Picking SNPs
Application to Association Studies

Dana Crawford, PhD

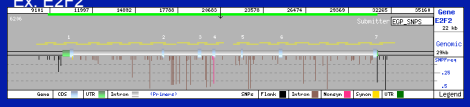
NIEHS
Environmental Genome Project
January 30, 2006

Outline of Tutorial

- Concepts of tagSNPs
- LD and haplotype definitions
- Haplotype blocks and definitions
- Tools to identify tagSNPs

Why Do We Need tagSNPs?

Ex: E2F2



Too Many SNPs to Genotype!

Whole Genome:

- **15,000,000 SNPs**
- **6,000,000 SNPs ≥ 5% MAF**

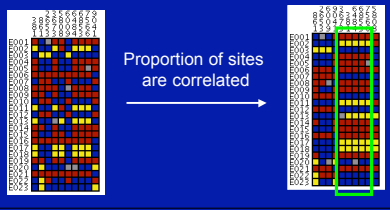
Average Gene:

- **26.5 kb**
- **130 SNPs**
- **44 SNPs ≥5% MAF**

SNPs Are Correlated
(aka linkage disequilibrium)

"the nonindependence of alleles at different sites." Pritchard and Przeworski 2001

Genotype at one site can predict genotype at another site



Measuring Pair-wise SNP Correlations

- SNP correlation described by linkage disequilibrium (LD)
- Pair-wise measures of LD: D´ and r^2

$D = p_{AB} - p_A p_B$; $D' = D/D_{max}$ **Recombination**

$$r^2 = \frac{D^2}{f(A_1)f(A_2)f(B_1)f(B_2)}$$ **Power**

LD Statistics: Practical Uses

- **r^2 is inversely related to power**

$1/r^2$

1,000 cases
1,000 controls } r^2=1.0

1,250 cases
1,250 controls } r^2 = 0.80

- **D´ is related to recombination history**

D´ = 1 no recombination
D´ < 1 historical recombination

Where to Find Population LD Statistics

For your gene or region of interest, search

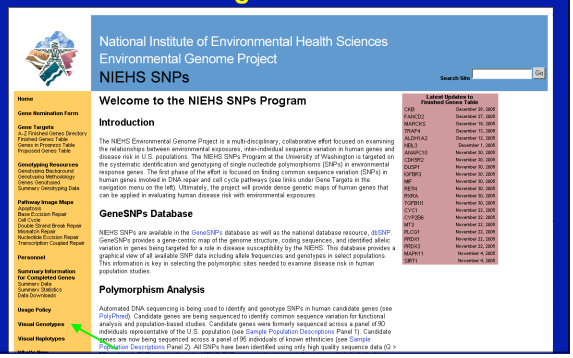
- HapMap — www.hapmap.org
- Perlegen — genome.perlegen.com
- SeattleSNPs PGA — pga.gs.washington.edu
- NIEHS SNPs — egp.gs.washington.edu

Where to Find Population LD Statistics

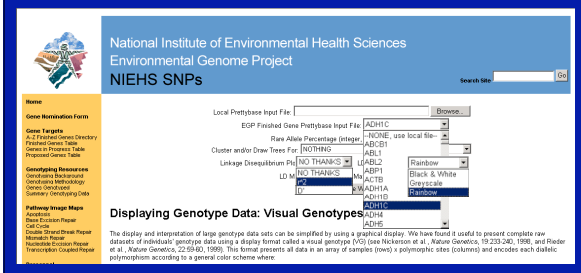
For your gene or region of interest, search

- HapMap — www.hapmap.org
- Perlegen — genome.perlegen.com
- SeattleSNPs PGA — pga.gs.washington.edu
- NIEHS SNPs — egp.gs.washington.edu

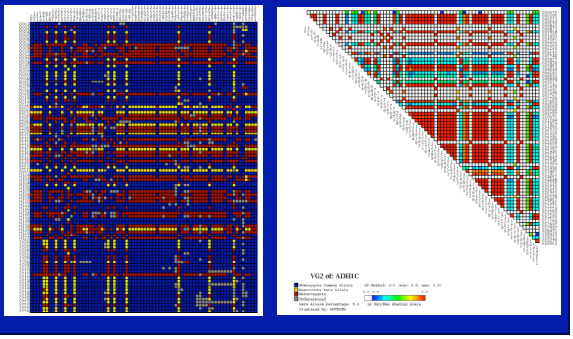
Visualizing Pair-wise LD



Visualizing Pair-wise LD

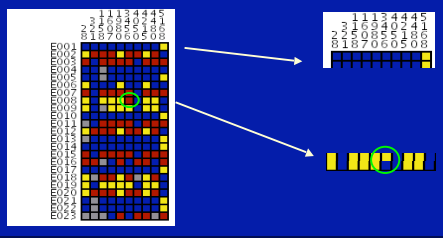


Visualizing Pair-wise LD



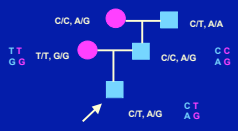
Multi-SNP Correlations (aka Haplotypes)

"…a unique combination of genetic markers present in a chromosome." pg 57 in Hartl & Clark, 1997



