CAP 5510: Introduction to Bioinformatics
CGS 5166: Bioinformatics Tools

Giri Narasimhan

ECS 254; Phone: x3748

giri@cis.fiu.edu

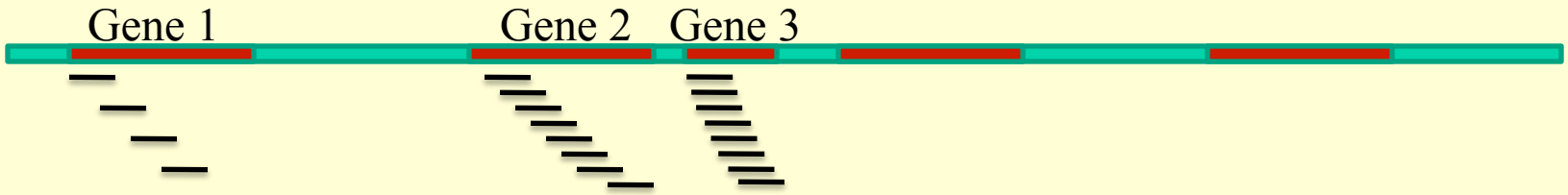
www.cis.fiu.edu/~giri/teach/BioinfS13.html

NGS Applications

Applications of NGS

- RNA-Seq
- ChIP-Seq
- SNP-Seq
- Metagenomics
- Alternative Splicing
- Copy Number Variations (CNV)
- …

RNA-Seq



- Align reads to genes and count
- Assume uniform sampling
  - Count of number of reads mapped per gene is a measure of its expression level
  - Expression of Gene 2 is twice that of Gene 1
  - Expression of Gene 3 is twice that of Gene 2

