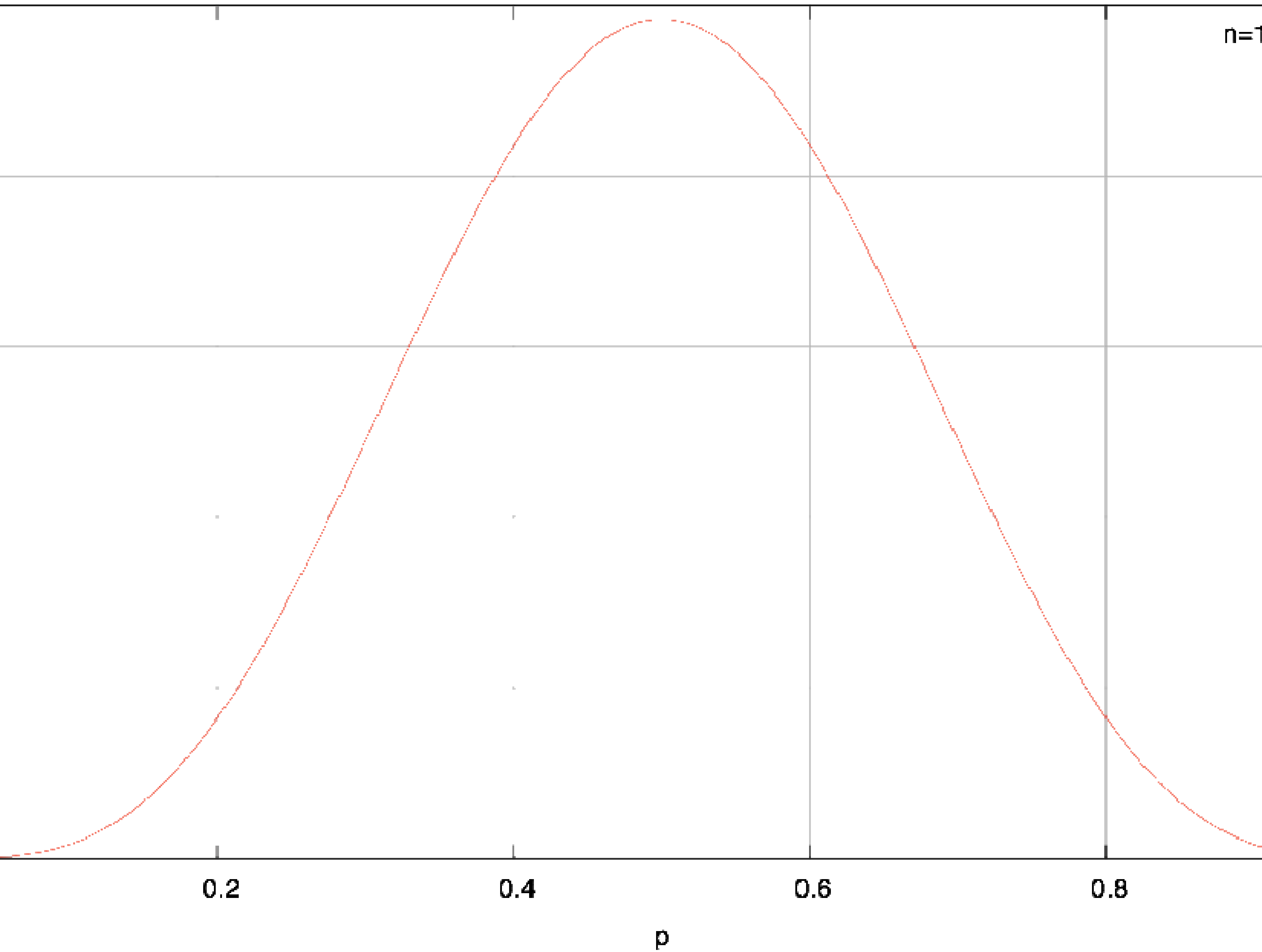
SNP Prediction

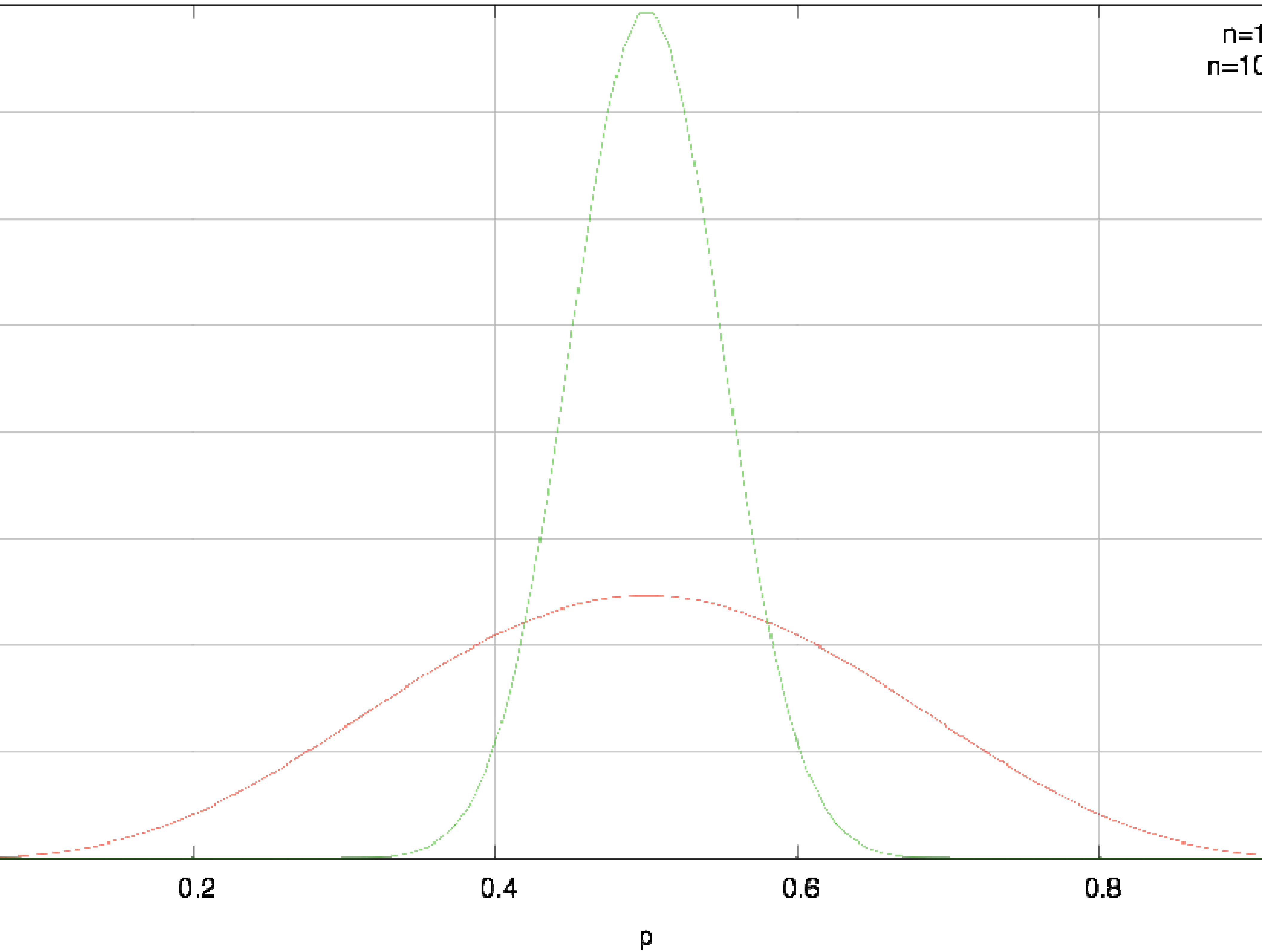
from short read sequence data (with reference sequence available)