Constructing Haplotypes

Collect pedigrees



Somatic cell hybrids

Allele-specific PCR

Constructing Haplotypes

Examples of Haplotype Inference Software:

EM Algorithm
Haploview
http://www.broad.mit.edu/mpg/haploview/
Arlequin
http://lgb.unige.ch/arlequin/

PHASE v2.1
http://www.stat.washington.edu/stephens/software.html

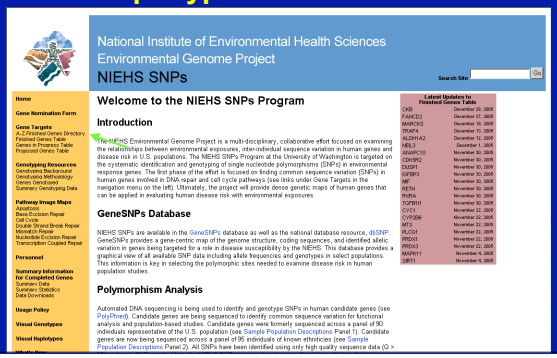
HAPLOTYPER
http://www.people.fas.harvard.edu/~junliu/Haplo/docMain.htm

Haplotypes in NIEHS SNPs

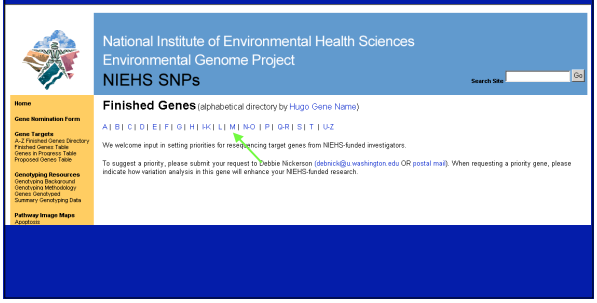
- >550 genes re-sequenced
  Cell cycle, DNA repair/replication, apoptosis
- 2 DNA panels
  1: Polymorphism Discovery Resource (PDR90)
  2: Europeans, Africans, Hispanics, and Asians
- PHASEv2.0 results posted on website
- Interactive tool (VH1) to visualize and sort haplotypes

