Missing Heritability & GWAS

Nick Martin

Shaun Purcell

Peter Visscher (in absentia)

And all the faculty….

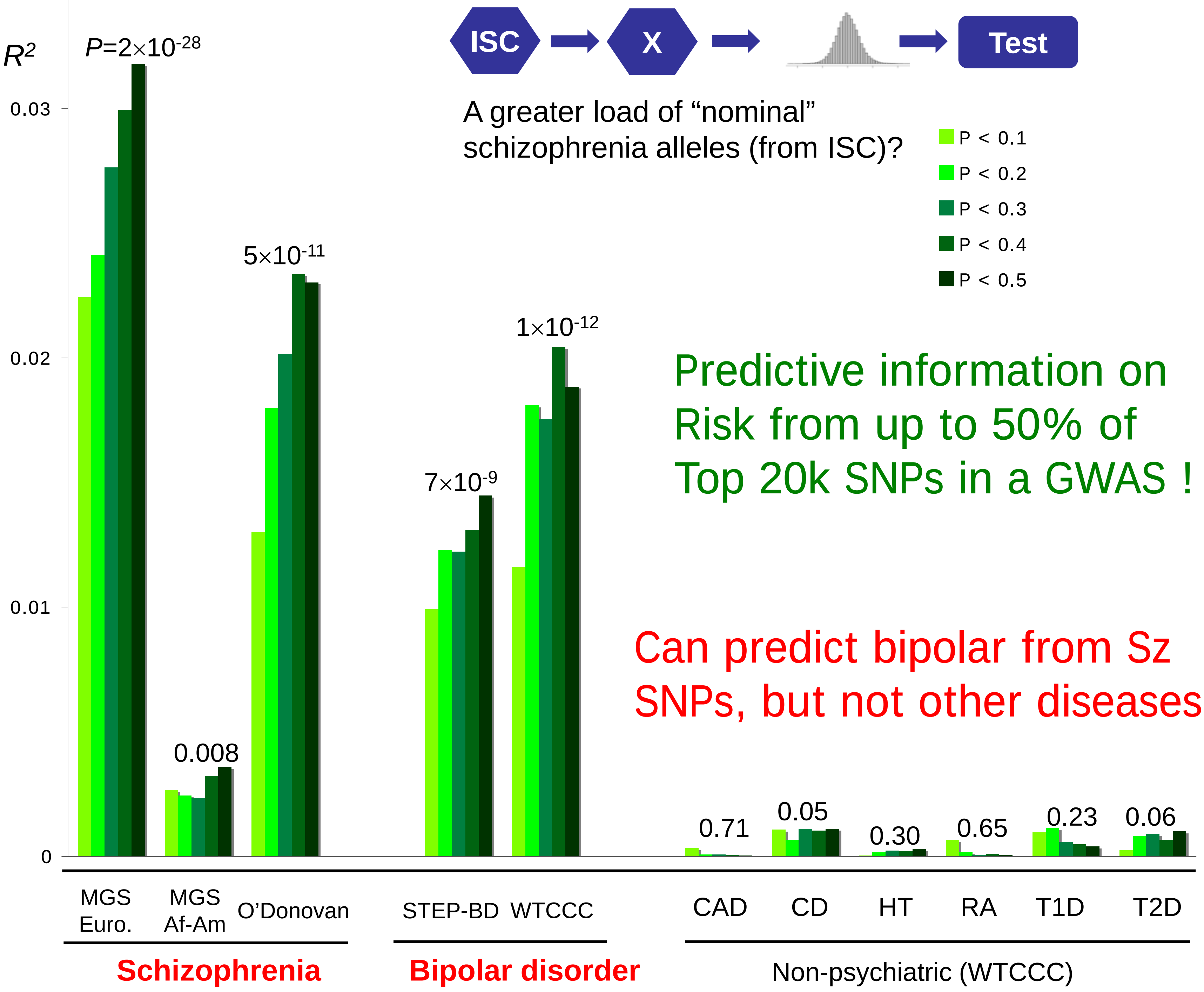
For most traits studies so far, GWAS is accounting for very little variance

Nat Genet. 2008 May;40(5):575-83

Genome-wide association analysis identifies 20 loci that influence adult height

Weedon MN, Lango H, Lindgren CM, Wallace C, Evans DM……..

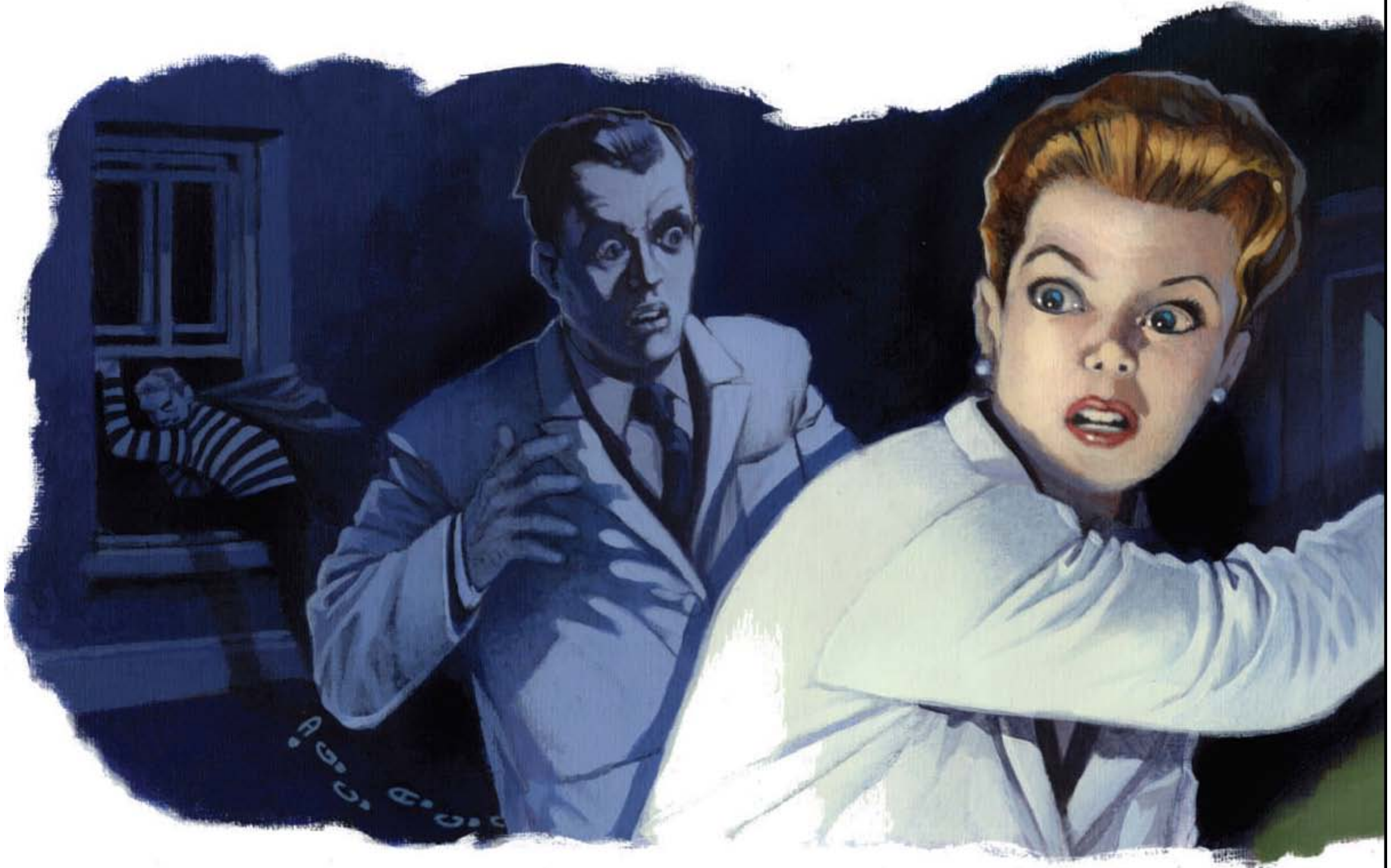
Adult height is a model polygenic trait, but there has been limited success in identifying the genes underlying its normal variation. To identify genetic variants influencing adult human height, we used genome-wide association data from 13,665 individuals and genotyped 39 variants in an additional 16,482 samples. We identified 20 variants associated with adult height (P < 5 x 10(-7), with 10 reaching P < 1 x 10(-10)). Combined, the 20 SNPs explain approximately 3% of height variation, with a approximately 5 cm difference between the 6.2% of people with 17 or fewer 'tall' alleles compared to the 5.5% with 27 or more 'tall' alleles. The loci we identified implicate genes in Hedgehog signaling (IHH, HHIP, PTCH1), extracellular matrix (EFEMP1, ADAMTSL3, ACAN) and cancer (CDK6, HMGA2, DLEU7) pathways, and provide new insights into human growth and developmental processes. Finally, our results provide insights into the genetic architecture of a classic quantitative trait.



A greater load of "nominal" schizophrenia alleles (from ISC)?

Predictive information on Risk from up to 50% of Top 20k SNPs in a GWAS !

Can predict bipolar from Sz SNPs, but not other diseases



The case of the missing heritability

When scientists opened up the human genome, they expected to find the genetic components of common traits and diseases. But they were nowhere to be seen. **Brendan Maher** shines a light on six places where the missing loot could be stashed away.

Possible explanations for missing heritability

(not mutually exclusive, but in order of increasing plausibility ?)

- Heritability estimates are wrong
- Nonadditivity of gene effects – epistasis, GxE
- Epigenetics – including parent-of-origin effects
- Low power for common small effects
- Disease heterogeneity – lots of different diseases with the same phenotype
- Poor tagging (1)
  - rare mutations of large effect (including CNVs)
- Poor tagging (2)
  - common variants in problematic genomic regions

Why do we care ?

- #1 biological question of the moment !
  - The death of genetic triumphalism?
  - But environmentalists should not crow – we are all ignorant
- Defines research agenda – what to do next ?
- Disease prediction – current best predictors are much worse than family history
- Intellectual curiosity
  - Fisher was right, but why ?

Possible explanations for missing heritability

(in order of increasing plausibility ?)

- Heritability estimates are wrong
- Nonadditivity of gene effects – epistasis, GxE
- Epigenetics – including parent-of-origin effects
- Low power for common small effects
- Disease heterogeneity – lots of different diseases with the same phenotype
- Poor tagging (1)
  - rare mutations of large effect (including CNVs)
- Poor tagging (2)
  - common variants in problematic genomic regions

Eaves LJ, Heath AC, Martin NG, Neale MC, Meyer JM, Silberg JL, Corey LA, Truett K, Walters E: Biological and cultural inheritance of stature and attitudes. In CR Cloninger Ed. *Personality and Psychopathology*, pp.269-308. American Psychiatric Press Inc., Washington, 1999

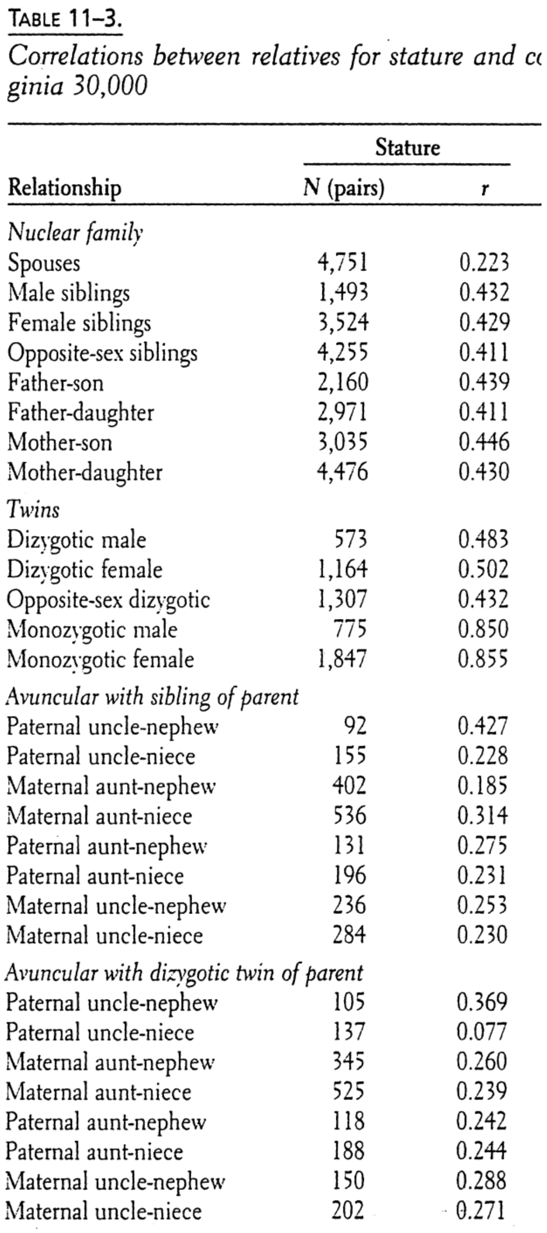


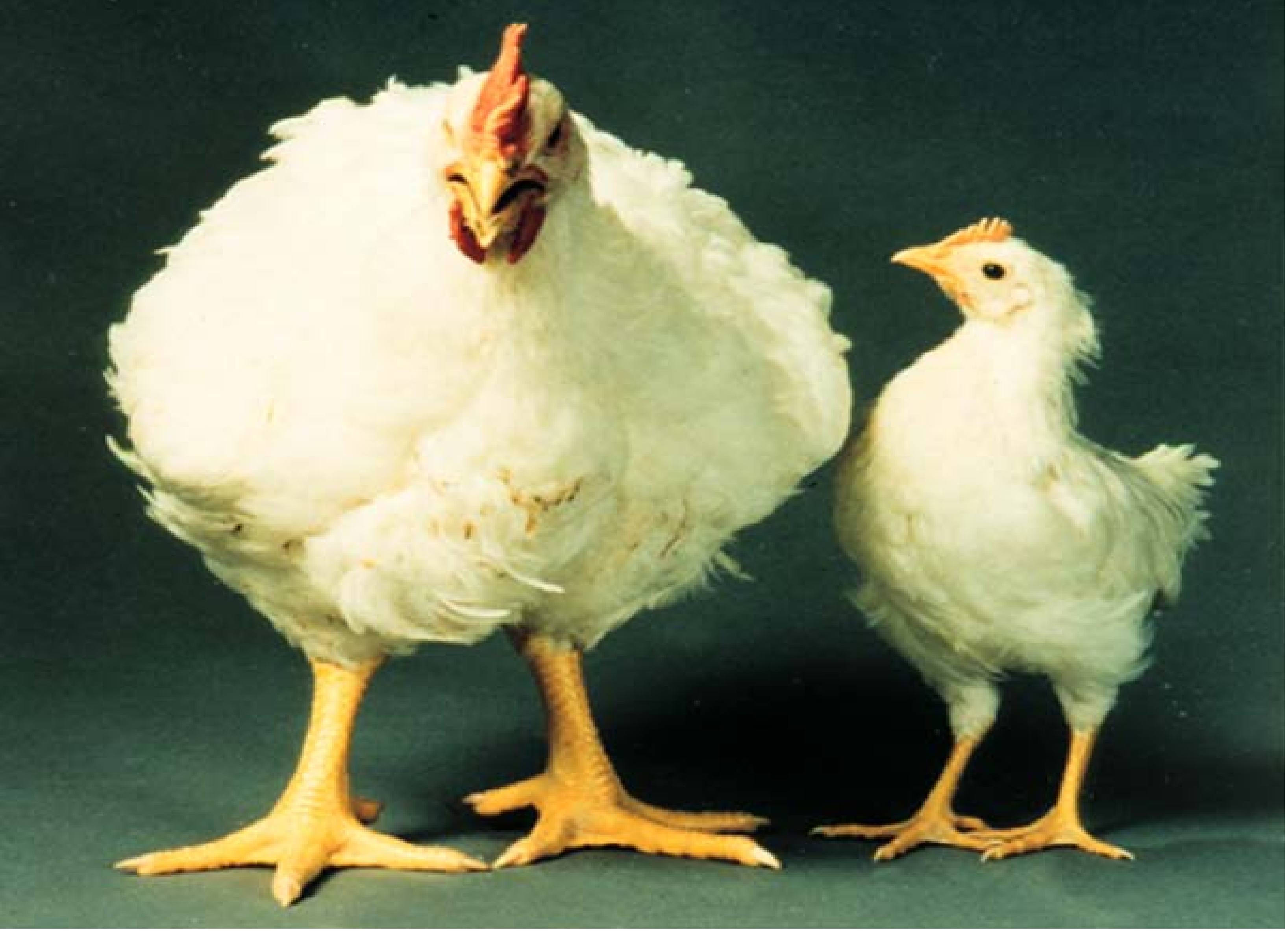
TABLE 11–3.

