Introduction to NGS data: Genotype and SNP calling

Matteo Fumagalli

Population genetics

Demographic inference

Adaptive evolution

Find the bioinformatician inside yourself

Forget about…

	A	B	C	D	E	F	G	H
1	DRIVER NAME		SETTLEMENT DATE		UNIT #		TRUCK	DRIVER
2	John Smith		01/21/07		3386			
3								
4	RUN NUMBER	RUN DATE	RUN MILES	RUN GROSS	DRIVER PAY			
5	PHL90748	11/06/07	346	$375.41	$356.63	$356.64	$489.59	
6	STL66360	11/07/07	353	$321.25	$305.18	$305.19	$448.31	
7	ABE04364	11/09/07	229	$411.59	$391.01	$391.01	$494.03	
8	RDU56336	11/12/07	1299	$1,651.03	$1,568.47	$1,568.48	$2,118.67	
9	RDU46336	11/12/07	1	$100.00	$95.00	$95.00	$100.00	
10				$0.00	$0.00	$0.00		
11				$0.00	$0.00	$0.00		
12				$0.00	$0.00	$0.00		
13	DRIVER TOTAL RUN GROSS PAY				$2,716.29	$2,716.32	$2,859.28	$2,716.31
14	FUEL SUR-CHARGE					AAFT	$142.99	
15	RUN NUMBER	RUN DATE	FSC $ PER MILE	RUN MILES	FUEL SUR-CHARGE			
16	PHL90748	11/06/07	$0.3300	346	$114.18			
17	STL66360	11/07/07	$0.3599	353	$127.06			
18	ABE04364	11/09/07	$0.3600	229	$82.44			
19	RDU56336	11/12/07	$0.3600	1299	$467.64			
20	RDU46336	11/12/07	$0.0000	1	$0.00			
21	0	12/30/99		0				
22	0	12/30/99		0				
23	0	12/30/99		0				$7,301.20

NOTES

Spreadsheet Example

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.

Applications

whole genome

ancient genomes

Exome capturing

		F5524885	F5524885 F5510208	F5524885 F5510208 F5510066	F5524885 F5510208 F5510066 F5522194	Any 3 of 4 F5524885 F5510208 F5510066 F5522194
Number of genes in which each affected has at least one...	Non-synonymous cSNP, splice site variant or coding indel (NS/SS/I)	4,510	3,284	2,765	2,479	3,768
	NS/SS/I not in dbSNP	513	128	71	59	119
	NS/SS/I not in eight HapMap exomes	799	168	53	21	160
	NS/SS/I neither in dbSNP nor eight HapMap exomes	360	38	8	1 (MYH3)	22
	...And predicted to be damaging	160	10	2	1 (MYH3)	3

RAD-sequencing

http://www.floragenex.com/

Pooled sequencing

...

Mapped reads

- **Depth**: number of reads mapped to a position
- **Coverage**: fraction of the genome with data

Workflow

Lab stuff
Sequencing
FastQ
recalibrate
Alignment
Genotype likelihoods
Allele frequencies
Haplotype frequencies
Genotype probabilities
Genotype calling
Variation detection
Population Genetics
other
Medical genetics

Workflow

Low-level data:

- Samples preparation + sequencing
- Call bases and quality scores

Workflow

Low-level data:

- Samples preparation + sequencing
- Call bases and quality scores

Genotype data:

- Call genotypes
- Estimate allele/haplotype frequencies
- SNPs detection

Workflow

Low-level data:

- Samples preparation + sequencing
- Call bases and quality scores

Genotype data:

- Call genotypes
- Estimate allele/haplotype frequencies
- SNPs detection

Analysis:

- Population genetics analysis
- Association studies

From genome 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
```

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
AACAGTGGGAGGCTGCAGCAGGAGGATTNCTGAAN
+
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
- $$Q_{score} = -10log_{10}(P)$$
- $$P = 10^{\frac{-Q}{10}}$$

!"#$%&'()*+,-./0123456789:;⇔@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~

Statistical inference (1)

Data (D)

Parameter *f* is frequency of G

Statistical inference (1)

Data (D)

Parameter *f* is frequency of G

Statistical inference (1)

Data (D)

Likelihood function
(proportional to the sampling distribution)

Parameter *f* is
frequency of G

Statistical inference (1)

Likelihood approach:

- All the information on the parameter is in the likelihood function (we use all the data!).
- More data leads to less bias and less variance.
- Suitable for hypothesis testing.

Statistical inference (1)

Likelihood approach:

- All the information on the parameter is in the likelihood function (we use all the data!).
- More data leads to less bias and less variance.
- Suitable for hypothesis testing.

Genotype likelihoods

$$L(Data \mid G = \{A_1, A_2\})$$

$$A_i \in \{A, C, G, T\}$$

How many genotype likelihoods do we have for each individual at each site?

Genotype likelihoods

$$L(Data \mid G = \{A_1, A_2\})$$

$$A_i \in \{A, C, G, T\}$$

How many genotype likelihoods do we have for each individual at each site?

3 if both alleles are known
10 if not

Genotype likelihoods

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

Base quality in Phred scale

Genotype likelihoods

- **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. (2011): type specific errors
- …

Calculating genotype likelihoods

$$P(X|G=bh)=\prod_{i=1}^{r}\left(\frac{L_b^{(i)}}{2}+\frac{L_h^{(i)}}{2}\right) \qquad b,h\in\{A,C,G,T\}$$

Calculating genotype likelihoods

$$P(X|G=bh)=\prod_{i=1}^{r}\left(\frac{L_b^{(i)}}{2}+\frac{L_h^{(i)}}{2}\right) \qquad b,h\in\{A,C,G,T\}$$

Example:

Chrom6 342

A T T T	Individual 1
T T	Individual 2
A A T T	Individual 3

Calculating genotype likelihoods

$$P(X|G=bh)=\prod_{i=1}^{r}\left(\frac{L_b^{(i)}}{2}+\frac{L_h^{(i)}}{2}\right) \qquad b,h\in\{A,C,G,T\}$$

Example:

Chrom6 342 A T T T

AA
AC
AG
AT
CC
CG
CT
GG
GT
TT

Iterate through every read for every genotypic configuration...

Calculating genotype likelihoods

$$P(X|G=bh)=\prod_{i=1}^{r}\left(\frac{L_b^{(i)}}{2}+\frac{L_h^{(i)}}{2}\right) \qquad b,h\in\{A,C,G,T\}$$

Example:

Chrom6 342 A T T T

AA
AC
AG
AT
CC
CG
CT
GG
GT
TT

Iterate through every read for every genotypic configuration...

Calculating genotype likelihoods

$$P(X|G=bh)=\prod_{i=1}^{r}\left(\frac{L_b^{(i)}}{2}+\frac{L_h^{(i)}}{2}\right) \qquad b,h\in\{A,C,G,T\}$$

Example:

Chrom6 342 A T T T

$$P(X|G=AC)=$$

Calculating genotype likelihoods

$$P(X|G=bh)=\prod_{i=1}^{r}\left(\frac{L_b^{(i)}}{2}+\frac{L_h^{(i)}}{2}\right) \qquad b,h\in\{A,C,G,T\}$$

Example:

Chrom6 342 A T T T

$$P(X|G=AC)=$$

Calculating genotype likelihoods

