A genomicist introduces quantitative genetics

Pradipta Ray,
BIOL 6385 / BMEN 6389,
The University of Texas at Dallas

(some material based on content by PR in Eric Xing’s 10-810 Carnegie Mellon class)
Father of modern genetics
What’s in a name?

Fluid definitions!

Study of whole genome: genomics

ACCTAGAACGGGACCTTTTTTATTA....

Study of single/few alleles (tractable)

Study of evolutionary relations across clades: phylogenetics

Study of evolutionary relations between individuals: population genetics

Clades

Individuals

Study of phylogenetics at whole genome scale: phylogenomics

Study of population genetics at whole genome scale: population genomics
Interconnected disciplines

- **Evolutionary process**: cooking pot
- **Alleles**: ingredients
- **Drift, mutation, selection, recombination, population structure and migration, stochasticity**: recipe
- **Changes in allele frequencies**: outcome of the process!
- **Often**, the goal is to observe the outcome and make evidence-driven guesses about **missing pieces** of the recipe

- **CLASSICAL GENETICS**: genotype – phenotype relationships: phenotype-associated loci, epistasis model, quantitative trait models, **pedigree based inference**
- **POPULATION GENETICS**: evolutionary forces: mutation rates, selectional model, recombination rate, **demography**: migratory model, population size, **population structure**: coalescents
What is a species phylogeny, really?

- It’s a jungle out there!

Pop gen variation should be modelled for a clearer insight into evolutionary dynamics.
Problems with phylogenetic inference

• ML inference: may not be representative

• Multi dimensionality (multiple genomic loci)
  • full joint: may not be tractable
  • full marginal: may not be consistent

\( \text{Allelic variation} \)

\( \text{ML/ Bayesian} \)

\( \text{Mouse} \)
\( \text{Rat} \)
\( \text{Human} \)

\( \text{REMEMBER POSTERIOR DECODING?} \)
Why not use phylo methods for pop genetics?

- Because populations are not well modelled by trees
- Typical genealogical BN for a diploid genome population

- Can we reverse time and use trees?
  - how to tackle situations like siblings?
  - or more complicated: inbreeding?
Why not use phylogenetic methods?

• Maybe we can just model chromosomes ...

• a forest of trees

• But recombination breaks the tree structure
Why not use phylogenetic methods?

• How about just modelling single alleles? (chances are recombination inside allele would be low)
  • even then tree models are not appropriate
    • migration (lateral movement of genes)
    • longer alleles or multi locus models: recombinants cannot be ignored
Why not use phylogenetic methods?

- In pop genetics, trees or genealogies are nuisance variables (marginalized out)
- Phylogenetics:

\[ P(D | \tau, \mu) \]

- Pop genetics:

\[ \sum_{\tau} P(D | \tau, \mu) P(\tau | \mu) \]

\[ = P(D, \mu) \left[ \text{sometimes} \right] \]

\[ \rightarrow \text{TYPICALLY, ALLELE FREQ. EVOLUTION IS} \]

\[ \text{MODELLED AS STOCH PROCESSES EVOLVING OVER} \]

\[ \text{TIME (GENERATIONS)} \]
Genetic variability of population

Due to polymorphisms at various genetic loci:

- SNPs
- Micro/mini satellites
- ALU
- Copy No Variation
- Segmental variation
- Tandem repeats
- Transposons
- Structural variation
- Aneuploidy

Chances of observing de novo variation higher

Chances of observing inherited variation higher

Chances of being neutral inc.

Chances of being deleterious increase
Nomenclature and brief explanations

• “Traits”
• Alleles, identity by state and by descent
• Phenotype, genotype, haplotype
Unit of pop gen evolution: alleles

• Mutually exclusive variants at a genetic locus

• Organism, wrt an allele:
  • hemizygous: only one copy of chromosome
  • homozygous: both copies have same allele
  • heterozygous: copies have different alleles
Haplotype vs genotype

- When we know the allelic composition of multiple alleles in an individual, can we partially reconstruct the chromosomes?

  | haplotype 1 | A | C | A | C | G | C | A |
  | haplotype 2 | + | A | G | G | C | G | T | A |
  | genotype    | AA CG AG CC GG CT AA |

  One individual

- How many possible haplotype pairs based on genotype of set of loci, n of which heterozygous?

