

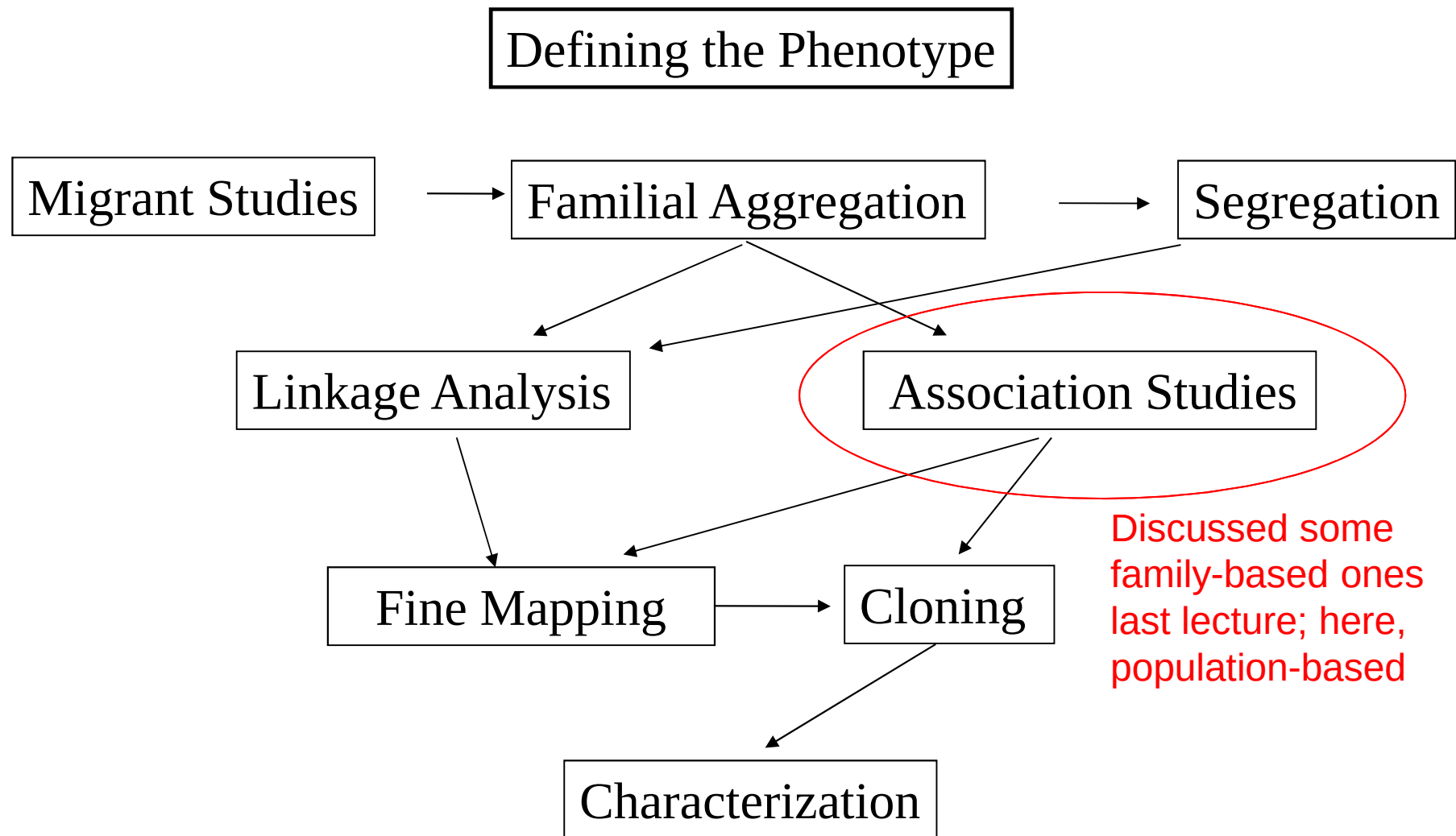
Association Studies



Overview

- Association Studies
 - Design
 - Analysis
- Candidate Gene Studies
 - Resources
 - Selecting 'tag' SNPs
- Pathways

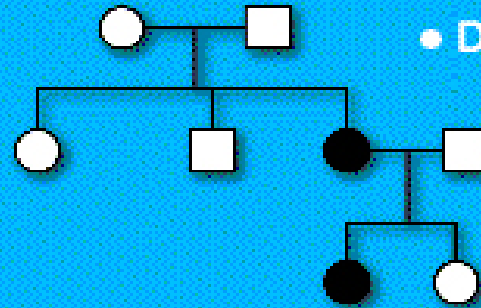
Process of Genetic Epidemiology



Association Studies

Approaches to Identifying Susceptibility Genes for Common Diseases

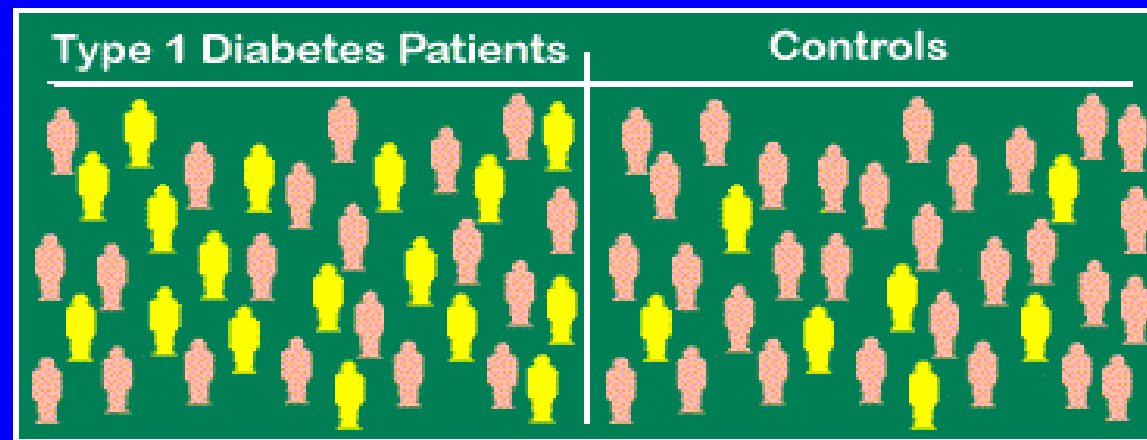
Linkage Studies:



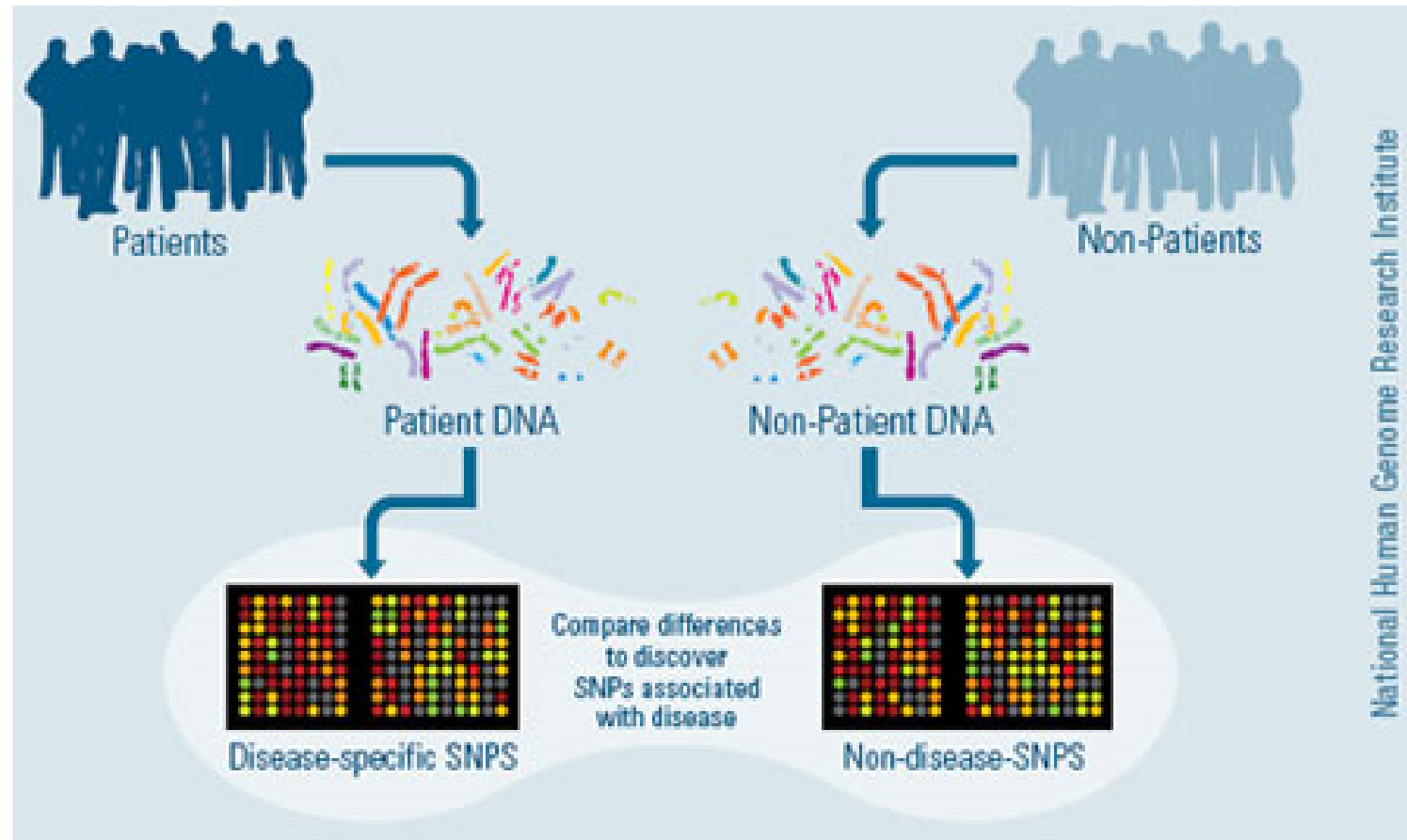
- A relation between loci
- Done within families

Association Studies:

- A relation between alleles
- Done in populations



Association Studies



Association Studies

- Use of (population-based) association studies is rapidly expanding, reflecting a number of good properties, including:
 - Ease, since one need not collect large pedigrees
 - Potential for being more powerful than conventional linkage-based approaches
- Common disease / common variant hypothesis