$$P(X|G=bh)=\prod_{i=1}^{r}\left(\frac{L_b^{(i)}}{2}+\frac{L_h^{(i)}}{2}\right) \qquad b,h\in\{A,C,G,T\}$$

Example:

Chrom6 342 $\boxed{\text{A}}$TTT

$$P(X|G=AC)=\left(\frac{L_A^{(1)}}{2}+\frac{L_C^{(1)}}{2}\right)*$$

Calculating genotype likelihoods

$$P(X|G=bh)=\prod_{i=1}^{r}\left(\frac{L_b^{(i)}}{2}+\frac{L_h^{(i)}}{2}\right) \qquad b,h\in\{A,C,G,T\}$$

Example:

Chrom6 342 A T T T

$$P(X|G=AC)=\left(\frac{L_A^{(1)}}{2}+\frac{L_C^{(1)}}{2}\right)*$$

$L_A^{(1)}=$

$L_C^{(1)}=$

Calculating genotype likelihoods

$$P(X|G=bh)=\prod_{i=1}^{r}\left(\frac{L_b^{(i)}}{2}+\frac{L_h^{(i)}}{2}\right) \qquad b,h\in\{A,C,G,T\}$$

Example:

Chrom6 342 A T T T

$$P(X|G=AC)=\left(\frac{L_A^{(1)}}{2}+\frac{L_C^{(1)}}{2}\right)*$$

$$L_A^{(1)}=$$

$$L_C^{(1)}=\frac{\varepsilon}{3}$$

Calculating genotype likelihoods

$$P(X|G=bh)=\prod_{i=1}^{r}\left(\frac{L_b^{(i)}}{2}+\frac{L_h^{(i)}}{2}\right) \qquad b,h\in\{A,C,G,T\}$$

Example:

Chrom6 342 A T T T

$$P(X|G=AC)=\left(\frac{L_A^{(1)}}{2}+\frac{L_C^{(1)}}{2}\right)*$$

$$L_A^{(1)}=1-\varepsilon$$

$$L_C^{(1)}=\frac{\varepsilon}{3}$$

$$P(X=A|G=AC)=\frac{1-\varepsilon}{2}+\frac{\varepsilon}{6}$$

Calculating genotype likelihoods

$$P(X|G=bh)=\prod_{i=1}^{r}\left(\frac{L_b^{(i)}}{2}+\frac{L_h^{(i)}}{2}\right) \qquad b,h\in\{A,C,G,T\}$$

Example:

Chrom6 342 A T T T

$$P(X|G=AC)=\left(\frac{L_A^{(1)}}{2}+\frac{L_C^{(1)}}{2}\right)*\left(\frac{L_A^{(2)}}{2}+\frac{L_C^{(2)}}{2}\right)*$$

Calculating genotype likelihoods

$$P(X|G=bh)=\prod_{i=1}^{r}\left(\frac{L_b^{(i)}}{2}+\frac{L_h^{(i)}}{2}\right) \qquad b,h\in\{A,C,G,T\}$$

Example:

Chrom6 342 A T T T

$$P(X|G=AC)=\left(\frac{L_A^{(1)}}{2}+\frac{L_C^{(1)}}{2}\right)*\left(\frac{L_A^{(2)}}{2}+\frac{L_C^{(2)}}{2}\right)*$$

$$\frac{\varepsilon}{3}$$

Calculating genotype likelihoods

$$P(X|G=bh)=\prod_{i=1}^{r}\left(\frac{L_b^{(i)}}{2}+\frac{L_h^{(i)}}{2}\right) \qquad b,h\in\{A,C,G,T\}$$

Example:

Chrom6 342 A T T T

$$P(X|G=AC)=\left(\frac{L_A^{(1)}}{2}+\frac{L_C^{(1)}}{2}\right)*\left(\frac{L_A^{(2)}}{2}+\frac{L_C^{(2)}}{2}\right)*\left(\frac{L_A^{(3)}}{2}+\frac{L_C^{(3)}}{2}\right)*\left(\frac{L_A^{(4)}}{2}+\frac{L_C^{(4)}}{2}\right)$$

$$=\left(\frac{1-\varepsilon}{2}+\frac{\varepsilon}{6}\right)*\frac{\varepsilon}{3}*\frac{\varepsilon}{3}*\frac{\varepsilon}{3}$$

Genotype likelihoods

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

$$ATTT$$

$$\varepsilon = 0.01$$

Genotype calling

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

$$ATTT$$

$$\varepsilon = 0.01$$

What is the genotype here?

Genotype calling

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

Simple genotype caller:
Maximum Likelihood

AT

Choose the genotype with the largest likelihood

Genotype calling

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

Simple genotype caller:
Maximum Likelihood

But **only** call the genotype if the largest likelihood is **much better** than the second best

Genotype calling

- Likelihood Ratio:

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

$$t = 1$$

The most likely genotype is at least **10 times** more likely than the second most likely one

