Genetic Variation and Genome-Wide Association Studies

Keyan Salari, MD/PhD Candidate
Department of Genetics

How many of you did the readings before class?

A. Yes, of course!

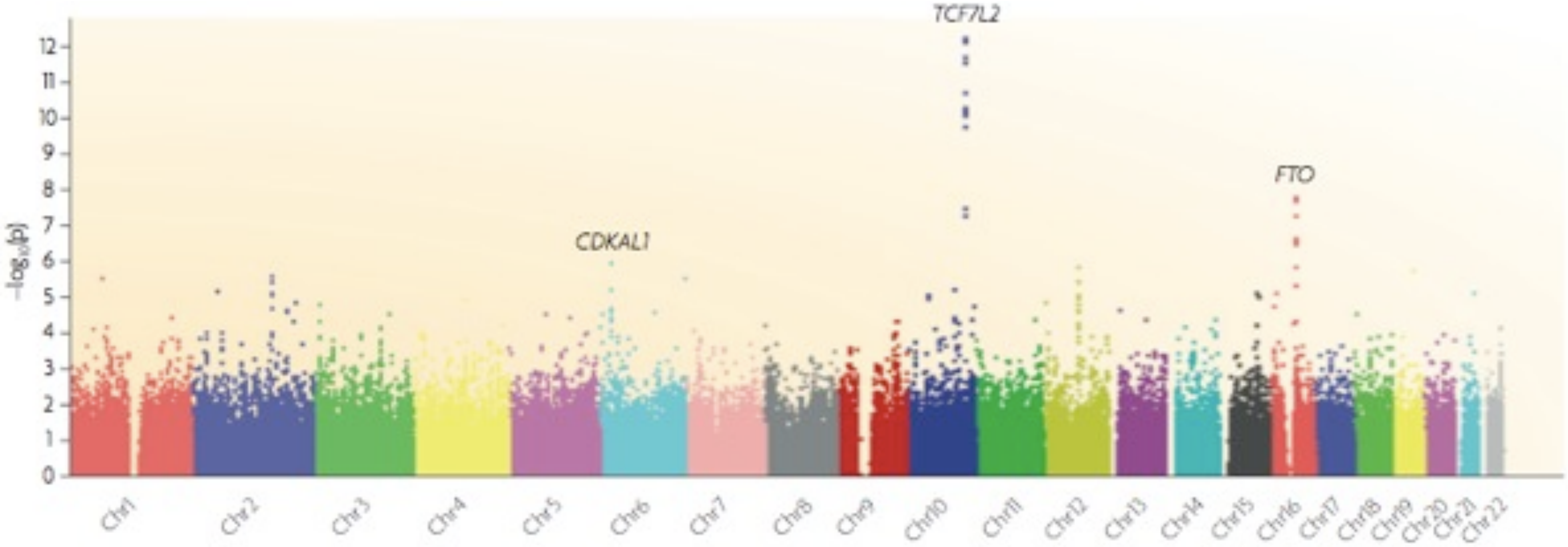
B. Started, but didn't get through them all

C. No, BBQs and fireworks ruled the weekend

1. Natural variation in the human genome

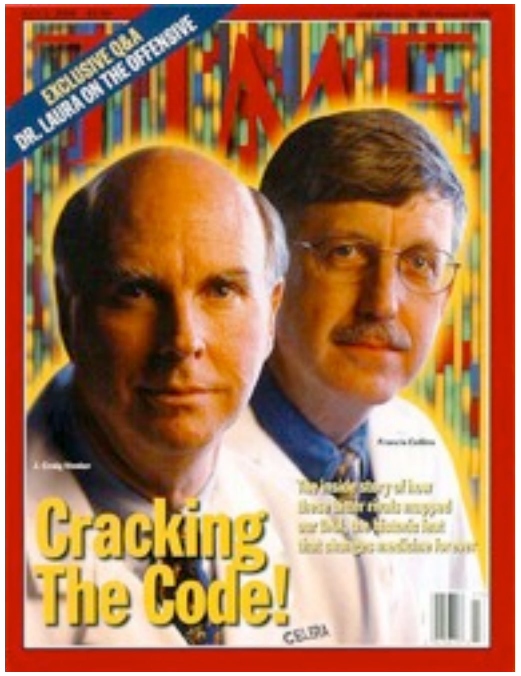
2. Genetic Association & Linkage Disequilibrium

3. Genome-wide association studies



Human Genetic Diversity

2000



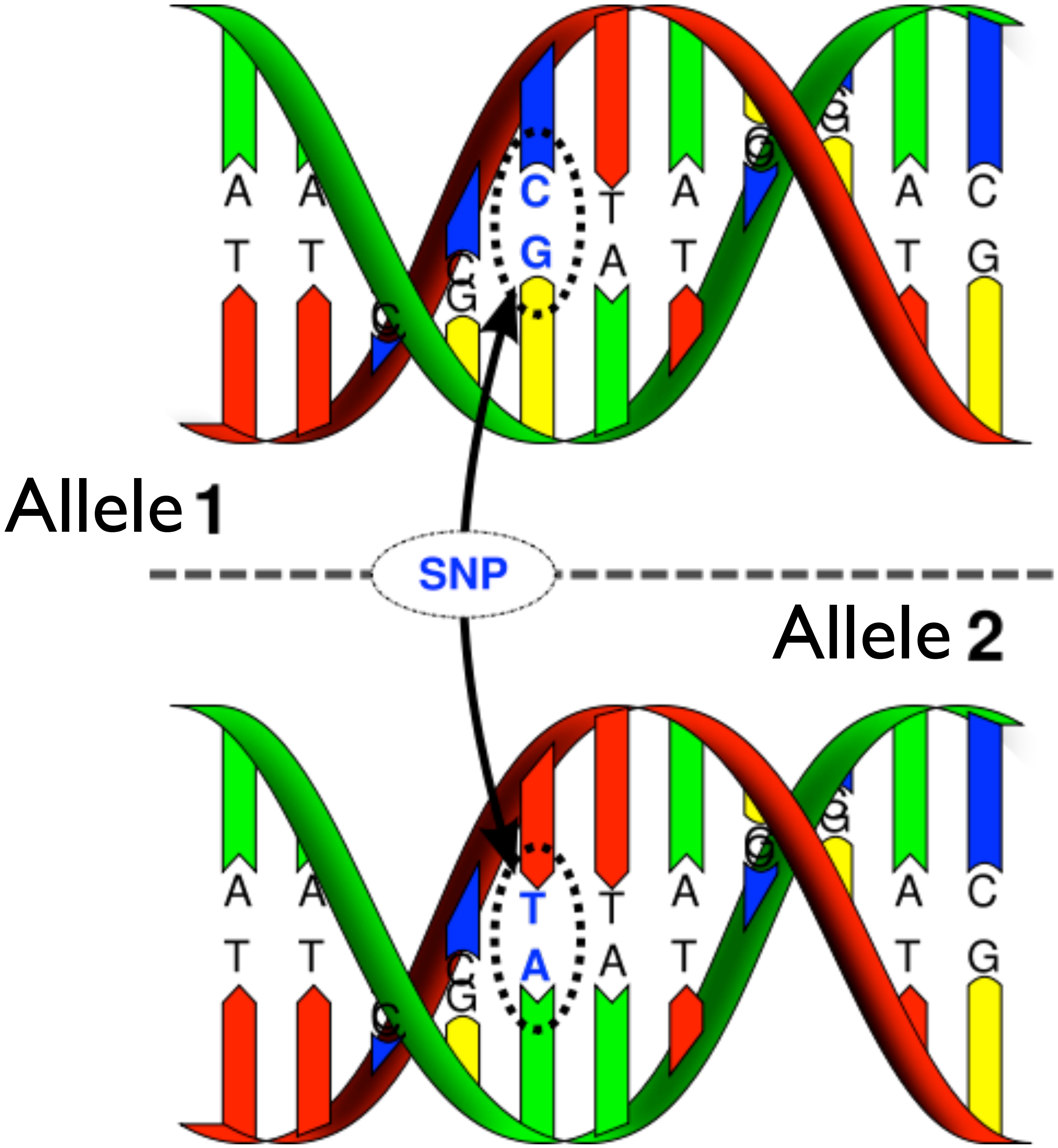
0.1% x 3.3 billion
= 3,300,000 bp of
differences

"I believe one of the great truths to emerge from this triumphant expedition inside the human genome is that in genetic terms, all human beings, regardless of race, are more than 99.9 percent the same."
President Bill Clinton, June 26, 2000, The White House East Room

Human Genetic Variation

- Differences or variations in the DNA sequences between 2 individual's genomes are infrequent
- At sites of variation, each different form or variant is called an allele
  - Common allele = major allele = wild-type allele
  - Variant allele = minor allele = mutant allele
- DNA differences where minor allele occurs $< 1\%$ of population are called mutations
- DNA differences where minor allele occurs $\geq 1\%$ of population are called polymorphisms

Human Genetic Variation



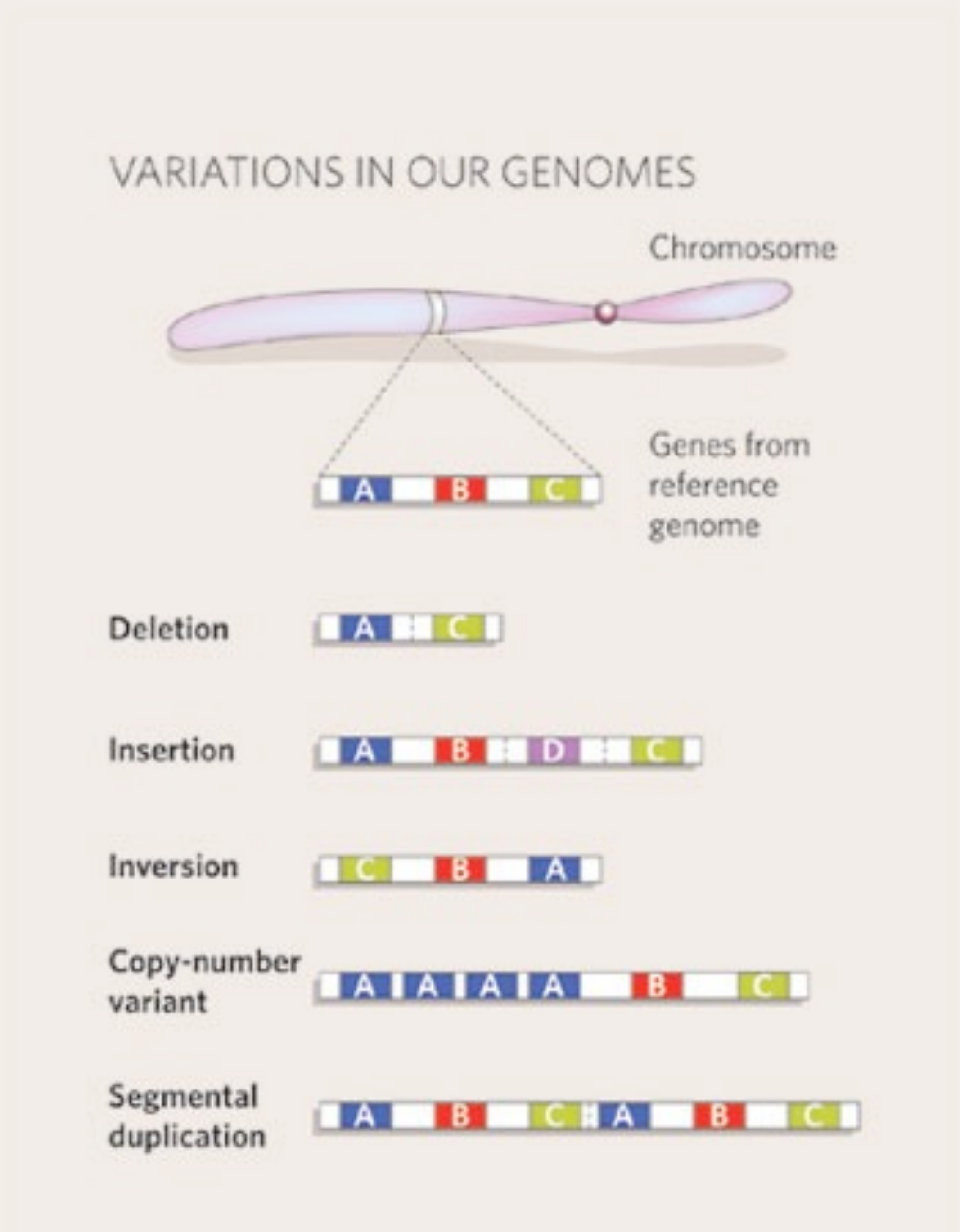
Single-nucleotide polymorphism (SNP)

Some are in parts of genes that are translated

- non-synonymous SNPs lead to a change in amino acid sequence of resultant protein
- synonymous SNPs do not result in amino acid change

Other SNPs are intergenic and may influence cell function through other means

Human Genetic Variation



Nature, 2005

Structural variation

- 12% of our genome
- thousands of genes, disease loci, functional elements
- likely role in phenotypic variation and human disease

