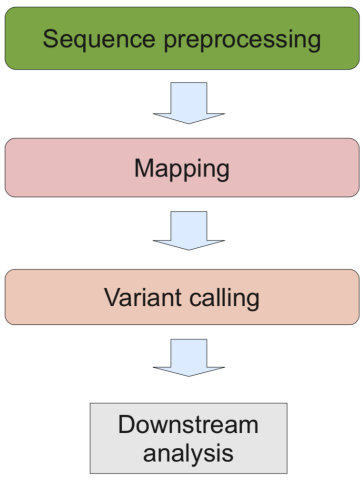
"To err is human, but to really foul things up you need a computer."
-- Paul Ehrlich

"To err is human– and to blame it on a computer is even more so."
-- Robert Orden

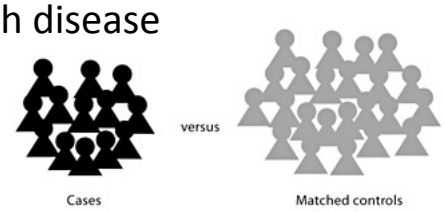
Resequencing

- Sequencing additional members of a species for which a reference genome is already available.
- Goal is to identify variations within the population.

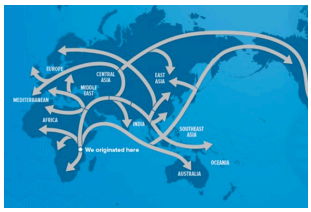


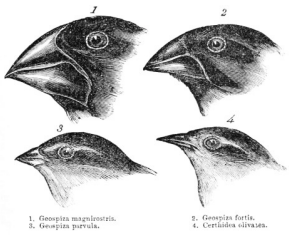
Why is this useful?

- Associate variations with disease



- Understand population dynamics

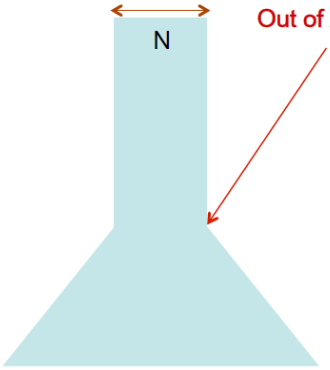


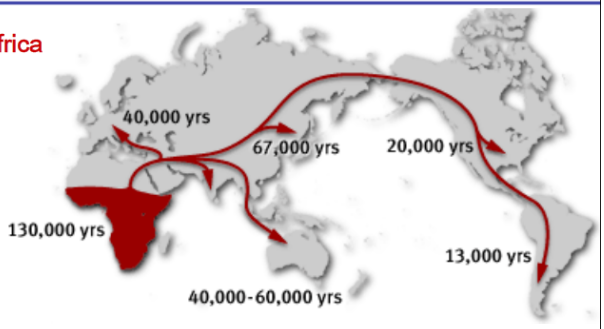


Variant calling

- Aim: produce **variant** calls (w.r.t. reference), and **genotype** calls
- True variants usually easy to spot
  - But: SNPs easier than indels which are easier than SVs
  - And: Sufficient coverage required
- Divergence / diversity often low (human: 0.1%)
  - False positives are an issue

Why humans are so similar





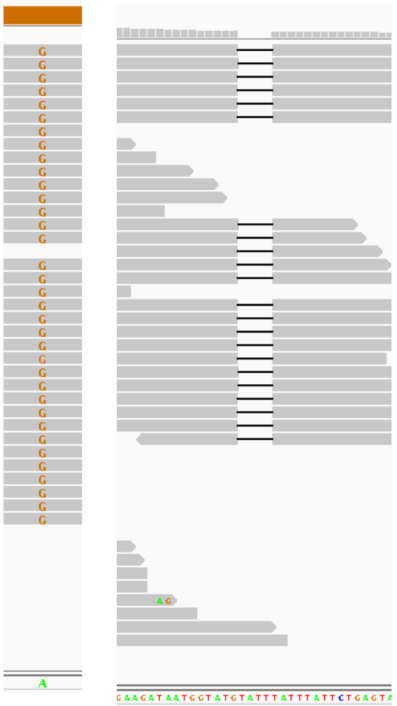
A small population that interbred reduced the genetic variation

Out of Africa ~ 40,000 years ago

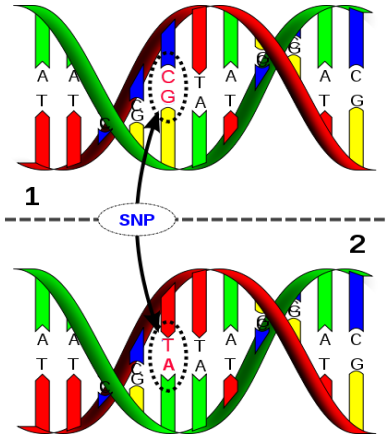
National Genographic project: https://genographic.nationalgeographic.com/

