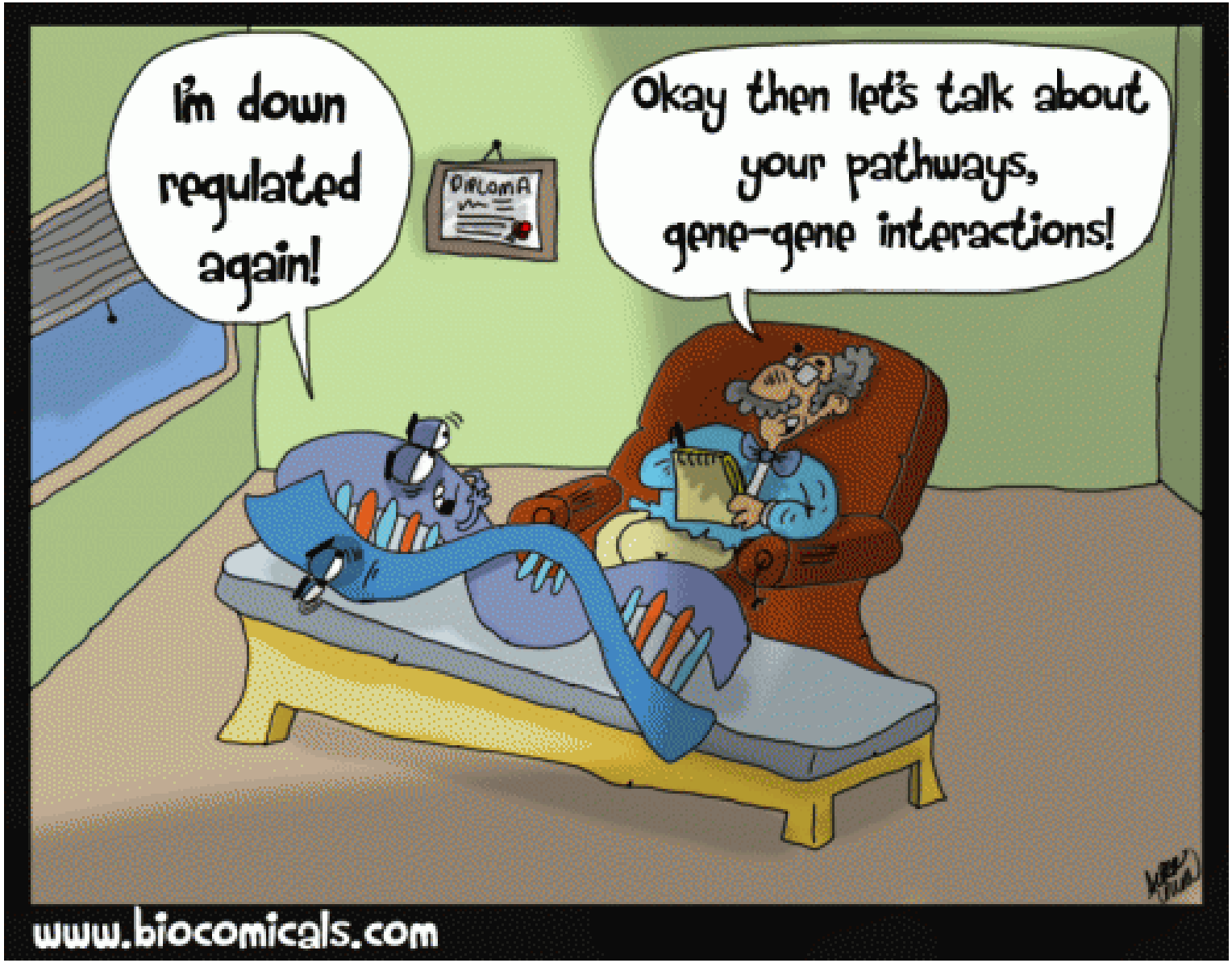
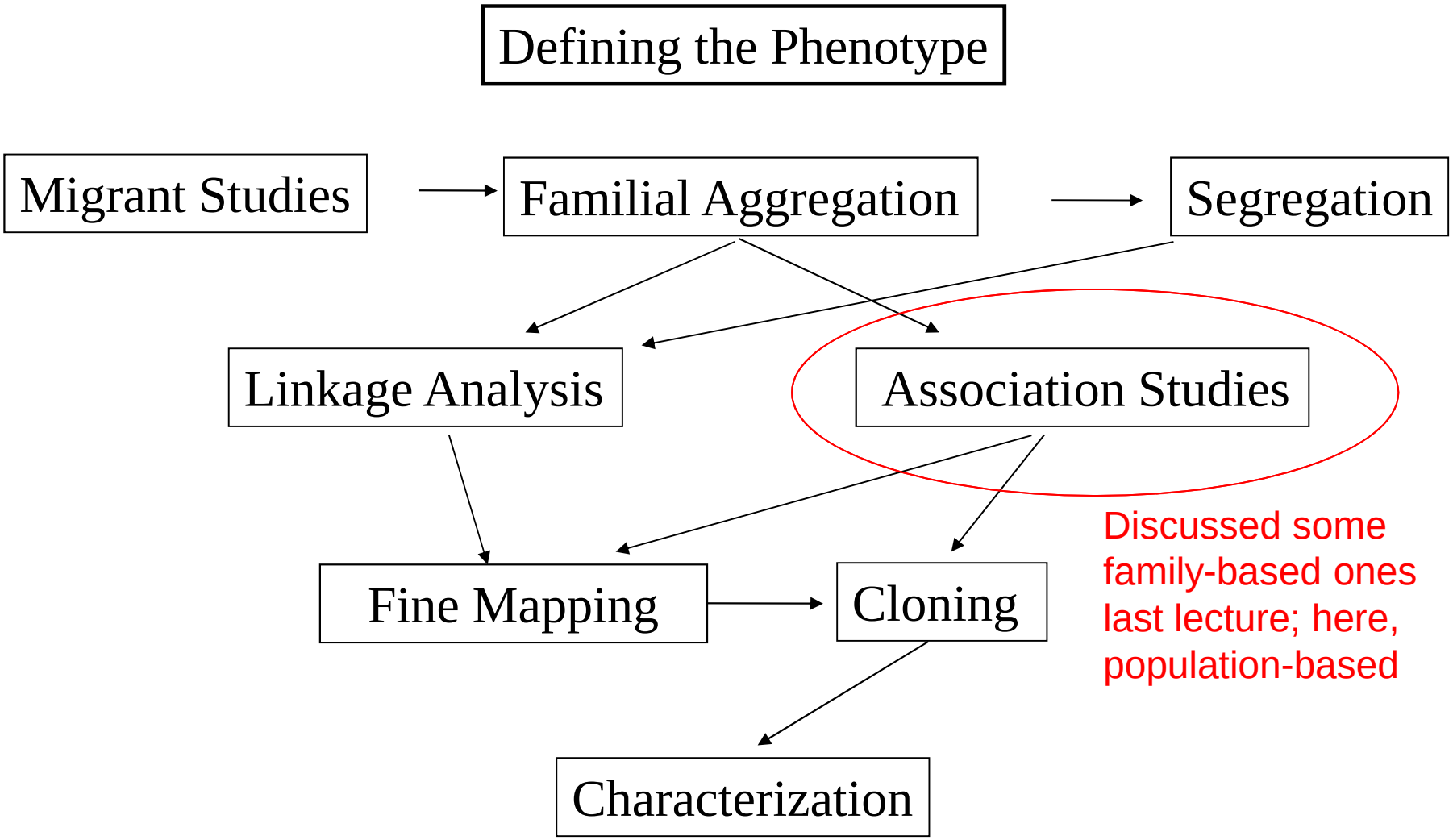
Association Studies



