

Variant Calling and Filtering for INDELs

Erik Garrison

SeqShop @ University of Michi

Overview

1. What are the causes of insertion/deletion (INDEL) mutations?
2. How do we detect INDELs using resequencing methods?
3. Bayesian variant calling. Haplotype-based calling.

An INDEL

A mutation that results from the gain or loss of sequence.

AATTAGCCATTA

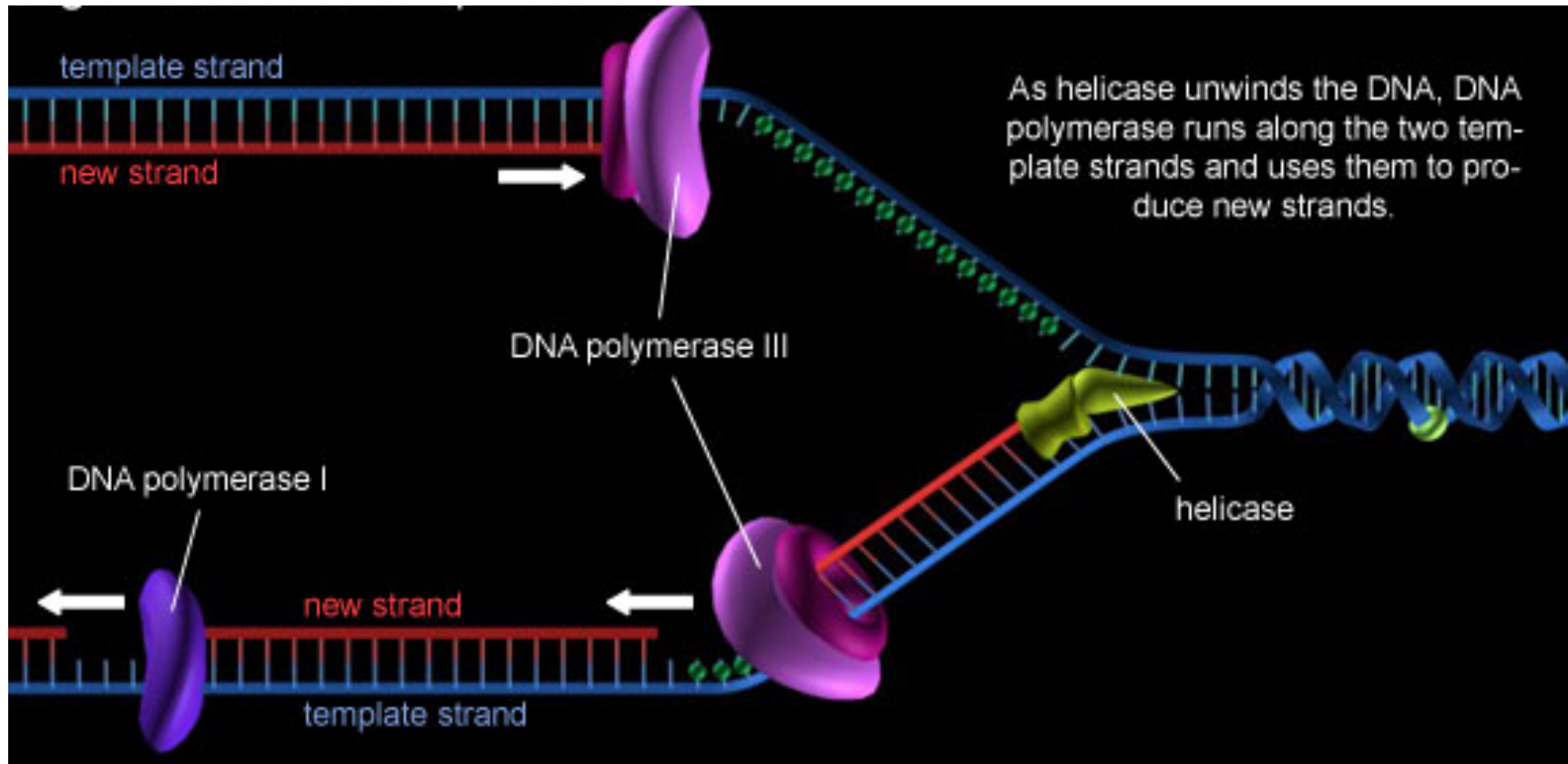
AATTA--CATTA

INDEL genesis

A number of processes are known to generate insertions and deletions in the process of DNA replication:

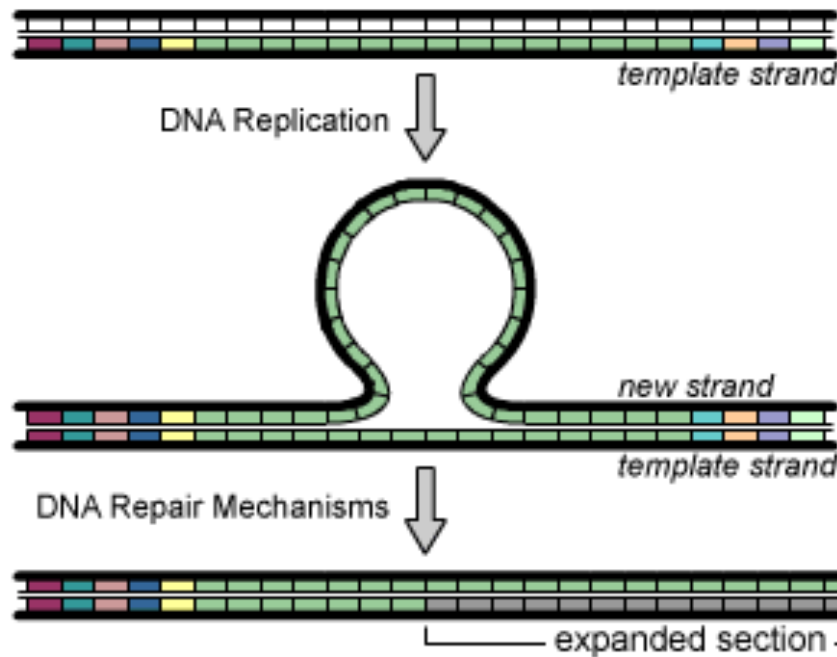
- Replication slippage
- Double-stranded break repair
- Structural variation (e.g. mobile element insertions, CNVs)

DNA replication



Polymerase *slippage*

A) Slippage Event



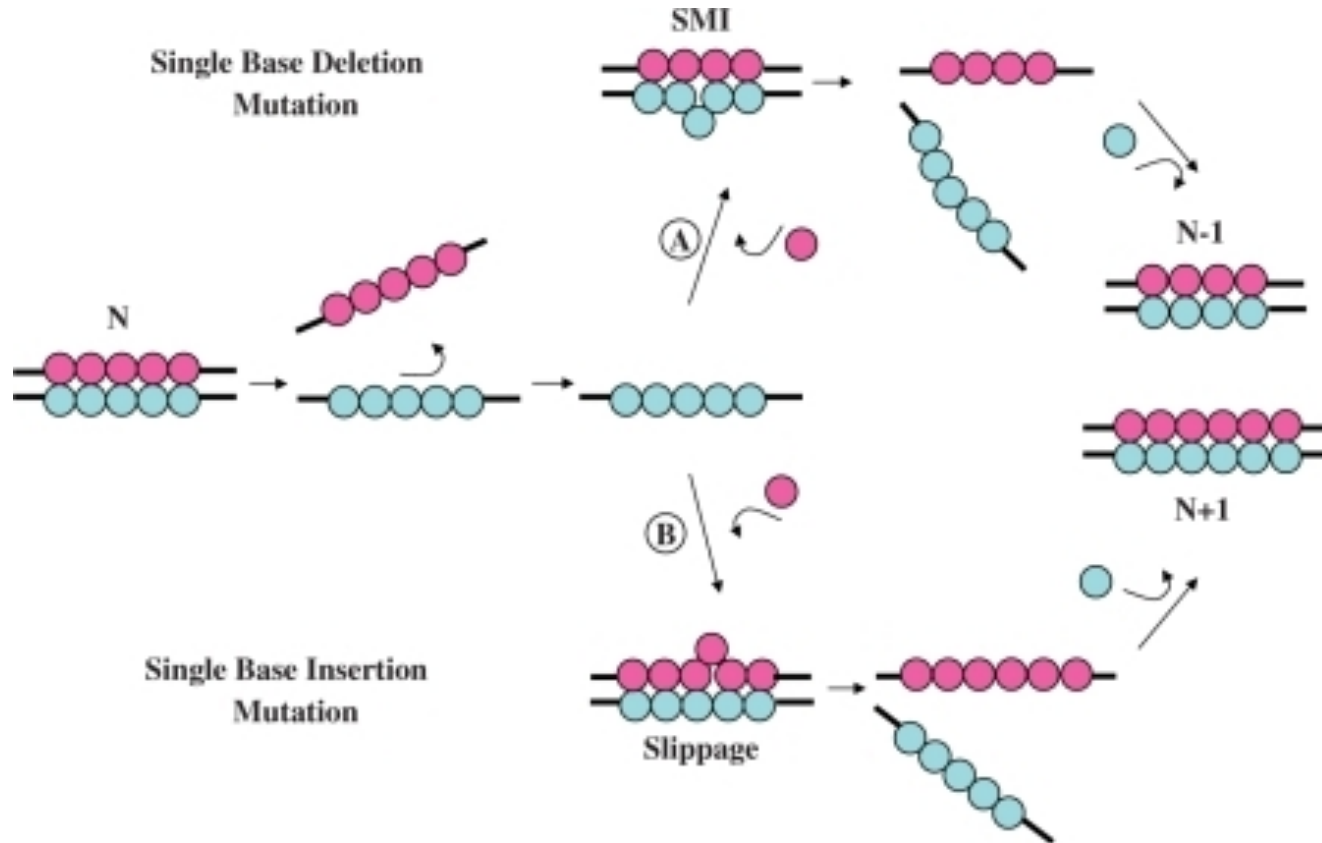
(A) During replication, polymerase slippage and subsequent reattachment may cause a bubble to form in the new strand. Slippage is thought to occur in sections of DNA with repeated patterns of bases (such as CAG), represented here by matching colors. Then, DNA repair mechanisms realign the template strand with the new strand and the bubble is straightened out. The resulting double helix is thus expanded.

B) No Slippage



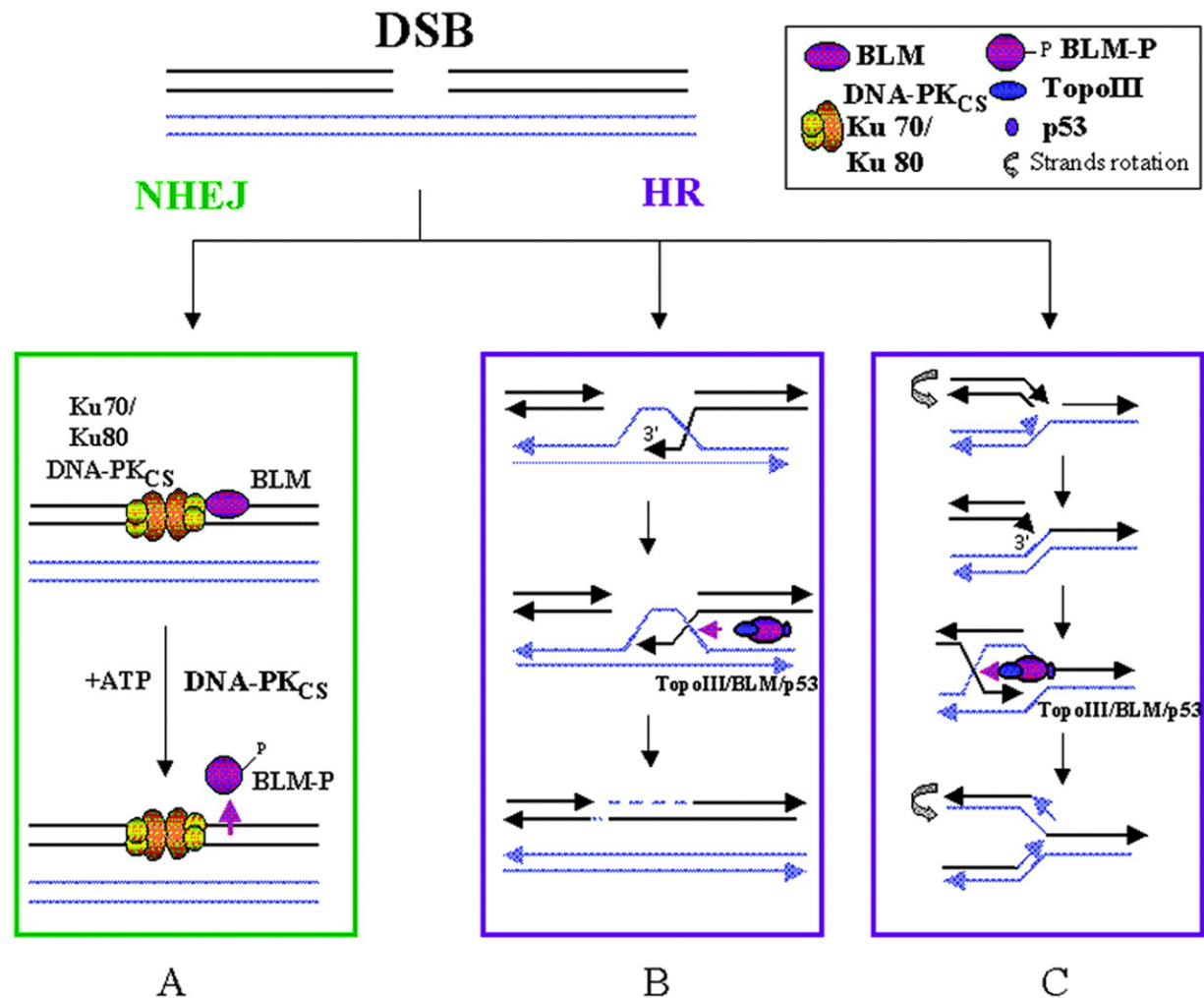
(B) Polymerase slippage, as theorized, cannot occur in DNA without repeating patterns of bases.

