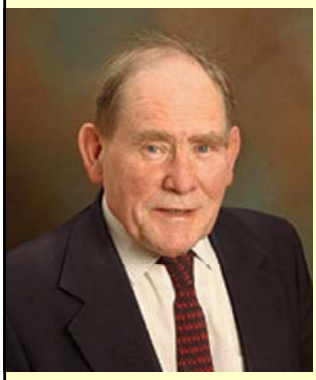
Què podem esperar dels estudis de genètica de malalties complexes i de la seva aplicació en psiquaitria?

Jaume Bertranpetit
Universitat Pompeu Fabra
Barcelona

INSTITUT de BIOLOGIA EVOLUTIVA
CSIC UPF

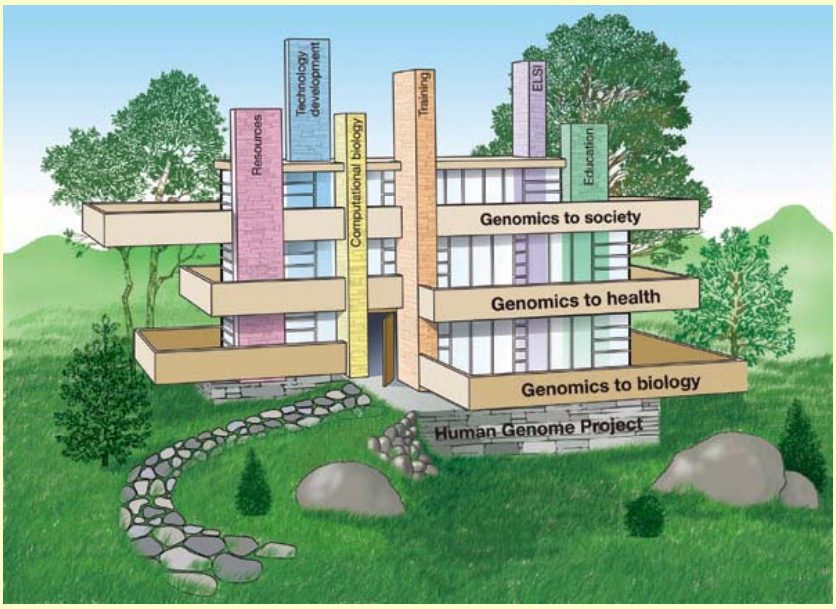
UNIVERSITAT POMPEU FABRA

The revolution in the Life Sciences



Living organisms may be viewed as the only part of the natural world whose members contain internal description of themselves.
This is why the whole biology must be rooted in the DNA, and our task is still to discover how these DNA sequences arose in evolution and how they are interpreted in to build the diversity of the living world, including disease.

Sydney Brenner, dec 2012, Science



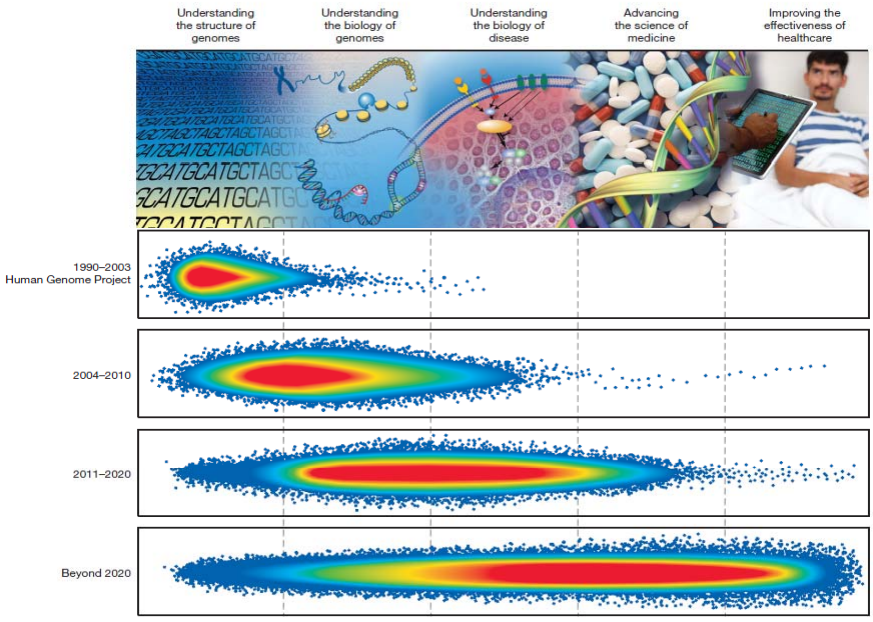
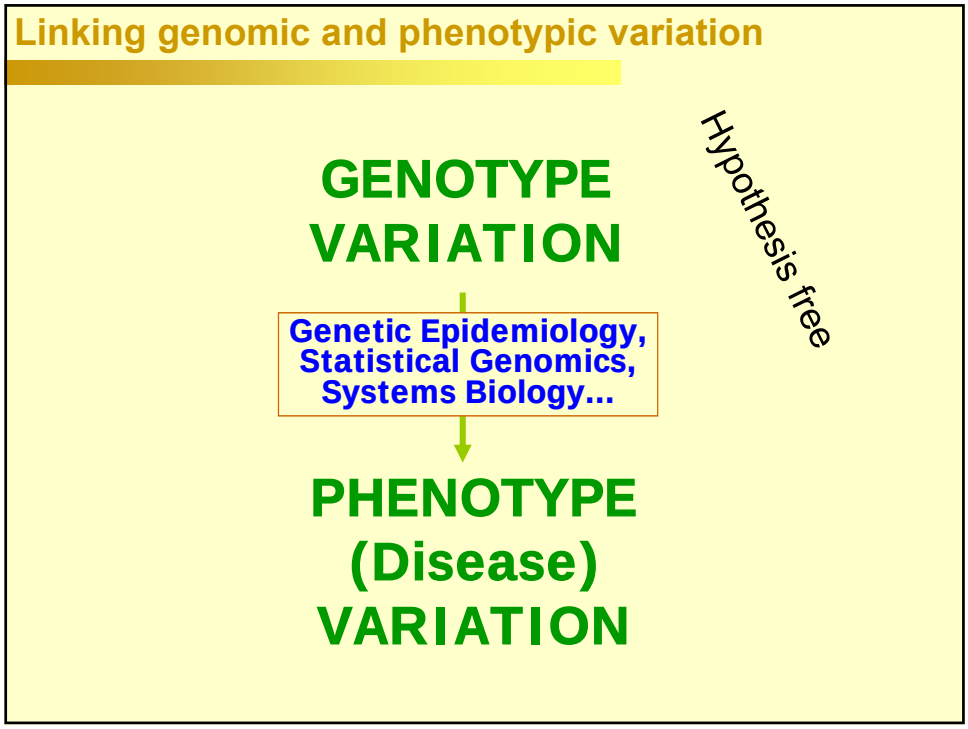
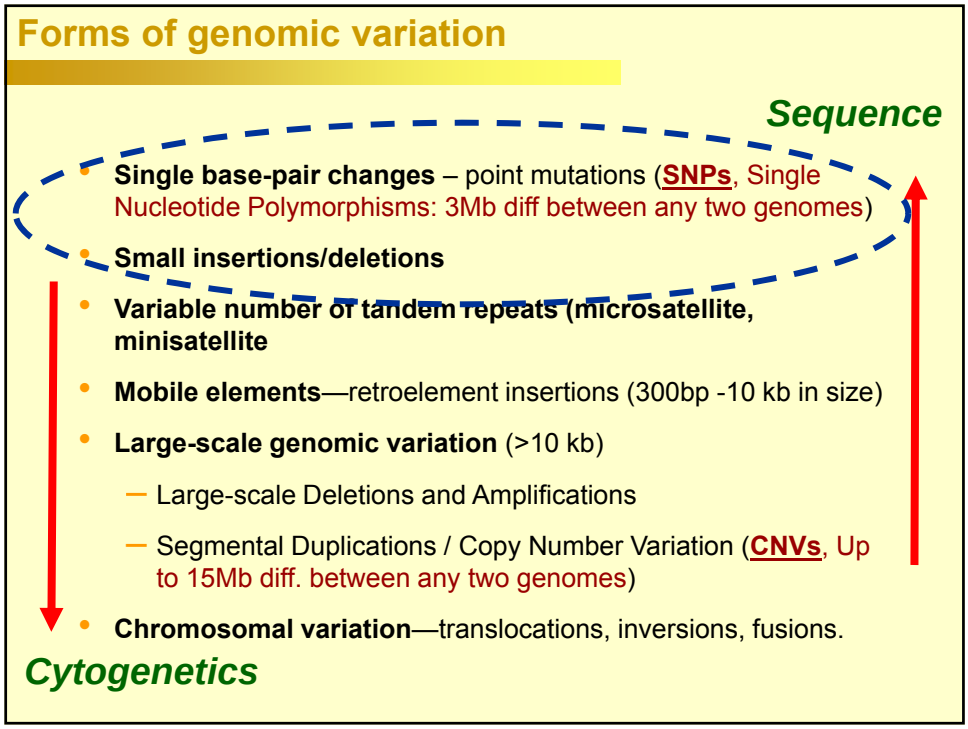


Figure 2 | Schematic representation of accomplishments across five domains of genomics research. The progression from base pairs to bedside is depicted in five sequential, overlapping domains (indicated along the top). Genomic accomplishments across the domains are portrayed by hypothetical, highly schematized density plots (each blue dot reflecting a single research accomplishment, with green, yellow and red areas reflecting sequentially higher densities of accomplishments). Separate plots are shown for four time intervals: the HGP; the period covered by the 2003 NHGRI vision for the future of genomics research[17]; the period described here (2011–2020); and the open-ended future beyond 2020.

