GWAS across Time and Space

What have we learned?

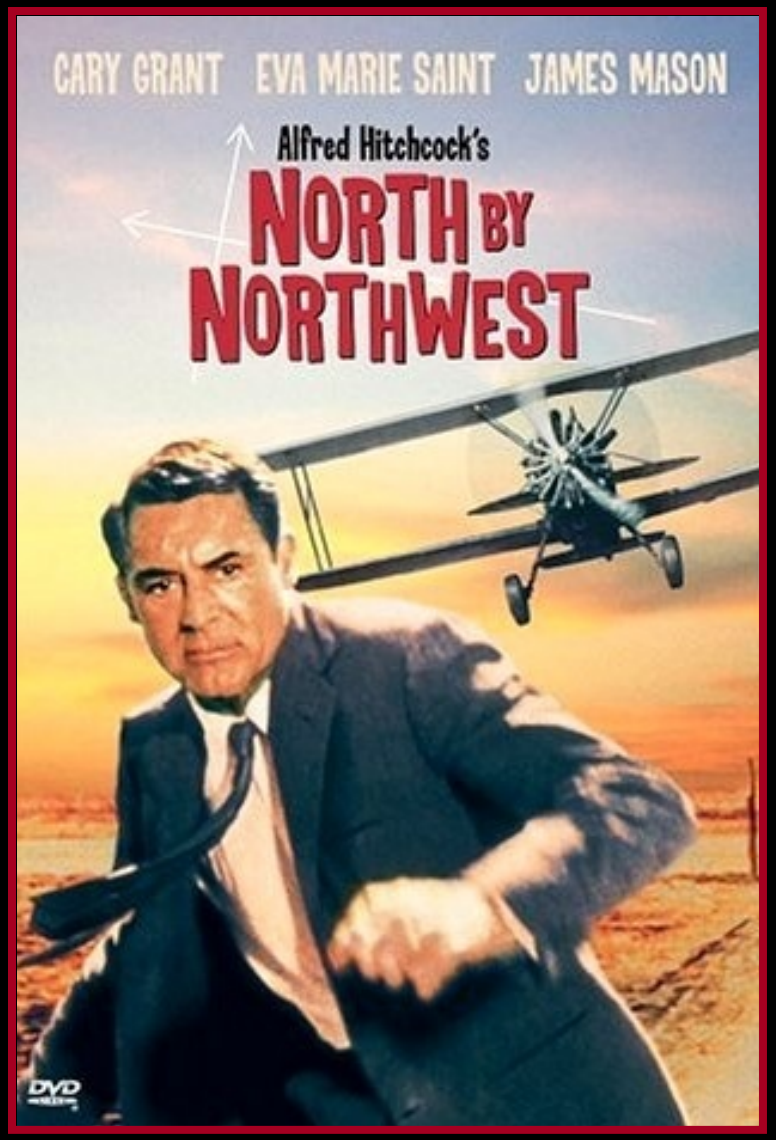
Arcadi Navarro

BIOSTEC

March the 4th 2014

*iCrea INSTITUCIÓ CATALANA DE RECERCA I ESTUDIS AVANÇATS
IBE CSIC UPF INSTITUT de BIOLOGIA EVOLUTIVA
INB
prbb Barcelona Biomedical Research Park
CRG Centre de Regulació Genòmica
UPF UNIVERSITAT POMPEU FABRA

Warning


CARY GRANT EVA MARIE SAINT JAMES MASON
Alfred Hitchcock's
NORTH BY NORTHWEST
DVD

Warning

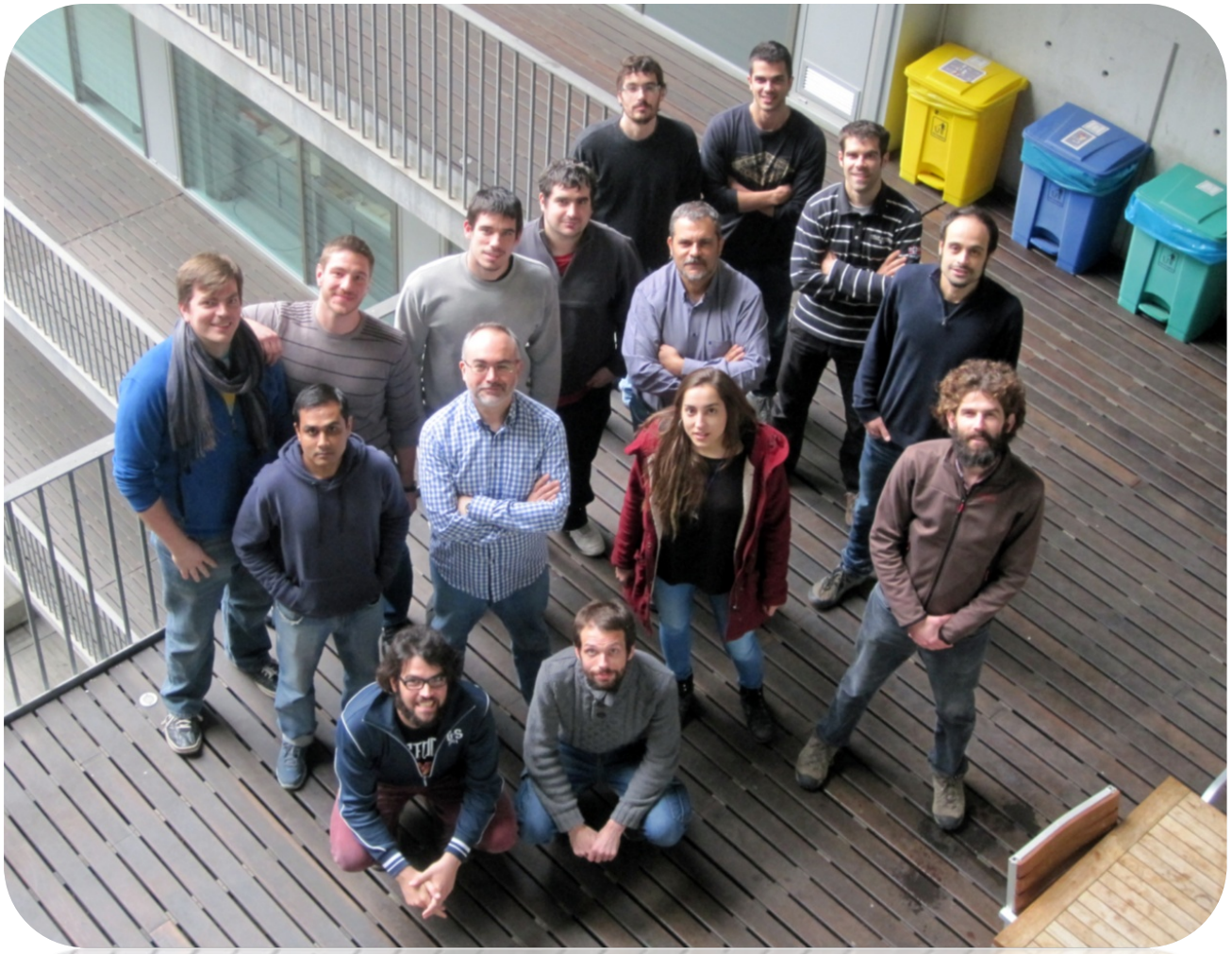
Warning

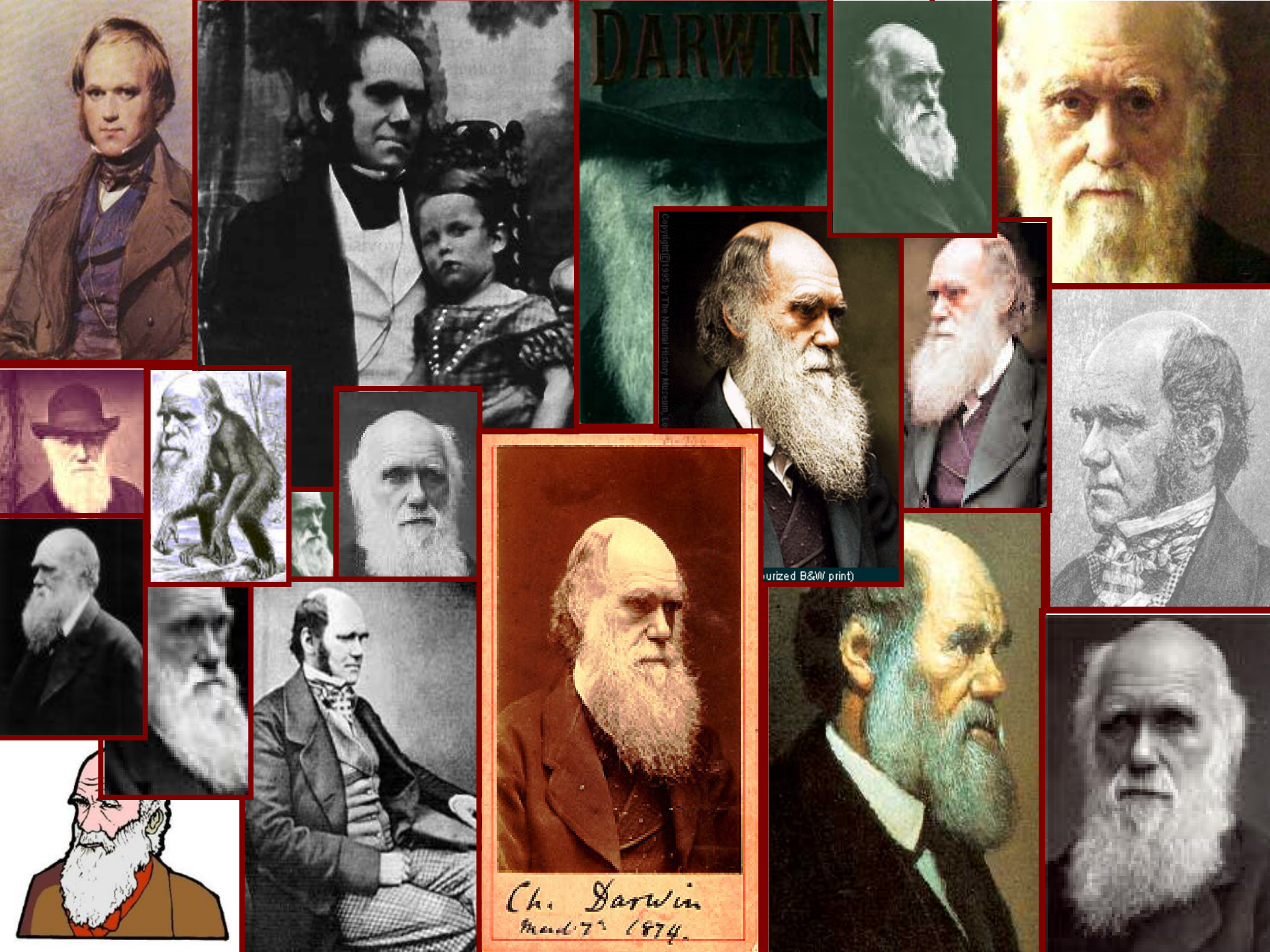
Where?

Where?

Who?

The Navarro lab, 2013




DARWIN
Ch. Darwin
Mar. 7th 1874.

What?

1. Longish Introduction. Why Genomics? Association studies? **Genome Wide Association Studies (GWAS)**
2. Candidate-Gene studies and their problems
3. Genome-Wide association studies
4. Whatever happened to personalized medicine

The Human Genome and Disease

The promise in the 90s: "In 10 years we will unravel the genetic bases of complex diseases!!" ... It was the reason under the Human Genome Project

The Human Genome and Disease

The promise in the 90s: "In 10 years we will unravel the genetic bases of complex diseases!!" ... It was the reason under the Human Genome Project

Table 1: Genomics Funding by Organization and Year, in US$ (millions)

Rank	Organization	2003	2004	2005	2006	Source
n/a	Department of Health and Human Services (DHHS) National Institutes of Health (NIH) – GENETICS[O]	$4236	$4535	$4840	$4878	[28]
1	NIH: National Cancer Institute (NCI) + National Human Genome Research Institute (NHGRI)	$562	$593	$593	$571	See Tables 2, 3 and text
2	European Commission[J]	$459	$462	$466	$468	Personal Communication, Indridi Benediktsson, September 2006
3	European Commission Matching Funds[J]	$459	$462	$466	$468	Personal Communication, Indridi Benediktsson, September 2006
4	United Kingdom Wellcome Trust[O]	$194	$194	$208	$199	[52,53]
5	Department of Energy (DOE) Office of Biological and Environmental Research[O]	$129	$152	$154	$158	[33-36]
6	National Science Foundation (NSF) Biological Sciences Directorate[O]	$124	$129	$134	$141	Personal Communication, Vernon Ross, April 2007
7	Japan Ministry of Education, Culture, Sports, Science and Technology (MEXT)[A]	$84.5	$99.2	$119	$125	Personal Communication, Kazuko Shinohara, January 2007
8	United Kingdom Biotechnology and Biological Sciences Research Council (BBSRC)[A]	$121	$127	$128	$117	Personal Communication, Clare Nixon, January 2007
9	Genome Canada[A]	$67.3	$65.7	$71.6	$106	Personal Communication, Genny Cardin, July 2006
10	China (Ministry of Science and Technology, National Natural Science Foundation of China, and Chinese Academy of Sciences)[J]	$80	$80	$80	no report	Personal Communication, Anonymous, October 2007
11	Germany Nationales Genomforschungsnetz (NGFN)[J]	$71.6	$61.4	$62.0	$64.8	Personal Communication, Uta Strasser, September 2006
12	Department of Defense (DOD) Congressional Directed Medical Research Programs (CDMRP)[O]	$102	$86.8	$53.5	$54.9	[45]
13	Cancer Research UK[A]	$34.1	$45.2	$48.1	$51.0	Personal Communication, Lynne Davies, January 2007
14	Netherlands Genomics Initiative[J]	$48.8	$18.1	$51.7	$45.8	[31,87,88]
15	South Korea Ministry of Science and Technology (MOST)[J]	$39.7	$35.0	$41.5	$44.3	Personal Communication, Jeongheui Lim, January 2007
16	United States Department of Agriculture (USDA) Agricultural Research Service[O]	$32.5	$38.8	$41.9	$43.1	Personal Communication, Peggy DelCollo and Joe Garbarino, November 2006
17	Ireland Higher Education Authority[J]	$34.1	$34.7	$34.5	$34.5	Personal Communication, Sorcha Carthy, August 2006
18	Canada Natural Sciences and Engineering Research Council (NSERC)[A]	$28.2	$30.4	$32.2	$34.1	Personal Communication, Barney Laciak, October 2007
19	Department of Homeland Security (DHS)[O]	$13.4	$25.9	$32.8	$27.2	Personal Communication, Elizabeth George, November 2006

BMC Genomics 2008, 9:472 http://www.biomedcentral.com/1471-2164/9/472

Table 1: Genomics Funding by Organization and Year, in US$ (millions) *(Continued)*

20	Japan Ministry of Agriculture, Forestry and Fisheries (MAFF)[A]	$17.9	$22.9	$21.7	$21.7	Personal Communication, Kazuko Shinohara, January 2007
21	Canadian Biotechnology Strategy (CBS)[A]	$16.1	$15.9	$15.9	$16.2	[55]
22	Howard Hughes Medical Institute (HHMI)[J]	$15.2	$13.8	$14.3	$15.6	Personal Communication, Sherry White, August 2006
23	Canada National Research Council (NRC) Genomics and Health Initiative[A]	$18.2	$15.8	$13.4	$15.3	Personal Communication, Gary Fudge, August 2007
24	Japan Ministry of Health, Labour and Welfare (MHLW)[A]	$12.8	$13.8	$18.6	$14.9	Personal Communication, Kazuko Shinohara, January 2007
25	American Cancer Society[J]	$5.95	$6.90	$5.61	$11.9	Personal Communication, Donella Wilson, July 2006
26	Spain Genoma Espana[J]	$9.98	$12.1	$13.5	$11.7	Personal Communications, Javier Montero Plata, June and September 2006
27	Australia National Health and Medical Research Council (NHMRC) *	$7.54	$6.34	$6.24	$7.65	Personal Communication, Marian Blake, July 2006
28	Department of Health and Human Services (DHHS) Centers for Disease Control and Prevention (CDC)[O]	$3.85	$4.53	$6.99	$6.95	[37,38]
29	Department of Defense (DOD) Defense Advanced Research Projects Agency (DARPA) Bio/Info/Micro Program[O]	$34.0	$13.4	$13.2	$6.5	[41-44]
30	South African Medical Research Council[J]	$1.73	$1.84	$2.04	$2.24	Personal Communication, Clive Glass, October 2007
31	Ireland Science Foundation[J]	$7.41	$7.96	$1.38	$2.24	Personal Communication, Tracy Moloney, September 2006
32	Ireland Health Research Board[J]	$1.29	$2.23	$1.50	$1.61	Personal Communication, Gillian Hastings, January 2007
33	South Africa National Research Foundation[J]	$0.658	$0.560	$0.999	no report	Personal Communication, Marna van Rooyen, October 2006
34	Japan Ministry of Economy, Trade and Industry (METI)[A]	$0	$4.51	$0	$0	Personal Communication, Kazuko Shinohara, January 2007
	TOTAL	**$2834**	**$2878**	**$2948**	**$2881**	

Genomics research funded, by organization, is shown in millions of US$ per year. The total funding values each year were determined by summing the values from each organization, with the exception of the first row (NIH – Genetics), which is described in the text. Rankings were determined by ordering the 2006 values, where the average of the three previous years was used as a substitute for 2006 values when the actual 2006 data was unavailable. The start of each fiscal year is indicated by the superscript character after each organization, where J = January 1, A = April 1, and O = October 1.

