PBG 200A Notes

Sam Fleischer

November 23, 2016

1 Genomics for Evolution and Ecology

Why?

- pop gen
- evo and eco genetics
- phylogeny reconstruction
- speciation genetics
- phylogeography
- molecular evolution
- comaprative genomics
- metagenomics

How? Tons of data. Because gathering data was difficult, a lot of thought was put into designing the perfect experiment. Today, biology is data-rich. Data is cheaper than brain capacity. A lot less work goes into experimental design.

It used to be really expensive.. 3 billion dollars. Now we can sequence a human genome in 2 weeks for 10,000 dollars. DNA is extremely dense information. It's cheaper to keep the DNA and sequence it than it is to house the DNA in a hard drive...

2 Sequencing technologies

what matters?

- output capacity
- scalability (smallest/biggest possible experiment)
- read length
- accuracy/error profiles
- speed
- ease of prep
- applications.compatibility
- cost
- availability
- tradeoffs - choose the best for what you actually need.

3 Workflow

- get material (DNA, RNA, etc)
- prepare sequencing library
- sequence it
- manage/store the data - convert to raw sequence
- pre-process the data
- data reduction / assembly
- primary analysis (identify basic units)
- do science

4 Types of sequencing libraries

- DNA/RNA
- RNA
  - normalized/unnormalized, full-length vs endtags, stranded or nonstranded
- DNA
  - single-end frags
  - paired-end frags
  - mate-pair
  - strobed
  - whole genome amplification?

5 Basic library prep

- fragment the DNA - chop up into manageable pieces (acoustic shearing - ultrasonic vibrations break it up into pieces) or (enzymatic - generates double-stranded breaks)
- size selection (decide what you need and get it) - we want the fragment distribution to be as tight as possible - we waste a lot of material
- end-repair (fragments have ragged ends - must repair them), A-tail, ligate adapters (tech specific) (must be able to manipulate the fragments) very very important
- adapters allow capture/manipulation, ID, sequencing
- enrich and/or capture adapter-ligated fragments
- quantify
- load on the machine

6 Basic Illumina approach

- requires adapter ligation and 2 PCR steps
  - pre sequencing (adapter enrichment)
  - on the machine (cluster generation)
- many enzymatic steps
- consequences: losses, duplicates, biases, errors

- sequencing by synthesis (multiple fluorophores)
- modifications: paired end, mate-pair libraries
- - start with RNA
  - convert to double stranded cDNA
  - add adapters on the end of the fragments (different adapters to each end)

7 Improved library prep: transposomes

- make transposomes - get fragmented DNA with adapters on the ends
- advantages
  - less input
  - single tube reaction
  - smaller volume
  - faster prep
  - fewer steps (fewer losses)

8 mate pair libraries

- necessary for complex templates with many repeats
  - many repeats are 10-12kb, but we sequence in 100 base pair size frags
- Illumina mate-pair approach can get around long repeats
- Cre/lox site-specific recombination

9 Illumina sequencing Bridge Amplification

- same principle as a florescent microscop
  - attach primers to fragments
  - attach fragments to machine
    - vast majority of adapters are empty
  - Convert to double stranded
  - repeat until you get clusters which can generate optical signals
- Error rate in Illumina technology gets larger with the length of the fragment.
- Not sequencing a single molecule - rather, a cluster of 200-basepair bits

10 Coverage - How much material do we need?

- 1pcg $\approx$ 1Gb
- Human
  - $\approx$ 6Gb (2n)
  - 100 ng $\approx$ 20,000 cells
  - 20 mcg - 1mg of tissue
  - coverage?
    - HiSeq (full run) $\approx$ 100x .. 2 weeks
    - MiSeq (full run) $\approx$ 2x .. 2 days
- Drosophila $\approx$ 0.18Gb
- Some plants > 100Gb

11 454 Sequencing

- First practical tech
- attach identical frags to a bead
- generate an emulsion?
- pipette the beads - every well is only big enough for a single bead
- end up with an environment where most are empty
- take a bunch of picture of the bead.. generate the sequence from that
- adding nucleotides generates a flash of light which is recorded.
- repeat nucleotides limit the length of reads - get a framechift mutation.. not good. Still, we can get up to 1000bp. (mode 700 bp, typical throughput 700 Mb)
- reads per run ≈ 1,000,000.
- run time 23 hours
- used to be competitive when Illumina was reading length 23bp. But now Illumina is preferred since it can read 200bp.

12 Ion Torrent

- High resolution Ph mirror?
- essentially measuring Ph in individual wells
- same kind of principle as 454.
- 50 Mb - 1.2 Gb chips (dispoable)
- read length 200-400 bases
- 2-6 hours (fast)
- cheaper and easier to operate
- 6 hr prep, 96 off the shelf barcodes
- load it on to a donkey and get out in the field
- all of the same problems as 454 (repeat basepairs)

13 PacBio: Single Molecule, Real-Tiem sequencing

- no longer has to generate clusters (always stays focused)
- DNA is not immobilized.. polymerase is immobilized
- in principle, a 100,000 basepair sequence can be read one nucleotide at a time
- based on the retension of nucleotides by the polymerase
- fragment long double-stranded DNA fragments
- tack on loop-shaped adapters to make a circle.
- needs lots of DNA (5 micrograms of DNA) this means a single Drosophila is not enough - need 30.

14 Coming up: single molecule nanopore sequencing

- motor protien breaks double-strand
- threads it through a pore
- reads electrical current
- scalable modules of 2000-8000 pores
  - 20 modules = human genome in 15 minutes
- $40 per Gigabase ($3600 per 30x human genom coverage)
- Prototype - current error rate 15-20%
- mini disposable minilop for $900 (plugs into USB port)

15 High-Capacity

- a blessing and a curse
- what do we do with a 600 Gb sequence? Or even 1 Gb?
- barcoding allows multiplexing
  - amplification primers or adapters
  - error-correcting
  - one or both directions, redundant or combinatorial
  - balancing adapters for quality calibration
  - adapter biases, empirical testing
  - 454 also uses gaskets to split the cells
- also useful for library titration and read quantification