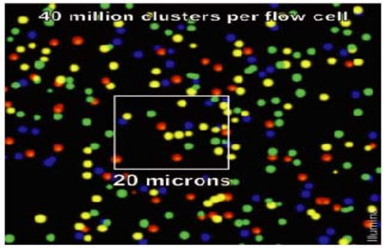
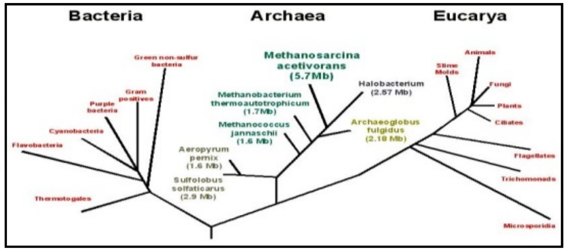
```
TAACCCTAACCCTAACCCTAACCCTAACCCTA
CCTAACCCTAACCCTAACCCTAACCCTAACCC
CCCTAACCCCTAACCCTAACCCTAACCCTAAC
AACCCTAACCCTAACCCTAACCCCTAACCCTA
ACCCTAACCCCAACCCCAACCCCAACCCCAAC
CTACCCTAACCCTAACCCTAACCCTAACCCTA
ACCCTAACCCTAACCCTAACCCCTAACCCTAA
```



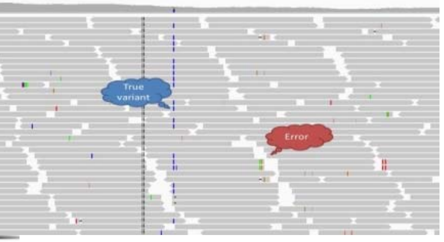

Bacteria
Archaea
Eucarya
Green non-sulfur bacteria
Gram positives
Purple bacteria
Cyanobacteria
Flavobacteria
Thermotogales
Methanosarcina acetivorans (5.7Mb)
Methanobacterium thermoautotrophicum (1.7Mb)
Methanococcus jannaschii (1.6 Mb)
Halobacterium (2.57 Mb)
Archaeoglobus fulgidus (2.18 Mb)
Aeropyrum pernix (1.6 Mb)
Sulfolobus solfataricus (2.9 Mb)
Animals
Slime Molds
Fungi
Plants
Ciliates
Flagellates
Trichomonads
Microsporidia

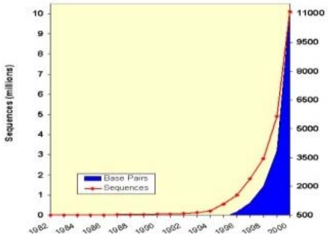
Supplementary Learning Materials

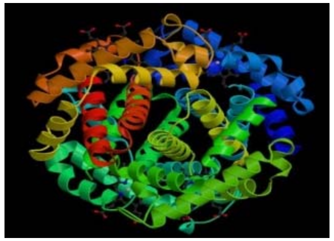
叶永鑫(Adam Y. Ye)

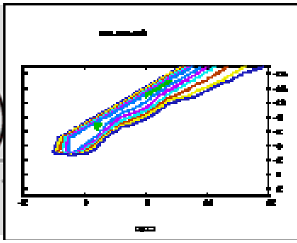
北京大学生物信息学中心

Center for Bioinformatics, Peking University









Outline

- Introduction of Likelihood and Bayesian approach

- Genotyper of MAQ and SNVMix

Likelihood & Bayesian

- Likelihood function
  - a function of the parameters of a statistical model
  - $L(\theta) = P(Data|\theta)$

- Bayesian approach
  - $P(\theta|Data) \propto P(\theta) * P(Data|\theta)$
  - posterior ∝ prior * likelihood

A Simple Demostration

- Toss a biased coin, let θ= P(Head) in one trial
- Probability for seeing HTHH?

$$\begin{aligned} L(\theta) &= P(Data|\theta) = P(HTHH|\theta) \\ &= \theta \cdot (1-\theta) \cdot \theta \cdot \theta = \theta^3(1-\theta) \end{aligned}$$

Bernoulli distribution

- Probability for seeing 3 Heads in 4 trials?