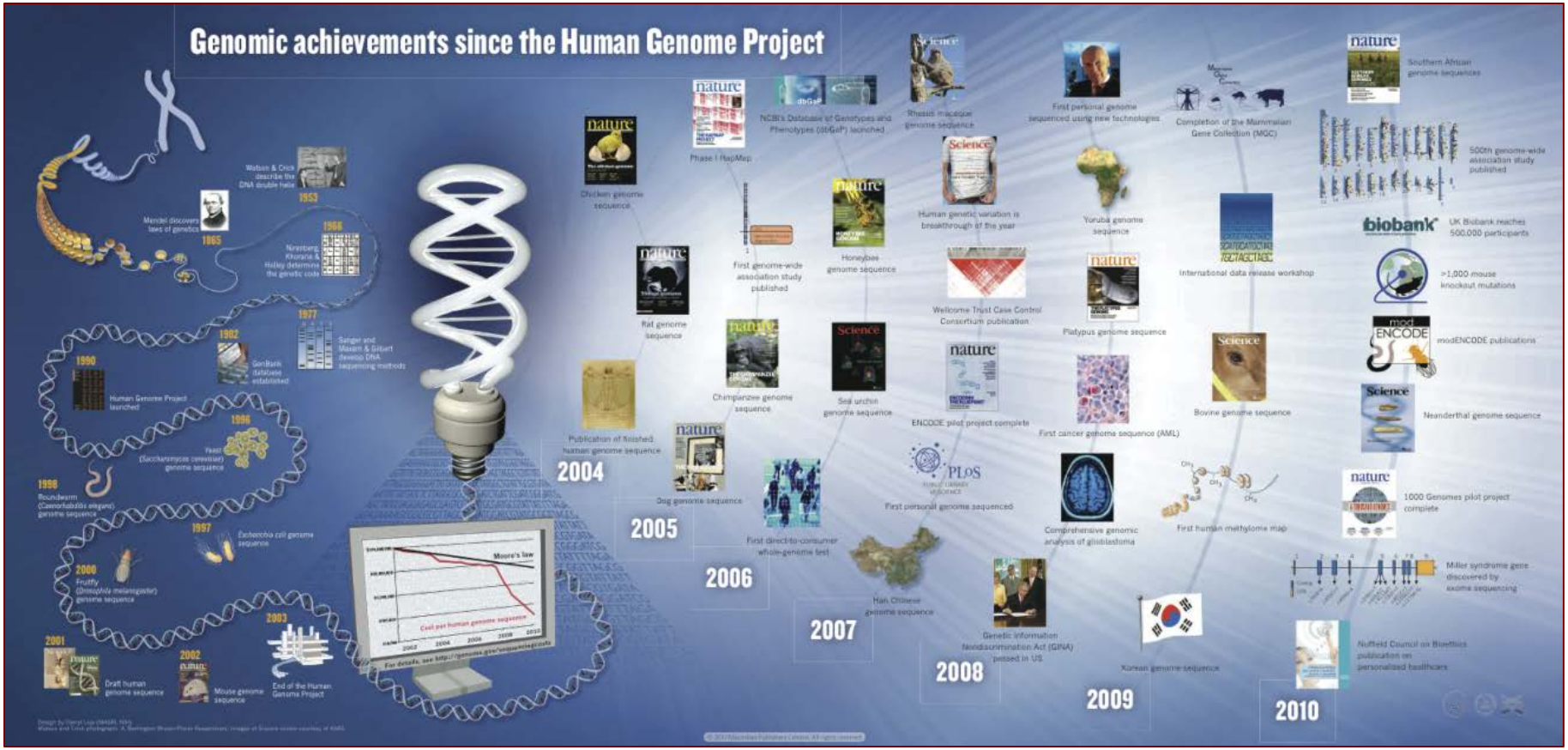
*Although the fiscal year for the Australian government begins July 1, the original data was reported by calendar year.

A deluge of developments since then

Mainly, we got 1000s of genomes...

A deluge of developments since then

The promise in the 90s: "In 10 years we will unravel the genetic bases of complex diseases!!" ... has been the motivation under many advances.

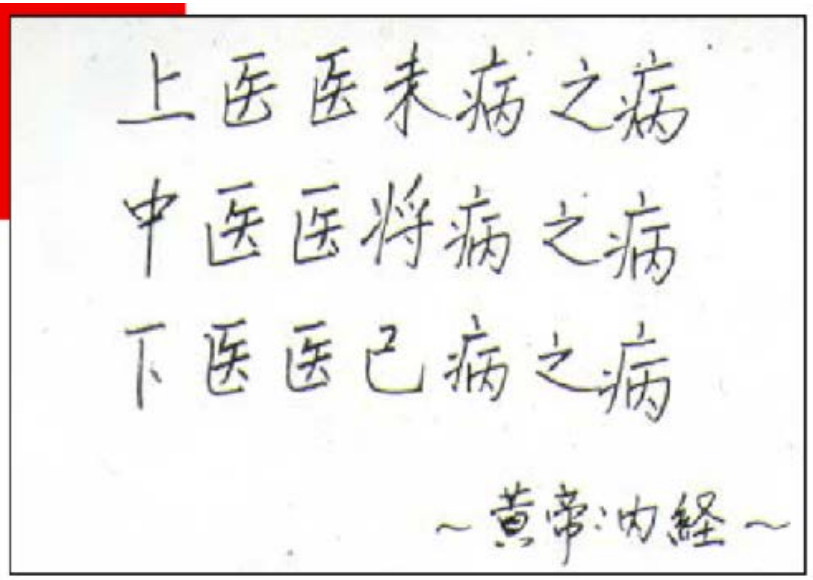


ED Green *et al. Nature* 470, 204-213 (2011) doi:10.1038/nature09764

And what about disease?

The promise in the 90s: "In 10 years we will unravel the genetic basis of complex diseases!!…"

(1) HapMap & 1000 Genomes Project	Patterns of LD / variation in many human populations
(2) Genotyping arrays	Dense coverage of human genome
(3) Large samples	Case-Control or Family Linkage



"Superior Doctors Prevent the Disease.
Mediocre Doctors Treat the Disease Before Evident.
Inferior Doctors Treat the Full Blown Disease."

-Huang Dee: Nai - Ching (2600 B.C. 1st Chinese Medical Text

And what about disease?

The promise in the 90s: "In 10 years we will unravel the genetic basis of complex diseases!!…"

(1) HapMap & 1000 Genomes Project	Patterns of LD / variation in many human populations
(2) Genotyping arrays	Dense coverage of human genome
(3) Large samples	Case-Control or Family Linkage

- Prevent disease from occurring
- Identify the cause of the disease
- Treat the cause of the disease rather than the symptoms
- Genomics may identify the cause of disease ("**All medicine may become pediatrics**" Paul Wise, Professor of Pediatrics, Stanford Medical School, 2008)
- Effects of environment, accidents, aging, penetrance …
- Health care costs can be greatly reduced if
  - invests in preventive medicine
  - one targets the cause of disease rather than symptoms

And what about disease?

The promise in the 90s: "In 10 years we will unravel the genetic basis of complex diseases!!…"

(1) HapMap & 100 Genomes Project – Patterns of LD / variation in many
s

(2) Genotyping ar f human genome

(3) Large samples amily Linkage

And what about disease?

The promise in the 90s: "In 10 years we will unravel the genetic basis of complex diseases!!…"

(1) HapMap & 1000 Genomes Project — Patterns of LD / variation in many human populations

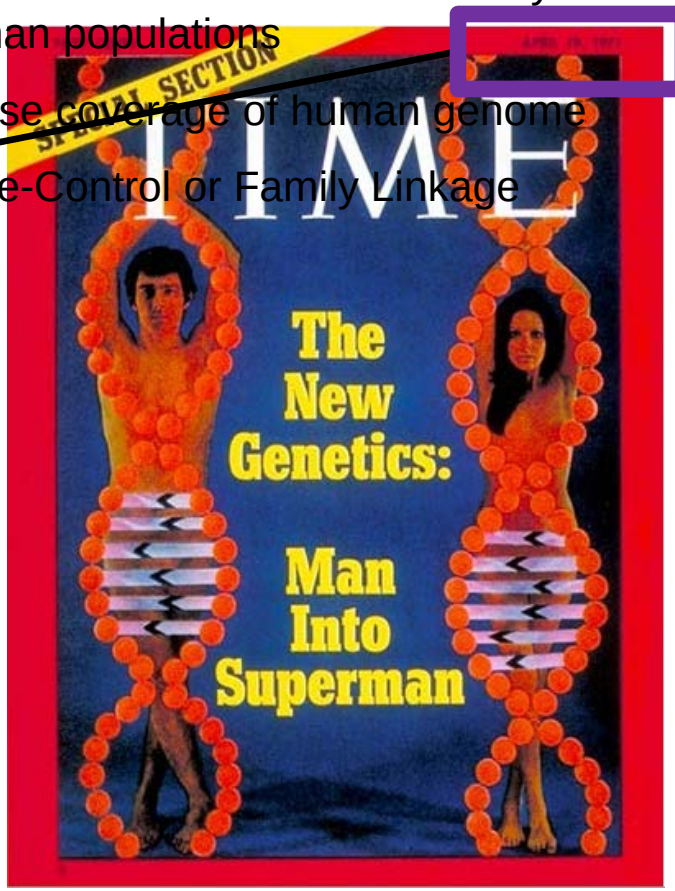
(2) Genotyping arrays — Dense coverage of human genome

(3) Large samples — Case-Control or Family Linkage

Apr. 19, 1971

The remarkable advances in molecular biology during the past two decades have given man an understanding of the basic processes that shape his life and have placed within the realm of possibility medical achievements undreamed of a scant few years ago.

The way is being opened not only for permanent cures of genetic disease but also for drastic changes in man's genetic makeup.

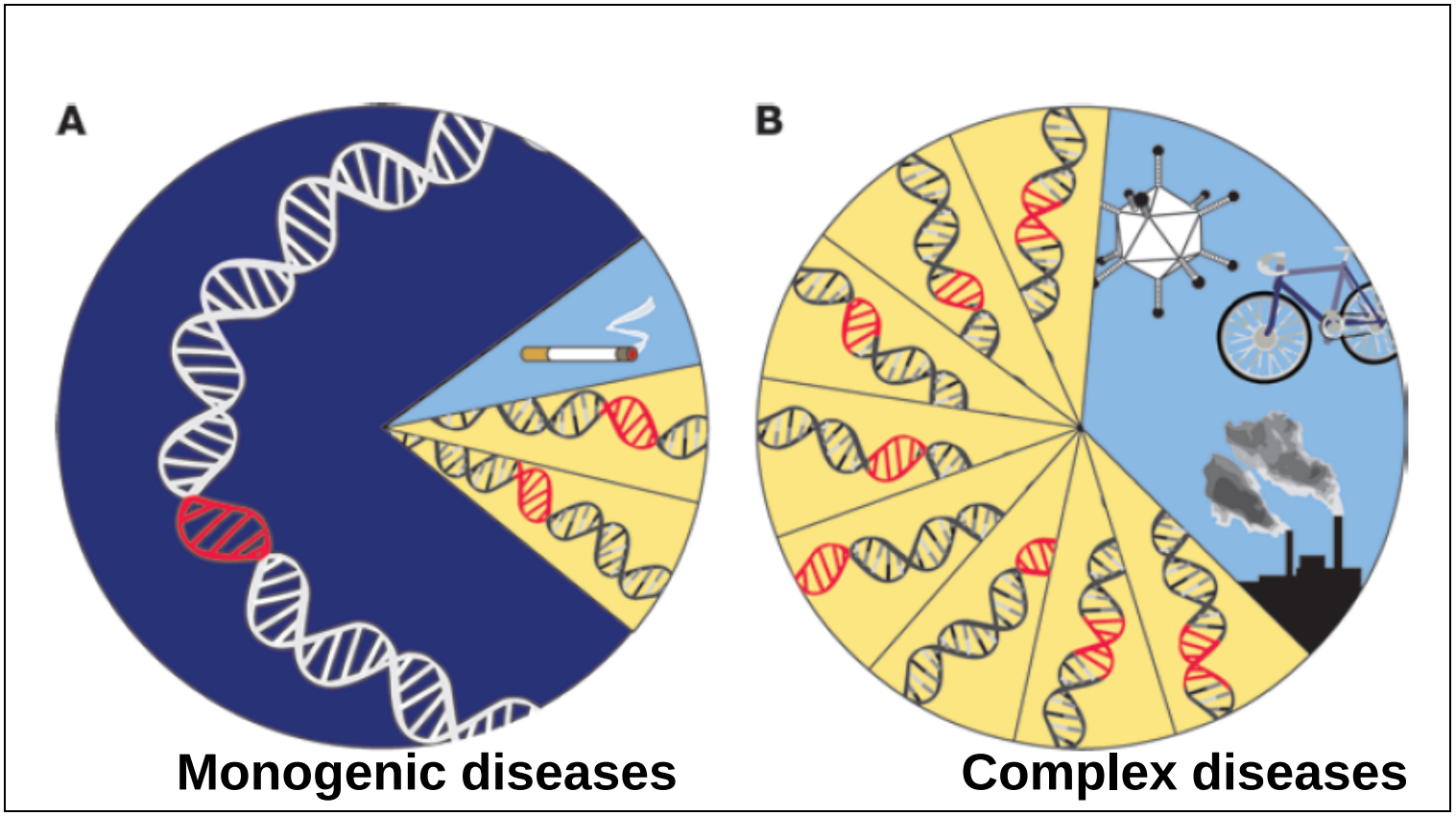


http://www.time.com/time/covers/0,16641,19710419,00.html

Buchanan AV et al, Int J Epidemiol, June 2006 (35: 593-596)

COMPLEX GENETIC ARCHITECTURE OF COMPLEX DISEASES

- ✓ *Polygenic + Environmental*
- ✓ *High heritability (h^2 > 50%) and Familial aggregation (λ_s >> 1)*
- ✓ *But lack of Mendelian inheritance (lots of sporadic cases)*



Manolio et al. (2008) *J Clin Invest.* 118(5):1590

COMPLEX GENETIC ARCHITECTURE OF COMPLEX DISEASES

The promise in the 90s: "In 10 years we will unravel the genetic bases of complex diseases!!". Everything seemed to go well...

... and people were doing this using diverse approaches.

✓ Linkage mapping:

Select an informative family.

More than 100,000 papers in 30 years (mostly in OMIM)

✓ Candidate genes:

Select a few variants in one or a few genes.

More than 84000 papers in 20 years (see the GAD database)

✓ Genome-Wide Association Studies:

Hypothesis-free approach. One looks as as many variables as possible

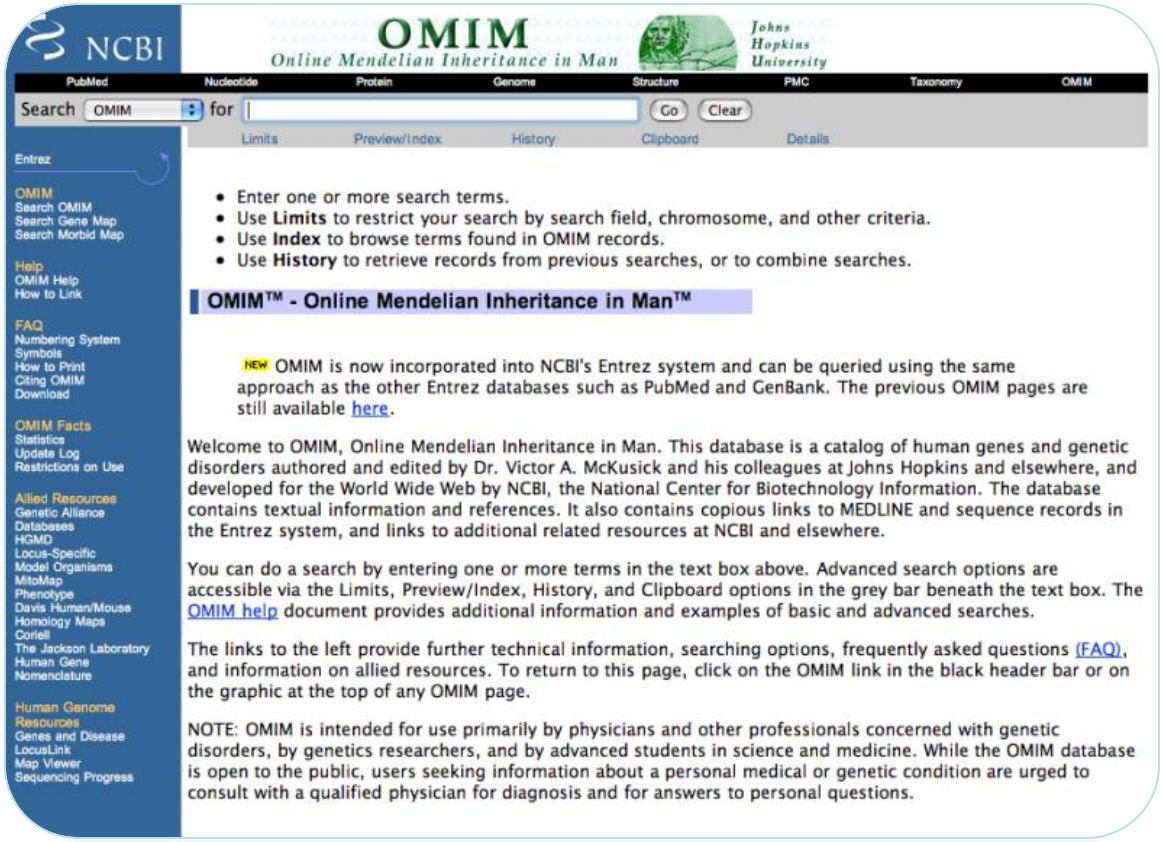
More than 1200 papers (see www.genome.gov)

THE GENETIC ARCHITECTURE OF MENDELIAN DISEASES

The promise in the 90s: "In 10 years we will unravel the genetic bases of complex diseases!!". Everything seemed to go well...

