

---

**RE-ALIGNMENT**

# Alignment errors during mapping require fix

		coor	12345678901234	5678901234567890123456
9	t	ttt	ref	aggttttataaaaac----aattaagtctacagagcaacta
10	a	aaaC	sample	aggttttataaaaacAAATaattaagtctacagagcaacta
11	a	aaaaaa	read1	aggttttataaaaac aaAtaa
12	a	aaaaaaa	read2	gtttttataaaaac aaAtaaTt
13	a	aaaaaaa	read3	ttataaaaac AAATaattaagtctaca
14	c	cccTTT	read4	CaaaT aattaagtctacagagcaac
15	a	aaaaaaa	read5	aaT aattaagtctacagagcaact
16	a	aaaaaaa	read6	T aattaagtctacagagcaacta
17	t	AAtttt	read1	aggttttataaaaacaaataa
18	t	tttttt	read2	gtttttataaaaacaataatt
19	a	aaaaaaa	read3	ttataaaaacaataattaagtctaca
20	a	aaaaaaa	read4	caaataattaagtctacagagcaac
21	g	Tgggg	read5	aataattaagtctacagagcaact
			read6	taattaagtctacagagcaacta

# Alignment of an insertion

Ref      A A A C A A T T A A G T

Sample      AAAT

Sample A A A C A A A T A A T T A A G T

Ref      A A A C - - - A A T T A A G T

Sample A A A C A A A T A A T T A A G T

Correct alignment

Sample read A A A C A A A T A A T T

Ref      A A A C - - - A A T T A A G T

Sample read A A A C A A A T A A T T

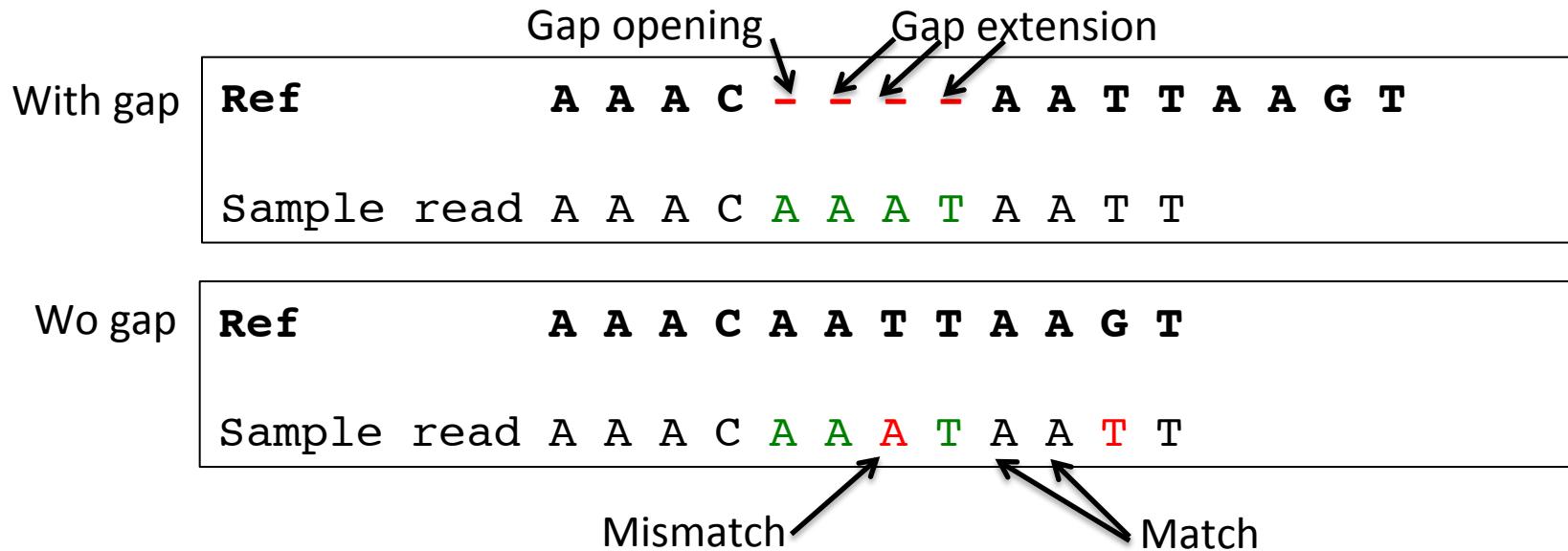
Ref      A A A C A A T T A A G T

Sample read A A A C A A A T A A T T

Correct alignment

Possible alignment

# Alignment



- Key component of alignment algorithm is the scoring
  - negative contribution to score
    - opening a gap
    - extending a gap
    - mismatches
  - positive contribution to score
    - matches
- The exact score contributions determine which alignment is chosen
- **Smith-Waterman** is an algorithm for finding optimal alignment given a scoring scheme without exhaustively enumerating and scoring all possible alignments

# Longer reads or multiple sequences

## Longer reads

With gap

**Ref** A A A C - - - A A T T | A A G T

Sample read A A A C **A A A T** A A T T | A A G T

Wo gap

**Ref** A A A C A A T T A A G T | C T A C

Sample read A A A C **A A A T** A A **T T** | **A A G T**

## Multiple reads

With gap

**Ref** A A A C - - - A A T T A A G T C T

Sample read A A A C **A A A T** A A T T

A C **A A A T** A A T T A A

A A T T A A G T C T

Wo gap

**Ref** A A A C A A T T A A G T C T A C

Sample read A A A C **A A A T** A A **T T**

A C **A A A T** A A **T T** **A A**

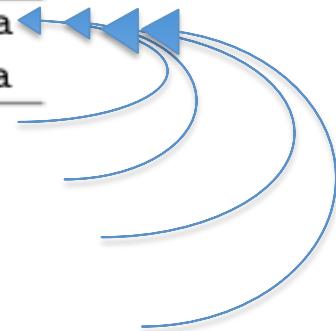
A A **T T** A A **G T C T**

# Few mismatches when considering one-to-one

Example: sample has insertion of AAAT relative to reference

## Base stacks

		coor	12345678901234	5678901234567890123456
9	t	ttt	ref	agg <del>ttttataaaac</del> ----aattaagtctacagagcaacta
10	a	aaaC	sample	agg <del>ttttataaaac</del> <ins>AAAT</ins> aattaagtctacagagcaacta
11	a	aaaaa	read1	agg <del>ttttataaaac</del> <ins>aaAt</ins> aa
12	a	aaaaaa	read2	g <del>ttttataaaac</del> <ins>aaAt</ins> aaTt
13	a	aaaaaaa	read3	ttataaaaac <ins>AAAT</ins> aattaagtctaca
14	c	cccTTT	read4	<ins>CaaaT</ins> aattaagtctacagagcaac
15	a	aaaaaa	read5	<ins>aaT</ins> aattaagtctacagagcaact
16	a	aaaaaa	read6	<ins>T</ins> aattaagtctacagagcaacta
17	t	AA <del>tttt</del>	read1	agg <del>ttttataaaac</del> <ins>aaataa</ins>
18	t	tttttt	read2	g <del>ttttataaaac</del> <ins>aaataa</ins> att
19	a	aaaaaa	read3	ttataaaaac <ins>aaataa</ins> attaagtctaca
20	a	aaaaaa	read4	<ins>caaataa</ins> attaagtctacagagcaac
21	g	Tgggg	read5	<ins>aataa</ins> attaagtctacagagcaact
			read6	<ins>taatt</ins> aagtctacagagcaacta



# Lots of mismatch in all-to-all if reads mismapped

## Base stacks

		coor	12345678901234	5678901234567890123456	
9	t	ttt	ref	agg <del>ttttataaaac</del> ----aattaagtctacagagcaacta	
10	a	aaaC	sample	agg <del>ttttataaaac</del> <ins>AAAT</ins> aattaagtctacagagcaacta	
11	a	aaaaa	read1	agg <del>ttttataaaac</del> <ins>aaAt</ins> aa	
12	a	aaaaaa	read2	ggttttataaaac <ins>aaAt</ins> aaTt	
13	a	aaaaaaa	read3	ttataaaaac <ins>AAAT</ins> aattaagtctaca	
14	c	cccTTT	read4	<ins>CaaaT</ins> aattaagtctacagagcaac	
15	a	aaaaaa	read5	<ins>aaT</ins> aattaagtctacagagcaact	
16	a	aaaaaa	read6	T aattaagtctacagagcaacta	
17	t	AA <del>tttt</del>	read1	agg <del>ttttataaaac</del> <ins>aaataa</ins>	
18	t	tttttt	read2	ggttttataaaac <ins>aaata</ins> att	
19	a	aaaaaa	read3	ttataaaaac <ins>aaata</ins> aattaagtctaca	No mismatches between reads
20	a	aaaaaa	read4	<ins>caaata</ins> aattaagtctacagagcaac	
21	g	Tgggg	read5	<ins>aata</ins> aattaagtctacagagcaact	
		read6		<ins>taatta</ins> aagtctacagagcaacta	

# Mapping vs. alignment

## Mapping vs. alignment

### Mapping

- A mapping is the region where a read sequence is placed.
- A mapping is regarded to be correct if it overlaps the true region.

### Alignment

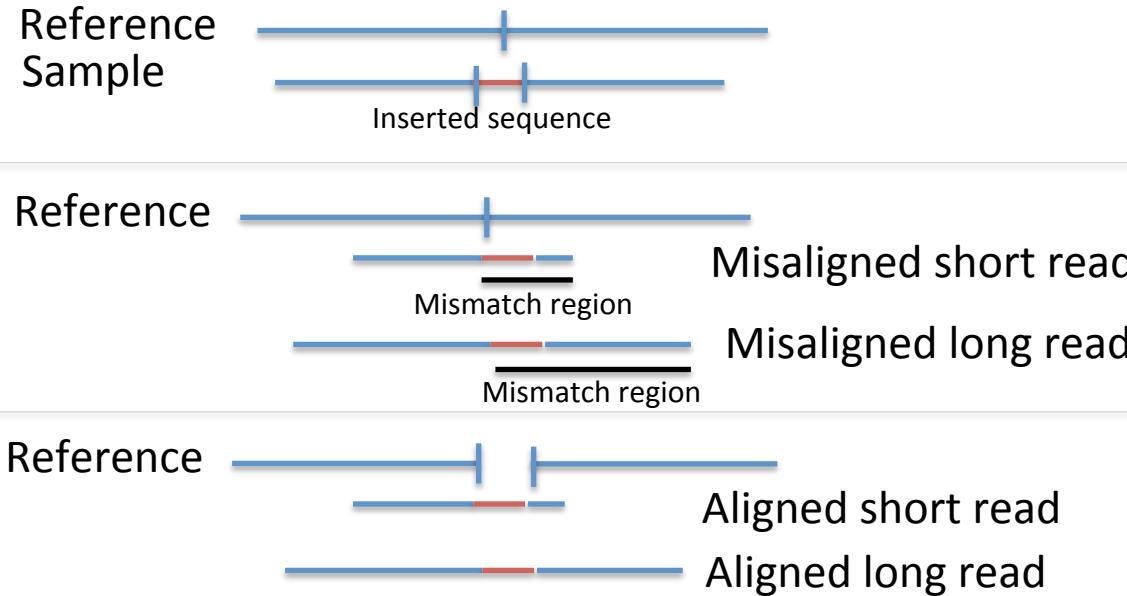
- An alignment is the detailed placement of each base in a read.
- An alignment is regarded to be correct only if each base is placed correctly.

### The problem

- A read mapper is fairly good at mapping, may not be good at alignment.
- This is because the true alignment minimizes differences between reads, but the read mapper only sees the reference.

