VCF FILE - BASICS

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                              FORMAT      NA00001        NA00
20    14370   rs6054257 G   A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51
20    17330   .         T   A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50
20    1110696 rs6040355 A   G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27
20    1230237 .         T   .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60
20    1234567 microsat1 GTCT G,GTACT 50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4
```

Meta data: definitions of tags used elsewhere in data lines

Header line

Data lines

Variant columns

Genotype columns

Source: The SAM format specification

Columns of data lines

- **CHROMO:** chromosome / contig
- **POS:** the reference position with the 1st 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 sepearated 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

Source: The SAM format specification

REF and ALT

Reference a t C g a >> C is reference base

										REF	**ALT**
Variant	a	t	**G**	g	a	>> C is a G	→	20	3	.	C	G
Variant	a	t	**–**	g	a	>> C is deleted	→	20	**2**	.	TC	T
Variant	a	t	C**a**g	a		>> A is inserted	→	20	3	.	C	CA

Source: The SAM format specification

REF and ALT

Reference a t C g a >> C is reference base

					REF	ALT
Variant	a t **G** g a >> C is a G	→	20	3	.	C	G
Variant	a t **–** g a >> C is deleted	→	**20**	**2**	.	TC	T

					REF	ALT
To represent both in the same record	→	**20**	**2**	.	TC	**T,TG**

Comparing VCF files: normalisation

- The VCF format is quite precise but still leaves room for representing one variant in multiple ways
  - VCF files need normalisation before comparison

- **Parsimony**
  - Pos: 5, Ref: ATC, Alt: AT
  - **Or** Pos: 6, Ref: TC, Alt: T >> most parsimonious
- **Left alignment**, suppose context: pos 8, ref: ATTTT, T deletion
  - Pos: 10, Ref: TT, Alt: T
  - **Or** Pos: 8, Ref: AT, Alt: A >> left aligned
- **MNP on separate lines**
  - 150 TCT CCC
  - Can be decomposed into two records: 150 T C AND 152 T C

- **One should also ensure that the same reference naming is used in both comparison files and that both files have the same sort order**

- More details at: https://github.com/chapmanb/bcbio.variation/wiki/Normalized-variant-representation and http://genome.sph.umich.edu/wiki/Variant_Normalization

021_generatingReports.bash (part II: VCF file)

- Time to take a look at the VCF part of the practical

Breaking indels into insertions and deletions
Counting SNPs, insertions and deletions
Use bcbio.variation to find concordant and discordant
Visualisation of FNs in IGV (insertions, deletions and SNPs)

Practical 022_workingWithFormats

- introduction to manipulation of:
  - SAM/BAM
  - VCF

RE-ALIGNMENT

Alignment errors during mapping require fix

9	t	ttt
10	a	aaaC
11	a	aaaaa
12	a	aaaaaa
13	a	aaaaaa
14	c	cccTTT
15	a	aaaaaa
16	a	aaaaaa
17	t	AAtttt
18	t	tttttt
19	a	aaaaaa
20	a	aaaaaa
21	g	Tgggg

```
coor      12345678901234     567890123456789012345 6
ref       aggttttataaaac----aattaagtctacagagcaacta
sample    aggttttataaaacAAATaattaagtctacagagcaacta
read1     aggttttataaaac     aaAtaa
read2      ggttttataaaac     aaAtaaTt
read3          ttataaaacAAATaattaagtctaca
read4              CaaaT     aattaagtctacagagcaac
read5                aaT     aattaagtctacagagcaact
read6                  T     aattaagtctacagagcaacta
read1     aggttttataaaacaaataa
read2      ggttttataaaacaaataatt
read3          ttataaaacaaataattaagtctaca
read4                  caaataattaagtctacagagcaac
read5                    aataattaagtctacagagcaact
read6                      taattaagtctacagagcaacta
```

Source: Heng Li, presentation at GSA workshop 2011

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
```

Correct alignment

```
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
```

Possible alignment

```
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
```

Alignment

Gap opening Gap extension

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
```

Wo 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
```

Mismatch Match

- 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
```

Match
Mismatch to ref
Mismatch to read

Few mismatches when considering one-to-one

Example: sample has insertion of AAAT relative to reference

Base stacks

9	t	ttt
10	a	aaaC
11	a	aaaaa
12	a	aaaaaa
13	a	aaaaaa
14	c	cccTTT
15	a	aaaaaa
16	a	aaaaaa
17	t	AAtttt
18	t	tttttt
19	a	aaaaaa
20	a	aaaaaa
21	g	Tgggg

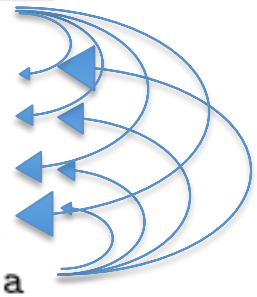
```
coor      12345678901234     567890123456789012345 6
ref       aggttttataaaac----aattaagtctacagagcaacta
sample    aggttttataaaacAAATaattaagtctacagagcaacta
read1     aggttttataaaac     aaAtaa
read2      ggttttataaaac     aaAtaaTt
read3          ttataaaacAAATaattaagtctaca
read4              CaaaT     aattaagtctacagagcaac
read5                aaT     aattaagtctacagagcaact
read6                  T     aattaagtctacagagcaacta
read1     aggttttataaaacaaataa
read2      ggttttataaaacaaataatt
read3          ttataaaacaaataattaagtctaca
read4                   caaataattaagtctacagagcaac
read5                     aataattaagtctacagagcaact
read6                       taattaagtctacagagcaacta
```

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

Base stacks

9	t	ttt
10	a	aaaC
11	a	aaaaa
12	a	aaaaaa
13	a	aaaaaa
14	c	cccTTT
15	a	aaaaaa
16	a	aaaaaa
17	t	AAtttt
18	t	tttttt
19	a	aaaaaa
20	a	aaaaaa
21	g	Tgggg