http://egp.gs.washington.edu

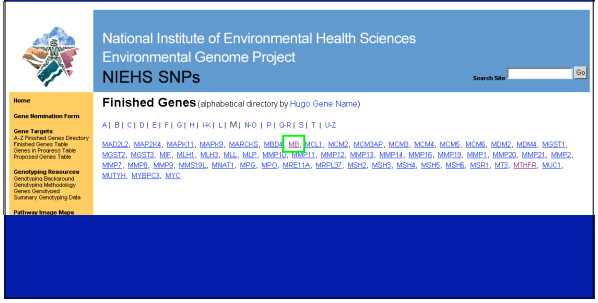
Haplotypes in NIEHS SNPs



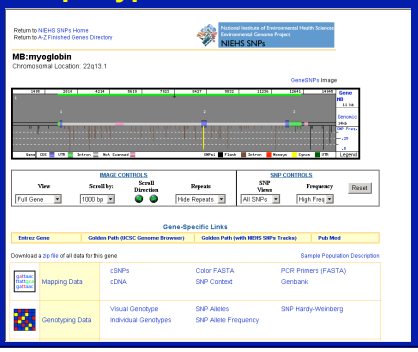
Haplotypes in NIEHS SNPs



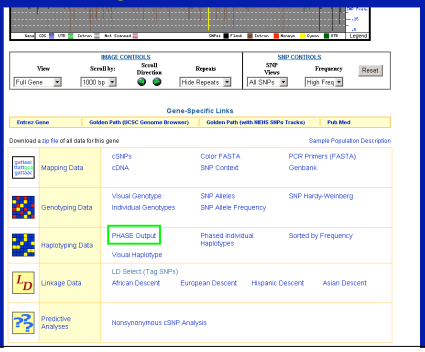
Haplotypes in NIEHS SNPs



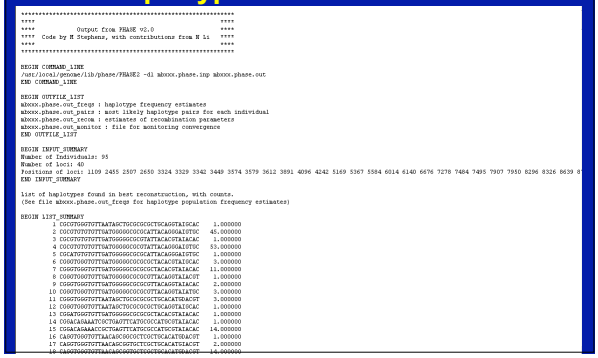
Haplotypes in NIEHS SNPs



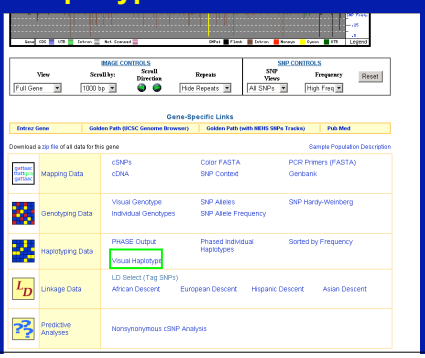
Haplotypes in NIEHS SNPs



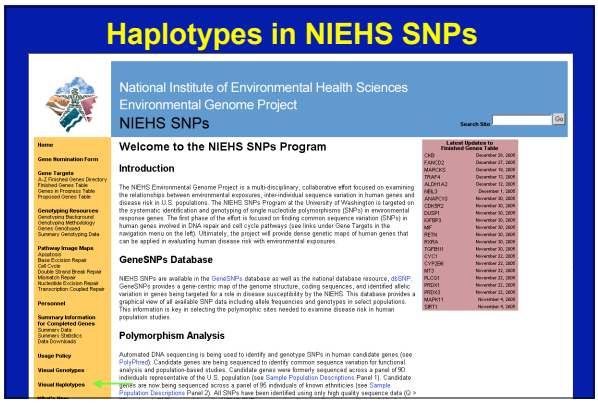
Haplotypes in NIEHS SNPs



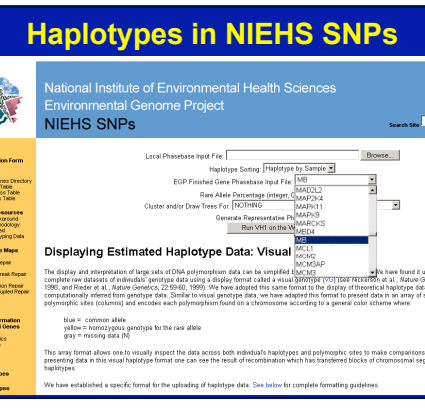
Haplotypes in NIEHS SNPs



Haplotypes in NIEHS SNPs

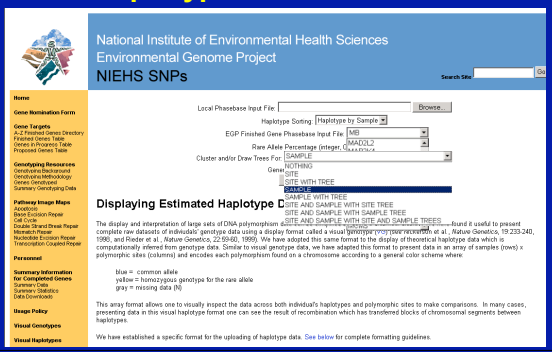


Haplotypes in NIEHS SNPs

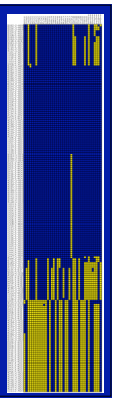


Haplotypes in NIEHS SNPs

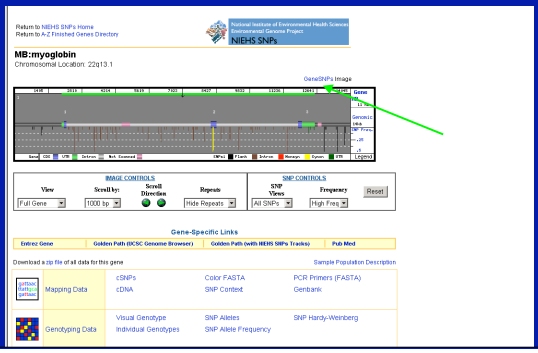
Haplotypes in NIEHS SNPs



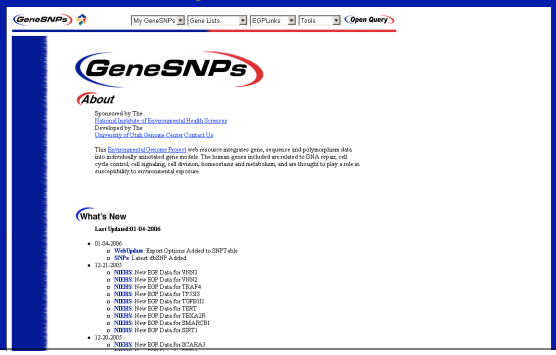
Haplotypes in NIEHS SNPs



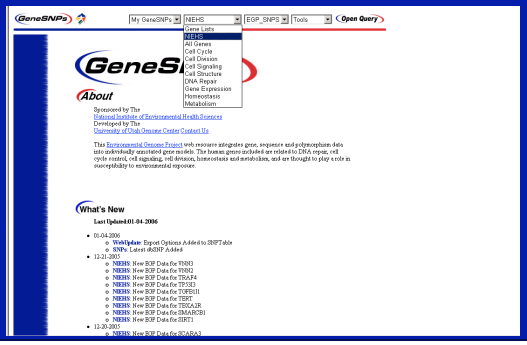
Haplotypes in GeneSNPs



Haplotypes in GeneSNPs

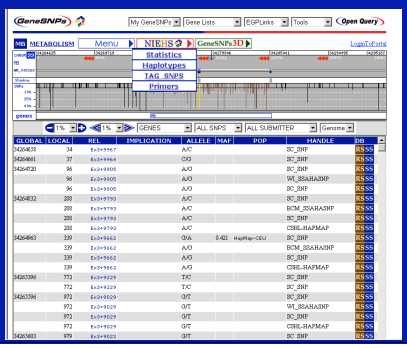


Haplotypes in GeneSNPs

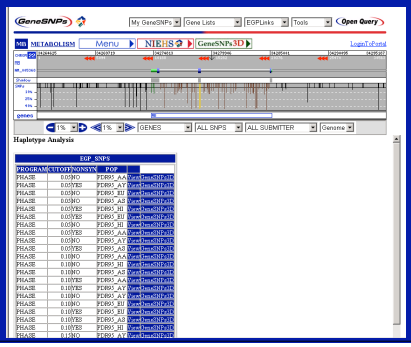


Haplotypes in GeneSNPs

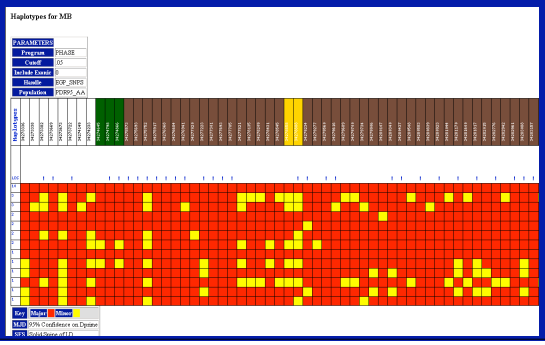
Haplotypes in GeneSNPs



Haplotypes in GeneSNPs



Haplotypes in GeneSNPs



Using LD and Haplotypes to Pick tagSNPs