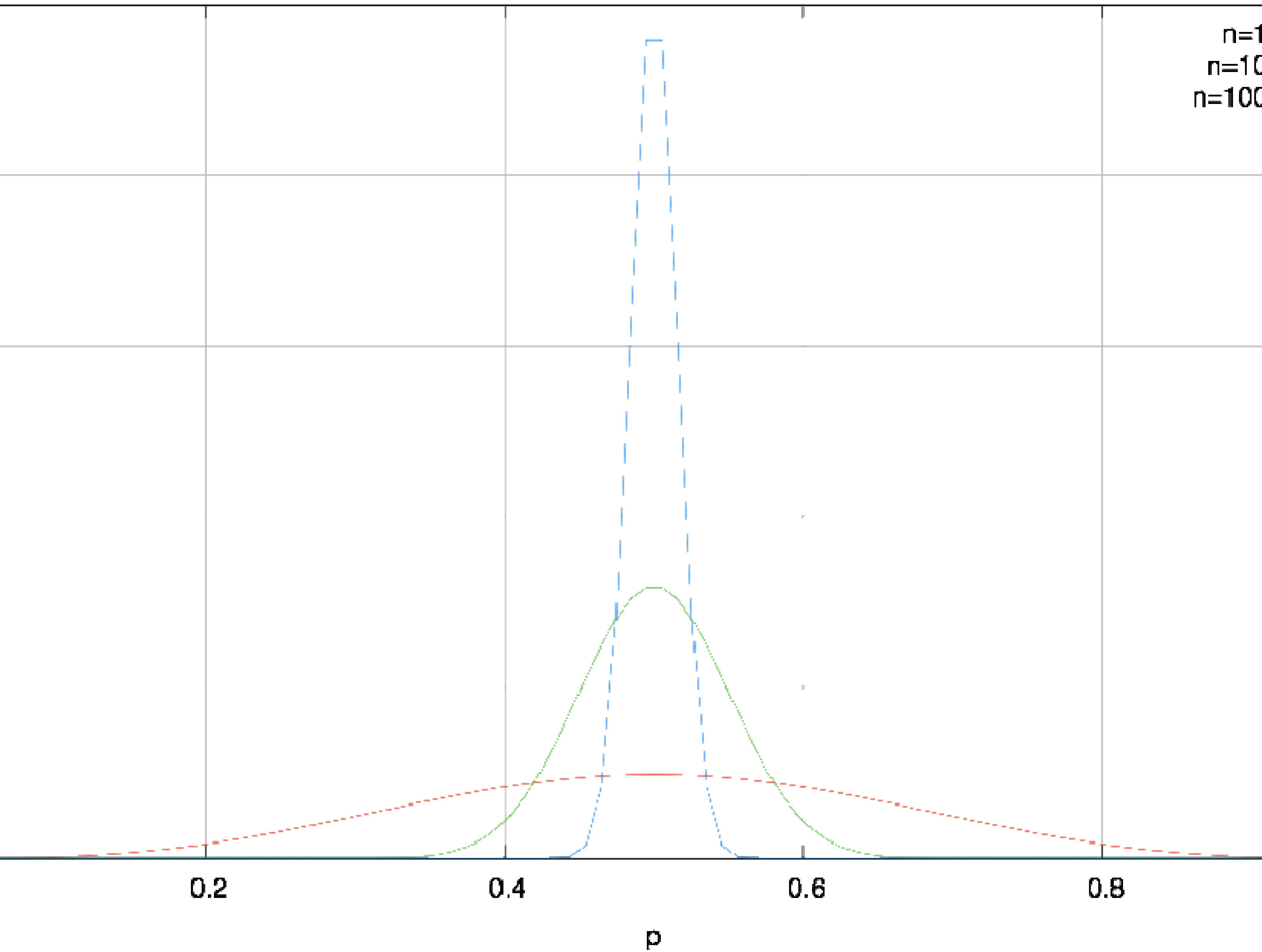
Simon Heath, CNG
May 2009

Principle

- Detect homozygous changes by comparison with the reference sequence
- At heterozygous positions, we should see 2 populations of reads containing the 2 alleles
- The ratio of reads containing the two alternate alleles should be (very) roughly 50:50