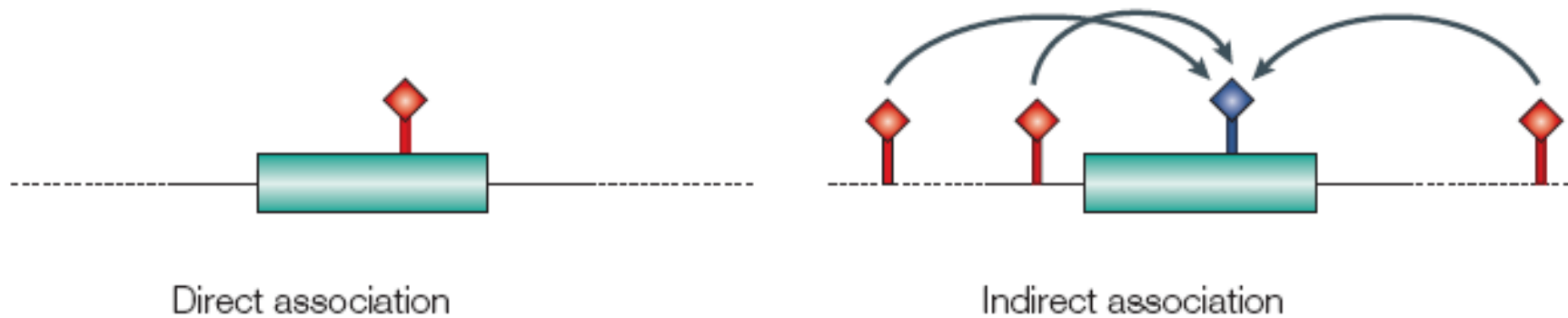
Linkage vs. Association

Linkage					Association			
Genotypic risk ratio (γ)	Frequency of disease allele A (p)	Probability of allele sharing (Y)	No. of families required (N)	Probability of transmitting disease allele A $P(\text{tr-A})$	Singletons		Sib pairs	
					Proportion of heterozygous parents (Het)	(N)	(Het)	(N)
4.0	0.01	0.520	4260	0.800	0.048	1098	0.112	235
	0.10	0.597	185	0.800	0.346	150	0.537	48
	0.50	0.576	297	0.800	0.500	103	0.424	61
	0.80	0.529	2013	0.800	0.235	222	0.163	161
2.0	0.01	0.502	296,710	0.667	0.029	5823	0.043	1970
	0.10	0.518	5382	0.667	0.245	695	0.323	264
	0.50	0.526	2498	0.667	0.500	340	0.474	180
	0.80	0.512	11,917	0.667	0.267	640	0.217	394
1.5	0.01	0.501	4,620,807	0.600	0.025	19,320	0.031	7776
	0.10	0.505	67,816	0.600	0.197	2218	0.253	941
	0.50	0.510	17,997	0.600	0.500	949	0.490	484
	0.80	0.505	67,816	0.600	0.286	1663	0.253	941

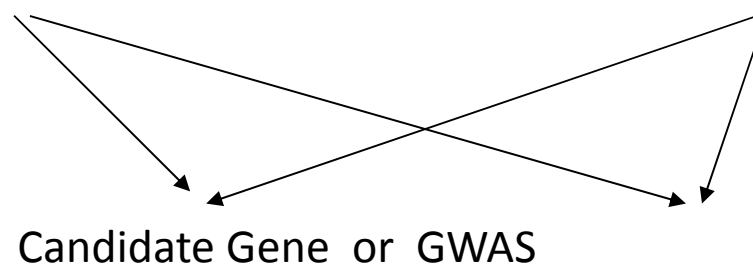
Comparison of linkage and association studies. Number of families needed for identification of a disease gene.

Risch & Merikangas, Science 1996

Association Studies



Hirschhorn & Daly, Nat Rev Genet 2005



Key concept: Linkage Disequilibrium

Implications on design?
Implications on analysis?

Association Study Approaches

- Direct vs Indirect (tagSNPs, later in lecture)
- Candidate genes:
 - Functional
 - All common variants
 - Exome Arrays
- All common variants in genome (GWAS)
- All variants in genes/genome (sequencing)
 - Expensive

Genomics Revolution

Human Genome Project:

13 years, \$3B for 1 sequence

Now: <1 week, ~\$1.5K

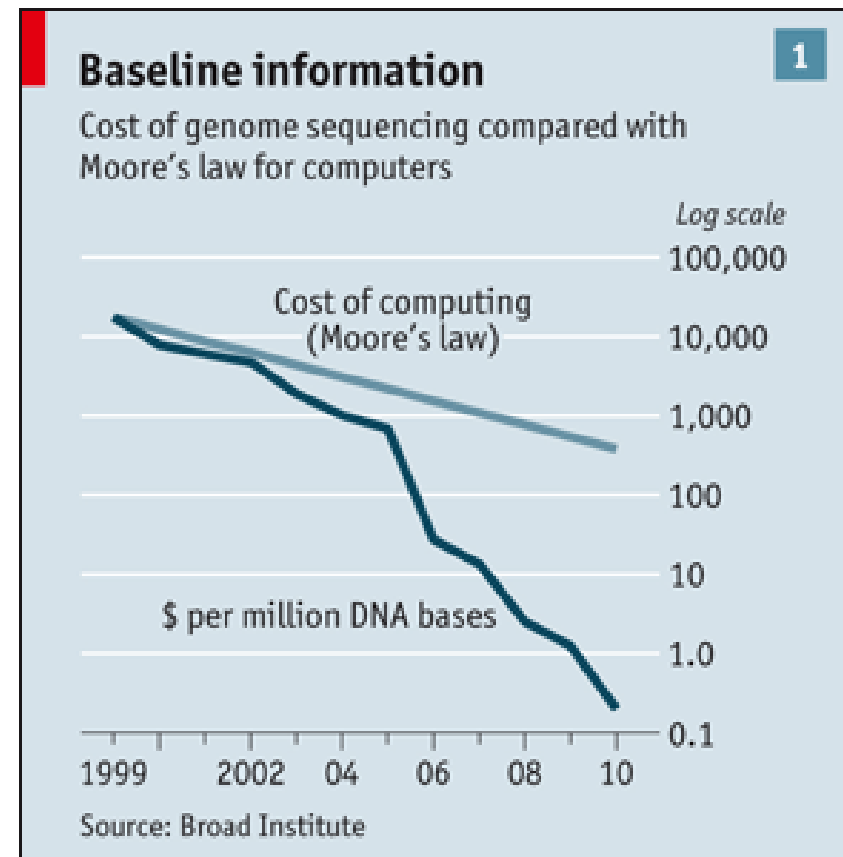
> 500 times faster

< one millionth the cost!

Still not quite there:

1 hour, \$1K

(#1 Innovation, 2010)



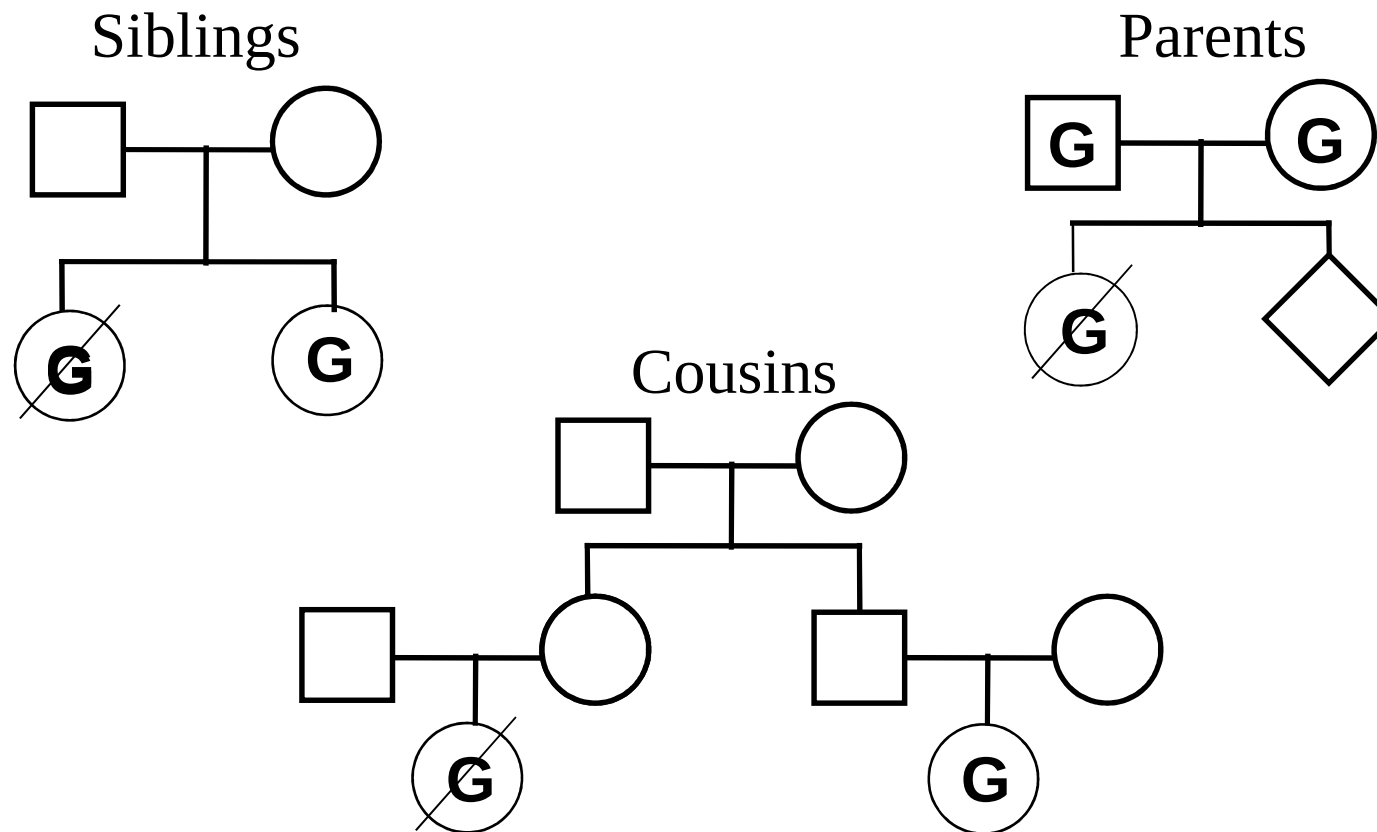
The Economist, 2010

Study Design: Control Selection

- A critical aspect of association studies is that controls should be selected from the cases' source population.
- That is, controls should be those individuals who, if they were diseased, would become cases.