Insertions and deletions via slippage



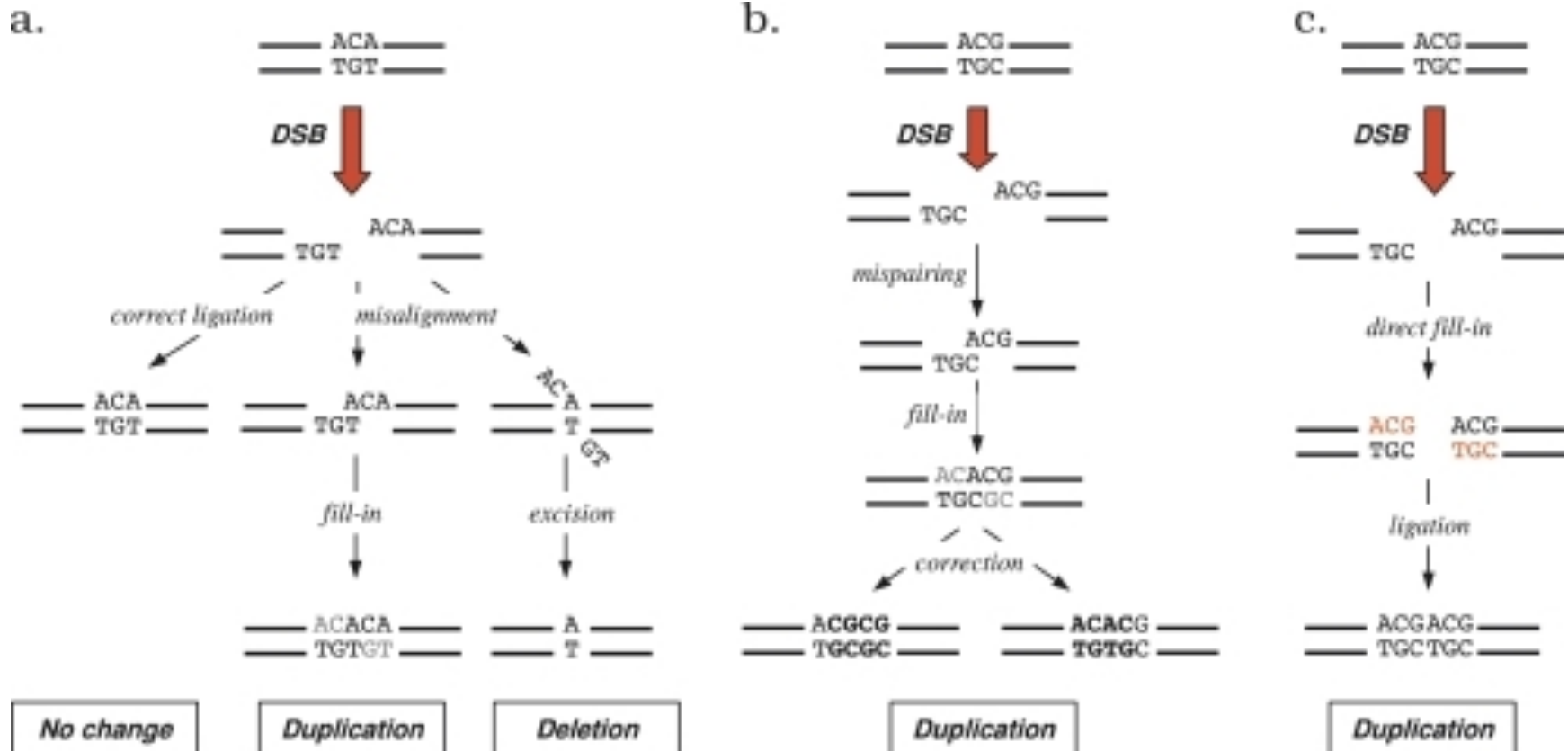
Energetic signatures of single base bulges: thermodynamic consequences and biological implications. Minetti CA, Remeta DP, Dickstein R, Breslauer KJ - Nucleic Acids Res. (2009)

Double-stranded break repair



Possible anti-recombinogenic role of Bloom's syndrome helicase in double-strand break processing. doi: [10.1093/nar/gkg834](https://doi.org/10.1093/nar/gkg834)

NHEJ-derived indels



DNA Slippage Occurs at Microsatellite Loci without Minimal Threshold Length in Humans: A Comparative Genomic Approach. Leclercq S, Rivals E, Jarne P - Genome Biol Evol (2010)

Calling INDEL variation

Design a process to detect indels from alignment data.

Take ~5 minutes with a partner (or two) and write down the steps your algorithm would execute to detect indels.

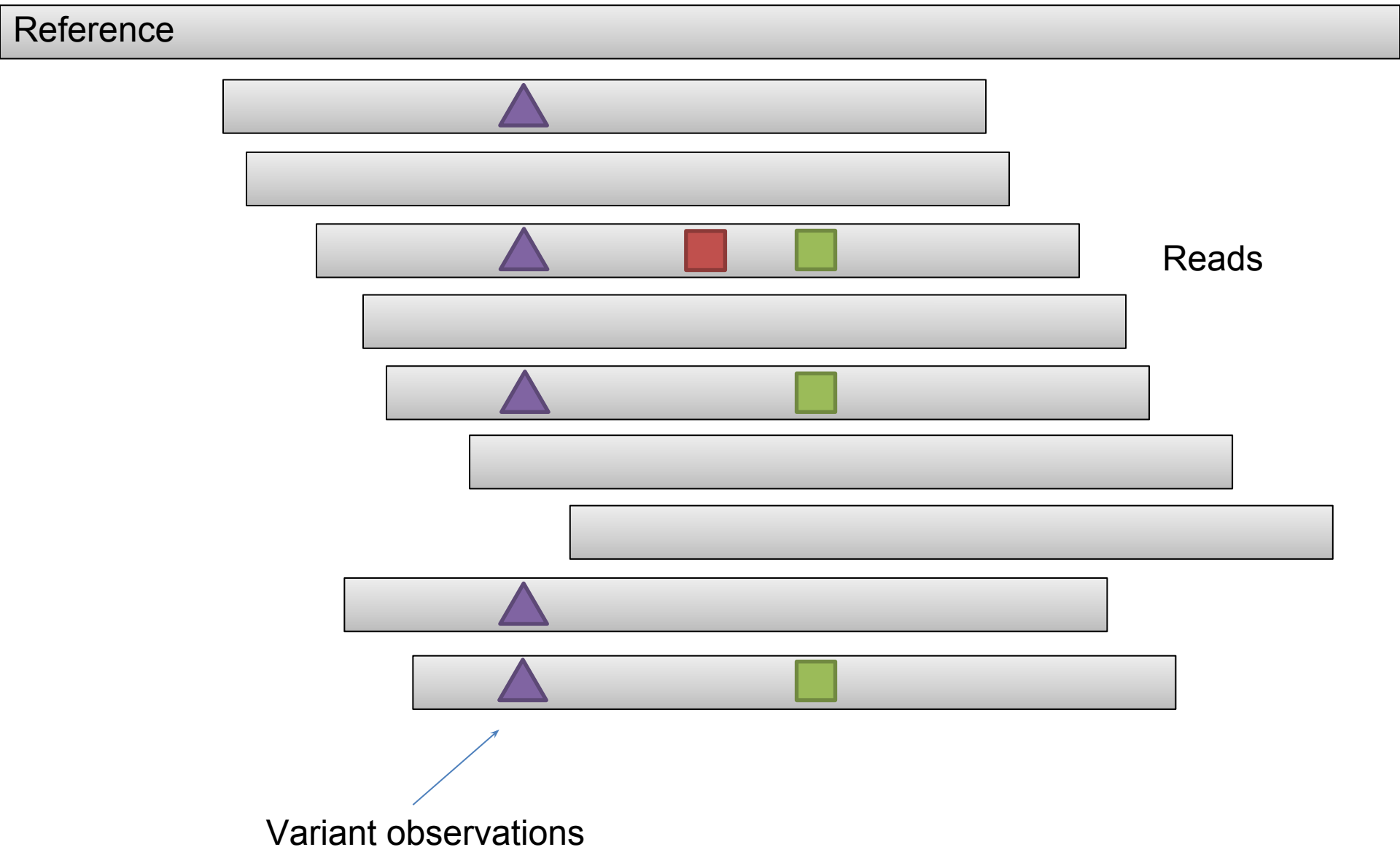
Genome (FASTA)

