**Glossary:**

**Ascertainment bias –** bias introduced by the sampling design that induces a nonrandom sample of observations. Examples: reference genome selection, genetic markers, selection of individuals

**Sympatric speciation –** process of divergence (speciation) between populations or species occupying the same geographical area and in the presence of gene flow

**Allopatric speciation –** process of divergence between populations or species that are geographically separated and in the absence of gene flow

**Linkage disequilibrium –** nonrandom association of alleles at different loci

**Islands of differentiation –** genomic regions of elevated differentiation owing to the action of natural selection

**Island model –** Model introduced by Sewall Wright to study population structure comprising multiple populations connected to each other through migration

**Metapopulation model –** idealized model in which several populations diverge without migration from a common ancestral gene poop (or metapopulation

**Allele frequency spectrum –** A distribution of the counts of SNPs with a given observed frequency in a single or multiple populations

**Genetic drift –** stochastic changes in gene frequency owing to finite size of populations, resulting from the random sampling of gametes from the parents at each generation

**Diversifying selection –** selection that favors different alleles in different parts of a species range

**Outline:**

**Sympatric speciation (presence of gene flow)**

* Diversifying selection can drive speciation but:
  + Gene flow (dispersal) brings alleles to where they “shouldn’t” be
  + If only looking at neutral alleles the population will appear homogenous
  + Recombination allows tightly linked genes to also experience selection if by an allele that is favored. Neutral alleles will also experience this if close to positively selection allele
* Overall, sympatric speciation can appear to be divergent via diversifying selection but also single gene pool for genes not under selection

**Benefits of NGS in inferring the history of divergence**

* Have information on all parts of the genome lets us have more accurate knowledge of demography
* Can ask whether some regions of the genome are exchanging genes more often than others (i.e. variation in gene flow levels is an indirect indicator of selection acting against gene flow in some regions compared to others)
* NGS allows us to get better estimations of recombination and linkage disequilibrium (used for population structure)
* Polymorphisms and LD across genome indicate selective sweeps and what genes are targets for diversifying selection

**Genome scans to detect divergence**

* Looking for islands of differentiation by looking at the distribution of summary statistics that measure genetic differentiation (Fst)
  + Regions with high Fst (high divergence) may be regions where genes are under divergent selection
  + Can also be used with metapopulations and island models
    - Nested island model (model where groups of populations have higher migration rates between them then other populations) NGS data can be analyzed using a similar method that also accounts for read depth
* Comparing gene tree to population tree to detect recent admixture using a specific site
  + Differences between population tree and gene tree can be due to incomplete lineage sorting (ILS) (shared ancestral allele) and gene flow
    - D calculates introgression of populations
      * Requires: genomes from both species, one from a potential source of introgression, and one for the outgroup to ID ancestral state
      * If ABBA and BABA topologies are approximately equal 🡪 D=0
      * D not equal to 0 is evidence of introgression
      * Distribution of D and scanning along genome can indicate regions that have experienced introgression
      * Demonstrates the importance of secondary contact (two divergent species can have a region of introgression caused by secondary contact)
* Limitations:
  + Summary statistics sets aside a lot of useful information
  + Same value can come from different causes (e.g. low Fst could be a shared ancestral allele or gene flow; D can be different from 0 and not because of admixture)

**Likelihood and model-based methods**

* Want to calculate the likelihood under a rich divergence model 🡪 allows estimation the most likely parameters given a model either with frequentist or Bayesian statistics
* Estimating likelihood of alternative models allows us to infer the most probable divergence model
* Two main approaches for studying divergence models:
  + Allele frequency spectrum
  + Sampling genealogies
* Allele frequency spectrum (AFS)
  + AFS has different patterns for different divergence models (especially how isolation differs from gene flow)
    - Absence of gene flow: frequency of SNPs in one population are different from the frequency of SNPs in the other population because genetic drift drives alleles to fixation
    - Presence of gene flow: many SNPs with similar frequencies between the two populations
    - Expected and observed AFS can be compared and the likelihood can be determined using a multinomial distribution
      * Limitations:
        + assumes all SNPs are independent but many are likely linked. Composite likelihoods address this but limited access to confidence intervals, etc.
        + reducing data to SNPs throws out information on linkage therefore these methods aren’t very sensitive to process that affect LD such as admixture and gene flow
        + Computationally challenging
        + Expensive for models with more than 3 populations
    - Has been success when considering SNPs only at synonymous sites and explicitly modeling genotype calling errors
  + New methods that don’t use mutational models but only gene flow (therefore only recent divergence) that implement simple diffusion process
    - Modeling branch lengths of the population and species trees as proportional to drift and by treating drift in different branches independently (no gene flow).
* Sampling genealogies
  + Recombination rate low so that recombination is unlikely to have happened since time of divergence then the history can be described by gene trees or genealogy (described by coalescent theory)
  + Integrates over all possible genealogies – advantage of genome wide rather
  + Sample a set of genealogies that are consistent with the data and can be used to obtain posterior probability in a Bayesian approach 🡪 estimate Ne, m, admixture contributions, dates of population declines, etc.
  + Likelihood ratio test can assess the fit of alternative models of divergence BUT this assumes that loci aren’t linked..
* Likelihood-free methods don’t calculate likelihoods and instead use direct simulations under the model of interest (Approximate Bayesian computation – straight forwards and include recombination in models)
  + Still in infancy because of cost of simulating population genomic data and can currently handle hundreds to a few thousand loci
* Product of approximate conditionals (PAC) approximates the likelihood as a sequence f conditional probabilities

**Historical gene flow and LD patterns**

* Movement of genes into a population can create LD in regions experiencing the gene flow 🡪 use **haplotype block lengths** 
  + As migrants enter a population it carries chromosomes that over time are broken into smaller fragments due to recombination
  + Distribution of fragment lengths depend on recombination rate and frequency at which population receives immigrants.
  + Older the migration, shorter the fragments
  + Limitations:
    - Other demographic events can generate blocks (i.e. bottlenecks, selection)
    - Unknown how sensitive these LD statistics are to the events

**Likelihood for models with recombination**

* Difficult due to complex expressions
* Full-likelihood methods that jointly estimate demography and recombination rates developed only for a single population (pairs of loci, as a function of recombination rates, assumes independence of pairs of loci)
  + All are limited to small segments of data (like genealogy sampling) because intermediate recombination rates are difficult to estimate
* Approximations of conditional likelihoods
  + Generate genealogies and ancestral recombination graphs (ARGs)
  + Can be used to compute likelihoods by importance sampling
* Hidden Markov models (HMMs) = Treating recombination as a spatial process across the genome
  + Estimates divergence times and ancestral Ne and changes in population sizes
  + For data comprising a pair or trio of hapoid genomes that are samples from the same or different populations
  + Allows obtaining likelihoods under complex models accounting for recombination and the correlation of genealogies of neighbouring sites
  + Coalescent times are treated as discrete intervals rather than continuous
  + Ancestry of each site modelled by a gene tree that changes at points of recombination as one moves along the genome as a function of recombination rates and some underlying demographic model
* Distribution of block haplotype block lengths as a function of immigration timing and rates
  + Composite likelihood method based on the distribution of immigrant haplotype blocks
  + Can infer changes in migration rates up to 1,000 generations ago
  + Assumes that migrant haplotype blocks can be correctly identified without error

**Conclusions**

* Available tools don’t take full advantage of population genomic data sets
* Need to develop methods to include recombination fully
* AFS and genealogy-sampling approaches assume different SNPs or loci are independently segretating and other methods that take fuller account of recombination are restricted to smaller portions of the genome