# Detection of indels

## Longer read length facilitates correct alignment

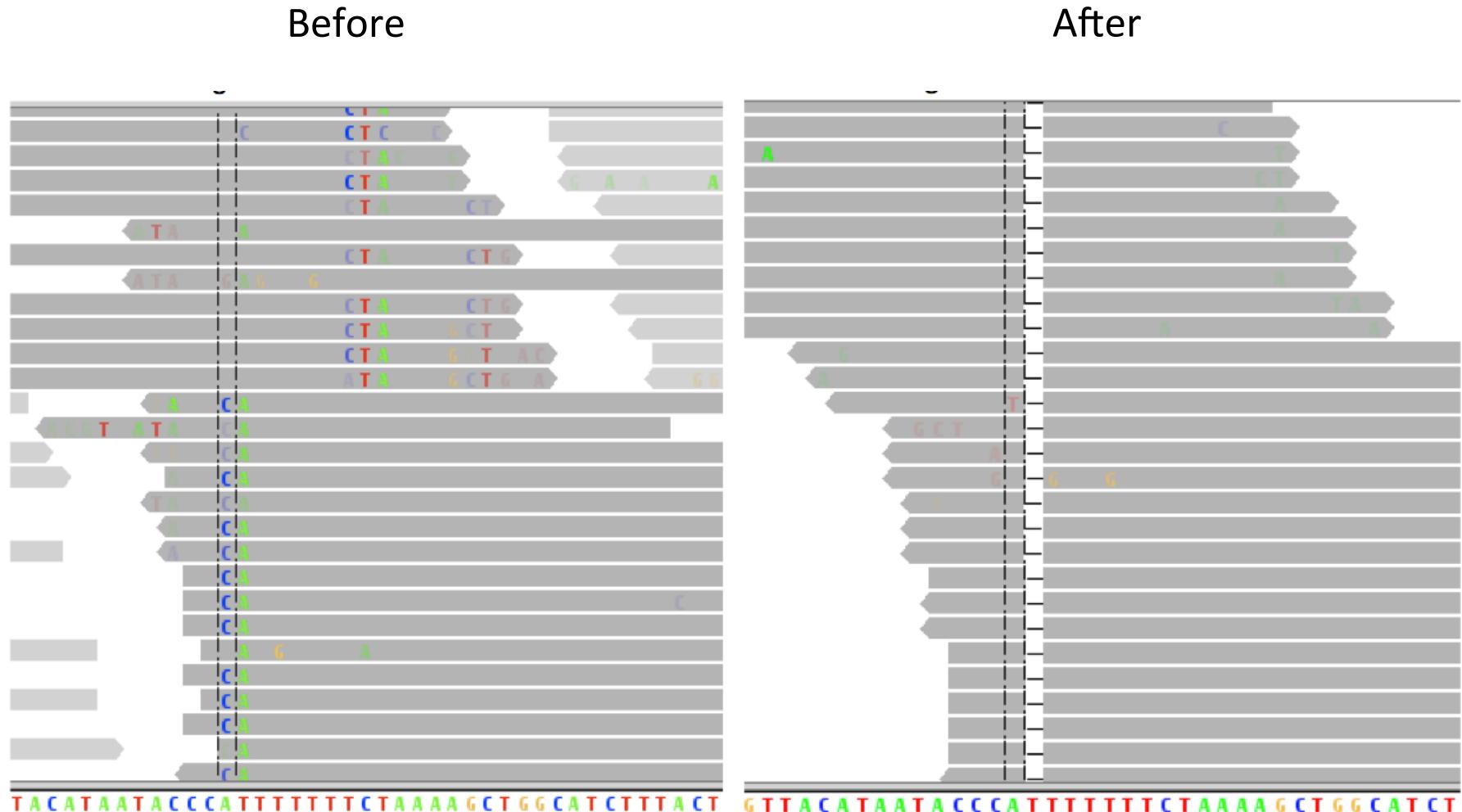


## Asymmetry between insertions and deletions



>> insertion and deletion of same size, but more likely to detect the deletion

# Local realignment around indels



---

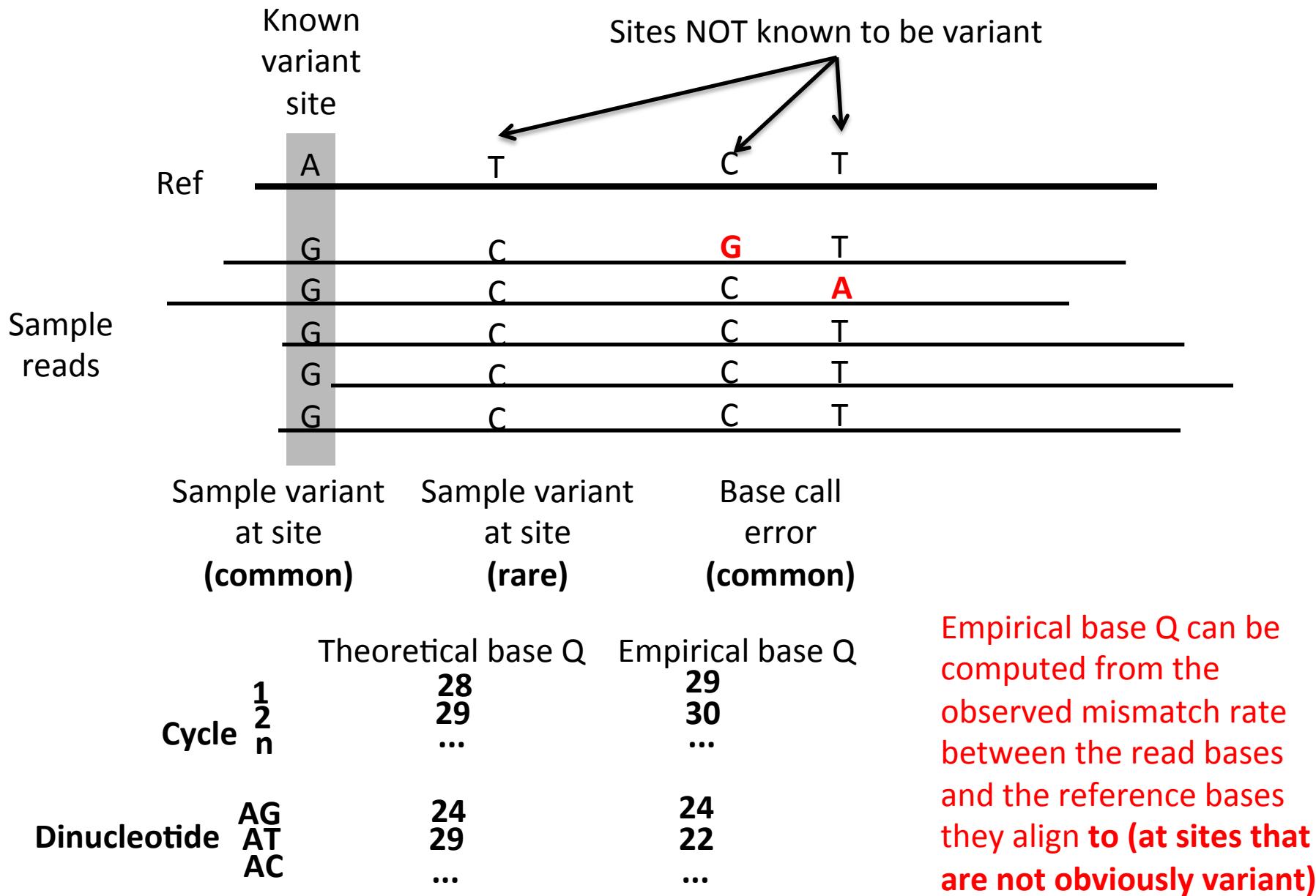
**BASE QUALITY SCORE RECALIB.  
(OPTIONAL)**

# Theoretical vs Empirical error rates / qualities

---

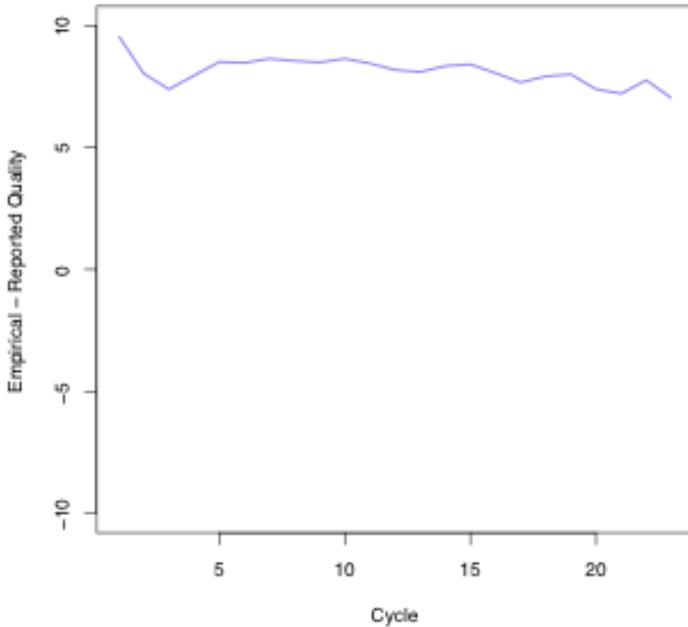
- The qualities in the fastq file are computed using a model
- This model is not perfect >> there are discrepancies between the model and the empirical error rate
- We can compute a good approximation of the empirical error rate by identifying all sites where there are mismatches between the read and the reference (**being careful to ignore sites with known SNPs**)
- We can analyse whether there are parameters of the bases that covary with the discrepancy
  - e.g. cycle
- We can use these quantified covariances to recalibrate the base qualities >> more accurate qualities

