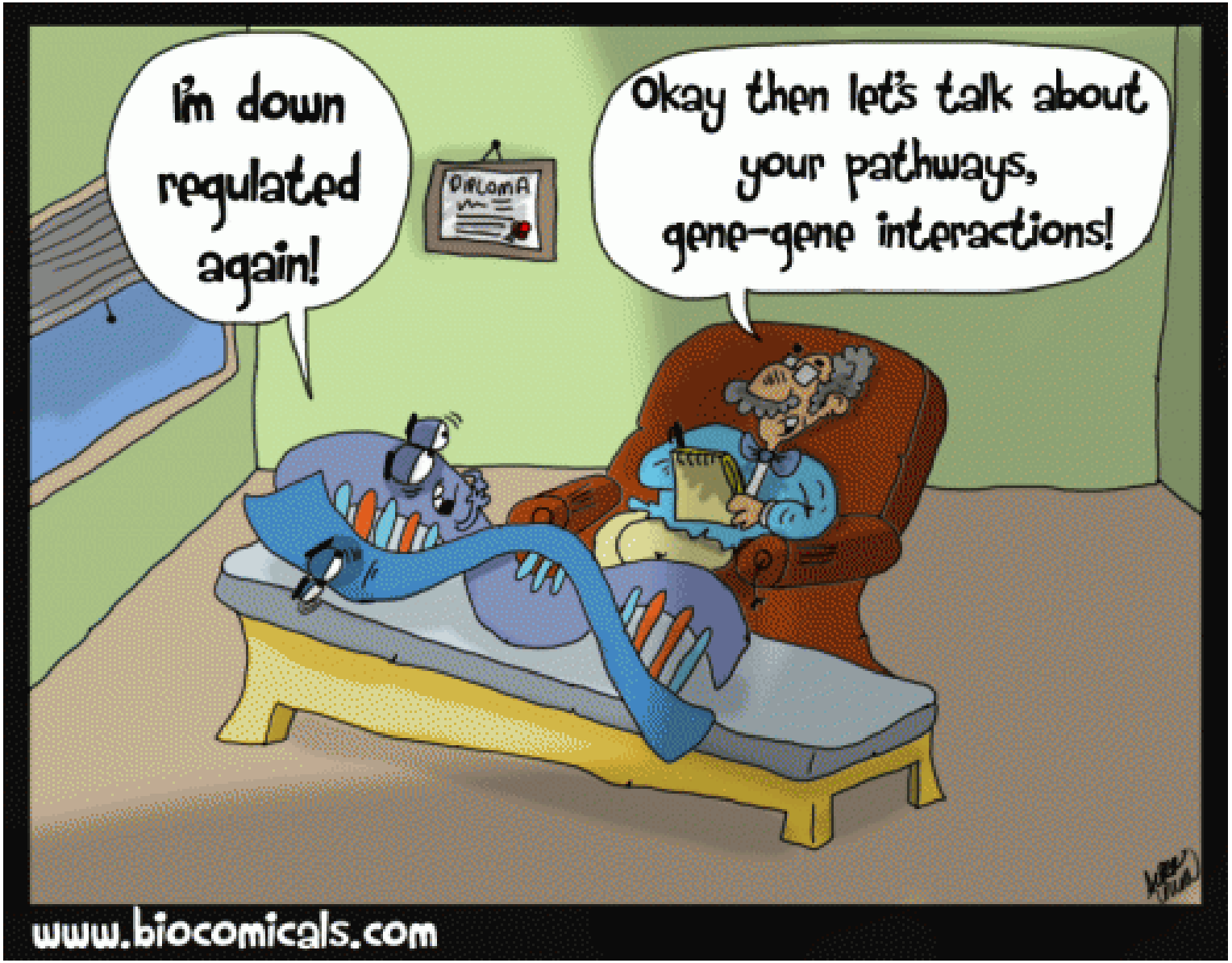
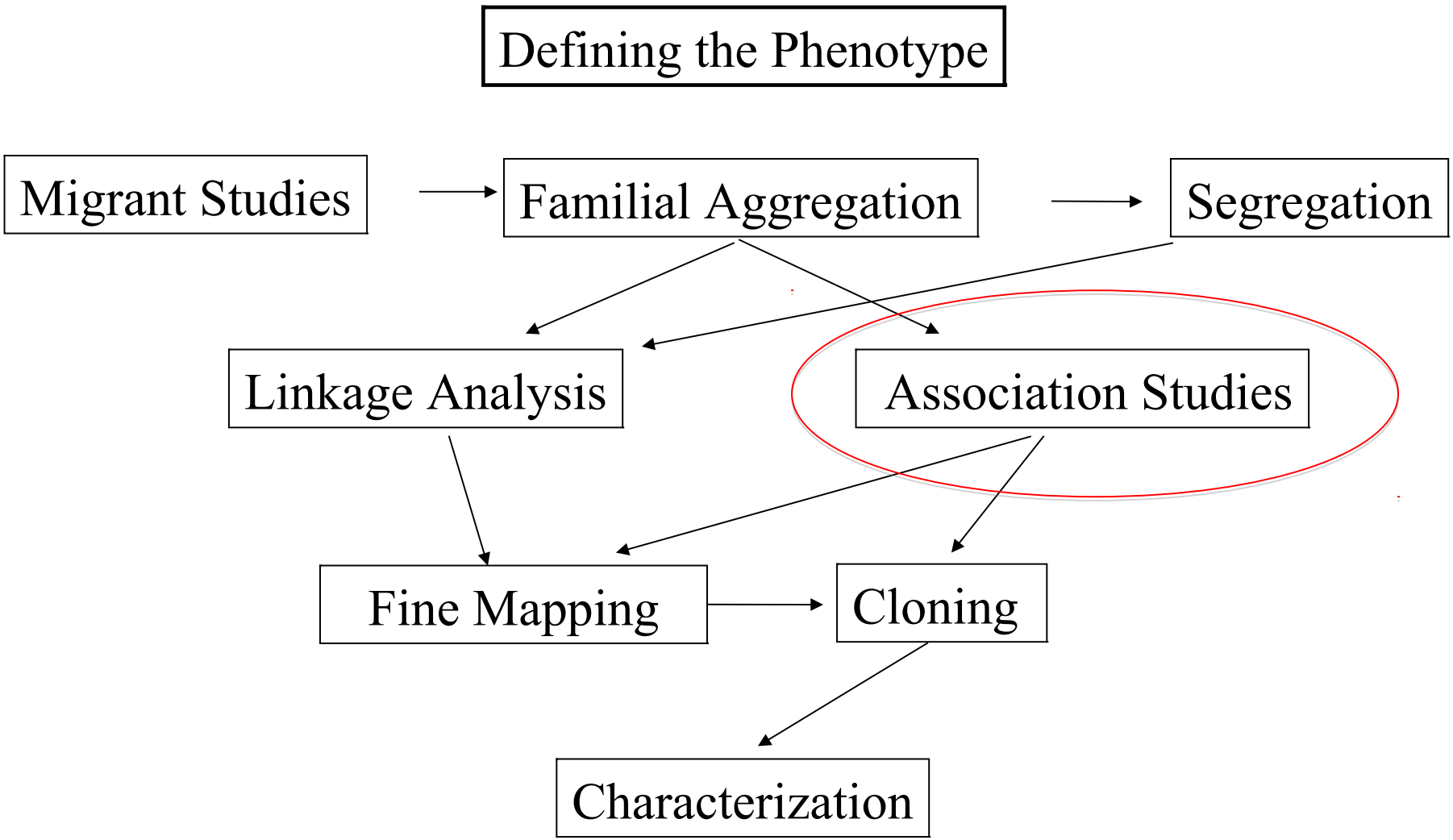
Association Studies