Redon *et al. Nature*, 2006

Human Genetic Variation

- Since we inherit 2 versions of each chromosome - one from mom, one from dad - we have 2 alleles of every polymorphism
- For a given polymorphism with 2 alleles (A and a), possible genotypes are:
  - Homozygous major allele, A/A
  - Heterozygous, A/a
  - Homozygous minor allele, a/a
- Possible models of inheritance: dominant, recessive, additive

dbSNP

www.ncbi.nlm.nih.gov/projects/SNP/

- Database of SNPs
- Build 130 (4/30/09) contained 6.5 million validated SNPs
- Build 131 (3/25/10) contained >12 million validated SNPs

International HapMap Project

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

- Multi-country effort to identify and catalog genetic similarities and differences in human beings
- Initial populations:
  - CEU - Utah residents with ancestry from northern and western Europe
  - CHB - Han Chinese in Beijing, China
  - JPT - Japanese in Tokyo, Japan
  - YRI - Yoruba in Ibadan, Nigeria

Genetic variation for a simple trait

Chr12:ALDH2 - SNP rs671

```
...GGGCTGCAGGCATACACTGAAGTGAAAACTGTGAGTGTG
...GGGCTGCAGGCATACACTGAAGTGAAAACTGTGAGTGTG
...  G  L  Q  A  Y  T  E  V  K  T  V  S  V
```

Genotype: G/G

Protein: functional

Phenotype: none

Chr12:ALDH2 - SNP rs671

```
... GGGCTGCAGGCATACACTGAAGTGAAAACTGTGAGTGTG
... GGGCTGCAGGCATACACTAAAGTGAAAACTGTGAGTGTG
...  G  L  Q  A  Y  T E/K V  K  T  V  S  V
```

Genotype: A/G

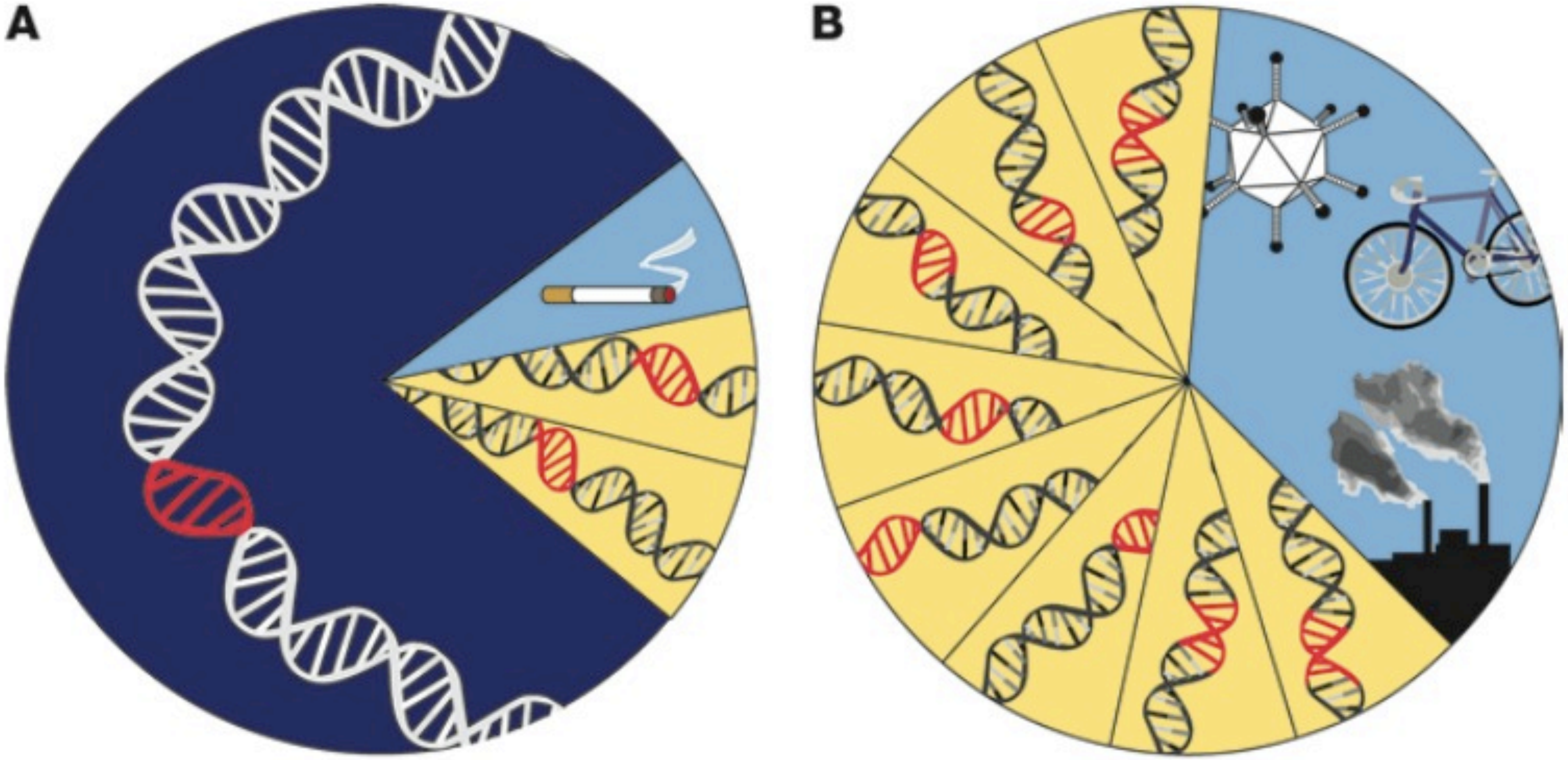
Protein: 1/2 functional

Phenotype: alcohol flush reaction

G allele functional
A allele missense (null)

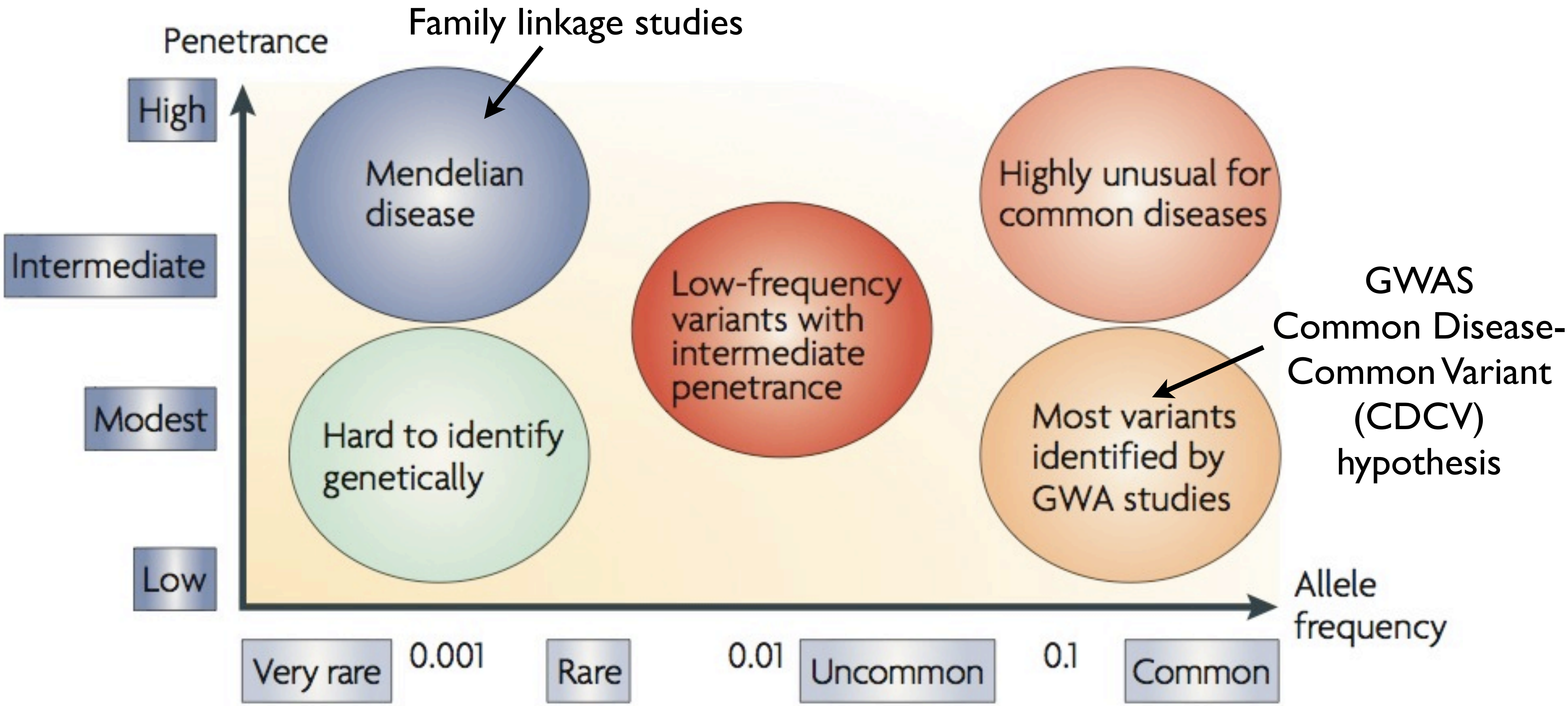
CEU 100% G
YRI 100% G
CHB/JPT 76-84% G

Mendelian vs Complex Traits



Manolio *et al. JCI*, 2008

Mendelian vs Complex Traits



Penetrance
High
Intermediate
Modest
Low
Mendelian disease
Hard to identify genetically
Low-frequency variants with intermediate penetrance
Highly unusual for common diseases
Most variants identified by GWA studies
Allele frequency
Very rare
0.001
Rare
0.01
Uncommon
0.1
Common

Family linkage studies

GWAS
Common Disease-
Common Variant
(CDCV)
hypothesis

McCarthy *et al. Nat Rev Genet*, 2008

Genetic epidemiology

- In a traditional epidemiological study, variation in an exposure is linked to an outcome (e.g., smoking and lung cancer, or cholesterol and myocardial infarction)
- In a genetic association study, variation in a gene is linked to an outcome

Genetic epidemiology

- Epidemiological studies ideally elucidate causality between a risk factor and an outcome
- Establishing causality requires isolating the effect of a given factor from other related or correlated factors (e.g., age + sun exposure ~ skin cancer)
- We use "adjusted" or multivariate analysis

Genetic epidemiology

- Establishing causality in a genetic association study would require isolating the function of a particular SNP from other correlated SNPs that may be nearby in the gene
- Because groups of alleles at neighboring genes or SNPs tend to be inherited together as a unit (called a haplotype), it becomes difficult to attribute causality
- Most genetic association studies identify SNPs *associated with* or correlated with the outcome

Why is using one’s genotype at rs2383207 to assess risk of MI problematic?

“The SNP is only correlated with MI, not causally implicated”

A. TRUE

B. FALSE

Linkage disequilibrium (LD)

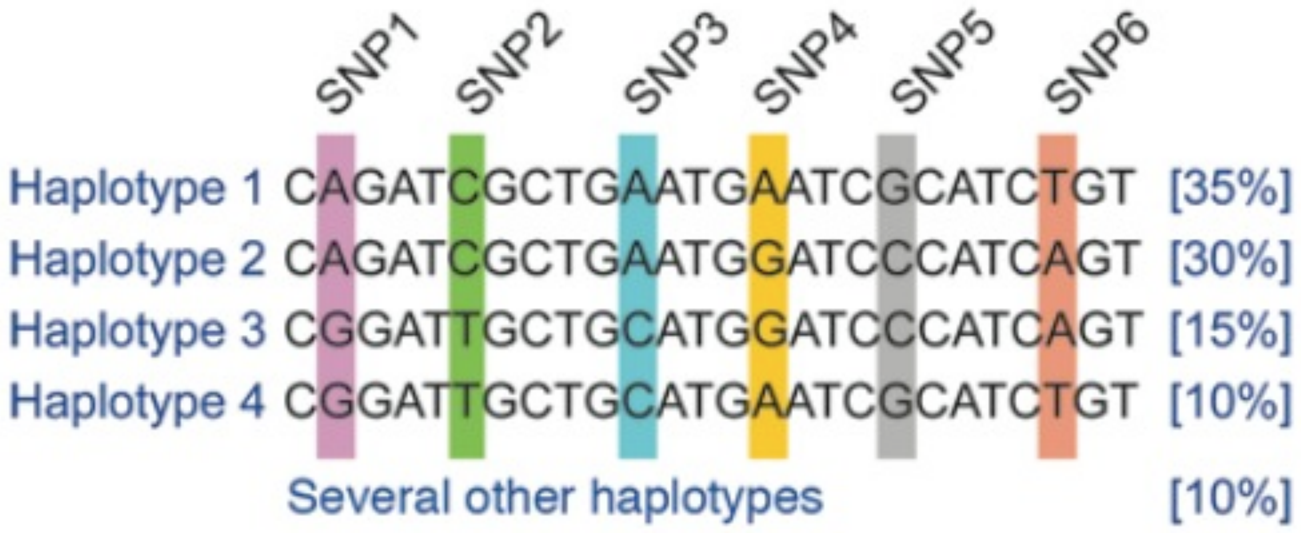
- Association between 2 alleles located near each other on a chromosome, such that they are inherited together *more frequently* than expected by chance
- Information about the allele of one SNP in an individual is strongly predictive of the allele of the other SNP on that chromosome
- LD persists because meiotic recombination does not occur at random, but is concentrated in hot spots

