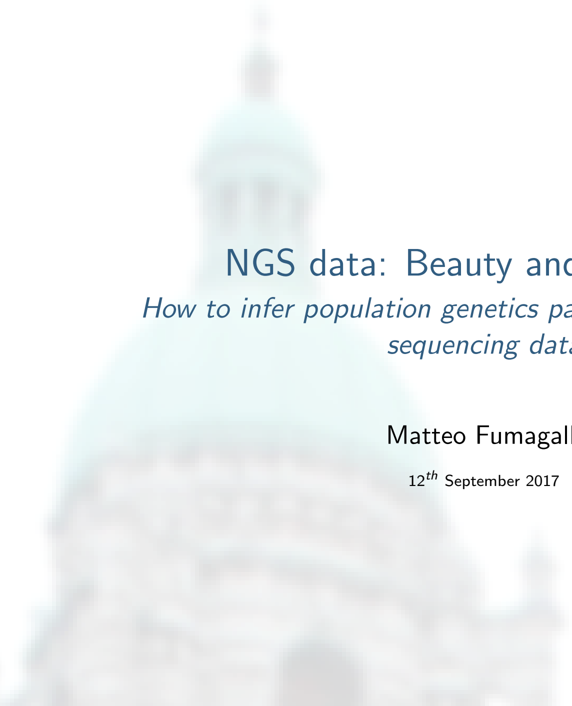
Imperial College London



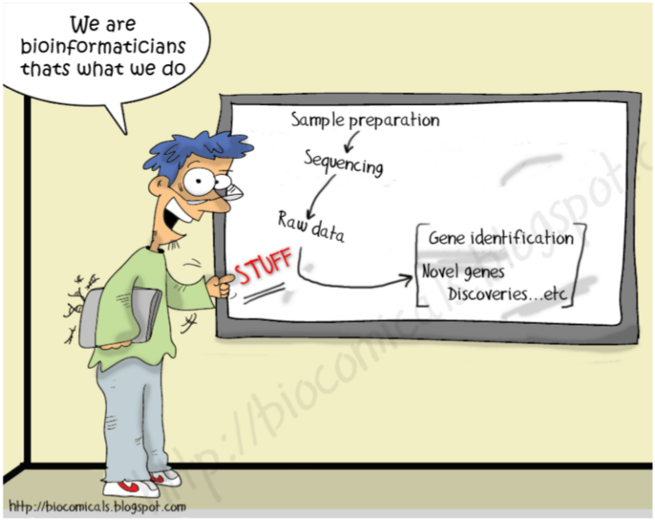
NGS data: Beauty and the Beast

How to infer population genetics parameters from messy sequencing data

Matteo Fumagalli

12^{th} September 2017

Goal of the day

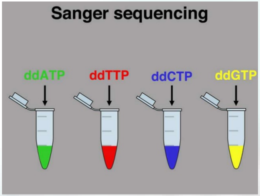


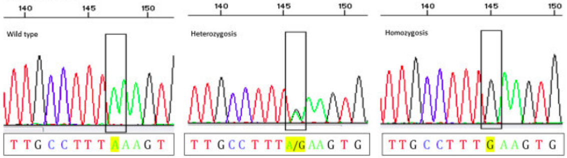
Presentation outline

1. Introduction
2. Genotype likelihoods
3. Genotype calling, really?
    - Low-depth (genotypes)
4. Allele frequencies
    - Low-depth (allele frequencies)
5. Experimental design

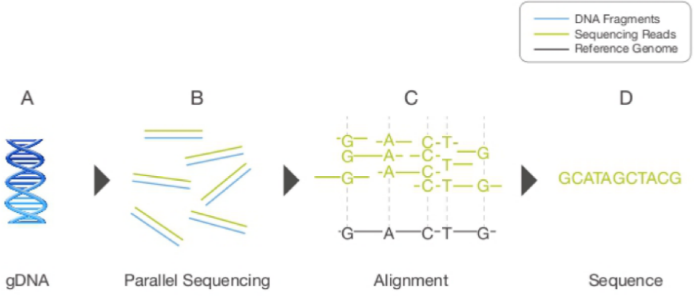
Sanger sequencing

aka first/former generation sequencing



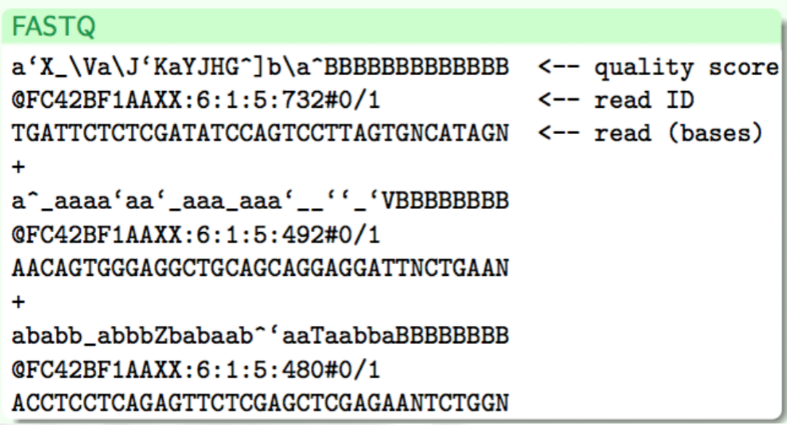


Next Generation Sequencing



A. Extracted gDNA
B. gDNA is fragmented into a library of small segments that are each sequenced in parallel.
C. Individual sequence reads are reassembled by aligning to a reference genome
D. The whole-genome sequence is derived from the consensus of aligned reads.

