Genotype likelhoods

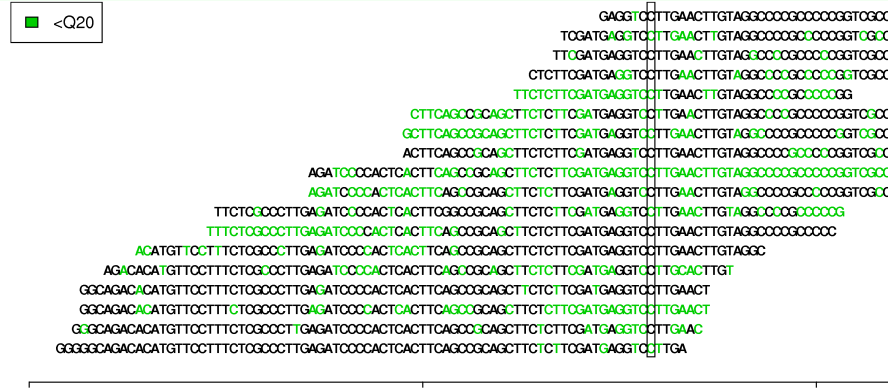
Anders Albrechtsen
The bioinformatic Centre, Copenhagen University

February 13, 2018

Mapped reads

My definitions (The literature is not consistent)

Depth	The number of reads that maps to a position
Counts	The number of different alleles mapped to a position
Coverage	The fraction of the genome (region) with data

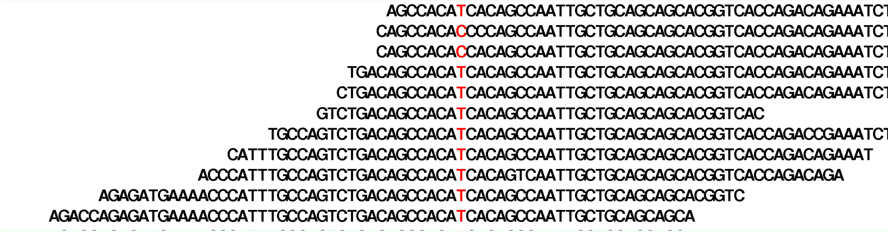

🟩 <Q20
GAGGTCCTTGAACTTGTAGGCCCCGCCCCCGGTCGCC
TCGATGAGGTCCTTGAACTTGTAGGCCCCGCCCCCGGTCGCC
TTCGATGAGGTCCTTGAACTTGTAGGCCCCGCCCCCGGTCGCC
CTCTTCGATGAGGTCCTTGAACTTGTAGGCCCCGCCCCCGGTCGCC
TTCTCTTCGATGAGGTCCTTGAACTTGTAGGCCCCGCCCCCGG
CTTCAGCCGCAGCTTCTCTTCGATGAGGTCCTTGAACTTGTAGGCCCCGCCCCCGGTCGCC
GCTTCAGCCGCAGCTTCTCTTCGATGAGGTCCTTGAACTTGTAGGCCCCGCCCCCGGTCGCC
ACTTCAGCCGCAGCTTCTCTTCGATGAGGTCCTTGAACTTGTAGGCCCCGCCCCCGGTCGCC
AGATCCCCACTCACTTCAGCCGCAGCTTCTCTTCGATGAGGTCCTTGAACTTGTAGGCCCCGCCCCCGGTCGCC
AGATCCCCACTCACTTCAGCCGCAGCTTCTCTTCGATGAGGTCCTTGAACTTGTAGGCCCCGCCCCCGGTCGCC
TTCTCGCCCTTGAGATCCCCACTCACTTCGGCCGCAGCTTCTCTTCGATGAGGTCCTTGAACTTGTAGGCCCCGCCCCCG
TTTCTCGCCCTTGAGATCCCCACTCACTTCAGCCGCAGCTTCTCTTCGATGAGGTCCTTGAACTTGTAGGCCCCGCCCCC
ACATGTTCCTTTCTCGCCCTTGAGATCCCCACTCACTTCAGCCGCAGCTTCTCTTCGATGAGGTCCTTGAACTTGTAGGC
AGACACATGTTCCTTTCTCGCCCTTGAGATCCCCACTCACTTCAGCCGCAGCTTCTCTTCGATGAGGTCCTTGCACTTGT
GGCAGACACATGTTCCTTTCTCGCCCTTGAGATCCCCACTCACTTCAGCCGCAGCTTCTCTTCGATGAGGTCCTTGAACT
GGCAGACACATGTTCCTTTCTCGCCCTTGAGATCCCCACTCACTTCAGCCGCAGCTTCTCTTCGATGAGGTCCTTGAACT
GGGCAGACACATGTTCCTTTCTCGCCCTTGAGATCCCCACTCACTTCAGCCGCAGCTTCTCTTCGATGAGGTCCTTGAAC
GGGGGCAGACACATGTTCCTTTCTCGCCCTTGAGATCCCCACTCACTTCAGCCGCAGCTTCTCTTCGATGAGGTCCTTGA

