CAP 5510: Introduction to Bioinformatics
CGS 5166: Bioinformatics Tools

Giri Narasimhan

ECS 254; Phone: x3748

giri@cis.fiu.edu

www.cis.fiu.edu/~giri/teach/BioinfS15.html

Genetics & GWAS

Basic Population Genetics

- Allele: one of two or more forms of DNA sequence of a particular gene
  - The word "allele" is a short form of allelomorph ('other form')
- Diploid: organisms with two sets of chromosomes
  - Homozygous alleles: if both copies of the allele are the same
  - Heterozygous alleles
- Alleles may be
  - Dominant: allele that is more often expressed in heterozygous individuals
  - Recessive
- Genotype: set of alleles in an individual, i.e., genetic composition

Genetic Characters

- Characters can be
  - Mendelian, i.e., single-gene effects, OR
  - Polygenic, i.e., caused by combined effect of multiple genetic factors, OR
  - Environmental
- Characters can be:
  - discrete (e.g., disease) or
  - continuous (e.g., height)
- Gene loci involved in continuous characters are called Quantitative Trait Loci (QTL)

Hardy-Weinberg Principle

- G.H. Hardy & Wilhelm Weinberg (1908)
  - Allele and genotype frequencies in a population remain constant.

		Females	
		A (p)	a (q)
Males	A (p)	AA (p^2)	Aa (pq)
	a (q)	Aa (pq)	aa (q^2)

  - Assumptions:
    - Diploid; sexual reproduction; non-overlapping generations
    - Biallelic loci; Allele frequencies independent of gender
    - Mating is random
    - Population size is infinite
    - Mutations can be ignored
    - Migration is negligible
    - Natural selection does not affect allele in question
    - Equilibrium attained in one generation

Genetic Linkage

- Meiosis: Cell division necessary for sexual reproduction
  - Produces gametes like sperm and egg cells.
- Meiosis: Starts with one diploid cell with 2 copies of each chromosome and produces four haploid cells, each with one copy of each chromosome. Each chromosome is recombined from the 2 copies.
  - At start of meiosis, chromosome pair recombine and exchange sections. Then they separate into two chromosomes.
  - Recombination: alleles on same chromosome may end up in different daughter cells
  - If two alleles are far apart, then there is a higher probability of a cross-over event between them putting them on different chromosomes.
  - Genetically linked traits are caused by alleles sufficiently close to each other. Used to produce genetic maps or linkage maps.

Linkage Disequilibrium (D)

- D = Difference between observed and expected allelic frequencies
- Given 2 bi-allelic loci A and B

AB	x_{11}
Ab	x_{12}
aB	x_{21}
ab	x_{22}

Allele	Frequency
A	$P_1 = x_{11} + x_{12}$
a	$P_2 = x_{21} + x_{22}$
B	$q_1 = x_{11} + x_{21}$
b	$q_2 = x_{12} + x_{22}$

- $D = x_{11} - p_1 q_1$

	A	a	Total
B	$x_{11} = p_1 q_1 + D$	$x_{21} = p_2 q_1 - D$	q_1
b	$x_{12} = p_1 q_2 - D$	$x_{22} = p_2 q_2 + D$	q_2
Total	P_1	P_2	1