$$\begin{aligned} L(\theta) &= P(Data|\theta) = P(3H \text{ in } 4|\theta) \\ &= \binom{4}{3}\theta^3(1-\theta) \end{aligned}$$

binomial distribution

Models for SNP Calling and Genotyping

- MAQ
  - Li, H., Ruan, J., and Durbin, R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Research 18, 1851–1858.
- samtools
  - Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27, 2987–2993.
- GATK
  - McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research 20, 1297–1303.
  - DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Genetics 43, 491–498.
- SNVMix
  - Goya, R., Sun, M.G.F., Morin, R.D., Leung, G., Ha, G., Wiegand, K.C., Senz, J., Crisan, A., Marra, M.A., Hirst, M., et al. (2010). SNVMix: predicting single nucleotide variants from next-generation sequencing of tumors. Bioinformatics 26, 730–736.
- …

Genotyping Model used in MAQ

- Data: a pile of bases, with baseQ
  - k nucleotide b and (n-k) nucleotide b' with error rate $\epsilon_1 \leq \cdots \leq \epsilon_k$ $\epsilon_{k+1} \leq \cdots \leq \epsilon_n$
- Goal: call genotype <b,b>, <b,b'>, <b',b'>
- For G=<b,b'>, $\Pr\{Data|G =< b, b' >\} \approx \frac{1}{2^n}\binom{n}{k}$

Genotyping Model used in MAQ

- For G=<b,b>,

$$\alpha_{nk} = \Pr\{\text{exactly } k \text{ errors in } n \text{ bases}\}$$

$$\bar{\alpha}_{nk}(\bar{\epsilon}) = \binom{n}{k} \bar{\epsilon}^{k} (1 - \bar{\epsilon})^{n-k}$$

Genotyping Model used in MAQ

$$\alpha_{nk} = \Pr\{\text{exactly } k \text{ errors in } n \text{ bases}\}$$

$$\beta_{nk} = \begin{cases} \Pr\{\text{more than } k \text{ errors} | \text{more than } k-1 \text{ errors in } n \text{ bases}\} & (k > 0) \\ \Pr\{\text{more than } 0 \text{ error in } n \text{ bases}\} & (k = 0) \end{cases}$$

$$\alpha_{nk} = (1 - \beta_{nk})\beta_{n(k-1)} \cdots \beta_{n2}\beta_{n1} = (1 - \beta_{nk}) \prod_{i=0}^{k-1} \beta_{ni} \qquad \sum_{k=0}^{n} \alpha_{nk} = 1$$

$$\beta_{nk} = \frac{\sum_{i=k+1}^{n} \alpha_{ni}}{\sum_{i=k}^{n} \alpha_{ni}} = \frac{1 - \sum_{i=0}^{k} \alpha_{ni}}{1 - \sum_{i=0}^{k-1} \alpha_{ni}} \qquad \beta_{nn} = 0$$

Genotyping Model used in MAQ

$$\bar{\alpha}_{nk}(\bar{\epsilon}) = \binom{n}{k}\bar{\epsilon}^{k}(1-\bar{\epsilon})^{n-k} \qquad \bar{\beta}_{nk}(\bar{\epsilon}) = \frac{1-\sum_{i=0}^{k}\bar{\alpha}_{ni}}{1-\sum_{i=0}^{k-1}\bar{\alpha}_{ni}}$$

$$\beta_{nk}(\bar{\epsilon}) = \bar{\beta}_{nk}^{f_k}(\bar{\epsilon}) \qquad 0 < f_k \leq 1$$

$$\alpha_{nk}(\bar{\epsilon}) = (1-\bar{\beta}_{nk}^{f_k})\prod_{i=0}^{k-1}\bar{\beta}_{ni}^{f_i} = (1-\bar{\beta}_{nk}^{f_k})\prod_{i=0}^{k-1}\left(\frac{\bar{\beta}_{ni}}{\bar{\epsilon}}\right)^{f_i}\cdot\bar{\epsilon}^{f_i} = c_{nk}(\bar{\epsilon})\cdot\prod_{i=0}^{k-1}\bar{\epsilon}^{f_i}$$

$$c_{nk}(\bar{\epsilon}) = (1-\bar{\beta}_{nk}^{f_k})\prod_{i=0}^{k-1}\left(\frac{\bar{\beta}_{ni}}{\bar{\epsilon}}\right)^{f_i}$$

