

# Variant Calling and Annotation

## Genomics Lecture #8/#9

Peter N Robinson

December 14, 2012

# Variant Calling

Variant  
Calling and  
Annotation

Peter N  
Robinson

Variant  
Calling

naive

MAP

MAQ

EM

SNVMix

Variant  
nomenclature

Annotation

Variant calling is an important procedure for whole-exome and whole-genome sequencing, and for some experiments also for RNA-seq.

Two major classes of variant

- Single-nucleotide variant (SNV)
- Structural variant

In this lecture, we will discuss issues and algorithms of SNV calling

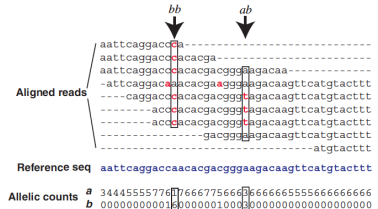
In the second half of the lecture, we will explain some of the issues and algorithms surrounding variant annotation.

## Germline variants

From read mapping, we get a series of aligned columns of nucleotides and have information about

- 1 mapping quality for each read
- 2 base call quality for each position
- 3 A stack of nucleotides

- $k$  wildtype nucleotides  $a$
- $n - k$  nucleotides  $b$
- $a, b \in \{a, c, g, t\}$  and  $a \neq b$



# Germline variants

Variant  
Calling and  
Annotation

Peter N  
Robinson

Variant  
Calling

naive

MAP

MAQ

EM

SNVMix

Variant  
nomenclature

Annotation

Ignoring for the moment everything but SNVs, our goal is to characterize each column of the alignment as homozygous wild-type ( $aa$ ), heterozygous ( $ab$ ), or homozygous variant ( $bb$ ).

# Germline variants

Variant  
Calling and  
Annotation

Peter N  
Robinson

Variant  
Calling

naive

MAP

MAQ

EM

SNVMix

Variant  
nomenclature

Annotation

Thus, if the true genotype is homozygous reference (a,a), and we observe  $k$  reference bases at such a position, then the remaining  $n - k$  bases must represent sequencing errors, and analogously for homozygous variant (b,b), positions

近似

True Genotype	Number of errors
a, a	$n - k$
b, b	$k$

If the true genotype is heterozygous, then we can approximate the probability of the genotype as

$$\text{dbinom}(n, k, p = 0.5) = \binom{n}{k} p^k (1 - p)^{n-k} = \binom{n}{k} \frac{1}{2^n}$$

# A naive algorithm

Variant  
Calling and  
Annotation

Peter N  
Robinson

Variant  
Calling

naive

MAP

MAQ

EM

SNVMix

Variant  
nomenclature

Annotation

Early NGS studies basically filtered base calls according to quality and then used a frequency filter.

Typically, a quality filter of PHRED Q20 was used (i.e., probability of error 1% ). Then, the following frequency thresholds were used according to the frequency of the non-ref base,  $f(b)$ :

$f(b)$	genotype call
$[0, 0.2)$	homozygous reference
$[0.2, 0.8]$	heterozygous
$(0.8, 1]$	homozygous variant

# A naive algorithm

Variant  
Calling and  
Annotation

Peter N  
Robinson

Variant  
Calling

naive

MAP

MAQ

EM

SNVMix

Variant  
nomenclature

Annotation

The frequency heuristic works well if the sequencing depth is high, so that the probability of a heterozygous nucleotide falling outside of the 20% – 80% region is low.

Problems with frequency heuristic:

- For low sequencing depth, leads to undercalling of heterozygous genotypes
- Use of quality threshold leads to loss of information on individual read/base qualities
- Does not provide a measure of confidence in the call

# A naive algorithm

Variant  
Calling and  
Annotation

Peter N  
Robinson

Variant  
Calling

naive

MAP

MAQ

EM

SNVMix

Variant  
nomenclature

Annotation

For these reasons, a number of probabilistic methods have been developed.

We will discuss two of them and provide some algorithmic background.

- MAQ: An early algorithm.
- SNVMix: A more flexible Bayesian algorithm

The MAQ SNV calling algorithm makes use of the MAP formalism, which will be explained in the following.



# MAP (Short review)

Variant  
Calling and  
Annotation

Peter N  
Robinson

Variant  
Calling

naive

MAP

MAQ

EM

SNVMix

Variant  
nomenclature

Annotation

Bayes' theorem follows from the definition of the conditional probability and relates the conditional probability  $P(A|B)$  to  $P(B|A)$  for two events  $A$  and  $B$  such that  $P(B) \neq 0$ :

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}. \quad (1)$$

# Bayes Theorem

Variant  
Calling and  
Annotation

Peter N  
Robinson

Bayes' theorem follows from the definition of the conditional probability and relates the conditional probability  $P(A|B)$  to  $P(B|A)$  for two events  $A$  and  $B$  such that  $P(B) \neq 0$ :

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

- posterior
- likelihood
- prior
- normalization constant

Variant  
Calling

naive

MAP

MAQ

EM

SNVMix

Variant  
nomenclature

Annotation

# Bayes Theorem

Variant  
Calling and  
Annotation

Peter N  
Robinson

Variant  
Calling

naive

MAP

MAQ

EM

SNVMix

Variant  
nomenclature

Annotation

Bayes' theorem is often used for a set of  $n$  mutually exclusive events  $E_1, E_2, \dots, E_n$  such that  $\sum_i P(E_i) = 1$ . Then, we have

$$P(E_i|B) = \frac{P(B|E_i)P(E_i)}{\sum_i P(B|E_i)P(E_i)}. \quad (2)$$

- This form of Bayes' theorem makes it clear why  $B = \sum_i P(B|E_i)P(E_i)$  is called the normalization constant, because it forces the sum of all  $P(E_i|B)$  to be equal to one, thus making  $P(\cdot|B)$  a real probability measure

# Bayes Theorem

Variant  
Calling and  
Annotation

Peter N  
Robinson

Variant  
Calling

naive

MAP

MAQ

EM

SNVMix

Variant  
nomenclature

Annotation

In the context of bioinformatics, Bayesian inference is often used to identify the most likely model: For instance, we observe a DNA sequence and would like to know if it is a gene ( $M_1$ ) or not ( $M_2$ ).

Often, the model is symbolized by  $M$  and the observed data by  $D$ . Then, Bayes' theorem can be given as:

$$P(M_1|D) = \frac{P(D|M_1)P(M_1)}{P(D|M_1)P(M_1) + P(D|M_2)P(M_2)} \quad (3)$$

# maximum a posteriori (MAP)

Variant  
Calling and  
Annotation

Peter N  
Robinson

Variant  
Calling

naive

MAP

MAQ

EM

SNVMix

Variant  
nomenclature

Annotation

In Bayesian statistics, maximum a posteriori (MAP) estimation is often used to generate an estimate of the maximum value of a probability distribution.

That is, if  $x$  is used to refer to the data ( $x$  can be an arbitrary expression), and  $\theta$  is used to refer to the parameters of a model, then Bayes' law states that:

$$P(\theta|x) = \frac{P(x|\theta)P(\theta)}{P(x)} \quad (4)$$

# maximum a posteriori (MAP)

Variant  
Calling and  
Annotation

Peter N  
Robinson

Variant  
Calling

naive

MAP

MAQ

EM

SNVMix

Variant  
nomenclature

Annotation

The term  $P(\theta|x)$  is referred to as the posterior probability, and specifies the probability of the parameters  $\theta$  given the observed data  $x$ . The denominator on the right-hand side can be regarded as a normalizing constant that does not depend on  $\theta$ , and so it can be disregarded for the maximization of  $\theta$ .