```
coor      12345678901234     5678901234567890123456
ref       aggttttataaaac----aattaagtctacagagcaacta
sample    aggttttataaaacAAATaattaagtctacagagcaacta
read1     aggttttataaaac     aaAtaa
read2      ggttttataaaac     aaAtaaTt
read3          ttataaaacAAATaattaagtctaca
read4              CaaaT     aattaagtctacagagcaac
read5                aaT     aattaagtctacagagcaact
read6                  T     aattaagtctacagagcaacta
read1     aggttttataaaacaaataa
read2      ggttttataaaacaaataatt
read3          ttataaaacaaataattaagtctaca
read4                   caaataattaagtctacagagcaac
read5                     aataattaagtctacagagcaact
read6                       taattaagtctacagagcaacta
```



No mismatches between reads

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.

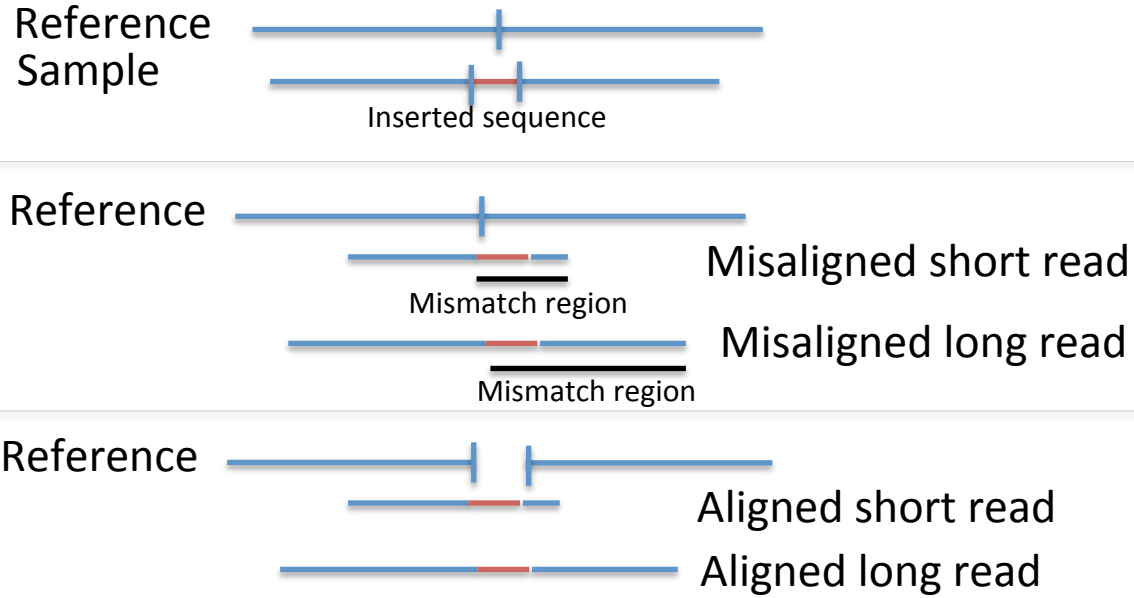
The problem

- 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.

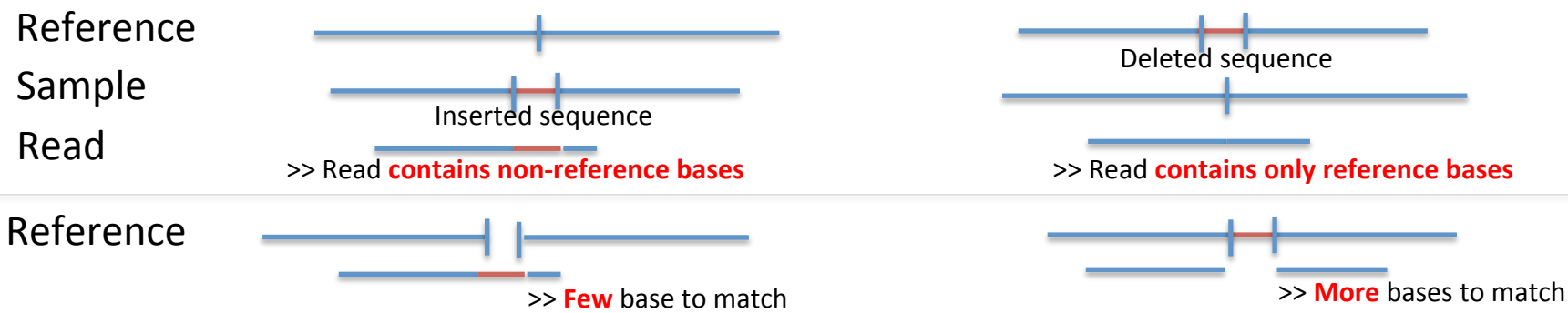
Source: Heng Li, presentation at GSA workshop 2011

Detection of indels

Longer read length facilitates correct alignment



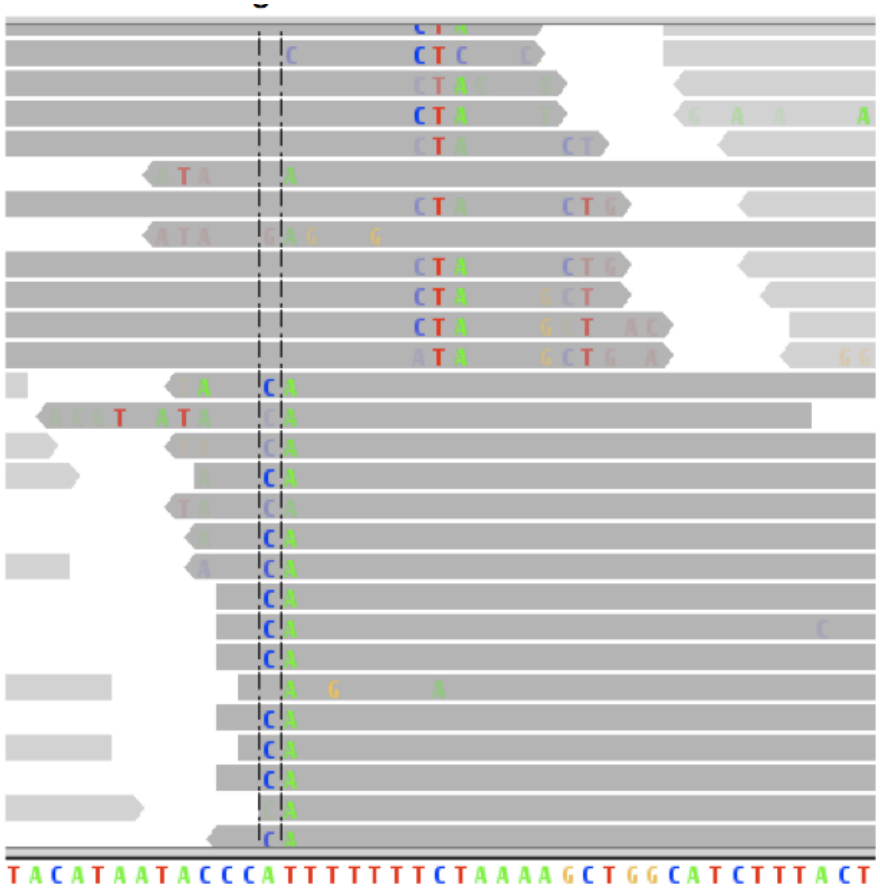
Asymetry between insertions and deletions



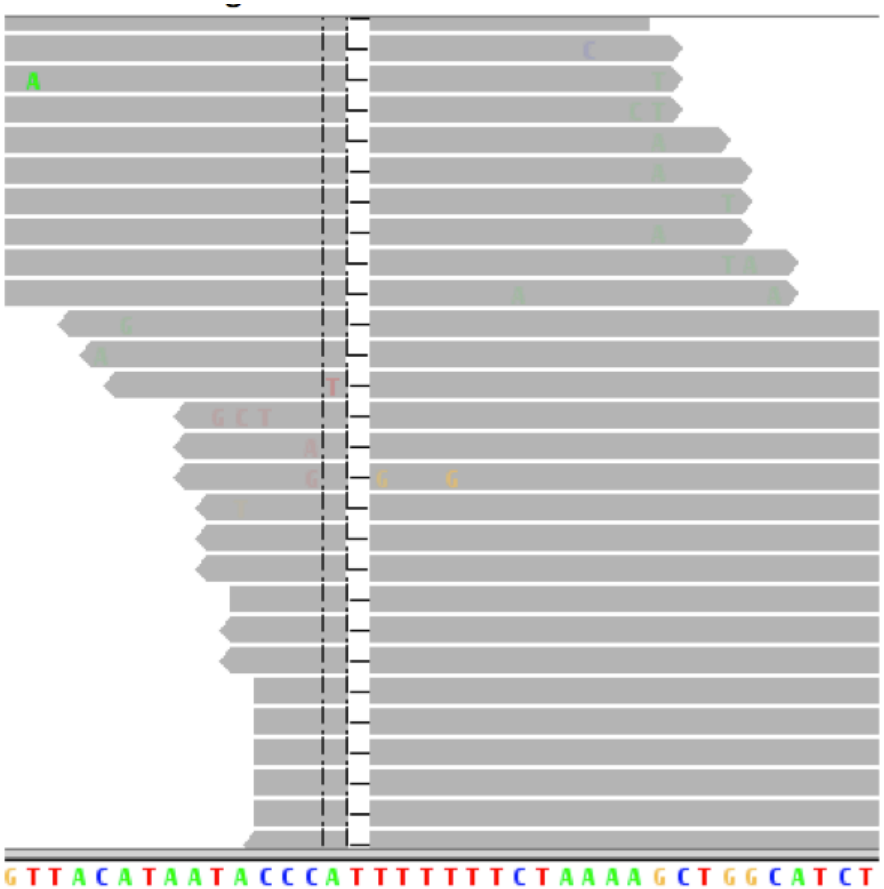
>> insertion and deletion of same size, but more likely to detect the deletion

Local realignment around indels

Before



After

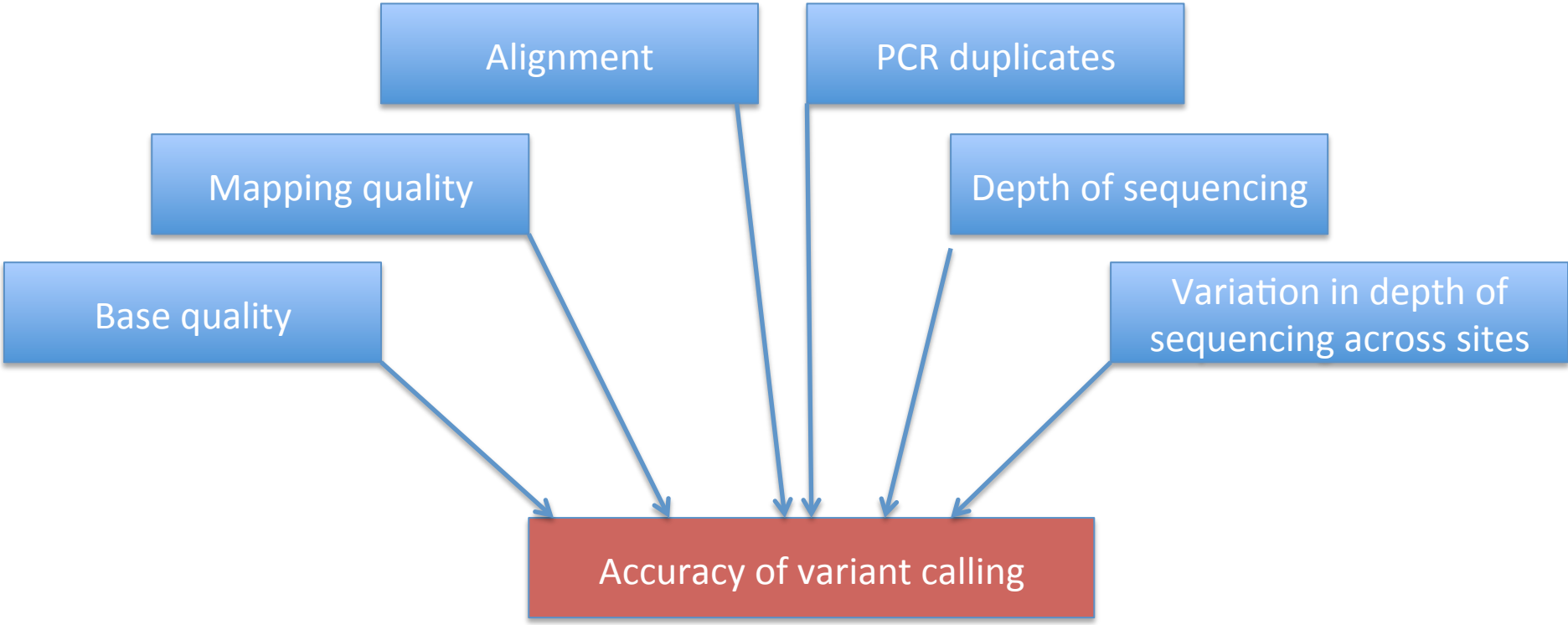


http://www.broadinstitute.org/files/shared/mpg/nextgen2010/nextgen_sivachenko.pdf

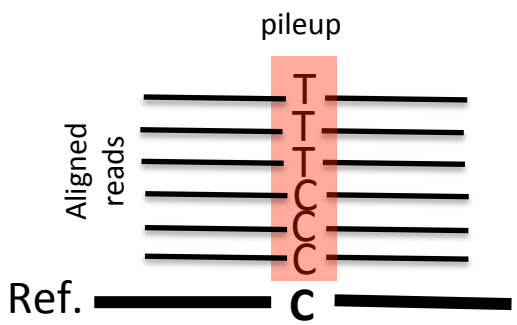
VARIANT CALLING

Recaping 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



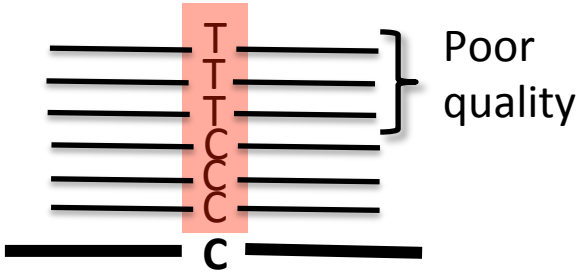
Variant sites (in a diploid genome)



The common and easy case

- Good mapping of reads
- Good base qualities
- Good depth

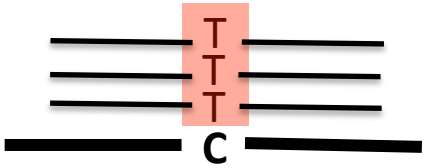
➡ **Genotype: T/C**



Poor quality

- of base calls
- of read mapping