Genotyping Model used in MAQ

$$\alpha_{nk}(\epsilon_1, \cdots, \epsilon_k; \epsilon_{k+1}, \cdots, \epsilon_n) \approx c_{nk}(\bar{\epsilon}) \cdot \prod_{i=0}^{k-1} \epsilon_{i+1}^{f_i}$$

$$\log \bar{\epsilon} = \frac{\sum_{i=0}^{k-1} f_i \log \epsilon_{i+1}}{\sum_{i=0}^{k-1} f_i} \qquad \prod_{i=0}^{k-1} \bar{\epsilon}^{f_i} = \prod_{i=0}^{k-1} \epsilon_{i+1}^{f_i}$$

$$f_k = 0.85^k$$

$$\alpha_{nk}(\epsilon_1, \cdots, \epsilon_k; \tilde{\epsilon}_1, \cdots, \tilde{\epsilon}_{\tilde{k}}; \epsilon_{k+1}, \cdots, \epsilon_n; \tilde{\epsilon}_{\tilde{k}+1}, \cdots, \tilde{\epsilon}_{\tilde{n}}) \approx c_{nk}(\bar{\epsilon}) \prod_{i=0}^{k-1} \epsilon_{i+1}^{f_i} \cdot c_{\tilde{n}\tilde{k}}(\bar{\tilde{\epsilon}}) \prod_{\tilde{i}=0}^{\tilde{k}-1} \tilde{\epsilon}_{\tilde{i}+1}^{f_{\tilde{i}}}$$

Genotyping Model used in MAQ

- For G=<b,b>,

$$\Pr\{Data|G=<b,b>\}=\alpha_{nk}(\epsilon_1,\cdots,\epsilon_k;\epsilon_{k+1},\cdots,\epsilon_n)$$

- For G=<b,b'>,

$$\Pr\{Data|G=<b,b'>\}\approx\frac{1}{2^n}\binom{n}{k}$$

- For G=<b',b'>,

$$\Pr\{Data|G=<b',b'>\}=\alpha_{n,n-k}(\epsilon_{k+1},\cdots,\epsilon_n;\epsilon_1,\cdots,\epsilon_k)$$

Genotyping Model used in MAQ

$$\Pr\{G|Data\} \propto \Pr\{G\} \cdot \Pr\{Data|G\}$$

- For G=<b,b>,

$$\Pr\{G =< b.b > |Data\} = \frac{\Pr\{G =< b.b >\} \cdot \Pr\{Data|G =< b.b >\}}{\Pr\{G =< b.b >\} \cdot \Pr\{Data|G =< b.b >\} + \Pr\{G =< b.b' >\} \cdot \Pr\{Data\ G =< b.b' >\} + \Pr\{G =< b'.b' >\} \cdot \Pr\{Data|G =< b'.b' >\}}$$

