PBG 200A Notes

Sam Fleischer

November 30, 2016

- Transcriptome sequencing - what matters?
  - want high-complexity: reads with varied starting points
  - want even coverage: low coefficient of variation.. all positions are covered evenly. Uneven coverage comes from bad RNA or bad priming.
  - there is potential to make stranded libraries
- Transcriptome assembly
  - what do we want to know?
    - genome annotation vs expression analysis
    - map the transcriptome to the genome sequence - find where the genes start and end - only a fraction of the genome is expressed - also, different sexes, and under different environmental conditions, express different genes.
    - It is easy to get most of the parts of most of the genes. It is very very "hideously" difficult to get all of the parts of all of the genes. Even the best model organism, Drosophila melanogaster, doesn't have a full gene sequence.
    - Do we need full-length transcripts? If a gene is only expressed in a single cell type, it is not important for all questions
  - de novo?
    - can be done in any nonmodel organism
    - Coverage at 5' ends will never be as good as at the 3' ends.
    - Assembly depends on coverage. The main problem is repeats. The repeats make up most of the genome, but very little of the transcripts. The problem is coverage. Some genes will be sequenced at very high/low coverage (differ in 4 orders of magnitude).
  - or reference mapping?
    - mapping the transcriptomes to existing DNA data.
    - gene structure
    - more easily get to transcriptome start site
    - better to do, if possible
- de novo transcriptome assembly
  - commonly used package: "Trinity"
  - assemble reads into contigs
  - another program tries piece contigs into clusters by finding overlaps
  - use reads or pairs of paired-end reads to get transcript isoforms. This is only as high-quality as the RNA sample.
- transcriptome assembly by refernece mapping
  - take RNA paired-end reads. either,
    - align then assemble
      - gapped alignment
      - use that to identify gene models

        - or assemble then align

- RNA to DNA mapping
    - the measure used to quantify abundance is (RPKM) “reads per kilobase per million reads” or (FPKM) “fragments per kilobase per million reads”
    - all of this assumes 2nd generation short fragment
    - 3rd gen will eliminate the process of assembly
- Anopheles phylogeny - sequencing and assembly
    - benefits of not normalizing
        - assembly is as good as coverage. This was seen as a limitation. However, the uneven representation is a feature when constructing a phylogeny
    - how much depth do we need? Maybe not that much. Dependent on the number of taxa? The more taxa, paradoxically, the more coverage you need.
    - orthology assignment
    - Even back in the middle ages (2010) they got more data than needed.
- Anopheles phylogeny - science
    - how many loci do you need? What do you want them for?
    - What’s the effect of using only highly conserved CDS?
    - Taxonomic range?
    - multispecies coalescence brings up gene-tree and species-tree resolution.
- Reduced Representation approaches
    - Sequenc capture methods
        - isolate only parts of the genome you want to resequence
        - must already know the sequence.. only suitable for resequencing applications. Must have a reference sequence.. decide which parts you want. Then you can capture the same parts from many other individuals or species or whatever.
        - If you know the sequence in a single individual, it is easy to capture the same gene in other individuals of the same species.
        - methods:
            - multiplex PCR
            - micro arry hybridization and release
            - molcular inversion probes
            - solution hybrid selection
        - factors: capacity, scalability, mismatch tolerance, cost, input
            - most of these don’t scale easily below 1 megabase
            - Is this something that will be screwed up with high levels of polymorphisms?
            - amount of imput material varies between micrograms and many milligrams
    - Multiplex PCR
        - target each individual and then mix, or do multiplex PCR
        - barcoding allows multiplexed sequencing.. disentangle sequences bioinformatically
        - PCR is a pain even for a single gene - 10,000 genes is especially difficult. Multiplex is even harder.
        - Hard to PCR something over 10 kilobases
    - RainStorm PCR
        - microfluidic device that can run many PCRs in parallel using microdroplets
        - Can get 10-20 thousand pair of loci in one run of the machine.
        - Either validate each one, or accept the high rate of failure.
        - Very costly

  - Molecular inversion probes
    - design probes to capture parts of DNA in a circular fashion
    - works well for shorter bases
  - Microarray-based genomic selection
    - first invented to study gene expression
    - Immobilize the probes, fragment the DNA
    - Then wash away everything else.. there's our library
    - This was done for a while until people realized they didn't need the microarray - you can do this in solution
  - Solution hybrid selection
    - instead of making immobilized probes, get a series of RNA molcules which correspond to the region of genomic DNA you are trying to caputre.
    - Then biotinylate the RNA probes
    - RNA probes will find desired DNA.
    - Add magnetic beads that bind to the RNA probes
    - wash away what you don't want (non-magnetized)
    - strip the RNA
    - This is by far the most commonly used approach - scalable, flexible
      - SureSelect™Target Enrichment System Capture Process
  - Pros and Cons
    - PCR is too much of a pain
    - Microarrays are extinct
    - Inversion probes are not yet good enough
    - That leaves solution hybrid selection, but it is expensive. One problem is that this is only "cheap" for big studies. Consider the cost of sequencing vs. cost of sample prep.
    - How much material? Micrograms.. but there are low-input methods for nanograms.. wet bench procedures are much more complicated

- Prum 2015 Technology
  - Sequence divergence is more than could be handled by straightforward applications - chickens and crows are too different.
  - Used parts of two genomes of distantly related birds but highly conserved. The fragmented DNA actually goes out from that conserved part.
  - Never getting pure isolation. Rather, an enriched sample - there is always strong contamination from nontargeted genomic regions.
  - Use reference mapping to separate the data set and then running de novo assembly for each of these references.

- Prum 2015 Science
  - Lots of organisms with large genomes
  - more taxa or more characters? Different question, different answer.
  - biological problems
    - add more characters and more taxa in the difficult part of the phylogeny.
    - lineage sorting / gene tree conflict - they just averaged over gene trees.. this could be disastrous (next quarter)
    - Math - talk to Bruce Rannala - not yet computationally feasible