The MAP estimate of  $\theta$  is defined as:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} P(\theta|x) = \underset{\theta}{\operatorname{argmax}} P(x|\theta)P(\theta) \quad (5)$$

# maximum a posteriori (MAP)

Variant  
Calling and  
Annotation

Peter N  
Robinson

Variant  
Calling

naive

MAP

MAQ

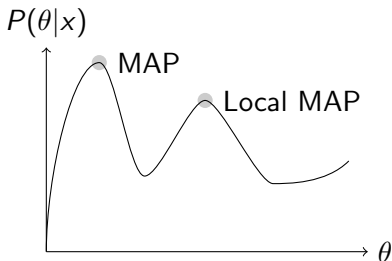
EM

SNVMix

Variant  
nomenclature

Annotation

One important issue about MAP estimation procedures (that we will not discuss further here), is that they tend to have the disadvantage that they “get stuck” in local maxima without being able to offer a guarantee of finding the global maximum.



# MAQ

Variant  
Calling and  
Annotation

Peter N  
Robinson

Variant  
Calling

naive

MAP

MAQ

EM

SNVMix

Variant  
nomenclature

Annotation

One of the first widely used read mappers and variant callers

Li H, Ruan J, Durbin R (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores  
Genome Research 18:1851–1858

- MAQ uses a number of interesting heuristics for read mapping and variant calling
- MAQ calls the genotype that maximizes the posterior probability



# MAQ

Variant  
Calling and  
Annotation

Peter N  
Robinson

Variant  
Calling

naive

MAP

MAQ

EM

SNVMix

Variant  
nomenclature

Annotation

We will not review the entire MAQ mapping algorithm, but just those parts that are relevant to the variant calling process

The Mapping Quality for the assigned alignment of a read  $s$  is denoted as  $Q_s$ , the PHRED-scaled probability that the read alignment is wrong.

$$Q_s = -10 \log_{10} \Pr[\text{read is wrongly mapped}]$$

For example  $Q_s = 30$  implies there is a 1:1000 probability that the read  $s$  has been wrongly mapped,  $Q_s = 20$  implies a 1:100 probability, and so on.

# FASTQ and PHRED-like Quality Scores

Variant  
Calling and  
Annotation

Peter N  
Robinson

Variant  
Calling

naive

MAP

MAQ

EM

SNVMix

Variant  
nomenclature

Annotation

Recall from lecture #1:

- Illumina sequences are reported in FASTQ format.

```
@My-Illu:6:73:941:1973#0/1
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTGTTCAACTCACAGTTT
+
!''*(((((***+))%%%+)) (%%%) . 1***-+*''))**55CCF>>>>>CCCCCCC65
```

- 1 Read identifier
- 2 sequence reported by the machine
- 3 '+' (can optionally include a sequence description)
- 4 ASCII encoded base quality scores

# PHRED Quality Scores

Variant  
Calling and  
Annotation

Peter N  
Robinson

- The PHRED quality score is defined as

$$Q_{PHRED} = -10 \log_{10} p$$

where  $p$  is the probability that the corresponding base call is **wrong**.

- The PHRED quality score is nothing more than a simple transformation.

$Q_{PHRED}$	$p$	Accuracy
10	$10^{-1}$	90%
20	$10^{-2}$	99%
30	$10^{-3}$	99.9%
40	$10^{-4}$	99.99%
50	$10^{-5}$	99.999%

Variant  
Calling

naive

MAP

MAQ

EM

SNVMix

Variant  
nomenclature

Annotation

# MAQ: Mapping quality

Variant  
Calling and  
Annotation

Peter N  
Robinson

Variant  
Calling

naive

MAP

MAQ

EM

SNVMix

Variant  
nomenclature

Annotation

- MAQ uses a system of templates and hashing to index the first 28 bp of each read
- Each 28 bp sequence of the reference is similarly hashed/indexed. If a hit is found to a read MAQ calculates the sum of qualities of mismatched bases  $q$
- Because of the templates/hash procedure, all reads with not more than two mismatches in the first 28 bp are found (pigeon hole principle)
- For simplicity, consider here quality values for an ungapped alignment of the entire read including the part of the read after the first 28 bp

# MAQ: Mapping quality

Variant  
Calling and  
Annotation

Peter N  
Robinson

We consider the probability that a read  $z$  comes from position  $u$  of a reference sequence  $\mathcal{R}$

$$p(z|\mathcal{R}, u) = \prod_{i \in \text{mismatched bases}} 10^{-\frac{q_i}{10}} = 10^{-\frac{\sum_i q_i}{10}}$$

That is, the probability that a read  $z$  comes from position  $u$  of reference sequence  $\mathcal{R}$  is modeled as the product of the PHRED quality scores for each of the bases that are mismatched in the alignment.

For instance, if the alignment at position  $u$  has one mismatch with PHRED base quality 20 and one with PHRED quality 10, then

$$p(z|\mathcal{R}, u) = 10^{-\frac{20+10}{10}} = 10^{-3} = 0.001$$

Variant  
Calling

naive

MAP

MAQ

EM

SNVMix

Variant  
nomenclature

Annotation

# MAQ: Mapping quality

Variant  
Calling and  
Annotation

Peter N  
Robinson

Variant  
Calling

naive

MAP

MAQ

EM

SNVMix

Variant  
nomenclature

Annotation

We now calculate the posterior probability of the mapping at position  $u$ ,  $p_s(u|\mathcal{R}, z)$  using Bayes law

$$p_s(u|\mathcal{R}, z) = \frac{p(z|\mathcal{R}, u)p(u|\mathcal{R})}{\sum_v p(z|\mathcal{R}, v)p(v|\mathcal{R})}$$

If we assume a uniform prior distribution  $p(u|\mathcal{R})$ , then the read is equally likely to begin at any position of the reference. The sum then goes over all positions from 1 to  $L - |z| + 1$ , where  $L$  is the length of  $\mathcal{R}$  and  $|z|$  is the length of a read.

$$p_s(u|\mathcal{R}, z) = \frac{p(z|\mathcal{R}, u)}{\sum_{v=1}^{L-|z|+1} p(z|\mathcal{R}, v)}$$

# MAQ: Mapping quality

Variant  
Calling and  
Annotation

Peter N  
Robinson

Variant  
Calling

naive

MAP

MAQ

EM

SNVMix

Variant  
nomenclature

Annotation

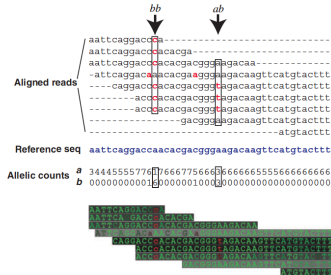
Finally, we can get the PHRED scaled mapping quality, reflecting the probability that the read is wrongly mapped

$$Q_s(u|\mathcal{R}, z) = -10 \log_{10}[1 - p_s(u|\mathcal{R}, z)]$$

Note that the explicit calculation of  $Q_s$  would require summing over the entire reference sequence, which would be impractical. MAQ therefore uses a **heuristic** to estimate  $Q_s$ , which will not be discussed further here.

If a read is incorrectly mapped, then genotypes inferred from that read may not be reliable. Therefore, MAQ does not allow base quality  $q_i$  to exceed the mapping quality of its read  $z$ .

$$q_i = \min(q_i, Q_s(u|\mathcal{R}, z))$$



Darker background  $\Leftrightarrow$  higher mapping quality. Higher contrast  $\Leftrightarrow$  higher base quality



# MAQ: Consensus Genotype Calling

Variant  
Calling and  
Annotation

Peter N  
Robinson

Variant  
Calling

naive

MAP

MAQ

EM

SNVMix

Variant  
nomenclature

Annotation

MAQ uses the quality values from mapping and of individual bases to call the most likely genotype. We assume we have a column of an alignment with  $k$  reference bases  $a$  and  $n - k$  variant bases  $b$ . Any other bases are ignored as being probably sequencing errors

True Genotype	# errors	Cond. Prob. of Genotype
$a, a$	$n - k$	$\alpha_{n, n-k}$
$b, b$	$k$	$\alpha_{n, k}$
$a, b$	$?$	$\binom{n}{k} \frac{1}{2^n}$

# MAQ: Consensus Genotype Calling

Variant  
Calling and  
Annotation

Peter N  
Robinson

Variant  
Calling

naive

MAP

MAQ

EM

SNVMix

Variant  
nomenclature

Annotation

The goal is now to decide which of the three possible genotypes has the highest posterior probability given the data (the mapping and alignment):  $p(g|D)$ .