why don’t we have genotypes?

This is not like Sanger sequencing

Sanger
: Both alleles are amplified and sequenced at the same time.

NGS
: Each allele is sequenced separately and the allele are sampled with replacement



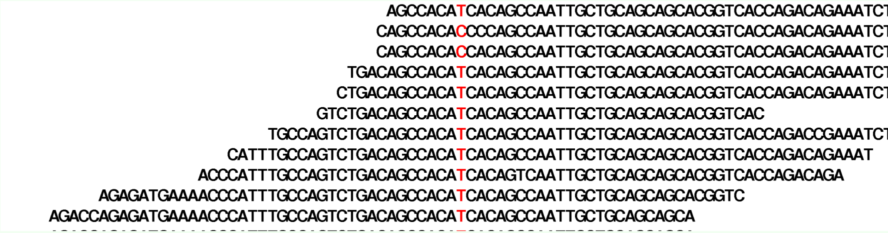
```
                                  AGCCACATCACAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAATCT
                                 CAGCCACACCCCAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAATCT
                                 CAGCCACACCACAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAATCT
                              TGACAGCCACATCACAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAATCT
                             CTGACAGCCACATCACAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAATCT
                           GTCTGACAGCCACATCACAGCCAATTGCTGCAGCAGCACGGTCAC
                      TGCCAGTCTGACAGCCACATCACAGCCAATTGCTGCAGCAGCACGGTCACCAGACCGAAATCT
                  CATTTGCCAGTCTGACAGCCACATCACAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAAT
               ACCCATTTGCCAGTCTGACAGCCACATCACAGTCAATTGCTGCAGCAGCACGGTCACCAGACAGA
     AGAGATGAAAACCCATTTGCCAGTCTGACAGCCACATCACAGCCAATTGCTGCAGCAGCACGGTC
AGACCAGAGATGAAAACCCATTTGCCAGTCTGACAGCCACATCACAGCCAATTGCTGCAGCAGCA
```

why don't we have genotypes?

Question?

Assuming an error rate of 1%

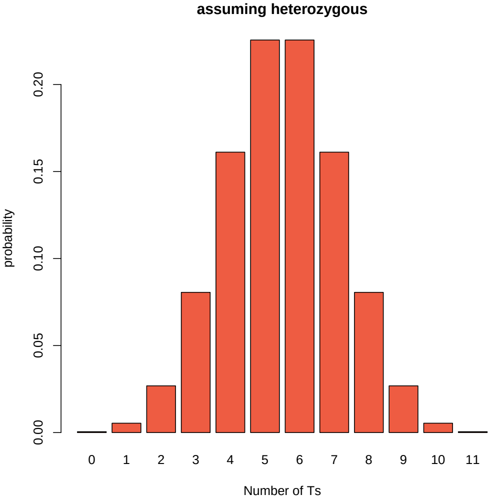
- Is the individual heterozygous C/T?



```
                                  AGCCACATCACAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAATCT
                                 CAGCCACACCCCAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAATCT
                                 CAGCCACACCACAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAATCT
                              TGACAGCCACATCACAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAATCT
                             CTGACAGCCACATCACAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAATCT
                           GTCTGACAGCCACATCACAGCCAATTGCTGCAGCAGCACGGTCAC
                      TGCCAGTCTGACAGCCACATCACAGCCAATTGCTGCAGCAGCACGGTCACCAGACCGAAATCT
                  CATTTGCCAGTCTGACAGCCACATCACAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAAT
               ACCCATTTGCCAGTCTGACAGCCACATCACAGTCAATTGCTGCAGCAGCACGGTCACCAGACAGA
     AGAGATGAAAACCCATTTGCCAGTCTGACAGCCACATCACAGCCAATTGCTGCAGCAGCACGGTC
AGACCAGAGATGAAAACCCATTTGCCAGTCTGACAGCCACATCACAGCCAATTGCTGCAGCAGCA
```

