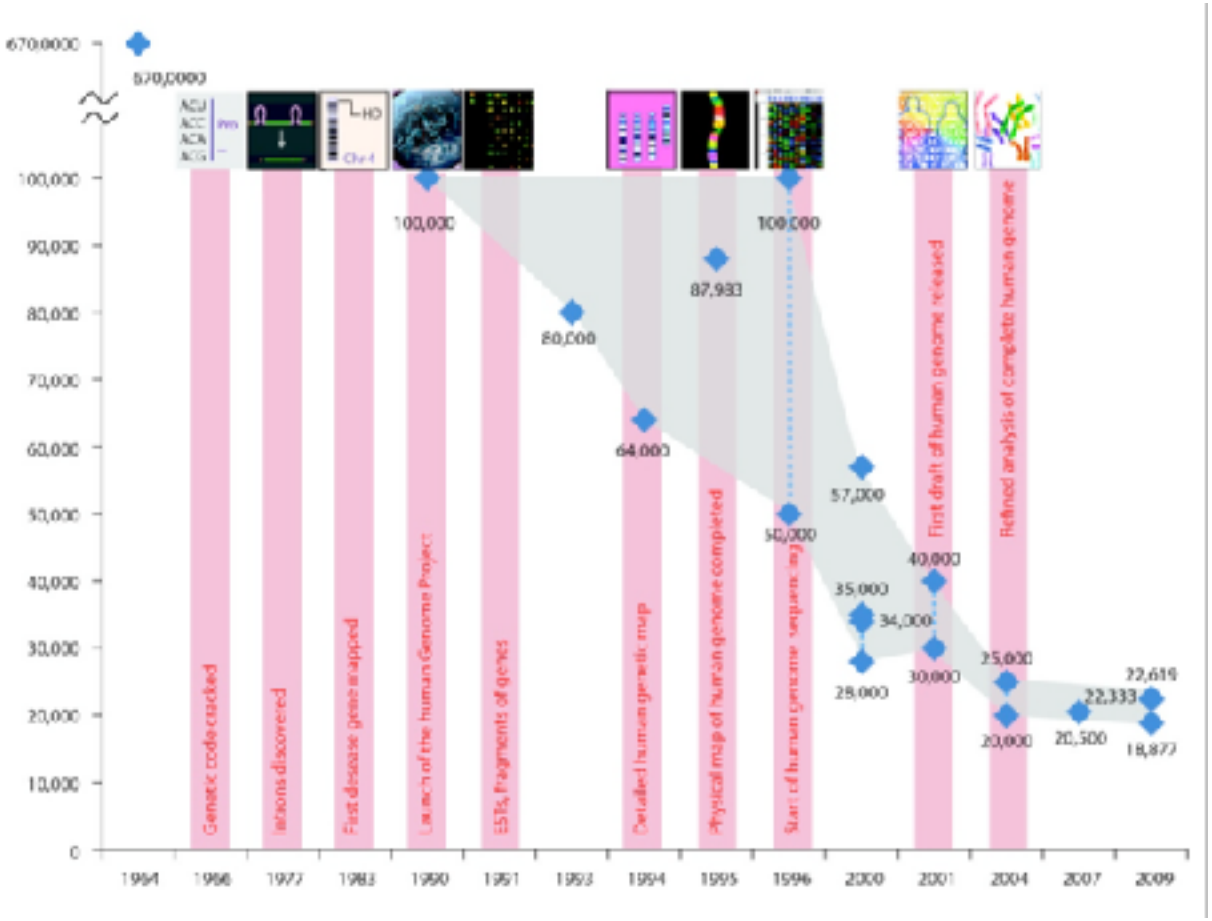
The Human Genome Project “paid forward” and paved the way for modern day genomics

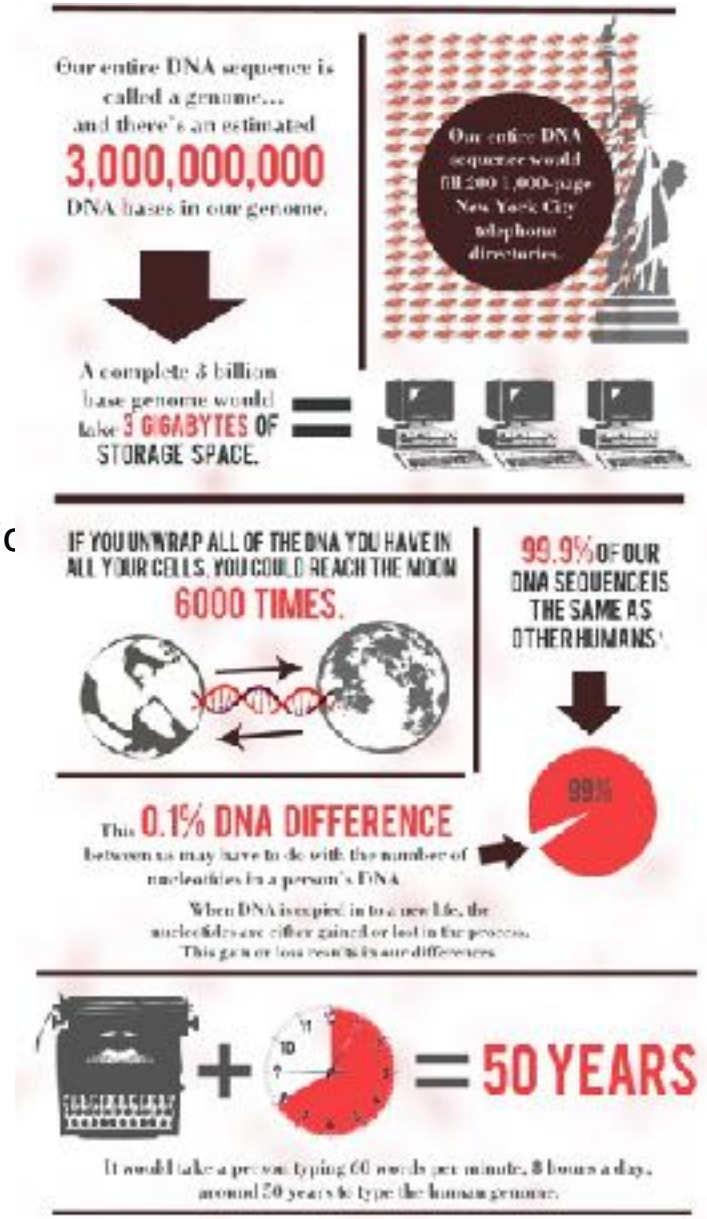
(Finally) a complete map



Pertea and Salzberg, Genome Biology 2010

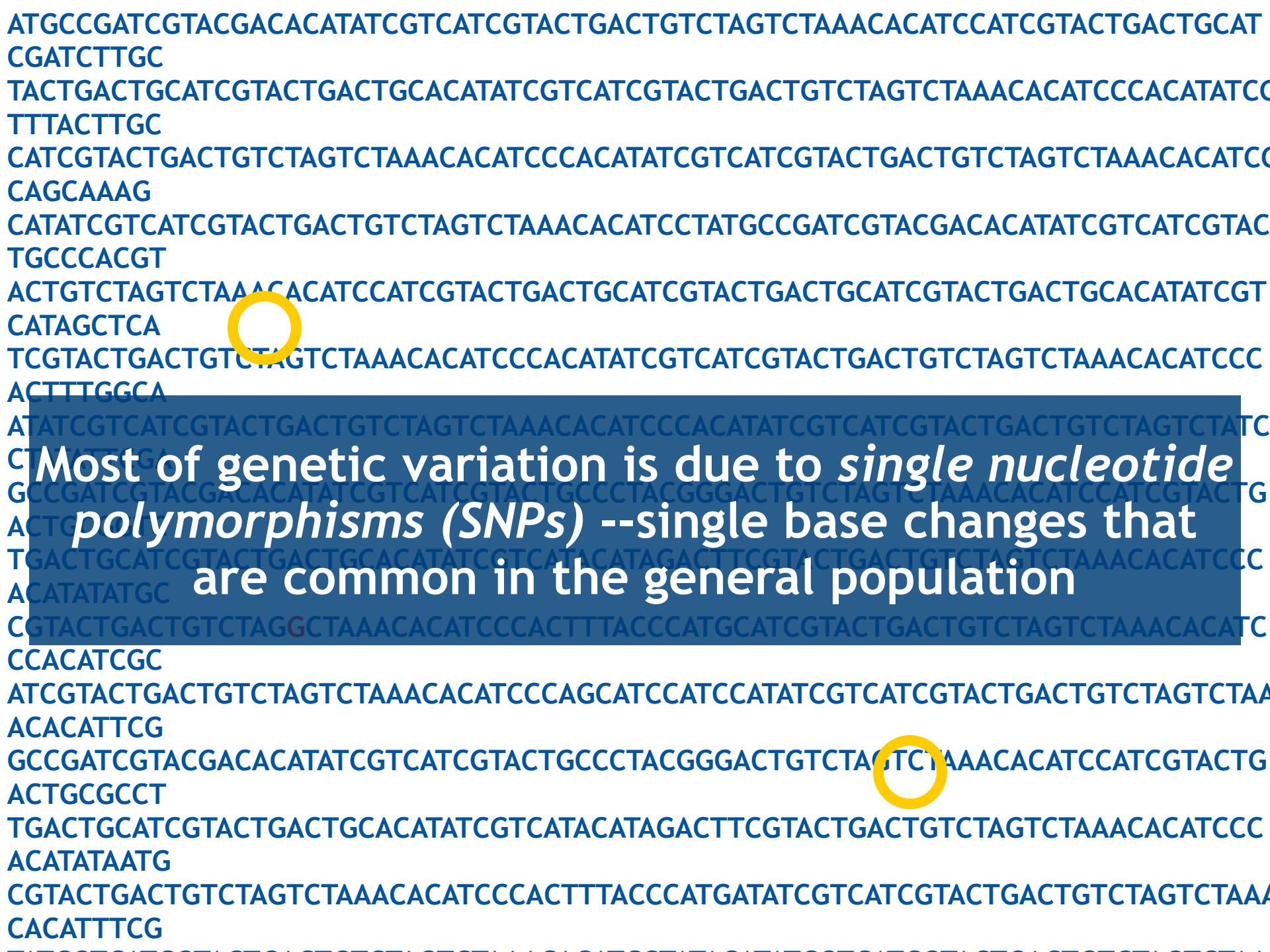
Human Genome: *some statistics*

- 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 regic
  - Repetitive elements and other variations
  - Transposable elements
- (So there's no such thing as "junk DNA"…)



ATGCCGATCGTACGACACATATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCATCGTACTGACTGCAT
CGATCTTGC
TACTGACTGCATCGTACTGACTGCACATATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCCACATATCC
TTTACTTGC
CATCGTACTGACTGTCTAGTCTAAACACATCCCACATATCGTCATCGTACTGACTGTCTAGTCTAAACACATCC
CAGCAAAG
CATATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCTATGCCGATCGTACGACACATATCGTCATCGTAC
TGCCCACGT
ACTGTCTAGTCTAAACACATCCATCGTACTGACTGCATCGTACTGACTGCATCGTACTGACTGCACATATCGT
CATAGCTCA
TCGTACTGACTGTCTAGTCTAAACACATCCCACATATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCC
ACTTTGGCA
ATATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCCACATATCGTCATCGTACTGACTGTCTAGTCTATC
CTATATTCGA
GCCGATCGTACGACACATATCGTCATCGTACTGCCCTACGGGACTGTCTAGTCTAAACACATCCATCGTACTG
ACTGCGGTC
TGACTGCATCGTACTGACTGCACATATCGTCATACATAGACTTCGTACTGACTGTCTAGTCTAAACACATCCC
ACATATATGC
CGTACTGACTGTCTAGACTAAACACATCCCACTTTACCCATGCATCGTACTGACTGTCTAGTCTAAACACATC
CCACATCGC
ATCGTACTGACTGTCTAGTCTAAACACATCCCAGCATCCATCCATATCGTCATCGTACTGACTGTCTAGTCTAA
ACACATTCG
GCCGATCGTACGACACATATCGTCATCGTACTGCCCTACGGGACTGTCTAGTCTAAACACATCCATCGTACTG
ACTGCGCCT
TGACTGCATCGTACTGACTGCACATATCGTCATACATAGACTTCGTACTGACTGTCTAGTCTAAACACATCCC
ACATATAATG
CGTACTGACTGTCTAGTCTAAACACATCCCACTTTACCCATGATATCGTCATCGTACTGACTGTCTAGTCTAAA
CACATTTCG

ATGCCGATCGTACGACACATATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCATCGTACTGACTGCAT
CGATCTTGC
TACTGACTGCATCGTACTGACTGCACATATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCCACATATCC
TTTACTTGC
CATCGTACTGACTGTCTAGTCTAAACACATCCCACATATCGTCATCGTACTGACTGTCTAGTCTAAACACATCC
CAGCAAAG
CATATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCTATGCCGATCGTACGACACATATCGTCATCGTAC
TGCCCACGT
ACTGTCTAGTCTAAACACATCCATCGTACTGACTGCATCGTACTGACTGCATCGTACTGACTGCACATATCGT
CATAGCTCA
TCGTACTGACTGTCTAGTCTAAACACATCCCACATATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCC
ACTTTGGCA
ATATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCCACATATCGTCATCGTACTGACTGTCTAGTCTATC
CTATATTCGA
GCCGATCGTACGACACATATCGTCATCGTACTGCCCTACGGGACTGTCTAGTCTAAACACATCCATCGTACTG
ACTGCGGTC
TGACTGCATCGTACTGACTGCACATATCGTCATACATAGACTTCGTACTGACTGTCTAGTCTAAACACATCCC
ACATATATGC
CGTACTGACTGTCTAGGCTAAACACATCCCACTTTACCCATGCATCGTACTGACTGTCTAGTCTAAACACATC
CCACATCGC
ATCGTACTGACTGTCTAGTCTAAACACATCCCAGCATCCATCCATATCGTCATCGTACTGACTGTCTAGTCTAA
ACACATTCG
GCCGATCGTACGACACATATCGTCATCGTACTGCCCTACGGGACTGTCTAGTCTAAACACATCCATCGTACTG
ACTGCGCCT
TGACTGCATCGTACTGACTGCACATATCGTCATACATAGACTTCGTACTGACTGTCTAGTCTAAACACATCCC
ACATATAATG
CGTACTGACTGTCTAGTCTAAACACATCCCACTTTACCCATGATATCGTCATCGTACTGACTGTCTAGTCTAAA
CACATTTCG


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

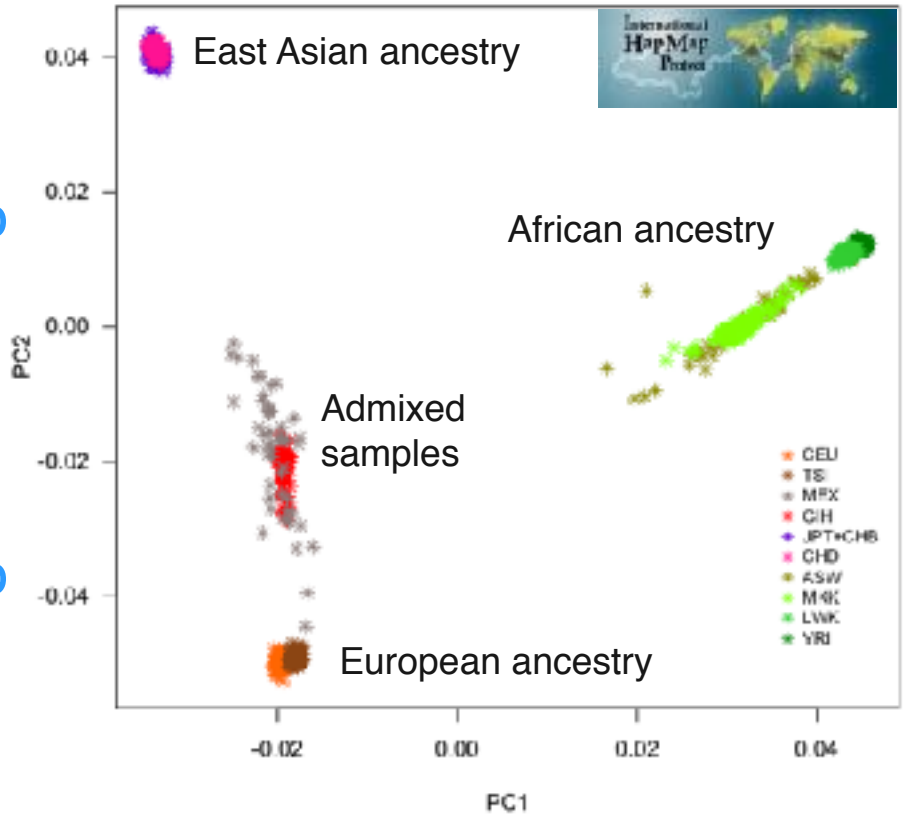
1.1 million SNPs
270 individuals from 4 populatio

Phase II

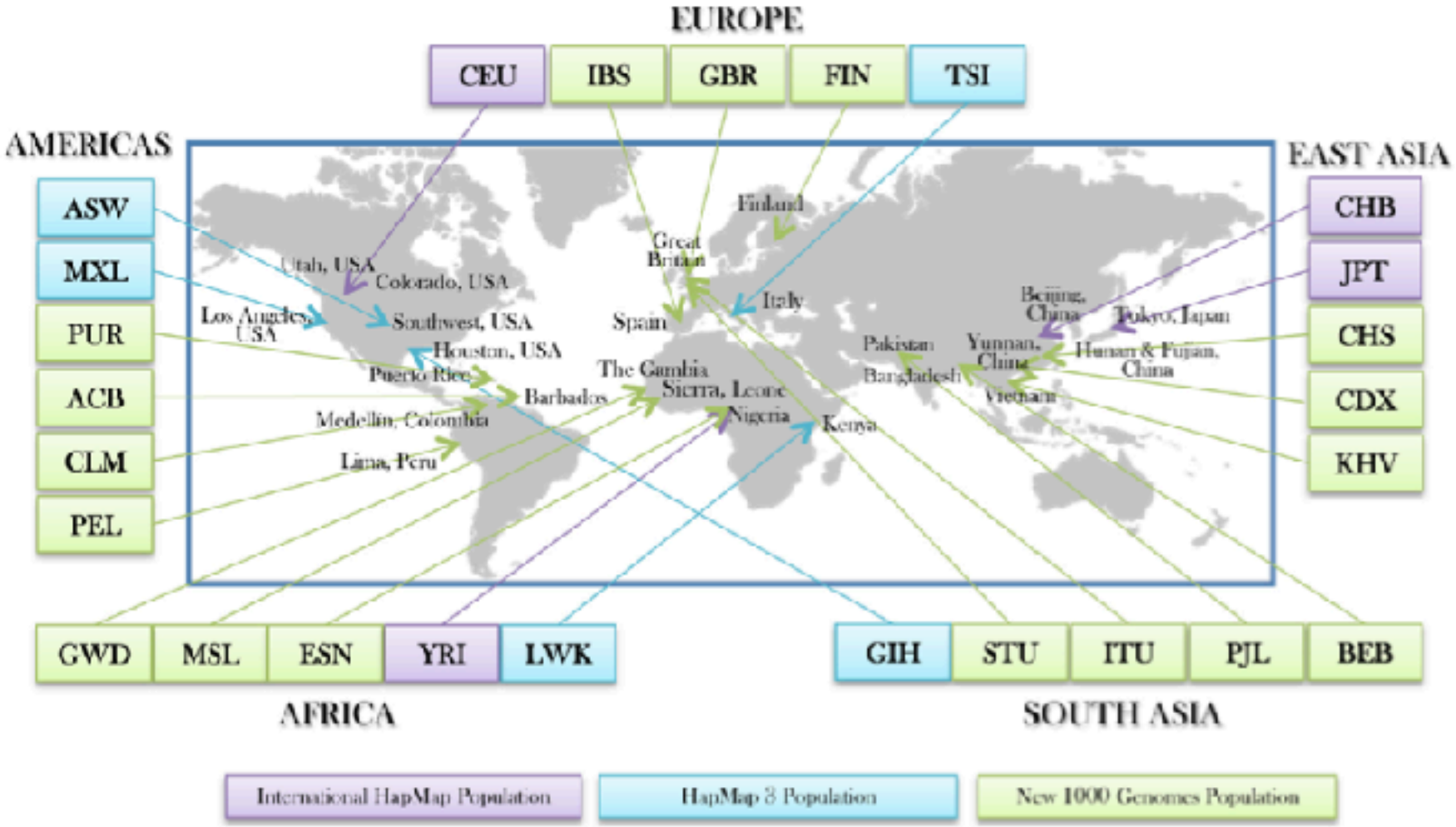
3.1 million SNPs
270 individuals from 4 populatio