MAQ now assumes the prior for the genotypes is

- $P(\langle a, a \rangle) = (1 - r)/2$
- $P(\langle b, b \rangle) = (1 - r)/2$
- $P(\langle a, b \rangle) = r$

Here,  $r$  is the probability of observing a heterozygous genotype. MAQ uses  $r = 0.001$  for new SNPs, and 0.2 for known SNPs, but site-specific values for  $r$  could also be used.

# MAQ: Consensus Genotype Calling

Variant  
Calling and  
Annotation

Peter N  
Robinson

Variant  
Calling

naive

MAP

**MAQ**

EM

SNVMix

Variant  
nomenclature

Annotation

MAQ thus calls the genotypes as

$$\hat{g} = \operatorname{argmax}_{g \in (\langle a, a \rangle, \langle a, b \rangle, \langle b, b \rangle)} p(g|D)$$

The quality of this genotype call can then be calculated as

$$Q_g = -10 \log_{10}[1 - P(\hat{g}|D)]$$

# MAQ: Consensus Genotype Calling

Variant  
Calling and  
Annotation

Peter N  
Robinson

Variant  
Calling

naive

MAP

MAQ

EM

SNVMix

Variant  
nomenclature

Annotation

We now need a way of calculating  $\alpha_{n,k}$ , the probability of observing  $k$  errors in  $n$  nucleotides in the alignment. If we assume that error rates arise independently, and error rates are identical for all bases, then we can use a binomial distribution

$$\text{dbinom}(n, k, p = \epsilon) = \binom{n}{k} \epsilon^k (1 - \epsilon)^{n-k}$$

For instance, the probability of observing 2 erroneous nucleotides in 20, if the per read error rate is  $\epsilon = 0.01$  can be calculated in R as

```
> dbinom(2,20,0.01)
[1] 0.01585576
```

# MAQ: Consensus Genotype Calling

Variant  
Calling and  
Annotation

Peter N  
Robinson

Variant  
Calling

naive

MAP

MAQ

EM

SNVMix

Variant  
nomenclature

Annotation

In practice, MAQ errors are correlated and are not identical for each base in the alignment. Therefore, MAQ does not use a binomial distribution, but a heuristic that reflects the probabilities of observing an alignment with the given pattern of per base error probabilities.

$$\alpha_{n,k} = c'_{n,k} \prod_{i=0}^{k-1} \epsilon_{i+1}^{\theta^i}$$

Here,  $\epsilon_i$  is the  $i^{\text{th}}$  smallest base error probability for the  $k$  observed errors,  $c'_{n,k}$  is a constant and  $\theta$  is a parameter that controls the dependency of errors.

This equation reflects the base errors. We will not go into further detail, but if desired see the Supplemental material of the MAQ paper.

# MAQ: Consensus Genotype Calling

Variant  
Calling and  
Annotation

Peter N  
Robinson

Variant  
Calling

naive

MAP

MAQ

EM

SNVMix

Variant  
nomenclature

Annotation

With all of this, we can now call the posterior probabilities of the three genotypes given the data  $D$ , that is a column with  $n$  aligned nucleotides and quality scores of which  $k$  correspond to the reference  $a$  and  $n - k$  to a variant nucleotide  $b$ .

$$\begin{aligned} p(G = \langle a, a \rangle | D) &\propto p(D | G = \langle a, a \rangle) p(G = \langle a, a \rangle) \\ &\propto \alpha_{n,k} \cdot (1 - r) / 2 \end{aligned}$$

$$\begin{aligned} p(G = \langle b, b \rangle | D) &\propto p(D | G = \langle b, b \rangle) p(G = \langle b, b \rangle) \\ &\propto \alpha_{n,n-k} \cdot (1 - r) / 2 \end{aligned}$$

$$\begin{aligned} p(G = \langle a, b \rangle | D) &\propto p(D | G = \langle a, b \rangle) p(G = \langle a, b \rangle) \\ &\propto \binom{n}{k} \frac{1}{2^n} \cdot r \end{aligned}$$

# MAQ: Consensus Genotype Calling

Variant  
Calling and  
Annotation

Peter N  
Robinson

Variant  
Calling

naive

MAP

MAQ

EM

SNVMix

Variant  
nomenclature

Annotation

Finally, the genotype with the highest posterior probability is chosen

$$\hat{g} = \operatorname{argmax}_{g \in (\langle a, a \rangle, \langle a, b \rangle, \langle b, b \rangle)} p(g|D)$$

The probability of this genotype is used as a measure of confidence in the call.

# What have we learned?

Variant  
Calling and  
Annotation

Peter N  
Robinson

Variant  
Calling

naive

MAP

MAQ

EM

SNVMix

Variant  
nomenclature

Annotation

The MAQ algorithm is typical for many in genomics in that a well known statistical or algorithmic framework is used with a number of heuristics that deliver reasonable values for the parameters needed for the framework to work.

Major aspects of MAQ SNV calling algorithm

- Integrates mapping and per base quality scores
- Bayesian (MAP) framework to integrate observations and priors on genotypes
- Provides estimation of reliability of genotype call.



# Expectation Maximization (EM)

Variant  
Calling and  
Annotation

Peter N  
Robinson

Variant  
Calling

naive

MAP

MAQ

EM

SNVMix

Variant  
nomenclature

Annotation

We will discuss how EM is used for mixture distributions. For ease of presentation, we will discuss in detail a mixture of Gaussians, but the principles are the same for other probability distributions

The basic framework is that we assume that a data point  $y_j$  is produced as follows

- First, choose one of  $i \in \{1, \dots, I\}$  components that produces the measurement
- Then, according to the parameters of component  $C = i$ , the actual measurement is generated

This is known as a mixture distribution, and the corresponding probability density function (pdf) is defined as

$$p(y_j|\theta) = \sum_{i=1}^I \alpha_i p(y_j|C=i, \beta_i) \quad (6)$$

# Expectation Maximization (EM)

Variant  
Calling and  
Annotation

Peter N  
Robinson

Variant  
Calling

naive

MAP

MAQ

EM

SNVMix

Variant  
nomenclature

Annotation

Note that in this notation,  $\theta$  comprises both the weight parameters for the probability of one of the  $I$  components generating the data, as well as the various parameters for each of the components,  $\beta_i$  (where in general  $\beta_i$  can be a vector of parameters)

- Of course,  $\sum_{i=1}^I \alpha_i = 1$
- The parameters  $\beta_i$  are associated with the PDF of component  $i$ .

We will now show how to perform maximum likelihood estimation using the Expectation Maximization (EM) framework to find values for the parameters  $\theta$  that maximize the probability of the data. This involves maximization of the log-likelihood for  $\theta$ .

$$\log L(\theta) = \log p(\mathbf{y}|\theta) \quad (7)$$

# Expectation Maximization (EM)

Variant  
Calling and  
Annotation

Peter N  
Robinson

Variant  
Calling

naive

MAP

MAQ

EM

SNVMix

Variant  
nomenclature

Annotation

We can flesh out the formula as follows

$$\begin{aligned}\log L(\theta) &= \log p(\mathbf{y}|\theta) \\ &= \log \left\{ \prod_{j=1}^J p(y_j|\theta) \right\} \\ &= \sum_{j=1}^J \log p(y_j|\theta) \\ &= \sum_{j=1}^J \log \left\{ \sum_{i=1}^I \alpha_i p(y_j|C=i, \beta_i) \right\}\end{aligned}$$

- Since the log is outside the sum in the last expression, there is no analytic (closed form) optimization.