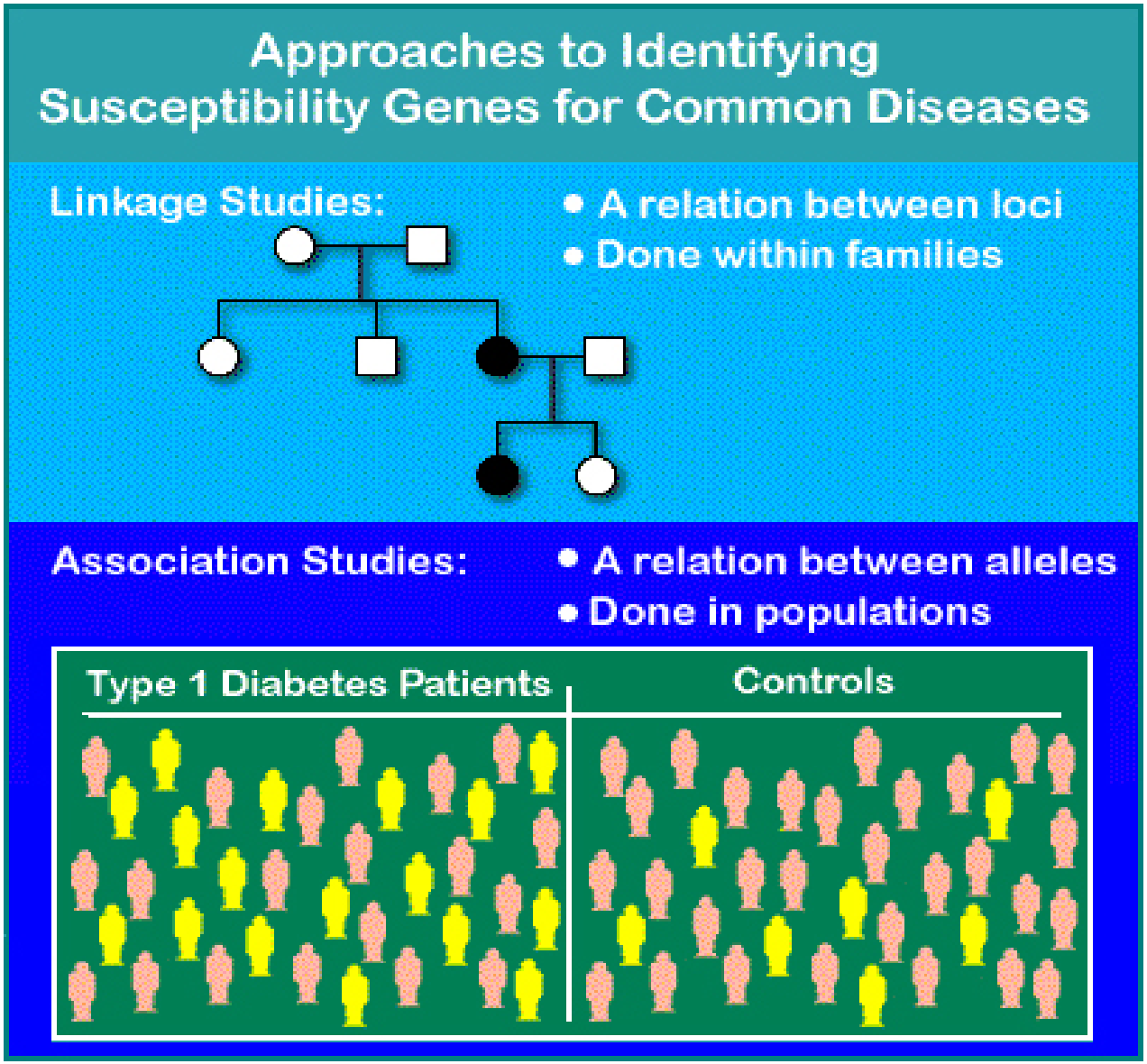
Overview

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

Process of Genetic Epidemiology

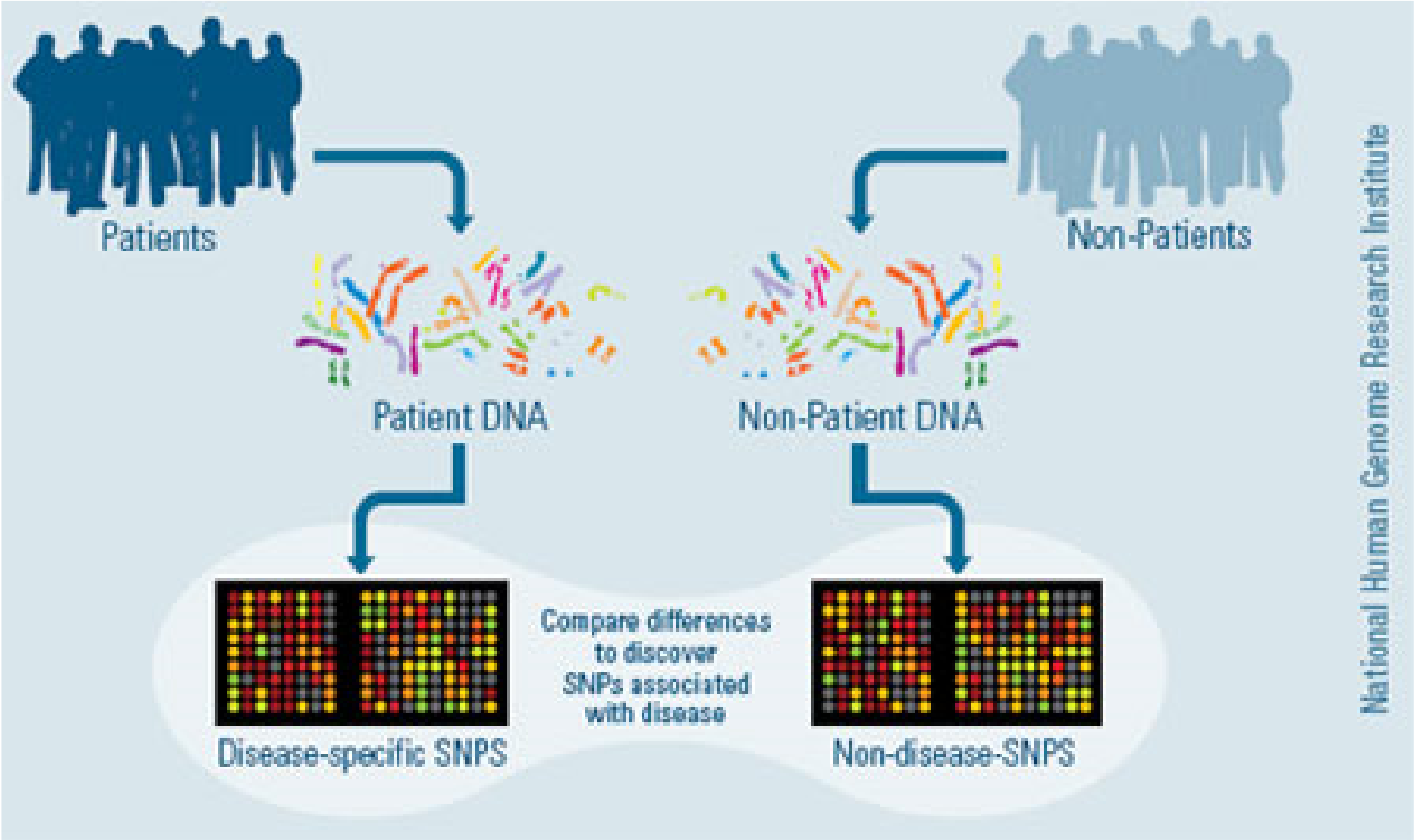


Association Studies



ROCHE Genetic Education (www)

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			
					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.

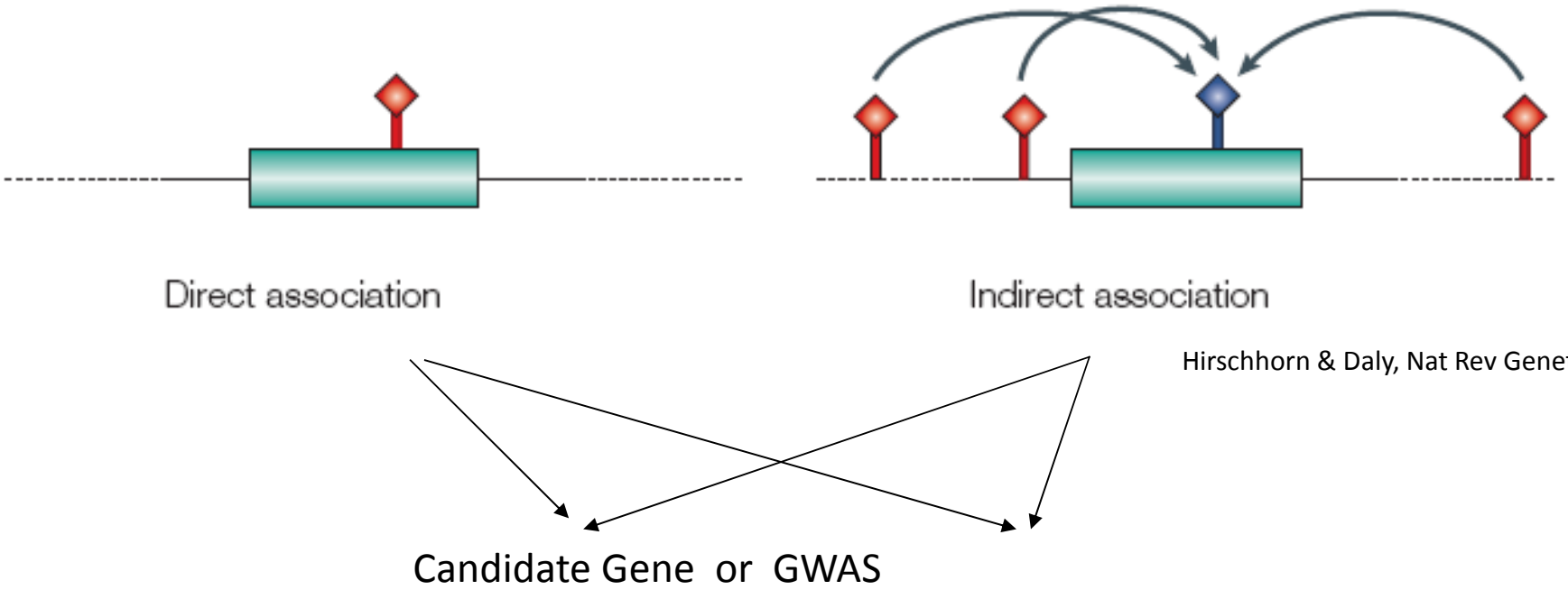
Risch & Merikangas, Science 1996

Association Studies – Educational Group Exercise!