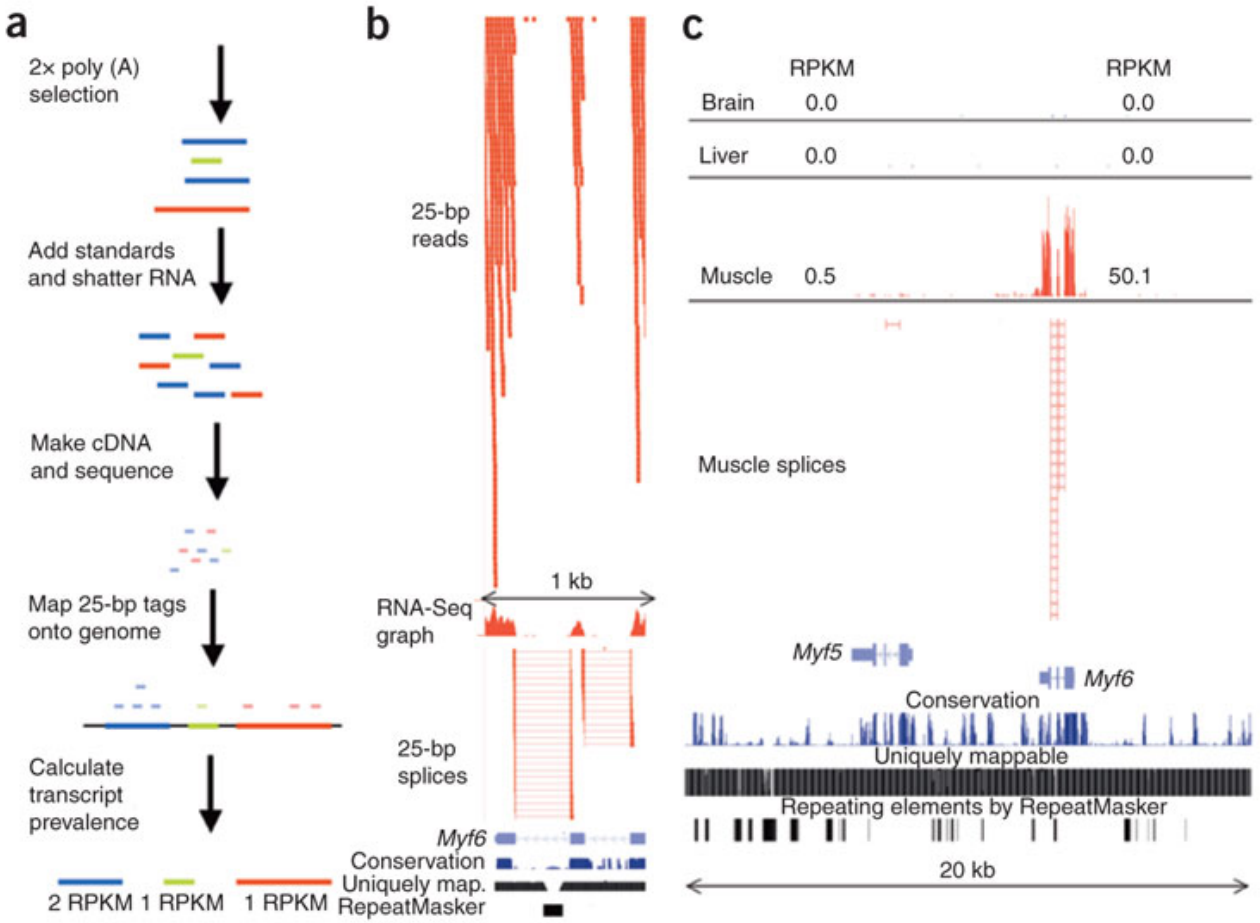
Expression Level of Gene

- RPKM = Ng / (N X L)
  - Ng = Number of reads mapped to gene
  - N = Total number of mapped reads (in millions)
  - L = Length of gene in KB
  - [Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B., Nat Methods. 2008 Jul;5(7):621-8. **Mapping and quantifying mammalian transcriptomes by RNA-Seq.**]