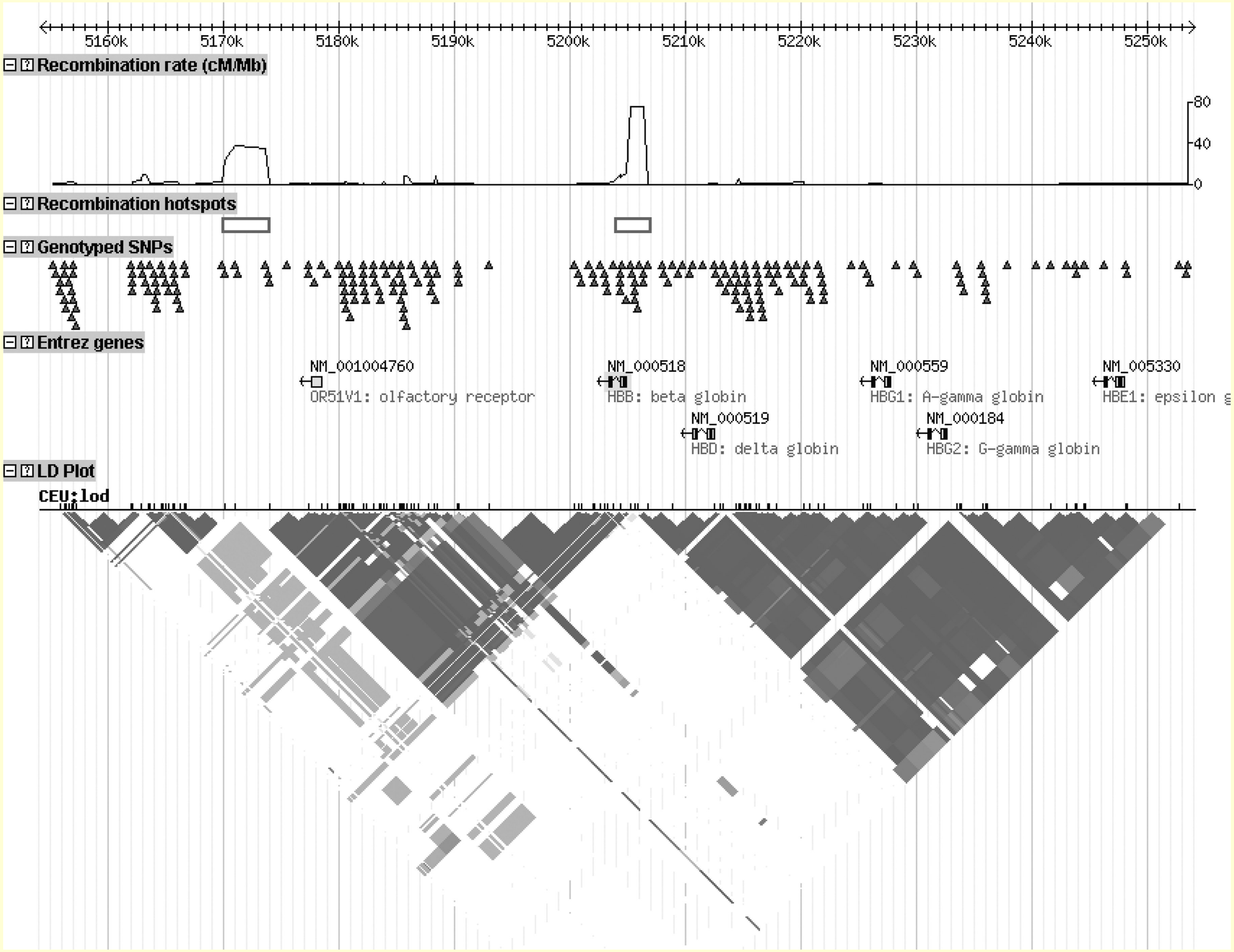
Linkage Disequilibrium

- Linkage (dis)equilibrium: when genotype at loci are (not) independent
- Assumptions of basic population genetics
  - Transmission of alleles (across generations) at two loci are independent
  - Fitness of genotypes at different loci are independent
- Both assumptions are not true in general
- There exists non-random associations of alleles at different loci
- The extent of these associations are measured by Linkage Disequilibrium

SNPs

- SNP: single nucleotide polymorphism
  - Mutations in single nucleotide position
  - Occurred once in human history
  - Passed on through heredity
  - ~10M SNPs in human genome
  - 1 SNP every 300 bp, most with a frequency of 10-50%
- Most variations within a population characterized by SNPs
- Want to correlate SNPs to human disease
- Genotype
  - Gives bases at each SNP for both copies of chromosome, but loses information as to the chromosome on which it appears. NO LABEL!
- Haplotype
  - Gives bases at each SNP for each chromosome. LABELED!

Fig 19.21 from Pevsner



Genotype vs Haplotype

- If the first locus is bi-allelic with two possible alleles (say, A & G)
  - Genotypes: AA, GG, AG
- If a second bi-allelic locus has alleles T & C
  - Genotypes: TT, CC, TC
- Genotypes & Haplotypes for the two loci are:

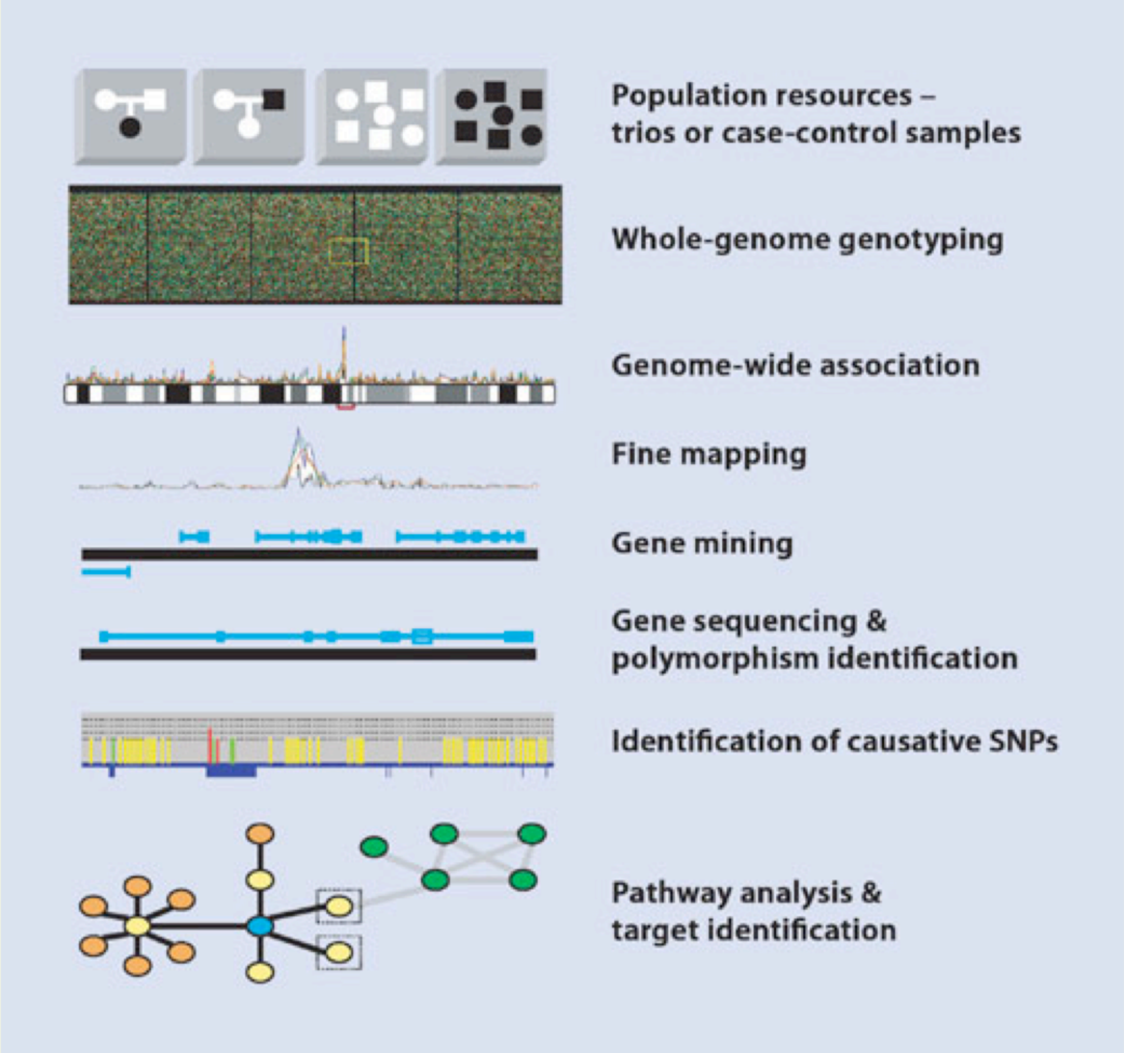
Haplotypes		Second Locus		
		TT	TC	CC
First Locus	AA	AT AT	AT AC	AC AC
	AG	AT GT	AT GC or AC GT	AC GC
	GG	GT GT	GT GC	GC GC

- Interesting problem: Haplotype Phasing
  - Given genotypes, resolve the haplotypes

Genome-wide Association Studies (GWAS)

- To identify patterns of polymorphisms that vary systematically between individuals with different disease states
  - To identify risk-enhancing or risk-decreasing alleles
- Examples of GWAS (900 studies; 3500 associations)
  - Prostate Cancer: Nature Genetics, 1 Apr 2007
  - Type 2 Diabetes: Science Express, 26 Apr 2007
  - Heart Diseases: Science Express, 3 May 2007
  - Breast Cancer, Nature & Nature Genetics, 27 May 2007
  - …
  - See: http://www.genome.gov/Pages/About/OD/ReportsPublications/GWASUpdateSlides-9-19-07.pdf
- Since variation is inherited in blocks / groups, it is enough to study a sample of the population, instead of looking at the whole population.
- GWA databases at NIH: dbGaP, caBIG, and CGEMS