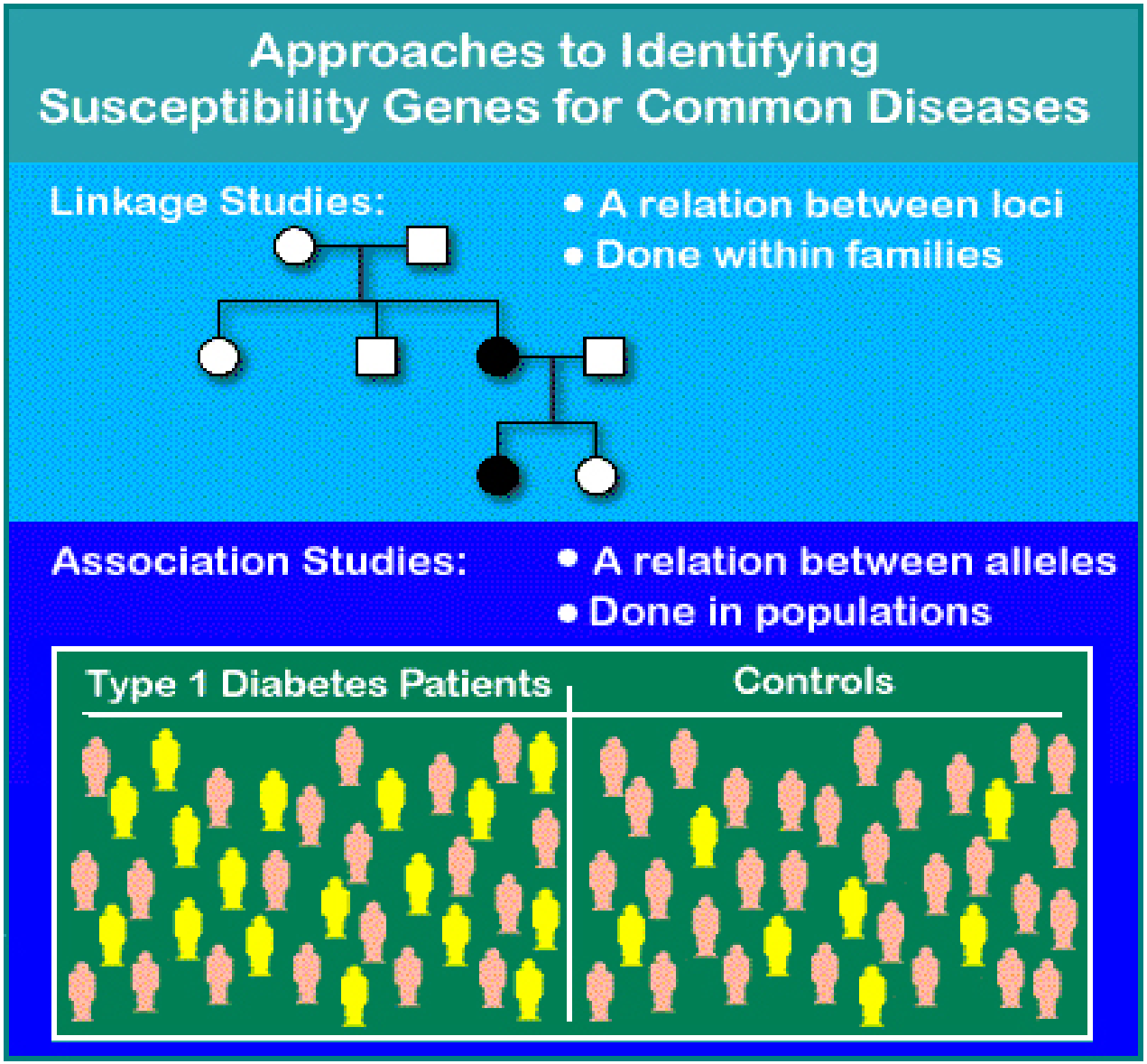
Overview

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

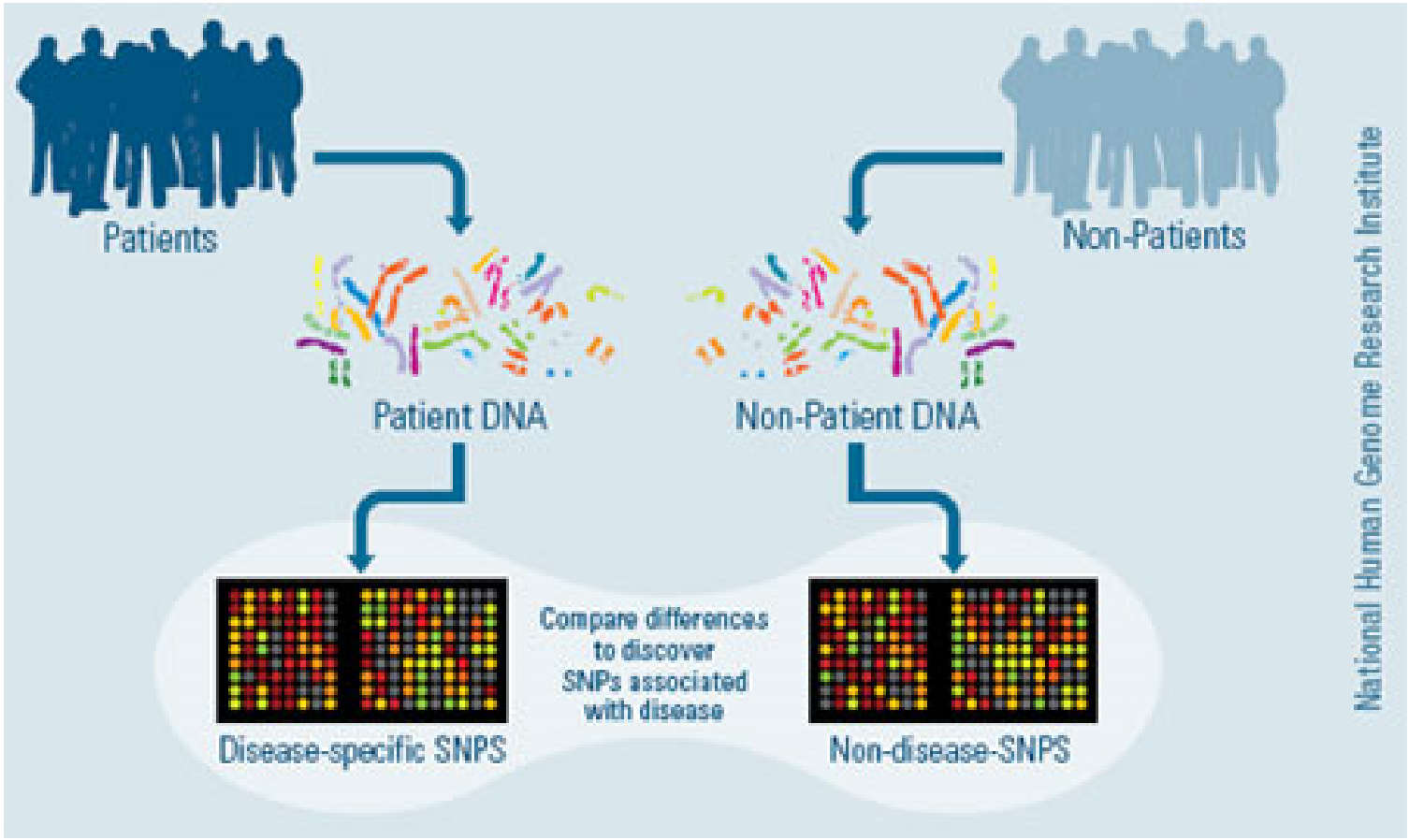
Process of Genetic Epidemiology



Association Studies



Association Studies



Association Studies

- Use of 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

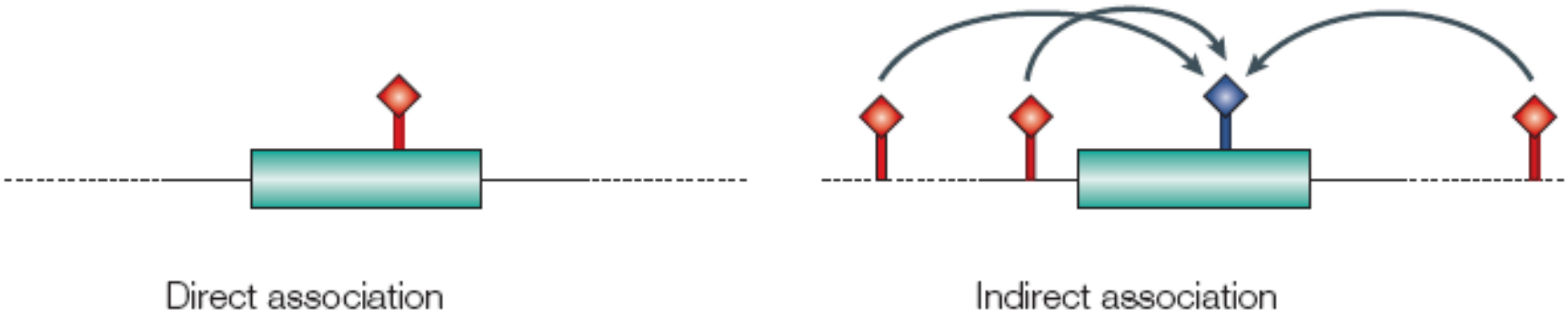
Linkage vs. Association

		Linkage			Association			
					Singletons		Sib pairs	
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(tr-A)	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

Candidate Gene or GWAS

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

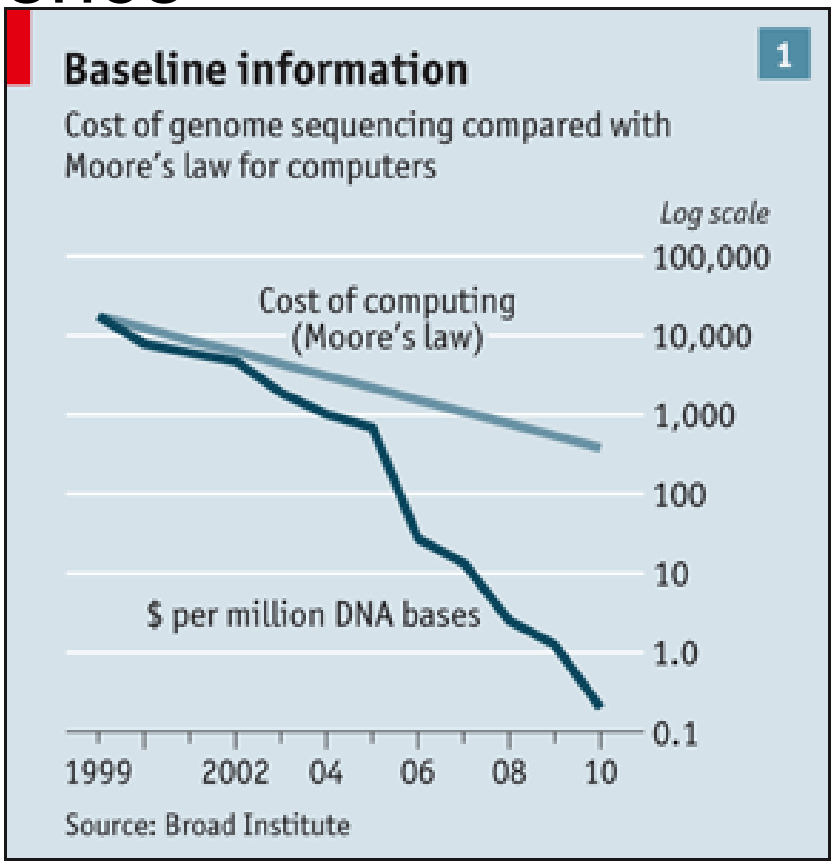
Now: 1 week, $1K

> 500 times faster

< one millionth the cost!

Soon: 1 hour, $1K

(#1 Innovation, 2010)



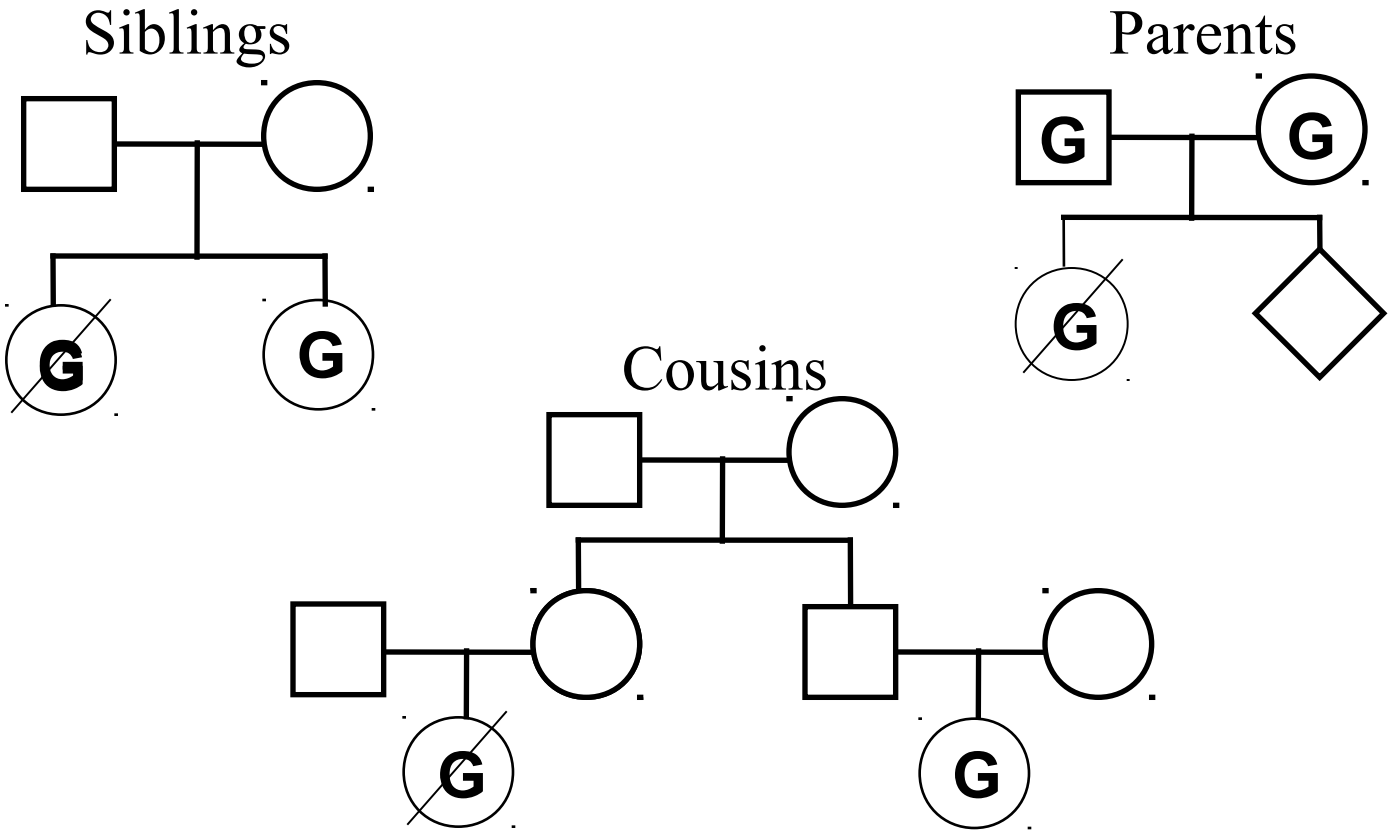
The Economist, 2010

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.