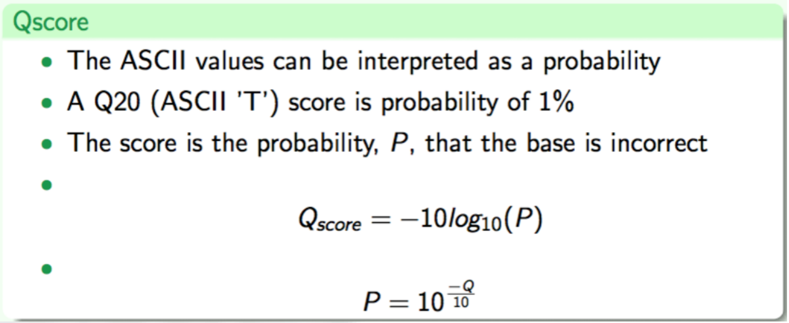
Low-level data



FASTQ

```
a‘X_\Va\J‘KaYJHG^]b\a^BBBBBBBBBBBBB  <-- quality score
@FC42BF1AAXX:6:1:5:732#0/1           <-- read ID
TGATTCTCTCGATATCCAGTCCTTAGTGNCATAGN  <-- read (bases)
+
a^_aaaa‘aa‘_aaa_aaa‘__‘‘_‘VBBBBBBBB
@FC42BF1AAXX:6:1:5:492#0/1
AACAGTGGAGGCTGCAGCAGGAGGATTNCTGAAN
+
ababb_abbbZbabaab^‘aaTaabbaBBBBBBBB
@FC42BF1AAXX:6:1:5:480#0/1
ACCTCCTCAGAGTTCTCGAGCTCGAGAANTCTGGN
```

Quality scores

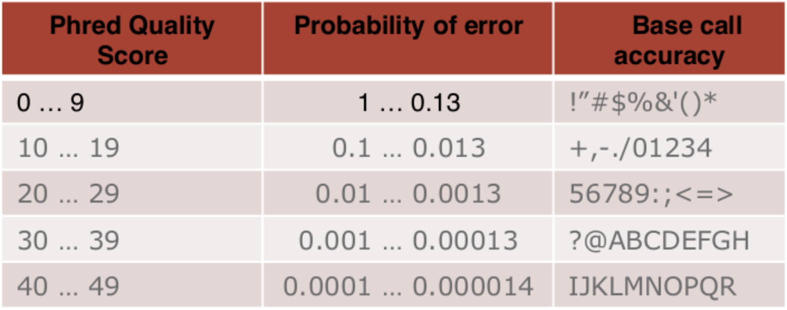


Qscore

- The ASCII values can be interpreted as a probability
- A Q20 (ASCII 'T') score is probability of 1%
- The score is the probability, P, that the base is incorrect
- $$Q_{score} = -10log_{10}(P)$$
- $$P = 10^{\frac{-Q}{10}}$$

Quality scores

The qscores are encoded as ASCII characters, and are shifted by +33 (now the standard) or +64.



Phred Quality Score	Probability of error	Base call accuracy
0 … 9	1 … 0.13	!”#$%&'()*
10 … 19	0.1 … 0.013	+,-./01234
20 … 29	0.01 … 0.0013	56789:;<=>
30 … 39	0.001 … 0.00013	?@ABCDEFGH
40 … 49	0.0001 … 0.000014	IJKLMNOPQR

Quality scores

Example

From a fastq file we observe an ‘A‘ with a qscore encoded as ‘7‘. What is the probability of being ‘A‘? What is the probability of geing ‘G‘?

1. We find that the corresponding ASCII value of ‘7‘ is ?
   (hint: `http://www.asciitable.com`)

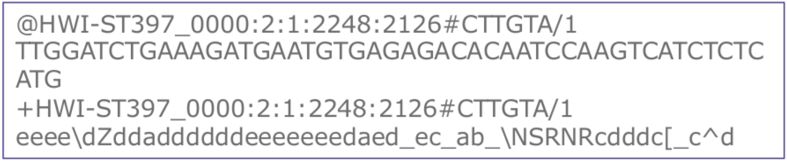
Quality scores

Example

From a fastq file we observe an ‘A‘ with a qscore encoded as ‘7‘. What is the probability of being ‘A‘? What is the probability of geing ‘G‘?