alignment and variant calling

```
##source=mutatrix population genome simulator
##seed=1373972756
##reference=chr0.fa
##phasing=true
##commandLine=mutatrix -S sample -p 2 -n 100 chr0.fa
##filters="AC > 0"
##INFO=ID=AC,Type=Number,A=Type=String,Descriptions="Type of each allele (snp, ins, del, mnp, complex)">
##INFO=ID=AN,Number=1,Type=Integer,Descriptions="Number of alternate alleles">
##INFO=ID=LEN,Number=A,Type=Integer,Descriptions="Length of each alternate allele">
##INFO=ID=CD,CROSSAT,Number=0,Type=Flag,Descriptions="Generate at a sequence repeat loci">
##FORMAT=AC=GT,Number=1,Type=String,Descriptions="Genotype">
##INFO=ID=AC,AC,Number=A,Type=Integer,Descriptions="Total number of alternate alleles in called genotypes">
##INFO=ID=AF,AF,Number=A,Type=Float,Descriptions="Estimated allele frequency in the range (0,1]">
##INFO=ID=AS,Number=1,Type=Integer,Descriptions="Number of samples with data">
##INFO=ID=AN,Number=1,Type=Integer,Descriptions="Total number of alleles in called genotypes"
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT sample1 sample2
chr0 1252 C A 99 AC=1;AF=0.25;AN=4;LEN=1;NA=1;NS=2;TYPE=snp
chr0 3646 T TC 99 AC=1;AF=0.25;AN=4;LEN=1;NA=1;NS=2;TYPE=ins
chr0 6283 C T 99 AC=1;AF=0.25;AN=4;LEN=1;NA=1;NS=2;TYPE=snp
chr0 7412 C T 99 AC=1;AF=0.25;AN=4;LEN=1;NA=1;NS=2;TYPE=snp
chr0 7935 T C 99 AC=1;AF=0.25;AN=4;LEN=1;NA=1;NS=2;TYPE=snp
chr0 8131 A T 99 AC=2;AF=0.5;AN=4;LEN=1;NA=1;NS=2;TYPE=snp
chr0 8682 AA TG 99 AC=1;AF=0.25;AN=4;LEN=1;NA=1;NS=2;TYPE=mnp
chr0 10926 T C 99 AC=1;AF=0.25;AN=4;LEN=1;NA=1;NS=2;TYPE=snp
chr0 11921 G GTT 99 AC=1;AF=0.25;AN=4;LEN=2;NA=1;NS=2;TYPE=ins
chr0 12955 T G 99 AC=1;AF=0.25;AN=4;LEN=1;NA=1;NS=2;TYPE=snp
chr0 13808 T TG 99 AC=1;AF=0.25;AN=4;LEN=1;NA=1;NS=2;TYPE=ins
chr0 15371 A T 99 AC=1;AF=0.25;AN=4;LEN=1;NA=1;NS=2;TYPE=snp
chr0 15407 A C 99 AC=1;AF=0.25;AN=4;LEN=1;NA=1;NS=2;TYPE=snp
chr0 16486 C G 99 AC=1;AF=0.25;AN=4;LEN=1;NA=1;NS=2;TYPE=snp
chr0 16563 T A 99 AC=1;AF=0.25;AN=4;LEN=1;NA=1;NS=2;TYPE=snp
chr0 16748 GTT G 99 AC=1;AF=0.25;AN=4;LEN=2;NA=2;NS=2;TYPE=del
chr0 17697 A G 99 AC=1;AF=0.25;AN=4;LEN=1;NA=1;NS=2;TYPE=snp
chr0 19548 A T 99 AC=1;AF=0.25;AN=4;LEN=1;NA=1;NS=2;TYPE=snp
chr0 20758 G A 99 AC=1;AF=0.25;AN=4;LEN=1;NA=1;NS=2;TYPE=snp
chr0 21532 T C 99 AC=1;AF=0.25;AN=4;LEN=1;NA=1;NS=2;TYPE=snp
chr0 22291 G T 99 AC=1;AF=0.25;AN=4;LEN=1;NA=1;NS=2;TYPE=snp
chr0 23193 C A 99 AC=1;AF=0.25;AN=4;LEN=1;NA=1;NS=2;TYPE=snp
chr0 23854 CTAA TTAA 99 AC=1;AF=0.25;AN=4;LEN=1;NA=1;NS=2;TYPE=mnp
chr0 24467 T A 99 AC=1;AF=0.25;AN=4;LEN=1;NA=1;NS=2;TYPE=snp
chr0 26108 G A 99 AC=1;AF=0.25;AN=4;LEN=1;NA=1;NS=2;TYPE=snp
chr0 29654 T A 99 AC=1;AF=0.25;AN=4;LEN=1;NA=1;NS=2;TYPE=snp
chr0 30062 T C 99 AC=1;AF=0.25;AN=4;LEN=1;NA=1;NS=2;TYPE=snp
chr0 31790 A G 99 AC=1;AF=0.25;AN=4;LEN=1;NA=1;NS=2;TYPE=snp
chr0 32792 T C 99 AC=1;AF=0.75;AN=4;LEN=1;NA=1;NS=2;TYPE=snp
chr0 33376 CC C 99 AC=2;AF=0.5;AN=4;LEN=1;NA=1;NS=2;TYPE=del
chr0 33403 T C 99 AC=2;AF=0.5;AN=4;LEN=1;NA=1;NS=2;TYPE=snp
chr0 33882 A G 99 AC=2;AF=0.5;AN=4;LEN=1;NA=1;NS=2;TYPE=snp
chr0 34454 C T 99 AC=4;AF=1;AN=4;LEN=1;NA=1;NS=2;TYPE=snp GT
chr0 34716 G A 99 AC=1;AF=0.25;AN=4;LEN=1;NA=1;NS=2;TYPE=snp
chr0 35404 T A 99 AC=1;AF=0.25;AN=4;LEN=1;NA=1;NS=2;TYPE=snp
chr0 36547 G A 99 AC=1;AF=0.25;AN=4;LEN=1;NA=1;NS=2;TYPE=snp
chr0 38015 T A 99 AC=1;AF=0.25;AN=4;LEN=1;NA=1;NS=2;TYPE=snp
chr0 38281 T C 99 AC=1;AF=0.25;AN=4;LEN=1;NA=1;NS=2;TYPE=snp
chr0 48047 A G 99 AC=1;AF=0.25;AN=4;LEN=1;NA=1;NS=2;TYPE=snp
chr0 49581 A G 99 AC=1;AF=0.25;AN=4;LEN=1;NA=1;NS=2;TYPE=snp
chr0 49601 G A 99 AC=1;AF=0.75;AN=4;LEN=1;NA=1;NS=2;TYPE=snp
chr0 49668 G A 99 AC=1;AF=0.75;AN=4;LEN=1;NA=1;NS=2;TYPE=snp
```

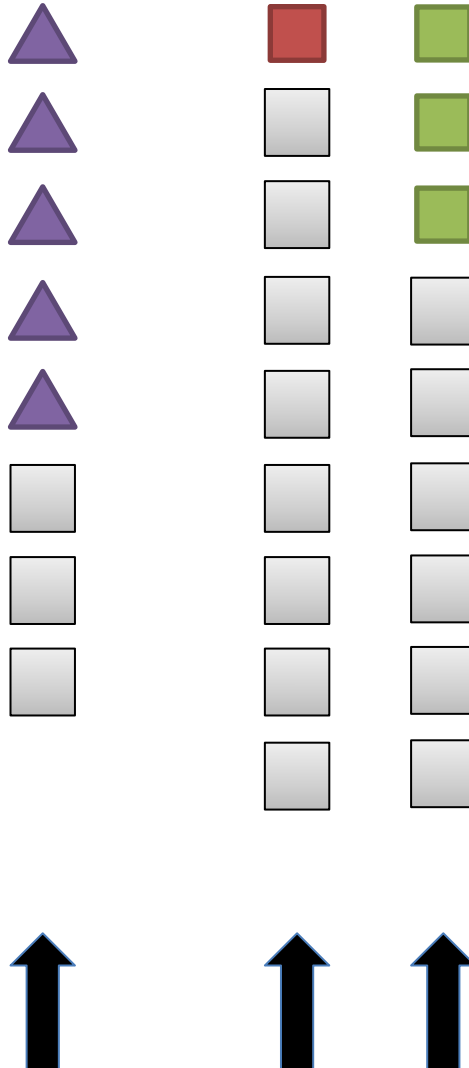
Variation (VCF)

Finding variation



Single-position calling

Reference



Haplotype information is lost.

INDEL normalization

Left alignment allows us to ensure that our representation is consistent across alignments and also variant calls.

CGTATGATCTAG**GCGCGC**TAGCTAGCTAGC ← Left aligned
CGTATGATCTA - - **GCGC**TAGCTAGCTAGC

CGTATGATCTAG**GCGCGC**TAGCTAGCTAGC
CGTATGATCTAG**C** - - **G**CTAGCTAGCTAGC

CGTATGATCTAG**GCGCGC**TAGCTAGCTAGC
CGTATGATCTAG**GCGC** - -TAGCTAGCTAGC

example: 1000G Phasel low coverage, chr15:81551110, ref:CTCTC alt:ATATA

[illegible]

Interpreted as 3 SNPs

[illegible]

Interpreted as microsatellite expansion/contraction

Complex allele realignment

example: 1000G Phase1 low coverage, chr20:708257, ref:AGC alt:CGA

ref: TATAGAGAGAGAGAGAGAGAGAGC GAGAGAGAGAGAGAGAGAGAGGGAGAGACGGAGTT
alt: TATAGAGAGAGAGAGAGAGAGC GAGAGAGAGAGAGAGAGAGAGAGGGAGAGACGGAGTT



ref: TATAGAGAGAGAGAGAGAGAGC--GAGAGAGAGAGAGAGAGAGGGAGAGACGGAGTT
alt: TATAGAGAGAGAGAGAGAG--CGAGAGAGAGAGAGAGAGAGGGAGAGACGGAGTT



Finding haplotype polymorphisms

Two
reads

AGAACCCAGTGCTCTTTCTGCT
AGAACCCAGTGGTCTTTCTGCT

a SNP

AGAACCCAGTGCTCTTTCTGCT
AGAACCCAGTGGTCTTTCTGCT

Their
alignmen
t

Another read
showing a SNP
on the same
haplotype as the
first

AGAACCCAGTGCTCTATCTGCT

AGAACCCAGTGCTCTATCTGCT
AGAACCCAGTGGTCTTTCTGCT

A variant
locus implied
by
alignments

Direct detection of haplotypes (FreeBayes)



Ref
Reads

	Variant Region	Variant Region
TACCGAT	CATTGGATCA	CGATTCC...GCATTGC
TACCGAT	CATTGGATCA	CGATTCC...GCATTGC
ACCGAT	TATTGCATCG	CGATTCC...GCATTGC
ACCGAT	CATTGGATCA	CGATTCC...GCATTGC
ACCGAT	TATTGGATCG	CGATTCC...GCATTGC
CCGAT	C-TTGGATCA	CGATTCC...GCATTGC
CCGAT	CATGGGATCA	CGATTCC...GCATTGC
...
		Variant Region
		AAAAAAAA-
		-AAAAAA-
		-AAAAAA-
		AAAAAA-A
		-AAAAAAA
		AAAAAAA-
		AAAAAAA A
		...

Observed Haplotypes

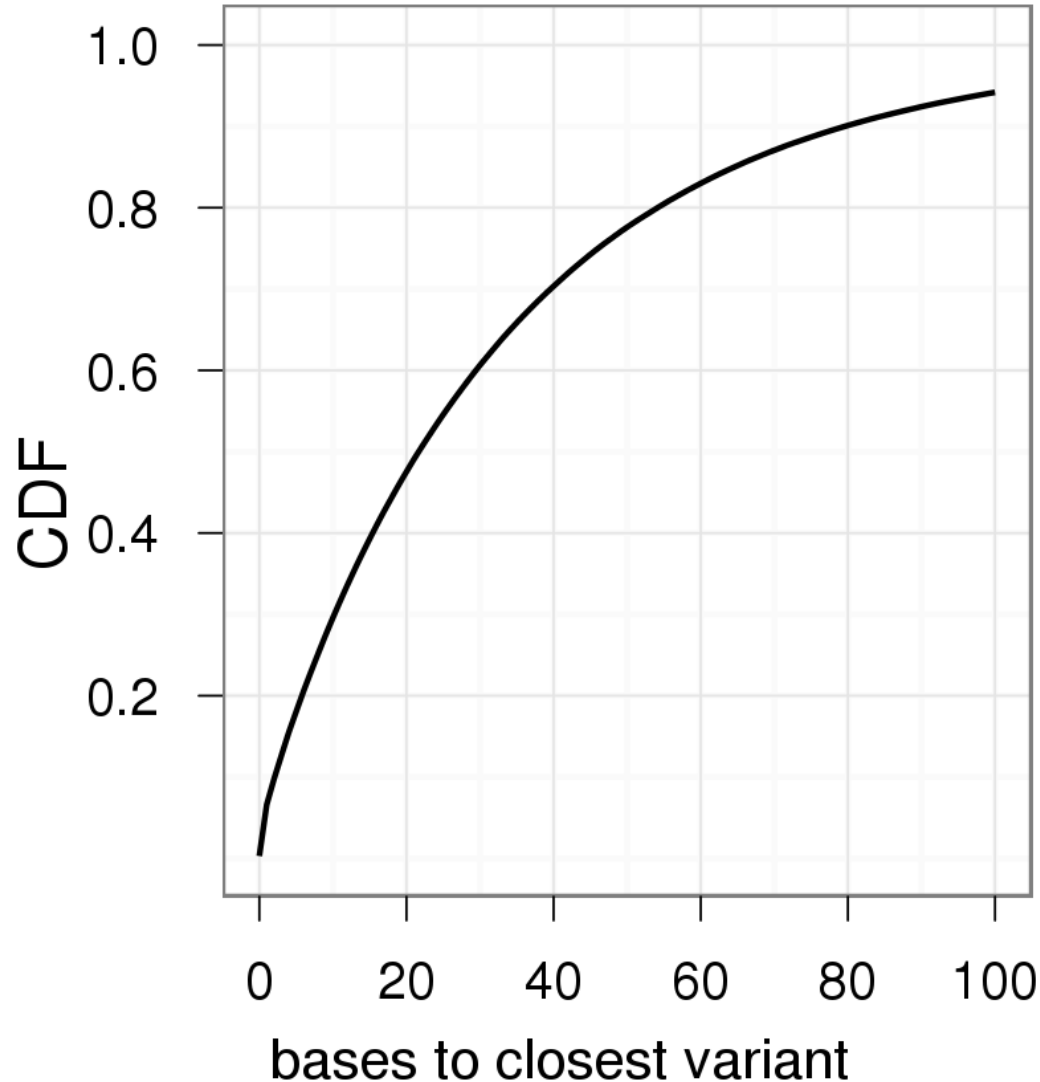
CATTGGATCA	x8
TATTGGATCG	x9
CTTGGATCA	x1
CATGGGATCA	x1
...	

(A) ₇	x10
(A) ₆	x7
(A) ₅	x1
(A) ₈	x1
...	

