# COMS 4761, Assignment 4

Jing He:jh3283

February 27, 2013

1. We covered in class rapidly searchable BWT of a genome R. This should save

a. The transformed genome (a string of length $\|R\| = 3 * 10^9$ over the alphabet A,C,G,T, requiring 2bits per character). This is a vector $T[1...3 * 10^9]$.

b. The first column of the sorted matrix of cyclic shifts, saved as the starting positions for each block of identical letters (4 numbers of magnitude $\|R\|$, each requiring $log\|R\| = 32$ bits). This is a vector F[1..4]

c. $\|R\|/log\|R\| = 3 * 10^9/32 = 93,750,000$ rows (each of length 4) of the matrix of ordinals (yellow matrix on the slides), each ordinal requiring 32 bits. This is a matrix M[1... 93,750,000,1...4]

d. $\|R\|$ deltas (the purple column on the slides) that list how many of the current letter have already been observe since the last saved row of M. These deltas are bounded by 32, so each require 5 bits. This is a vector $D[1...3 * 10^9]$.

All this was shown in class to allow searching for a read and knowing whether or not the read sequence is observed along the genome. It also finds letters of the read along the transformed genome, T. This still does not tell us where along R the read lies. Assuming reads are of length 32, propose an additional vector, of no more than $3 * 10^9$ bits in total, that you would save, in order to facilitate rapidly knowing where the reads lie. How will you use this vector?

Hint: Start with a vector of $3 * 10^9$ large numbers. Then save only some of these numbers.

**Sol:**

To find the index of original position in genome is in fact a tracing back question. To start with storing a vector of **??** $= [1..3 * 10^9]$, each number need 32 bits, and the total storage would be $S_0 = 32 * 3 * 10^9$. To save space, we adopt the similar thrifty storage strategy: index every I rows, so the total bits stored in total would be $log(3 * 10^9/I) * S_0/I$, so $log(3 * 10^9/I) * 3 * 10^9 \leq I * 3 * 10^9$

when the interval I ¿ 27, the storage space would be less than $3 * 10^9$, the each searching, we can tracing the index. the expectation # of index

one reads will hit is $32/28 = 8/7$

pseudo code:

```
step 1 store:
for ( 1:3*10^9)
if( i modular 28 == 0)
  store index into a vector S_{1} = {1...107142857}
step 2 trace:
for each base in a read (1:32)
if indexed, keep the index postition,
Pos original = Index * 28 (for index > 1) - search times
if indexed, keep search back until hit one index, store search times
Pos original = Index - search times-read length
```

2. Suppose you are looking at sequencing data from a single individual wildebeest (a mammal, hence a 3Gb genome). You are looking at a specific site, position 23456789 on chromosome 1. Suppose the local depth of coverage is 15 (15 reads map to positions that span this site). You are seeing 10 "A"s, 4 "C"s and one "T" at this site.

a. If sequencing error rate is 1% in each base, and all three errors are always equally likely (each with probability 1/300). Compute the likelihood ratio of this being a heterozygote A/C, vs. a homozygote A.

**Sol:**

$$Heterozygote : Pr(C) = Pr(A) = (1 - 1\%)/2 + 0.5 * \frac{1}{300} \quad (1)$$

$$= 0.4966667 \quad (2)$$

$$Pr(T) = Pr(G) = \frac{1}{2}\frac{1}{300} * 2 \quad (3)$$

$$= 0.003333333 \quad (4)$$

$$Homozygote : Pr(A) = (1 - 1\%) * 99\% + 1\% * \frac{1}{300} * \quad (5)$$

$$= 0.9801333 \quad (6)$$

$$Pr(T) = Pr(C) = Pr(G) = 1\% * 99\% + 99\% * \frac{1}{300} + 1\% * \frac{2}{300} \quad (7)$$

$$= 0.01326667 \quad (8)$$

$$= \frac{P(Data \| H1)}{P(Data \| H2)}$$

$$= \frac{\binom{15}{10} * 0.4966667^{10} * \binom{15}{4} * 0.4966667^4 * 0.003333333}{\binom{15}{10} * 0.9801333^{10} * \binom{15}{4} * 0.01326667^4 * 0.01326667}$$

$$= 1.117848e - 06$$

b. Suppose you looked up a database of wildebeest variation, and found that at this site, chromosome 1, position 23456789, 81% of wildebeest individuals are homozygote A, and 18% are heterozygote A/C (the rest are homozygote C). Given these priors, and the additional evidence from your sequencing work, what are the posterior probabilities of these two calls for your specific individual?

c. You observed the same read content (10 "A"s, 4 "C"s and one "T") also at position 3456789 on chromosome 2. However there, the reference genome is "A", and this site is not listed in the database as one of the 30 million sites that had been observed as a non-reference allele among wildebeest individuals that had so far been sequenced. You know that a typical wildebeest individual would demonstrate about 300,000 such novel, heterozygous alleles. Given this information for prior odds, and the evidence from sequencing, what are your posteriors for your individual at this site?
**Sol:**

d. You would like to report as many as you can of the novel alleles, with at most 3,000 expected false positives (1%). You decide to focus on the sites with local depth d, 10¡d¡30. You divide up this range of coverage values, i.e. the integers [11,29], into three ranges of coverage values [11,a],[a+1,b] and [b+1,29] , and require a novel allele in respective coverage ranges to be observed at least 4, 5 or 6 times in order to call it. What would be a good choice of a and b? (in your calculations you may ignore sites that have more than two alleles observed)

**Sol:**

e. If the entire genome is sequenced to a Poisson-average depth of 20X, what fraction of the genome are you ignoring by focusing at 10¡d¡30?

**Sol:**

f. In the coverage listed in (e), what fraction of the 300,000 novel, heterozygous alleles would you be able to call using the policy devised in (d)?

**Sol:**

g. How would your answer to (a) change if instead of assuming an error probability of 10-2.0 for each base, you have confidence information from the sequencing machine? Specifically, compute the likelihood ratio if this information is conveyed (in parenthesis) as log10 of the error probability for each of the 15 bases observed, as follows:

a (-3.7) a (-3.3) a (-3.0) a (-2.7) a (-2.3) a (-2.0) a (-1.7) a (-1.3) a (-1.0) a (-0.7) c (-1.7) c (-1.3) c (-1.0) c (-0.7) t (-0.3)

**Sol:**

3. While reads has been treated in class as independent observations, in fact, sequencing machines typically work with DNA fragments of roughly the same length, and produce paired reads from opposite ends of each such fragment. Reads in each pair are designated as such. One would expect them to map to positions with distance corresponding to the fragment length. Such paired end reads from a segment of prescribed length can indicate a

long insertion (or deletion) if they map to positions that are much closer (or farther, resp.) along the reference genome. Suppose you are sequencing with reads of length 100bp that are experimentally set to be paired-ends of segments of length 1kbp. Your experiment is imperfect, so in practice, the segment length only have a mean length of 1kbp, and are distributed normally around it, with std. deviation of 200bp.

a. You want to detect Alu (300bp) insertions. You compute, for each site, the distance between mapped read pairs, averaged across all read-pairs that span the site. You decide that if that distance exceeds a cutoff of 1250bp, you will call this a homozygote Alu insertion. How many read pairs do you need to average for this to have a false positive rate of less than 10-6?

**Sol:**

b. If coverage is 5X, what is the chance that a site will have less than this number of read pairs spanning it?

**Sol:**