n=1
0.2
0.4
0.6
0.8
p

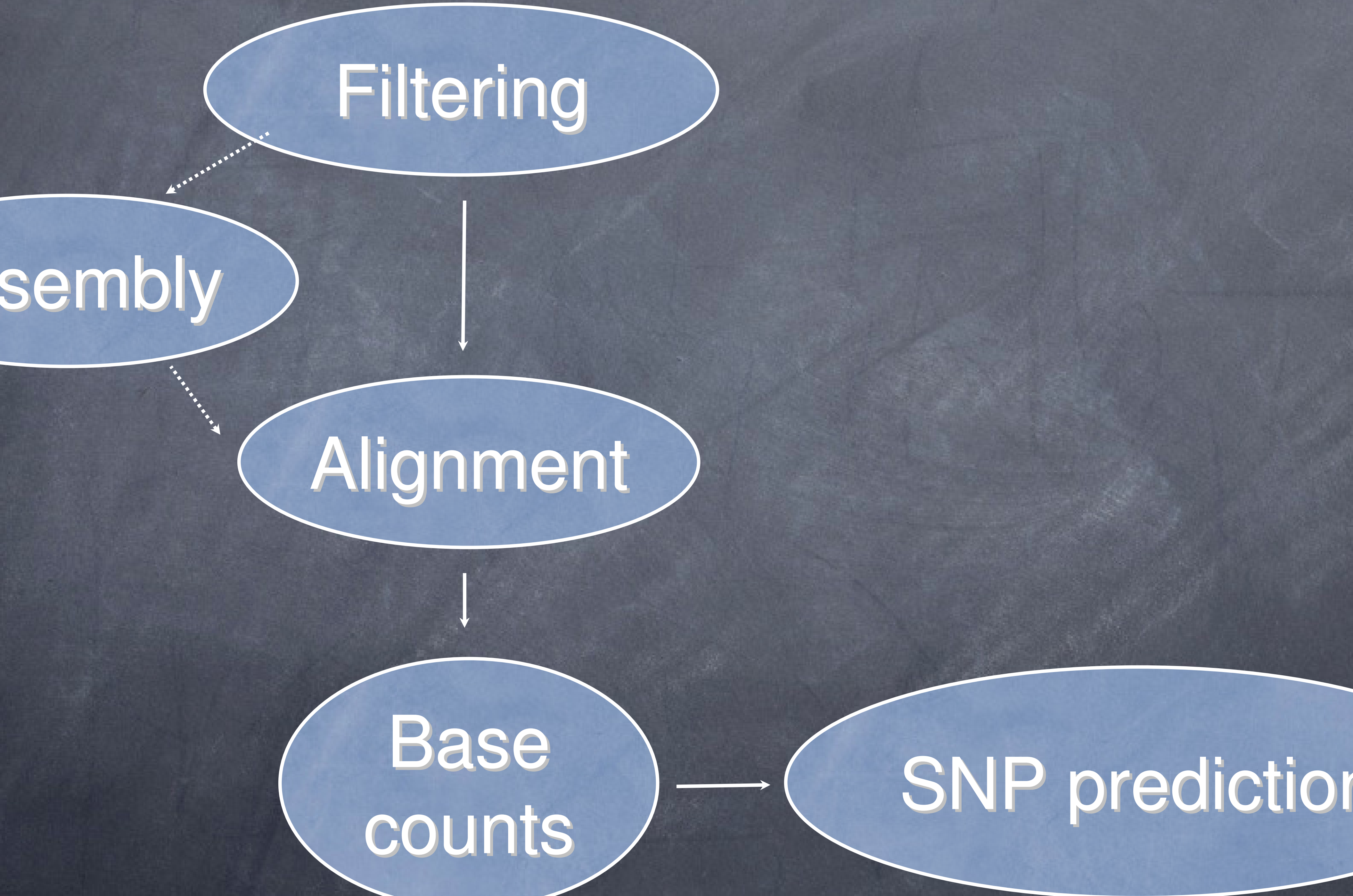

n=1
n=10
0.2
0.4
0.6
0.8
p


n=1
n=10
n=100
0.2
0.4
0.6
0.8
p

Objective of research

- Whole genome sequencing (mainly human)
- Detect SNPs by comparison with reference sequence and from heterozygous sites
- Detect structural variants from information on coverage, allelic ratios and paired end reads

SNP Analysis Pipeline



Filtering issues

- Short read alignments - can not afford to tolerate too many mis-matched bases
- e.g., Eland sets the limit at 2 mis-matches
- A read with a true variant starts with a disadvantage - more likely to be eliminated