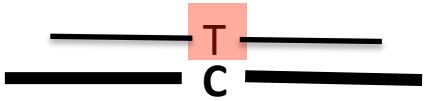
➡ **Genotype: T/C or CC (if Ts are base call errors)**



Poor depth

- May not have sampled both alleles

➡ **Genotype: T/C or T/T**



VERY poor depth

- Did not sample one allele

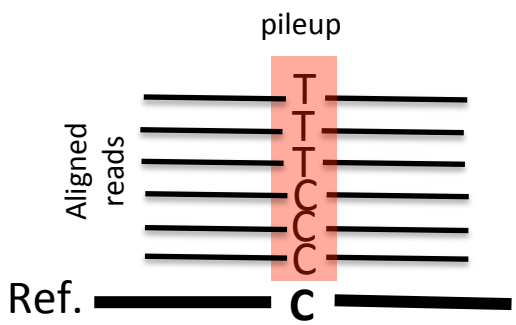
➡ **Genotype: T/T?**

NO depth

- No reads available

➡ **Genotype: ?/?**

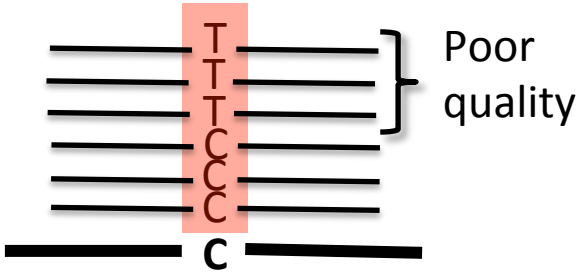
NB: Most sites are not variant



The common and easy case

- Good mapping of reads
- Good base qualities
- Good depth

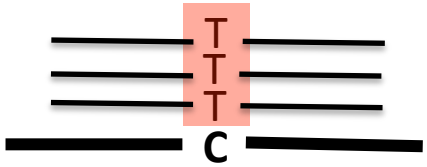
➡ **Genotype: T/C** **No change**



Poor quality

- of base calls
- of read mapping

➡ **Genotype: T/C or CC (if Ts are base call errors)** **Most likely CC**

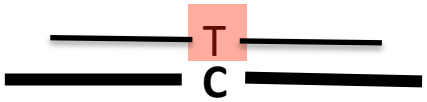


Poor depth

- May not have sampled both alleles

➡ **Genotype: T/C or T/T**

Most likely T/C



VERY poor depth

- Did not sample one allele

➡ **Genotype: T/T?**

Most likely T/C

NO depth

- No reads available

➡ **Genotype: ?/?**

Most likely C/C

Statistical approach must incorporate all available information

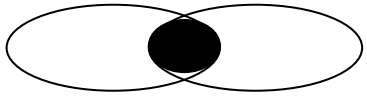
- The importance of prior information
- Combining prior information with new evidence
- We need to take a step back and introduce Bayes theorem.

Bayes rule

P(A/B) = conditional probability, probability of observing event A given that B is true

Definition: P(A/B) = P(A,B) / P(B)

Intuition: the sample space is restricted to B



P(B/A) = P(A,B) / P(A)

Intuition: the sample space is restricted to A

P(A, B) = P(A/B).P(B) = P(B/A) . P(A)

Bayes theorem **P(A/B)** = P(B/A) . P(A) / P(B)

P(B) = P(B,A) + P(B, not A)
P(B) = P(B / A) . P(A) + P(B / not A) . P(not A)

Exercise:

- Disease test is 99% sensitive >> P(test+/D) = 0.99
- Disease test is 99% specific >> P(test+/not D) = 0.01
- Suppose 0.5% of population has disease >> P(D) = 0.005

