Genetic analysis of complex traits
in the age
# of the genome-wide association study

David Duffy

Queensland Institute of Medical Research
Brisbane, Australia

Overview

- Complex genetic traits
- Complex diseases as quantitative traits
- The genetic architecture of quantitative traits
- Why are complex diseases heritable at all?
- Linkage disequilibrium and allelic association
- High-throughput genotyping
- Genome-wide association

What is a complex genetic trait?

This is a fuzzy concept, as everything in genetics is complex. For example, Retinitis Pigmentosa is due to mutations at 52 mapped and unmapped loci, but is not usually thought of as a complex disorder in that usually a single mutation is a **sufficient cause** in any one pedigree.

I would use it to refer to traits under the control of multiple genes and multiple environmental influences, where no individual genetic locus has a very large effect in its own right:

- Most common chronic diseases eg hypertension, cancers, diabetes
- Quantitative trait such as height, biochemical analytes

Complex genetic traits as quantitative traits

Most quantitative traits are complex genetically, and are under the control of many **quantitative trait loci**, each locus acting on a different part of a series of biochemical or physiological pathways or networks.

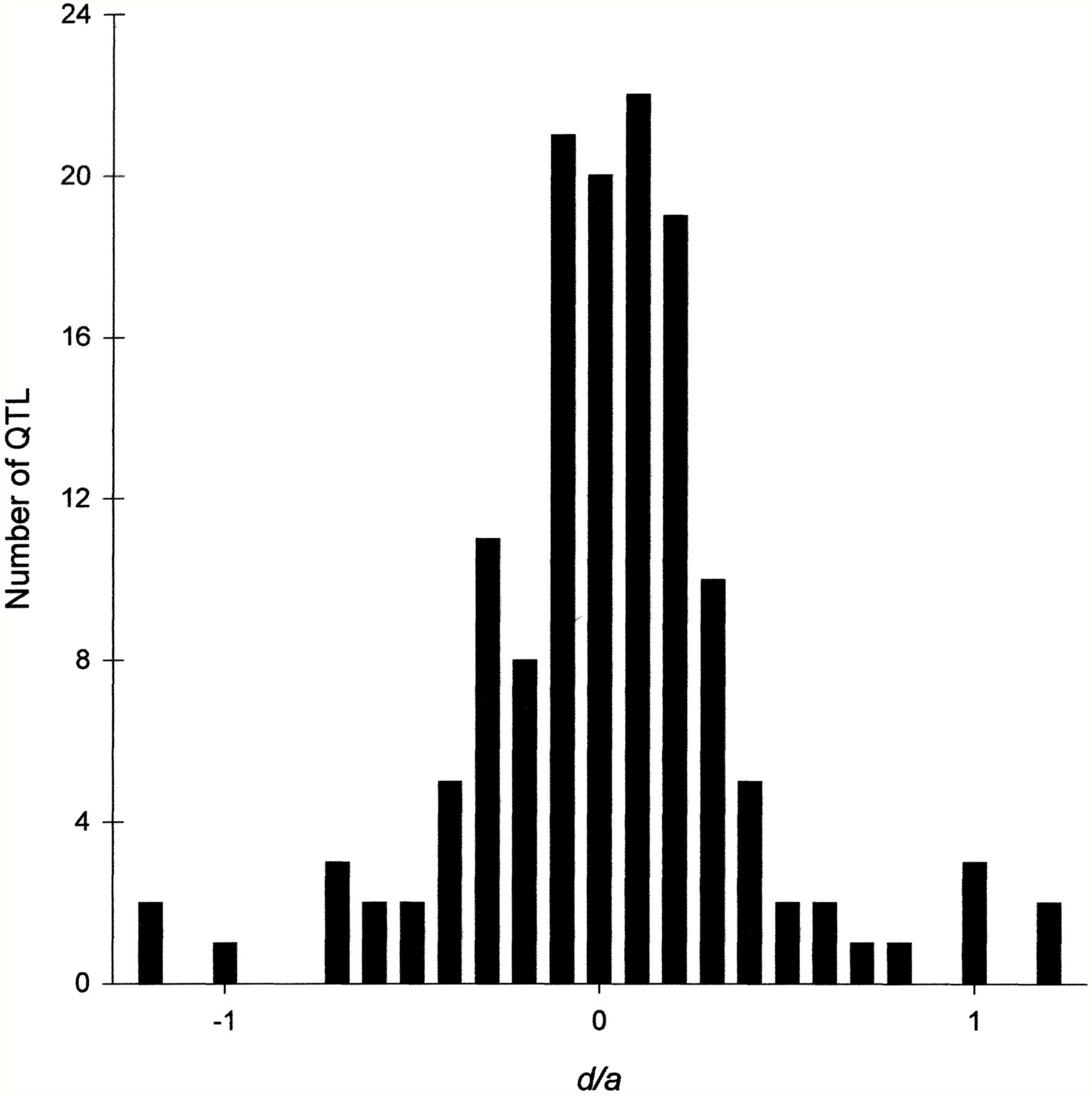
Many human diseases are characterized by important **endophenotypes** that are quantitative in nature, such as blood pressure, plasma glucose, airway responsiveness.

The genetic architecture of quantitative traits

- Multiple QTLs affect each trait
- Distribution of QTL effect sizes seem L-shaped or exponential
- Distribution of effect sizes of new mutations is also exponential
- QTLs interact with the environment of the organism
- Interaction between QTLs is common (**epistasis**)

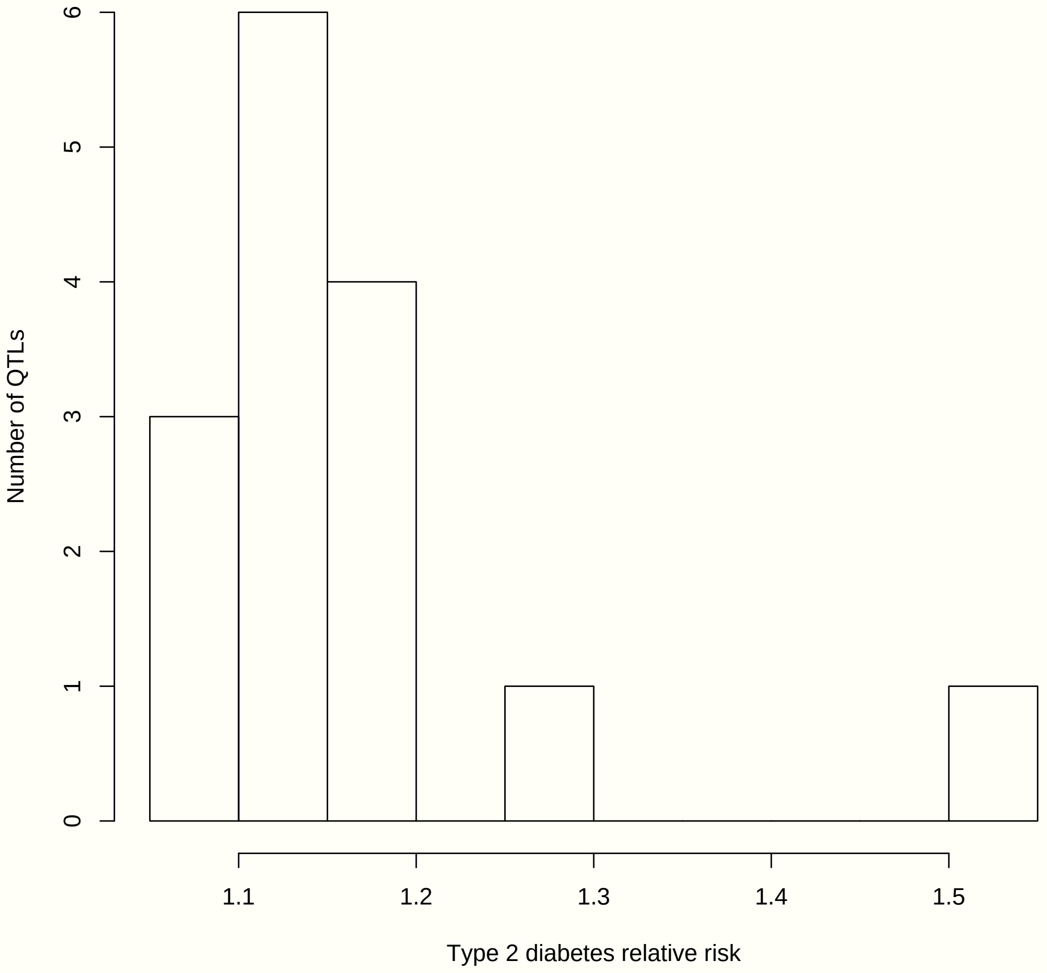
The genetic architecture of quantitative traits

Distribution of additive QTL effects on Drosophila sensory bristle number (Figure 6 from Dilda and Mackay, 2002).



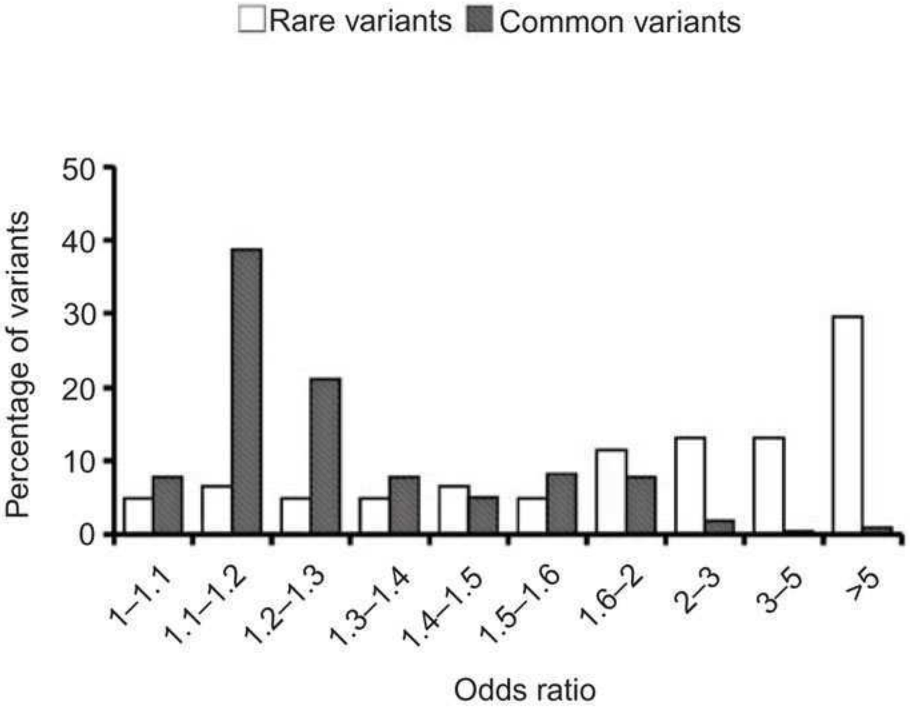
The genetic architecture of complex disease

Distribution of additive QTL effects on risk of Type 2 diabetes (from Doria et al, 2008).



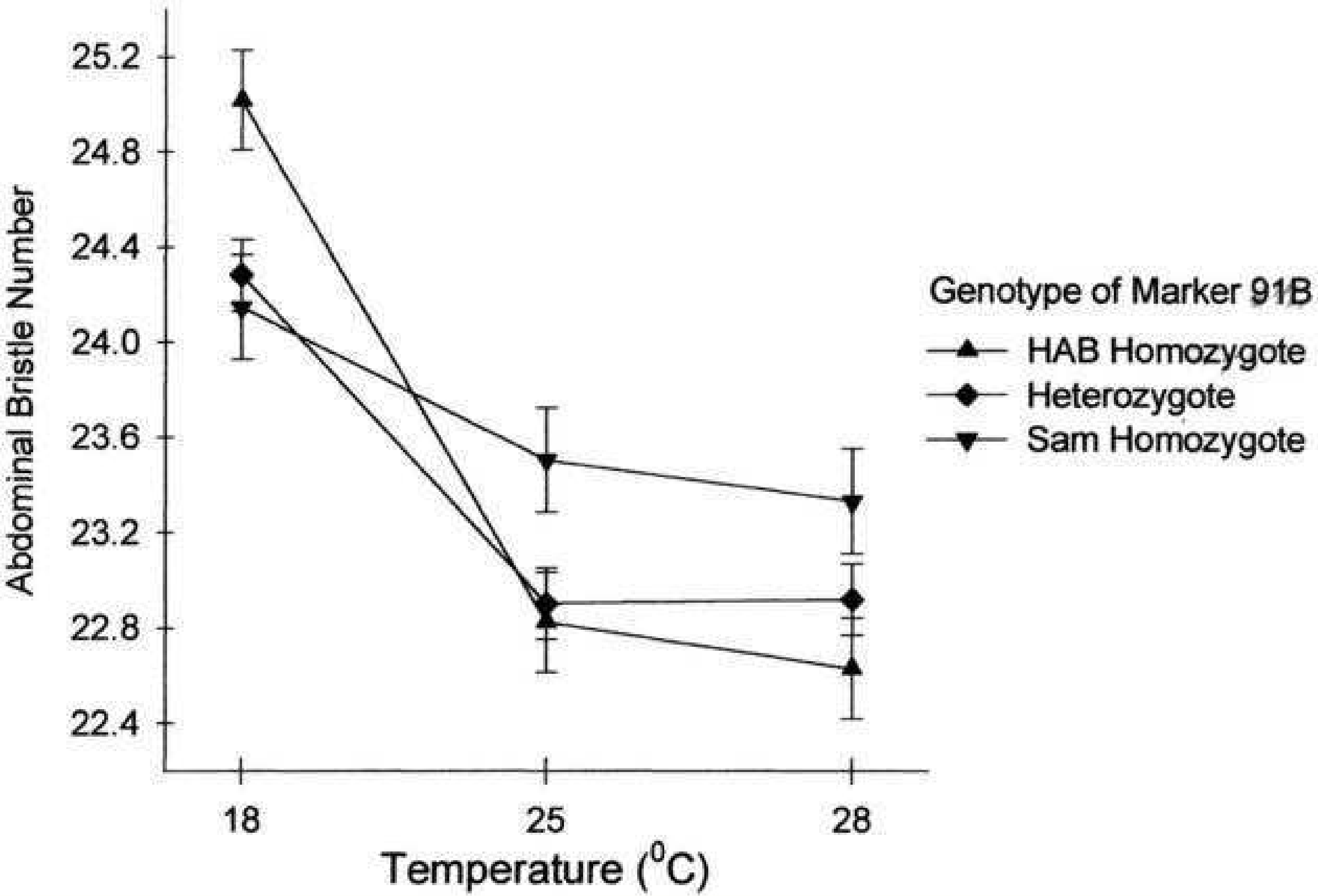
The genetic architecture of complex disease

Distribution of QTL effects on disease from 64 studies (from Bodmer and Bonilla, 2008).