Linking genomic and phenotypic variation



Forms of genomic variation



Forms of genomic variation

SNPs in the Human Genome

- Normally biallelic (two variants)
- Allelic frequencies present differences between populations.
- Non human specific
- Phenotypic effect: normally without effect

- Public databases for SNPs (you will see in other lessons)

http://www.nhgri.nih.gov, http://www.snp.cshl.org

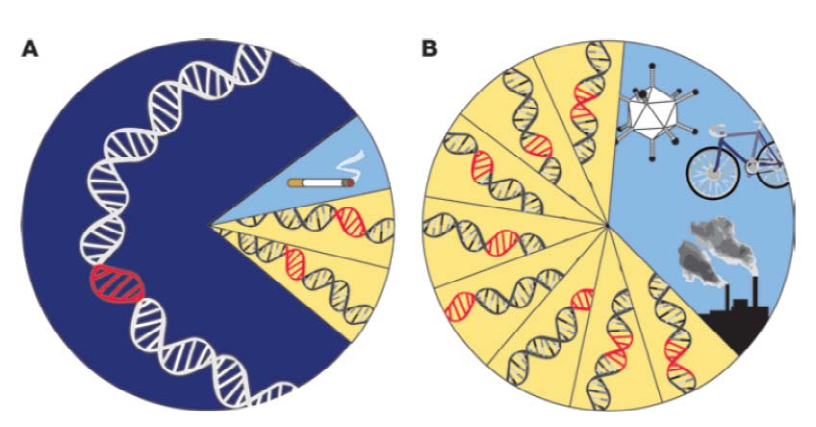
Disease

✓Complex diseases

- High prevalence
  - -Cardiopathy
  - -Reumatoid Arthritis
  - -Aunoimmune disease
  - -Psichiatic diseases
  - -Cancer
- Relevant genetic component
- No simple mendelian inheritance pattern
- Several genes with small contrintion (susceptibility)
- Multiple alleles interacting among themselves and with the environment.

Statistical genetic methods for disease gene identification

Genes + Environment



Enfermedades monogénicas **Enfermedades complejas**

Manolio, el al. *J. Clin. Invest*. (2008) 118:590-1605

Heritability

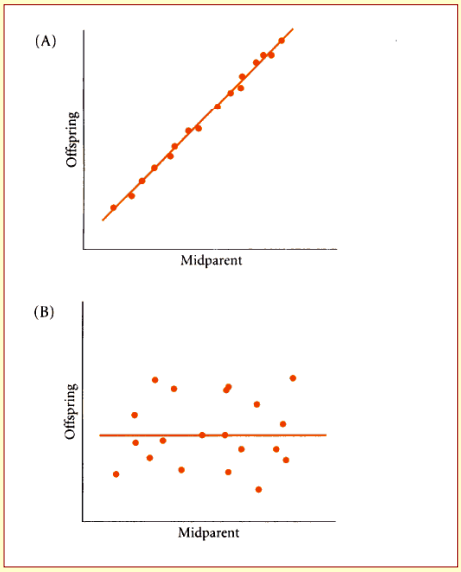
Heritability is the proportion of Phenotypic Variance that is accounted for by Genomic variance

$$h^2 = V_A/V_P$$

Heritability

Heritability is usually misinterpreted

$h^2 = 1$

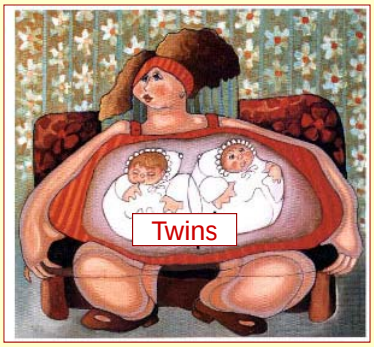


$V_A/V_P = 1$

$h^2 = 0$

$V_A/V_P = 0$

Heritability

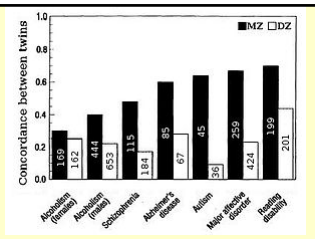


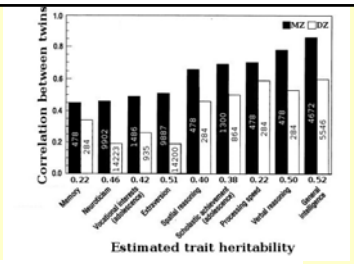
Heritability estimations: Twin classification

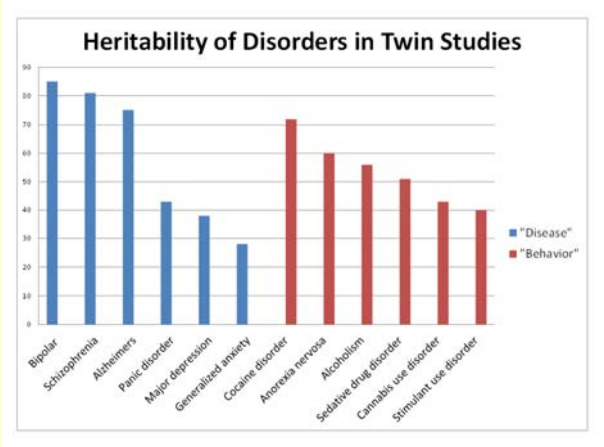
Monozygotic (MZ) or identical twins:
Share 100% of their genes and 100% of E (assumed)

Dizygotic (DZ) or fraternal twins:
Share 50 % of their genes 100% of environment (assumed)

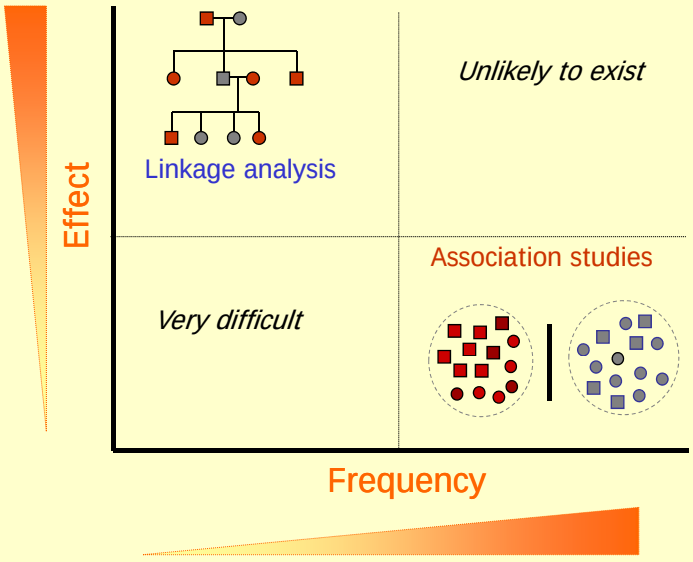
- Prevalence: 1/80 births approx.
  - 1/3 MZ
  - 1/3 same sex fraternal
  - 1/3 opposite sex fraternal







Statistical genetic methods for disease gene identification



Fishing associations with genetic markers

Linkage studies

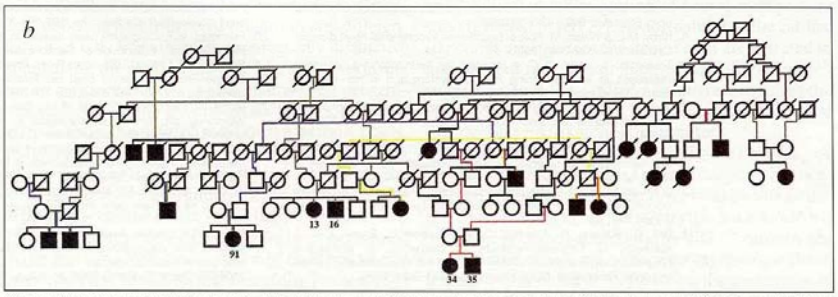


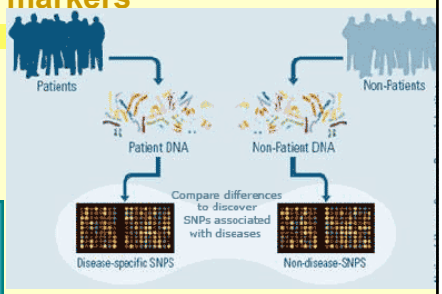
Fig. 1 *a*, Woman belonging to one of the large Bulgarian HMSNL kindreds. She has three affected sons, six affected grandchildren and, so far, one affected great-grandson. *b*, Simplified version of the Lom HMSNL kindred. The coloured lines follow the segregation of the core HMSNL haplotypes. Segment sharing analysis was conducted on individuals 91, 13 and 34.

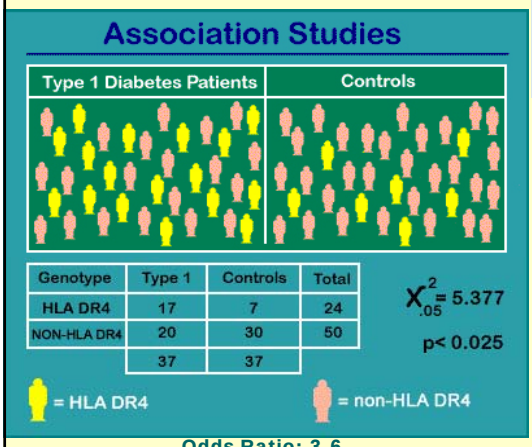
Linkage analysis works best with relatively rare, highly-heritable, well-defined phenotypes

... that is, Mendelian traits

Fishing associations with genetic markers

Association studies





Odds Ratio: 3.6
95% CI = 1.3 to 10.4

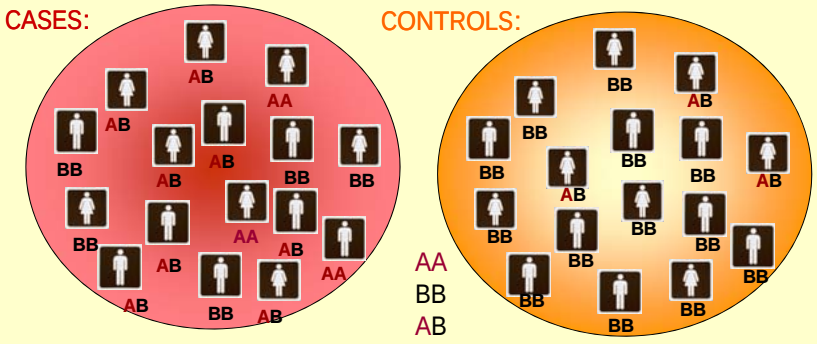
Association Studies Design — Steps

1. Sample collection
2. Marker selection
3. Genotyping
4. Quality control
5. Analysis
6. Follow up

Association Studies Design — Steps – Sample Collection

1. Sample collection
2. Marker selection
3. Genotyping
4. Quality control
5. Analysis
6. Follow up

Association Studies Design Steps – Sample Collection

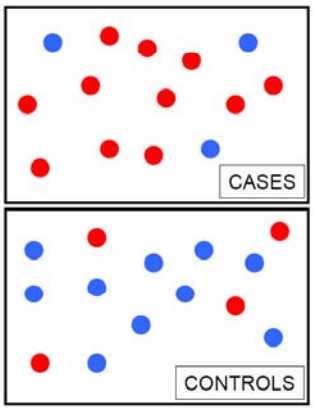


The most popular design for studying genetic association between a disease and a marker is to compare a set of cases with a set of controls (case-control design) collected from the same population. It is important to collect data on exposure to potential non-genetic (environmental) risk factors