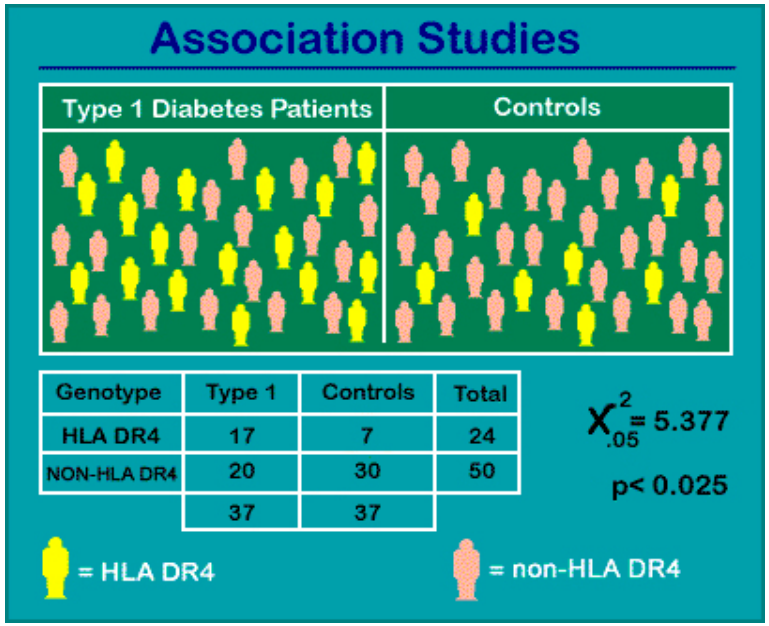
OMIM Home Page

http:/ / www.ncbi.nlm.nih.gov/ omim/

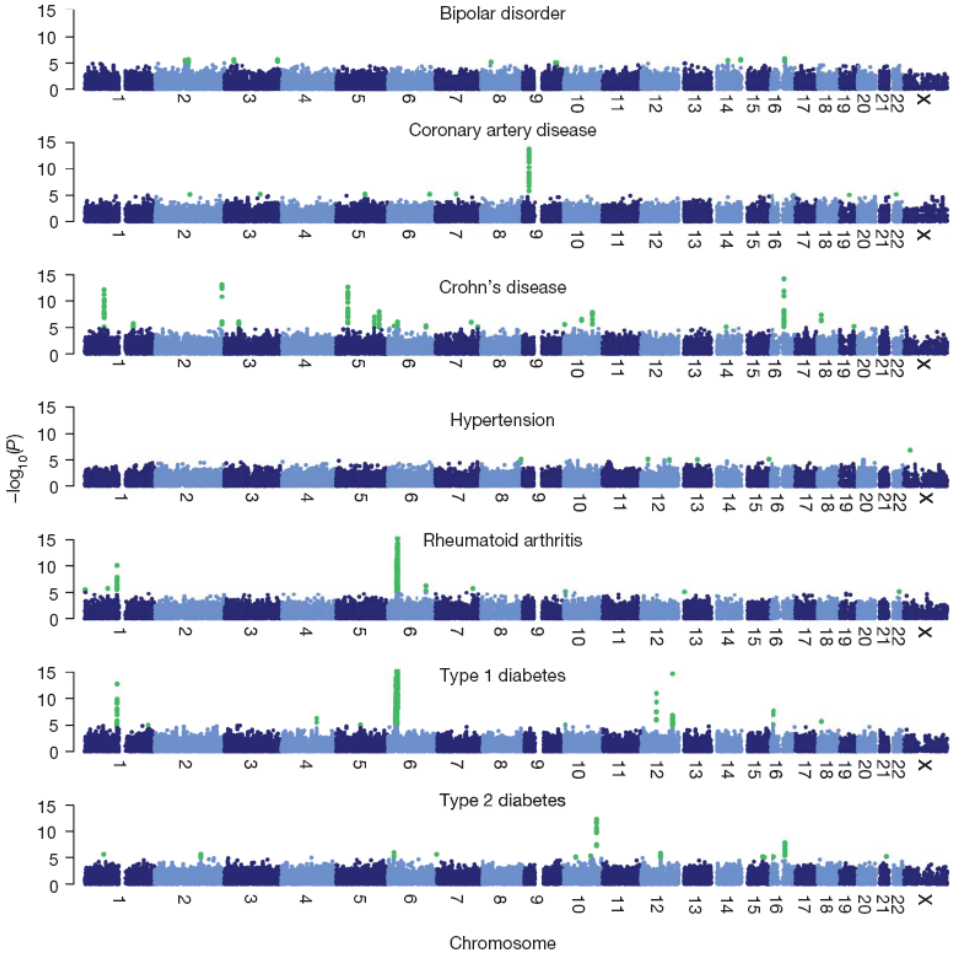


COMPLEX GENETIC ARCHITECTURE OF COMPLEX DISEASES

The promise in the 90s: "In 10 years we will unravel the genetic bases of complex diseases!!" ... And it seemed easy to fulfil...



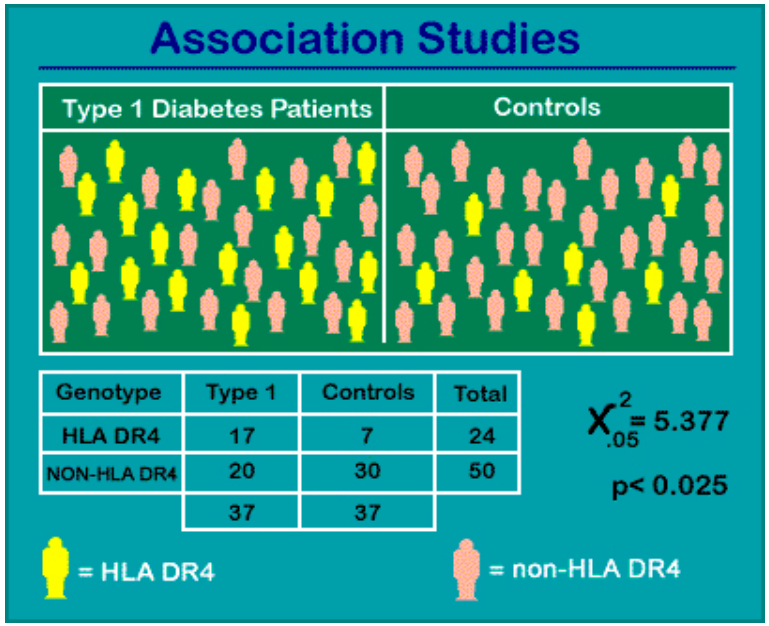
Odds Ratio: 3.6
95% CI = 1.3 to 10.4



The Wellcome Trust Case Control Consortium

Many Candidate Gene Studies

The promise in the 90s: "In 10 years we will unravel the genetic bases of complex diseases!!" ... And it seemed easy to fulfil...



Odds Ratio: 3.6
95% CI = 1.3 to 10.4

The Genetic Association Database (GAD)

Views:
Disease
Gene View
CH-SNP-HapMap
Reference
Environmental Factor Gene Interaction
All

The Genetic Association Database is an archive of human genetic association studies of complex diseases and disorders. The goal of this database is to allow the user to rapidly identify medically relevant polymorphism from the large volume of polymorphism and mutational data, in the context of standardized nomenclature.

The data is from published scientific papers. Study data is recorded in the context of official human gene nomenclature with additional molecular reference numbers and links. It is gene centered. That is, each record is a record of a gene or marker. If a study investigated 6 genes for a particular disorder, there will be 6 records.

Next 25 All View Search for DIABETES Record found: 500

	Last Update	Checked	CDC Index	1 - GAD 2 - CDC	Year	Assoc? Y or N	Gene Symbol	OMIM	Gene Expert	Broad Phenotype (Disease)	Disease Expert	MeSH Disease Terms	Disease Class	Chr	Ch-Band
view	13-APR-07		17626	2	2005		CTLA4	123890	Cx	diabetes, type 1			IMMUNE	2	2q33
view	13-APR-07		16890	2	2005		GCK	138079	Cx	diabetes, type 2			METABOLIC	7	7p15.3-p15.1
view	13-APR-07		8711	2	2006		APOA5	606368		myocardial infarct; diabetes,			CANCER	11	11q23
view	13-APR-07		16322	2	2005	N	DIO2	601413		body mass; diabetes, type 2; g		Diabetes Mellitus, Type 2|Insu	METABOLIC	14	14q24.2-q24.3
view	13-APR-07		11253	2	2005		CTLA4	123890	Cx	diabetes, type 1; thyroid dise			IMMUNE	2	2q33
view	13-APR-07		15560	2	2005		CAPN10	605286	Cx	diabetes, type 2; nephropathy,			METABOLIC	2	2q37.3

Candidate Gene Association Studies

The promise in the 90s: "In 10 years we will unravel the genetic bases of complex diseases!!". Everything seemed to go well...

(1) HapMap & 1000 Genomes Project	Patterns of LD / variation in many human populations
(2) Genotyping arrays	Dense coverage of human genome
(3) Large samples	Case-Control or Family Linkage

...However…

Statistical Problems	Sample size = Power
	Multiple testing
Genetic problems	Ascertainment bias
	Locus & Allelic heterogeneity
	Population substructure

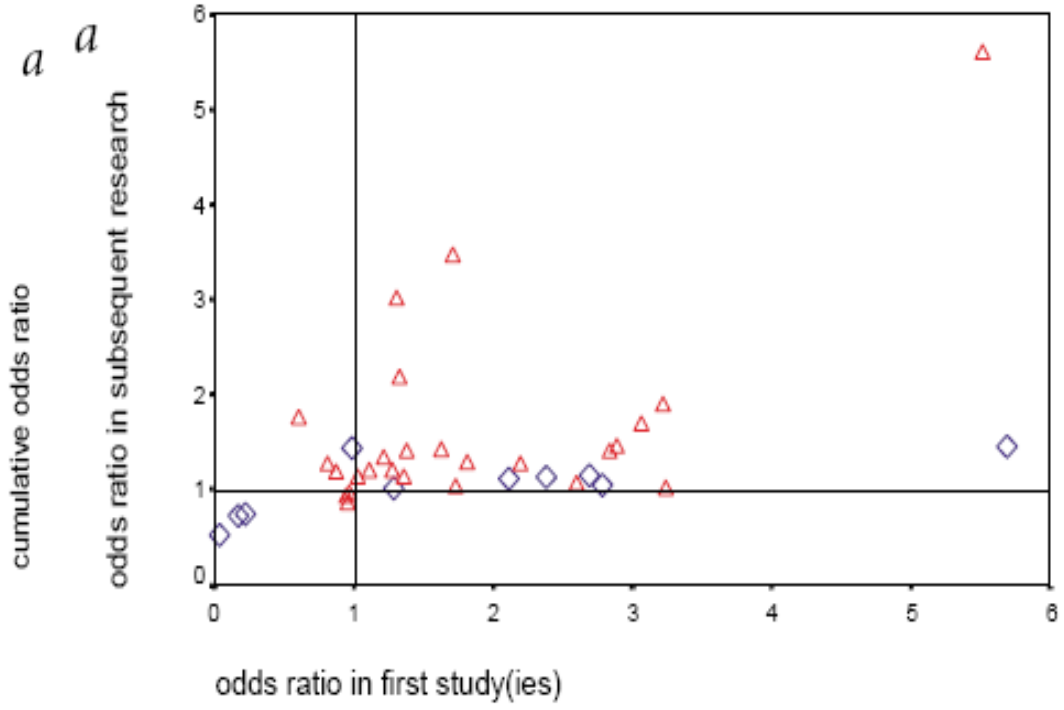
As a consequence, there were (are?) LACK OF REPLICATION problems!!

Candidate Gene Association Studies and Lack of Replication

Replication validity of genetic association studies

John P.A. Ioannidis[1–3], Evangelia E. Ntzani[1], Thomas A. Trikalinos[1] & Despina G. Contopoulos-Ioannidis[1,4]

nature genetics • volume 29 • november 2001



So the BULK of old candidate gene associations are not good

More than 100,000 papers registered in the GAD…but…
the field is dominated by false positives (getting better recently)

The precision of many old papers is similar to this:

"I have nothing to offer but blood, toil, tears and sweat."

Mahatma Gandhi

Candidate Gene Association Studies and Lack of Replication

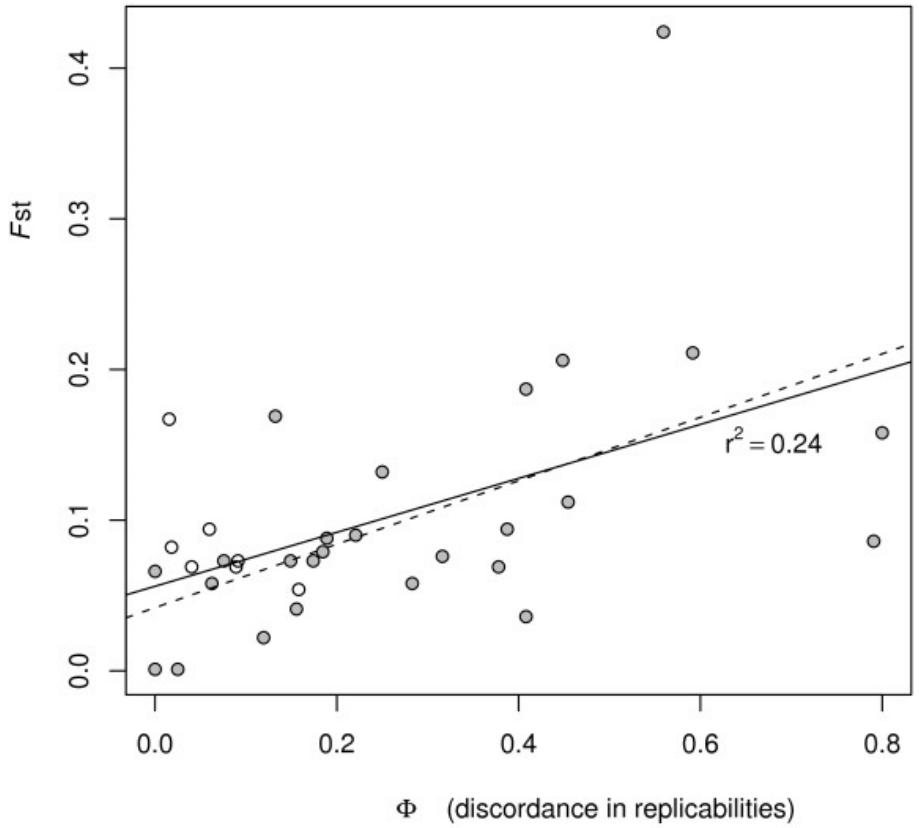
But not everything was bad...

What about focusing on patters of lack of replicability rather than on good replications?

Candidate Gene Association Studies and Lack of Replication

But not everything was bad...

Increased inconsistency of replicability (ϕ) with increased F_{ST} (i.e. higher consistency between Europe and Asia with lower genetic distance)



Correlation between discordance in replicability and F_{ST} for the 37 associations from the Continental Set.

Marigorta *et al. BMC Genomics* 2011, **12**:55
http://www.biomedcentral.com/1471-2164/12/55

BMC Genomics

RESEARCH ARTICLE Open Access

Recent human evolution has shaped geographical differences in susceptibility to disease

Urko M Marigorta[1], Oscar Lao[2], Ferran Casals[1], Francesc Calafell[1,3], Carlos Morcillo-Suárez[1,4], Rui Faria[1,5], Elena Bosch[1,3], François Serra[6], Jaume Bertranpetit[1,3], Hernán Dopazo[6], Arcadi Navarro[1,4,7*]

Abstract

Background: Searching for associations between genetic variants and complex diseases has been a very active area of research for over two decades. More than 51,000 potential associations have been studied and published, a figure that keeps increasing, especially with the recent explosion of array-based Genome-Wide Association Studies. Even if the number of true associations described so far is high, many of the putative risk variants detected so far have failed to be consistently replicated and are widely considered false positives. Here, we focus on the world-wide patterns of replicability of published association studies.

