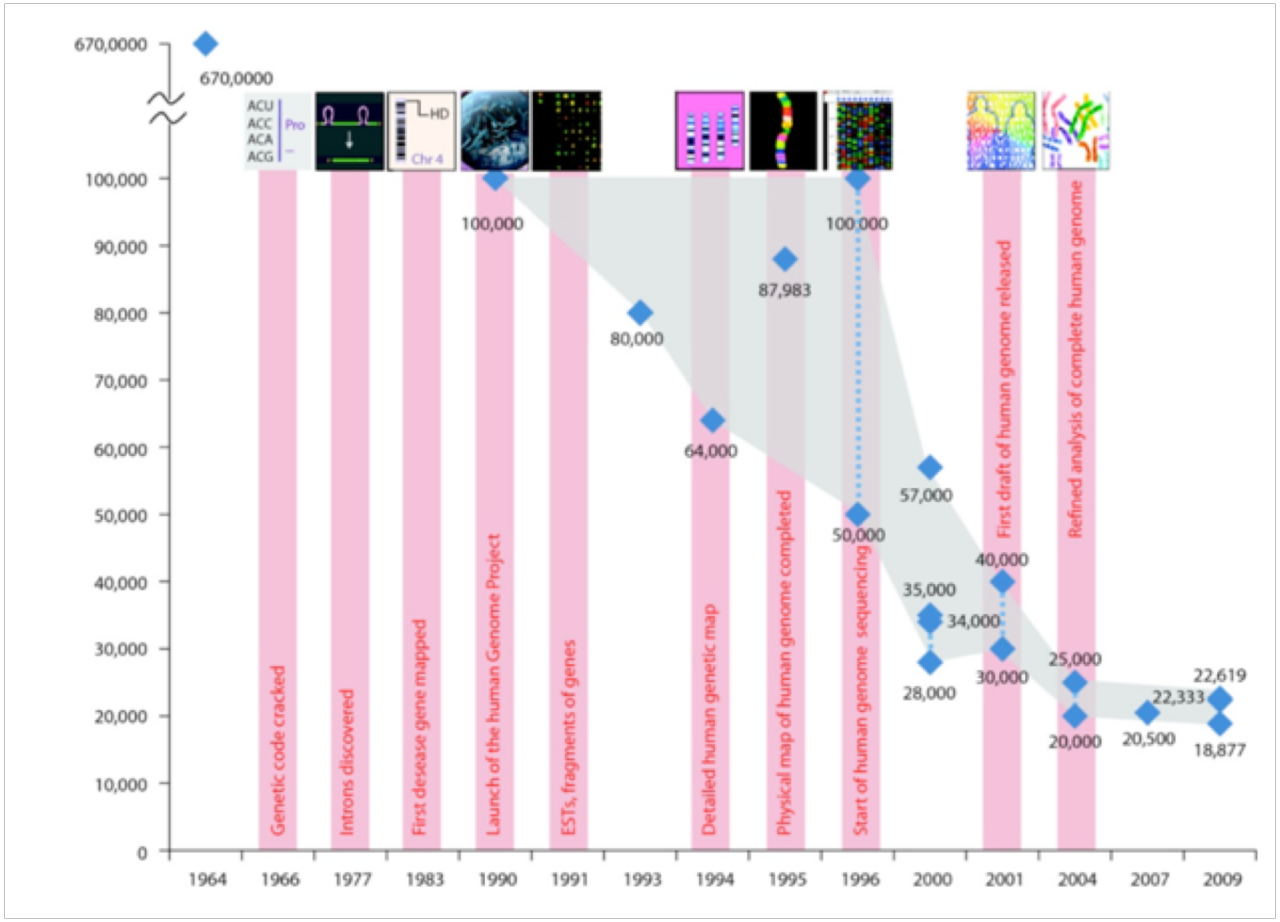
The Human Genome Project “paid forward” and paved the way for modern day genomics

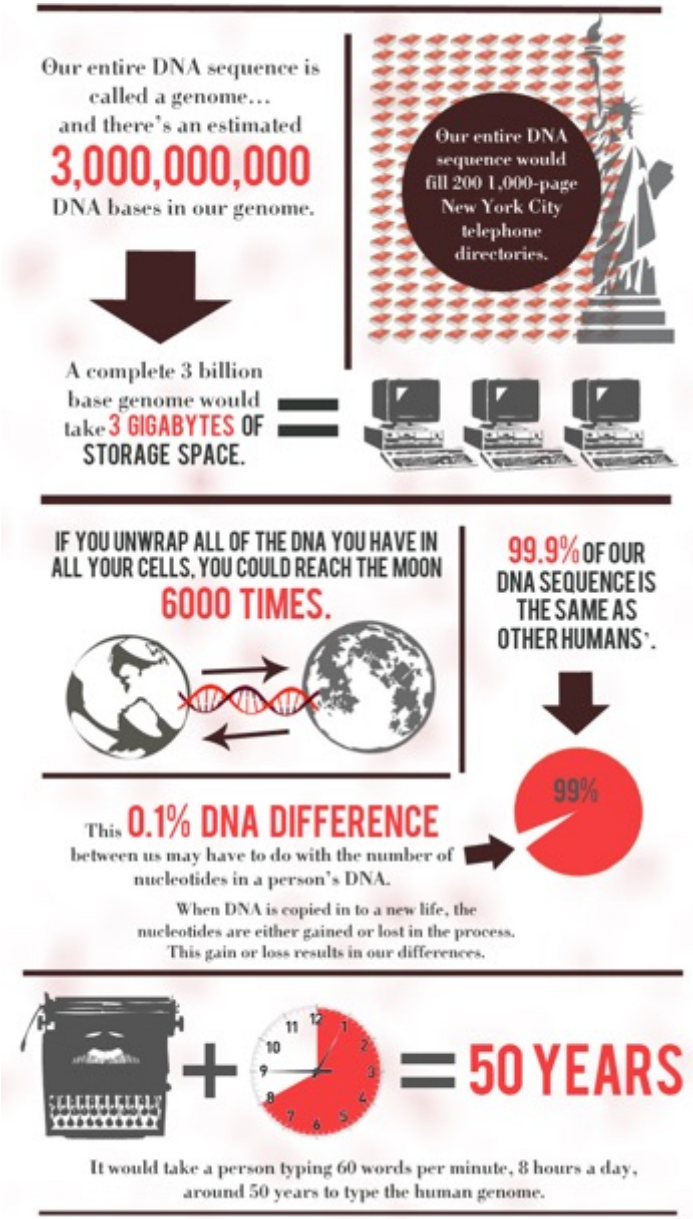
(Finally) a complete map

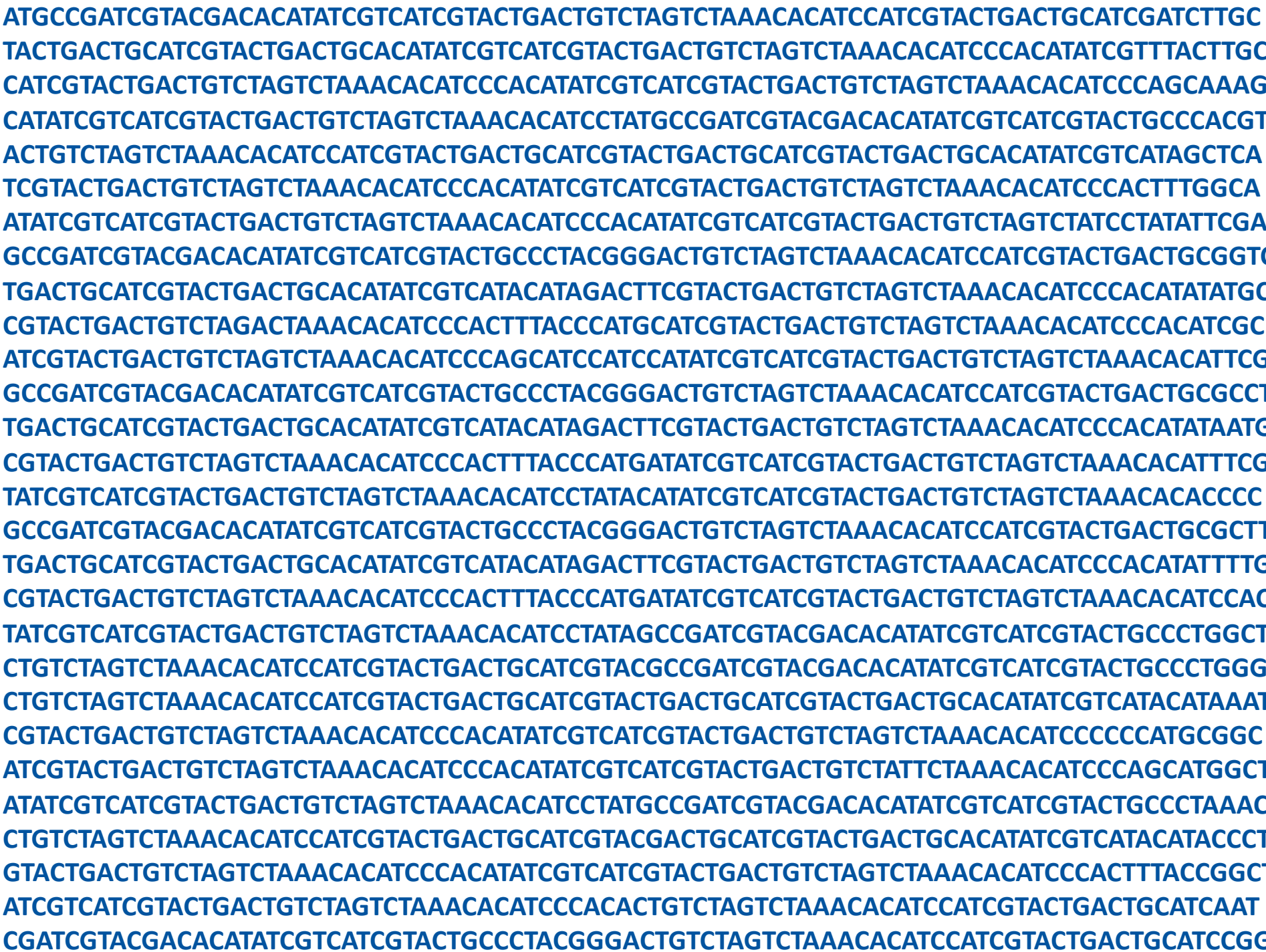


Pertea and Salzberg, Genome Biology 2010

Human Genome: *some statistics*

- 3.2 billion base pairs in the haploid genome
- ≈ 18,000-25,000 genes
  - ≈23,000 coding for proteins
  - Only 1.5% of the total genome
- Rest of the genome:
  - Non-coding RNA (rRNA, tRNA)
  - Regulatory sequences, e.g. promoter, enhancer regions
  - Repetitive elements and other variations
  - Transposable elements
- (So there's no such thing as "junk DNA"…)




ATGCCGATCGTACGACACATATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCATCGTACTGACTGCATCGATCTTGC
TACTGACTGCATCGTACTGACTGCACATATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCCACATATCGTTTACTTGC
CATCGTACTGACTGTCTAGTCTAAACACATCCCACATATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCCAGCAAAG
CATATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCTATGCCGATCGTACGACACATATCGTCATCGTACTGCCCACGT
ACTGTCTAGTCTAAACACATCCATCGTACTGACTGCATCGTACTGACTGCATCGTACTGACTGCACATATCGTCATAGCTCA
TCGTACTGACTGTCTAGTCTAAACACATCCCACATATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCCACTTTGGCA
ATATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCCACATATCGTCATCGTACTGACTGTCTAGTCTATCCTATATTCGA
GCCGATCGTACGACACATATCGTCATCGTACTGCCCTACGGGACTGTCTAGTCTAAACACATCCATCGTACTGACTGCGGTC
TGACTGCATCGTACTGACTGCACATATCGTCATACATAGACTTCGTACTGACTGTCTAGTCTAAACACATCCCACATATATGC
CGTACTGACTGTCTAGACTAAACACATCCCACTTTACCCATGCATCGTACTGACTGTCTAGTCTAAACACATCCCACATCGC
ATCGTACTGACTGTCTAGTCTAAACACATCCCAGCATCCATCCATATCGTCATCGTACTGACTGTCTAGTCTAAACACATTCG
GCCGATCGTACGACACATATCGTCATCGTACTGCCCTACGGGACTGTCTAGTCTAAACACATCCATCGTACTGACTGCGCCT
TGACTGCATCGTACTGACTGCACATATCGTCATACATAGACTTCGTACTGACTGTCTAGTCTAAACACATCCCACATATAATG
CGTACTGACTGTCTAGTCTAAACACATCCCACTTTACCCATGATATCGTCATCGTACTGACTGTCTAGTCTAAACACATTTCG
TATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCTATACATATCGTCATCGTACTGACTGTCTAGTCTAAACACACCCC
GCCGATCGTACGACACATATCGTCATCGTACTGCCCTACGGGACTGTCTAGTCTAAACACATCCATCGTACTGACTGCGCTT
TGACTGCATCGTACTGACTGCACATATCGTCATACATAGACTTCGTACTGACTGTCTAGTCTAAACACATCCCACATATTTTG
CGTACTGACTGTCTAGTCTAAACACATCCCACTTTACCCATGATATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCAC
TATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCTATAGCCGATCGTACGACACATATCGTCATCGTACTGCCCTGGCT
CTGTCTAGTCTAAACACATCCATCGTACTGACTGCATCGTACGCCGATCGTACGACACATATCGTCATCGTACTGCCCTGGG
CTGTCTAGTCTAAACACATCCATCGTACTGACTGCATCGTACTGACTGCATCGTACTGACTGCACATATCGTCATACATAAAT
CGTACTGACTGTCTAGTCTAAACACATCCCACATATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCCCCCATGCGGC
ATCGTACTGACTGTCTAGTCTAAACACATCCCACATATCGTCATCGTACTGACTGTCTATTCTAAACACATCCCAGCATGGCT
ATATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCTATGCCGATCGTACGACACATATCGTCATCGTACTGCCCTAAAC
CTGTCTAGTCTAAACACATCCATCGTACTGACTGCATCGTACGACTGCATCGTACTGACTGCACATATCGTCATACATACCCT
GTACTGACTGTCTAGTCTAAACACATCCCACATATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCCACTTTACCGGCT
ATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCCACACTGTCTAGTCTAAACACATCCATCGTACTGACTGCATCAAT
CGATCGTACGACACATATCGTCATCGTACTGCCCTACGGGACTGTCTAGTCTAAACACATCCATCGTACTGACTGCATCCGG


ATGCCGATCGTACGACACATATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCATCGTACTGACTGCATCGATCTTGC
TACTGACTGCATCGTACTGACTGCACATATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCCACATATCGTTTACTTGC
CATCGTACTGACTGTCTAGTCTAAACACATCCCACATATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCCAGCAAAG
CATATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCTATGCCGATCGTACGACACATATCGTCATCGTACTGCCCACGT
ACTGTCTAGTCTAAACACATCCATCGTACTGACTGCATCGTACTGACTGCATCGTACTGACTGCACATATCGTCATAGCTCA
TCGTACTGACTGTCTAGTCTAAACACATCCCACATATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCCACTTTGGCA
ATATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCCACATATCGTCATCGTACTGACTGTCTAGTCTATCCTATATTCGA
GCCGATCGTACGACACATATCGTCATCGTACTGCCCTACGGGACTGTCTAGTCTAAACACATCCATCGTACTGACTGCGGTC
TGACTGCATCGTACTGACTGCACATATCGTCATACATAGACTTCGTACTGACTGTCTAGTCTAAACACATCCCACATATATGC
CGTACTGACTGTCTAGGCTAAACACATCCCACTTTACCCATGCATCGTACTGACTGTCTAGTCTAAACACATCCCACATCGC
ATCGTACTGACTGTCTAGTCTAAACACATCCCAGCATCCATCCATATCGTCATCGTACTGACTGTCTAGTCTAAACACATTCG
GCCGATCGTACGACACATATCGTCATCGTACTGCCCTACGGGACTGTCTAGTCTAAACACATCCATCGTACTGACTGCGCCT
TGACTGCATCGTACTGACTGCACATATCGTCATACATAGACTTCGTACTGACTGTCTAGTCTAAACACATCCCACATATAATG
CGTACTGACTGTCTAGTCTAAACACATCCCACTTTACCCATGATATCGTCATCGTACTGACTGTCTAGTCTAAACACATTTCG
TATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCTATACATATCGTCATCGTACTGACTGTCTAGTCTAAACACACCCC
GCCGATCGTACGACACATATCGTCATCGTACTGCCCTACGGGACTGTCTAGTCTAAACACATCCATCGTACTGACTGCGCTT
TGACTGCATCGTACTGACTGCACATATCGTCATACATAGACTTCGTACTGACTGTCTAGTCTAAACACATCCCACATATTTTG
CGTACTGACTGTCTAGTCTAAACACATCCCACTTTACCCATGATATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCAC
TATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCTATAGCCGATCGTACGACACATATCGTCATCGTACTGCCCTGGCT
CTGTCTAGTCTAAACACATCCATCGTACTGACTGCATCGTACGCCGATCGTACGACACATATCGTCATCGTACTGCCCTGGG
CTGTCTAGTCTAAACACATCCATCGTACTGACTGCATCGTACTGACTGCATCGTACTGACTGCACATATCGTCATACATAAAT
CGTACTGACTGTCTAGTCTAAACACATCCCACATATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCCCCCATGCGGC
ATCGTACTGACTGTCTAGTCTAAACACATCCCACATATCGTCATCGTACTGACTGTCTACTCTAAACACATCCCAGCATGGCT
ATATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCTATGCCGATCGTACGACACATATCGTCATCGTACTGCCCTAAAC
CTGTCTAGTCTAAACACATCCATCGTACTGACTGCATCGTACGACTGCATCGTACTGACTGCACATATCGTCATACATACCCT
GTACTGACTGTCTAGTCTAAACACATCCCACATATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCCACTTTACCGGCT
ATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCCACACTGTCTAGTCTAAACACATCCATCGTACTGACTGCATCAAT
CGATCGTACGACACATATCGTCATCGTACTGCCCTACGGGACTGTCTAGTCTAAACACATCCATCGTACTGACTGCATCCGG


Most of genetic variation is due to *single nucleotide polymorphisms (SNPs)* --single base changes that are common in the general population

Human genome: *individual variations*

- Human genome is ~99 % similar between individuals
- 0.5-1% different

articles

A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms

The International SNP Map Working Group*

* *A full list of authors appears at the end of this paper.*

We describe a map of 1.42 million single nucleotide polymorphisms (SNPs) distributed throughout the human genome, providing an average density on available sequence of one SNP every 1.9 kilobases. These SNPs were primarily discovered by two projects: The SNP Consortium and the analysis of clone overlaps by the International Human Genome Sequencing Consortium. The map integrates all publicly available SNPs with described genes and other genomic features. We estimate that 60,000 SNPs fall within exon (coding and untranslated regions), and 85% of exons are within 5 kb of the nearest SNP. Nucleotide diversity varies greatly across the genome, in a manner broadly consistent with a standard population genetic model of human history. This high-density SNP map provides a public resource for defining haplotype variation across the genome, and should help to identify biomedically important genes for diagnosis and therapy.

The International HapMap Project

Phase I

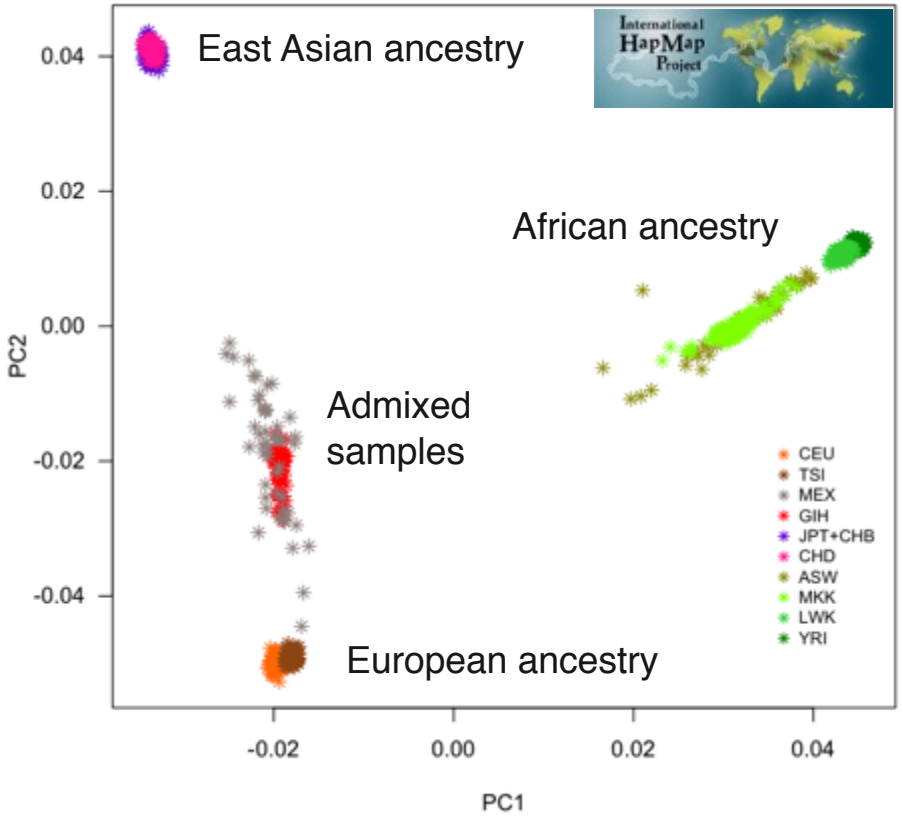
1.1 million SNPs
270 individuals from 4 populations

Phase II

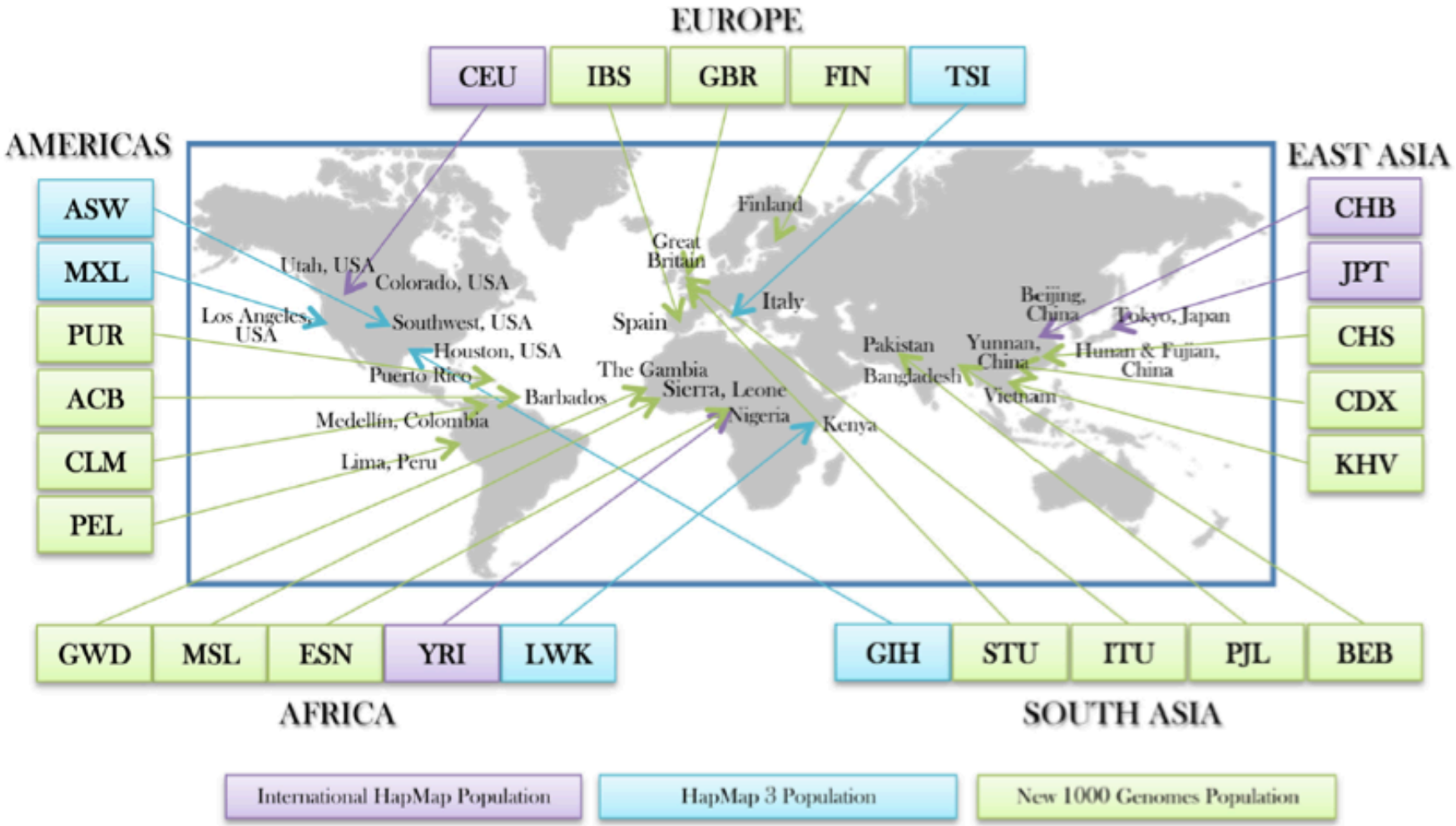
3.1 million SNPs
270 individuals from 4 populations