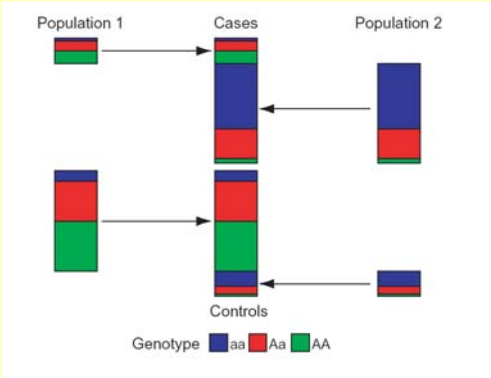
Association Studies Design Steps – Sample Collection

Population structure in genetic association studies

- Population consists of underlying subpopulations.
- Disease prevalence different between subpopulations.
- Cases preferentially ascertained from specific subpopulations.
- False positive evidence of association will occur at genetic markers that differ in genotype frequencies between the subpopulations.
- Traditionally, human geneticists have been skeptical of case-control studies for this reason.



Association Studies Design Steps – Sample Collection



BUT

In general populations there is a big problem with heterogeneity:

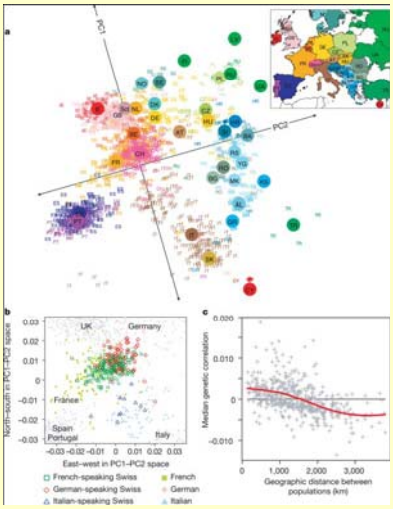
- Ambiguous phenotypes
- Multiple target genes in a functional unit
- Multiple target sites in a gene
- Multiple alleles in a site (recurrent mutations)

Solutions:
- Phenotype stringency
- Extreme phenotypes
- Big numbers

Association Studies Design Steps – Sample Collection

More benefits from HapMap and related projects

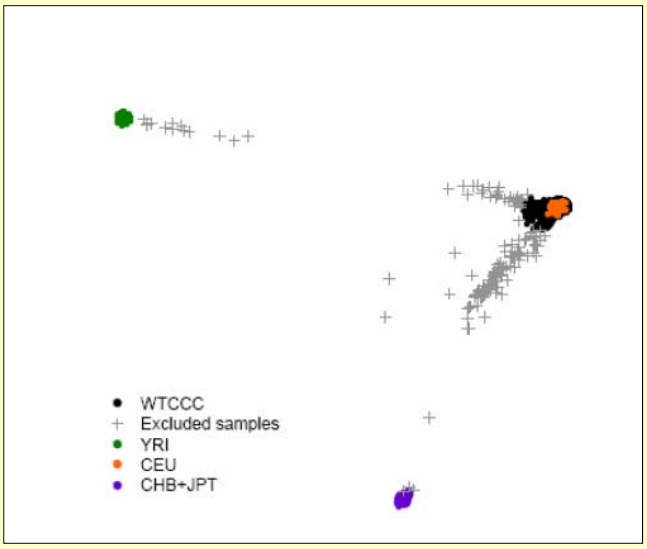
Population structure within Europe.



November et al Nature **456**, 98-101 (6 November 2008)

A posteriori control of stratification (brute force!!)

The WTCCC case: non-fitting subjects just removed



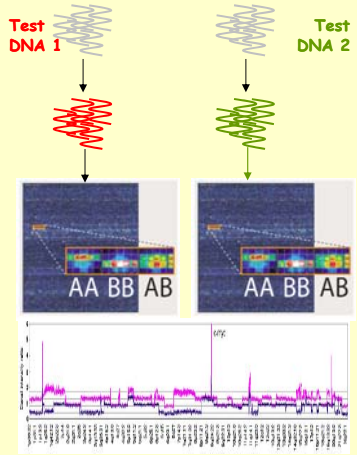
1. Sample collection
2. Marker selection
3. Genotyping
4. Quality control
5. Analysis
6. Follow up

The HapMap Project

The Hapmap Project allows proper WGASs

Different genotyping strategies are available…

- **Candidate Gene Approach.** You select tag-SNPs from a set of genes or regions of interest.
- **Whole Genome Scans.** You use an pre-designed, commercial array. Up to more than 2.5 million SNPs.



Association Studies Design Steps – Genotyping

1. Sample collection
2. Marker selection
3. Genotyping -sequencing
4. Quality control
5. Analysis
6. Follow up

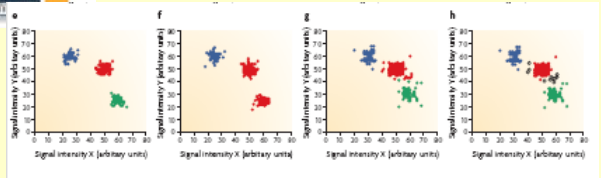
The HapMap Project

The Hapmap Project allows proper WGASs

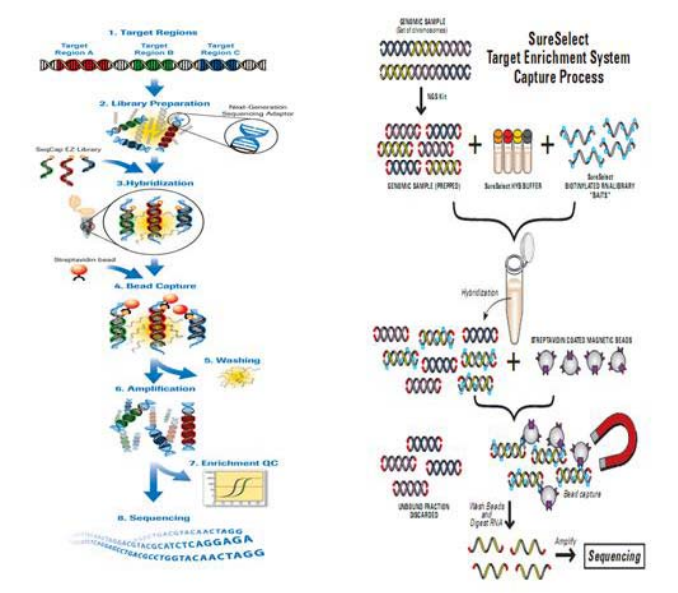
Different genotyping technologies are available…

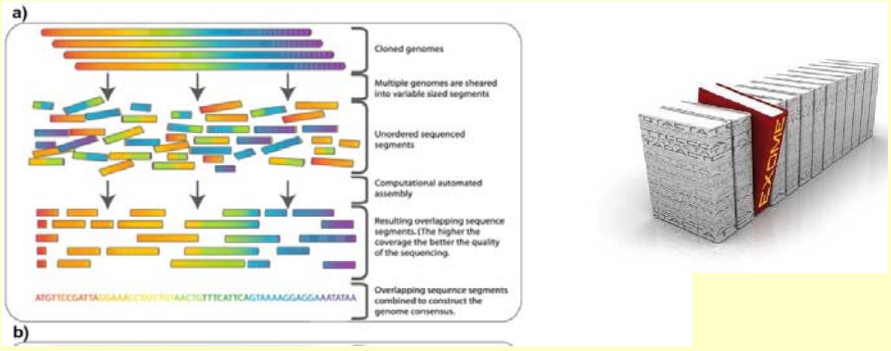
Association Studies Design Steps – Genotyping

GATTTAGATCGCGATAGAG
GATTTAGATCTCGATAGAG



Enrichment





Whole genome / exome shotgun sequencing

Genotyping Arrays

Technology moves fast....

Table 1. Genetic Progress through Technology.*

Scientific Advance	Technological Platform	Explanation	Reference
Sequencing of the human genome	Whole-genome expression arrays	Allows the expression of all genes to be determined by hybridization	Lander et al.,[12] Venter et al.,[13] Su et al.[14]
Human HapMap			
SNP technology		Demonstrates that individual SNPs predict adjacent SNPs and therefore suggests that genotyping of <500,000 SNPs may allow a nearly complete survey of all common genetic variability	The International HapMap Consortium[2]
Genome genotyping	Whole-genome SNP genotyping arrays	Allows whole-genome associations to be performed for common diseases, the commercial consequence of the HapMap	Sladek et al.[15]
High-throughput analysis	High-throughput sequencing techniques	Allows DNA sequencing that is faster and cheaper than conventional sequencing	Margulies et al.[16]
		Allows the expression of all RNA species, including different splice forms to be assessed in any tissue	Brenner et al.[17]
		Allows individual full-coding genome sequencing, together with whole-genome arrays that hybridize and bind to all exons	Olson[18]
Sequencing of the individual genome		Opens the way for personal genome sequencing	Wheeler et al.[19]
1000 Genomes Project		Allows the identification of comparatively rare polymorphic changes by placing the full genome sequences of 1000 anonymous subjects into the public domain	1000 Genomes Project[20]
Genotype-Tissue Expression (GTEx) project		Allows the creation of haplotypic gene-expression databases for many human tissues	NIH Roadmap for Medical Research[21]

* NIH denotes National Institutes of Health, and SNP single-nucleotide polymorphism.