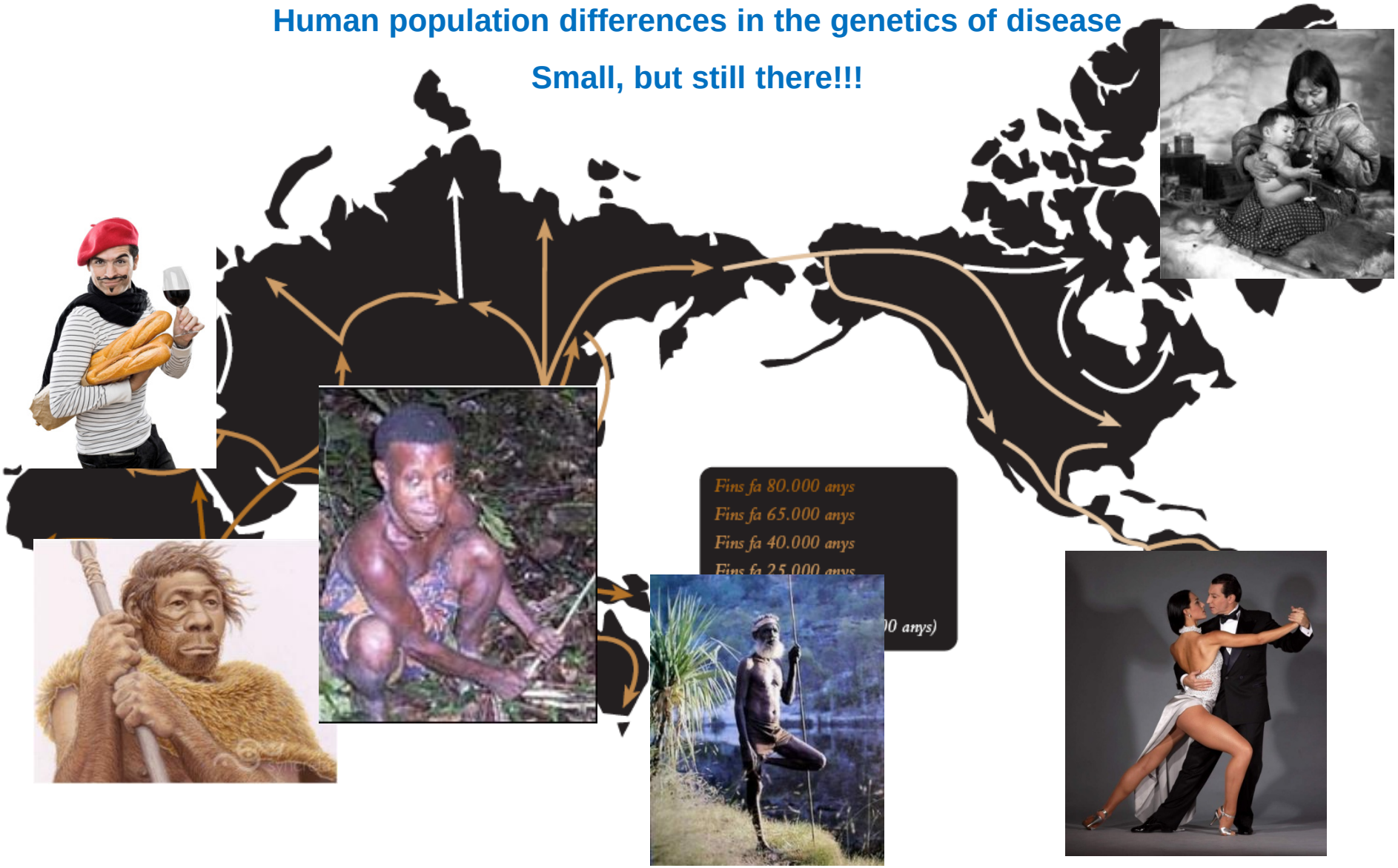
Results: We report three main findings. First, contrary to previous results, genes associated to complex diseases present lower degrees of genetic differentiation among human populations than average genome-wide levels. Second, also contrary to previous results, the differences in replicability of disease associated-loci between Europeans and East Asians are highly correlated with genetic differentiation between these populations. Finally, highly replicated genes present increased levels of high-frequency derived alleles in European and Asian populations when compared to African populations.

Conclusions: Our findings highlight the heterogeneous nature of the genetic etiology of complex disease, confirm the importance of the recent evolutionary history of our species in current patterns of disease susceptibility and could cast doubts on the status as false positives of some associations that have failed to replicate across populations.

WORLD COLONIZATION BY HUMANS (last ~100 kyears)

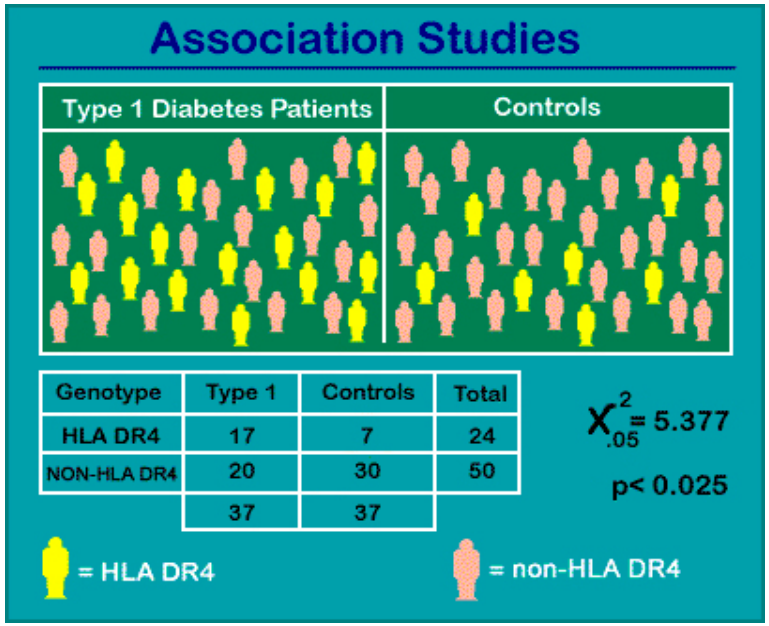
Human population differences in the genetics of disease

Small, but still there!!!

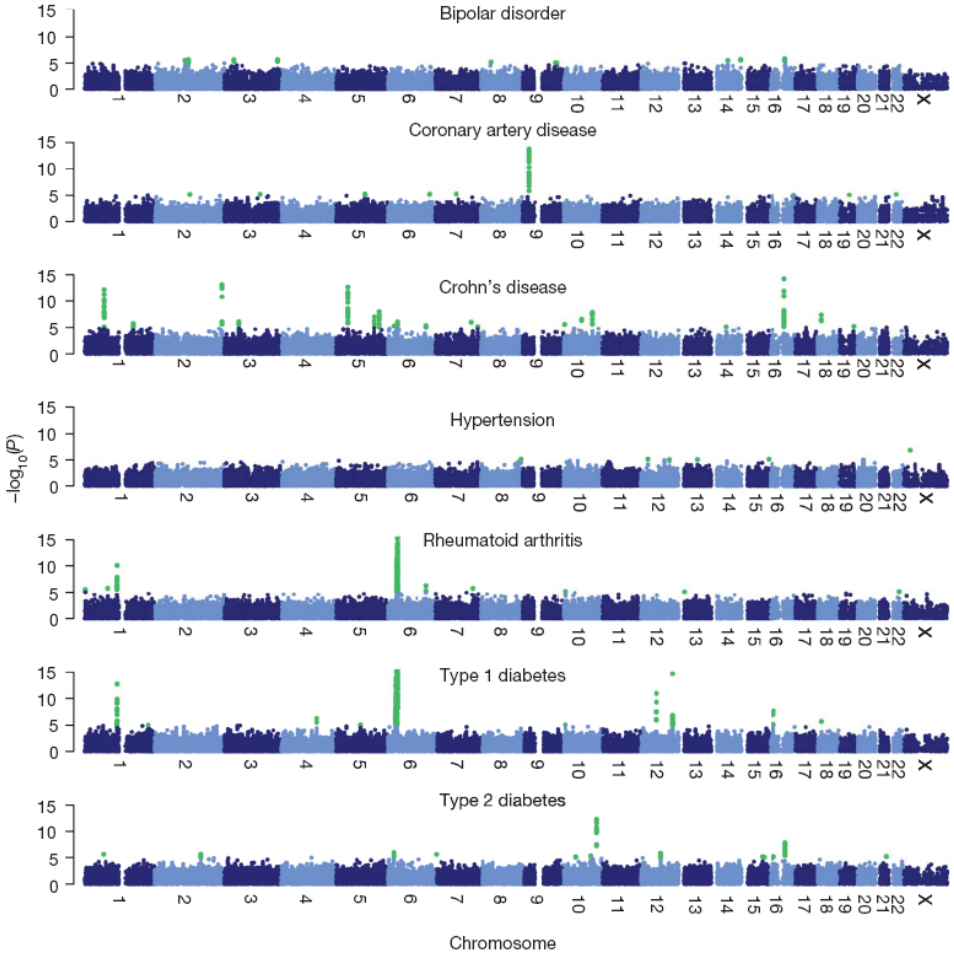


COMPLEX GENETIC ARCHITECTURE OF COMPLEX DISEASES

The promise in the 90s: "In 10 years we will unravel the genetic bases of complex diseases!!" ... And it seemed easy to fulfil...



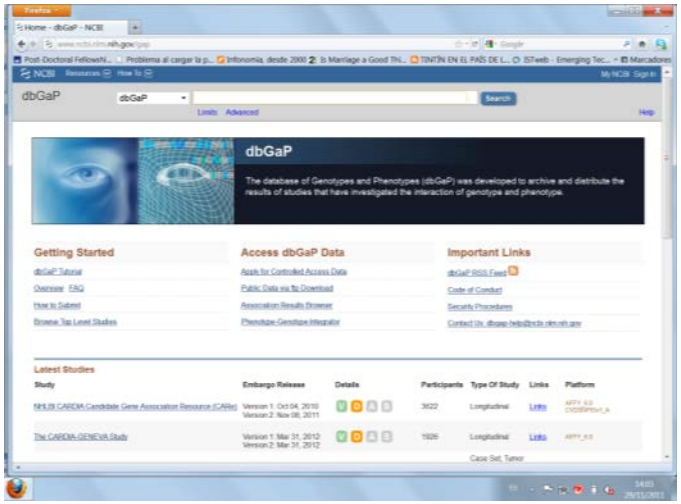
Odds Ratio: 3.6
95% CI = 1.3 to 10.4

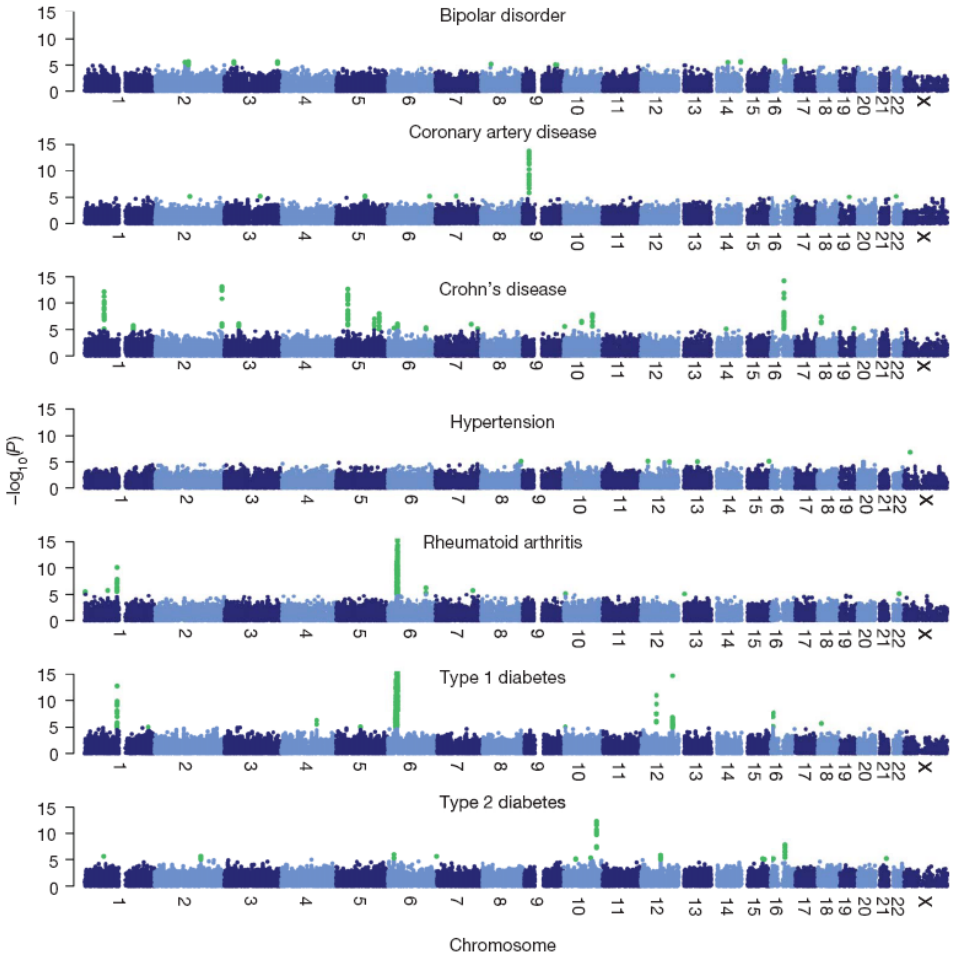


The Wellcome Trust Case Control Consortium

Plus many GWAS

The promise in the 90s: "In 10 years we will unravel the genetic bases of complex diseases!!" ... And it seemed easy to fulfil...