Phase III

1.6 million SNPs
1,184 individuals from 11 populations



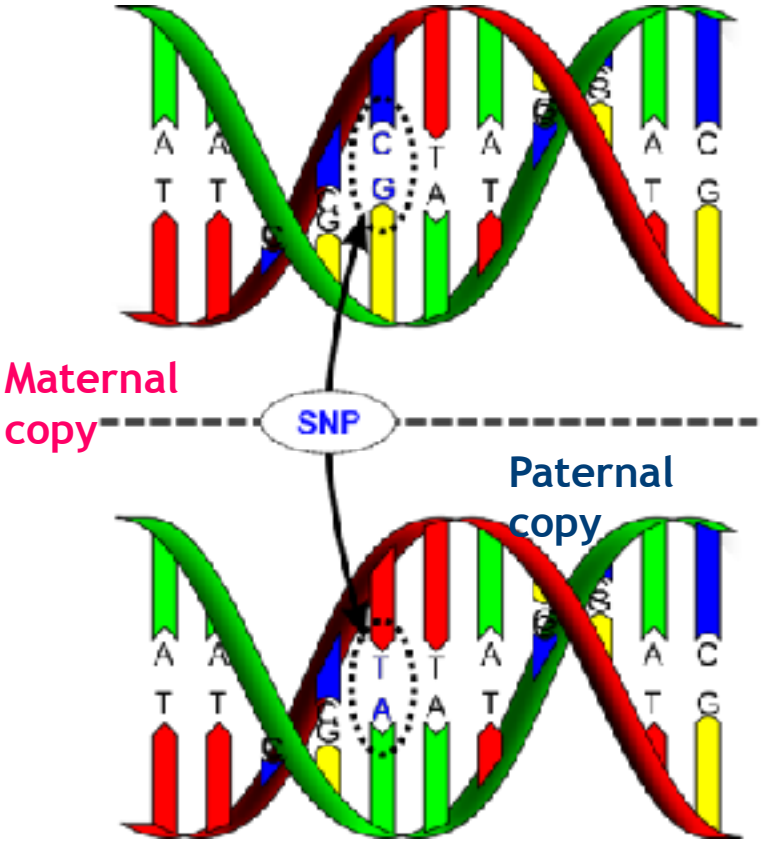
The 1000 Genomes Project



*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

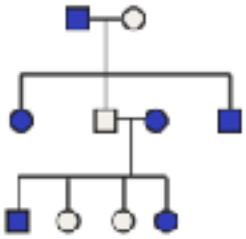
www.hapmap.org

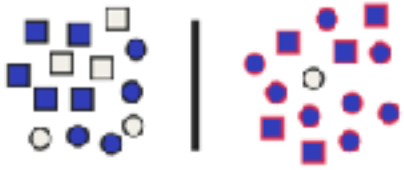


Linkage analysis

Candidate gene studies

GWAS





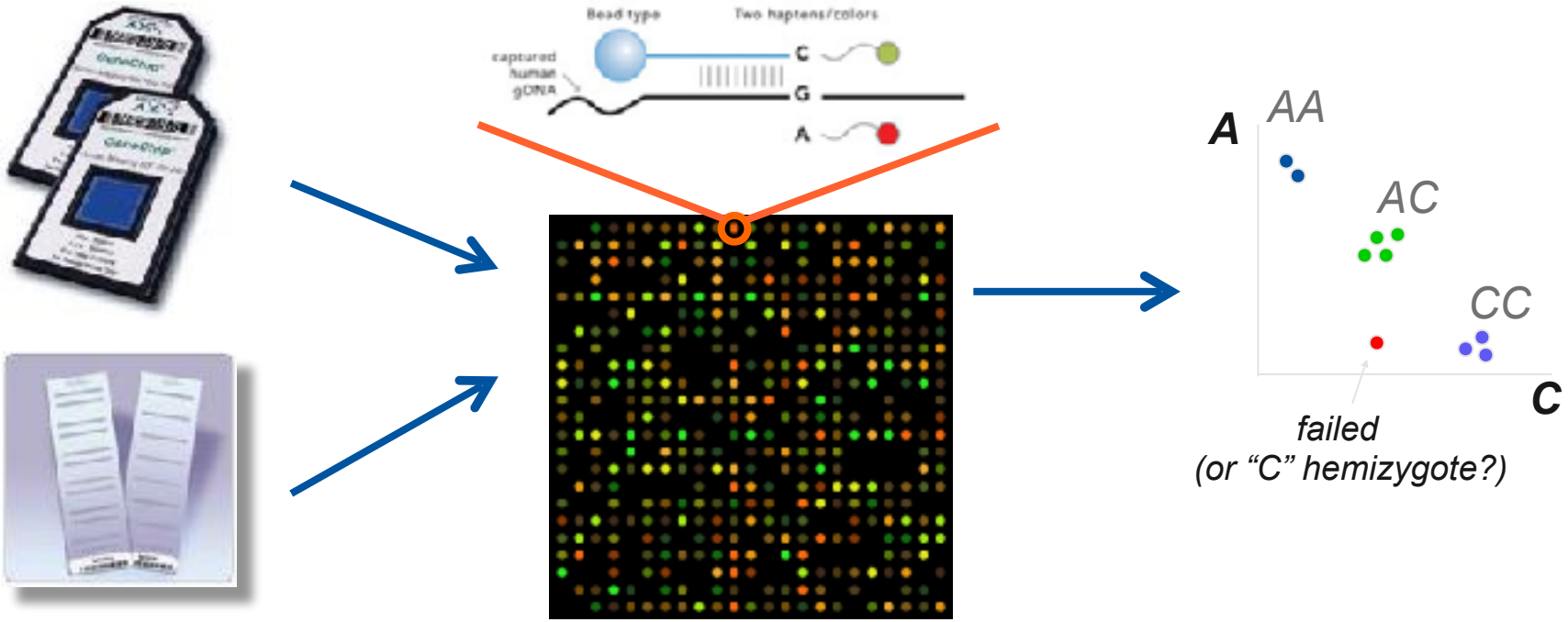
Common variant, common disease hypothesis

- 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

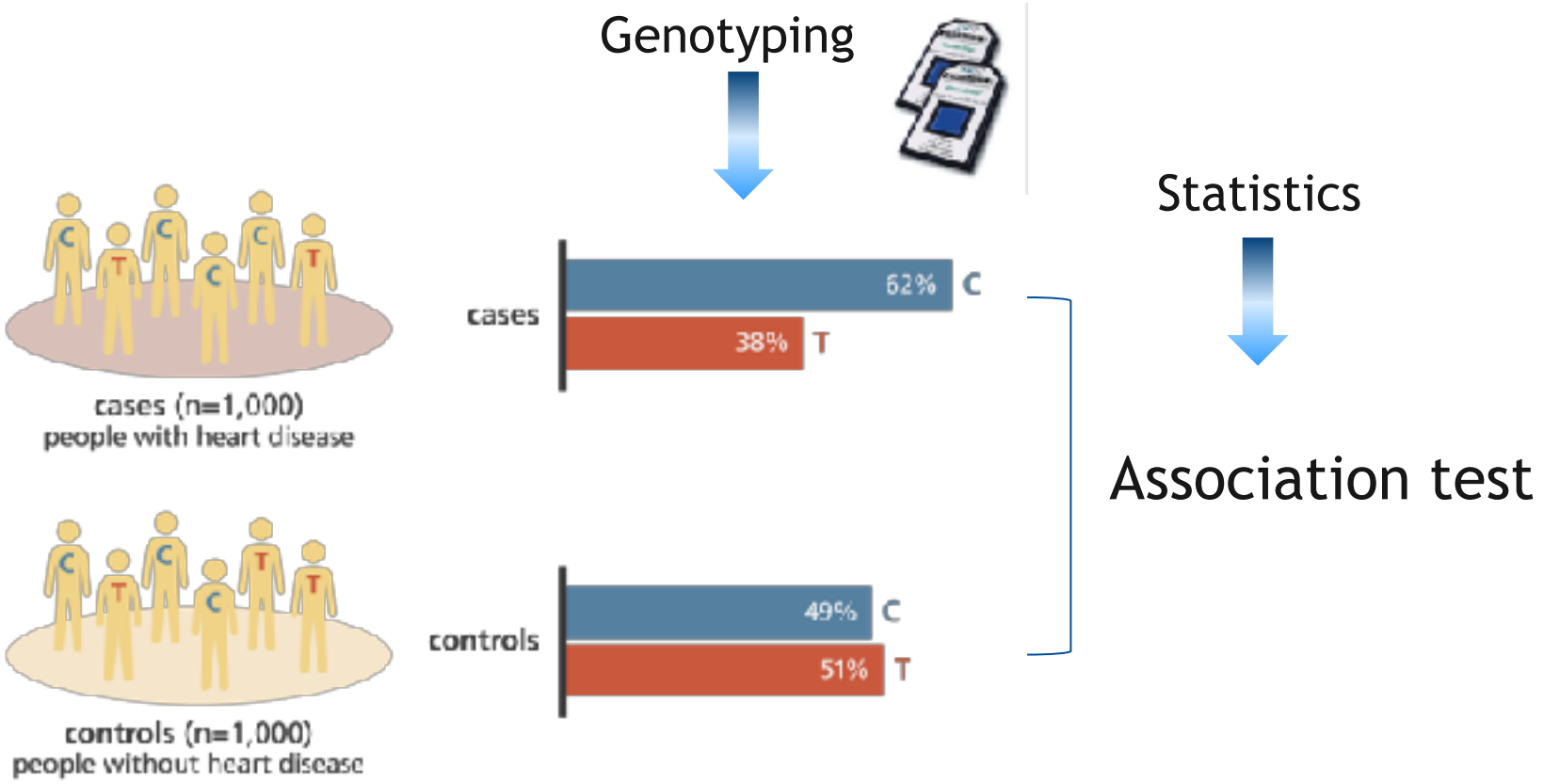
The beginnings of GWAS

Genotyping platforms

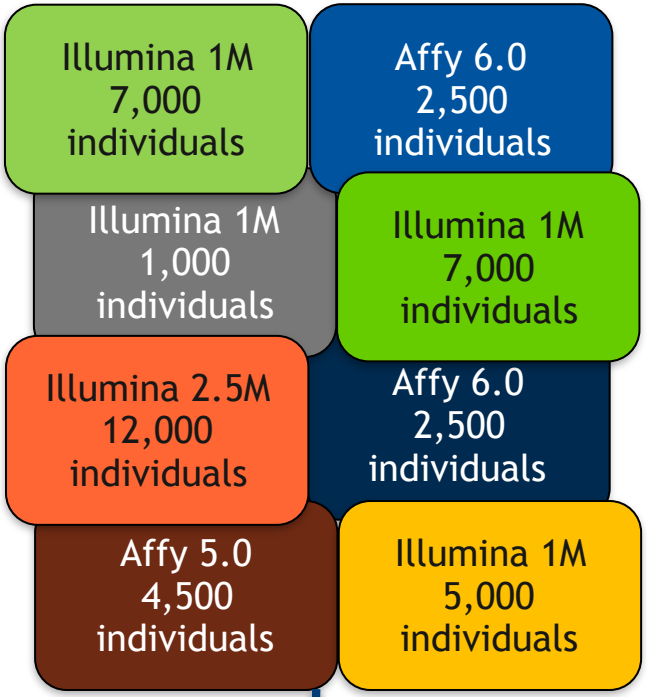
- Genome-wide SNP microarrays allow measurement of genotypes of 100,000's of SNPs in a single experiment
- Variety of microarrays (different SNP density, cost, etc) by Illumina and Affymetrix