Why haplotypes?

- Variants cluster.
- This has functional significance.
- Observing haplotypes lets us be more certain of the local structure of the genome.
- We can improve the detection process itself by using haplotypes rather than point mutations.

Sequence variants cluster



In ~1000 individuals, $\frac{1}{2}$ of variants are within ~22bp of another variant.

Variance to mean ratio (VMR) = 1.4.

The functional effect of variants depends on other nearby variants on the same haplotype

reference: AGG GAG CTG
Arg Glu Leu

OTOF gene – mutations
cause profound recessive
deafness

apparent: AGG **T**AG CTG
Arg **Ter** ---

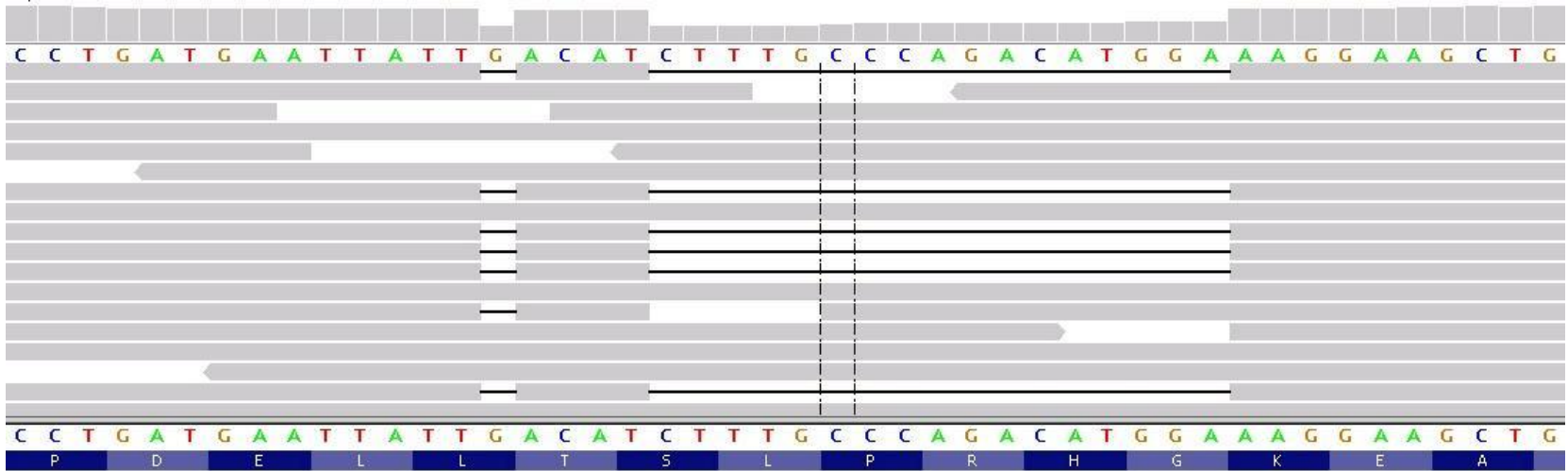
Apparent nonsense variant,
one YRI homozygote

actual: AGG **TT**G CTG
Arg **Leu** Leu

Actually a block substitution
that results in a missense
substitution

(Daniel MacArthur)

Importance of haplotype effects: frame-restoring indels



- Two apparent frameshift deletions in the *CASP8AP2* gene (one 17 bp, one 1 bp) on the same haplotype
- Overall effect is in-frame deletion of six amino acids

(Daniel MacArthur)

Frame-restoring indels in 1000 Genomes Phase I exomes

chr6:117113761, GPRC6A (~10% AF in 1000G)

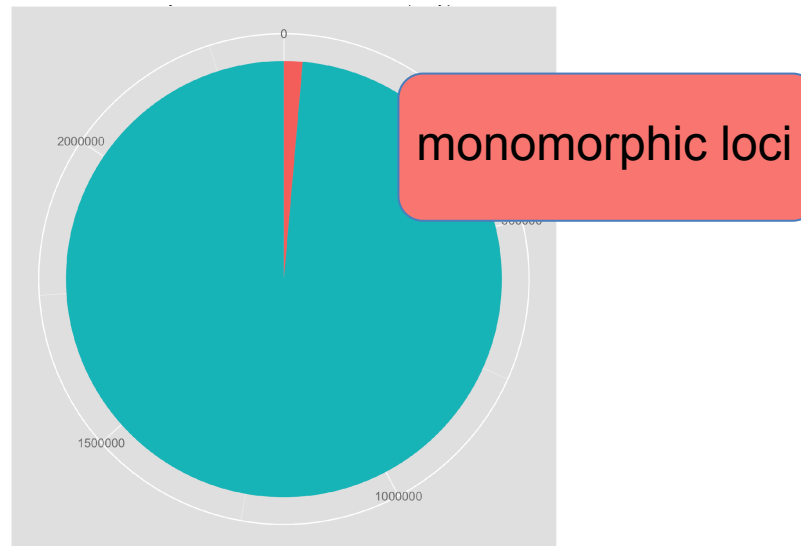
ref: A T T G T A A T T C T C A -- T A -- T T -- T G C C T T T G A A A G C
alt: A T T G T A A T T C T C A G G T A A T T T C C T G C C T T T G A A A G C

chr6:32551935, HLA-DRB1 (~11% AF in 1000G)

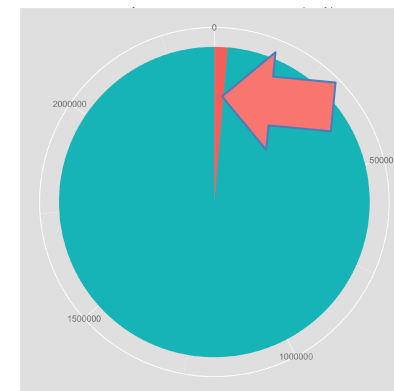
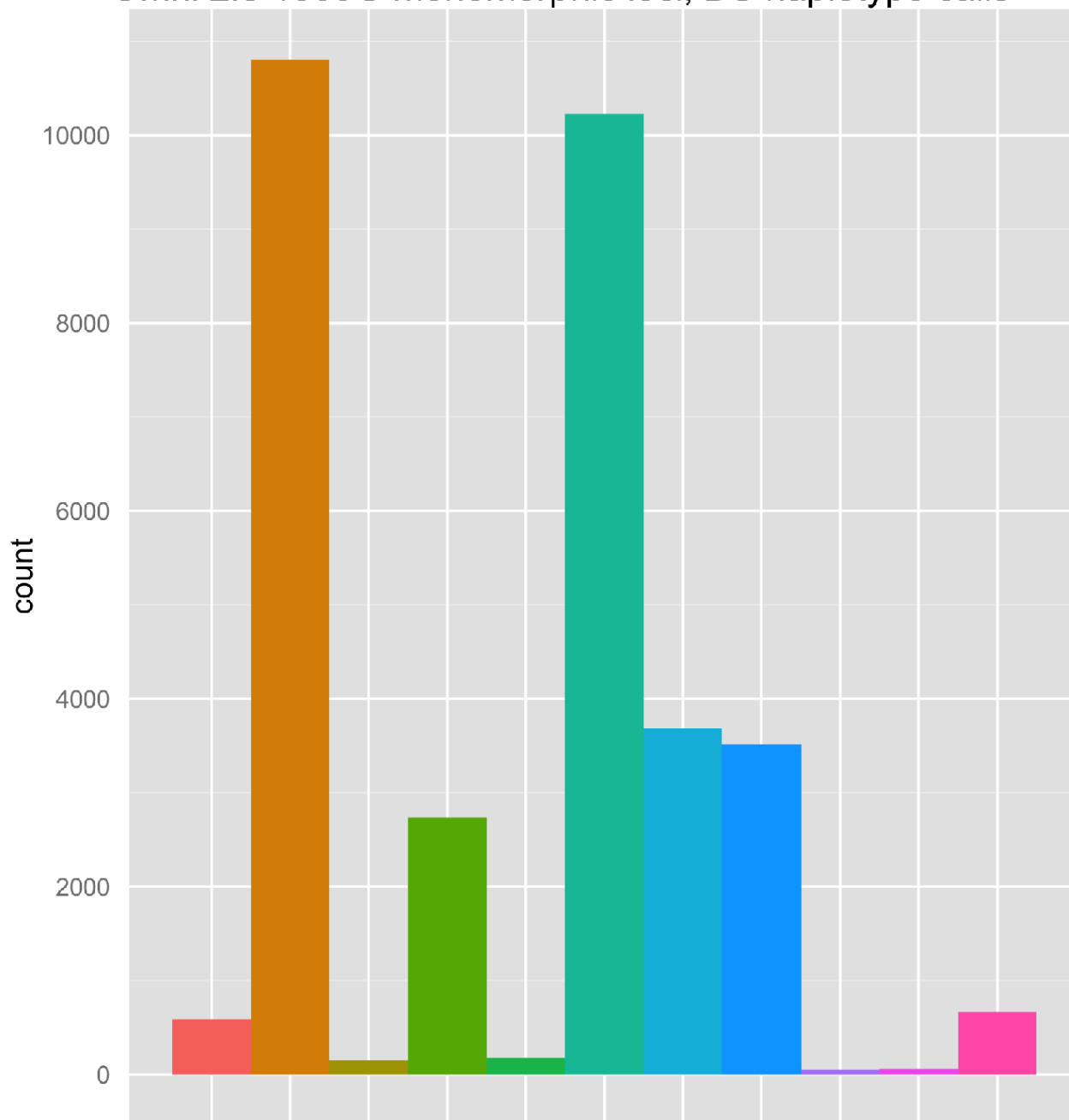
ref: C C A C C G C G G C C C G C G C C T G - C - T C C A G G A T G T C C
Alt: C C A C C G C G G -- C G C G C C T G T C T T C C A G G A G G T C C

Impact on genotyping chip design

- Biallelic SNPs detected during the 1000 Genomes Pilot project were used to design a genotyping microarray (Omni 2.5).
- When the 1000 Genomes samples were genotyped using the chip, 100k of the 2.5 million loci showed no polymorphism (monomorphs).



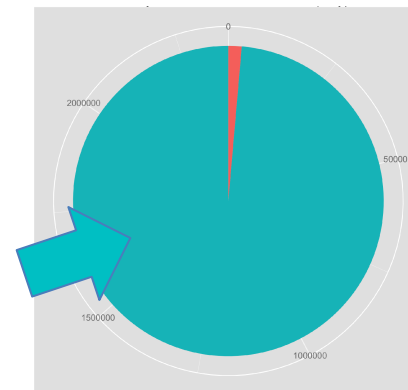
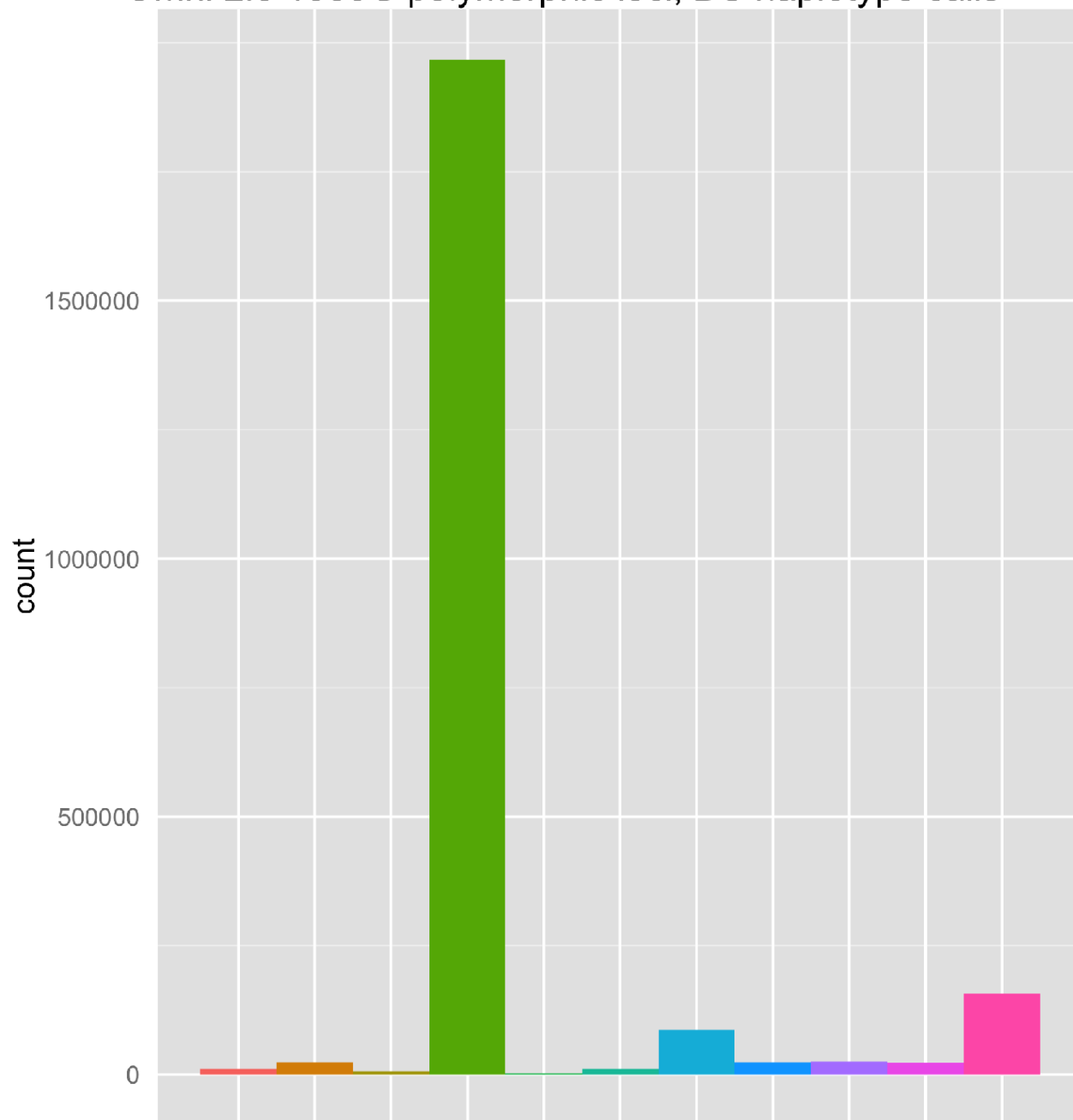
Omni 2.5 1000G monomorphic loci, BC haplotype calls



