PBG 200A Notes

Sam Fleischer

December 2, 2016

- RAD (restriction site associated DNA) tags - basic idea
  - can capture fairly reproducible parts of the genome without any a priori information
  - can roughly control the number of regions
  - radically simplifies genome assembly
    - drastically reduce complexity of sample
    - anchoring of sequencing reads at predictable sites
  - two levels of polymorphism
    - presence/absense of tag
    - sequence variotion in tag
  - Sticky ends

```
GAATTC
CTTAAG
  ↓
GAATT
C
```

- Experimiental design
  - main tradeoff: # of individuals vs. noumber of markers (for given depth $)
  - how much depth do you need?
  - length and configuration of reads
  - main cost: RAD adaptor oligos
  - multiplex at an early stage to save money (this vastly simplifies the process)
- Genotyping and Genetic mapping
  - goals:
    - identify the genes responsible for natural variation
    - examine distribution of variation in time/space
  - approaches:
    - marker identification and genotyping of recombinant progeny
    - direct genotyping by sequencing
    - individual genotyping vs Bulked Segregant Analysis
    - Choice of approaches depends on the model, question, and resources.
- Conventional High-throughput genotyping
  - basic strategy
    - identify genotypable polymorphisms prior to mapping
    - obtain individual recombinant progeny
    - genotype them for the polymorphisms of your choice
    - do conventional QTL mapping

  - In principle, does not require any prior resources (maps, genome sequence, organism-specific reasents, etc.)
  - Genetic design (how you get recombinant progeny) and genotypic approach (how you genotype) are completely separable

- Polymorphism identification
  - source of variable genotypes?
    * a pair of highly inbred parental strains
    * a pool of outbred individuals from the wild
    * anything in between
  - type of variable sequences?
    * complete genome sequences
    * partial transcriptomes
    * RAD-tags, reduced representation libraries, etc.
  - identification of variable sites and alleles
    * reference-map reads from each genotype to genome or transcriptome
    * assemble reads de novo, identify SNPs in assemblies
  - How many do you need?
    * limited by marker density or X-order density?
    * scale of LD
    * this determines sequencing strategy
  - selecting polymorphisms for genotyping
    * if sequencing parental lines of cross - choose fixed
    * if sequencing a population panel - choose high frequency
    * find real polymorphisms, not sequencing errors (effect of sequencing depth)
    * do we want to validate before use?
    * tech-specific requirements

- genetic design
  - controlled lab crosses
    * identify parental lines that differn in phenotype
    * cross, get lots of individual F2 or more advanced Xovers
    * genotype for SNPs known from parental lines
    * Pros: complete control, easy analysis, non hidden population structure
    * Cons: high LD, low resolution, have to do crosses
  - GWAS in natural populations
    * collect naturally polymorphic/recombinant genotypes
    * genotype for random naturally segregative markers
    * Pros: low LD, high resolution, no lab crosses
    * Cons: Messy data, false LD (population structure, drift, etc.)
  - Hybrid strategy
    * low-res QTL in lab cross and GWAS in nature: simplifies analysis of GWAS data

- high-throughput genotyping techs
  - SNP chips
    * largest number of marker loci (millions)
    * lowest cost per SNP per genotype, high total cost
    * suitable mainly for big experiments in model species, i.e. medical experiments in humans/mice
  - bead arrays
    * moderate number of marker loci

    - medium cost per SNP per genotype, medium total cost
    - equally useful in model and non-model species
  - mass-spec arrays
    - relatively small number of loci
    - cost per SNP per genotype roughly similar to bead arrays
    - can be used in any organism
    - useful for progressive genotyping strategies
    - can be scaled down more easily
- RAD-tag genotyping
  - start with some panel of recombinant progeny
  - RAD-tag individual progeny with personal barcodes
  - sequence them together in a pool (multiplex), assemble them separately
  - identify variable sites and call SNP alleles
  - sequence same RAD tags from both parents to identify parental haplotypes
  - detect recombinants, reconstruct linkage maps, map QTLs
  - does not require prior genome
  - requires investment in barcodes