Phase III

1.6 million SNPs
1,184 individuals from 11 populations



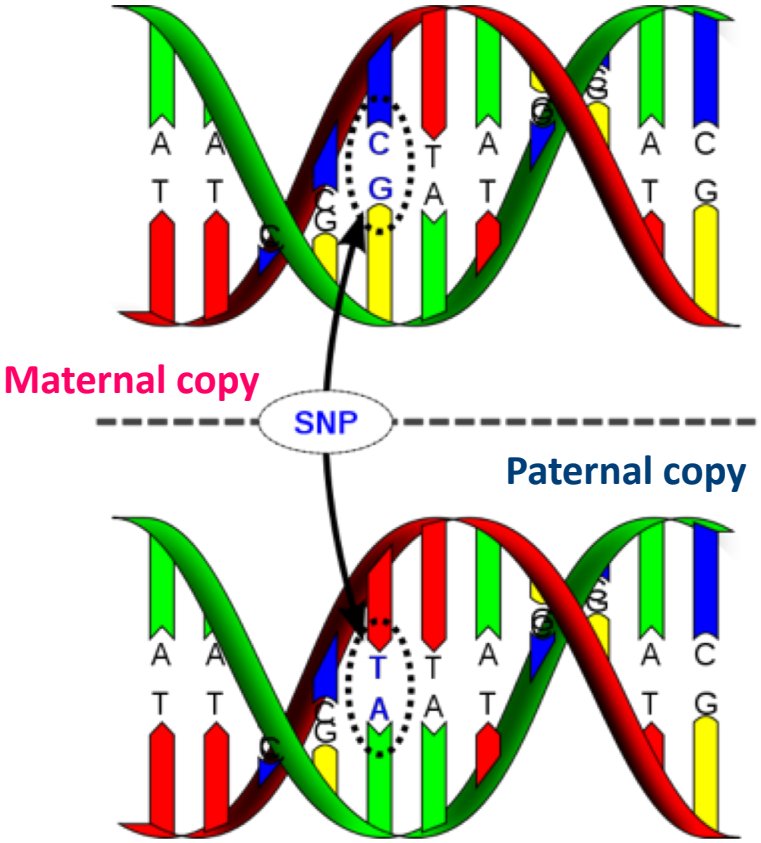
The 1000 Genomes Project

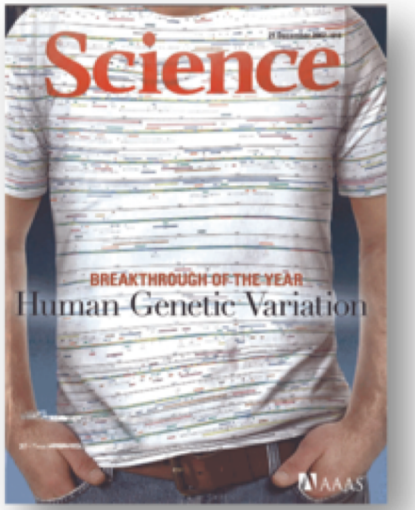


*S*ingle-*N*ucleotide *P*olymorphism

- "one base pair variation"
  - > 1% general population (common)
  - ≈10 million SNPs (≈0.25% genome)
  - Makes you and me unique
  - Most common type of genetic variation

www.hapmap.org

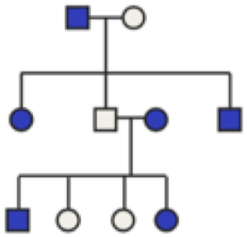


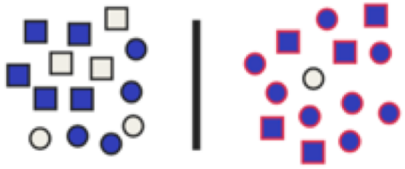


Linkage analysis

Candidate gene studies

GWAS





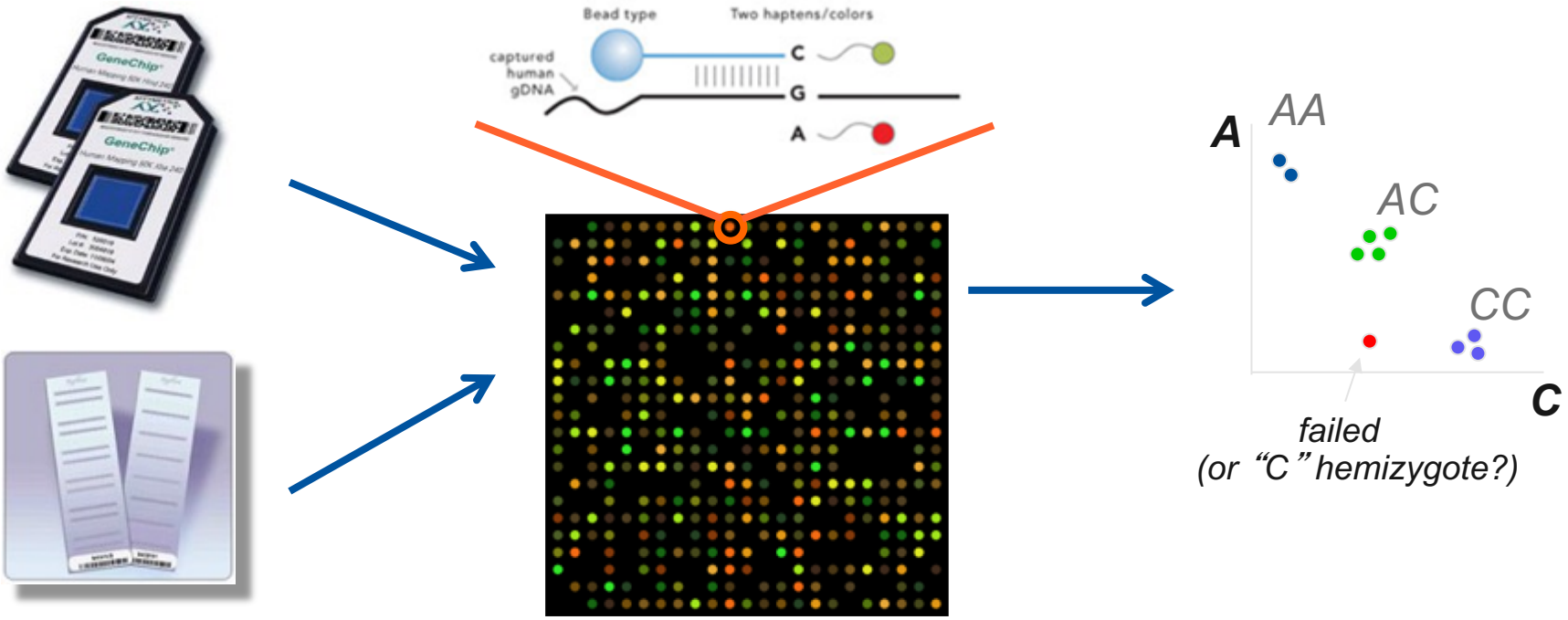
Common variant, common disease hypothesis

- Most common diseases happen later in life
- If common variants are not selected against, they may associate to late-onset (after reproduction) disease
- Common variants are easier to find and characterize

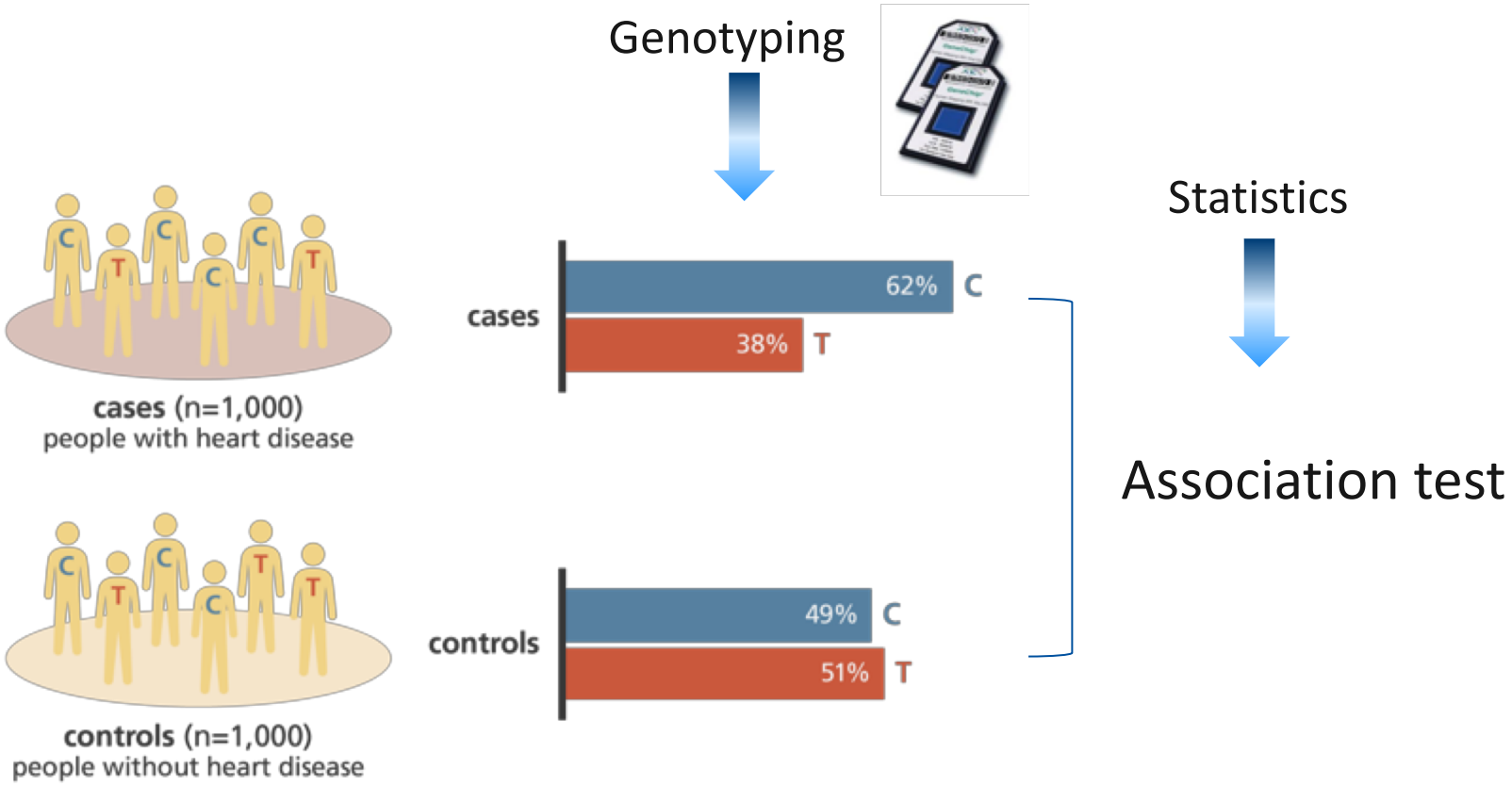
The beginnings of GWAS

Genotyping platforms

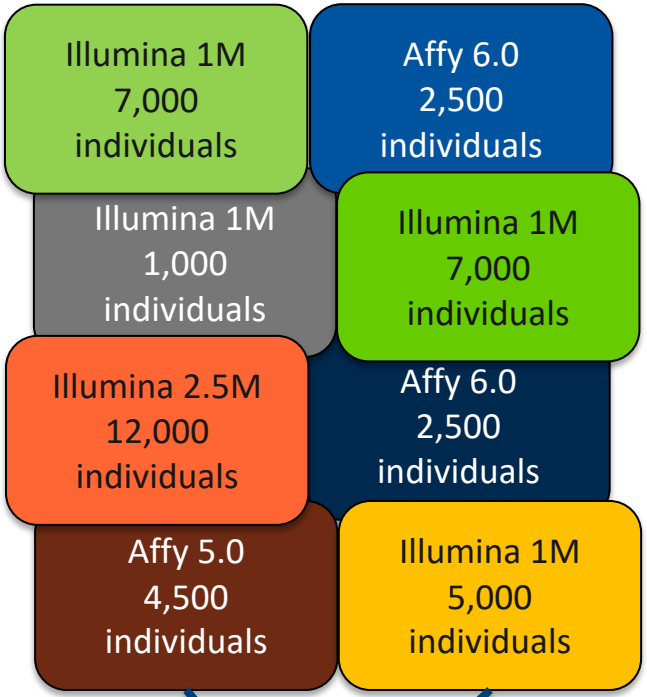
- Genome-wide SNP microarrays allow measurement of genotypes of 100,000's of SNPs in a single experiment
- Variety of microarrays (different SNP density, cost, etc) by Illumina and Affymetrix



GWAS (the big picture)