CLASS

- biallelic complex
- biallelic INDEL
- biallelic MNP
- biallelic SNP
- multiallelic complex
- multiallelic INDEL
- multiallelic INDEL, SNP, and MNP
- multiallelic mixed
- multiallelic SNP
- multiallelic SNP and MNP
- multiallelic SNP, MNP, and complex

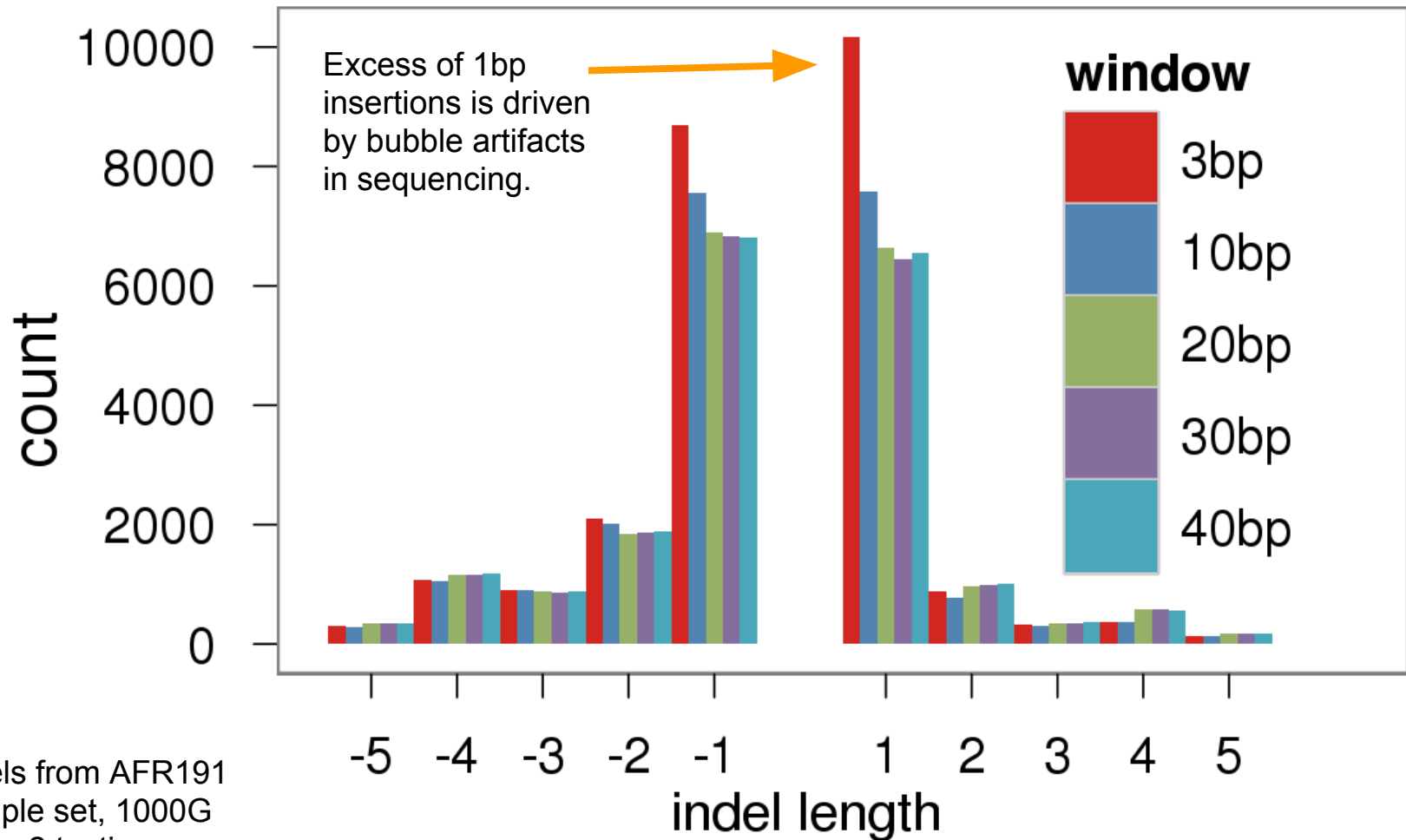
Omni 2.5 1000G polymorphic loci, BC haplotype calls



CLASS

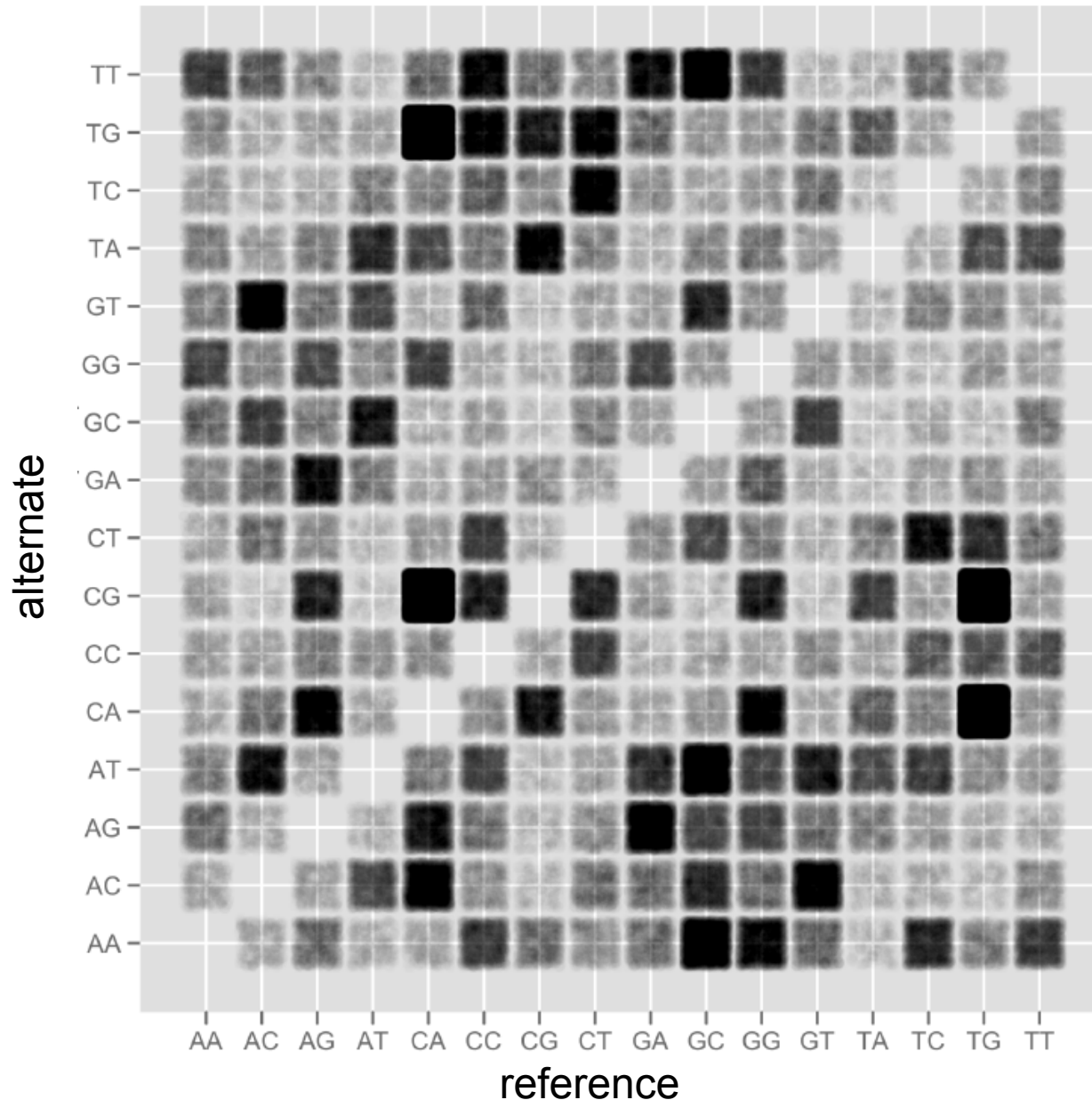
- biallelic complex
- biallelic INDEL
- biallelic MNP
- biallelic SNP
- multiallelic complex
- multiallelic INDEL
- multiallelic INDEL, SNP, and MNP
- multiallelic mixed
- multiallelic SNP
- multiallelic SNP and MNP
- multiallelic SNP, MNP, and complex

Measuring haplotypes improves specificity

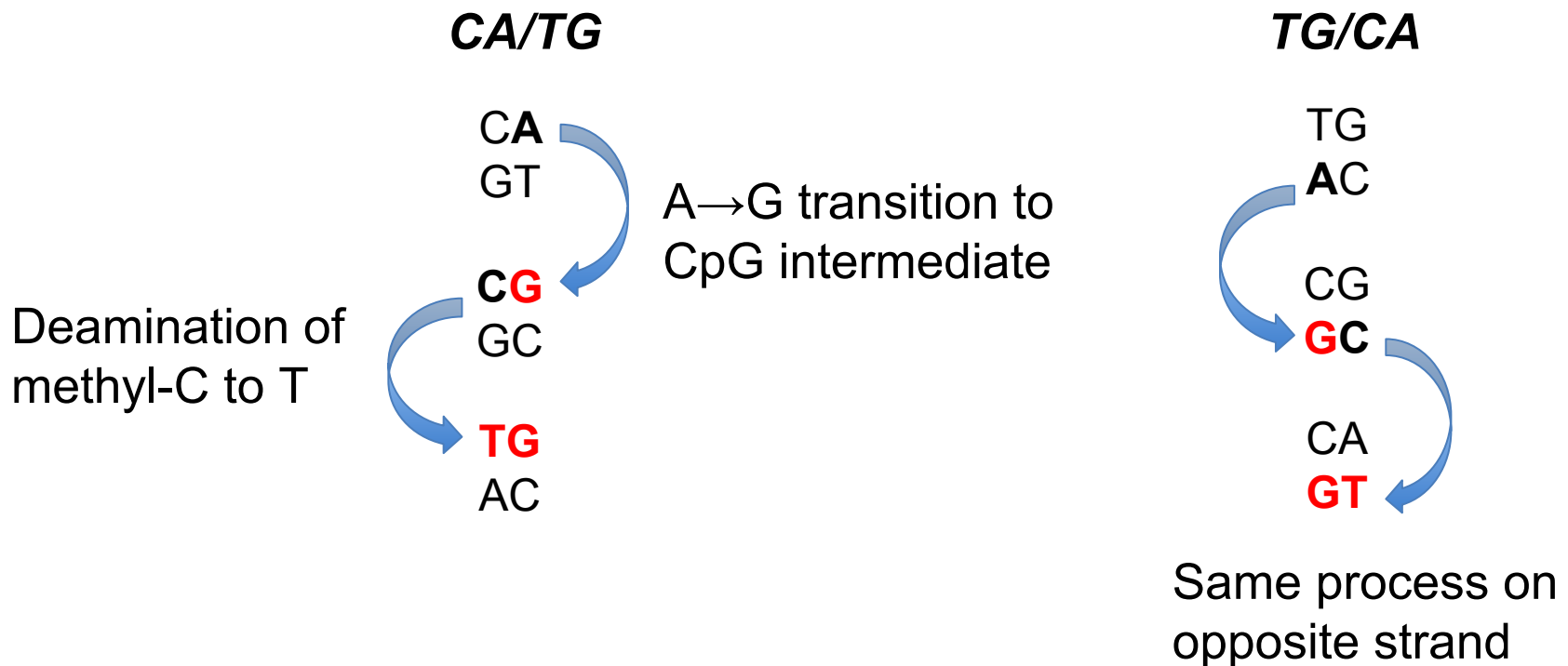


Indels from AFR191 sample set, 1000G phase2 testing.