- Come up with important aspects and determine the design AND how to analyse your genotype association study…
  - What aspects might influence where you draw your cases/controls?
  - How would you choose where in the genome?
    - Do you need all of that? Is there a way you could genotype less?
  - How might you conduct a test of if the allele frequencies differ in cases and controls at a specific SNP?

Not about getting things perfectly right, the exercise is to get you to think about the problem, as studies show that you will remember things better that way.

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

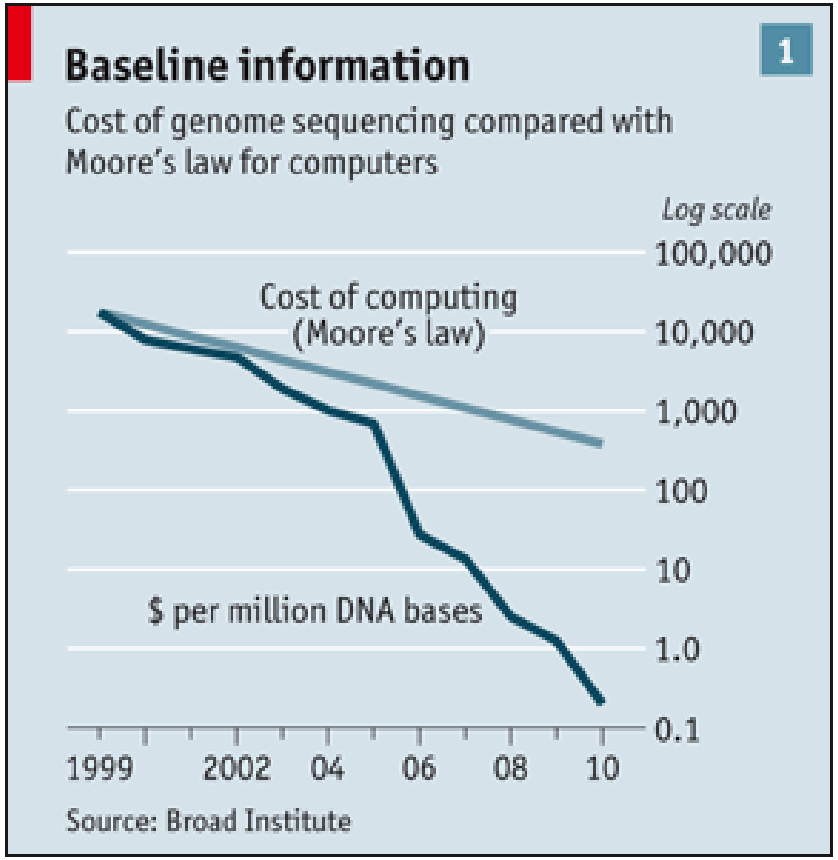
> 500 times faster

< one millionth the cost!

Still not quite there:

1 hour, $1K

(#1 Innovation, 2010)



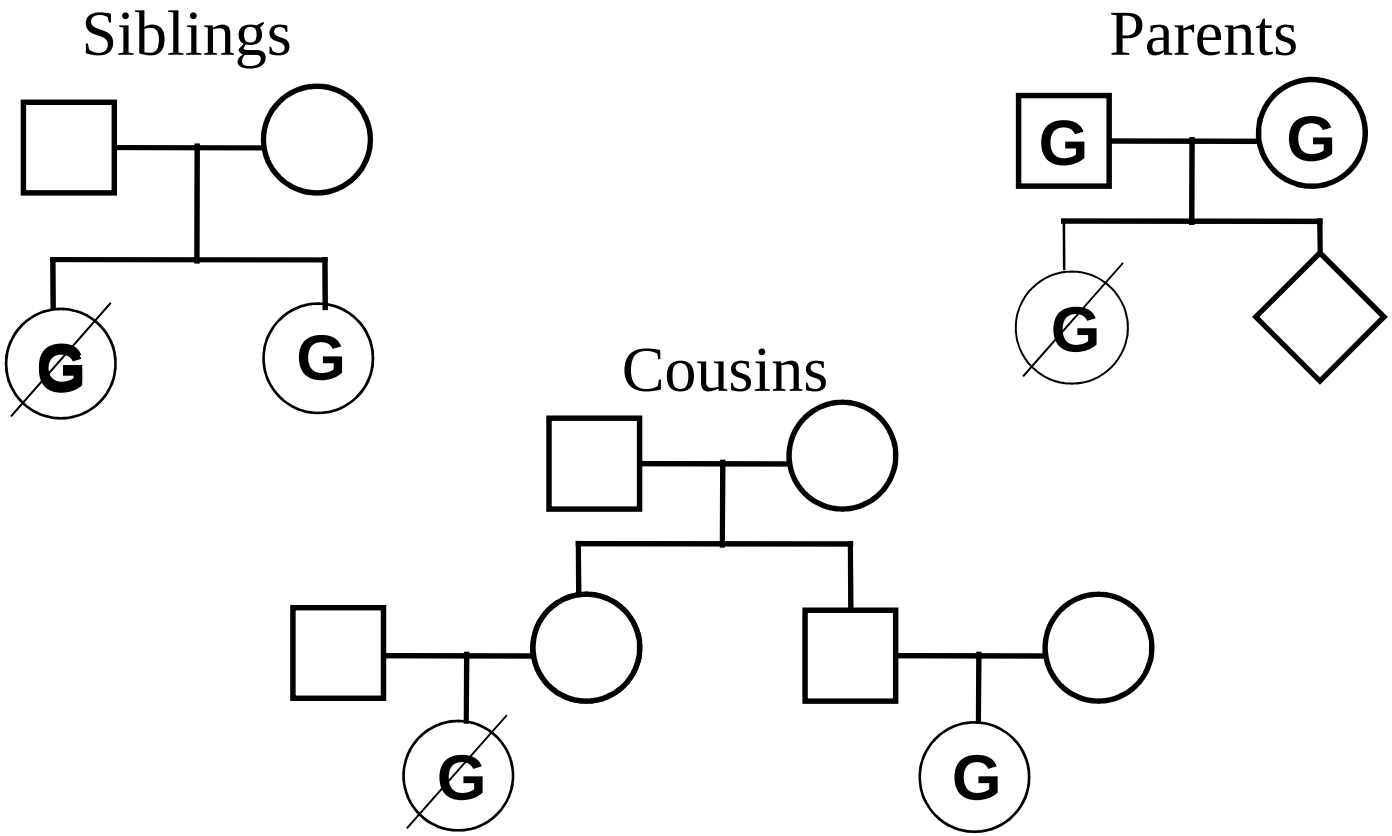
The Economist, 2010

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.

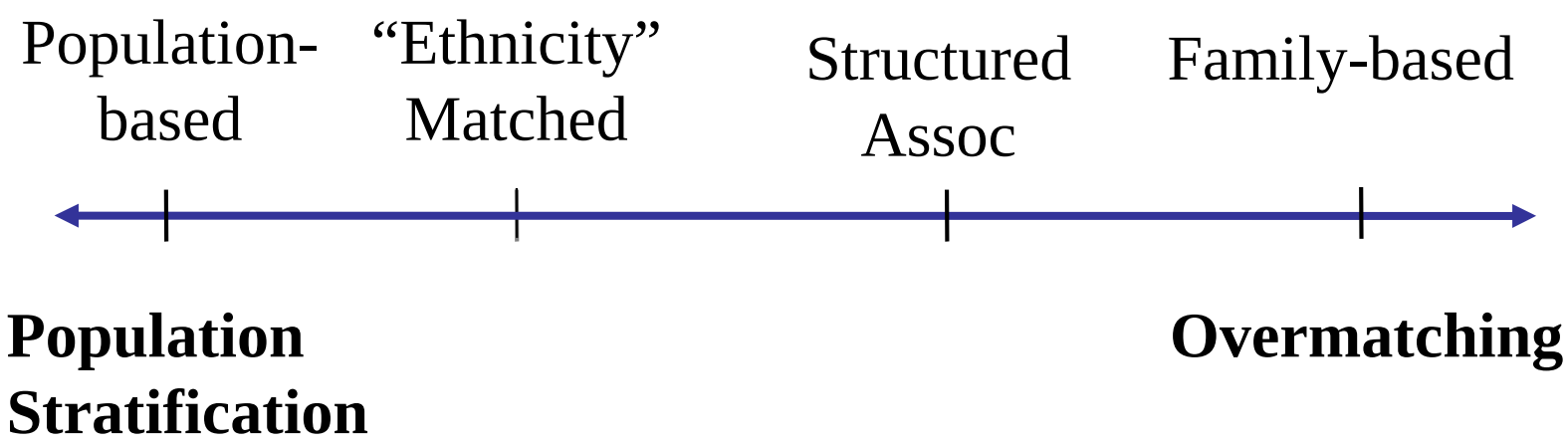
Family-Based Association Studies

(Covered in last lecture)

