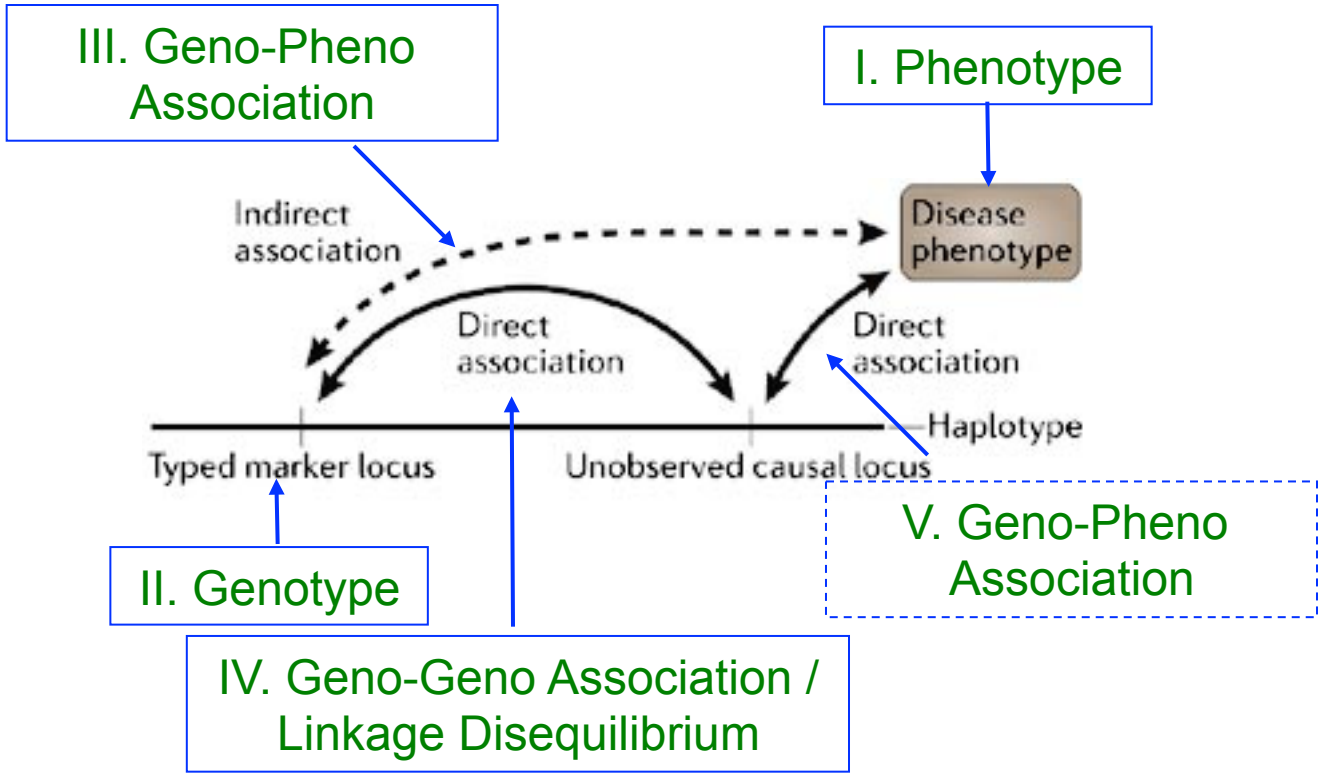
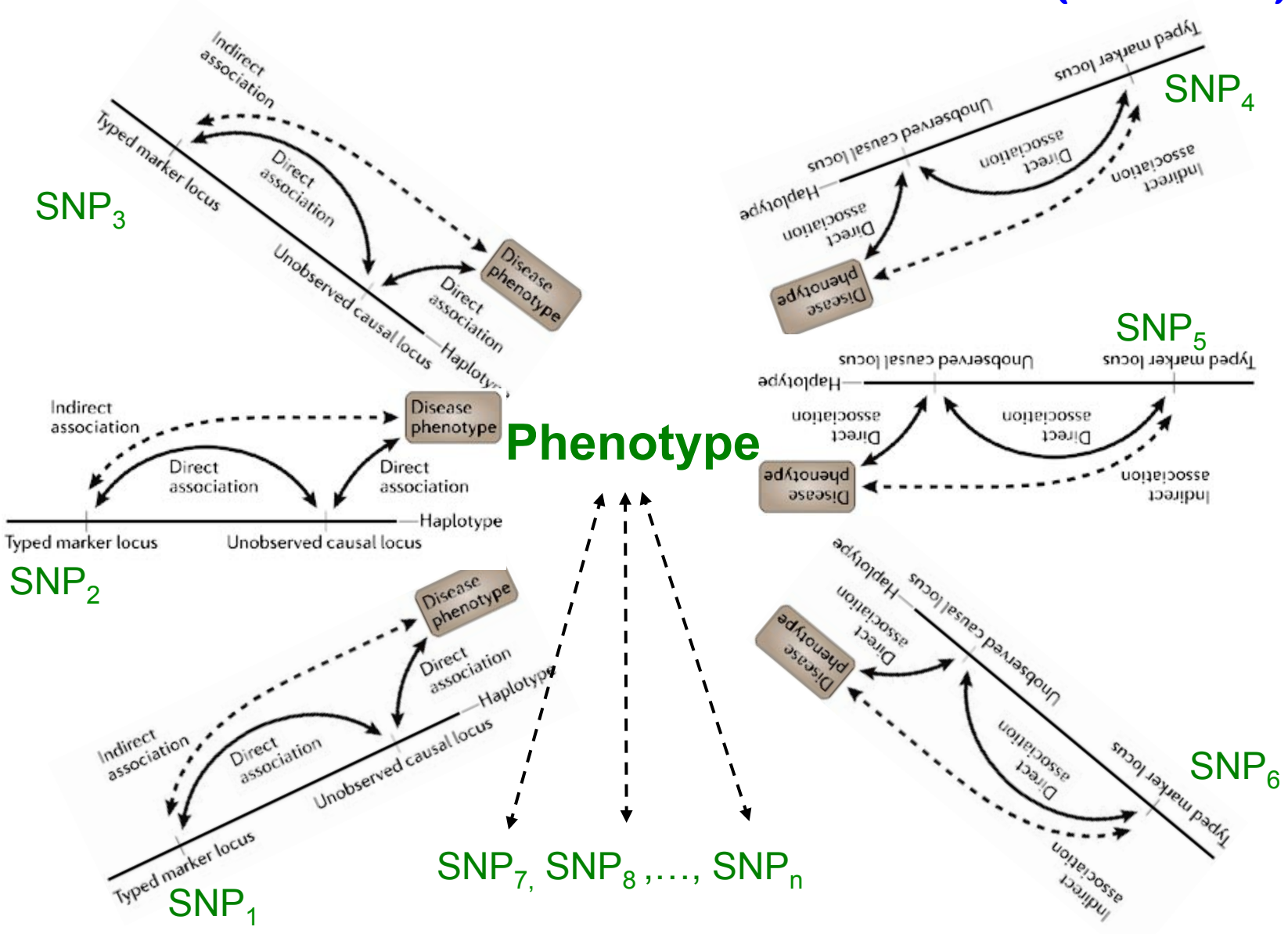
An Overview of Genome-Wide Association Study (GWAS): Rationale

Chao Xing, Ph.D.
McDermott Center for Human Genetics
Dept. of Bioinformatics & Clinical Sciences
chao.xing@utsouthwestern.edu
April 27th, 2017

Genetic Association Studies



Genome-Wide Association Studies (GWAS)



What is GWAS?

https://www.genome.gov/20019523/genomewide-association-studies-fact-sheet/

- A genome-wide association study is an approach that involves rapidly scanning markers across the complete sets of DNA, or genomes, of many people to find genetic variations associated with a particular disease.

  - Once new genetic associations are identified, researchers can use the information to develop better strategies to detect, treat and prevent the disease.
  - Such studies are particularly useful in finding genetic variations that contribute to common, complex diseases, such as asthma, cancer, diabetes, heart disease and mental illnesses.

An Early Example of Association (LD) Mapping

- Cystic Fibrosis
  - Autosomal recessive disease
  - Frequency = 1 per 2000 live births (Pr(d) = 0.022)

Science 1989, **245,** 1073

Identification of the Cystic Fibrosis Gene: Genetic Analysis

BAT-SHEVA KEREM, JOHANNA M. ROMMENS, JANET A. BUCHANAN, DANUTA MARKIEWICZ, TARA K. COX, ARAVINDA CHAKRAVARTI, MANUEL BUCHWALD, LAP-CHEE TSUI

Extensive linkage analysis provides evidence for the existence of a single CF locus on human chromosome 7 (region q31) (*7–10, 21*).

A candidate region of ~900 kb on chromosome 7q31 was identified via linkage analysis.

LD Mapping of CF Gene

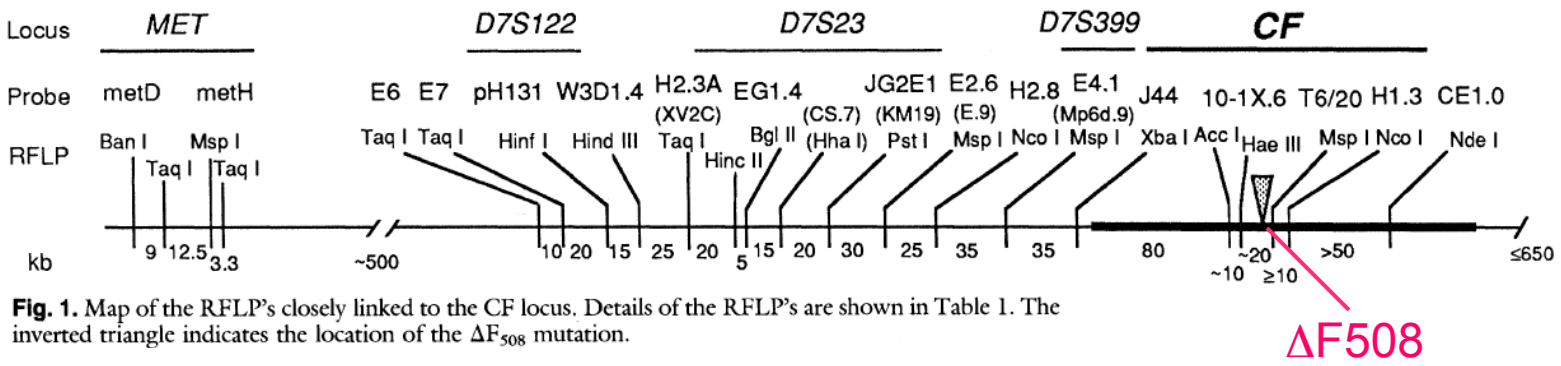
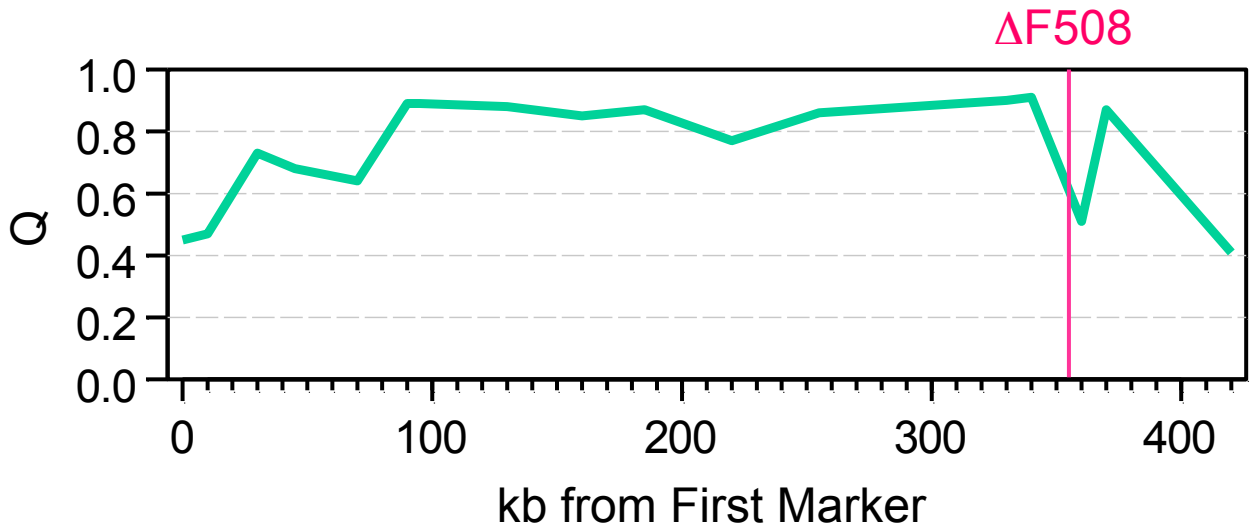
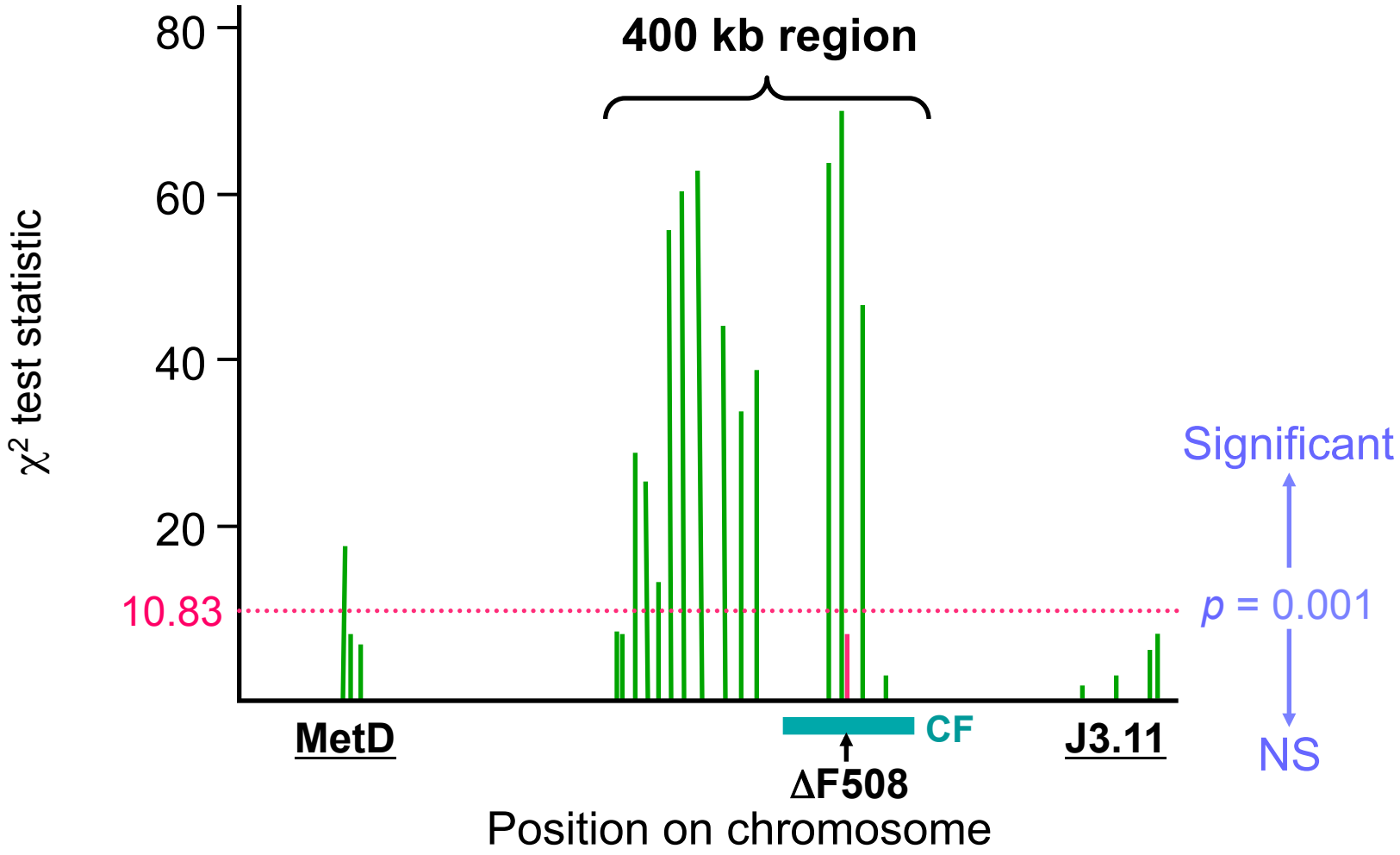