Types of Variations

- SNPs
- Indels
- Structural Variation



Single Nucleotide Polymorphisms (SNPs)



- 23 chromosome pairs
- 3 billion bases
- A single nucleotide change between pairs of chromosomes
- E.g.
  Haplotype1: AAGG**G**ATCCAC
  Haplotype2: AAGG**A**ATCCAC
- A/A or G/G homozygote
- A/G heterozygote

SNPs vs. SNVs

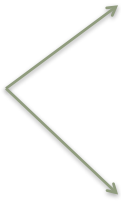
Both are aberrations at a single nucleotide

- **SNP**
  - Aberration expected at the position for any member in the species (well-characterized)
  - Occur in population at some frequency so expected at a given locus
  - Validated in population
  - Catalogued in dbSNP (http://www.ncbi.nlm.nih.gov/snp)
- **SNV**
  - Aberration seen in only one individual (not well characterized)
  - Occur at low frequency so not common
  - Not validated in population

SO … Really a matter of frequency of occurrence

Mapping choices must permit variation

Reference sequence:
actgtagattagccgagtagctagctagtcgat



Find best match for each read in a reference sequence.

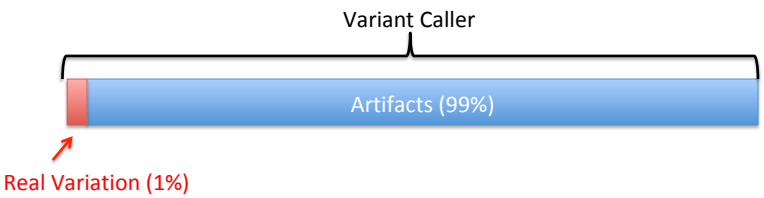
ccgagaagctag

BUT … differences may arise from:

- Sequencing errors
- Mapping problems (wrongly placed reads)
- Individual polymorphisms

Short read sequence data is noisy

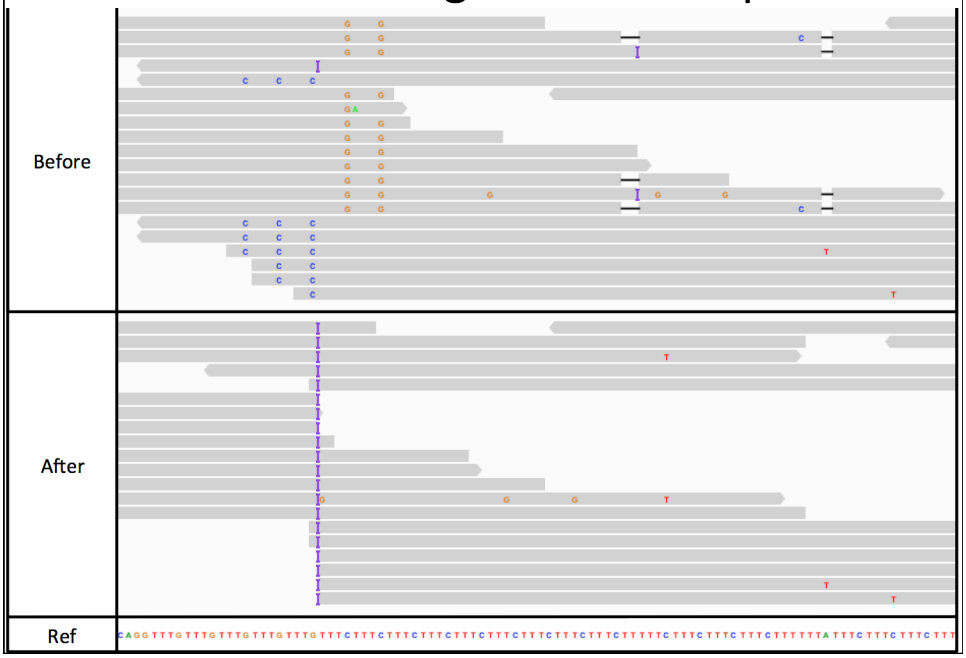
- In 2nd and 3rd-generation sequencing, ***most*** putative variants are typically artifacts (>99%).
- We must filter putative variants to remove artifacts or our analyses will be overwhelmed by noise.



Realignment

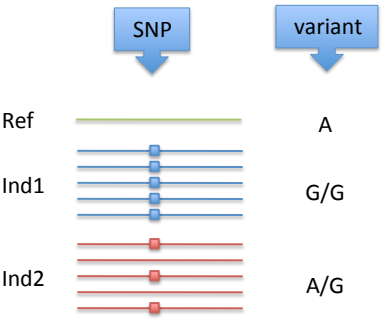
- Hashing and Burrows-Wheeler Transform are particularly not effective with indels.
- Instead of indels, often observe high SNP density.
- Realignment takes reads in these regions and does assembly followed by true alignment.