1. Sample collection
2. Marker selection
3. Genotyping
4. Quality control
5. Analysis
6. Follow up

Once we have our genotypes we have to clean our data from possible errors that could lead us to false positives or to losing power:

- technical genotyping errors
- cnv, segmental duplications interactions
- sample contaminations
- sample degradations
- populations structure

For Pedigree data

¿How to detect genotyping errors?

Mendelian errors

Replicates
Controls
X chromosome
Hardy-Weinberg

1. Sample collection
2. Marker selection
3. Genotyping
4. Quality control
5. Analysis
6. Follow up

Contingency Table:

Risk factor

	Yes	No	
Cases	n_{11}	n_{12}	$n_{1.}$
Controls	n_{21}	n_{22}	$n_{2.}$
	$n_{.1}$	$n_{.2}$	$n_{..}$

n_{11}= exposed with the disease
n_{21}= exposed without the disease
n_{12}= not exposed with the disease
n_{22}= not exposed without the disease

$n_{1.}$= total number with disease
$n_{2.}$= total number without disease

$n_{1.}$= total number of exposed
$n_{.1.}$= total number of non-exposed

Test of independence:

$$\chi_1^2 = \frac{\Sigma\ (O-E)^2}{E}$$

Statistic used to test for the significance of any differences

Association Studies Design Steps – Analysis

The association strength between a marker and the disease status is usually measured by odd-ratio (OR)

Risk factor

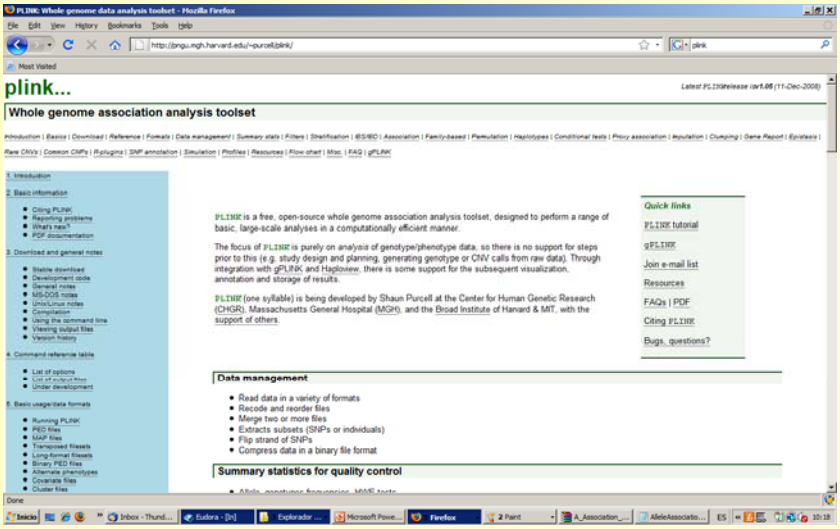
	Yes	No	
Cases	n_{11}	n_{12}	$n_{1.}$
Controls	n_{21}	n_{22}	$n_{2.}$
	$n_{.1}$	$n_{.2}$	$n_{..}$

$$\text{Odds-ratio (OR)} = \frac{n_{11} * n_{22}}{n_{21} * n_{12}}$$

OR=1 No difference in risk of exposed
OR>1 Increased risk of exposed to the risk factor
OR<1 Lower risk ("protective") of exposed to the risk factor

- **The 95% confidence interval (CI) of OR contains both information on "strength" and "significance"**
- **When the sample size is increased, typically the p-value can become even more significant, whereas OR usually stays the same (but 95% CI of OR becomes more narrow).**

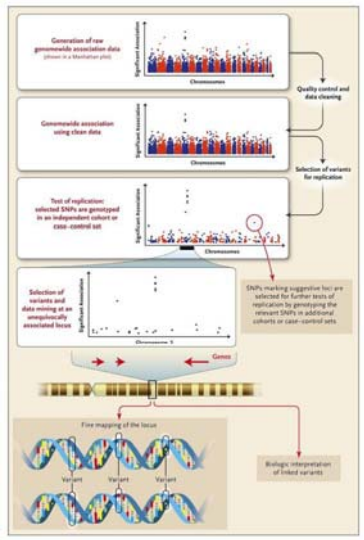
Association Studies Design Steps – Analysis

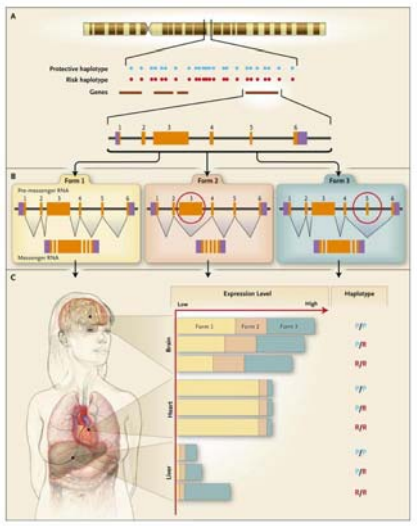


Association Studies Design Steps – Follow up

1. Sample collection
2. Marker selection
3. Genotyping
4. Quality control
5. Analysis
6. Follow up

Association Studies Design Steps – Follow up

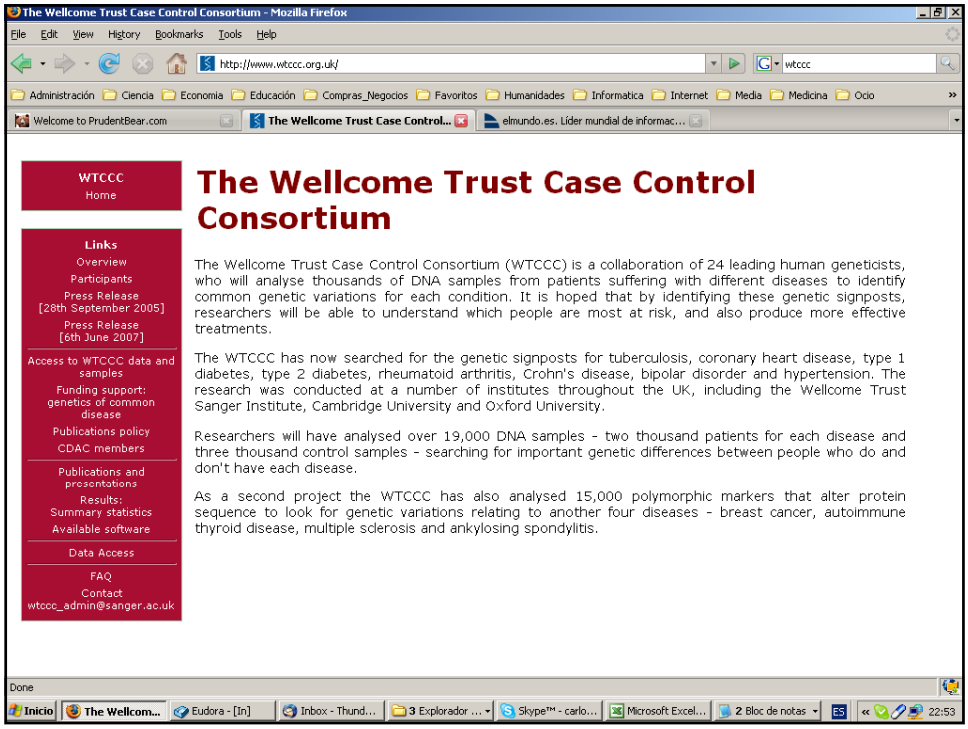




Association Studies Design Steps – Follow up

We have found an interesting association of some of our SNPs. What else should be done?

- Functional characterization. Looking around the place if there are genes, microRNA's, CNV's etc... (remember that SNPs are nothing more than markers)
- High density genotyping
- Resequencing
- Expression studies
- Replication (without replication there is no publication)



WTCCC

NATURE|Vol 447|7 June 2007 ARTICLES

Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls

The Wellcome Trust Case Control Consortium*

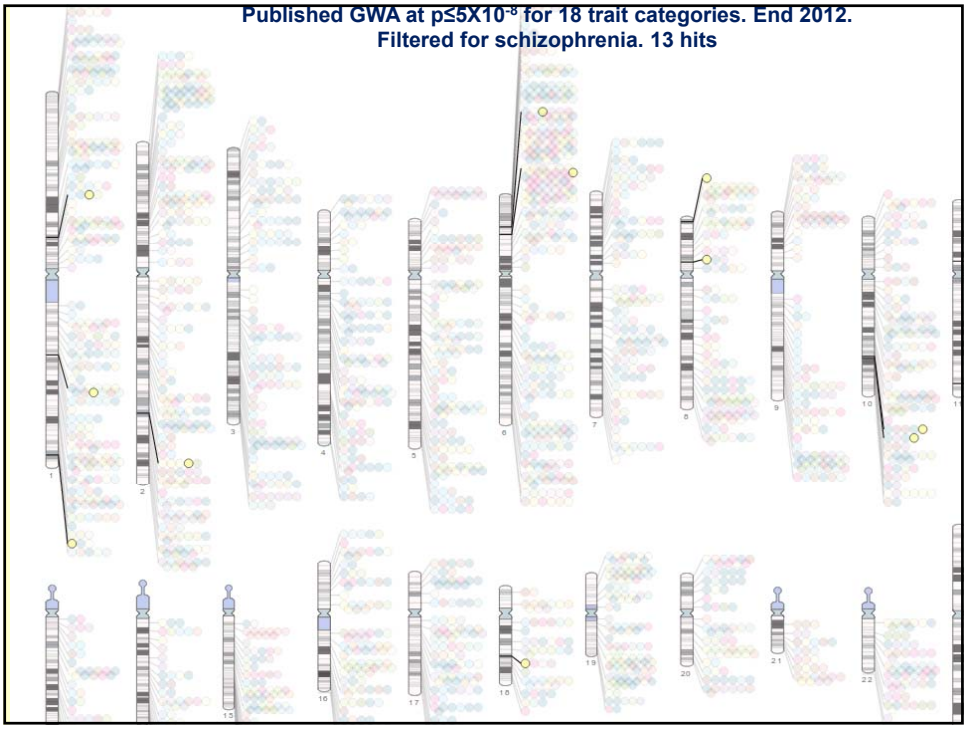
The Wellcome Trust Case Control Consortium (WTCCC) is a UK wide collaboration of over 50 research groups (clinical, experimental and analytical) aiming for a better understanding of common human diseases.