Correlations between relatives for stature and cc ginia 30,000

	Stature	
Relationship	N (pairs)	*r*
Nuclear family		
Spouses	4,751	0.223
Male siblings	1,493	0.432
Female siblings	3,524	0.429
Opposite-sex siblings	4,255	0.411
Father-son	2,160	0.439
Father-daughter	2,971	0.411
Mother-son	3,035	0.446
Mother-daughter	4,476	0.430
Twins		
Dizygotic male	573	0.483
Dizygotic female	1,164	0.502
Opposite-sex dizygotic	1,307	0.432
Monozygotic male	775	0.850
Monozygotic female	1,847	0.855
Avuncular with sibling of parent		
Paternal uncle-nephew	92	0.427
Paternal uncle-niece	155	0.228
Maternal aunt-nephew	402	0.185
Maternal aunt-niece	536	0.314
Paternal aunt-nephew	131	0.275
Paternal aunt-niece	196	0.231
Maternal uncle-nephew	236	0.253
Maternal uncle-niece	284	0.230
Avuncular with dizygotic twin of parent		
Paternal uncle-nephew	105	0.369
Paternal uncle-niece	137	0.077
Maternal aunt-nephew	345	0.260
Maternal aunt-niece	525	0.239
Paternal aunt-nephew	118	0.242
Paternal aunt-niece	188	0.244
Maternal uncle-nephew	150	0.288
Maternal uncle-niece	202	0.271

Heritability for height ~0.8

Little evidence for departure from additive model



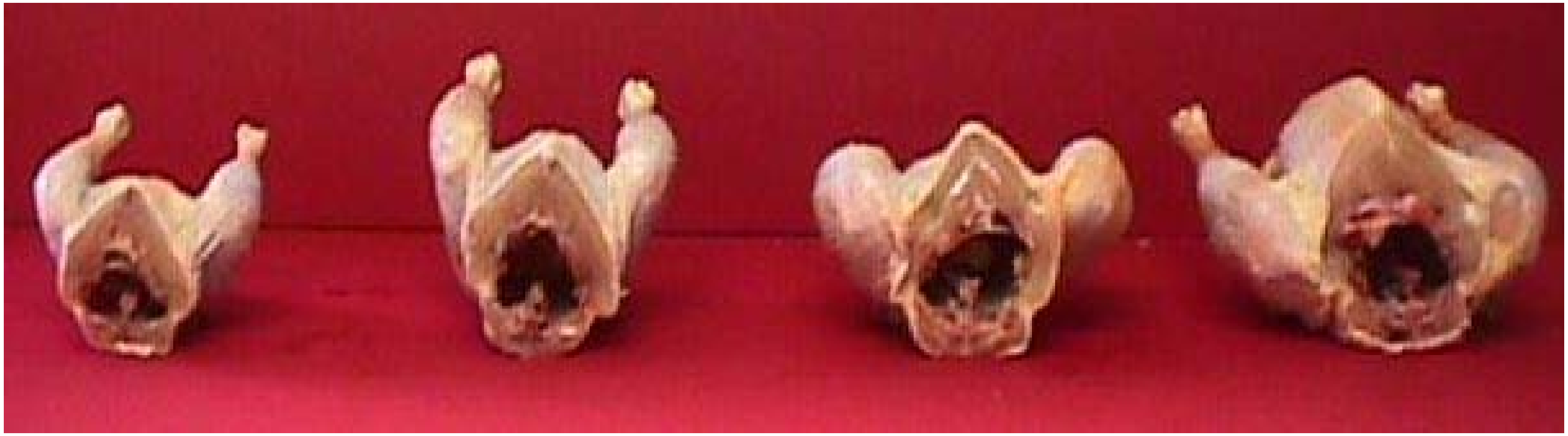
h^2 egg production and growth ~ 0.3

Common ancestor ~100 generations ago

©Roslin Institute

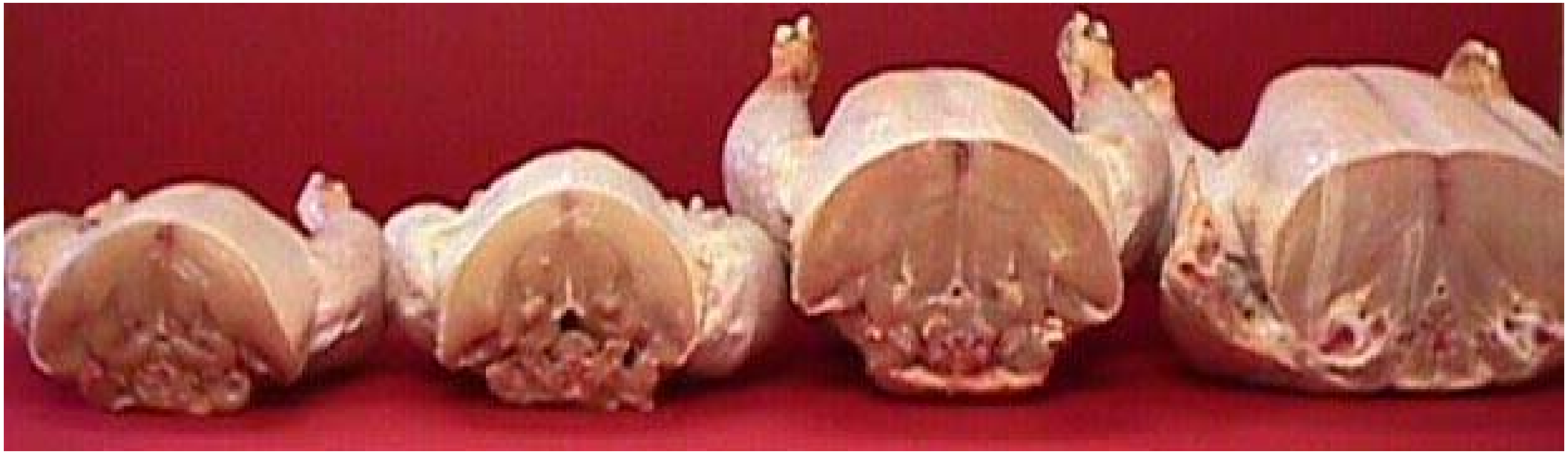
Broiler chickens

1957 Genetic control



h^2 ~ 0.3

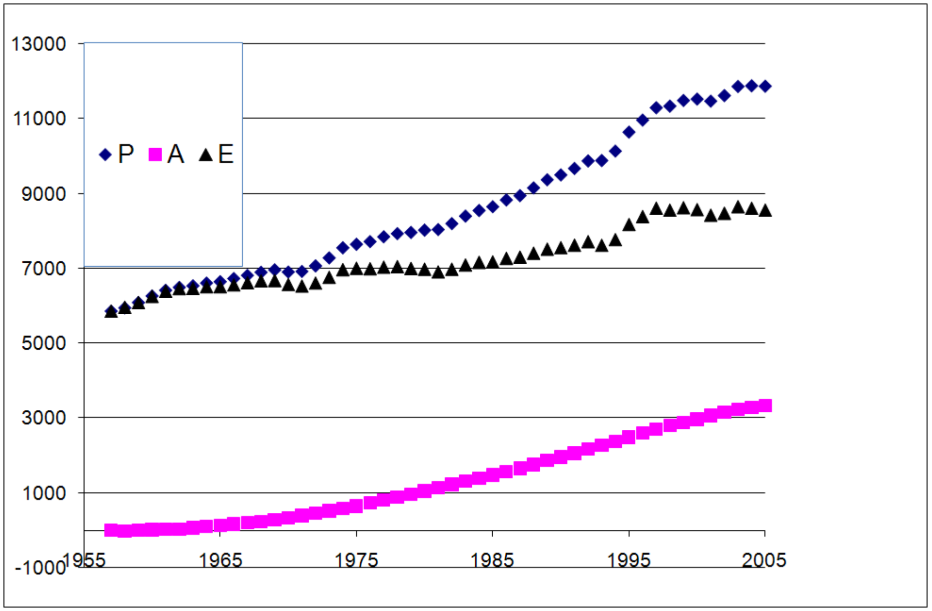
2001 Commercial line



Day 43 Day 57 Day 71 Day 85

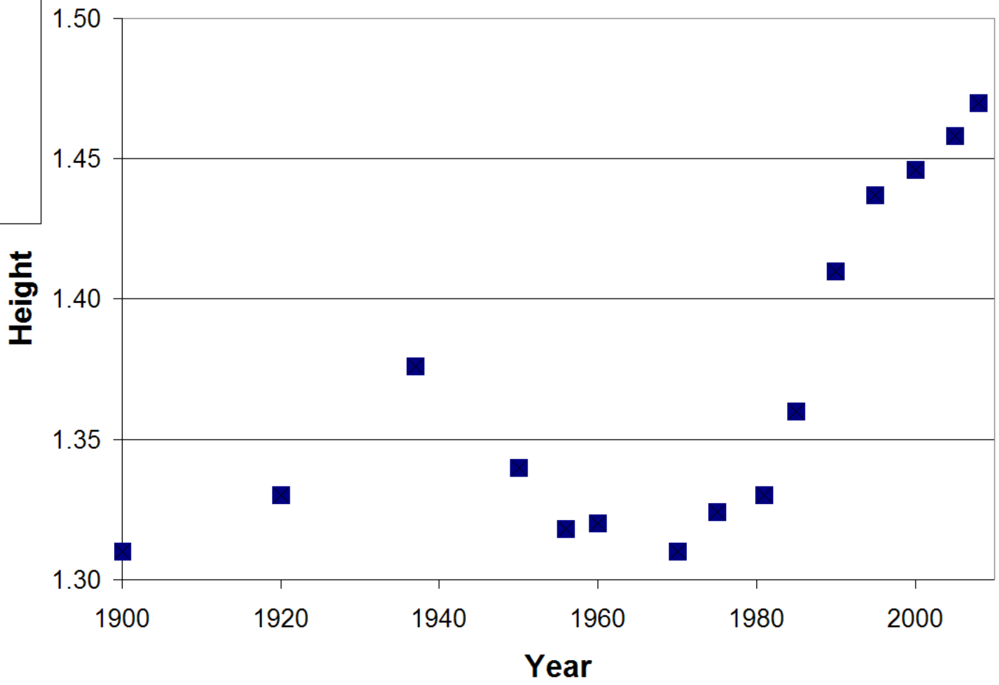
[Slide courtesy of Bill Hill]

(Outbred) dairy cattle



$h^2 \sim 0.3$

$h^2 \sim 0.6$



Observations on selection programmes in agriculture

- Additive genetic variation for most traits of interest, including diseases
- Continuing response in all species = exploitation of additive genetic variation
- No hard evidence of limits being reached
- Heritabilities falling little or not at all
- ***Selection response agrees with estimates of heritability***
- Similar conclusion for long-term selection experiments in model organisms

Human populations

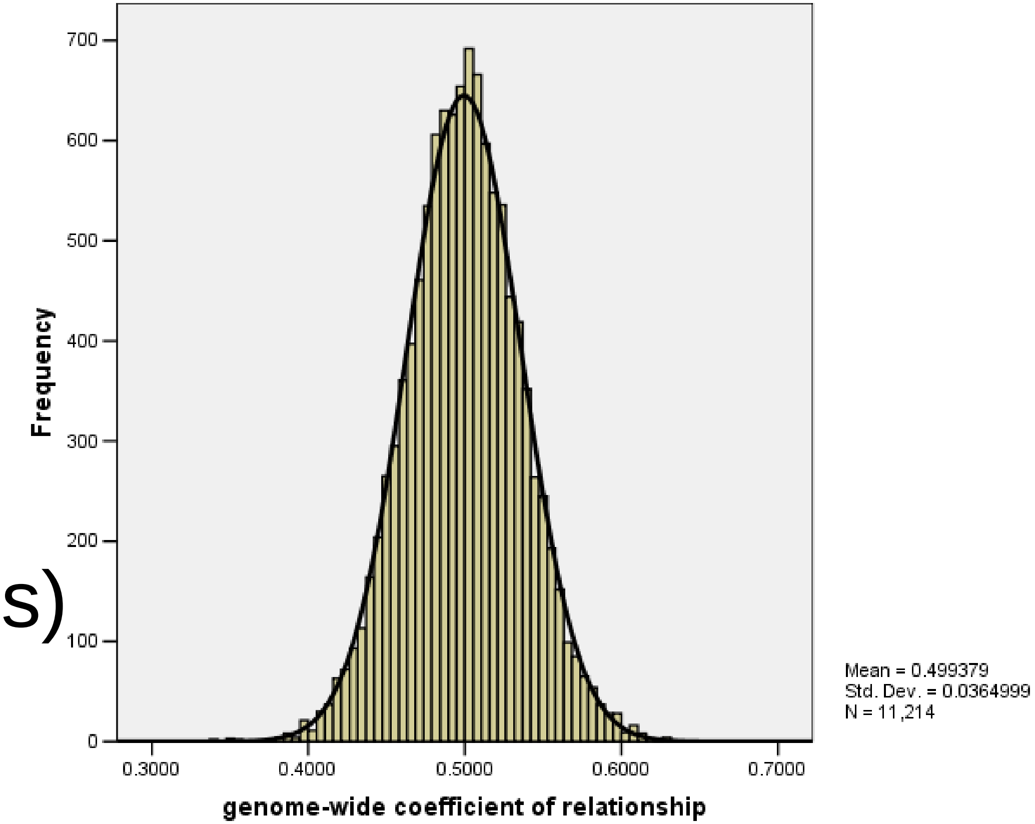
Estimating additive genetic variance within families:

Are fullsibs that share >50% of their genome IBD phenotypically more similar than those that share <50%?