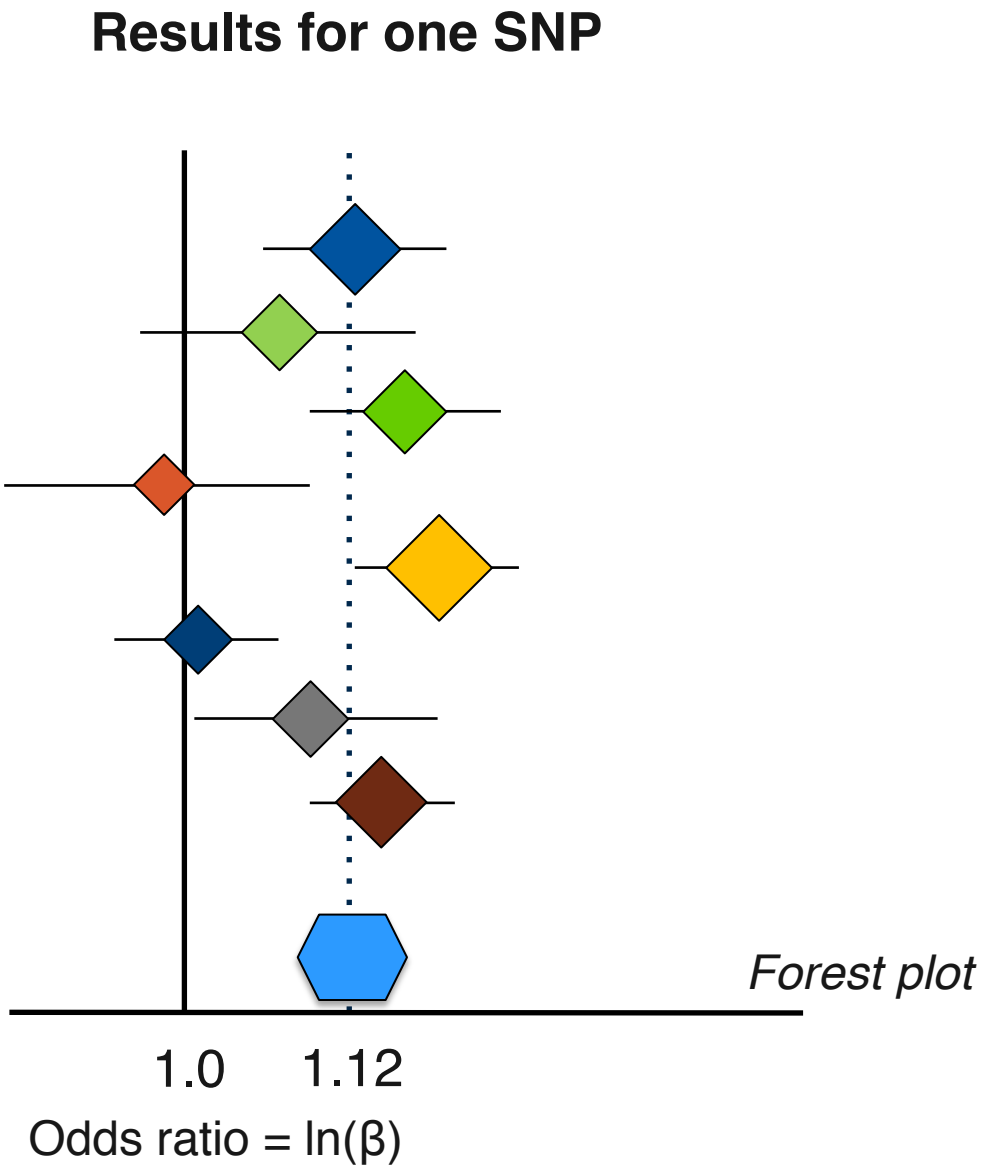


Combining GWAS datasets



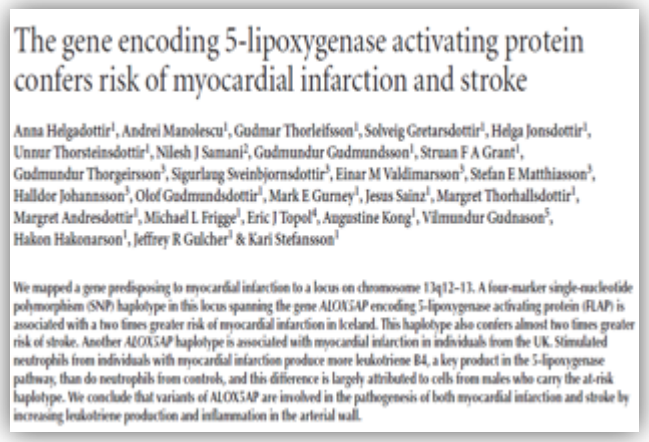
Imputation

Meta-analysis of GWAS



deCODE Genetics, Inc.

- >50% adult population of Iceland (>140,000) in biobank (blood)
- Pedigree information going back to the first settlements (≈1000 years ago)
- Extensive medical records & genotypic data
- Over 250 high-impact publications (Nature, Science, AJHG)
- 50 common diseases
  - Stroke (=CVA) association with *ALOX5AP*
  - MI association with *ALOX5AP*
  - Association of a variant on 9p21.1 with Abdominal aortic aneurysm (AAA), intracrainial aneurysm, stroke and MI

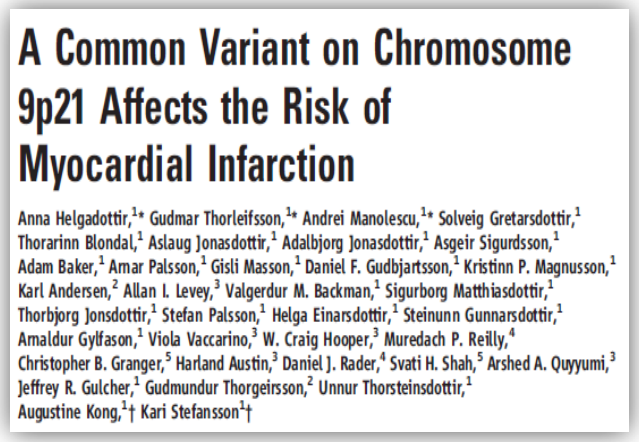


The gene encoding 5-lipoxygenase activating protein confers risk of myocardial infarction and stroke

Anna Helgadottir[1], Andrei Manolescu[1], Gudmar Thorleifsson[1], Solveig Gretarsdottir[1], Helga Jonsdottir[1], Unnur Thorsteinsdottir[1], Nilesh J Samani[2], Gudmundur Gudmundsson[1], Struan F A Grant[1], Gudmundur Thorgeirsson[3], Sigurlaug Sveinbjornsdottir[3], Einar M Valdimarsson[3], Stefan E Matthiasson[3], Halldor Johannsson[3], Olof Gudmundsdottir[3], Mark E Gurney[4], Jesus Sainz[1], Margret Thorhallsdottir[1], Margret Andresdottir[1], Michael L Frigge[3], Eric J Topol[4], Augustine Kong[1], Vilmundur Gudnason[5], Hakon Hakonarson[1], Jeffrey R Gulcher[1] & Kari Stefansson[1]

We mapped a gene predisposing to myocardial infarction to a locus on chromosome 13q12–13. A four-marker single-nucleotide polymorphism (SNP) haplotype in this locus spanning the gene *ALOX5AP* encoding 5-lipoxygenase activating protein (FLAP) is associated with a two times greater risk of myocardial infarction in Iceland. This haplotype also confers almost two times greater risk of stroke. Another *ALOX5AP* haplotype is associated with myocardial infarction in individuals from the UK. Stimulated neutrophils from individuals with myocardial infarction produce more leukotriene B4, a key product in the 5-lipoxygenase pathway, than do neutrophils from controls, and this difference is largely attributed to cells from males who carry the at-risk haplotype. We conclude that variants of *ALOX5AP* are involved in the pathogenesis of both myocardial infarction and stroke by increasing leukotriene production and inflammation in the arterial wall.

Helgadottir, A., *et al.* **Nature Genetics**; volume 36, 233; 2004



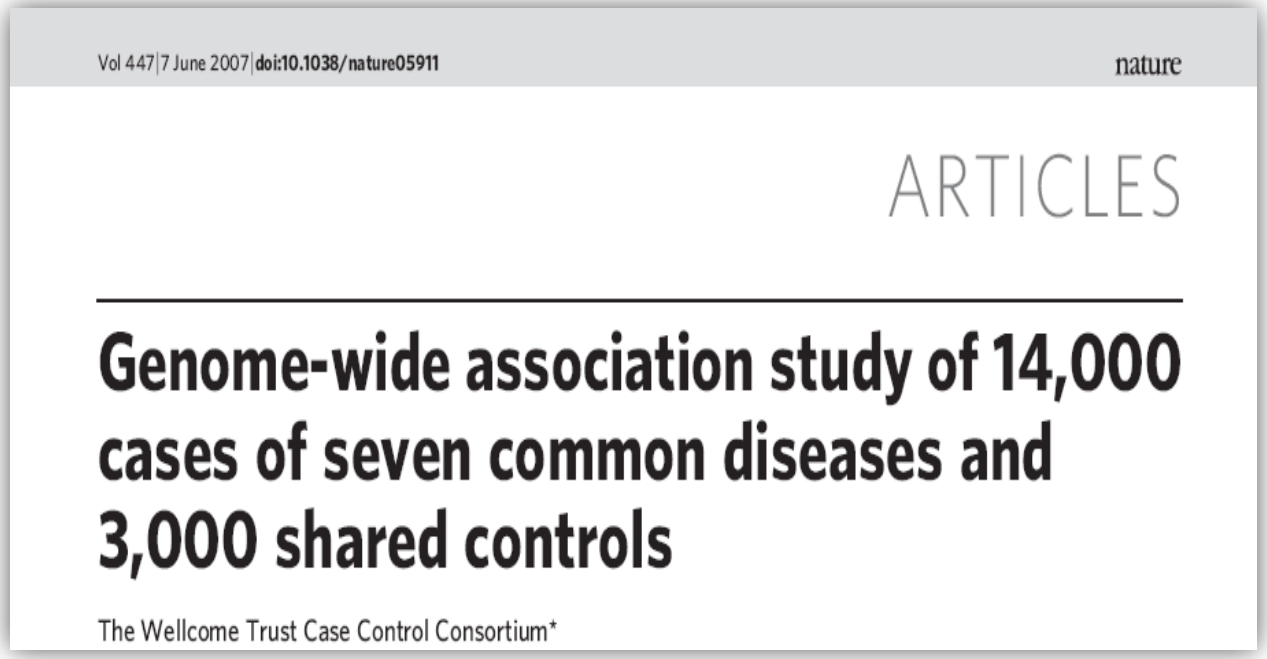
A Common Variant on Chromosome 9p21 Affects the Risk of Myocardial Infarction

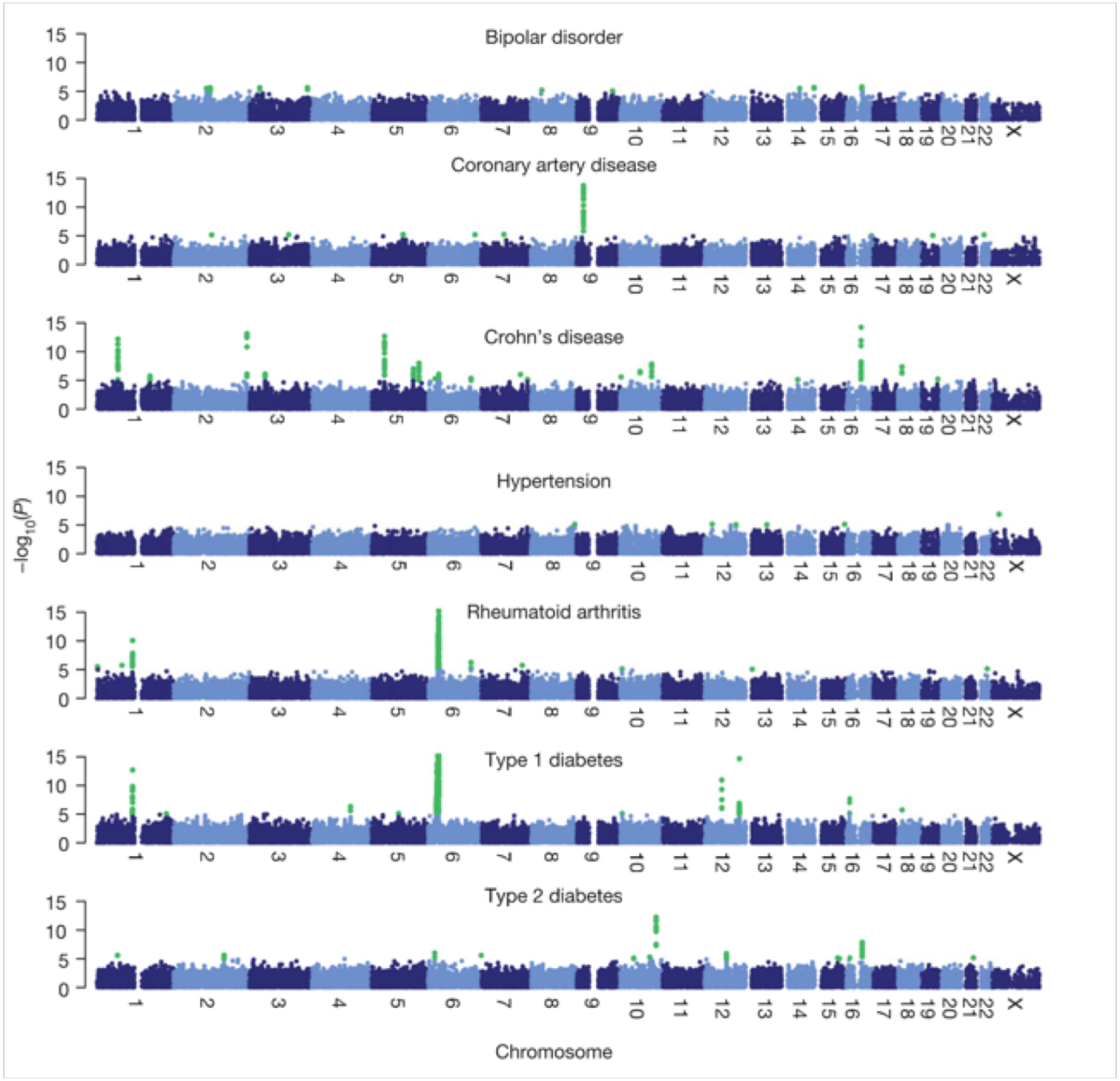
Anna Helgadottir,[1]* Gudmar Thorleifsson,[1]* Andrei Manolescu,[1]* Solveig Gretarsdottir,[1] Thorarinn Blondal,[1] Aslaug Jonasdottir,[1] Adalbjorg Jonasdottir,[1] Asgeir Sigurdsson,[1] Adam Baker,[1] Arnar Palsson,[1] Gisli Masson,[1] Daniel F. Gudbjartsson,[1] Kristinn P. Magnusson,[1] Karl Andersen,[2]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