(in our example *t*=1.27)

Genotype calling

- Likelihood Ratio:

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

$$t = 1$$

The most likely genotype is at least **10 times** more likely than the second most likely one

- Higher **confidence** of called genotypes
- More **missing** data

Statistical inference (2)

Statistical inference (2)

Statistical inference (2)

Bayesian inference

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)} = \frac{P(X|\theta)P(\theta)}{\sum_{\theta} P(X|\theta)P(\theta)}$$

$P(X|\theta)$ ← Likelihood of θ

$P(\theta)$ ← Prior probability distribution of θ

$P(\theta|X)$ ← Posterior probability distribution of θ

- We update our “belief” on the parameter after performing the experiment
- Parameter is not fixed (point estimate) but rather has a probability distribution
- As P(f|D) is a proper probability distribution, we can easily derive credible intervals

Genotype posterior probabilities

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

Simple genotype caller:
Bayesian

?

Genotype posterior probabilities

Genotype	Likelihood (log10)	Prior	Posterior probability
AA	-7.44	1/10	~ 0
AC	-7.74	1/10	~ 0
AG	-7.74	1/10	~ 0
AT	-1.22	1/10	0.94
CC	-9.91	1/10	~ 0
CG	-9.91	1/10	~ 0
CT	-3.38	1/10	0.006
GG	-9.91	1/10	~ 0
GT	-3.38	1/10	0.006
TT	-2.49	1/10	0.05

Simple genotype caller:
Bayesian

?

Genotype posterior probabilities

Genotype	Likelihood (log10)	Prior	Posterior probability
AA	-7.44	1/10	~ 0
AC	-7.74	1/10	~ 0
AG	-7.74	1/10	~ 0
AT	-1.22	1/10	**0.94**
CC	-9.91	1/10	~ 0
CG	-9.91	1/10	~ 0
CT	-3.38	1/10	0.006
GG	-9.91	1/10	~ 0
GT	-3.38	1/10	0.006
TT	-2.49	1/10	0.05

Simple genotype caller:
Bayesian

AT

Genotype posterior probabilities

Genotype	Likelihood (log10)	Prior	Posterior probability
AA	-7.44	1/10	~ 0
AC	-7.74	1/10	~ 0
AG	-7.74	1/10	~ 0
AT	-1.22	1/10	**0.94**
CC	-9.91	1/10	~ 0
CG	-9.91	1/10	~ 0
CT	-3.38	1/10	0.006
GG	-9.91	1/10	~ 0
GT	-3.38	1/10	0.006
TT	-2.49	1/10	0.05

Simple genotype caller: **Bayesian**

But **only** call the genotype if the largest probability is above a threshold (e.g. > 0.95)

Genotype posterior probabilities

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

Simple genotype caller:
Bayesian

Example: reference is T

→ AT (?)

Genotype posterior probabilities

Genotype	Likelihood (log10)	Prior	Posterior probability
AA	-7.44	0.01	~ 0
AC	-7.74	0.01	~ 0
AG	-7.74	0.01	~ 0
AT	-1.22	0.09	**0.67**
CC	-9.91	0.01	~ 0
CG	-9.91	0.01	~ 0
CT	-3.38	0.09	0.005
GG	-9.91	0.01	~ 0
GT	-3.38	0.09	0.0005
TT	-2.49	0.81	**0.32**

Simple genotype caller: **Bayesian**

P(A) = 0.9 if A is the **reference** allele;
P(A) = 0.1 otherwise

→ AT (?)

Example: reference is T

$P(TT) = P(A)^2$

e.g. Illumina Casava

Genotype posterior probabilities

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

Better genotype caller: **Bayesian**

P(A) = f

Where f (=0.75) is the **allele frequency** from a reference panel

Example: reference is T

P(TT) = …
P(AT) = …
P(AA) = …

Genotype posterior probabilities

Genotype	Likelihood (log10)	Prior	Posterior probability
AA	-7.44		
AC	-7.74		
AG	-7.74		