Family-Based Association Studies

(Covered in last lecture)



Termed “**pseudo**”-controls (the hypothetical other mendelian transmissions)

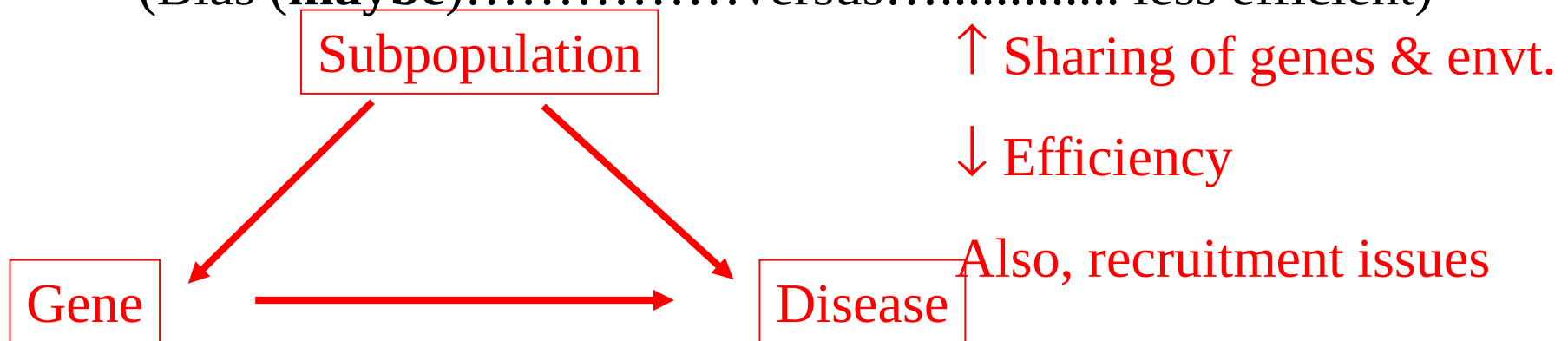
Continuum of Assoc Study Designs



Population Stratification

Overmatching

(Bias (**maybe**).....versus..... less efficient)



Association Analysis

Genotype	Cases	Controls	OR
TT	A	D	1
CT	B	E	BD/AE
CC	C	F	CD/AF

Simple chi-square test comparing genotype frequencies (2 d.f.)
Called a co-dominant analysis

Simple chi-quared test

Observed:

Geno	Case	Control	Total
TT	A	D	A+D=nCC
CT	B	E	B+E=nCT
CC	C	F	C+F=nTT
Total	A+B+C	D+E+F	A+B+C+D+E+F
	=nCase	=nCont	=n

Expected

Case	Control
$n_{TT} * n_{Case} / n$	$n_{TT} * n_{Cont} / n$
$n_{CT} * n_{Case} / n$	$n_{CT} * n_{Cont} / n$
$n_{CC} * n_{Case} / n$	$n_{CC} * n_{Cont} / n$

Sum $(\text{Observed} - \text{Expected})^2 / \text{Expected}$ as before. Chi squared with 2 degrees of freedom.

Hint: Expected cell count = row_total * column_total / total

Simple chi-squared test example

Observed:				Expected		
Geno	Case	Control	Total	Case	Control	OR (to TT genotype)
TT	5	15	20	$20 \times 35 / 65 = 10.8$	$20 \times 30 / 65 = 9.2$	1
CT	10	10	20	$20 \times 35 / 65 = 10.8$	$20 \times 30 / 65 = 9.2$	$(10 \times 15) / (5 \times 10) = 3$
CC	20	5	25	$25 \times 35 / 65 = 13.5$	$25 \times 30 / 65 = 11.5$	$(20 \times 15) / (5 \times 5) = 12$
Total	35	30	65			
	=nCase	=nCont	=n			

Sum $(\text{Observed} - \text{Expected})^2 / \text{Expected}$

$$\begin{aligned}
 &= (5 - 10.8)^2 / 10.8 + (10 - 10.8)^2 / 10.8 + (20 - 13.5)^2 / 13.5 \\
 &\quad + (15 - 9.2)^2 / 9.2 + (10 - 9.2)^2 / 9.2 + (5 - 11.5)^2 / 11.5 \\
 &= 13.7
 \end{aligned}$$

P-value = 0.0011

Genetic Model

Genotype	OR	ORs depend on genetic model
TT	1	$R = r = 1$ not risk allele
CT	r	$R > r = 1$ recessive
CC	R	$R = r > 1$ dominant
		$R = r^2 > 1$ log additive

(Assuming positive association)

Tests of association

If genetic model known:

- Collapse genotypes into 2x2 table, 1 d.f. test
- Trend test for log additive
- Use logistic regression: coding; covariates

- Rarely know genetic model
- Use all three models (dom, rec, log additive)
- **Compare fit with the co-dominant (2d.f.) model (LR test)** / it's two estimated coefficients
- Cannot use LR test to compare models with each other as not nested
- Model with best fit and smallest P is best?

Tests of Association

2 df Genotype			Recessive (G)			Dominant (G)		
Genotype	Case	Control	Case Control			Case Control		
GG	A	D	GG	A	D	GG or GT	A+B	D+E
GT	B	E	GT or TT	B+C	E+F	TT	C	F
TT	C	F						
~chi_sq(2df)			~chi_sq(1df)			~chi_sq(1df)		
Genotype	Case	Control	Case Control			Case Control		
TT	5	15	TT	5	15	TT or TG	15	25
GG	20	5	GG or GT	30	15	GG	20	5
GT	10	10						
P=0.0011			P=0.0045			P=0.0020		

Tests of Association – OR use logistic regression, covariates...

```
. * Load in the dataset, and describe it
```

```
.  
. cd c:\ttemp  
c:\ttemp
```

```
.  
. insheet using geno_ex.csv, clear  
(2 vars, 65 obs)
```

```
.  
. table geno_logadditive case
```

```
-----  
geno_log |      case  
additive |      0      1  
-----+-----  
      0 |     15     5  
      1 |     10    10  
      2 |      5    20  
-----
```

Tests of Association – OR use logistic regression, covariates...

```
. * handle a model fitting both coefficients
. generate geno1 = (geno_logadditive==1)
. generate geno2 = (geno_logadditive==2)
. logit case geno1 geno2
```

```
Iteration 0:  log likelihood = -44.862069
Iteration 1:  log likelihood = -37.630625
Iteration 2:  log likelihood = -37.619709
Iteration 3:  log likelihood = -37.619707
```

Logistic regression	Number of obs	=	65
	LR chi2(2)	=	14.48
	Prob > chi2	=	0.0007
Log likelihood = -37.619707	Pseudo R2	=	0.1614

case	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
geno1	1.098612	.68313	1.61	0.108	-.2402981	2.437522
geno2	2.484906	.7187953	3.46	0.001	1.076094	3.893719
_cons	-1.098612	.5163978	-2.13	0.033	-2.110733	-.0864911

Tests of Association – OR use logistic regression, covariates...

```
. * handle a model fitting both coefficients  
. generate geno1 = (geno_logadditive==1)  
. generate geno2 = (geno_logadditive==2)  
. logit case geno1 geno2
```

```
Iteration 0:  log likelihood = -44.862069  
Iteration 1:  log likelihood = -37.630625  
Iteration 2:  log likelihood = -37.619709  
Iteration 3:  log likelihood = -37.619707
```

```
Logistic regression               Number of obs   =          65  
                                LR chi2(2)         =          14.48  
                                Prob > chi2         =          0.0007  
Log likelihood = -37.619707       Pseudo R2        =          0.1614
```

case	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
-----+-----						
geno1	1.098612	.68313	1.61	0.108	-.2402981	2.437522
geno2	2.484906	.7187953	3.46	0.001	1.076094	3.893719
_cons	-1.098612	.5163978	-2.13	0.033	-2.110733	-.0864911
-----+-----						

**OR=exp(1.0986)
=3**

**OR=exp(2.4849)
=12**

**What model
does this
resemble?**

Tests of Association – OR use logistic regression, covariates...

```
. * recessive test for T / dominant test for G
. generate geno0 = (geno_logadditive==0)
. logit case geno0
```

```
Iteration 0:  log likelihood = -44.862069
Iteration 1:  log likelihood = -39.903582
Iteration 2:  log likelihood = -39.889842
Iteration 3:  log likelihood = -39.88984
```

Logistic regression	Number of obs	=	65
	LR chi2(1)	=	9.94
	Prob > chi2	=	0.0016
Log likelihood = -39.88984	Pseudo R2	=	0.1108

case	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
-----+-----					
geno0	-1.791759	.6055301	-2.96	0.003	-2.978576 - .6049422
_cons	.6931472	.3162278	2.19	0.028	.0733521 1.312942

Tests of Association – OR use logistic regression, covariates...

```
. * dominant test for G / recessive test for T  
. logit case geno2
```

```
Iteration 0:  log likelihood = -44.862069  
Iteration 1:  log likelihood = -39.003467  
Iteration 2:  log likelihood = -38.972594  
Iteration 3:  log likelihood = -38.97259
```

```
Logistic regression              Number of obs   =           65  
                                LR chi2(1)       =           11.78  
                                Prob > chi2       =           0.0006  
Log likelihood = -38.97259       Pseudo R2      =           0.1313
```

```
-----  
      case |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]  
-----+-----  
      geno2 |    1.89712   .5972157     3.18   0.001     .7265984    3.067641  
      _cons |   -.5108255  .3265986    -1.56   0.118    -1.150947    .129296  
-----
```

Tests of Association – OR use logistic regression, covariates...

. * log additive

```
. logit case geno
```

Logistic regression

Number of obs = 65

$$\text{LR } \chi^2(1) = 14.42$$

```
Prob > chi2      = 0.00014
```

case	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
geno	1.2451	.3592	3.47	0.0005	.5360	1.9440
_cons	-1.1577	.4664	-2.48	0.0131	-2.0461	-0.2359

Review: Population stratification

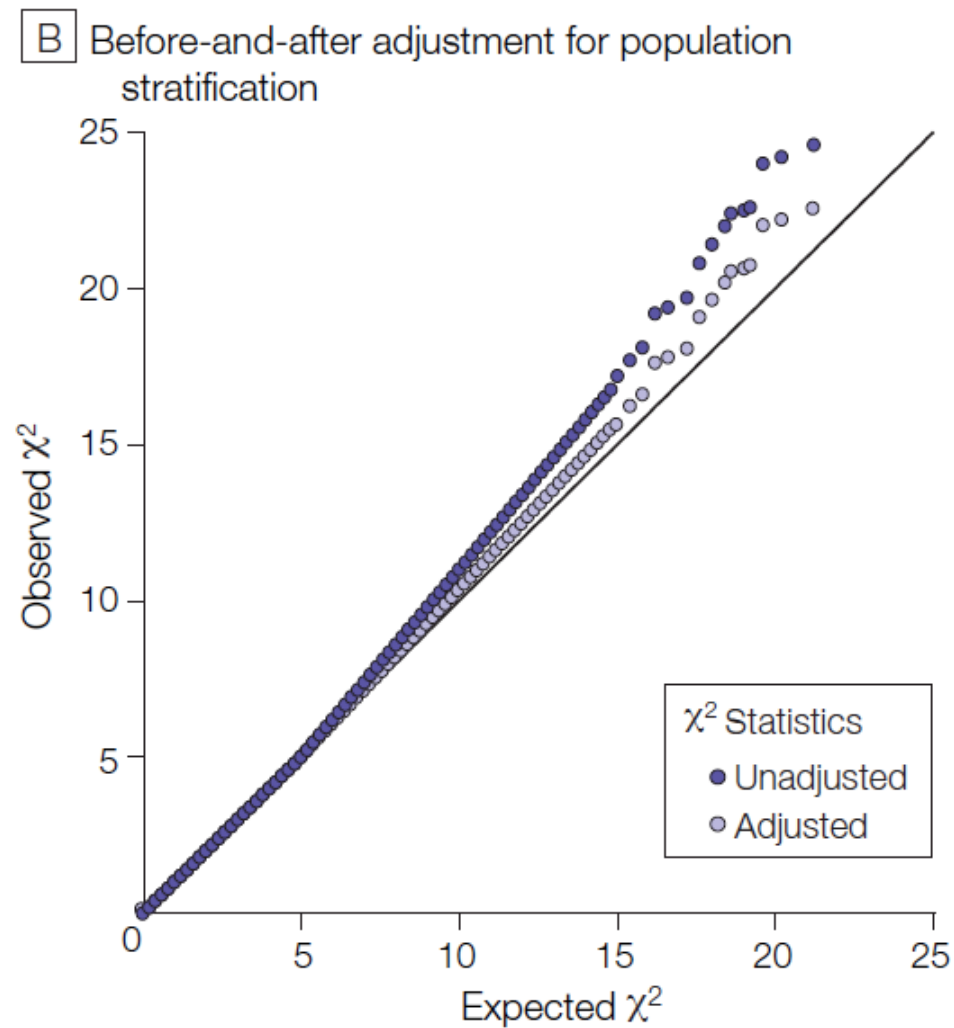
Population stratification is a form of confounding in genetic studies where a gene under study shows marked variation in allele frequency across subgroups of a population and these subgroups differ in their baseline risk of disease

(Pima and Papago example, significance indicates what population from, not association with trait.)