Helgadottir, A., *et al.* **Science** volume 316, 1491; 2007

Wellcome Trust Case-Control Consortium

- 1,500 1958 Birth Cohort Controls (58BC)
- 1,500 UK Blood Services Controls (UKBS)
- 14,000 cases of seven common diseases
  - Bipolar disorder
  - **Coronary artery disease**
  - Crohn's disease
  - **Hypertension**
  - Rheumatoid arthritis
  - **Type 1 diabetes**
  - **Type 2 diabetes**





WTCCC, Nature 2007

One famous example

- deCODE Genetics was the first to discover a SNP associated with myocardial infarction (MI) in 2007
- WTCCC, McPherson, and Samani were able to replicate the same finding in the same year, and many have reconfirmed it in different populations

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

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

The NEW ENGLAND JOURNAL of MEDICINE

ESTABLISHED IN 1812 AUGUST 2, 2007 VOL. 357 NO. 5

Genomewide Association Analysis of Coronary Artery Disease

Vol 447|7 June 2007|doi:10.1038/nature05911 nature

A Common Allele on Chromosome 9 Associated with Coronary Heart Disease

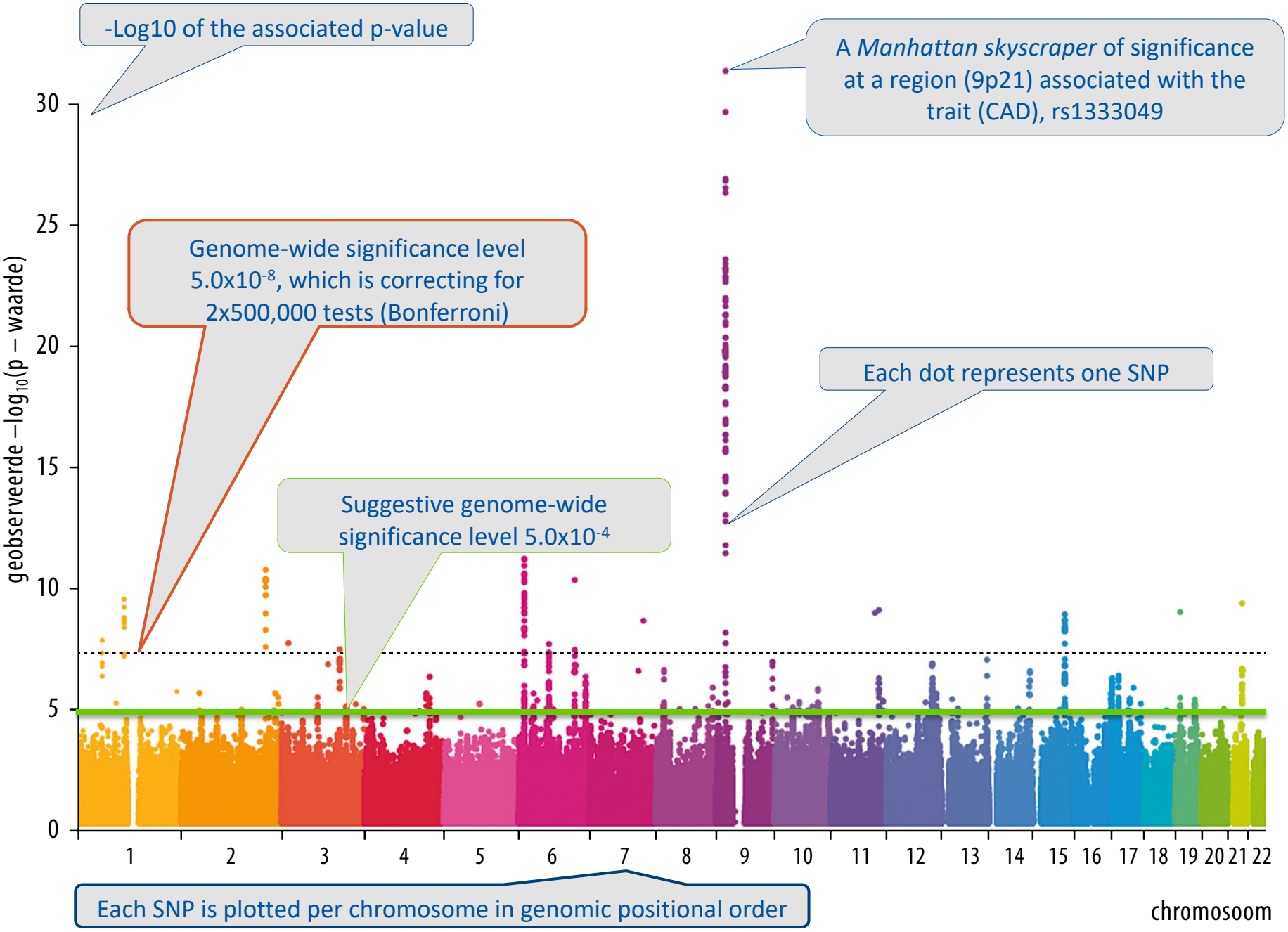
Ruth McPherson,[1]*† Alexander Pertsemlidis,[2]* Nihan Kavaslar,[1] Alexandre Stewart,[1] Robert Roberts,[1] David R. Cox,[3] David A. Hinds,[3] Len A. Pennacchio,[4,5] Anne Tybjaerg-Hansen,[6] Aaron R. Folsom,[7] Eric Boerwinkle,[8] Helen H. Hobbs,[2,9] Jonathan C. Cohen[2,10]†

ARTICLES

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

The Wellcome Trust Case Control Consortium*

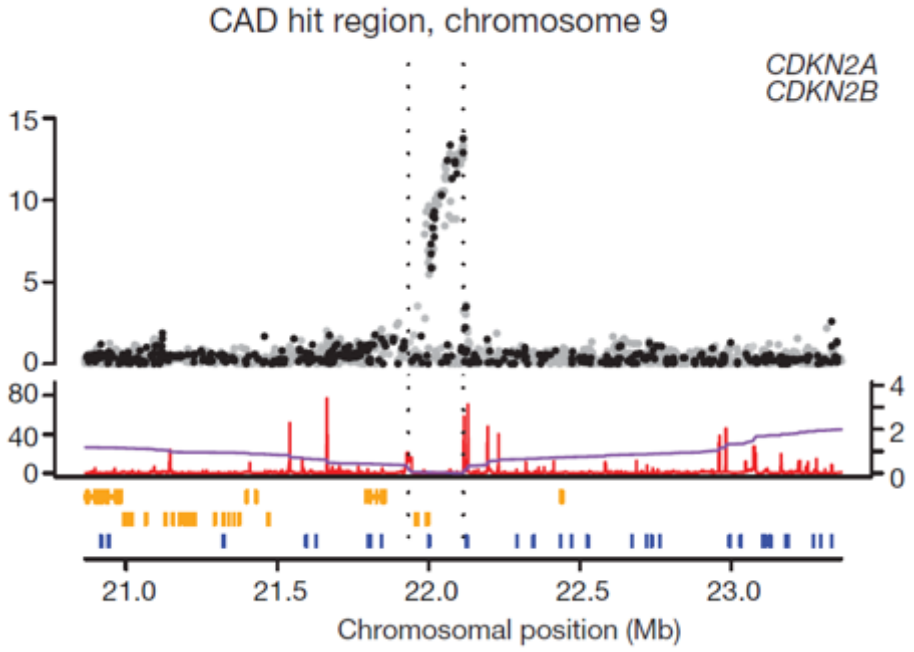
Helgadottir, A., *et al.* **Science**; 316(5830):1491-1493, 2007
McPherson, R., *et al.* **Science**; 316(5830):1488-1491, 2007
Wellcome Trust Case Control Consortium. **Nature**; 447(7145):661-678, 2007
Samani, N.J., *et al.* **N Engl J Med**; 357(5):443-453, 2007



Asselbergs F.W., van der Laan S.W., de Bakker P.I.W. **NTvG**; 159:A5821; 2014

9p21 and cardiovascular disease

- The SNPs associated with CAD on 9p21.1 are rs1333049, rs10757274, rs2383207, rs2891168, and rs10757278
- They are found in an *intergenic region*
- Genes nearby: *CDKN2A*, *CDKN2B*
  - also associated with *type 2 diabetes mellitus*
  - regulating cell proliferation, cell aging and the associated degeneration, and programmed cell death of many cell types



Wellcome Trust Case Control Consortium. **Nature**; 447(7145):661-678, 2007

A closer look at the results…

Table 3 | Regions of the genome showing the strongest association signals

Collection	Chromosome	Region (Mb)	SNP	Trend *P* value	Genotypic *P* value	$\log_{10}$(BF), additive	$\log_{10}$(BF), general	Risk allele	Minor allele	Heterozygote odds ratio	Homozygote odds ratio	Control MAF	Case MAF
				Standard analysis									
CAD	9p21	21.93–22.12	rs1333049	1.79×10^{-14}	1.16×10^{-13}	11.66	11.19	C	C	1.47 (1.27–1.70)	1.9 (1.61–2.24)	0.474	0.554

- **CAD:** coronary artery disease
- **9p21:** chromosome 9, short arm (p)
- **Region:** 21.93-22.12 megabase pairs
- **rs1333049:** official dbSNP ID

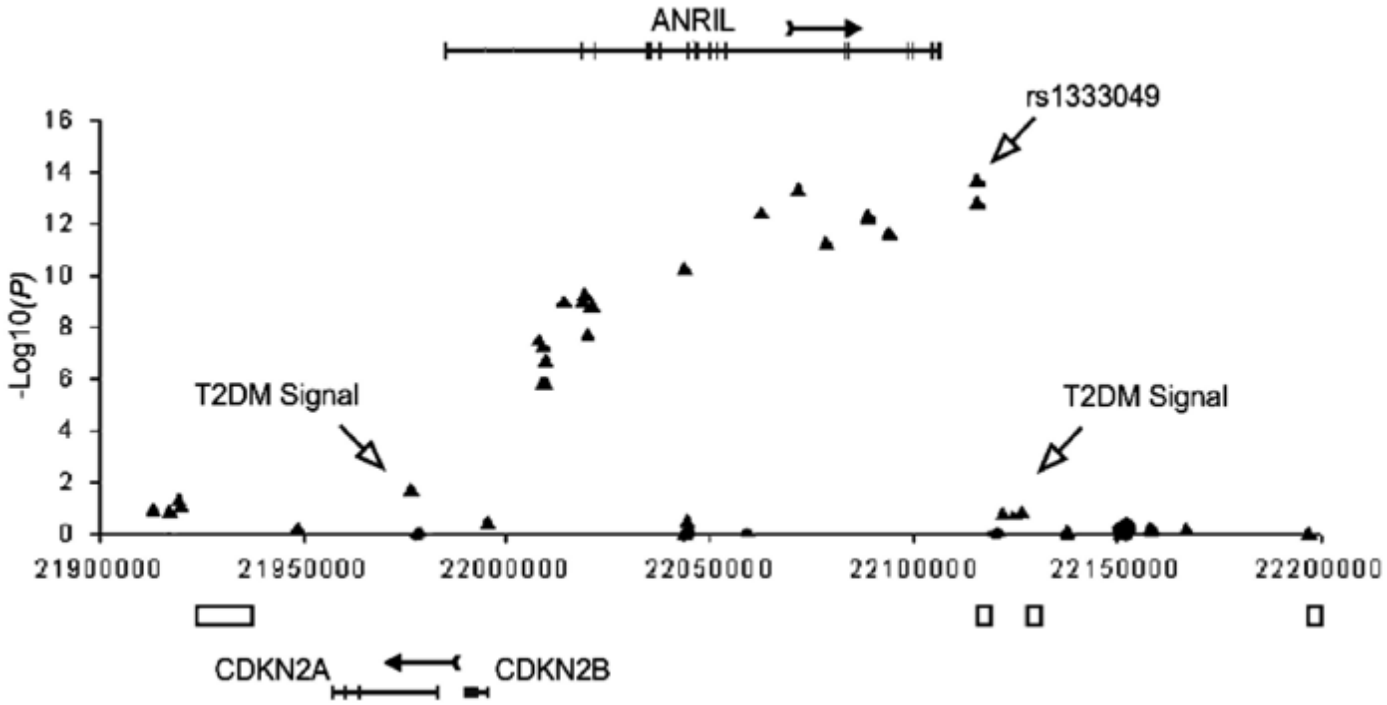
P-value of association test: AA vs. AB vs. BB