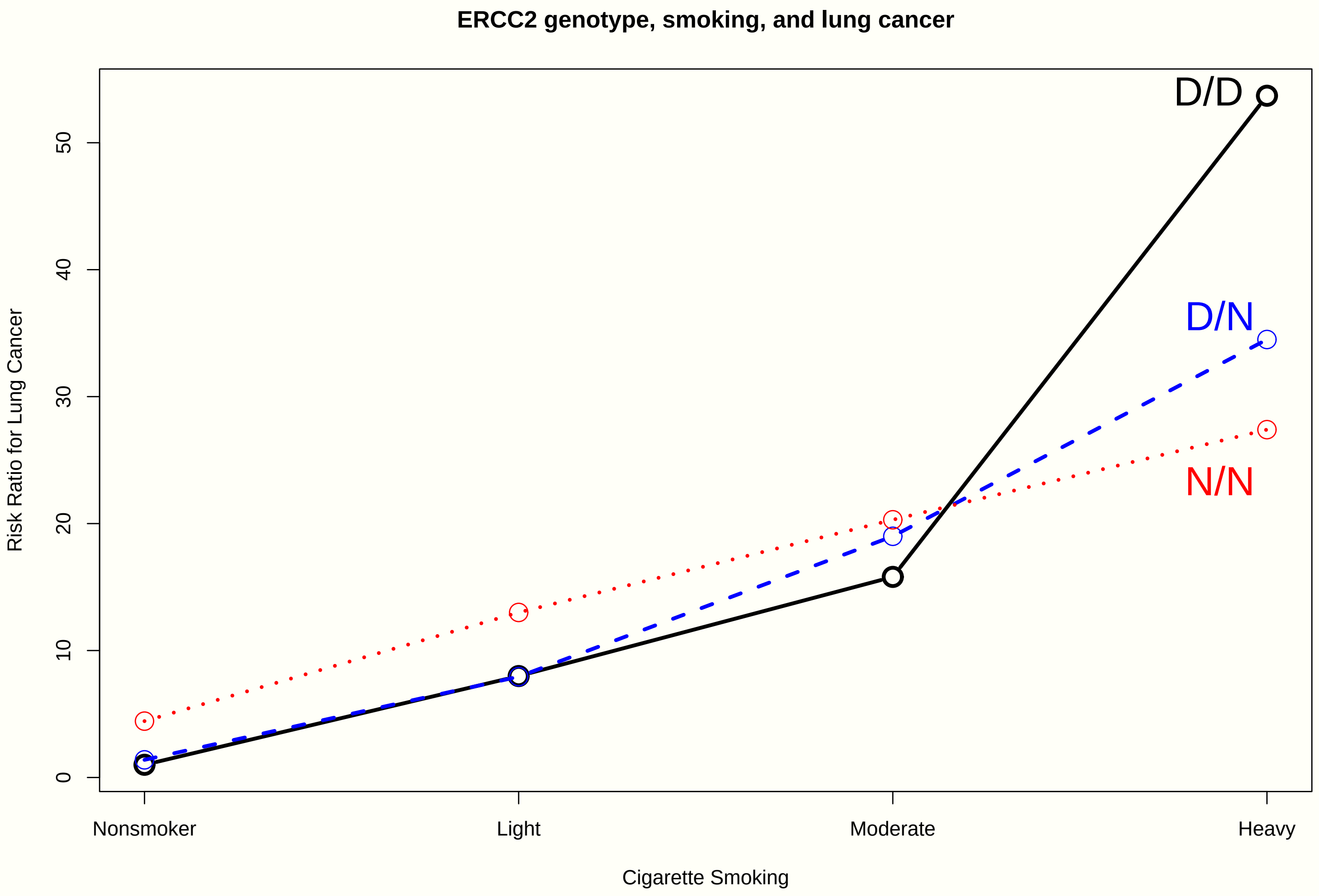
The genetic architecture of quantitative traits

Gene by environment interaction for a bristle number QTL (Figure 9 from Dilda and Mackay, 2002).



The genetic architecture of complex disease

Gene by environment interaction for ERCC2 and lung cancer (from Zhou et al, 2002).



Why are complex diseases heritable at all?

Most important human diseases aggregate within families. One might expect selection to purge risk genotypes from the population, but:

- Recurrent mutation gives rise to new disease alleles
- Selection operates weakly on recessive disorders
- Many diseases have only a small effect on reproductive success

Effect: Many rare disease alleles
(“Traditional” genetic load, mutation-selection)

- Pleiotropy plus overdominance can maintain polymorphism
- Modifier loci may arise

Effect: Higher frequency disease alleles with lower penetrances
(“common disease, common variants”)

Multiple rare alleles and schizophrenia

One type of rare mutation that can be screened for with current array technology is a microdeletion or duplication (CNV).

Walsh et al (2008): *De novo* deletions and duplications detected using Illumina 550K and Nimblegen 2.1M Genome-Wide SNP arrays.

	All Schizophrenia	Early-onset	Controls
N	150	76	268
New CN mutations	22 (14.8%)	15 (19.7%)	13 (4.9%)

Xu et al (2008): *De novo* microdeletions and duplications detected using the Affy Human Genome-Wide SNP array 5.0.

	"Sporadic" Scz	Familial Scz	Controls
N	152	48	159
New CN mutations	15 (9.9%)	0 (0%)	2 (1.2%)

Table 3 from Walsh et al (2008). Pathways and processes over-represented by genes disrupted in schizophrenia cases by deletions or insertions.

Pathway or process	P value
Signal transduction	0.012
Neuronal activities	0.049
Nitric oxide signaling	0.0002
Synaptic long term potentiation	0.0005
Glutamate receptor signaling	0.003
ERK/MAPK signaling	0.004
PTEN signaling	0.007
Neuregulin signaling	0.008
IGF-1 signaling	0.008
Axonal guidance signaling	0.015
Synaptic long term depression	0.017
G-protein coupled receptor signaling	0.034
Integrin signaling	0.036
Ephrin receptor signaling	0.042
Sonic hedgehog signaling	0.044

Recurrent mutation and schizophrenia

The multicentre study set up by deCODE Genetics, concentrated on just 66 *de novo* CNVs found by screening 7718 control families. Of these, 3 were increased in schizophrenics compared to controls:

Stefansson et al (2008): Recurrent microdeletions detected using the Illumina HumanHap300 and HumanCNV370 arrays.

Region	Coordinates (Mbp)	Schizophrenics	Controls
1q21.1	144.94-146.29	11/4718 (0.23%)	8/41199 (0.02%)
15q11.2	20.31-20.78	26/4718 (0.55%)	79/41194 (0.19%)
15q13.3	28.72-30.30	7/4213 (0.17%)	8/39800 (0.02%)

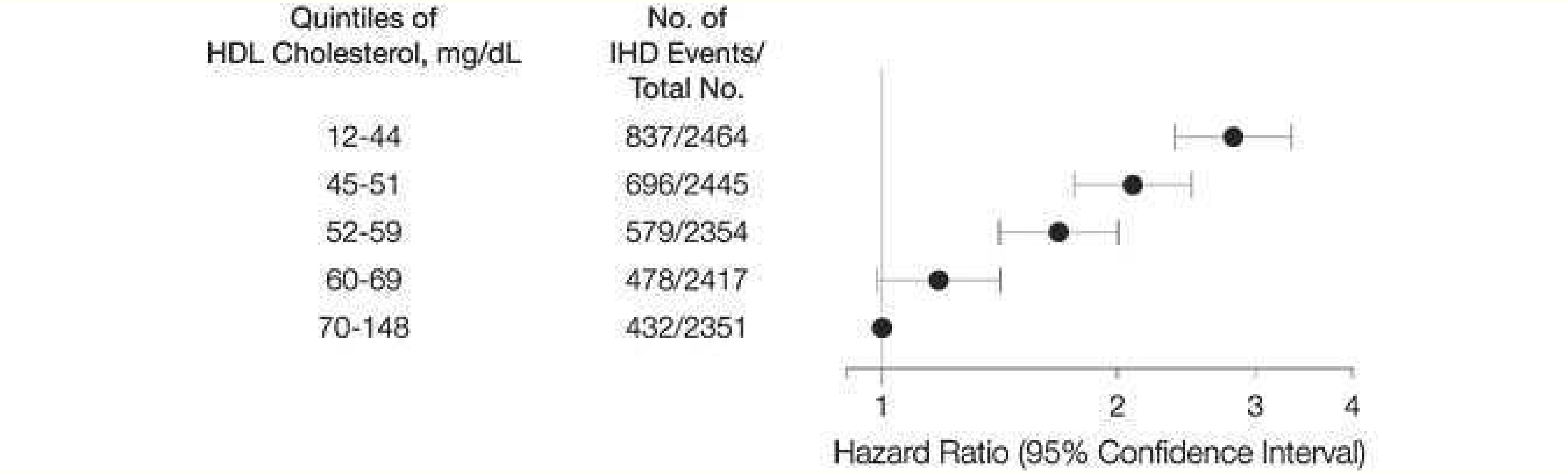
ApoE and Alzheimer’s Disease: “CDCV”

ApoE is one of the best examples of a common variant with a large effect on risk of a complex disorder - Alzheimer’s Disease. There is strong evidence for interactions with either other loci or environment.

Population	ApoE*4 frequency	Relative Risk for AD
Kenya	30%	1.0
Tanzania	25%	1.0
Yoruba	22%	1.0
African-Americans	20%	2.3
Europe	15%	2.5
Iran	6%	3.7

HDL and heart disease

Plasma HDL level is an important endophenotype/risk factor for atherosclerosis.



Quintiles of HDL Cholesterol, mg/dL	No. of IHD Events/ Total No.
12-44	837/2464
45-51	696/2445
52-59	579/2354
60-69	478/2417
70-148	432/2351

Rare alleles and Low HDL level

Cohen (2004) sequenced three genes (*ABCA1*, *APOA1*, *LCAT*) in 128 subjects with low HDL levels (lowest 5%) and 128 subjects with high HDL levels (highest 5%) from a population sample.

Low HDL group (21)		
*ABCA1**S198X (1)	*ABCA1**P248A (1)	*ABCA1**K401Q (1)
*ABCA1**W590S (1)	*ABCA1**R638Q (1)	*ABCA1**T774S (4)
*ABCA1**E815G (1)	*ABCA1**S1181F (1)	*ABCA1**R1341T (1)
*ABCA1**S1376G (1)	*ABCA1**R1615Q (1)	*ABCA1**A1670T (1)
*ABCA1**N1800H (1)	*ABCA1**D2243E (4)	***APOA1****R51T (1)
High HDL group (3)		
*ABCA1**R496W (1)	*ABCA1**R1680Q (1)	***LCAT****V114M (1)

ABCA1 is the Tangier disease gene and is a well-known cause of familial hypoalphalipoproteinemia (HDL < 10%’ile and positive family history).

All of these mutations are individually rare.

Rare ABCA1 alleles and heart disease

Two of the *ABCA1* mutations above have been characterized biochemically (Singaraja 2006) and lead to Tangier Disease (homozygotes):

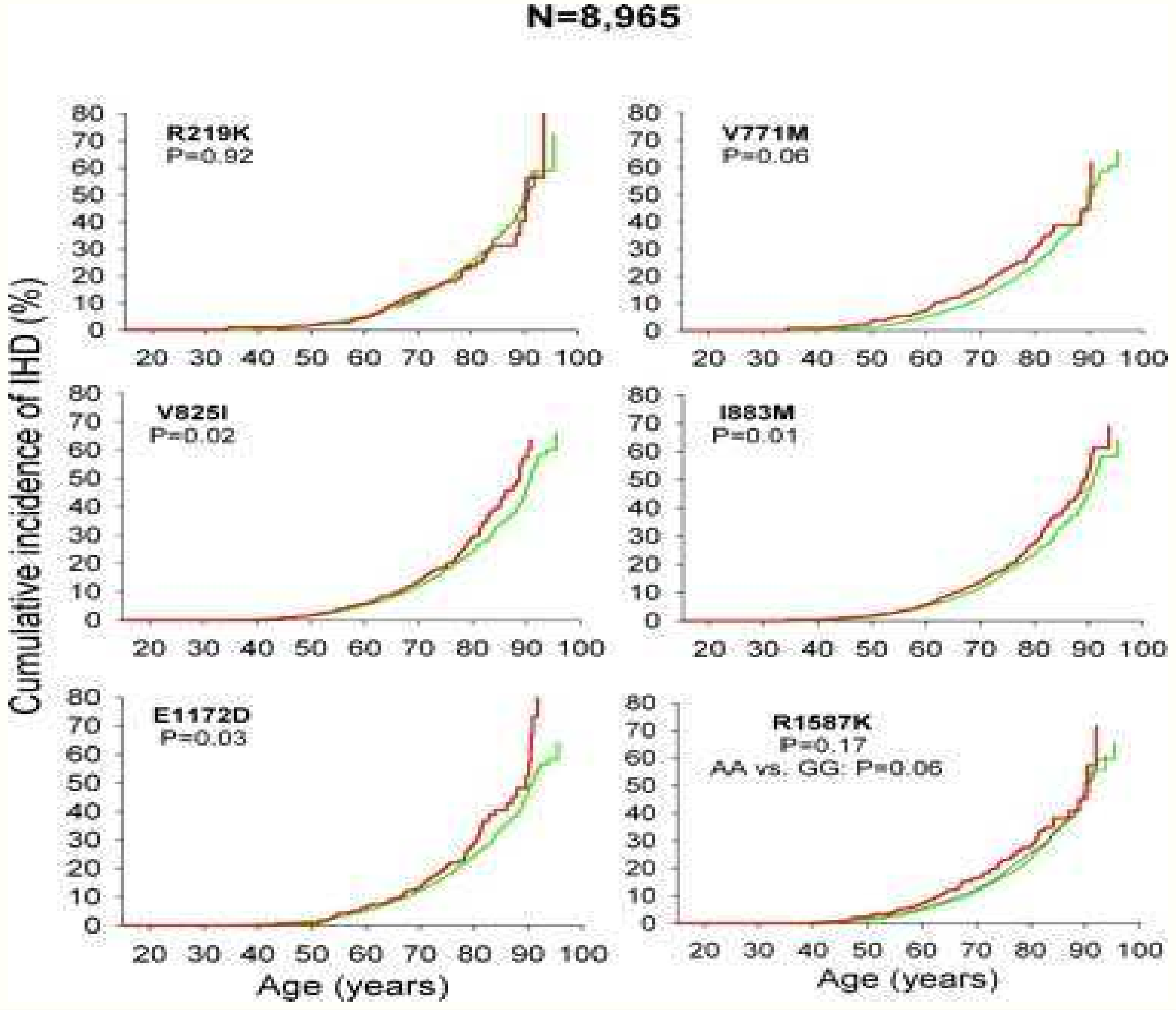
- W590S reduces Annexin V binding
- N1800H causes a failure of *ABCA1* to localize appropriately to the plasma membrane

Frikke-Schmidt et al (2008) studied 4 ABCA1 mutations in 42761 Danes, including N1800H:

Allele	Carriers	Relative risk of ischemic heart disease
P1065S	1 (0.0022%)	-
G1216V	7 (0.016%)	-
N1800H	95 (0.22%)	0.77 (0.41-1.45)
R2144X	6 (0.014%)	-
Any	109 (0.25%)	0.93 (0.53-1.62)

Common ABCA1 alleles and heart disease

Most studies have tested more common ABCA1 variants. In a subset of the same Danish sample (the Copenhagen City Heart Study), significant association with heart disease was detected. The alleles in question exhibited much smaller effects of HDL level than the rare alleles described earlier.