Correcting for population stratification

- “Homogeneous” group (and ignore)
- Genomic control (PMID: 11315092) [see next slide]
 - Non-central chi-square λ = median of all χ^2 tests in the sample
 - Adjust all test statistics for inflation due to empirical chi-square distribution ($\chi^2_{\text{new}} = \chi^2_{\text{old}}/\lambda$)
 - Critiques: An average across the genome, and may **over** or under correct for individual tests
 - For large samples, e.g. 100,000, polygenetic traits, should **NOT** equal 1!
- Ancestry Principal Components (more in GWAS lecture)
 - Eigenstrat (PMID: 16862161)
 - Adjust regression model for principal component values which serve as a proxy for ancestry
- Mixed model – more in GWAS lecture (PMID 20562875) [idea is to condition on the rest of the genome]

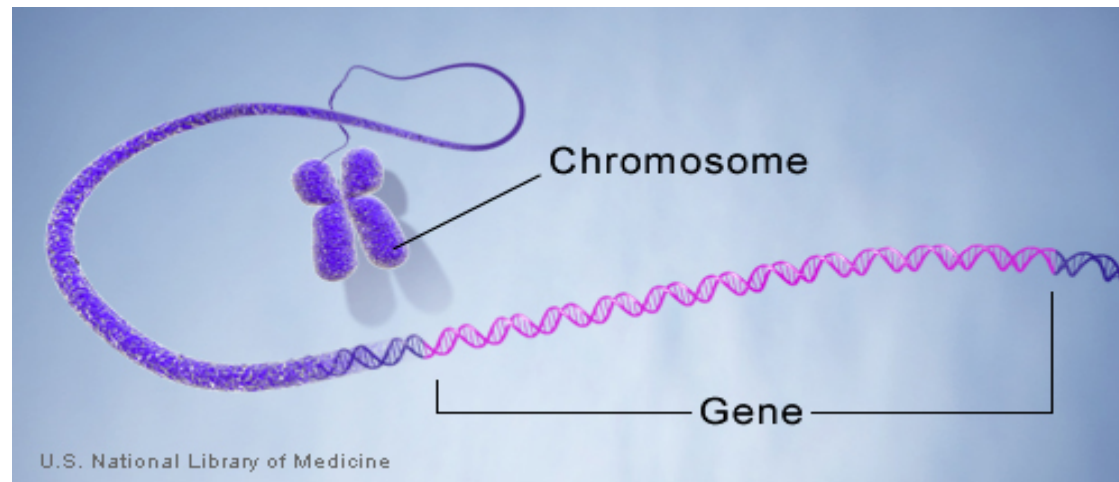
Quantile-Quantile (QQ) Plots



3. Candidate Gene Studies

- Selection of candidates
Linkage regions?
Biological support?

“I am interested in a candidate gene and have samples ready to study. What SNPs do I genotype?”



Candidate Gene: Where do I Start?

- Location:

What chromosome? What position on the chr?

- Exons/UTR:

How many exons? UTR regions?

- Size:

How large is the gene?

Use UCSC genome browser.

SNP Picking: Things to Consider

- Validation: What is the quality of the SNPs?
- Informativity: Are these SNPs informative in my population?
How common are they? Location?
- Potentially Functional: Do these SNPs have a potential biological impact? Missense variants?
- Previously Associated: Have previous studies found SNPs in the candidate gene associated with the outcome?

SNP Picking: Validation

http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=DisplayFiltered&DB=snp

dbSNP Go

Single Nucleotide Polymorphism (SNP)

NCBI

ENTREZ **SNP**
Single Nucleotide Polymorphism

PubMed Nucleotide Protein Genome

NP for Go Clear Save Search

Limits Preview/Index History Clipboard Details

Display Graphic Summary Show 20 Sort by Send to

All: 228 Human: 109 Mouse: 119 NEW: 107 Other Organisms: 0 UPDATE: 7

Items 1 - 20 of 109

☐ 1: [rs28484963](#) [*Homo sapiens*]

ACACTTTGCAATCCTGCCCCGCCCC [G/T] GTGATTTCTCTTTGGGCGGCAGGCA

MapView GeneView SeqView No 3D No OMIM

☐ 2: [rs17421560](#) [*Homo sapiens*]

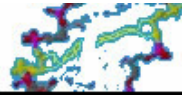
ATTTTCTATCATGAACAGCTGAGG [A/G] CTTTCTTTTTTAAATGTTTTTATT

MapView GeneView SeqView No 3D No OMIM

☐ 3: [rs17421511](#) [*Homo sapiens*]

SNP Picking: Validation

Single Nucleotide Polymorphism



Protein	Genome	Structure	PopSet	Taxonomy	OMIM	Books	SNP
Search for SNP on NCBI Reference Assembly							
for Go							
Reference SNP(refSNP) Cluster Report: rs28484963							
refSNP ID: rs28484963				Allele			
Organism: human (<i>Homo sapiens</i>)				Variation Class: SNP: single nucleotide polymorphism			
Molecule Type: Genomic				Alleles: G/T			
Created/Updated in build: 125/125				Ancestral Allele: T			
Map to Genome Build: 36.1							

SNP Details are organized in the following sections:

Submission	Fasta	Resource	GeneView	Map	Diversity	Validation
----------------------------	-----------------------	--------------------------	--------------------------	---------------------	---------------------------	----------------------------

Submitter records for this RefSNP Cluster

The submission [ss35207915](#) has the longest flanking sequence of all cluster members and w

NCBI Assay ID	Handle Submitter ID	Validation Status	Orientation /Strand	Alleles	5' Near Seq 30 bp
ss35207915	SSAHASNP TA-079.chr1_11782008		fwd/B	G/T	agaaaacactttgcattcctgccc

Validation Summary:

Validation status	Marker displays Mendelian segregation	PCR results confirmed in multiple reactions	Homozygotes detected in individual genotype data
	UNKNOWN	UNKNOWN	UNKNOWN

SNP Picking: Validation

 for	Go
Reference SNP(refSNP) Cluster Report: rs17421560	
refSNP ID: rs17421560	Allele
Organism: human (<i>Homo sapiens</i>)	<u>Variation Class:</u> SNP: single nucleotide polymorphism
Molecule Type: Genomic	Alleles: A/G
Created/Updated in build: 123/123	Ancestral Allele: G
Map to Genome Build: 36.1	

SNP Details are organized in the following sections:

Submission	Fasta	Resource	GeneView	Map	Diversity	Validation
----------------------------	-----------------------	--------------------------	--------------------------	---------------------	---------------------------	----------------------------

Submitter records for this RefSNP Cluster

The submission [ss24247001](#) has the longest flanking sequence of all cluster members and was used to instantiate sequence for rs.

NCBI Assay ID	Handle Submitter ID	<u>Validation Status</u>	<u>Orientation /Strand</u>	Allele	Validation status description
ss24247001	PERLEGEN afd0987736		fwd/T	A/G	 Validated by multiple, independent submissions to the refSNP cluster
					Validated by frequency or genotype data: minor alleles observed in at least two chromosomes.
					Validated by submitter confirmation
					All alleles have been observed in at least two chromosomes apiece
					Genotyped by HapMap project

SNP Picking: Informative

Population Diversity

Population Diversity

	Sample Ascertainment					<u>Genotype Detail</u> ^{NEW}				Alleles	
<u>ss2</u>	ss#	Population	Individual Group	Chrom. Sample Cnt.	Source	<u>A/A</u>	<u>A/G</u>	<u>G/G</u>	HWP	<u>A</u>	<u>G</u>
	rs24246985	AFD_EUR_PANEL	European	48	IG	0.375	0.625	0.273		0.188	0.812
		AFD_AFR_PANEL	African American	46	IG		1.000			1.000	
		AFD_CHN_PANEL	Asian	48	IG	0.042	0.958	1.000		0.021	0.979
		HapMap-CEU	European	120	IG	0.050	0.317	0.633	1.000	0.208	0.792
		HapMap-HCB	Asian	90	IG	0.200	0.800	0.479		0.100	0.900
		HapMap-JPT	Asian	90	IG	0.156	0.844	0.584		0.078	0.922
		HapMap-YRI	Sub-Saharan African	120	IG		1.000			1.000	
		CHMJ		74	IG				0.001	0.041	0.959

(rs17421511)

SNP Picking: Potentially Functional

Region	Contig position	mRNA pos	dbSNP rs#	Hetero-zygosity	Validation	3D	OMIM	Function	dbSNP allele	Protein residue	Codon pos	Amino acid pos
exon_12	6388294	1832	rs2274976	0.129	  H			nonsynonymous	A	Gln [Q]	2	594
				0.129	  H			contig reference	G	Arg [R]	2	594
exon_11	6388686	1748	rs2274974	N.D.		H		nonsynonymous	A	Glu [E]	2	566
				N.D.		H		contig reference	G	Gly [G]	2	566
exon_8	6391824	1356	rs4846051	0.164	  H			synonymous	T	Phe [F]	3	435
				0.164	  H			contig reference	C	Phe [F]	3	435
	6391843	1337	rs1801131	0.351	  H			nonsynonymous	C	Ala [A]	2	429
				0.351	  H	Yes		contig reference	A	Glu [E]	2	429
exon_7	6392263	1107	rs2066462	0.176	  H			synonymous	T	Ser [S]	3	352
				0.176	  H			contig reference	C	Ser [S]	3	352
exon_6	6392608	996	rs6664734	N.D.		H	Yes	synonymous	A	Val [V]	3	315
				N.D.		H	Yes	contig reference	G	Val [V]	3	315
	2647	957	rs13306555	N.D.		H	Yes	synonymous	T	Ala [A]	3	302
				N.D.		H	Yes	contig reference	C	Ala [A]	3	302
exon_5	6393745	716	rs1801133	0.403	  H	Yes		nonsynonymous	T	Val [V]	2	222
				0.403	  H	Yes	Yes	contig reference	C	Ala [A]	2	222
	6393804	657	rs13306554	N.D.	  H	Yes		synonymous	T	Pro [P]	3	202
				N.D.	  H	Yes		contig reference	C	Pro [P]	3	202
exon_3	6398643	468	rs2066466	0.050	  H	Yes		synonymous	A	Thr [T]	3	139
				0.050	  H	Yes		contig reference	G	Thr [T]	3	139
	6398715	396	rs2066461	0.026		H	Yes	synonymous	A	Thr [T]	3	115

SNP Picking: Previously Associated

www.ncbi.nlm.nih.gov/pubmed?term=mthfr cancer

Comics R TclTk Jobs Slashdot fbook News Articles PLINK New Tab UCSF THProject Google Sites uw e-mail

NCBI Resources How To

PubMed.gov
US National Library of Medicine
National Institutes of Health

PubMed mthfr cancer

RSS Save search Advanced

[Show additional filters](#)

Article types
Clinical Trial
Review
more ...