Risk allele: minor allele
Odds ratio: the odds of exposure between cases and controls

Minor allele frequency: the frequency of the risk (minor) allele in the population

9p21 points to a *RNA gene*

- Resequencing unveiled a RNA gene, *ANRIL*
- Current efforts are aimed to elucidate the role of *ANRIL* in (A)MI
- Might be involved in *early-onset MI* (before age of 50 years)



Samani, NJ., *et al.* **Circ Cardiovasc Genet**; 1:81-84, 2008

CARDIoGRAMplusC4D Study

- Coronary Artery Disease Genome–Wide Replication And Meta–Analysis Study: CARDIoGRAM
- > 63,000 cases and > 130,000 controls
  - Myocardial infarction (MI), coronary artery disease (CAD) or both
  - CAD: MI, CABG, PTCA, AP
  - Age limit: 45-66
- Sample size greatly influences power and effect size to discover new variants
- CARDIoGRAMplusC4D sought to solves this issue
- 55 susceptibility loci for CAD were discovered

Large-scale association analysis identifies new risk loci for coronary artery disease

The CARDIoGRAMplusC4D Consortium[1]

Coronary artery disease (CAD) is the commonest cause of death. Here, we report an association analysis in 63,746 CAD cases and 130,681 controls identifying 15 loci reaching genome-wide significance, taking the number of susceptibility loci for CAD to 46, and a further 104 independent variants ($r^2 < 0.2$) strongly associated with CAD at a 5% false discovery rate (FDR). Together, these variants explain approximately 10.6% of CAD heritability. Of the 46 genome-wide significant lead SNPs, 12 show a significant association with a lipid trait, and 5 show a significant association with blood pressure, but none is significantly associated with diabetes. Network analysis with 233 candidate genes (loci at 10% FDR) generated 5 interaction networks comprising 85% of these putative genes involved in CAD. The four most significant pathways mapping to these networks are linked to lipid metabolism and inflammation, underscoring the causal role of these activities in the genetic etiology of CAD. Our study provides insights into the genetic basis of CAD and identifies key biological pathways.

Coronary artery disease and its main complication, myocardial infarction, is the leading cause of death worldwide. Although, epidemiological studies have identified many risk factors for CAD, including plasma lipid concentrations, blood pressure, smoking, diabetes and markers of inflammation, a causal role has been proven only for some (for example, low-density lipoprotein (LDL) cholesterol and blood pressure), primarily through randomized clinical trials of drug therapy directed at the risk factor[1]. Twin and family studies have documented that a significant proportion (40–50%) of susceptibility to CAD is heritable (for a review, see ref. 2). Because genotypes are not confounded by environmental exposures, genetic analysis has the potential to define which risk factors are indeed causal and to identify pathways and therapeutic targets[3,4]. To date, genome-wide association studies (GWAS) have collectively reported a total of 31 loci, associated with CAD risk at genome-wide significance ($P < 5 \times 10^{-8}$)[5–13]. However, variants at these loci explain less than 10% of the heritability of CAD. One likely reason for this is that, given the polygenic nature of complex traits and the relatively small observed effect sizes of the loci identified, many genuinely associated variants do not reach the stringent *P*-value threshold for genome-wide significance. Indeed, there is increasing evidence that the genetic architecture of common traits involves a large number of causative alleles with very small effects[14]. Addressing this will require the discovery of additional loci while leveraging large-scale genomic data to identify the molecular pathways underlying the pathogenesis of CAD. Such discovery is facilitated by building molecular networks, on the basis of DNA, RNA and protein interactions, which have nodes of known biological function that also show evidence of association with risk variants for CAD and related metabolic traits.

In the largest GWAS meta-analysis of CAD undertaken to date by the Coronary ARtery DIsease Genome-wide Replication and Meta-analysis (CARDIoGRAM) Consortium[5], which involved 22,233 cases and 64,762 controls, in addition to loci reported at genome-wide significance, a linkage disequilibrium (LD)-pruned set of 6,222 variants achieved a nominal association *P* value of less than 0.01. Here, we test these 6,222 SNPs in a meta-analysis of over 190,000 individuals, with the primary aim of identifying additional susceptibility loci for CAD. To this end, we used the Metabochip array[15], which is a custom iSELECT chip (Illumina) containing 196,725 SNPs, designed to (i) follow-up putative associations in several cardiometabolic traits, including CAD, and (ii) fine map confirmed loci for these traits. All SNPs on the array with data in the CARDIoGRAM study were considered for analysis (79,138 SNPs, of which 6,222 were the replication SNPs and 20,876 were fine-mapping SNPs in the 22 CAD susceptibility loci identified at the time at which the array was designed; the remaining SNPs were submitted by the other consortia contributing to the Metabochip array[15]). In addition, we assess whether the genome-wide significant CAD risk alleles act through traditional risk factors by considering the available large GWAS for these traits[16–20]. Finally, we identify a broader set of SNPs passing a conservative FDR threshold for association with CAD and use this set to undertake network analysis to find key biological pathways underlying the pathogenesis of CAD.

RESULTS

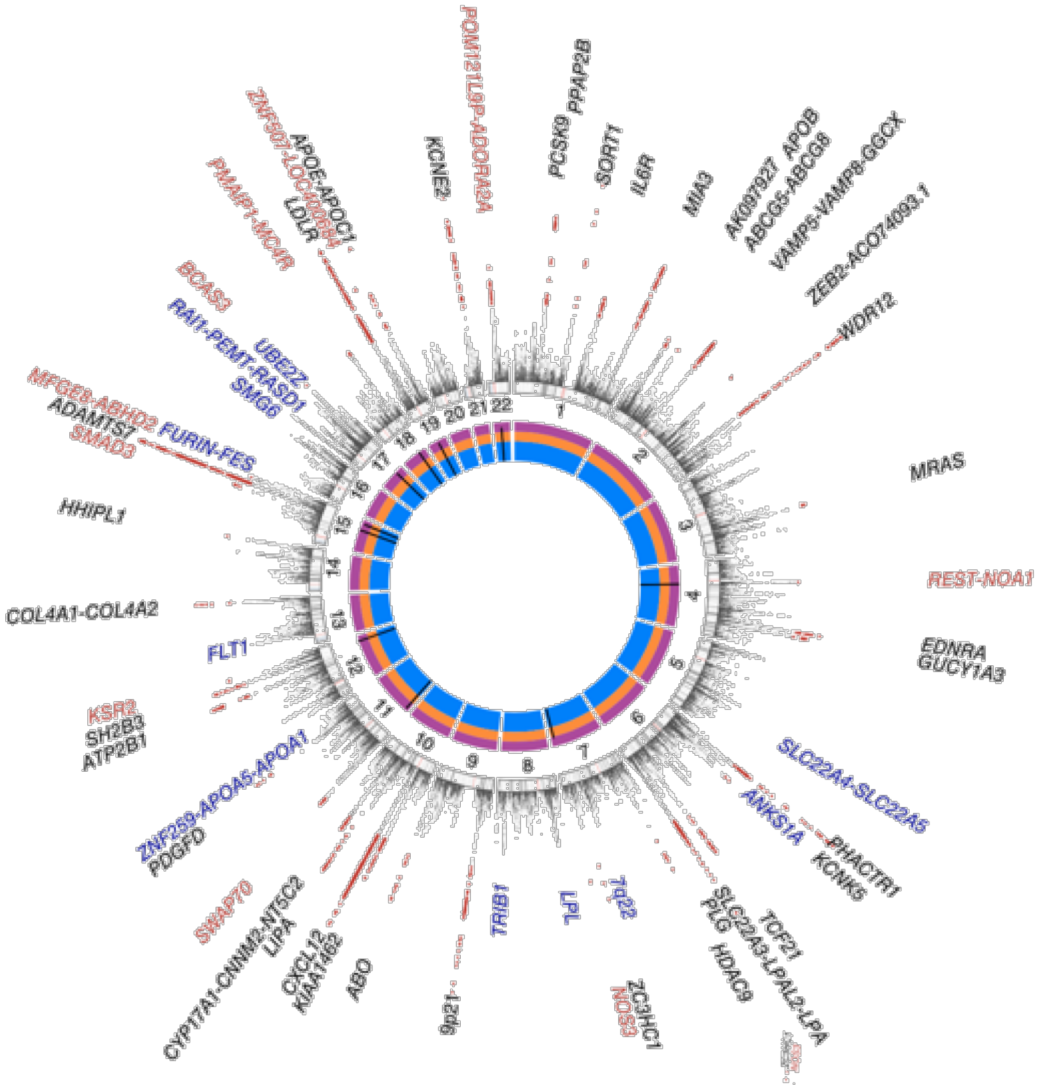
Study design

We expanded the CARDIoGRAM discovery data set (22,233 cases and 64,762 controls[5], stage 1) with 34 additional CAD sample collections (stage 2) of European or south Asian descent comprising 41,513 cases and 65,919 controls (study descriptions and sample characteristics are given in **Supplementary Tables 1a** and **2a**, respectively) and undertook a 2-stage meta-analysis to test SNPs on the Metabochip array

[1]A full list of authors and affiliations appears at the end of the paper.

Received 24 April; accepted 2 November; published online 2 December 2012; doi:10.1038/ng.2480

And 8 years later (>15 times more samples)

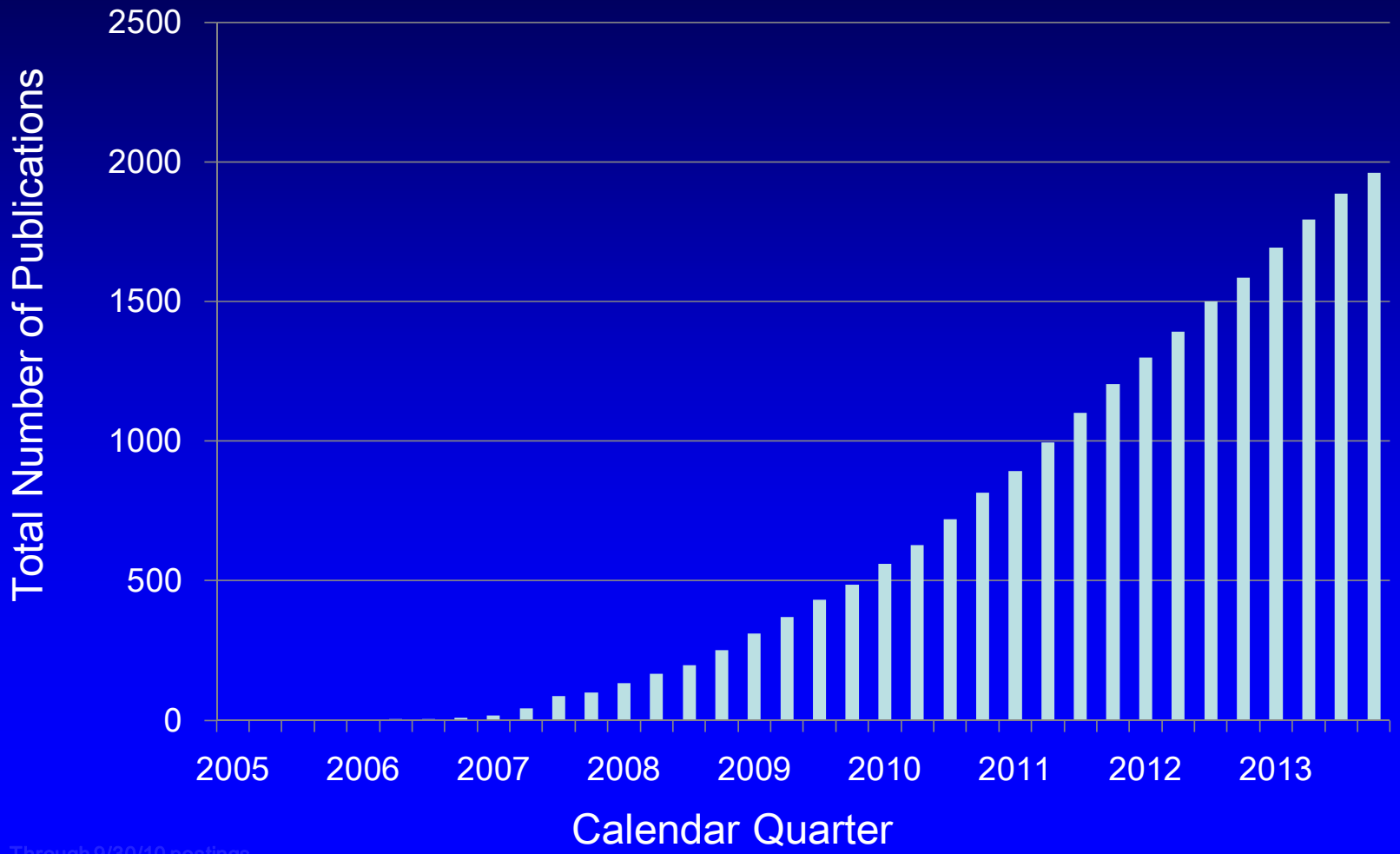


9p21 plus an additional 47 loci (!)