[Visscher *et al*. 2006, PLoS Genetics; 2008 AJHG]

Realised relationships

Mean	0.499
Range	0.31 – 0.64
SD	0.036



Height (N = 11,214 pairs)

h^2 = 0.86 (0.49-0.95)

Conclusions

- Estimates of additive genetic variation and narrow sense heritability unlikely to be out by order(s) of magnitude

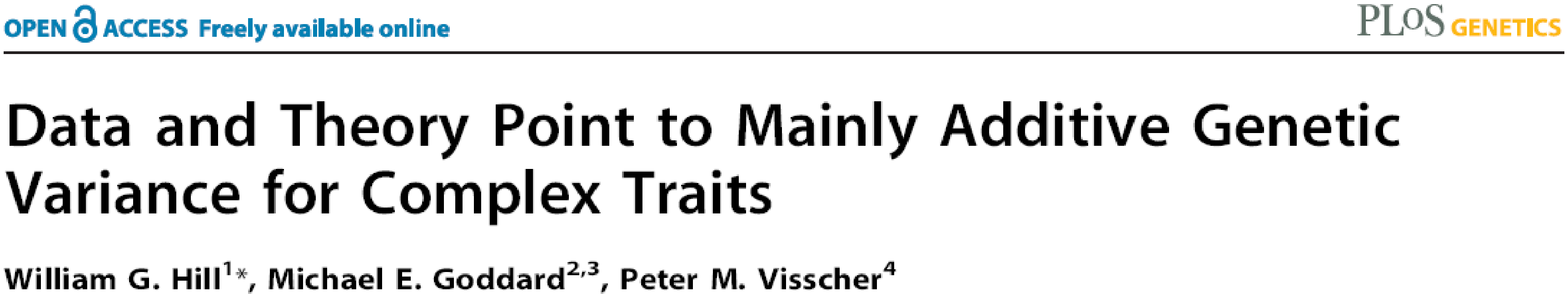
- GWAS data present new opportunities to estimate additive and non-additive genetic variance

Possible explanations for missing heritability

(in order of increasing plausibility ?)

- Heritability estimates are wrong
- Nonadditivity of gene effects – epistasis, GxE
- Epigenetics – including parent-of-origin effects
- Low power for common small effects
- Disease heterogeneity – lots of different diseases with the same phenotype
- Poor tagging (1)
  - rare mutations of large effect (including CNVs)
- Poor tagging (2)
  - common variants in problematic genomic regions

Non-additive variance?



OPEN ACCESS Freely available online

PLoS GENETICS

Data and Theory Point to Mainly Additive Genetic Variance for Complex Traits

William G. Hill[1]*, Michael E. Goddard[2,3], Peter M. Visscher[4]

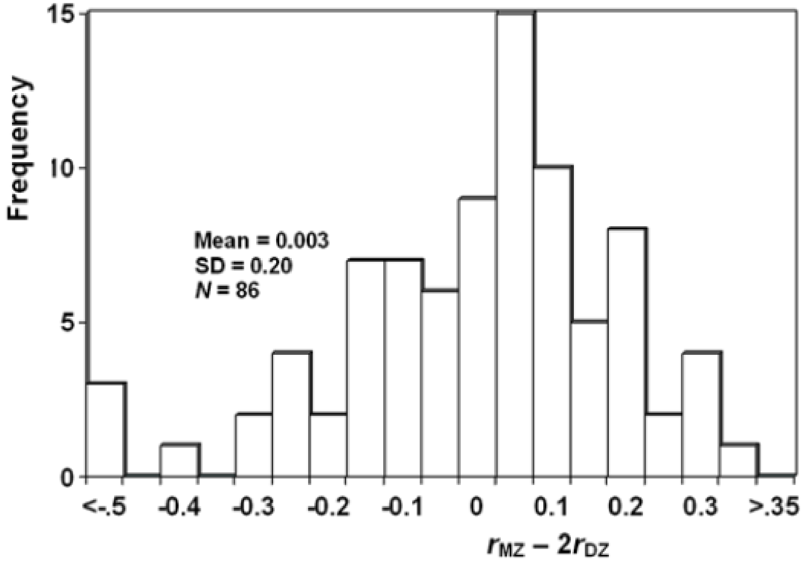
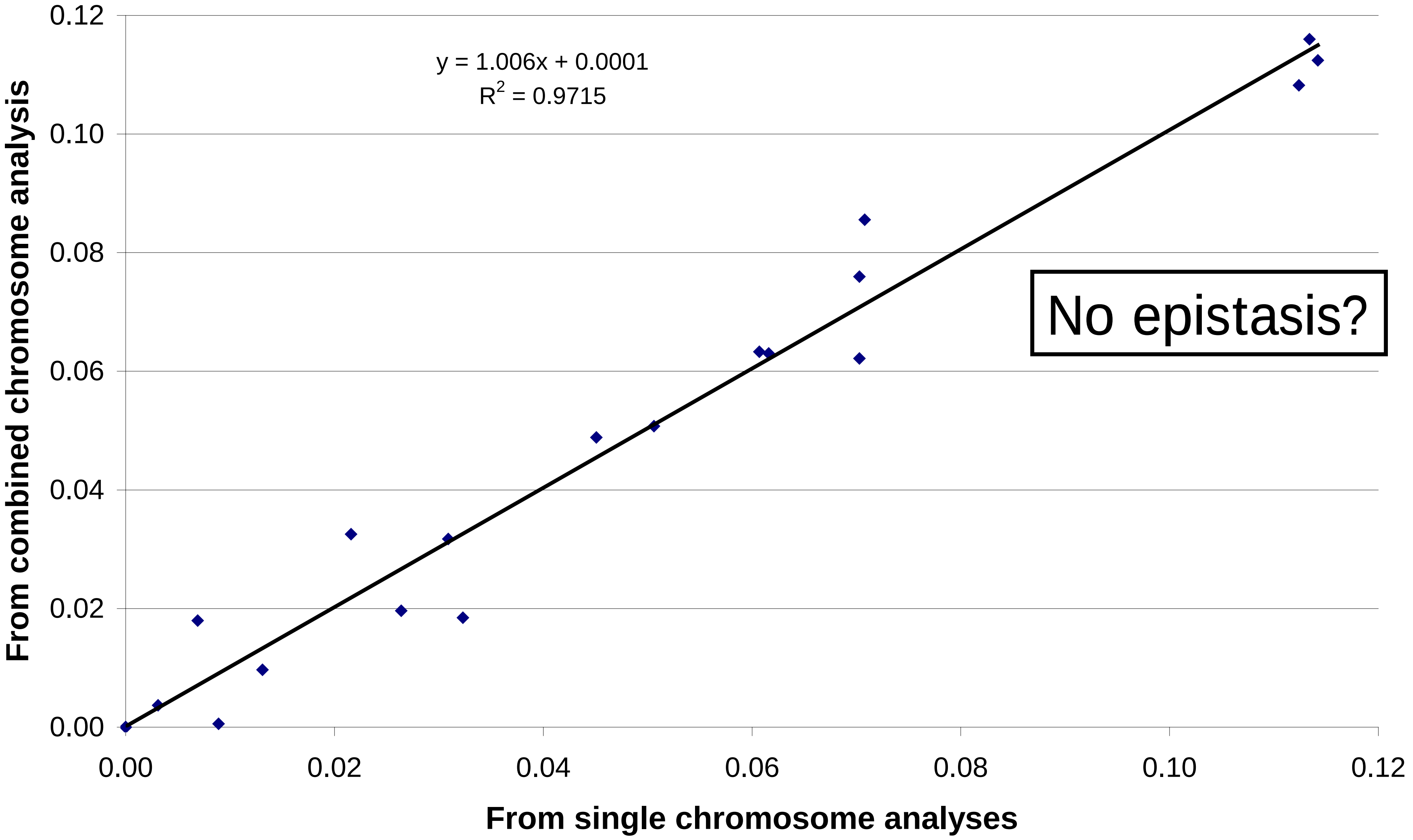


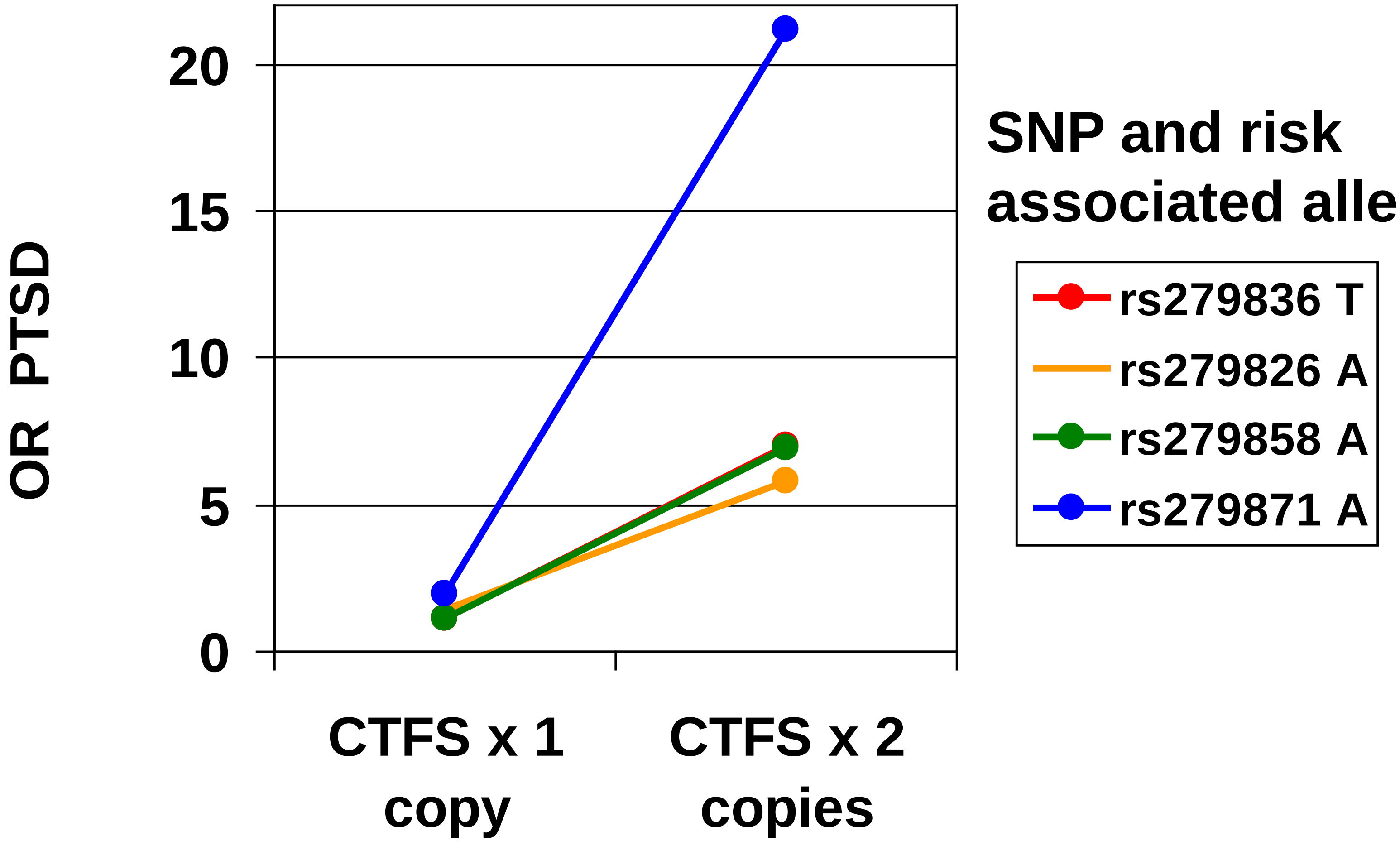
Figure 1. Distribution of $r_{MZ} - 2r_{DZ}$ for all traits on human twins.

Estimates of chromosomal heritabilities



G x E – possibly important, but not many examples

PTSD risk (OR) for interaction terms involving either 1 or 2 copies of at risk GABRA2 alleles with Childhood Trauma Factor Score (CTFS)



Association of childhood trauma exposure and GABRA2 polymorphisms with risk of posttraumatic stress disorder in adults Molecular Psychiatry (2009) 14, 234–238

Possible explanations for missing heritability

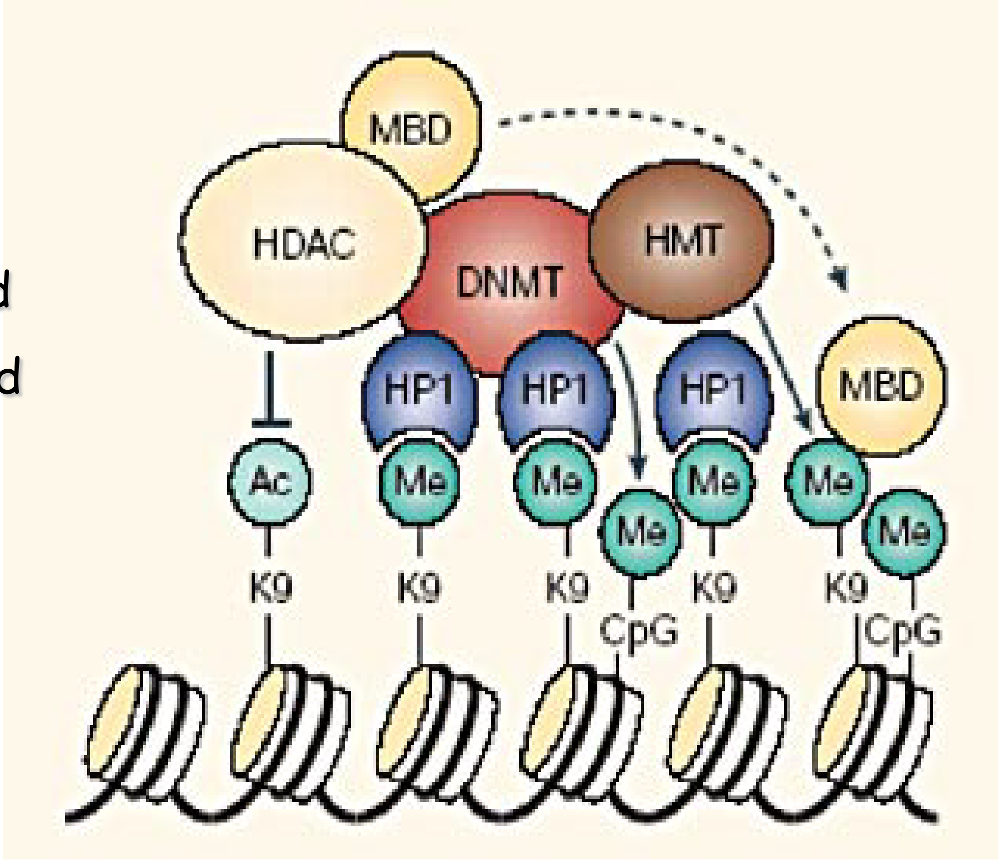
(in order of increasing plausibility ?)