Text availability
Abstract available
Free full text available
Full text available

Publication dates
5 years
10 years
Custom range...

Species
Humans
Other Animals

[Clear all](#)

[Show additional filters](#)

Display Settings: Summary, 20 per page, Sorted by Recently Added

Send to:

See 2454 articles about **MTHFR** gene function
See also: **MTHFR methylenetetrahydrofolate reductase (NAD(P)H)** in the Gene database
mthfr in [Homo sapiens](#) | [Mus musculus](#) | [Rattus norvegicus](#) | [All 29 Gene records](#)

Results: 1 to 20 of 1059 << First < Prev Page 1 of 53 Next > Last >>

☐ [The association between MTHFR C677T polymorphism and ovarian cancer risk: a meta-analysis of 18,628 individuals.](#)
Ma C, Liu Y, Zhang W, Liu P.
Mol Biol Rep. 2013 Jan 18. [Epub ahead of print]
PMID: 23329275 [PubMed - as supplied by publisher]
[Related citations](#)

☐ [Short-Term Folate Supplementation in Physiological Doses Has No Effect on ESR1 and MLH1 Methylation in Colonic Mucosa of Individuals with Adenoma.](#)
Al-Ghnam Abbadi R, Emery P, Pufulete M.
J Nutrigenet Nutrigenomics. 2013 Jan 16;5(6):327-338. [Epub ahead of print]
PMID: 23328702 [PubMed - as supplied by publisher]
[Related citations](#)

☐ [Pharmacogenetic analysis of adjuvant FOLFOX for Korean patients with colon cancer.](#)
Lee KH, Chang HJ, Han SW, Oh DY, Im SA, Bang YJ, Kim SY, Lee KW, Kim JH, Hong YS, Kim TW, Park YS, Kang WK, Shin SJ, Ahn JB, Kang GH, Jeong SY, Park KJ, Park JG, Kim TY.

MTHFR Summary

- **chr1**
- **Size: 20,329 bp**
- **Exons: 12**
- **Potentially Functional:**
5 missense of which 3 MAF >5%
- **Previously Associated:**
3 (C677T, A1298C, A2756G)

MTHFR Example

← browser.1000genomes.org/Homo_sapiens/Search/Results?site=ensembl&q=MTHFR

1000 Genomes

A Deep Catalog of Human Genetic Variation

Human ▾

Search 1000 Genomes

- New Search
- Results Summary**

⚙️ Configure this page

📄 Add your data

📄 Export data

📄 Get VCF data

🔖 Bookmark this page

Share this page

View in Ensembl

Results Summary

You searched for 'MTHFR'

Gene or Gene Product

3 entrie(s) matched your search strings.

- Gene:** [ENSG00000177000](#) Region in detail [\[1:11845780-11866977\]](#)
MTHFR - methylenetetrahydrofolate reductase (NAD(P)H) [Source:HGNC Symbol;Acc:7436]
- Variations in gene ENSG00000177000:** Variations in gene [\[1:11845780-11866977\]](#)
- Transcript:** [ENST00000376592](#) Region in detail [\[1:11845780-11863302\]](#)

Phenotype

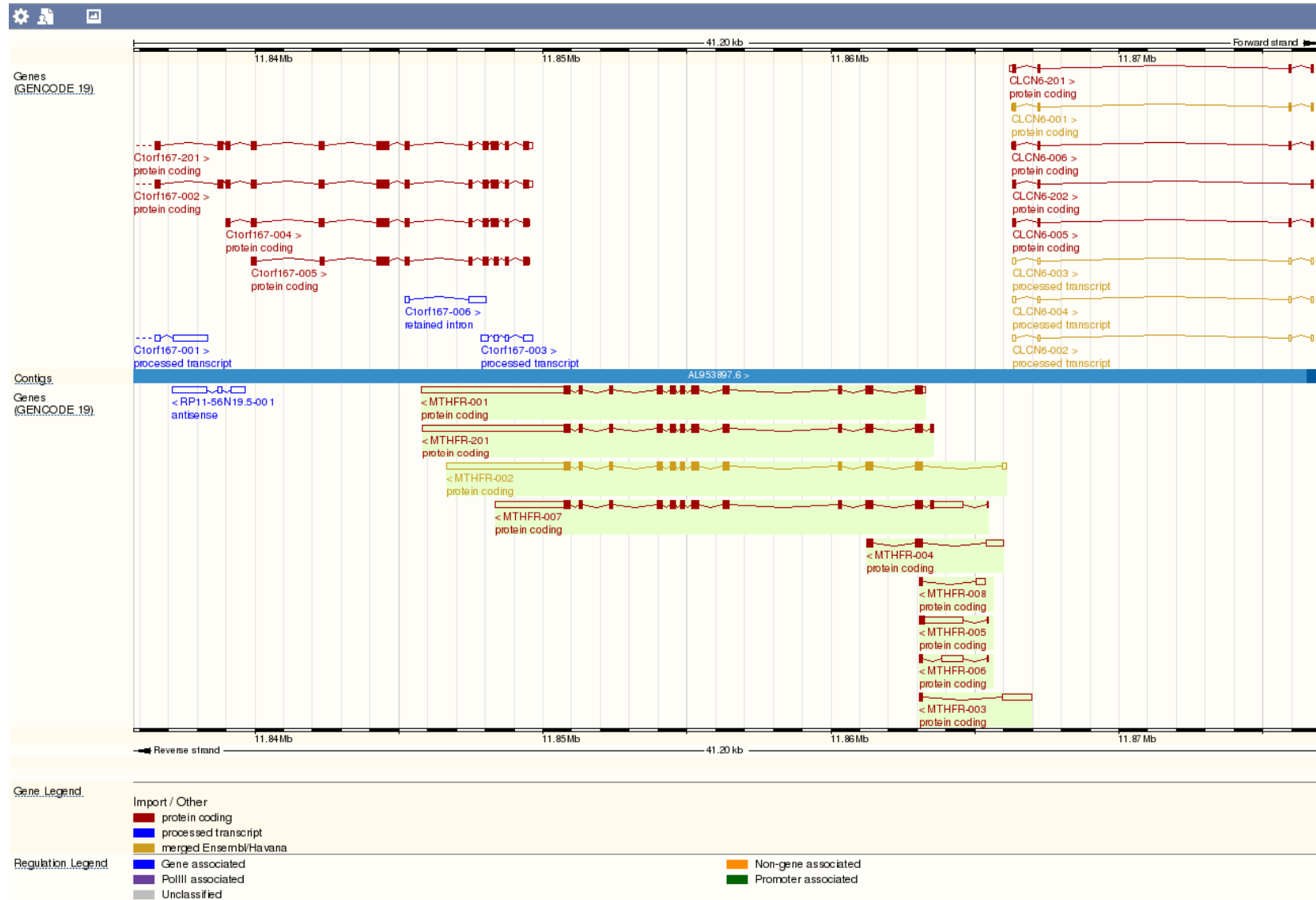
1 entrie(s) matched your search strings.

- ENSG00000177000 PH:** [11780](#)
methylenetetrahydrofolate reductase deficiency (MTHFRD)

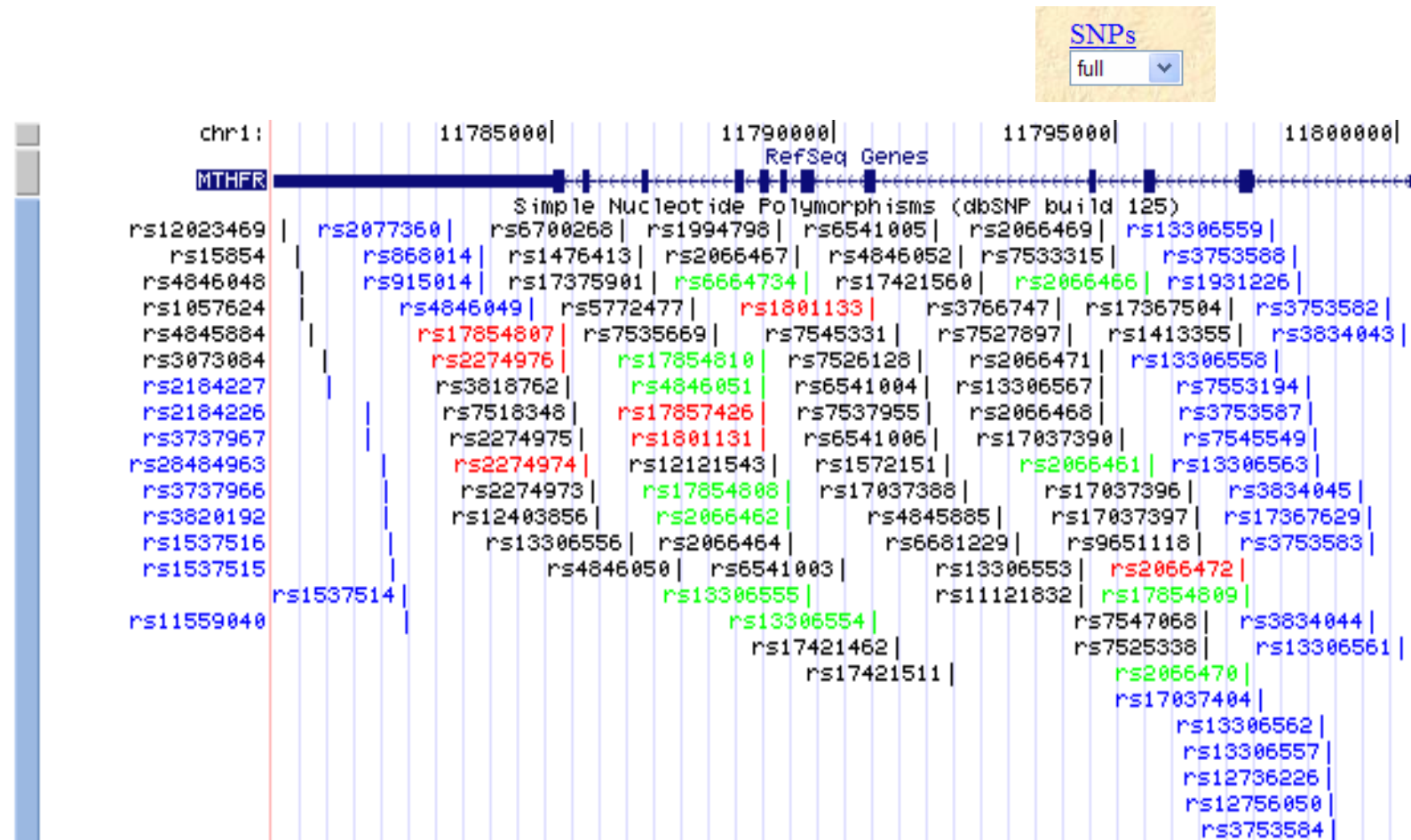
1000 Genomes release 16 - Oct 2014 © [EBI](#)