2bp MNPs and dinucleotide intermediates



A distinct mutational mechanism is responsible for the most frequent 2bp MNP



Filtering INDELs

As with SNPs, sequencing error rates are high.

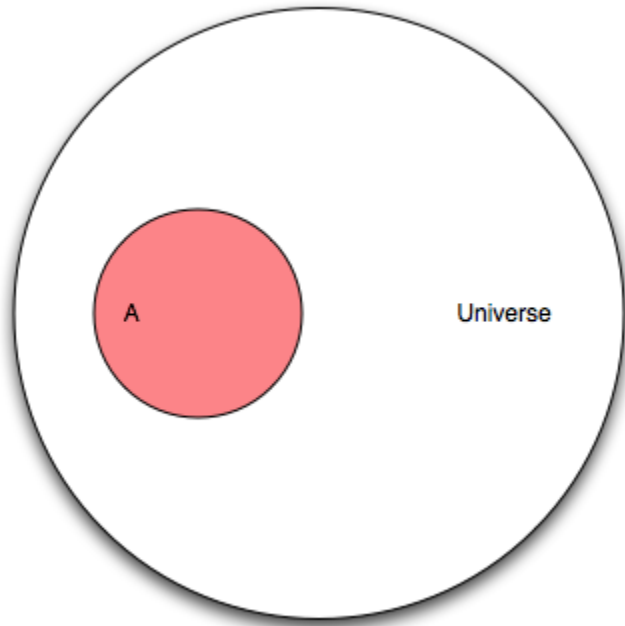
So, we need to filter.

The standard filter of NGS is the Bayesian variant caller.

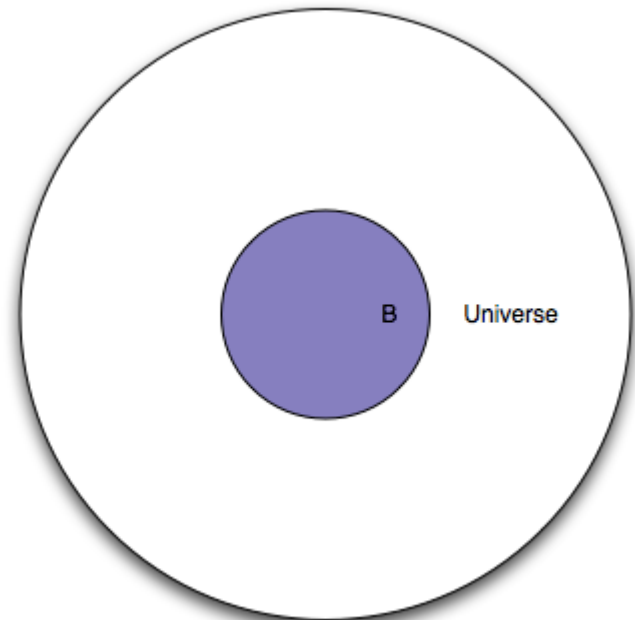
Combines population-based priors and data from many samples to make high-quality calls.

Bayesian (visual) intuition

We have a universe of individuals.



A = samples with a
variant at some locus



B = putative observations
of variant at some locus

probability(A|B)

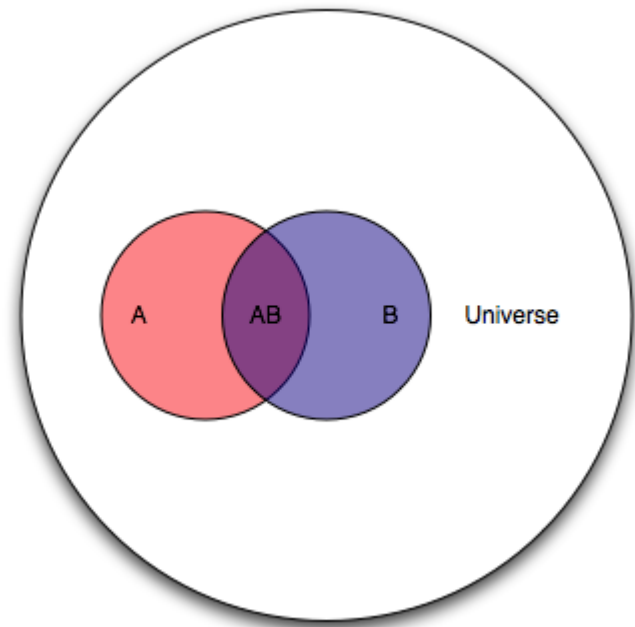
We want to estimate the probability that we have a real polymorphism "A" given "|" that we observed variants in our alignments "B".

$$P(A|B) = \frac{|AB|}{|B|}$$

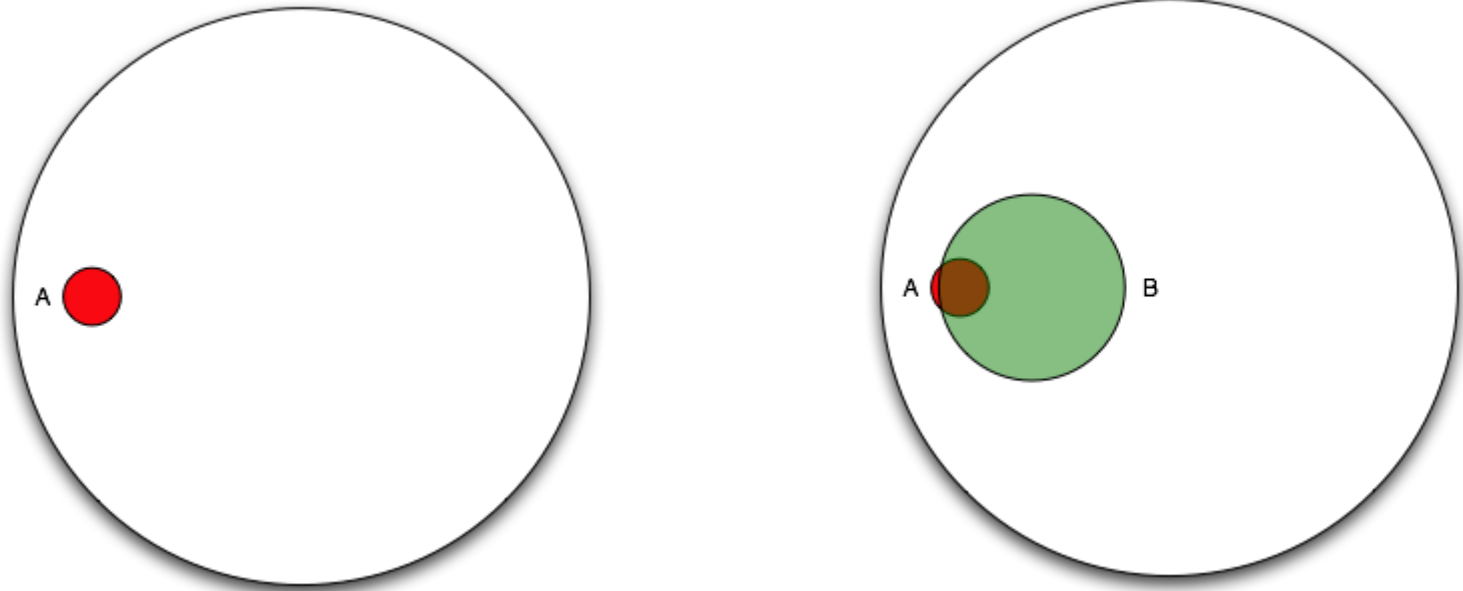
$$P(A|B) = \frac{P(AB)}{P(B)}$$

$$P(B|A) = \frac{P(AB)}{P(A)}$$

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

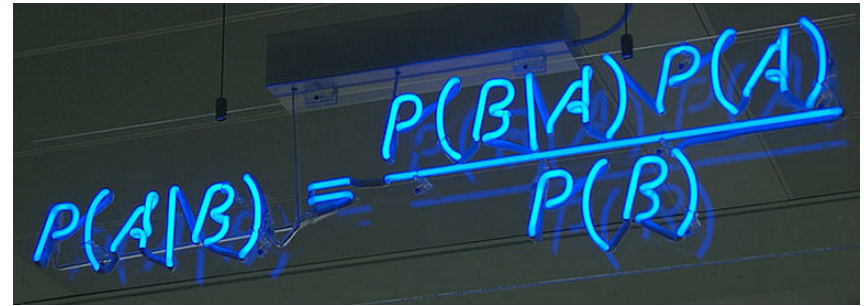


In our case it's a bit more like this...



Observations (B) provide pretty good sensitivity, but poor specificity.

The model



A photograph of a chalkboard with the formula $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$ written in blue chalk. The formula is written in a slightly messy, handwritten style.

- Bayesian model estimates the probability of polymorphism at a locus given input data and the population mutation rate (~pairwise heterozygosity) and assumption of “neutrality” (random mating).
- Following Bayes theorem, the probability of a specific set of genotypes over some number of samples is:
 - **$P(G|R) = (P(R|G) P(G)) / P(R)$**
- Which in FreeBayes we extend to:
 - **$P(G,S|R) = (P(R|G,S) P(G)P(S)) / P(R)$**
 - **G** = genotypes, **R** = reads, **S** = locus is well-characterized/mapped
 - **$P(R|G,S)$** is our data likelihood, **$P(G)$** is our prior estimate of the genotypes, **$P(S)$** is our prior estimate of the mappability of the locus, **$P(R)$** is a normalizer.

Handling non-biallelic/diploid cases

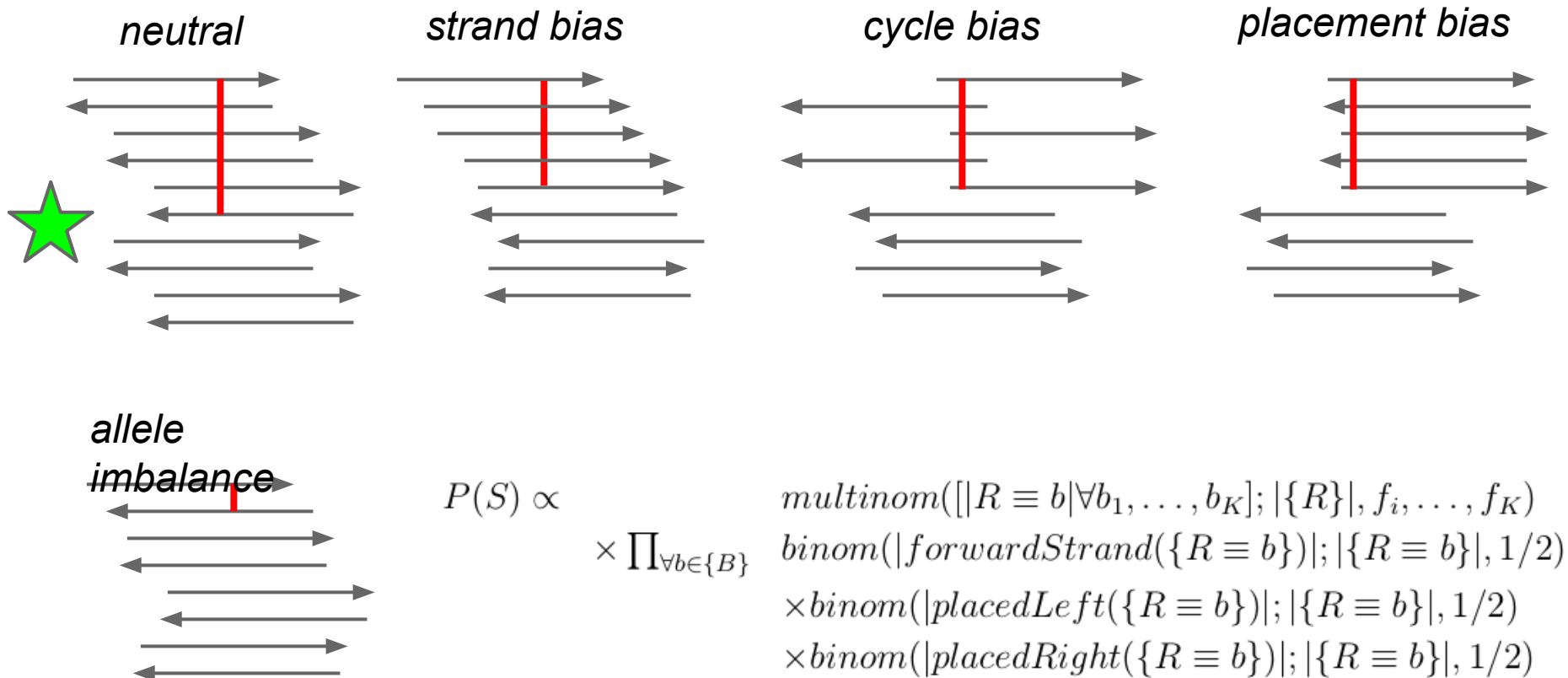
We compose our data likelihoods, **P(Reads|Genotype)** using a discrete multinomial sampling probability:

$$P(\text{reads}|\text{genotype}) = \binom{|\text{reads}|}{|\text{reads} = A|, |\text{reads} = B| \dots} \\ \times \prod_{\forall \text{alleles} \in \text{genotype}} \text{freq}(\text{allele} \in \text{genotype}) \\ \times \prod_{\forall \text{reads}} P(\text{correct}(\text{read}))$$

Our priors, **P(Genotypes)**, follow the Ewens Sampling Formula and the discrete sampling probability for genotypes.