- **r^2 is inversely related to power**

$1/r^2$

1,000 cases
1,000 controls } r^2=1.0

1,250 cases
1,250 controls } r^2 = 0.80

Example: LDSelect

- **D´ is related to recombination history**

D´ = 1 no recombination
D´ < 1 historical recombination

Example: Haplotype "blocks"

Using LD and Haplotypes to Pick tagSNPs

- **r^2 is inversely related to power**

$1/r^2$

1,000 cases
1,000 controls } r^2=1.0

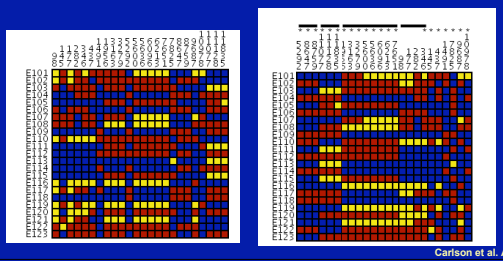
1,250 cases
1,250 controls } r^2 = 0.80

Example: LDSelect

Discovery genotype data → pair-wise LD → pick tagSNPs

LDSelect: Using LD to Pick tagSNPs

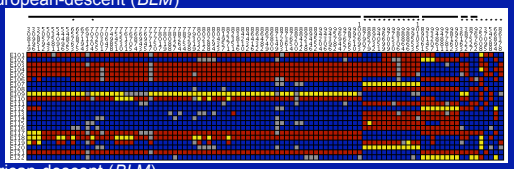
LDSelect
- Uses SNP discovery data (not haplotypes)
- Finds all correlated SNPs to minimize the total number
- Maintains genetic diversity of locus



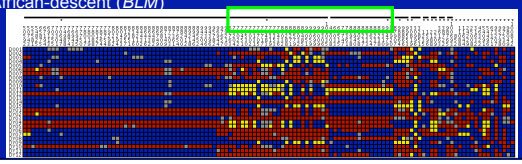
Carlson et al. AJHG (2004)

TagSNPs Are Population Specific

European-descent (*BLM*)

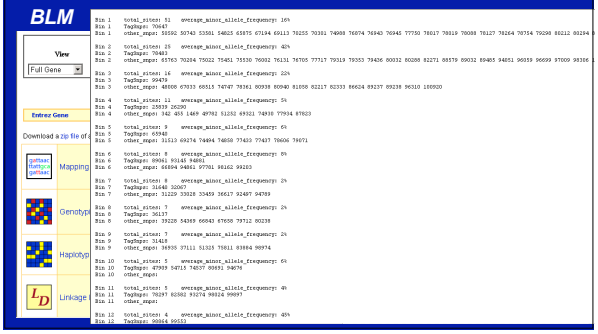


African-descent (*BLM*)



SNP Selection: tagSNP Data

BLM

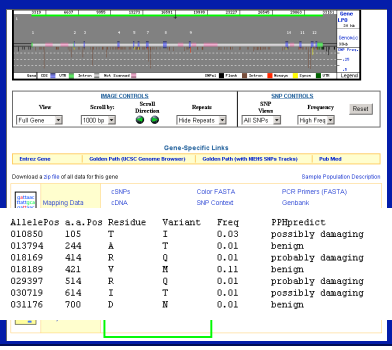


Side Note: Categorizing tagSNPs

- SNP context
  Nonrepetitive > repetitive
- Location of SNP
  Coding > noncoding
- Function
  Nonsynonymous > synonymous