N=8,965
Cumulative incidence of IHD (%)
R219K
P=0.92
V771M
P=0.06
V825I
P=0.02
I883M
P=0.01
E1172D
P=0.03
R1587K
P=0.17
AA vs. GG: P=0.06
0
10
20
30
40
50
60
70
80
20
30
40
50
60
70
80
90
100
Age (years)

Risk alleles for Type 1 Diabetes

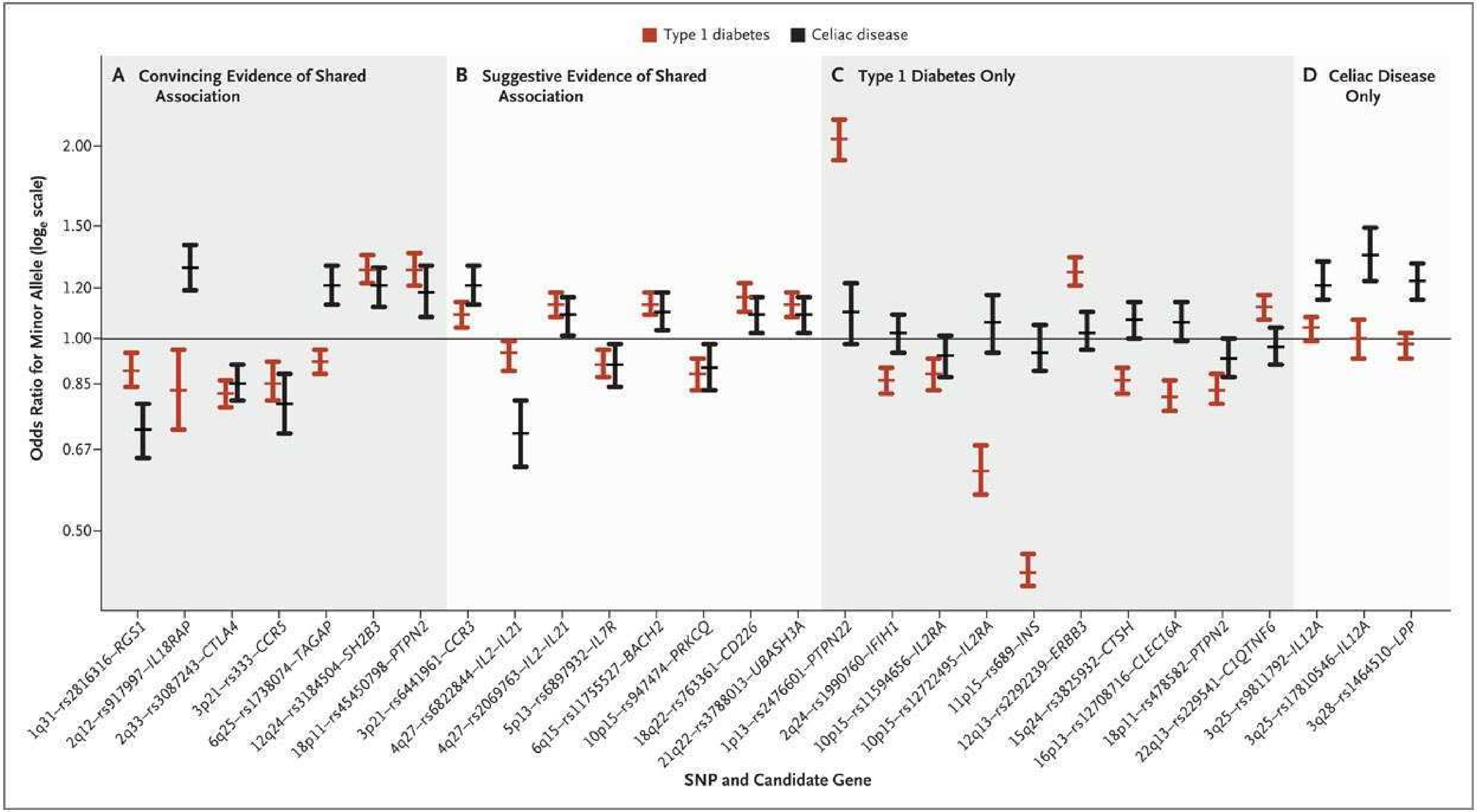
- 50% of T1D cases from 2% of population carrying high risk HLA genotypes
- 21 non-HLA risk loci confirmed
- Highest penetrance is 5.1% (baseline risk 0.3%)
- Pleiotropy for other autoimmune diseases and allergy

T1D susceptibility gene(s)	Chromosomal location (Name assigned via linkage analysis)	Other autoimmune diseases associated with locus	Other inflammatory diseases associated
DQA1, DQB1, DRB1	6p21 (IDDM1)	GE, RA, MS etc	Manifold but allelic heterogeneity
CTLA4 (CD28, ICOS)	2q33.2 (IDDM12)	AIH,GD	Atopy
CASP7	10q25 (IDDM17)	RA	
IFIH1	2q24 (IDDM19)	GD	
IL12B (?)	5q33.3 (IDDM18)		Atopy?, tuberculosis
IL2RA (CD25)	10p15 (IDDM10)	MS, GD	
PTPN22	1p13 (Idd10)	RA, GD, HT, SLE, AD, CD, MG, V	Endometriosis?
CCR5	3p21	Coeliac	
SH2B3	12q24	Coeliac	

Spectrum of risk alleles for Type 1 Diabetes

T1D Locus	Variant	Population frequency	Relative risk
DQA1, DQB1, DRB1	DR4-DQB1*0302	1%	20
	DR3-DQBG1*020	1%	20
TNF	rs1799964	22%	1.3
CTLA4 (CD28, ICOS)	A17T (rs231775)	71%	1.3
IFIH1	T946A (rs1990760)	30-60%	1.9
IL2	rs2069763	33%	1.1
IL2RA (CD25)	rs706778	45%	1.5
BACH2	rs11755527	45%	1.1
PTPN22	R620W	6-12%	1.8
CLEC16A	rs12708716	70%	1.2
SH2B3	rs3184504	40%	1.3

Spectrum of risk alleles for Type 1 Diabetes (Smyth et al 2008)



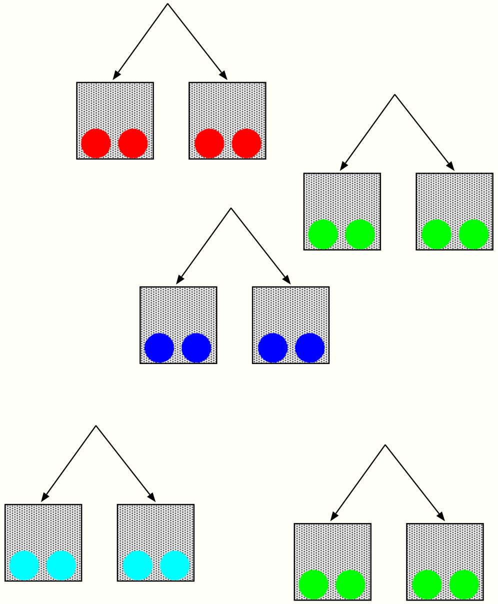
Linkage versus allelic association

Linkage analysis extracts information from co-transmission of traits and markers **between family members**. Localization of complex trait loci is usually at 1-10 Mbp resolution. The locus effect size needs to be more than 10% of the trait genetic variance to be detectable. Because of the natural randomization induced by segregation, linkage is robust to confounding.

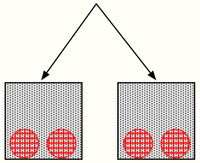
Allelic association analysis extracts information from co-occurrence of traits and markers **within individuals**. Localization of complex trait loci is usually at 0.01-0.1 Mbp resolution (in outbred populations). The locus effect size needs to be more than 1% of the trait genetic variance to be detectable. Association analysis is less robust to confounding than linkage analysis.

Linkage versus allelic association

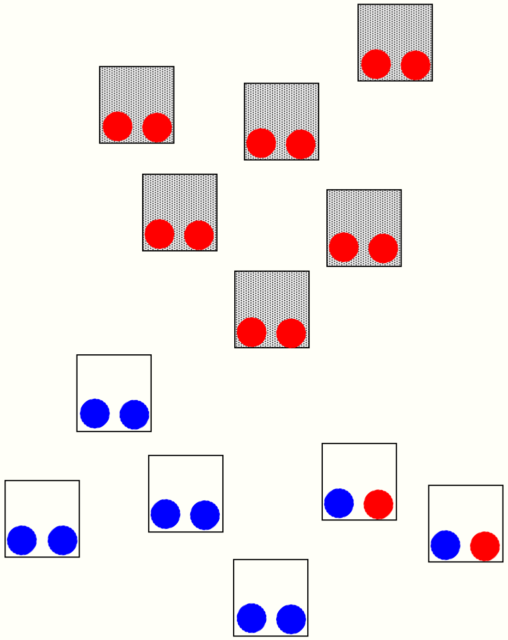
Affected Sib Pair Linkage



Mean IBD sharing = 100%
Expected sharing = 50%



Association



Case allele frequency = 100%
Expected frequency = 17%

Linkage disequilibrium and allelic association

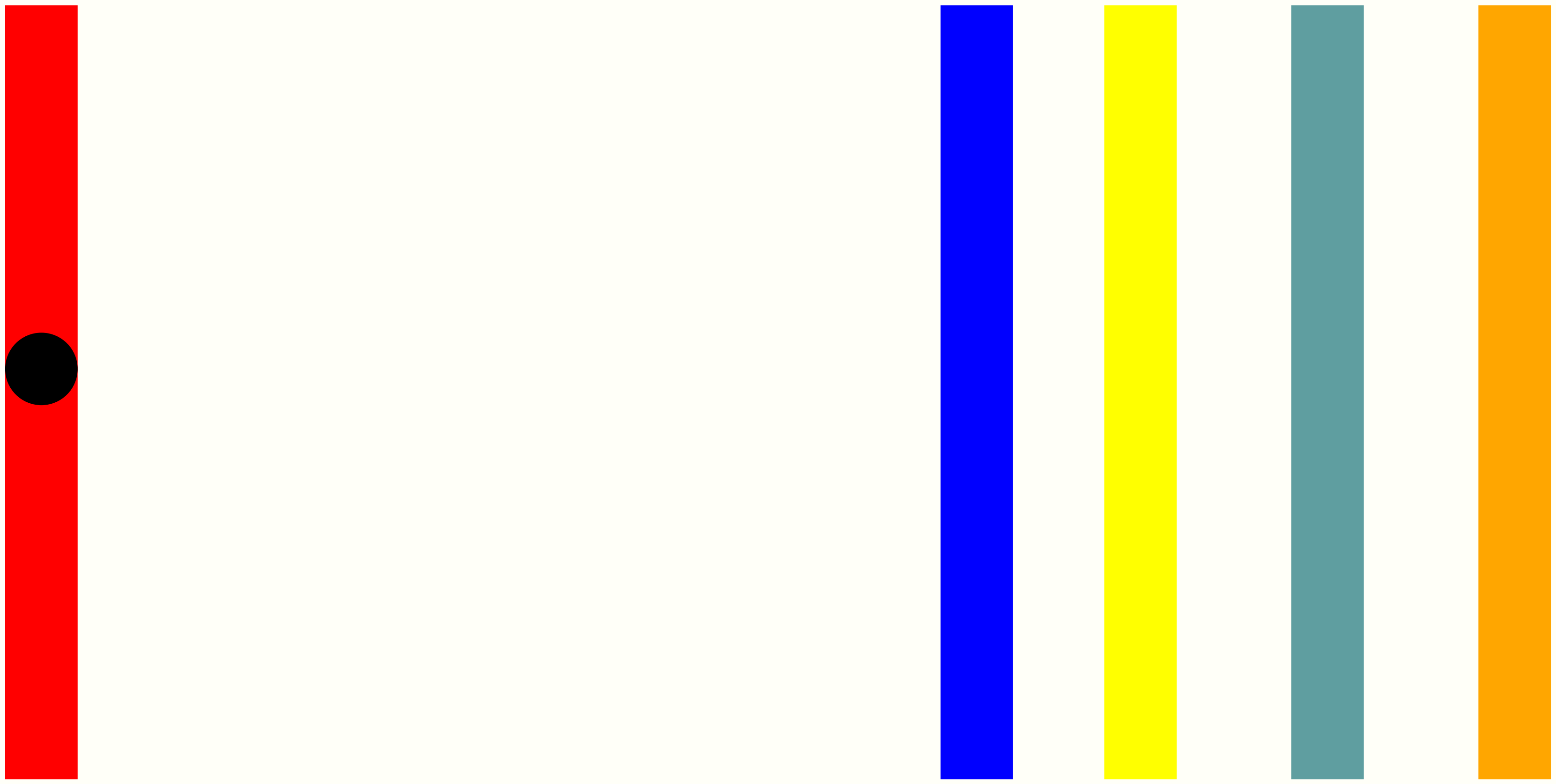
Allelic association between a trait and a gene variant occurs when:

- Direct relationship between variant and trait
- Linkage disequilibrium between variant and another directly associated allele
- Ethnic stratification

The most useful case is the second case, as it reduces the number of loci to be genotyped.

