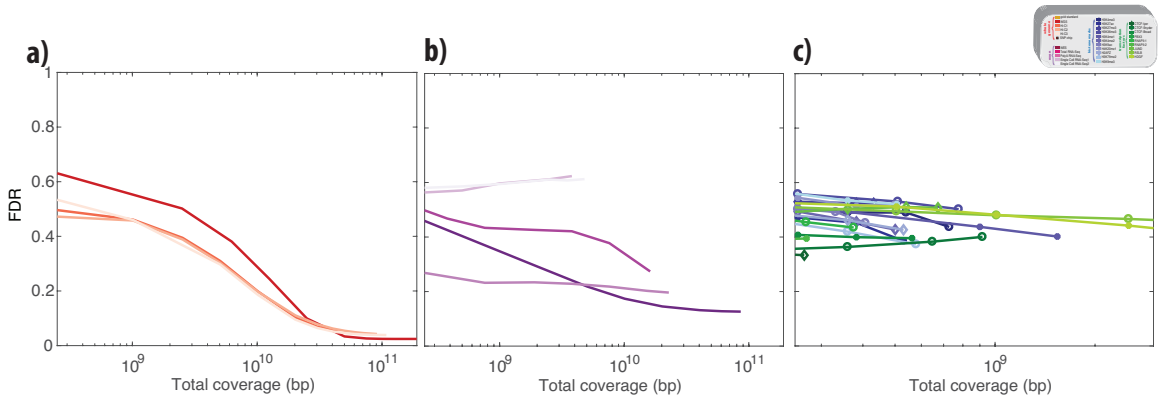
a)
b)
c)
FDR
1
0.8
0.6
0.4
0.2
0
$10^9$
$10^{10}$
$10^{11}$
Total coverage (bp)
Total coverage (bp)
Total coverage (bp)

Figure 4

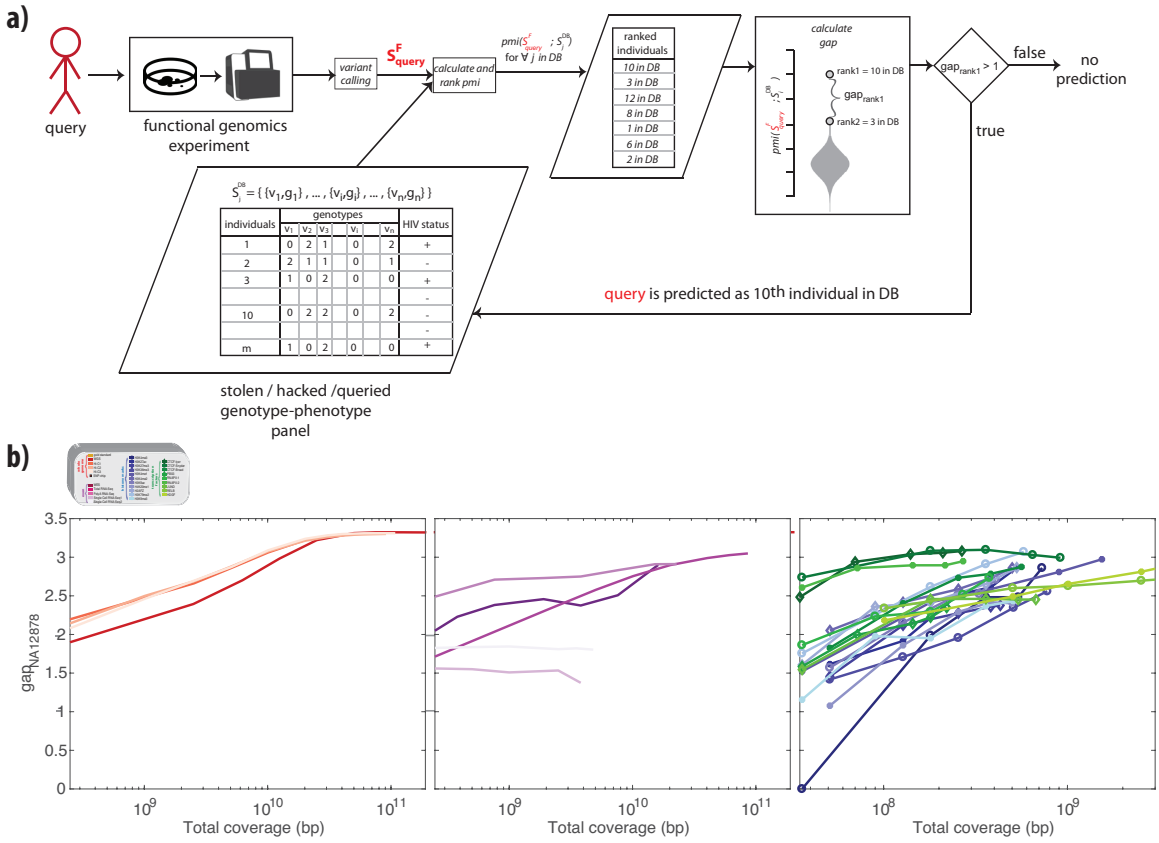
a)
query
functional genomics experiment
variant calling
calculate and rank pmi
ranked individuals
10 in DB
3 in DB
12 in DB
8 in DB
1 in DB
6 in DB
2 in DB
calculate gap
rank1 = 10 in DB
rank2 = 3 in DB
false
no prediction
true
query is predicted as 10th individual in DB
individuals
genotypes
HIV status
stolen / hacked /queried genotype-phenotype panel
b)
Total coverage (bp)
Total coverage (bp)
Total coverage (bp)