# Overview of Base Quality recalibration

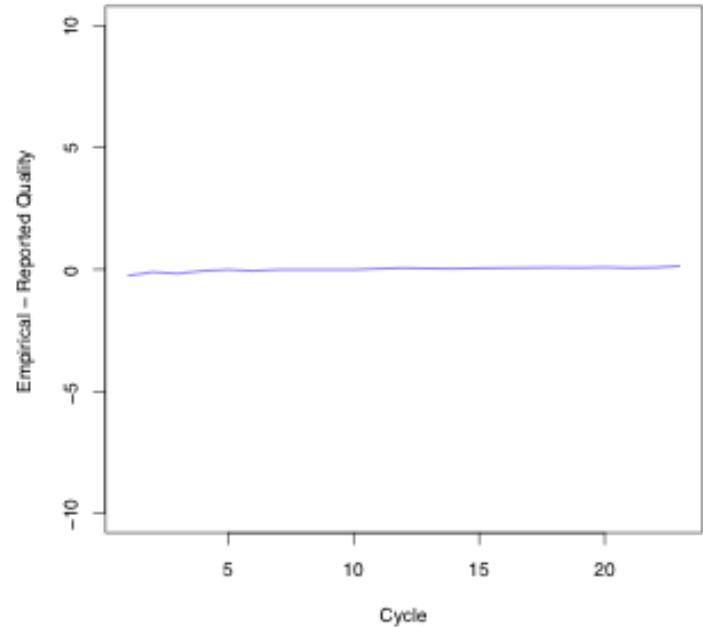


# Discrepancy and cycle

Original

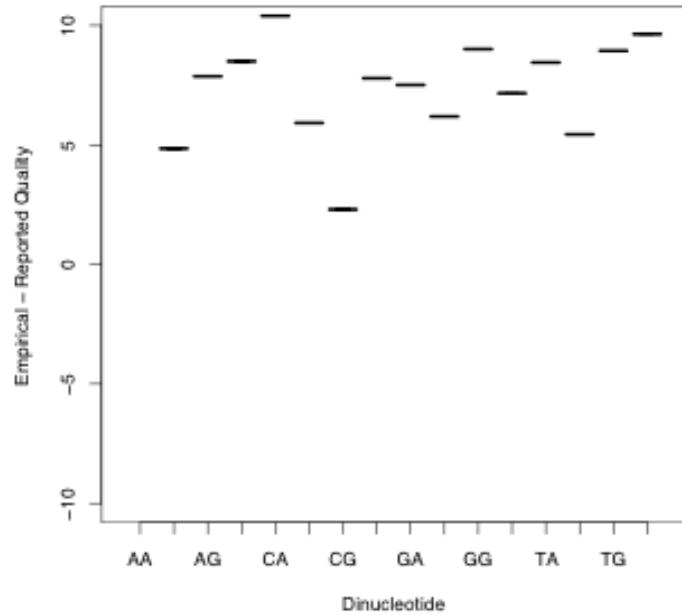


Recalibrated

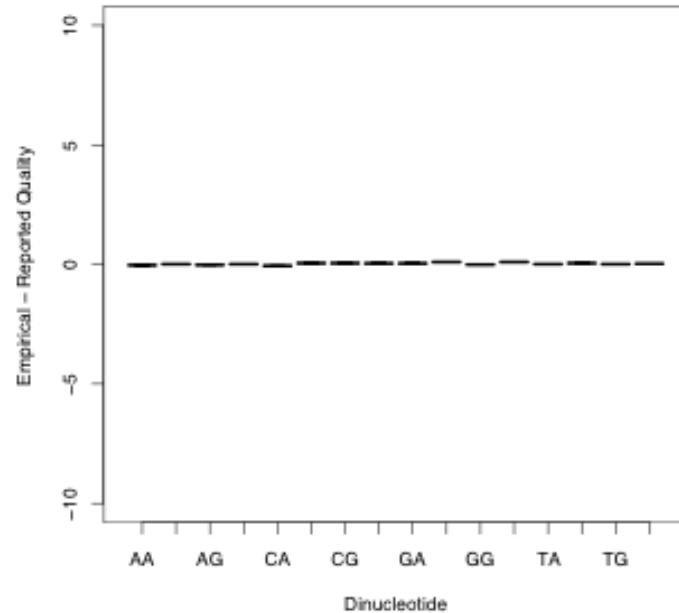


# Discrepancy and Dinuc context

Original



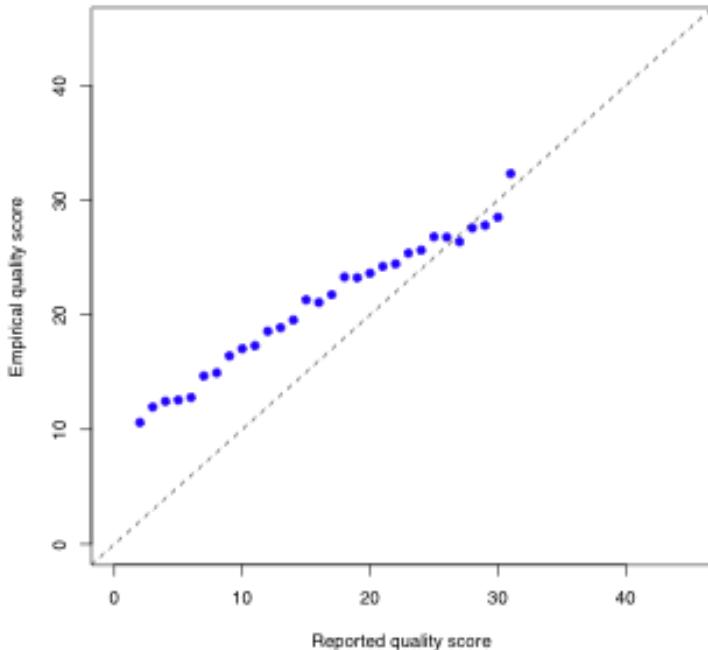
Recalibrated



# Result of recalibration

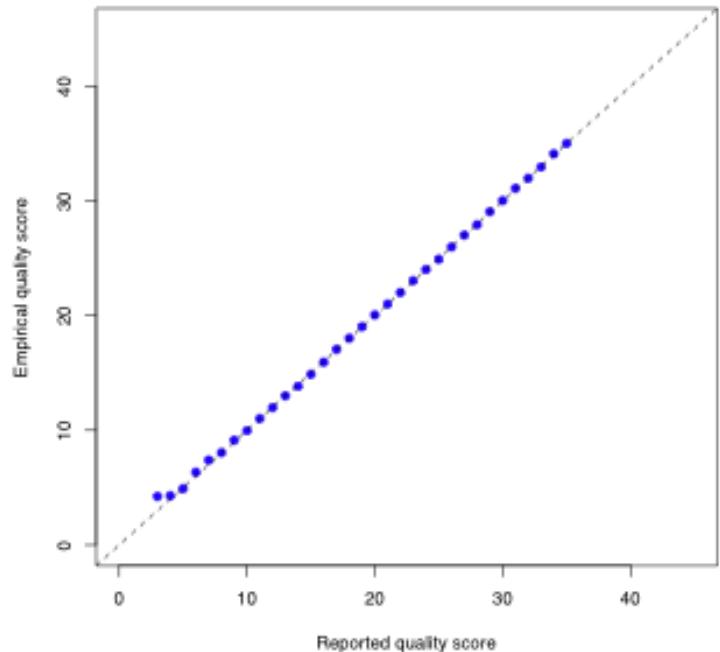
Original

Reported vs. empirical quality scores



Recalibrated

Reported vs. empirical quality scores



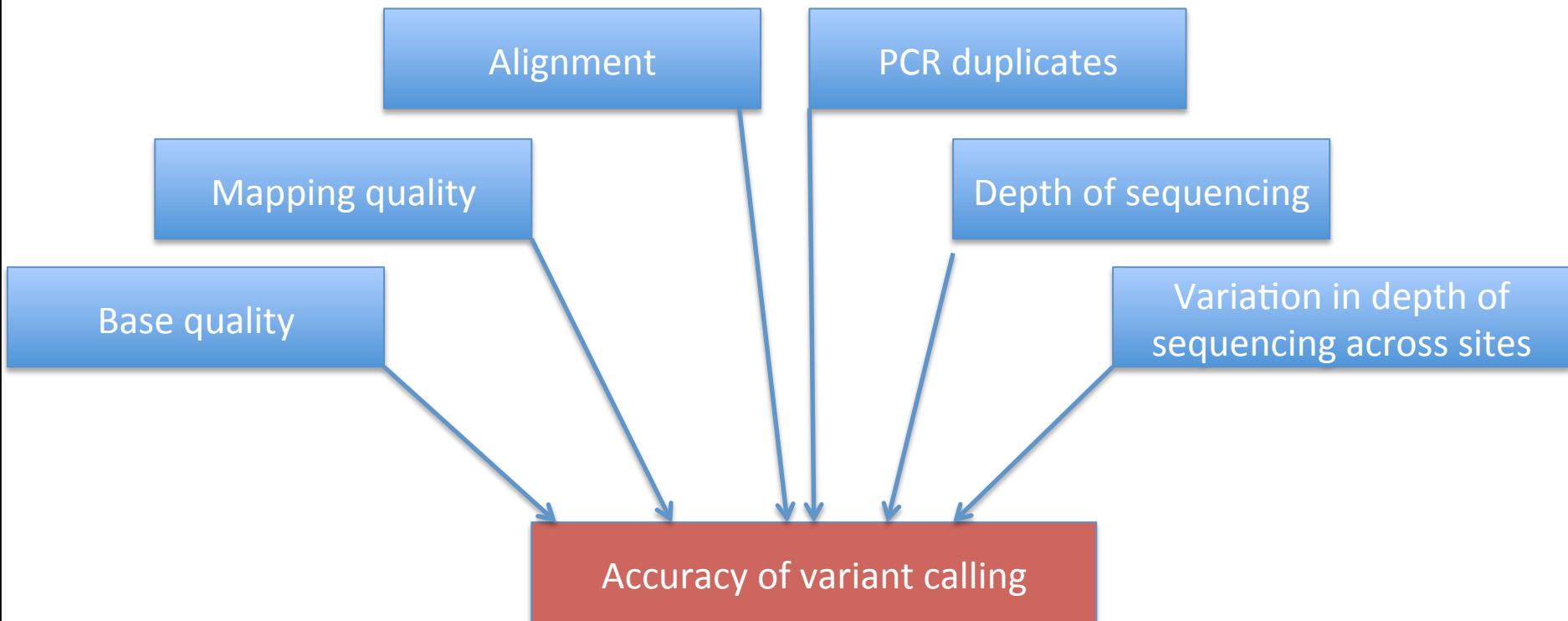
NB: The theoretical base qualities have become very good and there seems to be little to be gained in recalibrating “modern” fastq files

---

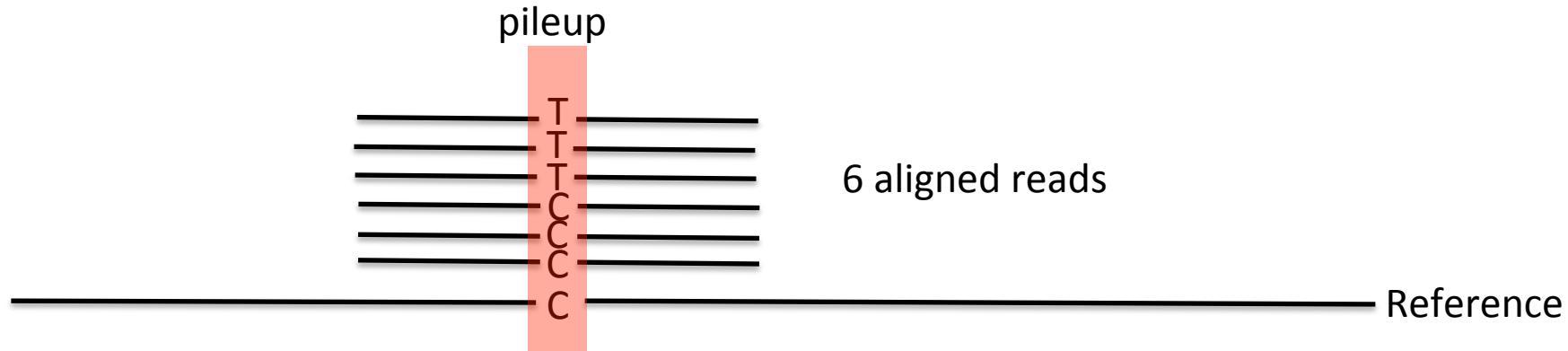
# **VARIANT CALLING**

# Recapping source of noise in variant calling

- If we had one long end-to-end read of the chromosome there would be no problem
- BUT we have short reads of imperfect quality



# A way of thinking of the variant calling process



- Compute:
  - the probability that the site is variant:  $P(\text{not CC} / \text{read data})$
  - the probability of the different genotypes:  $P(CC / \text{read data})$ ,  $P(CT / \text{read data})$ ,  $P(TT / \text{read data})$
- Intuition of difference between variant site and genotype:
  - **ref is C**, aligned bases are TTTTTTCC
  - highly likely that the site is variant
  - less clear what the genotype is: T/A or T/T?
- **Only good bases are considered in the pileup**
  - **Minimum base quality**
  - **Read mapping quality**
- The importance of depth of coverage
- Complex mathematical models involved in both allele frequency calculations and genotype likelihoods >> wise to use the recommended option settings in the tool documentation (as we have done in the practicals)

# Bayesian variant caller (optional)

## Input

Reference is C, observing 4C and 2T, all with base quality 30.