Fig. 1. Map of the RFLP's closely linked to the CF locus. Details of the RFLP's are shown in Table 1. The inverted triangle indicates the location of the ΔF_{508} mutation.



Kerem B et al. (1989) *Science* **245,** 1073

Association with CF Gene



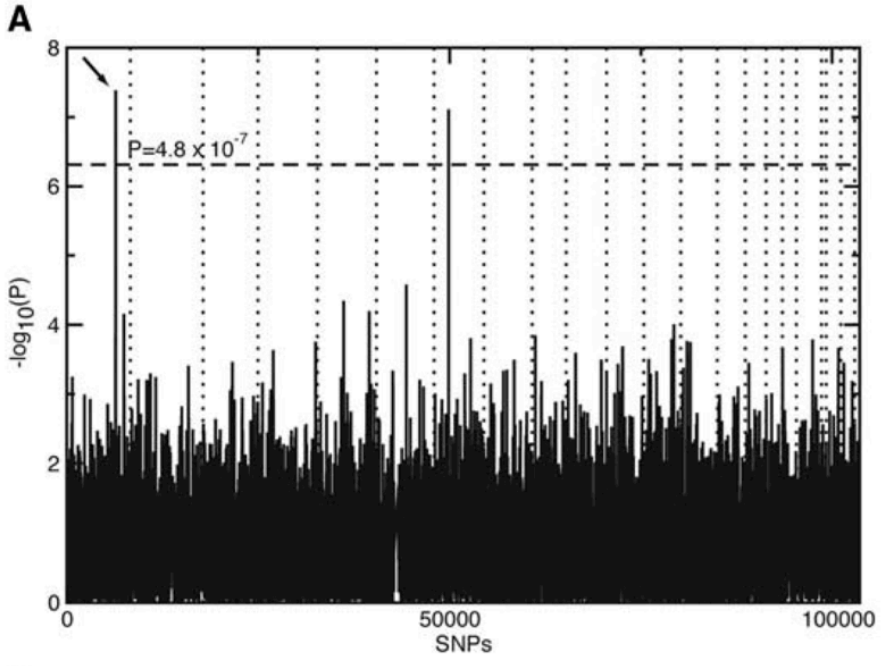
- RFLP nearest DF508 is not significant!

The First Genome-Wide Association Study

SCIENCE VOL 308 15 APRIL 2005

Complement Factor H Polymorphism in Age-Related Macular Degeneration

Age-related macular degeneration (AMD) is a major cause of blindness in the elderly. We report a genome-wide screen of 96 cases and 50 controls for polymorphisms associated with AMD. Among 116,204 single-nucleotide polymorphisms genotyped, an intronic and common variant in the comple-



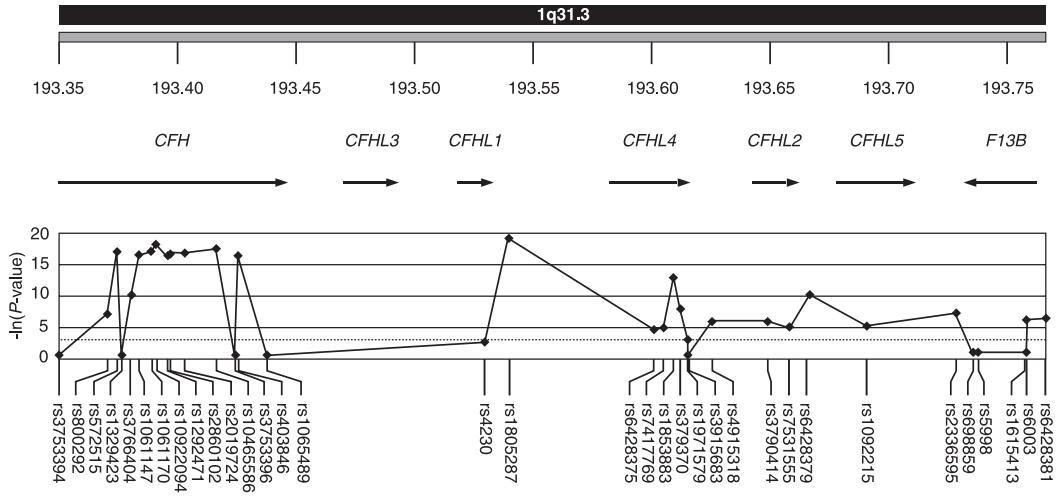
Accompanying The First GWAS

SCIENCE VOL 308 15 APRIL 2005 421

Complement Factor H Polymorphism and Age-Related Macular Degeneration

Albert O. Edwards,[1*†] Robert Ritter III,[1] Kenneth J. Abel,[2]
Alisa Manning,[3] Carolien Panhuysen,[3,6] Lindsay A. Farrer[3,4,5,6,7]

role (*3–9*). The first locus for AMD (*ARMD1*) was reported in a single extended family linked to chromosome 1q25.3-31.3 (*5*). Because there was strong evidence for linkage to this region

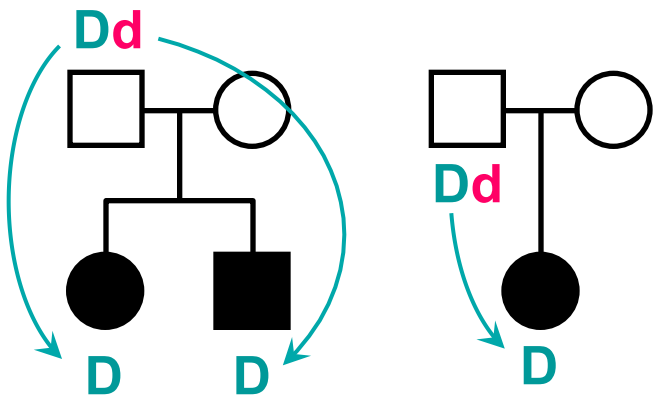


GWAS Made Possible

- Theoretical foundation
  - Rationale was presented by Risch & Merikangas (1996)
- A dense set of markers (SNPs/CNVs) capture a substantial proportion of common variation across the genome
  - HapMap Project, 1KG Project
  - Advance of technology

GWAS Rationale

Pr(share allele *ibd* from **one** parent) = 0.5 under H_0



Pr(one parent transmits D allele) = 0.5 under H_0

Linkage vs. Association

D = disease susceptibility allele
RR for Dd = γ ($\gamma > 1$) relative to dd
RR for DD = γ^2

- Looked at sample size required to detect linkage or association (TDT, McNemar's test) under a variety of disease models

Risch N & Merikangas K (1996) *Science* **273,** 1516

GWAS Rationale

		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.

- Sample sizes for 80% power; assumes parents are available; that the disease variant is directly genotyped ($r^2 = 1$); no allelic heterogeneity

Risch N & Merikangas K (1996) *Science* **273,** 1516

GWAS Rationale

The Future of Genetic Studies of Complex Human Diseases

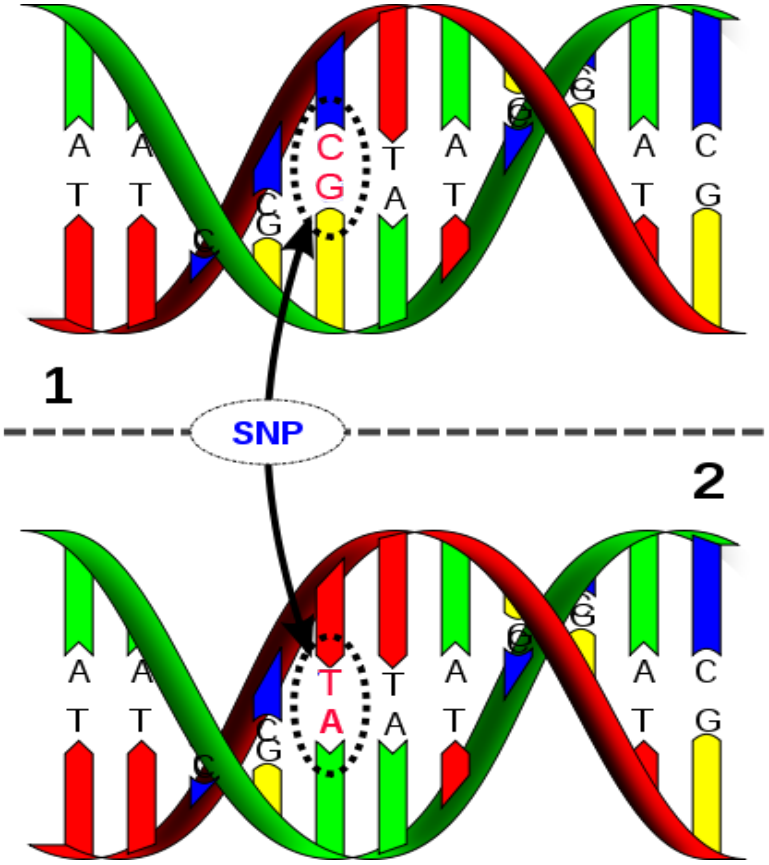
Neil Risch and Kathleen Merikangas

of affected siblings, we use a lod score (logarithm of the odds ratio for linkage) criterion of 3.0, which asymptotically corresponds to a type 1 error probability α of about 10^{-4}. In a linkage genome screen with 500 markers, this significance level gives a probability greater than 95% of no false positives. The equivalent false positive rate for 1,000,000 independent association tests can be obtained with a significance level $\alpha = 5 \times 10^{-8}$.

Thus, the primary limitation of genome-wide association tests is not a statistical one but a technological one. A large number of genes (up to 100,000) and polymorphisms (preferentially ones that create alterations in derived proteins or their expression) must first be identified, and an extremely large number of such polymorphisms will need to be tested.

Component II: Genetic Variations

Human Genome and Single Nucleotide Polymorphisms (SNPs)



- 23 chromosome pairs
- 3 billion bases
- A single nucleotide change between pairs of chromosomes
- E.g.