A combined genome wide association (GWA) study of 7 common diseases with British samples. 2,000 cases from each of seven diseases:

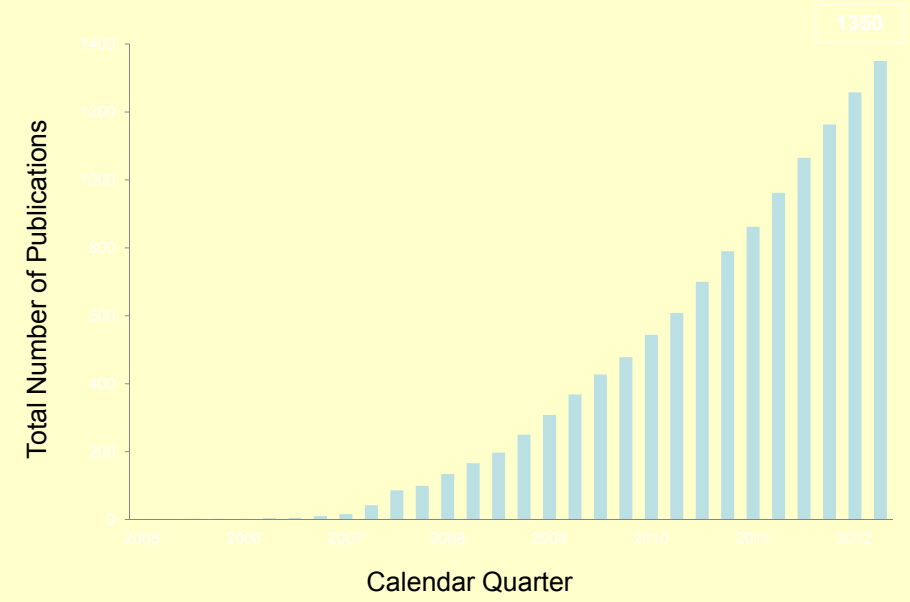
- bipolar disorder (BD)
- coronary artery disease (CAD)
- Crohn's disease (CD)
- hypertension (HT)
- rheumatoid arthritis (RA)
- type 1 diabetes (T1D)
- type 2 diabetes (T2D)

Published Genome-Wide Associations through 07/2012
Published GWA at p≤5X10^{-8} for 18 trait categories. End 2012

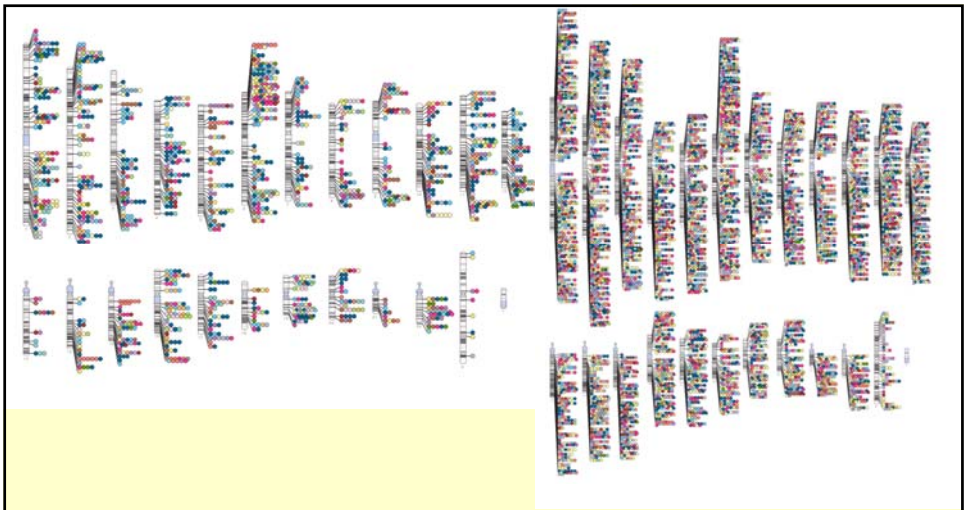
Published GWA at p≤5X10^{-8} for 18 trait categories. End 2012.
Filtered for schizophrenia. 13 hits



Published GWA Reports, 2005 – 6/2012

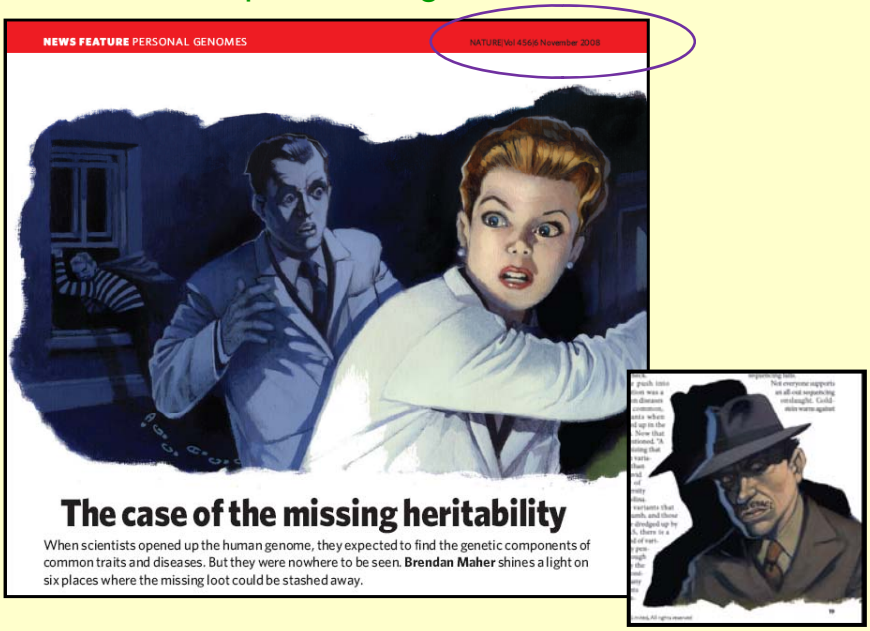


Through 6/30/12 postings



GWAS diagram July 2012, SNP-trait associations with p-value ≤ 1×10-11, PNG (left) and p-value ≤ 1×10^{-5} (right)

In spite of the great success….



In spite of the great success….

Vol 461|8 October 2009|doi:10.1038/nature08494 nature

REVIEWS

Finding the missing heritability of complex diseases

Teri A. Manolio[1], Francis S. Collins[2], Nancy J. Cox[3], David B. Goldstein[4], Lucia A. Hindorff[5], David J. Hunter[6], Mark I. McCarthy[7], Erin M. Ramos[5], Lon R. Cardon[8], Aravinda Chakravarti[9], Judy H. Cho[10], Alan E. Guttmacher[1], Augustine Kong[11], Leonid Kruglyak[12], Elaine Mardis[13], Charles N. Rotimi[14], Montgomery Slatkin[15], David Valle[9], Alice S. Whittemore[16], Michael Boehnke[17], Andrew G. Clark[18], Evan E. Eichler[19], Greg Gibson[20], Jonathan L. Haines[21], Trudy F. C. Mackay[22], Steven A. McCarroll[23] & Peter M. Visscher[24]

Genome-wide association studies have identified hundreds of genetic variants associated with complex human diseases and traits, and have provided valuable insights into their genetic architecture. Most variants identified so far confer relatively small increments in risk, and explain only a small proportion of familial clustering, leading many to question how the remaining, 'missing' heritability can be explained. Here we examine potential sources of missing heritability and propose research strategies, including and extending beyond current genome-wide association approaches, to illuminate the genetics of complex diseases and enhance its potential to enable effective disease prevention or treatment.

In spite of the great success….

Table 1 | Estimates of heritability and number of loci for several complex traits

Disease	Number of loci	Proportion of heritability explained	Heritability measure
Age-related macular degeneration[72]	5	50%	Sibling recurrence risk
Crohn's disease[21]	32	20%	Genetic risk (liability)
Systemic lupus erythematosus[73]	6	15%	Sibling recurrence risk
Type 2 diabetes[74]	18	6%	Sibling recurrence risk
HDL cholesterol[75]	7	5.2%	Residual* phenotypic varia
Height[15]	40	5%	Phenotypic variance
Early onset myocardial infarction[76]	9	2.8%	Phenotypic variance
Fasting glucose[77]	4	1.5%	Phenotypic variance

* Residual is after adjustment for age, gender, diabetes.

In spite of the great success….

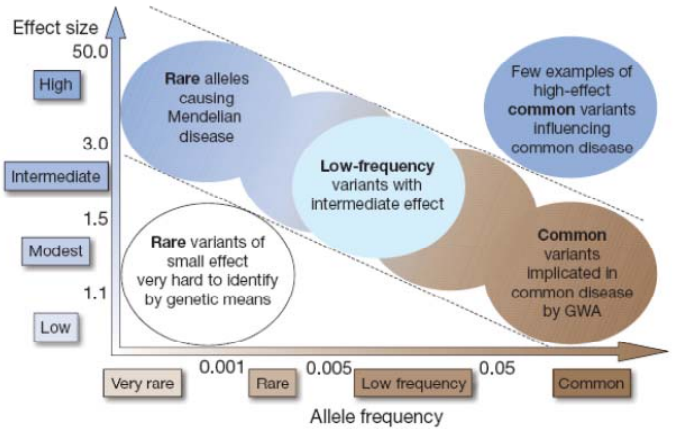


Figure 1 | Feasibility of identifying genetic variants by risk allele frequency and strength of genetic effect (odds ratio). Most emphasis and interest lies in identifying associations with characteristics shown within diagonal dotted lines. Adapted from ref. 42.