What is P(D / test+)?

P(D / test+) = P(test+/D) . P(D) / P(test+) **Bayes rule**

P(test+) = P(test+/D).P(D) + P(test+/not D).P(not D) = 0.99(0.005)+0.01(0.995) ≈ 0.005 + 0.01

P(D / test+) ≈ 0.005 / (0.005 + 0.01) ≈ 1/3

So only a 33% chance of having the disease if you test positive. Why is this?

Bayesian statistics

- H: the hypothesis
- D: the data
- Bayes theorem: $P(H_i/D) = P(D/H_i).P(H_i) / P(D)$
  - $P(H_i/D)$ = **the posterior probability** >> typically what a scientist wishes to quantify.
  - P(D/H) = the **likelihood**
  - P(H) = the **prior**
  - P(D) = sum over i[$P(D/H_i).P(H_i)$], thus a constant which is the same independent of the specific posterior $P(H_i/D)$ being calculated

A worked example

- Data (D): reference and sequence read bases
  - reference is C
  - sequenced bases are 4Cs and 2Ts (all with base quality of 30): **read stack is CCCCTT**

- Possible genotypes are CC, TT, and CT
  - These are our different hypotheses of what the genotype may be.

- Priors: P(CC) = 0.9985, P(CT) = 0.001, P(TT) = 0.0005
  - Arbitrarily set based on our existing knowledge of the human genome: most sites are not variant AND this site is a C in reference
  - **A weakness of bayesian statistics?**

- Likelihood of data
  - P(D / CC) = P(two Q30 errors) = $10^{-30/10}$. $10^{-30/10} = 10^{-6}$
  - P(D / TT) = P(four Q30 errors) = $10^{-30*4/10} = 10^{-12}$
  - P(D / CT) = P(sample 4Cs and 2Ts read from two chromosomes) = $15 * (1/2)^6 = 0.234$

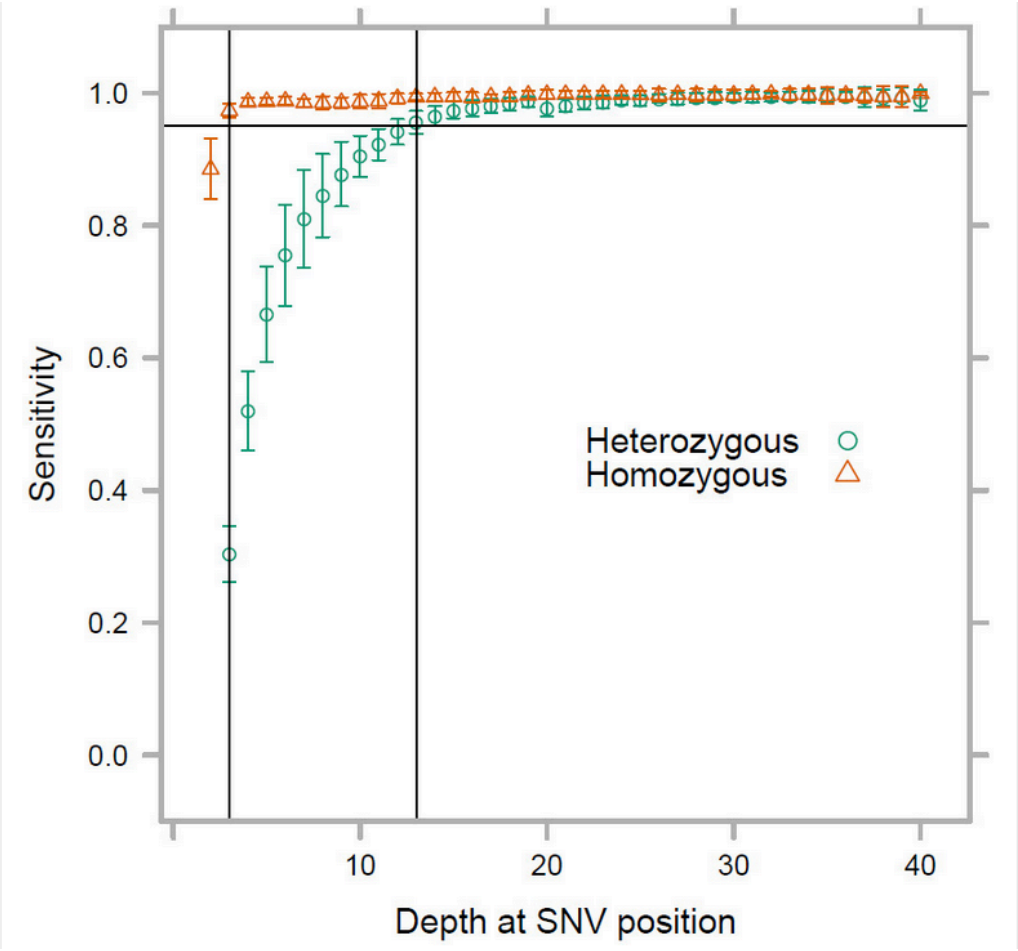
- P(D) = P(D / CC).P(CC) + P(D / CT).P(CT) + P(D / TT).P(D / TT)
  $= 10^{-6}$ (0.9985) + 0.234 (0.001) + 10^{-12} (0.0005) = 0.000235

- Posterior
  - P(CC / D) = P(D / CC) P(CC) / P(D) = 10^{-6} (0.9985) / 0.000235 = 4.242 (10^{-3})
  - P(CT / D) = P(D / CT) P(CT) / P(D) = 0.234 (0.001) / 0.000235 = 9.958 (10^{-1}) **= 0.99**