# Expectation Maximization (EM)

Variant  
Calling and  
Annotation

Peter N  
Robinson

Variant  
Calling

naive

MAP

MAQ

EM

SNVMix

Variant  
nomenclature

Annotation

The EM algorithm is essentially like a pushme-pullyou algorithm that goes back and forth between

- find an estimate for the likelihood function
- maximizing the whole term



# Expectation Maximization (EM)

Variant  
Calling and  
Annotation

Peter N  
Robinson

Variant  
Calling

naive

MAP

MAQ

EM

SNVMix

Variant  
nomenclature

Annotation

If there is time in the practical session, I will explain the derivation of the EM method and show in detail how the maximization expressions are derived for a simple distribution – mixture of Gaussians. For today, I will show only a high level summary. For the practical, you will be expected to implement a simplified version of EM – known as gene counting (will explain at end of this lecture)

# Expectation Maximization (EM)

Variant  
Calling and  
Annotation

Peter N  
Robinson

Variant  
Calling

naive

MAP

MAQ

EM

SNVMix

Variant  
nomenclature

Annotation

The EM algorithm will be explained using a mixture of Gaussians (SNVMix uses some slightly less familiar distributions). Recall the form of a multivariate Gaussian distribution for a  $k$ -dimensional vector  $\mathbf{x} = [x_1, x_2, \dots, x_k]$ :

$$\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \quad (8)$$

i.e. the probability density function is

$$f(x_1, x_2, \dots, x_k) = \frac{1}{(2\pi)^{k/2} |\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})} \quad (9)$$

# Expectation Maximization (EM)

Variant  
Calling and  
Annotation

Peter N  
Robinson

Variant  
Calling

naive

MAP

MAQ

EM

SNVMix

Variant  
nomenclature

Annotation

We thus wish to estimate the parameters for a mixture of Gaussians. We need to estimate both the mixture parameters  $\pi_1, \pi_2, \dots, \pi_c$  with  $\sum_{i=1}^c \pi_i = 1$ , but also the means and variances for each of the individual Gaussian distributions,  $\mu_1, \mu_2, \dots, \mu_c$ , and  $\Sigma_1, \Sigma_2, \dots, \Sigma_c$ .

We thus want to maximize the log likelihood given by

$$L(\theta | \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n) = \log \prod_{i=1}^n \sum_{k=1}^c \pi_k f(\mathbf{x}_i | \mu_k, \Sigma_k) \quad (10)$$

or equivalently

$$L(\theta | \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n) = \sum_{i=1}^n \log \sum_{k=1}^c \pi_k f(\mathbf{x}_i | \mu_k, \Sigma_k) \quad (11)$$

# Expectation Maximization (EM)

Variant  
Calling and  
Annotation

Peter N  
Robinson

Variant  
Calling

naive

MAP

MAQ

EM

SNVMix

Variant  
nomenclature

Annotation

We can now calculate the probability that a particular data point  $\mathbf{x}_j$  belongs to a particular component  $k$

We write the posterior probability that an observation  $\mathbf{x}_j$  belongs to component  $k$  as

$$\hat{\tau}_{jk} = \frac{f(\mathbf{x}_j | \mu_k, \Sigma_k) \hat{\pi}_k}{\sum_{i=1}^c f(\mathbf{x}_j | \mu_i, \Sigma_i) \hat{\pi}_i} \quad (12)$$

- The posterior probability  $\hat{\tau}_{jk}$  is unknown but can be easily **estimated** if we use the current values of the parameters for the Gaussians.
- $\hat{\tau}_{jk}$  is thus an estimate for the probability that observation  $j$  was generated by component  $k$  given the data and current parameter estimates
- This is the **Expectation** step of the EM algorithm



# Expectation Maximization (EM)

Variant  
Calling and  
Annotation

Peter N  
Robinson

Variant  
Calling

naive

MAP

MAQ

EM

SNVMix

Variant  
nomenclature

Annotation

Given our current estimates of the component membership for each of the datapoints, we can maximize the values of the mixture parameters as well as of the Gaussians by setting their first derivative to zero etc (individual steps not shown here). This leads to the following

$$\hat{\pi}_k = \frac{1}{n} \sum_{i=1}^n \hat{\tau}_{ik} \quad (13)$$

$$\hat{\mu}_k = \frac{1}{n} \sum_{i=1}^n \frac{\hat{\tau}_{ik} \mathbf{x}_i}{\hat{\pi}_k} \quad (14)$$

$$\hat{\Sigma}_k = \frac{1}{n} \sum_{i=1}^n \frac{\hat{\tau}_{ik} (\mathbf{x}_i - \hat{\mu}_k) (\mathbf{x}_i - \hat{\mu}_k)^T}{\hat{\pi}_k} \quad (15)$$

This is the **Maximization** step of the EM algorithm

# Expectation Maximization (EM)

Variant  
Calling and  
Annotation

Peter N  
Robinson

Variant  
Calling

naive

MAP

MAQ

EM

SNVMix

Variant  
nomenclature

Annotation

Thus, the individual steps of the EM algorithm are thus

- ➊ Initial component parameters (with a reasonable guess)
- ➋ For each data point, calculate posterior probability of membership to each component using the current parameter values
- ➌ Then, based on these estimates, maximize the log likelihood of the parameters given the data
- ➍ Repeat until convergence<sup>1</sup>

---

<sup>1</sup>and hope you have not landed in a local maximum.

# SNVMix

Variant  
Calling and  
Annotation

Peter N  
Robinson

Variant  
Calling

naive

MAP

MAQ

EM

SNVMix

Variant  
nomenclature

Annotation

We will now discuss an algorithm called SNVMix, that uses the EM framework to estimate optimal parameters for calling SNPs in a Bayesian framework.

- This algorithm was first described here:  
Shah SP et al. (2009) Mutational evolution in a lobular breast tumor profiled at single nucleotide resolution.  
*Nature* **461**:809-13.
- an improved version (which we will not discuss) was later presented here:  
Goya R et al. (2010) SNVMix: predicting single nucleotide variants from next-generation sequencing of tumors.  
*Bioinformatics* **26**:730-6.

# SNVMix: Model specification

Variant  
Calling and  
Annotation

Peter N  
Robinson

Variant  
Calling

naive

MAP

MAQ

EM

SNVMix

Variant  
nomenclature

Annotation

The core of the model is a specification of the genotypes and the conditional probabilities of the observed distribution of alleles – similar to MAQ.

- Let  $G_i = k, k \in \{\langle a, a \rangle, \langle a, b \rangle, \langle b, b \rangle\}$  be a multinomial random variable representing the genotype at nucleotide position  $i$  ( $a = \text{ref}, b$  is non-ref).
- Let the observed allele frequency  $X_i = [a_i, b_i]^T$ , i.e., a vector of counts of the reference and non-reference alleles at position  $i$
- Then  $N_i = a_i + b_i$  is the observed read depth at position  $i$ .

# SNVMix: Model specification

Variant  
Calling and  
Annotation

Peter N  
Robinson

Variant  
Calling

naive

MAP

MAQ

EM

SNVMix

Variant  
nomenclature

Annotation

The central idea is that we assume the allele counts are generated by a class conditional density

Thus

$$X_i \sim \text{Binom}(a_i | \mu_k, N_i) \quad (16)$$

The probability of the observed read counts (number of  $a_i$  bases amongst all  $N_i$  bases at position  $i$ ) is thus conditioned on the underlying genotype  $G_i = k$ , and  $\mu_k$  is the corresponding parameter of a Binomial distribution for genotype  $k$ .

# SNVMix: Model specification

Variant  
Calling and  
Annotation

Peter N  
Robinson

Variant  
Calling

naive

MAP

MAQ

EM

SNVMix

Variant  
nomenclature

Annotation

If we actually knew the genotype, then it is simplicissimo to calculate the probability of the allele counts using the binomial distribution.