CARDIoGRAMplusC4D Consortium, Nature Genetics 2015

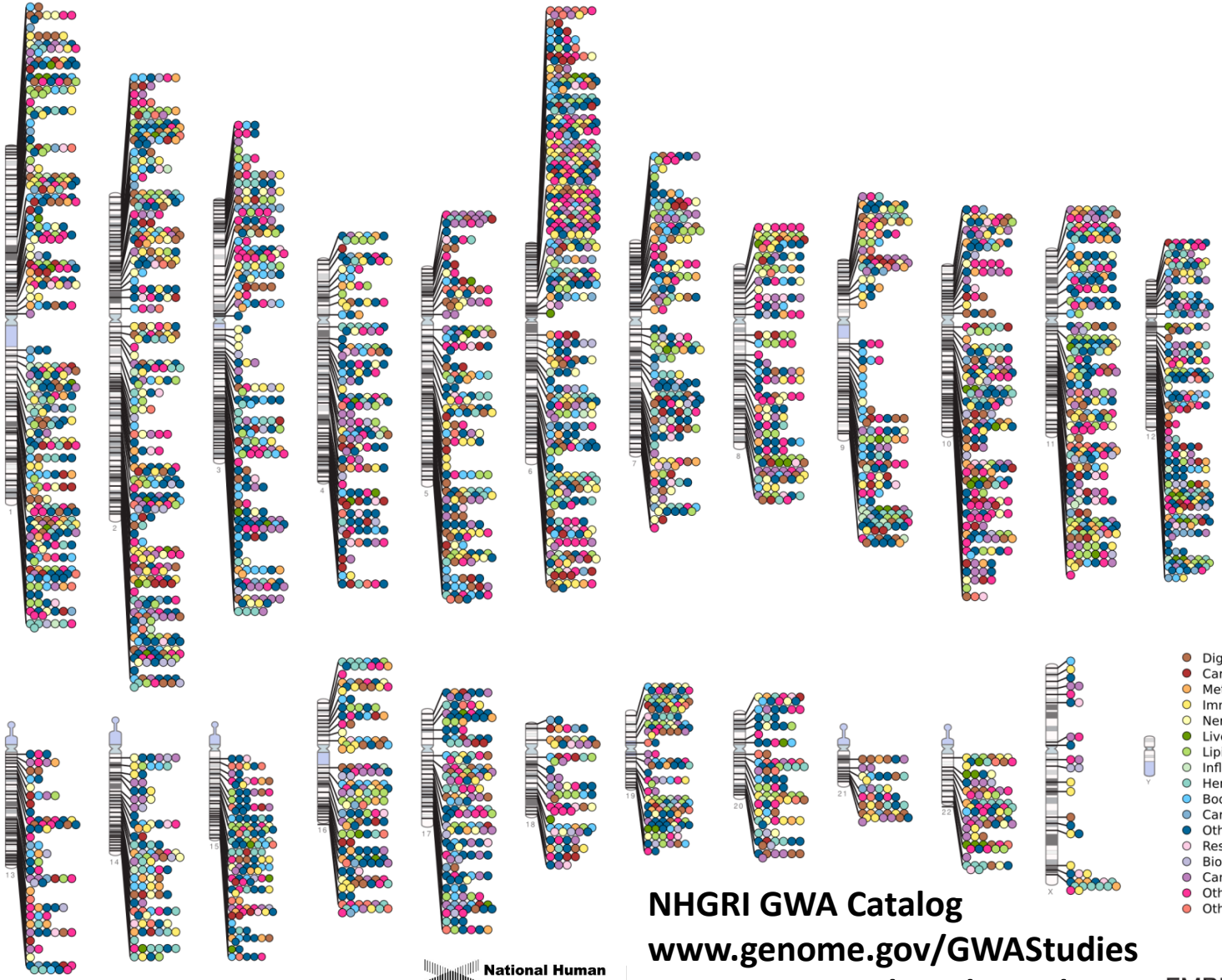
Published GWA Reports, 2005 – 2013

1960



Through 9/30/10 postings

Published Genome-Wide Associations through 12/2013
Published GWA at p≤5X10^{-8} for 17 trait categories



NHGRI GWA Catalog
www.genome.gov/GWAStudies
www.ebi.ac.uk/fgpt/gwas/

National Human Genome Research Institute

EMBL-EBI

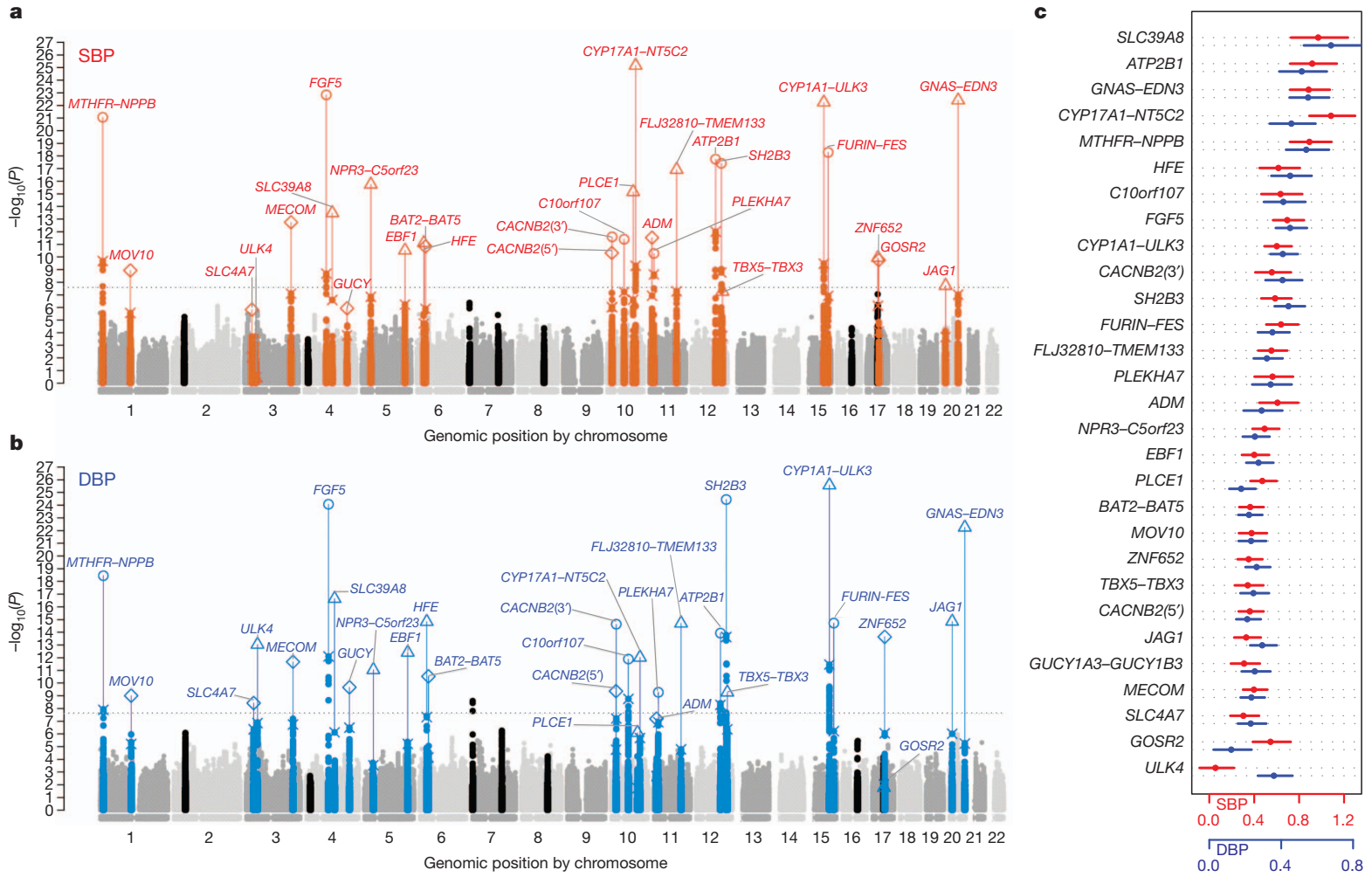


Figure 1 | Genome-wide $-\log_{10}$ *P*-value plots and effects for significant loci. a, **b**, Genome-wide $-\log_{10}$ *P*-value plots are shown for SBP (**a**) and DBP (**b**). SNPs within loci reaching genome-wide significance are labelled in red for SBP and blue for DBP (±2.5 Mb of lowest *P* value) and lowest *P* values in the initial genome-wide analysis as well as the results of analysis including validation data are labelled separately. The lowest *P* values in the initial GWAS are denoted with a X. The range of different sample sizes in the final meta-analysis including the validation data are indicated as: circle (96,000–140,000), triangle (>140,000–180,000) and diamond (>180,000–220,000). SNPs near unconfirmed loci are in black. The horizontal dotted line is $P = 2.5 \times 10^{-8}$. *GUCY* denotes *GUCY1A3–GUCY1B3*. **c**, Effect size estimates and 95% confidence bars per blood-pressure-increasing allele of the 29 significant variants for SBP (red) and DBP (blue). Effect sizes are expressed in mm Hg per allele.

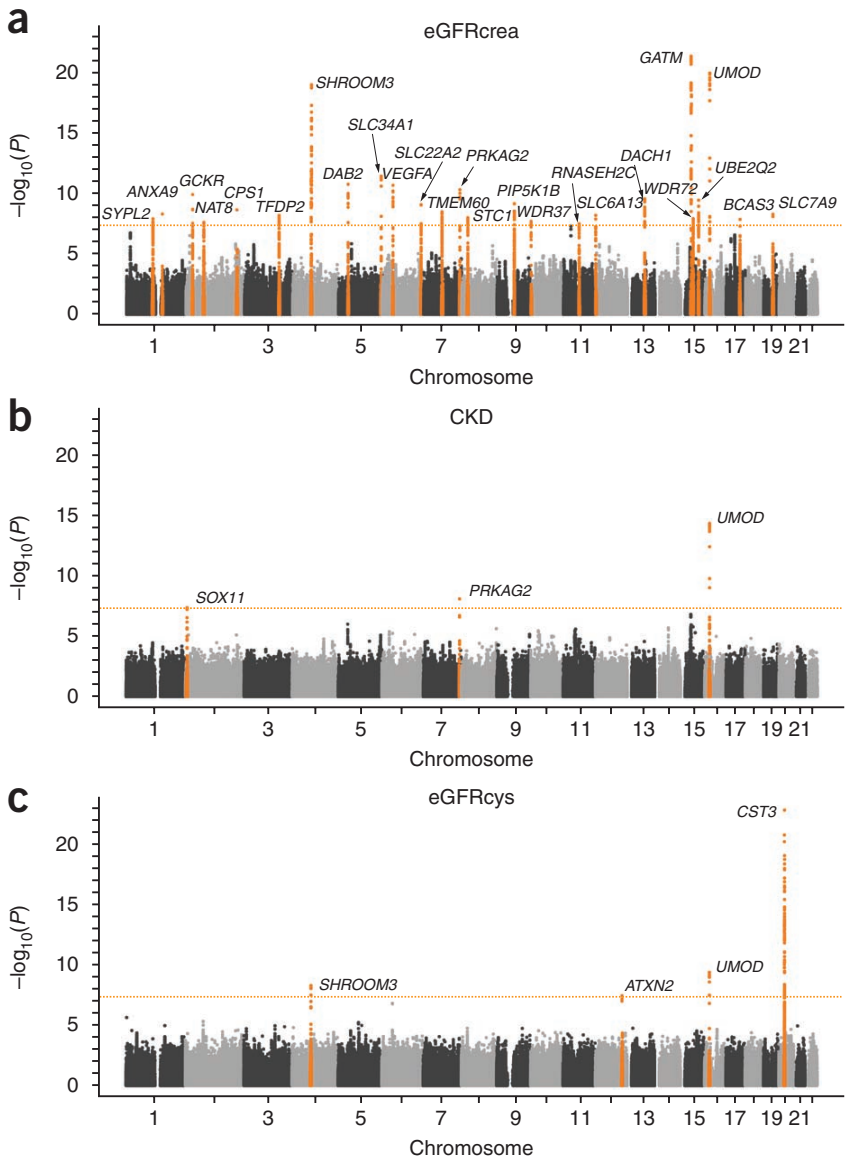


Figure 1 Genome-wide $-\log_{10} P$ value plot from stage 1. (**a**–**c**) Plots show discovery analysis of eGFRcrea (**a**), CKD (**b**) and eGFRcys (**c**). The dotted line indicates the genome-wide significance threshold at $P = 5 \times 10^{-8}$.

