PBG 200A Notes

Sam Fleischer

November 28, 2016

- Genome sequencing
  - physically break up the genome, clone them, order the clones, sequence each clone, made sense to everyone, but they only got 3% of the human genome after 10 years.
  - or shotgun sequencing, which is break up into random pieces.
- Sequence assembly
  - de novo assembly
    - requires longer reads, greater depth, low polymorphism
  - reference mapping
    - must assume you already have a genome sequence (reference genome)
    - use reference genome as a template
    - allows shorter reads, lower depth, tolerates polymorphisms
  - RNA to DNA mapping
    - quantification of gene expression
    - detection of splice junctions
    - alternative splicing
- Genome sequencing (shotgun) - basic ideas
  - high coverage (redundant)
  - if mean coverage is 1×, not going to happen since you need overlap. Even 10× is not enough.
  - the higher the × coverage, the greater the chances that every read will overlap with at last two other reads. For assembly to be contiguous, you need very high coverage (30×, 50×, and higher).
  - The shorter the reads, the higher coverage you need. The shorter the reads, the shorter the overlap, the higher the coverage you need.
    - the fundamental tradeoff: high coverage, short reads vs. low coverage, long reads
  - Eukaryotic genomes have repeates.. this is a problem. If the longest repeat is longer than the longest read, no way to get contiguous assmebly. There are some repeats which are 12 kilobases long.
- Second-generation technology
  - two steps
    - assemble unambiguous reads into contigs that end at repeat boundaries
    - combine contigs using bridging data
  - a scaffold is a collection of contigs linked by mate pairs. Gaps can represent repeats or true missing data.
  - You want long mate pairs
- de novo assembly
  - overlap consensus algorithms
    - compares reads, finds overlaps (first finds identical seeds) and extends if possible
  - de Bruijn graph algorithms (VELVET, ASySS, SOAPdenovo)
    - breaks up reads into shorter (every possible) k-mers.

        - every read gives an enormous database of k-mers.
        - connects reads which contain identical k-mers
        - collapses graphs into contigs
    - scaffolding
        - uses additional info from paired end and mate-pair libraries
        - this is where a lot of mis-assembly happens.
        - current tradeoff: contiguous vs. accurate.
    - issues
        - highly computationally intesive (not tractable for small k)
        - sensitive to polymorphisms, errors, non-random fragmentation, non-uniform coverage, repeats
        - polymorphisms are the biggest problems
            - slightly overcome this through inbreeding

- how good is the assembly (measure the quality)
    - N50
        - 50% of the entire assembly is contained in contigs that are at least as long as N50. The higher the N50, the more contiguous the assembly.
    - it is possible that 80% of the genome is missing
    - it is also possible that the assembly yields a longer contig than the size of the genome!
    - how good do we need the sample to be?
        - the difficulty and cost of genome assembly increases exponentially as we approach perfection.
            - must decide what the goal is BEFORE you decide the technology for the job

- reference mapping
    - much much easier than de novo, but assumes we already have a reference assembly
    - used to study intra-specific polymorphisms
    - not used much now since it is expensive and not much better (if at all) than de novo. Still may be better in certain circumstances.

- combining data
    - hybrid strategy
        - gets lots of short and some long reads

- whole genome amplification
    - processivity of up to 100kb
    - low error rates

- how to deal with high error rates
    - hybrid assembly
        - use lots of short reads to correct long reads
        - but use long reads to assembly
    - self-correcting assembly
        - overlapping long reads correct each other
        - PacBio is (basically) random error, so concensus will smooth out errors
    - additional tricks
        - multiple pass reads of PacBio “barbells” (doesn’t work that great in reality due to longevity of polymerase)
        - 2D reads with Nanopore (double strand connected to itself via loop on one end)

- long read assembly approaches
    - hierarchical

        - long read high error to long read low error by using short read low error
    - scaffolding/gap filing
    - read threading
        - graph structure resolves bubbles in scaffolding

- example - Drosophila serrata
    - Illumina
        - reads and coverage 200×
    - PacBio
        - reads and ocverage 63× (expensive for PacBio)

- Irys BioNano physical mapping
    - Get a visual fingerprint (light-based) to match up the molecules

- (from DNA to RNA) Transcriptome sequencing
    - RNA isolation
        - what sample do you want and why?
    - mRNA purification and cDNA synthesis
    - normalization (if needed)
        - nowadays sequencing is so cheap that we don't realy need this.. might as well just sequence the hell out of it.
    - non-stranded or strand-specific sequencing
        - determine which strand is expressed
        - avoid chimeric contigs from overlapping 3' UTRs.