- Stringent filtering will throw out many of the reads with true variants

de novo assembly

- Assemble short reads into contigs based on analyzing overlaps prior to alignment
- Reduces alignment errors and dependency on reference sequence
- Only currently practical for re-sequencing of target regions or very small genomes
- Need for this reduced by advances in technology?

Potential biases

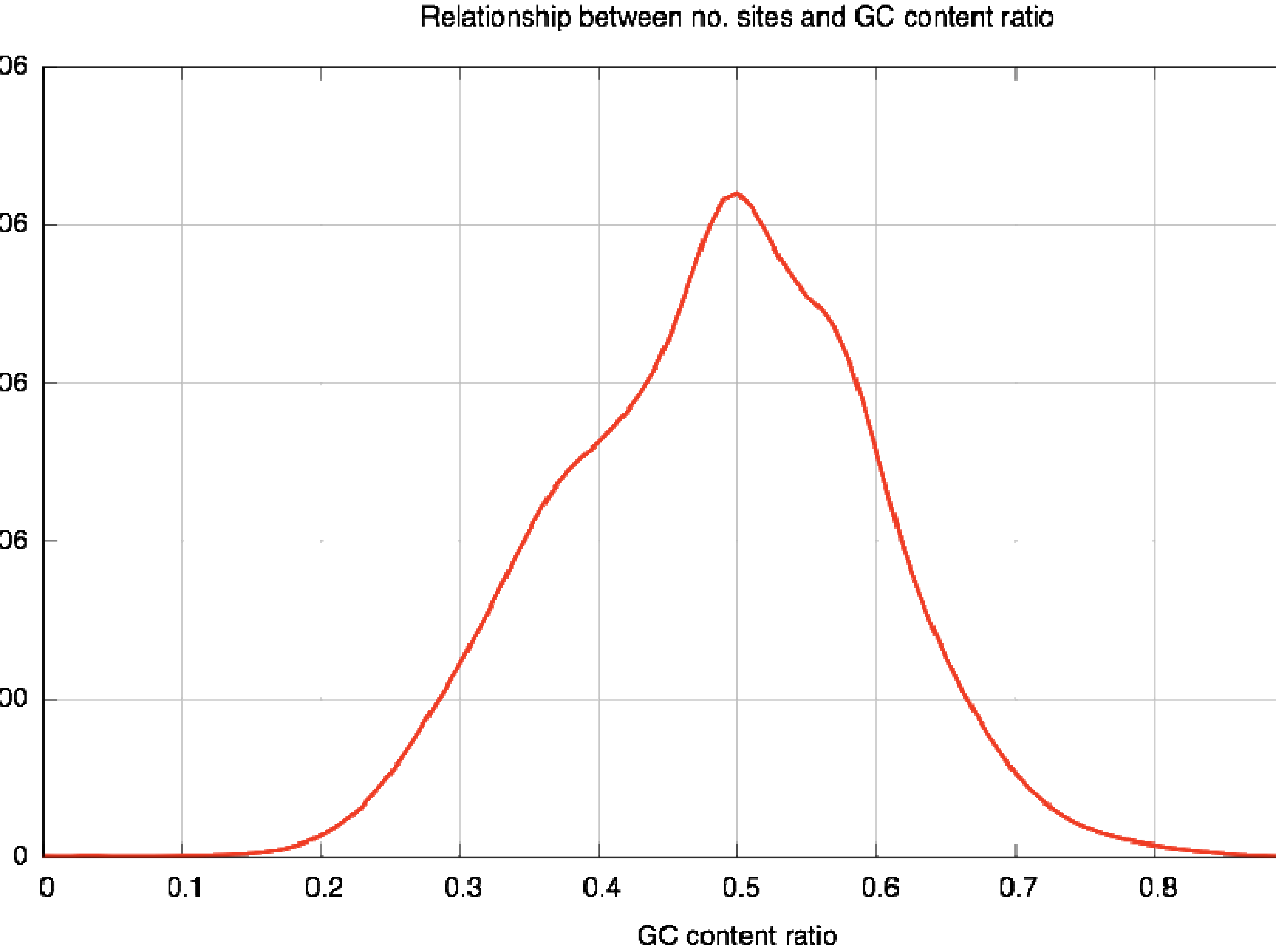
- Biases due to filtering and alignment
  - True differences between reads and reference may be filtered out
- Biases in sequencing technology
  - e.g. Solexa short reads have a biased base distribution for the first 2 bases