GWAS (the big picture)

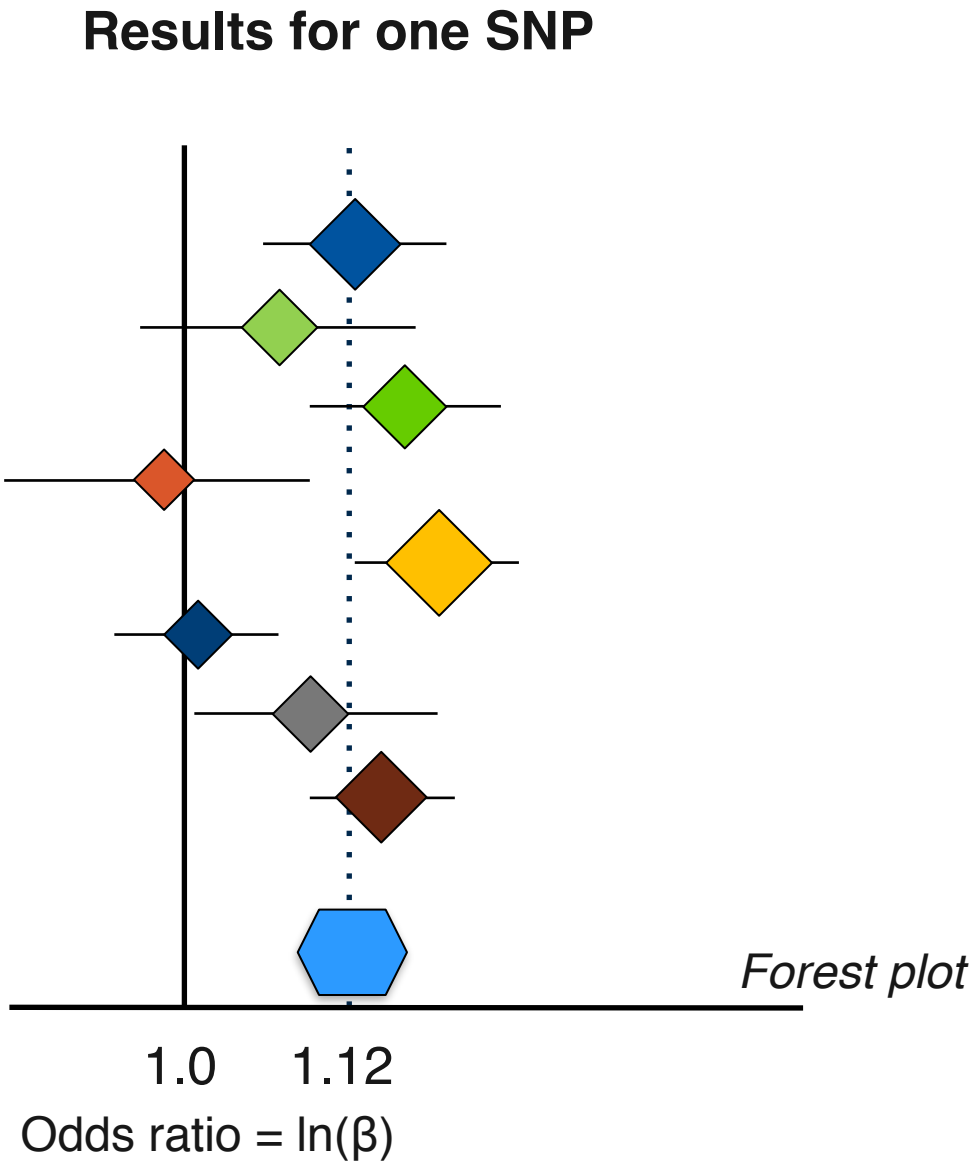


Combining GWAS datasets



Imputation

Meta-analysis of GWAS

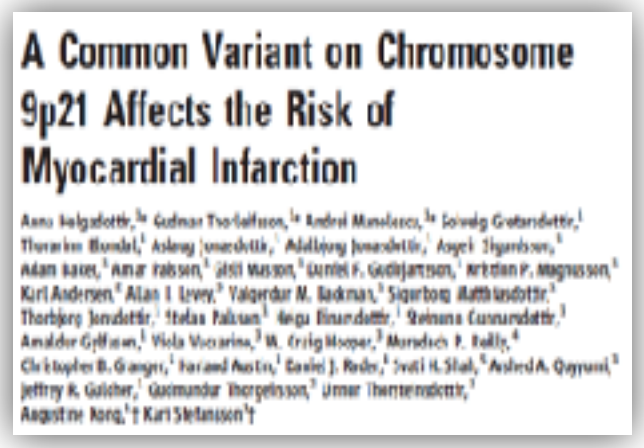


deCODE Genetics, Inc.

- >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,
- 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

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

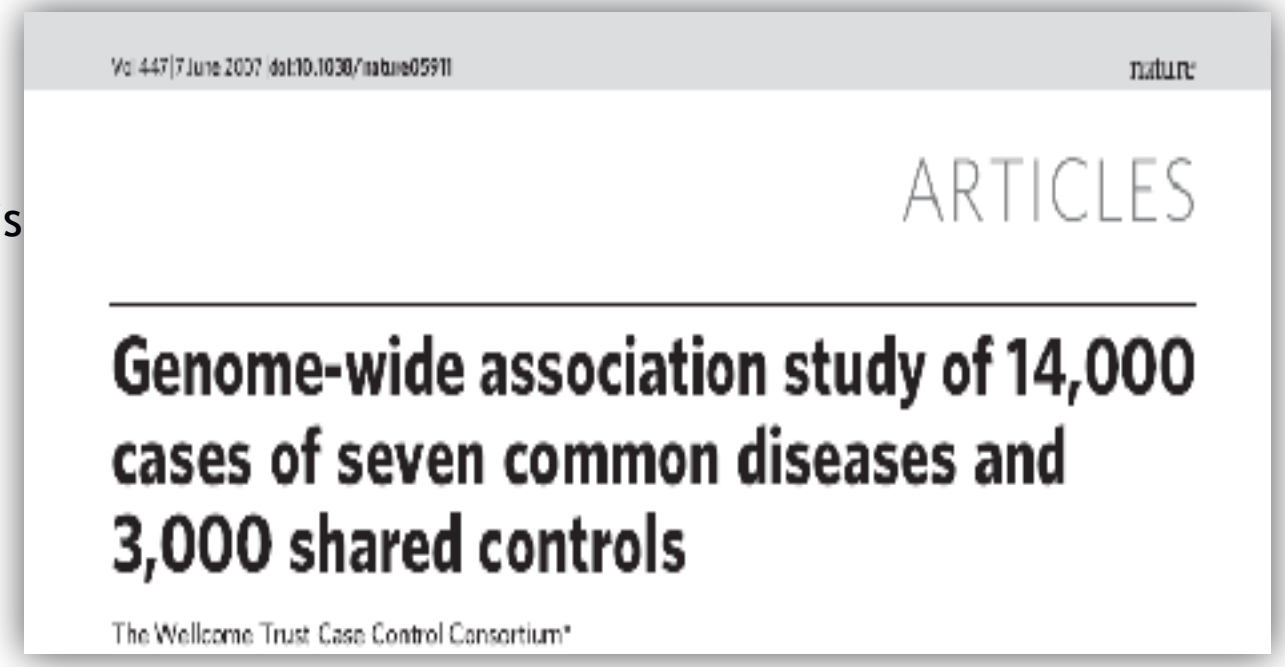


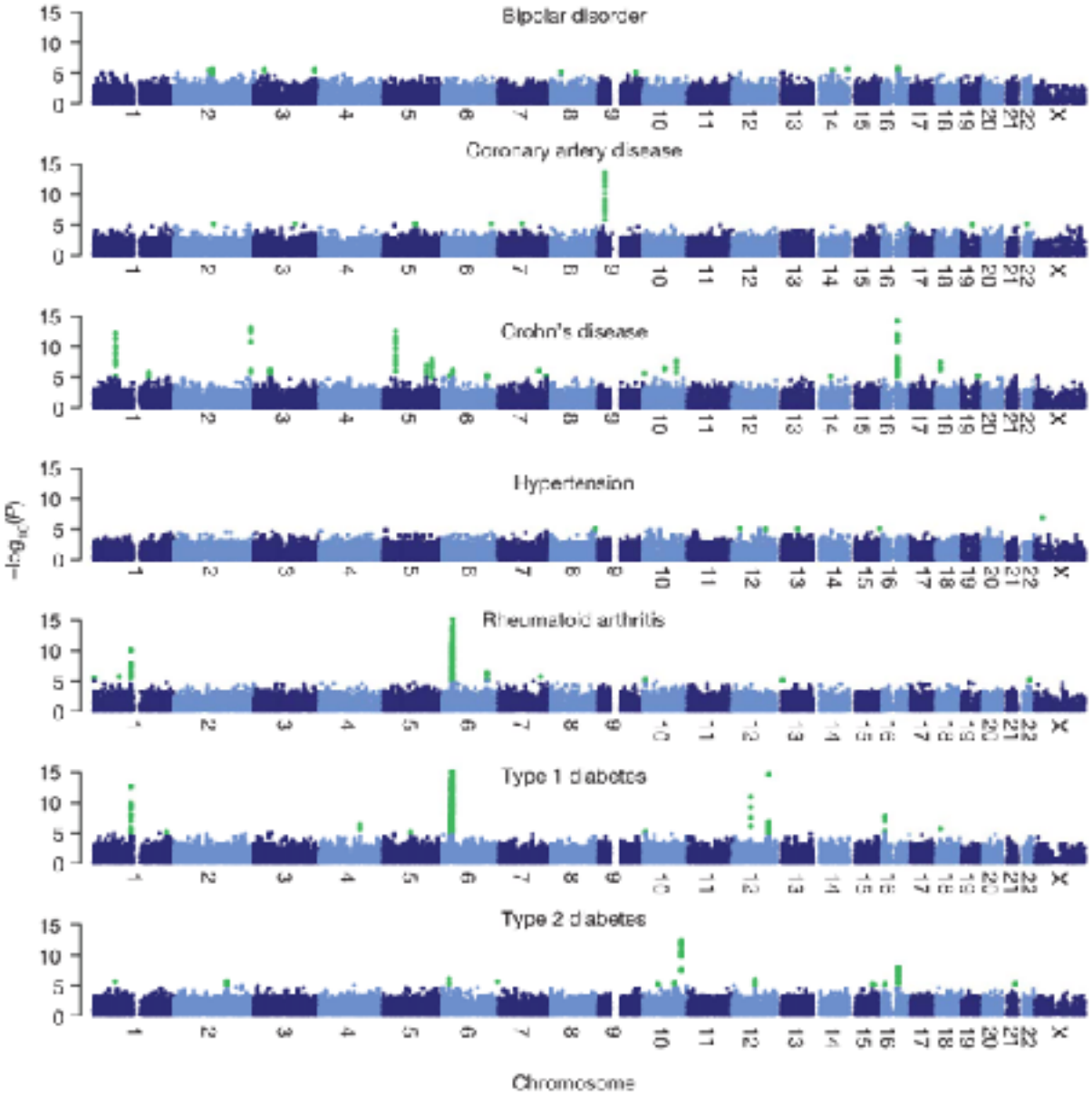
A Common Variant on Chromosome 9p21 Affects the Risk of Myocardial Infarction