1. We find that the corresponding ASCII value of ‘7‘ is ? (hint: `http://www.asciitable.com`)
2. We substract 33 to get a value of ?. This is our qscore.
3. The probability of ‘A‘ being incorrect is ? (hint: $p = 10^{\frac{-Q}{10}}$)
4. The probability of ‘A‘ being correct is ?
5. The probability of being ‘G‘ (or ‘C‘ or ‘T‘) is ?

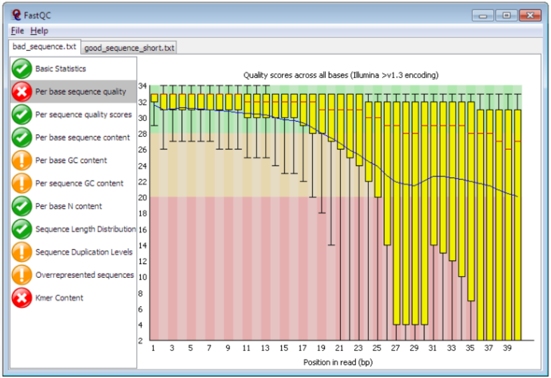
FastQ files



```
@HWI-ST397_0000:2:1:2248:2126#CTTGTA/1
TTGGATCTGAAAGATGAATGTGAGAGACACAATCCAAGTCATCTCTC
ATG
+HWI-ST397_0000:2:1:2248:2126#CTTGTA/1
eeee\dZddadddddeeeeeeeedaed_ec_ab_\NSRNRcdddc[_c^d
```

- sequencer
- flowcell
- lane/cell/tile
- position within the tile
- barcode id for pooling/multiplexing
- pair
- ...

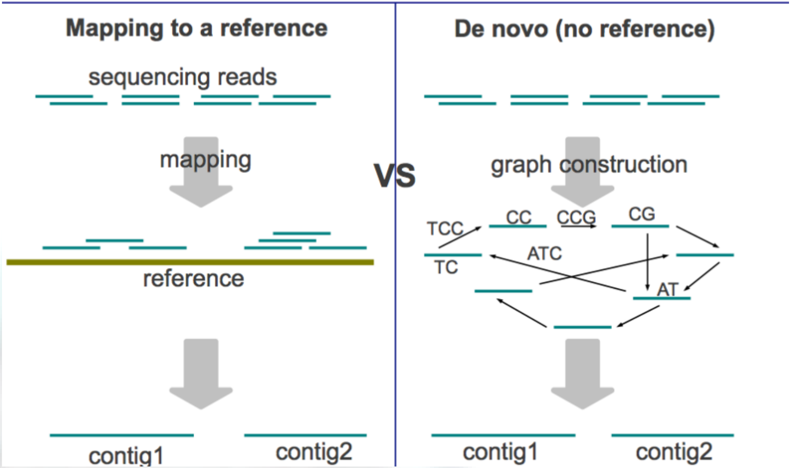
Quality check of fastq files: fastQC



- distribution of qscores over read
- overrepresented kmers
- ...

`http://www.bioinformatics.babraham.ac.uk/projects/fastqc/`

Alignment of reads



Mapping to a reference

Issues:

(i) Millions of short reads.

(ii) Blat/blast is too slow.

But:

Mate-pair gives additional information.

Many aligners exists:

- Soap/soap2 (BGI)
- Maq/bwa (Heng Li)
- Bowtie/bowtie2 (Langmead B)
- Eland, SSAHA2,RMAP,shrimp,zoom,GEM, snap, novoAlign.

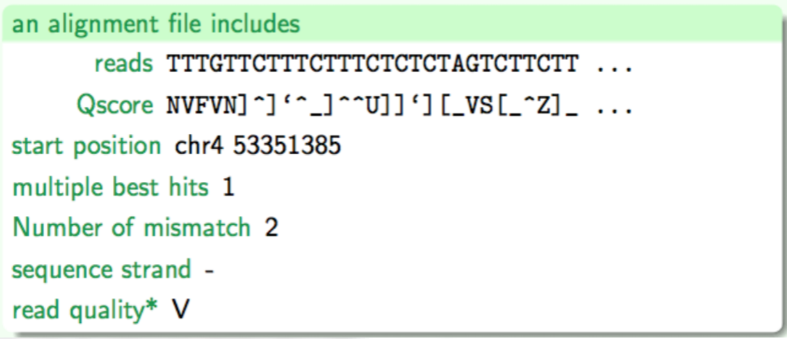
Most are based on burrows wheeler transform BTW (BWA,Bowtie,Soap2,...)

Mapping to a reference

Many different aligners, what's the difference?

- Memory usage
- Speed
- Gapped (indels)
- Using qscores (bwa pssm)
- Estimating a **mappability score**
- Multiple best hits
- Paired end data
- Output **formats** (SAM,...)