## Likelihood of data

- $P(D|CC) = \Pr\{\text{two Q30 errors}\} = 10^{-(30+30)/10} = 10^{-6}$
- $P(D|TT) = \Pr\{\text{four Q30 errors}\} = 10^{-(30*4)/10} = 10^{-12}$
- $P(D|CT) = \Pr\{\text{sample 6 reads from 2 chr}\} = 1/2^6 = 1.56 \times 10^{-2}$

## Posterior

- Prior:  $P(CC) = 0.9985$ ,  $P(CT) = 0.001$  and  $P(TT) = 0.0005$

$$P(CC|D) = \frac{P(D|CC)P(CC)}{P(D|CC)P(CC) + P(D|CT)P(CT) + P(D|TT)P(TT)}$$

- Get:  $P(CC|D) = 0.06$ ,  $P(CT|D) = 0.94$  and  $P(TT|D) = 3 \times 10^{-11}$

---

# **VCF FORMAT – MORE DETAILS**

# VCF format

```
##fileformat=VCFv4.0
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=1000GenomesPilot-NCBI36
##phasing=partial
##INFO<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO<ID=AF,Number=.,Type=Float,Description="Allele Frequency">
##INFO<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER<ID=q10,Description="Quality below 10">
##FILTER<ID=s50,Description="Less than 50% of samples have data">
##FORMAT<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS ID REF ALT QUAL FILTER INFO
20 14370 rs6054257 G A 29 PASS NS=3;DP=14;AF=0.5;DB;H2
20 17330 . T A 3 q10 NS=3;DP=11;AF=0.017
20 1110696 rs6040355 A G,T 67 PASS NS=2;DP=10;AF=0.333,0.667;AA=T;DB
20 1230237 . T . 47 PASS NS=3;DP=13;AA=T
20 1234567 microsat1 GTCT G,GTACT 50 PASS NS=3;DP=9;AA=G
```

**Meta data:**  
definitions of  
tags used  
elsewhere in  
data lines

**Header line**

FORMAT	NA00001	NA001
GT:GQ:DP:HQ	0 0:48:1:51,51	
GT:GQ:DP:HQ	0 0:49:3:58,50	
GT:GQ:DP:HQ	1 2:21:6:23,27	
GT:GQ:DP:HQ	0 0:54:7:56,60	
GT:GQ:DP	0/1:35:4	

**Data lines**

**Variant columns**

**Genotype columns**

# Columns of data lines

---

- **CHROMO**
- **POS:** the reference position with the 1<sup>st</sup> base having position 1
- **ID:** an id; rs number if dbSNP variant
- **REF:** reference base.
  - The value in POS refers to the position of the first base in the string
  - for indels, the reference string must include the base before the event (and this must be reflected in POS)
- **ALT:** comma separated list of alternate non-ref alleles called on at least one of the samples
  - if no alternate alleles then the missing value should be used “.”
- **QUAL:** phred-scaled quality score of the assertion made in ALT (whether variant or non-variant)
- **FILTER:** PASS if the position has passed all filters (defined in meta-data).
- **INFO:** additional information

# INFO, FORMAT, and genotypes

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	sample1
1	801943	rs7516866	C	T	9787.34	PASS			
AC=2;									
AF=1.00;									
AN=2;									
BaseQRankSum=1.009;									
DB;									
DP=556;									
FS=18.302;									
MQ=44.04;									
MQ0=38;									
MQRankSum=5.122;									
QD=17.60;									
ReadPosRankSum=3.375									
GT:AD:DP:GQ:PL									
1/1:37,518:556:99:9787,685,0									

We will explore  
these fields  
when we  
discuss filtering

# Genotype fields

---

- Format field specifies type of data present for each genotype
  - GT:AD:DP:GQ:PL
  - fields defined in metadata header
- GT: genotype, encoded as alleles separated by either | or /
  - 0 for the ref, 1 for the 1<sup>st</sup> allele listed in ALT, 2 for the second, etc
  - REF=A and ALT=T
    - genotype 0/1 means hetero A/T
    - genotype 1/1 means homo T/T
  - /: genotype unphased and | genotype phased
- DP: read depth at position for sample
- GQ: genotype quality encoded as a phred quality
- etc.....

# Homozygous SNP

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT
1	801943	rs7516866	C	T	9787.34	PASS		

AC=2;AF=1.00;AN=2;BaseQRankSum=1.009;DB;DP=556;DS;Dels=0.00;  
FS=18.302;HRun=1;HaplotypeScore=4.6410;MQ=44.04;MQ0=38;MQRankSum=5.122;QD=17.60;ReadPosRankSum=3.375

**GT:AD:DP:GQ:PL** **1/1**:37,518:556:99:9787,685,0

# Heterozygous SNP

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT
--------	-----	----	-----	-----	------	--------	------	--------

1	1918488	rs4350140	A	G	233.10	PASS		
---	---------	-----------	---	---	--------	------	--	--

AC=1;AF=0.50;AN=2;BaseQRankSum=1.349;DB;DP=33;DS;Dels=0.00;  
FS=0.000;HRun=0;HaplotypeScore=0.0000;MQ=68.18;MQ0=1;MQRankSum=0.436;QD=7.06;ReadPosRankSum=1.547

**GT:AD:DP:GQ:PL** **0/1:21,12:33:99:263,0,620**

# Homozygous deletion

---

**#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT**

1 1289367 rs35062587 **CTG** **C** 3139.27 PASS

AC=2;AF=1.00;AN=2;DB;DP=66;DS=0.000;HRun=0;HaplotypeScore=223.1329;MQ=68.34;MQ0=1;QD=47.56

**GT:AD:DP:GQ:PL** **1/1**:0,66:65:99:3181,196,0

# Heterozygous insertion

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT
--------	-----	----	-----	-----	------	--------	------	--------

1	17948305	.	G	GGGCCACAGCAG	3581.32	PASS		
---	----------	---	---	--------------	---------	------	--	--

AC=1;AF=0.50;AN=2;BaseQRankSum=-2.638;DP=54;DS;FS=0.000;HR  
un=0;HaplotypeScore=552.8152;MQ=70.65;MQ0=2;MQRankSum=3.  
258;QD=66.32;ReadPosRankSum=0.320

**GT:AD:DP:GQ:PL** **0/1:44,10:52:99:3581,0,3730**

---

# **VARIANT FILTERING**

# The rationale for filtering

---

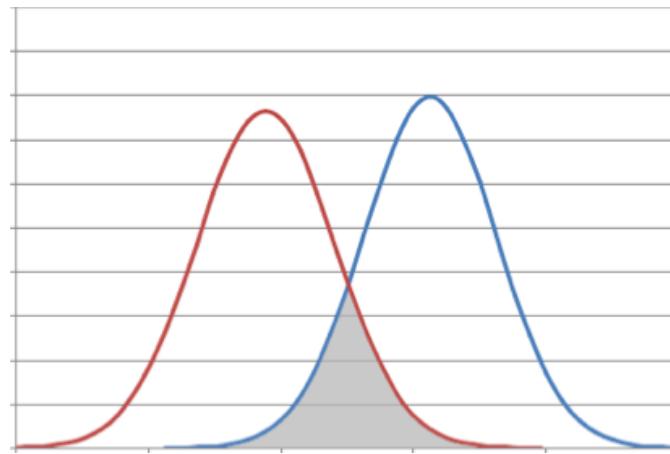
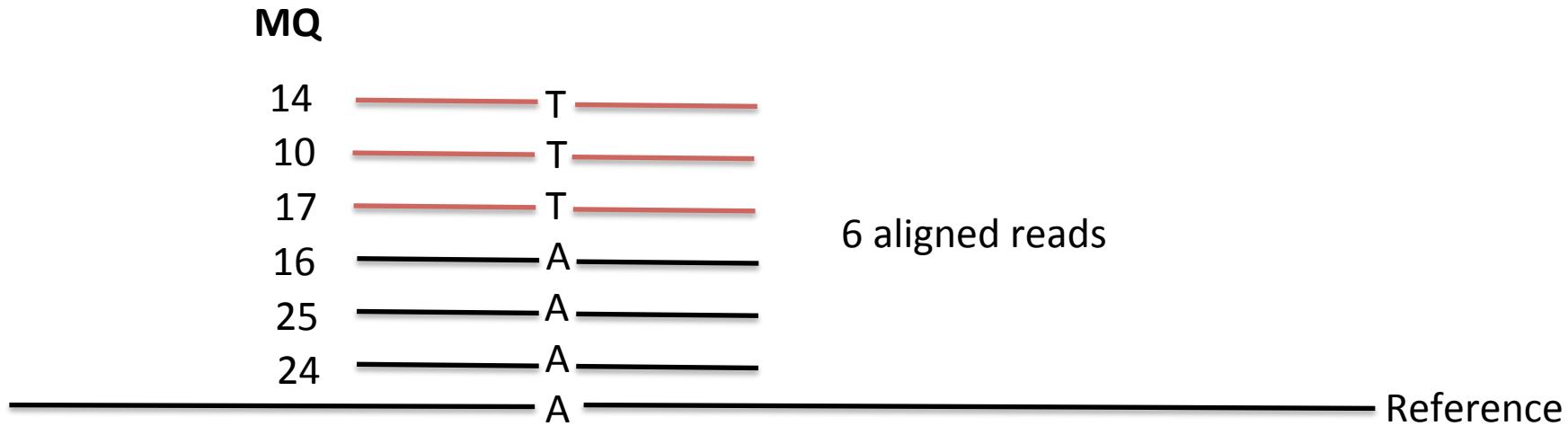
- QUAL is a basic measure of variant quality but it assumes that reads are correctly aligned and there are no systematic base call errors
- What causes errors in variant calling?
  - sequencing errors >> should be accounted for by base quality + recalibration + marking of duplicates
  - Incorrect alignment >> **Re-alignment step should have reduced this problem but not eliminated it**
- **Thus although QUAL (which depends on Mapping Quality of reads and Base qualities) is a useful measure, there will still be FP with high QUAL.**
  - and we wish to reduce the number of FP
- Tell tale signs of suspicious variants
  - poorly mapped reads (ambiguity)
    - MQ: Root Mean Square of MAPQ of all reads at locus
    - MQ0: Number of MAPQ 0 reads at locus
  - biased support for the **REF** and **ALT** alleles
    - MQRankSum: Mapping quality rank sum test
    - ReadPosRankSum: Read position rank sum test
    - Strand bias and FS

# INFO fields – important for filtering

---

- **QD:** variant quality score over depth of variant allele
  - Confidence in the site being variant should increase with increasing depth
- **MQ:** RMS MAPQ of all reads at locus
  - Regions of excessively low mapping quality are ambiguously mapped and variants called within are suspicious
- **MQ0:** number of MAPQ 0 reads at locus
- **MQRankSum:** Mapping quality rank sum test
  - If the alternate bases are more likely to be found on reads with lower MQ than reference bases then the site is likely mismapped
- **Haplotype score:** Probability that the reads in a window around the variant can be explained by at most two haplotypes
- **FS:** fisher exact test of read strand
  - If the reference-carrying reads are balanced between forward and reverse strands then the alternate-carrying reads should be as well
- **ReadPosRankSum:** Read position rank sum test
  - If the alternate bases are biased towards the beginning or end of the reads then the site is likely a mapping artifact

# Details of MQRankSum

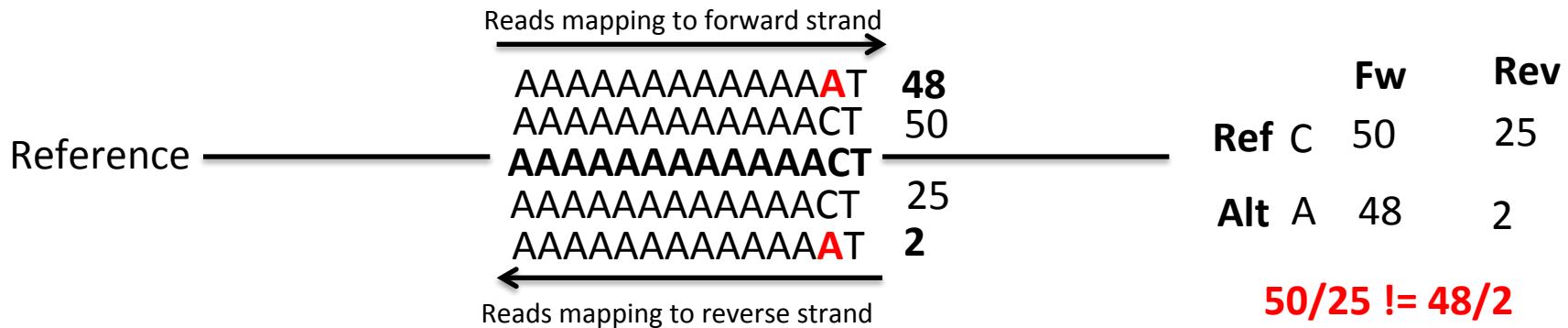
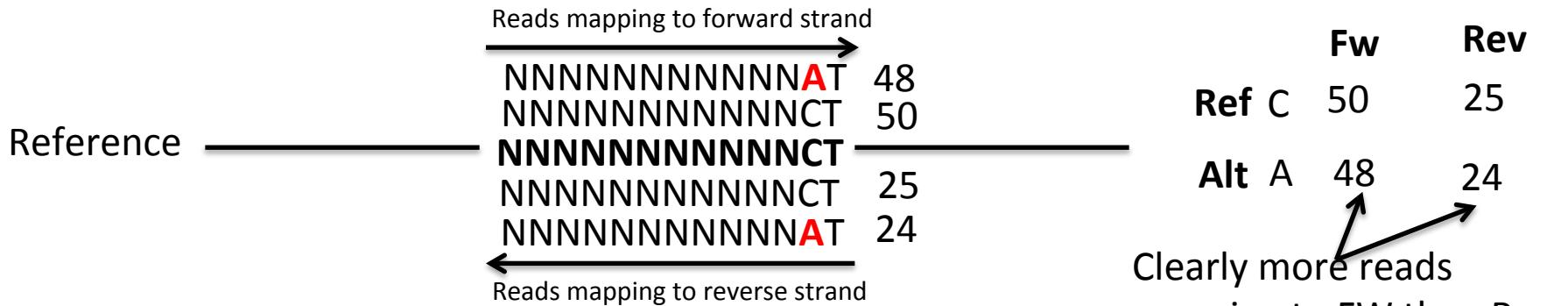


Rank sum test are used to test if two samples come from the same population when the distribution is unknown.