Thus

$$P(X_i) = \text{Binom}(a_i | \mu_k, N_i) = \binom{N_i}{a_i} \mu_k^{a_i} (1 - \mu_k)^{N_i - a_i} \quad (17)$$

Intuitively, we would expect the values of  $\mu_{aa}$  to be close to 1, those for  $\mu_{ab}$  to be close to 0.5 and those for  $\mu_{bb}$  to be near zero. However, we do not know the exact values for real data, which may depend on things such as the sequencing error rate<sup>2</sup>.

---

<sup>2</sup> And for cancer data, on the relative mixture of normal and cancerous tissue in a biopsy. ▶

# SNVMix: Model specification

Variant  
Calling and  
Annotation

Peter N  
Robinson

Variant  
Calling

naive

MAP

MAQ

EM

**SNVMix**

Variant  
nomenclature

Annotation

The prior probability of observing a genotype  $k$  at any position of the sequenced genome is represented as a multinomial variable  $\pi$ .

- $0 \leq \pi_k \leq 1, \quad \forall k$
- $\sum_{k=1}^3 \pi_k = 1$
- Note that in general we will expect the values of  $\pi$  to be highly skewed towards observing homozygous reference bases (since most genomic positions are not variant in any one individual)
- SNVMix is thus a classic generative mixture model to explain the observed data.

# SNVMix: Model specification

Variant  
Calling and  
Annotation

Peter N  
Robinson

Variant  
Calling

naive

MAP

MAQ

EM

SNVMix

Variant  
nomenclature

Annotation

The marginal distribution of  $X_i$  (in which we have marginalized – removed – the influence of the actual genotype) can then be calculated as the convex combination of the class conditional Binomial densities, weighted by the multinomial  $\pi$ :

$$p(X_i) = \sum_{k=1}^3 \pi_k \binom{N_i}{a_i} \mu_k^{a_i} (1 - \mu_k)^{N_i - a_i} \quad (18)$$

Again, the sum is taken over  $k$  representing the three genotypes  $aa$ ,  $ab$ , and  $bb$ .



# SNVMix: Model specification

Variant  
Calling and  
Annotation

Peter N  
Robinson

Variant  
Calling

naive

MAP

MAQ

EM

SNVMix

Variant  
nomenclature

Annotation

We can then use this equation to calculate the log likelihood of our entire dataset, which comprises positions  $1 \dots T$ .

$$\log p(X_{1:T} | \mu_{1:K}, \pi) = \sum_{i=1}^T \log \sum_{k=1}^3 \pi_k \binom{N_i}{a_i} \mu_k^{a_i} (1 - \mu_k)^{N_i - a_i} \quad (19)$$

Our problem is that the model parameters  $\theta = (\pi, \mu)$  are not known. If the true genotype were somehow known, we could simply calculate them from the training data. But, instead, we will learn (estimate) the parameters from data by using maximum a posteriori (MAP) expectation maximization (EM).

# SNVMix: Model specification

Variant  
Calling and  
Annotation

Peter N  
Robinson

Variant  
Calling

naive

MAP

MAQ

EM

SNVMix

Variant  
nomenclature

Annotation

Assuming we have solved for the parameters (we will get to that shortly), then we can easily calculate the posterior probability of any genotype using Bayes rule

$$p(G_i = k | X_{1:N}, \pi, \mu_k) = \frac{\pi_k \text{Binom}(X_i | \mu_k, N_i)}{\sum_{j=1}^3 \pi_j \text{Binom}(X_i | \mu_j, N_i)} \quad (20)$$

For notational simplicity, we will denote  $p(G_i = k | X_{1:N}, \pi, \mu_k)$  as  $\gamma_i(k)$ , the marginal probability of the genotype for position  $i$  given all the data and the model parameters.

# SNVMix: Prior Distributions

Variant  
Calling and  
Annotation

Peter N  
Robinson

Variant  
Calling

naive

MAP

MAQ

EM

SNVMix

Variant  
nomenclature

Annotation

Bayesian mixture models use hyperparameters, i.e., parameters of a prior distribution; the term is used to distinguish them from parameters of the model used for the final analysis. We will use two underlying distributions to calculate these hyperparameters for SNVMix.

- $\pi \sim \text{Dirichlet}(\pi|\delta)$
- $\mu \sim \text{Beta}(\mu_k|\alpha_k, \beta_k)$

# Beta distribution

Variant  
Calling and  
Annotation

Peter N  
Robinson

Variant  
Calling

naive

MAP

MAQ

EM

SNVMix

Variant  
nomenclature

Annotation

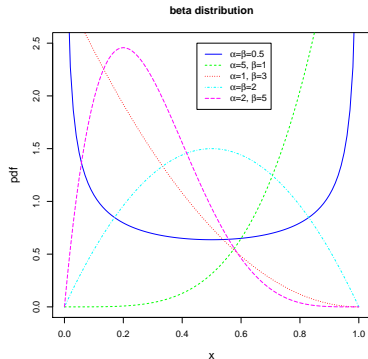
Beta distribution: A family of continuous distributions defined on  $[0, 1]$  and parametrized by two positive shape parameters,  $\alpha$  and  $\beta$

$$p(x) = \frac{1}{B(\alpha, \beta)} \cdot x^{\alpha-1} (1-x)^{\beta-1}$$

here,  $x \in [0, 1]$ , and

$$B(\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha) \cdot \Gamma(\beta)}$$

where  $\Gamma$  is the Gamma function  
(extension of factorial).



# Beta distribution

Variant  
Calling and  
Annotation

Peter N  
Robinson

Variant  
Calling

naive

MAP

MAQ

EM

SNVMix

Variant  
nomenclature

Annotation

Important in the current context is that the Beta distribution is the conjugate prior to the binomial distribution.

- That is, we can express our prior belief about the value of the  $\mu_k$  parameter of the binomial distribution for read counts using a Beta distribution
- We say that  $\mu_k$  is conjugately distributed according to a Beta distribution:  $\mu_k \sim \text{Beta}(\mu_k | \alpha_k, \beta_k)$ .
- This requires us to express our prior belief about  $\mu_{aa}, \mu_{ab}$ , and  $\mu_{bb}$  by specifying values for  $\alpha_k, \beta_k$

# Beta distribution

Variant  
Calling and  
Annotation

Peter N  
Robinson

Variant  
Calling

naive

MAP

MAQ

EM

SNVMix

Variant  
nomenclature

Annotation

For instance, let us say we are not very sure about what we think the value of  $\mu_{a,b}$  should be, perhaps because we do not know if the sample being sequenced contains tumorous or non-tumorous tissues<sup>3</sup>.

- We might then try  $\alpha_{ab} = \beta_{ab} = 3$

```
x <- seq(0.0, 1.0, 0.01)
y <- dbeta(x, 3, 3)
title <- expression(paste(alpha,"=",beta,"=3"))
plot(x, y, type="l",main=title,
xlab="x",ylab="pdf",col="blue",lty=1,cex.lab=1.25)
```

---

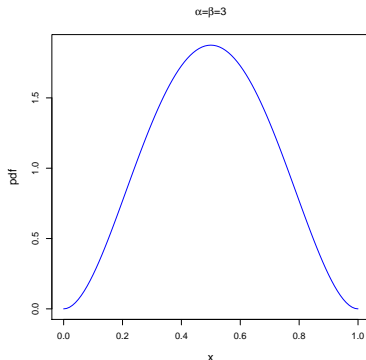
<sup>3</sup>Tumor tissue may be characterized by the loss of heterozygosity (LOH) of large chromosomal regions.