Alignment file



an alignment file includes

reads `TTTGTTCTTTCTTTCTCTCTAGTCTTCTT ...`

Qscore `NVFVN]^]‘^_]^^U]]‘][_VS[_^Z]_ ...`

start position chr4 53351385

multiple best hits 1

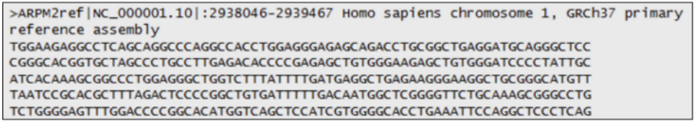
Number of mismatch 2

sequence strand -

read quality* V

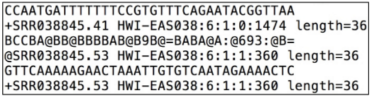
From genomes to variants

Genome (FASTA)



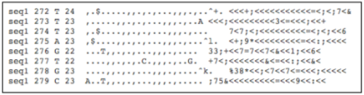
```
>ARPM2ref|NC_000001.10|:2938046-2939467 Homo sapiens chromosome 1, GRCh37 primary
reference assembly
TGGAAGAGGCCTCAGCAGGCCCAGGCCACCTGGAGGGAGAGCAGACCTGCGGCTGAGGATGCAGGGCTCC
CGGGCACGGTGCTAGCCCTGCCTTGAGACACCCCGAGAGCTGTGGGAAGAGCTGTGGGATCCCCTATTGC
ATCACAAAGCGGCCCTGGAGGGCTGGTCTTTATTTTGATGAGGCTGAGAAGGGAAGGCTGCGGGCATGTT
TAATCCGCACGCTTTAGACTCCCCGGCTGTGATTTTTGACAATGGCTCGGGGTTCTGCAAAGCGGGCCTG
TCTGGGGAGTTTGGACCCCGGCACATGGTCAGCTCCATCGTGGGGCACCTGAAATTCCAGGCTCCCTCAG
```

Reads (FASTQ)



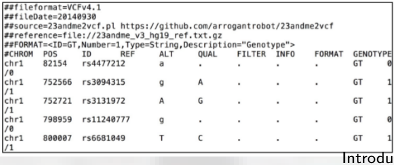
```
CCAATGATTTTTTTCCGTGTTTCAGAATACGGTTAA
+SRR038845.41 HWI-EAS038:6:1:0:1474 length=36
BCCBA@BB@BBBBAB@B9B@=BABA@A:@693:@B=
@SRR038845.53 HWI-EAS038:6:1:1:360 length=36
GTTCAAAAAGAACTAAATTGTGTCAATAGAAAACTC
+SRR038845.53 HWI-EAS038:6:1:1:360 length=36
```

Mapped Reads (mpileup, BAM)



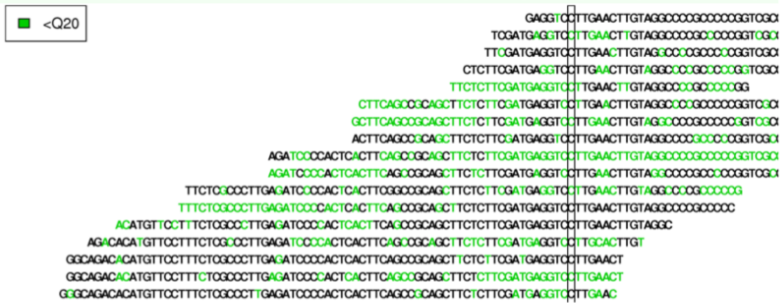
```
seq1 272 T 24  ,.$.....,,.,.,...,,,.,..^+. <<<+;<<<<<<<<<<<=<;<;7<&
seq1 273 T 23  ,.....,,.,.,...,,,.,..A <<<;<<<<<<<<<3<=<<<;<<+
seq1 274 T 23  ,.$....,,.,.,...,,,.,...    7<7;<;<<<<<<<<<=<;<;<<6
seq1 275 A 23  ,$....,,.,.,...,,,.,...^l.  <+;9*<<<<<<<<<=<<:;<<<<
seq1 276 G 22  ...T,,.,.,...,,,.,....  33;+<<7=7<<7<&<<1;<<6<
seq1 277 T 22  ....,,.,.,.C.,,,.,..G.  +7<;<<<<<<<&<=<<:;<<&<
seq1 278 G 23  ....,,.,.,...,,,.,....^k.   %38*<<;<7<<7<=<<<;<<<<<
seq1 279 C 23  A..T,,.,.,...,,,.,..... ;75&<<<<<<<<<=<<<9<<:<<
```

Variants (VCF)



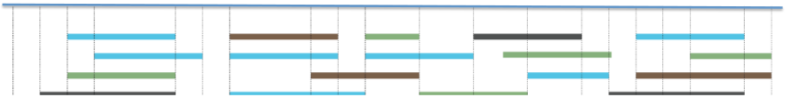
```
##fileformat=VCFv4.1
##fileDate=20140930
##source=23andme2vcf.pl https://github.com/arrogantrobot/23andme2vcf
##reference=file://23andme_v3_hg19_ref.txt.gz
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  GENOTYPE
chr1    82154   rs4477212       a       .       .       .       .       GT      0
/0
chr1    752566  rs3094315       g       A       .       .       .       GT      1
/1
chr1    752721  rs3131972       A       G       .       .       .       GT      1
/1
chr1    798959  rs11240777      g       .       .       .       .       GT      0
/0
chr1    800007  rs6681049       T       C       .       .       .       GT      1
/1
```