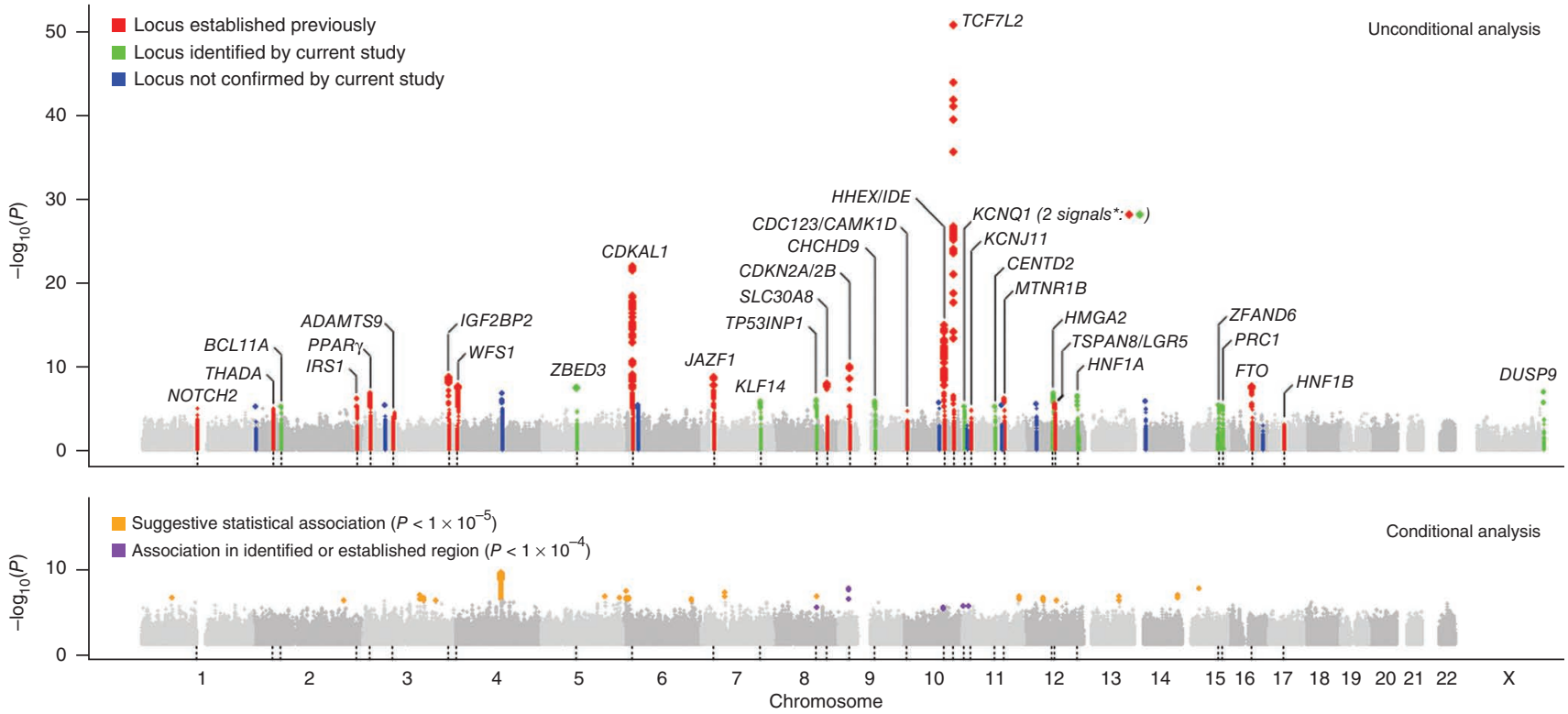
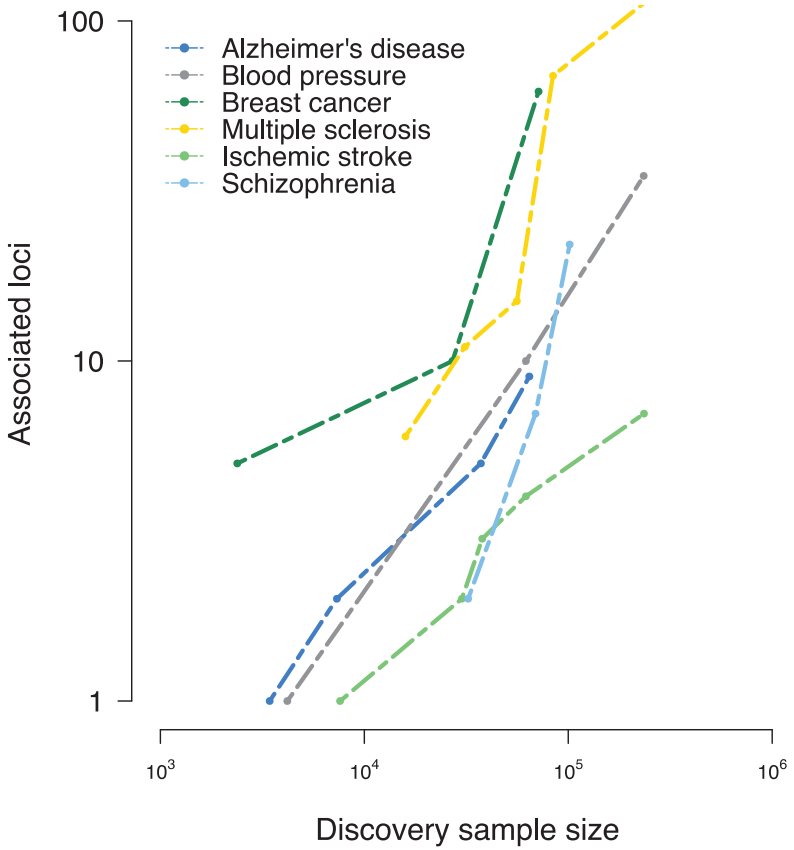
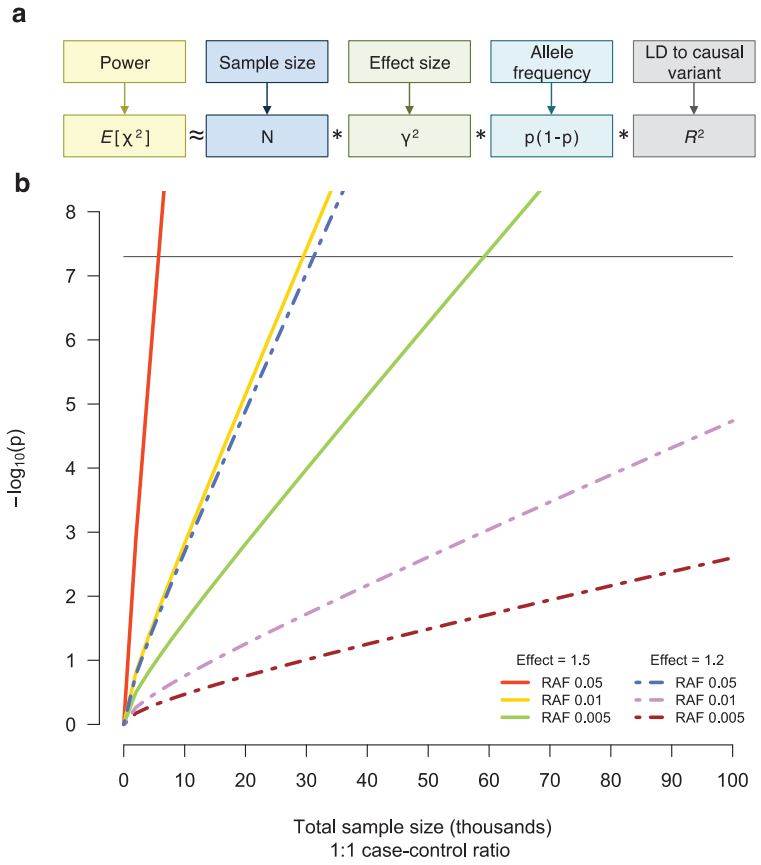


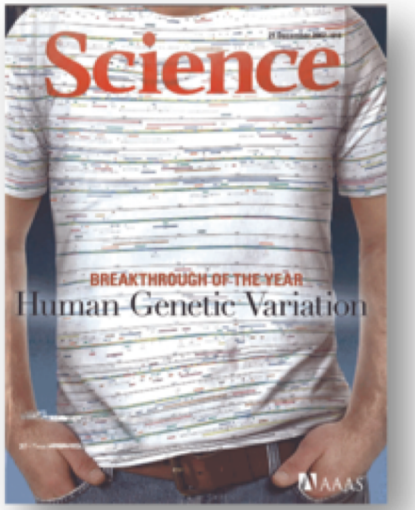
Figure 1 Genome-wide Manhattan plots for the DIAGRAM+ stage 1 meta-analysis. Top panel summarizes the results of the unconditional meta-analysis. Previously established loci are denoted in red and loci identified by the current study are denoted in green. The ten signals in blue are those taken forward but not confirmed in stage 2 analyses. The genes used to name signals have been chosen on the basis of proximity to the index SNP and should not be presumed to indicate causality. The lower panel summarizes the results of equivalent meta-analysis after conditioning on 30 previously established and newly identified autosomal T2D-associated SNPs (denoted by the dotted lines below these loci in the upper panel). Newly discovered conditional signals (outside established loci) are denoted with an orange dot if they show suggestive levels of significance ($P < 10^{-5}$), whereas secondary signals close to already confirmed T2D loci are shown in purple ($P < 10^{-4}$).

Power, Effect size, Sample size…





Pulit, S., *et al.* **Genes & Genomes**; 2014

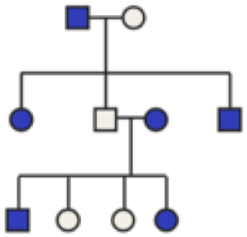


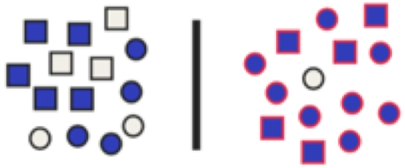
Linkage analysis

Candidate gene studies

GWAS

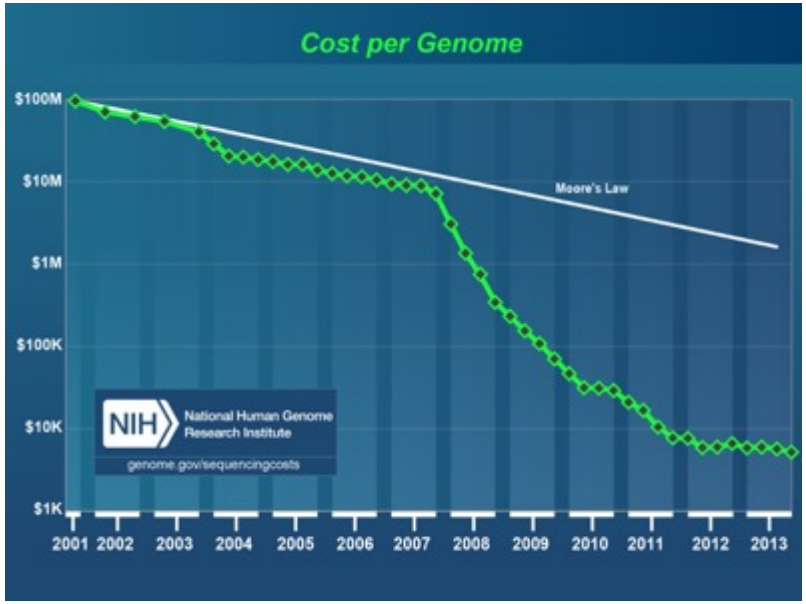
Sequencing



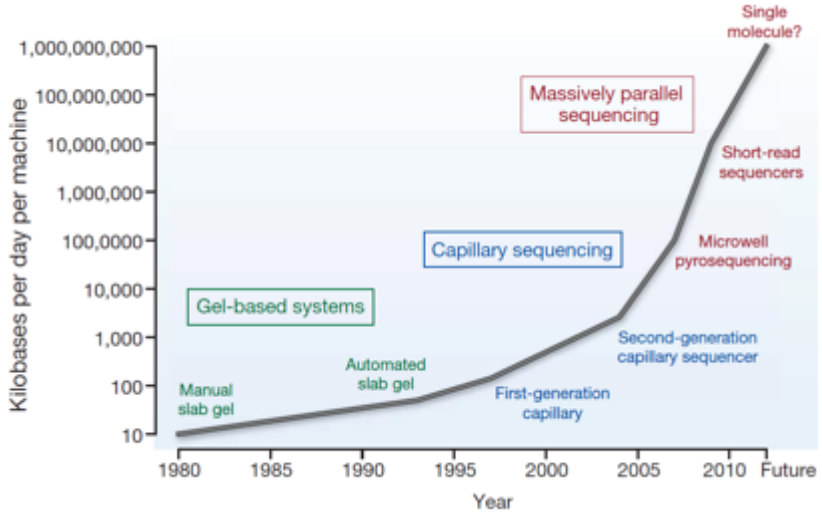


Next-generation sequencing

Milestone: $1000 dollar genome (2014, Illumina HiSeq X Ten Sequencer)



But how much money needs to be spend on annotation and (even more important) interpretation of the results?



Stratton et al. Nature 2009

Summary: what's been (being) done?

- Family-based linkage studies
  - Rare, Mendelian traits
- Candidate gene association studies
  - Many claims, few robust findings
  - Terrible track record in terms of reproducibility
- Genome-wide association studies (GWAS)
  - Complex traits and common diseases
- Whole-exome sequencing studies
  - Rare, Mendelian diseases (unsolved cases)
  - Complex traits and common diseases
- Whole-genome sequencing studies