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

Agnar Helgadottir, Gudmar Thorleifsson, Andrei Manolescu, Solveig Gretarsdottir, Thorarinn Blondal, Ashley Jonasdottir, Adalbjorg Jonasdottir, Asgeir Sigurdsson, Adam Baker, Arnar Palsson, Jesus Masson, Daniel F. Gudbjartsson, Kristian P. Magnusson, Kari Andersen, Allan I. Levey, Valgerdur M. Backman, Sigurborg Matthiasdottir, Thorbjorg Jonsdottir, Stefan Palsson, Hrefna Einarsdottir, Steinunn Gunnarsdottir, Amalder Gylfason, Viola Vaccarino, W. Craig Hooper, Monsheen P. Reilly, Christopher B. Granger, Farzad Austin, David J. Reich, Svati H. Shah, Ashraf A. Quyyumi, Jeffrey R. Gulcher, Gudmundur Thorgeirsson, Unnur Thorsteinsdottir, Augustine Kong, Kari Stefansson

The NEW ENGLAND JOURNAL of MEDICINE

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

Genomewide Association Analysis of Coronary Artery Disease

A Common Allele on Chromosome 9 Associated with Coronary Heart Disease

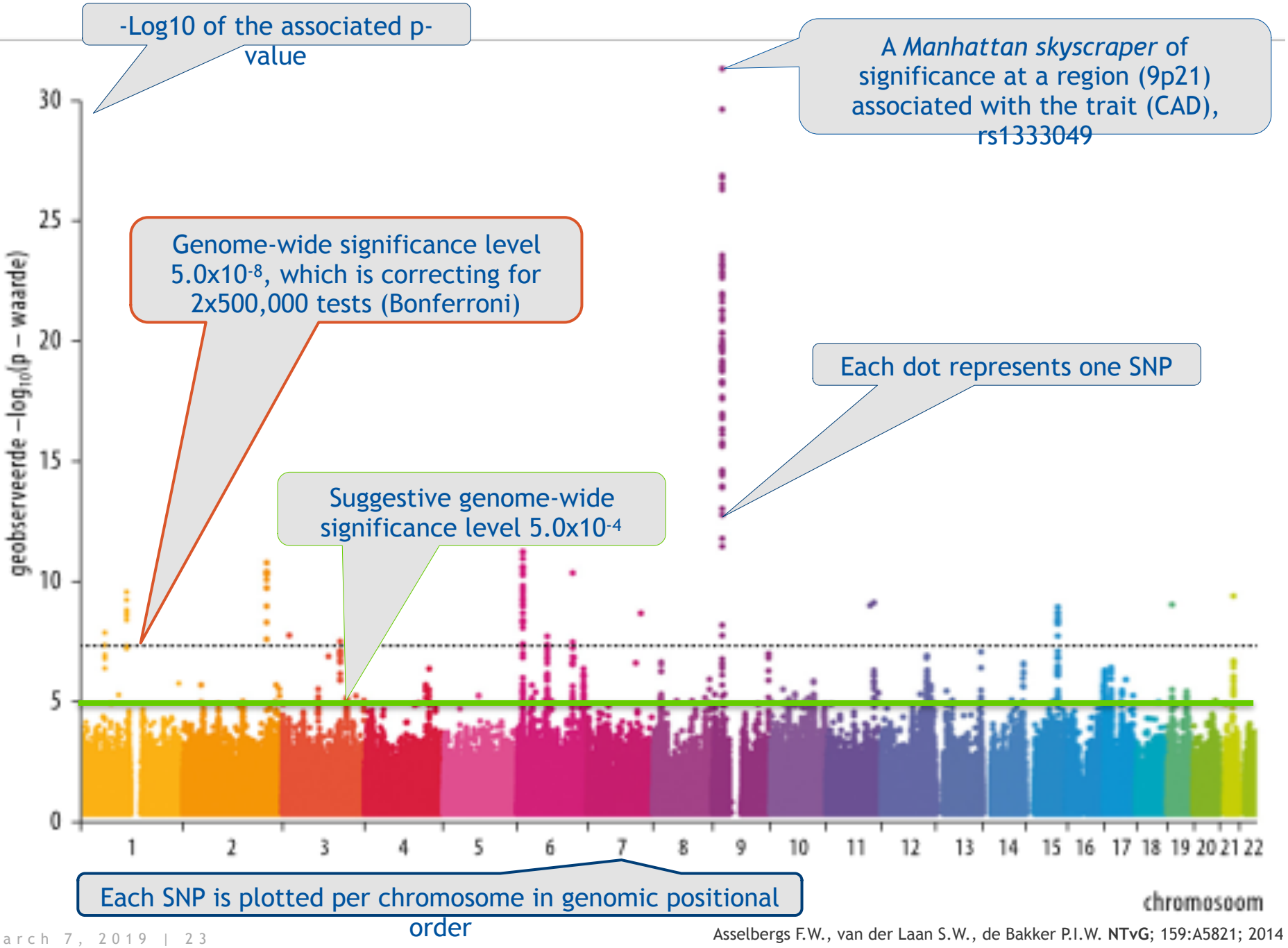
Ruth McPherson, Alexandre Pertsemlidis, Nihan Kavaslar, Alexandre Stewart, Robert Roberts, David R. Cox, David A. Hinds, Len A. Pennacchio, Anne Tybjaerg-Hansen, Aaron R. Folsom, Eric Boerwinkle, Helen H. Hobbs, Jonathan C. Cohen

ARTICLES

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

The Wellcome Trust Case Control Consortium*

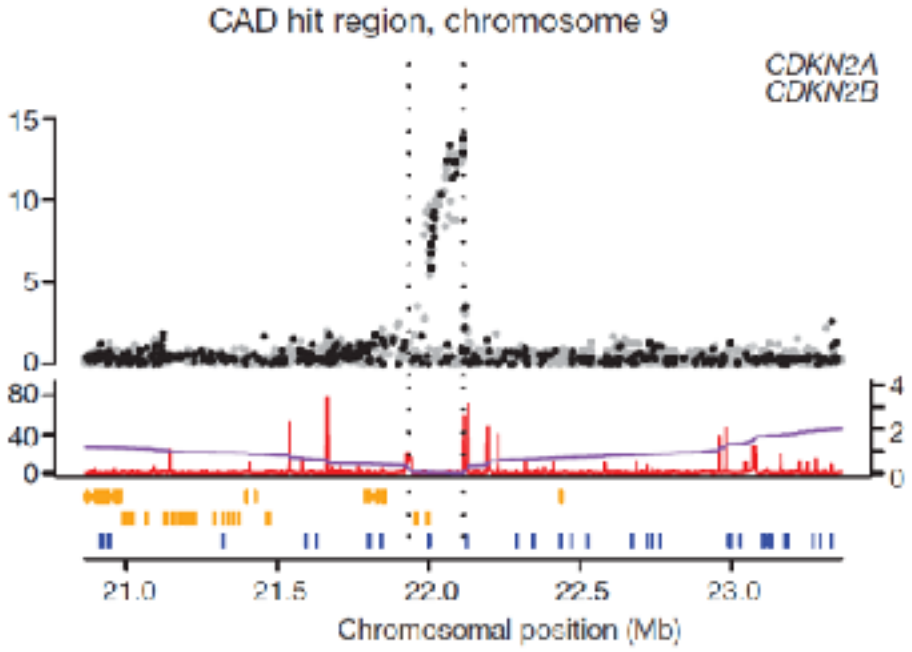
Helgadottir, A., *et al.* **Science**; 316(5830):1491-1493, 2007
McPherson, R., *et al.* **Science**; 316(5830):1488-1491, 2007
Wellcome Trust Case Control Consortium. **Nature**; 447(7145):661-678, 2007
Samani, N.J., *et al.* **N Engl J Med**; 357(5):443-453, 2007



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^{-16}	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

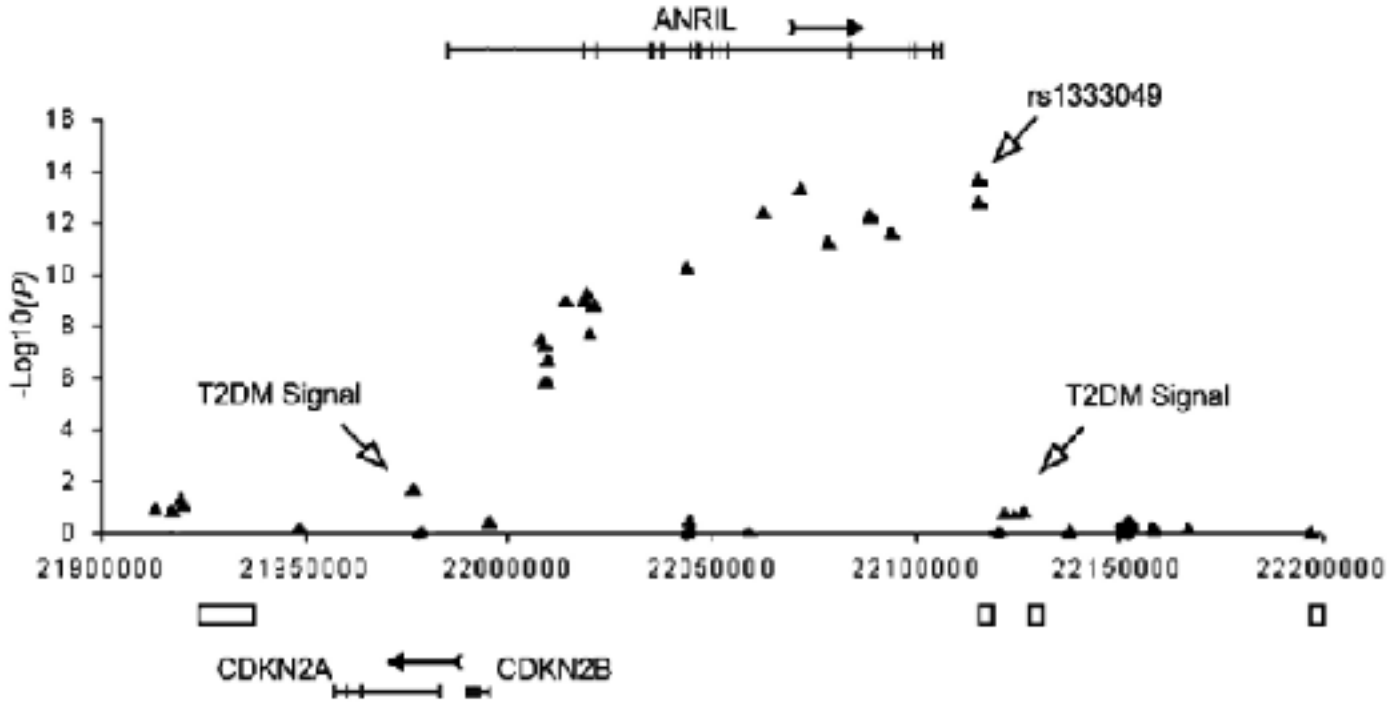
P-value of association test: AA vs. AB vs. BB