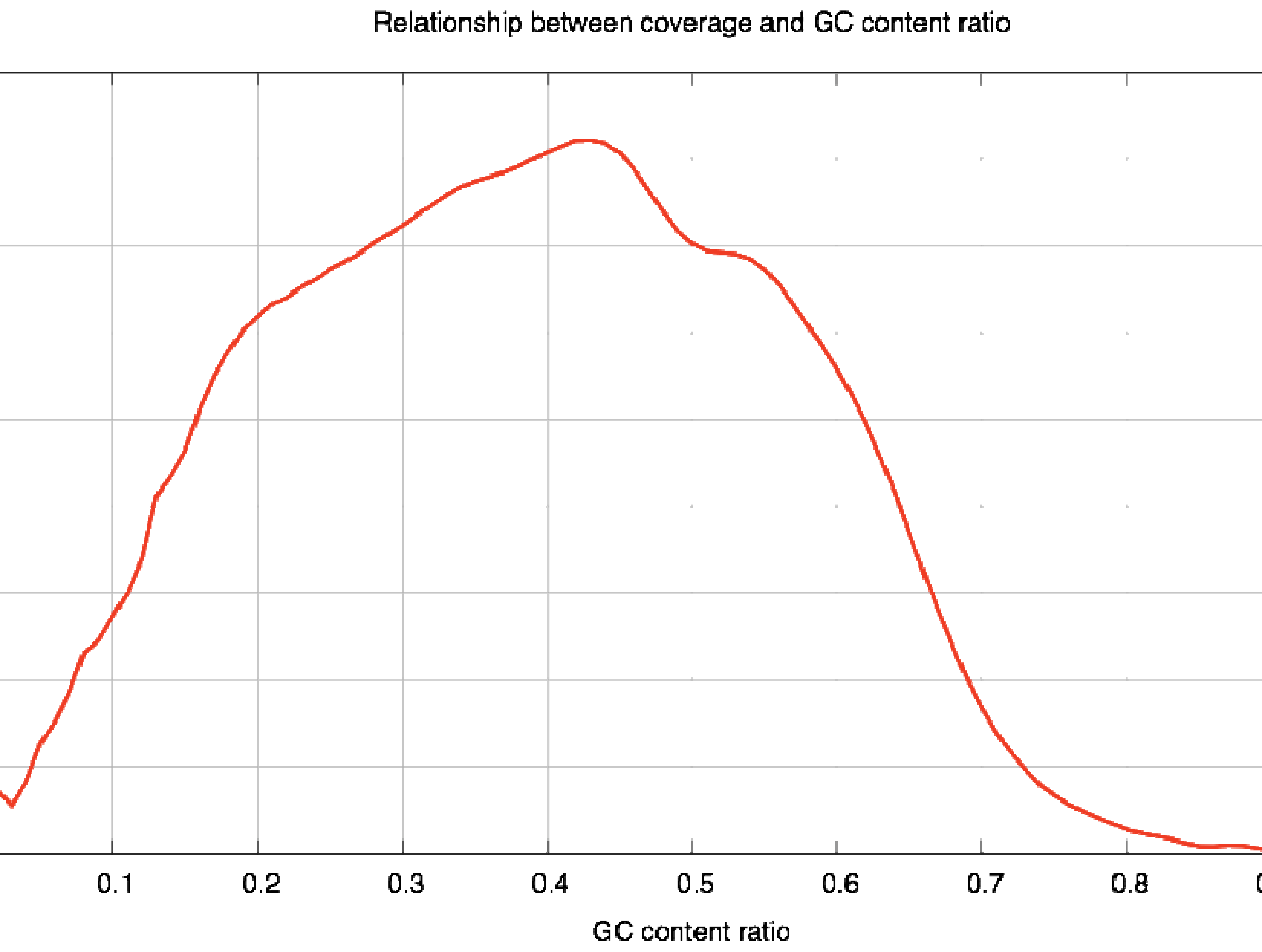
Predicting genotypes

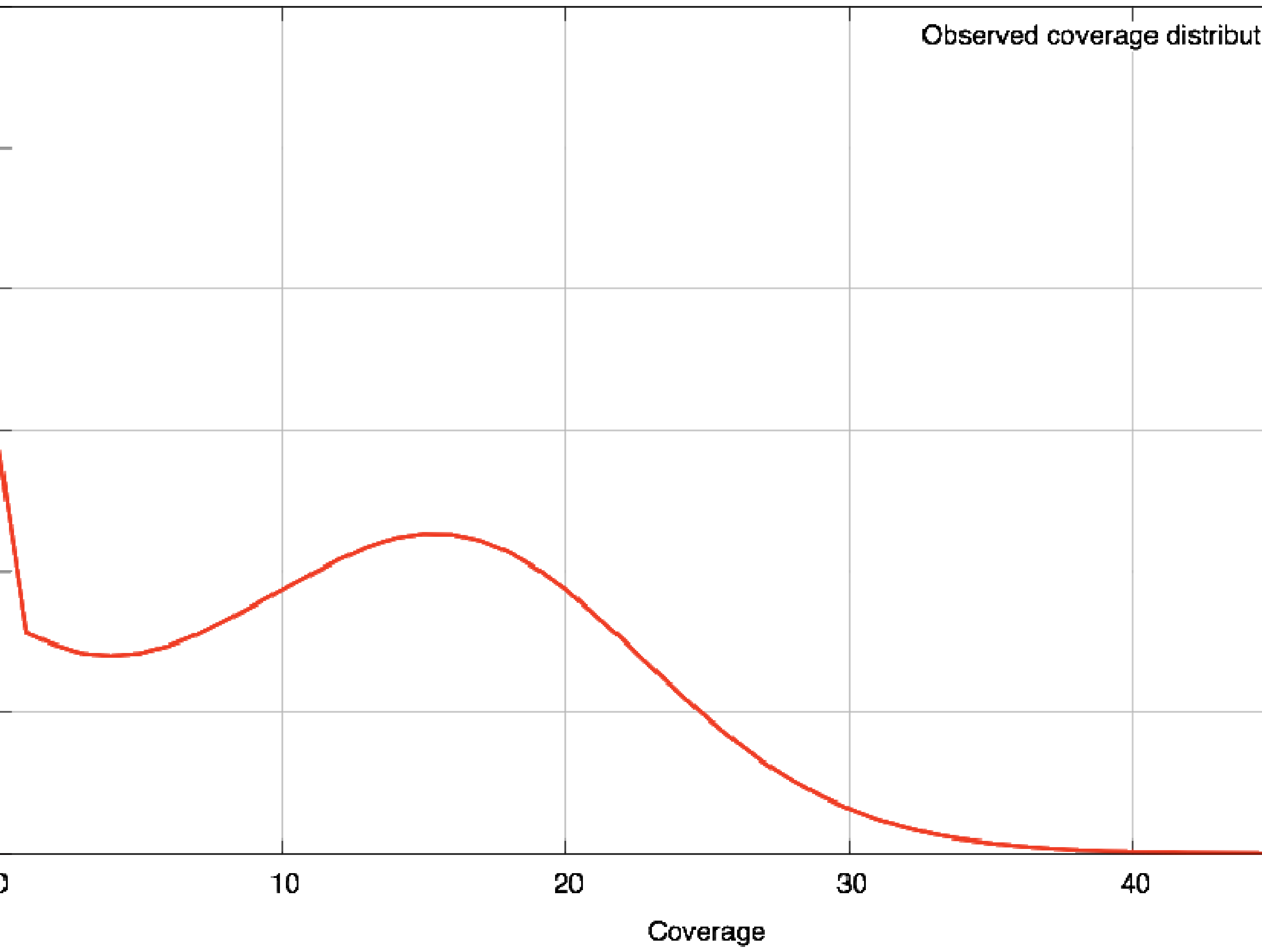
- Observations - short read sequences aligned to the genome
- For each position, get number of different bases seen i.e., A:8 G:6 C:0 T:1
- Simple model - use this information to infer genotype at the position (AG for example above)