GWAS Process



Population resources – trios or case-control samples

Whole-genome genotyping

Genome-wide association

Fine mapping

Gene mining

Gene sequencing & polymorphism identification

Identification of causative SNPs

Pathway analysis & target identification

Analysis

- Summary statistics for quality control
  - Allele, genotypes frequencies, missing genotype rates, inbreeding stats, non-Mendelian transmission in family data, Sex checks based on X chromosome SNPs
- Population stratification detection
  - Complete linkage hierarchical clustering
  - Multidimensional scaling analysis to visualise substructure
  - Significance test for whether two individuals belong to the same population
- Association Testing:
  - Case vs Control
    - Standard allelic test, Fisher's exact test, Cochran-Armitage trend test, Mantel-Haenszel and Breslow-Day tests for stratified samples, Dominant/recessive and general models, Model comparison tests
  - Family-based associations
  - QTLs
- …

Software

- PLINK: for analysis of genotype, phenotype data
- EIGENSOFT: for population structure analysis
- IMPUTE, SNPTEST, MACH, ProbABEL, BimBam, QUICKTEST

Genetics Software: STRUCTURE

Structure

- Use multi-locus genotype data to investigate population structure
  - Inferring presence of distinct populations
  - Assigning individuals to populations
  - Studying hybrid zones
  - Identifying migrants and admixed invididuals
  - Estimating allele frequencies in populations
- Types of markers
  - Microsatellites, RFLPs, SNPs
- Papers
  - http://pritch.bsd.uchicago.edu/publications/structure.pdf
    - Pritchard, Stephens, and Donnelly, *Genetics* 155:945-959, June 2000
  - http://pritch.bsd.uchicago.edu/publications/FalushEtAl03_Genetics.pdf
    - Falush, Stephens, Pritchard, *Genetics* 164:1567-1587, August 2003

Structure: Methods

- Model-based **clustering** method
- Assumptions
  - K populations (K may be unknown), each characterized by a set of allele frequencies at each locus
  - Within each population, loci are at Hardy-Weinberg equilibrium, and at linkage equilibrium
  - Objective is to assign individuals to populations to achieve the equilibria
  - Markers are not in LD within subpopulations (cannot handle markers extremely close together; weakly linked markers can be handled in Version 2.0)
  - Organisms may be diploid ot non-diploid
- Do not assume a particular mutation process

Data

- For diploid organisms, data for each individual can be
  - Stored in 2 successive rows with each locus in one column
    - George 1 -9 145 66 0 92
    - George 1 -9 -9 64 0 94
  - Or stored in 1 row with each locus in 2 consecutive columns
    - George 1 1 -9 -9 145 -9 66 64 0 0 92 94

Phase/Haplotype Information

- ❑ Phase may be given or unavailable.
- ❑ Two representations:
  - Maternal/paternal contributions are able (MARKOVP SE = 0)
  - Phase info relative to previous allele available (MARK PHASE = 1)

Missing data; e.g., no info on second X chr

From one parent, hence phased

102	156	165	101	143	105	104	101
100	148	163	101	143	-9	-9	-9
0.5	0.5	0.5	0.5	0.5	1.0	1.0	1.0

5 unphased (e.g., autosomal microsatellite) loci and 3 phased (e.g., X chr) loci

Perfectly in phase with previous allele

102	156	165	101	143	105	104	101
100	148	163	101	143	-9	-9	-9
0.5	0.5	0.5	0.5	0.5	0.5	1.0	1.0

Ancestry Models

- No admixture
  - Pure discrete populations
  - Output: Posterior probability that i is from population j
  - Occasionally better than admixture model at detecting subtle structure
- Admixture
  - Individuals with mixed ancestry
  - Output: Posterior mean estimates of fraction that i inherited from pop j
  - Flexible, realistic model and good starting point
  - Difficulty if there are very few representations of the parental populations
- Linkage
  - Generalizes the Admixture model

Ancestry Models (Cont'd)

- Linkage
  - Generalizes the Admixture model
  - Assumes an admixture event t generations in the past, at which time the chromosome inherited distinct chunks from ancestors
  - LD arises because linked alleles are often on the same chunk, and therefore come from ancestral population
  - Sizes of chunks are independent exponential random variables with mean length $1/t$
  - Recombination rate r dictates rate of switching from a chunk to a future chunk
  - MCMC algorithm integrates over the possible chunk sizes and break points
  - Needs location of markers (genetic map)
  - Reports ancestry of each individual
  - Slower computations, but practical for hundreds of loci & individuals