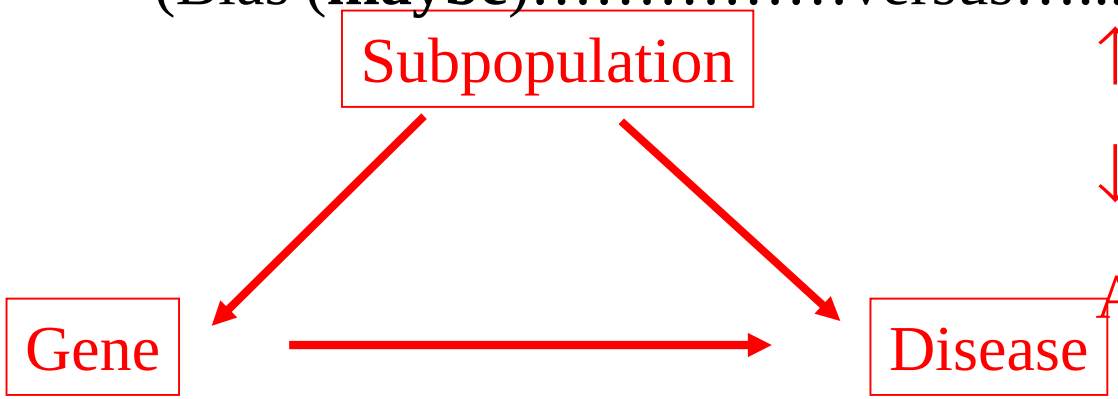


Termed "**pseudo**"-controls (the hypothetical other mendelian transmissions)

Continuum of Assoc Study Designs



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



↑ Sharing of genes & envt.

↓ Efficiency

Also, recruitment issues

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-squared 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 =nCase	D+E+F =nCont	A+B+C+D+E+F =n

Expected

Case	Control
nTT*nCase/n	nTT*nCont/n
nCT*nCase/n	nCT*nCont/n
nCC*nCase/n	nCC*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-squared test example

```
Observed:                        Expected
Geno Case  Control  Total    Case            Control          OR (to TT genotype)
  TT    5       15  20       20*35/65=10.8   20*30/65=9.2     1
  CT   10       10  20       20*35/65=10.8   20*30/65=9.2     (10x15)/(5x10) = 3
  CC   20        5  25       25*35/65=13.5   25*30/65=11.5    (20x15)/(5x5) = 12
Total  35       30  65
      =nCase =nCont  =n
Sum (Observed – Expected)^2/Expected
    = (5-10.8)^2/10.8 + (10-10.8)^2/10.8 + (20-13.5)^2/13.5
       + (15-9.2)^2/9.2 + (10-9.2)^2/9.2 + (5-11.5)^2/11.5
    = 13.7
P-value = 0.0011
```

Genetic Model

Genotype	OR
TT	1
CT	r
CC	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)** / 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

Genotype	Case	Control
GG	A	D
GT	B	E
TT	C	F

~chi_sq(2df)

Recessive (G)

	Case	Control
GG	A	D
GT or TT	B+C	E+F

~chi_sq(1df)

Dominant (G)

	Case	Control
GG or GT	A+B	D+E
TT	C	F

~chi_sq(1df)

Genotype	Case	Control
TT	5	15
GG	20	5
GT	10	10

P=0.0011

	Case	Control
TT	5	15
GG or GT	30	15

P=0.0045

	Case	Control
TT or TG	15	25
GG	20	5

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_loga |    case
  dditive   |    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
                                                  LR chi2(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

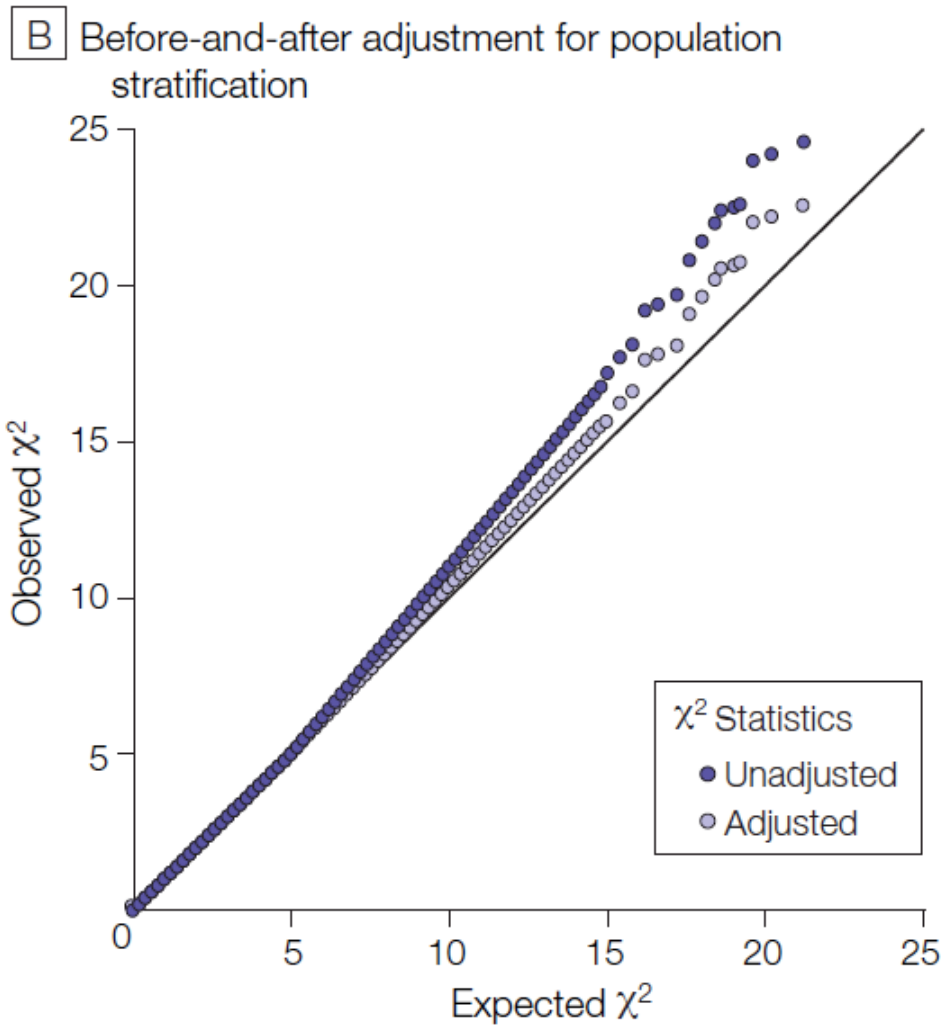
Population stratification is a form of confounding in genetic studies where a gene under study shows marked variation in allele frequency across subgroups of a population and these subgroups differ in their baseline risk of disease

(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 ($\chi^2_{new} = \chi^2_{old}/\lambda$)
  - 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!
- 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]

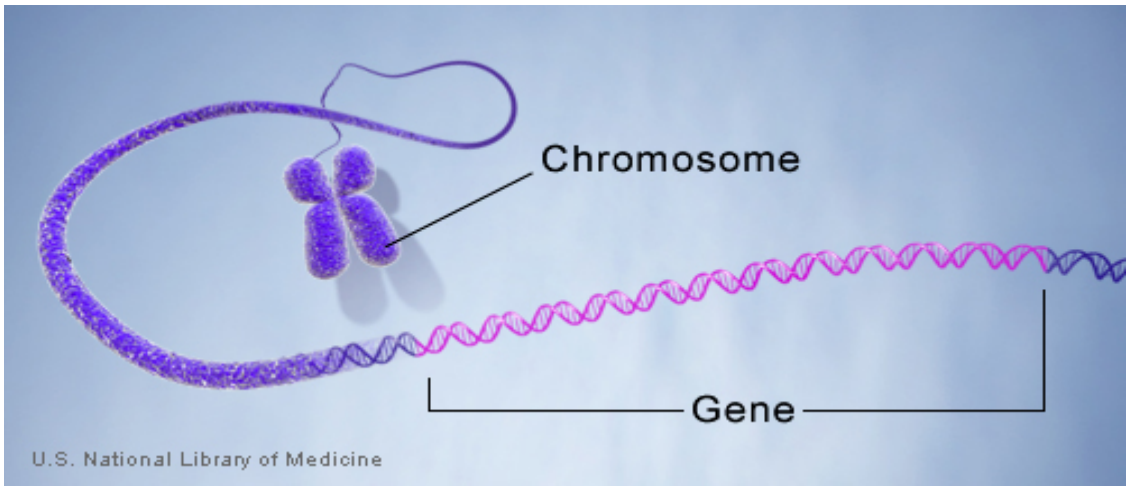
Quantile-Quantile (QQ) Plots



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?"



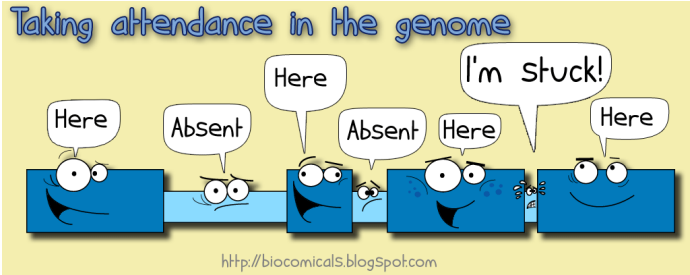
Candidate Gene: Where do I Start?