Our data: mapped reads with quality scores


🟩 <Q20
GAGGTCCTTGAACTTGTAGGCCCCGCCCCCGGTCGC
TCGATGAGGTCCTTGAACTTGTAGGCCCCGCCCCCGGTCGC
TTCGATGAGGTCCTTGAACTTGTAGGCCCCGCCCCCGGTCGC
CTCTTCGATGAGGTCCTTGAACTTGTAGGCCCCGCCCCCGGTCGC
TTCTCTTCGATGAGGTCCTTGAACTTGTAGGCCCCGCCCCCGG
CTTCAGCCGCAGCTTCTCTTCGATGAGGTCCTTGAACTTGTAGGCCCCGCCCCCGGTCGC
GCTTCAGCCGCAGCTTCTCTTCGATGAGGTCCTTGAACTTGTAGGCCCCGCCCCCGGTCGC
ACTTCAGCCGCAGCTTCTCTTCGATGAGGTCCTTGAACTTGTAGGCCCCGCCCCCGGTCGC
AGATCCCCACTCACTTCAGCCGCAGCTTCTCTTCGATGAGGTCCTTGAACTTGTAGGCCCCGCCCCCGGTCGC
AGATCCCCACTCACTTCAGCCGCAGCTTCTCTTCGATGAGGTCCTTGAACTTGTAGGCCCCGCCCCCGGTCGC
TTCTCGCCCTTGAGATCCCCACTCACTTCGGCCGCAGCTTCTCTTCGATGAGGTCCTTGAACTTGTAGGCCCCGCCCCCG
TTTCTCGCCCTTGAGATCCCCACTCACTTCAGCCGCAGCTTCTCTTCGATGAGGTCCTTGAACTTGTAGGCCCCGCCCCC
ACATGTTCCTTTCTCGCCCTTGAGATCCCCACTCACTTCAGCCGCAGCTTCTCTTCGATGAGGTCCTTGAACTTGTAGGC
AGACACATGTTCCTTTCTCGCCCTTGAGATCCCCACTCACTTCAGCCGCAGCTTCTCTTCGATGAGGTCCTTGCACTTGT
GGCAGACACATGTTCCTTTCTCGCCCTTGAGATCCCCACTCACTTCAGCCGCAGCTTCTCTTCGATGAGGTCCTTGAACT
GGCAGACACATGTTCCTTTCTCGCCCTTGAGATCCCCACTCACTTCAGCCGCAGCTTCTCTTCGATGAGGTCCTTGAACT
GGGCAGACACATGTTCCTTTCTCGCCCTTGAGATCCCCACTCACTTCAGCCGCAGCTTCTCTTCGATGAGGTCCTTGAAC

- Coverage: fraction of the genome with data
- Depth: number of reads mapped to a position
- Counts: number of different alleles mapped to a position
- Effective Base Depth: similar to the counts, but weighing for qscores and mapping quality

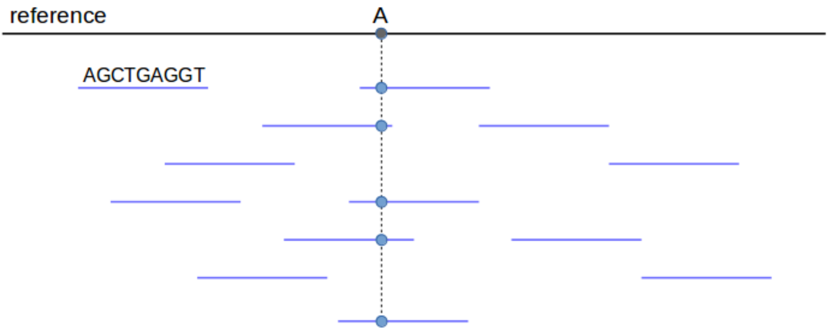
Challenges



- Variable and low depth
- High sequencing and mapping errors

Quality control filters

The data



● is a **nucleotide**/base/allele with a certain **quality** score

Genotype likelihoods

Likelihood

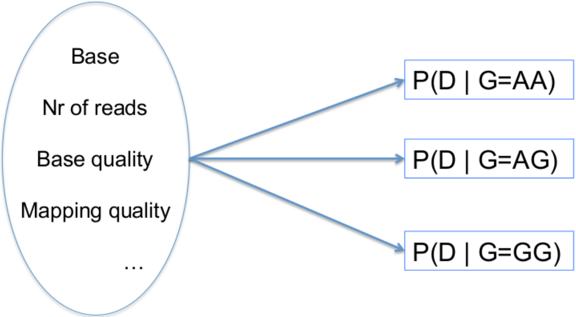
$P(D|G = \{A_1, A_2, ..., A_n\})$
with
$A_i \in \{A, C, G, T\}$ and n being the ploidy

How many genotypes likelihoods do we need to calculate for each each individual at each site?