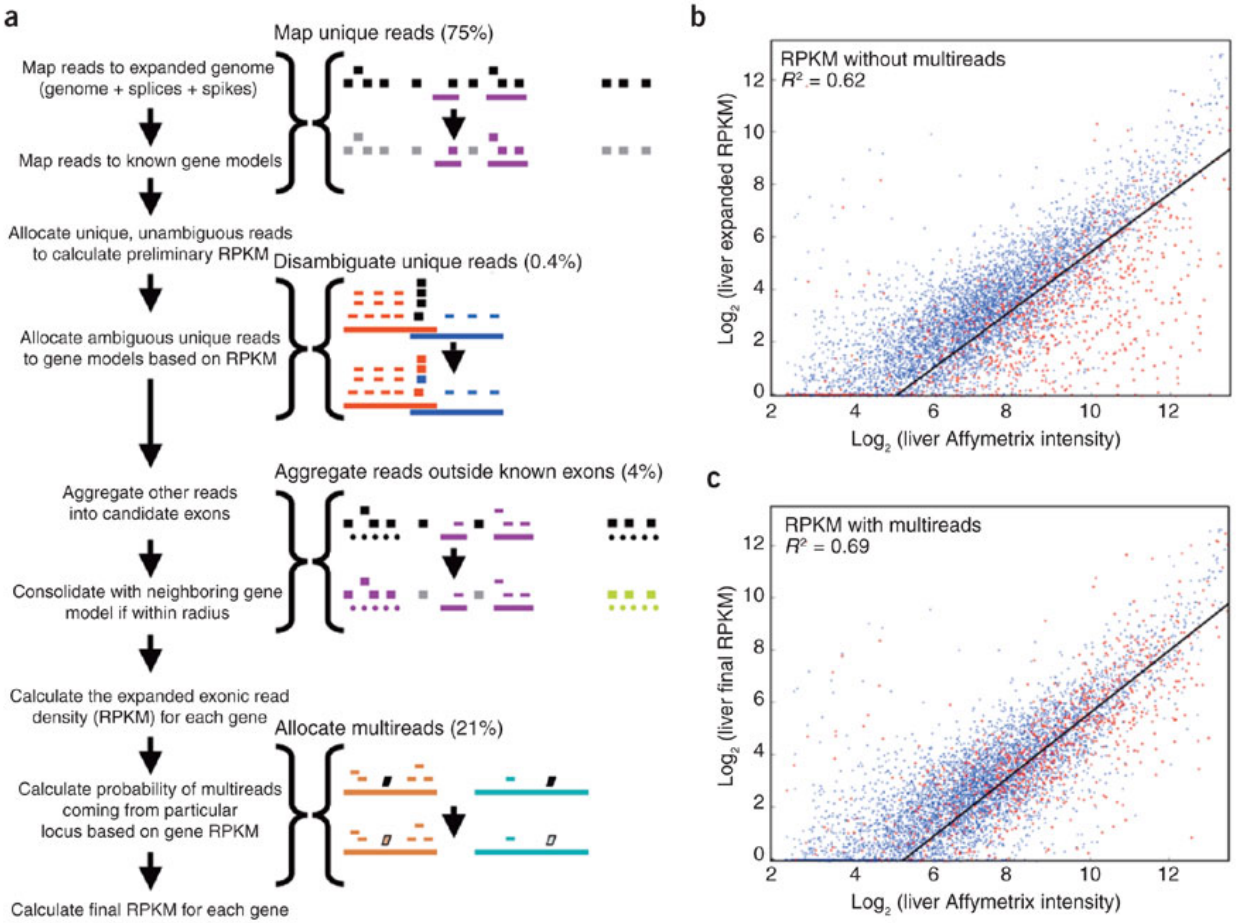
Complications

- Repeat regions
  - Paralogs and other homologous regions in genes
  - Ambiguities in maps
- Introns and Exons
  - Aligning reads to genome is more complex
- Alternative Splicing
- Transcription start site is upstream of ORFs
- Unknown ORFs and Small RNAs
- Other transcripts