  - P(TT / D) = P(D / TT) P(TT) / P(D) = 10^{-12} (0.0005) / 0.000235 = 2.124 (10^{-12})

- More complicated model in case of **multiple samples** AND **multi-allelic sites**

The effect of depth on errors

- Heterozygotes vs homozygote variant sites
- Equal sampling of alleles



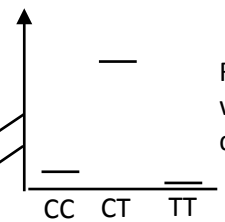
From "Quantifying single nucleotide variant detection sensitivity in exome sequencing"
BMC Bioinformatics 2013

Key parameters in the VCF file

6. QUAL - quality: Phred-scaled quality score for the assertion made in ALT. i.e. $-10log_{10}$ prob(call in ALT is wrong). If ALT is '.' (no variant) then this is $-10log_{10}$ prob(variant), and if ALT is not '.' this is $-10log_{10}$ prob(no variant). If unknown, the missing value should be specified. (Numeric)

P(CC / D) = P(D / CC) P(CC) / P(D) = 10^{-6} (0.9985) / 0.000235 = 4.242 (10^{-3}) << P(non-variant)
P(CT / D) = P(D / CT) P(CT) / P(D) = 0.234 (0.001) / 0.000235 = 9.958 (10^{-1})
P(TT / D) = P(D / TT) P(TT) / P(D) = 10^{-12} (0.0005) / 0.000235 = 2.124 (10^{-12})

Measures quality of the variant site call

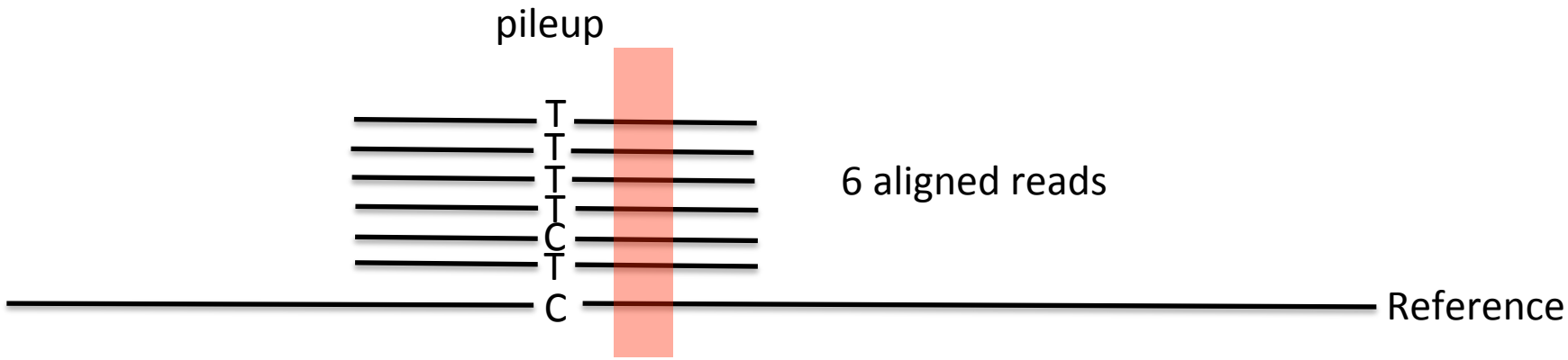


Relative weight of data evidence

- GL : genotype likelihoods comprised of comma separated floating point log_{10}-scaled likelihoods for all possible genotypes given the set of alleles defined in the REF and ALT fields. In presence of the GT field the same ploidy is expected and the canonical order is used; without GT field, diploidy is assumed. If A is the allele in REF and B,C,... are the alleles as ordered in ALT, the ordering of genotypes for the likelihoods is given by: F(j/k) = (k*(k+1)/2)+j. In other words, for biallelic sites the ordering is: AA,AB,BB; for triallelic sites the ordering is: AA,AB,BB,AC,BC,CC, etc. For example: GT:GL 0/1:-323.03,-99.29,-802.53 (Floats)