Genotype likelihoods

```
Chrom1    272    A    24    AAAAAGGAGAGGTAAG    <<<+;<<<<<<<<<<<=<;<;7<&
```

Base quality in Phred scale



Calculating genotype likelihoods

Likelihood function

$$P(D|G = \{A_1, A_2, ..., A_N\}) = \prod_{i=1}^{R} \sum_{j=1}^{N} \frac{L_{A_j,i}}{N}$$

- $L_{A_j,i} = P(D|A_G = A_j)$
- $A_i \in \{A, C, G, T\}$
- R is the depth (nr. of reads)
- N is the ploidy (nr. of chromosomes)

Example:
A
A
A
G
with all quality scores equal to 20 (in phred score)
$P(D|G = AC) =?$

Calculating genotype likelihoods

Likelihood function

$$P(D|G=\{A_1,A_2,...,A_N\})=\prod_{i=1}^{R}\sum_{j=1}^{N}\frac{L_{A_j,i}}{N}$$

A
A
A
G
& Q=20

$$P(D|G=\{A,C\})=...$$

Calculating genotype likelihoods

Likelihood function

$$P(D|G=\{A_1,A_2,...,A_N\})=\prod_{i=1}^{R}\sum_{j=1}^{N}\frac{L_{A_j,i}}{N}$$

A
A
A
G
& Q=20
$N=2; i=1; A_1=A; A_2=C$

$$P(D|G=\{A,C\})=(\frac{L_{A,1}}{2}+\frac{L_{C,1}}{2})\times ...$$

What are $L_{A,1}$ and $L_{C,1}$?

Calculating genotype likelihoods

Likelihood function

$$P(D|G=\{A_1,A_2,...,A_N\})=\prod_{i=1}^{R}\sum_{j=1}^{N}\frac{L_{A_j,i}}{N}$$

A
A
A
G
& Q=20

$$L_{C,1}=\frac{\epsilon_1}{3}$$

$$L_{A,1}=1-\epsilon_1$$

$$P(D|G=\{A,C\})=(\frac{1-\epsilon_1}{2}+\frac{\epsilon_1}{6})\times ...$$

Calculating genotype likelihoods

Likelihood function

$$P(D|G=\{A_1,A_2,...,A_N\})=\prod_{i=1}^{R}\sum_{j=1}^{N}\frac{L_{A_j,i}}{N}$$

A
A
A
G
& Q=20

$$L_{C,1}=\frac{\epsilon_1}{3}$$

$$L_{A,1}=1-\epsilon_1$$

$$P(D|G=\{A,C\})=(\frac{1-\epsilon_1}{2}+\frac{\epsilon_1}{6})\times(\frac{1-\epsilon_2}{2}+\frac{\epsilon_2}{6})\times(\frac{1-\epsilon_3}{2}+\frac{\epsilon_3}{6})\times\frac{\epsilon_4}{3}$$

What are ϵ_1, ϵ_2, ...?

Imperial College
London

Calculating genotype likelihoods

Genotype	Likelihood (log10)
AA	-2.49
AC	**-3.38**
AG	-1.22
AT	-3.38
CC	-9.91
CG	-7.74
CT	-9.91
GG	-7.44
GT	-7.74
TT	-9.91

A
A
A
G
& $\epsilon = 0.01$

Genotype calling

Genotype	Likelihood (log10)
AA	-2.49
AC	-3.38
AG	-1.22
AT	-3.38
CC	-9.91
CG	-7.74
CT	-9.91
GG	-7.44
GT	-7.74
TT	-9.91

AAAG & $\epsilon = 0.01$

What is the genotype here?

Genotype calling

Genotype	Likelihood (log10)
AA	-2.49
AC	-3.38
AG	**-1.22**
AT	-3.38
CC	-9.91
CG	-7.74
CT	-9.91
GG	-7.44
GT	-7.74
TT	-9.91

AAAG & $\epsilon = 0.01$
What is the genotype?
AG.

Maximum Likelihood

The simplest genotype caller: choose the genotype with the highest likelihood.

Major and minor alleles

Likelihood function

$$\log P(D|G = A) = \sum_{i=1}^{R} \log L_{A_j,i}$$

AAAG & $\epsilon = 0.01$

Allele	log-Likelihood
A	**-2.49**
C	-3.38
G	**-1.22**
T	-3.38

We can reduce the genotype space to 3 entries (from 10).

Genotype likelihoods

AAAG & ‘5555‘ & A,G alleles

Genotype	log-Likelihood
AA	-5.73
AG	-2.80
GG	-17.12

Examples varying qualities and reads...