A good idea to check this review

www.nature.com/reviews/genetics 356 | MAY 2008 | VOLUME 9

Genome-wide association studies for complex traits: consensus, uncertainty and challenges

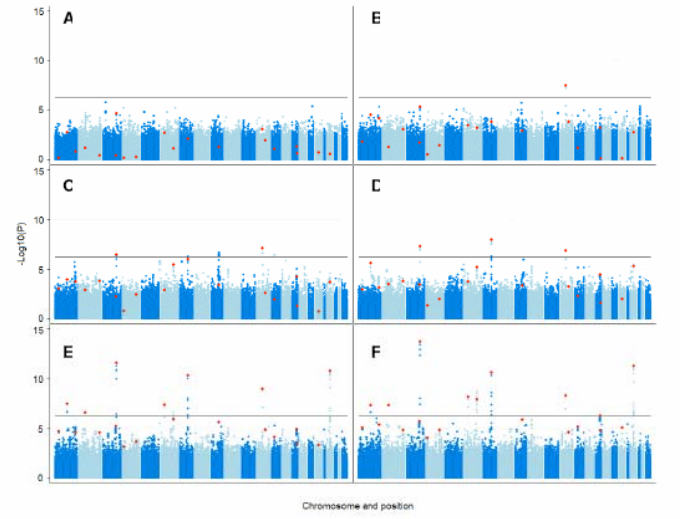
Mark I. McCarthy[*‡], *Gonçalo R. Abecasis*[§], *Lon R. Cardon*[*‖], *David B. Goldstein*[¶], *Julian Little*[#], *John P. A. Ioannidis*[**‡‡] *and Joel N. Hirschhorn*[§§‖‖¶¶]

Abstract | The past year has witnessed substantial advances in understanding the genetic basis of many common phenotypes of biomedical importance. These advances have been the result of systematic, well-powered, genome-wide surveys exploring the relationships between common sequence variation and disease predisposition. This approach has revealed over 50 disease-susceptibility loci and has provided insights into the allelic architecture of multifactorial traits. At the same time, much has been learned about the successful prosecution of association studies on such a scale. This Review highlights the knowledge gained, defines areas of emerging consensus, and describes the challenges that remain as researchers seek to obtain more complete descriptions of the susceptibility architecture of biomedical traits of interest and to translate the information gathered into improvements in clinical management.

52

A metanalysis on heigh

Metananalysis of human heigh.... From 1.9k to 19k...
And nothing until 18k individuals!!!
And not what we expected!!!



But GWAS are not free of problems

Table 2. Benefits, Misconceptions, and Limitations of the Genomewide Association Study.

Benefits
Does not require an initial hypothesis
Uses digital and additive data that can be mined and augmented without data degradation
Encourages the formation of collaborative consortia, which tend to continue their collaboration for subsequent analyses
Rules out specific genetic associations (e.g., by showing that no common alleles, other than *APOE*, are associated with Alzheimer's disease with a relative risk of more than 2)
Provides data on the ancestry of each subject, which assists in matching case subjects with control subjects
Provides data on both sequence and copy-number variations
Misconceptions
Thought to provide data on all genetic variability associated with disease, when in reality only common alleles with large effects are identified
Thought to screen out alleles with a small effect size, when in reality such findings may still be very useful in determining pathogenic biochemical pathways, even though low-risk alleles may be of little predictive value
Limitations
Requires samples from a large number of case subjects and control subjects and therefore can be challenging to organize
Finds loci, not genes, which can complicate the identification of pathogenic changes on an associated haplotype
Detects only alleles that are common (>5%) in a population
Requires replication in a similarly large number of samples

Issues generating false associations:

Population stratification

- Need to be careful.

Multiple Comparisons

- Need smarter ways of analyzing data.

Rare variants

- Need to get more data…

Allelic heterogeneity

- When multiple disease variants exist at the same gene, a single marker may not capture them well enough.
- Haplotype-based association analysis is good theoretically, but it hasn't shown its advantage in practice.

Locus heterogeneity

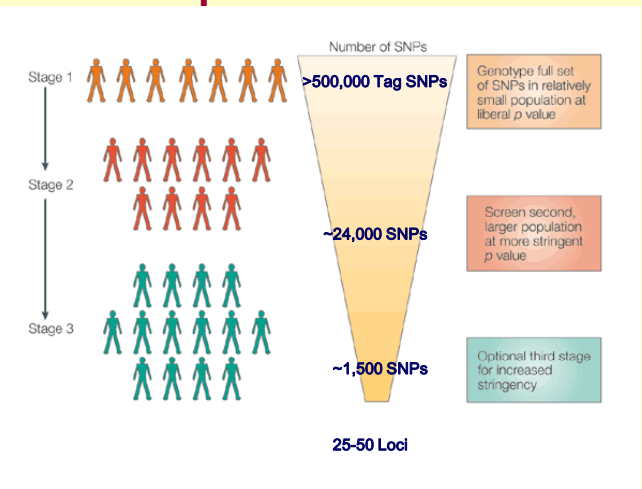
- Multiple genes may influence the disease risk independently. As a result, for any single gene, a fraction of the cases may be no different from the controls.

Effect modification (interaction) between two genes may exist with weak/no marginal effects.

- It is unknown how often this happens in reality. But when this happens, analyses that only look at marginal effects won't be useful.
- It often requires larger sample size to have reasonable power to detect interaction effects than the sample size needed to detect marginal effects.

Replication is necessary

Replication Studies are a must



Replica

Replica

Replica

Hirschhorn & Daly *Nat. Genet. Rev.* 6: 95, 2005
NCI-NHGRI Working Group on Replication *Nature* 447: 655, 2007

Many more GWAS

And lots of information has accumulated

The NCBI dbGaP database of genotypes and phenotypes

Matthew D Mailman, Michael Feolo, Yumi Jin, Masato Kimura, Kimberly Tryka, Rinat Bagoutdinov, Luning Hao, Anne Kiang, Justin Paschall, Lon Phan, Natalia Popova, Stephanie Pretel, Lora Ziyabari, Moira Lee, Yu Shao, Zhen Y Wang, Karl Sirotkin, Minghong Ward, Michael Kholodov, Kerry Zbicz, Jeffrey Beck, Michael Kimelman, Sergey Shevelev, Don Preuss, Eugene Yaschenko, Alan Graeff, James Ostell & Stephen T Sherry

The National Center for Biotechnology Information has created the dbGaP public repository for individual-level phenotype, exposure, genotype and sequence data and the associations between them. dbGaP assigns stable, unique identifiers to studies and subsets of information from those studies, including documents, individual phenotypic variables, tables of trait data, sets of genotype data, computed phenotype-genotype associations, and groups of study subjects who have given similar consents for use of their data.

NATURE GENETICS | VOLUME 39 | NUMBER 10 | OCTOBER 2007

http://www.nature.com/ng/journal/v39/n10/abs/ng1007-1181.html

NCBI created database (2007!....*well, december 2006*)

The main aim is to uniformely store primary data from published studies

57

Other genetic elements
- Chromosome deletions
- Copy number variants
- Micro RNA
- Metilation
- Rare variants
-…….

58

Rare vs. Frequent variants

Table 2 Characteristics of common and rare disease variants compared

Common disease variants	Rare disease variants
Discovery by population association, case-control studies, using genome-wide markers (WGA)	Discovery by DNA resequencing of candidate genes, preferably in early onset cases with one or more relatives affected
Mostly MAF > 5%	MAF > 0.1% to 2–3% Higher than rare familial mutations, lower than polymorphisms. Often population specific.
Explained by LD with functional variant	Not detected by WGA
OR mostly between 1.2 and 1.5 Higher ORs could be due to recent natural selection	OR mostly ≥ 2
No familial concentration	No familial concentration
Need large studies with control for ethnic heterogeneity to achieve statistical significance and minimize false positives	Assess significance by increased frequencies in cases vs. controls and by functional analysis of variant
Make substantial contribution to PAR	Summation of effects of several variants make significant contribution to PAR
Low penetrance makes prophylactic intervention unlikely	Penetrance often high enough to justify prophylactic interventions
Hard to find functionally relevant variant	Variants identified are functionally relevant
Contribution to disease etiology questionable	Make a contribution to understanding disease etiology
May suggest candidates for rare variant search	Effect may be modified by common variants

Rare variants and CNVs

Table 2 | Selected disease associations with rare CNVs and common CNPs

Disease	Locus	Type of CNV	Size (kb)	Population frequency	Case frequency	Effect size (OR)
Rare CNVs						
Autism/IMR[59]	16p11.2	*De novo* deletion	600	1×10^{-4}	1%	100
Autism[59]	16p11.2	Rare duplication	600	3×10^{-4}	0.50%	16
Schizophrenia[60,78]	1q21.1	Rare deletion	1,400	2×10^{-4}	0.30%	15
IMR[79]	1q21.1	Rare deletion	1,400	2×10^{-4}	0.47%	Not observed in 4,737 controls
Schizophrenia[60,78]	15q13.3	Rare deletion	1,600	2×10^{-4}	0.20%	12
Epilepsy[80]	15q13.3	Rare deletion	1,600	2×10^{-4}	1.0%	Not observed in 3,699 controls
IMR[79,81]	15q13.3	Rare deletion	1,600	2×10^{-4}	0.30%	Not observed in 960 controls
Schizophrenia[82]	22q11.2	Rare deletion	3,000	2.5×10^{-4}	1%	40
Common CNPs						
Crohn's disease[83]	*IRGM*	Deletion polymorphism	20	7%	11%	1.5
Body mass index[61]	*NEGR1*	Deletion polymorphism	45	65%	Quantitative trait	<1 kg
Psoriasis[84]	*LCE3C*	Deletion polymorphism	30	55%	65%	1.3

IMR, idiopathic mental retardation.