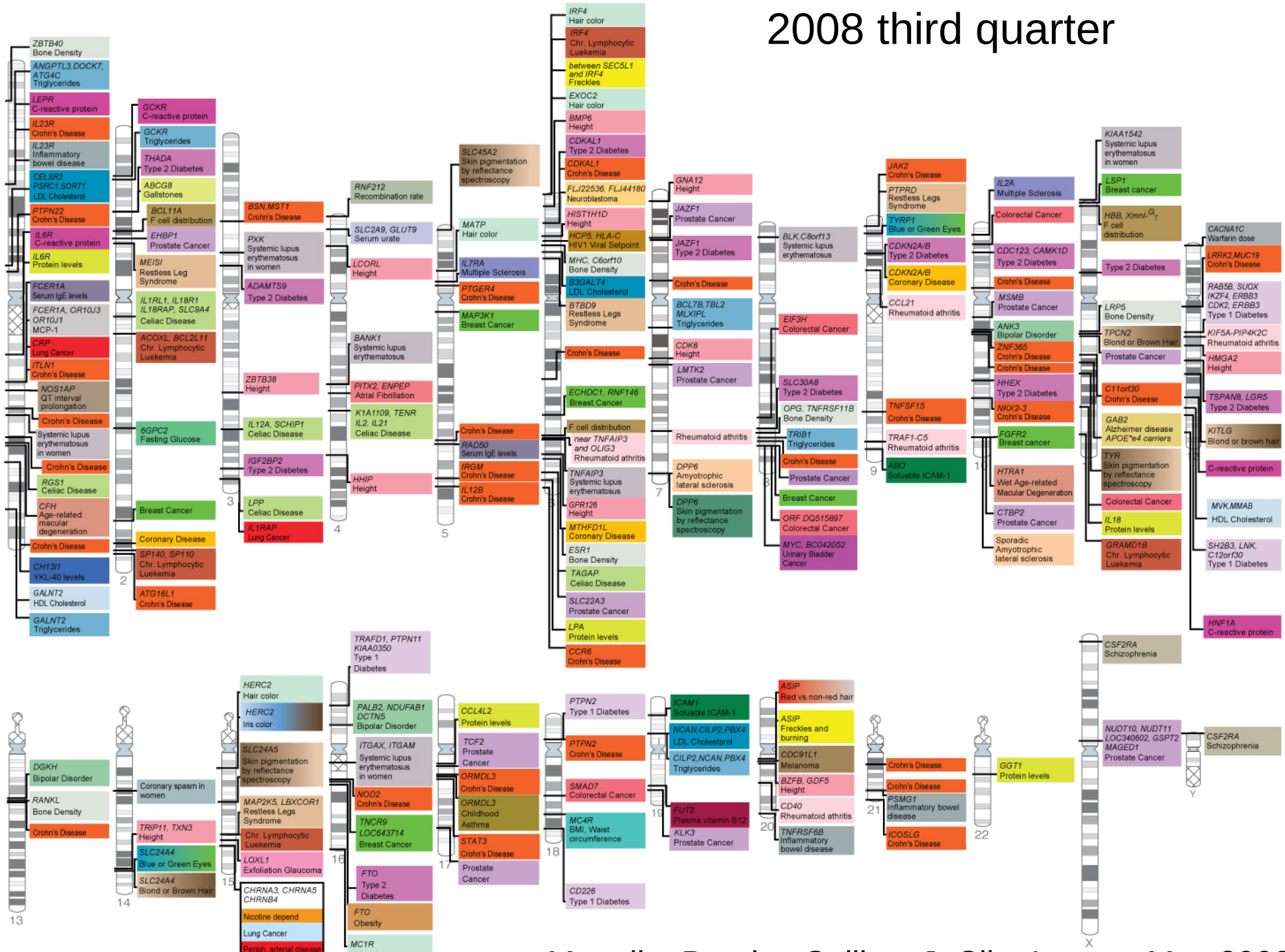




WTCCC
Home

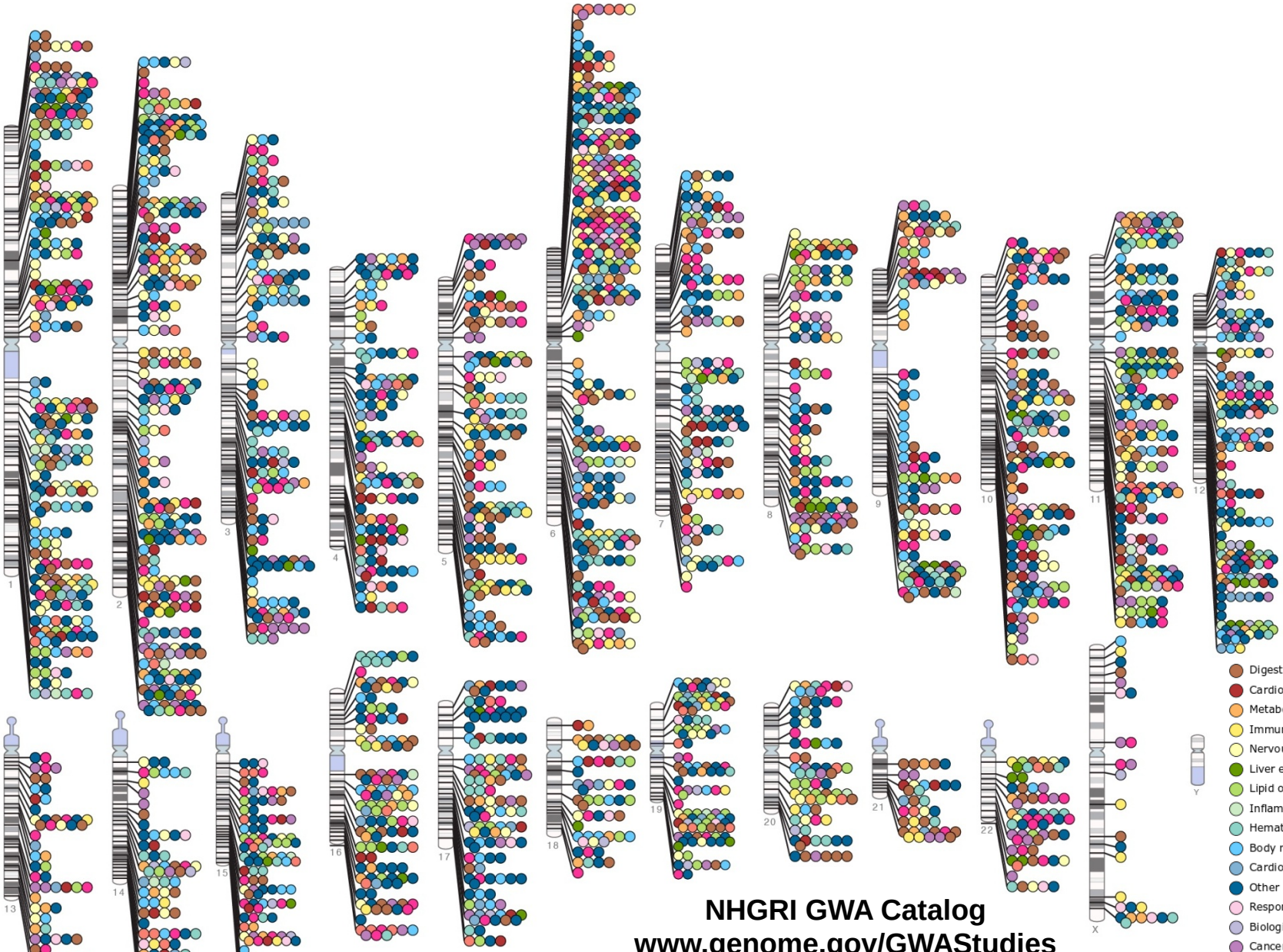
The Wellcome Trust Case Control Consortium

2008 third quarter



Manolio, Brooks, Collins, J. Clin. Invest., May 2008

Published Genome-Wide Associations through 12/2012
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

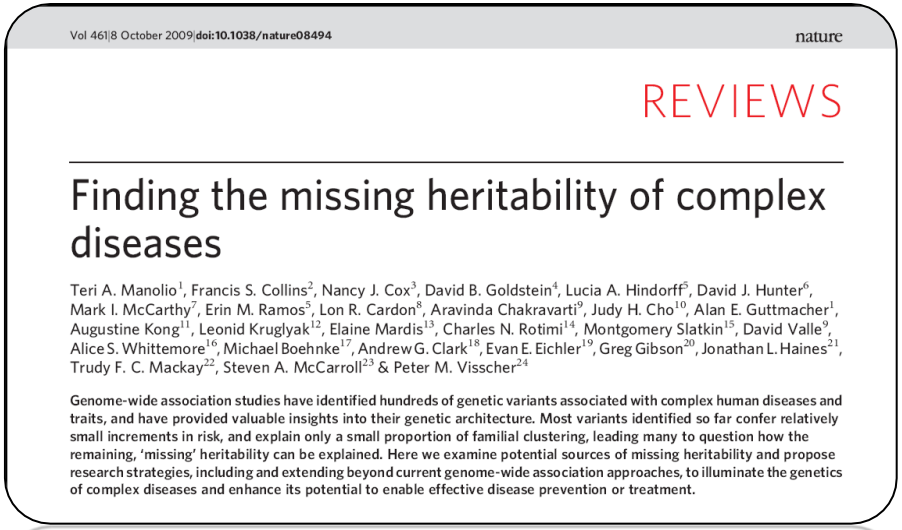

Abdominal aortic aneurysm
Acute lymphoblastic leukemia
Adhesion molecules
Adiponectin levels
Age-related macular degeneration
AIDS progression
Alcohol dependence
Alopecia areata
Alzheimer disease
Amyloid A levels
Amyotrophic lateral sclerosis
Angiotensin-converting enzyme activity
Ankylosing spondylitis
Arterial stiffness
Asparagus anosmia
Asthma
Atherosclerosis in HIV
Atrial fibrillation
Attention deficit hyperactivity disorder
Autism
Basal cell cancer
Behcet's disease
Bipolar disorder
Biliary atresia
Bilirubin
Bitter taste response
Birth weight
Bladder cancer
Bleomycin sensitivity
Blond or brown hair
Blood pressure
Blue or green eyes
BMI, waist circumference
Bone density
Breast cancer
Butyrylcholinesterase levels
C-reactive protein
Calcium levels
Cardiac structure/function
Cardiovascular risk factors
Carnitine levels
Carotenoid/tocopherol levels
Carotid atherosclerosis
Celiac disease
Celiac disease and rheumatoid arthritis
Cerebral atrophy measures
Chronic lymphocytic leukemia
Chronic myeloid leukemia
Cleft lip/palate
Coffee consumption
Cognitive function
Conduct disorder
Colorectal cancer
Corneal thickness
Coronary disease
Cortical thickness
Creutzfeldt-Jakob disease
Crohn's disease
Crohn's disease and celiac disease
Cutaneous nevi
Cystic fibrosis severity
Dermatitis
DHEA-s levels
Diabetic retinopathy
Dilated cardiomyopathy
Drug-induced liver injury
Drug-induced liver injury (amoxicillin-clavulanate)
Endometrial cancer
Endometriosis
Eosinophil count
Eosinophilic esophagitis
Epirubicin-induced leukopenia
Erectile dysfunction and prostate cancer treatment
Erythrocyte parameters
Esophageal cancer
Essential tremor
Exfoliation glaucoma
Eye color traits
F cell distribution
Fibrinogen levels
Folate pathway vitamins
Follicular lymphoma
Fuch's corneal dystrophy
Freckles and burning
Gallstones
Gastric cancer
Glioma
Glycemic traits
Graves disease
Hair color
Hair morphology
Handedness in dyslexia
HDL cholesterol
Heart failure
Heart rate
Height
Hemostasis parameters
Hepatic steatosis
Hepatitis
Hepatitis B vaccine response
Hepatocellular carcinoma
Hirschsprung's disease
HIV-1 control
Hodgkin's lymphoma
Homocysteine levels
HPV seropositivity
Hypospadias
Idiopathic pulmonary fibrosis
IFN-related cytopeni
IgA levels
IgE levels
Inflammatory bowel disease
Insulin-like growth factors
Intracranial aneurysm
Iris color
Iron status markers
Ischemic stroke
Juvenile idiopathic arthritis
Keloid
Kidney stones
LDL cholesterol
Leprosy
Leptin receptor levels
Liver enzymes
Longevity
LP (a) levels
LpPLA(2) activity and mass
Lung cancer
Magnesium levels
Major mood disorders
Malaria
Male pattern baldness
Mammographic density
Matrix metalloproteinase levels
MCP-1
Melanoma
Menarche & menopause
Meningioma
Meningococcal disease
Metabolic syndrome
Migraine
Moyamoya disease
Multiple sclerosis
Myeloproliferative neoplasms
Myopia (pathological)
N-glycan levels
Narcolepsy
Nasopharyngeal cancer
Natriuretic peptide levels
Neuroblastoma
Nicotine dependence
Obesity
Open angle glaucoma
Open personality
Optic disc parameters
Osteoarthritis
Osteoporosis
Otosclerosis
Other metabolic traits
Ovarian cancer
Pancreatic cancer
Pain
Paget's disease
Panic disorder
Parkinson's disease
Periodontitis
Peripheral arterial disease
Personality dimensions
Phosphatidylcholine levels
Phosphorus levels
Photic sneeze
Phytosterol levels
Platelet count
Polycystic ovary syndrome
Primary biliary cirrhosis
Primary sclerosing cholangitis
PR interval
Progranulin levels
Progressive supranuclear palsy
Prostate cancer
Protein levels
PSA levels
Psoriasis
Psoriatic arthritis
Pulmonary funct. COPD
QRS interval
QT interval
Quantitative traits
Recombination rate
Red vs.non-red hair
Refractive error
Renal cell carcinoma
Renal function
Response to antidepressants
Response to antipsychotic therapy
Response to carbamazepine
Response to clopidogrel therapy
Response to hepatitis C treat
Response to interferon beta therapy
Response to metaformin
Response to statin therapy
Restless legs syndrome
Retinal vascular caliber
Retinol levels
Rheumatoid arthritis
Ribavirin-induced anemia
Schizophrenia
Serum metabolites
Skin pigmentation
Smoking behavior
Speech perception
Sphingolipid levels
Statin-induced myopathy
Stevens-Johnson syndrome
Stroke
Sudden cardiac arrest
Suicide attempts
Systemic lupus erythematosus
Systemic sclerosis
T-tau levels
Tau AB1-42 levels
Telomere length
Testicular germ cell tumor
Thyroid cancer
Thyroid volume
Tooth development
Total cholesterol
Triglycerides
Tuberculosis
Type 1 diabetes
Type 2 diabetes
Ulcerative colitis
Urate
Urinary albumin excretion
Urinary metabolites
Uterine fibroids
Venous thromboembolism
Ventricular conduction
VEGF levels
Vertical cup-disc ratio
Vitamin B12 levels
Vitamin D insuffiency
Vitamin E levels
Vitiligo
Warfarin dose
Weight
White cell count
White matter hyperintensity
YKL-40 levels

Problem. Where did the heritability go?

The missing heritability problem

Nature Feature Nov 2008

Manolio et al Nature October 2009



Problem. Where did the heritability go?

The missing heritability problem

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

Disease	Number of loci	Proportion of heritability explained
Age-related macular degeneration[72]	5	50%
Crohn's disease[21]	32	20%
Systemic lupus erythematosus[73]	6	15%
Type 2 diabetes[74]	18	6%
HDL cholesterol[75]	7	5.2%
Height[15]	40	5%
Early onset myocardial infarction[76]	9	2.8%
Fasting glucose[77]	4	1.5%

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

Problem. Where did the heritability go?

Consequences of the "missing heritability" problem

Manolio Review Oct 2009

Nature Review Nov 2008

Part of the heritability has been discovered…but most remains "missing"!

Does this challenge the CV/CD paradigm?
Are they rare variants? Are GWAS results artifacts?

How to explain all this?

Think Complex Diseases:

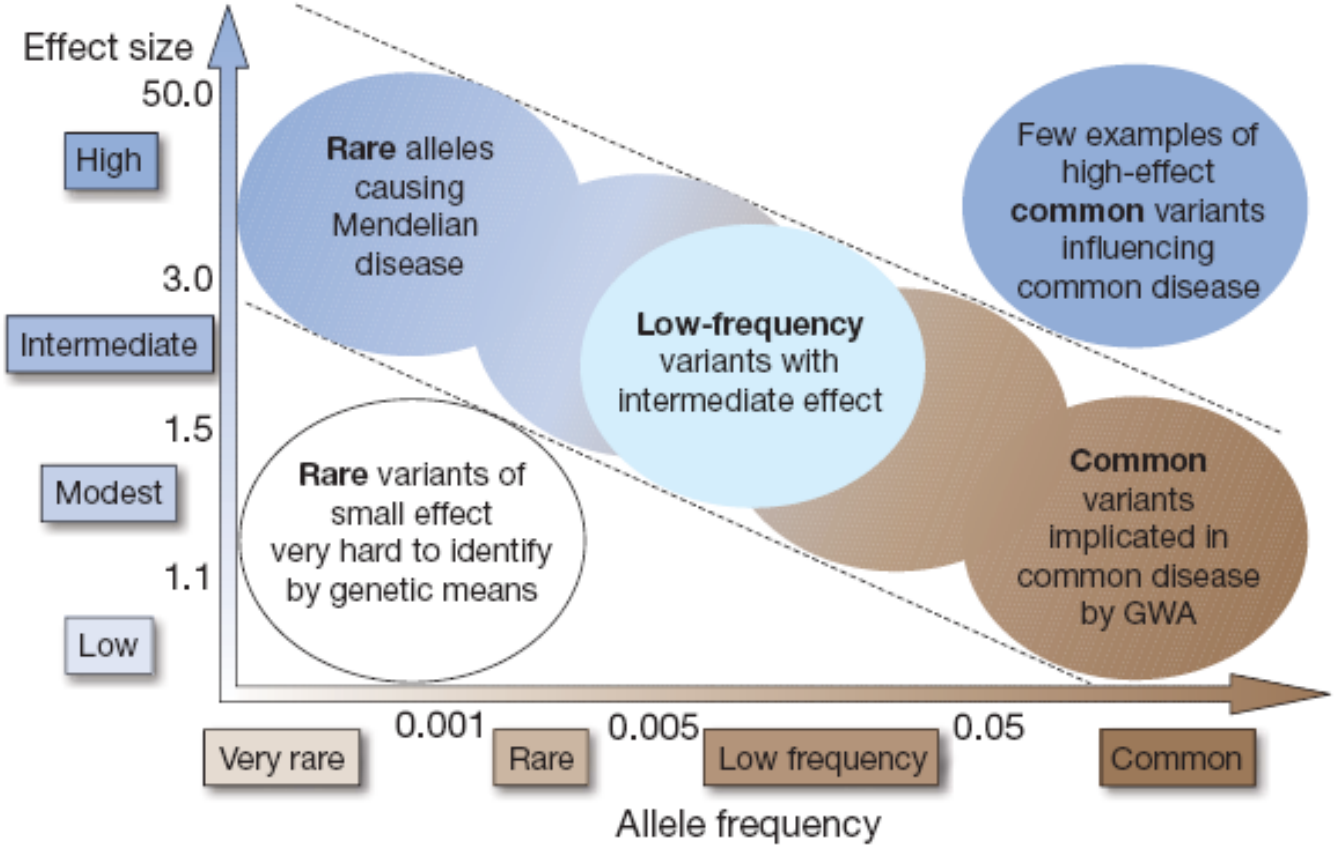


Figure 1 | Feasibility of identifying genetic variants by risk allele frequency and strength of genetic effect (odds ratio). Most emphasis and interest lies in identifying associations with characteristics shown within diagonal dotted lines. Adapted from ref. 42.

How to explain all this?

Rare Variants, rare CNVs, epigenetics….?

... but the signal may be there!! Heritability may be **Hidden, not Missing**

ANALYSIS

nature genetics

Common SNPs explain a large proportion of the heritability for human height

Jian Yang[1], Beben Benyamin[1], Brian P McEvoy[1], Scott Gordon[1], Anjali K Henders[1], Dale R Nyholt[1], Pamela A Madden[2], Andrew C Heath[2], Nicholas G Martin[1], Grant W Montgomery[1], Michael E Goddard[3] & Peter M Visscher[1]

SNPs discovered by genome-wide association studies (C
account for only a small fraction of the genetic variatio
complex traits in human populations. Where is the rema
heritability? We estimated the proportion of variance fo
human height explained by 294,831 SNPs genotyped on
3,925 unrelated individuals using a linear model analysi
validated the estimation method with simulations based
the observed genotype data. We show that 45% of varia
can be explained by considering all SNPs simultaneousl
most of the heritability is not missing but has not previo
been detected because the individual effects are too sm
to pass stringent significance tests. We provide evidenc
that the remaining heritability is due to incomplete link
disequilibrium between causal variants and genotyped
exacerbated by causal variants having lower minor allel
frequency than the SNPs explored to date.

ANALYSIS

nature genetics

Genome partitioning of genetic variation for complex traits using common SNPs

Jian Yang[1]*, Teri A Manolio[2], Louis R Pasquale[3], Eric Boerwinkle[4], Neil Caporaso[5], Julie M Cunningham[6], Mariza de Andrade[7], Bjarke Feenstra[8], Eleanor Feingold[9], M Geoffrey Hayes[10], William G Hill[11], Maria Teresa Landi[12], Alvaro Alonso[13], Guillaume Lettre[14], Peng Lin[15], Hua Ling[16], William Lowe[17], Rasika A Mathias[18], Mads Melbye[8], Elizabeth Pugh[16], Marilyn C Cornelis[19], Bruce S Weir[20], Michael E Goddard[21,22] & Peter M Visscher[1]

From almost no heritability to a lot in a 3 years!