Categorizing tagSNPs

LPO



Hugo	AllelePos	a.a.Pos	Residue	Variant	Freq	PPHpredict	SIFTpredict
lpo	010850	105	T	I	0.03	possibly damaging	TOLERATED
lpo	013794	244	A	T	0.01	benign	TOLERATED
lpo	018169	414	R	Q	0.01	probably damaging	INTOLERANT
lpo	018189	421	V	M	0.11	benign	TOLERATED
lpo	029397	514	R	Q	0.01	probably damaging	INTOLERANT
lpo	030719	614	I	T	0.01	possibly damaging	INTOLERANT
lpo	031176	700	D	N	0.01	benign	INTOLERANT

Haplotypes in Genetic Association Studies

Two main approaches with haplotypes:

Haplotypes → Pick tagSNPs → Genotype samples

Pick tagSNPs → Infer haplotypes → Test for association

Haplotypes in Genetic Association Studies

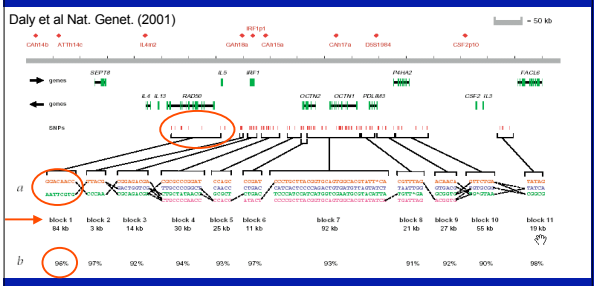
Two main approaches with haplotypes:

Haplotypes → Pick tagSNPs → Genotype samples

Recombination
Natural selection → Haplotype block definition
Population history
Population demography

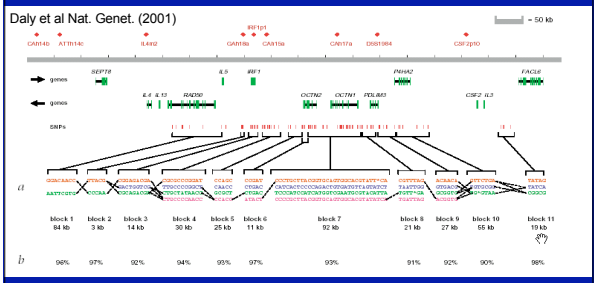
Pick tagSNPs → Infer haplotypes → Test for association

Haplotype "Blocks"



Strong LD **Few Haplotypes** **Represent most chromosomes**

Block Definitions



D´ [Gabriel et al Science (2002)]

Block Definitions

Four-gamete test:

<4 haplotypes, D´=1 → block

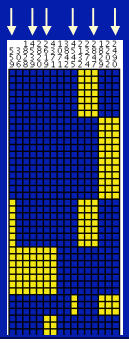
4 haplotypes, D´<1 → boundary

Haplotype Blocks and tagSNPs

Identifying blocks and tagSNPs:

- **Manually**
  Visual haplotype
- **Algorithms**
  HapMap and Haploview

Haplotype Blocks and tagSNPs



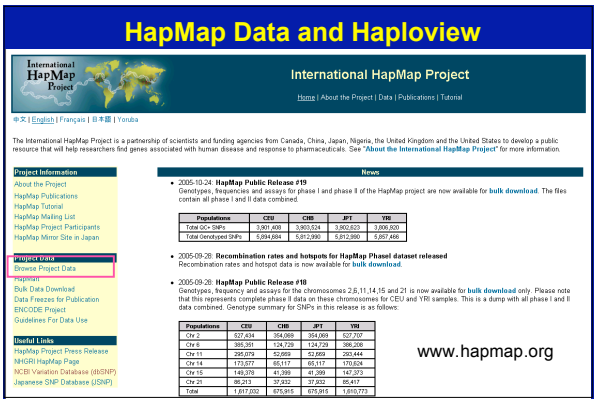
***LTA*:**
16 SNPs (MAF >10%)
6 "common" haplotypes

Haplotype Blocks and tagSNPs

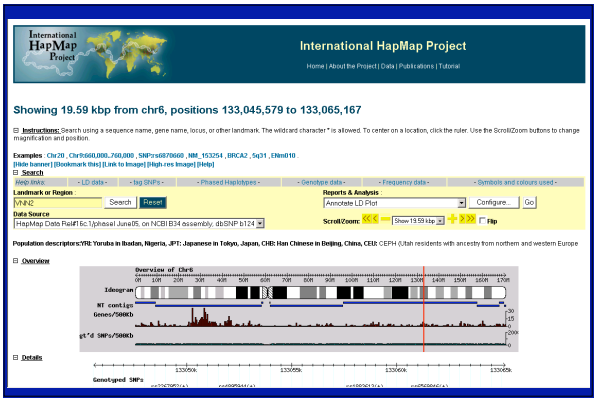
Identifying blocks and tagSNPs:

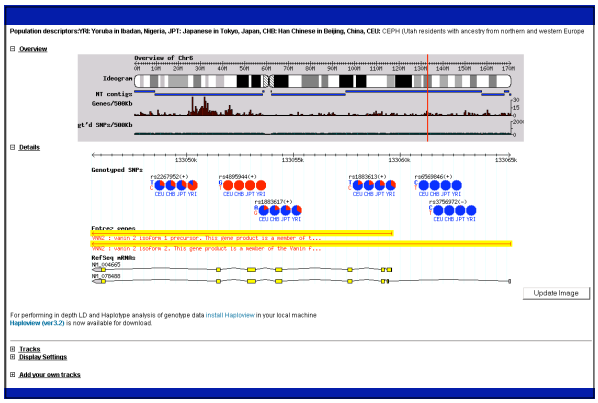
- Manually
  - Visual Haplotype
- Algorithms
  - HapMap and HaploView

HapMap Data and Haploview

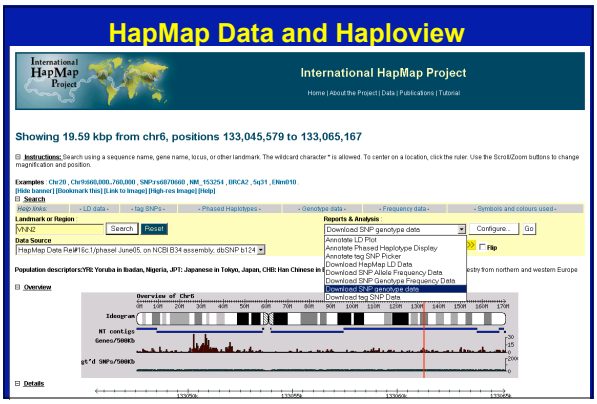


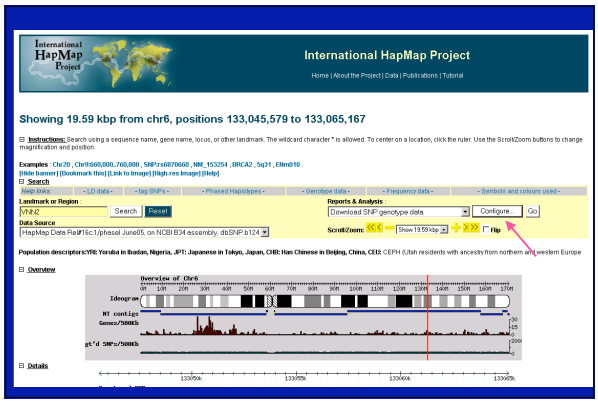
www.hapmap.org



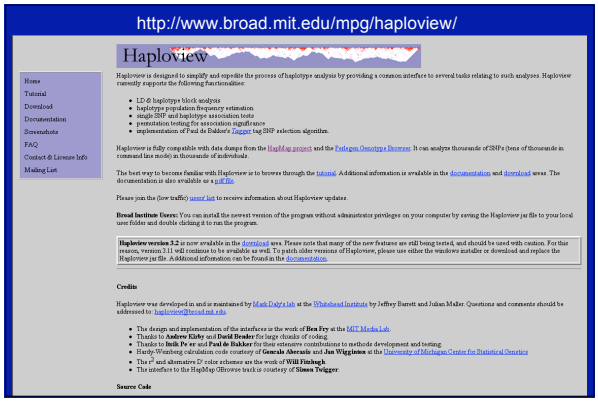


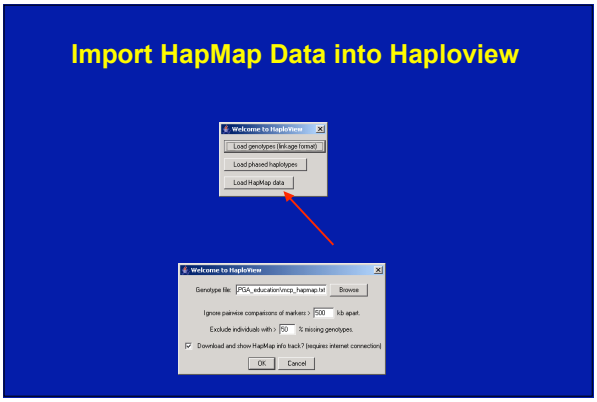
HapMap Data and Haploview

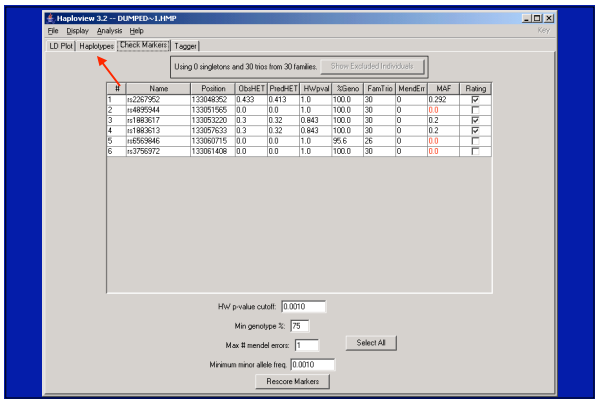


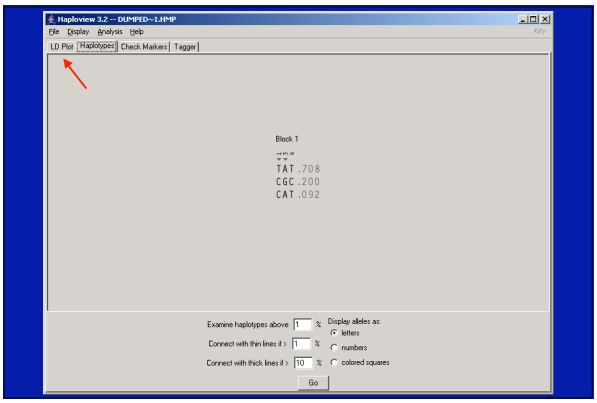

International HapMap Project
Showing 19.59 kbp from chr6, positions 133,045,579 to 133,065,167