Another realignment example



Variant calling methods

- > 15 different algorithms
- Three categories
  - Allele counting (simplest)
  - Probabilistic methods, e.g. Bayesian model
    - to quantify statistical uncertainty
    - Assign priors based on observed allele frequency of multiple samples
    - Effective leverage genotype calling
  - Heuristic approach
    - Based on thresholds for read depth, base quality, variant allele frequency, statistical significance



Nielsen R, Paul JS, Albrechtsen A, Song YS. Genotype and SNP calling from next-generation sequencing data. Nat Rev Genet. 2011 Jun;12(6):443-51. PMID: 21587300.

http://seqanswers.com/wiki/Software/list

SNP and Indel calling is a large-scale Bayesian Inference problem.

Bayes Model

$$\Pr\{G \mid D\} = \frac{\Pr\{G\}\Pr\{D \mid G\}}{\sum_i \Pr\{G_i\}\Pr\{D \mid G_i\}}$$

Prior of genotype

Likelihood of genotype

$$\Pr\{D \mid G\} = \prod_j \left(\frac{\Pr\{D_j \mid H_1\}}{2} + \frac{\Pr\{D_j \mid H_2\}}{2} \right)$$

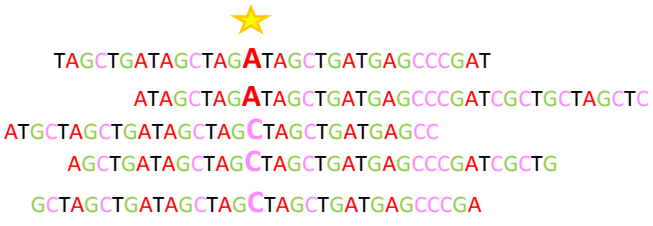
where

$G = H_1 H_2$
Is diploid assumption

$\Pr\{D \mid H\}$
Is the haploid likelihood function

- Inference: What is the genotype G of each sample given read data D for each sample?
- Calculated via Bayes' rule the probability of each position G
- Assumes reads are independent
- Relies on the likelihood function to estimate probability of sample data given proposed haplotype

Genotype Calling from Sequence Data



Sequence Reads

5'-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3'

Reference Genome

2A and 3C — Observed Data

A/C or A/A or C/C — Predicted Genotype

Initially we have no reads

The only information we really have is the reference genome identity.

Sequence Reads

5'-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3'

Reference Genome

P(reads|A/A)= ?

P(reads|A/C)= ?

P(reads|C/C)= ?

Possible Genotypes

Inference with short read data

GCTAGCTGATAGCTAGCTAGCTGATGAGCCCGA

Sequence Reads

5'-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3'

Reference Genome

P(reads|A/A)= P(C observed, read maps |A/A)

P(reads|A/C)= P(C observed, read maps |A/C)

P(reads|C/C)= P(C observed, read maps |C/C)

Possible Genotypes

Inference assuming seq error of 1%

GCTAGCTGATAGCTAGCTAGCTGATGAGCCCGA

Sequence Reads

5'-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3'

Reference Genome

P(reads|A/A)= 0.01 e.g. our read is an error/mistake

P(reads|A/C)= 0.50 we would expect half reads to be C

P(reads|C/C)= 0.99 Only exception is mistakes in sequencing.

Possible Genotypes

Note that per read base quality scores are typically factored into these probabilities.

Therefore most variant callers do base quality recalibration.

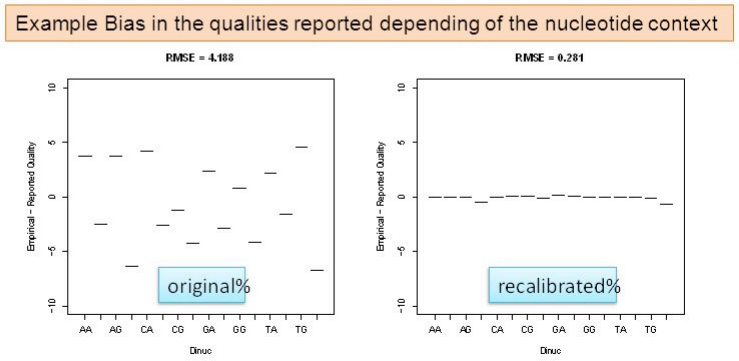


image: broadinstitute.org/gatk

As data accumulate …

AGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTG

GCTAGCTGATAGCTAGCTAGCTGATGAGCCCGA

Sequence Reads

5’-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3’

Reference Genome

P(reads|A/A)= 0.0001

P(reads|A/C)= 0.25

P(reads|C/C)= 0.98

Possible Genotypes

As data accumulate …

Sequence Reads

5'-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3'