- We test the null hypothesis that the two samples of MQ are the same
- If we reject this null hypothesis, the variant site is likely to be a False Positive

# Strand bias (assume heterozygote) - optional

Strand bias is **NOT** about more reads mapping to one of the strands than the other

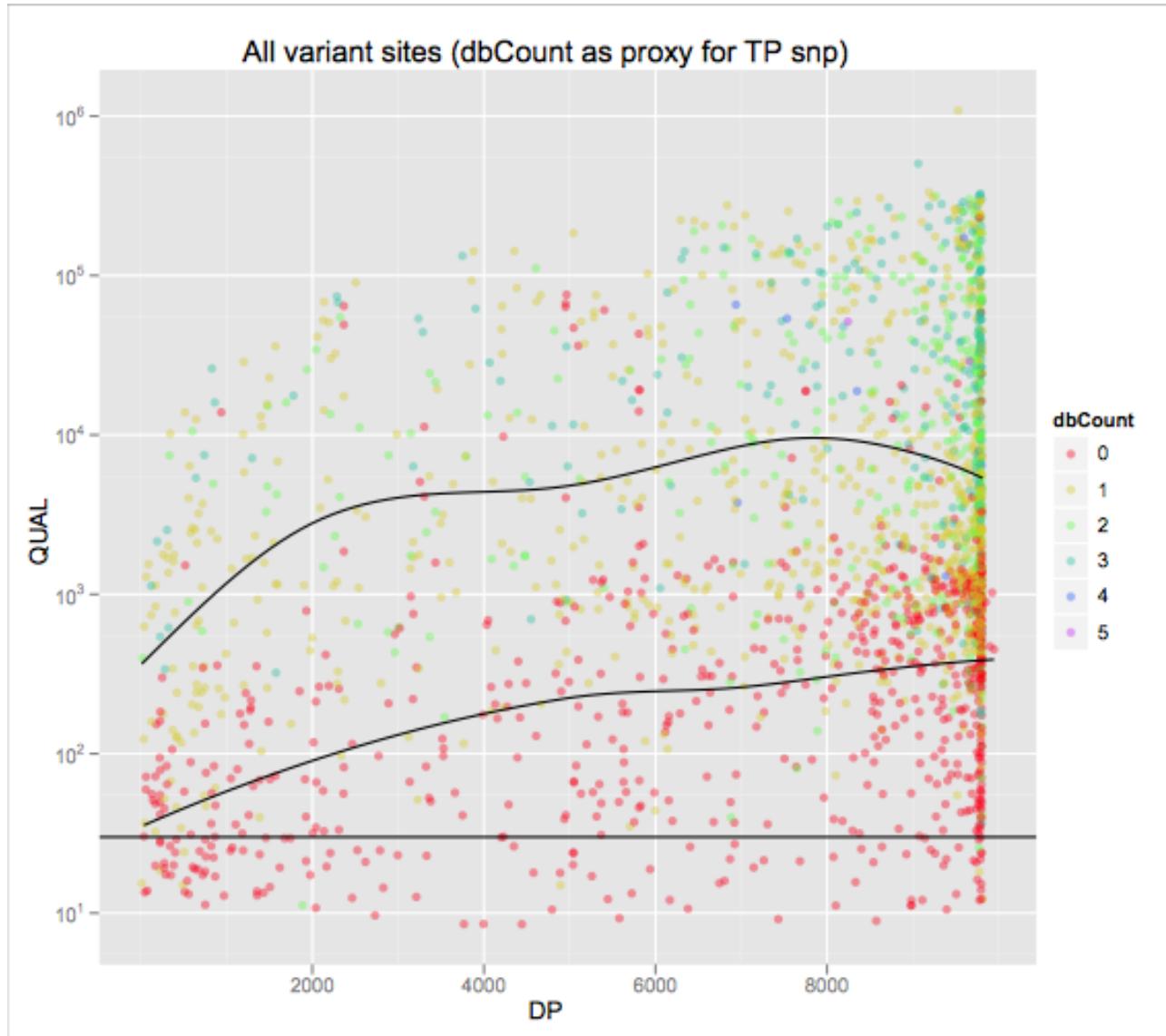


# Hard vs. soft filtering

---

- Can set thresholds for these INFO fields and request that all thresholds are passed for a variant to be considered valid
- Which fields do you use and where do you set the thresholds?
  - use datasets of known SNPs and compare their INFO fields to those likely FP variants
  - fields that provide a good separation can be used as filters
- Disadvantage of **hard filtering**
  - works with hard cut-offs
- Variant Quality Score Recalibration (GATK) or **soft filtering**
  - use machine learning to learn the features of true variants and distinguish them from false positives

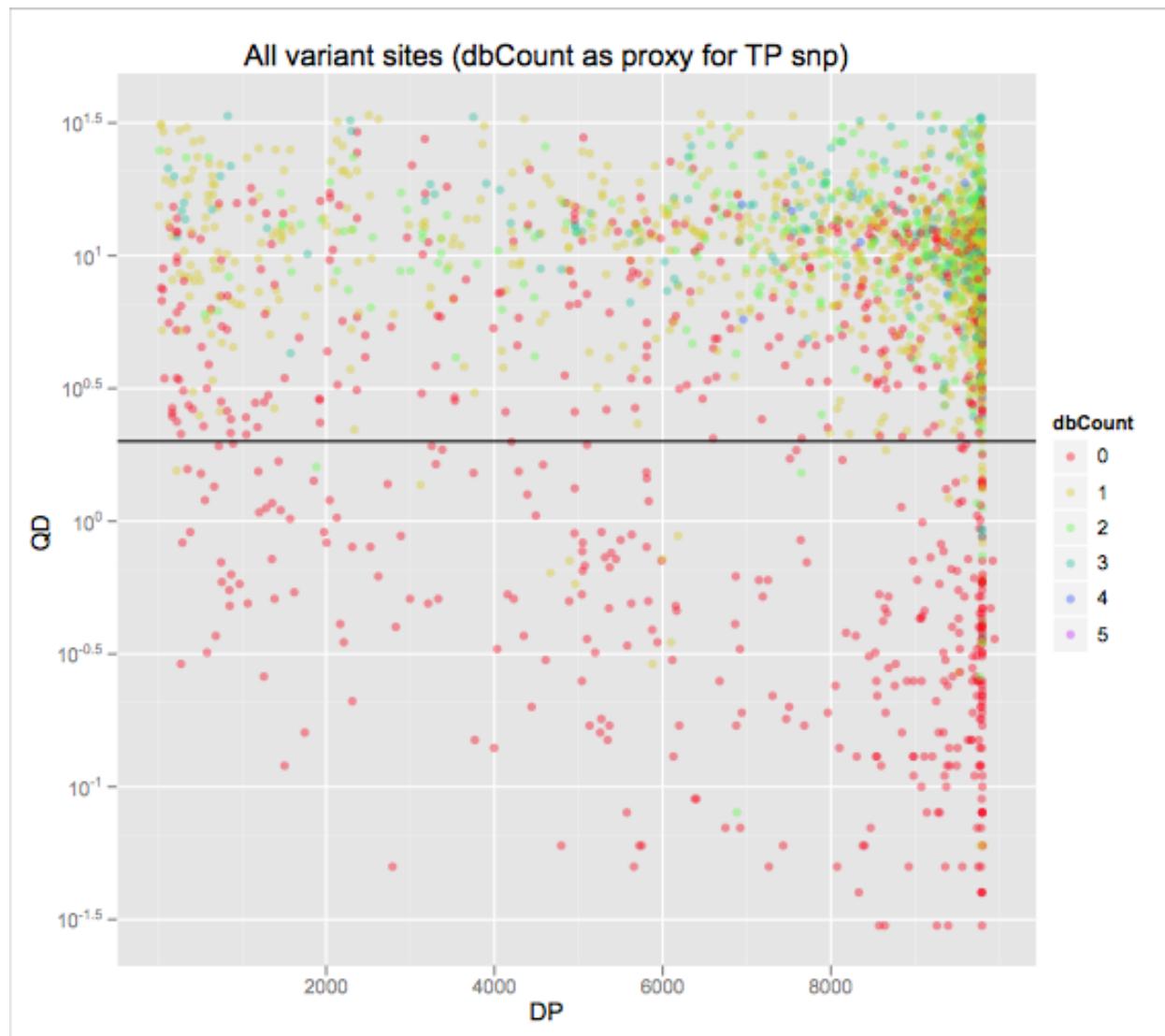
# QUAL provides OK but not great separation



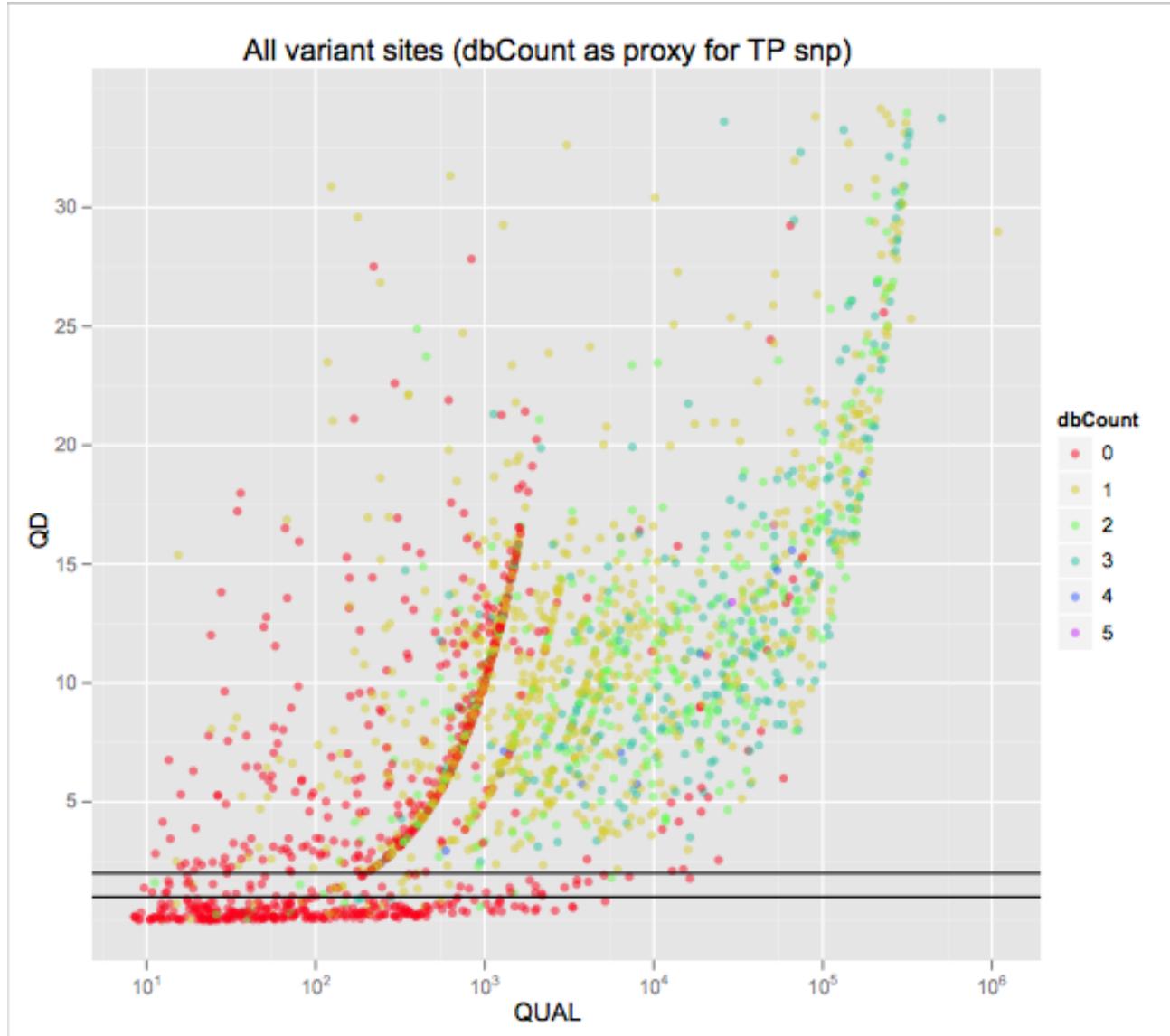
Note how:

- QUAL increases with depth
- QUAL for known SNPs (green) is higher than for unknown SNPs (red) at any given depth

# QD vs depth: much better separation



# Qual vs QD

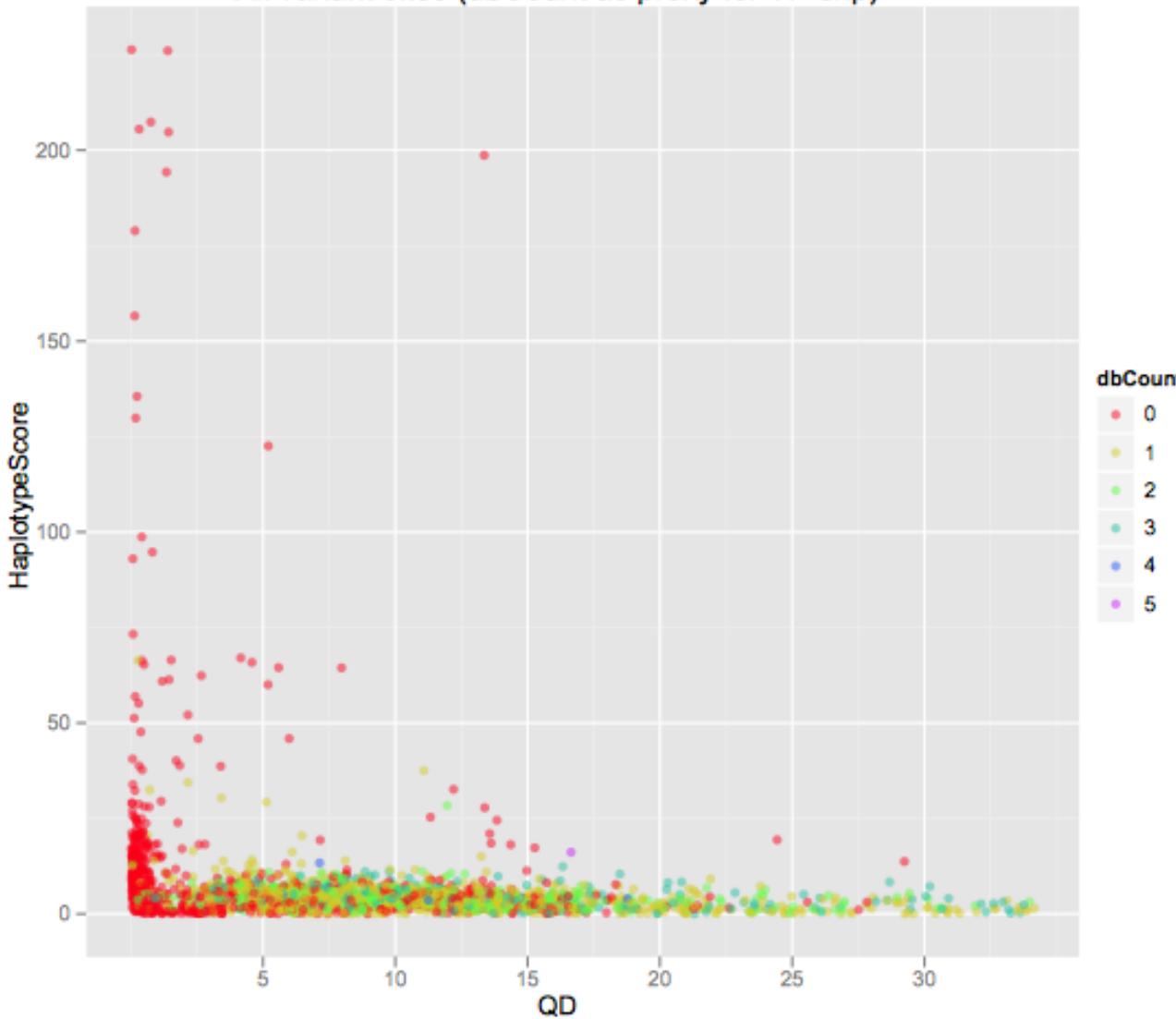


Illustrates advantages of soft filtering:

- learning features of true SNPs
- no fixed level cutoffs

# Haplotype Score vs QD: advantage of combining

All variant sites (dbCount as proxy for TP SNP)



No known SNPs even with  $QD > 3$  have high HaplotypeScore, so can remove more likely FP SNPs by filtering on HaplotypeScore

Can require SNPs to have HaplotypeScore less than 20.

# practical 03\_advancedPipeline.bash

---

- Now we can try out all the refinements we have discussed

# practical 031\_generatingReports.bash

---

- Let us see if all those refinements made any difference.

---

# **FUNCTIONAL ANNOTATION**

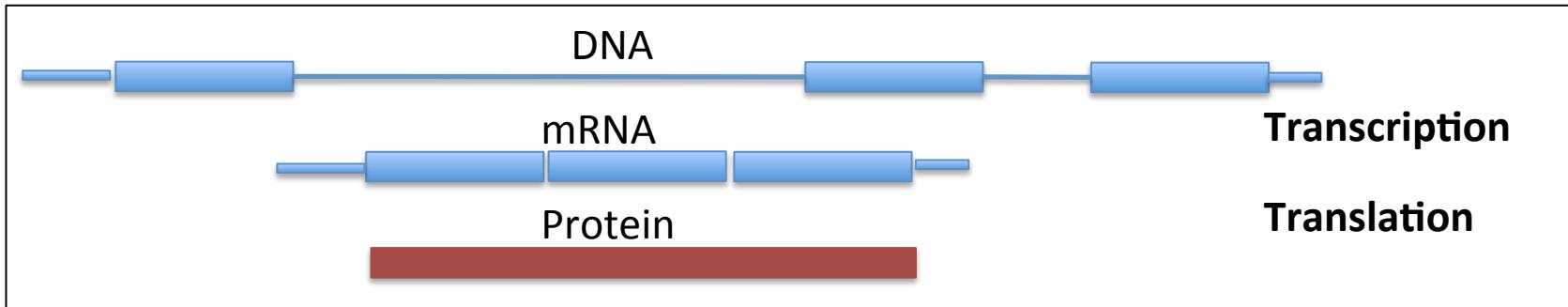
# Warning

---

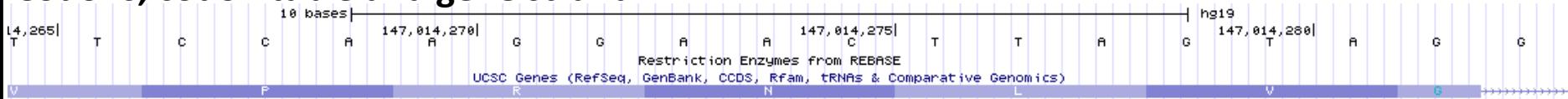
- You will be **rapidly** introduced for many functional annotations in this section
- **Unless you have a good prior background in molecular biology, you will struggle to understand all of these annotations**
- Keep in mind that the main goal is to use these annotations to distinguish variants with a big phenotypic impact from those with no effect or a very mild effect
  - in the case of monogenic disease, a single variant can be responsible for the entire clinical phenotype.

# What is a gene?

## Central Dogma



## Codons, codon table and gene strand



- There are large numbers of genes in multicellular organisms
- Not all organisms have multiple exons per gene and thus do not need splicing
- Outside genes are regulatory elements that control when and where genes are expressed

# Synonymous and non-synonymous mutations

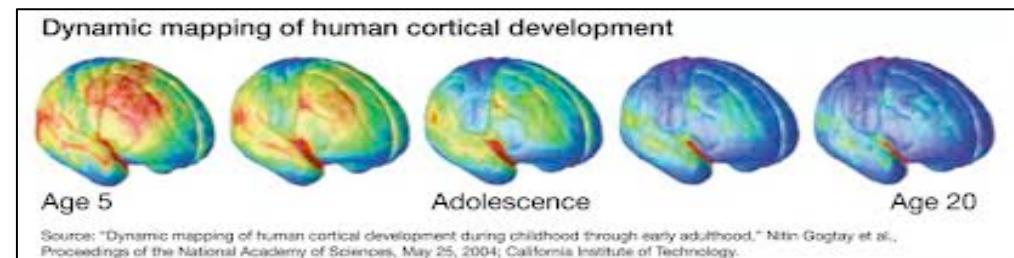
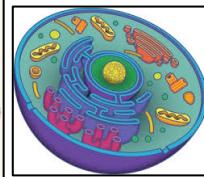
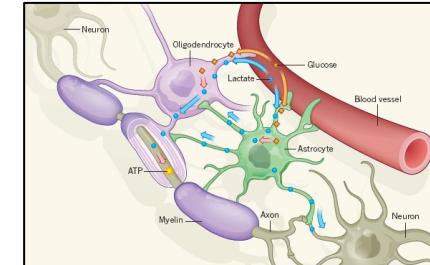
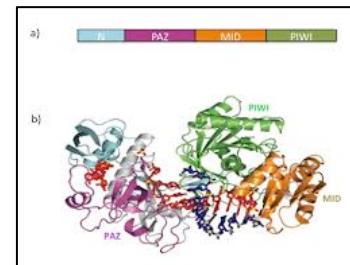
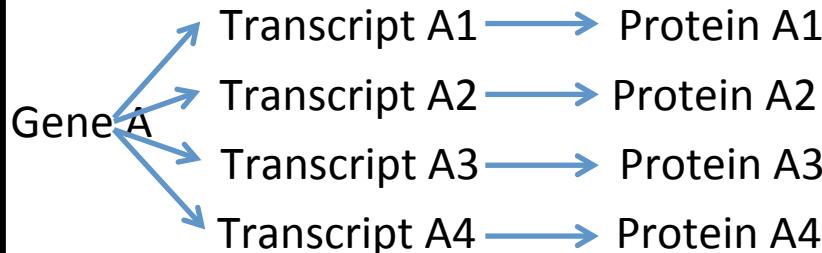
		Second letter					
		U	C	A	G		
First letter	U	UUU } Phe UUC } UUU } Leu UUG }	UCU } UCC } Ser UCA } UCG }	UAU } Tyr UAC } UAA Stop UAG Stop	UGU } Cys UGC } UGA Trp UGG Trp	U C A G	Third letter
	C	CUU } CUC } Leu CUA } CUG }	CCU } CCC } CCA } Pro CCG }	CAU } His CAC } CAA } Gln CAG }	CGU } CGC } Arg CGA } CGG }	U C A G	
	A	AUU } Ile AUC } AUA Red } Met AUG }	ACU } ACC } ACA } Thr ACG }	AAU } Asn AAC } AAA } Lys AAG }	AGU } Ser AGC } AGA Stop AGG Stop	U C A G	
	G	GUU } GUC } Val GUA } GUG }	GCU } GCC } GCA } Ala GCG }	GAU } Asp GAC } GAA } Glu GAG }	GGU } GGC } Gly GGA } GGG }	U C A G	