Rare variants may add noise

Rare Variants Create Synthetic Genome-Wide Associations

Samuel P. Dickson[1,2], Kai Wang[3], Ian Krantz[3,4,5], Hakon Hakonarson[3,4,5], David B. Goldstein[1]*

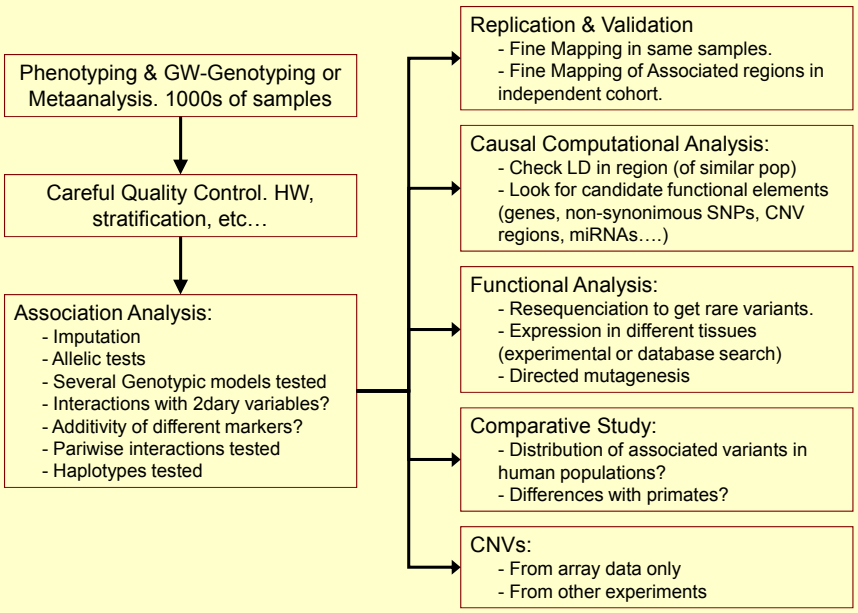
1 Institute for Genome Sciences and Policy, Center for Human Genome Variation, Duke University, Durham, North Carolina, United States of America, **2** Bioinformatics Research Center, North Carolina State University, Raleigh, North Carolina, United States of America, **3** Center for Applied Genomics, Children's Hospital of Pennsylvania, Philadelphia, Pennsylvania, United States of America, **4** Division of Human Genetics, Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, United States of America, **5** Department of Pediatrics, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania, United States of America

Abstract

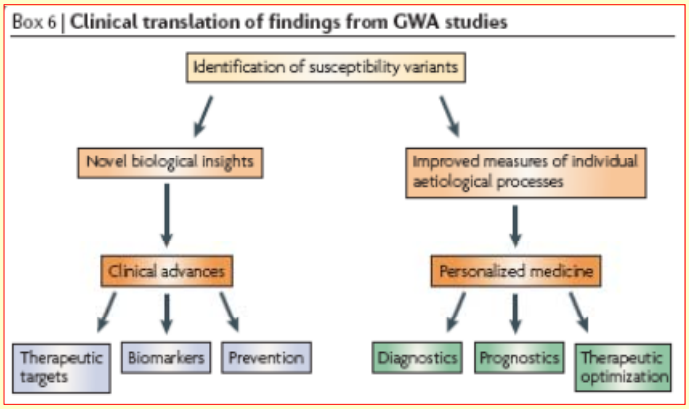
Genome-wide association studies (GWAS) have now identified at least 2,000 common variants that appear associated with common diseases or related traits (http://www.genome.gov/gwastudies), hundreds of which have been convincingly replicated. It is generally thought that the associated markers reflect the effect of a nearby common (minor allele frequency >0.05) causal site, which is associated with the marker, leading to extensive resequencing efforts to find causal sites. We propose as an alternative explanation that variants much less common than the associated one may create "synthetic associations" by occurring, stochastically, more often in association with one of the alleles at the common site versus the other allele. Although synthetic associations are an obvious theoretical possibility, they have never been systematically explored as a possible explanation for GWAS findings. Here, we use simple computer simulations to show the conditions under which such synthetic associations will arise and how they may be recognized. We show that they are not only possible, but inevitable, and that under simple but reasonable genetic models, they are likely to account for or contribute to many of the recently identified signals reported in genome-wide association studies. We also illustrate the behavior of synthetic associations in real datasets by showing that rare causal mutations responsible for both hearing loss and sickle cell anemia create genome-wide significant synthetic associations, in the latter case extending over a 2.5-Mb interval encompassing scores of "blocks" of associated variants. In conclusion, uncommon or rare genetic variants can easily create synthetic associations that are credited to common variants, and this possibility requires careful consideration in the interpretation and follow up of GWAS signals.

PLoS Biol January 2010 8(1): e1000294

What can be done?



…and the goals…



63

DISEASE MECHANISMS

Genetic architectures of psychiatric disorders: the emerging picture and its implications

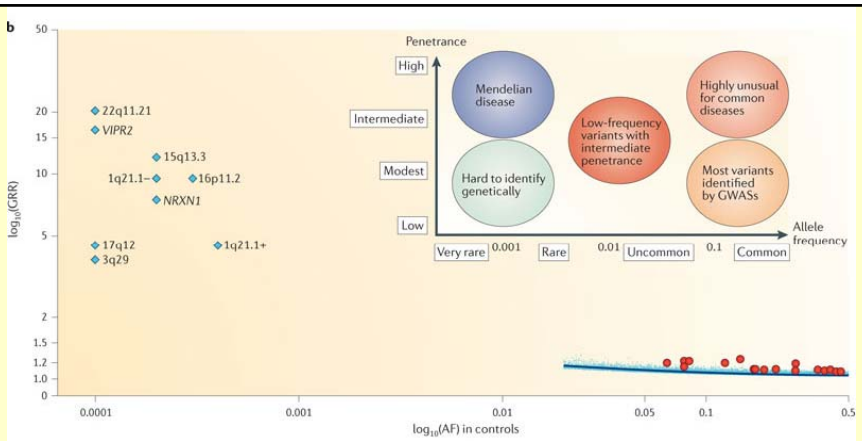
Patrick F. Sullivan[1], Mark J. Daly[2] and Michael O'Donovan[3]

Abstract | Psychiatric disorders are among the most intractable enigmas in medicine. In the past 5 years, there has been unprecedented progress on the genetics of many of these conditions. In this Review, we discuss the genetics of nine cardinal psychiatric disorders (namely, Alzheimer's disease, attention-deficit hyperactivity disorder, alcohol dependence, anorexia nervosa, autism spectrum disorder, bipolar disorder, major depressive disorder, nicotine dependence and schizophrenia). Empirical approaches have yielded new hypotheses about aetiology and now provide data on the often debated genetic architectures of these conditions, which have implications for future research strategies. Further study using a balanced portfolio of methods to assess multiple forms

Table 1 | **Defining features of nine psychiatric disorders***

Name	Life prevalence	Heritability	Essential characteristics
Alzheimer's disease	0.132	0.58	Dementia, defining neuropathology
Attention-deficit hyperactivity disorder (ADHD)	0.053	0.75	Persistent inattention, hyperactivity, impulsivity
Alcohol dependence (ALC)	0.178	0.57	Persistent ethanol use despite tolerance, withdrawal, dysfunction
Anorexia nervosa	0.006	0.56	Dangerously low weight from self-starvation
Autism spectrum disorder (ASD)	0.001	0.80	Markedly abnormal social interaction and communication beginning before age 3
Bipolar disorder (BIP)	0.007	0.75	Manic-depressive illness, episodes of mania, usually with major depressive disorder
Major depressive disorder (MDD)	0.130	0.37	Unipolar depression, marked and persistent dysphoria with physical and cognitive symptoms
Nicotine dependence (NIC)	0.240	0.67	Persistent nicotine use with physical dependence (usually cigarettes)
Schizophrenia (SCZ)	0.004	0.81	Long-standing delusions and hallucinations

*Most of these definitions are made more restrictive by requiring persistence over time (for example ≥6 months of symptoms), substantial impairment and presence across multiple different contexts. See S1 (table) for more detail. Additional sources are REFS 1,2,181–185).



b | Allelic spectrum of schizophrenia (SCZ). The inset is a conceptual schematic from a 2008 *Nature Reviews Genetics* article10. The lower part of the figure depicts empirical results for SCZ. The *x*-axis is log10(allele frequency (AF)) in controls. The *y*-axis is the point estimate for log10(genotypic relative risk (GRR). For clarity, confidence intervals are not shown. There are no known Mendelian variants for SCZ (AF << 0.0001, GRR >> 50). There are no known common variants (AF > 0.1) with GRR > 1.5, and these can be excluded with >99% statistical power. Nine structural variants associated with SCZ are shown as light blue diamonds (TABLE 2; 1q21.1– is the deletion and 1q21.1+ is the duplication). If AF in controls was 0, AF was set to 0.0001. These structural variants do not have a corresponding region in the inset. Seventeen common variants have been associated with SCZ (red circles; TABLE 3). SNPs contributing to the Psychiatric Genomics Consortium SCZ risk profile score59

Table 2 | **Structural variation associated with psychiatric disorders**

Structural variant	Location (Mb)	Genes	Type	Disorder	Frequency in cases	Frequency in controls	Odds ratio	*P* value	Other associations	Refs
1q21.1	chr1: 145.0–148.0	34	Deletion	SCZ	0.0018	0.0002	9.5	8×10^{-6}	Developmental delay, intellectual disability, micro- and macrocephaly, dysmorphia, epilepsy, cataracts, cardiac defects, possibly ASD[185], thrombocytopenia-absent radius syndrome[44,159,184–186]	184
			Duplication	SCZ	0.0013	0.0004	4.5	0.02		184
2p16.3	chr2: 50.1–51.2	*NRXN1* exons	Deletion	ASD				0.004	Developmental delay, intellectual disability, epilepsy, Pitt–Hopkins-like syndrome 2	81
			Deletion	SCZ	0.0018	0.0002	7.5	1×10^{-6}		184
3q29	chr3: 195.7–197.3	19	Deletion	SCZ	0.0010	0.0	3.8	4×10^{-4}	Developmental delay, intellectual disability, possibly ASD	184
7q11.23	chr7: 72.7–74.1	25	Duplication	ASD	0.0011			0.003	Developmental delay, intellectual disability. Deletion: Williams–Beuren syndrome	81
7q36.3	chr7: 158.8–158.9	*VIPR2*	Duplication	SCZ	0.0024	0.0001	16.4	4×10^{-5}		44, 184
15q11.2	chr15: 23.6–28.4	70	Duplication	ASD	0.0018			4×10^{-9}	Developmental delay, intellectual disability, Prader–Willi and Angelman syndromes[188]	81
15q13.3	chr15: 30.9–33.5	12	Duplication	ADHD	0.0125	0.0061	2.1	2×10^{-4}	Developmental delay, intellectual disability, epilepsy[188,189]	120
			Duplication	ASD	0.0013			2×10^{-5}		81
			Deletion	SCZ	0.0019	0.0002	12.1	7×10^{-7}		184
16p13.11	chr16: 15.4–16.3	8	Duplication	ADHD	0.0164	0.0009	13.9	8×10^{-4}	Deletion: developmental delay, epilepsy[188,189]	119
16p11.2	chr16: 29.5–30.2	29	Deletion	ASD	0.0037			5×10^{-20}	Developmental delay, intellectual disability, epilepsy, macrocephaly, obesity[190,191]	81
			Duplication	ASD	0.0013			2×10^{-5}	Developmental delay, intellectual disability, epilepsy, microcephaly, low body mass index[190,191]	81
			Duplication	SCZ	0.0031	0.0003	9.5	3×10^{-8}		184
17q12	chr17: 34.8–36.2	18	Deletion	ASD	0.0017	0.0	6.12	9×10^{-4}		192
			Deletion	SCZ	0.0006	0.0	4.49	3×10^{-4}		
22q11.21	chr22: 18.7–21.8	53	Deletion or duplication	ASD	0.0013			0.002	Developmental delay, intellectual disability, velocard-iofacial-DiGeorge syndrome	81
			Deletion	SCZ	0.0031	0.0	20.3	7×10^{-13}		184