- For G=<b,b'>,

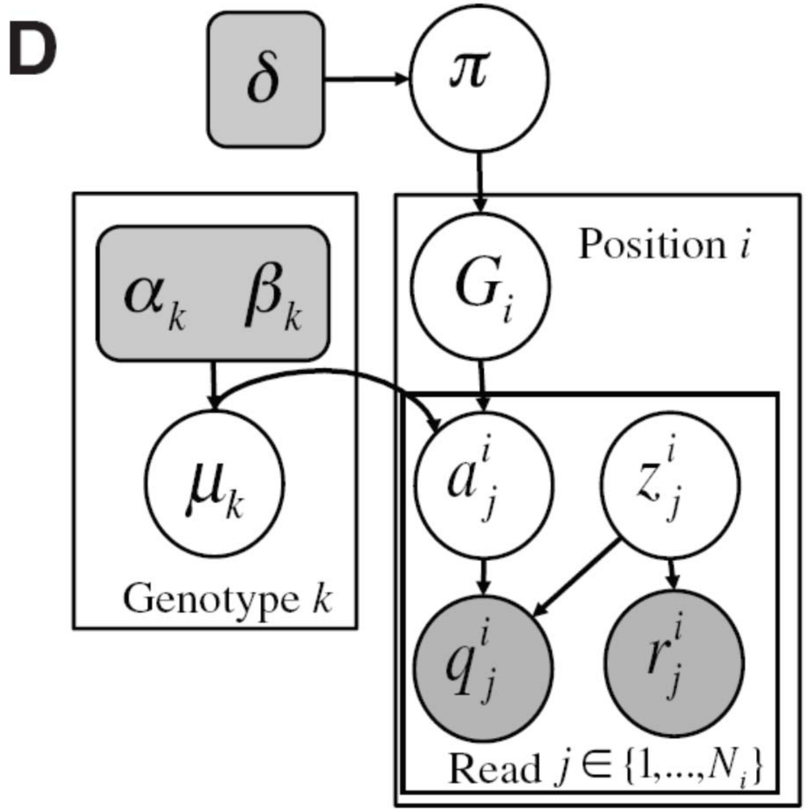
$$\Pr\{G =< b.b' > |Data\} = \frac{\Pr\{G =< b.b' >\} \cdot \Pr\{Data|G =< b.b' >\}}{\Pr\{G =< b.b >\} \cdot \Pr\{Data|G =< b.b >\} + \Pr\{G =< b.b' >\} \cdot \Pr\{Data|G =< b.b' >\} - \Pr\{G =< b'.b' >\} \cdot \Pr\{Data|G =< b'.b' >\}}$$

- For G=<b',b'>,

$$\Pr\{G =< b'.b' > |Data\} = \frac{\Pr\{G =< b'.b' >\} \cdot \Pr\{Data|G =< b'.b' >\}}{\Pr\{G =< b.b >\} \cdot \Pr\{Data|G =< b.b >\} + \Pr\{G =< b.b' >\} \cdot \Pr\{Data|G =< b.b' >\} + \Pr\{G =< b'.b' >\} \cdot \Pr\{Data|G =< b'.b' >\}}$$

Genotyping Model used in SNVMix

- Probabilistic Graphical Model
  - position i, read j, genotype k
  - G_i: genotype
  - a_j^i: match reference allele or not?
  - q_j^i: prob. of correct base calling
  - z_j^i: alignment correct or not?
  - r_j^i: prob. of correct mapping
  - μ_k: parameter of binomial for genotype k



SNVMix2 model

Goya, R., et al. (2010). SNVMix: predicting single nucleotide variants from next-generation sequencing of tumors. Bioinformatics 26, 730–736.