Zhou and Wang BMC Bioinformatics 2007 8:484
Pedigree: modelling relations between individuals
IBD and IBS in alleles

• Identity by state (IBS) : if two alleles are identical based on their sequence

• Identity by descent (IBD) : if two alleles are identical based on common ancestor, with no recombination or mutation affecting descent

• IBD implies IBS, but not the other way around
Coalescent

- Shows evolutionary history of a population of alleles

https://frederikleliaert.wordpress.com/
Phasing

• The process of identifying the haplotypes
  • Long read sequencing
  • Observed patterns in short read sequencing
  • Sequencing related individuals

Patterns of genetic inheritance
Phasing
Identify recombination crossover points
Contrast with relatives’ genomes
Epistasis

• How does the genotype affect the phenotype?
• Mendelian genetics: allele → phenotype mapping is relatively simple
• Notion of dominance
  • Really a special case of a G2P function
  • Other variations: incomplete dominance, co-dominance, multi-locus (complex) trait
G2P models : epistasis

- Phenotype corresponding to a particular locus is modified by alleles / genotype at other loci
  - may or may not be additive : epistasis models
  - which is the primary locus ?
G2P models: pure dominance

- Two allele, two phenotype model
- Phenotype of one will mask the phenotype of the other: Mendelian model: dominant & recessive

Mendelian model: parental, first filial (F1) and F2 generation

A → dominant
a → recessive
G2P models: pure dominance

- Before genotyping / haplotyping / genome sequencing
  - allele frequency of pure dominant / recessive bi allelic traits were frequently determined by analyzing the phenotype frequencies
G2P models: co-dominance

- Both alleles equally affect phenotype
Real valued traits

• So far, we have talked about binary or meristic (countable) traits

• How about continuous phenotypes
  • continuous valued phenotype (height, weight, ...)
  • continuous liability (e.g., disease susceptibility)
Dominance models for real valued traits

• Typically, neither co-dominant or pure dominant / recessive, somewhere in the middle
  • Weakly dominant (dominance models): weighted additive effect models for simplicity
Explaining phenotypic variance

\[ \sigma^2_p = \sigma^2_g + \sigma^2_d + \sigma^2_i + \sigma^2_e \]

- \( \sigma^2_g \): Genetic variance
- \( \sigma^2_d \): Variance in dominance
- \( \sigma^2_i \): Variance in epistatic model
- \( \sigma^2_e \): Environmental factors
Predicting the phenotype: QTL

- Complex, (usually) real valued traits:
  - Quantitative trait locus
    - additive model to explain trait based on the loci
- Multiple loci implicated

Wikipedia, from PLoS
Implicating loci: association studies

• How to find the implicated loci in 1st place?
  • find loci which are jointly discriminative wrt phenotype
    (independently discriminative is a sp case)

• Notion is to pick loci that reduce entropy in \( P(\text{phenotype} \mid \text{genotype}) \)

• Discrimination is a necessary condition for implication, is it sufficient?
Forces shaping allele frequency

Variation increasing changes

- Mutation: de novo random change across generation

Variation reducing changes

- Selection: directed change across generation
- Drift: random change across generation not explained by mutation & selection (finite sampling)

Variation increasing or reducing

- Genetic flow / migration: one time change of frequencies and population size
- Population: effect of finite (small/large) population size
Modelling a population

- MULT LOC1
- MULT CHROM
- TIME

- A2G
- SINGLE LOCUS GENOTYPE
- MAPPED ALLELES

- GENOTYPE
- HAPLOTYPEx (in the limit, 2 homology Chrom)

- EVOLUTION

- MULTI LOCS
- ALLELE SETS
Recombination

• When modelling multiple loci
  • recombination increases variation
A few other relevant concepts

• Linkage disequilibrium (LD)
  • How to calculate LD?
  • What causes LD?
  • Why is it useful?

• Recombination rate
  • Estimating it from population genetic data