Breakdown of linkage disequilibrium

Generation 0

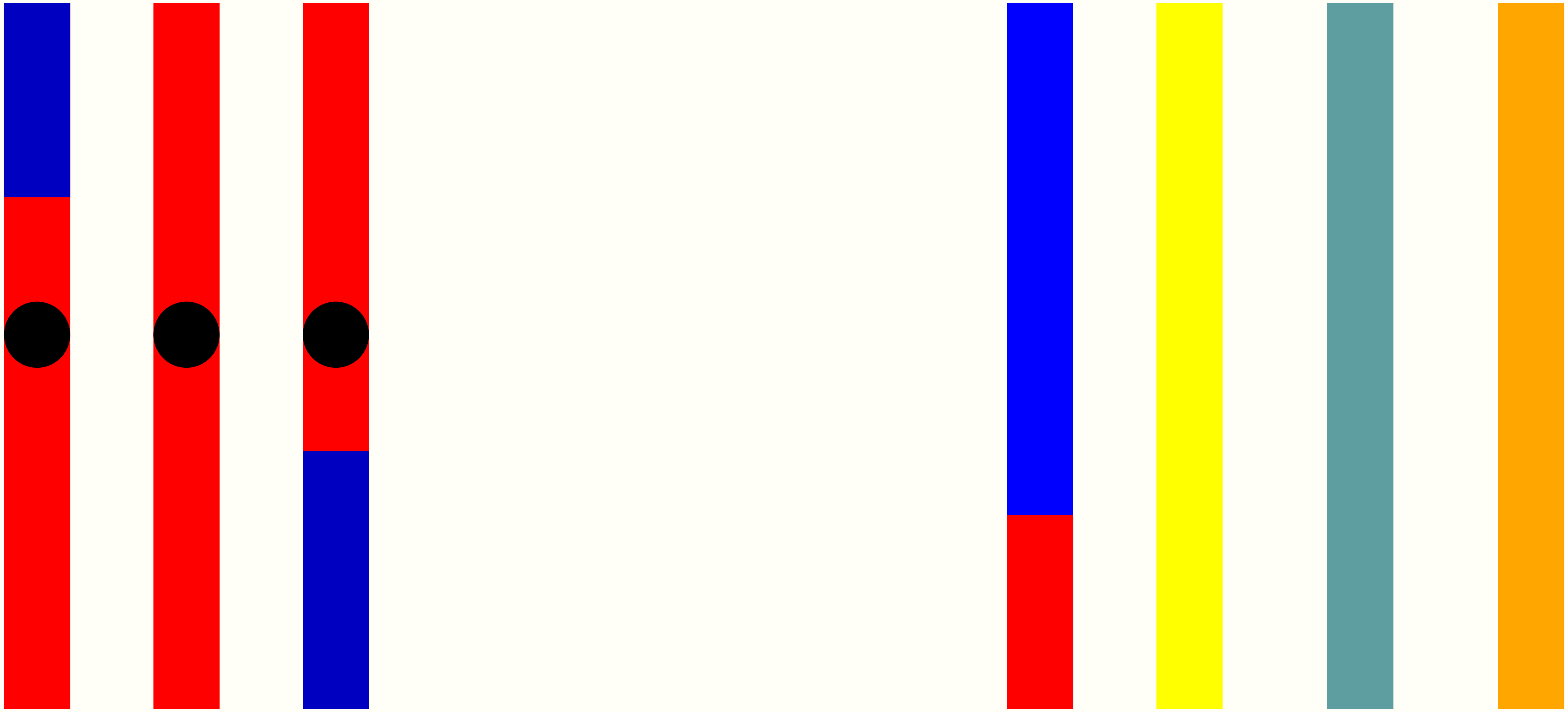


Case

Controls

Breakdown of linkage disequilibrium

Generation 1

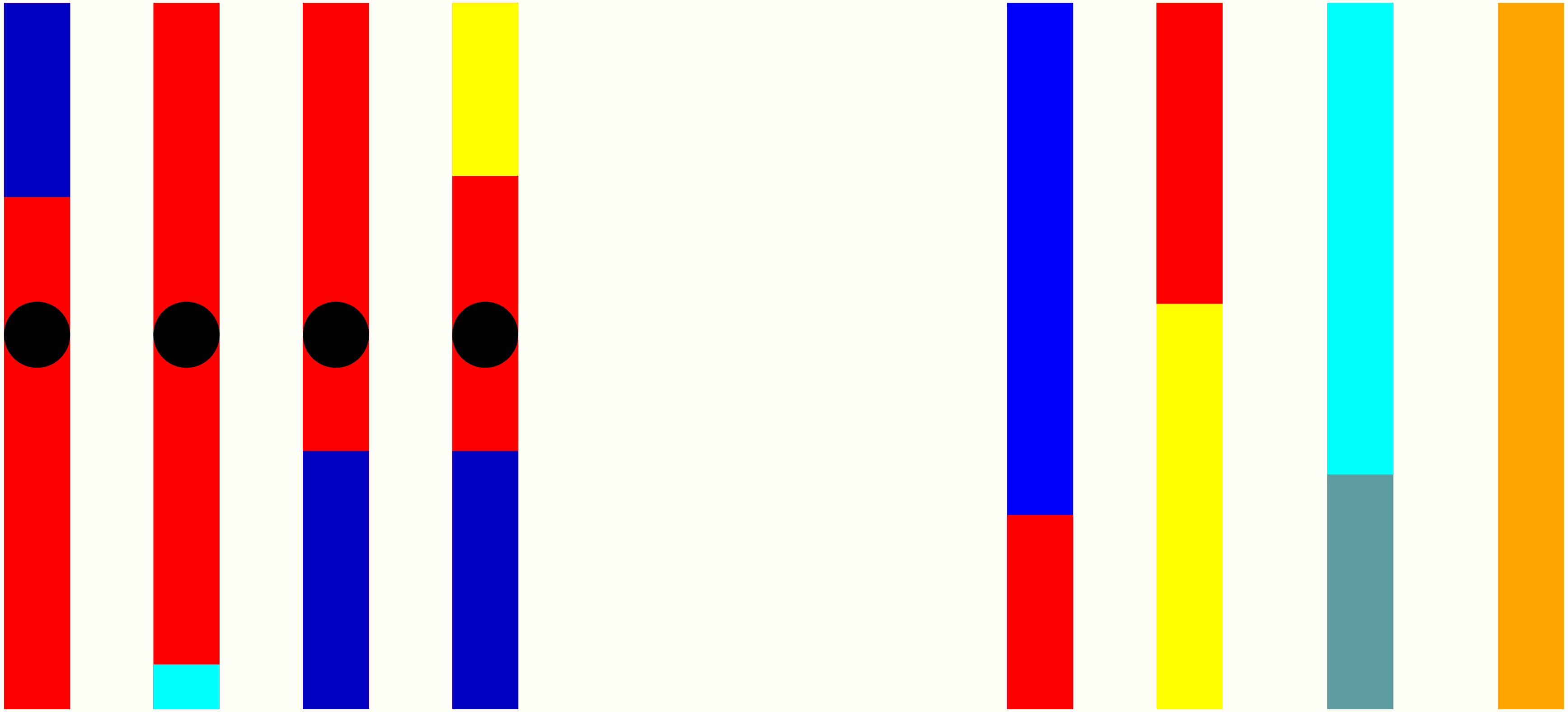


Cases

Controls

Breakdown of linkage disequilibrium

Generation 5

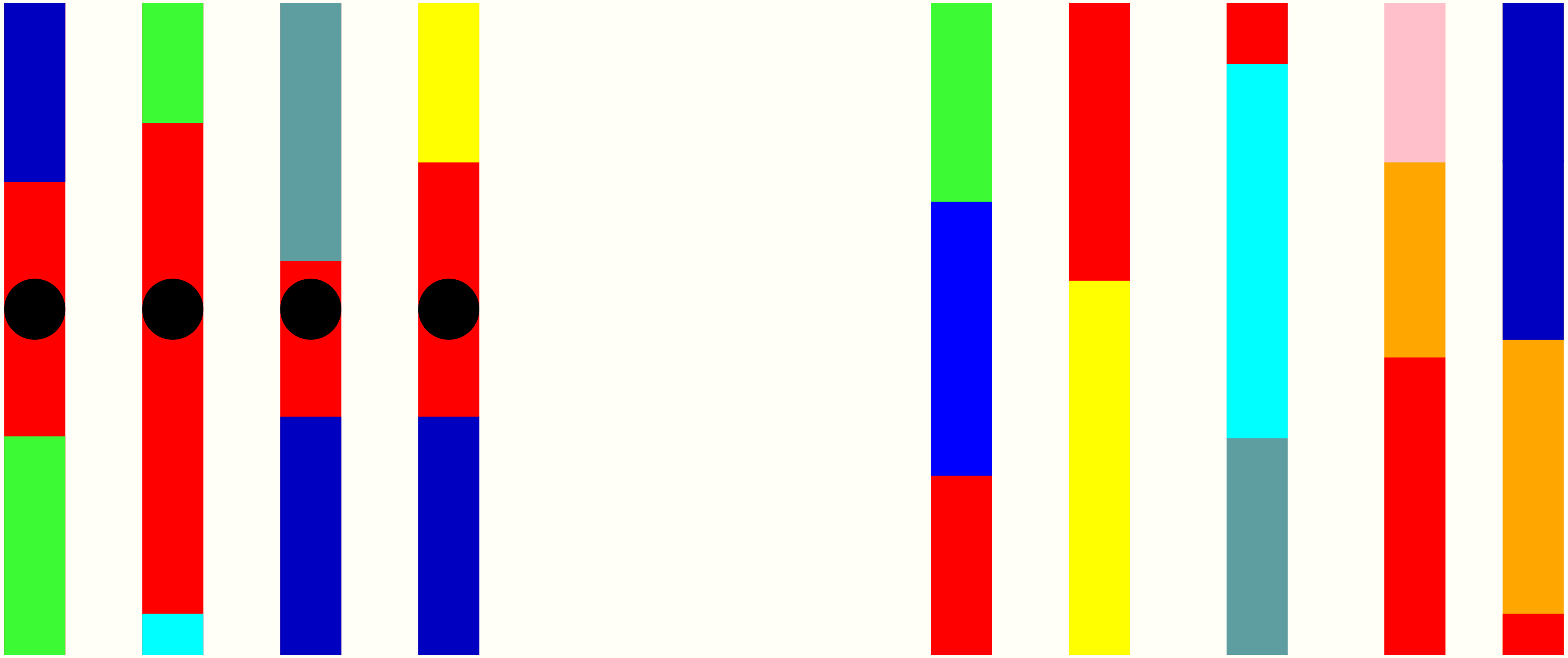


Cases

Controls

Breakdown of linkage disequilibrium

Generation 10

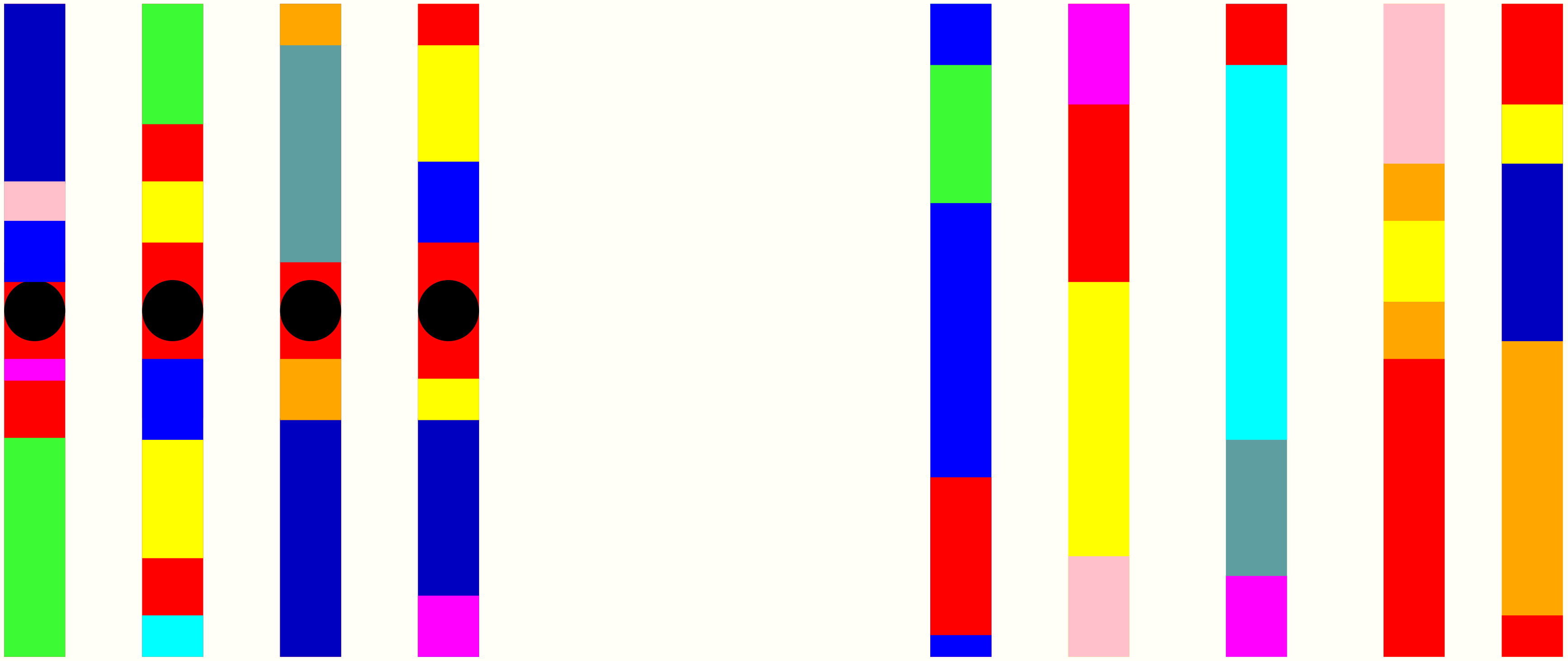


Cases

Controls

Breakdown of linkage disequilibrium

Generation 100

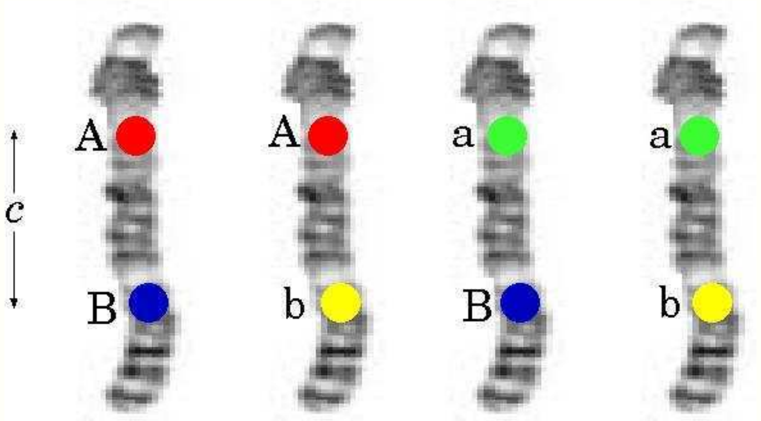


Cases

Controls

Expected length of disease haplotype ~ 1/G

Linkage disequilibrium: two diallelic loci



	B	b	Total
A	x_1	x_2	P_A
a	x_3	x_4	P_a
Total	P_B	P_b	1.0

The usual measure of linkage disequilibrium is:

$$D = x_1 - P_A P_B.$$

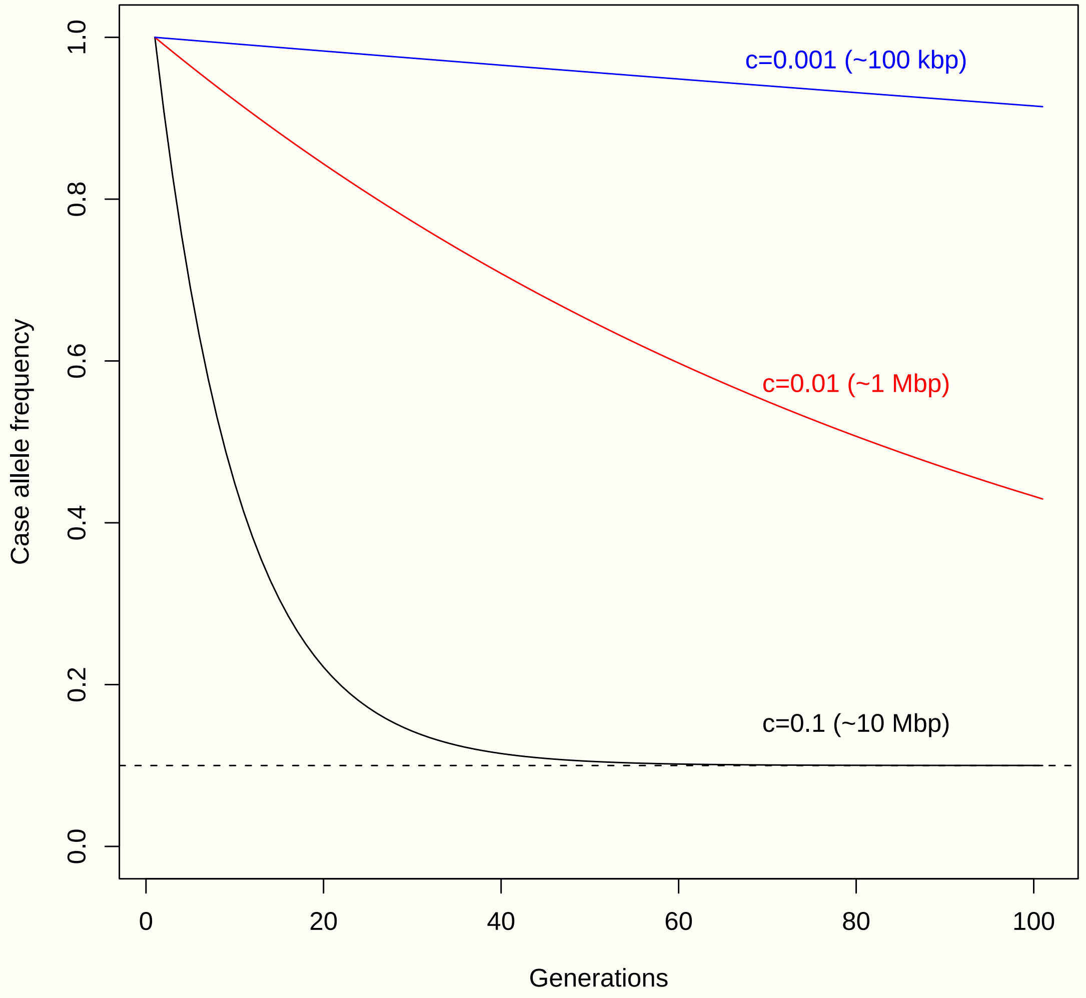
With each generation, *D* diminishes [Jennings 1917],

$$D^{(t)} = (1 - c)^t D^{(0)}$$

For loci separated by a recombination distance (*c*) of 1%, a 50% decrease in *D* will take 69 generations.