From 2009 to 2012 (Manolio et al to Visscher et al)

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

Disease	Number of loci	Proportion of heritability explained
Age-related macular degeneration[72]	5	50%
Crohn's disease[21]	32	20%
Systemic lupus erythematosus[73]	6	15%
Type 2 diabetes[74]	18	6%
HDL cholesterol[75]	7	5.2%
Height[15]	40	5%
Early onset myocardial infarction[76]	9	2.8%
Fasting glucose[77]	4	1.5%

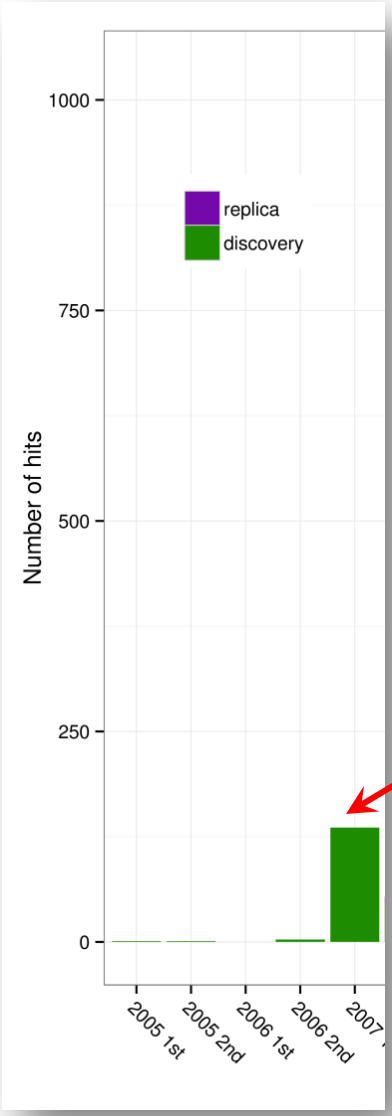
* Residual is after adjustment for age, gender, diabetes.

Table 1. Population Variation Explained by GWAS for a Selected Number of Complex Traits

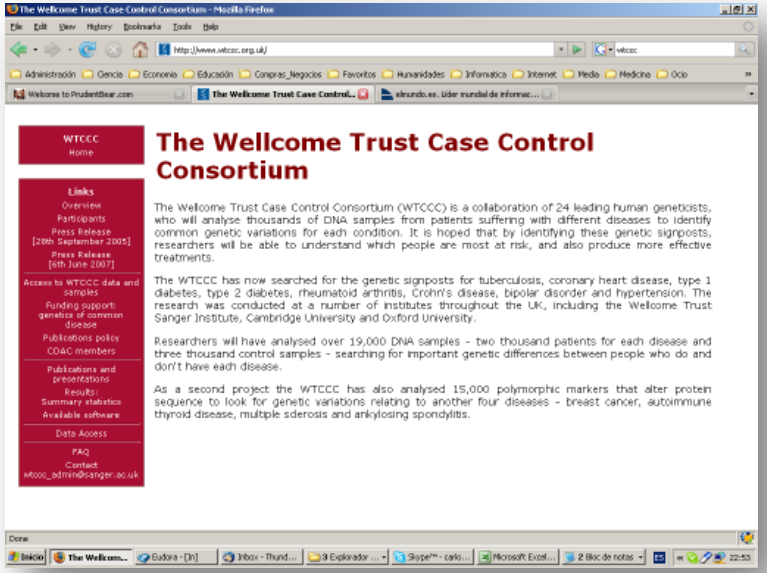
Trait or Disease	h^2 Pedigree Studies	h^2 GWAS Hits[a]	h^2 All GWAS SNPs[b]
Type 1 diabetes	0.9[98]	0.6[99,c]	0.3[12]
Type 2 diabetes	0.3–0.6[100]	0.05-0.10[34]	
Obesity (BMI)	0.4–0.6[101,102]	0.01-0.02[36]	0.2[14]
Crohn's disease	0.6–0.8[103]	0.1[11]	0.4[12]
Ulcerative colitis	0.5[103]	0.05[12]	
Multiple sclerosis	0.3–0.8[104]	0.1[45]	
Ankylosing spondylitis	>0.90[105]	0.2[106]	
Rheumatoid arthritis	0.6[107]		
Schizophrenia	0.7–0.8[108]	0.01[79]	0.3[109]
Bipolar disorder	0.6–0.7[108]	0.02[79]	0.4[12]
Breast cancer	0.3[110]	0.08[111]	
Von Willebrand factor	0.66–0.75[112,113]	0.13[114]	0.25[14]
Height	0.8[115,116]	0.1[13]	0.5[13,14]
Bone mineral density	0.6-0.8[117]	0.05[118]	
QT interval	0.37–0.60[119,120]	0.07[121]	0.2[14]
HDL cholesterol	0.5[122]	0.1[57]	
Platelet count	0.8[123]	0.05–0.1[58]	

[a] Proportion of phenotypic variance or variance in liability explained by genome-wide-significant and validated SNPs. For a number of diseases, other parameters were reported, and these were converted and approximated to the scale of total variation explained. Blank cells indicate that these parameters have not been reported in the literature.
[b] Proportion of phenotypic variance or variance in liability explained when all GWAS SNPs are considered simultaneously. Blank cell indicate that these parameters have not been reported in the literature.
[c] Includes pre-GWAS loci with large effects.

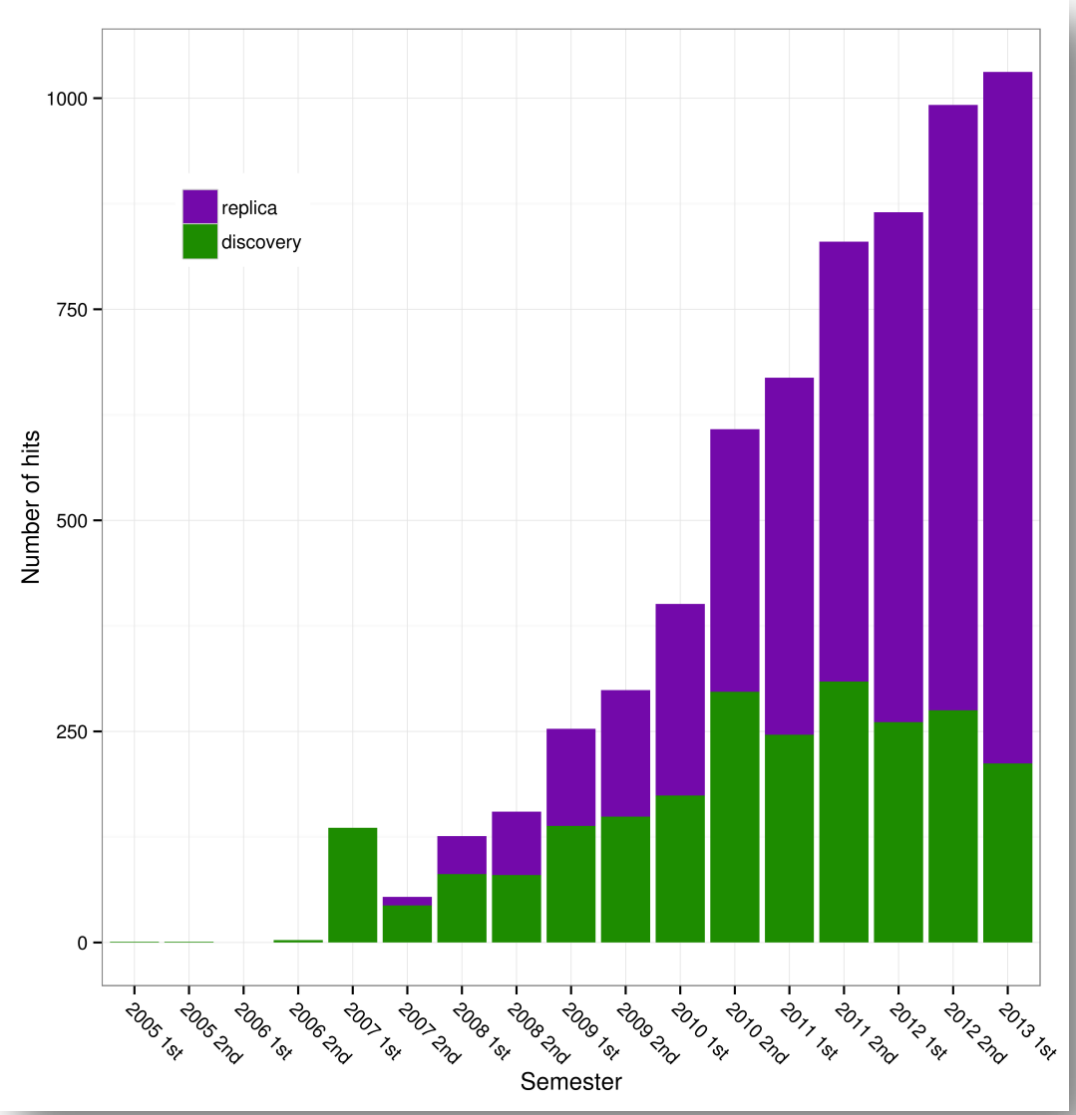
GWAS across Time: increasing replication



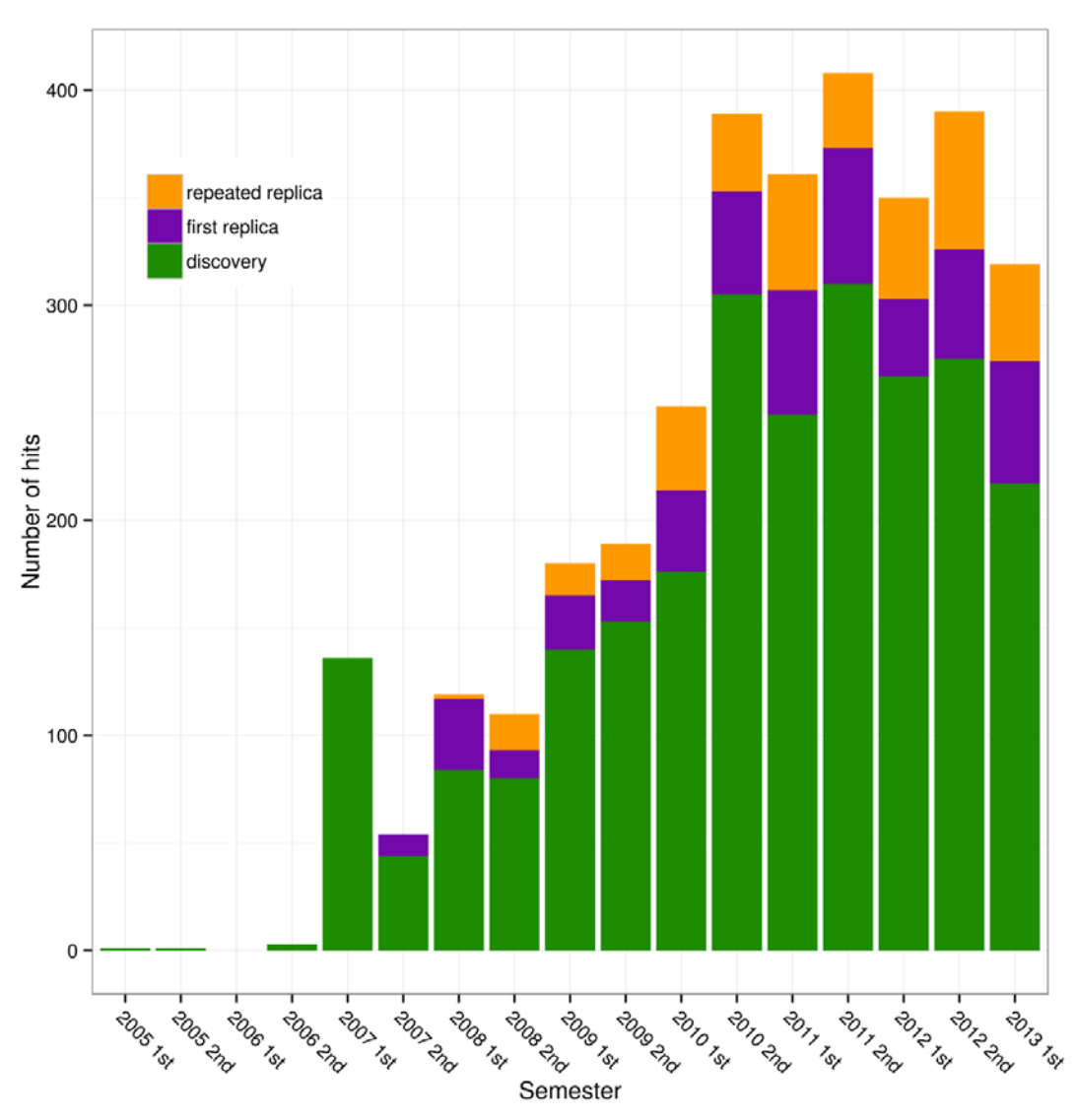
June 2007, publication of the WTCCC



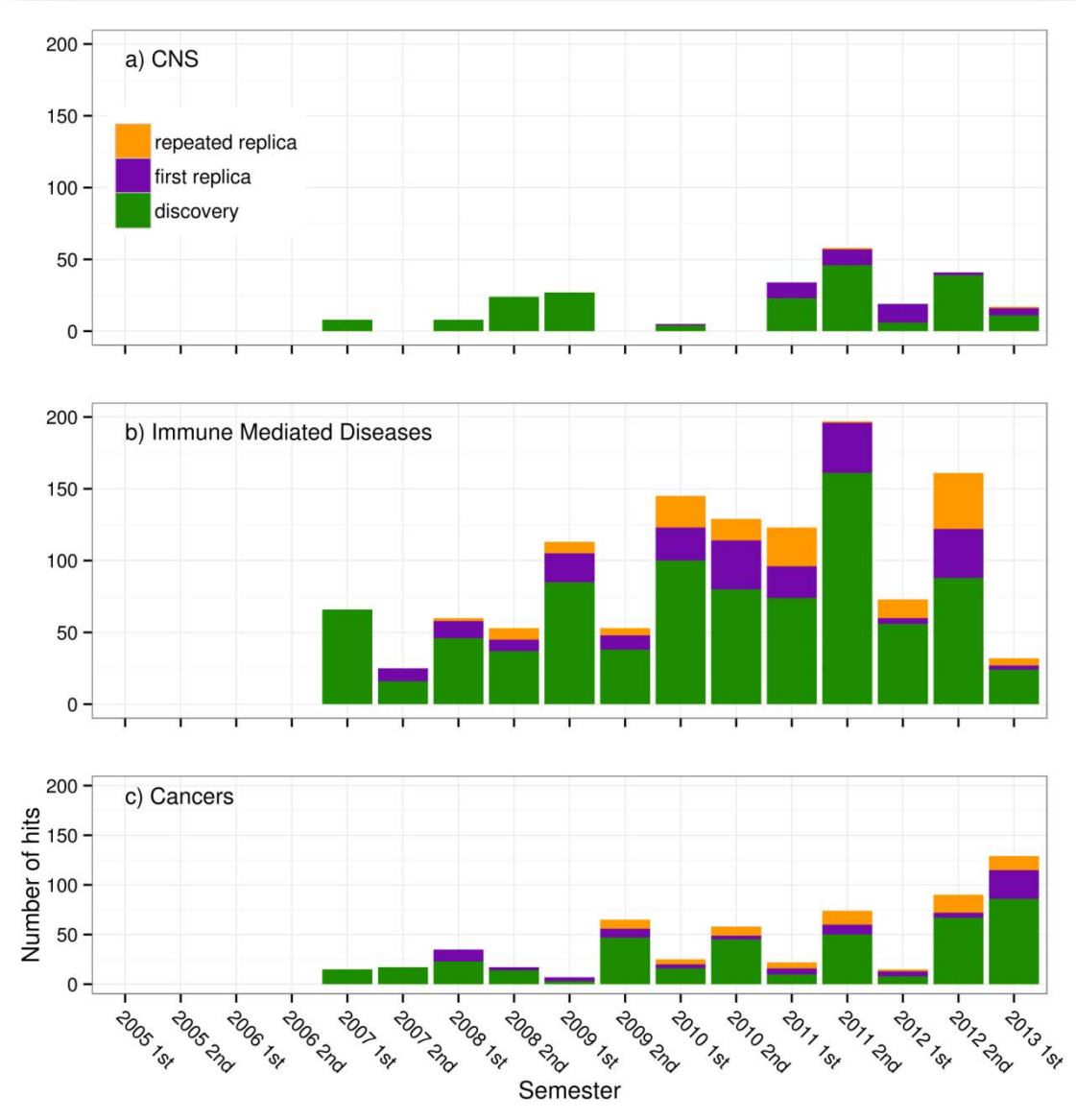
GWAS across Time: increasing replication