Figure 5

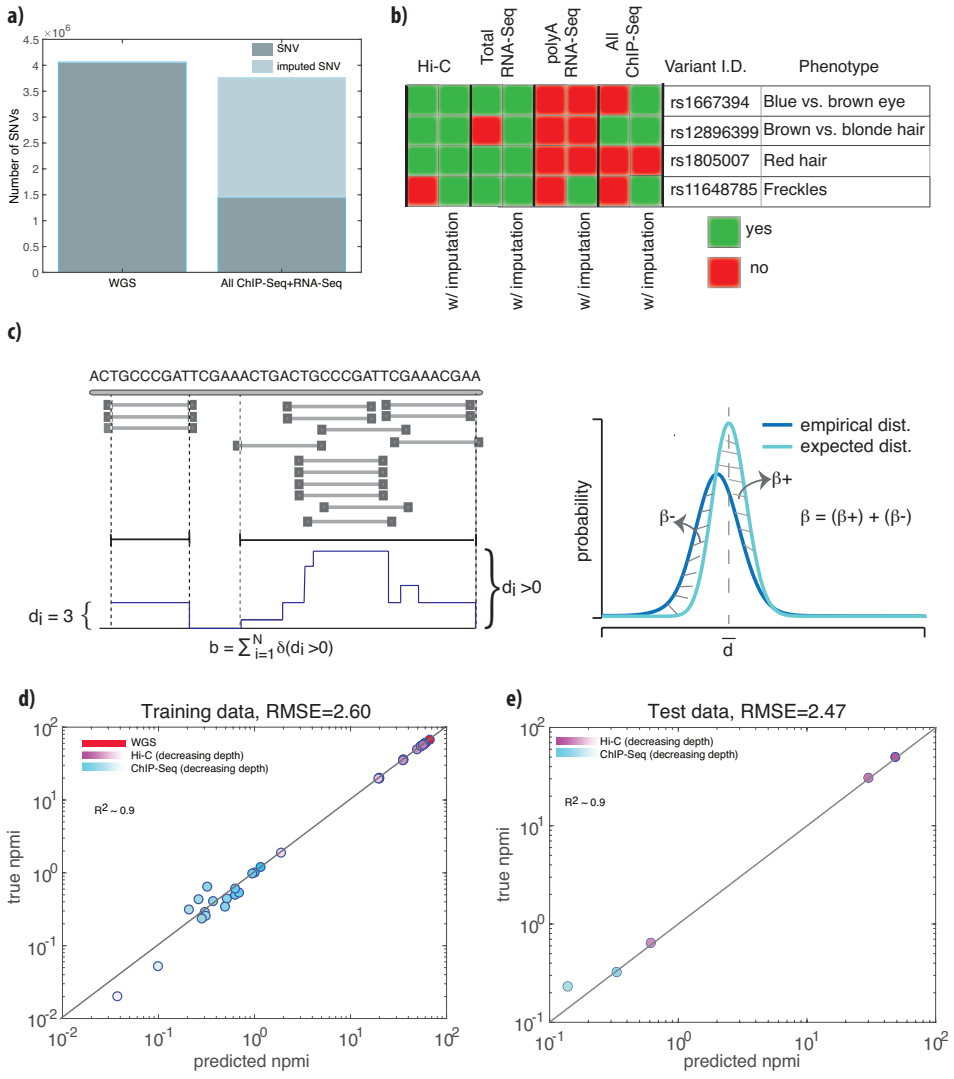
a)
Number of SNVs
SNV
imputed SNV
WGS
All ChIP-Seq+RNA-Seq
b)
Hi-C
Total RNA-Seq
polyA RNA-Seq
All ChIP-Seq
Variant I.D.
Phenotype
rs1667394
Blue vs. brown eye
rs12896399
Brown vs. blonde hair
rs1805007
Red hair
rs11648785
Freckles
w/ imputation
yes
no
c)
ACTGCCCGATTCGAAACTGACTGCCCGATTCGAAACGAA
di = 3
di >0
b = ∑ N i=1 δ(di >0)
empirical dist.
expected dist.
probability
β+
β-
β = (β+) + (β-)
d
d)
Training data, RMSE=2.60
WGS
Hi-C (decreasing depth)
ChIP-Seq (decreasing depth)
R2 ~ 0.9
true npmi
predicted npmi
e)
Test data, RMSE=2.47
Hi-C (decreasing depth)
ChIP-Seq (decreasing depth)
R2 ~ 0.9
true npmi
predicted npmi

