PBG 200A Notes

Sam Fleischer

December 7, 2016

- Allele-specific expression (Coolon 2014) - technology
  - advantage of RNA-seq over other techniques for ASE
  - Sequenced different experiments on different lanes - maybe confounded biological and technological differences
  - The longer the reads, the greater the chances that any given read spans a site that has fixed alleles in two different species
  - they did controls and validation with pyrosequencing since it was the first paper of its kind
- Coolon 2014 - Science
  - assumptions of cis/trans decomposition
    - no transvection
    - no allele-specific cis-trans interactions
  - using mixed tissues - effects on gene expression levels - allometric changes?
  - only females since the X chromosome is the 20% of the genome
  - Taxonomic range - a balance between hybrids being possible and having distinct alleles
- Metagenomics and Community Genomics
  - approach: characterize the structure and function of ecological communities based on the genomes of their members, without explicitly working with or identifying the actual organisms
  - advantages
    - rapid and high throughput
    - this is only appraoch for non-culturable organisms
  - primarily used for cryptic (microscopic) members of ecosystems
- Approaches to community genomics
  - culture-independent census
    - identify all community members without knowing what any of them might be
    - compare different communities
    - identify candidate factors that may be shaping these communities
    - monitor changes in community composition in response to biotic or abiotic changes
  - should be completely unbiased and quantitative - this hard to achieve - usually focus on a particular lineage of bugs rather than entire community
  - Metagenomics - What can we do?
    - assess the collective metabolic capability of the community
    - compare it between different communities
    - identiy ecological correlates of metabolic capabilities
  - caveats: hard to get unbiased data, can miss rare taxa, hard to assign functions to specific organisms, many functions are unknown ("no freaking clue what they do!")
- culture independent census

  - most microscopic members of most ecosystems are unknown and unculturable, can't find them because don't know how to look for them, and couldn't isolate them even if we knew them. Instead, use a culture independent approach:
    - isolate total DNA of who community
    - amplify (standard PCR) a "universal" gene that is present in all taxa
    - identify community members based on sequence
    - quantify abundance, determine community composition
  - Caveats:
    - no truly universal markers actually exist
    - quantitative biases in DNA isoltation (different isolation procedures produce different outcomes), PCR, etc.
    - many sequences are novel, cannot be identified

- 16S rRNA sequencing
  - closest thing to a universal bacterial marker (we think)
  - base primers on conserved regions, identify taxa from variable regions
  - extensive databases and many previous surveys enable easy identification of most taxa from their 16S sequences
  - read length is main limitation: typical tradeoff with depth important for quantitative ecological analysis
  - can't assemble since there are 1000s of taxa in a single read - the unit is a fragment
  - have to barcode and multiplex for good design / lower cost
  - quantitative biases in DNA isolation and amplification and varition in copy number
  - 16S sequence identity does not reveal functional capabilities
    - Genome and functional capabilities can vary greatly within OTUs defined by ribosomal sequences (same caveat for any other single gene)

- metagenomic sequencing
  - shotgun genome sequencing of environmental DNA samples
  - advantages
    - identifies many of the functional capabilities of the ecosystem without relying on culturing
    - important because many of these capabilities are provided by the unculturable members of the community
    - dynamic pictures of changing ecological communities
  - problems
    - hard to assemble genomes from shotgun environmental sequences - especially rare taxa
    - hard to match functional capabilities to their owners (somebody here has genes for nitrogen fixation, but I don't know who it is)
    - many genes are novel so their functions remain unknown
    - usually not enough DNA - have to use WGA (more biases)

- annotation of metagenomic data
  - identifying OTUs based on universal genes
  - identifying biochemical capabilities
    - clusters of orthologous groups (COGs)
    - non-supervised orthologous groups (NOGs)
    - KEGG pathways

- genome assembly from metagenomic sequences
  - the most abundant and distinctive members can be assembled directly; closely related taxa are difficult to distinguish - long tail of low per-nucleotide coverage
  - for the remaining reads, try binning them before assembly
    - by GC content

        - by tetranucleotide frequencies
        - by read coverage
        - by BLAST similarity to known taxa
    - then try to do a de novo assembly on each bin
    - hybrid approach: generate reference genomes for some members of the community; imited to cultuable taxaa but yields the best insights in the end

- metatranscriptomes
    - metatranscriptome will be even more dynamic than metagenome (what genes are being expressed now, here, in this environment)
    - can study rapid changes in response to biotic and abiotic factors
    - best if combined with metagenomic data - do metatranscriptome changes reflect a changing community, or functional changess in the same community?

- Permafrost metagenome - technology
    - PCR results in sequence bias - emulsion PCR minimizes competition - really important since different taxa have drastically different GC content - resulting in huge PCR bias
    - 95% of the data remained unassembled
    - today, we would be able to assemble 10% instead of 5%. We would need 100% assembly to get all the taxa.
    - 3rd generation technology will be useful if it can be made cheap enough. 40Gb of data would be "bloody expensive" - but theoretically this is the data we need.

- Permafrost metagenome - science
    - guess the ecology of specific taxa - someone can do carbon and nitrogen fixation - also have methanotrphic bateria
    - community similarity / changes in composition under changes in environmental conditions
    - role of uncultivated taxa
    - correlating community function and composition
    - *functions* converged, not taxonomy - essentially, different taxa capable of the same function increased in abundance

- HGT study - technology
    - source of data / sampling strategy - everything they do in this paper is completely dependent on someone, somewhere, culturing all of these taxa
    - recency of HGT

- HGT study - science
    - unlike the last paper, this entire completely depends on 100% confidence of assigning taxa to gene fragments
    - ecological similarity means the opportunity to share genes