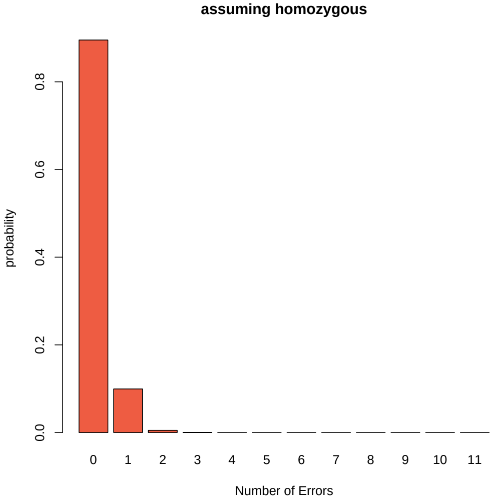
What do we expect

P(2 or less minor bases | heterozygous) = 0.065



What do we expect

$P(2 \text{ or more errors} \mid \text{homozygous}) = 0.00015$

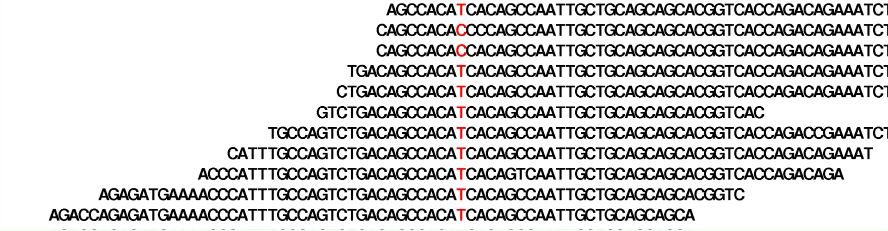


why don't we have genotypes?

Question?

Assuming an error rate of 1%

- Is the individual heterozygous C/T?
- P(2 or more errors | homozygous) = 0.00015
- P(2 or less minor bases | heterozygous) = 0.065



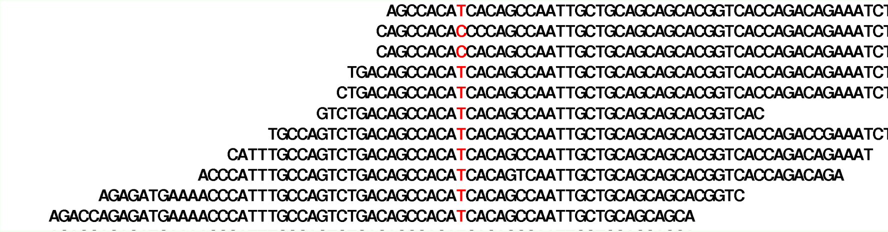
```
                                  AGCCACATCACAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAATCT
                                 CAGCCACACCCCAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAATCT
                                 CAGCCACACCACAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAATCT
                              TGACAGCCACATCACAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAATCT
                             CTGACAGCCACATCACAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAATCT
                           GTCTGACAGCCACATCACAGCCAATTGCTGCAGCAGCACGGTCAC
                      TGCCAGTCTGACAGCCACATCACAGCCAATTGCTGCAGCAGCACGGTCACCAGACCGAAATCT
                  CATTTGCCAGTCTGACAGCCACATCACAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAAT
               ACCCATTTGCCAGTCTGACAGCCACATCACAGTCAATTGCTGCAGCAGCACGGTCACCAGACAGA
     AGAGATGAAAACCCATTTGCCAGTCTGACAGCCACATCACAGCCAATTGCTGCAGCAGCACGGTC
AGACCAGAGATGAAAACCCATTTGCCAGTCTGACAGCCACATCACAGCCAATTGCTGCAGCAGCA
```

why don’t we have genotypes?

Question?