Reference Genome

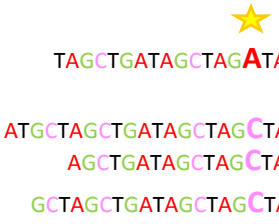
P(reads|A/A)= 0.000001

P(reads|A/C)= 0.125

P(reads|C/C)= 0.97

Possible Genotypes

As data accumulate …



Sequence Reads

5'-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3'

Reference Genome

P(reads|A/A)= 0.00000099

P(reads|A/C)= 0.0625

P(reads|C/C)= 0.0097

Possible Genotypes

P(reads|genotype)

Sequence Reads

5'-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3'

Reference Genome

P(reads|A/A)= 0.00000098

P(reads|A/C)= 0.03125

P(reads|C/C)= 0.000097

Not the "end" yet

Sequence Reads

5'-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3'

Reference Genome

P(reads|A/A) = 0.00000098
P(reads|A/C) = 0.03125
P(reads|C/C) = 0.000097

Making a genotype call requires
combining sequence data with prior information

$$P(Genotype|reads) = \frac{P(reads|Genotype)Prior(Genotype)}{\sum_G P(reads|G)Prior(G)}$$

Sources of prior information

- Individual based prior
  - Equal probability of showing polymorphism
  - 1/1000 bases different from reference (generic mutation rate)
  - Error Free and Poisson distribution
  - Single sample, single site
- Population based prior
  - Estimate frequency from many individuals
  - Multiple sample, single site
- Haplotype/Imputation based prior
  - Jointly model flanking SNPs, use haplotype information
  - Important for low coverage sequence data
  - Multiple samples, multiple sites

Individual based prior

Sequence Reads

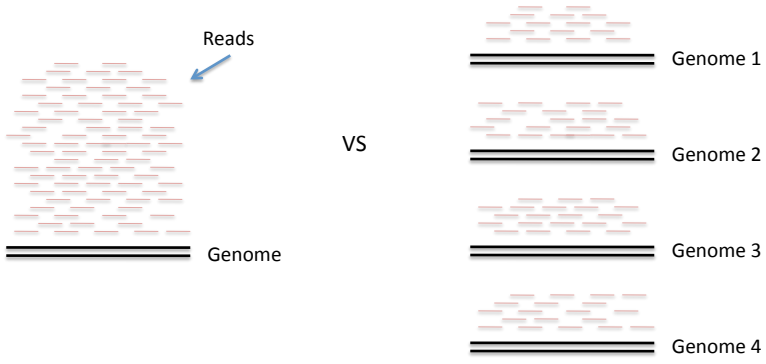
5'-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3'

Reference Genome

P(reads|A/A)= 0.00000098	Prior(A/A) = 0.00034	P(A/A|reads) < 0.01
P(reads|A/C)= 0.03125	Prior(A/C) = 0.00066	P(A/C|reads) = 0.175
P(reads|C/C)= 0.000097	Prior(C/C) = 0.99900	P(C/C|reads) = 0.825

Base Prior: every site has 1/1000 probability of varying

Coverage (High vs Low)



- Which one has more power to detect variations?

Population Based Prior

Sequence Reads

5'-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3'

Reference Genome

P(reads|A/A)= 0.00000098	Prior(A/A) = 0.04	P(A/A|reads) < .001
P(reads|A/C)= 0.03125	Prior(A/C) = 0.32	P(A/C|reads) = 0.999
P(reads|C/C)= 0.000097	Prior(C/C) = 0.64	P(C/C|reads) = <.001

Population Based Prior: Use frequency information from examining other datasets at the same site. E.g. P(A) = 0.2

Catalogs of human genetic variation

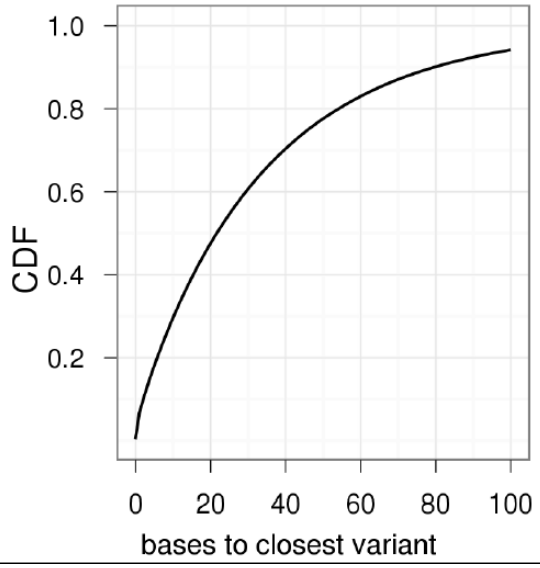
- **The 1000 Genomes Project**
  - http://www.1000genomes.org/
  - SNPs and structural variants from 3000 individuals from about 25 populations
- **HapMap**
  - http://hapmap.ncbi.nlm.nih.gov/
  - identify and catalog genetic similarities and differences
- **dbSNP**
  - http://www.ncbi.nlm.nih.gov/snp/
  - Database of SNPs and multiple small-scale variations
- **COSMIC**
  - http://www.sanger.ac.uk/genetics/CGP/cosmic/
  - Catalog of Somatic Mutations in Cancer
- **TCGA**
  - http://cancergenome.nih.gov/
  - The Cancer Genome Atlas researchers are mapping the genetic changes in 20 selected cancers
- **ClinVar**
  - http://www.ncbi.nlm.nih.gov/clinvar/
  - aggregates information about sequence variation and its relationship to human health

Haplotype based prior

(Why haplotypes?)

