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

BioPerl

Example 1: Convert SwissProt to fasta format

```
#! /local/bin/perl -w

use strict;
use Bio::SeqIO;
my $in  = Bio::SeqIO->newFh ( -file   => '<seqs.html',
                      -format => 'swiss' );
my $out = Bio::SeqIO->newFh ( -file   => '>seqs.fasta',
                      -format => 'fasta' );

print $out $_ while <$in>;

exit; #bioperl1.pl
```

Example 2 : Load sequence from remote server

```
#!/usr/bin/perl -w
use Bio::DB::SwissProt;

$database = new Bio::DB::SwissProt;

$seq = $database->get_Seq_by_id('MALK_ECOLI');

my $out = Bio::SeqIO->newFh(-fh => STDOUT,
             -format => 'fasta');

print $out $seq;

exit;
```

```
#!/local/bin/perl -w

use Bio::DB::GenBank;

my $gb =
    new Bio::DB::GenBank(
    -retrievaltype=>'tempfile',
    -format=>'Fasta');

my ($seq) = $seq =
    $gb->get_Seq_by_id("5802612");
print $seq->id, "\n";
print $seq->desc(), "Sequence: \n";
print $seq->seq(), "\n";
exit;
```

Sequence Formats in BioPerl

```
#! /local/bin/perl -w
use strict;
use Bio::SeqIO;
my $in  = Bio::SeqIO->new ( -file   => 'seqs.html', -format => 'swiss' );
my $out = Bio::SeqIO->new ( -file   => 'seqs.fas', -format => 'fasta' );

while ($seq = $in->next_seq()) {
    $accNum = $seq->accession_number();
    print “Accession# = $accNum\n”;
    $out->write_seq($seq);
}

exit; #bioperl2.pl
```

BioPerl

```
#!/usr/bin/perl –w
# define a DNA sequence object with given sequence
$seq = Bio::Seq->new('-seq'=>'actgtggcgtcaact',
    '-desc'=>'Sample Bio::Seq object',
    '-display_id' => 'somethingxxx',
    '-accession_number' => 'accnumxxx',
    '-alphabet' => 'dna' );
$gb = new Bio::DB::GenBank();

$seq = $gb->get_Seq_by_id('MUSIGHBA1'); #returns Seq object
$seq = $gb->get_Seq_by_acc('AF303112'); #returns Seq object
# this returns a SeqIO object :
$seqio = $gb->get_Stream_by_batch([ qw(J00522 AF303112)]));
exit; #bioperl3.pl
```

Sequence Manipulations

```
#!/local/bin/perl -w

use Bio::DB::GenBank;

$gb = new Bio::DB::GenBank();

$seq1 = $gb->get_Seq_by_acc('AF303112');
$seq2=$seq1->trunc(1,90);
$seq2 = $seq2->revcom();

print $seq2->seq(), "\n";
$seq3=$seq2->translate;
print $seq3->seq(), "\n";
exit; #bioperl4.pl
```

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	$\mathbf{x_{11}}$
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	$q_2 = x_{12} + x_{22}$

- $D = x_{11} - p_1q_1$

	A	**a**	**Total**
B	$x_{11} = p_1q_1 + D$	$x_{21} = p_2q_1 - D$	q_1
b	$x_{12} = p_1q_2 - D$	$x_{22} = p_2q_2 + D$	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!

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:

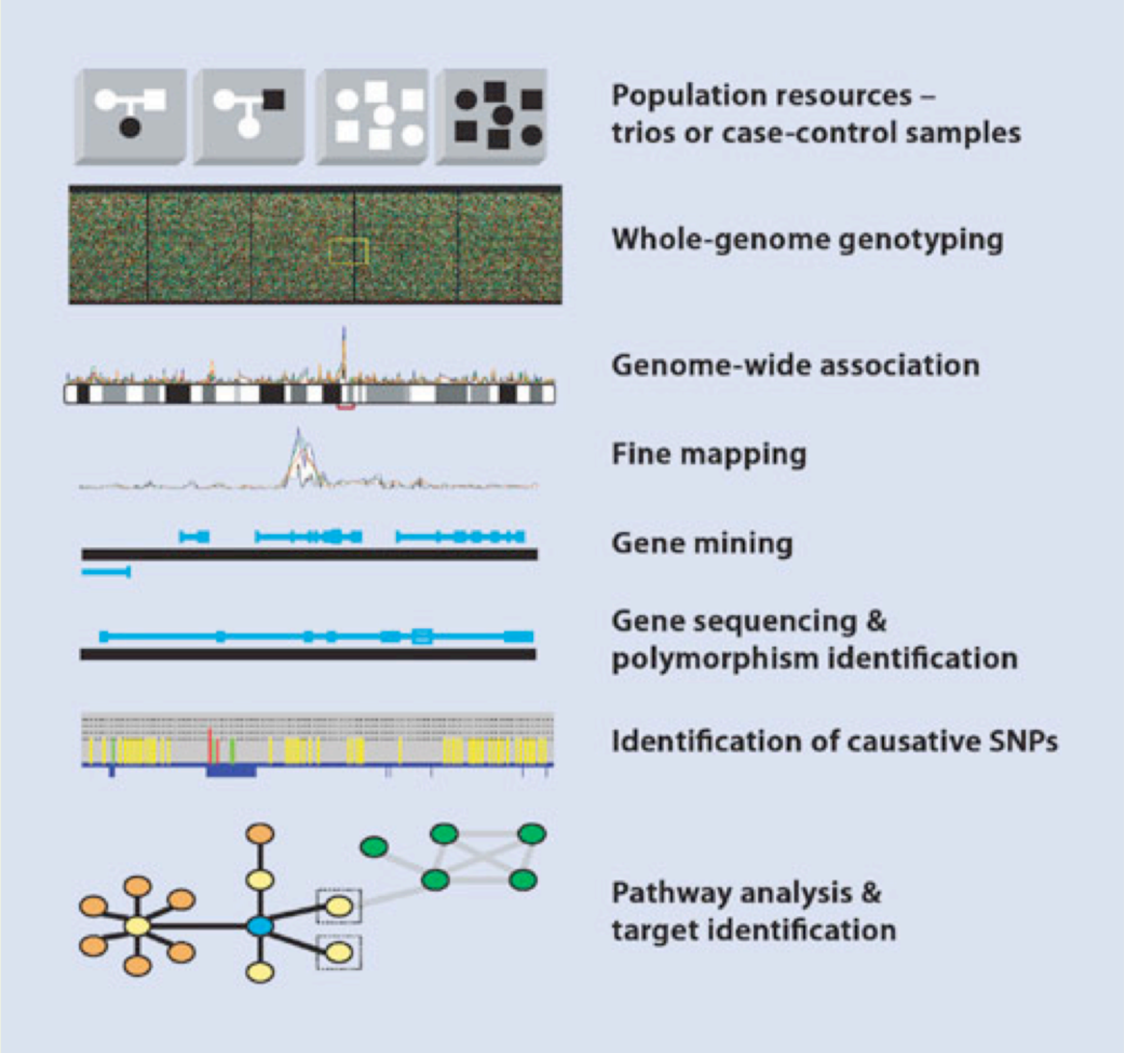
Haplotypes		Second Locus		
		TT	TC	CC
First Locus	AA	AT AT	AT AC	AC AC
	AG	AT GT	AT GC or AC GT	AC GC
	GG	GT GT	GT GC	GC GC

- 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