Locations are US National Center for Biotechnology Information (NCBI) Build 37 and University of California, Santa Cruz (UCSC) hg19. The positions of these structural variants are denoted in Supplementary information S3 (figure) with yellow circles. For succinctness, the citations refer to the most comprehensive study rather than to an initial report. 'Genes' refers to the number from the UCSC Known Genes data set. ADHD, attention-deficit hyperactivity disorder; ASD, autism spectrum disorder;

Table 3 | **Genome-wide association study findings for psychiatric disorders**

Phenotype	SNP	Location	Discovery GWAS (cases/controls)	Largest meta-analysis (cases/controls)	*P* value	Odds ratio	Nearest gene
Alzheimer's disease	rs3818361	chr1:207784968	2,018/5,324 (REF. 34)	<19,870/30,846 (REF. 35)	3.7×10^{-14}	1.18	*CR1*
	rs744373	chr2:127894615	3,006/14,642 (REF. 193)	<19,870/30,846 (REF. 35)	2.6×10^{-14}	1.17	*BIN1*
	rs9349407	chr6:47453378	8,309/7,366 (REF. 36)	18,762/29,827 (REF. 36)	8.6×10^{-9}	1.11	*CD2AP*
	rs11767557	chr7:143109139	8,309/7,366 (REF. 36)	18,762/39,997 (REF. 36)	6.0×10^{-10}	1.11	*EPHA1*
	rs11136000	chr8:27464510	3,941/7,848 (REF. 33)	8,371/26,065 (REF. 193)	1.6×10^{-19}	1.18	*CLU*
	rs610932	chr11:59930307	6,688/13,251 (REF. 35)	>19,000/38,000 (REF. 35)	1.2×10^{-19}	1.10	*MS4A cluster*
	rs3851179	chr11:85868640	3,941/7,848 (REF. 33)	8,371/26,065 (REF. 193)	3.2×10^{-17}	1.15	*PICALM*
	rs3764650	chr19:1046520	5,509/11,531 (REF. 35)	>17,000/34,000 (REF. 35)	5.0×10^{-21}	1.23	*ABCA7*
	rs2075650	chr19:45395619		8,371/26,065 (REF. 193)	1×10^{-205}	2.53	*APOE, TOMM40*
	rs3865444	chr19:51727962	8,309/7,366 (REF. 36)	18,762/29,827 (REF. 36)	1.6×10^{-8}	1.10	*CD33*
Alcohol consumption	rs1229984	chr4:100239319	REF. 102		1.3×10^{-11}		*ADH1B*
	rs6943555	chr7:50806023	REF. 101		4.1×10^{-8}		*AUTS2*
	rs671	chr12:112241766	REF. 100		3×10^{-111}		*ALDH2*
Bipolar disorder	rs12576775	chr11:79077193	7,481/9,251 (REF. 60)	11,974/51,793 (REF. 60)	4.4×10^{-8}	1.14	*ODZ4*
	rs4765913	chr12:2419896	7,481/9,250 (REF. 60)	11,974/51,792 (REF. 60)	1.5×10^{-8}	1.14	*CACNA1C*
	rs1064395	chr19:19361735	682/1300 (REF. 194)	8,441/35,362 (REF. 194)	2.1×10^{-8}	1.17	*NCAN*
Nicotine consumption	rs1329650	chr10:93345120	38,181 (REF. 93)	73,853 (REF. 93)	5.7×10^{-10}		*LOC100188947*
	rs1051730	chr15:78804339	38,181 (REF. 93)	73,853 (REF. 93)	2.8×10^{-23}		*CHRNA3*
	rs3733829	chr19:41310571	38,181 (REF. 93)	73,853 (REF. 93)	1.0×10^{-8}		*EGLN2, CYP2A6*
Smoking cessation	rs3025343	chr9:136478355	41,278 (REF. 93)	64,924 (REF. 93)	3.6×10^{-8}	1.13	*DBH*
Smoking initiation	rs6265	chr11:27679916	74,035 (REF. 93)	143,023 (REF. 93)	1.8×10^{-8}	0.94	*BDNF*
Schizophrenia	rs1625579	chr1:98607034	9,394/12,462 (REF. 59)	17,839/33,859 (REF. 59)	1.6×10^{-11}	1.12	*MIR137*
	rs2312147	chr2:58222928		18,206/42,536 (REF. 195)	1.9×10^{-8}	1.09	*VRK2*
	rs1344706	chr2:165778428	479/2,937 (REF. 174)	18,945/38,675 (REF. 196)	2.5×10^{-11}	1.10	*ZNF804A*
	rs17662626	chr2:193984621	9,394/12,463 (REF. 59)	17,839/33,860 (REF. 59)	4.6×10^{-8}	1.20	*PCGEM1*
	rs13211507	chr6:28257377	3,322/3,587 (REF. 70)	18,206/42,536 (REF. 195)	1.4×10^{-13}	1.22	*MHC*
	rs7004635	chr8:3360967	9,394/12,465 (REF. 59)	17,839/33,862 (REF. 59)	2.7×10^{-8}	1.10	*MMP16*
	rs1050253	chr8:4180844	9,394/12,464 (REF. 59)	17,839/33,861 (REF. 59)	4.1×10^{-8}	1.11	*CSMD1*
	rs16887244	chr8:38031345	3,750/6,468 (REF. 68)	8,133/11,007 (REF. 68)	1.3×10^{-10}	1.19	*LSM1*
	rs7014558	chr10:104775008	9,394/12,466 (REF. 59)	17,839/33,863 (REF. 59)	1.8×10^{-8}	1.10	*CNNM2*
	rs11101580	chr10:104906211	9,394/12,467 (REF. 59)	17,839/33,864 (REF. 59)	1.1×10^{-8}	1.15	*NT5C2*
	rs11810869	chr11:46580680	1,160/3,714 (REF. 197)	3,738/7,802 (REF. 197)	3.0×10^{-8}	1.25	*AMBRA1*
	rs12807809	chr11:124606285		18,206/42,536 (REF. 195)	2.8×10^{-8}	1.12	*NRGN*
	rs12966547	chr18:52752017	9,394/12,468 (REF. 59)	17,839/33,865 (REF. 59)	2.6×10^{-10}	1.09	*CCDC68*
	rs9960767	chr18:53155002		18,206/42,537 (REF. 195)	4.2×10^{-8}	1.20	*TCF4*

The emerging picture

-Multiple high confidence structural variants

-Rare exonic variants

-Increasing number of robustly significant and replicated common variants

Novel biological hypothesis

-Cholesterol metabolism in Alzheimer disease

Reinforce previous hypothesis

-- synaptic biology for schizophrenia or autism

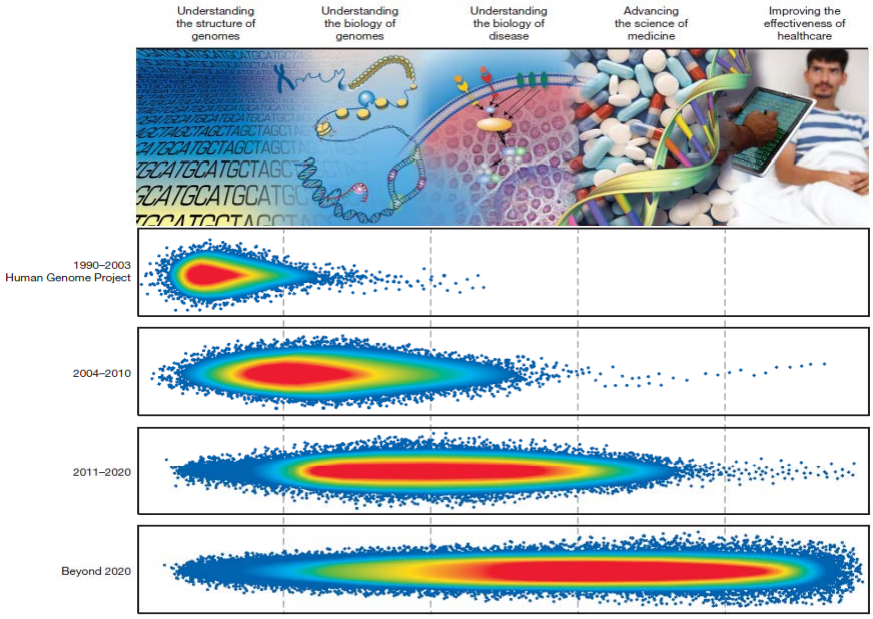


Figure 2 | Schematic representation of accomplishments across five domains of genomics research. The progression from base pairs to bedside is depicted in five sequential, overlapping domains (indicated along the top). Genomic accomplishments across the domains are portrayed by hypothetical, highly schematized density plots (each blue dot reflecting a single research accomplishment, with green, yellow and red areas reflecting sequentially higher densities of accomplishments). Separate plots are shown for four time intervals: the HGP; the period covered by the 2003 NHGRI vision for the future of genomics research[17]; the period described here (2011–2020); and the open-ended future beyond 2020.