Linkage disequilibrium (LD)

- Regions that lack hot spots are likely to be in strong LD
- A commonly used measure of LD is r^2, the proportion of variation in one SNP explained by another, or the proportion of observations in which two specific pairs of their alleles occur together

Linkage disequilibrium (LD)

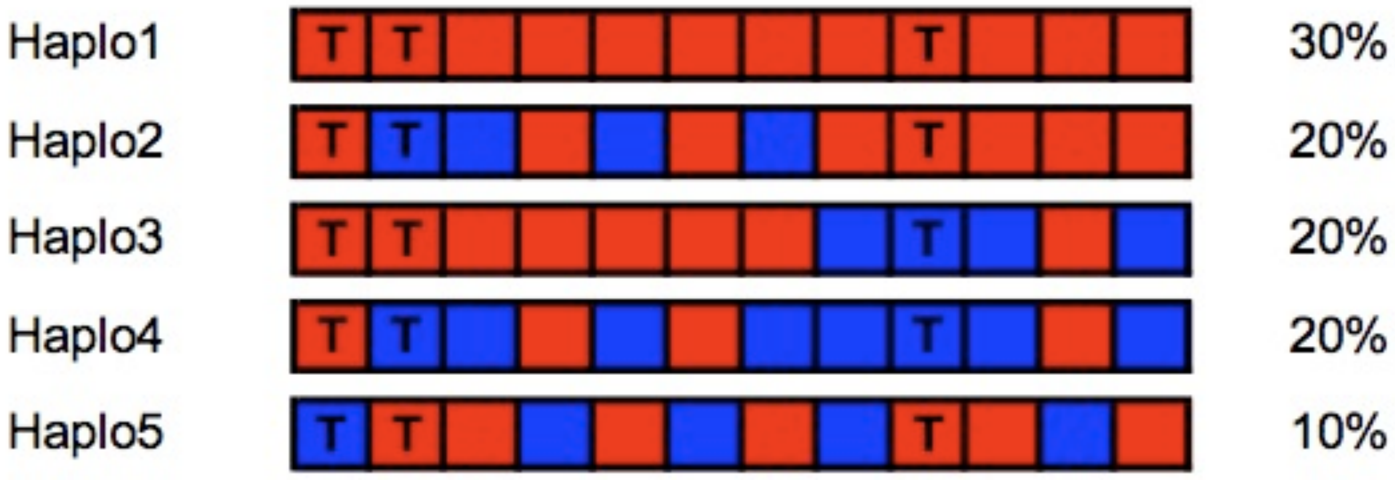
- Two SNPs that are perfectly correlated have an r^2 of 1.0, e.g. allele A of SNP1 is always observed with allele C of SNP2 (and vice versa)
- r^2 of 0 would be interpreted as an observation of of allele A of SNP1 providing no information at all about which allele of SNP4 is present



Haplotype 1 CAGATCGCTGAATGAATCGCATCTGT [35%]
Haplotype 2 CAGATCGCTGAATGGATCCCATCAGT [30%]
Haplotype 3 CGGATTGCTGCATGGATCCCATCAGT [15%]
Haplotype 4 CGGATTGCTGCATGAATCGCATCTGT [10%]
Several other haplotypes [10%]

tag SNPs

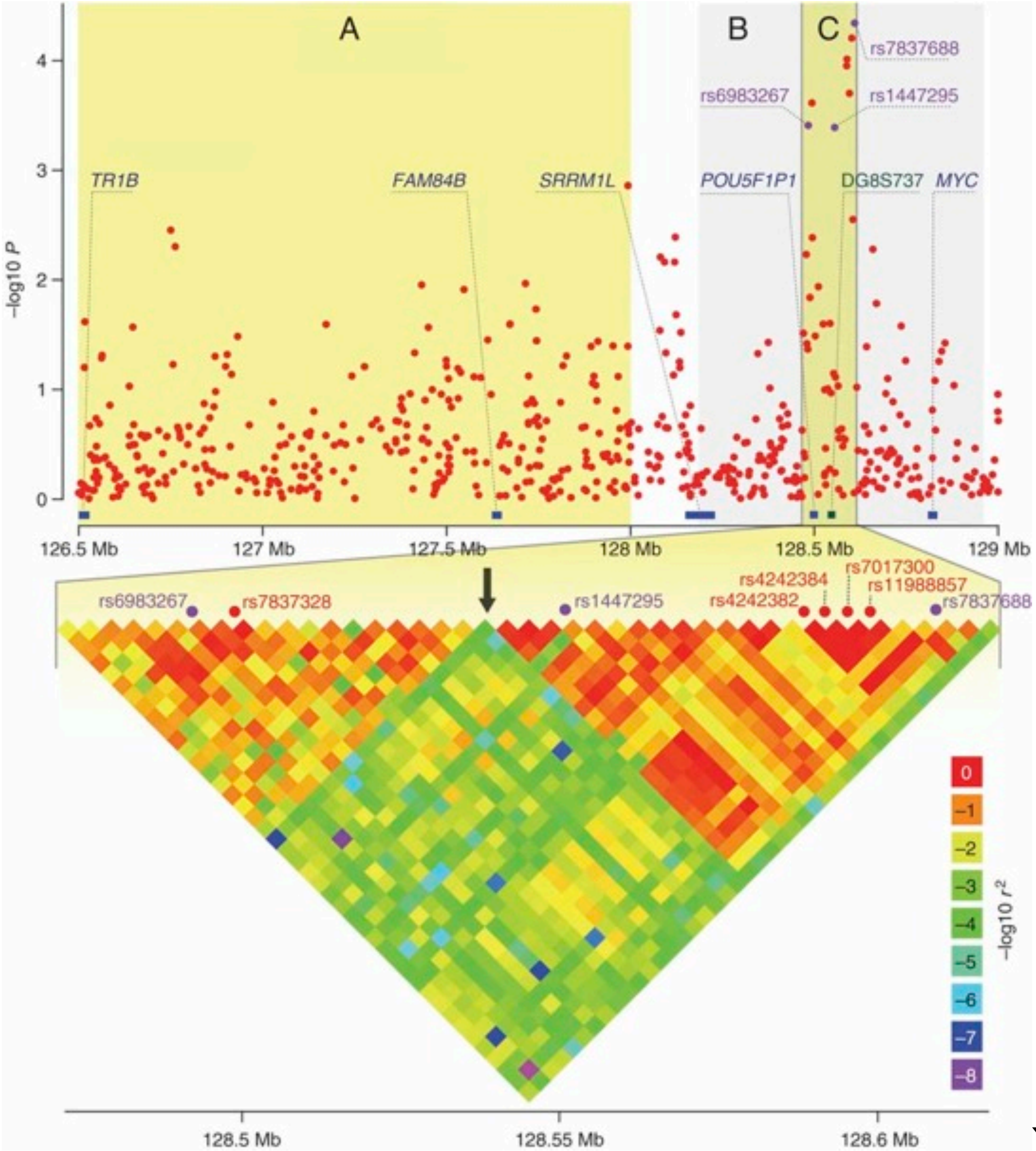
- *n* biallelic SNPs could generate 2^n haplotypes in theory
- Because humans are a relatively "young" species, and due to non-random recombination, far fewer combinations make up bulk of haplotypes in population
- Thus, a few carefully selected SNPs (called tag SNPs need to be genotyped to predict variants at rest of SNPs in each region



Haplo1	🟥T	🟥T	🟥	🟥	🟥	🟥	🟥	🟥	🟥T	🟥	🟥	🟥	30%
Haplo2	🟥T	🟦T	🟦	🟥	🟦	🟥	🟦	🟥	🟥T	🟥	🟥	🟥	20%
Haplo3	🟥T	🟥T	🟥	🟥	🟥	🟥	🟥	🟦	🟦T	🟦	🟥	🟦	20%
Haplo4	🟥T	🟦T	🟦	🟥	🟦	🟥	🟦	🟦	🟦T	🟦	🟥	🟦	20%
Haplo5	🟦T	🟥T	🟥	🟦	🟥	🟦	🟥	🟦	🟥T	🟥	🟦	🟥	10%