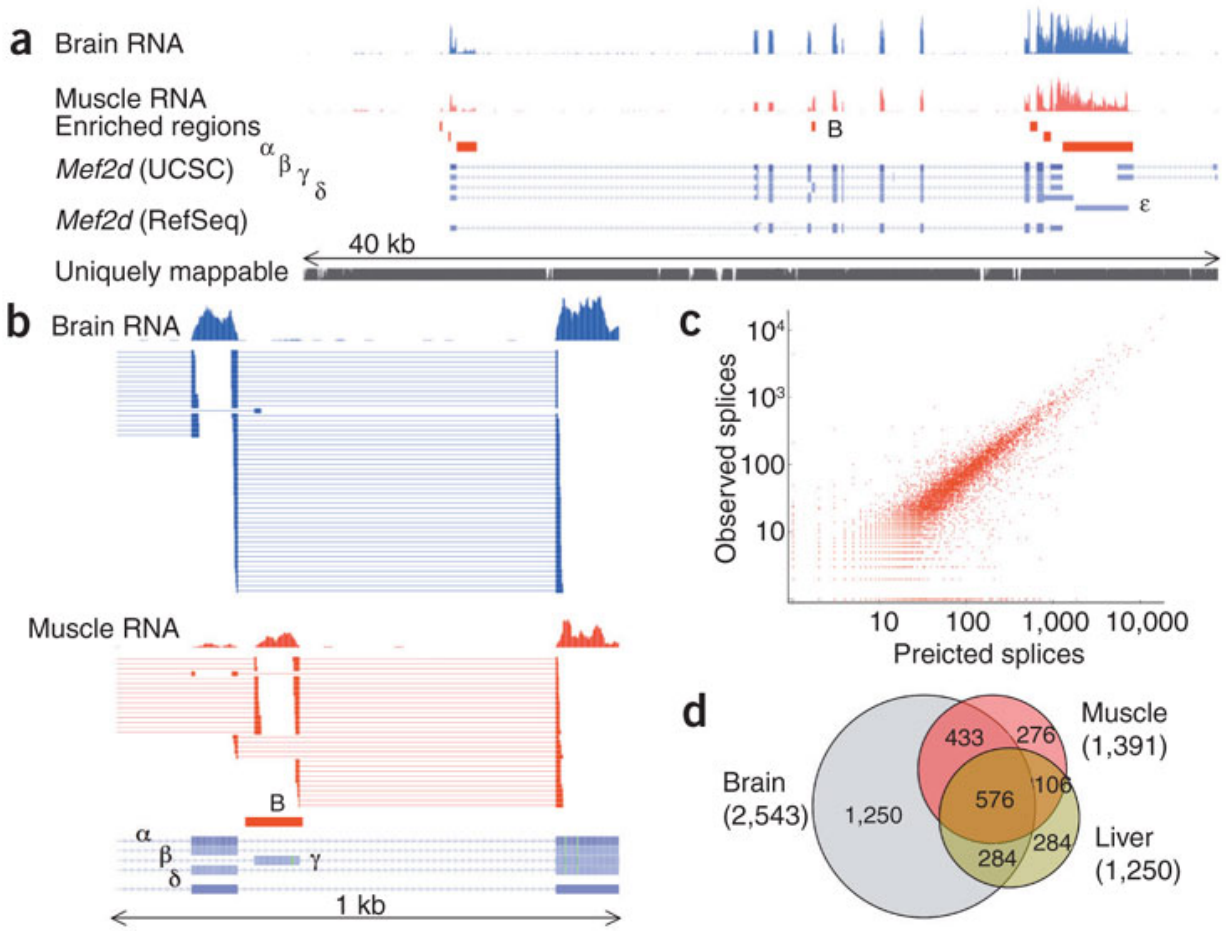
RNA-Seq Procedure



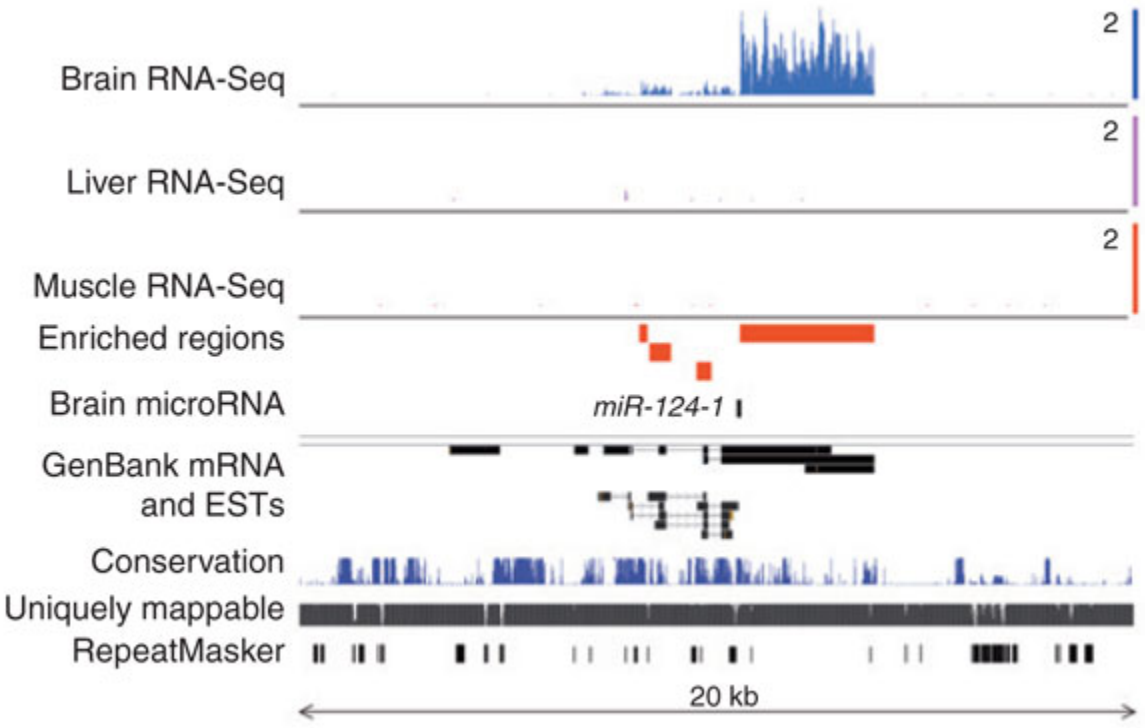
Mapping Reads to Reference



Alternative Splicing



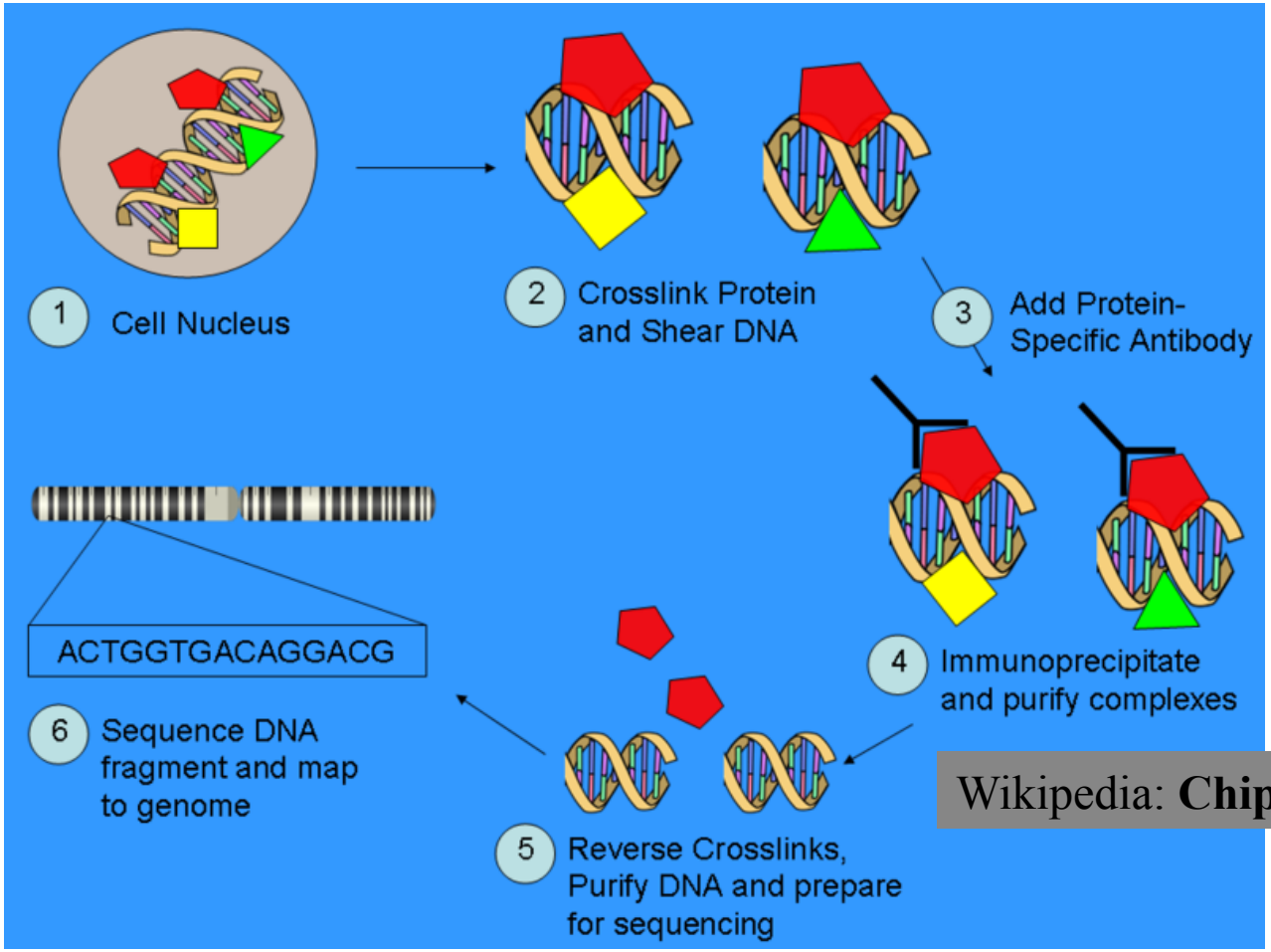
microRNA


Brain RNA-Seq
2
Liver RNA-Seq
2
Muscle RNA-Seq
2
Enriched regions
Brain microRNA
miR-124-1
GenBank mRNA
and ESTs
Conservation
Uniquely mappable
RepeatMasker
20 kb