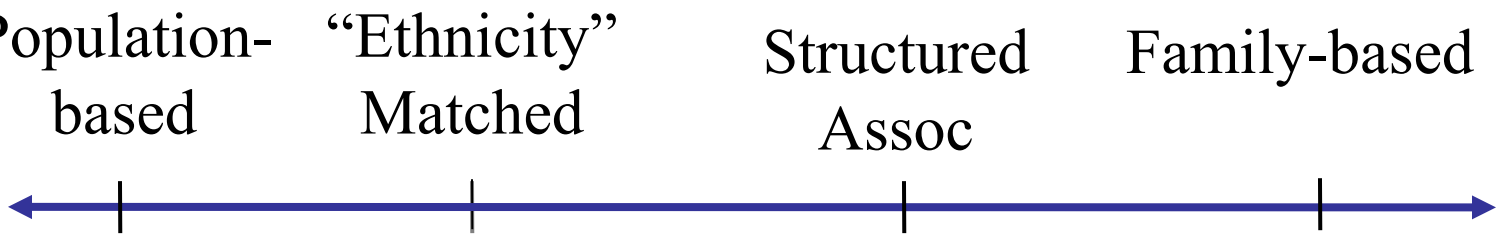
Family-Based Association Studies

(Covered in last lecture)



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

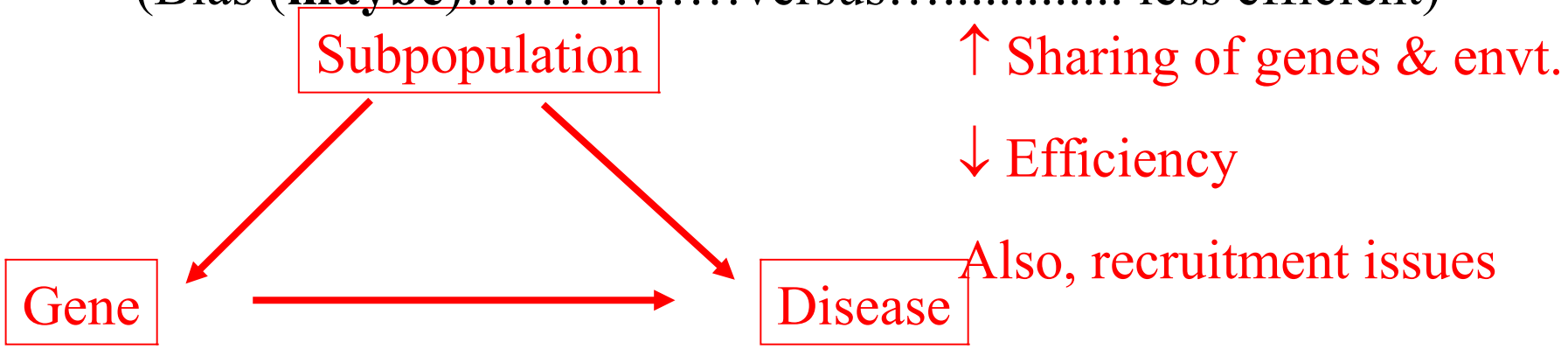
Continuum of Assoc Study Designs



Population Stratification

Overmatching

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



↑ Sharing of genes & envt.

↓ Efficiency

Also, recruitment issues

Association Analysis

Genotype	Cases	Controls	OR
GG	A	D	1
GT	B	E	BD/AE
TT	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
CC	A	D	A+D=nCC
CT	B	E	B+E=nCT
TT	C	F	C+F=nTT
Total	A+B+C =nCase	D+E+F =nCont	A+B+C+D+E+F =n

Expected

Case	Control
nCC*nCase/n	nCC*nCont/n
nCT*nCase/n	nCT*nCont/n
nTT*nCase/n	nTT*nCont/n

Sum (Observed – Expected)^2/Expected as before. Chi squared with 2 degrees of freedom.

Hint: Expected cell count = row_total * column_total / total

Simple chi-quared test example

Observed:				Expected	
Geno	Case	Control	Total	Case	Control
CC	20	5	25	25*35/65=13.5	25*30/65=11.5
CT	10	10	20	20*35/65=10.8	20*30/65=9.2
TT	5	15	20	20*35/65=10.8	20*30/65=9.2
Total	35	30	65		
	=nCase	=nCont	=n		

Sum (Observed – Expected)^2/Expected

= (20-13.5)^2/13.5 + (10-10.8)^2/10.8 + (5-10.8)^2/10.8
+ (5-11.5)^2/11.5 + (10-9.2)^2/9.2 + (15-9.2)^2/9.2

= 13.7

P-value = 0.0011

Genetic Model

Genotype	OR
GG	1
GT	r
TT	R

ORs depend on genetic model

R = r = 1 not risk allele

R > r = 1 recessive

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)
- 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)**			**Additive**	
Genotype	Case	Control		Case	Control		Case	Control		Case	Control
GG	A	D	GG	A	D	GG or GT	A+B	D+E	G	2*A+B	2*D+E
GT	B	E	GT or TT	B+C	E+F	TT	C	F	T	2*C+B	2*F+E
TT	C	F									
~chi_sq(2df)				~chi_sq(1df)			~chi_sq(1df)			~chi_sq(1df)	
Genotype	Case	Control		Case	Control		Case	Control		Case	Control
GG	20	5	GG	20	5	GG or GT	30	15	G	50	20
GT	10	10	GT or TT	15	25	TT	5	15	T	20	40
TT	5	15									
	P=0.0011			P=0.0020			P=0.0045			P=0.000031	

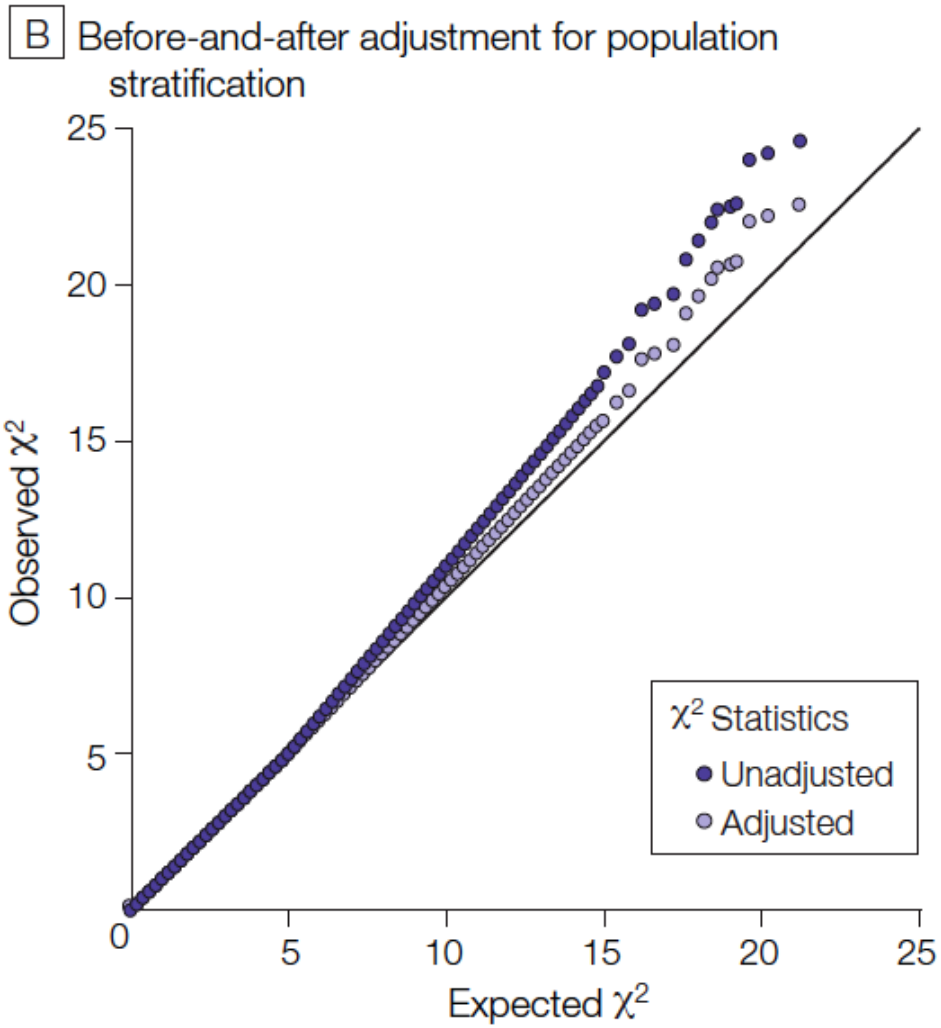
Review: Population stratification

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

Correcting for population stratification

- "Homogeneous" group (and ignore)
- Genomic control (PMID: 11315092)
  - 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 ($\chi^2_{new} = \chi^2_{old}/\lambda$)
  - 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!
- 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)

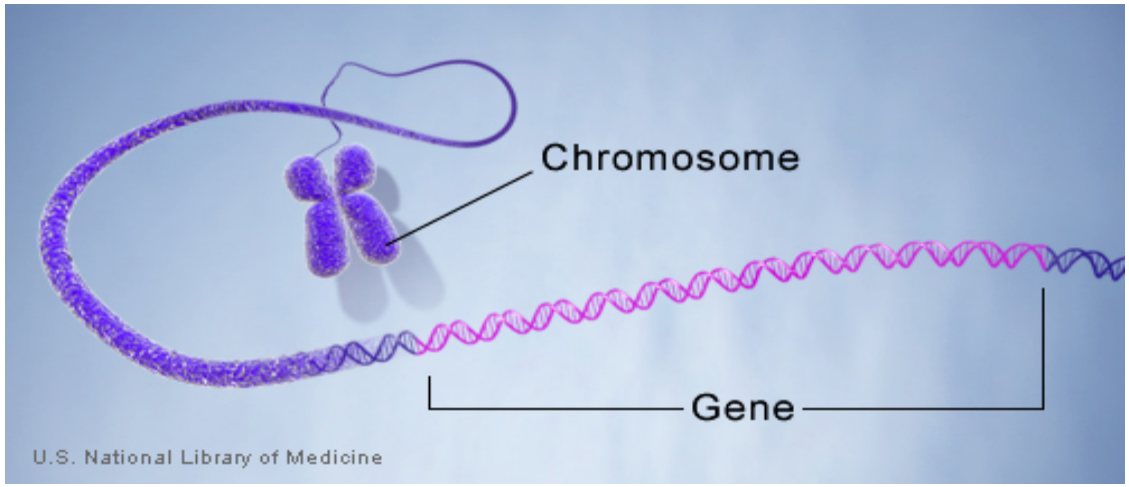
Quantile-Quantile (QQ) Plots



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?

