Next Generation Sequencing Technologies

Rob Mitra
2/3/21
How much do we need to sequence to call a genome?

Two considerations:

1. Error rate of the platform

2. Diploid Samples
Diploid Samples

- At a genomic location, a sample may be homozygous for the reference allele, heterozygous (1 reference allele plus 1 variant allele), or homozygous for the variant allele.

- Q: If we read out, for example, 4 reads that cover a genomic location and are reference, have we sequenced enough to rule out a heterozygous genotype at that location?
The Binomial Distribution

Let’s motivate this distribution by our example:
Q: If we read out, for example, 4 reads that cover a genomic location and are reference, have we sequenced enough to rule out a heterozygous genotype at that location?

Assume the sample is heterozygous: R, M
What is the probability of drawing 4 bases and coming up with "R" each time? It’s a coin flip: \((1/2)^4\)
The Binomial Distribution

• But what if we want to ensure that we read out at least 3 copies of each alleles at a given position? Now the math gets a little trickier, and we need to use the binomial distribution.

• What we want know first is if we flip a coin n times, what is the probability of getting exactly k heads?
Binomial Distribution

- Binomial distribution $P(k,n,p) = \frac{n!}{k!(n-k)!} (p)^k (1-p)^{(n-k)}$

- Derivation
Our original question

• But what if we want to ensure that we read out at least 3 copies of each allele at a given position?
Oxford Nanopore

Chang Liu, Arnav Moudgil
Jessica Hoisington-Lopez
Oxford Nanopore
Oxford Nanopore
Oxford Nanopore
Data Workflow – (5-mer example)

- Hidden Markov model
- Only four options per transition
- Pore type = distinct kmer length

ONTI: CCGACTCCGGTTACCCGCGTGGATTGCTTGGGCCAGGGCCG
REF: CCGACTCCGGTTACCCGCGTGGATTGCTTGGGCCAGGGCCG

- Form probabilistic path through measured states, currents and transitions
  - e.g. Viterbi algorithm
Oxford Nanopore

- ONT now has > 99% base accuracy

- Long reads 20-50kb, but some very long reads are possible (hundreds of thousands of bases or even megabases)

- Small footprint.

- Many real-world applications: Covid testing, HLA typing, field tests etc.
Applications of Next-Gen Sequencing

Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets

Graphical Abstract

Drop-seq single cell analysis

Authors
Evan Z. Macosko, Anindita Basu, ..., Aviv Regev, Steven A. McCarroll
B Barcoded primer bead

PCR handle  Cell barcode  UMI TTT(T27)
Rest of protocol on board

• Template switching (MMLV adds CCC) oligo rGrGrGrG

• Smart Smarter trick.
Some issues/questions

• Drop out

• How many transcripts per cell?

• Differential Expression

• Cluster and win
Conclusions

• Most applications -> Illumina

• De novo assembly -> PacBio

• Understand Poisson

• Understand how reads are mapped back
  – Hashing
  – Arrays
  – DP