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

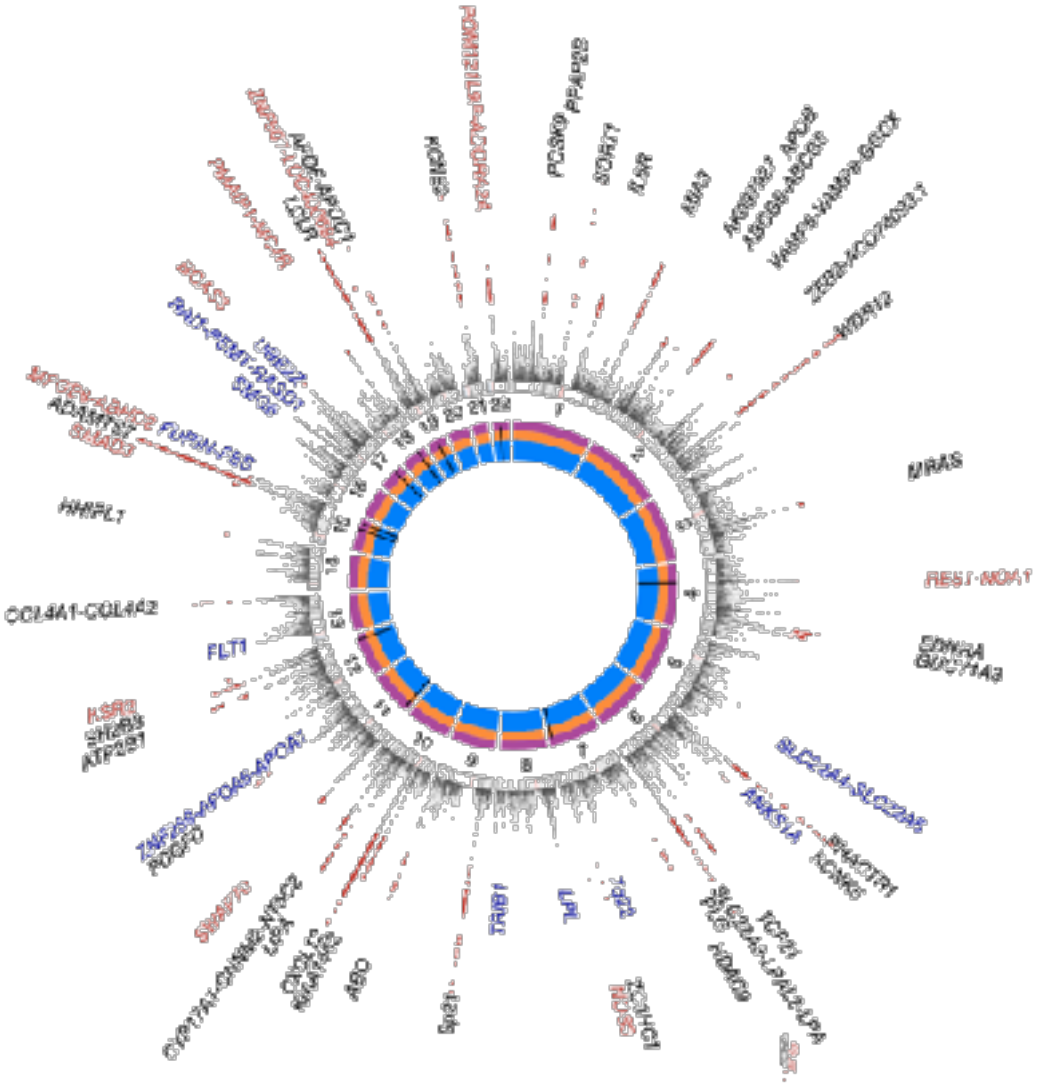
- Coronary Artery Disease Genome-Wide Replication And Meta-Analysis Study: CARDIoGRAM
- > 63,000 cases and > 130,000 controls
  - Myocardial infarction (MI), coronary artery disease (CAD) or both
  - CAD: MI, CABG, PTCA, AP
  - Age limit: 45-66
- Sample size greatly influences power and effect size to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

ARTICLES

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

The CARDIoGRAMplusC4D Consortium*

And 8 years later (>15 times more samples)

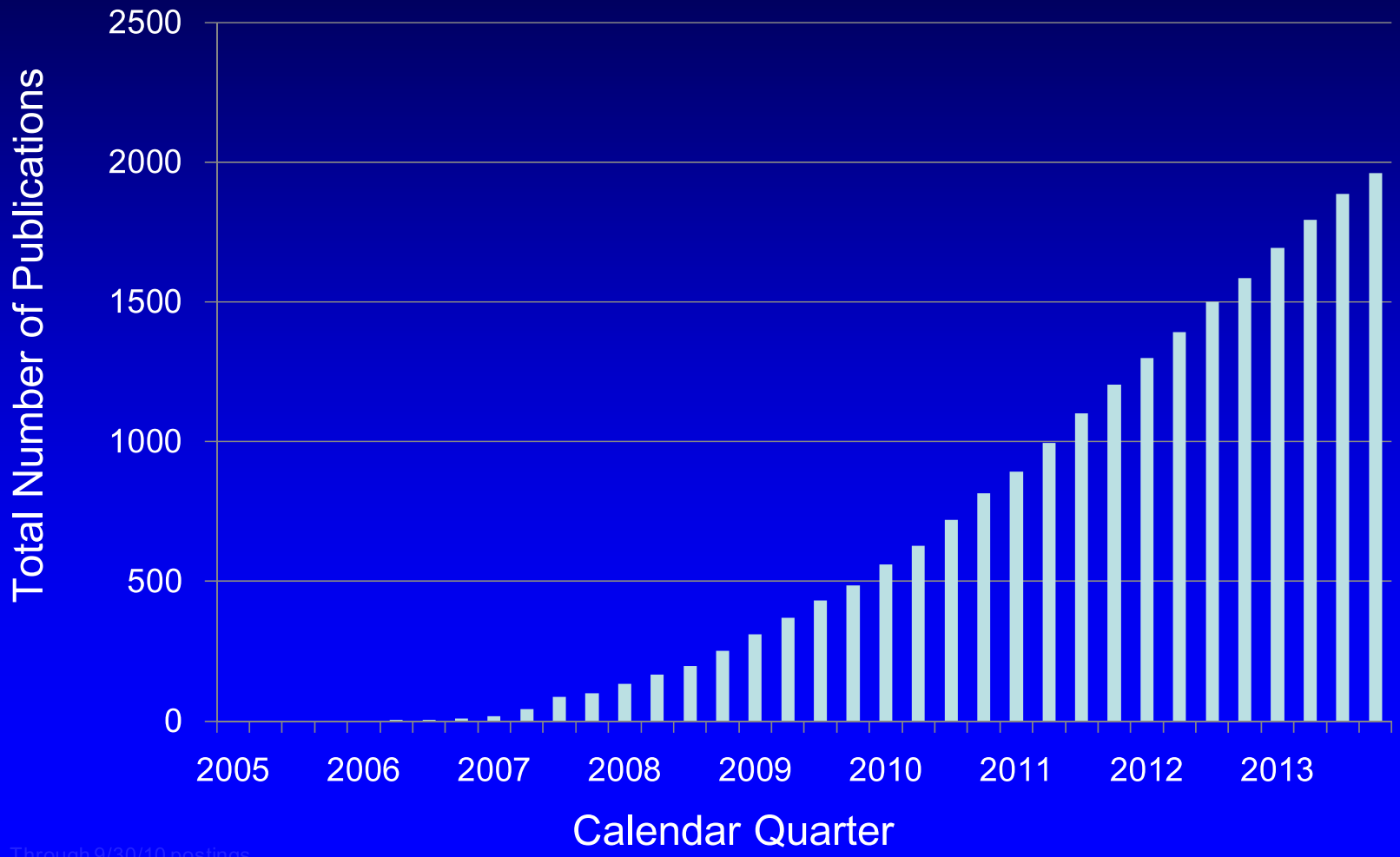


9p21 plus an additional 47 loci (!)