- Variants cluster
- Observing haplotypes lets us be more certain of local genome structure
- We improve the SNP detection process by using haplotypes rather than point mutations

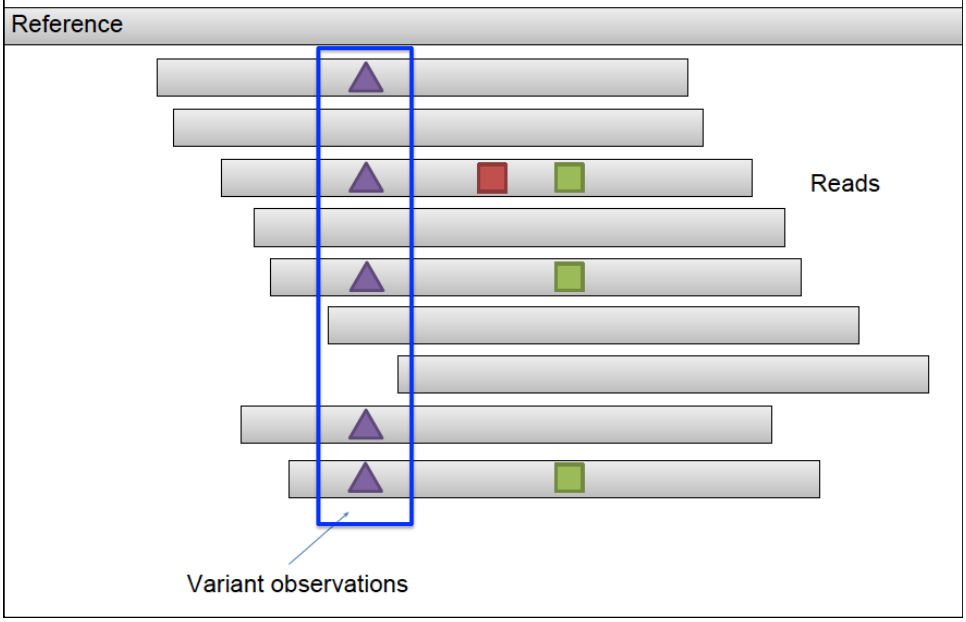
Sequence variants cluster

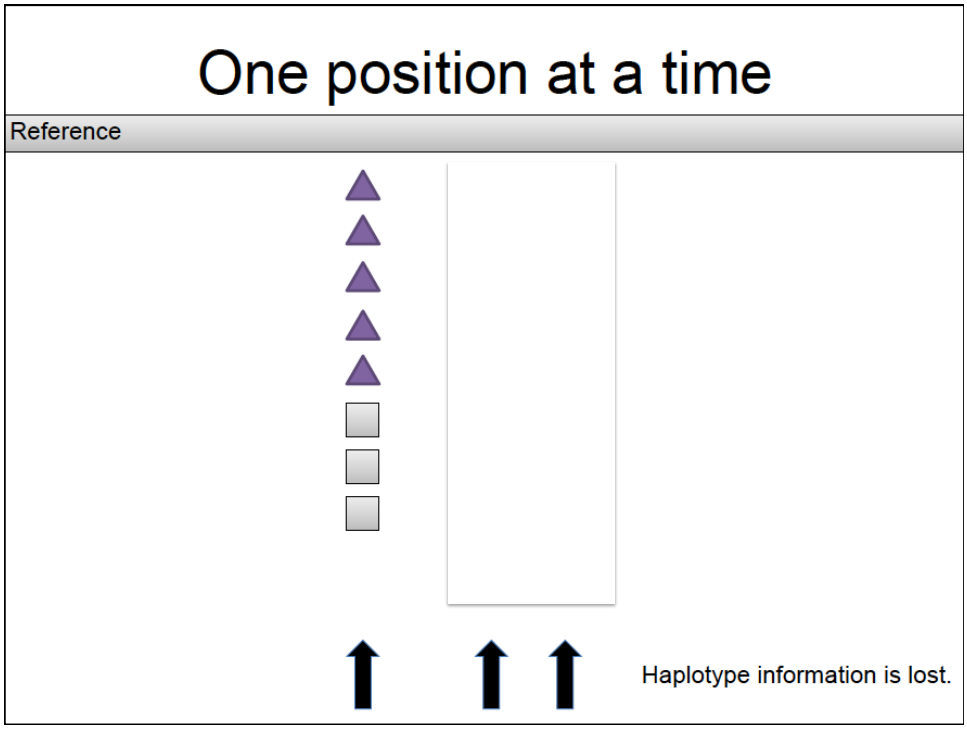


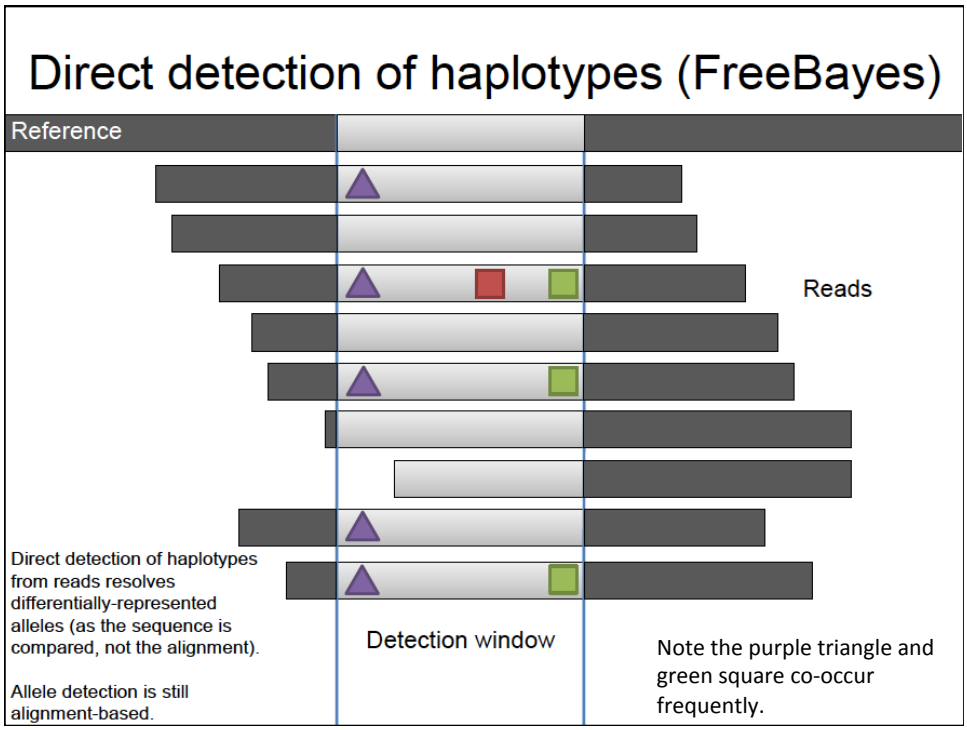
In ~1000 individuals, ½ of variants are within ~22bp of another variant.

Variance to mean ratio (VMR) = 1.4.

One position at a time




One position at a time
Reference
Haplotype information is lost.