- Heritability estimates are wrong
- Nonadditivity of gene effects – epistasis, GxE
- Epigenetics – including parent-of-origin effects
- Low power for common small effects
- Disease heterogeneity – lots of different diseases with the same phenotype
- Poor tagging (1)
  - rare mutations of large effect (including CNVs)
- Poor tagging (2)
  - common variants in problematic genomic regions

Chromatin modifications are complex

Ac - acetylated

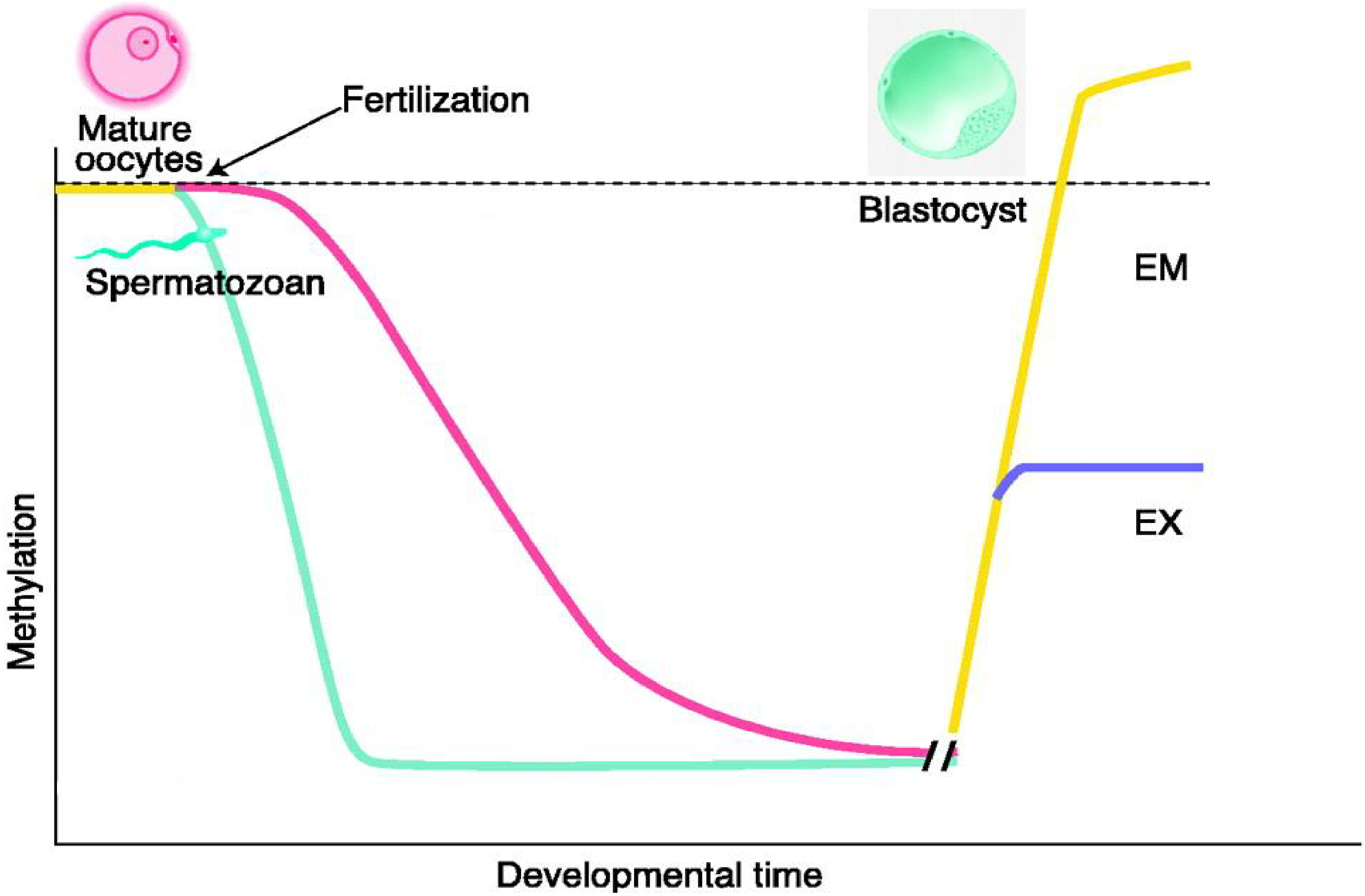
Me- methylated



Greatly simplified schematic

When are the marks laid down?

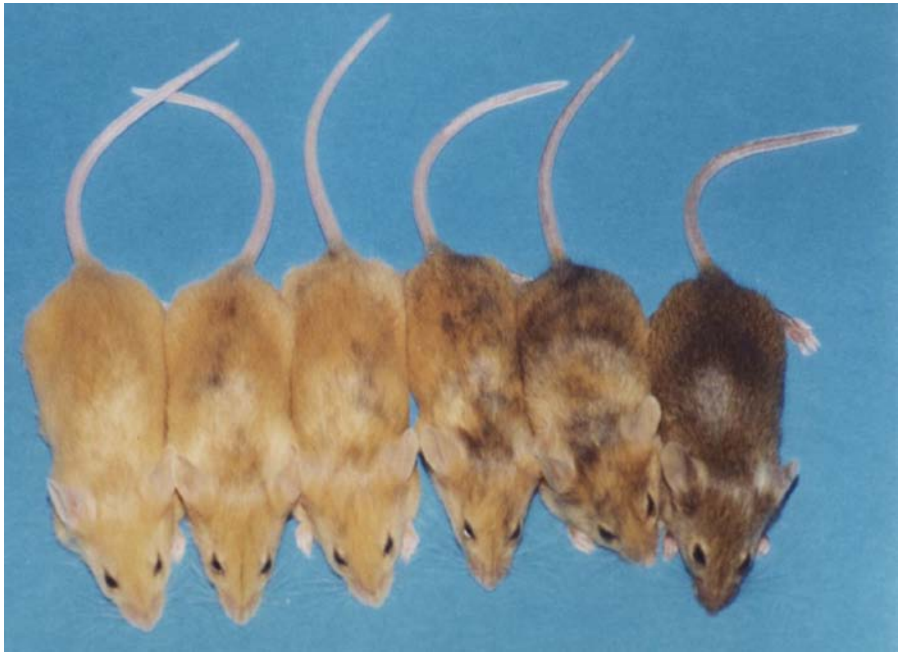
concept of totipotency



Reik *et al.*, Science 293,1089

Intangible variation

Genetically identical mice (same environment) can display different phenotypes



Agouti viable yellow

Different phenotypes correlate with differences in epigenetic state - detectable, laid down in early development

Epigenetic factors

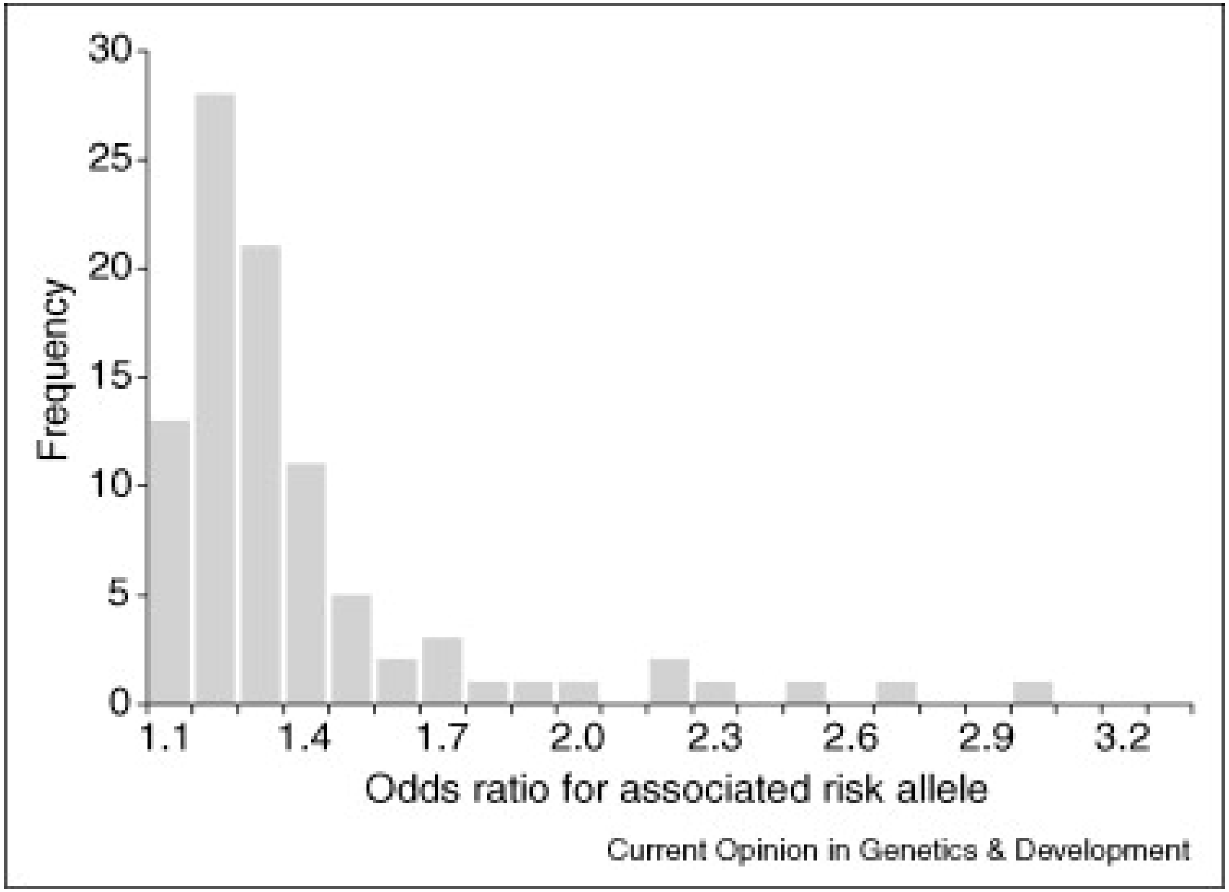
- ‘Stable heritable epimutations’
  - If inherited then like any DNA sequence change
- ‘Unstable heritable epimutations’
  - Decay in family resemblance larger than predicted by additive genetic model
- Non-heritable epigenetic factors
  - Individual environmental effects
  - May increase MZ twin similarity (but why?)

Possible explanations for missing heritability

(in order of increasing plausibility ?)

- Heritability estimates are wrong
- Nonadditivity of gene effects – epistasis, GxE
- Epigenetics – including parent-of-origin effects
- Low power for common small effects
- Disease heterogeneity – lots of different diseases with the same phenotype
- Poor tagging (1)
  - rare mutations of large effect (including CNVs)
- Poor tagging (2)
  - common variants in problematic genomic regions

Effects sizes of validated variants from 1st 16 GWAS studies



Current Opinion in Genetics & Development

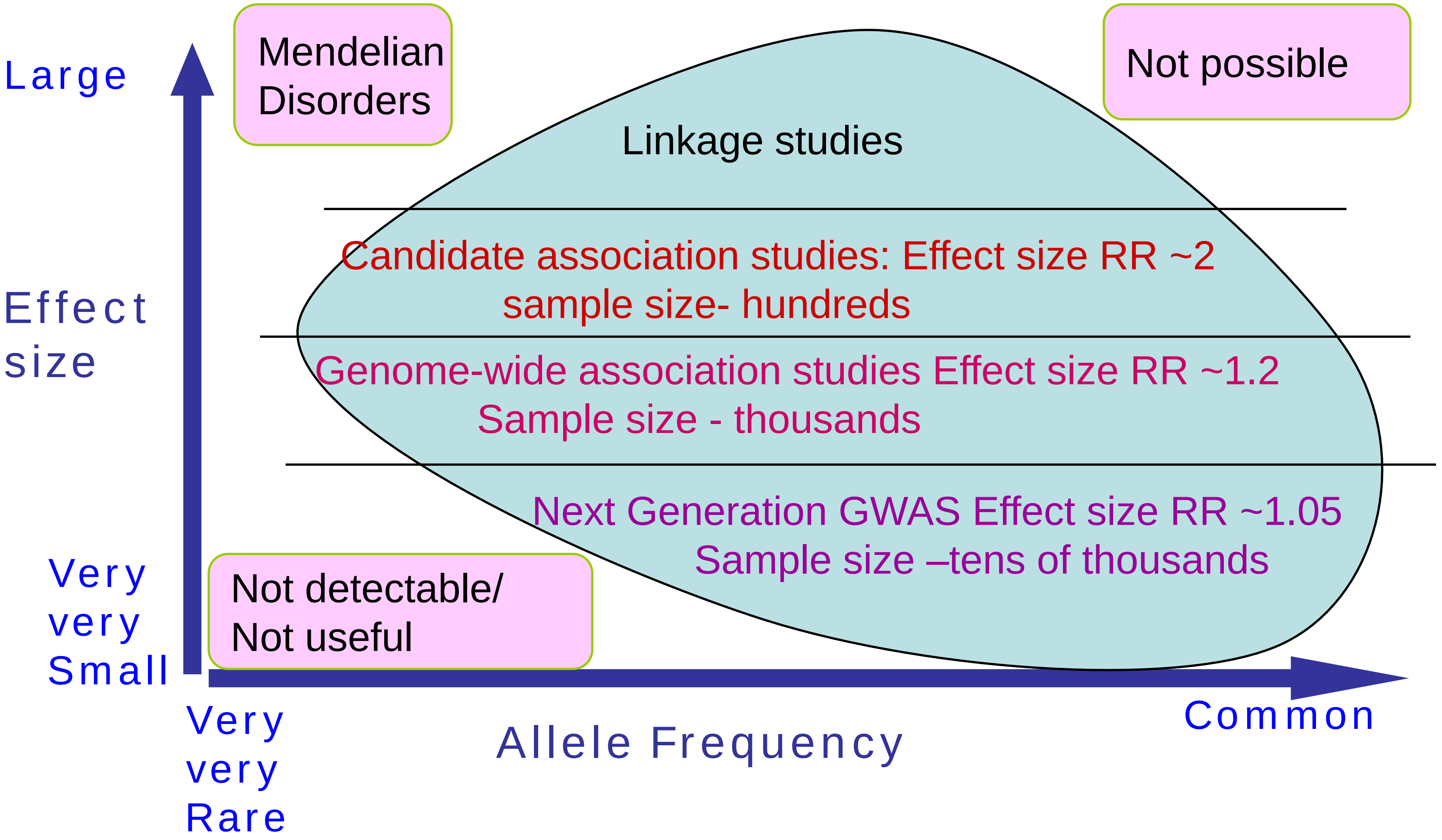
Most effect sizes are very small <1.1

Prediction of individual genetic risk of complex disease

Naomi R Wray[1], Michael E Goddard[2] and Peter M Visscher[1]

Current Opinion in Genetics & Development 2008, 18:257–263

…and will need huge sample sizes to detect



Large

Effect size

Very very Small

Mendelian Disorders

Not possible

Linkage studies

Candidate association studies: Effect size RR ~2
sample size- hundreds

Genome-wide association studies Effect size RR ~1.2
Sample size - thousands

Next Generation GWAS Effect size RR ~1.05
Sample size –tens of thousands

Not detectable/
Not useful

Very very Rare

Allele Frequency

Common

- Under a neutral model we expect a U-shaped distribution of allele frequencies (i.e. most SNPs will have very small MAF and will therefore be poorly tagged by current chips

- Under a stabilising selection & mutation balance large effects will have lower MAF (Zhang & Hill 2005)
- Shaun to expand on this !