AT	-1.22		
CC	-9.91		
CG	-9.91		
CT	-3.38		
GG	-9.91		
GT	-3.38		
TT	-2.49	0.56	

Better genotype caller: **Bayesian**

P(A) = f

Where f (=0.75) is the **allele frequency** from a reference panel

Example: reference is T

$P(TT) = f^2$
P(AT) = …
P(AA) = …

Genotype posterior probabilities

Genotype	Likelihood (log10)	Prior	Posterior probability
AA	-7.44		
AC	-7.74		
AG	-7.74		
AT	-1.22	0.38	
CC	-9.91		
CG	-9.91		
CT	-3.38		
GG	-9.91		
GT	-3.38		
TT	-2.49	0.56	

Better genotype caller: **Bayesian**

P(A) = f

Where f (=0.75) is the **allele frequency** from a reference panel

Example: reference is T

$P(TT) = f^2$
$P(AT) = 2f(1-f)$
$P(AA) = \ldots$

Genotype posterior probabilities

Genotype	Likelihood (log10)	Prior	Posterior probability
AA	-7.44	0.06	~ 0
AC	-7.74	0	0
AG	-7.74	0	0
AT	-1.22	0.38	**0.93**
CC	-9.91	0	0
CG	-9.91	0	0
CT	-3.38	0	0
GG	-9.91	0	0
GT	-3.38	0	0
TT	-2.49	0.56	0.07

Better genotype caller: **Bayesian**

P(A) = f

Where f is the **allele frequency** from a reference panel

Example: reference is T

$P(TT) = f^2$
$P(AT) = 2f(1-f)$
$P(AA) = (1-f)^2$

Assuming **f=0.75** and only **A and T** alleles

Statistical inference (3)

Genotype posterior probabilities

Genotype	Likelihood (log10)	Prior	Posterior probability
AA	-7.44	0.16	~ 0
AC	-7.74	0	0
AG	-7.74	0	0
AT	-1.22	0.48	**0.96**
CC	-9.91	0	0
CG	-9.91	0	0
CT	-3.38	0	0
GG	-9.91	0	0
GT	-3.38	0	0
TT	-2.49	0.36	0.38

Better genotype caller: **Empirical Bayesian**

P(A) = f

Where f is the **allele frequency** estimated from the data itself

With **f=0.6**

Workflow

Low-level data:

- Samples preparation + sequencing
- Call bases and quality scores

Genotype data:

- Call genotypes
- **Estimate allele frequencies**
- **SNPs detection**

Analysis:

- Population genetics analysis
- Association studies

SNP calling procedures

- Alignment-based caller

We completely rely on how reads have been mapped

Figure from Erik Garrison

SNP calling procedures

- Assembly-based caller (as in GATK)

Local re-alignment around putative variants; better resolution for INDELs detection.

- Haplotype-based caller (as in freebayes)

Figure from Erik Garrison

Estimating allele frequencies

Individual	True genotype	Reads allele A	Reads allele G
1	AA		
2	AA		
3	AG		
4	AG		
5	GG		
6	GG		
Tot.			

Assume only 2 allelic types

True allele frequency is 0.50

Estimating allele frequencies

Individual	True genotype	Reads allele A	Reads 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
Tot.		41	14

Assume only 2 allelic types

True allele frequency is 0.50

Estimating allele frequencies

Individual	True genotype	Reads allele A	Reads 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
Tot.		41	14

Simple allele frequency estimator:

from **reads counts**

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

Estimating allele frequencies

Individual	True genotype	Reads allele A	Reads 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
Tot.		41	14

Simple allele frequency estimator:

from **reads counts**

$$\hat{f} = \frac{\sum_{i=1}^{N} n_{(A,i)}}{\sum_{i=1}^{N} (n_{(A,i)} + n_{(G,i)})} = 0.75$$

