Tag-based Error correction

Dougals Wu

December 1, 2015

This is a note for myself on the tag-based error correction program that I coded up.

Grouping

For a given read 1, the first 13 bases is the index as below.

[GGAAGAGCACACG] TCT GAA CTC CAG TCA CAC TGA TAT CTC GTA TGC CGT CTT CTG CTT GAA AAA AAA AAGG GGG G

Reads with same index are grouped together in a dictionary (python).

Concensus base

For a cluster of reads (group), bases at a single position were extracted, concensus base was predicted using maximum likelihood.

For a give position, assume the bases A, C, T, G are observed j, k, l, m times respectively. The likelihood of the concensus base is A would be computed as:

$$L(base = A|jA, kC, lT, mG) = P(jA, kC, lT, mG|base = A) = \prod_{b_i \in \text{(all bases)}} P(b_i|base = A)$$

$$P(jA, kC, lT, mG|base = A) = P(A|base = A)^j \times P(C|base = A)^k \times P(T|base = A)^l \times P(G|base = A)^m$$

And sequencing error was estimated at 0.01.

$$P(jA, kC, lT, mG|base = A) = (1 - 0.01)^j \times 0.01^k \times 0.01^l \times 0.01^m$$

Likelihood is calculated for all four base

$$\theta \in \{L(base = A|jA, kC, lT, mG), L(base = T|jA, kC, lT, mG), L(base = C|jA, kC, lT, mG), L(base = G|jA, kC, lT, mG)\}$$

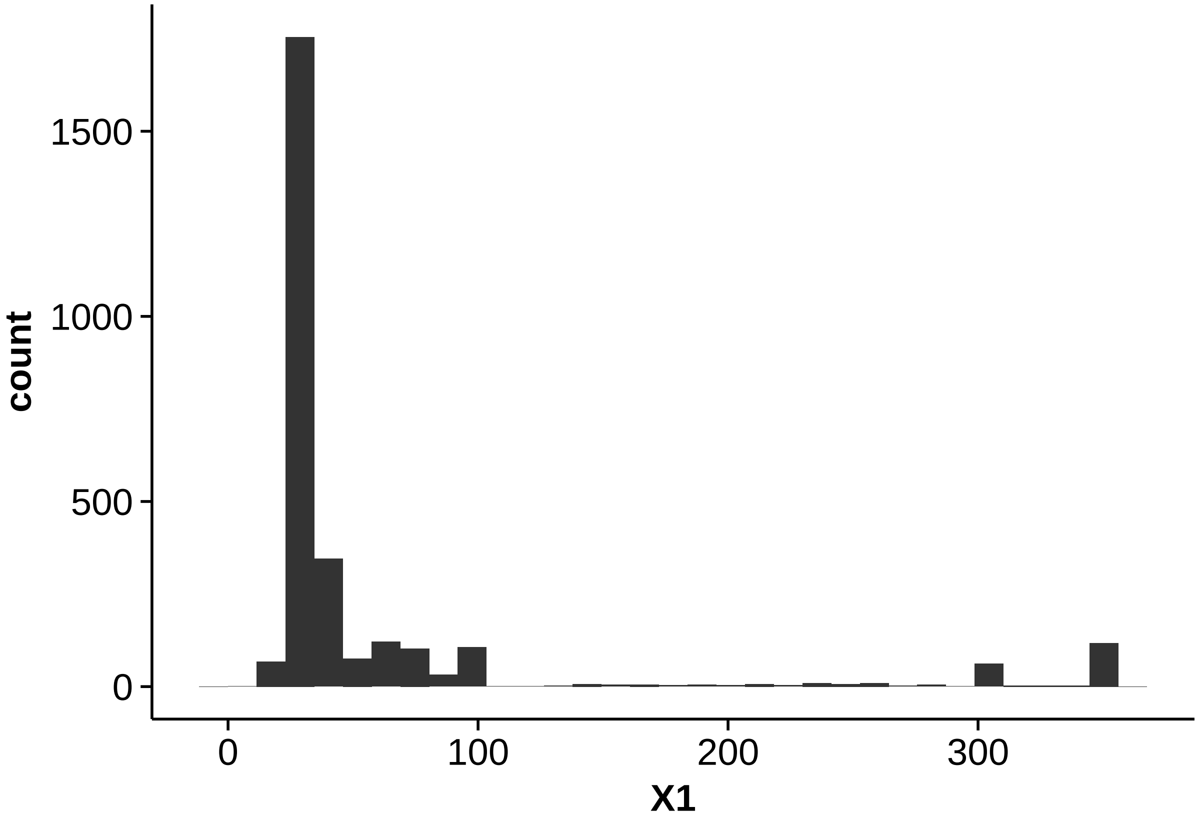
likelihood ratio test was performed by:

$$log(\Lambda) = log(\frac{max(\theta)}{\sum \theta - max(\theta)})$$

If the $log(\Lambda)$ is greater than some threshold, the concensus base is determined to be the base that has maximum likelihood.

log likelihood ratio threshold

To choose the optimal cut off for likelihood ratio, the data was subsampled to 250000 reads and distribution of $log(\Lambda)$ was plotted.



Since the distribution is almost a bimodal distribution, a threshold of 100 was chosen.