# Why are there different isoforms?

Approx. 25,000 genes in the human genome

**DNA >transcription> RNA >translation> PROTEIN**

**BUT** 1 gene  $\neq$  1 protein



**Some transcripts expressed ubiquitously, some highly specific (cell type AND timepoint)**

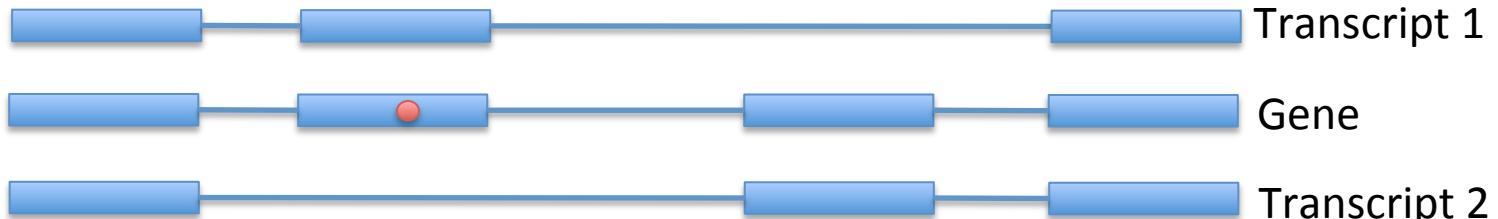
Other domains      Ankyrin repeats  
(spectrin binding)    (membrane binding)

chr10: 61,900,000 62,000,000 62,100,000 62,200,000 62,300,000 62,400,000 62,500,000

ANK3/NM\_0011149  
ANK3/NM\_020987  
ANK3/NM\_001204404  
ANK3/NM\_001204403

# What is annotation?

- Adding information about the variants
- Two broad categories of annotations
  - annotation that **depend on gene models**
    - coding/non-coding
    - if coding: synonymous / non-synonymous
    - if non-synonymous >> what is the impact on protein structure (Polyphen, SIFT, etc)
  - annotations that **do not depend on gene models**
    - variant frequency in different database / different populations
    - degree of conservation across species
- Considerable complications caused by different gene models



- Two approaches to problem
  - decide **ex-ante** what which transcript to use for each gene
  - annotate with all transcript for a given gene and pick the **highest impact effect**

# Ensembl – Basis for variant effect prediction

**e!Ensembl** BLAST/BLAT | BioMart | Tools | Downloads | Help & Documentation | Blog | Mirrors

Search:  for    
e.g. [BRCA2](#) or [rat 5:62797383-63627669](#) or [coronary heart disease](#)

**Browse a Genome**  
The Ensembl project produces genome databases for vertebrates and other eukaryotic species, and makes this information freely available online.

**Popular genomes**

 <b>Human</b> GRCh38.p3	 <b>Mouse</b> GRCm38.p4
 <b>Zebrafish</b> GRCz10	

★ [Log in to customize this list](#)

**All genomes**  
-- Select a species --

[View full list of all Ensembl species](#)

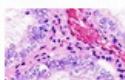
Other species are available in [Ensembl Pre!](#) and [EnsemblGenomes](#).

**Did you know...?**  
"id": "I Java, Perl, Python and Ruby, oh my! Try our [REST API](#) for quick access to Ensembl data.  
"seq": "LPSSLSVI"

Ensembl supports data from external projects through [Track hubs](#) 

**ENCODE data in Ensembl**  

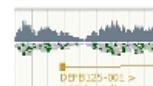

**Variant Effect Predictor**  


**Gene expression in different tissues**  


**Find SNPs and other variants for my gene**  


**Retrieve gene sequence**  
GCCTGACTTCGGGTGG  
GGGCTTGCGCGAGC  
GGGCTCTGCTCGCCCT  
AGGGGACAGATTGGGA  
CACCTCTGAAACCGCTT  
CCGAGTCAAGCGGGCG

**Compare genes across species**  


**Use my own data in Ensembl**  


**Learn about a disease or phenotype**  


• • • • • ► II

# Annotation software

---

- Two sets of software
  - Annovar
    - provides a wide range of annotations that can be applied with one tool
    - we have experienced some inconsistencies in the results e.g. non-synonymous SNPs without polyphen score
  - SNPEff and dbNSFP (Non-Synonymous Functional Prediction)
- Both tested by GATK team
  - recommended snpEff, but with strict requirements
  - **snpEff version 2.0.5** (not 2.0.5d)
  - db should be **GRCh37.64** (which is the **Ensembl database version 64**)
  - should use the option **-onlyCoding true** (using false can cause erroneous annotation)
- GATKs VariantAnnotator to pick the highest impact.
- Finally, also annotate with **dbNSFP, which contains:**
  - variant frequencies
  - conservation scores
  - protein function effect (Polyphen, SIFT)

# snpEff annotation get placed into INFO field

31942920 . G T 683.93 PASS

AC=1;AF=0.50;AN=2;BaseQRankSum=4.358;DP=73;DS;Dels=0.00;FS=0.000;HRun=0;  
HaplotypeScore=1.7876;MQ=69.76;MQ0=0;MQRankSum=0.977;QD=9.37;ReadPosR  
ankSum=0.508; VQSLOD=1.6292;culprit=QD

**SNPEFF\_AMINO\_ACID\_CHANGE=E114\*;**  
**SNPEFF\_CODON\_CHANGE=Gag/Tag;**  
**SNPEFF\_EFFECT=STOP\_GAINED;**  
**SNPEFF\_EXON\_ID=exon\_22\_31942847\_31942957;**  
**SNPEFF\_FUNCTIONAL\_CLASS=NONSENSE;**  
**SNPEFF\_GENE\_BIOTYPE=processed\_transcript;**  
**SNPEFF\_GENE\_NAME=SFI1;**  
**SNPEFF\_IMPACT=HIGH;**  
**SNPEFF\_TRANSCRIPT\_ID=ENST00000421060;**

GT:AD:DP:GQ:PL 0/1:42,31:73:99:714,0,981

# Explanation of snpEff fields (bold are important)

Field name	Example	Description
SNPEFF_EFFECT	NON_SYNONYMOUS_CODING	The highest-impact effect resulting from the current variant (or one of the highest-impact effects, if there is a tie)
SNPEFF_IMPACT	MODERATE	Impact of the highest-impact effect resulting from the current variant ( <b>HIGH</b> , MODERATE, LOW, or MODIFIER)
SNPEFF_FUNCTIONAL_CLASS	MISSENSE	Functional class of the highest-impact effect resulting from the current variant (NONE, SILENT, MISSENSE, or NONSENSE)
SNPEFF_CODON_CHANGE	Tgc/Agc	Old/New codon for the highest-impact effect resulting from the current variant
SNPEFF_AMINO_ACID_CHANGE	C12S	Old/New amino acid for the highest-impact effect resulting from the current variant
SNPEFF_GENE_NAME	SLC6A18	Gene name for the highest-impact effect resulting from the current variant
SNPEFF_GENE_BIOTYPE	protein_coding	Gene biotype for the highest-impact effect resulting from the current variant
SNPEFF_TRANSCRIPT_ID	ENST00000296821	Transcript ID for the highest-impact effect resulting from the current variant
SNPEFF_EXON_ID	exon_5_1225470_1225752	Exon ID for the highest-impact effect resulting from the current variant

## Impact classification of all effects

### High-Impact effects      Moderate-Impact effects      Low-Impact effects

SPlice_Site_Acceptor	NON_SYNONYMOUS_CODING	SYNONYMOUS_START
SPlice_Site_Donor	CODON_CHANGE (note: this effect is not yet fully implemented)	NON_SYNONYMOUS_START
START_LOST	CODON_INSERTION	START_GAINED
EXON_DELETED	CODON_CHANGE_PLUS_CODON	SYNONYMOUS_CODING
FRAME_SHIFT	CODON_DELETION	SYNONYMOUS_STOP
STOP_GAINED	CODON_CHANGE_PLUS_CODON	NON_SYNONYMOUS_STOP
STOP_LOST	UTR_5_DELETED	
	UTR_3_DELETED	

# All vs. top impact

---

- SnpEff uses Ensembl gene models and annotates initially with the effect in **all** different transcripts.
- Multiple functional annotations for the same variant make interpretation difficult
- One can then use GATK to pull out the TOP impact ie the most damaging effect

# A second source of functional annotation: dbNSFP

---

- NSFP = Non-synonymous functional prediction
- Limited to non-synonymous variants
- Has many data fields. We use only:
  - dbnsfpSIFT\_score
  - dbnsfpPolyphen2\_HVAR\_pred
  - dbnsfp29way\_logOdds
  - dbnsfp1000Gp1\_AF

# Example of annotation with dbNSFP

766910 rs1809933 C T 556.42 PASS

AC=1;AF=0.50;AN=2;BaseQRankSum=1.366;DB;DP=30;Dels=0.00;FS=0.000;HRun=0;HaplotypeScore=1.8675;MQ=47.46;MQ0=0;MQRankSum=-0.651;QD=18.55;ReadPosRankSum=-1.757;SB=-109.24;

SNPEFF\_AMINO\_ACID\_CHANGE=R42Q;SNPEFF\_CODON\_CHANGE=cGg/  
cAg;SNPEFF\_EFFECT=NON\_SYNONYMOUS\_CODING;SNPEFF\_EXON\_ID=exon\_5\_766813\_767034;SNPEFF\_FUNCTIONAL\_CLASS=MISSENSE;SNPEFF\_GENE BIOTYPE=processed\_transcript;SNPEFF\_GENE\_NAME=ZDHHC11B;SNPEFF\_IMPACT=MODERATE;SNPEFF\_TRANSCRIPT\_ID=ENST00000382776;

**dbnsfp29way\_logOdds=3.0289;** SiPhy score based on 29 mammals genomes. The larger the score, the more conserved the site.

**dbnsfp1000Gp1\_AF=0.76;** Alt. allele frequency in the whole 1000Gp1 data.

**dbNSFP\_Polyphen2\_HVAR\_pred=B;** Polyphen2 prediction based on HumVar, "D" ("probably damaging"), "P" ("possibly damaging") and "B" ("benign"). Multiple entries separated by ";".

**dbNSFP\_SIFT\_score=0.560000;** SIFT score, If a score is smaller than 0.05 the corresponding NS is predicted as "D(amaging)"; otherwise it is predicted as "T(olerated)". SIFT predicts whether an amino acid substitution affects protein function.

GT:AD:DP:GQ:PL 0/1:5,25:30:98:586,0,98

# Some human variant databases

## ExAC Browser (Beta) | Exome Aggregation Consortium

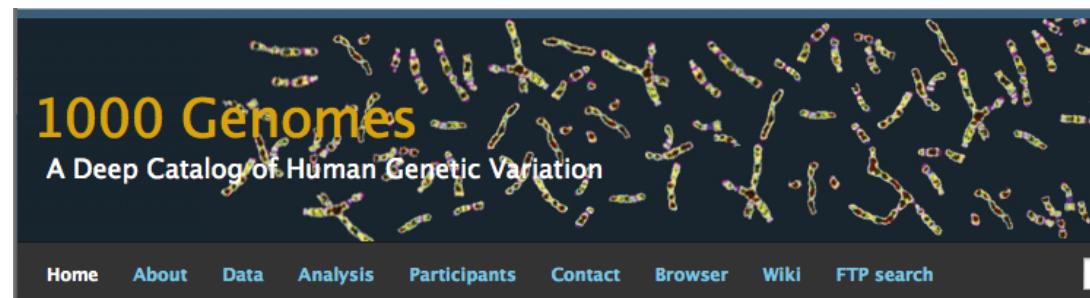
Search for a gene or variant or region

Examples - Gene: [PCSK9](#), Transcript: [ENST00000407236](#), Variant: [22-46615880-T-C](#), Multi-allelic variant: [rs1800234](#), Region: [22:46615715-46615880](#)

### About ExAC

The [Exome Aggregation Consortium](#) (ExAC) is a coalition of investigators seeking to aggregate and harmonize exome sequencing data from a wide variety of large-scale sequencing projects, and to make summary data available for the wider scientific community.