GWAS across Time: increasing replication



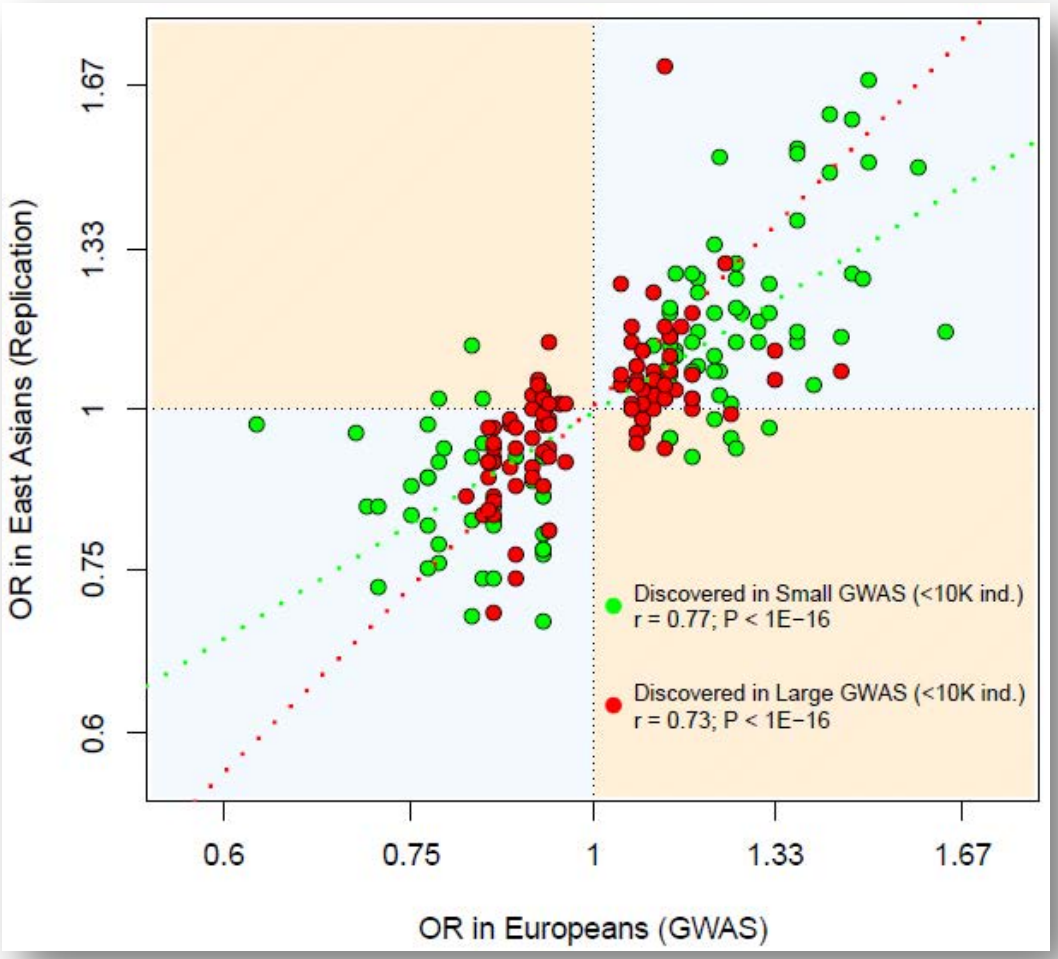
GWAS across Time: increasing replication



GWAS across Space: must be Shared Variants

We just proved high correlation between European / East Asian variants

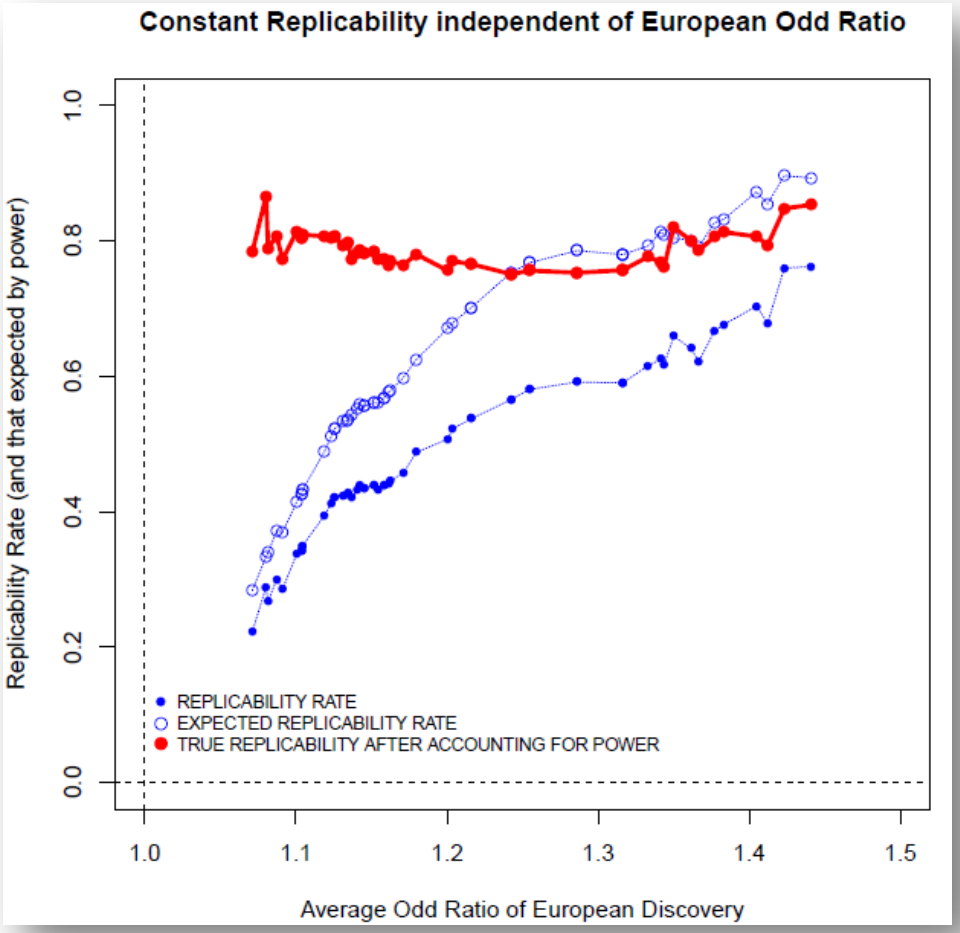
Marigorta and Navarro, Plos Genetics 2013



GWAS across Space: must be Shared Variants

And that effective replicability depends on statistical power

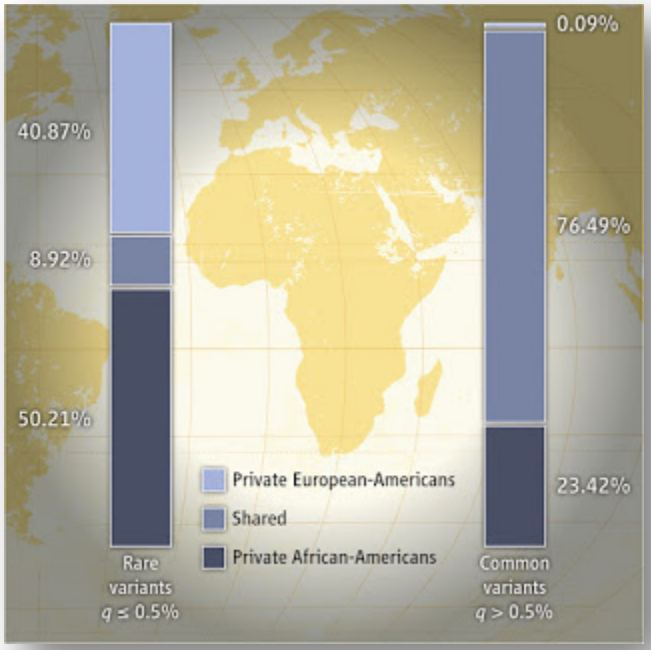
Marigorta and Navarro, Plos Genetics 2013



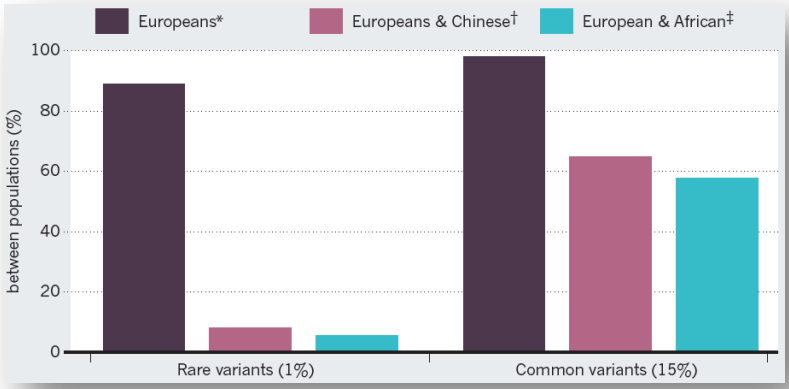
A raging debate!!

Which means that GWAS result so far must be due to high-frequency disease variants that are shared by all humankind

Casals and Bertranpetit *Science* 6090 (2012)

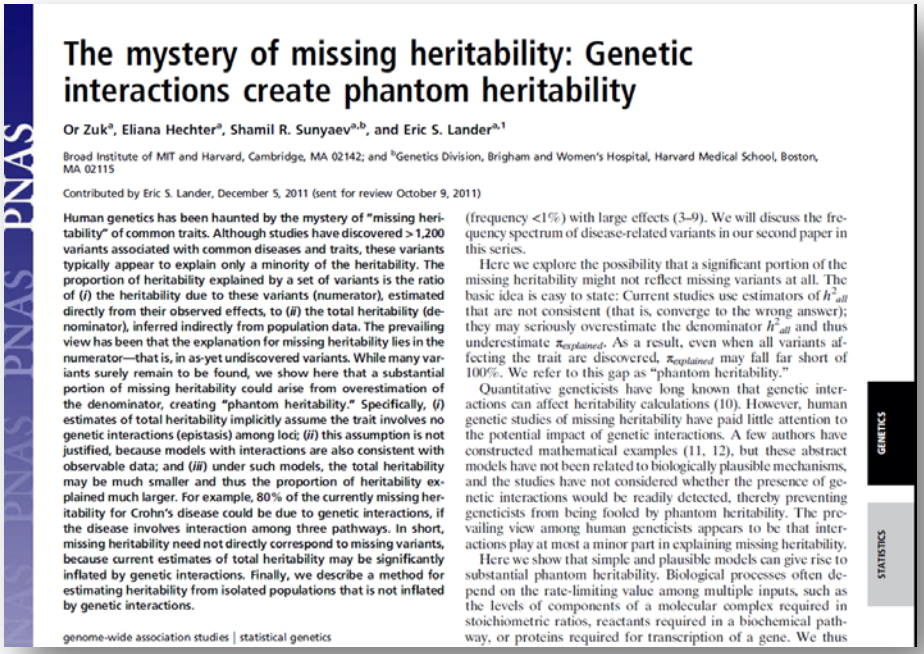


CD Bustamante *et al. Nature* 475 (2011)



- **Common variants** *are (usually) shared among continental populations (MAF > 0.1)*
- **Rare variants** *are mostly population-specific (MAF < 0.01)*

But the raging debate is still open in many aspects!!



PNAS PNAS PNAS

The mystery of missing heritability: Genetic interactions create phantom heritability

Or Zuk[a], Eliana Hechter[a], Shamil R. Sunyaev[a,b], and Eric S. Lander[a,1]

Broad Institute of MIT and Harvard, Cambridge, MA 02142; and [b]Genetics Division, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115

Contributed by Eric S. Lander, December 5, 2011 (sent for review October 9, 2011)

Human genetics has been haunted by the mystery of "missing heritability" of common traits. Although studies have discovered >1,200 variants associated with common diseases and traits, these variants typically appear to explain only a minority of the heritability. The proportion of heritability explained by a set of variants is the ratio of (*i*) the heritability due to these variants (numerator), estimated directly from their observed effects, to (*ii*) the total heritability (denominator), inferred indirectly from population data. The prevailing view has been that the explanation for missing heritability lies in the numerator—that is, in as-yet undiscovered variants. While many variants surely remain to be found, we show here that a substantial portion of missing heritability could arise from overestimation of the denominator, creating "phantom heritability." Specifically, (*i*) estimates of total heritability implicitly assume the trait involves no genetic interactions (epistasis) among loci; (*ii*) this assumption is not justified, because models with interactions are also consistent with observable data; and (*iii*) under such models, the total heritability may be much smaller and thus the proportion of heritability explained much larger. For example, 80% of the currently missing heritability for Crohn's disease could be due to genetic interactions, if the disease involves interaction among three pathways. In short, missing heritability need not directly correspond to missing variants, because current estimates of total heritability may be significantly inflated by genetic interactions. Finally, we describe a method for estimating heritability from isolated populations that is not inflated by genetic interactions.

genome-wide association studies | statistical genetics

(frequency <1%) with large effects (3–9). We will discuss the frequency spectrum of disease-related variants in our second paper in this series.

Here we explore the possibility that a significant portion of the missing heritability might not reflect missing variants at all. The basic idea is easy to state: Current studies use estimators of h^2_{all} that are not consistent (that is, converge to the wrong answer); they may seriously overestimate the denominator h^2_{all} and thus underestimate $\pi_{explained}$. As a result, even when all variants affecting the trait are discovered, $\pi_{explained}$ may fall far short of 100%. We refer to this gap as "phantom heritability."

Quantitative geneticists have long known that genetic interactions can affect heritability calculations (10). However, human genetic studies of missing heritability have paid little attention to the potential impact of genetic interactions. A few authors have constructed mathematical examples (11, 12), but these abstract models have not been related to biologically plausible mechanisms, and the studies have not considered whether the presence of genetic interactions would be readily detected, thereby preventing geneticists from being fooled by phantom heritability. The prevailing view among human geneticists appears to be that interactions play at most a minor part in explaining missing heritability.

Here we show that simple and plausible models can give rise to substantial phantom heritability. Biological processes often depend on the rate-limiting value among multiple inputs, such as the levels of components of a molecular complex required in stoichiometric ratios, reactants required in a biochemical pathway, or proteins required for transcription of a gene. We thus

GENETICS

STATISTICS

On top of that, the "Missing heritability" may be just non-existent. It may be "Phantom heritability" caused by ignoring epistasis, which inflates family based estimates of heritability

Estimated heritability

- If real heritability = estimated heritability:

Real heritability?

- If real heritability < estimated heritability:

A raging debate!!

IN SUMMARY:

As to risk prediction:

(1) Missing heritability

(2) Hidden heritability

(3) Phantom heritability

(4) ...

As to the causal variants:

(1) Common variants

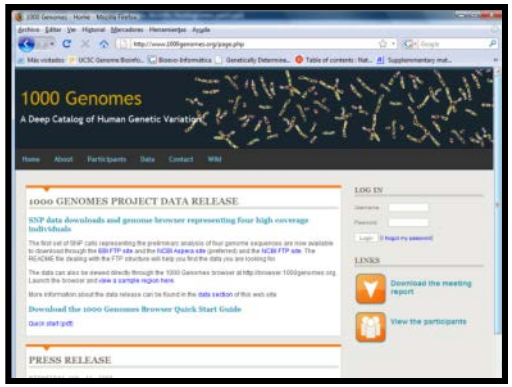
(2) Rare variants

(3) Epigenetics

(4) ...

So the information is there. What to do now?

(1) New information about variants



(1) Larger sample sizes and families. Increasing pressure for a

Global Alliance for Genomics and Health

(2) New arrays and/or NGS sequencing

(Human Omni5_Quad and/or Hiseq2500)

The current SOTA is either large samples or exomics

Sequencing Evolution/Revolution




1990: thousand bases/day




2000: million bases/day




2010: billion bases/day

The current SOTA is either large samples or exomics

nature genetics

Exome sequencing identifies the cause of a mendelian disorder

Sarah B Ng[1,10], Kati J Buckingham[2,10], Choli Lee[1], Abigail W Bigham[2], Holly K Tabor[2,3], Karin M Dent[4], Chad D Huff[5], Paul T Shannon[6], Ethylin Wang Jabs[7,8], Deborah A Nickerson[1], Jay Shendure[1] & Michael J Bamshad[1,2,9]

We demonstrate the first successful application of exome sequencing to discover the gene for a rare mendelian disorder of unknown cause, Miller syndrome (MIM%263750). For four affected individuals in three independent kindreds, we captured and sequenced coding regions to a mean coverage of 40× and sufficient depth to call variants at ~97% of each targeted exome. Filtering against public SNP databases and eight HapMap exomes for genes with two previously unknown variants in each of the four individuals identified a single candidate gene, *DHODH*, which encodes a key enzyme in the pyrimidine *de novo* biosynthesis pathway. Sanger sequencing confirmed the presence of *DHODH* mutations in three additional families with Miller syndrome. Exome sequencing of a small number of unrelated affected individuals is a powerful, efficient strategy for identifying the genes underlying rare mendelian disorders and will likely transform the genetic analysis of monogenic traits.

Received 2 October; accepted 9 November; published online 13 November; corrected online 22 November 2009 (details online); doi:10.1038/ng.499

The current SOTA is either large samples or exomics... or other designs

LETTERS

nature genetics

Exome sequencing in sporadic autism spectrum disorders identifies severe *de novo* mutations

Brian J O'Roak[1], Pelagia Deriziotis[2], Choli Lee[1], Laura Vives[1], Jerrod J Schwartz[1], Santhosh Girirajan[1], Emre Karakoc[1], Alexandra P MacKenzie[1], Sarah B Ng[1], Carl Baker[1], Mark J Rieder[1], Deborah A Nickerson[1], Raphael Bernier[3], Simon E Fisher[2,4], Jay Shendure[1] & Evan E Eichler[1,5]

Evidence for the etiology of autism spectrum disorders (ASDs) has consistently pointed to a strong genetic component complicated by substantial locus heterogeneity[1,2]. We sequenced the exomes of 20 individuals with sporadic ASD (cases) and their parents, reasoning that these families would be enriched for *de novo* mutations of major effect. We identified 21 *de novo* mutations, 11 of which were protein altering. Protein-altering mutations were significantly enriched for changes at highly conserved residues. We identified potentially causative *de novo* events in 4 out of 20 probands, particularly among more severely affected individuals, in *FOXP1*, *GRIN2B*, *SCN1A* and *LAMC3*. In the *FOXP1* mutation carrier, we also observed a rare inherited *CNTNAP2* missense variant, and we provide functional support for a multi-hit model for disease risk[3]. Our results show that trio-based exome sequencing is a powerful approach for identifying new candidate genes for ASDs and suggest that *de novo* mutations may contribute substantially to the genetic etiology of ASDs.