CARDIoGRAMplusC4D Consortium, Nature Genetics 2015

Published GWA Reports, 2005 – 2013

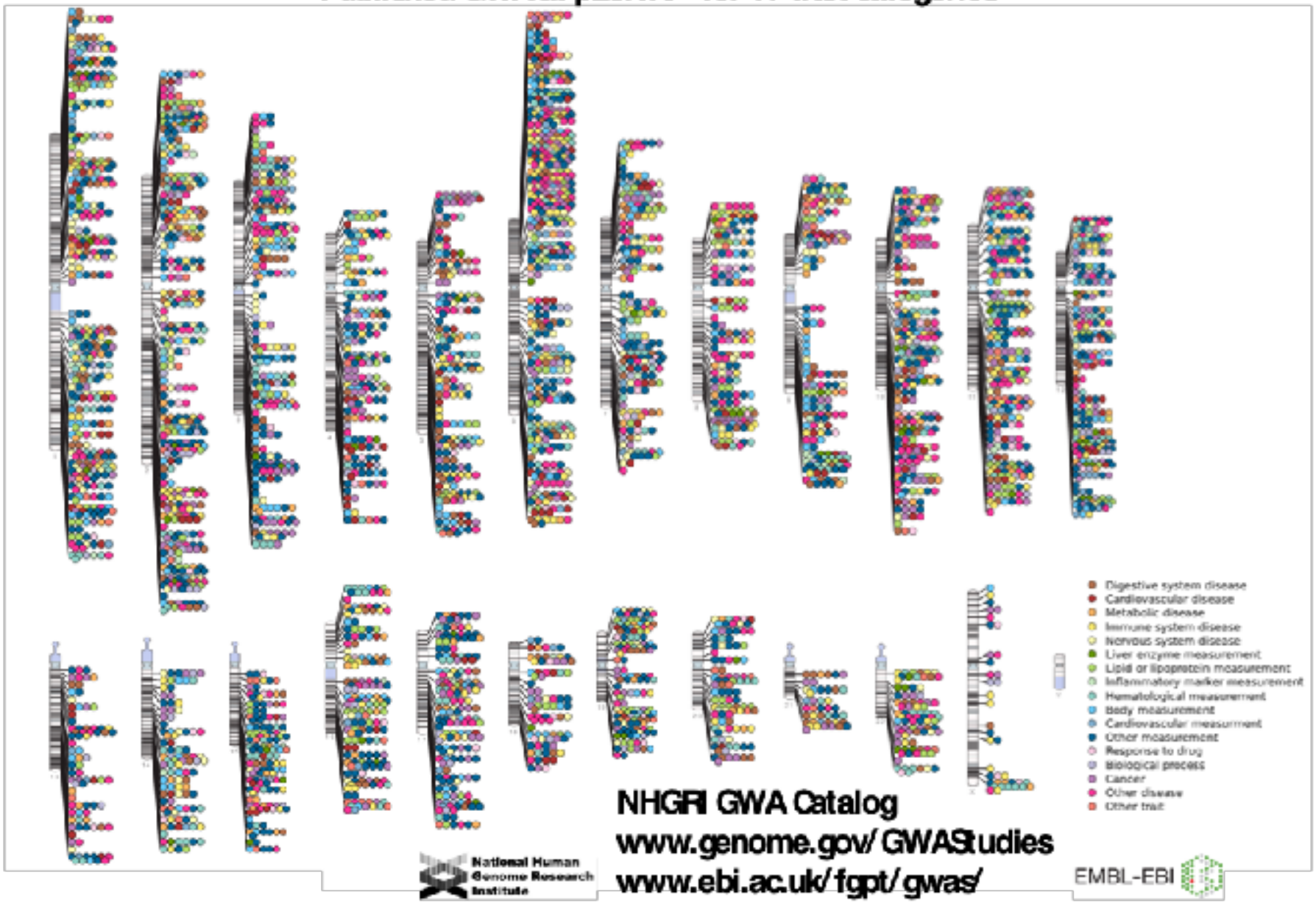
1960



Through 9/30/10 postings

Published Genome-Wide Associations through 12/2013

Published GWA at $p \leq 5X10^{-8}$ for 17 trait categories



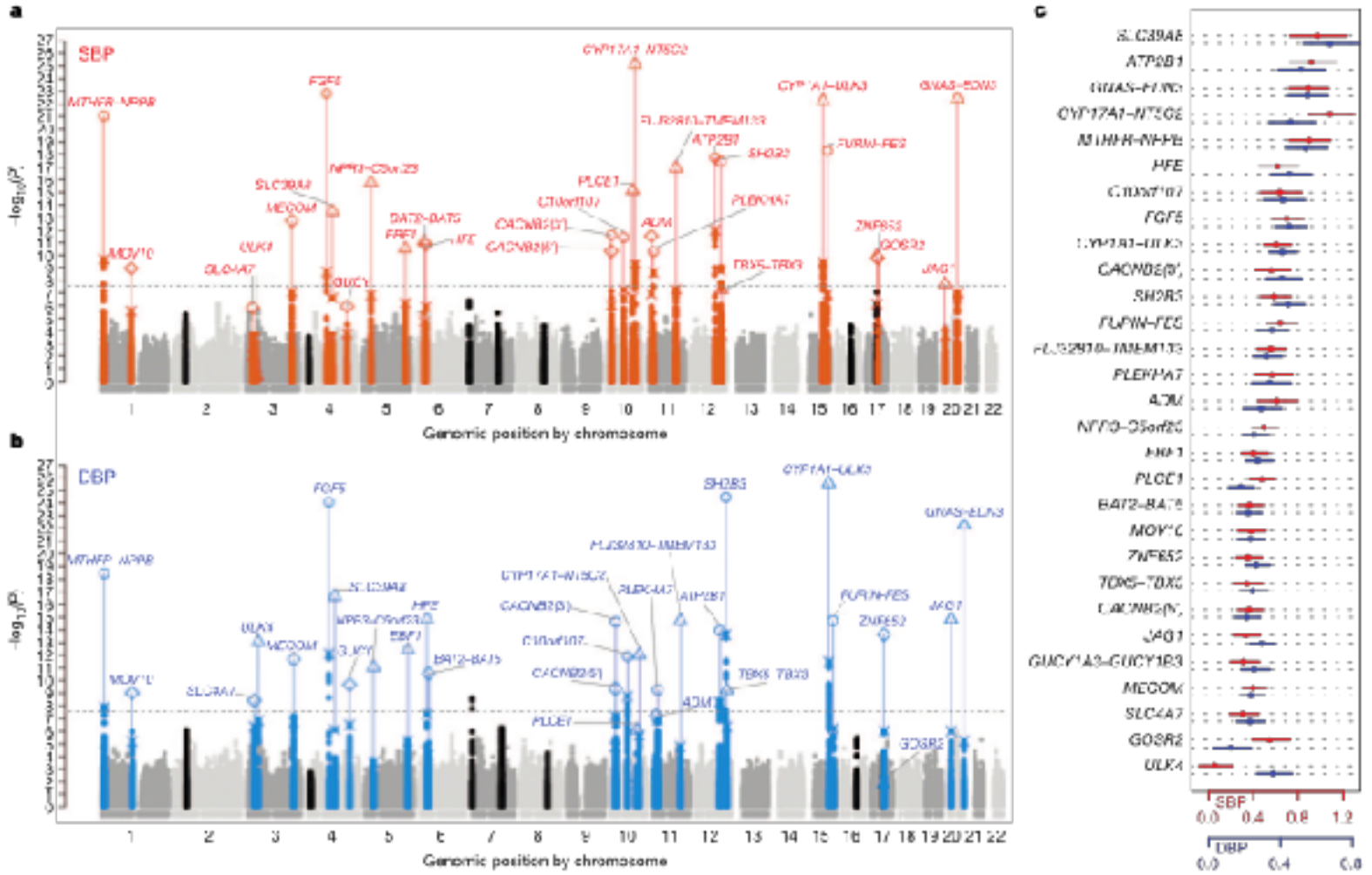


Figure 1 Genome-wide $-\log_{10}$ *P*-value plots and effects for significant loci. **a**, **b**, Genome-wide $\log_{10}$ *P*-value plots are shown for SBP (**a**) and DBP (**b**). SNPs within loci reaching genome-wide significance are labelled in red for SBP and blue for DBP (±2.5 Mb of lowest *P* value) 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.

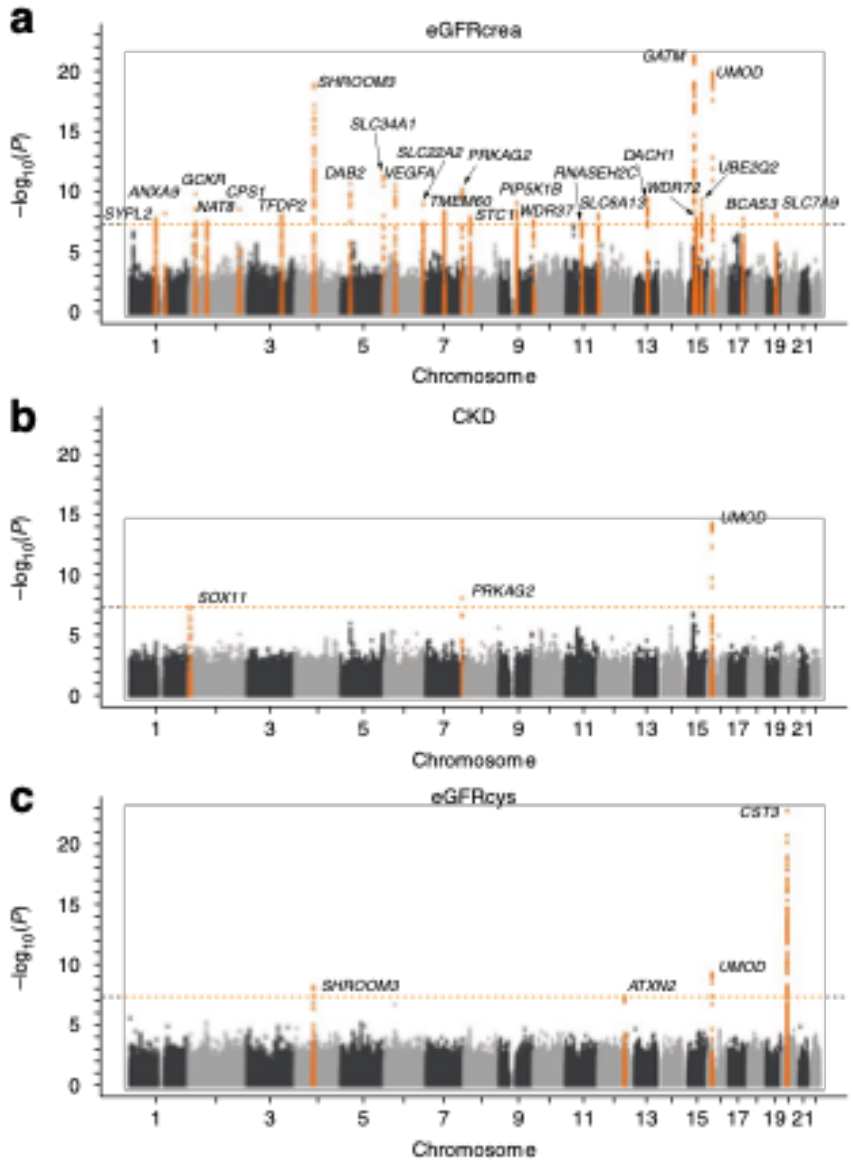


Figure 1 Genome-wide $-\log_{10}$ *P* value plot from stage 1. (**a**–**c**) Plots show discovery analysis of eGFRcrea (**a**), CKD (**b**) and eGFRcys (**c**). The dotted line indicates the genome-wide significance threshold at $P = 5 \times 10^{-8}$.