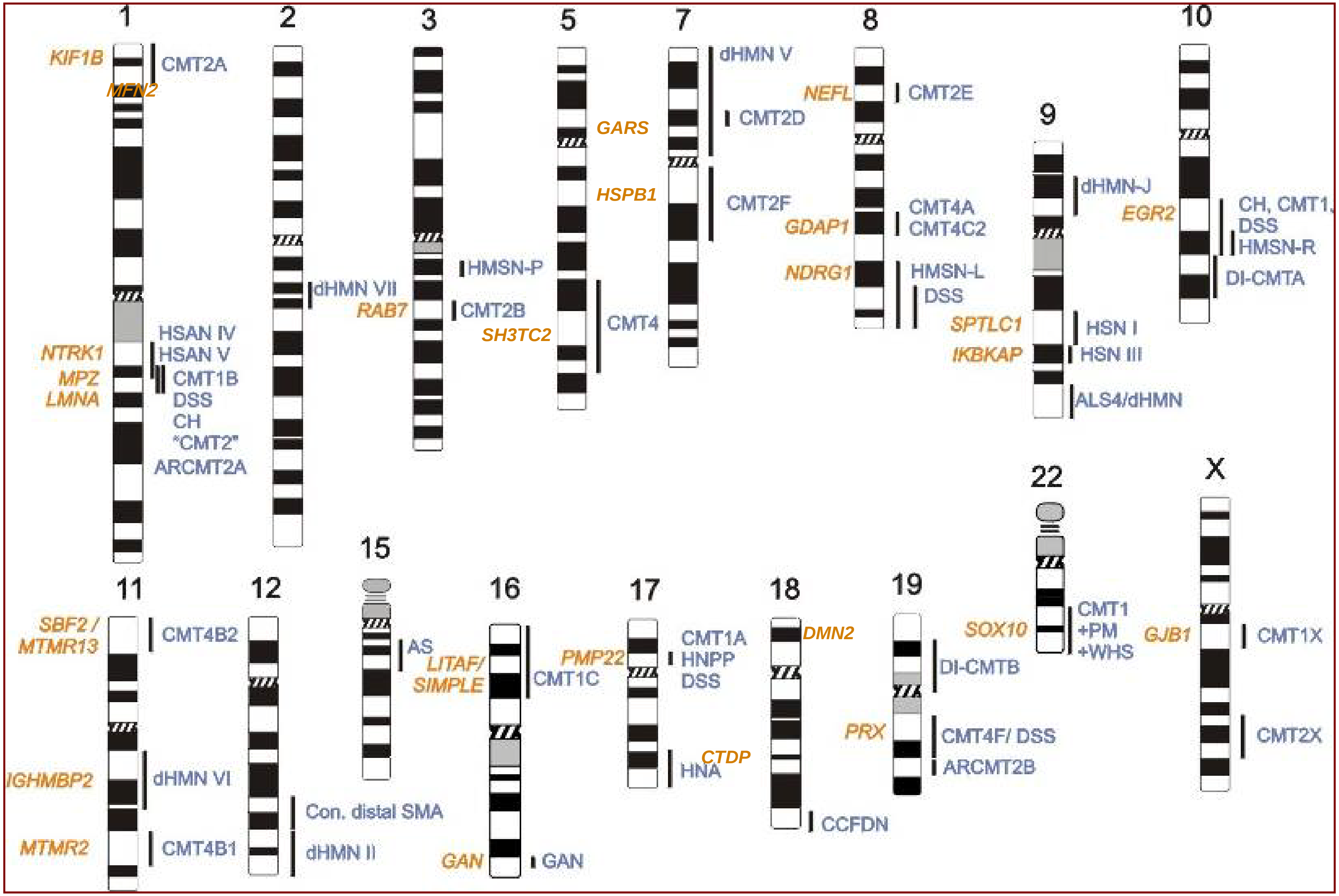
Possible explanations for missing heritability

(in order of increasing plausibility ?)

- Heritability estimates are wrong
- Nonadditivity of gene effects – epistasis, GxE
- Epigenetics – including parent-of-origin effects
- Low power for common small effects
- Disease heterogeneity – lots of different diseases with the same phenotype
- Poor tagging (1)
  - rare mutations of large effect (including CNVs)
- Poor tagging (2)
  - common variants in problematic genomic regions

What if our “disease” is actually dozens (hundreds, thousands) of different diseases that all look the same?

Loci for Inherited Peripheral Neuropathies
Multiple causal loci for Charcot Marie Tooth disease (CMT)



Possible explanations for missing heritability

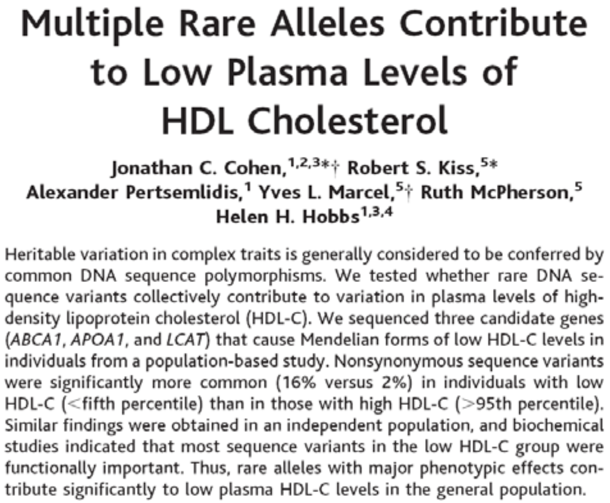
(in order of increasing plausibility ?)

- Heritability estimates are wrong
- Nonadditivity of gene effects – epistasis, GxE
- Epigenetics – including parent-of-origin effects
- Low power for common small effects
- Disease heterogeneity – lots of different diseases with the same phenotype
- Poor tagging (1)
  - rare mutations of large effect (including CNVs)
- Poor tagging (2)
  - common variants in problematic genomic regions

Even for “simple” diseases the number of alleles is large

- Ischaemic heart disease (LDR) >190
- Breast cancer (BRAC1) >300
- Colorectal cancer (MLN1) >140

Complex disease: common or rare alleles?



Multiple Rare Alleles Contribute to Low Plasma Levels of HDL Cholesterol

Jonathan C. Cohen,[1,2,3]*† Robert S. Kiss,[5]*
Alexander Pertsemlidis,[1] Yves L. Marcel,[5]† Ruth McPherson,[5]
Helen H. Hobbs[1,3,4]

Heritable variation in complex traits is generally considered to be conferred by common DNA sequence polymorphisms. We tested whether rare DNA sequence variants collectively contribute to variation in plasma levels of high-density lipoprotein cholesterol (HDL-C). We sequenced three candidate genes (*ABCA1*, *APOA1*, and *LCAT*) that cause Mendelian forms of low HDL-C levels in individuals from a population-based study. Nonsynonymous sequence variants were significantly more common (16% versus 2%) in individuals with low HDL-C (<fifth percentile) than in those with high HDL-C (>95th percentile). Similar findings were obtained in an independent population, and biochemical studies indicated that most sequence variants in the low HDL-C group were functionally important. Thus, rare alleles with major phenotypic effects contribute significantly to low plasma HDL-C levels in the general population.

Increasing evidence for Common Disease – Rare Variant hypothesis (CDRV)

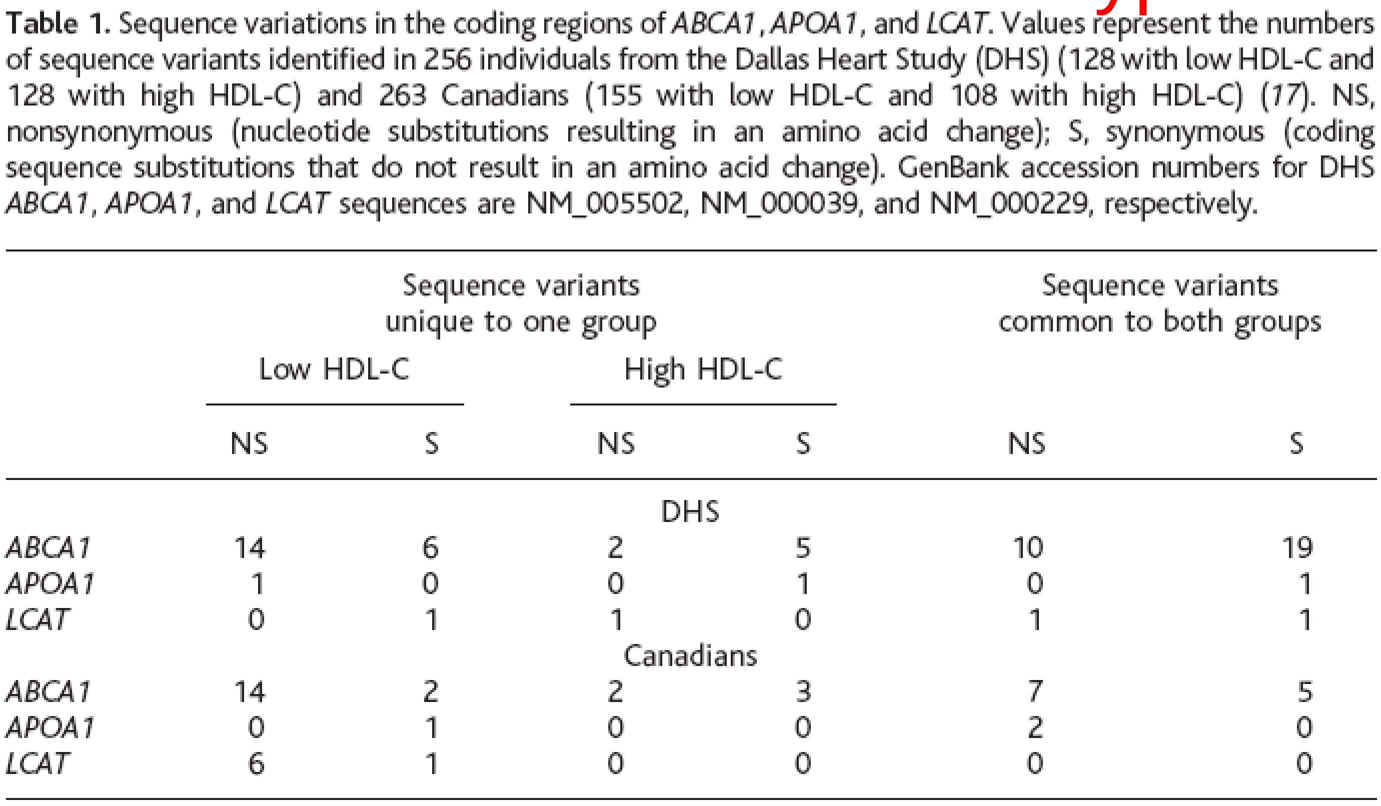
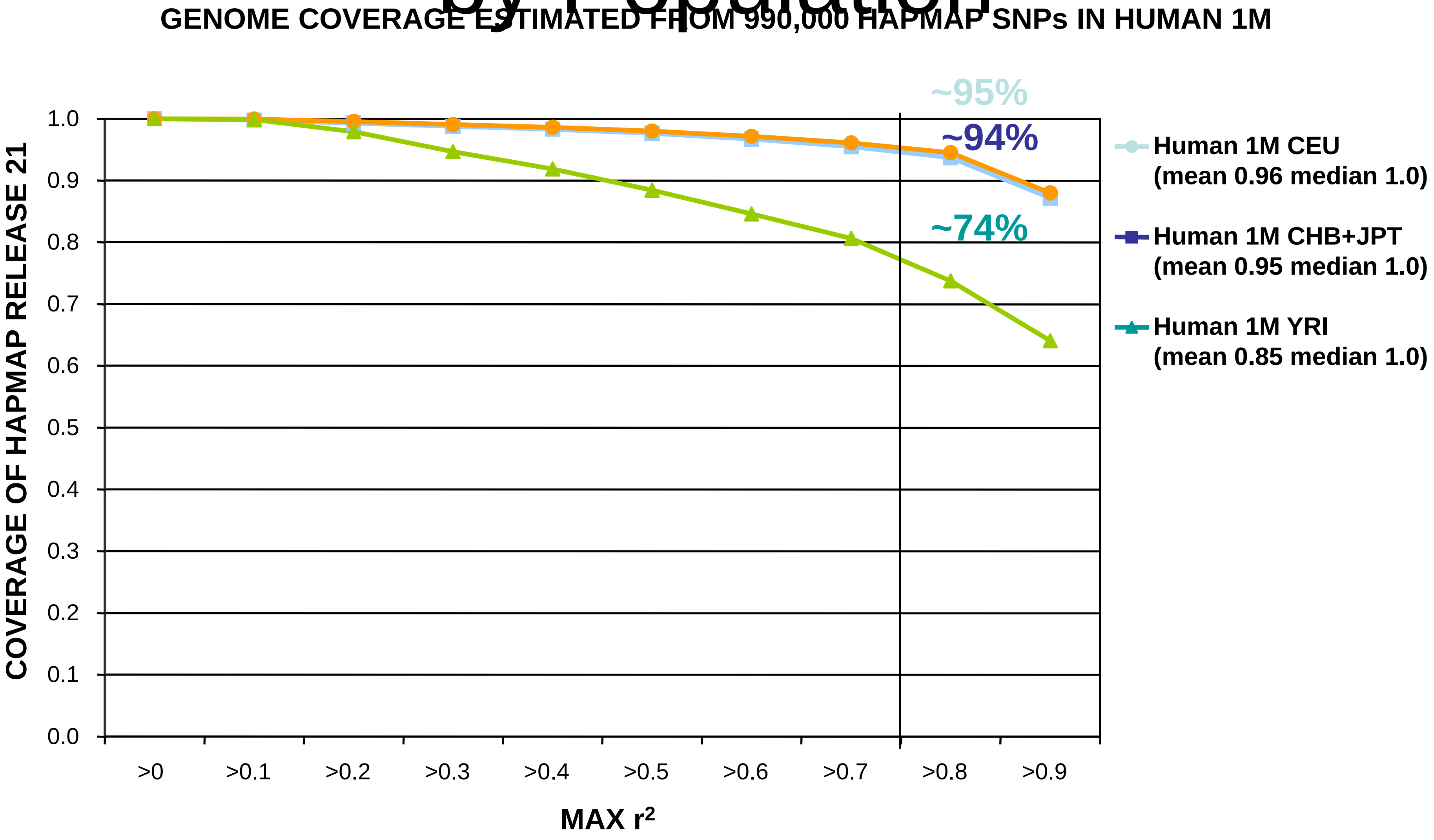