- GLE : genotype likelihoods of heterogeneous ploidy, used in presence of uncertain copy number. For example: GLE=0:-75.22,1:-223.42,0/0:-323.03,1/0:-99.29,1/1:-802.53 (String)
- PL : the phred-scaled genotype likelihoods rounded to the closest integer (and otherwise defined precisely as the GL field) (Integers)
- GP : the phred-scaled genotype posterior probabilities (and otherwise defined precisely as the GL field); intended to store imputed genotype probabilities (Floats)
- GQ : conditional genotype quality, encoded as a phred quality $-10log_{10}$ p(genotype call is wrong, conditioned on the site's being variant) (Integer)

Measures quality of the genotype call

Intuition of difference between variant and genotype calling



- In the VCF file there is QUAL value and a GQ value

- Intuition of difference between variant site and genotype:
  - **ref is C**, aligned bases are TTTTTC
  - highly likely that the site is variant
  - less clear what the genotype is: T/C or T/T?

- **Only good bases are considered in the pileup**
  - **Minimum base quality**
  - **Read mapping quality**

bayesianStatistic_example.xlsx

- Explore the effect of changing:
  - Priors
  - Qualities
  - Sequence depth

- For those that are curious about Bayesian statistics.

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                              FORMAT      NA00001        NA00
20    14370   rs6054257 G   A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51
20    17330   .         T   A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50
20    1110696 rs6040355 A   G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27
20    1230237 .         T   .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60
20    1234567 microsat1 GTCT G,GTACT 50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4
```