Chromatin Immunoprecipitation

- Useful for pinpointing location of TFBS for TF
- High-throughput method to find all binding sites for a specific TF under specific conditions
- Identify sites using
  - ChIP-on-chip (Microarray technique)
  - ChIP-Seq (Sequencing technique)
- Problems: TFs bind to specific TFBS only under specific conditions – hard to predict

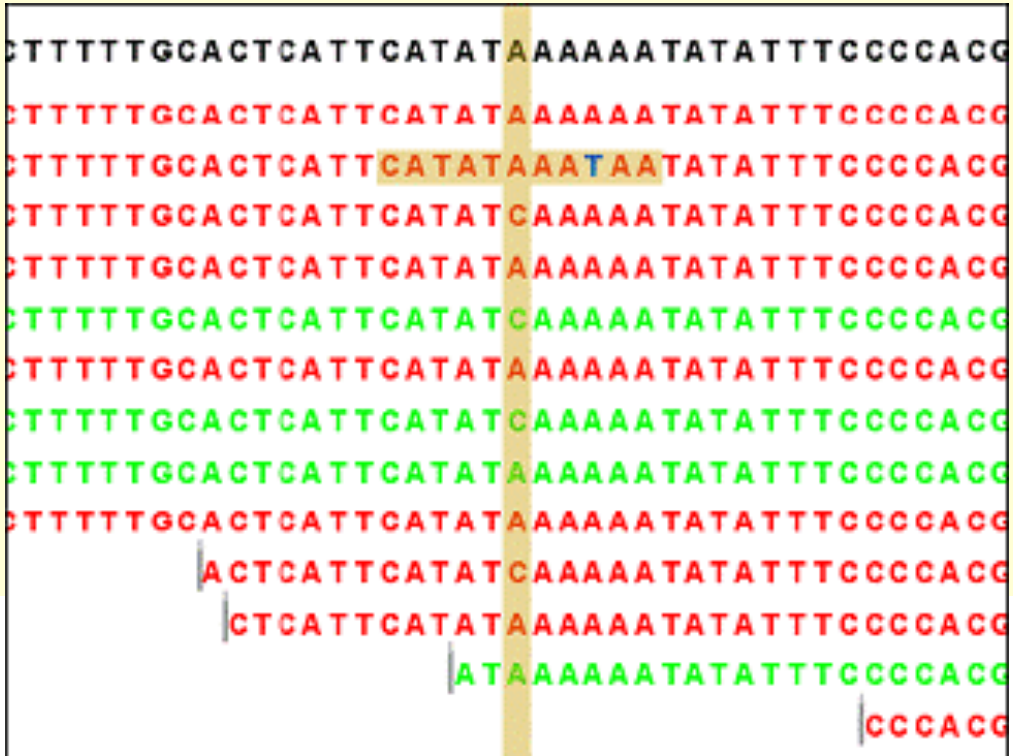
ChIP-Seq



Wikipedia: **Chip-Sequencing**

SNP-Seq

- Align reads and look for differences
  - Differences to reference
    - Align reads to reference sequence first
  - Differences within reads
  - Differences between samples or sets of reads



Environmental Microbiology

- Conventional methods
  - Culture, then identify
    - Slow, expensive, labor intensive, unculturable microbes
  - PCR-based length heterogeneity studies
- Microarray-based methods
  - Unique probes for organisms (e.g., Virochip)
    - Only works for sequenced regions of known organisms
- NGS-based methods

Metagenomics

- Detect known pathogens
- Diversity
  - Identity of individual species not needed
- Functional profile of community

NGS-based method

- Map reads against appropriate database
- Identify closest hits for each read
- Generate contigs
- Generate abundance information
- Clustering of reads can be beneficial to estimate abundance

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 invidividuals
  - 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)