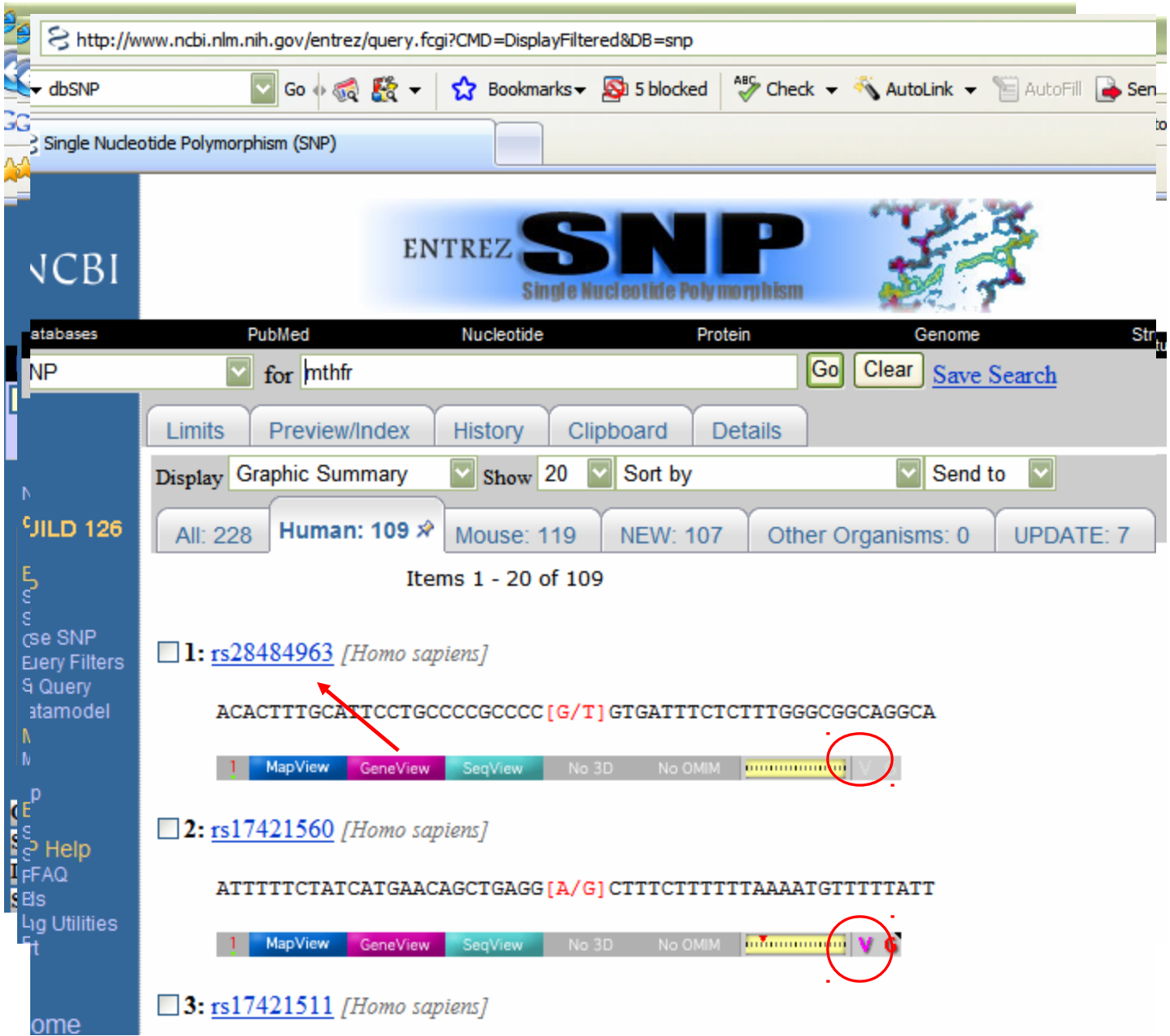
- Size:
  How large is the gene?

Use UCSC genome browser.

SNP Picking: Things to Consider

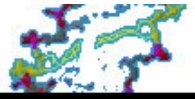
- Validation: What is the quality of the SNPs?
- Informativity: Are these SNPs informative in my population? How common are they? Location?
- Potentially Functional: Do these SNPs have a potential biological impact? Missense variants?
- Previously Associated: Have previous studies found SNPs in the candidate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

SNP Picking: Validation



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	
Organism:	human (*Homo sapiens*)
Molecule Type:	Genomic
Created/Updated in build:	125/125
Map to Genome Build:	36.1

Allele	
Variation Class:	SNP: single nucleotide polymorphism
Alleles:	G/T
Ancestral Allele:	T

SNP Details are organized in the following sections:

Submission | Fasta | Resource | GeneView | Map | Diversity | Validation

Submitter records for this RefSNP Cluster

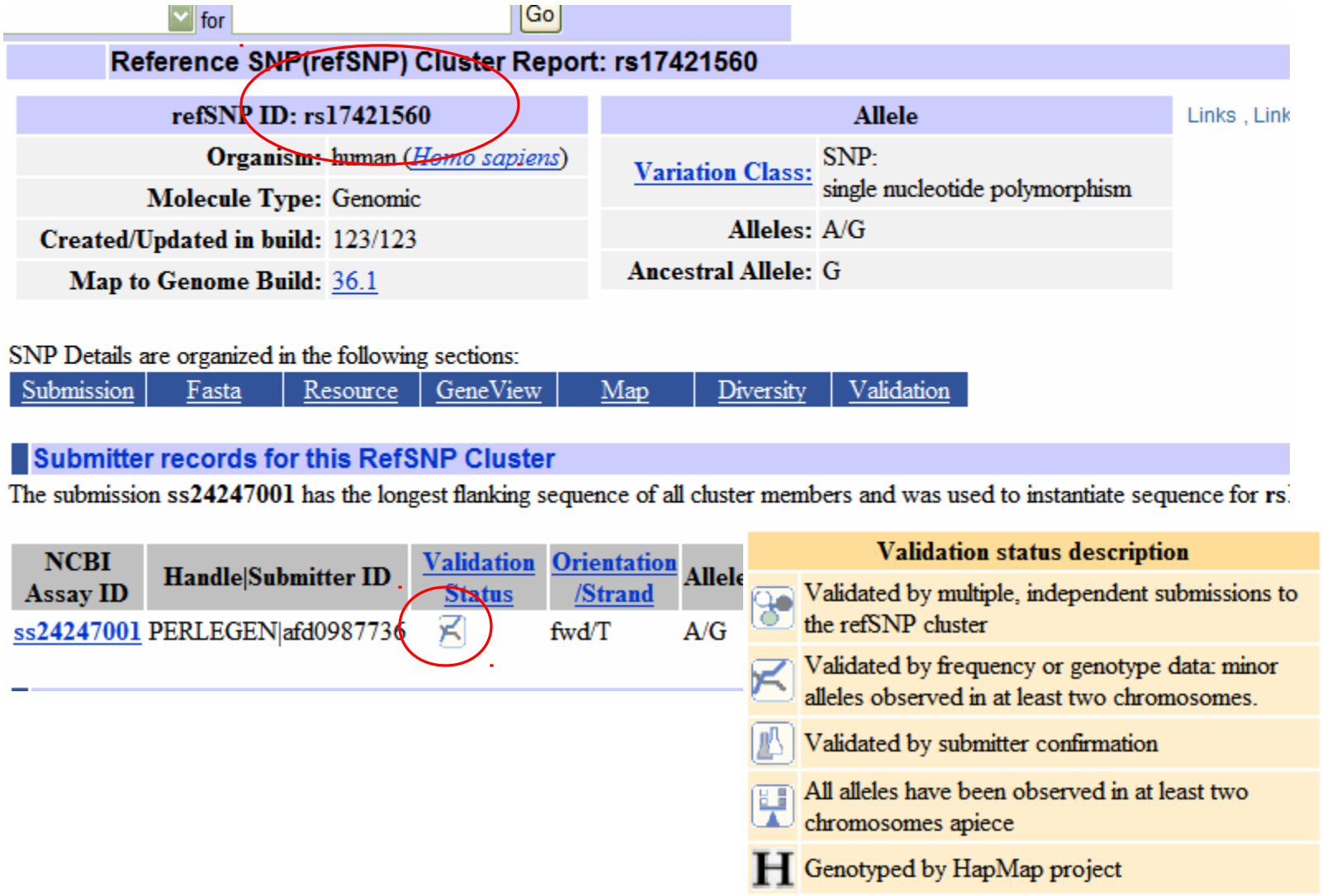
The submission **ss35207915** has the longest flanking sequence of all cluster members and w

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

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



SNP Picking: Informative

Population Diversity

Population Diversity

	Sample Ascertainment				Genotype Detail NEW				Alleles	
ss#	Population	Individual Group	Chrom. Sample Cnt.	Source	A/A	A/G	G/G	HWP	A	G
s24246985	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

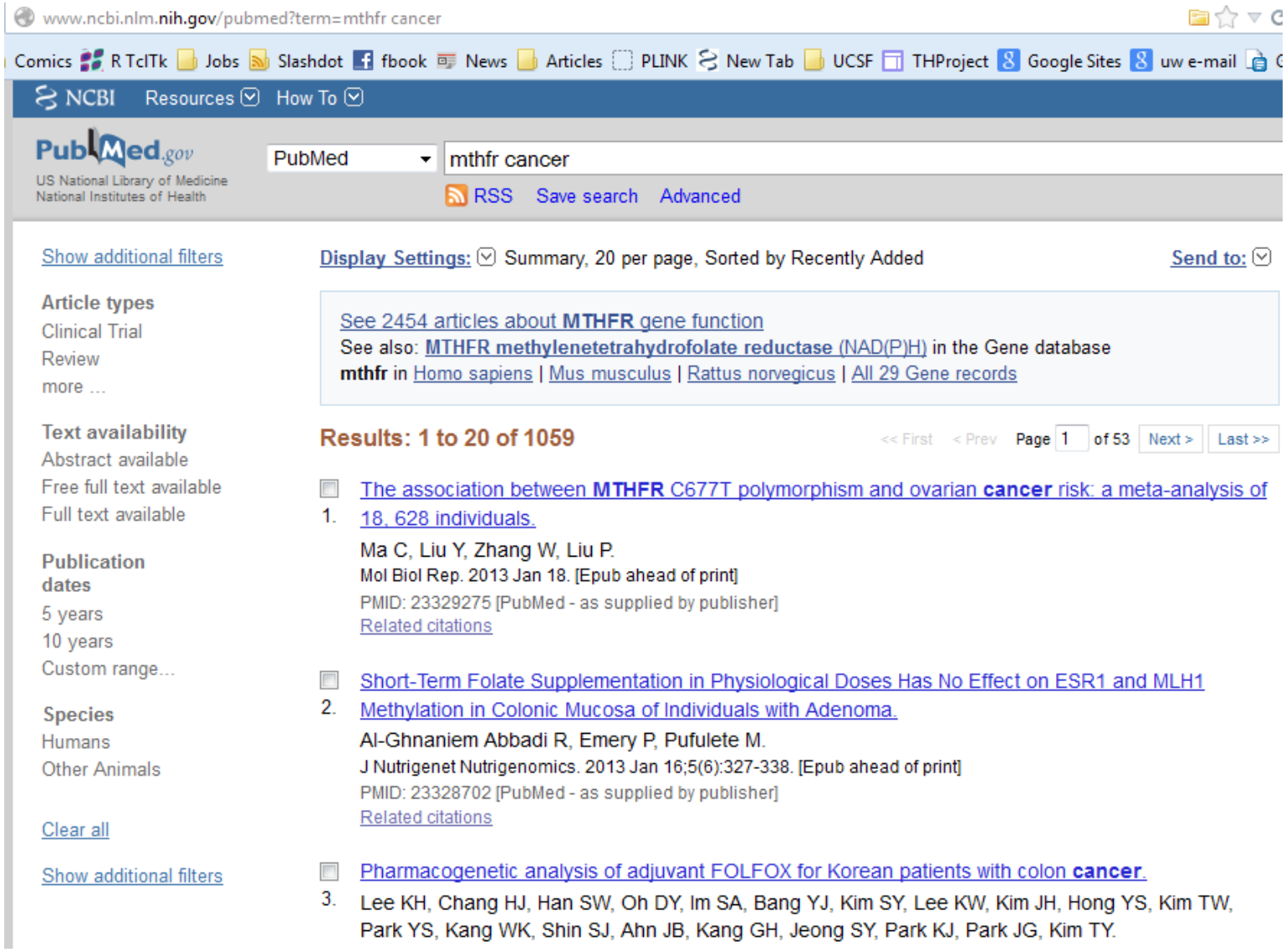
(rs17421511)