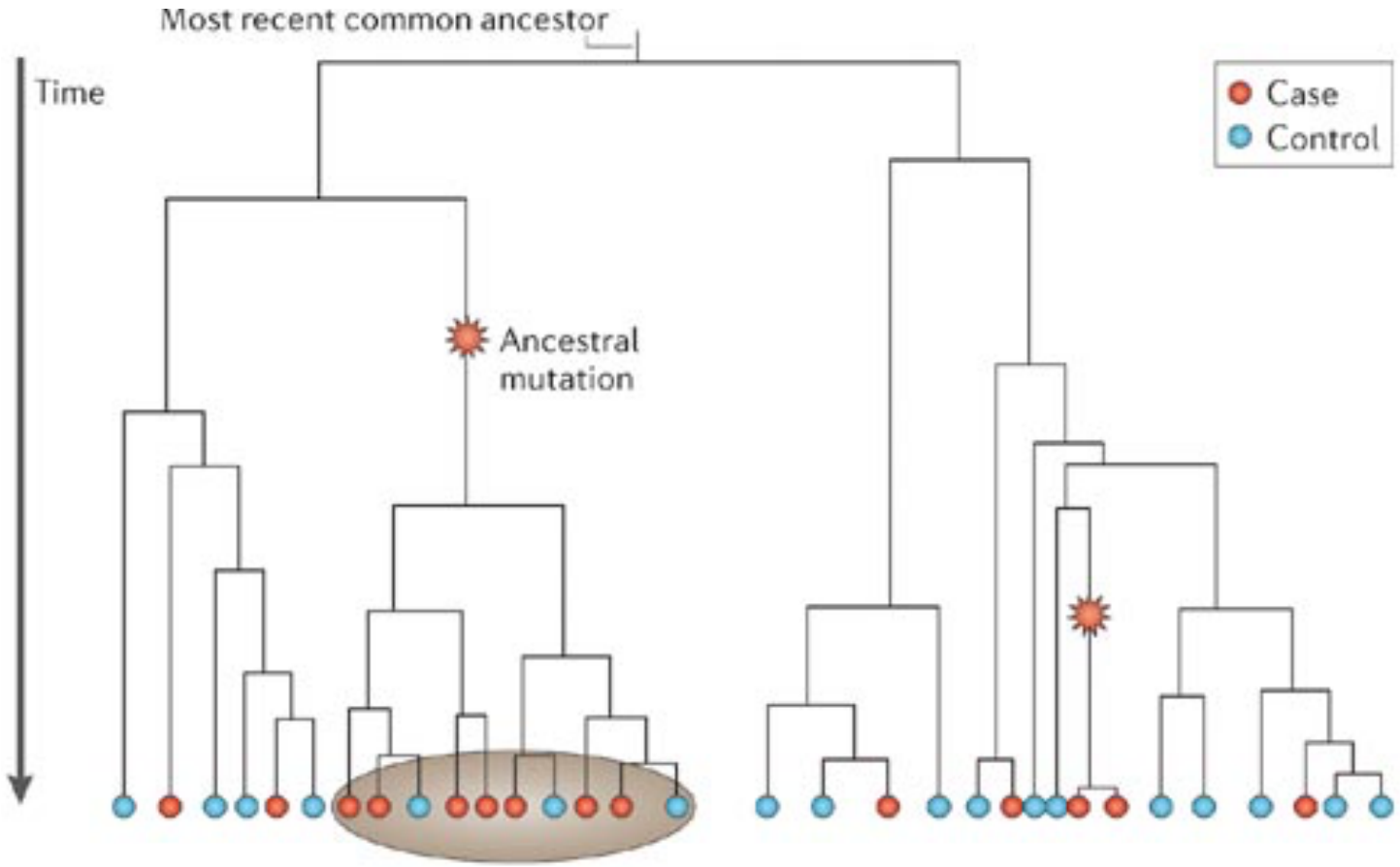
  Haplotype1: AAGG**G**ATCCAC
  Haplotype2: AAGG**A**ATCCAC
- A/A or G/G homozygote
- A/G heterozygote

Component II: Genetic Variations

- **Single nucleotide polymorphisms** (**SNPs**) due to point mutations.
- **Structural variation** due to deletions, duplications, insertions, inversions, and translocations.
  - Microscopic variation (more than 3Mb).
  - Submicroscopic variation (less than 3Mb).
    - **Copy number variants** (**CNVs**) are submicroscopic structural variations that are due to deletion, duplication, and replicative transposition.
    - If the variation in copy number occurs in tandem, it is referred to as **variable number of tandem repeats** (**VNTRs**).

Component III: Association Test

Rationale of genetic association study



Balding (2006) *Nat Rev Genet* **7,** 781

Component III: Association Test

- **Standard χ^2 test on alleles (1 d.f.)**

	A	a	Total
Cases	r_A	r_a	$2R$
Controls	s_A	s_a	$2S$
Total	n_A	n_a	$2N$

H_0: $p_{A,\,case} = p_{A,\,ctrl}$

$$X^2 = \frac{2N(r_A s_a - r_a s_A)^2}{(2R)(2S)(n_A)(n_a)}$$

$$= 2N \cdot r^2$$

- N individuals, R cases and S controls; $2N$ alleles total
- Warning! Requires HWE in cases *and* controls *together* for validity
- Tends to be anticonservative when there is an excess of homozygotes

Sasieni (1997) *Biometrics* **53:**1253

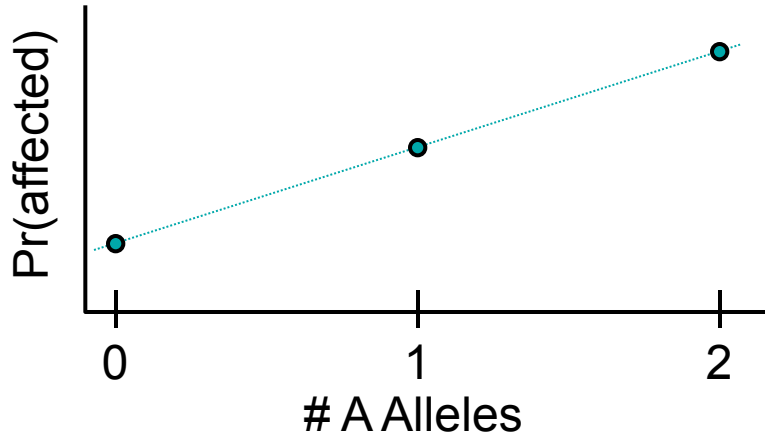
Component III: Association Test

- **Standard χ^2 test on genotypes (2 d.f.)**

	aa	Aa	AA	Total
Cases	r_0	r_1	r_2	R
Controls	s_0	s_1	s_2	S
Total	n_0	n_1	n_2	N

H_0: genotype frequencies same in cases & controls

- **Cochran-Armitage trend test (1 d.f.)**
  - Assumes linear relationship in Pr(case | genotype)
  - Tests H_0: slope of the line is zero
  - Equivalent to the score test under the logistic model



Balding (2006) *Nat Rev Genet* **7,** 781

Component III: Association Test

- **Logistic regression**

$$\ln\left(\frac{p}{1-p}\right) = \alpha + \sum_{i=1}^{k}\beta_i x_i + \boldsymbol{\beta}_g \mathbf{x}_g$$

p = Pr(case)

Covariate effects

Genotype effects

  - Additive model (on the log scale)

$$x_g(\mathrm{aa}) = 0,\ x_g(\mathrm{Aa}) = 1,\ x_g(\mathrm{AA}) = 2$$

  - Dominant model (in A)

$$x_g(\mathrm{aa}) = 0,\ \ x_g(\mathrm{Aa}) = x_g(\mathrm{AA}) = 1$$

  - Genotype model

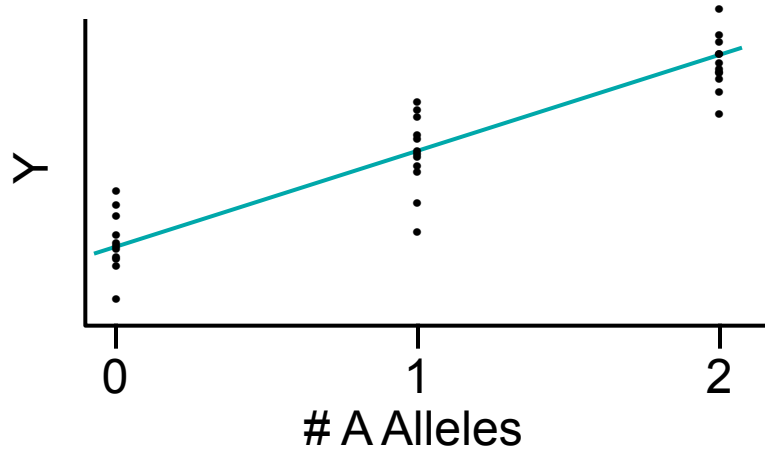
$$\boldsymbol{\beta}_g \mathbf{x}_g = \beta_{Aa} x_{Aa} + \beta_{AA} x_{AA},\ \ x_g = \begin{cases} 1 \text{ if genotype is } g \\ 0 \text{ otherwise} \end{cases}$$

Component III: Association Test

- **Linear regression**

$$Y = \alpha + \sum_i \beta_i x_i + \beta_A x_A + e$$

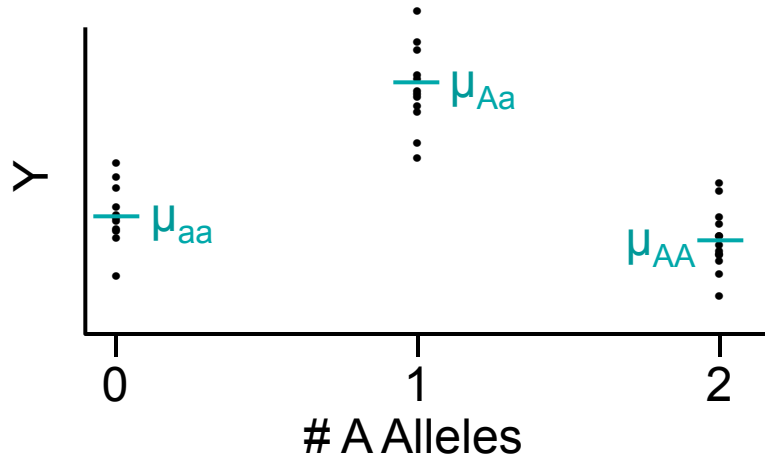
  - For dominant/recessive model w/o covariates, becomes a *t* test



- **ANOVA**

$$Y = \alpha + \sum_i \beta_i x_i + \beta_{Aa} x_{Aa} + \beta_{AA} x_{AA} + e$$

  - Analogous to the 2-d.f. χ^2 test for a binary trait
  - Assumptions for linear regression (normality, etc.) apply here & above



Component IV: Linkage Disequilibrium (LD)

Genotype-Phenotype Association

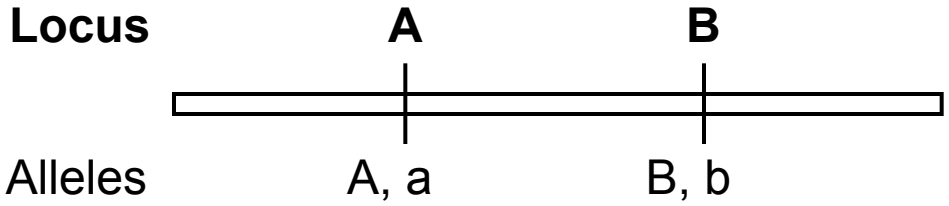
	A	**a**	**Total**
Cases	r_A	r_a	$2R$
Controls	s_A	s_a	$2S$
Total	n_A	n_a	$2N$

Genotype-Genotype Association (**LD**)

	A	**a**	**Total**
D ~~**Cases**~~	r_A	r_a	$2R$
d ~~**Controls**~~	s_A	s_a	$2S$
Total	n_A	n_a	$2N$

Linkage Disequilibrium (LD)

- **Linkage Disequilibrium:** Association between the alleles of two loci due to their close proximity on a chromosome
- Under linkage *equilibrium,* the alleles of two loci are independent



Haplotypes: AB, Ab, aB, ab

Under equilibrium: $Pr(AB) = Pr(A)Pr(B)$
$Pr(ab) = Pr(a)Pr(b)$, etc.
$Pr(A \mid B) = Pr(A)$
$Pr(B \mid A) = Pr(B)$

Measures of LD

	B	b	Total
A	n_{AB}	n_{Ab}	n_A
a	n_{aB}	n_{ab}	n_a
Total	n_B	n_b	n

	B	b	Total
A	p_{AB}	p_{Ab}	p_A
a	p_{aB}	p_{ab}	p_a
Total	p_B	p_b	1

- Let $p_{ij} = n_{ij} / n$ be the *observed* proportion of each haplotype
- Let π_{ij} be the (unknown) "true" population proportion
  - Definitions are given in terms of the true proportions, but are calculated using the estimates from samples.

1. **D** = difference between observed (true) proportions and their expected values under equilibrium.

$$D = p_{AB} - p_A p_B = p_{ab} - p_a p_b$$
$$= -(p_{Ab} - p_A p_b) = -(p_{aB} - p_a p_B)$$
$$= p_{AB} p_{ab} - p_{Ab} p_{aB}$$

[Pr(AB) – Pr(A)Pr(B)]

Devlin B & Risch N (1995) *Genomics* **29,** 311

Measures of LD

2. Lewontin's **D′** = D/D_{max}: standardized for allele frequencies

$$D' = \begin{cases} \dfrac{p_{AB}p_{ab} - p_{Ab}p_{aB}}{\min(p_A p_b, p_a p_B)} & D > 0 \\ \dfrac{p_{AB}p_{ab} - p_{Ab}p_{aB}}{\min(p_A p_B, p_a p_b)} & D < 0 \end{cases}$$

$-1 \le D' \le 1$

3. $\Delta^2 = r^2$ = square of the standardized measure

$$\Delta^2 = r^2 = \frac{(p_{AB}p_{ab} - p_{aB}p_{Ab})^2}{p_A p_a p_B p_b} = \frac{D^2}{p_A p_a p_B p_b}$$

Note: Δ is the correlation coefficient for a 2 × 2 table

Devlin B & Risch N (1995) *Genomics* **29,** 311

Measures of LD

4. Levin's population attributable risk, δ*

$$\delta^* = \frac{p_A(\phi - 1)}{1 + p_A(\phi - 1)}, \text{ where } \phi = \frac{p_{AB}/p_A}{p_{aB}/p_a} \text{ is the relative risk}$$

5. Approximation to δ* that is robust to preferential sampling of disease haplotypes in case-control sampling

$$\delta = P_{excess} = \frac{p_{AB}p_{ab} - p_{Ab}p_{aB}}{p_B p_{ab}}$$

Devlin B & Risch N (1995) *Genomics* **29,** 311

Measures of LD

6. Difference in proportions, ***d***

$$d = \frac{p_{AB}}{p_B} - \frac{p_{Ab}}{p_b} = \frac{p_{AB}p_{ab} - p_{Ab}p_{aB}}{p_B p_b}$$

7. Odds ratio, **λ**

$$\lambda = \frac{p_{AB}p_{ab}}{p_{Ab}p_{aB}} \qquad 0 \le \lambda < \infty$$

- Does not depend on marker allele frequencies

8. Yule's **Q**

$$Q = \frac{\lambda - 1}{\lambda + 1} = \frac{p_{AB}p_{ab} - p_{Ab}p_{aB}}{p_{AB}p_{ab} + p_{Ab}p_{aB}} \qquad -1 \le Q \le 1$$

Devlin B & Risch N (1995) *Genomics* **29,** 311

Which measure is best for mapping?

- For disease gene **localization**, Devlin and Risch (1995) recommended δ, the modified population attributable risk
  - E(D′) and E(δ) *maximum* at disease locus
  - δ is robust to case-control sampling
- D′ may often be used in place of δ
  - D′ $\propto$ δ in common inheritance models
  - Estimates of D′ are biased upwards in small samples or when one allele is rare
- r^2 gives indication of *power* to detect disease locus
  - but r^2 depends on allele frequencies; may not have max. expectation at disease locus
  - Estimates of r^2 may also be inflated in small samples

Lewontin (1964) *Genetics* **49,** 49

https://hapmap.ncbi.nlm.nih.gov/

- "The International HapMap Project is a multi-country effort to identify and catalog genetic similarities and differences in human beings."
- "The goal of the International HapMap Project is to compare the genetic sequences of different individuals to identify chromosomal regions where genetic variants are shared."