- 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?

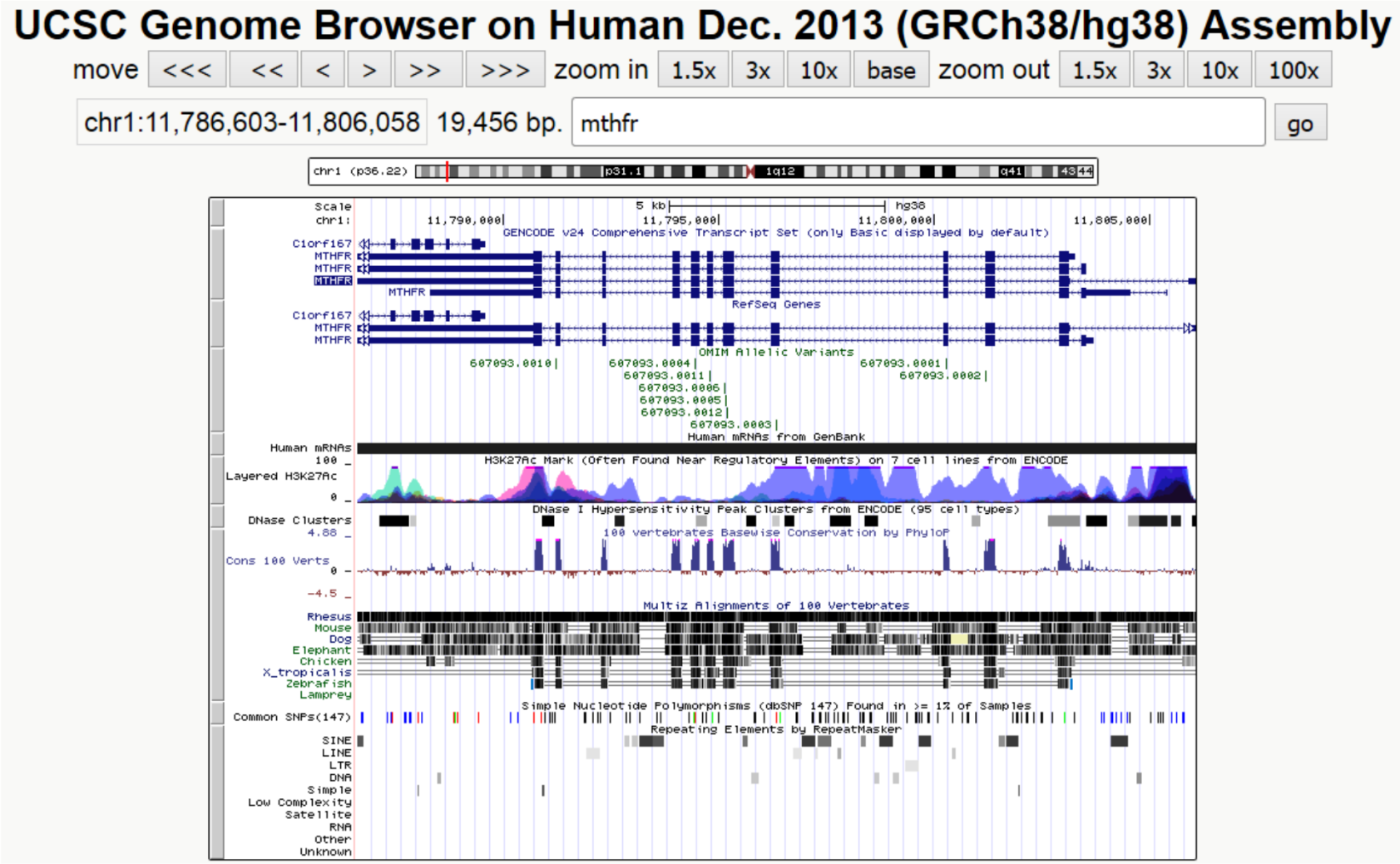
- Potentially Functional: Do these SNPs have a potential biological impact? Missense variants?

- Previously Associated: Have previous studies found SNPs in the candidate gene associated with the outcome?

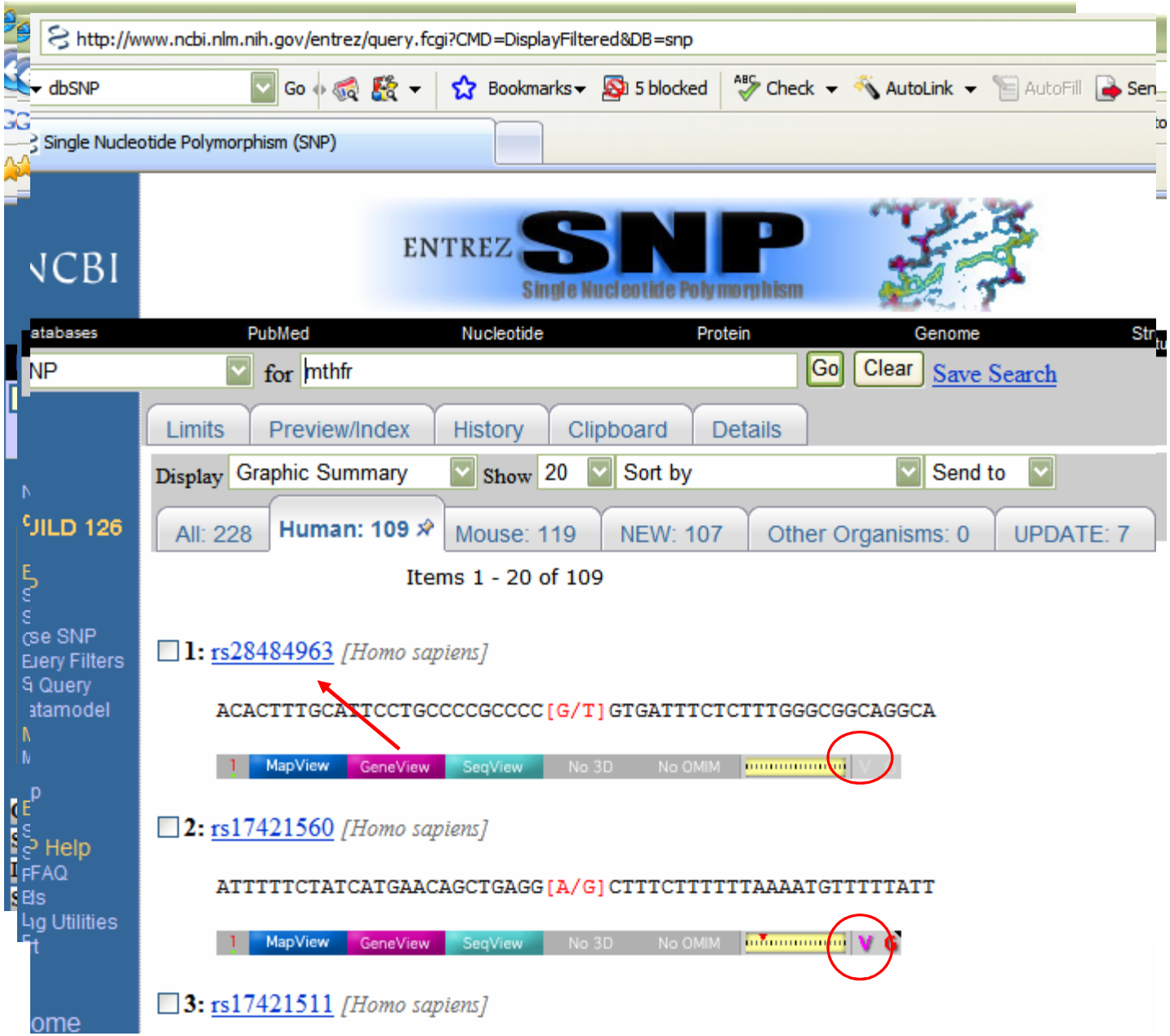
Learning activity! Try these websites with these terms

- Links:
  - UCSC Genome browser / 1000 Genomes Browser
  - dbSNP
  - Pubmed
- Example terms: mthfr (a gene), mthfr cancer
- Things to find out: Chromosome? # Exons? # Introns? Size of gene? Functional variants? Previously associated variants? SNP validation status?

UCSC Genome Browser



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

for Go

Reference SNP(refSNP) Cluster Report: rs17421560

refSNP ID: rs17421560	
Organism:	human (*Homo sapiens*)
Molecule Type:	Genomic
Created/Updated in build:	123/123
Map to Genome Build:	36.1

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

Links , Link

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	Validation Status	Orientation /Strand	Allele
ss24247001	PERLEGEN|afd0987736		fwd/T	A/G

Validation status description	
	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
H	Genotyped by HapMap project

SNP Picking: Informative

Population Diversity

Population Diversity

	Sample Ascertainment					Genotype Detail NEW				Alleles	
ss2	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

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 >>

1. 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

2. Short-Term Folate Supplementation in Physiological Doses Has No Effect on ESR1 and MLH1 Methylation in Colonic Mucosa of Individuals with Adenoma.
   Al-Ghnaniem 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

3. 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.