# Beta distribution

Variant  
Calling and  
Annotation

Peter N  
Robinson

$$\alpha_{ab} = \beta_{ab} = 3$$



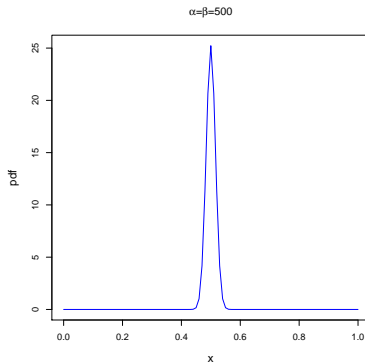
- $x$  is here representing a value for  $\mu_{ab}$ , and the  $y$  axis reflects our belief about the prior probability of this value
- Question: Are we very sure about  $\mu_{ab}$ ? 

# Beta distribution

Variant  
Calling and  
Annotation

Peter N  
Robinson

$$\alpha_{ab} = \beta_{ab} = 500.$$



- Question: How sure are we now about  $\mu_{ab}$ ?



# Beta distribution in SNVMix

Variant  
Calling and  
Annotation

Peter N  
Robinson

Variant  
Calling

naive

MAP

MAQ

EM

SNVMix

Variant  
nomenclature

Annotation

The expected value of a  $\text{Beta}(\alpha + \beta)$  distribution is simply

$$\frac{\alpha}{\alpha + \beta}$$

In SNVMix, these values are defined as

- $\alpha_{aa} = 1000$ ,  $\beta_{aa} = 1$ , that is, our prior belief in reference reads given homozygous reference sequence is 0.999001
- $\alpha_{ab} = 500$ ,  $\beta_{ab} = 500$ , that is, our prior belief in reference reads given a het true sequence is 0.5
- $\alpha_{bb} = 1$ ,  $\beta_{bb} = 1000$  (vice versa to  $\alpha_{aa} = 1000$ ,  $\beta_{aa} = 1$ )

# SNVMix: M-step updating equation for $\mu$

Variant  
Calling and  
Annotation

Peter N  
Robinson

The maximization step updating equation basically adds to observed counts for a certain true genotype to our prior.

$$\mu_k^{new} = \frac{\sum_{i=1}^T a_i^{I(G_i=k)} + \alpha_k}{\sum_{i=1}^T N_i^{I(G_i=k)} + \alpha_k + \beta_k - 2} \quad (21)$$

- Note that in this notation,  $I(G_i = k)$  is an indicator function so that the expression is zero unless  $G_i = k$
- The update is simply the proportion of the observed reference reads with “pseudocounts” added from the Beta prior (amongst all positions called to genotype  $k$ ).

Variant  
Calling

naive

MAP

MAQ

EM

SNVMix

Variant  
nomenclature

Annotation

# Dirichlet

Variant  
Calling and  
Annotation

Peter N  
Robinson

Variant  
Calling

naive

MAP

MAQ

EM

SNVMix

Variant  
nomenclature

Annotation

The Dirichlet distribution is the multivariate generalization of the beta distribution and represent the conjugate prior of the multinomial distribution. Thus, just as SNVMix used the Beta distribution as a prior for  $\mu$  (binomial distribution of read counts), it uses the Dirichlet as a prior for  $\pi$  (multinomial distribution for the three possible genotypes).

In SNVMix, the values for the prior are set to

$\delta(\langle a, a \rangle)$	$\delta(\langle b, b \rangle)$	$\delta(\langle b, b \rangle)$
1000	100	100

# Dirichlet

Variant  
Calling and  
Annotation

Peter N  
Robinson

Variant  
Calling

naive

MAP

MAQ

EM

SNVMix

Variant  
nomenclature

Annotation

Thus, the prior is skewed toward  $\pi_{aa}$  assuming that most positions will be homozygous for the reference allele. The pseudocounts are essentially equivalent to having seen  $1000 + 100 + 100 = 1200$  positions with the distribution 83.3%  $\langle a, a \rangle$ , and 8.3% each for  $\langle b, a \rangle$  and  $\langle a, b \rangle$ .

The weight of the prior belief is reflected in the number of pseudocounts. For instance, the following counts result in the same proportion but there is much less weight of prior belief

$\delta(\langle a, a \rangle)$	$\delta(\langle b, a \rangle)$	$\delta(\langle a, b \rangle)$
10	1	1

# Dirichlet

Variant  
Calling and  
Annotation

Peter N  
Robinson

Variant  
Calling

naive

MAP

MAQ

EM

SNVMix

Variant  
nomenclature

Annotation

A Dirichlet distribution of order  $k$  (3 in our example) is a PDF that represents the belief (“probability”) that the probabilities of  $k$  distinct events (in our case, the genotypes  $\langle a, a \rangle, \langle a, b \rangle, \langle b, b \rangle$ ) are  $x_i$  given that each event has been observed  $\alpha_i - 1$  times.

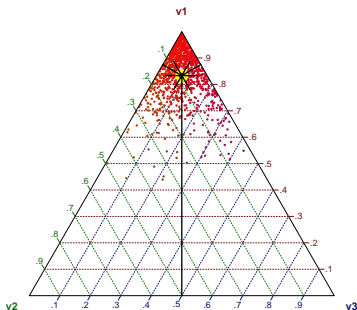
$$f(x_1, x_2, \dots, x_{k-1}; \alpha_1, \alpha_2, \dots, \alpha_k) = \frac{1}{B(\alpha)} \prod_{i=1}^k x_i^{\alpha_i-1} \quad (22)$$

- Note that by convention  $f$  has  $k - 1$  arguments. Since  $\sum_{i=1}^k x_i = 1$  there is no need to show the  $k^{\text{th}}$  argument.
- $B(\alpha)$  is the Beta function

$$B(\alpha) = \frac{\prod_{i=1}^k \Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^k \alpha_i\right)} \quad (23)$$

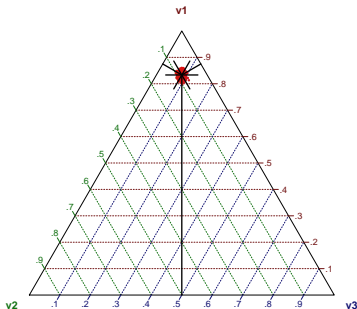
Dirichlet distribution with a low number of pseudocounts (relatively weak prior):

- $\delta(\langle a, a \rangle) = 10$
- $\delta(\langle a, b \rangle) = 1$
- $\delta(\langle b, b \rangle) = 1$



Dirichlet distribution with a higher number of pseudocounts  
(relatively strong prior):

- $\delta(\langle a, a \rangle) = 1000$
- $\delta(\langle a, b \rangle) = 100$
- $\delta(\langle b, b \rangle) = 100$



# SNVMix: M-step updating equation for $\pi$

Variant  
Calling and  
Annotation

Peter N  
Robinson

Variant  
Calling

naive

MAP

MAQ

EM

**SNVMix**

Variant  
nomenclature

Annotation

The update equation for  $\pi$  is similar to that for  $\mu$

$$\pi_k^{new} = \frac{\sum_{i=1}^T I(G_i = k) + \delta(k)}{\sum_{j \in \{\langle a,a \rangle, \langle a,b \rangle, \langle b,b \rangle\}} \sum_{i=1}^T I(G_i = j) + \delta(j)} \quad (24)$$



# SNVMix: Initialize EM

Variant  
Calling and  
Annotation

Peter N  
Robinson

We are now in a position to initialize the EM

We need:

- Mapped NGS reads comprising  $i = 1, \dots, T$  genomic positions, each of which has  $N_i$  reads with  $a_i$  reference and  $b_i$  nonreference bases.
- Initialize  $\pi_k = \frac{\delta(k)}{N_\delta}$  where  $N_\delta = \frac{\delta(k)}{\sum_j \delta(j)}$
- Initialize  $\mu_k = \frac{\alpha_k}{\alpha_k + \beta_k}$
- pick a tolerance to judge convergence

Variant  
Calling

naive

MAP

MAQ

EM

SNVMix

Variant  
nomenclature

Annotation

# SNVMix: Run EM

Variant  
Calling and  
Annotation

Peter N  
Robinson

Variant  
Calling

naive

MAP

MAQ

EM

**SNVMix**

Variant  
nomenclature

Annotation

The EM algorithm iterates between the E-step where we assign the genotypes using Equation (20) and the M-step where we re-estimate the model parameters with equations (24) for  $\pi$ , (21) for  $\mu$ .

