Day 5: Intro to Population Genomics

Monday, September 8, 2025 9:37 PM

Population Genomics -- *the study of genomic diversity within species and the processes that shape it*

- **Some big picture questions population genomicists ask:**
  - *How much diversity is present within a species?*
  - *How have historical events like bottlenecks, range expansions, or fragmentation shaped that diversity?*
  - *Do populations from different environments experience different selection pressures, and has this led to different patterns of adaptation?*
  - *What is the genomic basis of adaptive differentiation between populations (i.e., genetic architecture)?*

- Generally distinguish between **neutral processes and non-neutral (selective) ones**
  - Neutral:
    - Population structure (spatially distinct pops)
    - Gene flow
    - Demographic history (bottlenecks; range expansion)
    - Admixture/hybridization
  - Selective:
    - Directional selection (selection on new mutations in large and well mixed pops --> leads to rapid fixation of adaptive alleles)
    - Local adaptation (spatially varying selection that leads to pop-divergence in adaptive alleles/traits)
    - Other types of selection
      - Balancing selection, Purifying or negative selection
- Pop size controls the amount of diversity and the rate at which it is either lost or becomes fixed
  - Small populations are slower to introduce new mutations *and* are more subject to genetic drift (allele loss or fixation) and less responsive to natural selection.
  - But what is a "population"? This means different things to an ecologist or demographer vs. a geneticist
    - In ecology, a population is a group of individuals of the same species living in the same place (N)
    - In genetics, a population is a group of individuals with a shared history of past and ongoing genetic exchange and experiencing similar processes of drift and selection
    - "effective population size" or "Ne" -- this is the size the population behaves in terms of evolutionary processes of drift and selection, and is almost always smaller than the census size (N)
- **How has genomics transformed the study of these processes?**
  - Greater precision for detecting drift/demographic events
    - Evolutionary processes are noisy, and loci provide the statistical power in studies of demographic history (many samples of the evolutionary process); under NGS, the number of loci has increased by orders of magnitude
    - Linkage and linkage disequilibrium (LD) now becomes something to contend with…how independent are loci in providing samples of the genome? It depends on how much LD exists…
  - Detecting the genetic basis of adaptive traits!
    - Before NGS, only studies of model organisms could glimpse what genes or genomic regions were involved in adaptation or ecologically relevant traits
    - Now, we can use association studies to predict adaptation variation using thousands or millions of genome-wide loci
      - Richer understanding of the types of genes and variants involved (it's not just protein about coding sequence!)

Population genomics in practice: SNP and genotype calling from NGS data

- Population genomics requires findings SNPs (Single Nucleotide Polymorphisms) across the genome -- generally aim for thousands to millions of SNPs, depending on study's goals

- SNP calling vs. Genotype calling
  - SNPs are sites that are polymorphic (variable) within the sample -- use NGS and bioinformatics to try and detect which sites are polymorphic
  - Genotypes are what alleles an individual carries.

. .
        - For diploids, there are 3 genotype classes
            - Hom1, Het, Hom2; AA, Aa, aa
        - Generally when aligning reads to a reference genome, we call genotypes as Ref (reference allele) and Alt (alternate allele):
            - Ref/Ref, Ref/Alt, Alt/Alt; 0, 1, 2 coded as numbers of copies of the alternate allele
- Pipeline steps:
    - (1) Clean reads (fastq files) to get rid of sequence adapters and low quality bases
        - Base calling errors during sequencing
        - Base Q-score (-10log(10)Probability of erroneous base call
            - Q10 = 10% error; Q20 = 1% error; Q30 = 0.1% error
    - (2) Map reads (sequence alignment files: sam and their binary version: bam)
        - map to reference if available (also called alignment)
        - map to de-novo assembly (need to make if reference unavailable)
    - (3) Sort and index mapped reads (sorted.bam, sorted.bam.fai)
    - (4) Call genotypes (variant call format files: VCF) or use GLs
        - Incorporates quality measures from mapped reads
            - Mapping quality (mapQ) -- calculated same as base Q-score (see above)
        - Sequencing depth (aka, coverage) is important!
            - DP is how many sequence reads mapped to a given site in the genome
            - Want high coverage (DP>10) to accurately call genotypes, but that's not always feasible
            - Medium coverage (DP 5-10) is often acceptable, but may contain some errors
            - Low coverage (DP 0.5-5) can be analyzed if using appropriate methods that don't try and determine genotype precisely but rather give a probability of each genotype occurring.
            - Genotypes called based on multinomial probability distribution
                - For diploids, expect roughly 1:1 ratio of counts for each allele
                - But for typical DP values (~10), it's easy to have 4:6 or 3:7 or 2:8 (or even 1:9) with some probability .
    - (5) Filtering
        - Goal is to quality-control the SNP data to filter out false SNPs or mis-called genotypes
        - Nuanced process -- trying to strike balance between getting lots of SNPs genotyped across most of the samples (issue of missing data)
            - Good metrics to filter on:
                - DP (depth),
                - Site (SNP) missingness (at a given SNP locus, what % of samples have data?)
                - Sample (Ind) missingness (for a given individual, what % of its SNPs have data?)
    - (6) Downstream Analyses
        - Diversity and divergence
            - Nucleotide diversity, heterozygosity
            - Site Frequency Spectrum (SFS) and Tajima's D
            - Fst, Dxy
            - LD (r2)
        - Genetic structure / population differentiation
            - Genetic PCA
            - Admixture
        - Selection
            - Fst outliers
            - Genotype-environment association (GEA)
            - Selective sweeps