Estimating allele frequencies

Individual	True genotype	Reads allele A	Reads 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
Tot.		41	14

Simple allele frequency estimator:

from **reads counts**

$$\hat{f} = \frac{\sum_{i=1}^{N} n_{(A,i)}}{\sum_{i=1}^{N} (n_{(A,i)} + n_{(G,i)})} = 0.75$$

Estimating allele frequencies

Individual	True genotype	Reads allele A	Reads 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
Tot.		41	14

Simple allele frequency estimator:

from **reads counts with error**

$$\hat{f} = \frac{\sum_{i=1}^{N} (n_{(A,i)} - \varepsilon(n_{(A,i)} + n_{(G,i)}))}{\sum_{i=1}^{N} (n_{(A,i)} + n_{(G,i)})(1 - 2\varepsilon)}$$

Estimating allele frequencies

Individual	True genotype	Reads allele A	Reads 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
Tot.		41	14

Simple allele frequency estimator:

from **reads counts with error**

$$\hat{f} = \frac{\sum_{i=1}^{N}(n_{(A,i)} - \varepsilon(n_{(A,i)} + n_{(G,i)}))}{\sum_{i=1}^{N}(n_{(A,i)} + n_{(G,i)})(1-2\varepsilon)} = 0.77$$

Estimating allele frequencies

Individual	True genotype	Reads allele A	Reads 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
Tot.		41	14

Simple allele frequency estimator:

from **reads counts with error**

$$\hat{f} = \frac{\sum_{i=1}^{N}(n_{(A,i)} - \varepsilon(n_{(A,i)} + n_{(G,i)}))}{\sum_{i=1}^{N}(n_{(A,i)} + n_{(G,i)})(1-2\varepsilon)} = 0.77$$

Estimating allele frequencies

Individual	True genotype	Reads allele A	Reads 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
Tot.		41	14

Simple allele frequency estimator: from **reads counts with error and weights** (Y Li et al. 2010)

$$p_i = \frac{n_{(A,i)} - \varepsilon(n_{(A,i)} + n_{(G,i)})}{(n_{(A,i)} + n_{(G,i)})(1 - 2\varepsilon)}$$

Weighting function

$$\hat{f} = \frac{1}{\sum_{i=1}^{N} w_i} \sum_{i=1}^{N} p_i w_i = 0.57$$

Estimating allele frequencies

Individual	True genotype	Reads allele A	Reads 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
Tot.		41	14

Maximum Likelihood (ML) estimator (Kim et al. 2011)

$$L = \prod_{i=1}^{N} p(D_i \mid f)$$

Estimating allele frequencies

Maximum Likelihood (ML) estimator (Kim et al. 2011)

$$L = \prod_{i=1}^{N} p(D_i \mid f)$$

$$p(D_i \mid f) = \sum_{g \in \{0,1,2\}} p(D \mid G = g) p(G = g \mid f)$$

Estimating allele frequencies

Maximum Likelihood (ML) estimator (Kim et al. 2011)

$$L = \prod_{i=1}^{N} p(D_i \mid f)$$

$$p(D_i \mid f) = \sum_{g \in \{0,1,2\}} p(D \mid G = g) p(G = g \mid f)$$

Estimating allele frequencies

Maximum Likelihood (ML) estimator (Kim et al. 2011)

$$L = \prod_{i=1}^{N} p(D_i \mid f)$$

Genotype likelihoods

$$p(D_i \mid f) = \sum_{g \in \{0,1,2\}} p(D \mid G = g) p(G = g \mid f)$$

Estimating allele frequencies

Maximum Likelihood (ML) estimator (Kim et al. 2011)

$$L = \prod_{i=1}^{N} p(D_i \mid f)$$

Genotype likelihoods

$$p(D_i \mid f) = \sum_{g \in \{0,1,2\}} p(D \mid G = g) p(G = g \mid f)$$

If we assume HWE:

$$p(G = AA \mid f) = f^2$$

$$p(G = AG \mid f) = 2f(1-f)$$

$$p(G = GG \mid f) = (1-f)^2$$

Estimating allele frequencies