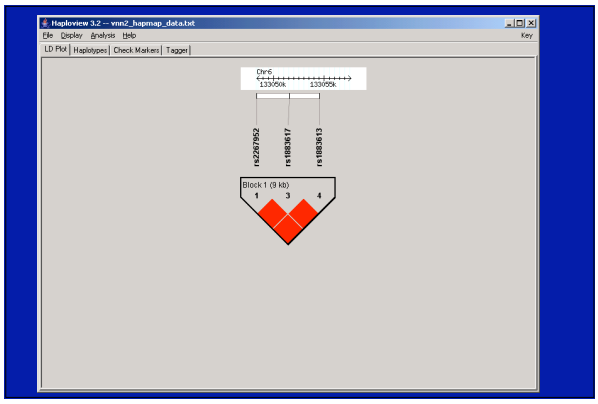

Configure... SNP genotype data

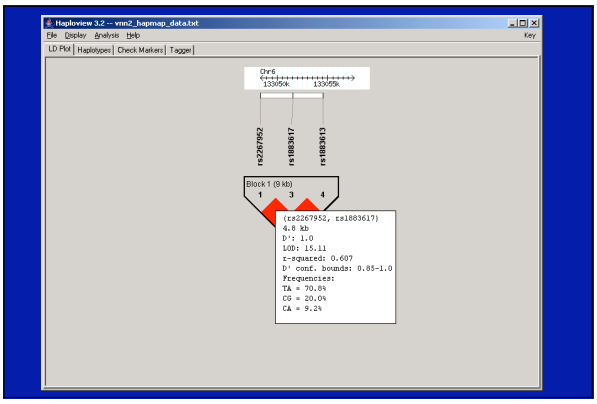

http://www.broad.mit.edu/mpg/haploview/
Haploview

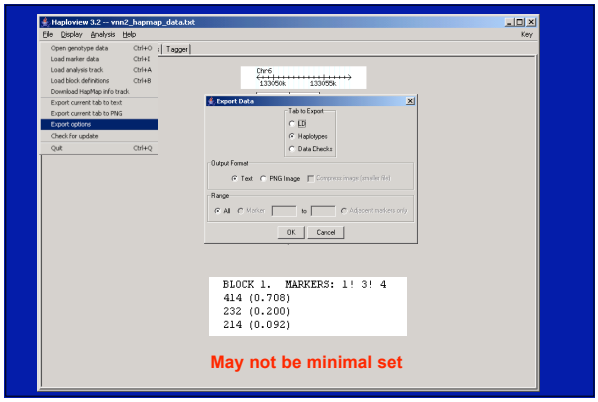

Import HapMap Data into Haploview

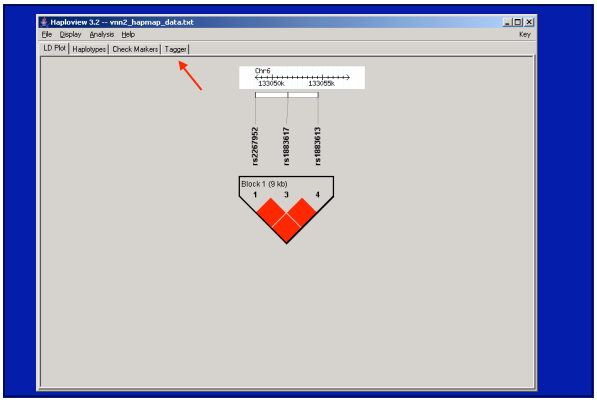

Haploview 3.2 -- DUMPED~1.HMP

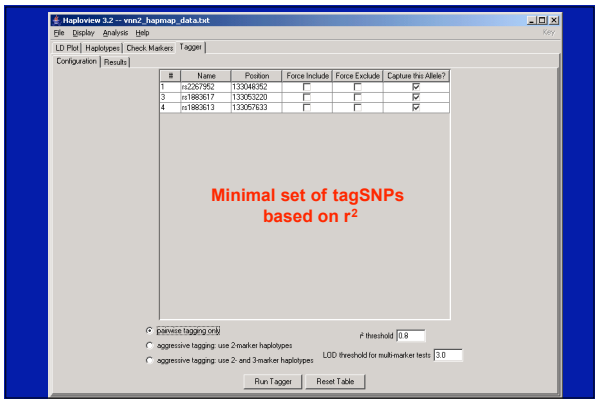

Haploview 3.2 -- DUMPED~1.HMP

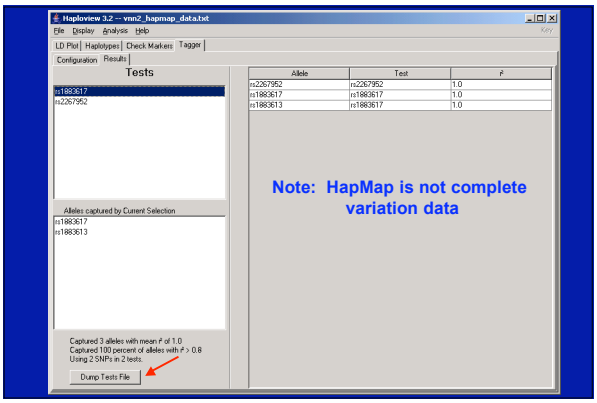

Haploview 3.2 -- vnn2_hapmap_data.txt
LD Plot | Haplotypes | Check Markers | Tagger
Chr6
133050k 133055k
rs2267952
rs1883617
rs1883613
Block 1 (9 kb)
1 3 4


Haploview 3.2 -- vnn2_hapmap_data.txt
LD Plot | Haplotypes | Check Markers | Tagger
Chr6
133050k 133055k
rs2267952
rs1883617
rs1883613
Block 1 (9 kb)
1 3 4
(rs2267952, rs1883617)
4.8 kb
D': 1.0
LOD: 15.11
r-squared: 0.607
D' conf. bounds: 0.85-1.0
Frequencies:
TA = 70.8%
CG = 20.0%
CA = 9.2%

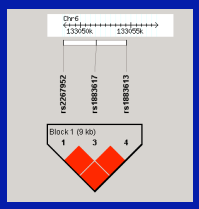

Haploview 3.2 -- vnn2_hapmap_data.txt
Export Data
Tab to Export
LD
Haplotypes
Data Checks
Output Format
Text
PNG Image
Range
All
OK
Cancel
BLOCK 1. MARKERS: 1! 3! 4
414 (0.708)
232 (0.200)
214 (0.092)
May not be minimal set


Haploview 3.2 -- vnn2_hapmap_data.txt
LD Plot | Haplotypes | Check Markers | Tagger
Chr6
133050k 133055k
rs2267952
rs1883617
rs1883613
Block 1 (9 kb)
1 3 4

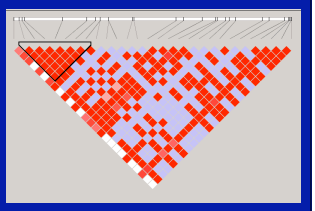

Haploview 3.2 -- vnn2_hapmap_data.txt
LD Plot | Haplotypes | Check Markers | Tagger
Configuration | Results
Name
Position
Force Include
Force Exclude
Capture this Allele?
rs2267952 133048352
rs1883617 133053220
rs1883613 133057633
Minimal set of tagSNPs based on r2
pairwise tagging only
aggressive tagging: use 2-marker haplotypes
aggressive tagging: use 2- and 3-marker haplotypes
r² threshold 0.8
LOD threshold for multi-marker tests 3.0
Run Tagger
Reset Table


Haploview 3.2 -- vnn2_hapmap_data.txt