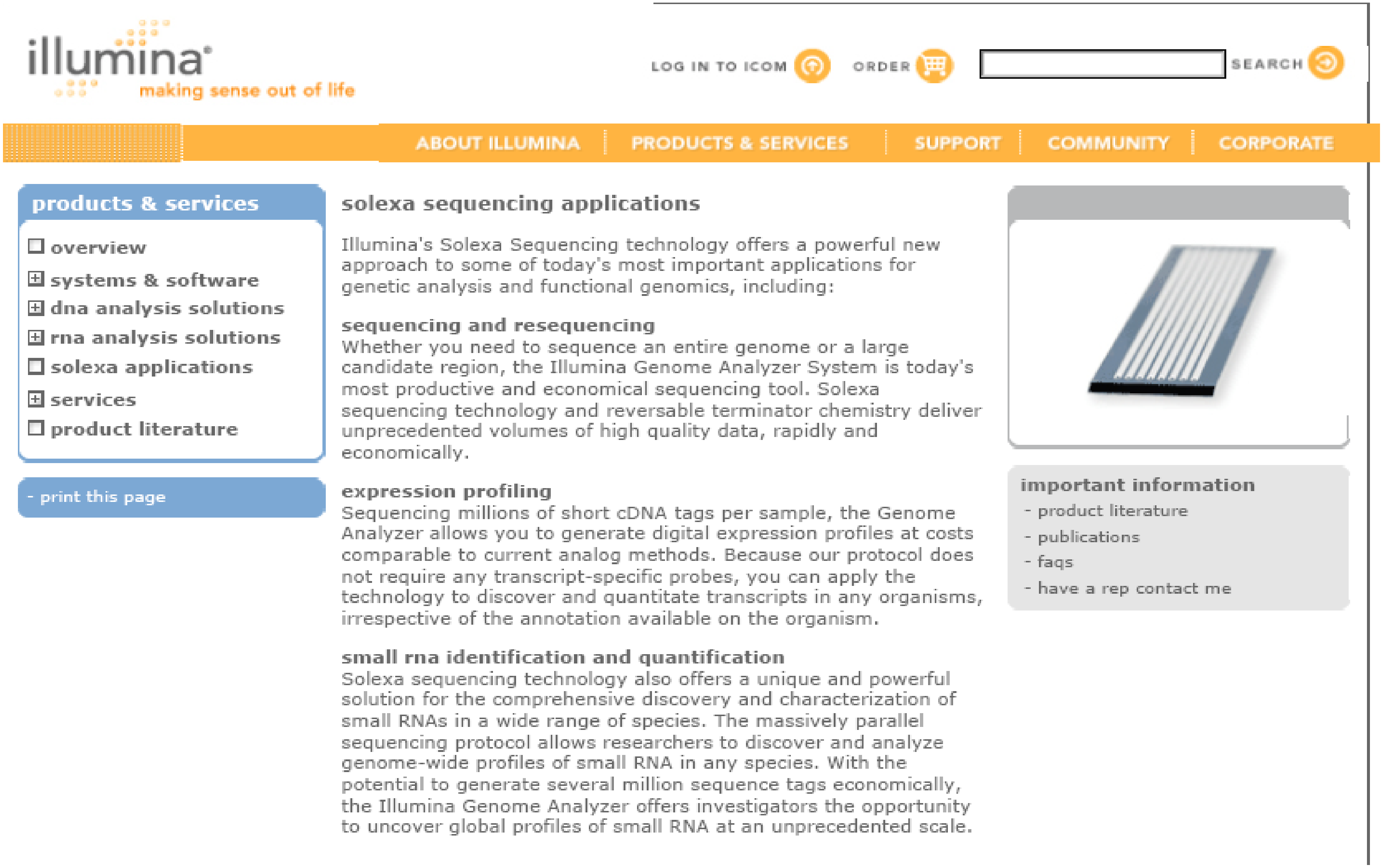
Table 1. Sequence variations in the coding regions of *ABCA1*, *APOA1*, and *LCAT*. Values represent the numbers of sequence variants identified in 256 individuals from the Dallas Heart Study (DHS) (128 with low HDL-C and 128 with high HDL-C) and 263 Canadians (155 with low HDL-C and 108 with high HDL-C) (*17*). NS, nonsynonymous (nucleotide substitutions resulting in an amino acid change); S, synonymous (coding sequence substitutions that do not result in an amino acid change). GenBank accession numbers for DHS *ABCA1*, *APOA1*, and *LCAT* sequences are NM_005502, NM_000039, and NM_000229, respectively.

	Sequence variants unique to one group				Sequence variants common to both groups	
	Low HDL-C		High HDL-C			
	NS	S	NS	S	NS	S
			DHS			
ABCA1	14	6	2	5	10	19
APOA1	1	0	0	1	0	1
LCAT	0	1	1	0	1	1
			Canadians			
ABCA1	14	2	2	3	7	5
APOA1	0	1	0	0	2	0
LCAT	6	1	0	0	0	0

[Science 2004]

Human 1M HapMap Coverage by Population





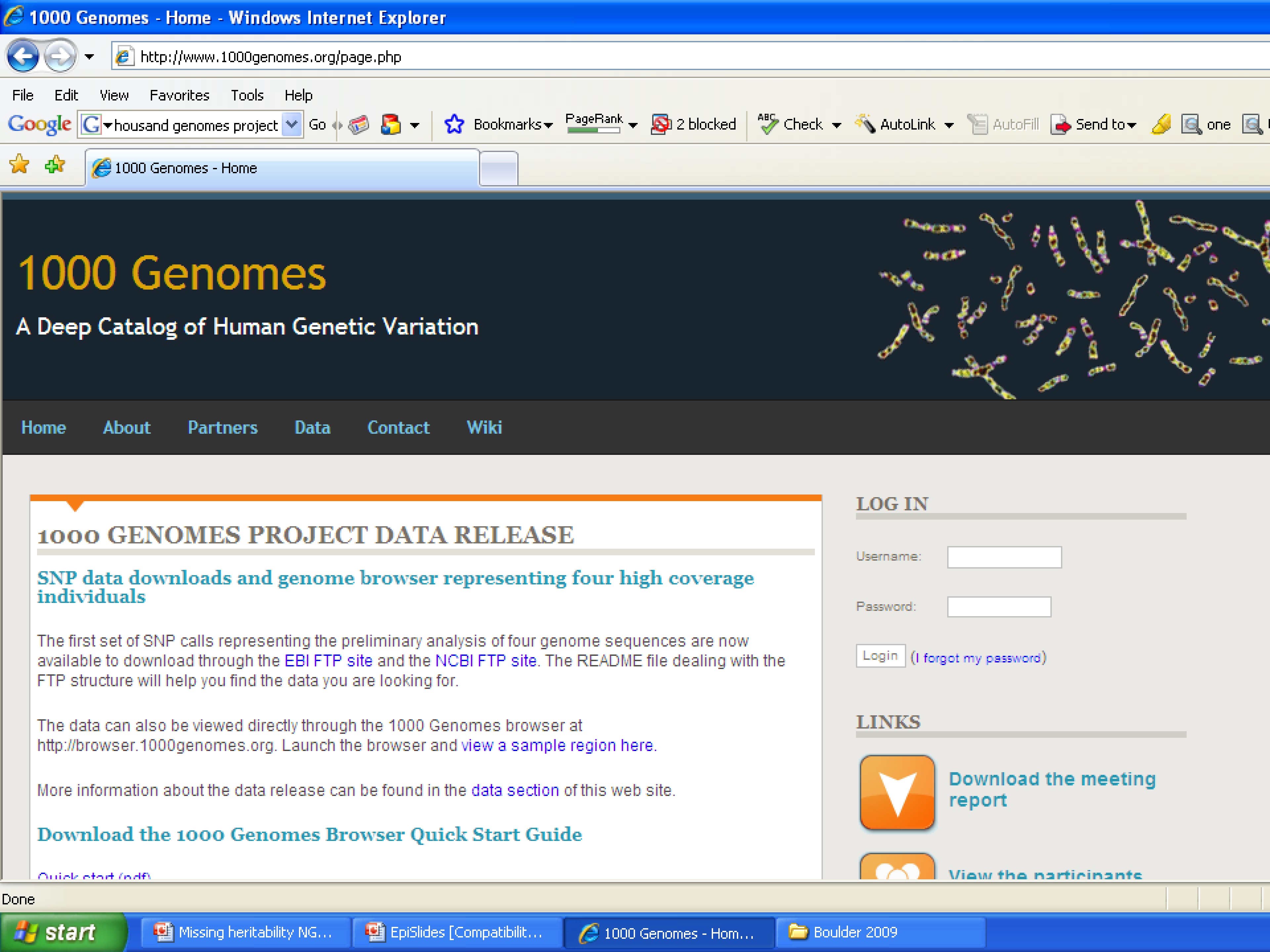
solexa sequencing applications

Illumina's Solexa Sequencing technology offers a powerful new approach to some of today's most important applications for genetic analysis and functional genomics, including:

sequencing and resequencing
Whether you need to sequence an entire genome or a large candidate region, the Illumina Genome Analyzer System is today's most productive and economical sequencing tool. Solexa sequencing technology and reversable terminator chemistry deliver unprecedented volumes of high quality data, rapidly and economically.

expression profiling
Sequencing millions of short cDNA tags per sample, the Genome Analyzer allows you to generate digital expression profiles at costs comparable to current analog methods. Because our protocol does not require any transcript-specific probes, you can apply the technology to discover and quantitate transcripts in any organisms, irrespective of the annotation available on the organism.

small rna identification and quantification
Solexa sequencing technology also offers a unique and powerful solution for the comprehensive discovery and characterization of small RNAs in a wide range of species. The massively parallel sequencing protocol allows researchers to discover and analyze genome-wide profiles of small RNA in any species. With the potential to generate several million sequence tags economically, the Illumina Genome Analyzer offers investigators the opportunity to uncover global profiles of small RNA at an unprecedented scale.



1000 Genomes

A Deep Catalog of Human Genetic Variation

1000 GENOMES PROJECT DATA RELEASE

SNP data downloads and genome browser representing four high coverage individuals

The first set of SNP calls representing the preliminary analysis of four genome sequences are now available to download through the EBI FTP site and the NCBI FTP site. The README file dealing with the FTP structure will help you find the data you are looking for.

The data can also be viewed directly through the 1000 Genomes browser at http://browser.1000genomes.org. Launch the browser and view a sample region here.

More information about the data release can be found in the data section of this web site.

Download the 1000 Genomes Browser Quick Start Guide

Quick start (pdf)

LOG IN

Username:

Password:

(I forgot my password)

LINKS

Download the meeting report

View the participants

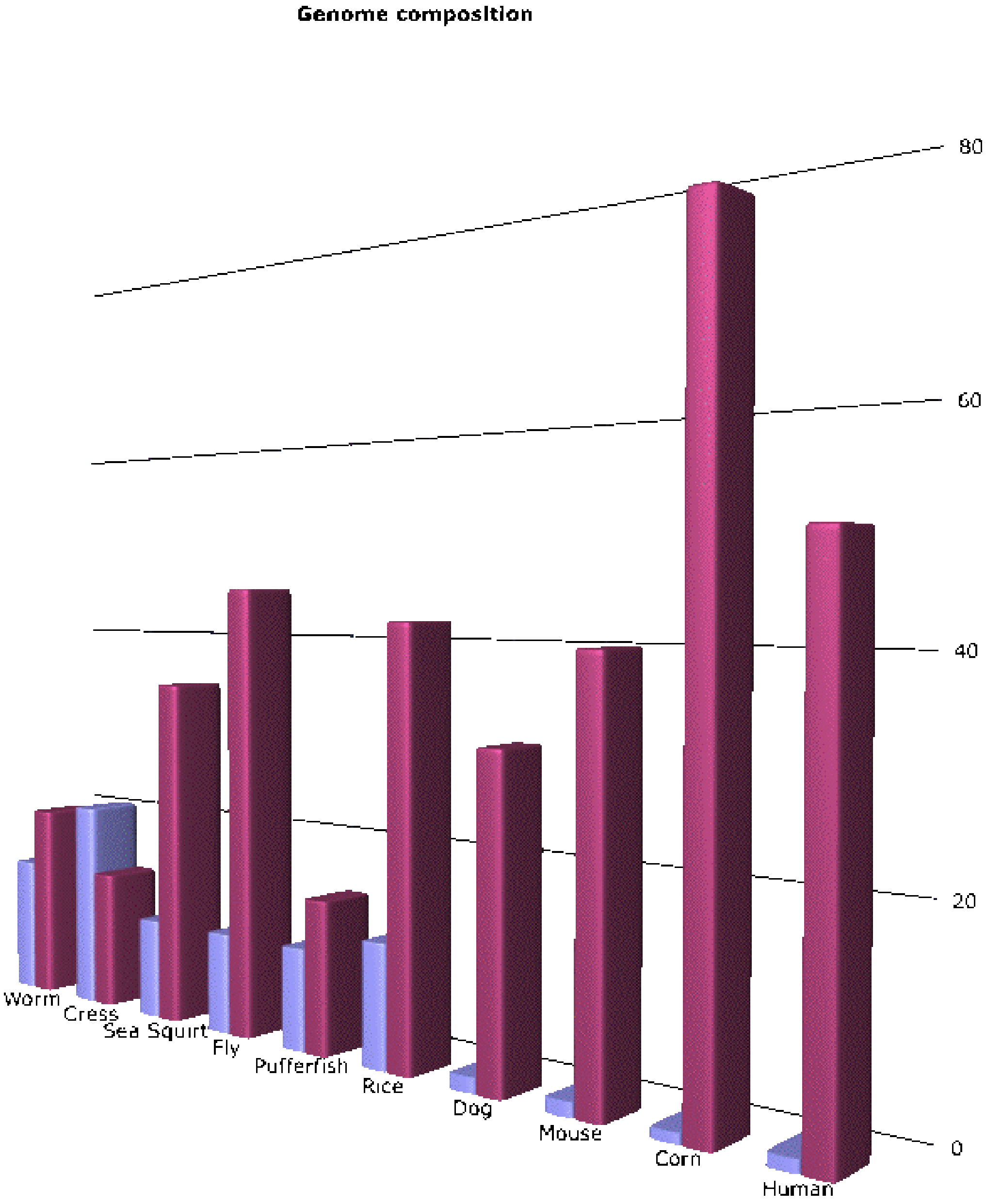
Possible explanations for missing heritability

(in order of increasing plausibility ?)