Linkage disequilibrium: two diallelic loci

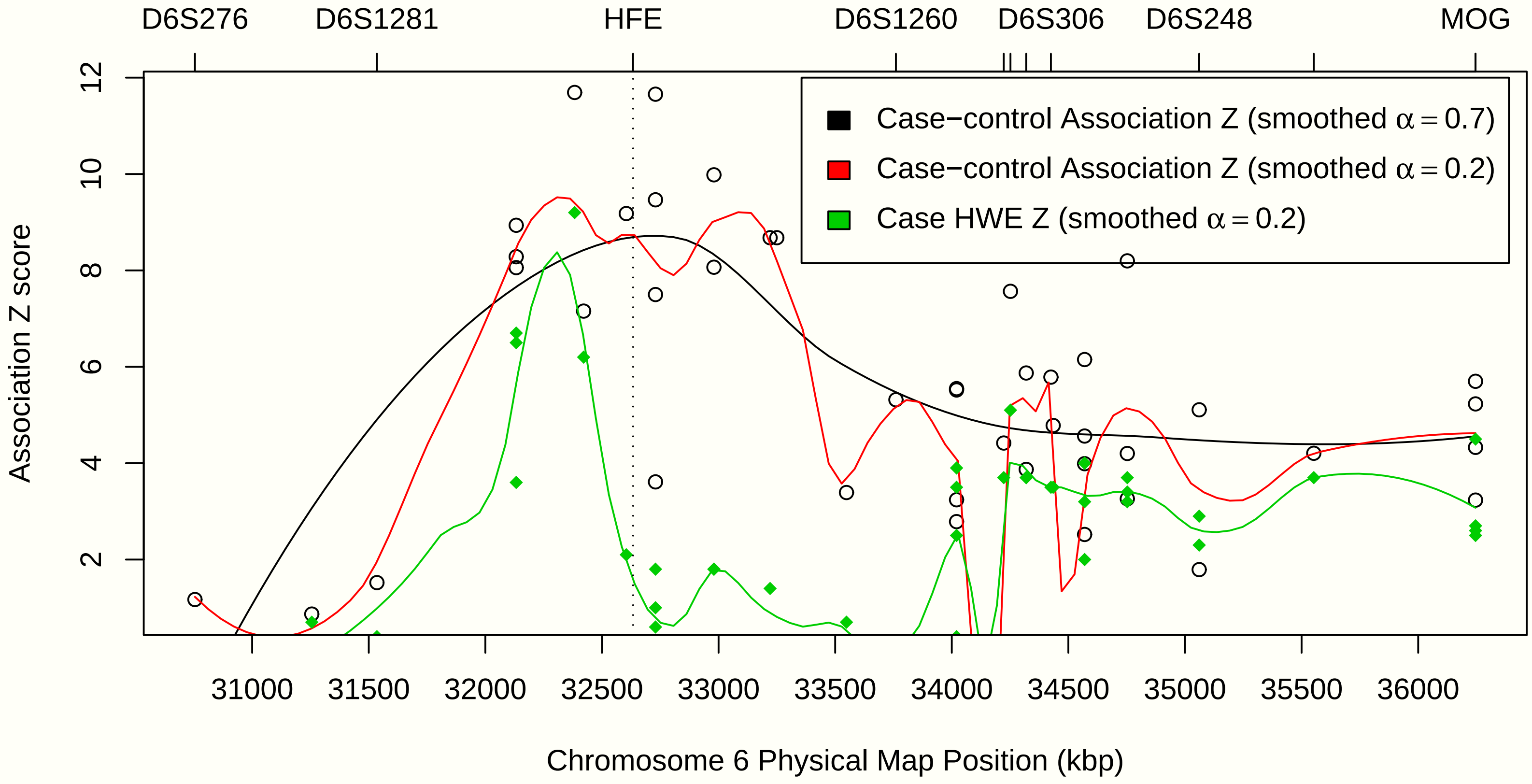
Relationship between marker frequency in cases and generation. Model assumes marker allele frequency 10%, and a rare dominant gene.



Linkage disequilibrium: marker locus and a trait

At a practical level, this is straightforward. We usually ignore the fact that the marker allele is not the causative variant, and test the strength of the relationship between the phenotype value and individual genotype.

Generally, the closer the marker is to the trait locus, the stronger the association to the phenotype.



Association analysis

Phenotype	Data	Model	Association measure	Test Statistic
Dichotomous	Cross-classified counts of affecteds versus genotype	Log-linear model	Risk ratio	Contingency chi-square test
		Logistic Regression	Odds ratio	Likelihood Ratio Test
Categorical	Cross-classified counts of trait class versus genotype class	Log-linear model	Risk ratio	Contingency chi-square test
Quantitative	Trait mean and standard error for each genotype class	Linear model	Genotype or allele deviation	F-test
Time to event (eg age at diagnosis)	Survival curve for each genotype class	CPH survival analysis	Hazard ratio	LRT

Ethnic Stratification

Population or **ethnic stratification** refers to the fact that frequencies of alleles at many loci differ between (human) populations originating from different geographical regions.

In a mixture of populations, alleles at different loci that are increased together in particular subpopulations will exhibit overall **extragametic allelic association**.

If a trait is associated with the culture or environment of a particular subpopulation, this too will give rise to overall extragametic association.

Given that most of the QTL effect sizes detected to date are relatively small (eg relative risk of 1.1-1.3), this means that **confounding** of this type can be a real problem.

Lactase persistence alleles and height

Campbell et al (2005) describe an example of stratification effects, the association between LCT-13910C>T and stature in a US population sample

	All	Subdivided by Grandparental Ancestry		
		Four US-born	Southeastern Europe	Northwestern Europe
Tall	65.6% (N=1123)	69.2% (N=645)	35.8% (N=127)	66.5% (N=351)
Short	57.1% (N=1056)	66.2% (N=637)	24.7% (N=227)	65.4% (N=192)
P-value	3.6×10^{-7}	0.098	0.0016	0.71

The association failed to replicate in more ethnically homogenous European samples or using family-based tests (which test for linkage *and* association).

This particular SNP (rs4988235) is known to vary markedly in frequency across ethnic groups.

LCT around the world

Population	LCT -13910C>T
Scandinavia	81.5%
Orkney Islands	68.8%
Basque	66.7%
French	43.1%
Balochi (Pakistan)	36.0%
North Italian	35.7%
Russian	24.0%
Mozabite (Algeria)	21.7%
Hazara (Pakistan)	8.0%
Sardinian	7.1%
Tuscan (Italy)	6.3%
Yoruba (Nigeria)	0.0%

Dealing with stratification

- Adjustment on reported ancestry
- Adjustment on marker-derived ancestry scores
- Genomic control
- Family based association analysis

If population stratification is a problem, then one approach to correcting for its effects is to include the individual’s ancestry as a covariate in the analysis.

One estimate of ancestry is based on asking the individual about the ancestry of each of their grandparents.

Alternatively, either a population genetic analysis of the study data, or an external dataset, can be used to identify genetic markers that are informative for ancestry (so-called “AIMs”).

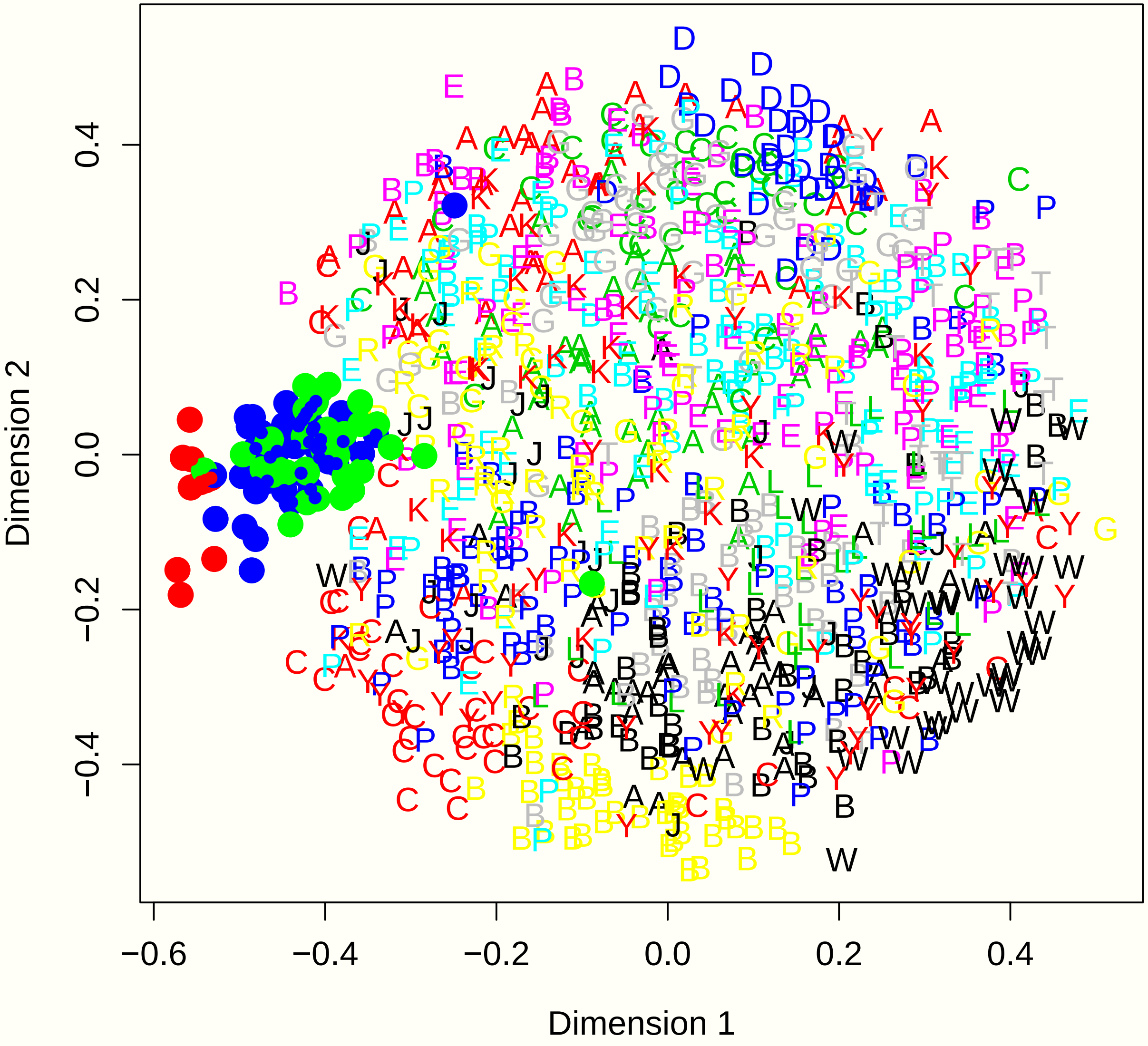
Multidimensional scaling analysis of multilocus identity-by-state

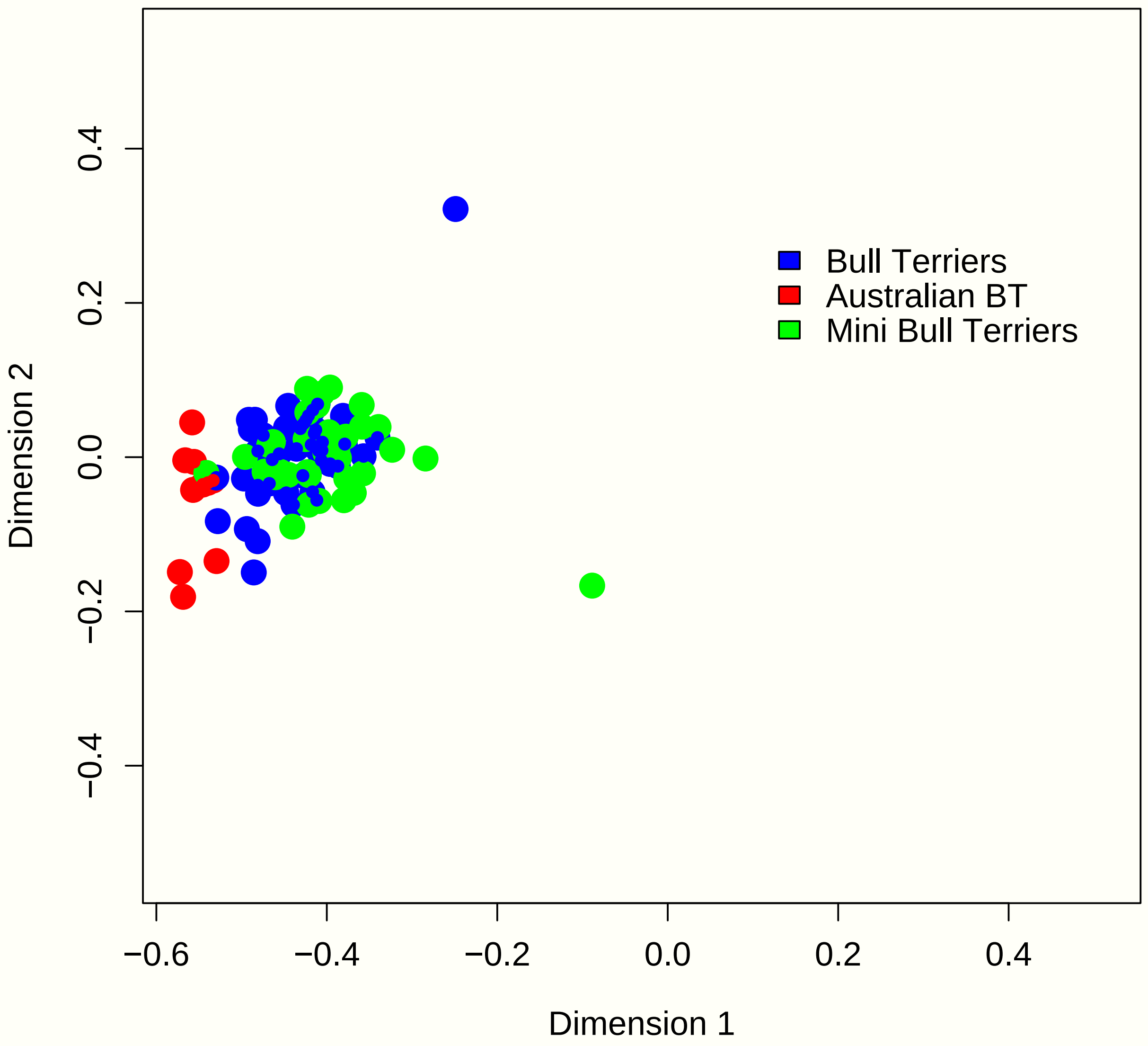
The average sharing of alleles at a large number of markers between pairs of individuals is a measure of relatedness. This empirical kinship matrix can be used to estimate genetic distances between all genotyped individuals, and from these positions of each individual in a relationship space. These can then be tested for the presence of clustering, where each cluster represents a subpopulation.

If membership of particular populations is already known, the clusters can be checked to see whether they successfully represent the genetic structure of the population.

Either a cluster membership probability score can be generated, or the coordinates of each individual on the first few principal dimensions of the genetic relationship space can be used as covariates in a association analysis.