- **To produce a genome-wide map of common variation**
  - The International HapMap Project. Nature 426, 789-796. **2003**.
  - A Haplotype Map of the Human Genome. Nature 437, 1299-1320. **2005**.
  - A second generation human haplotype map of over 3.1 million SNPs. Nature 449, 851-861. **2007**.
  - Integrating common and rare genetic variation in diverse human populations. Nature 467, 52-58. **2010**.

HapMap Phases

- Phase I
  - Goal: One common (MAF ≥ 0.05) SNP every 5 kb
  - Over 1 million SNPs genotyped
  - 90 CEU (30 trios), 90 YRI (30 trios), 45 CHB, 44 JPT
  - ENCODE: Ten 500-kb regions were sequenced in 48 individuals
- Phase II
  - Extension of Phase I with 2.1 million additional SNPs
  - More uncommon SNPs
- Phase III
  - Extension to other populations
  - Genotyped for Affymetrix 6.0 and Illumina 1M chips; ca. 1.6 million SNPs

HapMap Phase 3 Samples

International HapMap Consortium (2010) *Nature* 467, 52

label	population sample	#samples	QC+ Draft 1
ASW*	African ancestry in Southwest USA	90	71
CEU*	Utah residents with Northern and Western European ancestry from the CEPH collection	180	162
CHB	Han Chinese in Beijing, China	90	82
CHD	Chinese in Metropolitan Denver, Colorado	100	70
GIH	Gujarati Indians in Houston, Texas	100	83
JPT	Japanese in Tokyo, Japan	91	82
LWK	Luhya in Webuye, Kenya	100	83
MEX*	Mexican ancestry in Los Angeles, California	90	71
MKK*	Maasai in Kinyawa, Kenya	180	171
TSI	Toscans in Italy	100	77
YRI*	Yoruba in Ibadan, Nigeria	180	163
		1,301	**1,115**

* Population is made of family trios

1000 Genomes Phase 1 Samples

Table 1 Samples in the 1000 Genomes Project phase I integrated variant set

Full Population Name	Abbreviation	No. Samples
African Ancestry in Southwest US	ASW	61
Luhya in Webuye, Kenya	LWK	97
Yoruba in Ibadan, Nigeria	YRI	88
Total African ancestry	**AFR**	**246**
Colombian in Medellin, Colombia	CLM	60
Mexican Ancestry in Los Angeles, CA	MXL	66
Puerto Rican in Puerto Rico	PUR	55
Total American ancestry	**AMR**	**181**
Han Chinese in Beijing, China	CHB	97
Han Chinese South, China	CHS	100
Japanese in Tokyo, Japan	JPT	89
Total Asian ancestry	**ASN**	**286**
Utah residents (CEPH) with Northern and Western European ancestry	CEU	85
Toscani in Italia	TSI	98
British in England and Scotland	GBR	89
Finnish in Finland	FIN	93
Iberian populations in Spain	IBS	14
Total European ancestry	**EUR**	**379**

The Project has grouped these 1092 samples into four ancestry groups representing the "predominant component of ancestry": African (AFR), American (AMR), Asian (ASN), and European (EUR) (Abecasis *et al.* 2012).

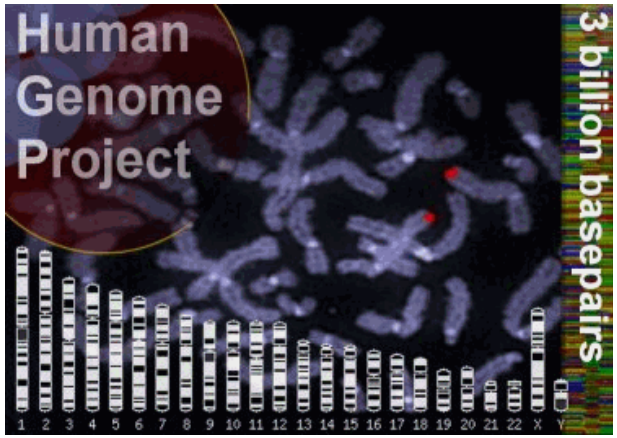
Nelson et al. (2013) *Genes Genet Genomics* **3,** 1795-1807

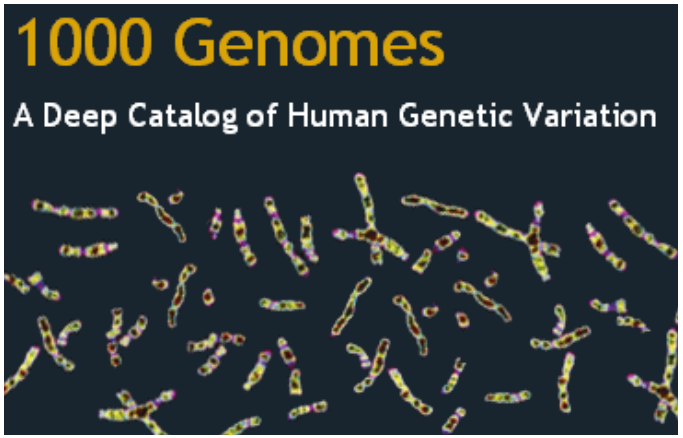
1000 Genomes Phase 3 Samples

Population	Trios	*n*
Chinese Dai in Xishuangbanna (**CDX**)	Y	93
Han Chinese in Beijing (**CHB**)	N	103
Japanese in Tokyo (**JPT**)	N	104
Kinh in Ho Chi Minh City, Vietnam (**KHV**)	Y	99
Southern Han Chinese (**CHS**)	Y	105
Total East Asian		**504**
Bengali in Bangladesh (**BEB**)	Y	86
Gujarati Indian in Houston (**GIH**)	Y	103
Indian Telugu in the UK (**ITU**)	Y	102
Punjabi in Lahore, Pakistan (**PJL**)	Y	103
Sri Lankan Tamil in the UK (**STU**)	Y	102
Total South Asian		**489**
African Ancenstry in SW USA (**ASW**)	Y	61
African Caribbean in Barbados (**ACB**)	Y	96
Esan in Nigeria (**ESN**)	Y	99
Gambian in Western Division (**GWD**)	Y	113
Luhya in Webuye, Kenya (**GWD**)	Y	99
Mende in Serra Leone (**MSL**)	Y	85
Yoruba in Ibadan, Nigeria (**YRI**)	Y	108
Total African		**661**
British in UK (**GBR**)	Y	91
Finnish in Finland	N	99
Iberian in Spain (**IBS**)	Y	107
Toscani in Italy (**TSI**)	N	107
CEPH Utah (**CEU**)	Y	99
Total European		**503**
Colombian in Medellín (**CLM**)	Y	94
Mexican in Los Angeles (**MXL**)	Y	64
Peruvian in Lima (**PEL**)	Y	85
Puerto Rican (**PUR**)	Y	104
Total Americas		**347**
Grand Total		**2504**

1000 Genomes Project Web site

Exploring the human genome





2002 2008

Sanger sequencing, targeted genotyping

Genome-wide genotyping (GWAS)

Exome sequencing

Genome sequencing

HapMap

- **Main Goal: "to determine the common patterns of DNA sequence variation"**

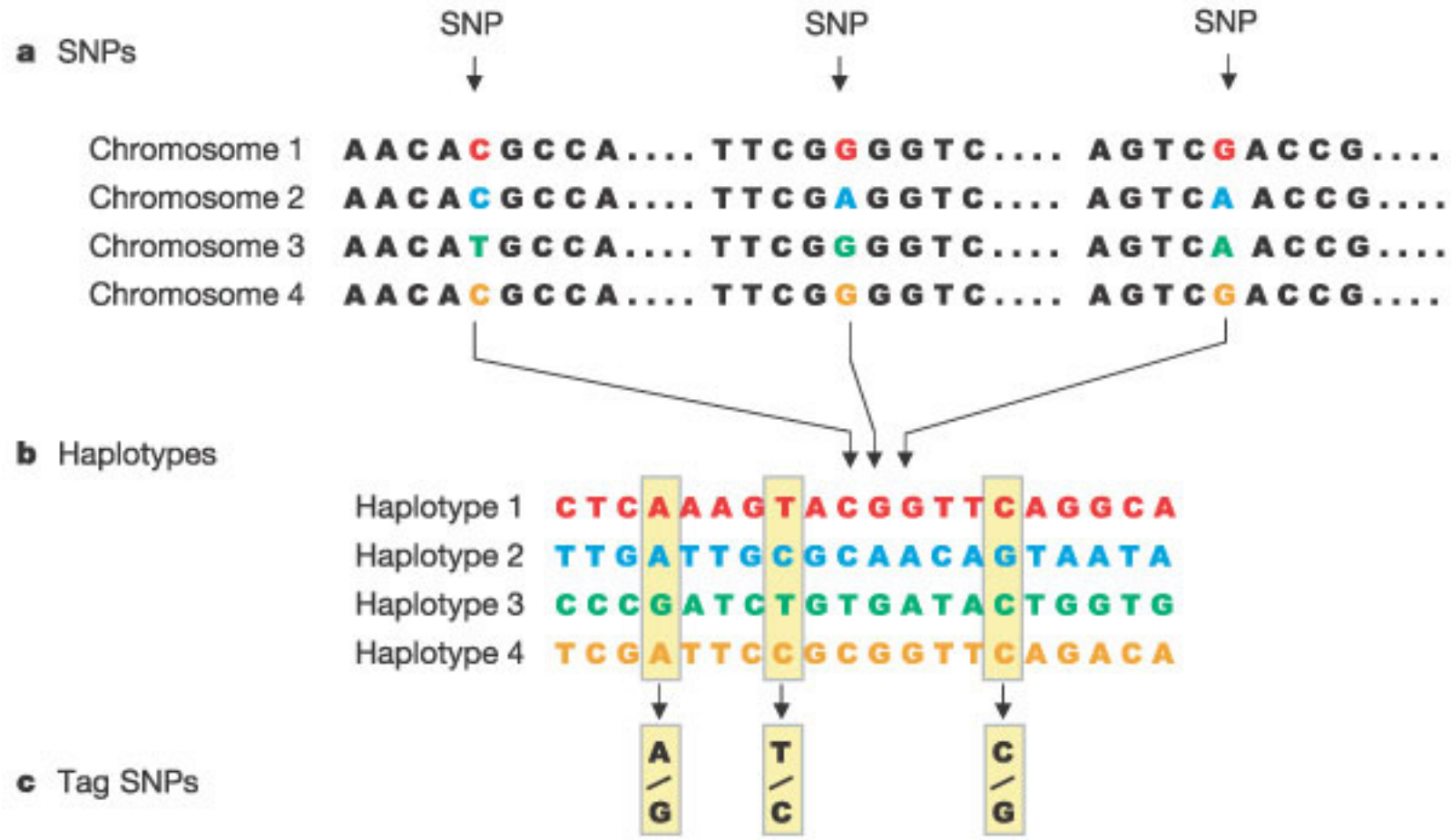
➢ LD "block" structure

- In small genomic regions, a small number of haplotypes accounts for most of the overall variability

➢ Tagging

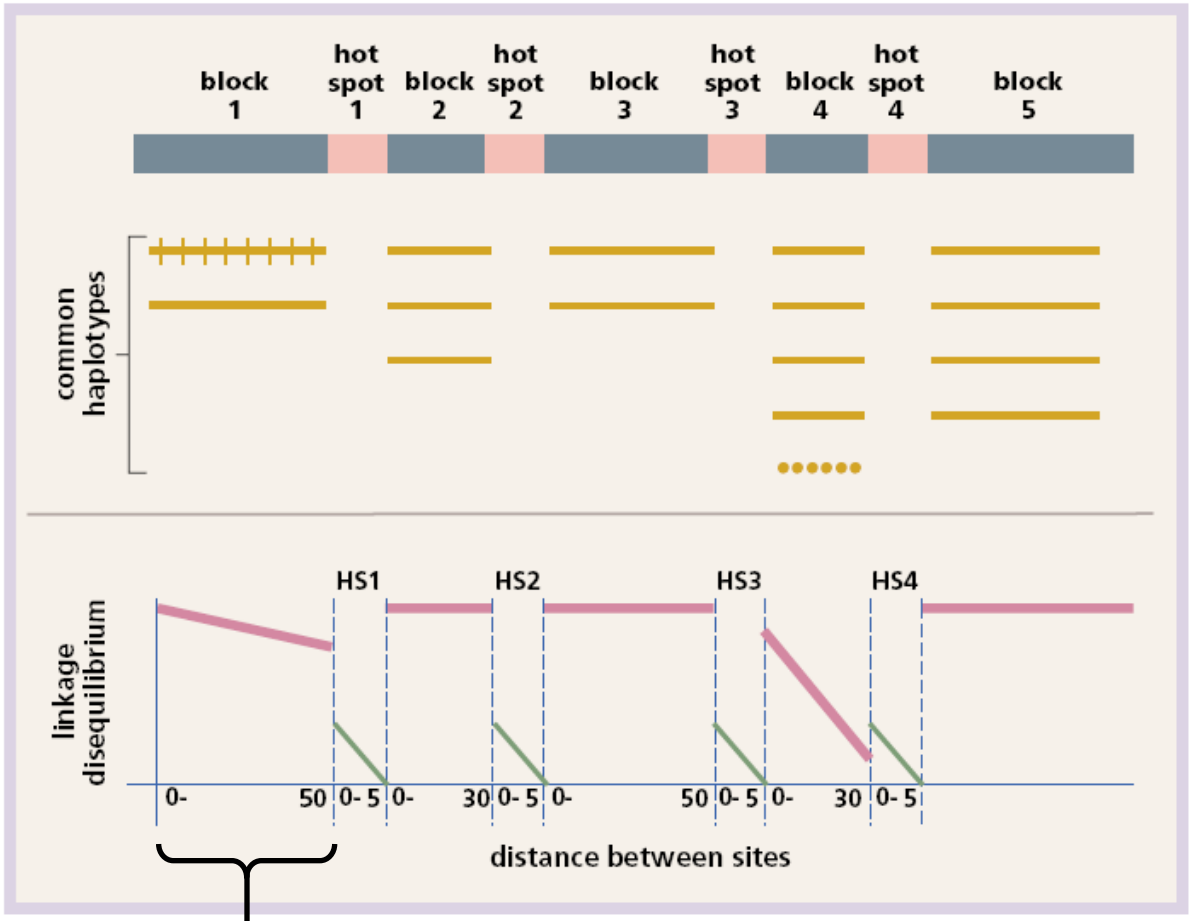
- Thus, we can use "tag" SNPs to capture the variation in haplotypes