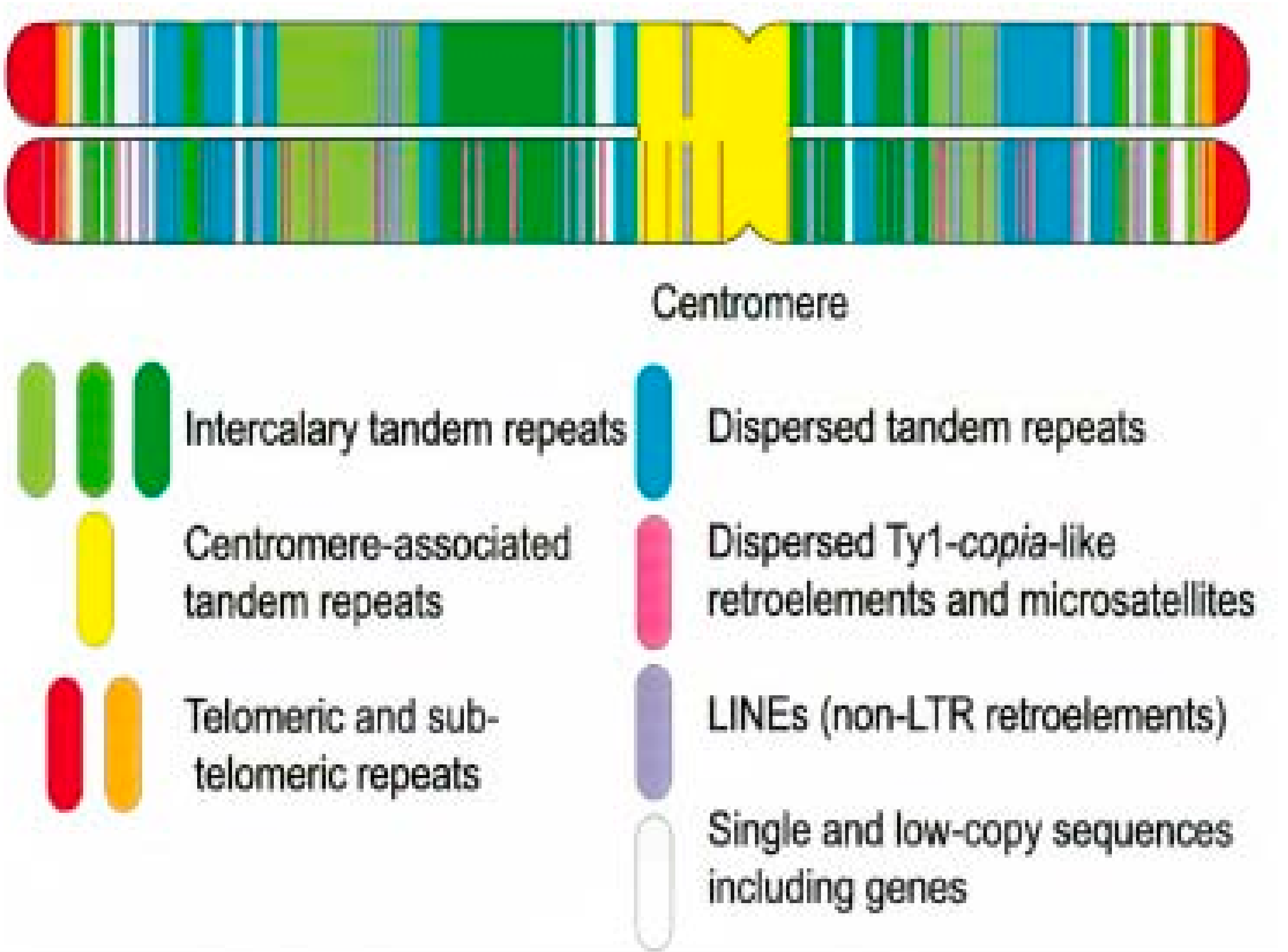
- Heritability estimates are wrong
- Nonadditivity of gene effects – epistasis, GxE
- Epigenetics – including parent-of-origin effects
- Low power for common small effects
- Disease heterogeneity – lots of different diseases with the same phenotype
- Poor tagging (1)
  - rare mutations of large effect (including CNVs)
- Poor tagging (2)
  - common variants in problematic genomic regions

50% of human genome is repetitive DNA.

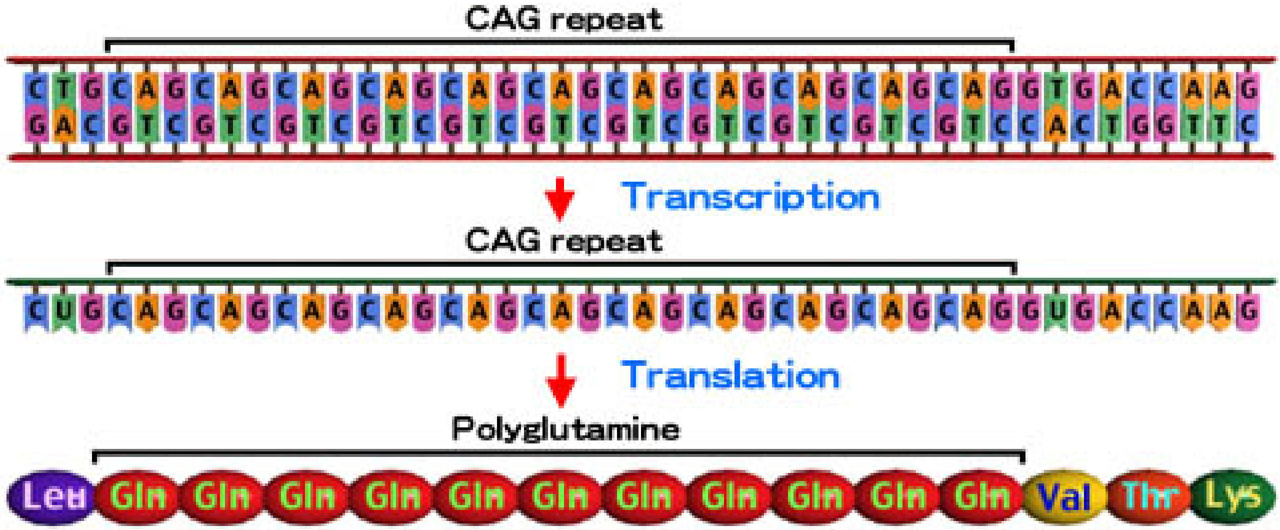
Only 1.2% is coding

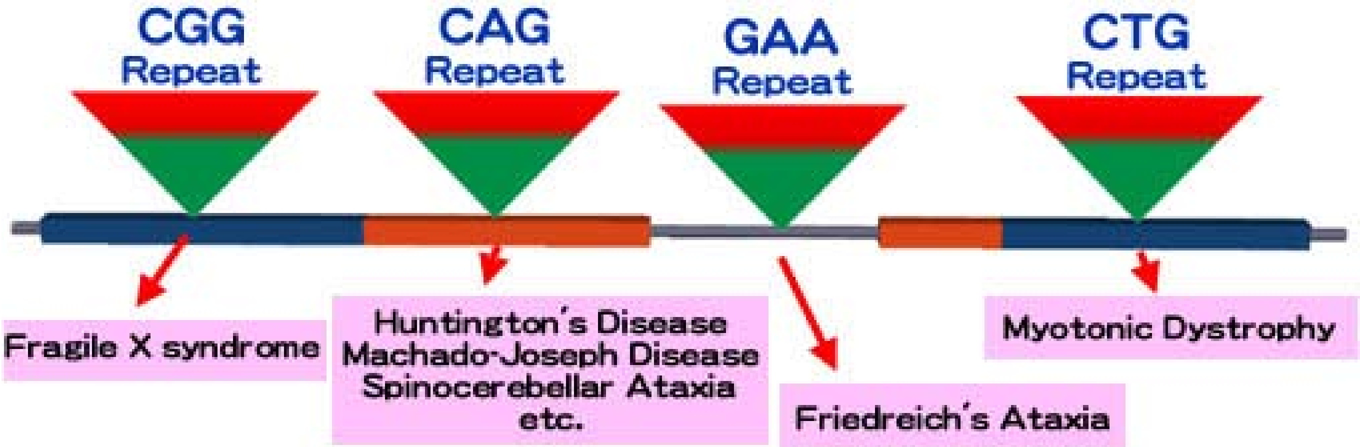


Types of repetitive elements and their chromosomal locations

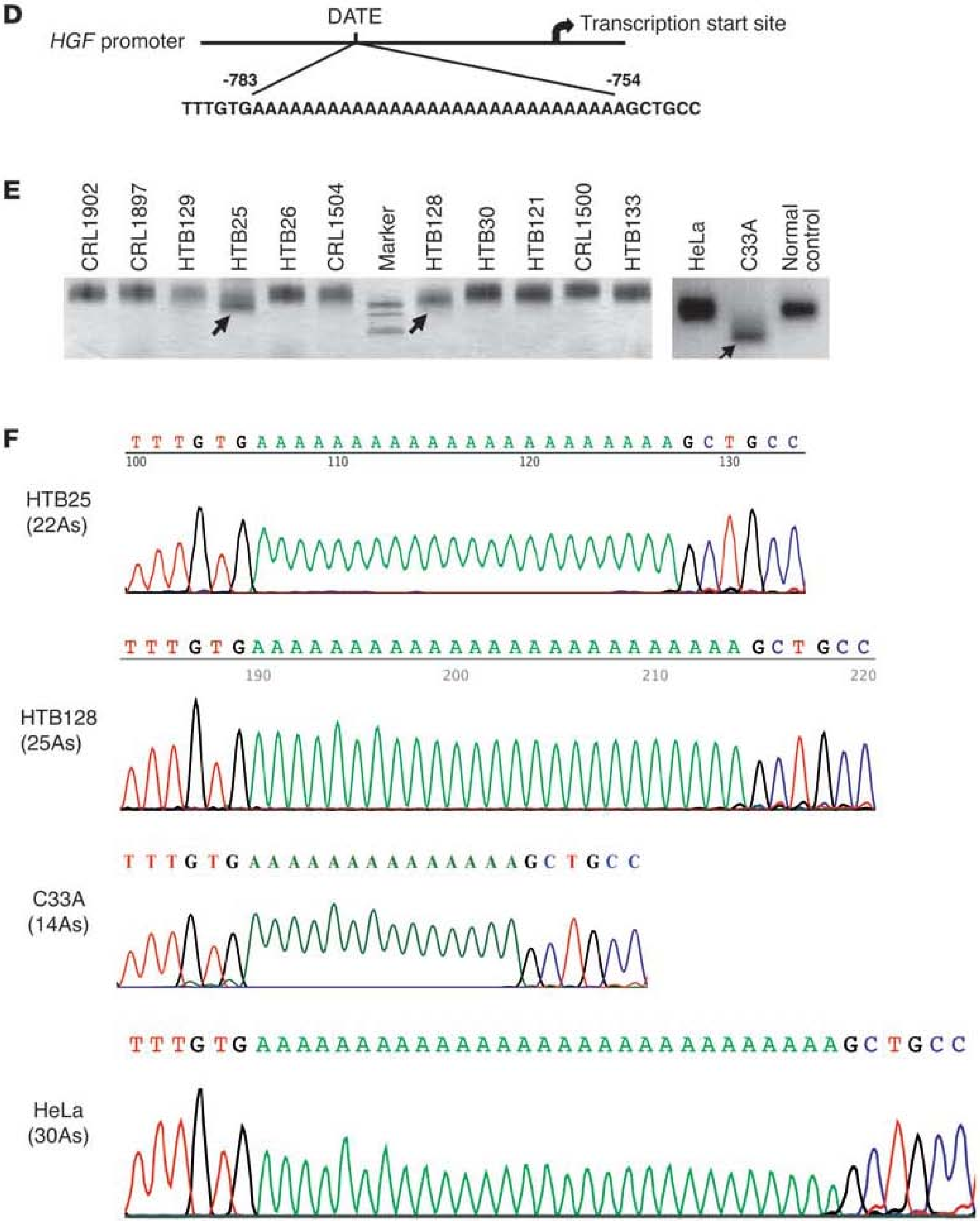

Centromere
Intercalary tandem repeats
Dispersed tandem repeats
Centromere-associated tandem repeats
Dispersed Ty1-*copia*-like retroelements and microsatellites
Telomeric and sub-telomeric repeats
LINEs (non-LTR retroelements)
Single and low-copy sequences including genes

Triplet repeat diseases


CAG repeat
CTGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGGTGACCAAG
GACGTCGTCGTCGTCGTCGTCGTCGTCGTCGTCGTCGTCCACTGGTTC
Transcription
CAG repeat
CUGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGGUGACCAAG
Translation
Polyglutamine
Leu
Gln
Gln
Gln
Gln
Gln
Gln
Gln
Gln
Gln
Gln
Gln
Val
Thr
Lys



Simple repeat polymorphism has major effect on gene expression and breast cancer risk. Poorly tagged by SNPs ?