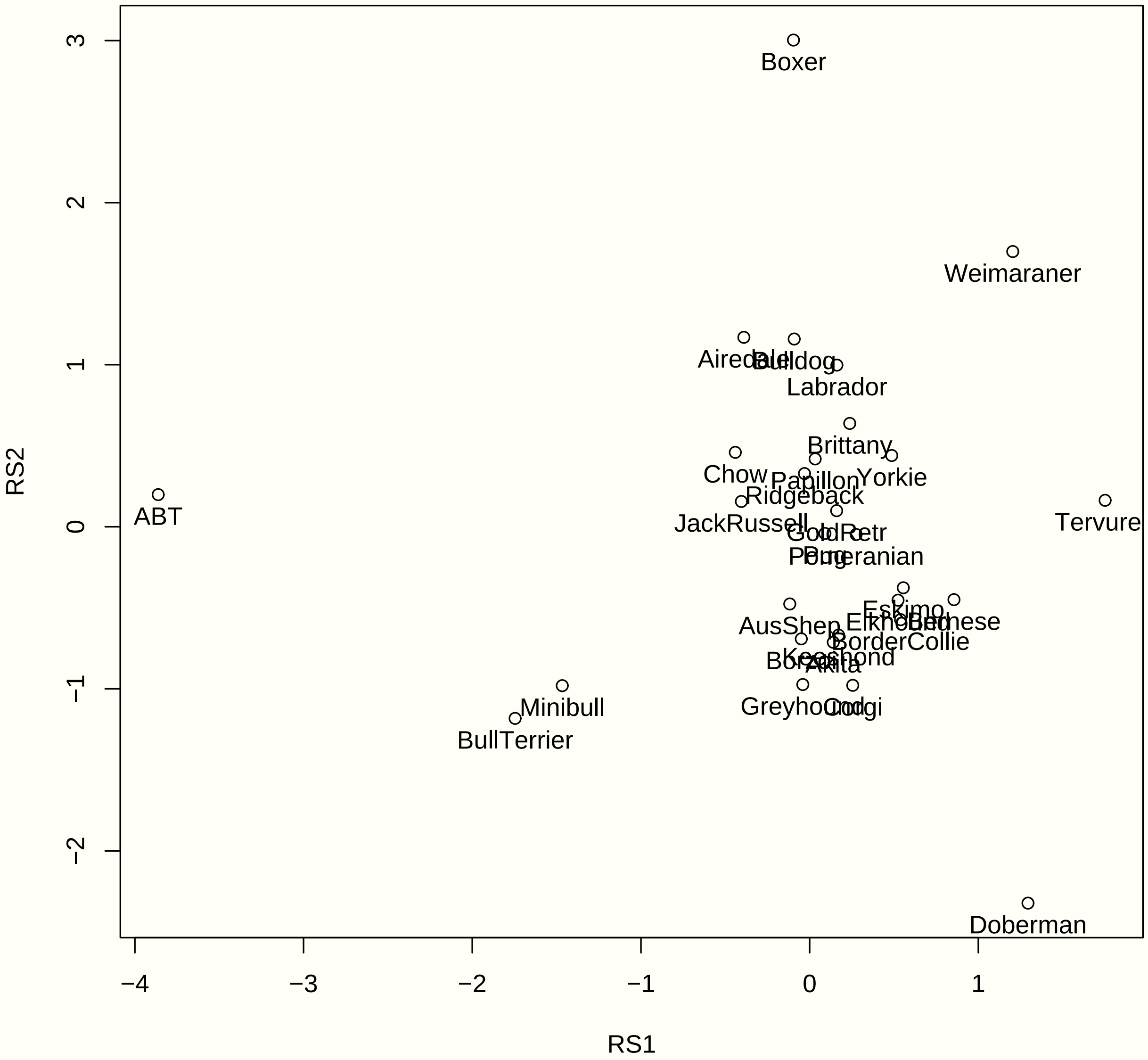
MDS Plot for different dog breeds



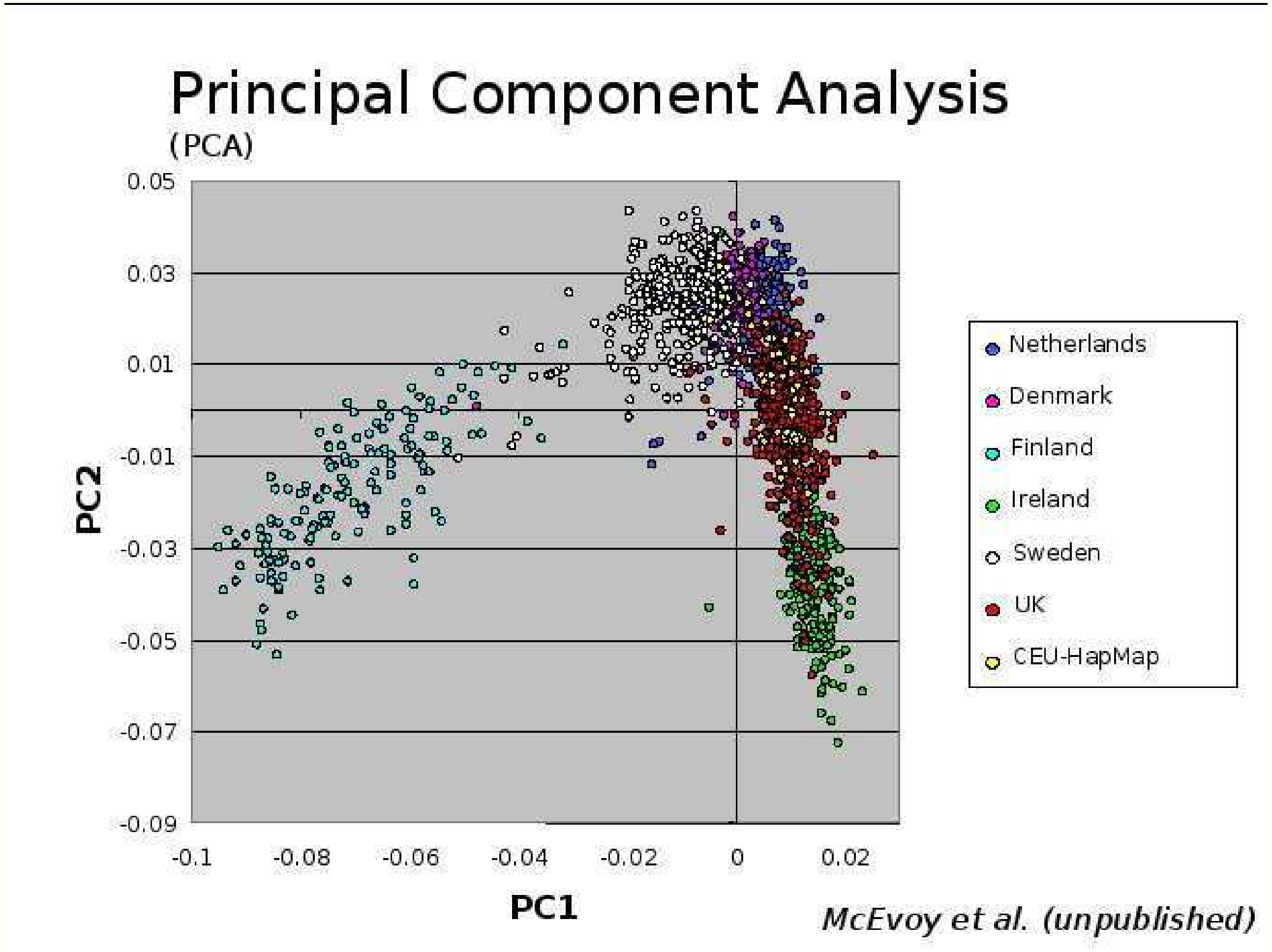


Bull Terriers
Australian BT
Mini Bull Terriers
Dimension 1
Dimension 2
−0.6
−0.4
−0.2
0.0
0.2
0.4
−0.4
−0.2
0.0
0.2
0.4

Plot of breed scores on first two principal components
extracted from interbreed genetic distances at 16 microsatellite markers



MDS Plot for different European populations

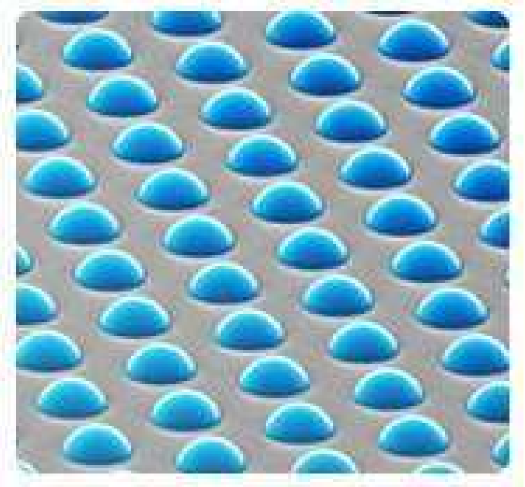


McEvoy et al. (unpublished)

High-throughput genotyping

Moore’s Law states that the number of transistors that can be placed inexpensively on an integrated circuit increases exponentially, doubling approximately every two years.

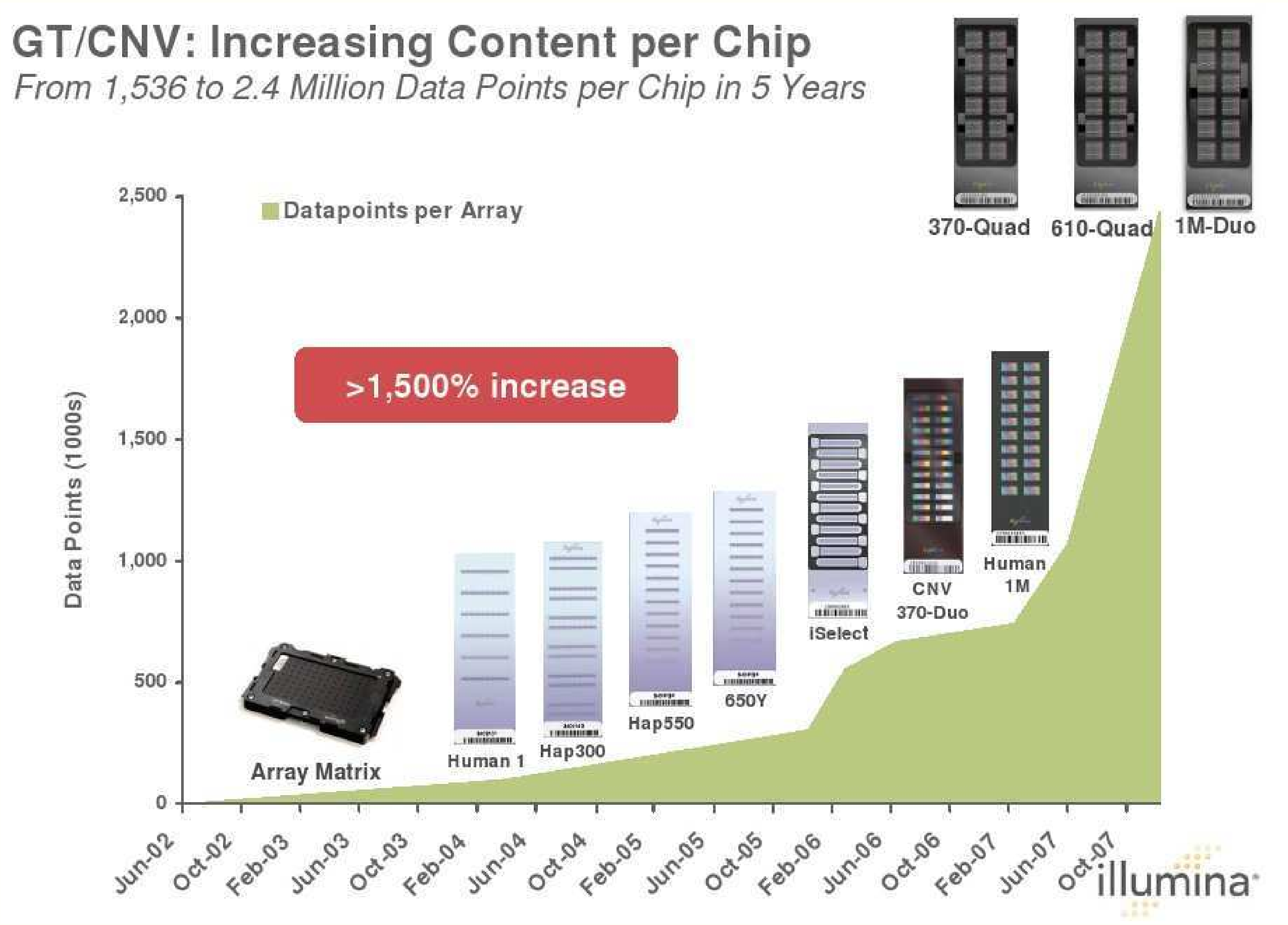
The same miniaturization trends are currently affecting genotyping technology.



Illumina BeadArray Technology is based on 3-micron silica beads that self assemble in microwells on silica slides, with a uniform spacing of 5.7 microns.

Each bead is covered with hundreds of thousands of copies of a specific oligonucleotide that act as the capture sequences for a particular STS.

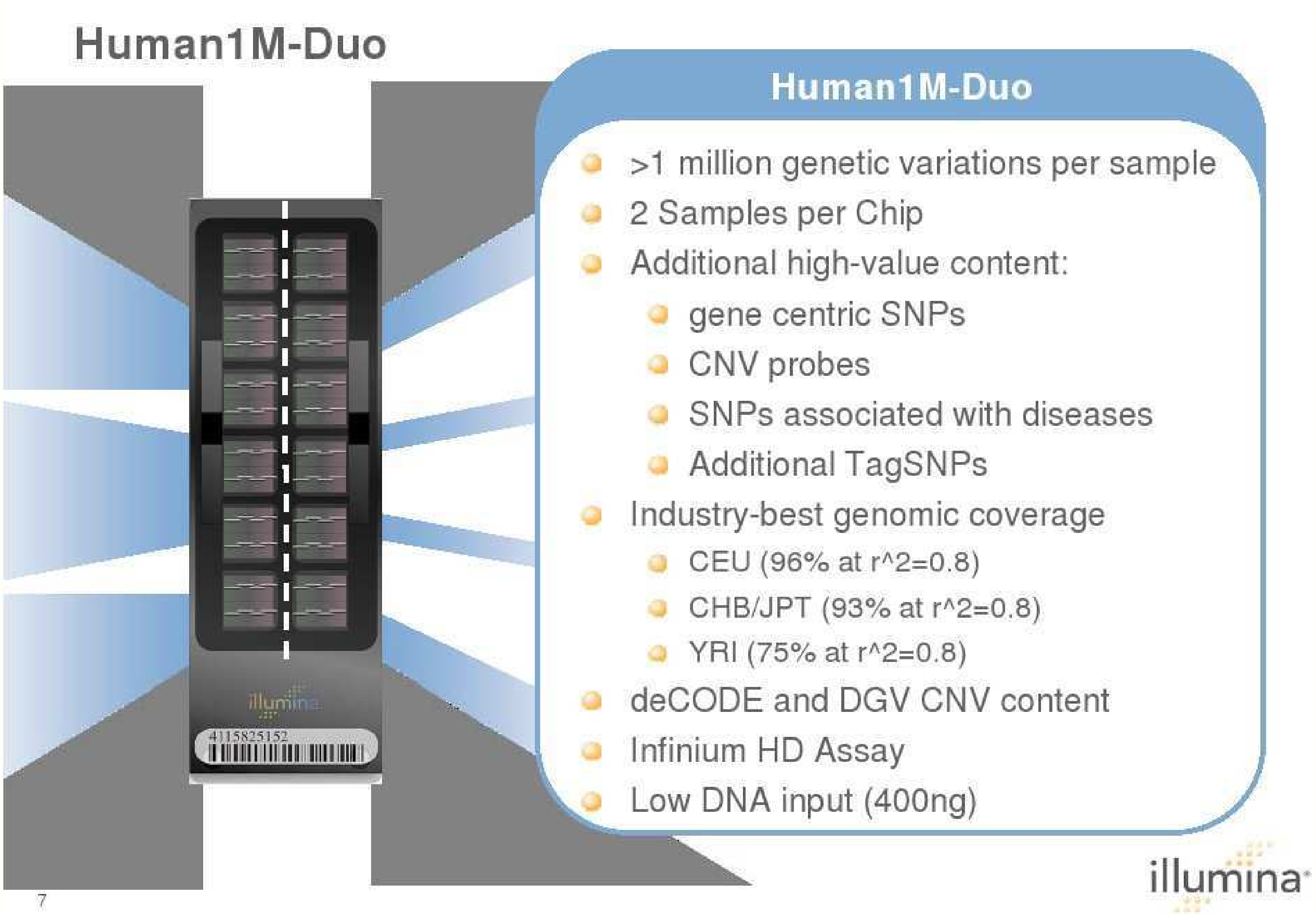
High-throughput genotyping



GT/CNV: Increasing Content per Chip

From 1,536 to 2.4 Million Data Points per Chip in 5 Years

High-throughput genotyping



Human1M-Duo

Human1M-Duo

- >1 million genetic variations per sample
- 2 Samples per Chip
- Additional high-value content:
  - gene centric SNPs
  - CNV probes
  - SNPs associated with diseases
  - Additional TagSNPs
- Industry-best genomic coverage
  - CEU (96% at r^2=0.8)
  - CHB/JPT (93% at r^2=0.8)
  - YRI (75% at r^2=0.8)
- deCODE and DGV CNV content
- Infinium HD Assay
- Low DNA input (400ng)