LD Plot | Haplotypes | Check Markers | Tagger
Configuration | Results
Tests
rs1883617
rs2267952
Alleles captured by Current Selection
rs1883617
rs1883613
Captured 3 alleles with mean r² of 1.0
Captured 100 percent of alleles with r² > 0.8
Using 2 SNPs in 2 tests.
Dump Tests File
Allele
Test
r²
rs2267952 rs2267952 1.0
rs1883617 rs1883617 1.0
rs1883613 rs1883617 1.0
Note: HapMap is not complete variation data

Variation data, LD, and tagSNPs for *VNN2* in European-Americans



HapMap
3 tagSNPs



NIEHS SNPs
9 tagSNPs

Where to Find Tagging Software

HaploBlockFinder	http://cgi.uc.edu/cgi-bin/kzhang/haploBlockFinder.cgi
Haploview	http://www.broad.mit.edu/mpg/haploview/
LDSelect	http://droog.gs.washington.edu/ldSelect.html
SNPtagger	http://www.well.ox.ac.uk/~xiayi/haplotype/index.html
TagIT	http://popgen.biol.ucl.ac.uk/software.html
tagSNPs	http://www-rcf.usc.edu/~stram/tagSNPs.html

Haplotypes, TagSNPs, and Caveats

- Haplotypes are inferred
- Block-like structure assumed for some software
- Different block definitions
- Block boundaries sensitive to marker density
- Genotype savings may not be great (recombination)

Common Errors in Association Studies

Bell and Cardon (2001)

- Small sample size
- Subgroup analysis and multiple testing
- Random error
- Poorly matched control group
- Failure to attempt study replication — e.g., Second case/control study, Gene expression studies
- ✓ Failure to detect LD with adjacent loci
- Over interpreting results and positive publication bias
- Unwarranted 'candidate gene' declaration after identifying association in arbitrary genetic region

Picking SNPs Application to Association Studies Summary

- Resources available for pair-wise LD and haplotypes
- Software for tagSNP selection available
- Be aware the limitations of the approach you choose
- Replication required by several journals