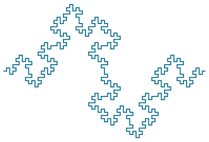
Bios 824: HTS Module

Bios 824: HTS Statistical Model

Biostatistics and Bioinformatics

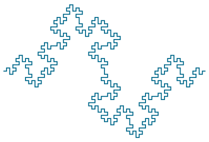


Spring 2019

Bios 824: HTS Module

Bios 824: HTS Statistical Model

Biostatistics and Bioinformatics



Spring 2019

Section 1

Outline

- Model to infer genotypes for germline variants (SNPs) from DNA-Seq

- Regression model for RNA-Seq counts

- Error mapping model for alignment

Section 2

Germline Variant Calling

Introduction

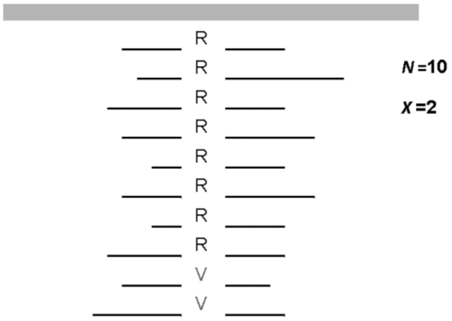
- We will consider an approach called `seqEM` propose by Martin *et al.*, [1]
- I personally do *not* recommend this approach for data analysis
- It is a good approach, in my opinion, for teaching purposes

Pile up: Example

http://samtools.sourceforge.net/pileup.shtml

```
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<<:<<
```

Simplified Pileup Illustration (from Martin *et al.*, [1])



Model: Notation

- Let B_j denote the allele at locus j
- For simplicity, assume that B_j is either R (the "reference" allele) or V (the "variant" allele)
- The number of sequencing base calls at locus j is D_j
- The D_j base calls are $\tilde{B}_{1j}, \ldots, \tilde{B}_{1D_j}$
- D_j is the depth at the locus

Model

- Let G_j denote the genotype at locus j
- G_j is either $RRRVandVV$.
- In absence of alignment or sequencing errors
  - $G_j = RR\tilde{B}_{1j} =, \ldots, = \tilde{B}_{1D_j} = R$
  - $G_j = RV$ about half of the D_j base calls $\tilde{B}_{1j}, \ldots, \tilde{B}_{1D_j}$ are R and half are V.
  - $G_j = VV\tilde{B}_{1j} =, \ldots, = \tilde{B}_{1D_j} = V$

Model: Error

- It is unrealistic to assume that there are neither alignment nor sequencing errors
- Assume that the errors are symmetric

$$\alpha = \mathbb{P}[\tilde{B}_{ij} = R|B_j = V] = \mathbb{P}[\tilde{B}_{ij} = R|B_j = V]$$

Distribution of number of variant calls

- Let

$$S_j = \sum_{i=1}^{D_j} I[\tilde{B}_{ij} = V$$

- This is number of variant calls (the number of bases at this locus called as V) among the D_j reads
- The distribution of S_j is not binomial.
- The distribution of S_j given the genotype is binomial
- Recall that we have assumed that error probability is the same for each read (not indexed by i) and symmetric

Distribution of S_j given $G_j = g$

$$\mathbb{P}[S_j = s | G_j = g, D_j = d] = \begin{cases} \binom{d}{s}\alpha^s(1-\alpha)^{d-s} & g = RR \\ \binom{d}{s}\frac{1}{2}^d & g = RV \\ \binom{d}{s}(1-\alpha)^s(\alpha)^{d-s} & g = VV \end{cases}$$

Joint Distribution of S_j and G_j

- p_{VV}: Prior genotypic probability for VV
- p_{RV}: Prior genotypic probability for RV
- $p_{RR} = 1 - p_{VV} - p_{RV}$: Prior genotypic probability for RR

$$\mathbb{P}[S_j = s, G_j = g | D_j = d; \theta] = \begin{cases} \binom{d}{s}\alpha^s(1-\alpha)^{d-s}(1 - p_{VV} - p_{RV}) & g = RR \\ \binom{d}{s}\frac{1}{2}^d p_{RV} & g = RV \\ \binom{d}{s}(1-\alpha)^s(\alpha)^{d-s} p_{VV} & g = VV \end{cases}$$

where

$$\theta = (\alpha, p_{VV}, p_{RR})$$

Likelihood

- Observations: $\tilde{B}_{1j}, \dots, \tilde{B}_{1D_j}$
- Observed statistic: S_j
- Latent variable $G_j, B_{1j}, \dots B_{D_j j}$
- Model parameter $\theta = (\alpha, p_{VV}, p_{RR})$
- Distribution of S_j

$$\mathbb{P}[S_j = s | D_j = d; \theta] = \sum_{g \in \{VV, RV, RR\}} \mathbb{P}[S_j = s, G_j = g | D_j = d; \theta]$$

Likelihood

Data from n patients

$$\ell[\theta] = \sum_{k=1}^{n} \mathbb{P}[S_j = s_{kj} | D_j = d_{kj}; \theta]$$

Infer Genotype

- Let $\hat{G}_j$ be the genotype call at locus j
- The call is wrong if $\hat{G}_j \neq G_j$
- This is called the Baye's error
- The Bayes' decision rule minimized the probability of Bayes' error

$$\hat{G}_j = \mathrm{argmax}_g \mathbb{P}[S_j = s, G_j = g | D_j = d; \theta]$$

- We cannot calculate this so we will use the plugin decision rule

$$\hat{G}_j = \mathrm{argmax}_g \mathbb{P}[S_j = s, G_j = g | D_j = d; \hat{\theta}_n]$$

θ_n is the vector of parameter estimates

E. R. Martin, D. D. Kinnamon, M. A. Schmidt, E. H. Powell, S. Zuchner, and R. W. Morris.
SeqEM: an adaptive genotype-calling approach for next-generation sequencing studies.
Bioinformatics, 26(22):2803–2810, 09 2010.