Meta data: definitions of tags used elsewhere in data lines

Header line

Data lines

Variant columns

Genotype columns

Columns of data lines

- **CHROMO**
- **POS**: the reference position with the 1st 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 sepearated 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 1st 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;FS=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		

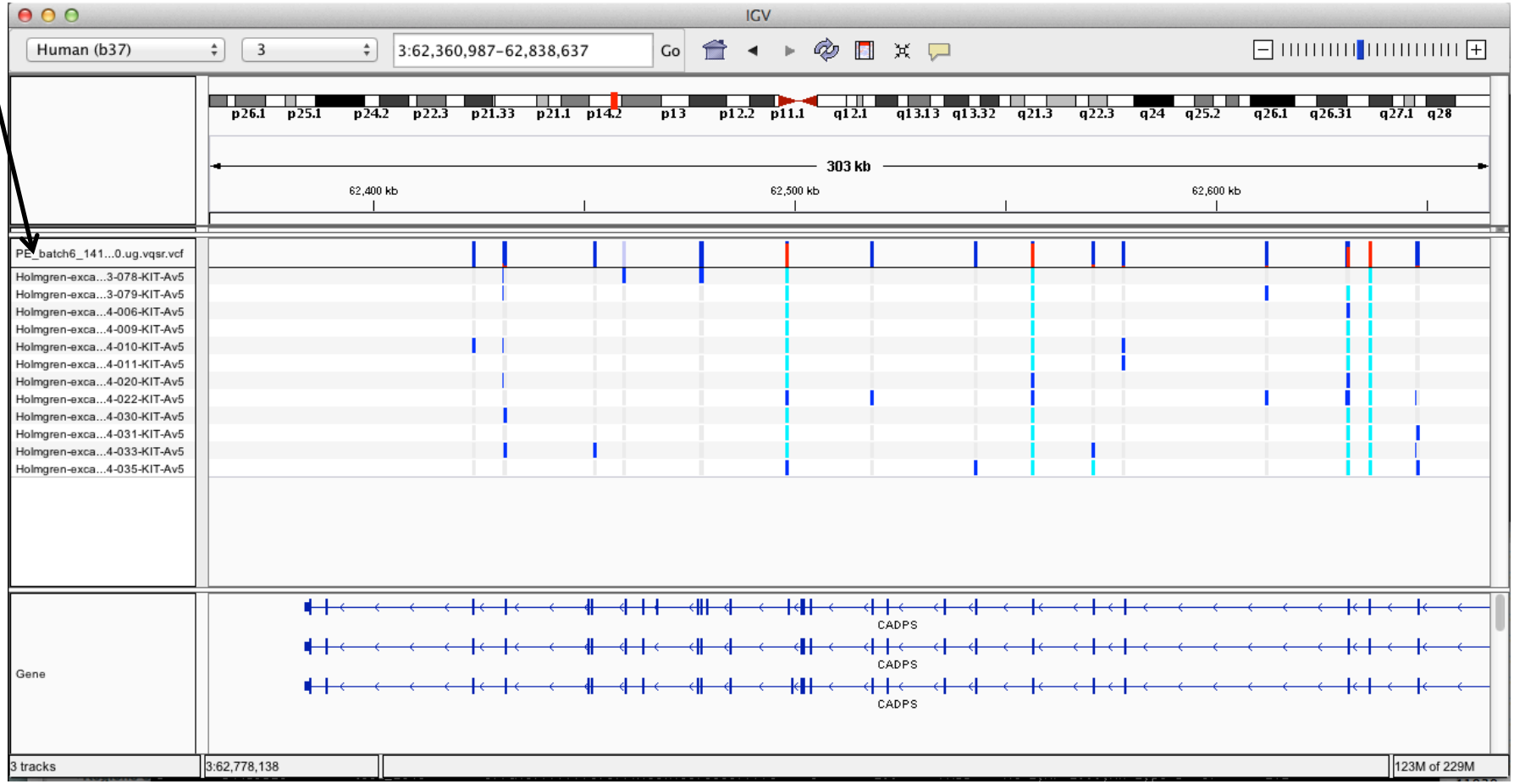
AC=1;AF=0.50;AN=2;BaseQRankSum=-2.638;DP=54;DS;FS=0.000;HRun=0;HaplotypeScore=552.8152;MQ=70.65;MQ0=2;MQRankSum=3.258;QD=66.32;ReadPosRankSum=0.320