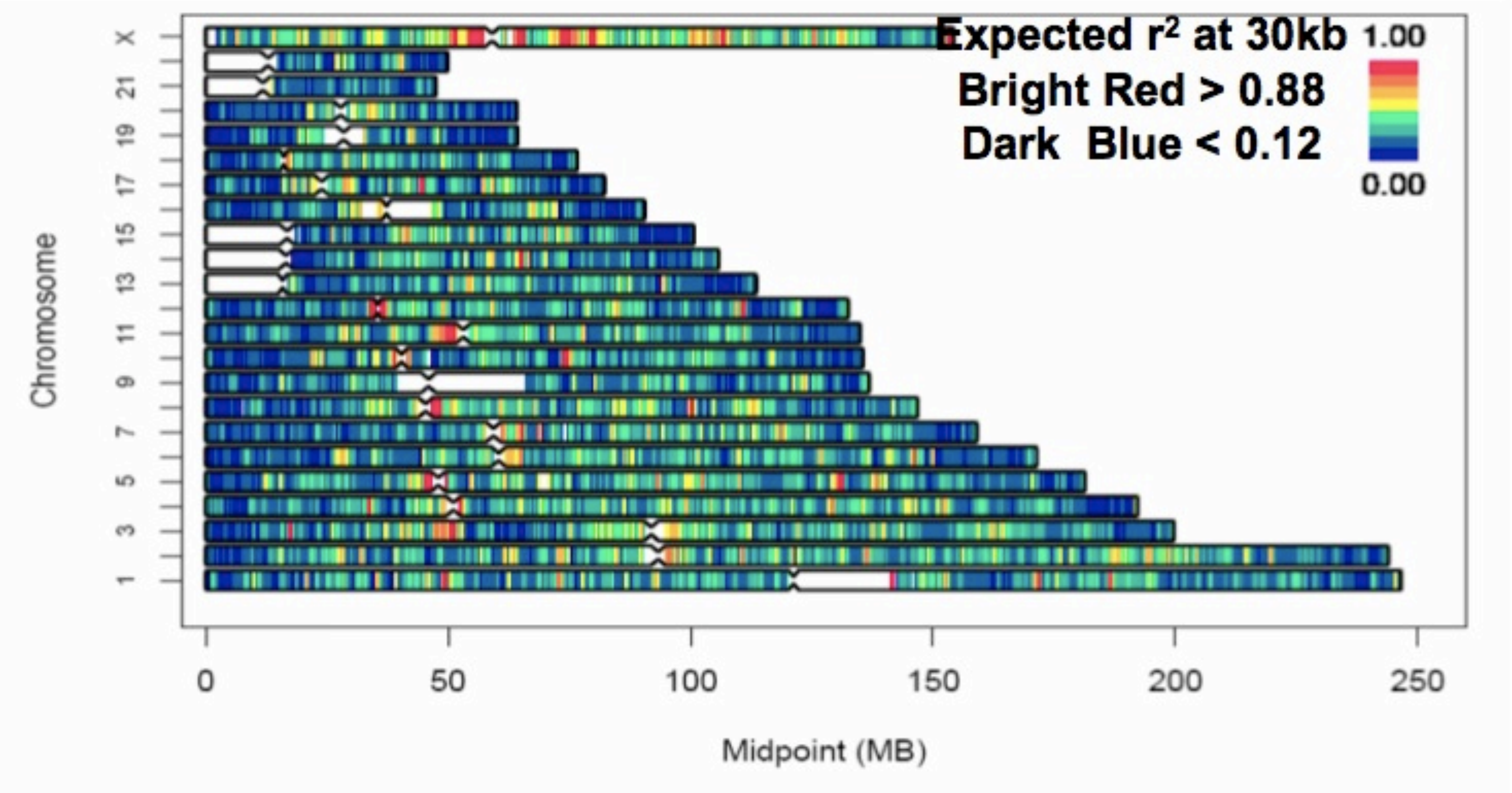
12 biallelic SNPs

8q24 and prostate cancer

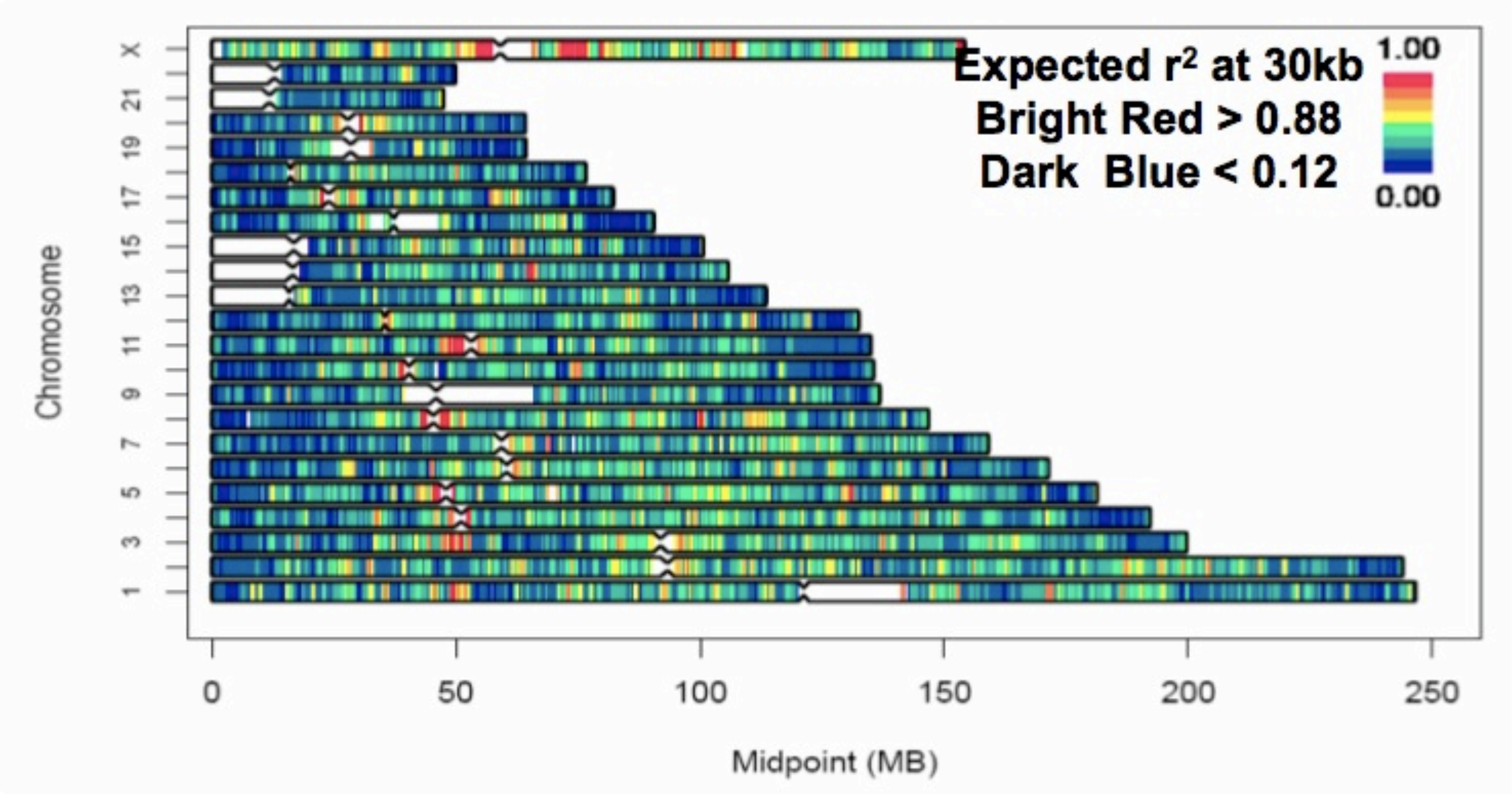


Yeager *et al. Nat Genet*, 2007

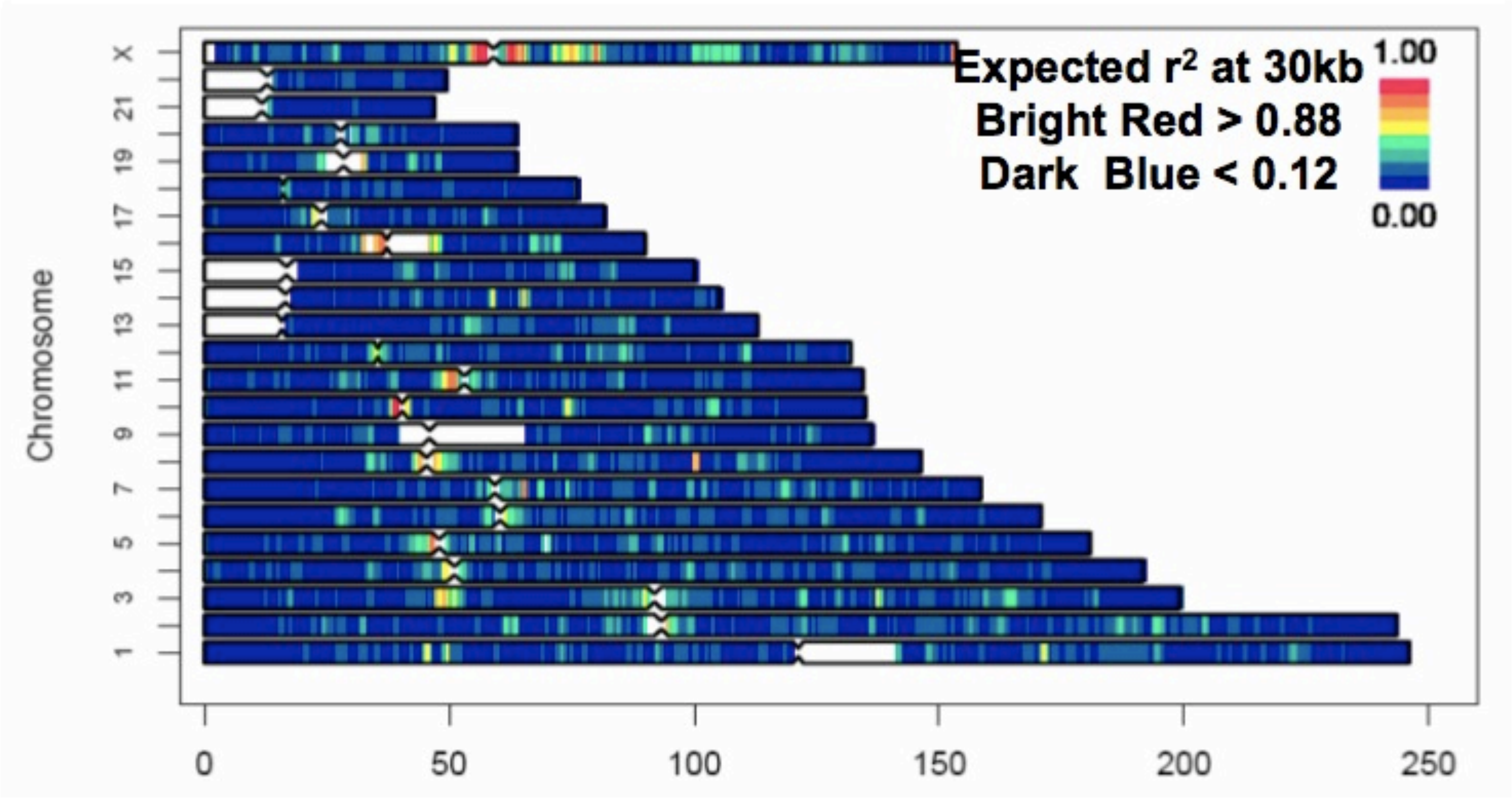
Genome-wide LD structure in Europeans (CEU)



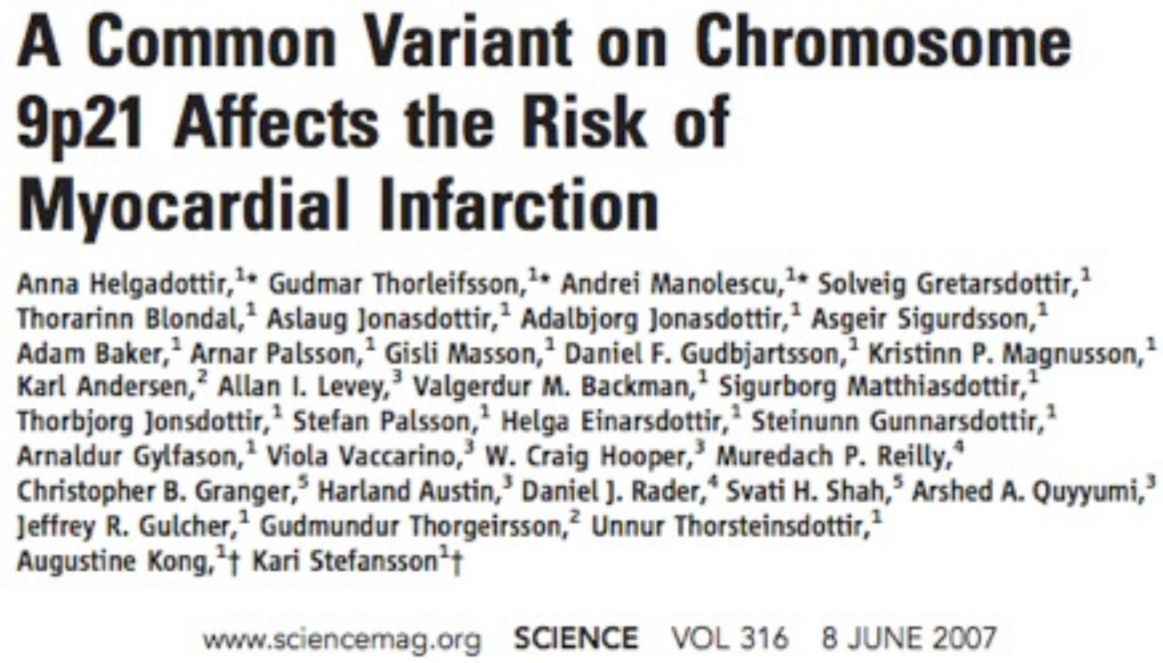
Genome-wide LD structure in Asians (CHB+JPT)



Genome-wide LD structure in Africans (YRI)



LD exercise



A Common Variant on Chromosome 9p21 Affects the Risk of Myocardial Infarction

Anna Helgadottir,[1]* Gudmar Thorleifsson,[1]* Andrei Manolescu,[1]* Solveig Gretarsdottir,[1] Thorarinn Blondal,[1] Aslaug Jonasdottir,[1] Adalbjorg Jonasdottir,[1] Asgeir Sigurdsson,[1] Adam Baker,[1] Arnar Palsson,[1] Gisli Masson,[1] Daniel F. Gudbjartsson,[1] Kristinn P. Magnusson,[1] Karl Andersen,[2] Allan I. Levey,[3] Valgerdur M. Backman,[1] Sigurborg Matthiasdottir,[1] Thorbjorg Jonsdottir,[1] Stefan Palsson,[1] Helga Einarsdottir,[1] Steinunn Gunnarsdottir,[1] Arnaldur Gylfason,[1] Viola Vaccarino,[3] W. Craig Hooper,[3] Muredach P. Reilly,[4] Christopher B. Granger,[5] Harland Austin,[3] Daniel J. Rader,[4] Svati H. Shah,[5] Arshed A. Quyyumi,[3] Jeffrey R. Gulcher,[1] Gudmundur Thorgeirsson,[2] Unnur Thorsteinsdottir,[1] Augustine Kong,[1]† Kari Stefansson[1]†

www.sciencemag.org SCIENCE VOL 316 8 JUNE 2007

"The strongest association with MI was observed with three correlated SNPs - rs1333040, rs2383207, and rs10116277. Each had an odds ratio around 1.22 and P-value approximately 1 x 10-6. All three SNPs are located within a 190-kb LD block on chromosome 9p21."

Using the 2 provided genomes - one European individual and one African individual - impute the genotype of rs1333040 in each individual. rs2383207 and rs10116277 have been measured directly.

LD exercise

What is the genotype of at rs1333040 for the European individual (Patient1)?

A. CC

B. CT

C. TT

What is the genotype of at rs1333040 for the African individual (Patient5)?