SNP Picking: Potentially Functional

Region	Contig position	mRNA pos	dbSNP rs# cluster id	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		Yes		synonymous	A	Thr [T]	3	139
				0.050		Yes		contig reference	G	Thr [T]	3	139
	6398715	396	rs2066461	0.026		Yes		synonymous	A	Thr [T]	3	115

C677T

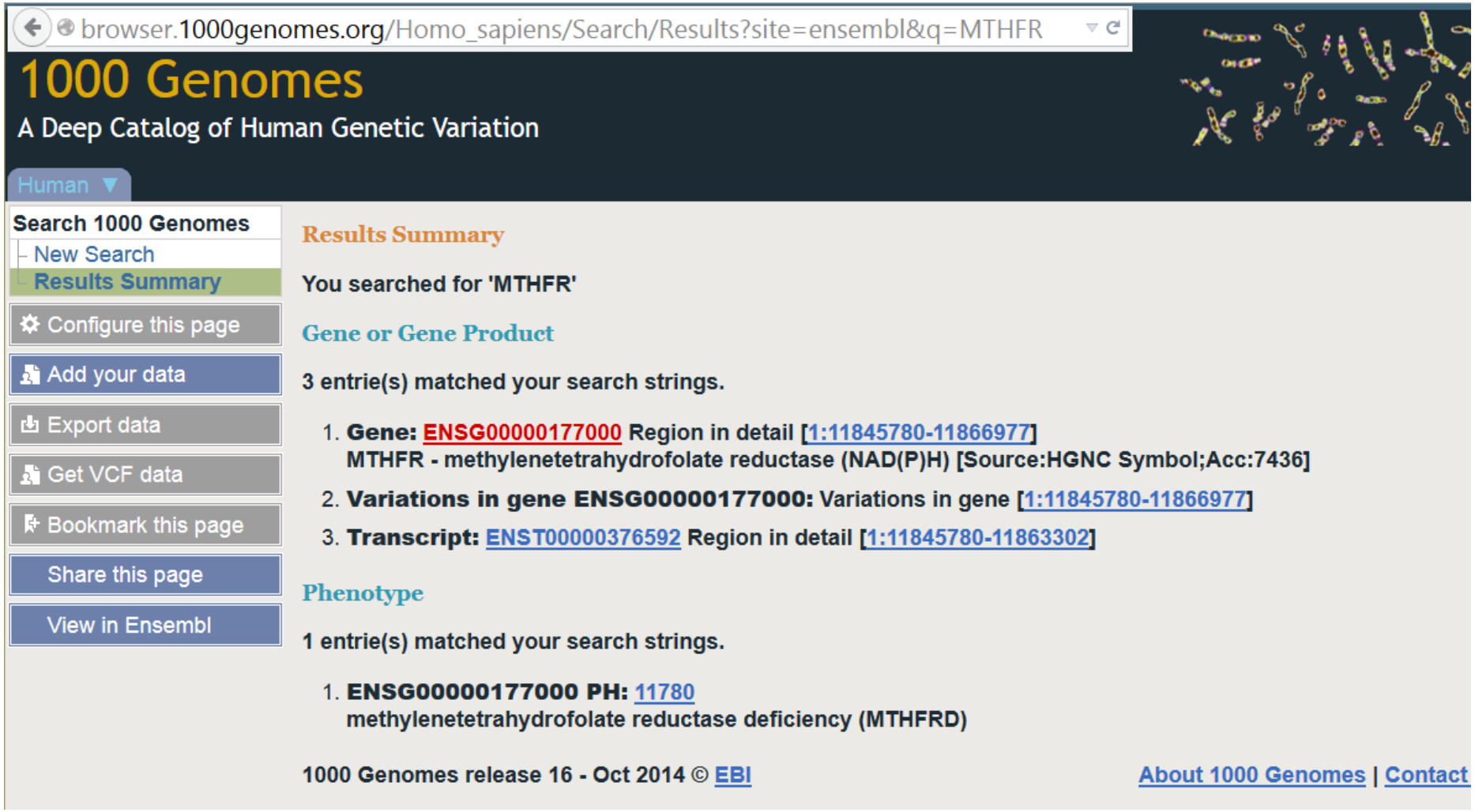
SNP Picking: Previously Associated



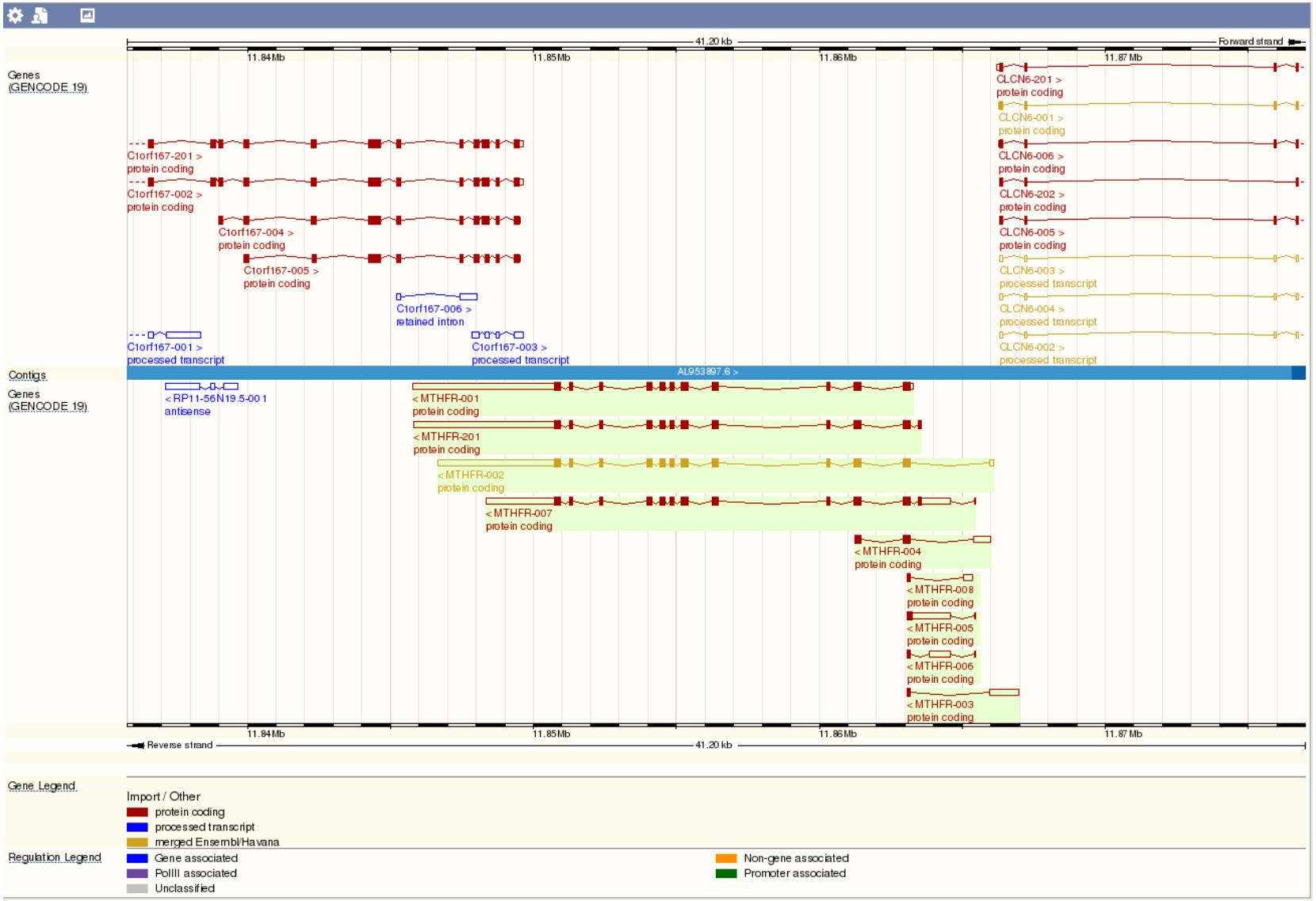
MTHFR Summary

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

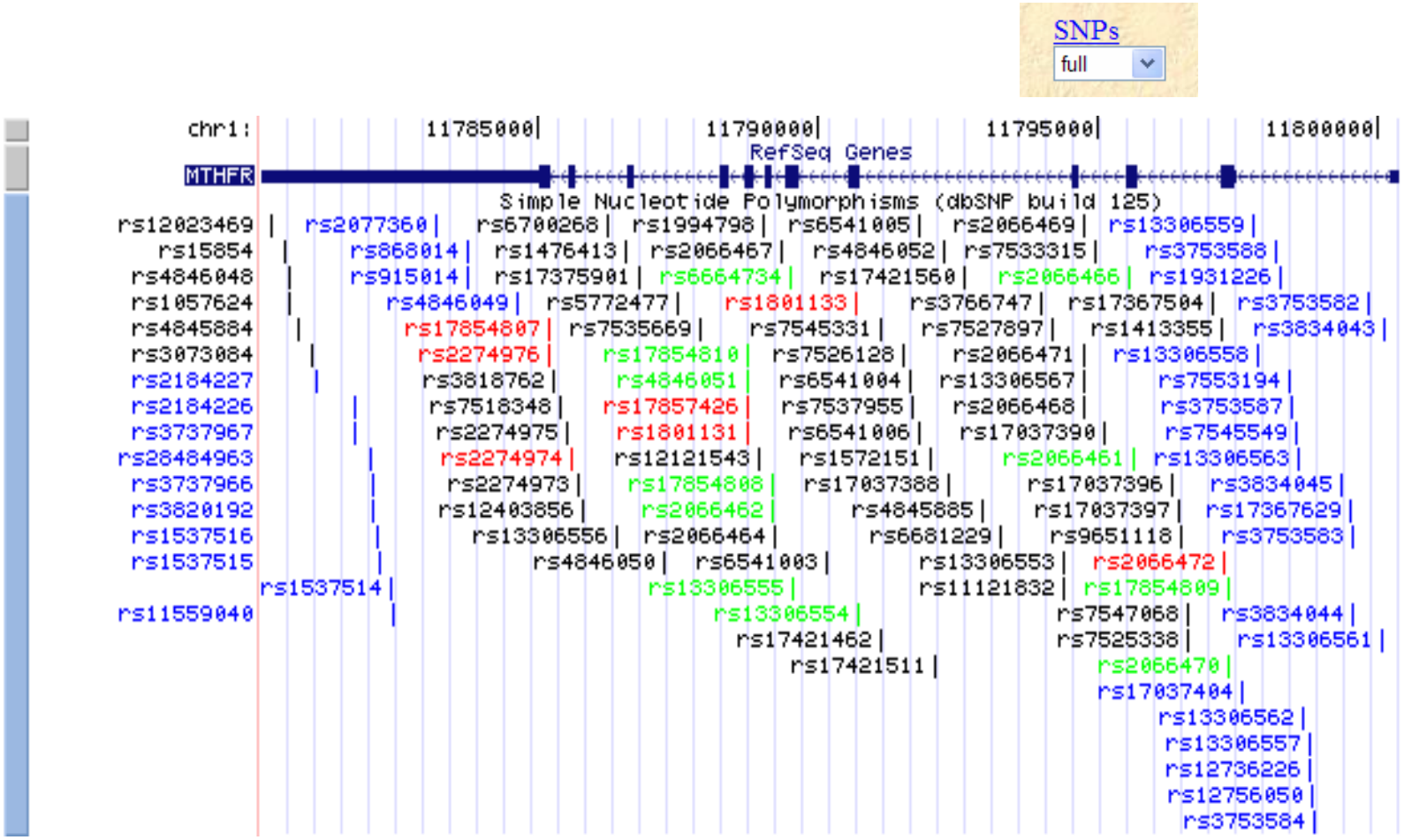
MTHFR Example



MTHFR Example



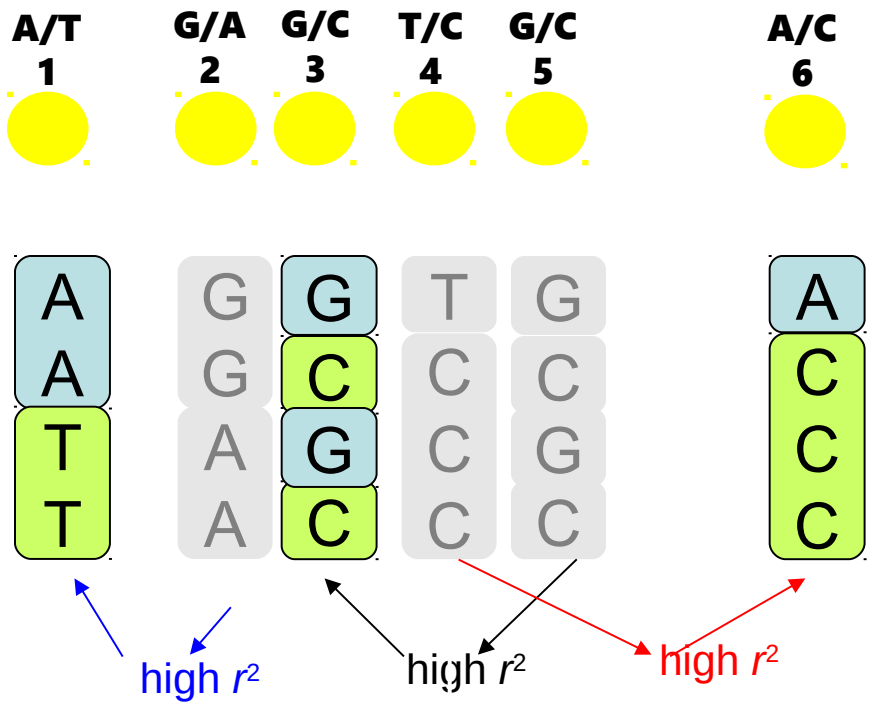
MTHFR SNPs



102 SNPs across MTHFR → Too Many SNPs to Genotype!

Too many MTHFR SNPs
Solution: Tag SNP Selection

- SNPs are correlated (aka Linkage Disequilibrium)



Pairwise Tagging:

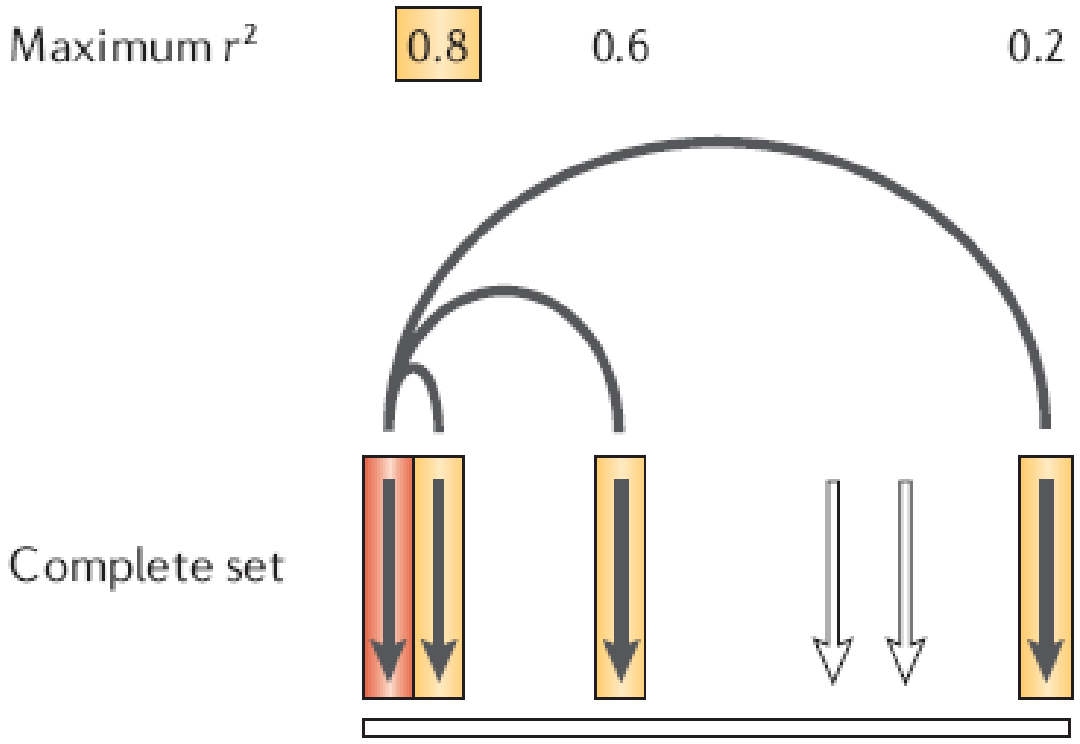
SNP 1
SNP 3
SNP 6

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

SNAP
SNP Annotation and Proxy Search

www.broadinstitute.org/mpg/snap/

BROAD INSTITUTE

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