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

Multi-sample VCF file

Variant sites

Genotypes



Variant sites

Blue: proportion of ref alleles

Red: proportion of variant alleles

Genotypes (**coloured by genotype**)

Grey: reference

Dark blue: heterozygous variant

Cyan: homozygous variant

Multi-sample VCF file - **Close-up**

Variant sites

Blue: proportion of ref alleles

Red: proportion of variant alleles

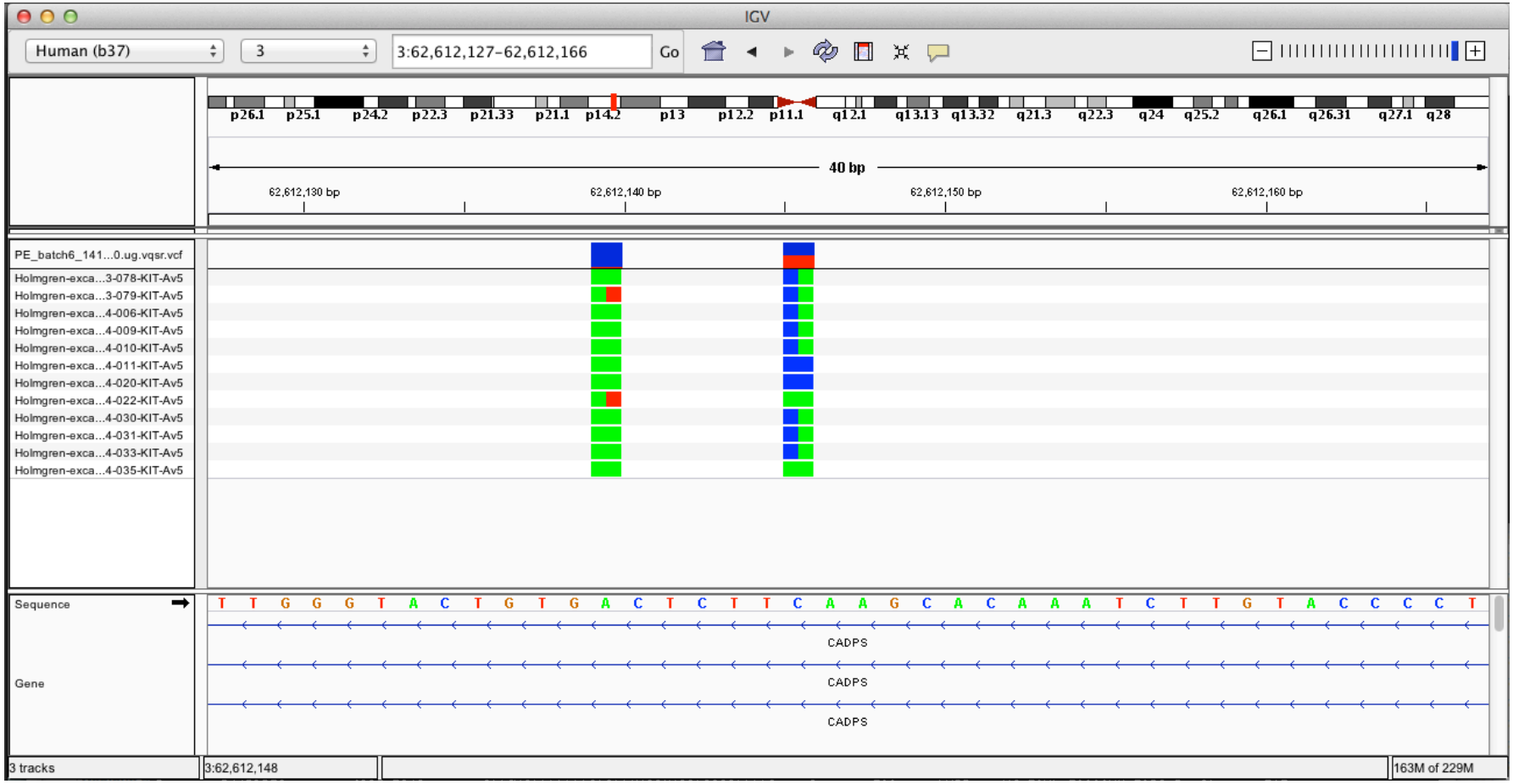
Genotypes (**coloured by genotype**)

Grey: reference

Dark blue: heterozygous variant

Cyan: homozygous variant

Multi-sample VCF file - **Genotypes coloured by allele**



<u>Variant sites</u>

Blue: proportion of ref alleles

Red: proportion of variant alleles

Appendix

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 $P(TT|D) = 3 \times 10^{-11}$

Source: Heng Li, presentation at GSA workshop 2011