The data set provided on this website spans [60,706 unrelated individuals](#) sequenced as part of various disease-specific and population genetic studies. The ExAC Principal Investigators and groups that have contributed data to the current release are listed [here](#).



# What are the most useful fields

---

- The tradeoff
  - Use very strong filters and get very short lists but risk that the causal variant has been excluded
  - Use weaker filters to keep the causal variant “in” but risk getting very long lists
- A typical strategy is to start with very strong filters
  - See whether you can identify a good candidate
  - If not loosen some of the filters
- Strong filters are for example:
  - SNPEFF IMPACT: HIGH or HIGH and MEDIUM
  - Low frequency in 1000G or in Exac
  - Polyphen damaging

---

**HAVING A GO ON YOUR OWN  
WITH THE REAL DATASET**

# A full pipeline from fastq to annotated VCF

---

- 040\_functionalAnnotation.bash
- This is a recap of everything we have done in the course
- This is your last chance to ask questions.
- Do not copy and paste commands that you do not understand

# Hunt for the causal mutant

---

- Maybe the most fun part of the course
- You will find the instructions in:  
`041_findCausalVariantExercise.txt`

---

**MULTIPLE SAMPLES**  
**FROM HERE ON IT IS JUST USEFUL INFO**

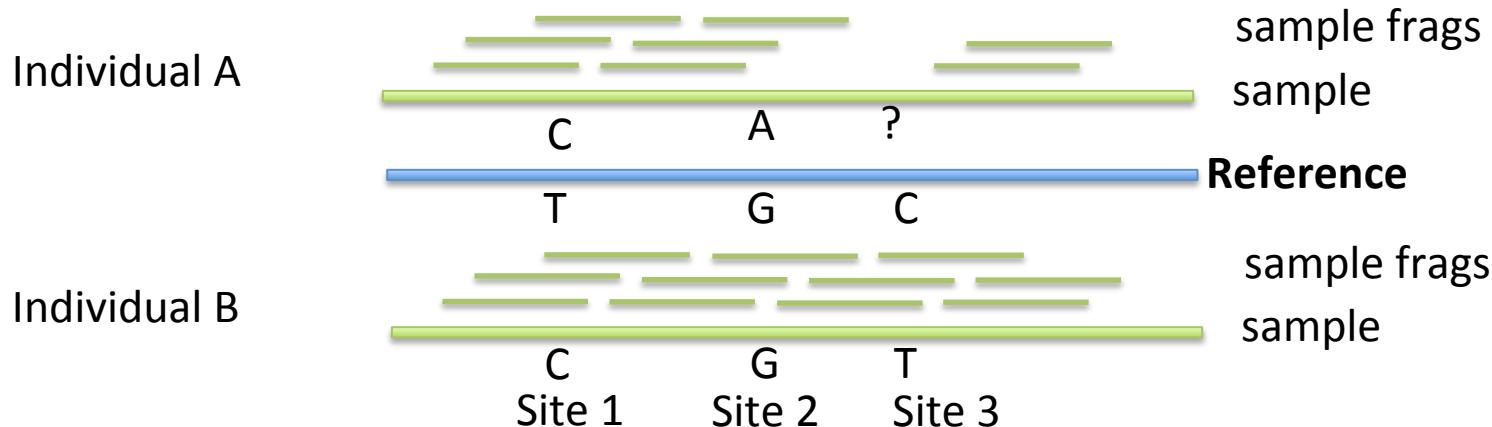
# Batch calling / Joint calling

---

- Better specificity through filtering
  - more variant sites to train on in soft filtering
  - increased power of tests like FS
- Better sensitivity for regions of low coverage (Need to graphically illustrate)
- **Squaring up**
  - genotyping all samples at all sites where at least one of the samples is variant (see next slide)

# The squaring off problem

**Variant files only contain variant sites!**



**Merging singles**

	Indiv. A	Indiv. B
Site 1	Variant	Variant
Site 2	Variant	
Site 3		Variant

**Joint calling**

	Indiv. A	Indiv. B
Site 1	Variant	Variant
Site 2	Variant	Ref
Site 3	No call	Variant

**Both sites lack information  
but we have information on  
individual B at site 2 (REF G)**

**ALL individuals characterised  
at ALL sites where at least  
ONE individual is variant**

# Additional important applications and methods

---

- Somatic variant calling (e.g. cancer cells)
- Pooled sequencing and allele count estimation
- Trios and denovo variants
- Bioplanet website for direct comparison of different methods (different read technologies, different mappers, different SAM processing, different Callers, different variant filtering)
- Pipelines like nextgen bcbio
- Structural variation programs like XHMM and BreakDancer

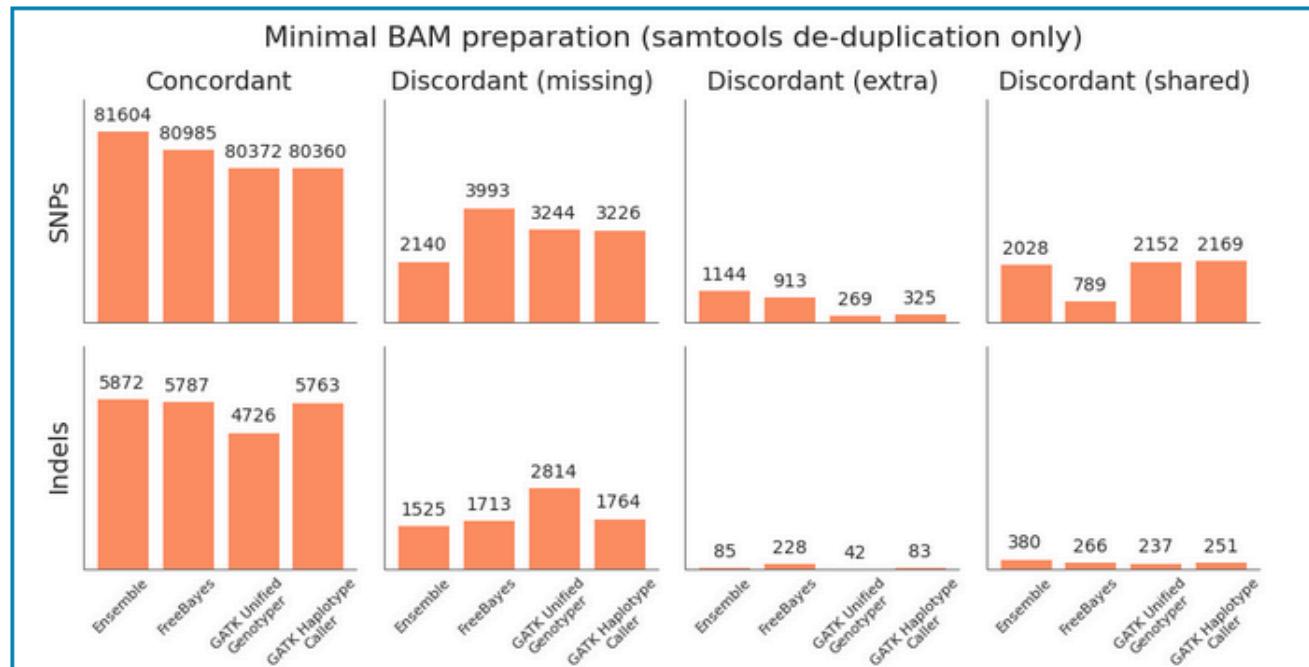
---

# **RECENT ADVANCES**

# HaplotypeCaller and FreeBayes

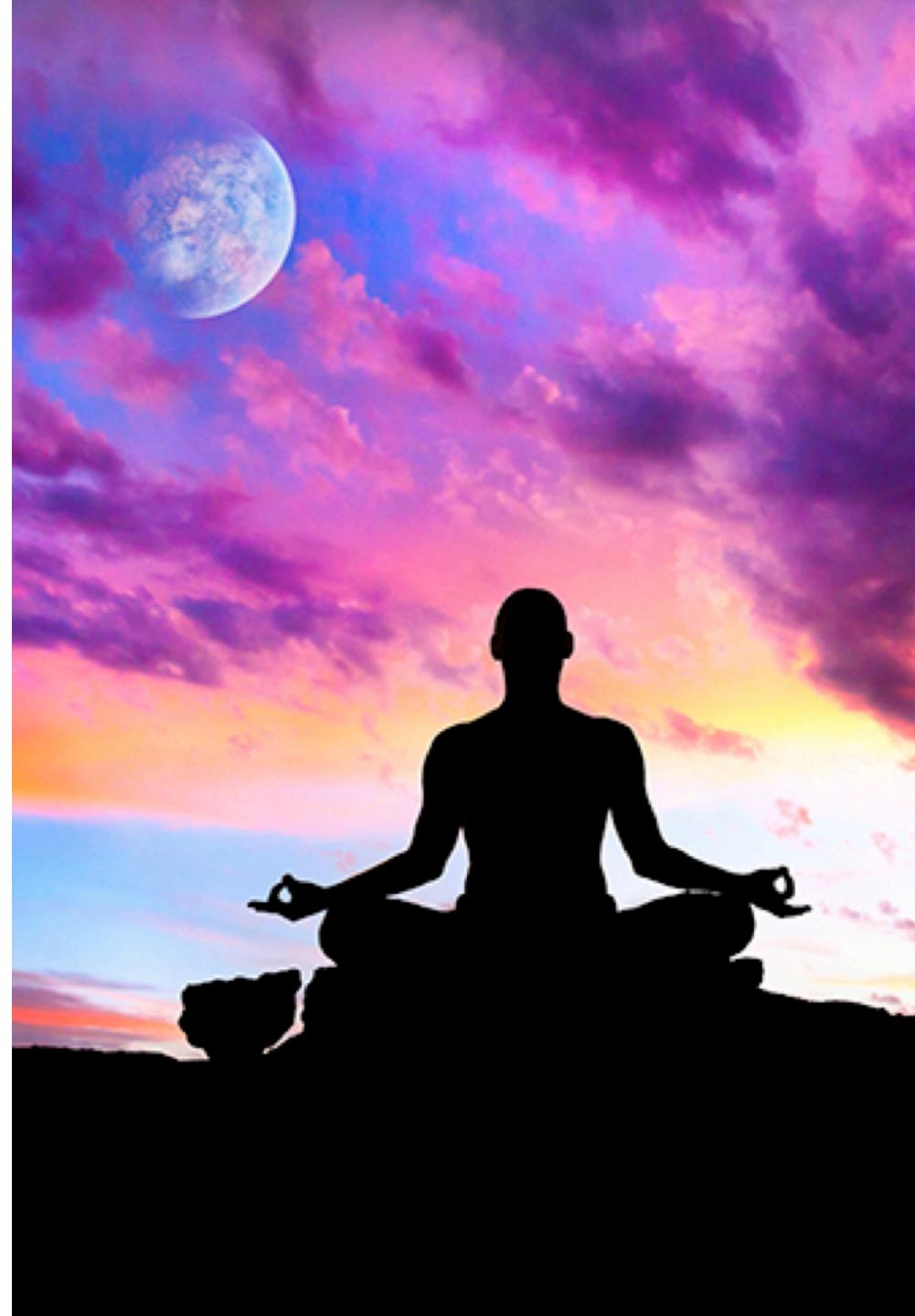
- HaplotypeCaller (in GATK) is an alternative to UnifiedGenotyper which is better at detecting insertions and deletions
  - performs local assembly
- Possibility of dropping the realignment and the recalibration
- FreeBayes is another similar variant caller that does well without realignment and base quality score recalibration.

From bcbio.wordpress.com (Brad Chapman)



---

# **CONCLUDING REMARKS**



---

Summary in plenum of the whole process and all the key concepts



Good luck with the other mountains!!