SNPs, Haplotypes, & Tag SNPs



Int'l HapMap Consortium (2003) *Nature* **426,** 789

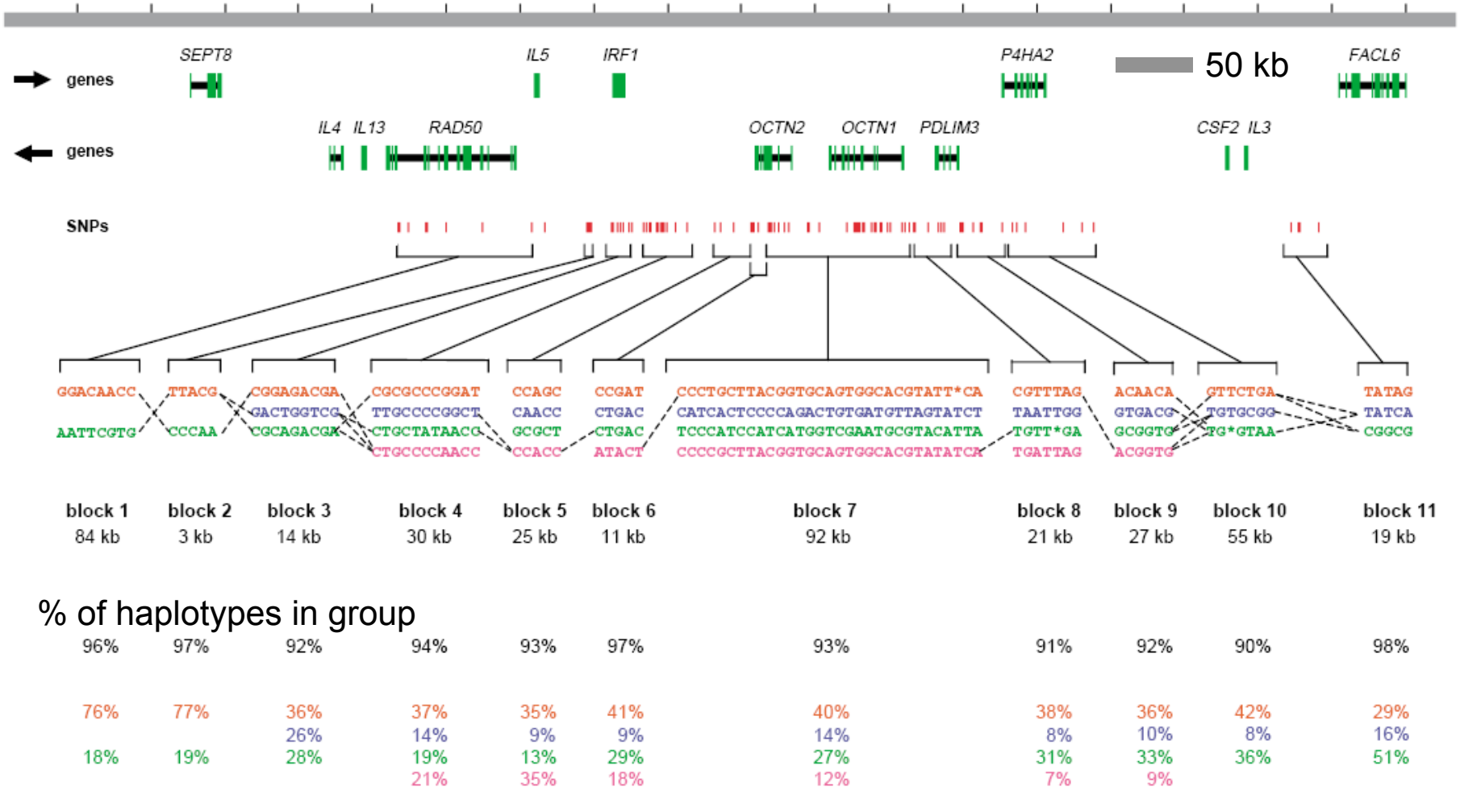
Haplotype Blocks



Decay of LD with distance: 0 kb (left)-50 kb (right)

Goldstein DB (2001) *Nat Genet* **29,** 109

LD Blocks on Chromosome 5q31

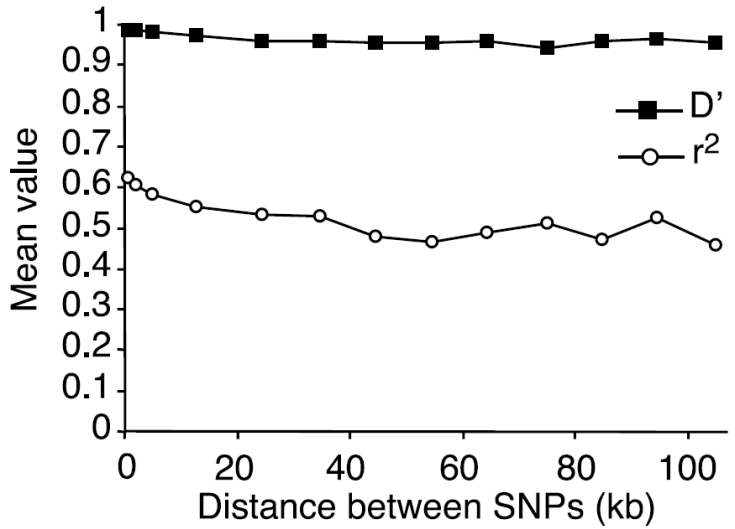


Sample size ~ 500 chromosomes

Daly MJ (2001) *Nat Genet* **29,** 229

Defining Haplotype Blocks

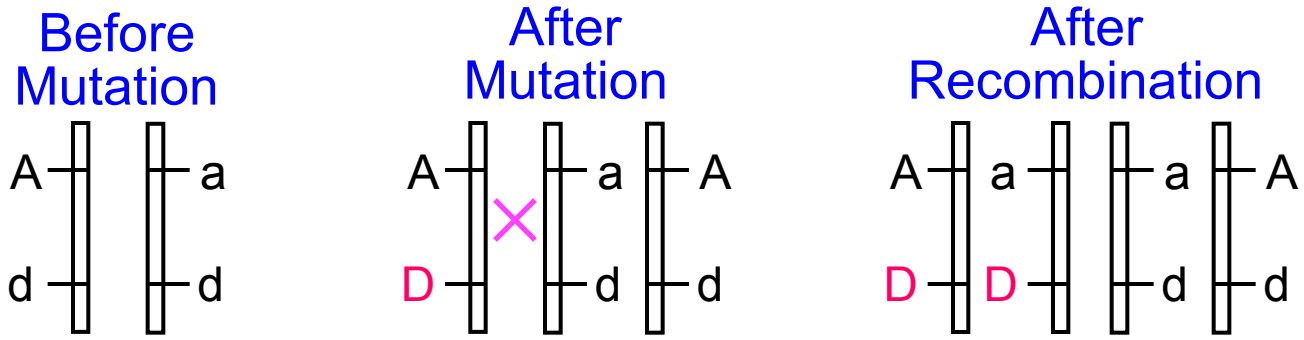
- **Gabriel et al** (2002): Examine 95% CI for | D′ | pairwise
  - "Strong LD" (no evidence for historical recombination): 95% CI = (> 0.7, > 0.98)
  - "Strong evidence for historical recombination": upper bound < 0.9
  - **Haplotype Block:** Region in which < 5% of pairs of SNPs show strong evidence for historical recombination
  - LD within blocks is very strong (*right*)
  - Implemented in Haploview



Gabriel SB (2002) *Science* **296,** 2225

Defining Haplotype Blocks

- **Four-gamete Rule (Hudson & Kaplan 1985)**
  - *Infinite-sites model:* Assume
    - a particular mutation cannot occur twice independently, and
    - reverse mutations don't occur
  - Then, all four haplotypes of two SNPs are found *only if recombination has occurred.*



  - *Implementation:* Within a block, all pairs of markers must pass the four-gamete test
    - Haplotype frequency of at least one of the four must be < 0.01 (HaploView).

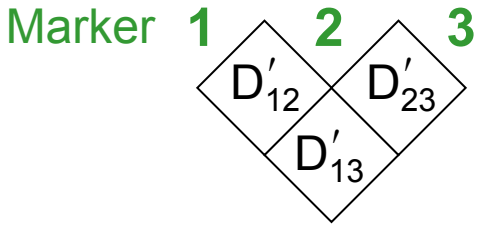
Wang N (2002) *Am J Hum Genet* **71,** 1227

Defining Haplotype Blocks

- **Solid Spine of LD** (HaploView)
  - First and last markers in block are in strong LD with intermediate markers
  - However, intermediate markers may not be in strong LD with each other

Example HaploView LD Plot

Darker red = higher | D′ |



Blocks and SNP Density

- Increasing density of the SNP scans changes the size of LD blocks!

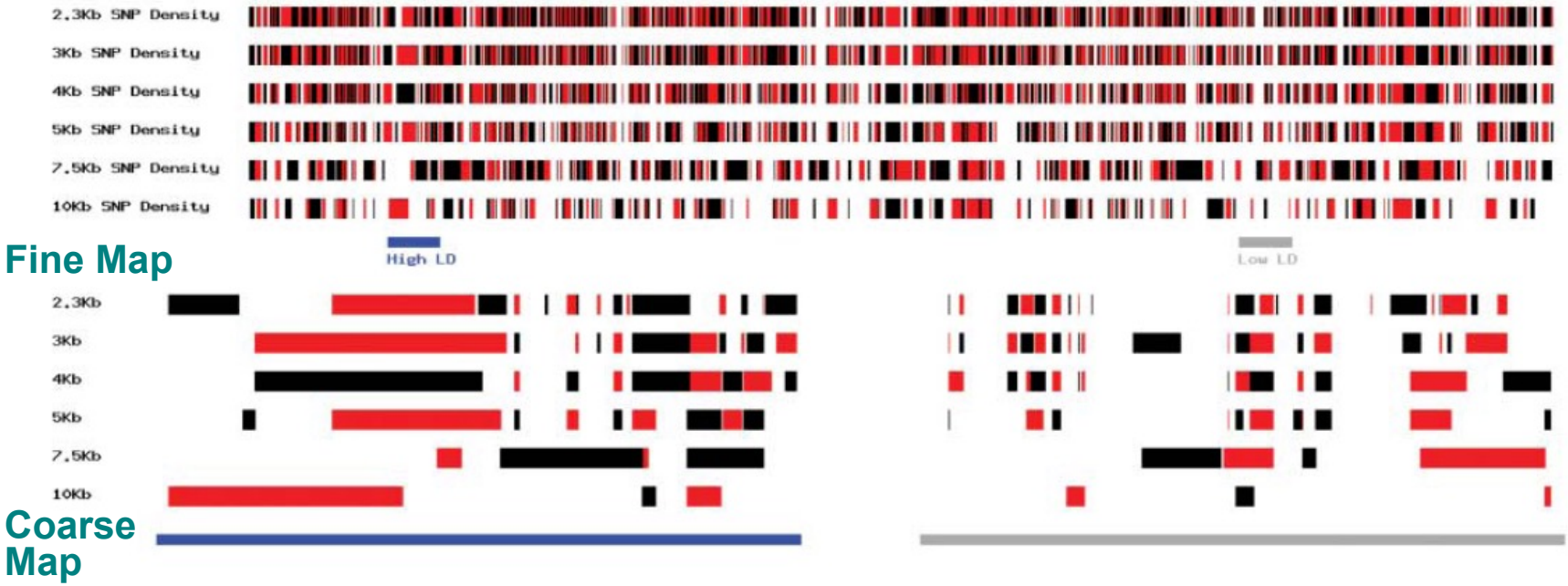


Figure 7. Average block sizes and sequence coverage for the 96 unrelated UK Caucasian and 97 African American individuals. Blocks were identified using the Gabriel *et al.* (3) approach at an overall marker density of 10 kb (one marker per 10 kb), 7.5 kb or 5, 4, 3 and 2.3 kb for UK Caucasian (**A**) and African American (**B**), respectively. The top two panels show the percentage of genomic sequence covered by blocks on the *y*-axis (red lines) and the average block size on the *z*-axis (blue lines). (**C**) A graphic depiction of blocks across the entire 10 Mb region. Within each marker density (row) blocks are alternately coloured black and red to show the points of change. Below the full region image, two illustrative 400 kb segments of high (rs2066906–rs967083) and low (rs2183794–rs932675) LD are expanded in greater detail to indicate specific block compositions.