[About 1000 Genomes](#) | [Contact](#)

MTHFR Example



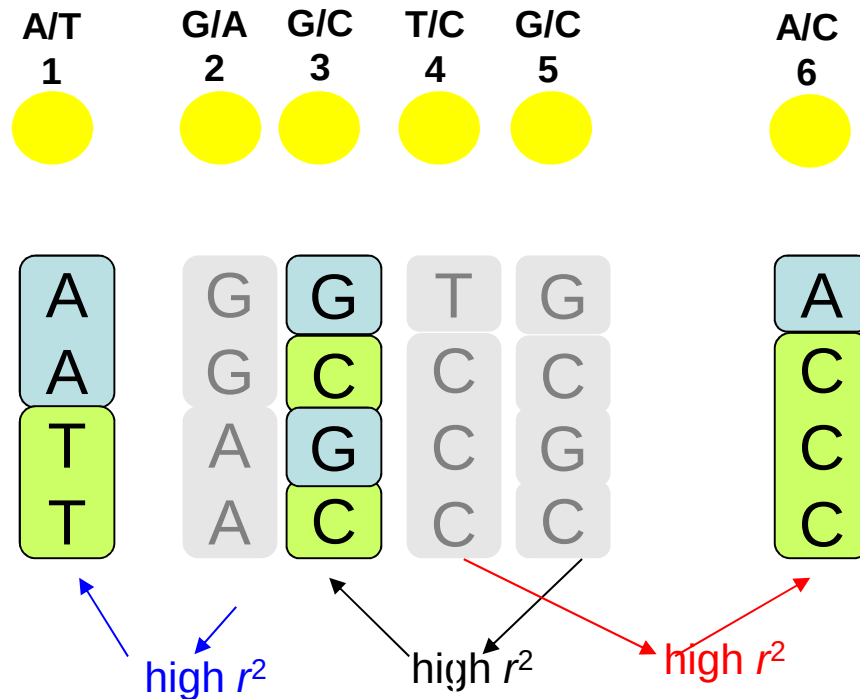
MTHFR SNPs



Too many MTHFR SNPs

Solution: Tag SNP Selection

- SNPs are correlated (aka Linkage Disequilibrium)



Pairwise Tagging:

SNP 1

SNP 3

SNP 6

3 tags in total

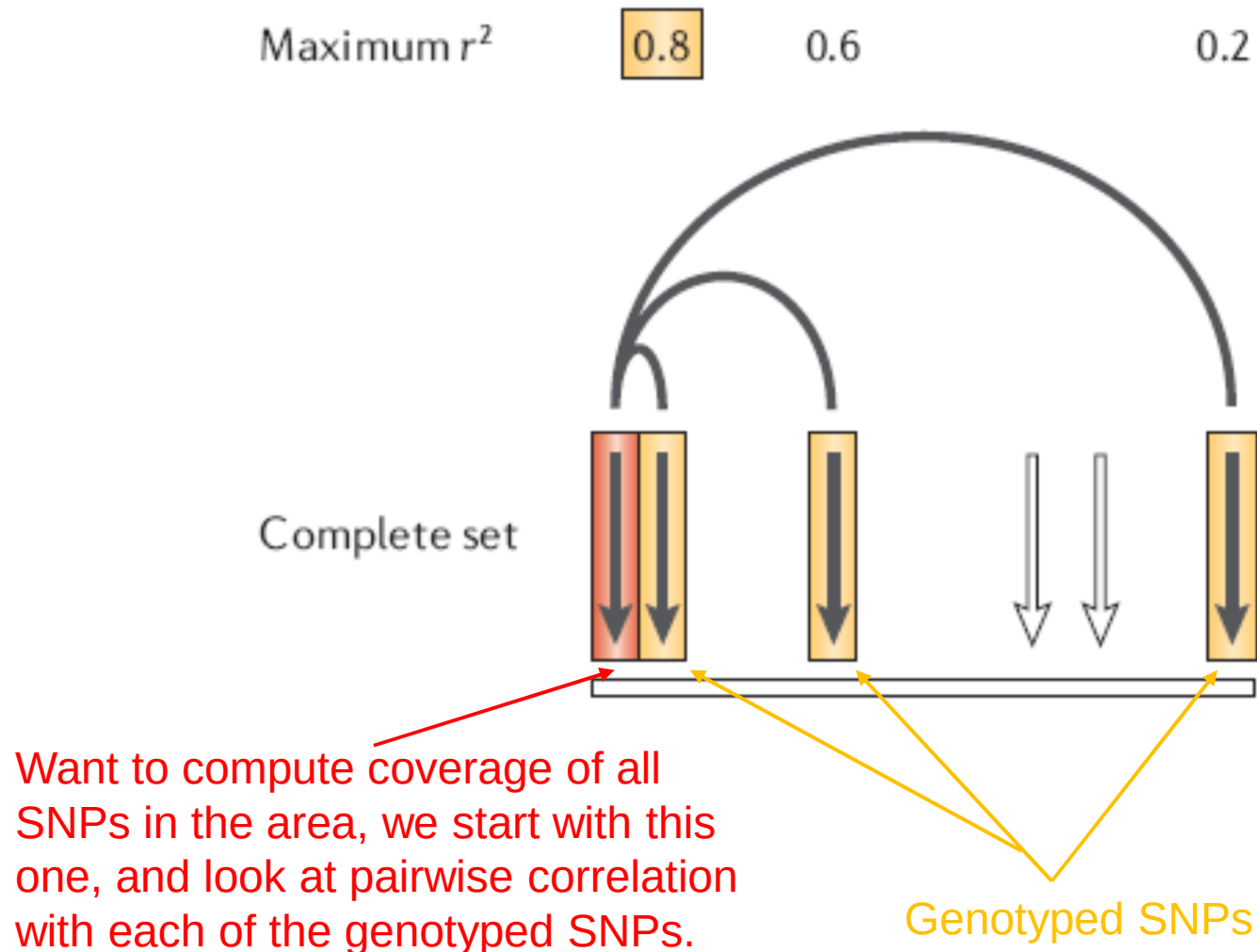
Test for association:

SNP 1

SNP 3

SNP 6

Coverage: Measurement Error in TagSNPs



Common Measures of Coverage

- **Threshold Measures**

- e.g., 73% of SNPs in the complete set are in LD with at least one SNP in the genotyping set at $r^2 > 0.8$

- **Average Measures**

- e.g., Average maximum $r^2 = 0.84$

Coverage and Sample Size

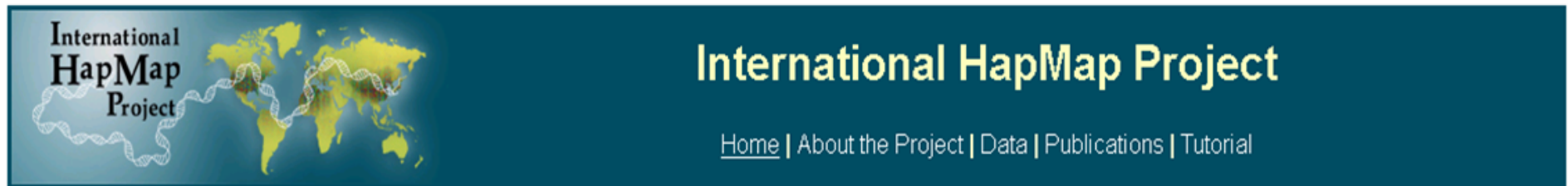
- Sample size required for Direct Association, n
- Sample size for Indirect Association

$$n^* = n / r^2$$

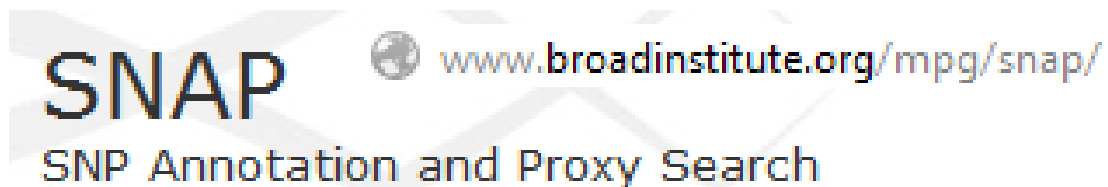
- For $r^2 = 0.8$, increase is 25%
- For $r^2 = 0.5$, increase is 100%

Tag SNPs Database Resources

<http://www.hapmap.org>



<http://browser.1000genomes.org/index.html>



HapMap

- Re-sequencing to discover millions of additional SNPs; deposited to dbSNP.
- SNPs from dbSNP were genotyped
- Looked for 1 SNP every 5kb
- SNP Validation
 - Polymorphic
 - Frequency
- Haplotype and Linkage Disequilibrium Estimation
 - LD tagging SNPs

HapMap Phase III Populations

- ASW African ancestry in Southwest USA
- CEU Utah residents with Northern and Western European ancestry from the CEPH collection
- CHB Han Chinese in Beijing, China
- CHD Chinese in Metropolitan Denver, Colorado
- GIH Gujarati Indians in Houston, Texas
- JPT Japanese in Tokyo, Japan
- LWK Luhya in Webuye, Kenya
- MEX Mexican ancestry in Los Angeles, California
- MKK Maasai in Kinyawa, Kenya
- TSI Toscani in Italia
- YRI Yoruba in Ibadan, Nigeria

Tag SNPs: HapMap



International HapMap Project

[Home](#) | [About the Project](#) | [Data](#) | [Publications](#) | [Tutorial](#)

[中文](#) | [English](#) | [Français](#) | [日本語](#) | [Yoruba](#)

The International HapMap Project is a partnership of scientists and funding agencies from Canada, China, Japan, Nigeria, the United Kingdom and the United States. Researchers find genes associated with human disease and response to pharmaceuticals. See "[About the International HapMap Project](#)" for more information.