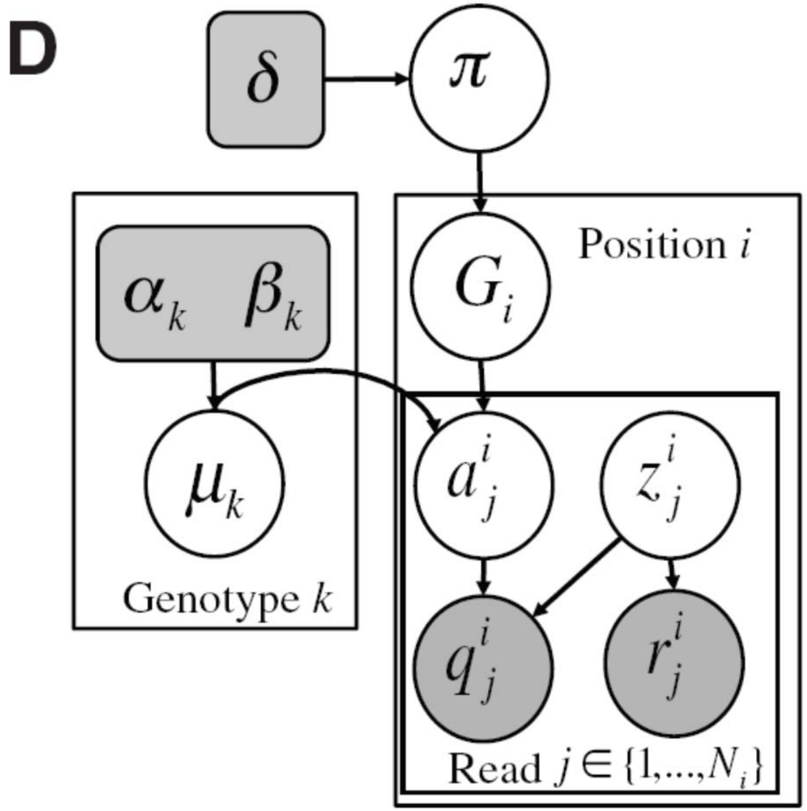
Genotyping Model used in SNVMix

$$p(G_i|\pi) = \text{Multinomial}(G_i|\pi, 1)$$

$$p(\pi|\delta) = \text{Dirichlet}(\pi|\delta)$$

$$p(a^i_j|G_i = k, \mu_k) = \text{Bernoulli}(a^i_j|\mu_k)$$

$$p(\mu_k|\alpha_k, \beta_k) = \text{Gamma}(\mu_k|\alpha_k, \beta_k)$$



SNVMix2 model

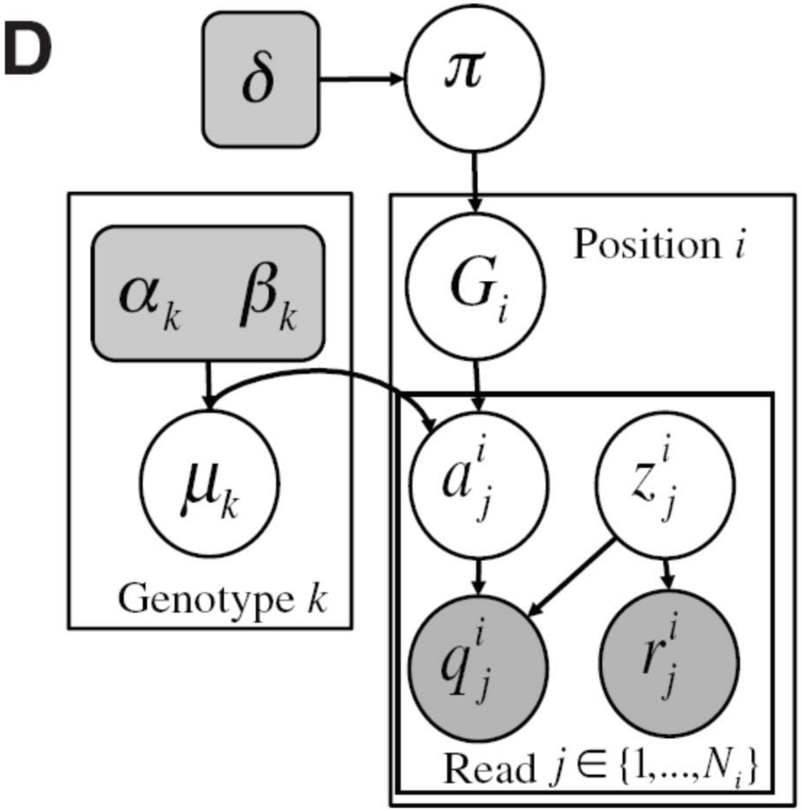
Goya, R., et al. (2010). SNVMix: predicting single nucleotide variants from next-generation sequencing of tumors. Bioinformatics 26, 730–736.

Genotyping Model used in SNVMix

$$p(z_j^i) = \text{Bernoulli}(z_j^i|0.5)$$

$$p(q_j^i|a_j^i. z_j^i) = \begin{cases} q_j^i & \text{if } a_j^i = 1. z_j^i = 1 \\ 1 - q_j^i & \text{if } a_j^i = 0. z_j^i = 1 \\ 0.5 & \text{if } z_j^i = 0 \end{cases}$$

$$p(r_j^i|z_j^i) = \begin{cases} r_j^i & \text{if } z_j^i = 1 \\ 1 - r_j^i & \text{if } z_j^i = 0 \end{cases}$$



SNVMix2 model

Goya, R., et al. (2010). SNVMix: predicting single nucleotide variants from next-generation sequencing of tumors. Bioinformatics 26, 730–736.

Thank you for your attention

https://www.coursera.org/course/pkubioinfo