At each iteration we evaluate the complete data log-likelihood as given by Equation (19) and the algorithm terminates when this quantity no longer increases

# SNVMix1 vs SNVMix2

Variant  
Calling and  
Annotation

Peter N  
Robinson

Variant  
Calling

naive

MAP

MAQ

EM

SNVMix

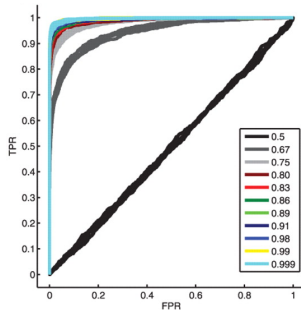
Variant  
nomenclature

Annotation

The SNVMix algorithm was later extended to include mapping and base qualities into the same Bayesian framework, primarily by adapting the formulas used for the EM equations. We will not discuss this here<sup>a</sup>

<sup>a</sup>Goya R et al. (2010) *Bioinformatics* **26**:730–736.

Performance of SNVMix2 algorithm on simulated data with increasing levels of certainty in the base call



# Annotation and Variant Nomenclature

Variant  
Calling and  
Annotation

Peter N  
Robinson

Variant  
Calling

naive

MAP

MAQ

EM

SNVMix

Variant  
nomenclature

Annotation

The output of variant calling programs is a set of calls with chromosomal coordinates, e.g.

10 72535007 C T

- i.e., position 72614421 on chromosome 10, which in the reference sequence has a thymine (T) is found to have an alternate cytosine (C) base
- This is fine, but biologists or MDs generally need gene-based annotations such as

HGVS=C10orf27:NM\_152710:exon8:c.710G>A:p.R237Q

# Annotation

Variant  
Calling and  
Annotation

Peter N  
Robinson

Variant  
Calling

naive

MAP

MAQ

EM

SNVMix

Variant  
nomenclature

Annotation

Annotation algorithms take genomic sequences and transcripts, and metadata about genes (e.g., positions of exons on a chromosome), and infer the transcript-level annotation for chromosomal variants.

- You are expected to know the basic variant nomenclature from the HGVS
- If you are not familiar with this, study how to represent missense, nonsense, and deletion mutations in cDNA and protein at <http://www.hgvs.org/mutnomen/>
- DNA: A,C,G,T
  - c.435C>A
- Protein: 1- or 3-letter code
  - p.A212P, Ala212Pro
- Use HGNC<sup>4</sup> Gene symbols, e.g., *FBN1* for Fibrillin-1
- Nice Tool: <http://www.humgen.nl/mutalyzer/1.0.1/>

<sup>4</sup>HUGO Gene Nomenclature Committee

# Standardized Nomenclature for Mutations

Variant  
Calling and  
Annotation

Peter N  
Robinson

Variant  
Calling

naive

MAP

MAQ

EM

SNVMix

Variant  
nomenclature

Annotation

## Short review of HGVS nomenclature for mutations

DNA ...

- Single-nucleotide Substitution c.123A>G
- Deletion c.123de1A
- Duplication c.123dupA
- Insertion c.123\_124insC

# Deletions & Insertions

Variant  
Calling and  
Annotation

Peter N  
Robinson

Variant  
Calling

naive

MAP

MAQ

EM

SNVMix

Variant  
nomenclature

Annotation

- c.546delT
- c.546del
- c.586\_591del
- c.586\_591delTGGTCA oder c.586\_591del6
- c.546\_547insT (not c.546insT – which is ambiguous)
- c.1086\_1087insGCGTGA

# Alleles

Variant  
Calling and  
Annotation

Peter N  
Robinson

Variant  
Calling

naive

MAP

MAQ

EM

SNVMix

Variant  
nomenclature

Annotation

- Nomenclature for two alleles in one gene may be needed for autosomal recessive diseases:
- c . [546C>T] + [2398delT]



# Frameshift

Variant  
Calling and  
Annotation

Peter N  
Robinson

Variant  
Calling

naive

MAP

MAQ

EM

SNVMix

Variant  
nomenclature

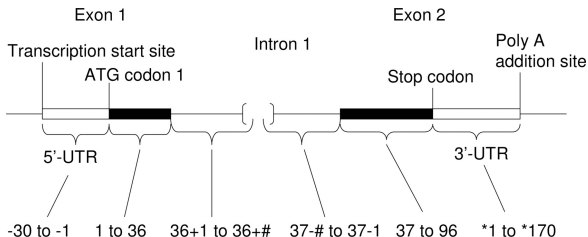
Annotation

- Short form p.Arg83fs
- Long form: p.Arg83SerfsX15
  - The frameshift has led to an amino acid substitution (Arg83Ser), which as a rule is located at the first amino acid affected by the frameshift mutation (“fs”)
  - We then see the length of the remaining protein sequence up to the next stop codon (X15, i.e., 15 amino acids)
  - Note that the amino acid sequence following the frameshift is almost always different from the wildtype sequence
  - We expect to find a premature termination codon not very far downstream from frameshift mutations ([why?](#))

# Numbering

Variant  
Calling and  
Annotation

Peter N  
Robinson



# indicates any positive integer number

- Note: splice mutations are numbered according to the closest exonic nucleotide position: z.B. 36+1G>C, 37-2A>G

# Annotation

Variant  
Calling and  
Annotation

Peter N  
Robinson

Variant  
Calling

naive

MAP

MAQ

EM

SNVMix

Variant  
nomenclature

Annotation

We will now show how to take the output of variant calling programs ( calls with chromosomal coordinates) and to produce gene/transcript-based annotations

There are a number of programs out there that are commonly used to process VCF files. Two of the best are

- Annovar: Wang K et al., (2010) *Nucleic Acids Res* **38**:e164.
- Variant Effect Predictor (At the EBI)

The latter can be used online (google it)

# Annotation

Variant  
Calling and  
Annotation

Peter N  
Robinson

Variant  
Calling

naive

MAP

MAQ

EM

SNVMix

Variant  
nomenclature

Annotation

The algorithms involve a lot of heuristics, which have to deal with the “weirdness” of our genome: overlapping genes, genes located in introns of other genes, multiple coding/non-coding transcripts of a single gene, antisense transcripts etc etc.

We will not attempt to demonstrate the complete process, but will present a simplified version of how a deletion variant gets annotated.

# Deletion

Variant  
Calling and  
Annotation

Peter N  
Robinson

Variant  
Calling

naive

MAP

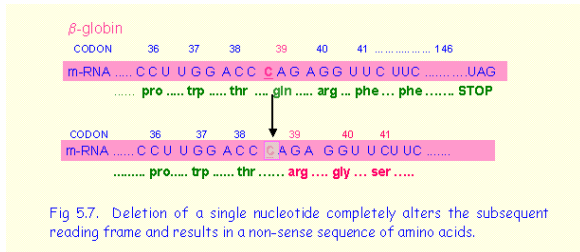
MAQ

EM

SNVMix

Variant  
nomenclature

Annotation



- Example of a deletion of a single nucleotide causing a frameshift

# Deletion

Variant  
Calling and  
Annotation

Peter N  
Robinson

Variant  
Calling

naive

MAP

MAQ

EM

SNVMix

Variant  
nomenclature

Annotation

On the chromosome, a deletion mutation is represented as

1    1887113    1887113    C    -

That is,

- Deletion of a “C” at position 1887113 of chromosome 1

Our goal is now to find out whether this deletion occurs within an exon of a gene, and if so, the determine the effect on the cDNA sequence

- KIAA1751:uc001aim.1:exon18:c.2193delG:p.Q731fs

# Annotation: Where does the data come from

Variant  
Calling and  
Annotation

Peter N  
Robinson

Variant  
Calling

naive

MAP

MAQ

EM

SNVMix

Variant  
nomenclature

Annotation

The read mapping algorithms use a reference genome sequence, usually hg18 (March 2006) or hg19 (Feb 2009). One can download annotation files from various sources including Ensembl and UCSC. These files include information about genes and transcripts, and are used to infer the effects of variants.