Direct detection of haplotypes (FreeBayes)
Reference
Reads
Direct detection of haplotypes from reads resolves differentially-represented alleles (as the sequence is compared, not the alignment).
Allele detection is still alignment-based.
Detection window
Note the purple triangle and green square co-occur frequently.

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

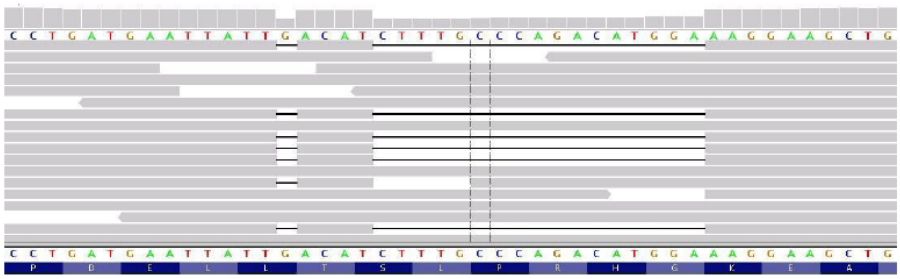
```
reference: AGG GAG CTG
           Arg Glu Leu

apparent:  AGG TAG CTG
           Arg Ter ---

actual:    AGG TTG CTG
           Arg Leu Leu
```