A. CC

B. CT

C. TT

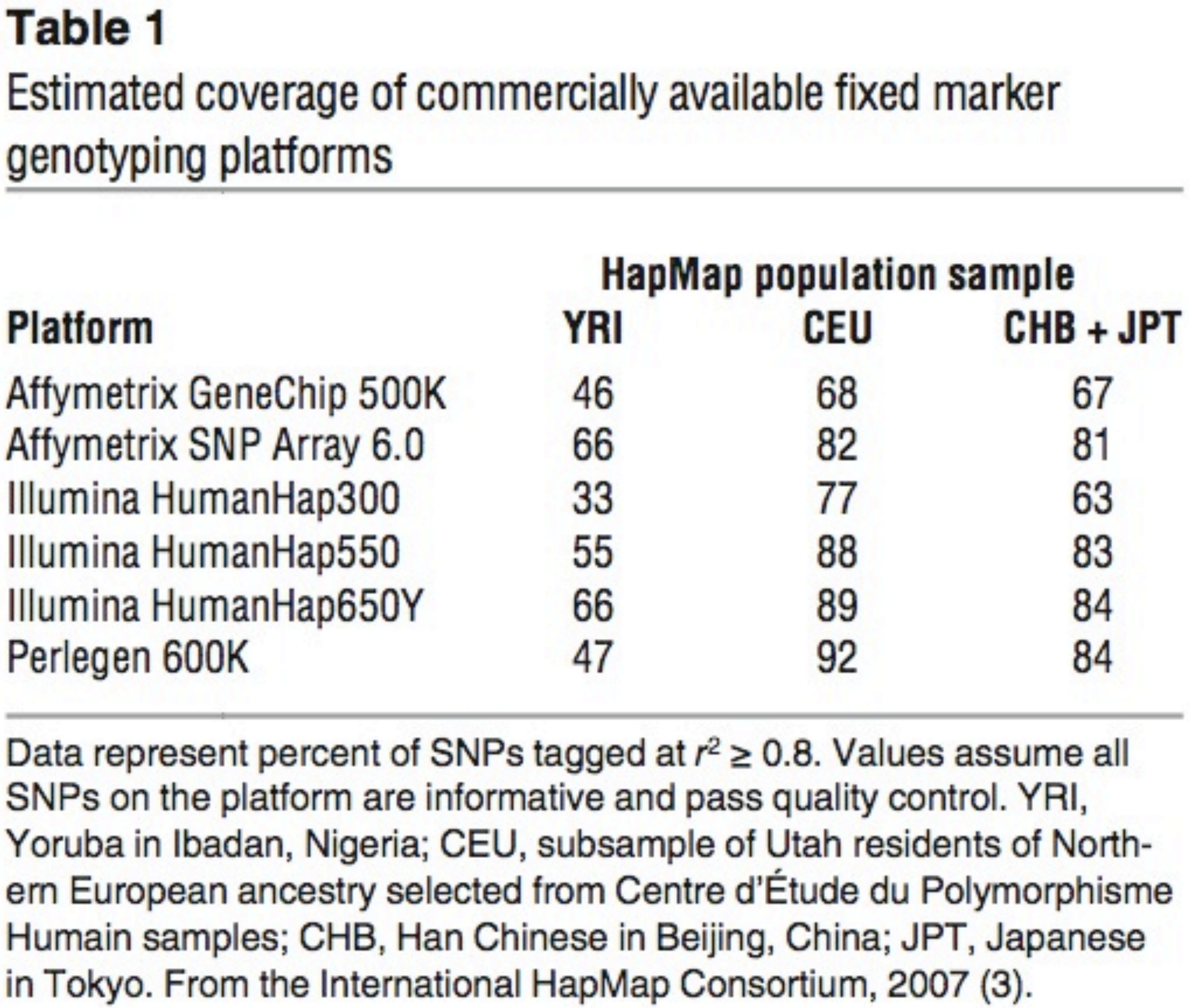


Table 1

Estimated coverage of commercially available fixed marker genotyping platforms

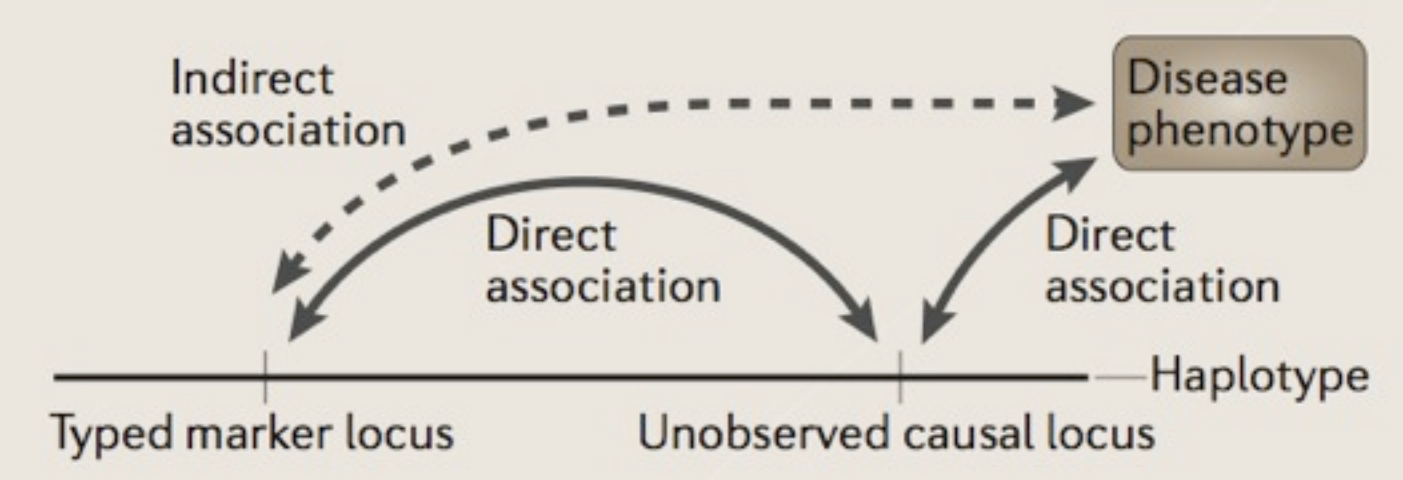
	HapMap population sample		
Platform	**YRI**	**CEU**	**CHB + JPT**
Affymetrix GeneChip 500K	46	68	67
Affymetrix SNP Array 6.0	66	82	81
Illumina HumanHap300	33	77	63
Illumina HumanHap550	55	88	83
Illumina HumanHap650Y	66	89	84
Perlegen 600K	47	92	84

Data represent percent of SNPs tagged at $r^2 \geq 0.8$. Values assume all SNPs on the platform are informative and pass quality control. YRI, Yoruba in Ibadan, Nigeria; CEU, subsample of Utah residents of Northern European ancestry selected from Centre d'Étude du Polymorphisme Humain samples; CHB, Han Chinese in Beijing, China; JPT, Japanese in Tokyo. From the International HapMap Consortium, 2007 (3).

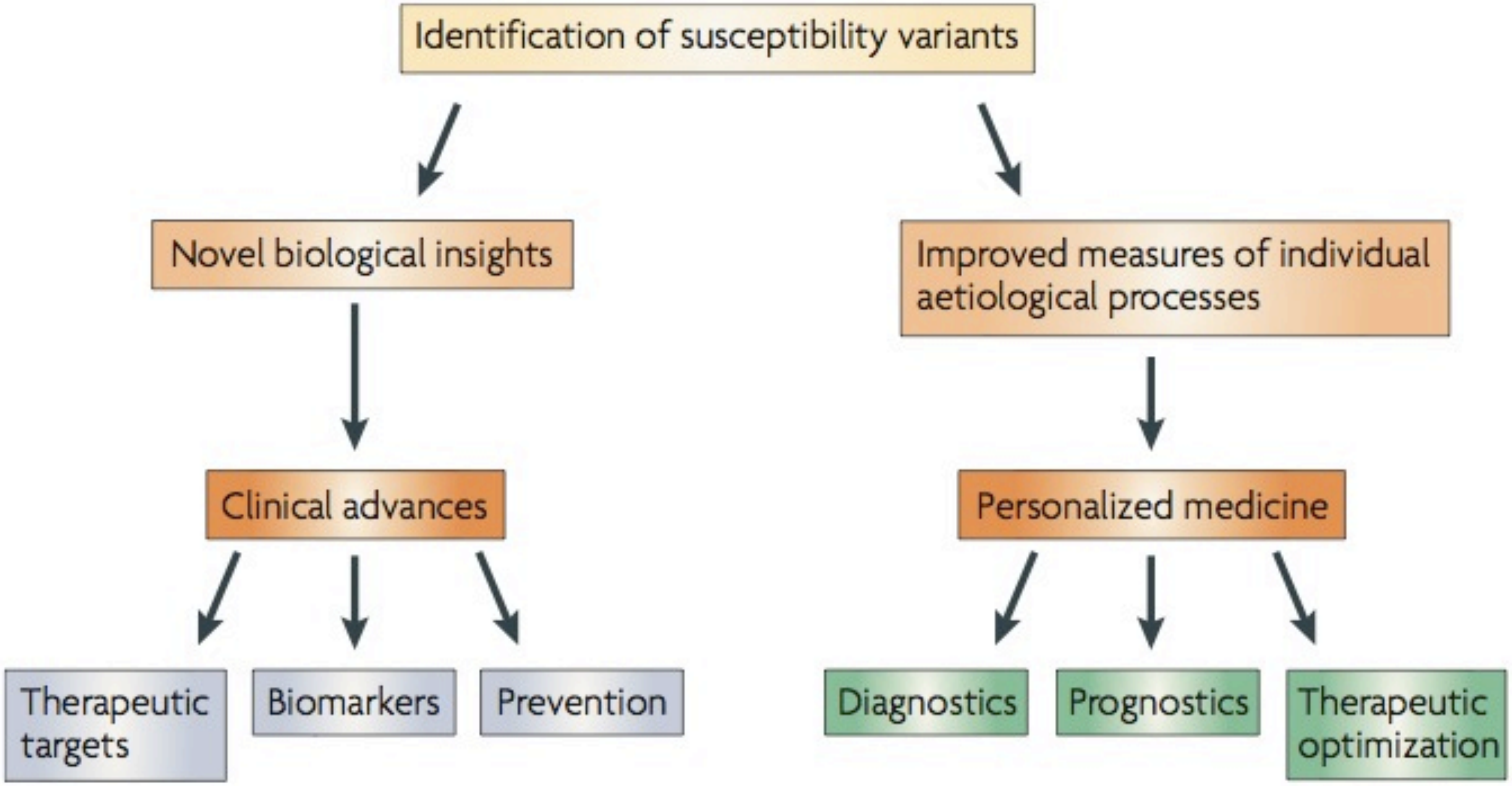
Manolio *et al. JCI*, 2008

LD

- Significant LD exists in the human genome
- Extent and structure of LD varies greatly across the regions of the genome and across different populations
- Association studies rely on LD to tag haplotypes in chromosomal regions
- As such, reported SNP associations are presumed to be *not causative*, but rather in LD with a causative variant (which is OK!)



Genetic association



McCarthy *et al. Nat Rev Genet*, 2008

Genetic association studies

- Objective: identify a genetic variant (SNP) where one allele is observed more often with the phenotype (disease) than the other allele
- Candidate gene association studies are guided by known/postulated biology or previous results
- Alternative is screening entire genome for associations, i.e. genome-wide association study

Study Design

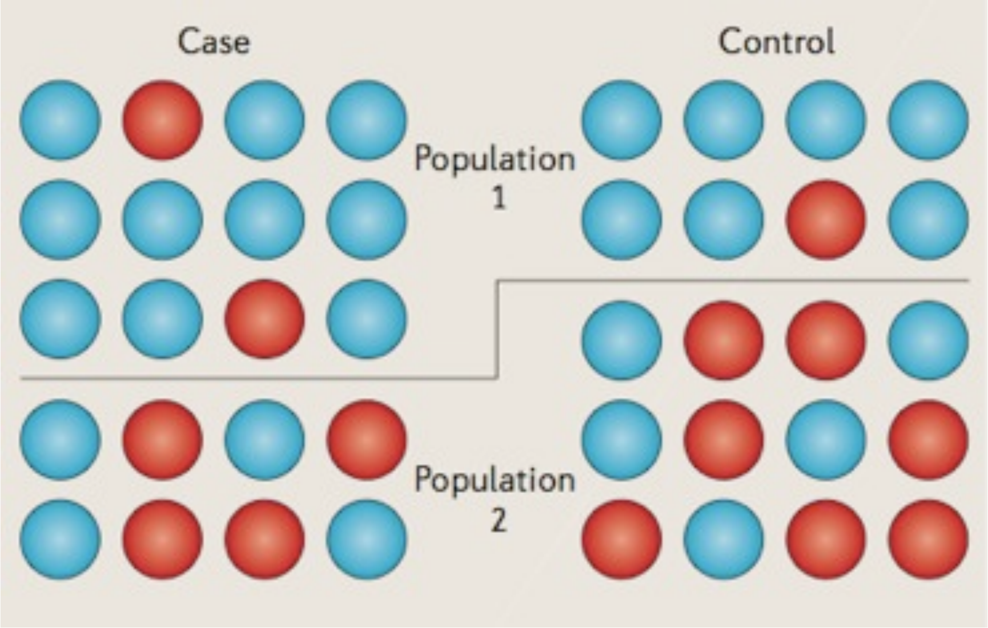
	Case-Control	Cohort
Advantages	• Shorter time frame • Easier to study rare diseases • Large number of cases/controls can be assembled	• Cases are incident • Direct measure of risk • Fewer biases than case-control
Disadvantages	• Prone to biases incl population stratification • Cases are prevalent (not incident) • Overestimates relative risk for common diseases	• Large sample size needed if incidence is low • Expensive/lengthy follow-up • Poorly suited for studying rare diseases

Selecting cases and controls

- Misclassification of case and control participants can drastically reduce the power of a GWAS and bias results toward no association
- Ensure cases are truly affected
- Ensure controls are truly unaffected
- E.g. diabetes - self-report? 2X fasting glucose >125mg/dL? OGTT? pre-diabetics? undiagnosed diabetes? MODY/early-onset patients?

