PBG 200A Notes

Sam Fleischer

December 5, 2016

- SNP array genotyping
  - based on DNA-DNA hybridization
  - need to know where the variable sites are in an organism
    - multiple prob pairs for each variable nucleotide
  - the individual who carries a C(A) will hybridize better to probes with G(T). Measuring differences in signal intensities.
  - very precise, very cost-effective, on a per-individual, per-SNP basis
  - only is useful in model organisms
  - For $600, get 1 million alleles in 1 individual.
- Bead array genotyping
  - most common for nonmodel organisms,
  - but have to have a reference genome and know the variable sites and alternative alleles at a site
  - need the same info for nonmodel and model organisms, but for nonmodel organisms you need to make your own reagents.
  - primers attach to reverse-primers which have their own barcode. Reverse primers attach to beads, so each type of bead corresponds to a single allele. Plate-reader fishes out the beads and immobilizes them in different locations on a plate. Image the plate for the different tags.
  - advantage: flexibility.
    - $.2 - $.5 per SNP per genotype (48-384 SNPs)
    - minimum experiment: approx 500 individuals, $10,000-$15,000.
    - high requirements for DNA quantity and quality
- Sequenom MassArray genotyping
  - have to know a reference sequence (something)
  - need to know 100-200 variable sites
  - do PCR for every locus
  - no tags, no beads, directly reading the sequence with a mass spectrometer (must be able to determine each allele from each allele in all sites)
  - main limitation is SNP compatibility
  - 30-36 SNPs per plex.
  - flexible: suitable for gradual experiments (feeds information a little bit at a time, can improve experimental design on the fly)
  - main limitation is SNP compatibility
  - $.15-$.25 per SNP per genotype
  - requires less DNA and less quality
- Genotyping by Sequence
  - Do reads for k individuals in a single pool
  - separate bioinformatically later on

  - assign ancestry to each individual at each location

- Multiplexed shotgun genotyping
  - essentially getting unlimited info
  - assuming the individuals have high LD
  - if the density of variate sites is much less than the density of recombination break points.
  - this means we don't need high coverage - we don't have complete genome
  - can be very cheap on the per individual basis - can multiplex multiple individuals on a single Illumina lane.
  - very cheap - $20-$40 per individual fo thousands of markers
  - disadvantages:
    - need to have fairly well-assembled reference genomes, and info on each parent,
    - high upfront costs(barcode adaotors - need a single barcide for each individual). The same 100 or so barcodes can be reused. Scalaing up is easy.
  - Key point: not using genetic information to infer a genetic map. Using physical genome as reference.

- Bulked segregant analysis
  - Basic principle: identify differences in the frequencies of marker alleles between pools of individuals with different phenotypes
  - In the absence of population structure, this is useful (do we want most or least diverse group?)
  - can be done in lab or in natural populations
  - limit by scale of LD
  - Genetic designs for BSA
    - controlled lab cross
    - hitchhiking mapping (select for phenotypes, see which genotypes come with it)
    - introgression mapping
    - Natual poulaition

- Genotyping by sequencing (Gar paper): technology
  - a lot easier and simpler - can't use it for genome mapping without a refernce equilibrium
  - this was an excellent example of how to exploit the available technological resources
  - pseudogenomes (updated genome) - mapping to a semi-related genome - assume the genomes are close enough, and that they are more or less colinear. Map the reads to the semi-related genome - very high error rates, but fine for their purpose
  - masking polymorphisms - throwing out over 90% of data, still end up with hundreds of variable site per Mb.

- Measuring gene expression: microarrays vs. RNA-seq
  - what genes are expressed where, when, and in what amount
  - determines the function of the genome
  - look at gene expression to determine the link between genotype and phenotype
  - Microarrays
    - immobilize probes on solid support
    - "analog" technology
    - cost-effective when good arrays already available,
    - but custom arrays can be made for any organism
    - has almost completely been displaced by...
  - RNA-seq
    - "digital" technology
    - can be done for any organism, but requires a reference for mapping reads
    - assuming the number of reads you observe in an RNA sample is proportional to expression

    - don't need an organism-specifi reagent, but do need a reference genome,
    - but can make a custom reference (which is a helluva lot easier than making a custom array)
    - got cheaper fast, now even cheaper and easier

- Design and analysis of expression experiments
  - interested in differences between treatments/categories
    - brain/liver, grain/rice, male/female, etc.
    - must account for variation within treatment first!
  - biological variation within treatment
    - genetic differences, environmental conditions, etc.
  - technical variance
    - effect of experimental procedures, dissection, batch of arrays/dyes,reagents, ozone levels that day, etc.
    - Artyom knows a statistician who swears she can look at two data sets and determine which Illumina machine it came from..... wtf.
  - biological replicates necessary to detect differences
    - avoid confounding biological and technical variation
    - know the technical properties of your method
    - anticipate sources of biological and technical variance

- design of RNA-seq experiments
  - unlike microarrays, RNA-seq provides categorical (counts) data
  - in principle, can detect "differential expression" from a single replicate per treatment (conditioned on depth)
  - does not tell you whether you are looking at true biological variation
  - for quantitative experiments, need multiple biological replicates per treatment, just a with arrays
  - minimize environmental and technical variation
  - *avoid confounding biological and technical variation*
  - example of a bad design:
    - using different reagents, different Illumina machines - completely confounding biological and technical variation
  - example of a good design:
    - use barcodes!
    - isolate RNA - use the same batch of reagents for each library
    - can mix all six biological replicates because of barcodes
    - put the mix on six different lanes, minimizing technical variation
  - the importance of this depends on the level of variation you are measuring
  - may be looking at subtle quantitative difference

- problems with RNA-seq quantification
  - standard Illumina procedure includes many enzymatic steps with purification in between
  - losses and biases at every step
  - OK if looking for big differences; may be a problem if looking for subtle quantative variation

- Allele-specific expression analysis
  - Differences in gene expression between genotpes can be due either to mutation in that gene (cis-regulatory) or any changes in its upstream regulators (trans-regulatory)
  - For some applications, we don't care (health)
  - In evolutionary terms, however, it makes a big difference. How do we figure out where the divergence is based on cis- or transregulatory variation
  - F1 hybrids can be used to infer cisregulatory variation
  - By comparing two parents and their F1, can decompose gene expression divergence into cis- and trans0regulatory components