Figure 6

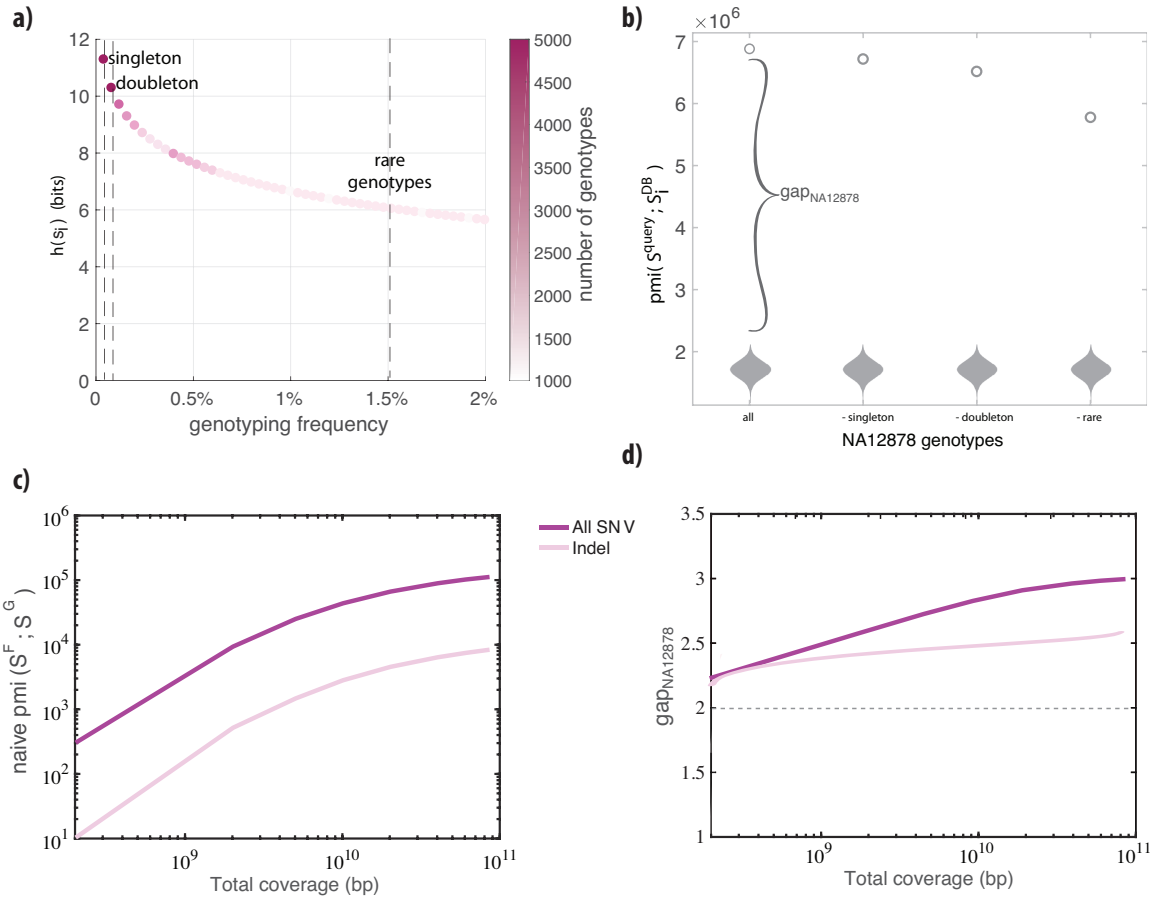
a)
singleton
doubleton
rare
genotypes
h(sj) (bits)
genotyping frequency
number of genotypes
b)
×10^6
gap_NA12878
all
- singleton
- doubleton
- rare
NA12878 genotypes
c)
All SNV
Indel
naive pmi (S^F ; S^G)
Total coverage (bp)
d)
gap_NA12878
Total coverage (bp)