Selecting cases and controls

- Are cases and controls drawn from same population?
- Case-control studies are particularly prone to population stratification, a form of confounding caused by genetic differences between cases and controls unrelated to disease but due to sampling them from populations of different ancestries



Power

- GWA studies to date have identified variants with modest odds ratios or relative risks (1.3-1.5)
- A GWAS needs enough subjects to be sufficiently powered for detecting such modest effect sizes - typically this means 1000s of cases and controls
- Independent population samples are needed for replication
- Initial GWAS have tendency to overestimate effect size (odds ratio) - this is called the "winner's curse"

Genotyping

- Genotyping platform should be sufficiently dense to capture a large proportion of the variation in the population studied

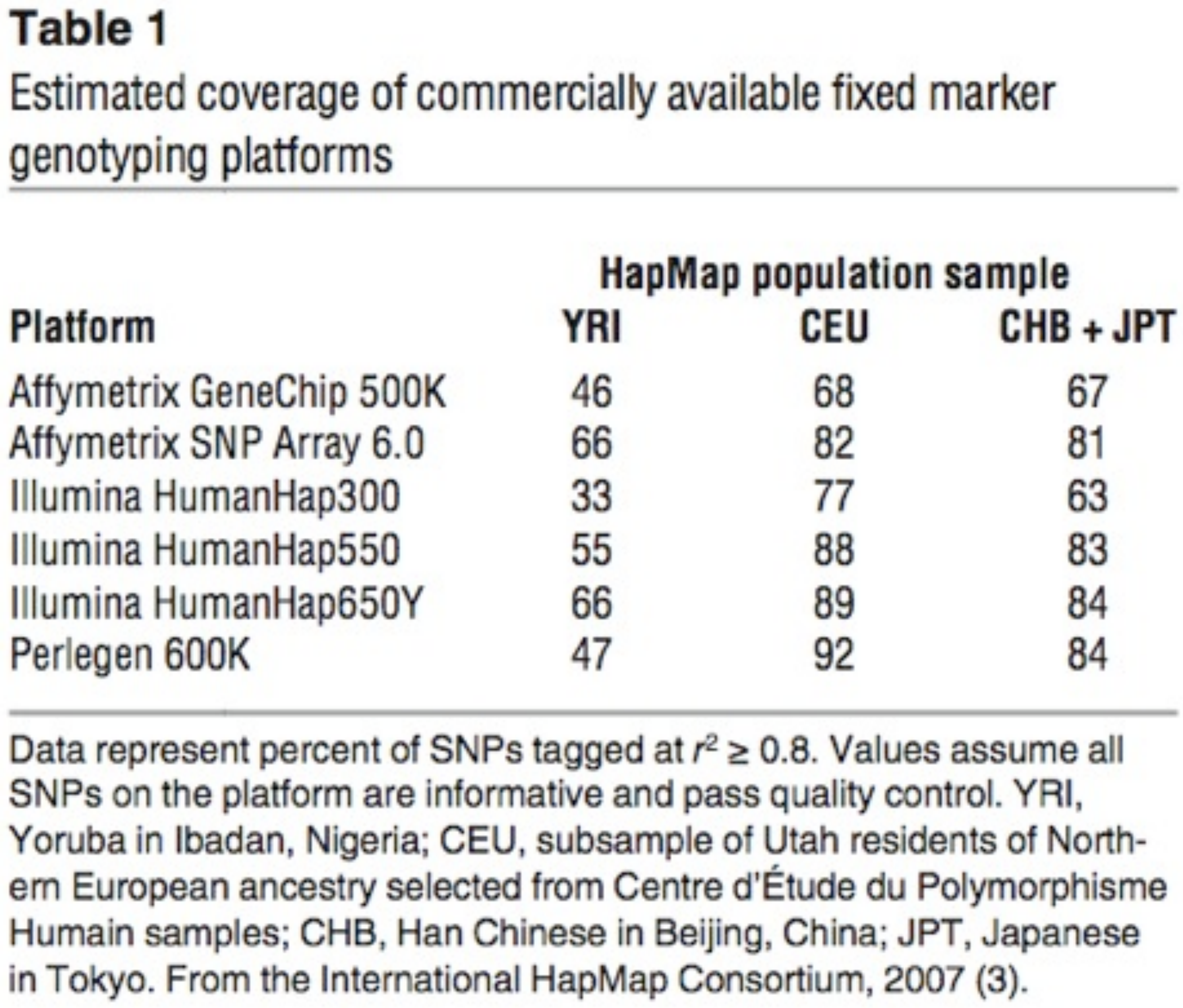


Table 1
Estimated coverage of commercially available fixed marker genotyping platforms

	HapMap population sample		
Platform	**YRI**	**CEU**	**CHB + JPT**
Affymetrix GeneChip 500K	46	68	67
Affymetrix SNP Array 6.0	66	82	81
Illumina HumanHap300	33	77	63
Illumina HumanHap550	55	88	83
Illumina HumanHap650Y	66	89	84
Perlegen 600K	47	92	84

Data represent percent of SNPs tagged at $r^2 \geq 0.8$. Values assume all SNPs on the platform are informative and pass quality control. YRI, Yoruba in Ibadan, Nigeria; CEU, subsample of Utah residents of Northern European ancestry selected from Centre d'Étude du Polymorphisme Humain samples; CHB, Han Chinese in Beijing, China; JPT, Japanese in Tokyo. From the International HapMap Consortium, 2007 (3).

Genotyping

- Genotype data must go through quality control - errors in calling genotypes is a threat to the validity of genetic association studies
- Look at "call rate" of genotyping - proportion of samples successfully typed for a given SNP; avoid analyzing SNPs with call rates <90-95%
- Check that genotype data observes Hardy-Weinberg Equilibrium

Hardy-Weinberg Equilibrium

- $p + q = 1$
- $p^2 + 2pq + q^2 = 1$
- Deviations may be due to:
  - non-random mating (inbreeding)
  - genetic drift
  - migration
  - new mutations
  - selection

Cohort: test all subjects

Case-control: test controls

Analysis

- Each SNP is an independent test
- Associations are tested by comparing the frequency of each allele in cases and controls
- The frequency of each of 3 possible genotypes can also be compared

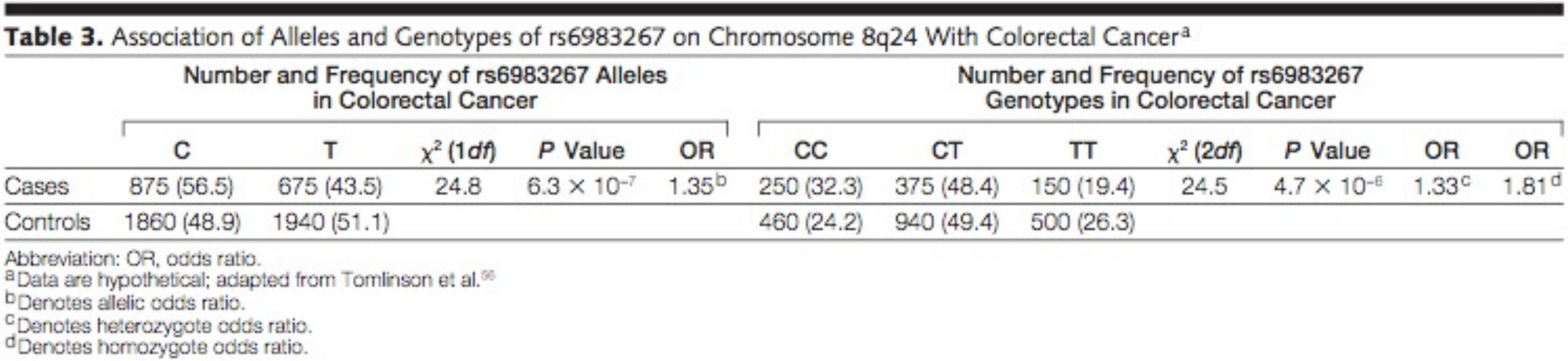


Table 3. Association of Alleles and Genotypes of rs6983267 on Chromosome 8q24 With Colorectal Cancer[a]

	Number and Frequency of rs6983267 Alleles in Colorectal Cancer					Number and Frequency of rs6983267 Genotypes in Colorectal Cancer						
	C	T	χ^2 (1*df*)	*P* Value	OR	CC	CT	TT	χ^2 (2*df*)	*P* Value	OR	OR
Cases	875 (56.5)	675 (43.5)	24.8	6.3×10^{-7}	1.35[b]	250 (32.3)	375 (48.4)	150 (19.4)	24.5	4.7×10^{-6}	1.33[c]	1.81[d]
Controls	1860 (48.9)	1940 (51.1)				460 (24.2)	940 (49.4)	500 (26.3)				

Abbreviation: OR, odds ratio.

[a] Data are hypothetical; adapted from Tomlinson et al.[96]

[b] Denotes allelic odds ratio.

[c] Denotes heterozygote odds ratio.

[d] Denotes homozygote odds ratio.

Pearson *et al. JAMA*, 2008

Odds ratios

- measure of effect size, or strength of association
- odds = P / (1-P)
- Probability of winning is 50%:
  - odds is 0.5 / (1-0.5) = 1 (1 to 1, 50:50, "even money")
- If probably of winning is 75%
  - odds is 0.75 / (1-0.75) = 3
- Odds ratio = $\dfrac{\text{odds(event | exposure)}}{\text{odds(event | lack of exposure)}}$

Odds ratios

- P (D | genotype “AT”) = 0.8
- P (D | genotype “TT”) = 0.2
- OR for getting the disease with genotype AT compared to TT?
  - OR = (0.8 / 0.2) / (0.2 / 0.8) = 16
- What’s the OR for AT individuals relative to an average population risk of 25%?
  - OR = (0.8 / 0.2) / (0.25 / 0.75) = 12

Analyzing a SNP for association

Genotype Counts

Association of rs6983267 on 8q24 with colorectal cancer

	CC	CT	TT
Cases	250	375	150
Controls	460	940	500

OR_{CT} = odds(disease | CT) / odds(disease | CC) = 250*940 / 460*375 = 1.36

OR_{TT} = odds(disease | TT) / odds(disease | CC) 250*500 / 460*150 = 1.81

$(1.36)^2 = 1.85$ (approximate additive model)

Analyzing a SNP for association

Allele Counts

Association of rs6983267 on 8q24 with colorectal cancer

	C	T
Cases	875 (56.5)	675 (43.5)
Controls	1860 (48.9)	1940 (51.1)

OR$_T$ = odds(disease | T) / odds(disease | C)
= 875*1940 / 1860*675 = 1.35

Cases

C alleles = 2 * 250 “CC” + 375 “CT” = 875
T alleles = 2 * 150 “TT” + 375 “CT” = 675

Controls

C alleles = 2 * 460 “CC” + 940 “CT” = 1860
T alleles = 2 * 500 “TT” + 940 “CT” = 1940

Odds ratios

- GWAS most often report ORs relative to the low-risk allele or lowest-risk genotype
- To turn this into a meaning risk estimate, the prevalence of the disease and the genotype frequencies must be taken into account
- P(D) = prevalence

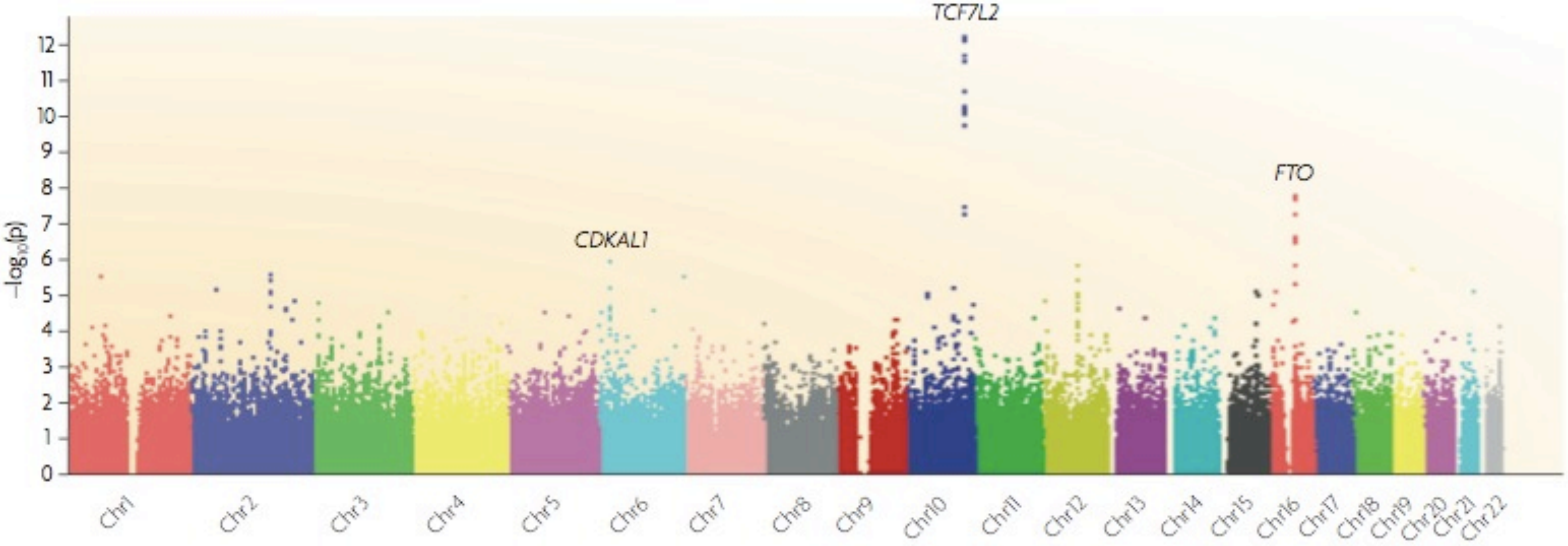
$$= P(D|AA)P(AA) + P(D|Aa)P(Aa) + P(D|aa)P(aa)$$