Project Information

[About the Project](#)
[HapMap Publications](#)
[HapMap Tutorial](#)
[HapMap Mailing List](#)
[HapMap Project Participants](#)
[HapMap Mirror Site in Japan](#)

Project Data

[Browse Project Data](#)
[HapMap](#)
[Bulk Data Download](#)
[Data Freezes for Publication](#)
[ENCODE Project](#)
[Guidelines For Data Use](#)

Useful Links

[HapMap Project Press Release](#)
[NHGRI HapMap Page](#)

News

• 2007-01-17: **Reactome links updated**

Links to the [Reactome](#) pathway database from the HapMap genome browser have been updated and

• 2007-01-14: **HapMap Public Release #21a**

Genotype frequencies and assays for phase I and phase II of the HapMap project are now available for

This release contains all processed data from the HapMap project, and includes all data from phase I genotypes from the Illumina Infinium 100k and 300k genotyping arrays, Affymetrix nsSNPs, and SNPs (She et al. 2006) of the extended MHC locus.

Populations	CEU	CHB+JPT	YRI
Total Non-Redundant	3,904,218	3,936,482	3,846,092
Total QC+ SNPs	4,871,127	4,881,441	4,774,448
Total Genotyped SNPs	6,838,923	6,799,238	6,798,546

New features in this release include Segmental Duplications and Copy Number Variation (CNV). [Depth Celera Reads](#) (She et al. 2004; Bailey et al. 2001 & 2002). Structural Variation datasets include

[CNV regions determined in HapMap samples \(Bader et al. 2006\)](#)

Tag SNPs: HapMap

Showing 19.3 kbp from chr1, positions 11,780,944 to 11,800,247

Instructions

Search using a sequence name, gene name, locus, or other landmark. The wildcard character * is allowed. To center on a location, click the ruler. Use the Scroll/Zoom buttons to center on a position.

Examples : Chr20 , Chr9:660,000..760,000 , SNP:rs6870660 , NM_153254 , BRCA2 , 5q31 , ENM010 .

[Hide banner] [Bookmark this] [Link to Image] [SNP genotype data] [tag SNP Data] [HapMap LD Data] [Phased Haplotype Data] [High-res Image] [Help] [Reset]

Search

Help links: - LD - - tagSNPs - - Phased Haplotype - - Genotype data - - Frequency data - - Symbols

Landmark or Region :

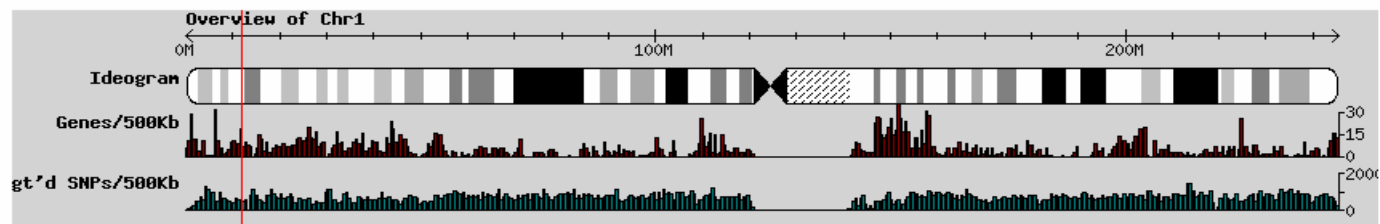
Data Source : HapMap Data Rel 21a/phaseII Ene07, on NCBI B35 assembly, dbSNP b125

Reports & Analysis :

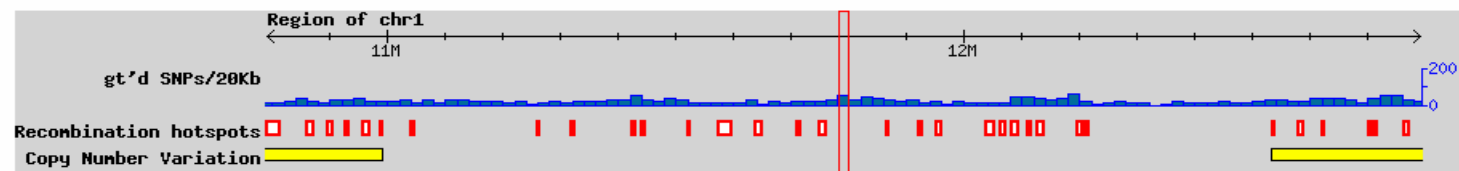
Scroll/Zoom: <<< << < > >> >>> Show 19.3 kbp

Population descriptors: YRI: Yoruba in Ibadan, Nigeria, JPT: Japanese in Tokyo, Japan, CHB: Han Chinese in Beijing, China, CEU: CEPH (Utah residents with ancestry from Europe)

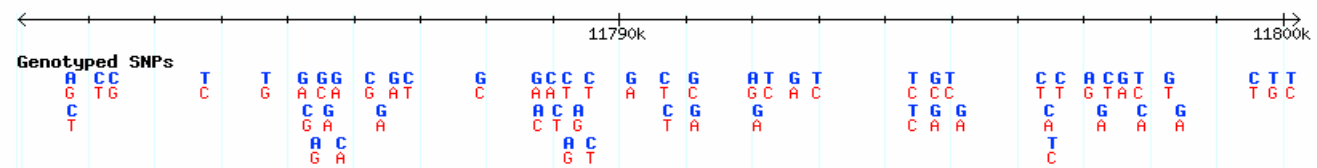
Overview



Region



Details



Tag SNPs: HapMap & Haploview

Configure... SNP genotype data

Population

Strand

Output format

CEU

rs

☐ text

☐ Save to Disk

☒ Open directly in HaploView [NB doesn't work on all OS platforms or browsers]

Cancel

Configure

Go



<http://www.broad.mit.edu/mpg/haploview/>

OR Just choose “HapMap Download” directly from Haploview (instead of “HapMap Format”, and choose Chromosome 1, from 11,768Kb-11,789Kb

Tag SNPs: HapMap & Haploview

File Display Analysis Help

LD Plot Haplotypes **Check Markers** Tagger

#	Name	Position	Obs...	Pre...	HW...	%G...	Fam...	Me...	MAF	Rating
1	rs2184226	11781702	0.156	0.167	1.0	100.0	30	0	0.092	☒
2	rs3737967	11781715	0.094	0.101	0.263	94.4	27	1	0.054	☒
3	rs1537516	11782127	0.218	0.185	1.0	96.7	27	0	0.103	☒
4	rs1537514	11782334	0.247	0.206	0.849	98.9	29	0	0.117	☒
5	rs868014	11783713	0.0	0.0	1.0	100.0	30	0	0.0	☐
6	rs4846049	11784631	0.449	0.469	1.0	98.9	29	0	0.375	☒
7	rs2274976	11785193	0.122	0.11	1.0	100.0	30	0	0.058	☒
8	rs3818762	11785269	0.448	0.428	1.0	96.7	27	0	0.31	☒
9	rs7518348	11785384	0.0	0.0	1.0	95.6	26	0	0.0	☐
10	rs2274975	11785483	0.0	0.0	1.0	98.9	29	0	0.0	☐
11	rs2274974	11785585	0.011	0.017	1.0	97.8	28	0	0.0090	☐
12	rs2274973	11785690	0.0	0.0	1.0	100.0	30	0	0.0	☐
13	rs12403856	11785777	0.0	0.0	1.0	100.0	30	0	0.0	☐
14	rs6700268	11786216	0.0	0.0	1.0	98.9	29	0	0.0	☐
15	rs13306556	11786376	0.233	0.206	0.849	100.0	30	0	0.117	☒
16	rs1476413	11786566	0.506	0.45	0.83	98.9	29	0	0.342	☒
17	rs17375901	11786782	0.122	0.11	1.0	100.0	30	0	0.058	☒
18	rs7535669	11787865	0.0	0.0	1.0	98.9	29	0	0.0	☐
19	rs4846051	11788723	0.0	0.0	1.0	100.0	30	0	0.0	☐
20	rs1801131	11788742	0.402	0.462	0.988	96.7	29	1	0.362	☒
21	rs12121543	11788937	0.404	0.399	0.552	98.9	29	0	0.275	☒
22	rs1994798	11789021	0.513	0.498	0.609	86.7	20	0	0.471	☒
23	rs2066462	11789162	0.259	0.215	0.807	94.4	25	0	0.123	☒
24	rs2066464	11789192	0.0	0.0	1.0	100.0	30	0	0.0	☐
25	rs2066467	11789338	0.0	0.0	1.0	95.6	26	0	0.0	☐