Ke et al. (2004) *Hum Mol Genet* **13,** 577-588

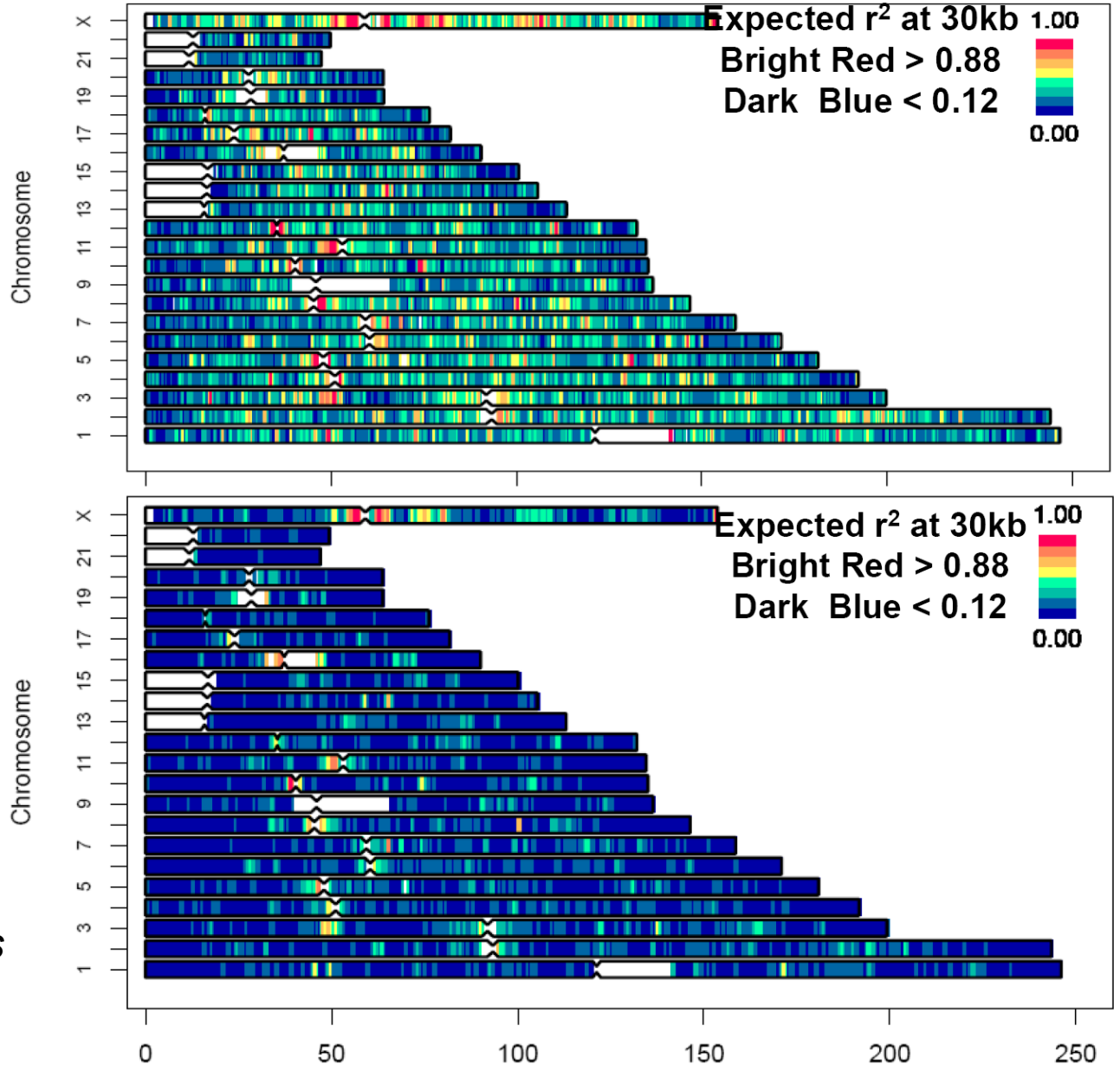
The Extent of LD

- The extent of LD depends on
  - Age of mutation (time to most recent common ancestor)
  - Population history
    - Bottlenecks
    - Rapid expansion
    - Selection
- Implication for association study
  - **Is there enough LD in the population to detect associations?**

Genomic Variation in LD Extent

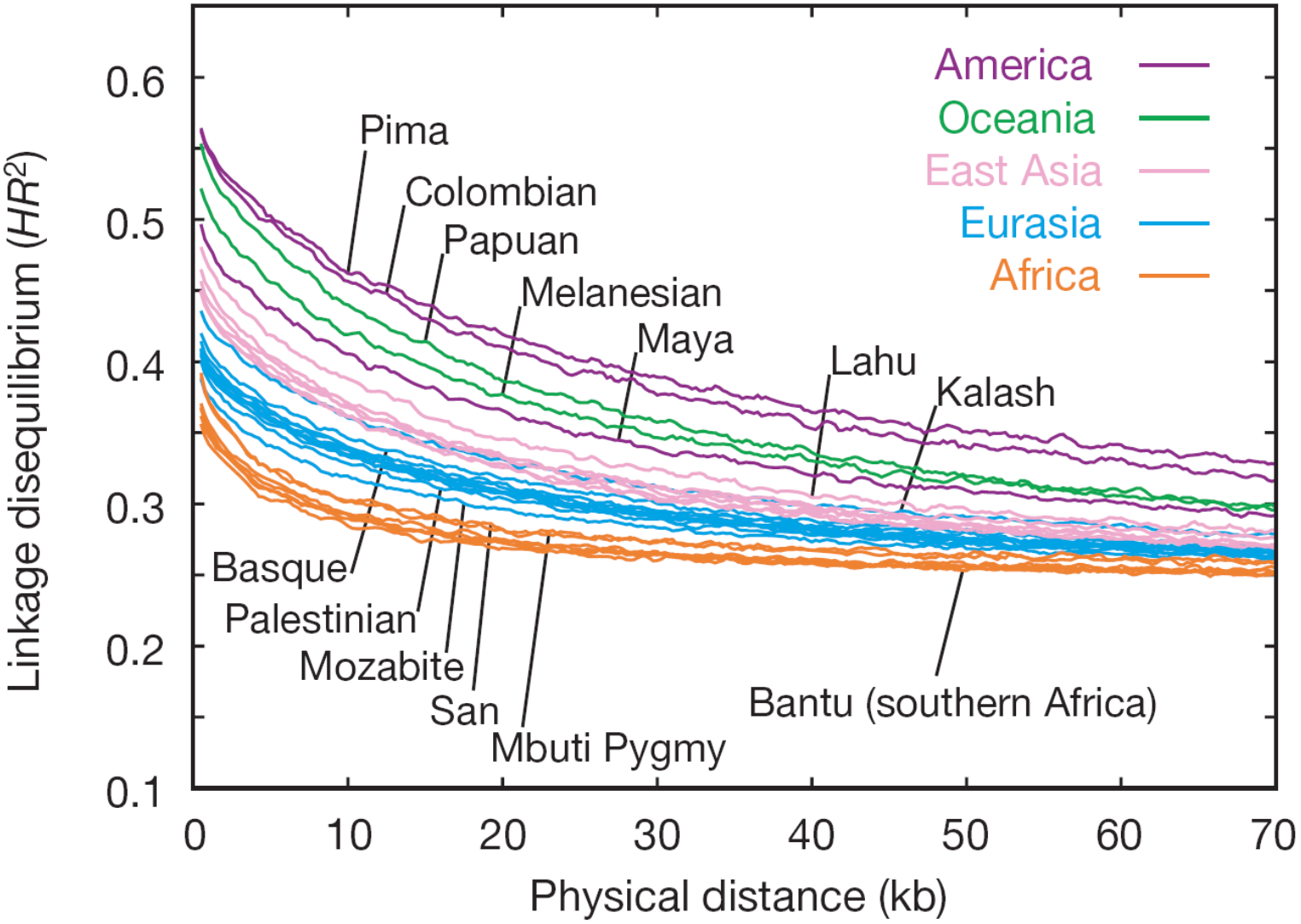
CEU

YRI



Smith (2005)
Genome Res
15, 1519

LD Extent in 29 World Populations



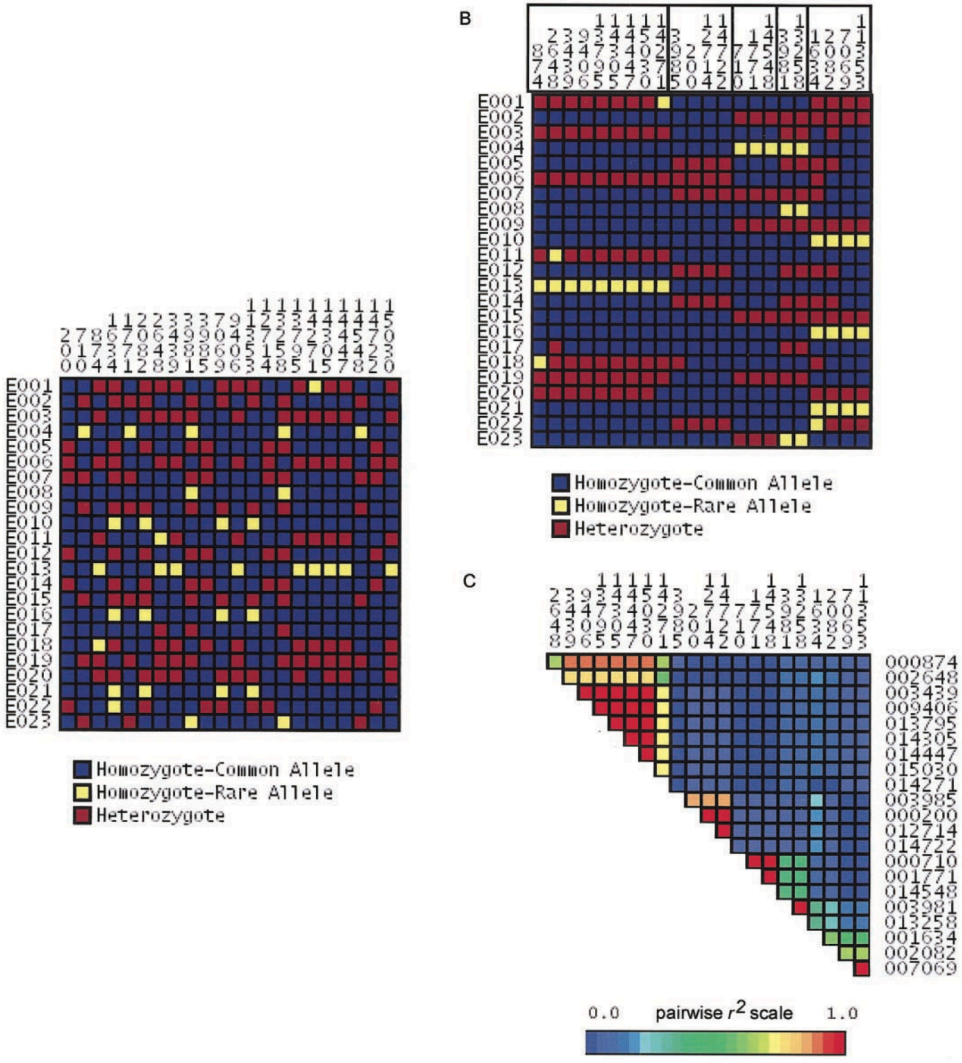
Jakobsson (2008) *Nature* **451,** 998

SNP Tagging

- *Goal:* To find minimal set of SNPs that capture the variation in all SNPs in a region
  - Reduce multiple testing
  - Reduce cost
- Simple pairwise method
  - Choose set of SNPs such that every SNP of interest is in strong LD (say, $r^2 > 0.8$) with at least one tag SNP
  - "Greedy" algorithm*: Sort SNPs into bins based on LD threshold, largest first, choose tag SNP within each bin
  - Can extend to pairs of SNPs (haplotype tag)
  - Implemented in HaploView and Tagger**
- Many other SNP tagging methods exist, largely based on LD blocks

*Carlson CS (2004) *Am J Hum Genet* **74,** 106
de Bakker PIW (2005) *Nat Genet* **37, 1217

SNP Tagging



Carlson CS (2004) *Am J Hum Genet* **74,** 106

Coverage of Commercial Arrays

Table 1 Global coverage (%) by SNP chips ($r^2 \geq 0.8$)

	SNP chip	*CEU*	*CHB+JPT*	*YRI*
Affymetrix	SNP Array 5.0	64	66	41
Affymetrix	SNP Array 6.0	83	84	62
Illumina	HumanHap300	77	66	29
Illumina	HumanHap550	87	83	50
Illumina	HumanHap650Y	87	84	60
Illumina	Human1M	93	92	68

Li et al. (2008) *Eur J Hum Genet* **16,** 635
HapMap II; MAF ≥ 0.05

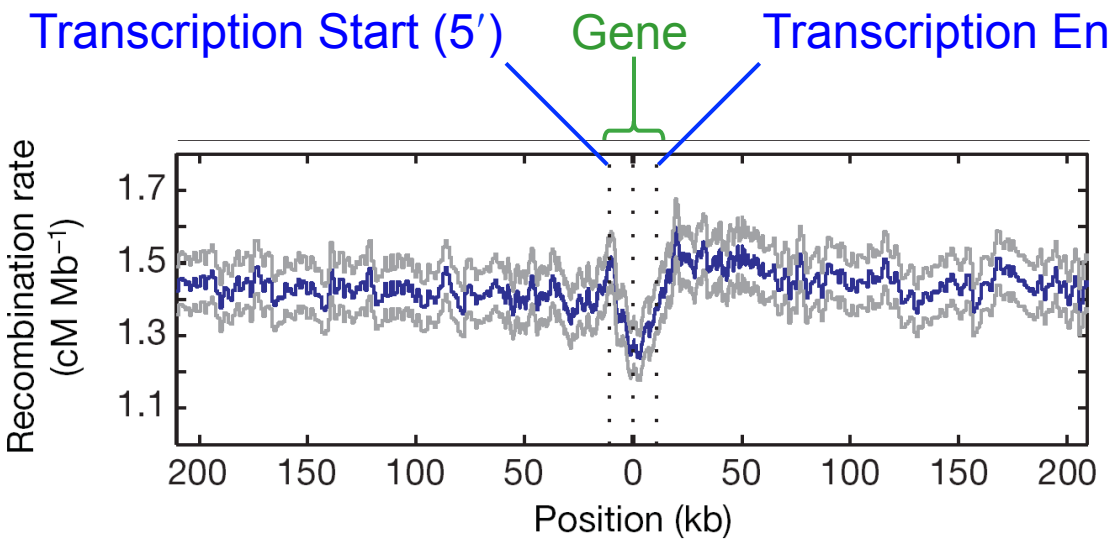
Coverage of 1000 Genomes at $r^2 \geq 0.8$

	CEU		**CHB + JPT**		**YRI**	
MAF:	*> 5%*	*> 1%*	*> 5%*	*> 1%*	*> 5%*	*> 1%*
Omni 2.5	0.83	0.73	0.83	0.73	0.65	0.51
Omni 5	0.87	0.83	0.85	0.76	0.71	0.58

Illumina product documentation

Component V: "Direct" Pheno-Geno Association

Recombination Rates near Genes



The LD and tag trick will work!