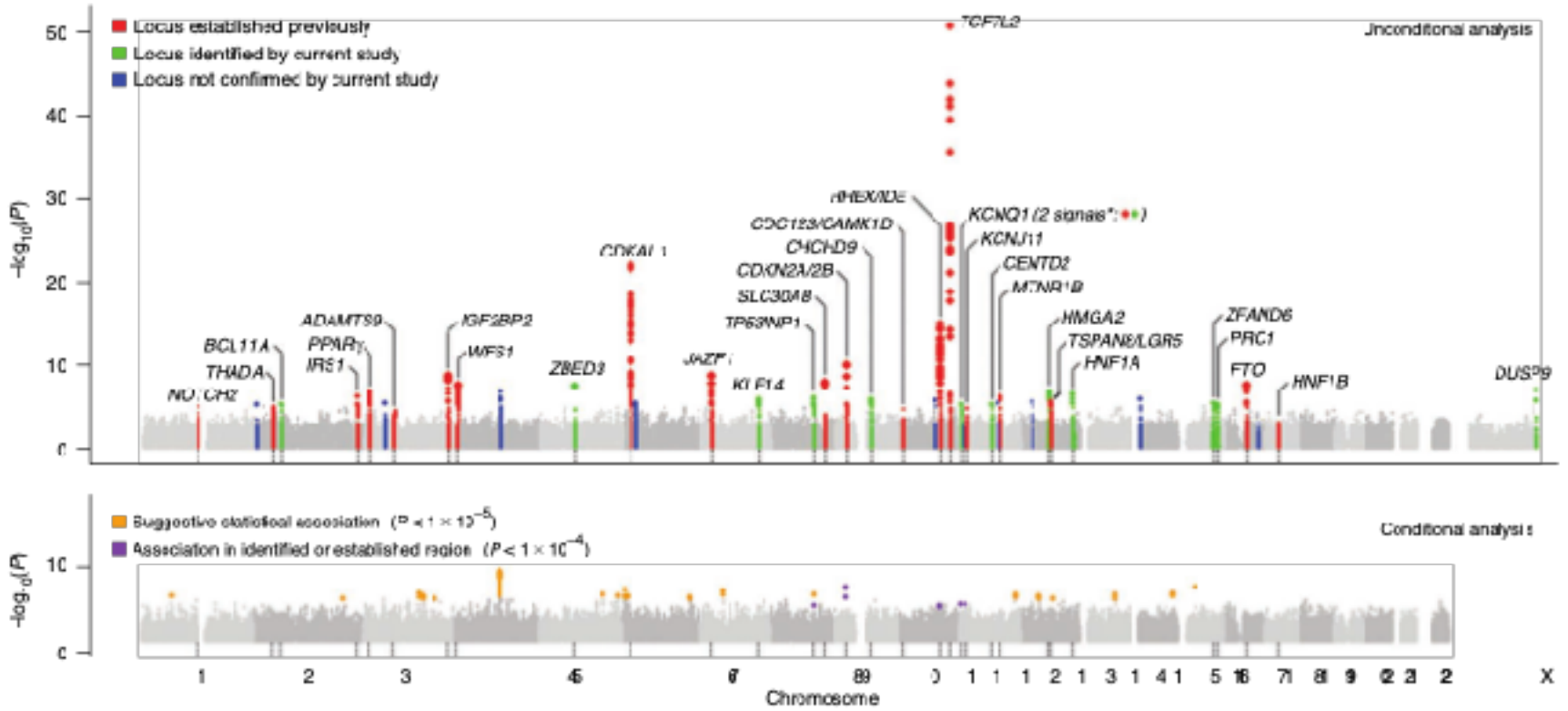
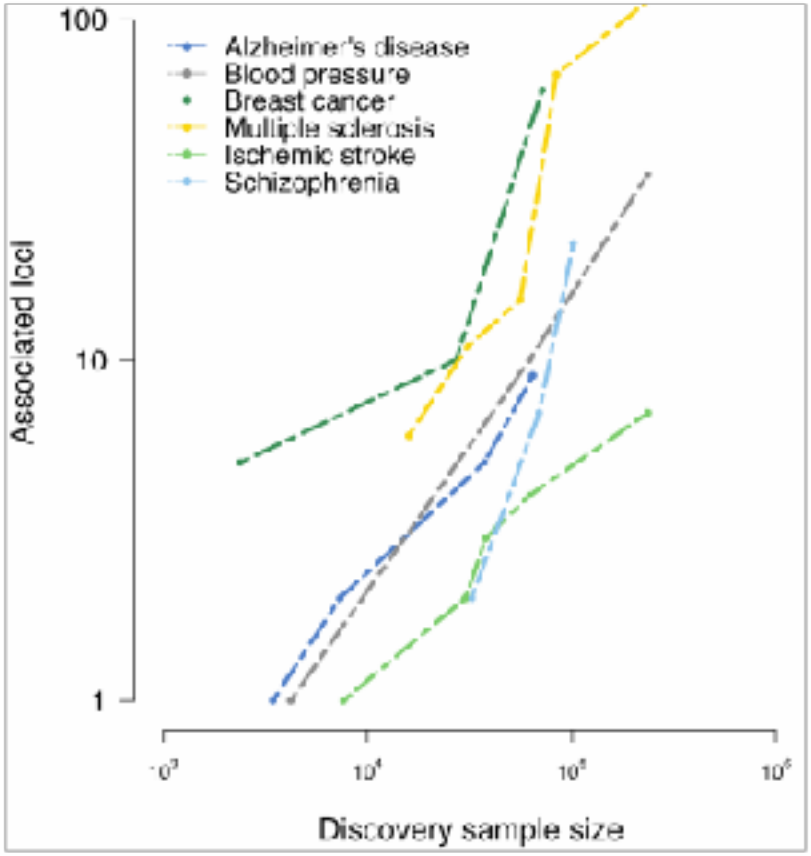
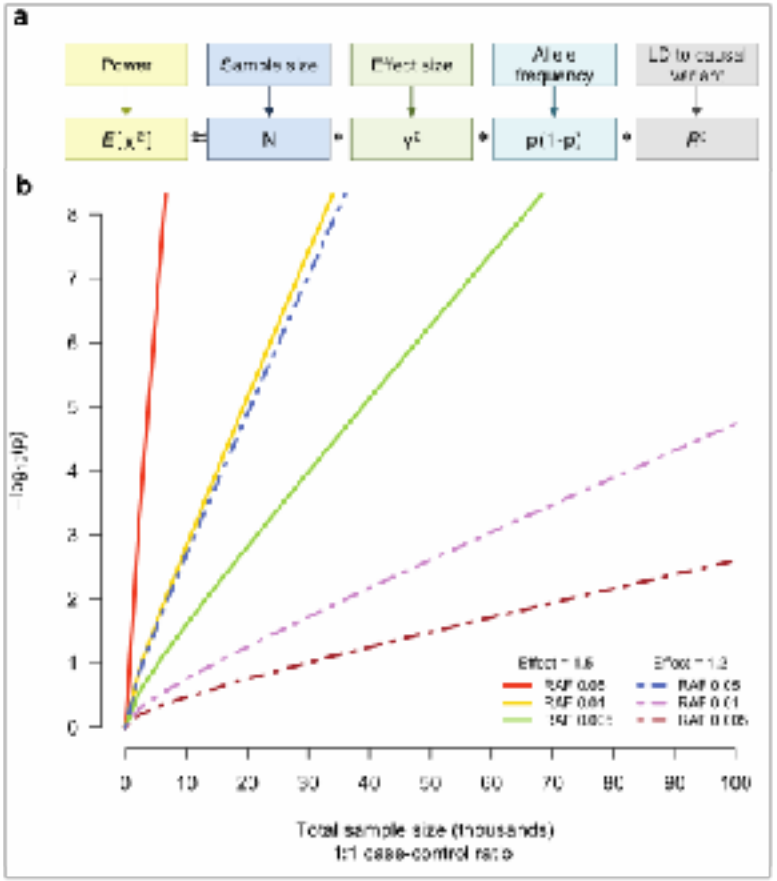


Figure 1 Genome-wide Manhattan plots for the DIAGRAM+ stage 1 meta-analysis. Top panel summarizes the results of the unconditional meta-analysis. Previously established loci are denoted in red and loci identified by the current study are denoted in green. The ten signals in blue are those taken forward 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…





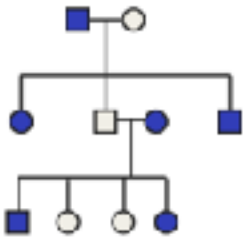
Pulit, S., *et al.* **Genes & Genomes**; 2014

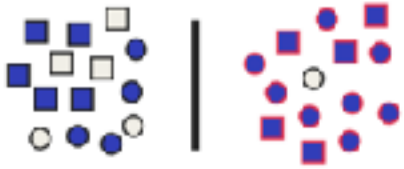
Linkage analysis

Candidate gene studies

GWAS

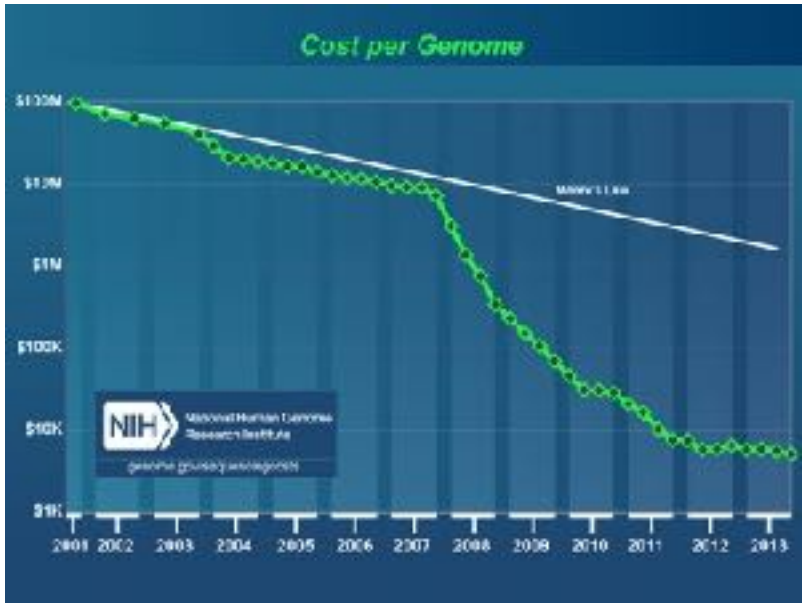
Sequencing



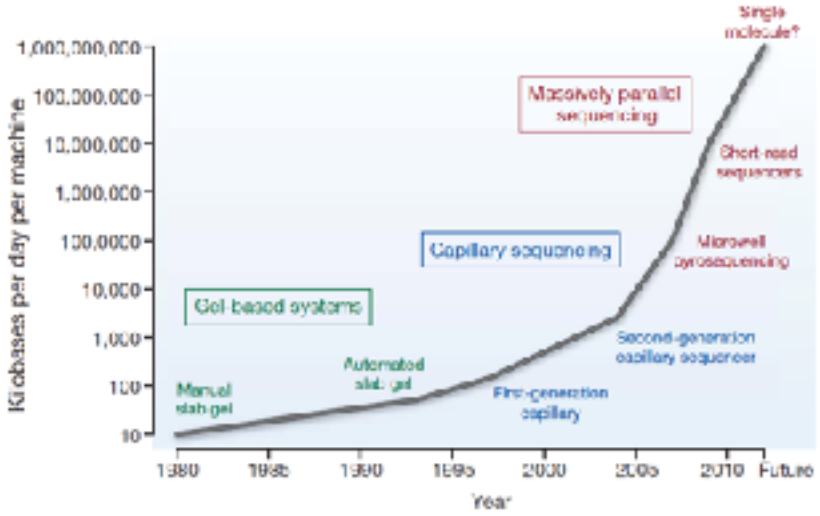


Next-generation sequencing

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