OTOF gene – mutations cause profound recessive deafness

Apparent nonsense variant, one YRI homozygote

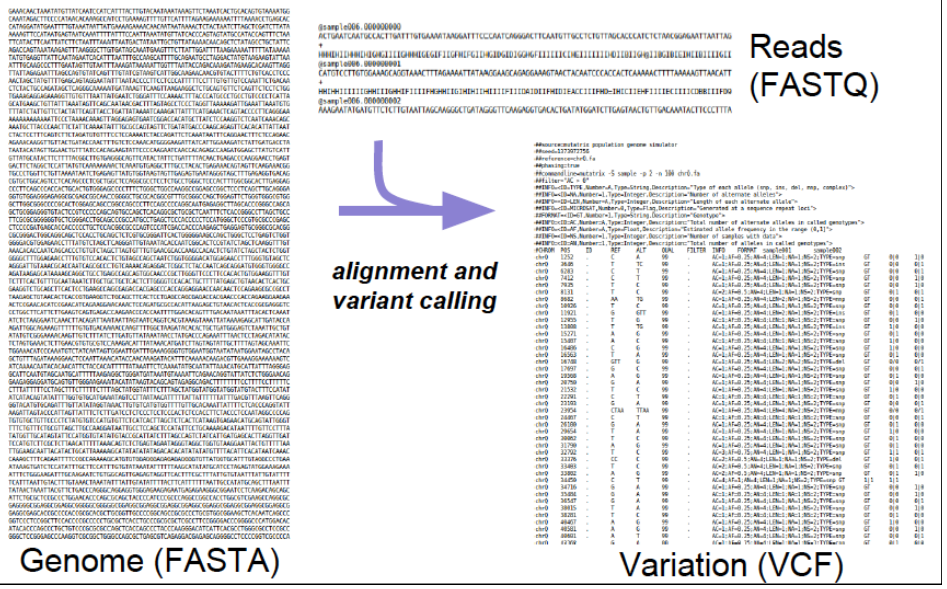
Actually a block substitution that results in a missense substitution

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

Working with sequences, finding variations



How to describe variants: Variant Call Format (VCF)

HEADER LINES: start with "##", describe all symbols found later on in FORMAT and ANNOTATIONS, e.g.,

```
##fileformat=VCFv4.1
##FORMAT=<ID=AD,Number=.,Type=Integer,Description="Allelic depths for the ref and alt alleles in the order listed">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth (reads with MQ=255 or with bad mates are filtered)">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
……
```

SITE RECORDS:

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	ZW155	ZW177
chr2R	2926	.	C	A	345.03	PASS	[ANNOTATIONS]	GT:AD:DP:GQ:PL	0/1:4,9:13:80:216,0,80	0/0:6,0:6:18:0,18,166
chr2R	9862	.	TA	T	180.73	.	[ANNOTATIONS]	GT:AD:DP:GQ:PL	1/1:0,5:5:15:97,15,0	1/1:0,4:4:12:80,12,0
chr2R	10834	.	A	ACTG	173.04	.	[ANNOTATIONS]	GT:AD:DP:GQ:PL	0/0:14,0:14:33:0,33,495	0/1:6,3:9:99:105,0,315

ID: some ID for the variant, if known (e.g., dbSNP)

REF, ALT: reference and alternative alleles (on forward strand of reference)

QUAL = -10*log(1-p), where p is the probability of variant being present given the read data

FILTER: whether the variant failed a filter (filters defined by the user or program processing the file)

How to describe variants: Variant Call Format (VCF)

```
[HEADER LINES]
#CHROM  POS    ID  REF   ALT   QUAL    FILTER    INFO           FORMAT          ZW155                  ZW177
chr2R   2926   .    C     A    345.03    PASS  [ANNOTATIONS]  GT:AD:DP:GQ:PL  0/1:4,9:13:80:216,0,80  0/0:6,0:6:18:0,18,166
chr2R   9862   .   TA     T    180.73    .     [ANNOTATIONS]  GT:AD:DP:GQ:PL  1/1:0,5:5:15:97,15,0    1/1:0,4:4:12:80,12,0
chr2R 10834    .    A  ACTG   173.04    .     [ANNOTATIONS]  GT:AD:DP:GQ:PL  0/0:14,0:14:33:0,33,495      ./.
```

GT (genotype):

- 0/0 reference homozygote
- 0/1 reference-alternative heterozygote
- 1/1 alternative homozygote
- 0/2, 1/2, 2/2, etc. - possible if more than one alternative allele present
- ./. missing data

AD: allele depths
DP: total depth (may be different from sum of AD depths, a the latter include only reads significantly supporting alleles)

PL: genotype likelihoods (phred-scaled), normalized to the best genotype, e.g.,
PL(0/1) = -10*log[Prob(data|0/1) / Prob(data|best_genotype)]

GQ: genotype quality – this is just PL of the second-best genotype

How to describe variants: Variant Call Format (VCF)

```
[HEADER LINES]
#CHROM  POS    ID  REF   ALT   QUAL    FILTER    INFO           FORMAT          ZW155                  ZW177
chr2R   2926   .    C     A    345.03    PASS  [ANNOTATIONS]  GT:AD:DP:GQ:PL  0/1:4,9:13:80:216,0,80  0/0:6,0:6:18:0,18,166
chr2R   9862   .   TA     T    180.73    .     [ANNOTATIONS]  GT:AD:DP:GQ:PL  1/1:0,5:5:15:97,15,0    1/1:0,4:4:12:80,12,0
chr2R 10834    .    A  ACTG   173.04    .     [ANNOTATIONS]  GT:AD:DP:GQ:PL  0/0:14,0:14:33:0,33,495  0/1:6,3:9:99:105,0,315
```