HW p-value cutoff: 0.0010

Min genotype %: 75

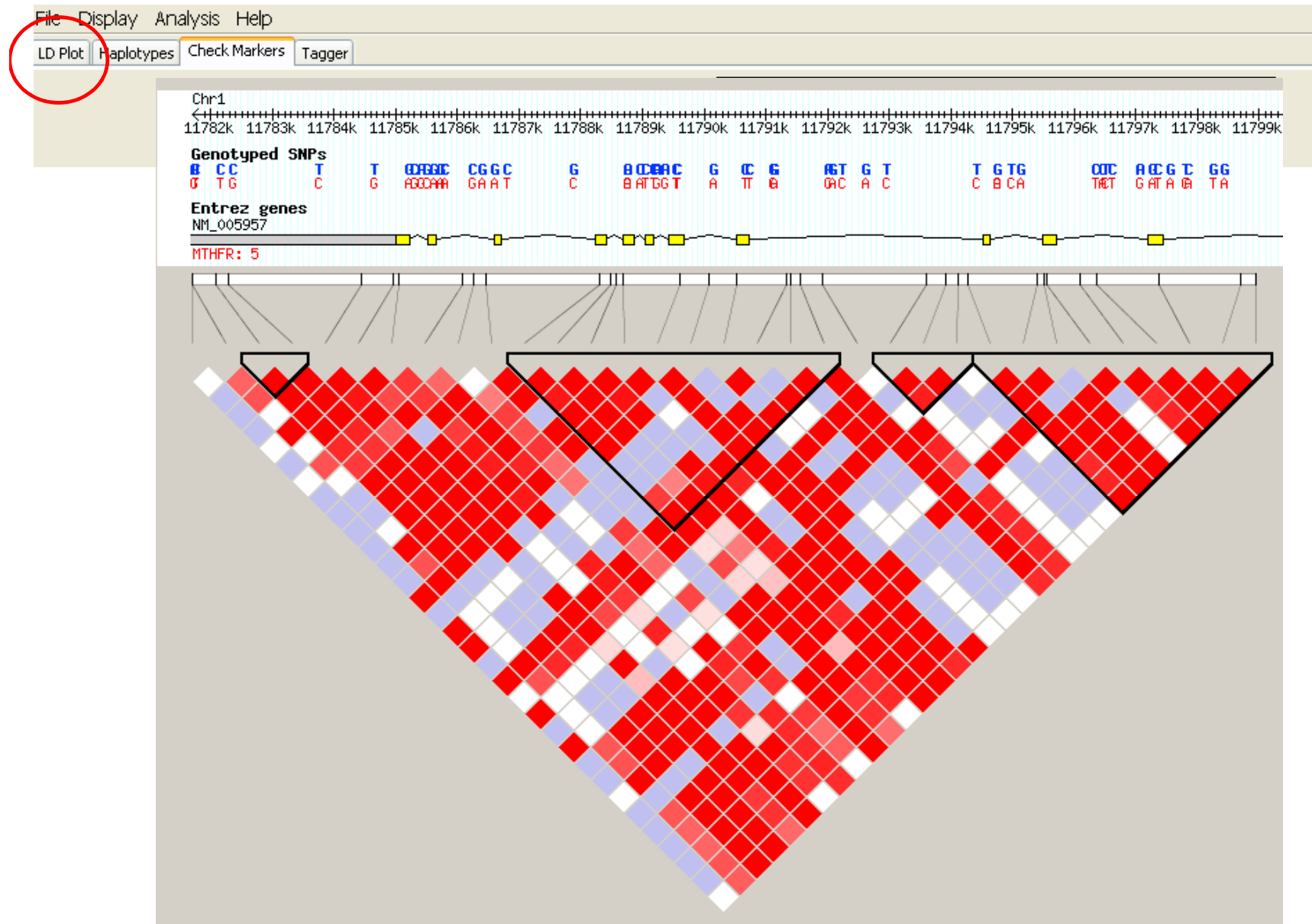
Max # mendel errors: 1

Minimum minor allele freq: 0.05

Select All

Rescore Markers

Tag SNPs: HapMap & Haploview



Tag SNPs: HapMap & Haploview

File Display Analysis Help

LD Plot Haplotypes Check Markers **Tagger**

Configuration Results

#	Name	Position	Force Incl...	Force Excl...	Capture this All...
1	rs2184226	11781702	☐	☐	☒
2	rs3737967	11781715	☐	☐	☒
3	rs1537516	11782127	☐	☐	☒
4	rs1537514	11782334	☐	☐	☒
6	rs4846049	11784631	☐	☐	☒
7	rs2274976	11785193	☒	☐	☒
8	rs3818762	11785269	☐	☐	☒
15	rs13306556	11786376	☐	☐	☒
16	rs1476413	11786566	☐	☐	☒
17	rs17375901	11786782	☐	☐	☒
20	rs1801131	11788742	☒	☐	☒
21	rs12121543	11788937	☐	☐	☒
22	rs1994798	11789021	☐	☐	☒
23	rs2066462	11789162	☐	☐	☒
28	rs6541003	11790133	☐	☐	☒
29	rs1801133	11790644	☒	☐	☒
32	rs17421462	11791113	☐	☐	☒
33	rs1572151	11791977	☐	☐	☒
34	rs17421511	11792054	☐	☐	☒
35	rs4846052	11792217	☐	☐	☒
36	rs17421560	11792590	☐	☐	☒
38	rs11121832	11794386	☐	☐	☒
39	rs2066471	11794724	☐	☐	☒
41	rs7533315	11794949	☐	☐	☒
42	rs17037390	11795109	☐	☐	☒

☒ pairwise tagging only r² threshold 0.8

☐ aggressive tagging: use 2-marker haplotypes

☐ aggressive tagging: use 2- and 3-marker haplotypes LOD threshold for multi-marker tests 3.0

Run Tagger **Reset Table**

Tag SNPs: HapMap & Haploview

File Display Analysis Help

LD Plot Haplotypes Check Markers **Tagger**

Configuration **Results**

rs2274976

rs1801131

rs1801133

rs2066462

rs2184226

rs4846052

rs17037390

rs11121832

rs2066471

rs13306561

rs12121543

rs3818762

rs1476413

rs17375901

rs9651118

Alleles captured by Current S

rs3737967

rs2274976

rs17037397

Allele	Test	r ²
rs2184226	rs2184226	1.0
rs3737967	rs2274976	0.826
rs1537516	rs2066462	1.0
rs1537514	rs2066462	1.0
rs4846049	rs1801131	0.929
rs2274976	rs2274976	1.0
rs3818762	rs3818762	1.0
rs13306556	rs2066462	0.844
rs1476413	rs1476413	1.0
rs17375901	rs17375901	1.0
rs1801131	rs1801131	1.0
rs12121543	rs12121543	1.0
rs1994798	rs4846052	0.924
rs2066462	rs2066462	1.0
rs6541003	rs4846052	1.0
rs1801133	rs1801133	1.0
rs17421462	rs2184226	1.0
rs1572151	rs2184226	1.0
rs17421511	rs2066471	1.0
rs4846052	rs4846052	1.0
rs17421560	rs2066462	0.844
rs11121832	rs11121832	1.0
rs2066471	rs2066471	1.0
rs7533315	rs11121832	0.815
rs17037390	rs17037390	1.0
rs17037396	rs2066462	1.0
rs17037397	rs2274976	1.0
rs9651118	rs9651118	1.0
rs17367504	rs17037390	1.0
rs2066470	rs2066462	1.0
rs7553194	rs2066462	0.914
rs3753582	rs2066462	0.924
rs13306561	rs13306561	1.0

Captured 33 alleles with mean r² of 0.97
 Captured 100 percent of alleles with r² > 0.8
 Using 15 SNPs in 15 tests.

Dump Tests File

Tag SNPs: HapMap Summary

- Identified 33 common MTHR SNPs (MAF > 5%) among Caucasians
- Forced in 3 potentially functional/previously associated SNPs
- Identified tag based on pairwise tagging
- 15 tags SNPs could capture all 33 MTHR SNPs (mean $r^2 = 97\%$)
- Note: number of SNPs required varies from gene to gene and from population to population

Extra Slides

Haploview using 1000 Genomes Project data

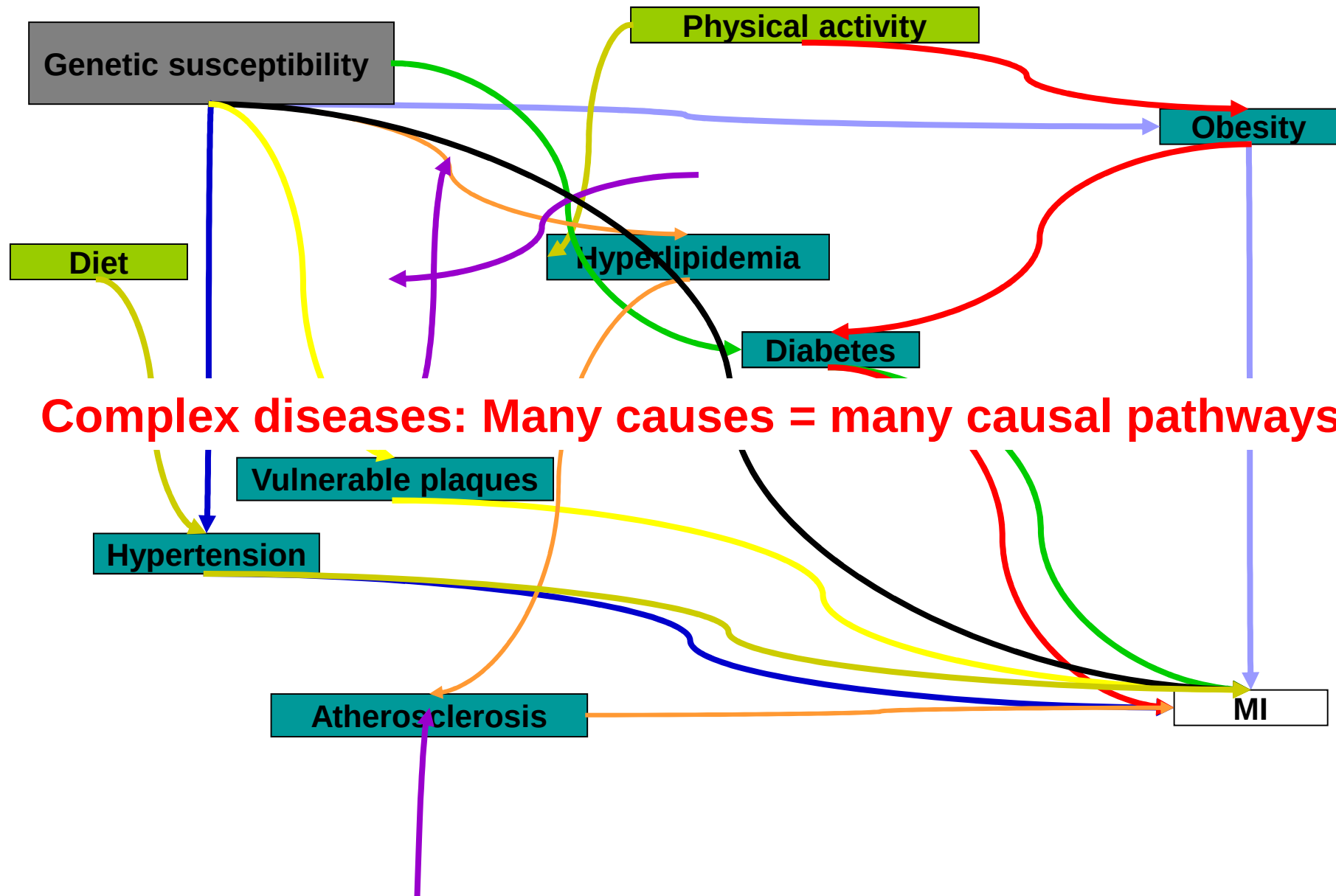
“First, determine the genomic coordinates of region you are interested in. This needs to be in hg19 coordinates. If you have hg18 coordinates, [liftOver](#) is a useful tool to convert coordinates from one human genome build to another (liftOver format is chr:start-end, for example: chr8:1000-50000).

Next, go to this 1000G [website](#), and plug in your genomic coordinates of interest. Here the coordinates should not include chr in the chromosome name (for example: 8:1000-50000). Then on the next page select ancestral populations you are interested in (you can select multiple populations by holding down Ctrl). Give the website a few seconds to generate the files. Eventually a link to a marker information file (.info) and linkage pedigree file (.ped) will appear. Right click on each of these files and save them to your computer.

Now, fire up Haploview and select Open new data. Go to the Linkage Format tab and browse for your .ped file in the Data File field and your .info file in the Locus Information File field (the .info file field is usually automatically generated after selecting your .ped file if your .ped and .info files have the same prefix). Haploview will load the files and you should be ready to visualize the LD structure. Enjoy using 1000G data in Haploview!”

genometoolbox.blogspot.com/2013/05/download-one-thousand-genome-data-for.html

4. Pathways



Pathways

- May be interested in potential joint and/or interaction effects of multiple genes in one pathway
- Many sites provide models of molecular and biochemical pathways.
- Example: BioCarta: <http://www.biocarta.com/>

Polygenic Models

- Many weak associations combine to $\rightarrow$ risk?
- Score model:

$$x_j = \frac{\sum_{i=1}^m \ln(OR_i) \times SNP_{ij}}{m}$$

where

- $\ln(OR_i)$ = 'score' for SNP_i from 'discovery' sample
- SNP_{ij} = # of alleles (0,1,2) for SNP_i , person j in 'validation' sample.
- Large number of SNPs (m)
- x_j associated with disease?

Application of Model