Genotype likelihoods - example

AAAG & ‘555**0**‘ & A,G alleles

Genotype	log-Likelihood

Genotype likelihoods - example

AAAG & ‘555**0**‘ & A,G alleles

Genotype	log-Likelihood
AA	-4.58
AG	-2.81
GG	-17.14

Genotype likelihoods - example

AAAG & ‘555**K**‘ & A,G alleles

Genotype	log-Likelihood

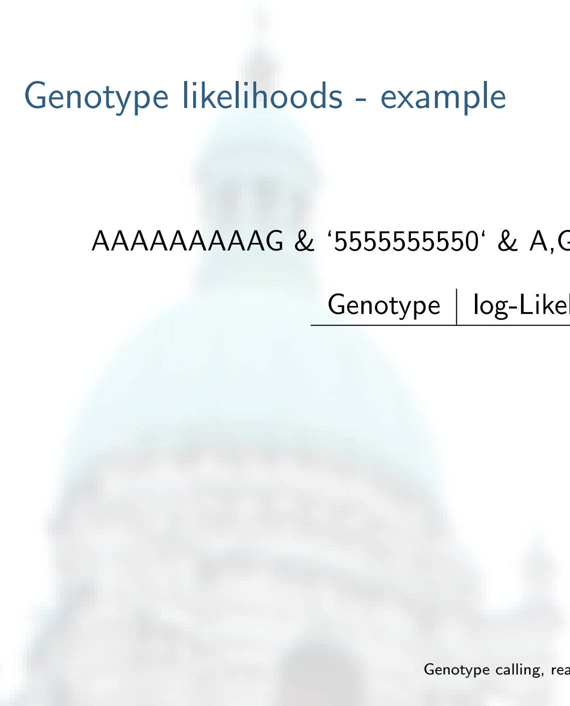
Genotype likelihoods - example

AAAG & ‘555**K**‘ & A,G alleles

Genotype	log-Likelihood
AA	-10.80
AG	-2.80
GG	-17.11

Imperial College
London

Genotype likelihoods - example



AAAAAAAAAG & ‘5555555550‘ & A,G alleles

Genotype	log-Likelihood

Genotype likelihoods - example

AAAAAAAAAG & ‘5555555550‘ & A,G alleles

Genotype	log-Likelihood
AA	-4.64
AG	-7.01
GG	-51.37

NGS data uncertainty

Issue

There is a notable amount of statistical uncertainty in assigning individual genotypes depending on the number of reads and their quality. How can we deal with that when doing population genetics analysis (e.g. estimating genetic diversity)?

Solutions

1. let's pretend we don't have such uncertainty

NGS data uncertainty

Issue

There is a notable amount of statistical uncertainty in assigning individual genotypes depending on the number of reads and their quality. How can we deal with that when doing population genetics analysis (e.g. estimating genetic diversity)?

Solutions

1. let's pretend we don't have such uncertainty
2. genotype filtering
3. (a third way)

Genotype likelihood ratio

$$\log_{10} \frac{L_{G(1)}}{L_{G(2)}} > t$$

i.e. $t = 1$ meaning that the most likely genotype is 10 times more likely than the second most likely one

Genotype posterior probability

AAAG & $\epsilon = 0.01$ & A,G alleles

Genotype	Likelihood (log)	Prior	Posterior
AA	-5.73	1/3	0.05
AG	-2.80	1/3	0.95
GG	-17.12	1/3	0

Only call genotypes if the largest probability is above a certain threshold (e.g. 0.95).

Pros and cons?

Genotype posterior probability

AAAG & $\epsilon = 0.01$ & A,G alleles

Genotype	Likelihood (log)	Prior	Posterior
AA	-5.73	1/3	0.05
AG	-2.80	1/3	0.95
GG	-17.12	1/3	0

Only call genotypes if the largest probability is above a certain threshold (e.g. 0.95).

Pros and cons?

- Yes: genotype are called with higher **confidence**
- No: more **missing** data

Exercise 1

Simulate NGS data and calculate genotype likelihoods and probabilities.
Assess the amount of uncertainty and data missingness.

nr of hetero individuals
assuming known genotypes with example of 4 samples
do not show real AA AG AG GG real
data A AAAG AGGG GG
unknown genotypes:
calculate by sampling distribution of nr of hetero from geno post probs unif
then show expected value

genetic distances

pca

Exercise 2

PCA with both methods
Alone: best PCA with filtering (perhaps later when introducing snp calling)

Estimating allele frequencies

Assuming 2 alleles (A,G) with true allele frequency of 0.50

Sample	True genotype	Reads allele A	Read allele G
1	AA	7	0
2	AA	25	1
3	AG	5	3
4	AG	4	4
5	GG	0	2
6	GG	0	4

What is the simplest estimator of allele frequencies?

Estimating allele frequencies

Assuming 2 alleles (A,G) with true allele frequency of 0.50

Sample	True genotype	Reads allele A	Read allele G
1	AA	7	0
2	AA	25	1
3	AG	5	3
4	AG	4	4
5	GG	0	2
6	GG	0	4
Total		41	14

$$\hat{f} = \frac{\sum_{i=1}^{N} n_{A,i}}{\sum_{i=1}^{N}(n_{A,i} + n_{G,i})}$$

$\hat{f} = 0.75$

What is wrong with this estimator?