Int'l HapMap Consortium (2007) *Nature* **449,** 851

Component V: "Direct" Pheno-Geno Association

Sample Size

- Let A be a typed SNP and B be a functional variant in LD with A with true r^2 of ρ^2_{AB}
- Let N_A and N_B be the sample size required to detect association between the phenotype and A or B (respectively). Then

$$N_B = \frac{N_A}{\rho^2_{AB}}$$

  - Implicitly Assuming that $r^2_{BC} = \rho^2_{AB} r^2_{AC}$,
    where r^2_{AC} and r^2_{BC} are r^2 between markers and the phenotype C

Pritchard & Przeworski (2001) *AJHG* **69**, 1
Terwilliger JD & Hiekkalinna T (2006) *EJHG* **14,** 426
Bochdanovits Z et al (2008) *EJHG* **16,** 525

Component I: Phenotype and Study Design

Table 1. Study Designs Used in Genome-wide Association Studies

	Case-Control	Cohort	Trio
Assumptions	Case and control participants are drawn from the same population Case participants are representative of all cases of the disease, or limitations on diagnostic specificity and representativeness are clearly specified Genomic and epidemiologic data are collected similarly in cases and controls Differences in allele frequencies relate to the outcome of interest rather than differences in background population between cases and controls	Participants under study are more representative of the population from which they are drawn Diseases and traits are ascertained similarly in individuals with and without the gene variant	Disease-related alleles are transmitted in excess of 50% to affected offspring from heterozygous parents
Advantages	Short time frame Large numbers of case and control participants can be assembled Optimal epidemiologic design for studying rare diseases	Cases are incident (developing during observation) and free of survival bias Direct measure of risk Fewer biases than case-control studies Continuum of health-related measures available in population samples not selected for presence of disease	Controls for population structure; immune to population stratification Allows checks for Mendelian inheritance patterns in genotyping quality control Logistically simpler for studies of children's conditions Does not require phenotyping of parents
Disadvantages	Prone to a number of biases including population stratification Cases are usually prevalent cases, may exclude fatal or short episodes, or mild or silent cases Overestimate relative risk for common diseases	Large sample size needed for genotyping if incidence is low Expensive and lengthy follow-up Existing consent may be insufficient for GWA genotyping or data sharing Requires variation in trait being studied Poorly suited for studying rare diseases	May be difficult to assemble both parents and offspring, especially in disorders with older ages of onset Highly sensitive to genotyping error

Pearson, T. A. et al. JAMA 2008;299:1335-1344

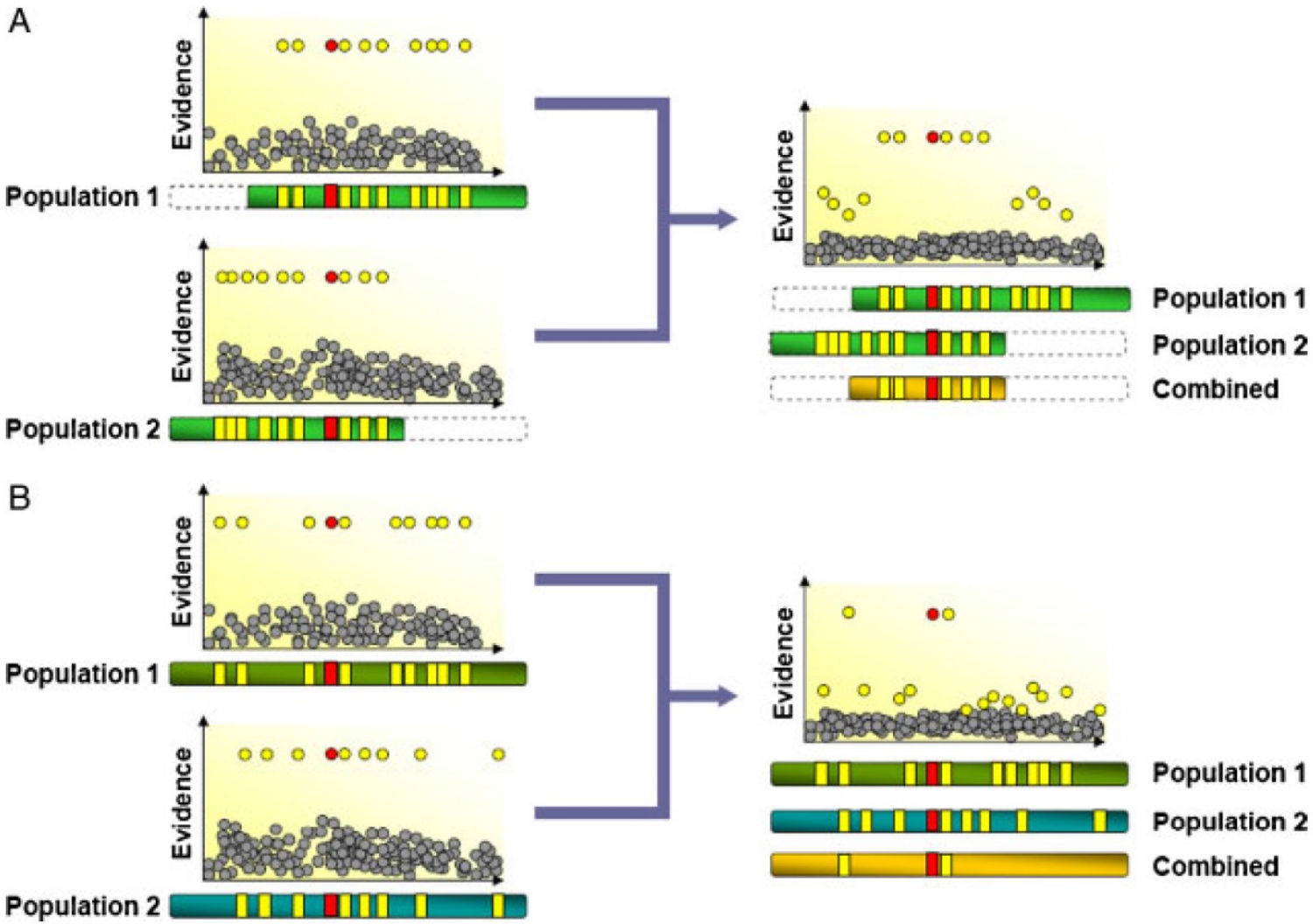

© Francis Collins, 2005


© Francis Collins, 2005

Multiethnic Studies

- Because LD structures are different in widely diverged populations, an association that replicates across populations will be *very precisely located*
  - Replication in a different population can improve power if a causal variant exists in both populations
  - Transethnic analysis can fine map and localize the causal variant.

LD Structure in Two Populations



Teo Y-Y et al. (2010) *Genet Epidemiol* **34,** 653

Multiethnic GWAS for Local Replication

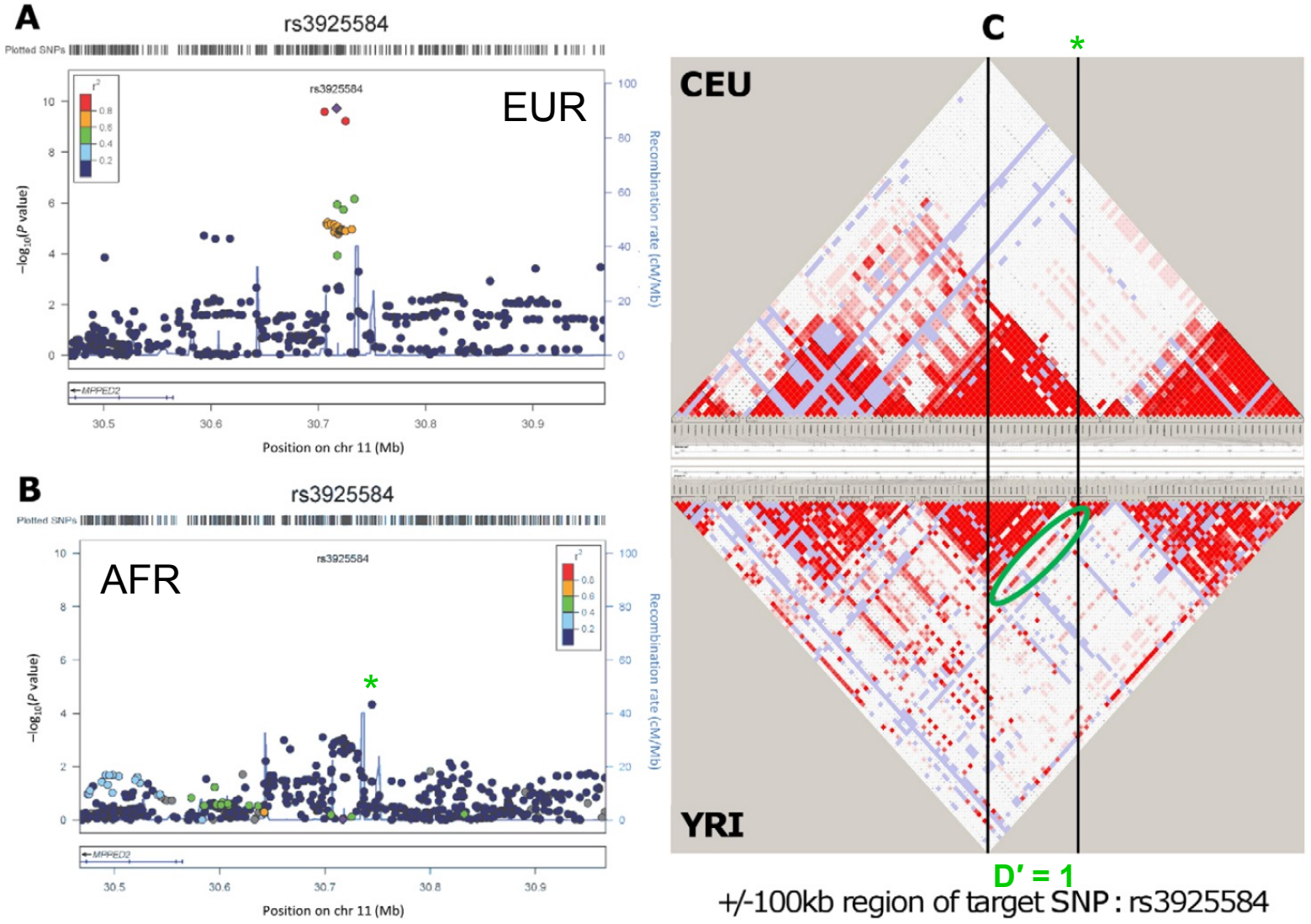


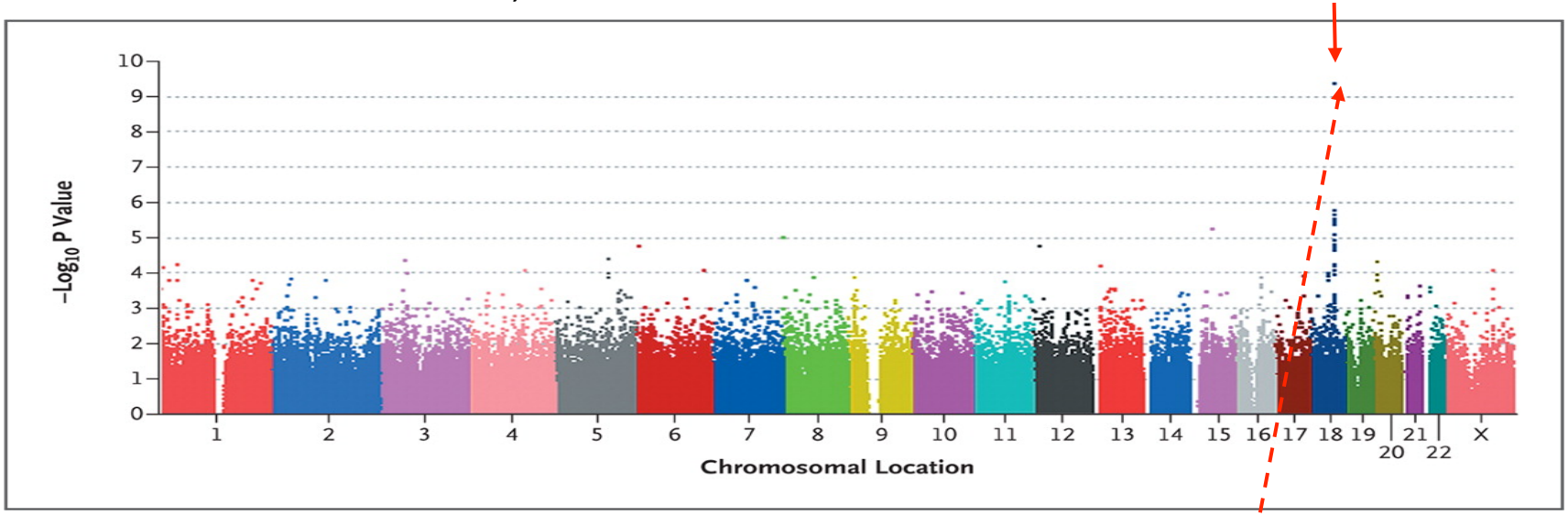
Figure 1. Genetic association and LD distribution of the *MPPED2* gene locus in European and African ancestry populations. Regional

Pattaro C et al. (2012) *PLoS Genet* **8,** e1002584

Who Causes Fuchs's Corneal Dystrophy

Baratz KH et al. NEJM 2010;363:1016-1024

rs613872 in TCF4



A Common Trinucleotide Repeat Expansion within the Transcription Factor 4 (*TCF4*, E2-2) Gene Predicts Fuchs Corneal Dystrophy

Eric D. Wieben[1], Ross A. Aleff[1], Nirubol Tosakulwong[2], Malinda L. Butz[3], W. Edward Highsmith[3], Albert O. Edwards[4], Keith H. Baratz[5]*

Who Is the Criminal?