Are our locus and alleles sequenceable?

In WGS, biases in the way we observe an allele (placement, position, strand, cycle, or balance in heterozygotes) are often correlated with error. We include this in our posterior $\mathbf{P}(\mathbf{G}, \mathbf{S} | \mathbf{R})$, and to do so we need an estimator of $\mathbf{P}(\mathbf{S})$.



The detection process

1. Parse alleles (small haplotypes) from alignments using CIGAR strings.
2. Pick suitable alleles (very weak input filters to improve runtime).
3. Build haplotypes across target locus.
4. Generate genotype likelihoods.
5. Sample a posterior space around the data-likelihood maximum
 - a. update genotype estimates and iterate (hill-climbing posterior search) until convergence on maximum *a posteriori* genotyping over all samples
6. Output a record, and do it again

SVM filtering

INDEL detection is hard.

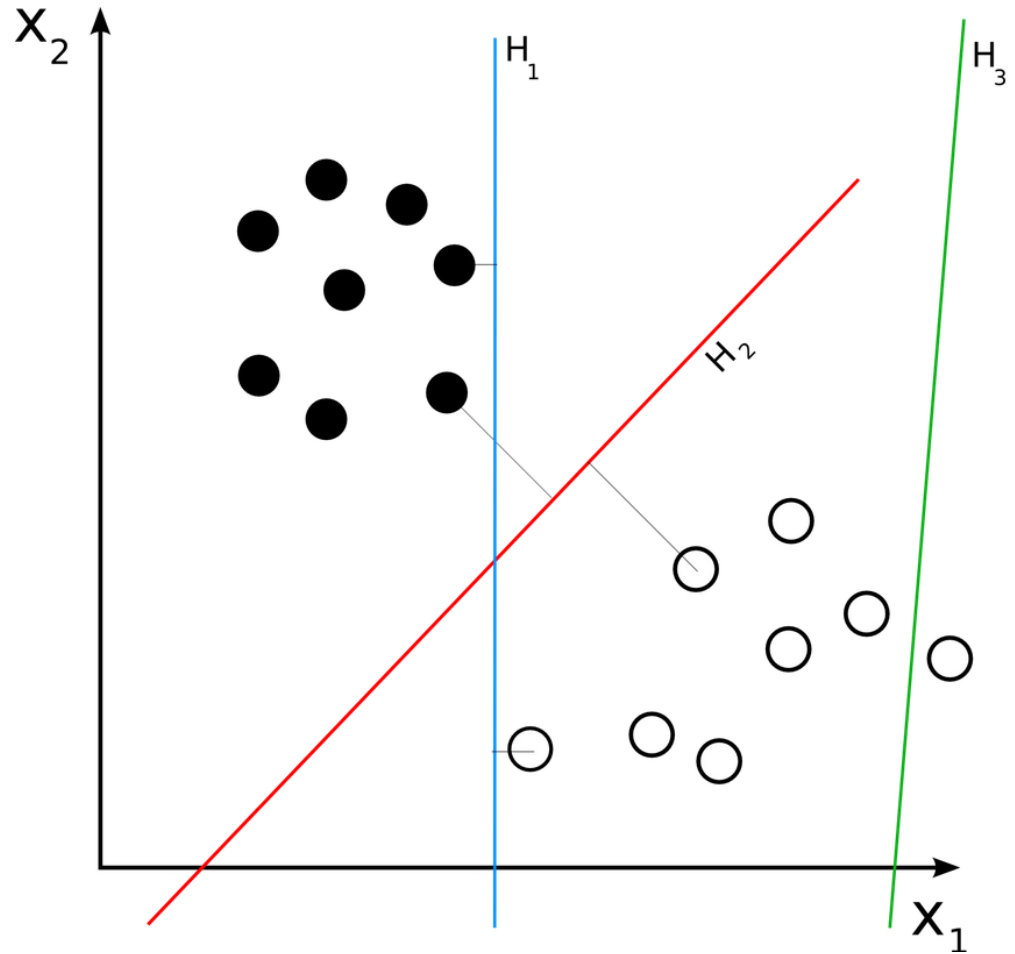
A priori models can't capture all types of error.

It's especially difficult when we try to make a consensus set from lots of input variant callers.

We can use classifiers like Support Vector Machines (SVM) to further improve results.

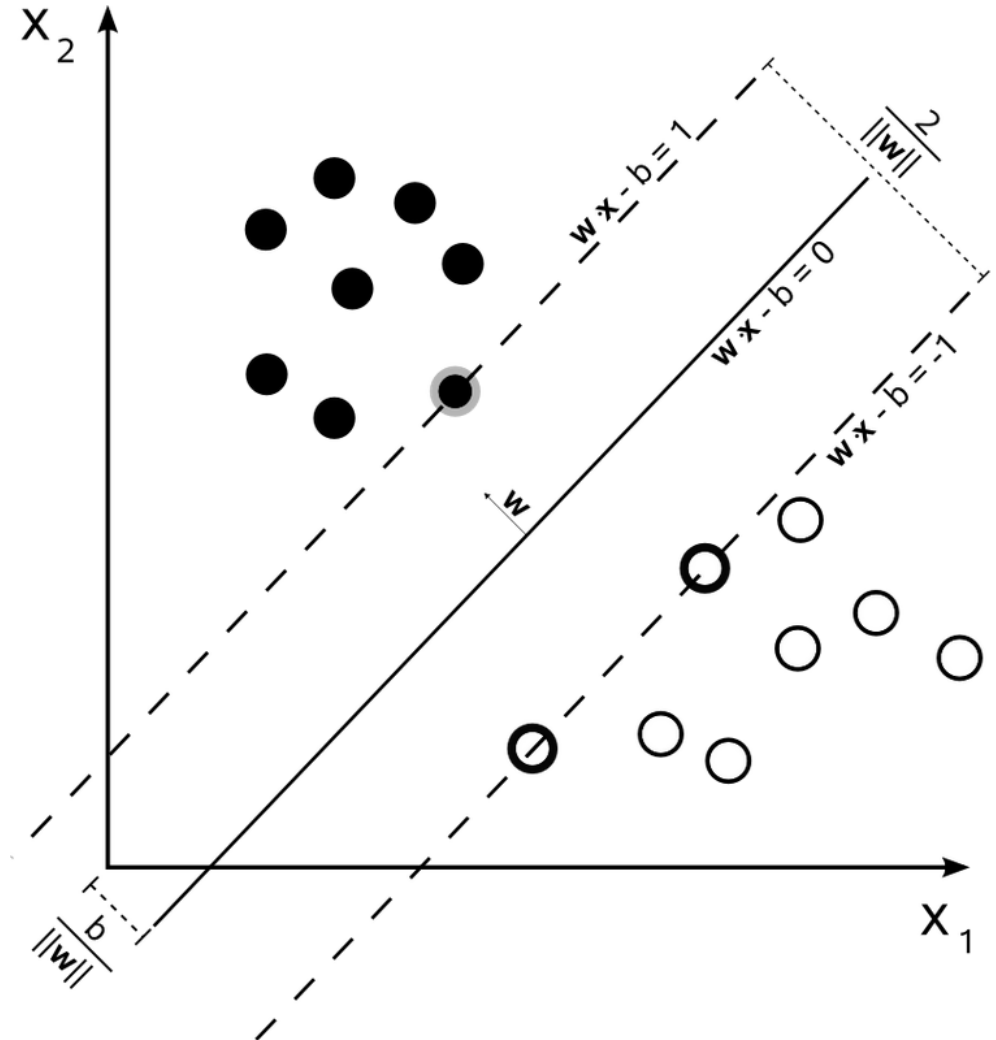
SVM classifier

Find a hyperplane (here a line in 2D) which separates observations.



SVM classifier

The best separating hyperplane is determined by maximum margin between groups we want to classify.



SVM approach for INDEL filtering

Extract features that tend to vary with respect to call quality:

- call QUALity
- read depth
- sum of base qualities
- inbreeding coefficient (heterozygosity)
- entropy of sequence at locus
- mapping quality
- allele frequency in population
- read pairing rate
- etc.

SVM approach for INDEL filtering

Now, use overlaps in validation samples or sites to determine likely errors and true calls.

Use this list + annotations of the calls to train an SVM model.

Apply the model to all the calls, filter, and measure validation rate of the whole set.