Assuming an error rate of 1%

- Is the individual heterozygous C/T?
- P(2 or more errors | homozygous) = 0.00015
- P(2 or less minor bases | heterozygous) = 0.065
- on average there is about 1 heterozygous site per 1000 bases



```
                                  AGCCACATCACAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAATCT
                                 CAGCCACACCCCAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAATCT
                                 CAGCCACACCACAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAATCT
                              TGACAGCCACATCACAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAATCT
                             CTGACAGCCACATCACAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAATCT
                           GTCTGACAGCCACATCACAGCCAATTGCTGCAGCAGCACGGTCAC
                      TGCCAGTCTGACAGCCACATCACAGCCAATTGCTGCAGCAGCACGGTCACCAGACCGAAATCT
                  CATTTGCCAGTCTGACAGCCACATCACAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAAT
               ACCCATTTGCCAGTCTGACAGCCACATCACAGTCAATTGCTGCAGCAGCACGGTCACCAGACAGA
     AGAGATGAAAACCCATTTGCCAGTCTGACAGCCACATCACAGCCAATTGCTGCAGCAGCACGGTC
AGACCAGAGATGAAAACCCATTTGCCAGTCTGACAGCCACATCACAGCCAATTGCTGCAGCAGCA
```

Genotype likelihoods

Summarise the data in 10 genotype likelihoods

bases (b):
TCCTTTTTTTT
quality scores (Q):
GHSSBBTTTTG

$\rightarrowtail$

	A	C	G	T
A	1	2	3	4
C		5	6	7
G			8	9
T				10

The likelihood

$P(Data|G = \{A_1, A_2\}) \propto P(X|G = \{A_1, A_2\}) = P(X|G)$
where $A \in \{A, C, G, T\}$

Estimating genotype likelihoods

GATK (McKenna et al. 2010)

$$P(X|G) \propto \prod_{i=0}^{n} P(b_i|A_1, A_2) = \prod_{i=0}^{n} \left(\frac{1}{2} P(b_i|A_1) + \frac{1}{2} P(b_i|A_2) \right)$$

where $P(b|A) = \left\{ \begin{array}{cc} \frac{\epsilon}{3} & b \neq A \\ 1 - \epsilon & b = A \end{array} \right.$,

where $G = \{A_1, A_2\}$, b is the observed base and ϵ is the probability of error from the quality score.

Example of genotype likelihood calculations

b	Qasci	Qscore	ϵ	$p(b_i\|T)$	$p(b_i\|C)$	$p(b_i\|G/A)$
T	G	38	0.00016	1 - 0.00016	5.3e-05	5.3e-05
C	H	39	0.00013	4.2e-05	1 - 0.00013	4.2e-05
C	S	50	1e-05	3.3e-06	1 - 1e-05	3.3e-06
T	S	50	1e-05	1 - 1e-05	3.3e-06	3.3e-06
T	B	33	5e-04	1 - 5e-04	0.00017	0.00017
T	B	33	5e-04	1 - 5e-04	0.00017	0.00017
T	T	51	7.9e-06	1 - 7.9e-06	2.6e-06	2.6e-06
T	T	51	7.9e-06	1 - 7.9e-06	2.6e-06	2.6e-06
T	T	51	7.9e-06	1 - 7.9e-06	2.6e-06	2.6e-06
T	T	51	7.9e-06	1 - 7.9e-06	2.6e-06	2.6e-06
T	G	38	0.00016	1 - 0.00016	5.3e-05	5.3e-05

$$P(Data|G = TC) \propto \prod_{i=0}^{n} P(b_i|T,C) = \prod_{i=0}^{n} \left(\frac{1}{2} P(b_i|T) + \frac{1}{2} P(b_i|C) \right)$$

Genotype likelihoods

Other methods

samtools/H. Li et al. 2008 quality scores, quality dependency

soapSNP/R. Li et al. 2009 quality scores, quality dependency

GATK/McKenna et al. 2010 quality scores

Kim et al. 2010? type specific errors

Genotype calling

10 genotype likelihoods

	A	C	G	T
A	0.0	0.001	0.0	0.01
C		0.02	0.001	0.12
G			0.0	0.003
T				0.001

simple genotype callers - Maximum likelihood

ML I Choose the genotype with the largest likelihood $\arg\max_G P(X|G)$

ML II only call a genotype if the likelihood with much better than the second best e.g. a likelihood ratio > 2