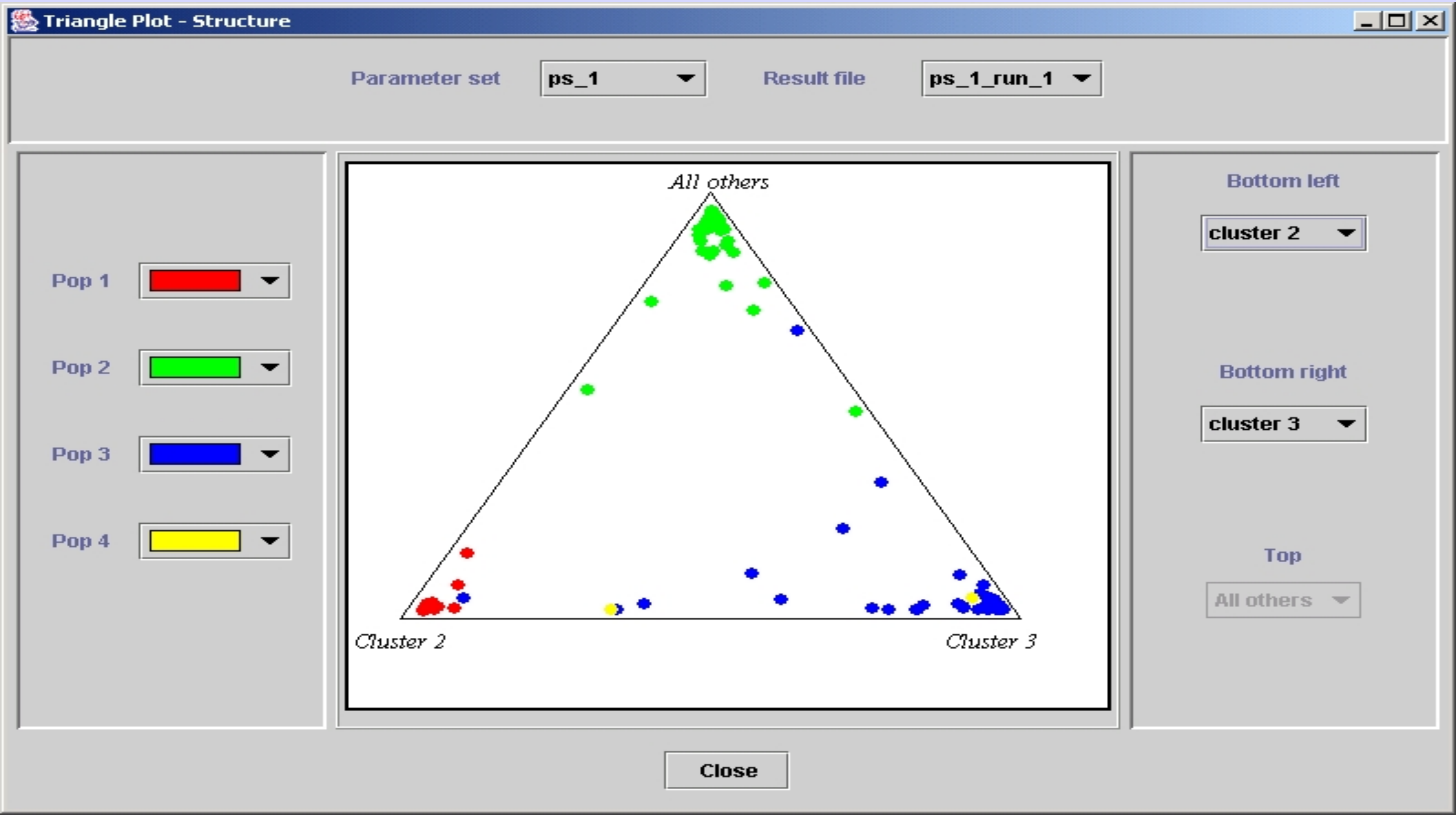
Variants

- Can handle prior info on population
  - Useful to test if an individual is an immigrant to that population or has recent immigrant ancestors
  - Useful to incorporate training data and to classify individuals of unknown origin
  - Parameter called MIGPRIOR to allow for limited misclassification
- Can handle 2 models for allele frequencies
  - Allele frequency in each population are independently drawn from a distribution with parameter λ
  - Can be determined by fixing K = 1, and then estimating λ
  - Allele frequencies are correlated, i.e., different populations may have similar allele frequencies
- K has to be estimated carefully.

Miscellaneous

- ❑ Missing data (as long as it is independent of the allele)
- ❑ Dominant Loci

Results



Triangle Plot - Structure

Parameter set: ps_1

Result file: ps_1_run_1

Pop 1 🟥

Pop 2 🟩

Pop 3 🟦

Pop 4 🟨

Bottom left: cluster 2

Bottom right: cluster 3

Top: All others

Applications

- Diversity and introgression in Scottish wildcats (Beaumont et al., *Mol Ecol*, 10:319-336)
- Study of 20 chicken breeds (Rosenberg et al., *Genetics*, 159:699-713)