7

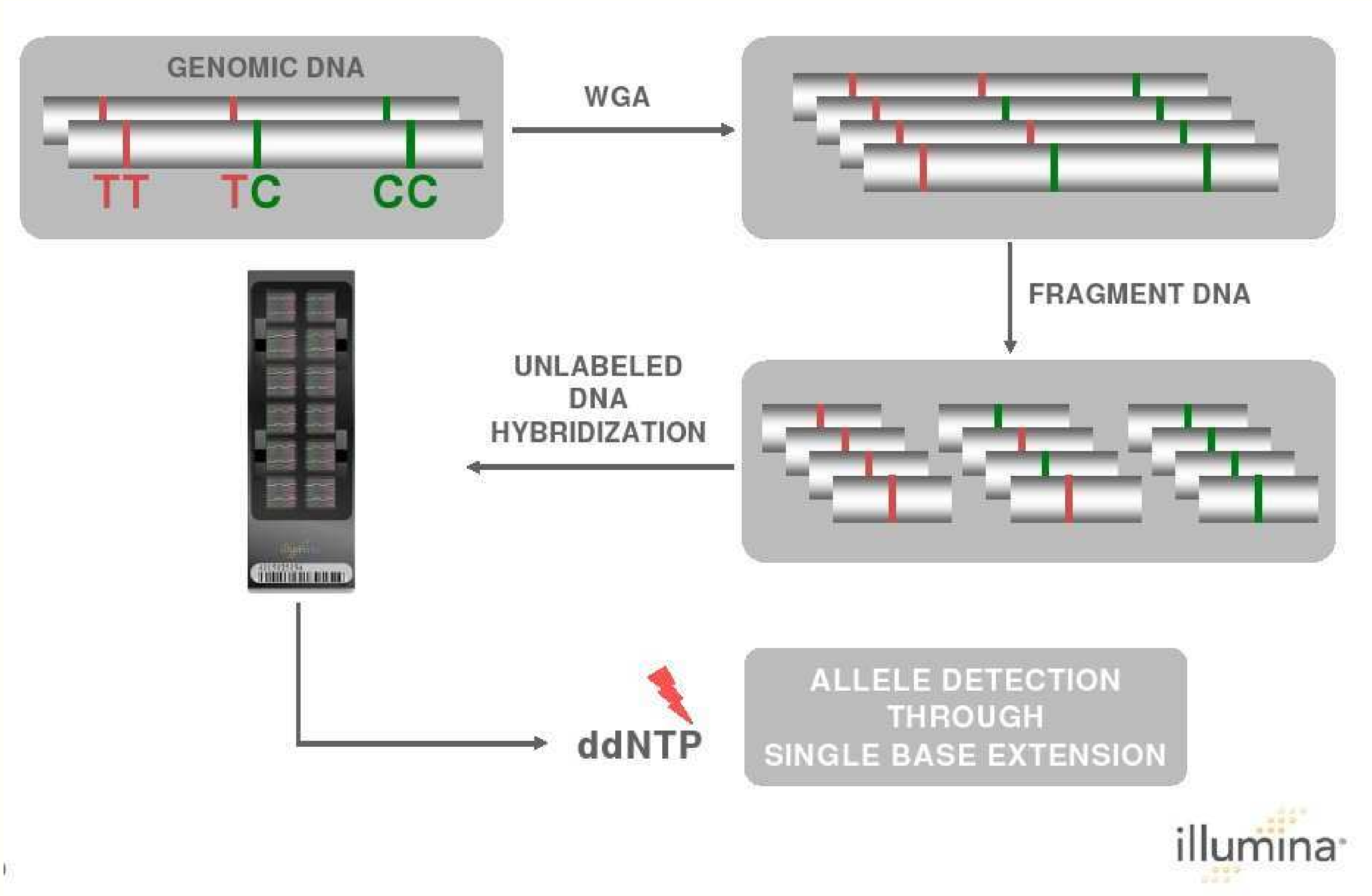
illumina

High-throughput genotyping

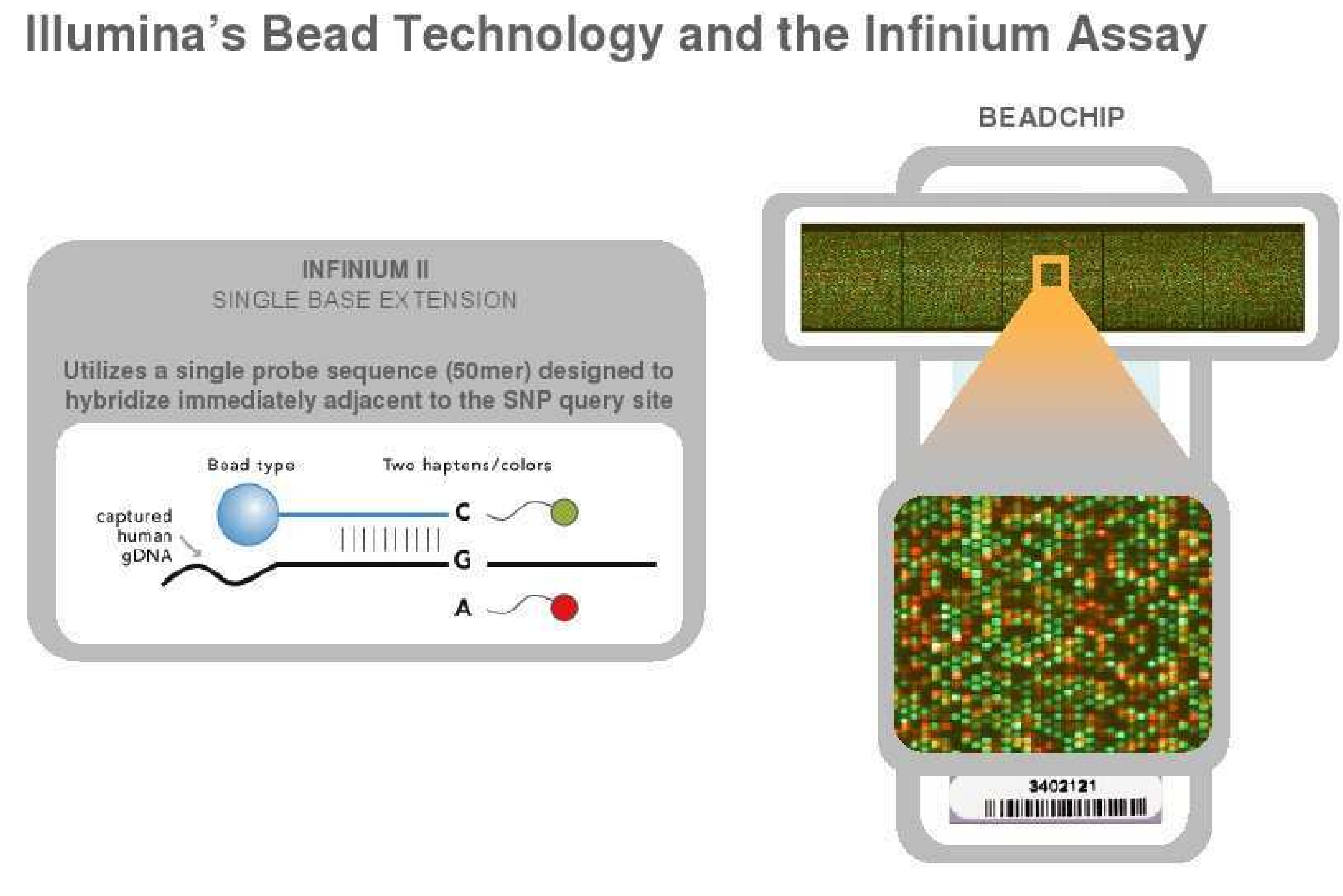
Affymetrix Genome-Wide Human SNP Array 6.0

- 906000 SNPs
- 946000 probes for CNVs
- 99.8% call rates
- Low DNA input (500 ng)

High-throughput genotyping



Illumina high-throughput genotyping



Illumina's Bead Technology and the Infinium Assay

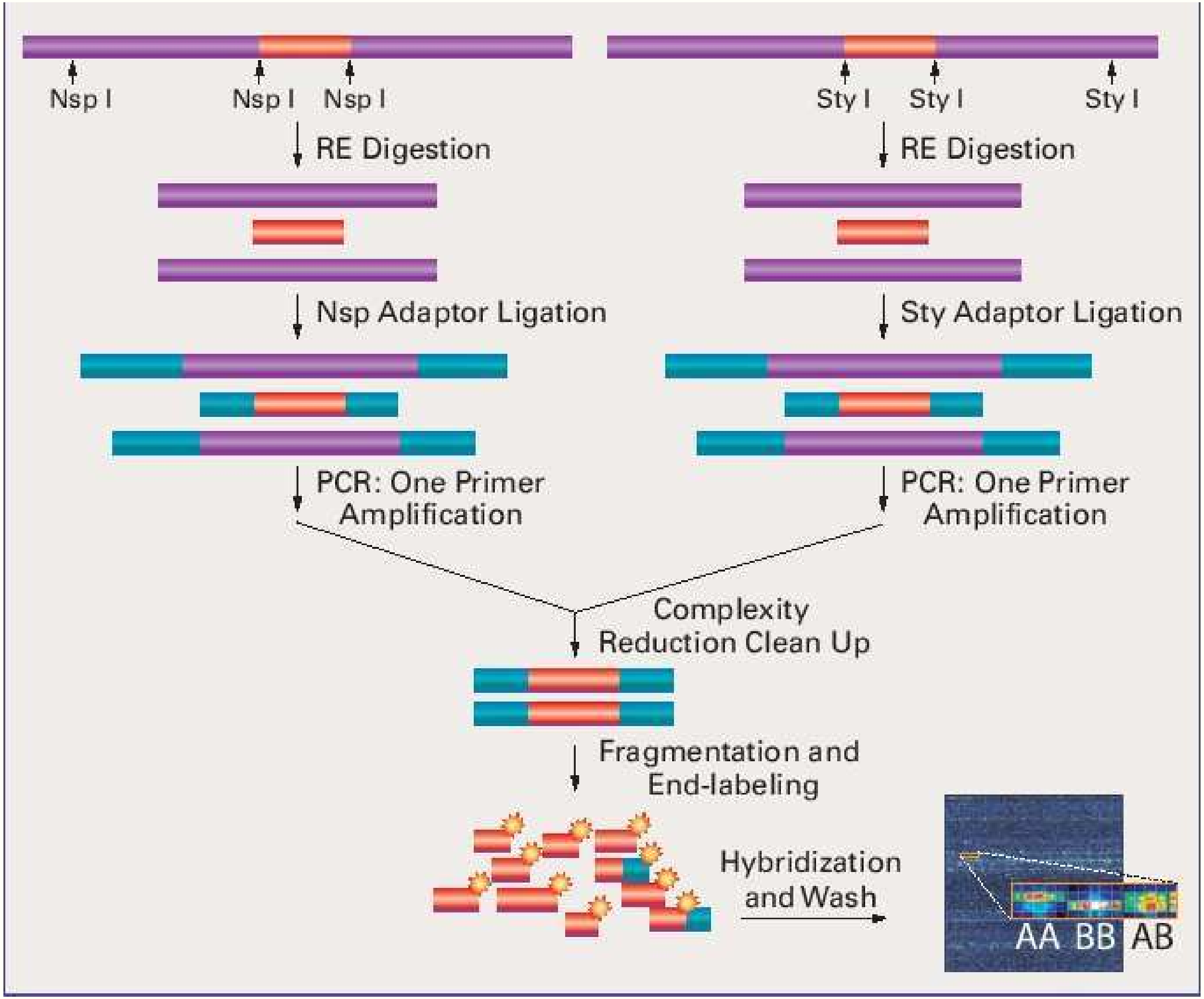
INFINIUM II

SINGLE BASE EXTENSION

Utilizes a single probe sequence (50mer) designed to hybridize immediately adjacent to the SNP query site

BEADCHIP

Affymetrix high-throughput genotyping



Genome-wide association

Over 240 GWAS publications to date (see http://www.genome.gov).

Appearing in January 2009 (according to Pubmed):

Phenotype	Reference	N Individuals	N SNPs
Alzheimers	Dement Geriatr Cogn Disord. 27: 59-68.	1088	2578
Alzheimers	Nat Genet. 2009 Jan 11.	2099	313K
Alzheimers	Am J Hum Genet.84:35-43.	1000	550K
Alzheimers	Mol Psychiatry. 2009 Jan 6.	2099	313K
Kawasaki Disease	PLoS Genet 5(1):e1000319	254 (+ 585)	250K
Lp(a)	J Lipid Res. 2009 Jan 5.	386	250K
Ulcerative Colitis	Nat Genet. 2009 Jan 4.	3600	250K
Prostate Cancer	Cancer Res 69:10-5.		
Juvenile idiopathic arthritis	Arthritis Rheum. 60:258-63	400	
Hypertension	PNAS 106:226-31	542	100K
Mean Platelet Volume	Am J Hum Genet. 84:66-71.	1644	500K
Transferrin level	Am J Hum Genet. 84:60-65.	1200	300K

Characteristics of GWAS

Genome-wide

- Large amounts of data
- Large numbers of markers
- Large numbers of statistical tests

Association

- Confounding by ethnic stratification
- Localization of causative variants

Data cleaning and validation

Always important in genetics, but what to do with 500K markers?

Use strict criteria to discard all data for suspicious markers: often 10-20% of the entire dataset. Since dense genotyping, usually have alternative marker from any given map interval.

- Assay failure rate (by marker, by individual)
- Hardy-Weinberg Disequilibrium, usually in controls (by marker)
- Mendelian inconsistencies (by marker, by individual)
- Agreement with appropriate population allele frequencies (by marker)
- Agreement with appropriate population haplotype frequencies (by marker)
- Rare minor allele (by marker) !?

Sources of error

- Poor quality of individual DNA samples: arrays require good quality DNA
- Laboratory or fieldwork sample mixups [there are always some]
- Pedigree errors: nonpaternities, informant confusion
- Poorly designed SNP assays
- SNP mapping errors: note realization about extent of duplications
- Misclassified phenotypes
- Data handling problems [where I usually err]

Assays problems often lead to miscalling of a heterozygote as one or other homozygote. This is why testing for HWE is informative.

The multiple testing problem

We usually assess believability of results of a study by calculating P-values, where if

> *T* is the measure of effect size of a particular SNP on a trait, say,
>
> *P* = Probability of a result greater than or equal to *T*, **if** the given SNP does not really have any effect.

That is, any difference between *T* and 0 is just due to “noise” in the experiment. Mendelism is one source of such noise in observational studies.

So, the P-value is an estimate of a false positive result (“**Type I error** rate”) given that the SNP is not truly associated.

By common consent, a 5% chance of following up on a false positive is regarded as an acceptable risk. Equivalently, setting a **critical P-value** of 5% means that we expect 5 out of 100 tests to be a false positive.