Figure 7

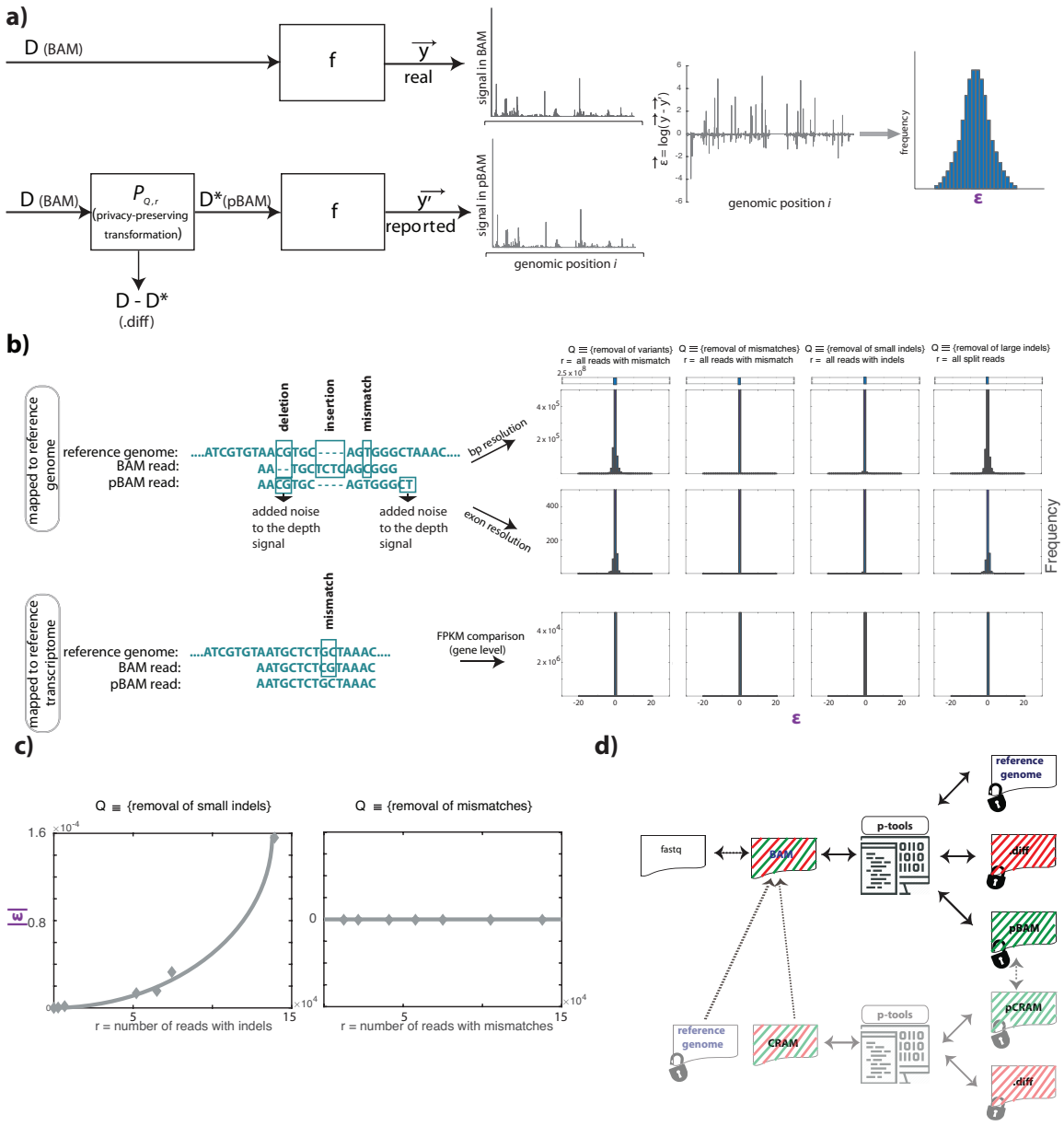
a)
D (BAM)
f
real
signal in BAM
D (BAM)
D*(pBAM)
(privacy-preserving transformation)
reported
signal in pBAM
genomic position i
D - D*
(.diff)
frequency
b)
mapped to reference genome
reference genome:
BAM read:
pBAM read:
deletion
insertion
mismatch
bp resolution
exon resolution
added noise to the depth signal
mapped to reference transcriptome
FPKM comparison (gene level)
Frequency
c)
Q ≡ {removal of small indels}
Q ≡ {removal of mismatches}
r = number of reads with indels
r = number of reads with mismatches
d)
fastq
BAM
p-tools
reference genome
diff
pBAM
pCRAM
CRAM

Figure 8