Association and Familial Segregation of CTG18.1 Trinucleotide Repeat Expansion of *TCF4* Gene in Fuchs' Endothelial Corneal Dystrophy

IOVS | January 2014 | Vol. 55 | No. 1

V. Vinod Mootha,[1,2] Xin Gong,[1] Hung-Chih Ku,[2] and Chao Xing[2]

Table 2. Demographic Information, and *TCF4* CTG18.1 and rs613872 Genotyping Results of Caucasian FECD Cases and Controls

Characteristic	Cases, $n = 120$	Controls, $n = 100$	*P* Value*
Men/women	41/79	40/60	4.0×10^{-1}
Age ± SD, y	70.6 ± 10.7	67.3 ± 11.6	3.1×10^{-2}
rs613872			
GG	17	3	3.1×10^{-17}
GT	78	18	
TT	25	79	
CTG18.1†			
XX	7	1	6.5×10^{-25}
SX	81	6	
SS	32	93	
Haplotype			
G-X	0.382	0.040	5.9×10^{-19}
T-X	0.014	0.000	
G-S	0.085	0.080	
T-S	0.519	0.880	

indicated their association with the disease status.[38] The LD measured by r^2 between the two loci was estimated to be 0.65 and 0.31 in cases and controls, respectively, and LD measured by D' was estimated to be 0.93 and 1.00 in cases and controls,

Caucasians

Who Is the Criminal?

Association of *TCF4* Gene Polymorphisms with Fuchs' Corneal Dystrophy in the Chinese

Anbupalam Thalamuthu,[1,2] *Chiea Chuen Khor,*[2,3,4] *Divya Venkataraman,*[5] *Li Wei Koh,*[5] *Donald T. H. Tan,*[5,6,7] *Tin Aung,*[5,6,7] *Jodhbir S. Mehta,*[5,6,7,8] *and Eranga N. Vithana*[5,7]

Investigative Ophthalmology & Visual Science, July 2011, Vol. 52, No. 8

Weinberg equilibrium. Single-locus analysis conducted on the 18 SNPs within *TCF4* revealed four of them to be nonpolymorphic in Chinese. This included the most significantly associated SNP from the original report (rs613872).[22] After accounting for

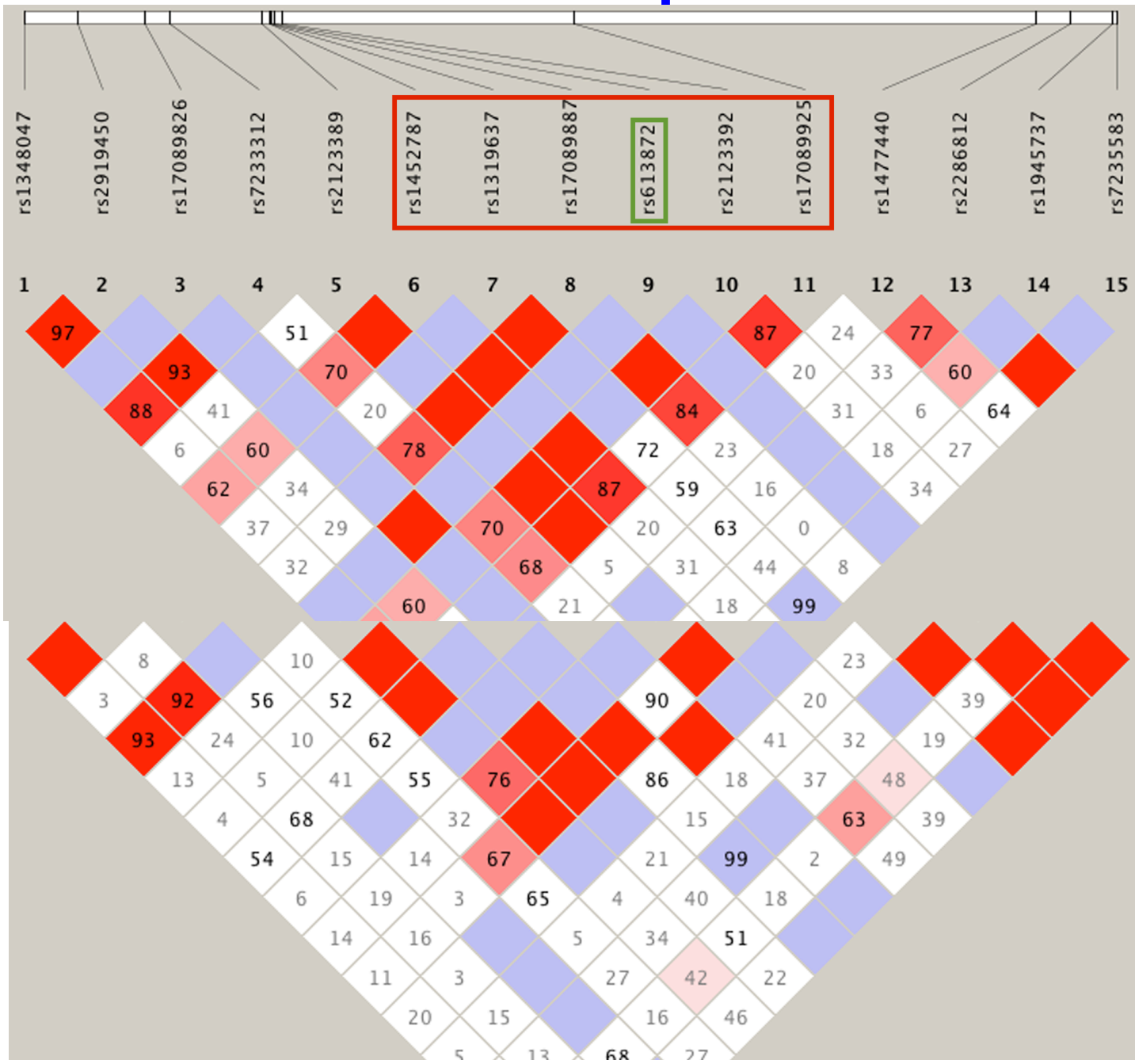
!

Population	MAF
EUR	0.177
EAS	0.004
AFR	0.005
AMR	0.084
SAS	0.104

Transethnic Replication

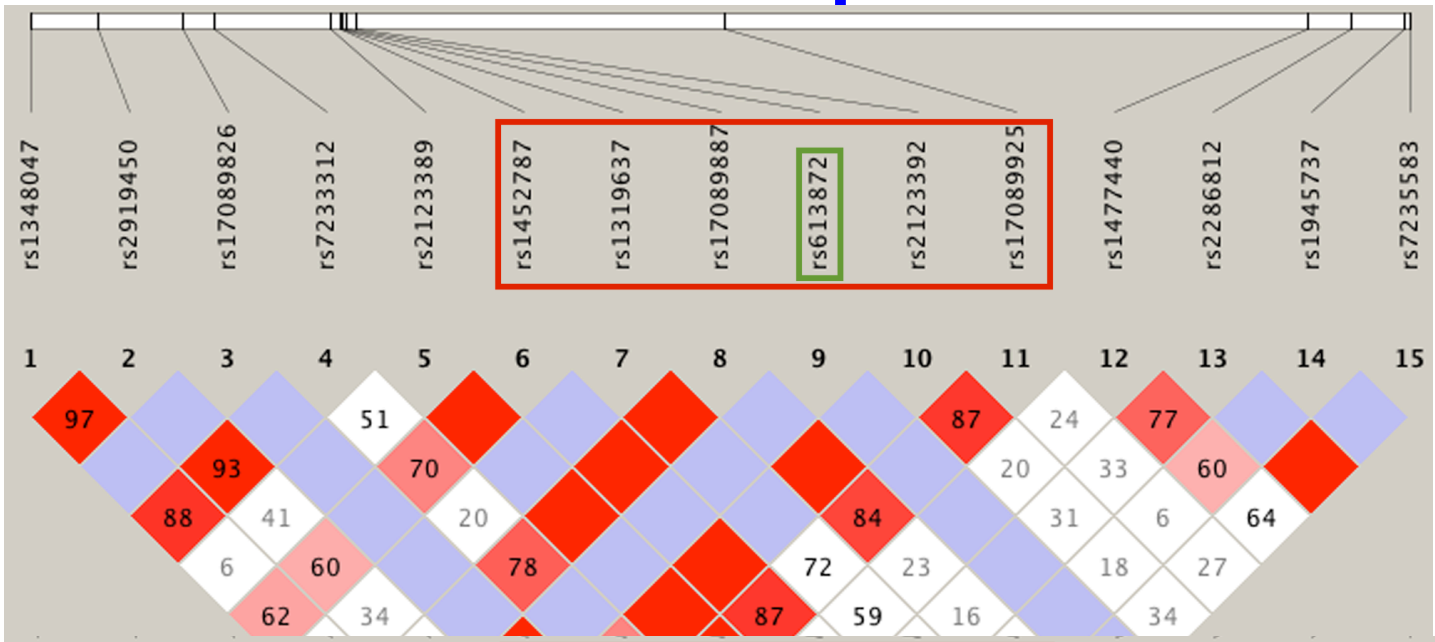
1KG CHB

1KG CEU

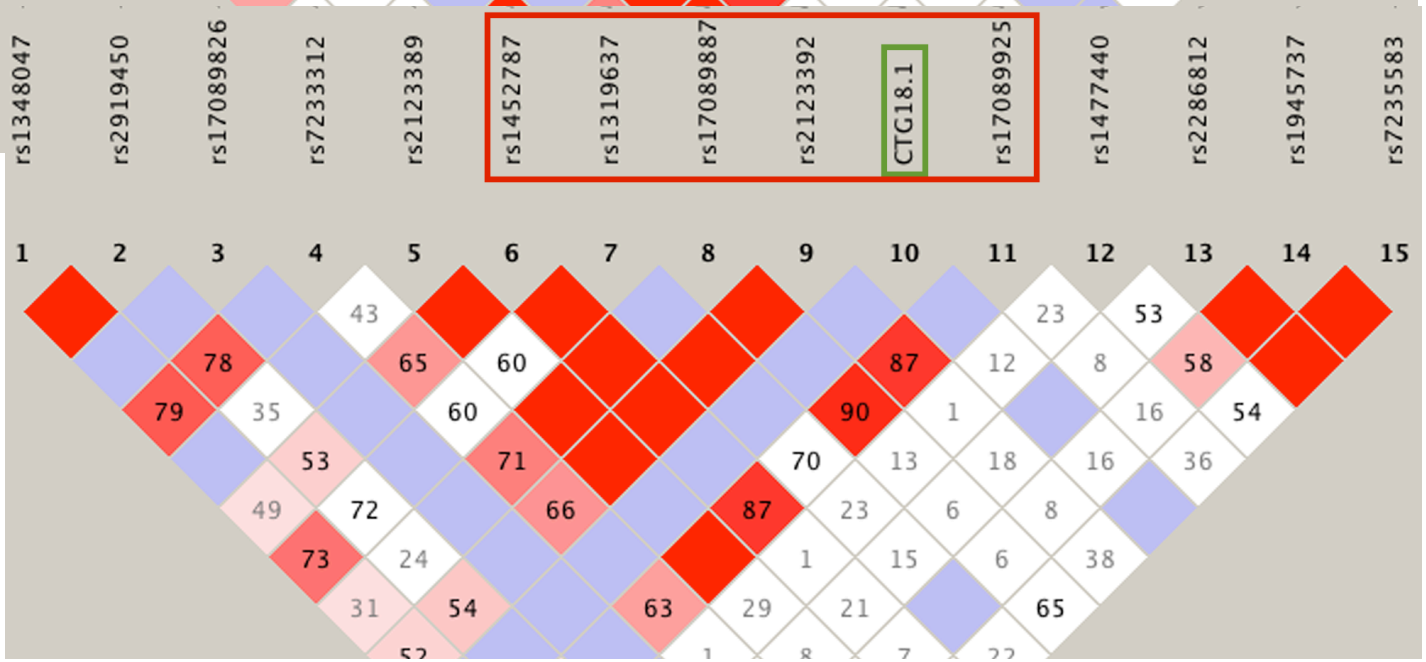


Transethnic Replication

1KG CHB



Singapore Controls



Transethnic Replication

Transethnic Replication of Association of CTG18.1 Repeat Expansion of *TCF4* Gene With Fuchs' Corneal Dystrophy in Chinese Implies Common Causal Variant

Chao Xing,[1] Xin Gong,[2] Imran Hussain,[2] Chiea-Chuen Khor,[3] Donald T. H. Tan,[4–6] Tin Aung,[4–6] Jodhbir S. Mehta,[4–7] Eranga N. Vithana,[4,6,8] and V. Vinod Mootha[1,2]

TABLE 1. Demographic Information and *TCF4* CTG18.1 Genotype Distribution* in Chinese FECD Cases and Controls

Characteristic	Cases, $n = 57$	Controls, $n = 121$	*P* Value
CTG18.1*			
XX	3	0	4.7×10^{-14}†
SX	22	2	
SS	32	119	

TABLE 3. Haplotype Association* of *TCF4* Polymorphisms With FECD in a Chinese Population

					Haplotype Frequencies			
rs1452787	rs17089887	rs2123392	CTG18.1	rs17089925	Cases	Control	*P* Value	Global *P* Value
A	C	T	S	T	0.394	0.387	-	
G	T	C	S	C	0.272	0.375	2.5×10^{-1}	
A	T	T	S	C	0.050	0.148	1.7×10^{-1}	1.5×10^{-9}
A	C	T	X	T	0.211	0.009	2.1×10^{-5}	
A	C	T	S	C	0.038	0.055	5.7×10^{-1}	

Linkage Format

Family structure

FID	ID	P1	P2	Sex	Affection	Marker1a	Marker1b	Marker2a	Marker2b
IBD054	430	0	0	1	0	1	3	3	1
IBD054	412	430	431	2	2	1	3	1	3
IBD054	431	0	0	2	0	3	3	3	3
IBD058	438	0	0	1	0	3	3	3	3
IBD058	470	438	444	2	2	3	3	3	3
IBD058	444	0	0	2	0	3	3	3	3

FDC001	FDC001	0	0	1	2	G	T	C	A
FDC002	FDC002	0	0	2	2	G	T	C	A
FDC003	FDC003	0	0	2	2	G	G	C	C
FDC004	FDC004	0	0	2	2	G	T	C	A
FDC005	FDC005	0	0	2	2	G	T	C	A
FDC006	FDC006	0	0	2	2	G	G	C	C

A Course on Statistic Genetics Is Coming

An Introduction to Statistical Genetics

- A 8-weeks course with systemic and comprehensive introduction of statistical genetics from concept to practice.

Instructors:
Julia Kozlitina, Ph.D.
Chao Xing, Ph.D.