Experiment-wise error

If our experiment involves 500000 independent tests,

Critical threshold	Expected False Positives
0.05	25000
0.01	5000
0.001	500
1×10^{-4}	50
1×10^{-5}	5
1×10^{-6}	0.5
5×10^{-7}	0.25
1×10^{-7}	0.05

Currently, the consensus is that we want to keep the number of expected false positives per GWAS well below even 1, so a critical P-value of 5×10^{-7} is commonly used.

The effective number of tests

Because of linkage disequilibrium, results of association tests of adjacent SNPs are correlated.

That is, if one SNP in a region gives a false positive result, then you will obtain false positives for all other SNPs in the same LD block. Therefore, we are actually performing fewer tests than the nominal 500000.

Moskvina and Schmidt (2008) for instance, estimated that a 500K Affy scan is equivalent to 277000 independent tests. Based on this analysis, a critical P-value of 1.8×10^{-7} gives a genome-wide Type I error rate of 5%.

Power of a GWAS

Power refers to the **true positive** probability, for a effect of a specified size. As we choose stricter thresholds to minimize the false positive rate, this also decreases the true positive rate.

The false positive rate is uncorrelated with the number of individuals in an association study.

The true positive rate increases with the number of individuals in the study, but so do the study costs.

To control costs, we can use a **two-stage design**:

- Screen all the SNPs in a subset of the sample
- Genotype the most significant SNPs in the rest of the sample.
- Combine the data and analyse together

This gives close to the same power as just genotyping all the SNPs in all the study participants.

Example power calculations

If there are 100 QTLs controlling a binary trait, each with a relative risk of 1.2, and we study 2000 cases and 2000 controls,

Critical threshold	Expected False Positives	Expected True Positives (out of 100)		
		Risk allele 20% frequency	Risk allele 10% frequency	Risk allele 5% frequency
0.05	25000	99	82	50
0.01	5000	96	61	27
0.001	500	85	33	9
1×10^{-4}	50	67	15	3
1×10^{-5}	5	46	6	0.7
1×10^{-6}	0.5	28	2	0.2
5×10^{-7}	0.25	24	1.5	0.1
1×10^{-7}	0.05	16	0.7	0.03

Example power calculation in R

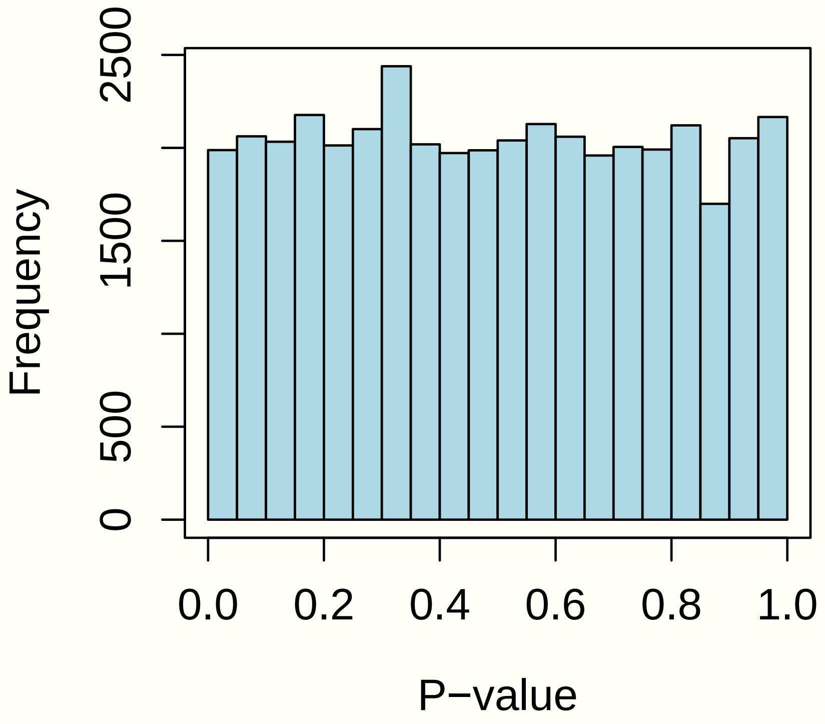
The results in the above table were generated using R:

```
rr <- 1.2

freq <- 0.05

alpha <- c(0.05, 0.01, 0.001, 1e-4, 1e-5, 1e-6, 5e-7, 1e-7)

power.prop.test(p1=freq,    # control allele frequency
                p2=rr*freq, # case allele frequency
                n=4000,     # chromosomes
                sig.level=alpha)
```

The empirical distribution of test results

We can compare the observed distribution of our 500000 test statistics to that under the **null hypothesis** of no QTLs.

Under that null hypothesis, all the P-values come from the uniform distribution, or the test statistics come from the appropriate equivalent distribution, such as the central chi-square.



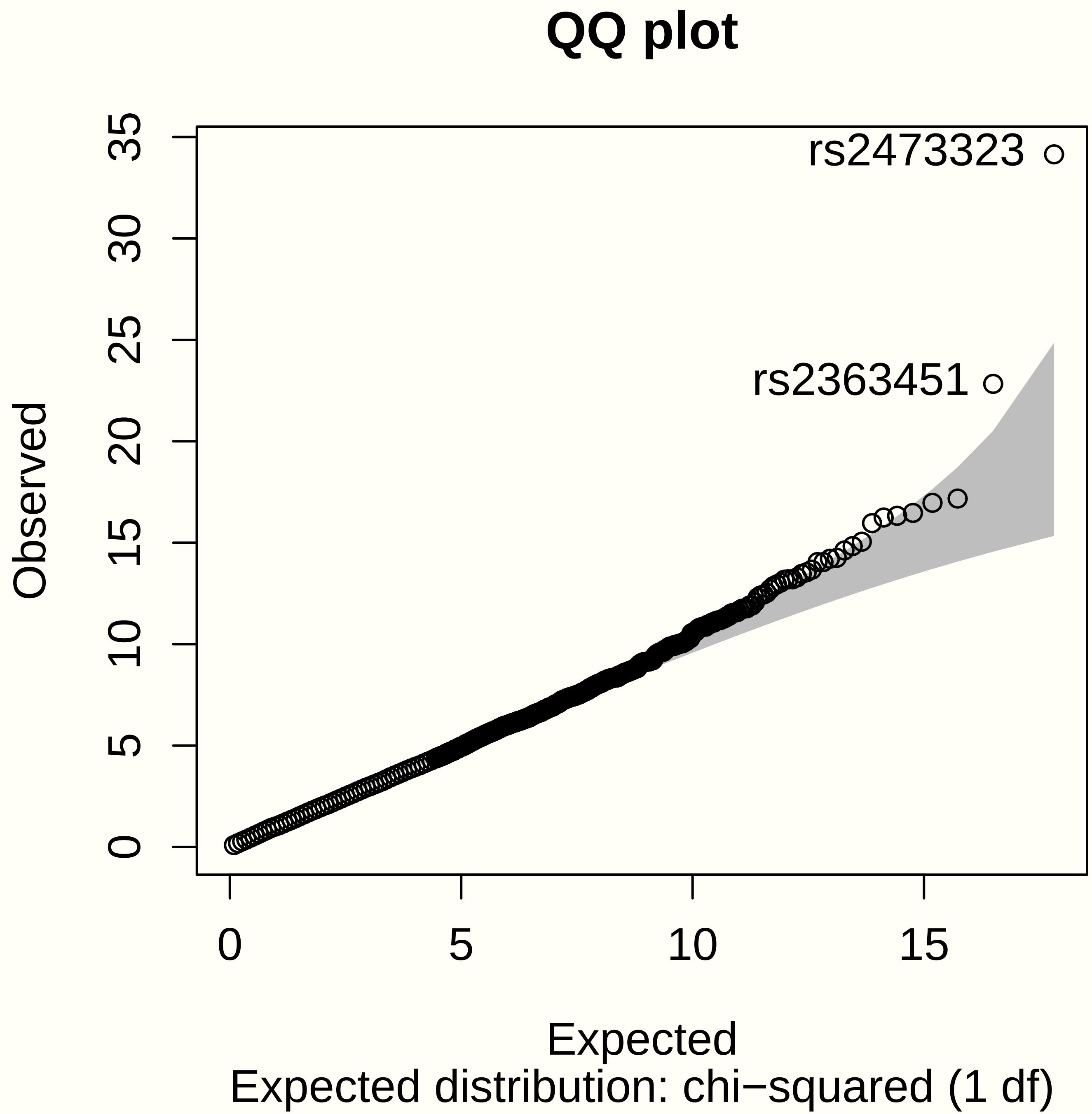
The Quantile-Quantile plot of test statistics

A nice graphical representation of all the test results is the Q-Q plot of the observed statistics distribution versus the expected distribution under the null. To get this, we order the results or P-values by size.

For example, the expected value for the 200th out of 500000 P-values would be 200/500000 and this is compared to the observed 200th best P-value. For a chi-square, it will be the chi-square value corresponding to a P-value of 200/500000.

The observed and expected results should fall along a straight line. We can put a **confidence envelope** around this line to highlight any interesting results.

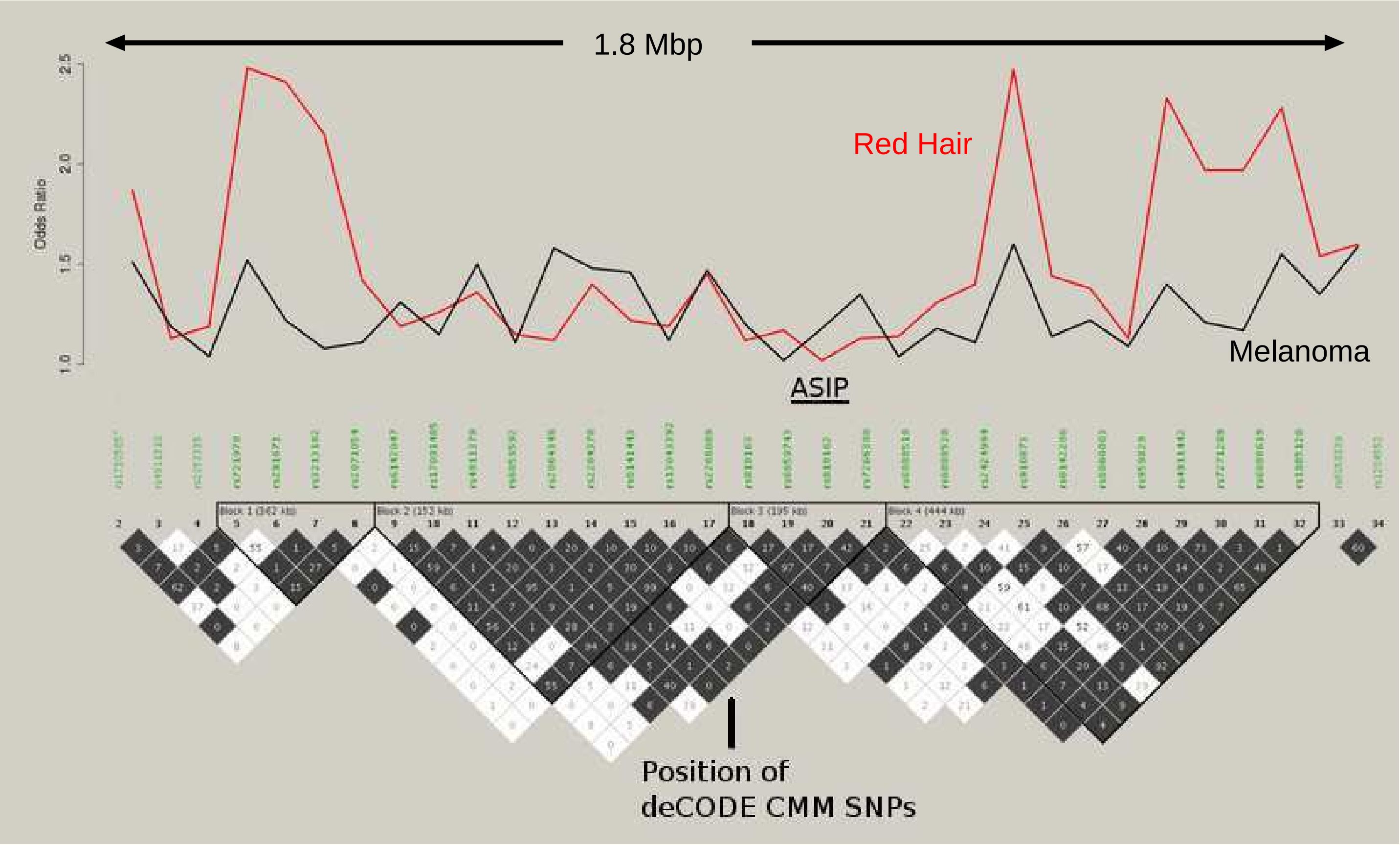
Ideally, we will see a few results that are higher than expected under the null hypothesis up at the top of the distribution. If we saw a large number of outliers, we might suspect ethnic stratification.


QQ plot
Observed
0
5
10
15
20
25
30
35
rs2473323
rs2363451
0
5
10
15
Expected
Expected distribution: chi−squared (1 df)

Linkage disequilibrium between SNPs

Given the density of SNPs in a modern GWAS, the intermarker distances are small, and so significant linkage disequilibrium is common. In some regions, LD extends over long regions, so a number of adjacent SNPs may be associated to a trait.

This can make it difficult to localize the causative locus or variant within a large gene.



Long haplotypes and disease association

Brown et al (2008) carried out a DNA pooling GWAS for cutaneous malignant melanoma.

The best and second best P-values were obtained from SNPs on chromosome 20, and additional SNPs in that region were subsequently genotyped.

Association to other SNPs in the same region were reported independently by Gudbjartsson et al (2008). I was able to show that these are in strong LD with the SNPs reported by our group.

Haplotype rs17305657-rs1885120	Haplotype Frequency		Association
(1.77 Mb long)	Cases	Controls	P-value
*C*AC*AC*TCCGATCTCAATGAACC*T*TCTA*CA*T*C*	0.073	0.040	7.9e-6
TACGTTTCGATCTCAATAAATCCCCTGTGTG	0.052	0.048	0.60
TATGTTTTGCTCCCCGTGAACTCCTCATGCG	0.041	0.049	0.24
TACGTTTTGCTCCCCGTGAACTCCTCATGTG	0.033	0.036	0.56
TGCGTTTCGATCTCAATAAATCCCCTGTGTG	0.024	0.026	0.71
TATGTTTCGATCTCAATAAATCCCCTGTGTG	0.024	0.025	0.91
TATGTTCTGCTCCCCGTGAACTCCTCATGTG	0.018	0.026	0.11
TACGTTTTGCTCCCCGTGAACTCCTCATGCG	0.021	0.024	0.56
TGCGTTTCGATTTCAATAAATCCCCTGTGTG	0.021	0.022	0.75
TGCGTTTTGCTCCCCGTGAACTCCTCATGTG	0.018	0.021	0.44
TATGTTTTGCTCCCCGTGAACTCCTCATGTG	0.021	0.018	0.51
TGCGTTCCCCTCCCAGGATACC*T*CCTA*CA*TG	0.016	0.021	0.19
TATGTTTCGATTTCAATAAATCCCCTGTGTG	0.017	0.018	0.88
TATGTTTTGCTCCGAGTGAACTCCTCATGTG	0.014	0.018	0.35
TACGTTTCGATCTCAATGAACC*T*TCTA*CA*T*C*	0.018	0.014	0.28
TGCGTTTTGCTCCCCGTGAACTCCTCATGCG	0.013	0.018	0.25
TATGTGTTGCTCCCCGTGAACTCCTCATGTG	0.012	0.017	0.29
TATGTGTTGCTCCCCGTGAACTCCTCATGCG	0.014	0.014	0.99
TATGTGTCGATTTCAATAAATCCCCTGTGTG	0.013	0.013	0.96
Other rare haplotypes	0.537	0.532	