- More on this next week

Genome-wide analysis

- An odds ratio and associated p-value are calculated for each SNP (100,000 - 1M *P*-values!)
- $-\log_{10}$(*P*-value) : stronger p-value, bigger number

Manhattan Plot

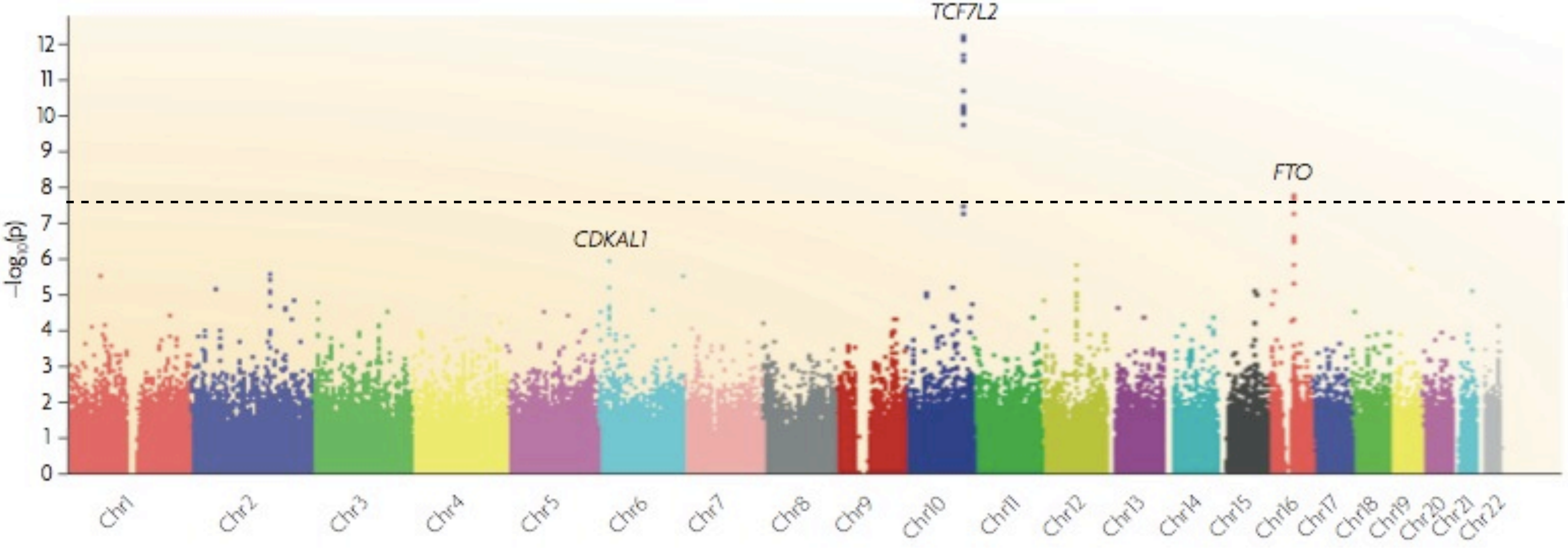


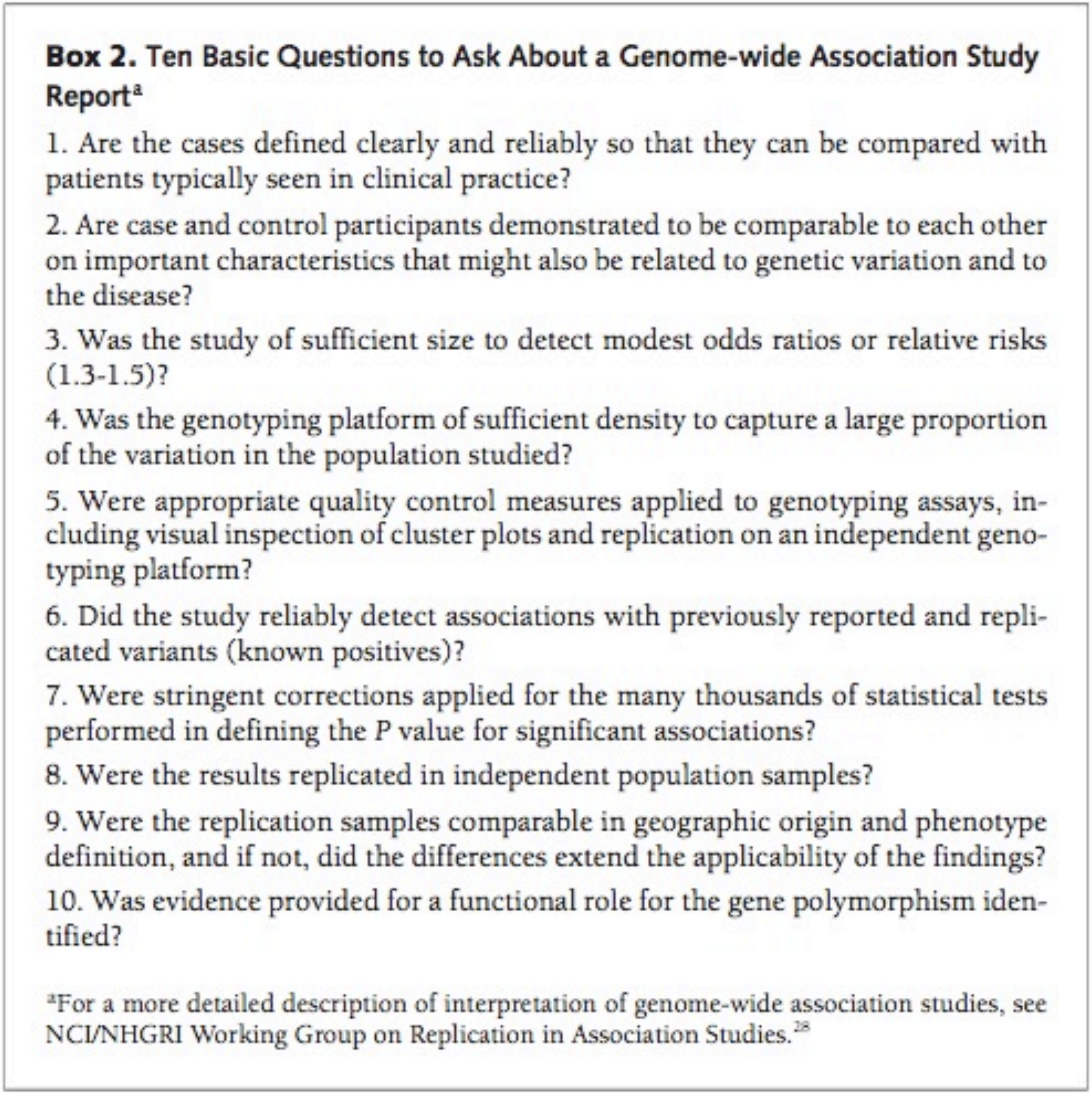
Multiple Hypothesis Testing

- A conventional p = 0.05 threshold assumes a 5% chance of a false-positive finding due to chance
- When performing one test, this is reasonable
- When performing 1,000,000 tests, this will lead to *many false-positives* (1x10^6 * 5% = 50,000 significant SNPs just by chance)
- Addressed most commonly by Bonferroni correction: threshold $P<0.5 / 10^6 = 5 \times 10^{-8}$

Multiple Hypothesis Testing

Manhattan Plot





Box 2. Ten Basic Questions to Ask About a Genome-wide Association Study Report[a]

1. Are the cases defined clearly and reliably so that they can be compared with patients typically seen in clinical practice?

2. Are case and control participants demonstrated to be comparable to each other on important characteristics that might also be related to genetic variation and to the disease?

3. Was the study of sufficient size to detect modest odds ratios or relative risks (1.3-1.5)?

4. Was the genotyping platform of sufficient density to capture a large proportion of the variation in the population studied?

5. Were appropriate quality control measures applied to genotyping assays, including visual inspection of cluster plots and replication on an independent genotyping platform?

6. Did the study reliably detect associations with previously reported and replicated variants (known positives)?

7. Were stringent corrections applied for the many thousands of statistical tests performed in defining the *P* value for significant associations?

8. Were the results replicated in independent population samples?

9. Were the replication samples comparable in geographic origin and phenotype definition, and if not, did the differences extend the applicability of the findings?

10. Was evidence provided for a functional role for the gene polymorphism identified?

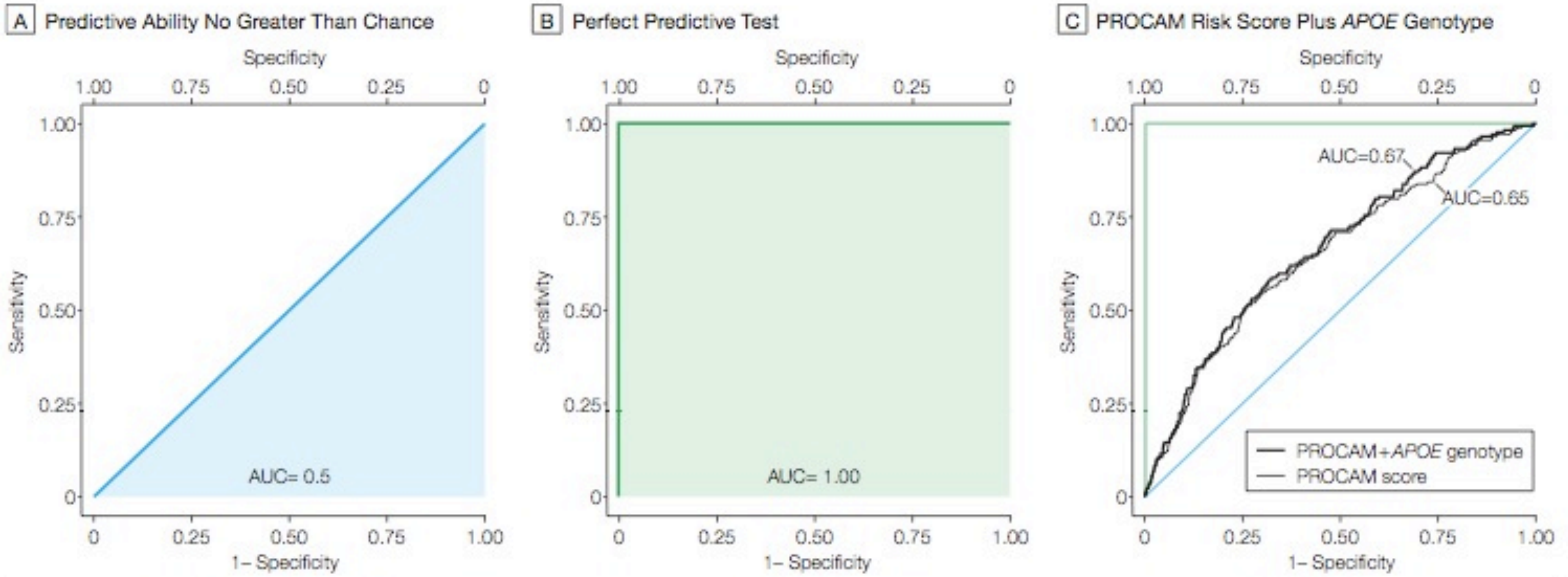
[a] For a more detailed description of interpretation of genome-wide association studies, see NCI/NHGRI Working Group on Replication in Association Studies.[28]

Pearson *et al. JAMA*, 2008

Is there clinical utility in the findings?

Sensitivity, specificity, PPV, NPV
These studies for the most part have not been done yet!

Figure. Example of a Receiver Operating Characteristic (ROC) Curve for Cardiovascular Risk Related to *APOE*



A, Example of an ROC curve for a test that performs no better than chance. B, Example of an ROC curve for a test with perfect predictive ability (sensitivity = 100%; specificity = 100%). C, ROC curves for cardiovascular disease calculated using PROCAM (Prospective Cardiovascular Munster study) risk score plus *APOE* genotype. Based on 2451 men (of 3012 eligible) who had complete data for PROCAM and *APOE* genotyping. *APOE* genotype was fitted as a class variable with 3 categories 33, 22/23, and 34/44. Factors included age, body mass index, total cholesterol, triglycerides, systolic blood pressure, and family history. Other factors in PROCAM were not measured in all men. For the PROCAM score, the ROC value (95% confidence interval) was 0.65 (0.61-0.70), with a detection rate of 11.7% for a false-positive rate of 5.0%. In univariate analysis, *APOE* genotype was significant at $P=.01$. In multivariate analysis, the area under the curve increased to 0.67 (0.63-0.71) (detection rate,14.0%), but this improvement was not significant ($P=.11$). Panel C data based on Humphries et al.[12]

Is there clinical utility in the findings?