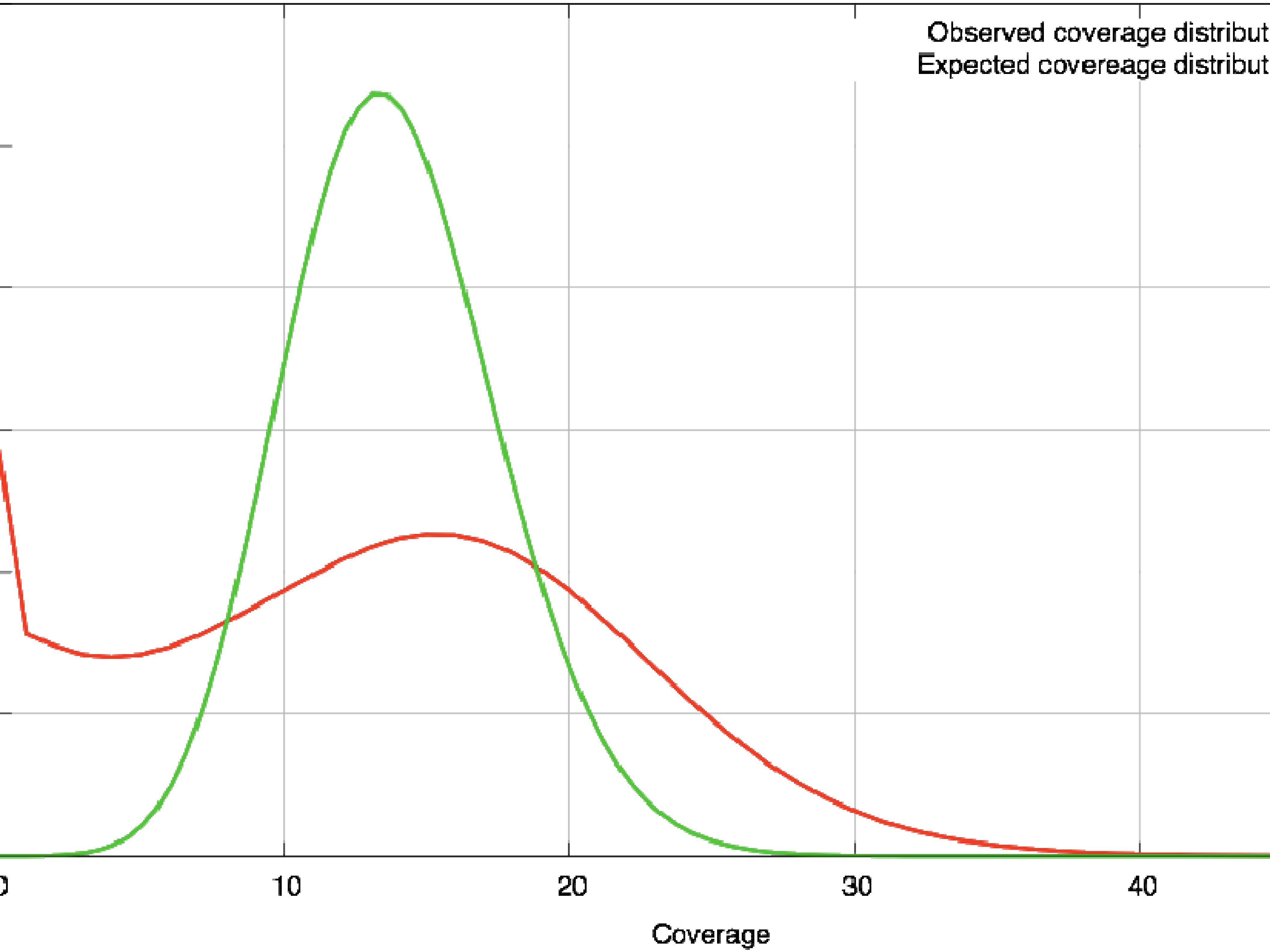
Biases in sequence composition

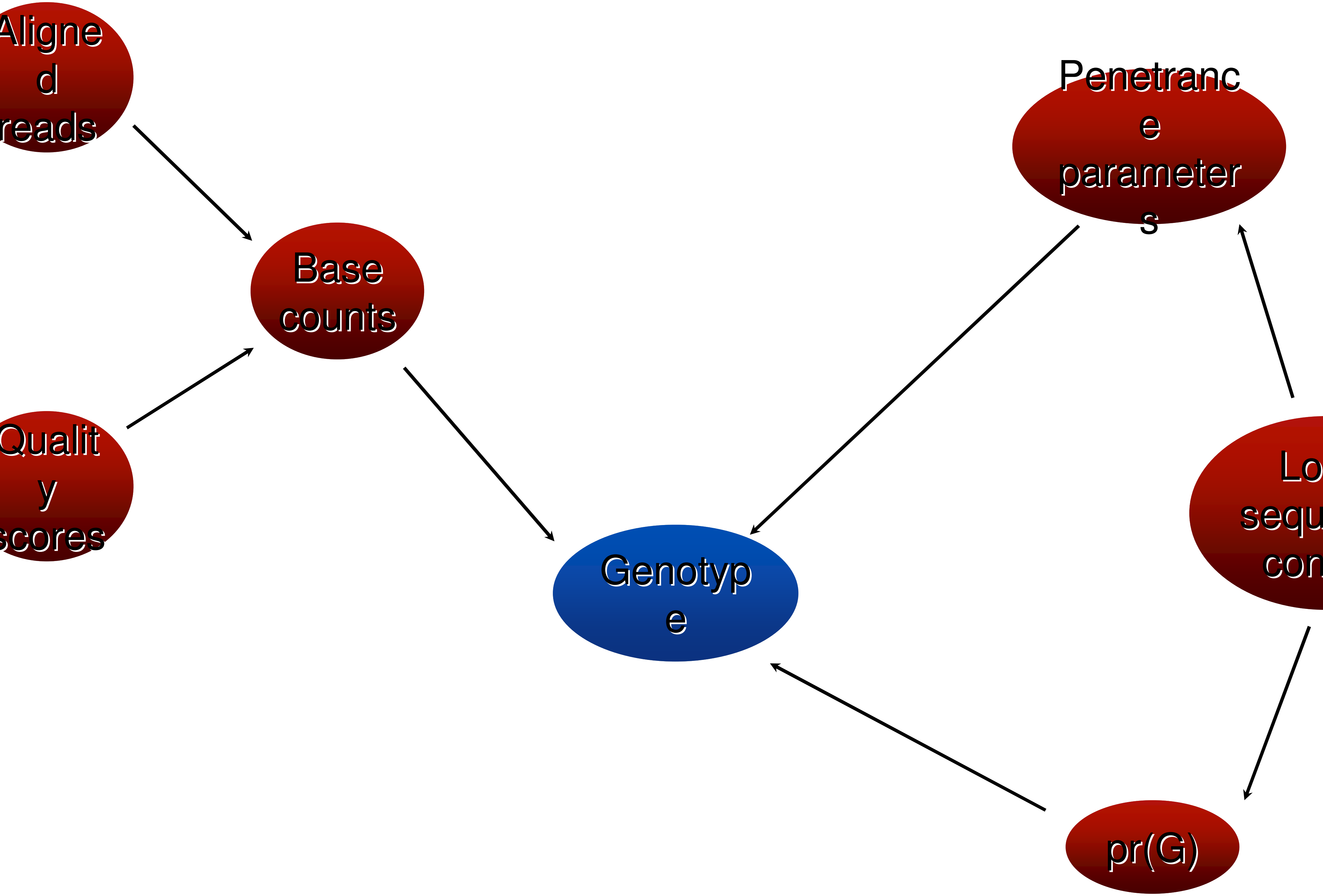
- Ratio of GC:AT bases varies along the genome
- Mutation rate can depend on GC:AT ratio
- Sequencing error rate may also be correlated with GC:AT
- The probability of a section being sequenced can be correlated with GC:AT


Relationship between no. sites and GC content ratio
06
06
06
06
00
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
GC content ratio


Relationship between coverage and GC content ratio
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0
GC content ratio


Observed coverage distribut
0
10
20
30
40
Coverage


Observed coverage distribut
Expected covereage distribut
0
10
20
30
40
Coverage



Aligned reads

Quality scores

Base counts

Genotype

Penetrance parameters

Lo
sequ
con

pr(G)

Example data

- Whole genome sequencing of 1 (human) sample by Solexa sequencers
    - ~10-20 fold coverage across the genome
    - Mix between 36 bp and 76 bp reads
- ~65 Gbases of generated sequence

Example data

- Mapped short reads individually to reference sequence using bowtie alignment software
- Can align all sequences for 1 individual in < 8 hours on a single computer using bowtie and software developed by Mario Foglio
- Extracted reads mapping to chromosome 19 for further analysis

Performance of simple model

- Looked at sequence data and ‘known’ genotype data from chromosome 19 for one individual

- Compared predicted genotypes from sequence data to ‘known’ genotypes (~9500 markers)

- 90% of markers were called, 0.5% discordancy rate

Extra information

- Error rates of bases
- Paired end information
- Local sequence context
- Allele frequencies of known SNPs
- LD relationships between known SNPs
- Previously typed SNPs on the individual
- Data on close relatives

Future work

- Improvement in call rate required - although much of this would simply require increasing coverage
- Systematic analysis of paired end information to detect structural variants
- Collection of phase information from SNPs occurring on the same reads