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

Phenotype

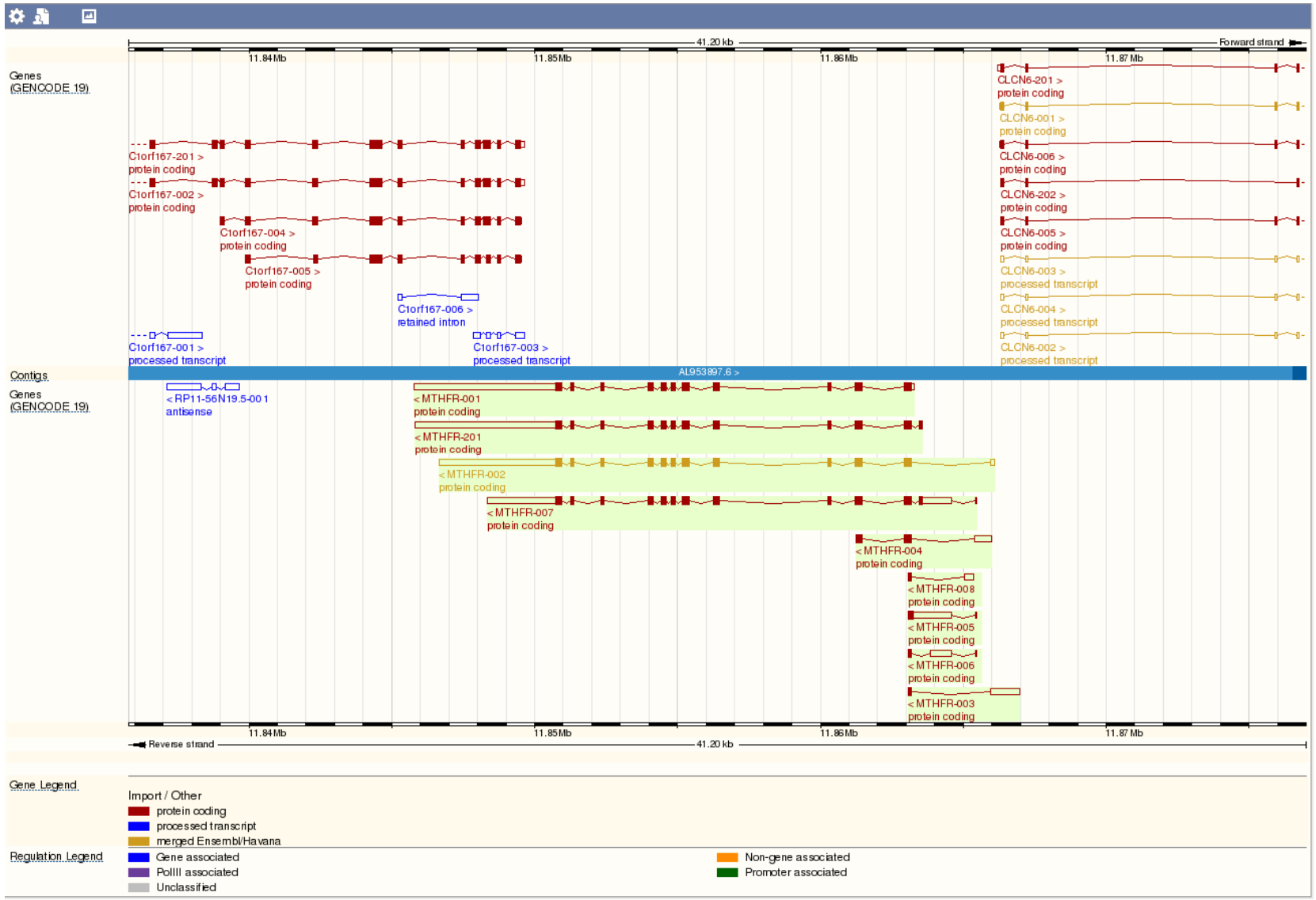
1 entrie(s) matched your search strings.

1. **ENSG00000177000 PH:** 11780
   methylenetetrahydrofolate reductase deficiency (MTHFRD)

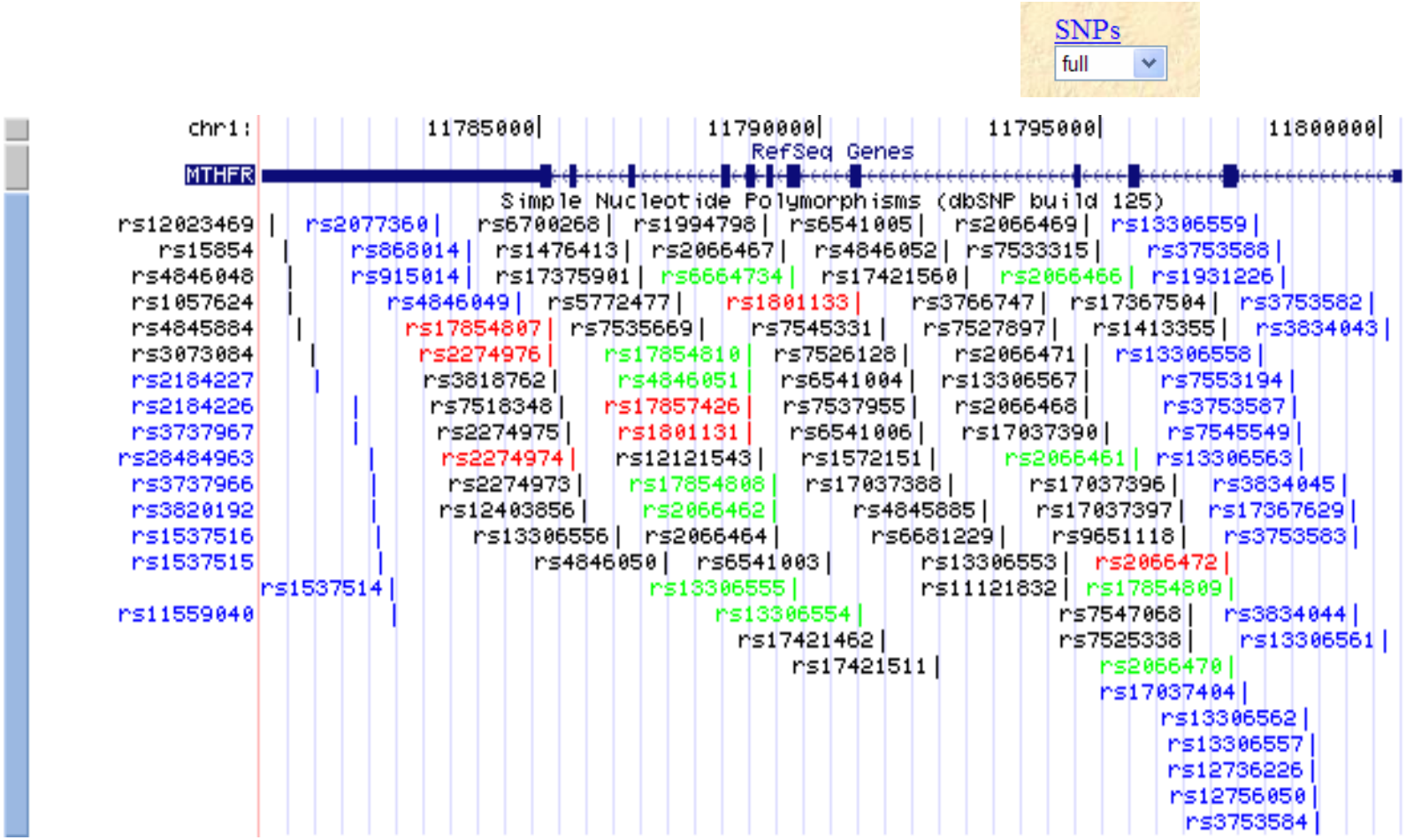
1000 Genomes release 16 - Oct 2014 © EBI

About 1000 Genomes | Contact

MTHFR Example



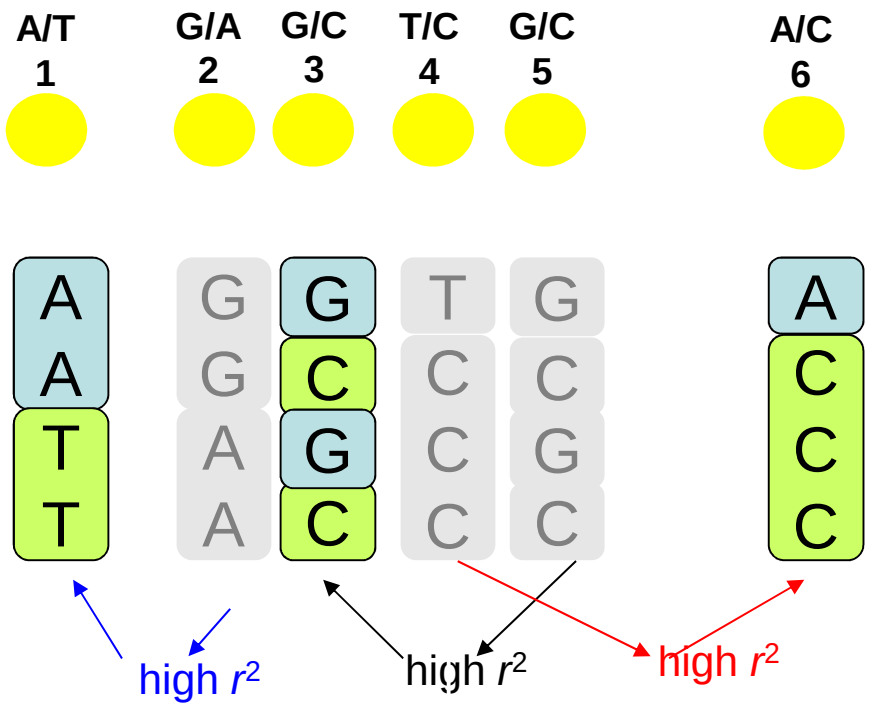
MTHFR SNPs



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:

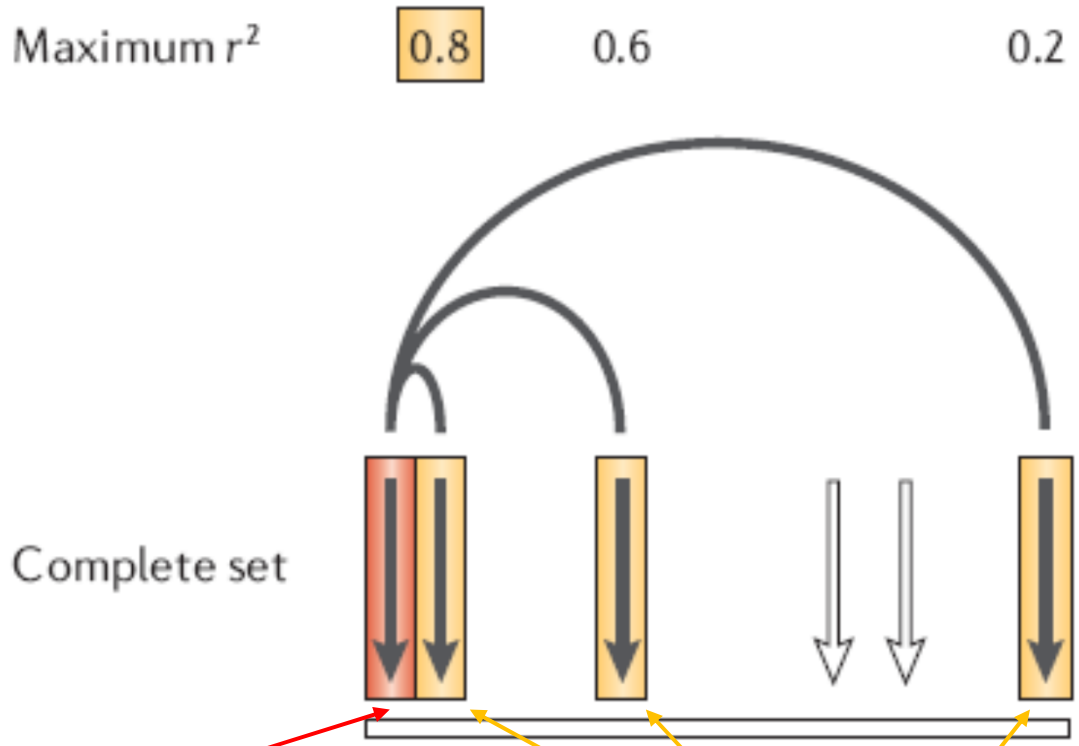
SNP 1
SNP 3
SNP 6

3 tags in total

Test for association:

SNP 1
SNP 3
SNP 6

Coverage: Measurement Error in TagSNPs



Want to compute coverage of all SNPs in the area, we start with this one, and look at pairwise correlation with each of the genotyped SNPs.

Genotyped SNPs

Common Measures of Coverage

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

- **<u>Average Measures</u>**
  - 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 www.broadinstitute.org/mpg/snap/

SNP Annotation and Proxy Search

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 was taken offline due to security issues & laziness (newer database)

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

1000 Genomes Project

- http://www.internationalgenome.org/about
- ~2-5x whole-genome sequencing coverage, some whole exome
- 2,504 individuals diverse genetic ancestries:

1000 Genomes Project Populations

Supergroup

Subgroup

	AFR	AMR	EAS	EUR	SAS
ACB	96	0	0	0	0
ASW	61	0	0	0	0
BEB	0	0	0	0	86
CDX	0	0	93	0	0
CEU	0	0	0	99	0
CHB	0	0	103	0	0
CHS	0	0	105	0	0
CLM	0	94	0	0	0
ESN	99	0	0	0	0
FIN	0	0	0	99	0
GBR	0	0	0	91	0
GIH	0	0	0	0	103
GWD	113	0	0	0	0

	AFR	AMR	EAS	EUR	SAS
IBS	0	0	0	107	0
ITU	0	0	0	0	102
JPT	0	0	104	0	0
KHV	0	0	99	0	0
LWK	99	0	0	0	0
MSL	85	0	0	0	0
MXL	0	64	0	0	0
PEL	0	85	0	0	0
PJL	0	0	0	0	96
PUR	0	104	0	0	0
STU	0	0	0	0	102
TSI	0	0	0	107	0
YRI	108	0	0	0	0

Haploview using 1000 Genomes Project data

(or just type in the name of the gene)

Go to the 1000 Genomes project website, and go to the 1000 Genomes Project Browser.

Enter gene name (e.g., MTHFR), to get the genomic coordinates (b37) of interest (e.g., 1:11845780-11866977) and view the region in the browser.

Click “Get VCF data” on LHS. Since you already searched for the gene (MTHFR), it will have the relevant coordinates already filled in for you.

We don’t want VCF format. Click the VCF to PED converter link on the window that pops up.

Click Next (the “Region” should have gene:coordinatesBEGIN-coordinatesEND)

Choose populations of interest – match what you are trying to target. So, for European ancestry (EUR), select CEU + FIN + GBR + IBS + TSI.

Click next, then right click on “Marker Information File” and “Linkage Pedigree File” to save the data set.

Finally, open Haploview, choose linkage format, and navigate to the pedigree information file (.ped), and then the marker information file (.info) should be auto-filled in. Click OK, and you can use Haploview!!!

Tag SNPs: HapMap & Haploview

This Column only happens because of old HapMap data (which is why I am leaving it in!) – also missing genotype rate

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

Select All

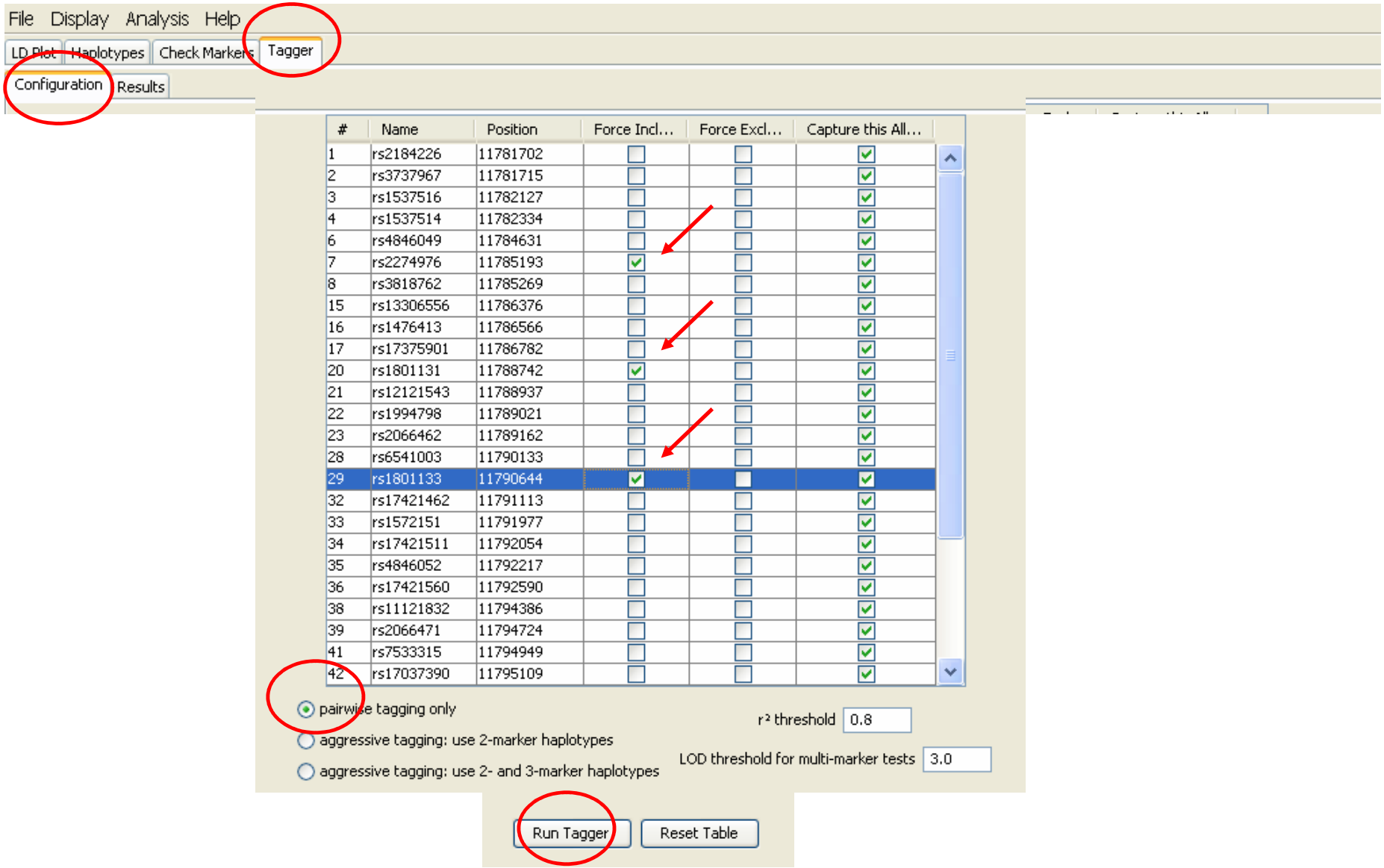
Minimum minor allele freq. 0.05

Rescore Markers

Tag SNPs: HapMap & Haploview

Can't get anymore ☹

Tag SNPs: HapMap & Haploview



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	r²
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

- How many tag SNPs did we need, to tag all SNPs with $r^2>0.8$?