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 St researchers find genes associated with human disease and response to pharmaceuticals. See "**About the International HapMap Project**" for more informatic

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
HapMart
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 ar

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

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

  This release contains all processed data from the HapMap project, and includes all data from phase genotypes from the Illumina Infinium 100k and 300k genotyping arrays, Affymetrix nsSNPs, and SNP 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). S Depth Celera Reads (She et al. 2004; Bailey et al. 2001 & 2002). Structural Variation datasets includ

CNV regions determined in HapMap samples (Redon 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 c 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 :

mthfr Search

Data Source

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

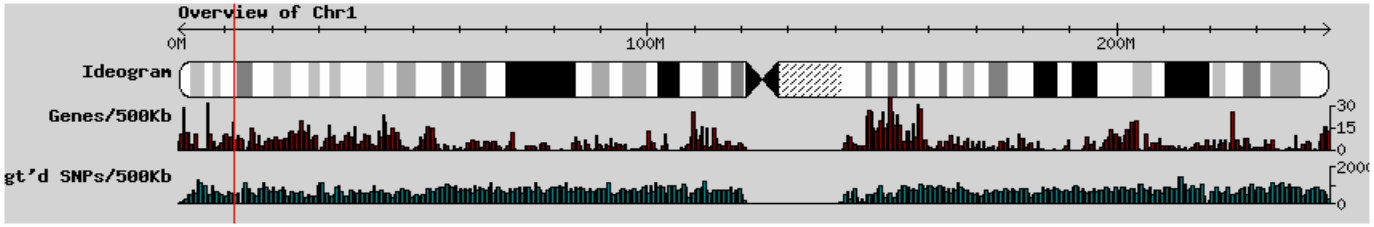
Reports & Analysis :