Somatic mutation and functional polymorphism of a novel regulatory element in the *HGF* gene promoter causes its aberrant expression in human breast cancer

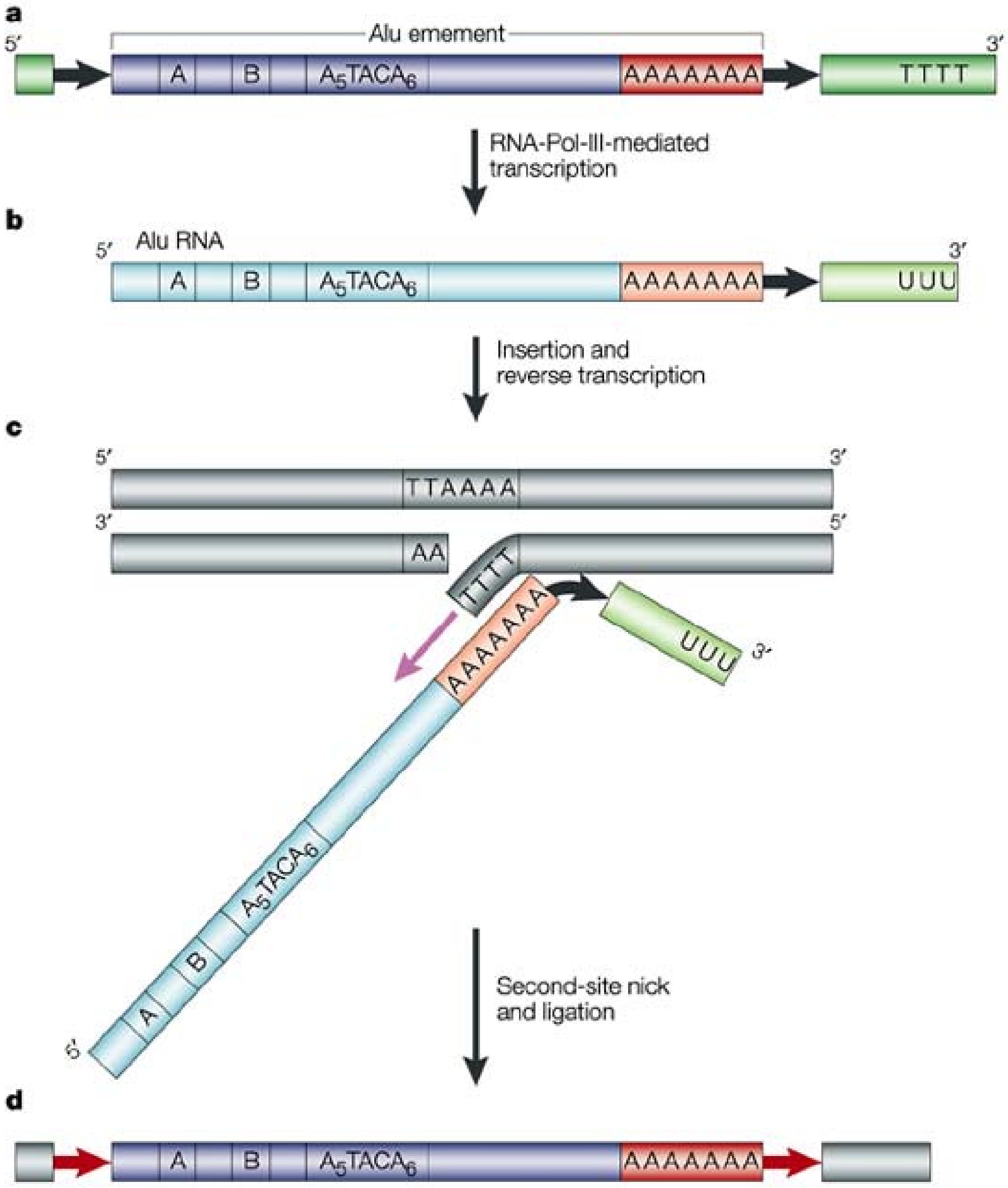
Jihong Ma,[1] Marie C. DeFrances,[1] Chunbin Zou,[1] Carla Johnson,[1]
Robert Ferrell,[2] and Reza Zarnegar[1]

[1]Division of Experimental Pathology, Department of Pathology, School of Medicine, and
[2]Department of Human Genetics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, Pennsylvania, USA.

: *J. Clin. Invest.* doi:10.1172/JCI36640.

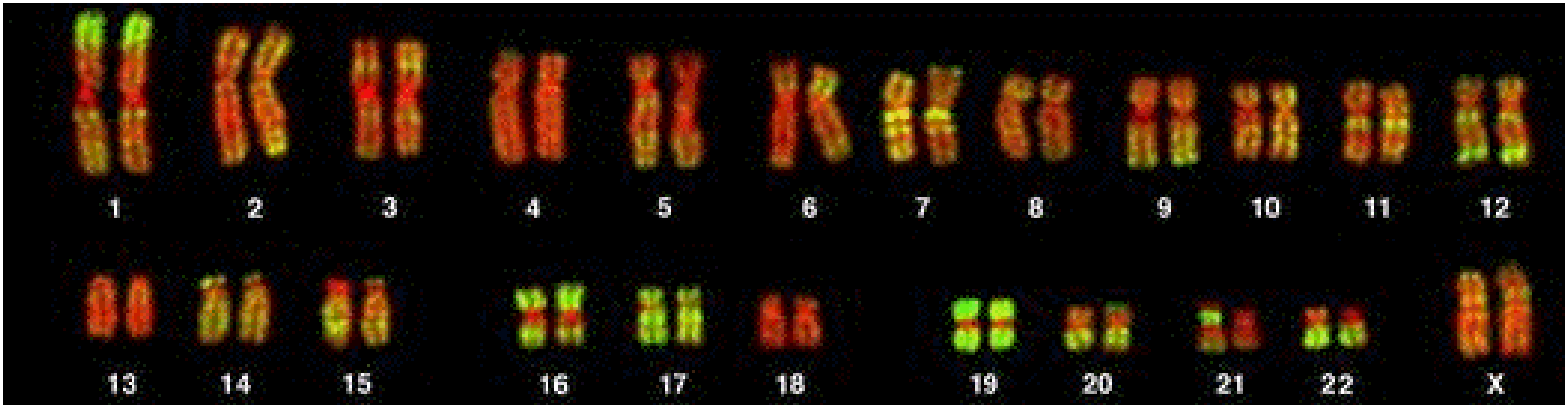
Alu elements

The structure of each Alu element is bi-partite, with the 3' half containing an additional 31-bp insertion (not shown) relative to the 5' half. The total length of each Alu sequence is 300 bp, depending on the length of the 3' oligo(dA)-rich tail. The elements also contain a central A-rich region and are flanked by short intact direct repeats that are derived from the site of insertion (black arrows). The 5' half of each sequence contains an RNA-polymerase-III promoter (A and B boxes). The 3' terminus of the Alu element almost always consists of a run of As that is only occasionally interspersed with other bases (**a**).



Nature Reviews | Genetics

The abundant Alu transposable element, a member of the middle repetitive DNA sequences, is present in all human chromosomes (the Alu element is stained green, while the remainder of the DNA in the chromosomes is stained red).

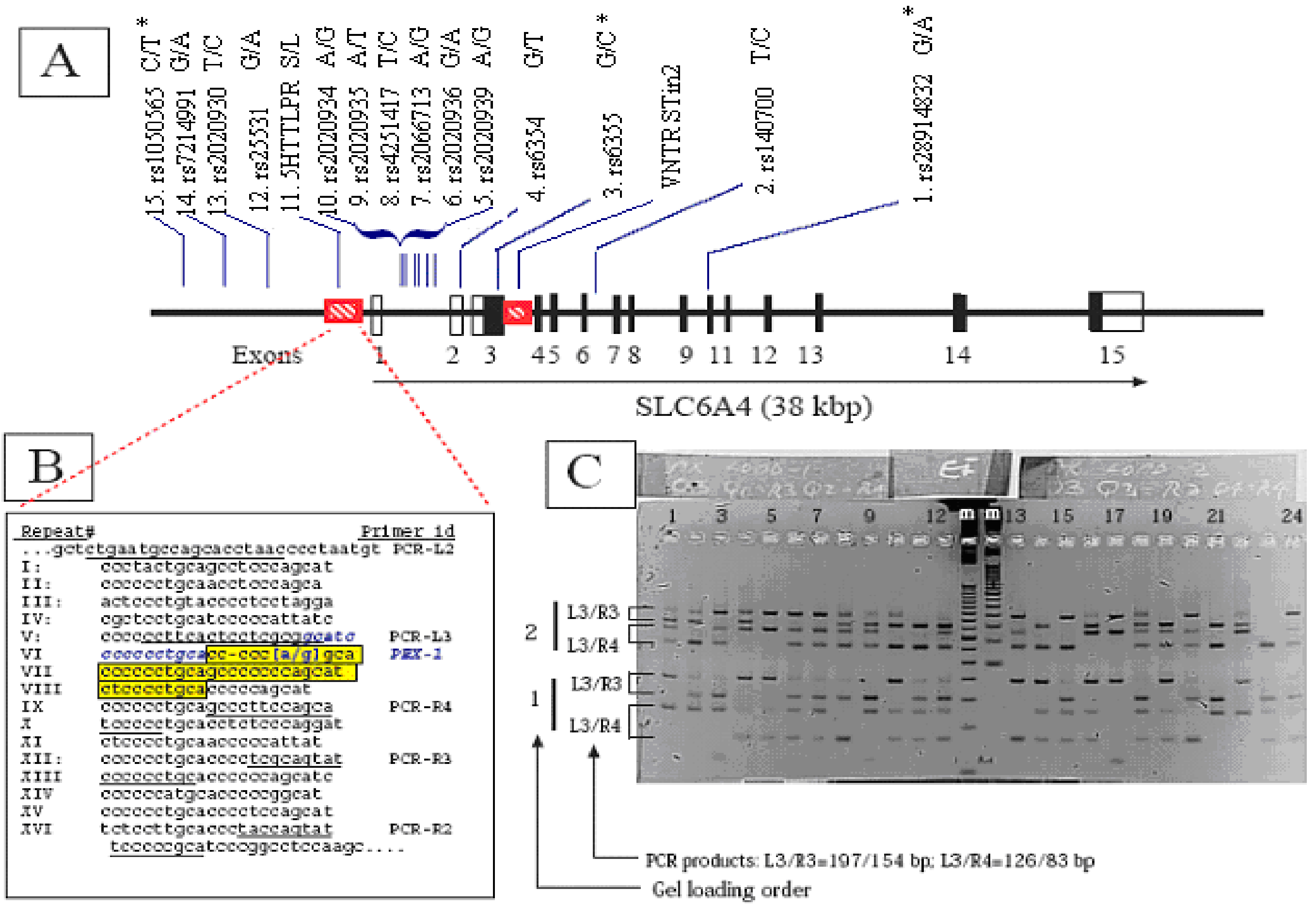


- > 1 million in genome – unique to humans
- Involved in RNA editing – functional ?
- How well are they tagged ??????

Example – 5HTLPR

- Serotonin transporter length polymorphism (5HTLPR – one (short) or two (long) 44bp repeat units
- Has been widely associated with psychiatric outcomes +/- interaction with environment (Caspi)
- How well is it tagged by available SNPs?

Attempting to SNP tag 5HTLPR


A
15. rs1050565 C/T *
14. rs7214991 G/A
13. rs2020930 T/C
12. rs25531 G/A
11. 5HTTLPR S/L
10. rs2020934 A/G
9. rs2020935 A/T
8. rs4251417 T/C
7. rs2066713 A/G
6. rs2020936 G/A
5. rs2020939 A/G
4. rs6354 G/T
3. rs6355 G/C *
VNTR STin2
2. rs140700 T/C
1. rs28914832 G/A *
Exons
1
2
3
4 5
6
7 8
9
11
12
13
14
15
SLC6A4 (38 kbp)
B
Repeat#
Primer id
...gctctgaatgccagcacctaacccctaatgt PCR-L2
I: ccctactgcagcctcccagcat
II: ccccctgcaacctcccagca
III: actccctgtacccctcctagga
IV: cgctcctgcatccccattatc
V: cccccttcactcctcgcggcatc PCR-L3
VI cccccctgcacc-ccc[a/g]gca PEX-1
VII ccccctgcagcccccccagcat
VIII ctcccctgcacccccagcat
IX cccccctgcagcccttccagca PCR-R4
X tcccccctgcacctctcccaggat
XI ctcccctgcaacccccattat
XII: cccccctgcacccctcgcagtat PCR-R3
XIII cccccctgcaccccccagcatc
XIV cccccatgcacccccggcat
XV cccccctgcacccctccagcat
XVI tctccttgcaccctaccagtat PCR-R2
tcccccgcatcccggcctccaagc....
C
1
3
5
7
9
12
m
m
13
15
17
19
21
24
2
L3/R3
L3/R4
1
L3/R3
L3/R4
PCR products: L3/R3=197/154 bp; L3/R4=126/83 bp
Gel loading order

5HTLPR is badly tagged by adjacent SNPs

a)

Marker #	Marker	MAF	1	2	3	4	5	6	7	**8**	9	**10**	**11**	13	14	15
1	rs28914832	0.002	\	0.85	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.06	1.00	0.07
2	rs140700	0.100	0.00	\	1.00	0.95	0.90	0.95	1.00	1.00	0.30	0.81	0.54	0.53	0.64	0.61
3	rs6355	0.021	0.00	0.00	\	1.00	1.00	1.00	1.00	1.00	1.00	0.86	0.77	1.00	0.82	0.83
4	rs6354	0.198	0.01	0.41	0.01	\	1.00	0.99	0.98	1.00	0.81	0.31	0.20	0.72	0.36	0.17
5	rs2020939	0.412	0.00	0.06	0.02	0.17	\	1.00	1.00	1.00	0.84	0.67	0.42	0.88	0.85	0.85
6	rs2020936	0.196	0.01	0.41	0.01	0.98	0.17	\	1.00	1.00	0.80	0.31	0.20	0.71	0.36	0.17
7	rs2066713	0.388	0.00	0.07	0.03	0.15	0.44	0.16	\	1.00	0.74	0.59	0.50	0.55	0.38	0.46
8	**rs4251417**	0.091	0.00	0.01	0.00	0.03	0.14	0.03	0.06	\	1.00	0.87	0.91	0.95	0.95	0.95
9	rs2020935	0.064	0.03	0.05	0.00	0.18	0.03	0.18	0.02	0.01	\	1.00	0.94	0.96	0.92	0.08
10	**rs2020934**	0.489	0.00	0.07	0.02	0.02	0.33	0.02	0.21	0.08	0.07	\	0.79	1.00	0.92	0.91
11	**5HTTLPR**	0.429	**0.00**	**0.03**	**0.01**	**0.01**	**0.16**	**0.01**	**0.12**	**0.06**	**0.05**	**0.49**	\	0.97	0.91	0.90
13	rs2020930	0.036	0.00	0.00	0.00	0.08	0.02	0.08	0.01	0.00	0.50	0.04	0.03	\	1.00	0.95
14	rs7214991	0.374	0.00	0.08	0.02	0.05	0.30	0.05	0.14	0.05	0.10	0.48	0.38	0.06	\	1.00
15	rs1050565	0.325	0.00	0.09	0.03	0.02	0.25	0.01	0.16	0.04	0.00	0.39	0.30	0.02	0.81	\

b)

Haplotype frequencies

C-A-S	0.381
C-G-S	0.045
C-A-L	0.024
C-G-L	0.459
T-A-L	0.081

Summary

- Huge amount of repetitive sequence
- Highly polymorphic
- Some evidence that it has functional significance
- Earlier studies too small (100s) to detect effect sizes now known to be realistic
- Much (most?) such variation poorly tagged with current chips
- Current CNV arrays only detect large variants; no systematic coverage of the vast number of small CNVs (including microsatellites)