- In practice forced in potentially functional/previously associated SNPs

- Identified tag based on pairwise tagging

- Note: number of SNPs required varies from gene to gene <u>and from population to population</u>
  - <u>Why?</u>

Extra Slides

(if time)

Pathways

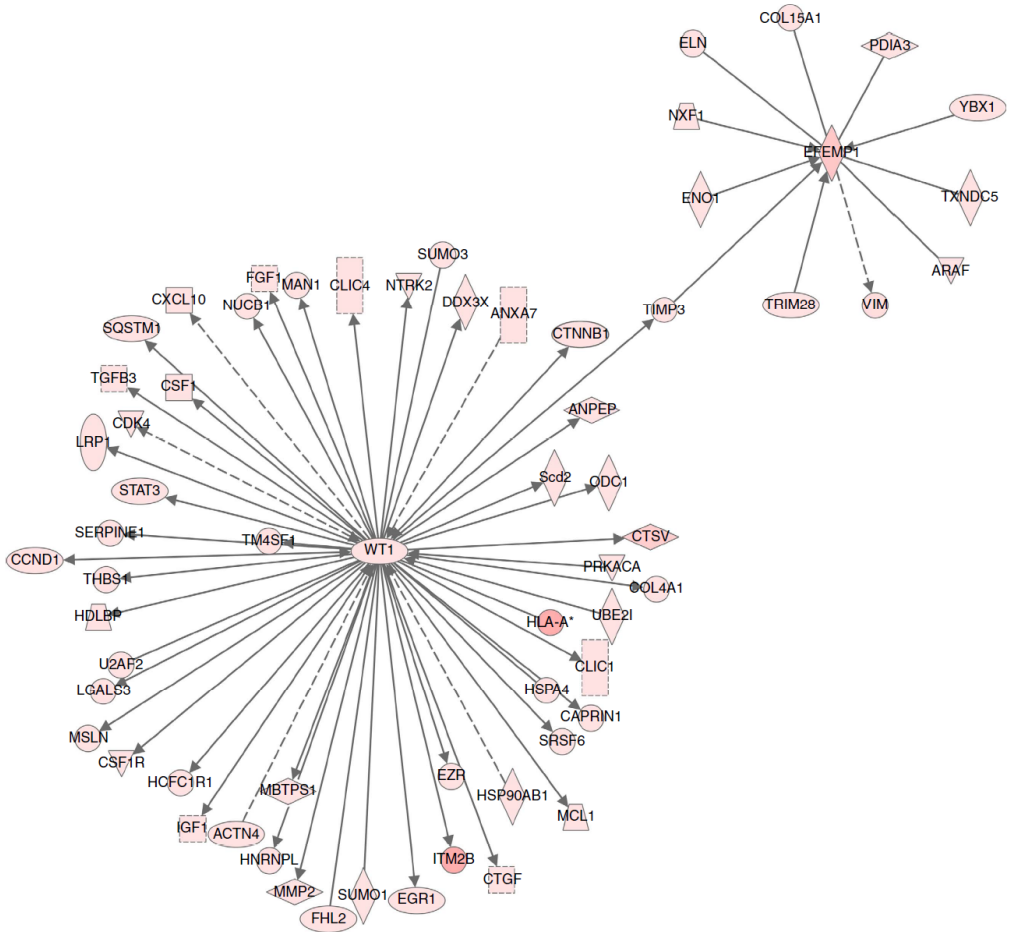


Figure 3 | Ingenuity Pathway Analysis outlines potential regulatory networks around EFEMP1 and WT1. Network Analysis for EFEMP1 and WT1 was carried out using the RNA-seq FPKM > 30 gene list (see Methods section). WT1 regulates many extracellular matrix genes, including MMP2 (matrix metalloproteinase-2) and CTGF (connective tissue growth factor). EFEMP1 directly interacts with ELASTIN, a component of elastic fibres in the ECM. TIMP3 (tissue inhibitor of metalloproteinase-3), which is activated by WT1 and interacts with EFEMP1 and was found to connect between the two networks.

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:

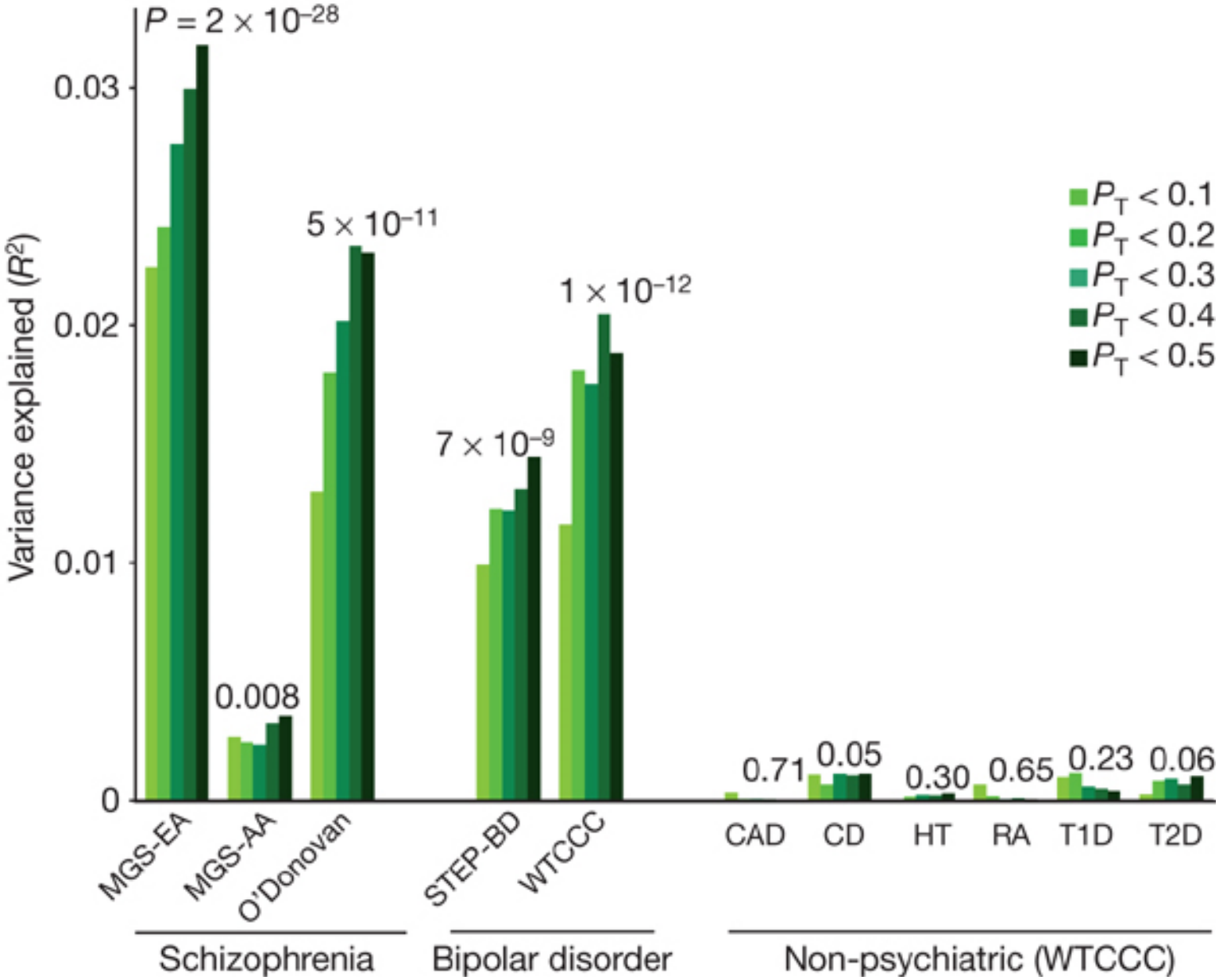
$$x_j = \frac{\sum_{i=1}^{m} \ln\left(OR_i\right) \times SNP_{ij}}{m}$$

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