Individual	True genotype	Reads allele A	Reads 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
Tot.		41	14

Maximum Likelihood (ML) estimator (Kim et al. 2011)

$$\hat{f} = \arg\max_p \prod_{i=1}^{N} p(D_i \mid f)$$

$$\hat{f} = 0.46$$

Workflow

Low-level data:

- Samples preparation + sequencing
- Call bases and quality scores

Genotype data:

- Call genotypes
- Estimate allele frequencies
- **SNPs detection**

Analysis:

- Population genetics analysis
- Association studies

SNP calling

- A lot of missing data if calling genotypes at low depth (heterozygotes can be lost!)
- Rare variants are hard to detect
- Trade-off between False Positives and False Negatives

SNP calling – effect of errors

Calling SNPs if 2 alternate alleles are observed (5X and 100 samples and error rate of 0.01):

False positive rate?

SNP calling – effect of errors

Calling SNPs if 2 alternate alleles are observed (5X and 100 samples and error rate of 0.01):

False positive rate? >99%

SNP calling – effect of errors

Calling SNPs if 2 alternate alleles are observed (5X and 100 samples and error rate of 0.01):

False positive rate? >99%

Heavy filtering of data (error rate of 0.001):

False positive rate?

SNP calling – effect of errors

Calling SNPs if 2 alternate alleles are observed (5X and 100 samples and error rate of 0.01):

False positive rate? >99%

Heavy filtering of data (error rate of 0.001):

False positive rate? 60%

Numbers from R. Nielsen

SNP calling

- What is the most straightforward method for SNP calling?

SNP calling

- What is the most straightforward method for SNP calling?
  - Assign as SNPs sites where at least one heterozygote has been called
  - …

SNP calling

- What is the most straightforward method for SNP calling?
  - Assign as SNPs sites where at least one heterozygote has been called
  - Assign as SNPs sites where the estimated allele frequency is above a certain threshold (e.g. ?)

SNP calling

- MLE of allele frequency at each site:

Call a SNP if

$$\hat{f}_{MLE} > t$$

Where *t* can be defined as the minimum sample allele frequency detectable (e.g. with 10 samples *t* can be set to 0.05)

Likelihood Ratio Test

- Compare the goodness of fit of the null and alternative model

- Null Model: frequency=0
- Alternative Model: frequency>0

The model with more parameters “tends” to fit better.

Whether or not this fit is “significantly better” is assessed by the comparison of the two likelihoods.

SNP calling

- Likelihood Ratio Test (**LRT**): test statistical hypotheses based on comparing the maximum likelihood under 2 different models.

$$T = -2\ln\left(\frac{L(f=0)}{L(f\neq 0)}\right)$$

T is chi-squared distributed with 1 degree of freedom -> assign a *p*-value

Workflow

Low-level data:

- Samples preparation + sequencing
- Call bases and quality scores

Genotype data:

- Call genotypes
- Estimate allele frequencies
- SNPs detection

Analysis:

- Population genetics analysis
- Association studies

Practical

- Basic filtering
- Estimation of allele frequencies and SNP calling
- Genotype calling
- Advanced methods to estimate SFS

Raghavan et al. 2015 Science