Diabetes 16-SNP model

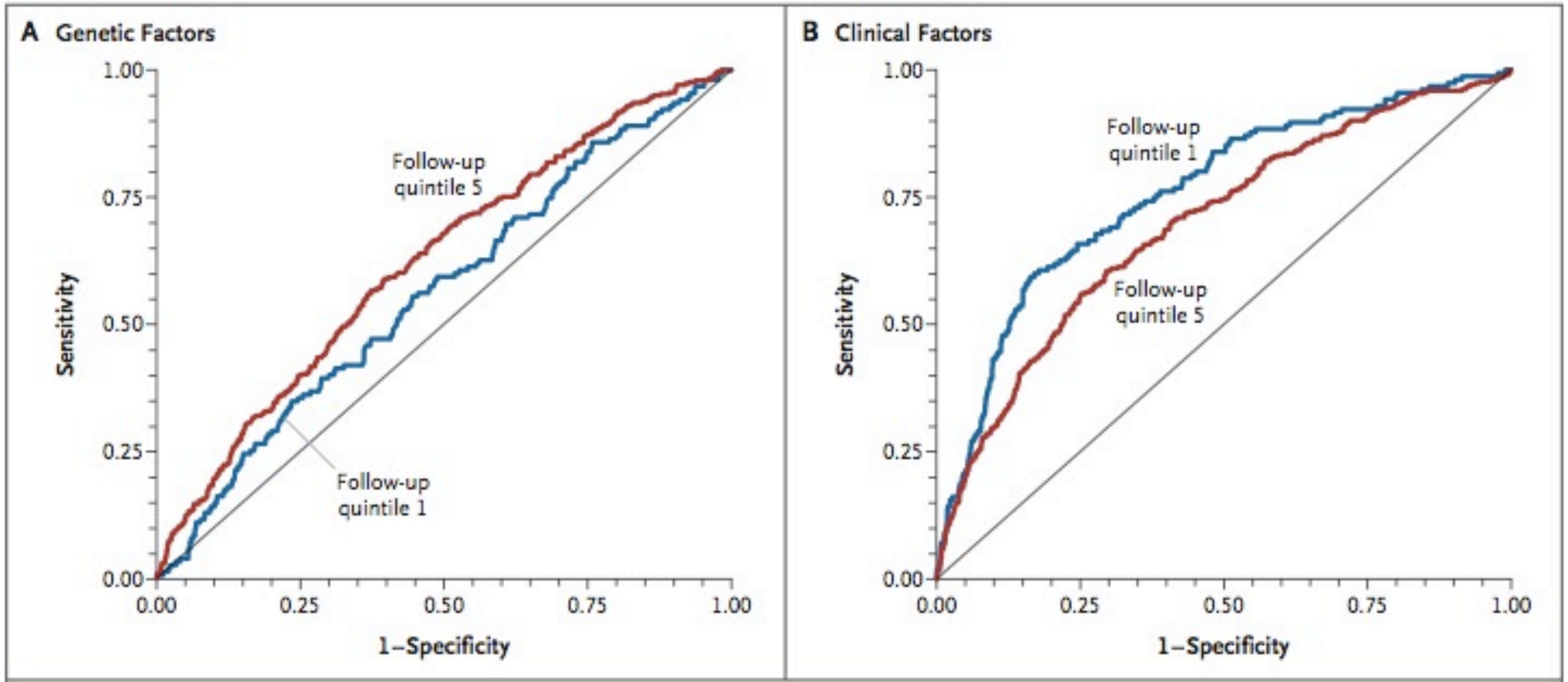


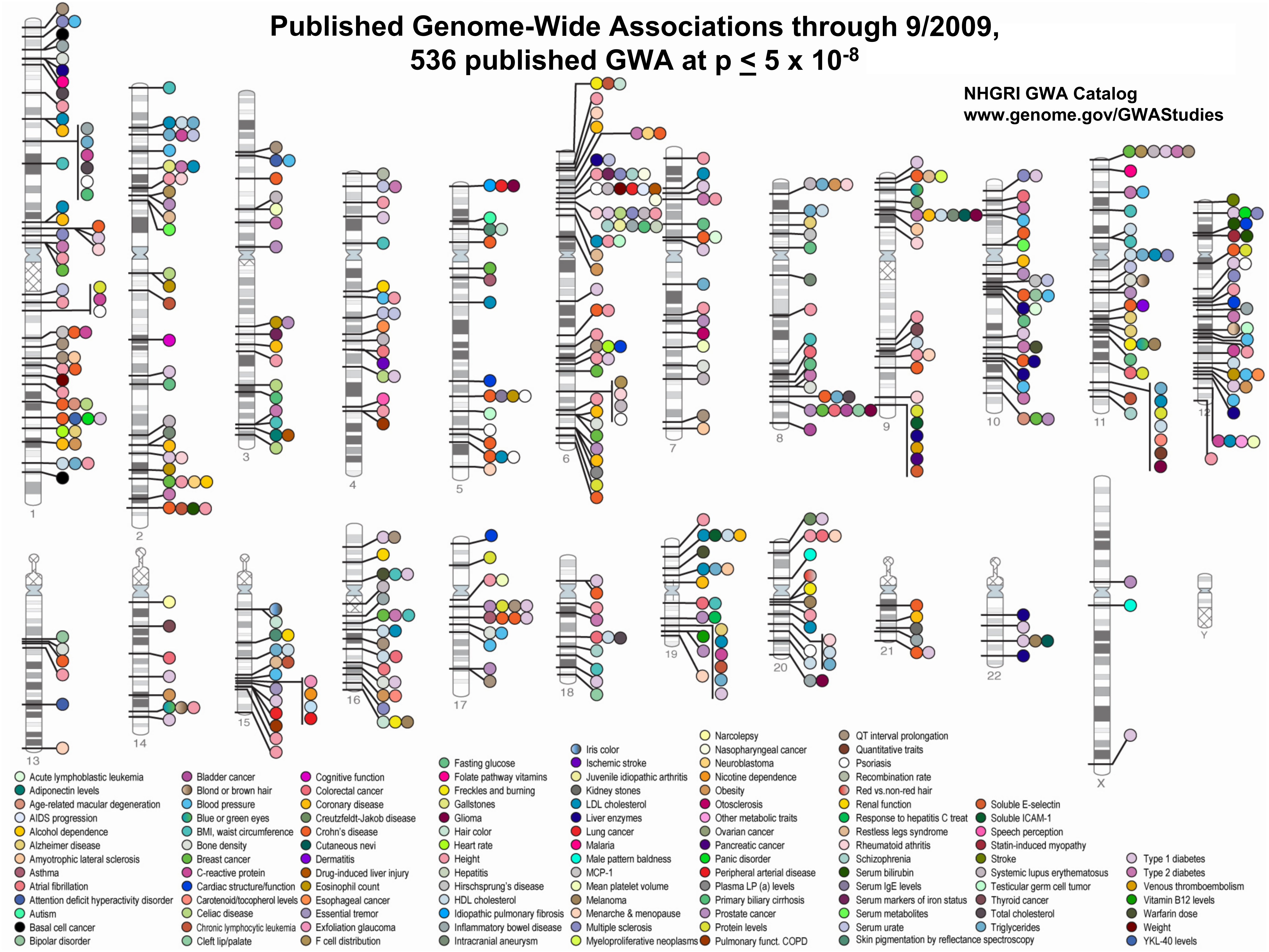
Figure 4. Area under the ROC Curve (C Statistic) for Clinical and Genetic Models Predicting Type 2 Diabetes, According to the Duration of Follow-up.

The effect of genetic risk factors increases with the duration of follow-up, with an area under the ROC curve (AUC) of 0.56 in quintile 1 (blue) and 0.62 in quintile 5 (red) (P=0.01) (Panel A), whereas the effect of clinical risk factors decreased with the duration of follow-up, with an AUC of 0.75 in quintile 1 and 0.67 in quintile 5 (P=0.01) (Panel B). The black line indicates reference values.

Lyssenko et al, *NEJM* 2010

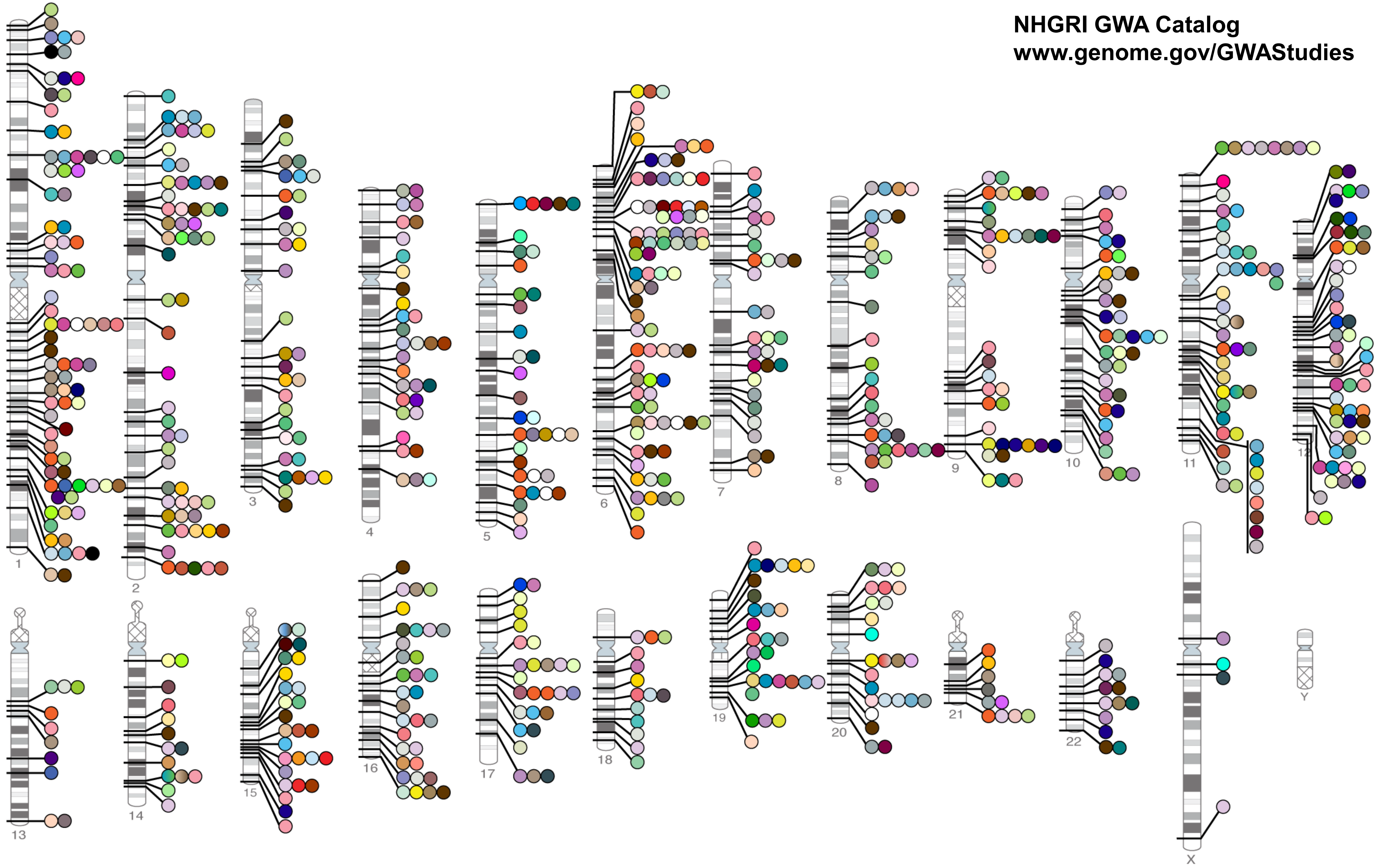
Published Genome-Wide Associations through 9/2009,
536 published GWA at p $\leq 5 \times 10^{-8}$

NHGRI GWA Catalog
www.genome.gov/GWAStudies



Published Genome-Wide Associations through 3/2010, 779 published GWA at $p \leq 5x10^{-8}$ for 148 traits

NHGRI GWA Catalog
www.genome.gov/GWAStudies



Wednesday, July 7, 2010

Future of GWAS

- Addressing missing heritability
  - Common variation not fully explored - 25% of genes (51% of drug targets) have SNPs not measured directly or imputable on commercial genotyping platforms (Rong Chen *et al*. personal communication)
  - Rare variants unexplored - sequencing-based methods
  - Structural variants and other non-SNP polymorphisms, epigenomics