Estimating allele frequencies

Assuming 2 alleles (A,G) with true allele frequency of 0.50

Sample	True genotype	Reads allele A	Read allele G
1	AA	7	0
2	AA	25	1
3	AG	5	3
4	AG	4	4
5	GG	0	2
6	GG	0	4
Total		41	14

$$\hat{n_A} = \sum_{i=1}^{N} (1-\epsilon) n_{A,i} + \epsilon n_{G,i} - \epsilon n_{A,i} - (1-\epsilon) n_{G,i}$$

$\hat{f} = 0.77$

Estimating allele frequencies

Maximum Likelihood estimator

$$P(D|f) = \prod_{i=1}^{N} \sum_{g \in \{0,1,2\}} P(D|G=g)P(G=g|f)$$

Estimating allele frequencies

Maximum Likelihood estimator

$$P(D|f) = \prod_{i=1}^{N} \sum_{g\in\{0,1,2\}} P(D|G=g)P(G=g|f)$$

$P(D|G=g)$ is the genotype likelihood and $P(G=g|f)$ is given by HWE (for instance).

In our previous example, $\hat{f} = 0.46$ which is much closer to the true value than previous estimators.

SNP calling

Challenges

- If high levels of missing data, then genotypes can be lost.
- Rare variants are hard to detect.
- Trade off between false positive and false negative rates.

How to call SNPs?

- If at least one heterozygous genotype has been called.
- If the estimated allele frequency is above a certain threshold.

SNP calling

Call a SNP if

$$\hat{f} \geq t$$

where t can be the minimum sample allele frequency detectable (e.g. $t = 1/2N$ with N diploids).

Likelihood Ratio Test

A Likelihood Ratio Test (LRT) compares the goodness of fit between the null and the alternative model:

- Null model: $f = 0$
- Alternative model: $f \neq 0$

$$T = -2 \log \frac{L(f = 0)}{L(f = \hat{f}_{MLE})}$$

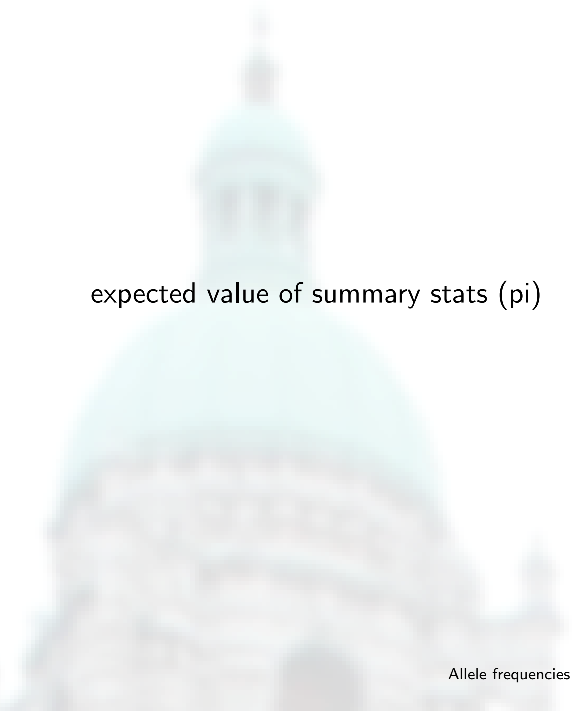
where T is χ^2 distributed with 1 degree of freedom.

Exercise 3

Estimate allele frequencies and call SNPs
calculate PCA with some filtering

expected value of proba being variable

expected value of nr snps

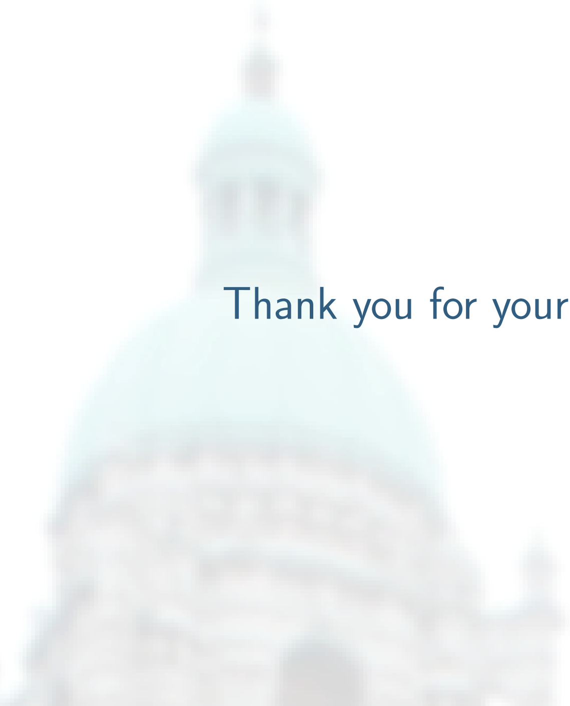


expected value of summary stats (pi)

Exercise - 4

calculate SFS
calculate summary stats 1 pop
alone: sliding windows
alone: 2 pops (2d-sfs and fst)

exp design with bigfoot



Thank you for your attention