We will demonstrate this with the UCSC known genes data.<sup>5</sup> One can download a series of files with the entire genome sequence and information about each transcript, its exon structure, and the nucleotide sequence of the splice mRNA (cDNA).

---

<sup>5</sup>see <http://genome.ucsc.edu>

# UCSC KnownGenes

Variant  
Calling and  
Annotation

Peter N  
Robinson

Variant  
Calling

naive

MAP

MAQ

EM

SNVMix

Variant  
nomenclature

Annotation

field	example	explanation
name	uc001aaa.3	Name of gene
chrom	chr1	ref. chromosome
strand	+	+ or - strand
txStart	11873	Transcription start
txEnd	14409	Transcription end
cdsStart	11873	Coding region start
cdsEnd	11873	Coding region end
exonCount	3	Number of exons
exonStarts	11873,12612,13220,...	Exon start positions
exonEnds	12227,12721,14409,...	Exon start end positions



# Annotation: bird's eye view

Variant  
Calling and  
Annotation

Peter N  
Robinson

Variant  
Calling

naive

MAP

MAQ

EM

SNVMix

Variant  
nomenclature

Annotation

- ① Identify the gene(s) that overlaps with the variant<sup>6</sup>
- ② Determine whether the variant is located in an exon<sup>7</sup>
- ③ Determine whether the variant is located in the coding sequence<sup>8</sup>
- ④ If the variant is a SNV, determine whether the encoded amino acid is changed<sup>9</sup>, if so annotate as **missense**
- ⑤ If the variant is located right before or after an exon/intron boundary, annotate as **splicing**
- ⑥ If the variant removes/adds nucleotides from the CDS, annotate as **deletion/insertion**<sup>10</sup>

---

<sup>6</sup> If you do not find such a gene, the variant is **upstream**, **downstream**, or **intergenic**.

<sup>7</sup> If not, the variant is **intronic**

<sup>8</sup> If not, the variant is **3' or 5' UTR**

<sup>9</sup> If not, the variant is a **synonymous substitution**

<sup>10</sup> If the variant removes/adds  $2n+1$  or  $2n+2$  nucleotides, also annotate as **frameshift**

# Annotation: Overview

Variant  
Calling and  
Annotation

Peter N  
Robinson

Variant  
Calling

naive

MAP

MAQ

EM

SNVMix

Variant  
nomenclature

Annotation

---

## Algorithm 1 Create annotation

---

```
1: get_annotation(chr, pos, ref, var)
2: gene  $\leftarrow$  get_overlapping_gene(chr, pos)
3: if gene == NULL then
4:   return ‘‘variant non genic’’
5: end if
6: if ann  $\leftarrow$  get_deletion(chr, pos, ref, var) != NULL then
7:   return ann
8: else
9:   // Other functions for other mutation classes
10: end if
```

---

# Annotation: Get overlapping gene

Variant  
Calling and  
Annotation

Peter N  
Robinson

Variant  
Calling

naive

MAP

MAQ

EM

SNVMix

Variant  
nomenclature

Annotation

---

## Algorithm 2 Get overlapping gene

---

```
1: gene  $\leftarrow$  get_overlapping_gene(chr, pos)
2: for all  $g \in \mathcal{G}$  do //loop over candidate genes
3:   txStart  $\leftarrow$  g.txStart
4:   txEnd  $\leftarrow$  g.txEnd
5:   if pos < txStart then
6:     continue //variant is 5' to gene
7:   else if pos > txEnd then
8:     continue //variant is 3' to gene
9:   end if
10:  return g // variant is located within gene model
11: end for
```

---

● Note that for genes on “-”, the meaning of txStart and txEnd is “reversed”

# Annotation: Get rvarstart

---

## Algorithm 3 Get rvarstart

---

```
1: get_rvarstart(gene g, pos)
2: cumlenintron  $\leftarrow$  0 // cumulative length of introns
3: cumlenexon  $\leftarrow$  0 // cumulative length of exons
4: for  $k \in g.getExons()$  do //loop over exons
5:     cumlenexon  $\leftarrow$  cumlenexon + g.getExonLen(k)
6:     if  $k > 0$  then //no intron before first exon!
7:         cumlenintron  $\leftarrow$  cumlenintron + g.getIntronLen(k)
8:     end if
9:     if pos located in exon  $k$  then
10:         rvarstart  $\leftarrow$  start - txStart - cumlenintron + 1;
11:         return rvarstart // start of variant within mRNA
12:     end if
13: end for
```

---

# Annotation: Get deletion

---

## Algorithm 4 Get rvarstart

---

```
1: get_deletion(gene g, pos,ref)
2: rvarstart  $\leftarrow$  get_rvarstart(g, pos)
3: cdspos  $\leftarrow$  rvarstart - g.getCDSstart() + 1
4: mut  $\leftarrow$  sprintf("c%ddel%c",cdspos,ref)
5: return mut
```

---

- This will return something like c.2193delG, a deletion of a "G" at position 2193 of the coding sequence
- N.B.: This function is highly simplified and will only work for the + strand.

# Annotation: Final words

Variant  
Calling and  
Annotation

Peter N  
Robinson

This should be enough to give you some idea of how genomic variants are annotated. We have of course left out a large number of functionalities

- Figure out effect on protein based on translating the DNA sequence
- Functions for other kinds of variants
- Figuring out how to number variants in genes on the minus strand
- Reporting either all consequences of a given variant (there may be a different consequence for each of the isoforms of a gene)
- Alternatively, prioritize the variants to report only the most relevant variant

Variant  
Calling

naive

MAP

MAQ

EM

SNVMix

Variant  
nomenclature

Annotation

# Summary

Variant  
Calling and  
Annotation

Peter N  
Robinson

Variant  
Calling

naive

MAP

MAQ

EM

SNVMix

Variant  
nomenclature

Annotation

In these two lectures we have examined how to call variants. We have studied two classes of algorithms, MAP estimation and expectation maximization, and how they are used in two variant calling algorithms, MAQ and SNVMix. Finally, we have examined how called variants are annotated as to their consequences for genes and transcripts, which is usually necessary for biological or medical interpretation of variants.

What you should now know:

- The kinds of data used in variant calling (mapping quality, base quality, depth, ...)
- How this data is exploited to improve variant calling
- Bayes' law and how it can be used to estimate parameters
- Be able to interpret the major formulae of MAP and EM (no need to memorize)
- Variant nomenclature and basic ideas of variant annotation ("bird's eye view")

# The End of the Lecture as We Know It

Variant  
Calling and  
Annotation

Peter N  
Robinson

Variant  
Calling

naive

MAP

MAQ

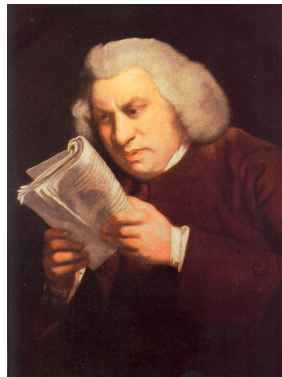
EM

SNVMix

Variant  
nomenclature

Annotation

- Email: [peter.robinson@charite.de](mailto:peter.robinson@charite.de)
- Office hours by appointment



Lectures were once useful; but now, when all can read, and books are so numerous, lectures are unnecessary. If your attention fails, and you miss a part of a lecture, it is lost; you cannot go back as you do upon a book... People have nowadays got a strange opinion that everything should be taught by lectures. Now, I cannot see that lectures can do as much good as reading the books from which the lectures are taken. I know nothing that can be best taught by lectures, except where experiments are to be shown. You may teach chymistry by lectures. You might teach making shoes by lectures!