1000G variant integration pipeline

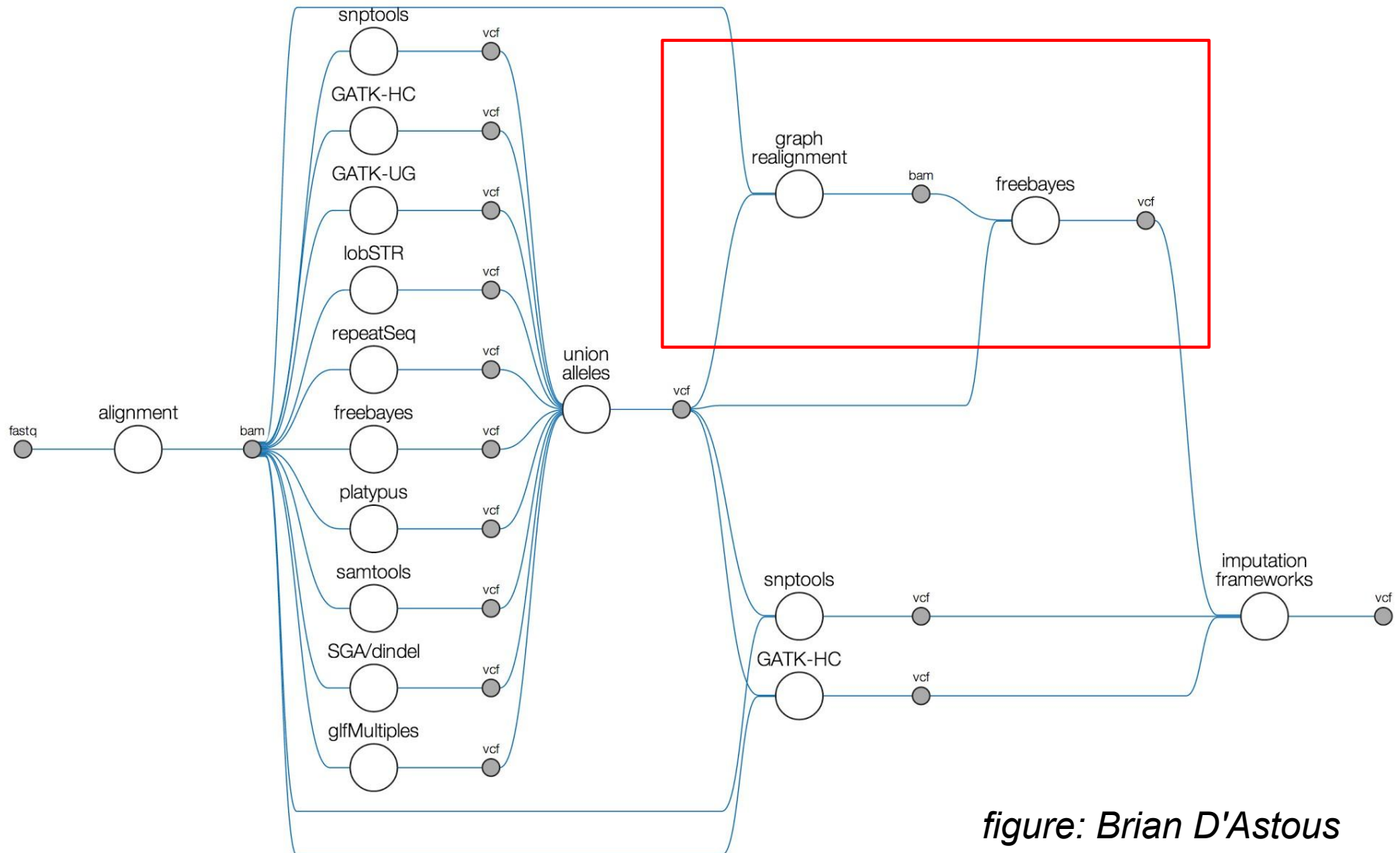
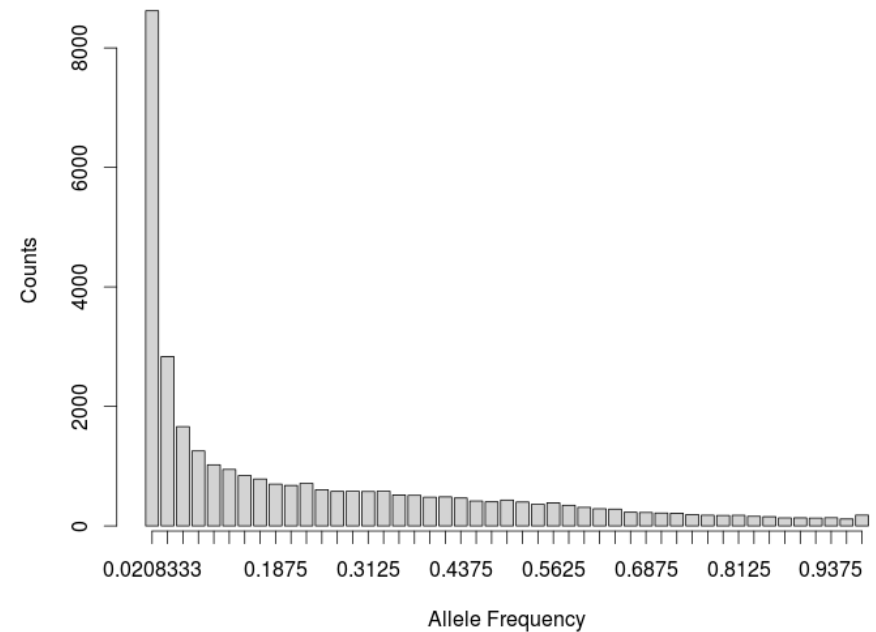
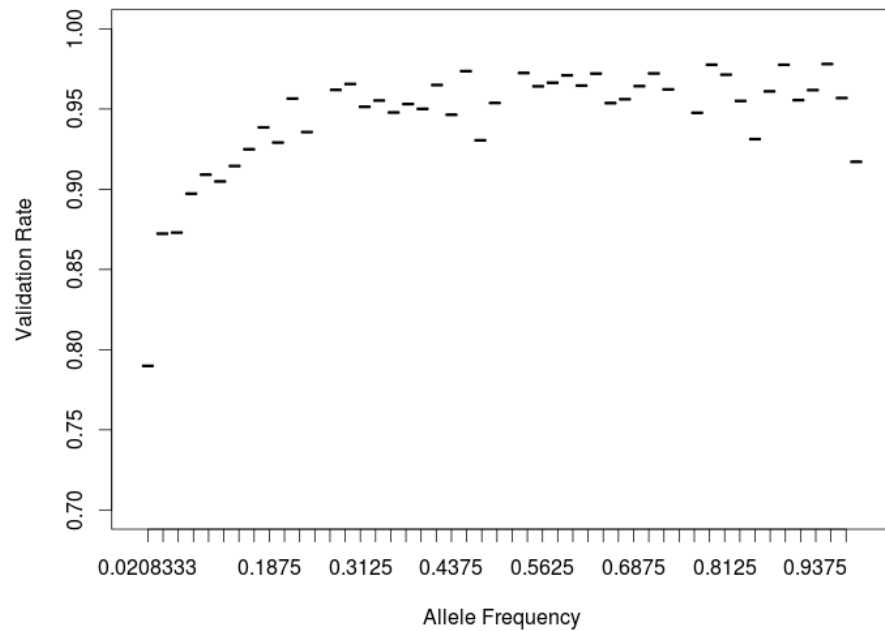


figure: Brian D'Astous

Application of SVM to 1000G INDELs

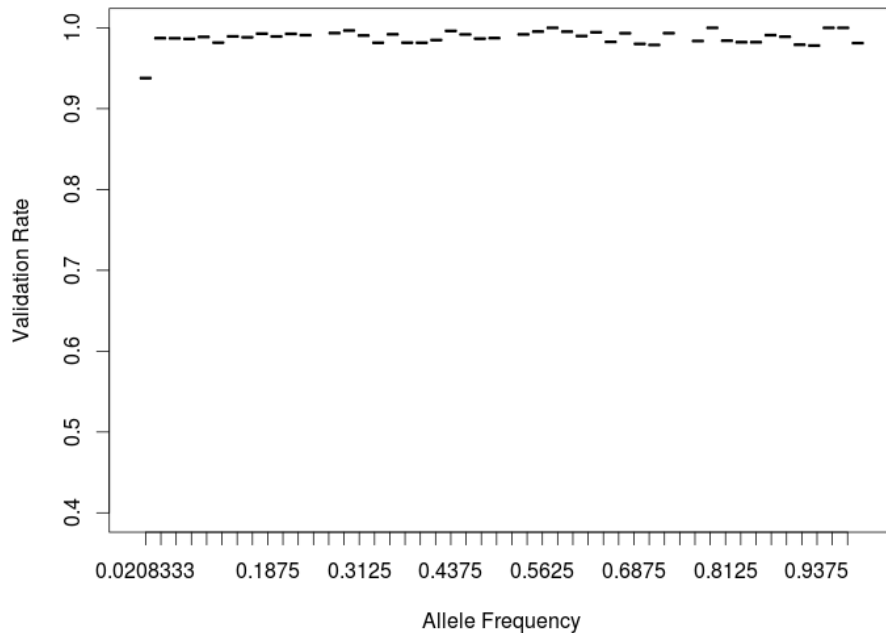


Raw validation rates of indels in 1000G phase 3, “MVNCall” set.

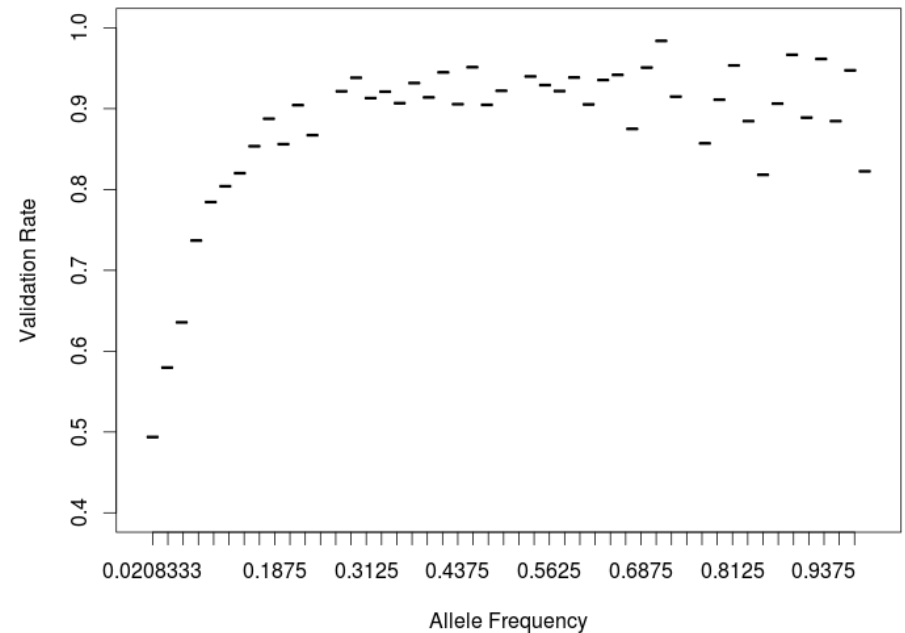
Tony Marcketta and Adam Auton

Application of SVM to 1000G INDELs

Passing SVM



Failing SVM

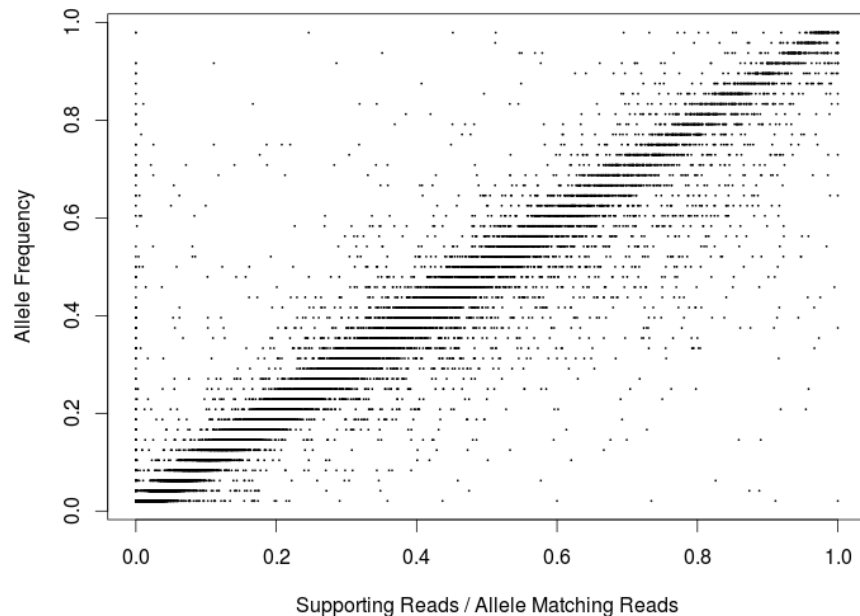


Filtering results, using SVM-based method.

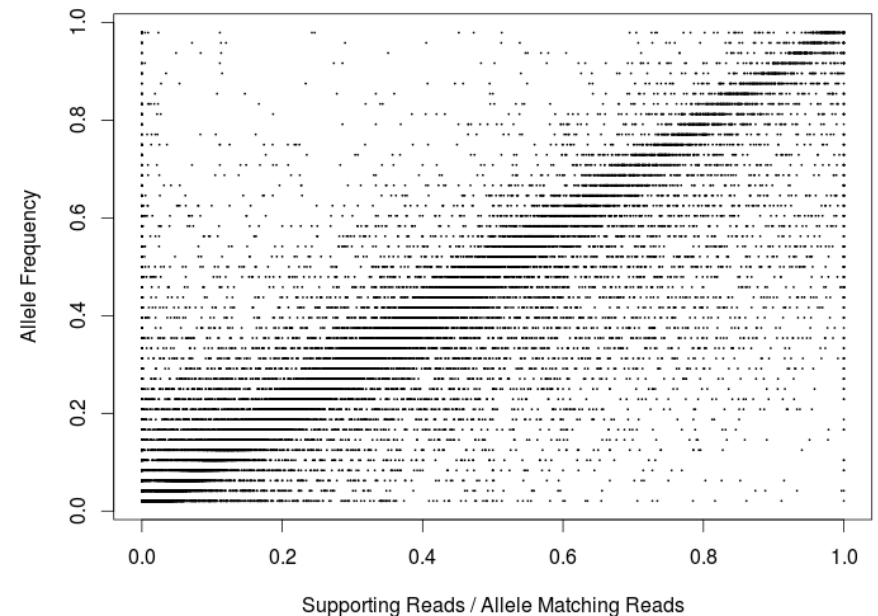
Anthony Marcketta and Adam Auton

Application of SVM to 1000G INDELs

Passing SVM



Failing SVM



Correlation between allele frequency and observation counts.

Anthony Marcketta and Adam Auton

Questions?

...