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

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

From one parent, hence phased

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

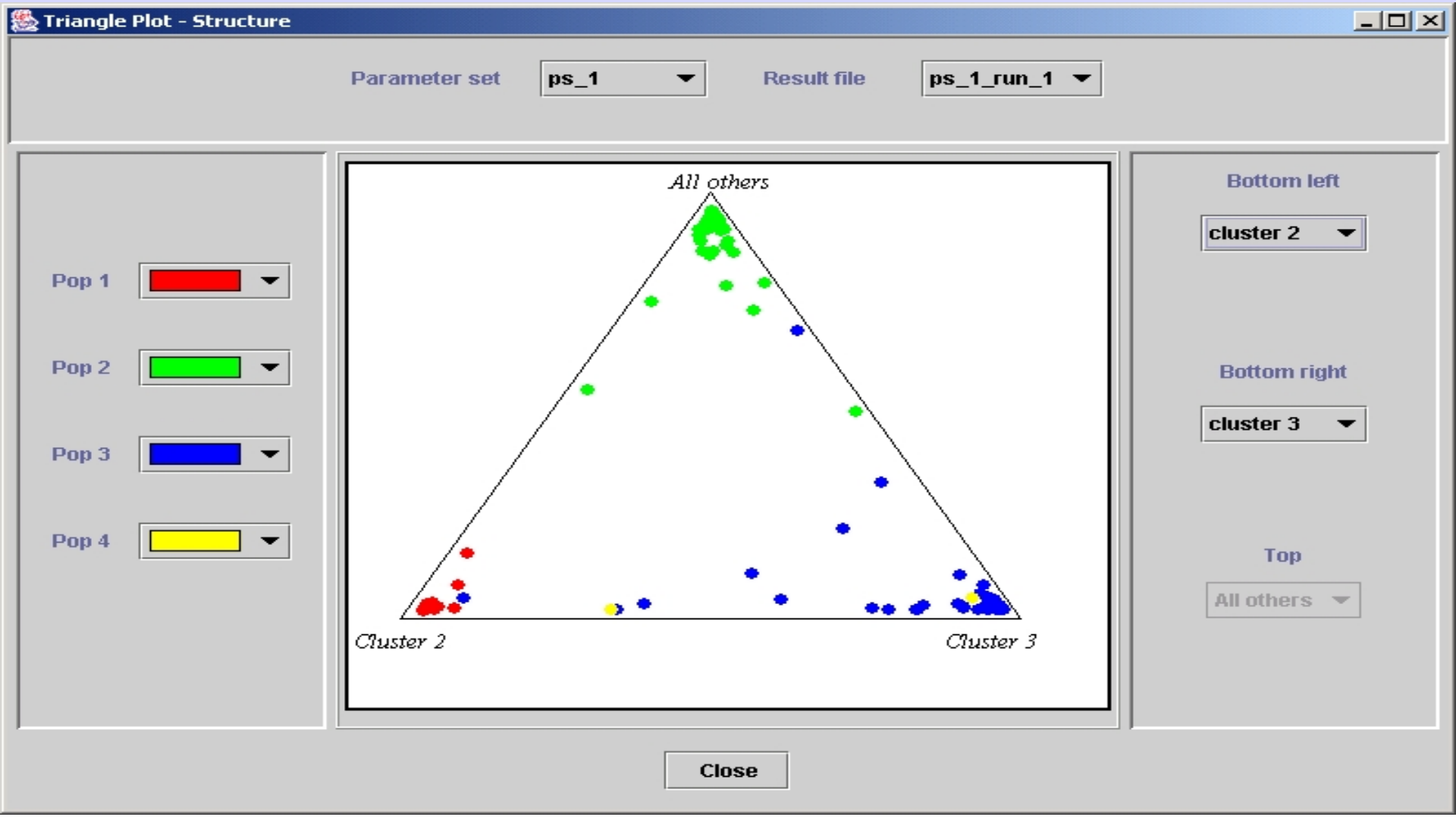
Variants

- 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
All others
Cluster 2
Cluster 3
Bottom left
cluster 2
Bottom right
cluster 3
Top
All others
Close

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)