[ANNOTATIONS]: all kinds of quantities and flags that characterize the variant; supplied by the variant caller (different callers will do it differently)

Example:

```
AC=2;AF=0.333;AN=6;DP=16;FS=0.000;GQ_MEAN=16.00;GQ_STDDEV=10.54;MLEAC=2;MLEAF=0.33
3;MQ=25.00;MQ0=0;NCC=1;QD=22.51;SOR=3.611
```

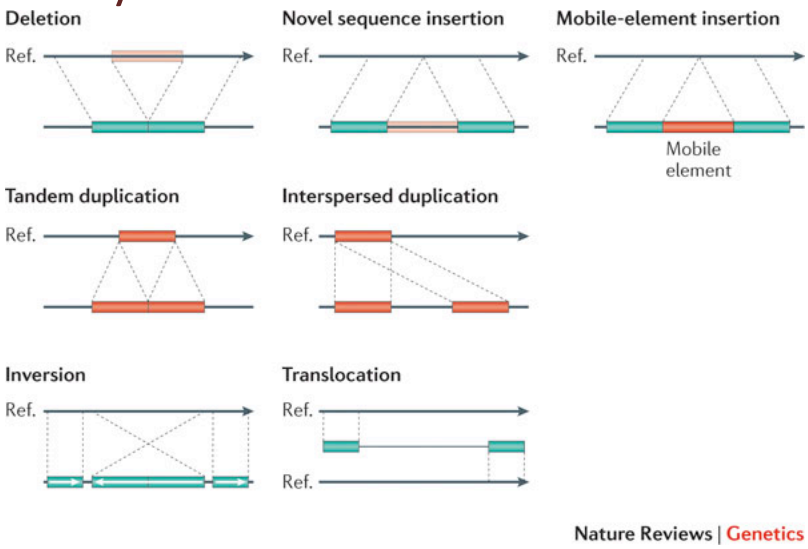
All ANNOTATION parameters are defined in the **HEADER LINES** on top of the file

```
…
##INFO=<ID=AC,Number=A,Type=Integer,Description="Allele count in genotypes, for each ALT allele, in the same order as listed">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency, for each ALT allele, in the same order as listed">
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth; some reads may have been filtered">
##INFO=<ID=FS,Number=1,Type=Float,Description="Phred-scaled p-value using Fisher's exact test to detect strand bias">
##INFO=<ID=GQ_MEAN,Number=1,Type=Float,Description="Mean of all GQ values">
##INFO=<ID=MQ,Number=1,Type=Float,Description="RMS Mapping Quality">
##INFO=<ID=NCC,Number=1,Type=Integer,Description="Number of no-called samples">
##INFO=<ID=QD,Number=1,Type=Float,Description="Variant Confidence/Quality by Depth">
##INFO=<ID=SOR,Number=1,Type=Float,Description="Symmetric Odds Ratio of 2x2 contingency table to detect strand bias">
…
```

Types of Variations

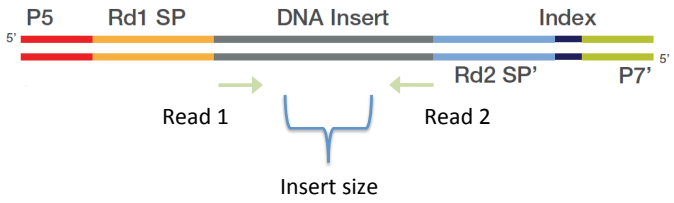
- SNPs
- Indels
- Structural Variation

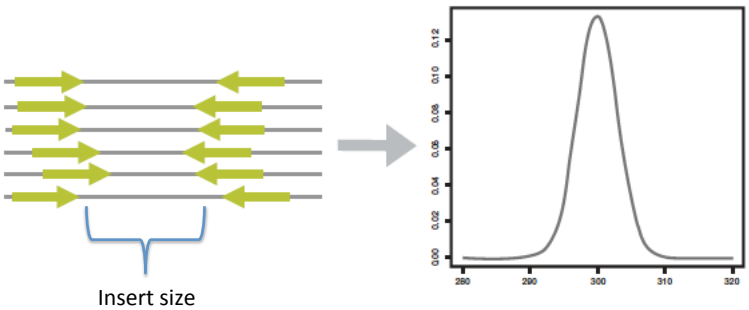
Many classes of structural elements.



Alkan, C. et al. Genome structural variation discovery and genotyping. Nature Reviews Genetics 12, 363-376 (2011).

Detecting most structural variants depends on paired end reads.





Insert size
within specified range