Download SNP genotype data Configure... Go

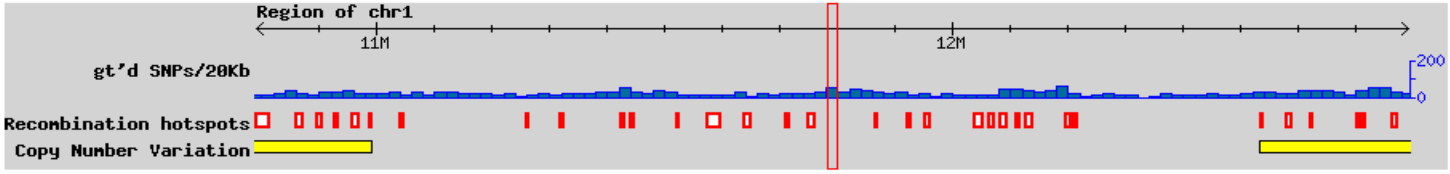
Scroll/Zoom: Show 19.3 kbp Flip

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

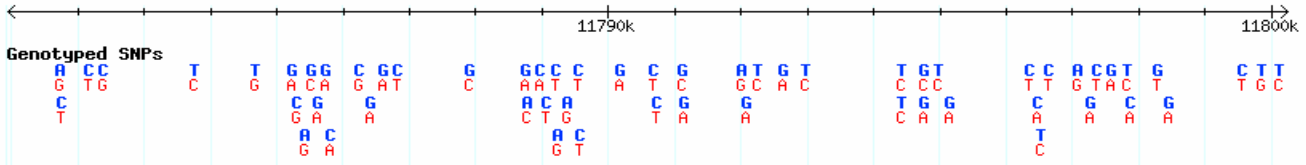
Overview



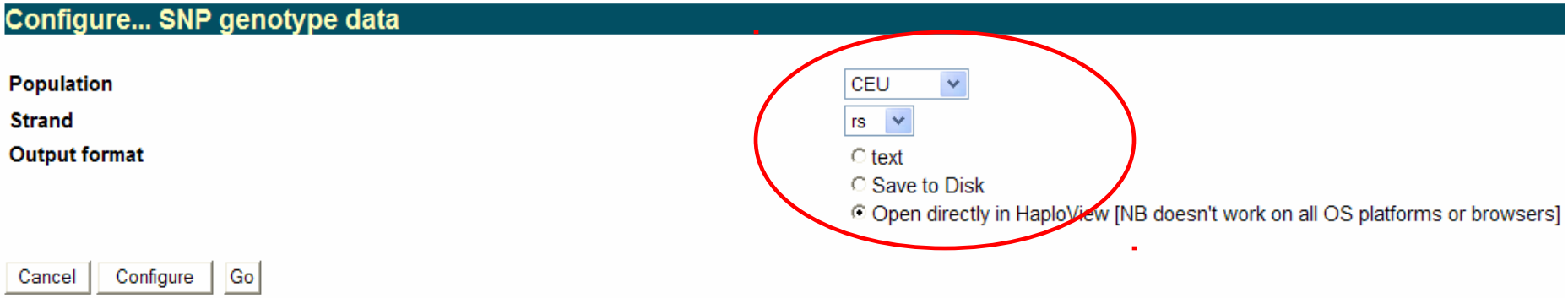
Region

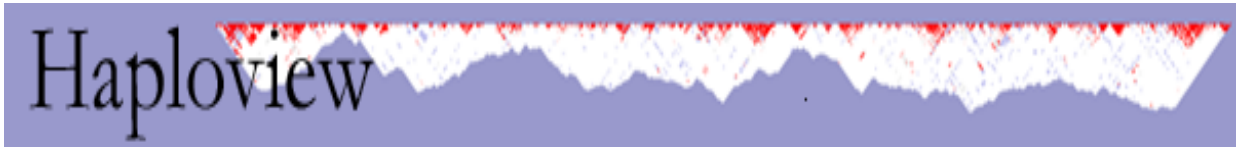


Details



Tag SNPs: HapMap & Haploview

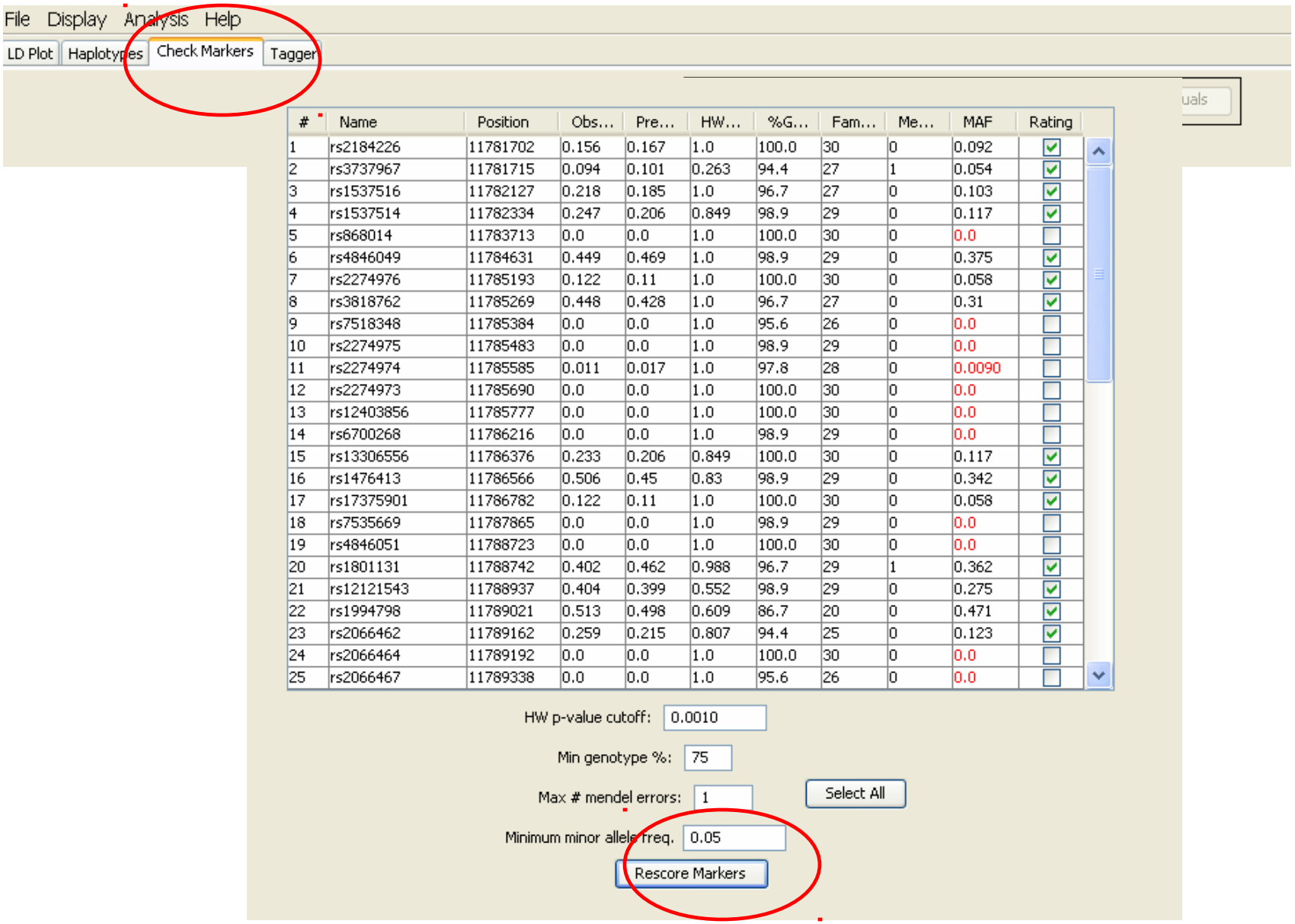




http://www.broad.mit.edu/mpg/haploview/

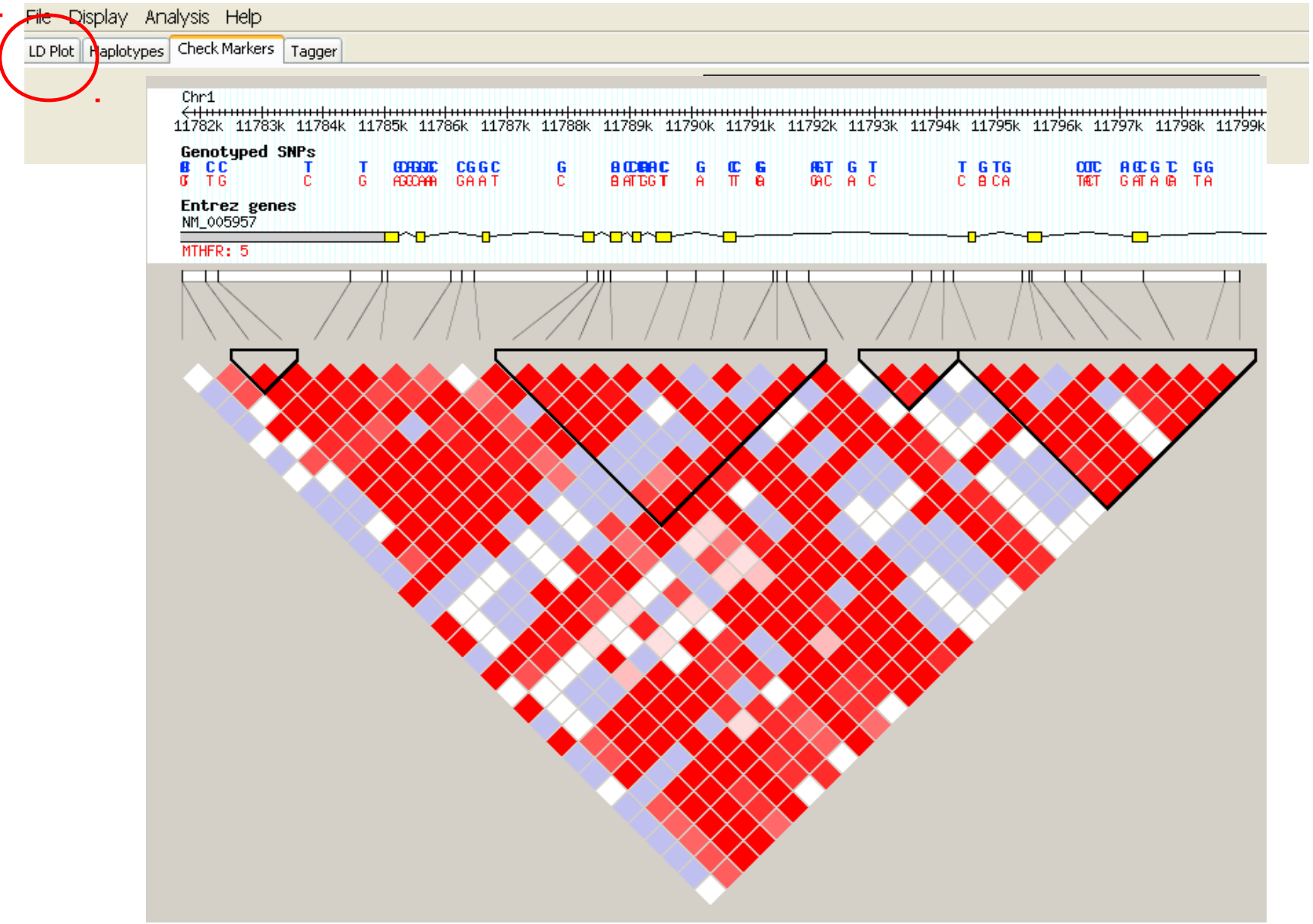
OR Just choose “HapMap Download” directly from Haploview (instead of “HapMap Format”, and choose Chromosome 1, from 11,768Kb-11,789Kb

Tag SNPs: HapMap & Haploview

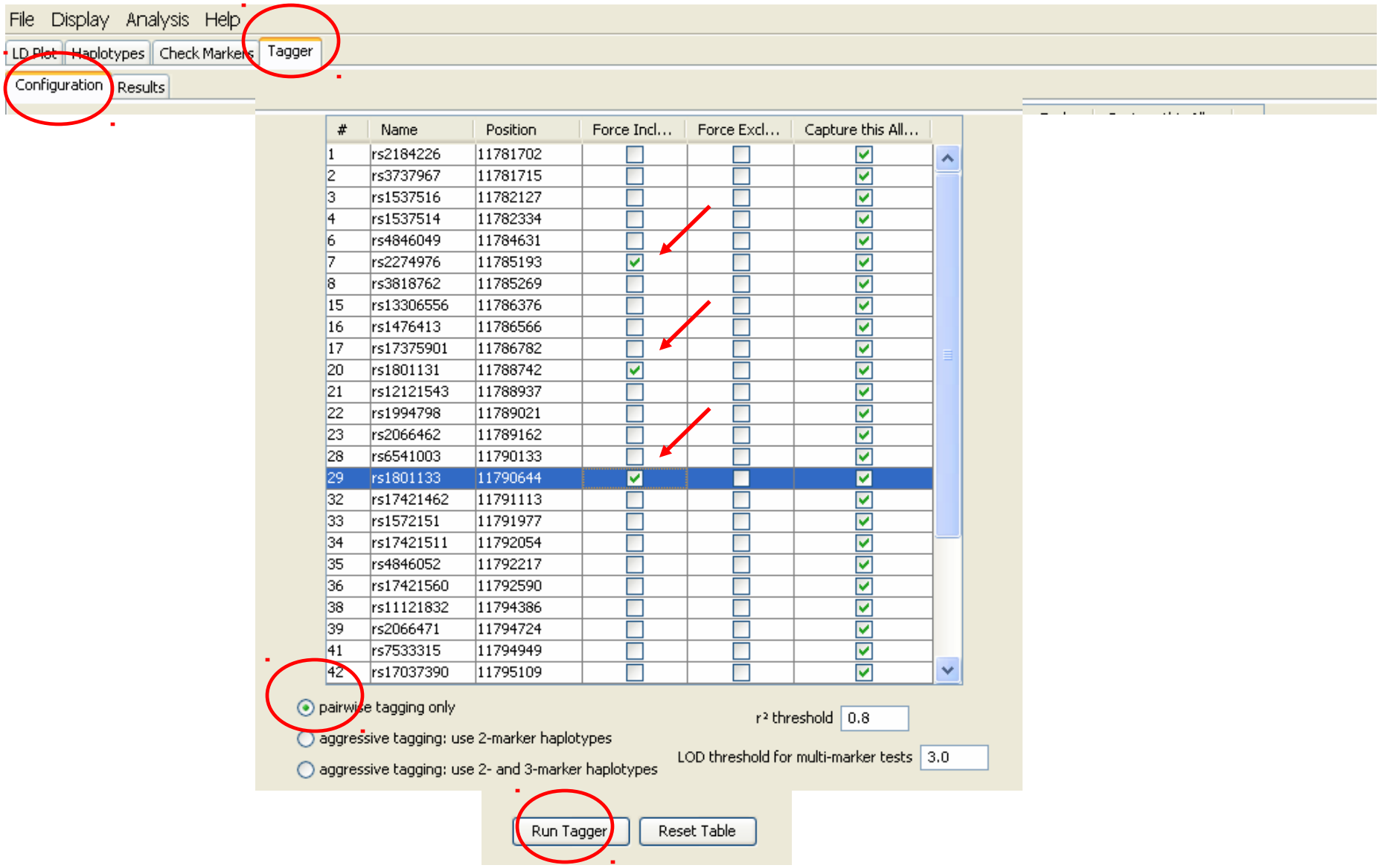


#	Name	Position	Obs...	Pre...	HW...	%G...	Fam...	Me...	MAF	Rating
1	rs2184226	11781702	0.156	0.167	1.0	100.0	30	0	0.092	☑
2	rs3737967	11781715	0.094	0.101	0.263	94.4	27	1	0.054	☑
3	rs1537516	11782127	0.218	0.185	1.0	96.7	27	0	0.103	☑
4	rs1537514	11782334	0.247	0.206	0.849	98.9	29	0	0.117	☑
5	rs868014	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	☐

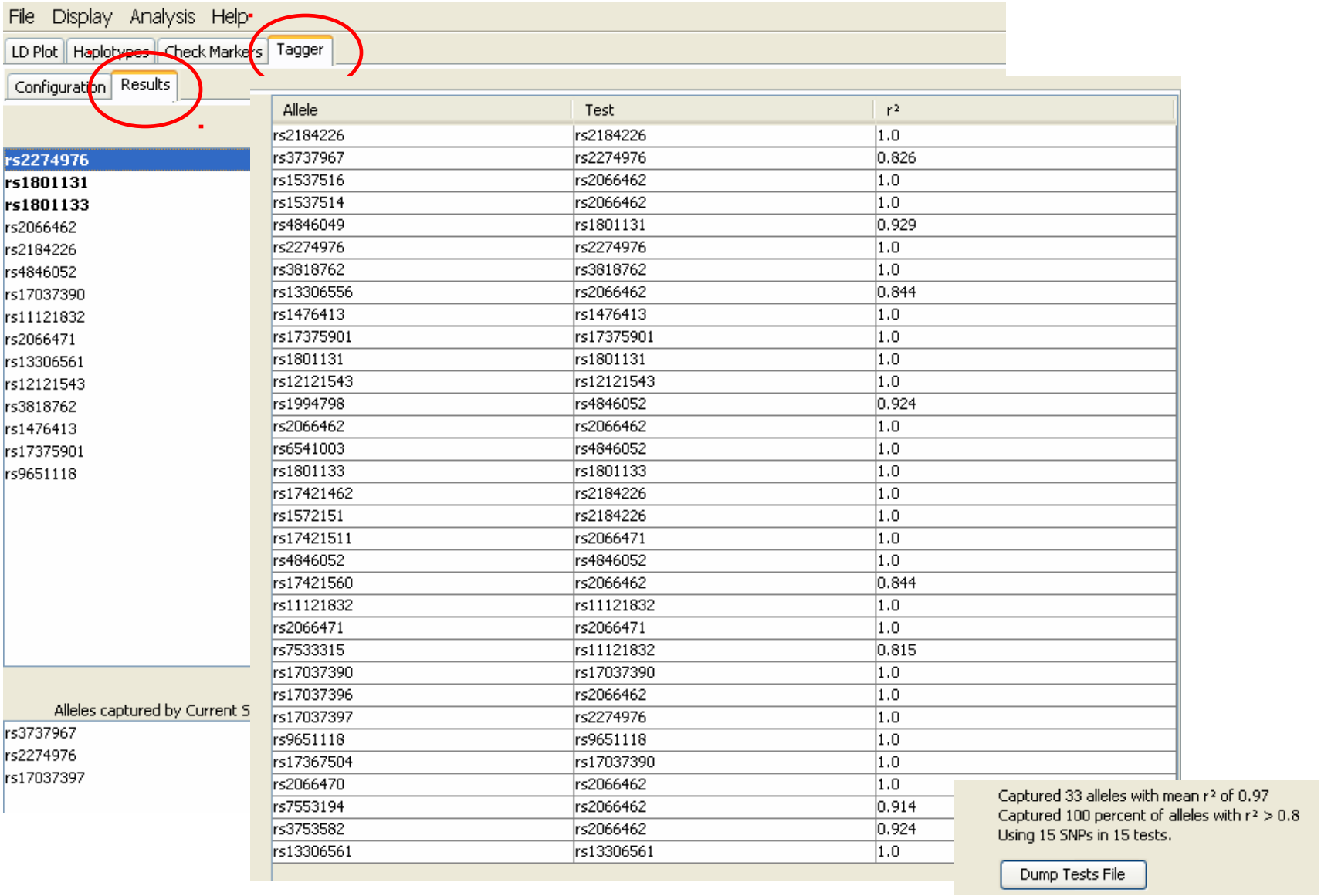
Tag SNPs: HapMap & Haploview



Tag SNPs: HapMap & Haploview



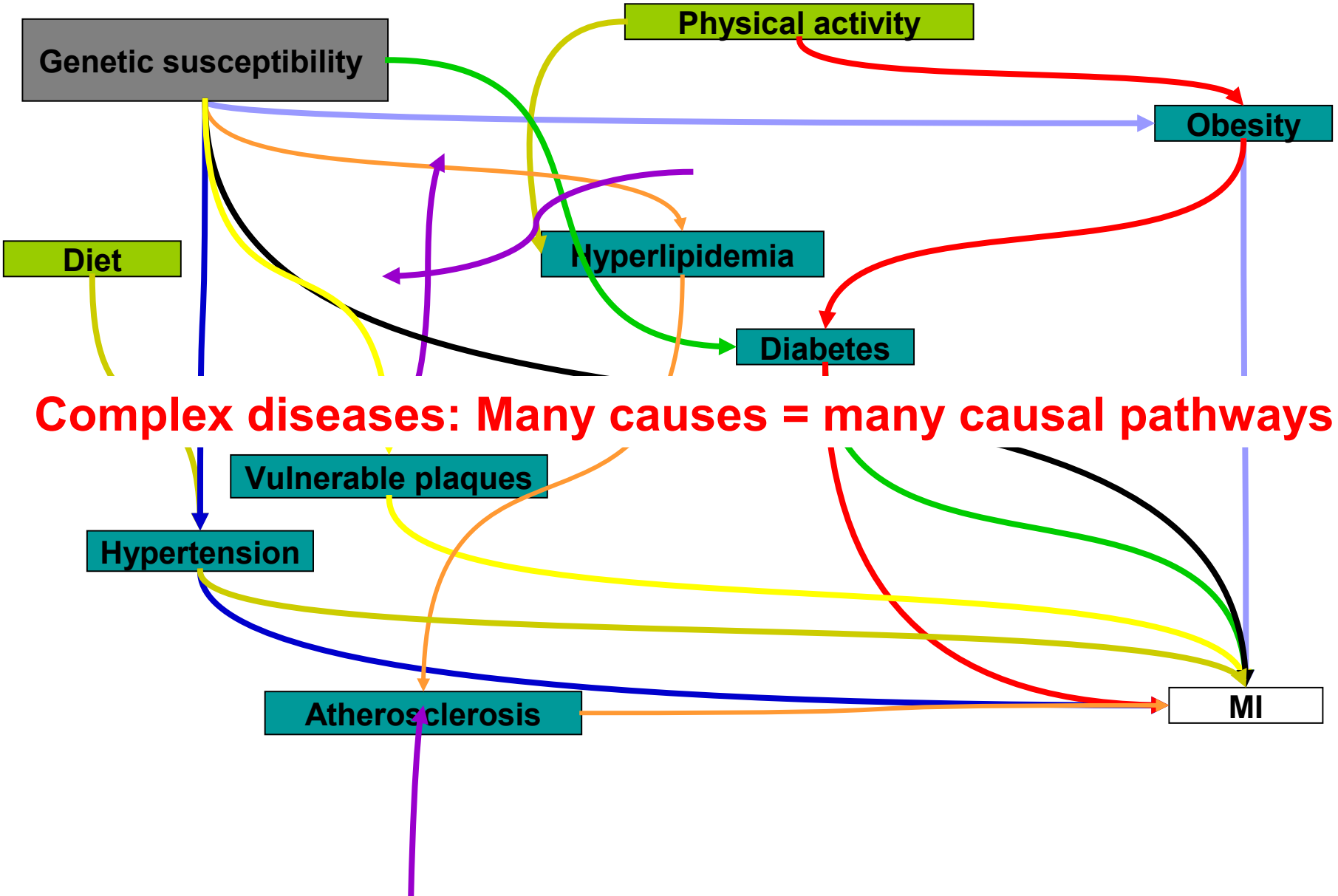
Tag SNPs: HapMap & Haploview



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 r2 = 97%)
- Note: number of SNPs required varies from gene to gene and from population to population

4. Pathways


Genetic susceptibility
Physical activity
Obesity
Diet
Hyperlipidemia
Diabetes
Complex diseases: Many causes = many causal pathways!
Vulnerable plaques
Hypertension
Atherosclerosis
MI

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 → risk?
- Score model:

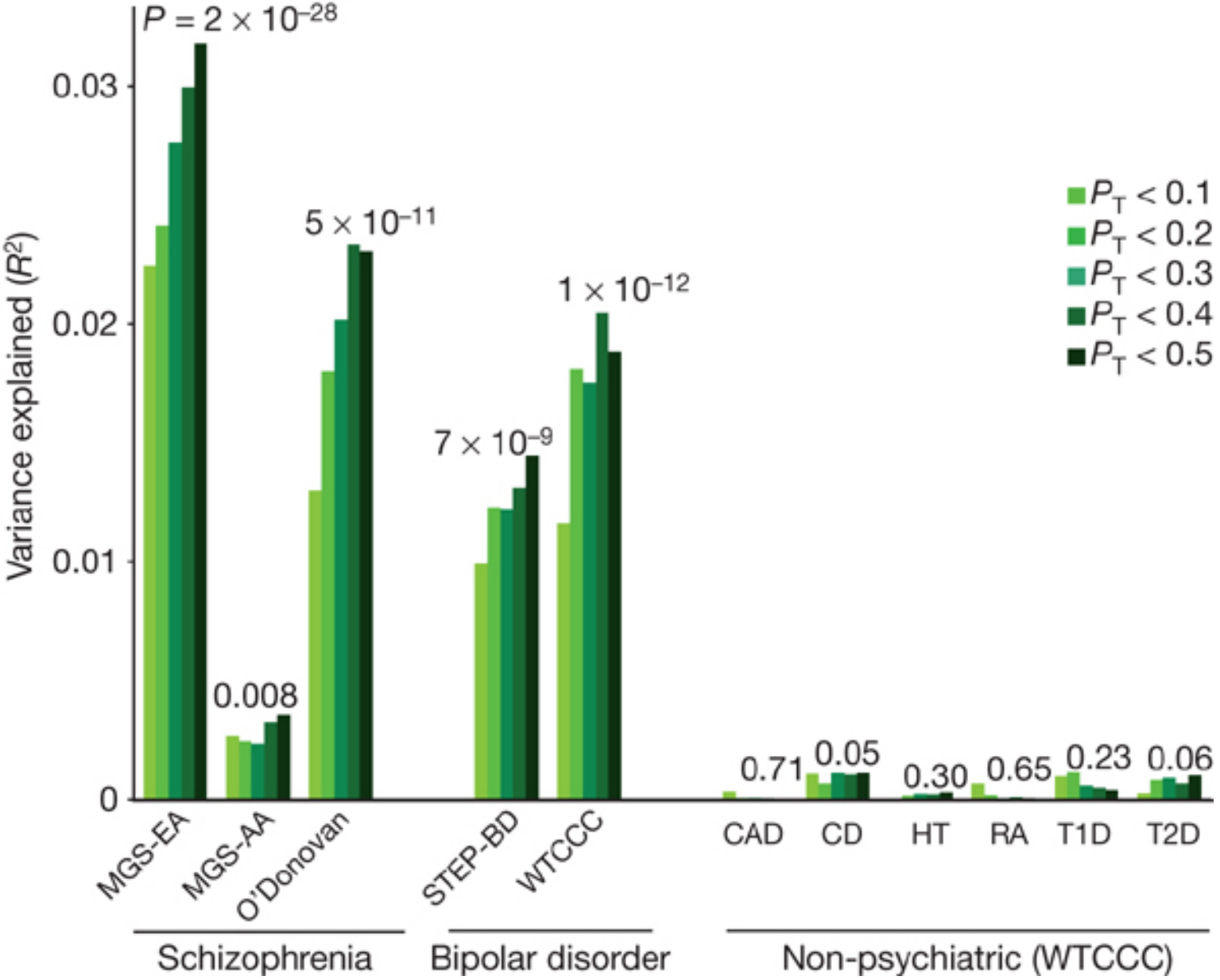
$$x_j = \frac{\sum_{i=1}^{m} \ln(OR_i) \times SNP_{ij}}{m}$$

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?

ISC / Purcell et al. *Nature* 2009

Application of Model



Purcell / ISC et al. *Nature* 2009