ASDs are characterized by pervasive impairment in language, communication and social reciprocity and restricted interests or stereotyped behaviors[1]. Several new candidate loci for ASDs have recently been identified using genome-wide approaches that discover individually rare events of major effect[2]. A number of genetic syndromes with features of the ASD phenotype, collectively referred to as syndromic autism, have also been described[4]. Despite this progress, the genetic basis for the vast majority of cases remains unknown. Several observations support the hypothesis that the genetic basis for ASDs in sporadic cases may differ from that of families with multiple affected individuals, with the former being more likely to result from *de novo* mutation events rather than inherited variants[1,5–7]. In this study, we sequenced the protein-coding regions of the genome (the exome)[8] to test the hypothesis that *de novo* protein-altering mutations contribute substantially to the genetic basis of sporadic ASDs.

In contrast with array-based analysis of large *de novo* copy number variants (CNVs), this approach has greater potential to implicate single genes in ASDs.

We selected 20 trios with an idiopathic ASD, each consistent with a sporadic ASD based on clinical evaluations (**Supplementary Table 1**), pedigree structure, familial phenotypic evaluation, family history and/or elevated parental age. Each family was initially screened by array comparative genomic hybridization (CGH) using a customized microarray[9]. We identified no large (>250 kb) *de novo* CNVs but did identify a maternally inherited deletion (~350 kb) at 15q11.2 in one family (**Supplementary Fig. 1**). This deletion has been associated with increased risk for epilepsy[10] and schizophrenia[11,12] but has not been considered causal for autism.

Similar to researchers from a previous study[13], who reported exome sequencing on ten parent–child trios with sporadic cases of moderate to severe intellectual disability, we performed exome sequencing on each of the 60 individuals separately by subjecting whole-blood derived genomic DNA to in-solution hybrid capture and Illumina sequencing (Online Methods). We obtained sufficient coverage to call variants for ~90% of the primary target (26.4 Mb) (**Table 1**). Genotype concordance with SNP microarray data was high (99.7%) (**Supplementary Table 2**), and on average, 96% of proband variant sites were also called in both parents (**Supplementary Table 3**). Given the expected rarity of true *de novo* events in the targeted exome (<1 per trio) (**Supplementary Table 4**)[14], we reasoned that most apparently *de novo* variants would result from under calling in parents or systematic false positive calls in the proband. We therefore filtered variants previously observed in the dbSNP database, 1000 Genomes Pilot Project data[15] and 1,490 other exomes sequenced at the University of Washington (**Supplementary Fig. 2**). We performed Sanger sequencing on the remaining *de novo* candidates (<5 per trio), validating 18 events within coding sequence and three additional events mapping to 3′ untranslated regions (**Table 2**). A list of predicted variant sites within these genes from the 1000 Genomes Pilot Project data[15] is provided for comparison (**Supplementary Table 5**).

[1]Department of Genome Sciences, University of Washington School of Medicine, Seattle, Washington, USA. [2]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK. [3]Department of Psychiatry and Behavioral Sciences, University of Washington, Seattle, Washington, USA. [4]Language and Genetics Department, Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands. [5]Howard Hughes Medical Institute, Seattle, Washington, USA. Correspondence should be addressed to E.E.E. (eee@gs.washington.edu) or J.S. (shendure@uw.edu).

Received 22 February; accepted 21 April; published online 15 May 2011; doi:10.1038/ng.835

LETTER

doi:10.1038/nature10113

Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia

Xose S. Puente[1], Magda Pinyol[2], Victor Quesada[1], Laura Conde[3], Gonzalo R. Ordóñez[1], Neus Villamor[2], Georgia Escaramis[4], Pedro Jares[5], Silvia Beà[2], Marcos González-Díaz[3], Laia Bassaganyas[6], Tycho Baumann[6], Manel Juan[7], Mónica López-Guerra[7], Dolors Colomer[7], José M. C. Tubío[8,9], Cristina López[2], Alba Navarro[3], Cristian Tornador[4], Marta Aymerich[3], Maria Rozman[3], Jesús M. Hernández[10], Diana A. Puente[1], José M. P. Freije[1], Gloria Velasco[1], Ana Gutiérrez-Fernández[1], Dolors Costa[3], Anna Carrió[3], Sara Guijarro[3], Anna Enjuanes[3], Lluis Hernández[3], Jordi Yagüe[7], Pilar Nicolás[9], Carlos M. Romeo-Casabona[9], Heinz Himmelbauer[10], Ester Castillo[10], Juliane C. Dohm[10], Silvia de Sanjosé[11], Miguel A. Piris[12], Enrique de Alava[5], Jesús San Miguel[13], Romina Royo[14], Josep L. Gelpi[14], David Torrents[14], Modesto Orozco[15], David G. Pisano[16], Alfonso Valencia[14], Roderic Guigó[17], Mónica Bayés[18], Simon Heath[19], Marta Gut[19], Peter Klatt[17], John Marshall[18], Keiran Raine[18], Lucy A. Stebbings[18], P. Andrew Futreal[18], Michael R. Stratton[18], Peter J. Campbell[18], Ivo Gut[19], Armando López-Guillermo[6], Xavier Estivill[4], Emili Montserrat[6], Carlos López-Otín[1]* & Elías Campo[2]*

Chronic lymphocytic leukaemia (CLL), the most frequent leukaemia in adults in Western countries, is a heterogeneous disease with variable clinical presentation and evolution[1,2]. Two major molecular subtypes can be distinguished, characterized respectively by a high or low number of somatic hypermutations in the variable region of immunoglobulin genes[3,4]. The molecular changes leading to the pathogenesis of the disease are still poorly understood. Here we performed whole-genome sequencing of four cases of CLL and identified 46 somatic mutations that potentially affect gene function. Further analysis of these mutations in 363 patients with CLL identified four genes that are recurrently mutated: notch 1 (*NOTCH1*), exportin 1 (*XPO1*), myeloid differentiation primary response gene 88 (*MYD88*) and kelch-like 6 (*KLHL6*). Mutations in *MYD88* and *KLHL6* are predominant in cases of CLL with mutated immunoglobulin genes, whereas *NOTCH1* and *XPO1* mutations are mainly detected in patients with unmutated immunoglobulins. The patterns of somatic mutation, supported by functional and clinical analyses, strongly indicate that the recurrent *NOTCH1*, *MYD88* and *XPO1* mutations are oncogenic changes that contribute to the clinical evolution of the disease. To our knowledge, this is the first comprehensive analysis of CLL combining whole-genome sequencing with clinical characteristics and clinical outcomes. It highlights the usefulness of this approach for the identification of clinically relevant mutations in cancer.

To gain insights into the molecular alterations that cause CLL, we performed whole-genome sequencing of four cases representative of different forms of the disease: two cases, CLL1 and CLL2, with no mutations in the immunoglobulin genes (*IGHV*-unmutated) and two cases, CLL3 and CLL4, with mutations in these genes (*IGHV*-mutated) (Supplementary Table 1 and Supplementary Information). We used a combination of whole-genome sequencing and exome sequencing, as well as long-insert paired-end libraries, to detect variants in chromosomal structure (Supplementary Fig. 1 and Supplementary Tables 2–5).

We obtained more than 99.7% concordance between whole-genome sequencing calls and genotyping data, indicating that the coverage and parameters used were sufficient to detect most of the sequence variants in these samples (Supplementary Information). We detected about 1,000 somatic mutations per tumour in non-repetitive regions (Fig. 1a, Supplementary Fig. 2 and Supplementary Table 6). These numbers of somatic mutations were lower than the numbers in melanoma and lung carcinoma[5,6], but in agreement with previous estimates of less than one mutation per megabase (Mb) for leukaemias[7]. The most common substitution was the transition G>A/C>T, usually occurring in a CpG context (Fig. 1b and Supplementary Fig. 2). We also detected marked differences in the mutation pattern between CLL samples and these differences were associated with tumour subtype (Fig. 1b). Thus, *IGHV*-mutated cases showed a higher proportion of A>G/T>G mutations than cases with unmutated *IGHV* (16 ± 0.2% versus 6.2 ± 0.1%). The base preceding the adenine in A to C transversions showed an over-representation of thymine, when compared to the prevalence expected from its representation in non-repetitive sequences in the wild-type genome ($P < 0.001$, Fig. 1c), and there were fewer A to C substitutions at GpA dinucleotides than would be expected by chance ($P < 0.001$). These differences between CLL subtypes might reflect the molecular mechanisms implicated in their respective development. The pattern and context of mutations are consistent with their being introduced by the error-prone polymerase η during somatic hypermutation in immunoglobulin genes[8]. This indicates that polymerase η could contribute to the high frequency of A > T to C > G transversions in cases with *IGHV*-mutated. It also extends the differences observed between these two CLL subtypes to the genomic level.

We classified the somatic mutations into three different classes according to their potential functional effect (Supplementary Information). We also searched for small insertions and deletions (indels) in coding regions: we found and validated five somatic indels, which caused frameshifts in protein-coding regions (Supplementary Table 7).

[1]Departamento de Bioquímica y Biología Molecular, Instituto Universitario de Oncología, Universidad de Oviedo, 33006 Oviedo, Spain. [2]Unidad de Genómica, Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), 08036 Barcelona, Spain. [3]Unidad de Hematopatología, Servicio de Anatomía Patológica, Hospital Clínic, Universitat de Barcelona, IDIBAPS, 08036 Barcelona, Spain. [4]Genes and Disease Programme, Center for Genomic Regulation, Pompeu Fabra University (CRG-UPF), 08003 Barcelona, Spain. [5]Servicio de Hematología, Hospital Universitario Centro de Investigación del Cáncer-IBMCC (USAL-CSIC), Universidad de Salamanca, 37007 Salamanca, Spain. [6]Servicio de Hematología, Hospital Clínic, IDIBAPS, Universidad de Barcelona, 08036 Barcelona, Spain. [7]Servicio de Inmunología, Hospital Clínic, IDIBAPS, 08036 Barcelona, Spain. [8]Hospital Clínico Universitario de Santiago de Compostela, 15 706 Santiago de Compostela, Spain. [9]Cátedra Inter-Universitaria de Derecho y Genoma Humano, Universidad de Deusto, Universidad del País Vasco, 48007 Bilbao, Spain. [10]Sequencing Unit, Center for Genomic Regulation, Pompeu Fabra University (CRG-UPF), 08003 Barcelona, Spain. [11]Unidad de Infecciones y Cáncer, Institut Català de Oncologia-IDIBELL, 08907 Hospitalet de Llobregat, Spain. [12]Molecular Pathology Programme, Spanish National Cancer Research Centre (CNIO), 28029 Madrid, Spain. [13]Programa Conjunto de Biología Computacional, Barcelona Supercomputing Center (BSC), Institut de Recerca Biomèdica (IRB), Spanish National Bioinformatics Institute, Universitat de Barcelona, 08028 Barcelona, Spain. [14]Structural Biology and Biocomputing Programme, Spanish National Cancer Research Centre (CNIO), Spanish National Bioinformatics Institute, 28029 Madrid, Spain. [15]Genomic Bioinformatics Programme, Center for Genomic Regulation, Pompeu Fabra University (CRG-UPF), Spanish National Bioinformatics Institute, 08003 Barcelona, Spain. [16]Centro Nacional de Análisis Genómico, Parc Científic de Barcelona, 08028 Barcelona, Spain. [17]Centro Nacional de Biotecnología, Consejo Superior de Investigaciones Científicas, 28049 Madrid, Spain. [18]Wellcome Trust Sanger Institute, Hinxton CB10 1SA, UK.
*These authors contributed equally to this work.

The current SOTA is either large samples or exomics...or other designs

Or some ingenious comparisons

Vol 464|29 April 2010|doi:10.1038/nature08990 nature

LETTERS

Genome, epigenome and RNA sequences of monozygotic twins discordant for multiple sclerosis

Sergio E. Baranzini[1], Joann Mudge[2], Jennifer C. van Velkinburgh[2], Pouya Khankhanian[1], Irina Khrebtukova[3], Neil A. Miller[2], Lu Zhang[3], Andrew D. Farmer[2], Callum J. Bell[2], Ryan W. Kim[2], Gregory D. May[2], Jimmy E. Woodward[2], Stacy J. Caillier[1], Joseph P. McElroy[1], Refujia Gomez[1], Marcelo J. Pando[4], Leonda E. Clendenen[2], Elena E. Ganusova[2], Faye D. Schilkey[2], Thiruvarangan Ramaraj[2], Omar A. Khan[5], Jim J. Huntley[3], Shujun Luo[3], Pui-yan Kwok[6,7], Thomas D. Wu[8], Gary P. Schroth[3], Jorge R. Oksenberg[1,7], Stephen L. Hauser[1,7] & Stephen F. Kingsmore[2]

Table 1 | SNP and indel genotypes and differences between siblings in three twin pairs

Genotype change and individual	Platform	Twin pair 041896 SNP genotypes	Twin pair 041896 Replicated SNP genotype differences§	Twin pair 041896 Indel genotypes	Twin pair 041896 Replicated indel genotype difference	Twin pair 230178 SNP genotypes	Twin pair 230178 Replicated SNP genotype difference	Twin pair 230178 Indel genotypes	Twin pair 041907 SNP genotypes	Twin pair 041907 Replicated SNP genotype difference	Twin pair 041907 Indel genotypes
No change	Genome-Seq*	1,086,309	79,209	26,908	91 (91.9%)	ND	NA	ND	ND	NA	ND
	SNP array (×2)	736,782	1,638 (98.3%)	NA	NA	783,189	888 (95.3%)	NA	796,870	385 (98.0%)	NA
	mRNA-Seq‡	51,201	8,816 (98.2%)	1,314	91 (91.9%)	39,816		1,034	18,123		397
Ref in -001 → het in -101	Genome-Seq*†	202	0	3	0	ND	NA	ND	ND	NA	ND
	SNP array (×2)	32	0	NA	NA	36	0	NA	32	0	NA
	mRNA-Seq†‡	12	0	0	0	6	0	0	2	0	0
Het in -001 → ref in -101	Genome-Seq*†	134	0	1	0	ND	NA	ND	ND	NA	ND
	SNP array (×2)	49	0	NA	NA	31	0	NA	11	0	NA
	mRNA-Seq†‡	5	0	0	0	9	0	0	16	0	0
Het in -001 → hom in -101	Genome-Seq*†	1,513	0	128	0	ND	NA	ND	ND	NA	ND
	SNP array (×2)	29	0	NA	NA	24	0	NA	17	0	NA
	mRNA-Seq†‡	203	0	7	0	573	0	23	170	0	5
Hom in -001 → het in -101	Genome-Seq*†	1,392	0	81	0	ND	NA	ND	ND	NA	ND
	SNP array (×2)	16	0	NA	NA	62	0	NA	0	0	NA
	mRNA-Seq†‡	102	0	1	0	429	0	0	0	0	0

Genotype categories: homozygous reference (ref), heterozygous variant (het) and homozygous variant (hom). NA, not app
* Nucleotide genotyped if 11–44× coverage and $Q \geq 20$.
† Genotypes determined according to frequency cutoffs in Supplementary Table 8 and differences called if frequencies diffe
‡ Genotyped if present in >2 reads, >1 uniquely aligning read and $Q \geq 20$.
§ Detected by platform on corresponding row, replicated by platform listed on row below.

Table 2 | CpG sites and clusters in monozygotic twins, normal and cancer samples

Genomic DNA sample	CpG sites*	CpG clusters	Ratio of CpGs to clusters	CpGs shared	CpG clusters shared	mCpG unique to one sample†	Between sample comparison†	CpGs shared	CpG clusters shared	mCpG unique to one sample†
041896-001 T cell	2,146,620	1,230,241	1.74	98.1%	98.2%	2‡	041896- & 230178-001 T cell	97.4%	97.7%	522
041896-101 T cell	2,033,078	1,190,741	1.71			0				305
230178-001 T cell	1,636,285	1,038,787	1.58	97.8%	97.9%	3	041896-001 & 230178-101 T cell	96.5%	96.9%	445
230178-101 T cell	1,917,131	1,155,024	1.66			7				362
041907-001 T cell	1,779,140	1,094,361	1.63	90.6%	92.7%	174	041896- & 041907-001 T cell	97.5%	98.1%	304
041907-101 T cell	1,642,200	1,038,090	1.58			2				282
Normal breast	1,829,855	1,086,405	1.68	96.7%	97.9%	696	041896-001 T cell & normal breast	97.3%	98.0%	5,620
Breast cancer	2,010,173	1,192,180	1.69			861				1,560
Normal lung	2,096,524	1,216,046	1.72	97.9%	98.8%	6,891	041896-001 T cell & normal lung	96.1%	97.0%	3,329
Lung cancer	1,619,178	956,760	1.69			9,618				926

CpG sites and clusters were compared between $CD4^+$ lymphocytes from three pairs of monozygotic twins, breast and lung cancer and normal tissue samples.
* >10 RRBS reads aligned by ELAND-extended and $Q > 20$.
† CpG >80% methylated in one sample and <20% in other.
‡ Not replicated after RRBS read alignment with GSNAP.

And this is only the scientific part of the problem

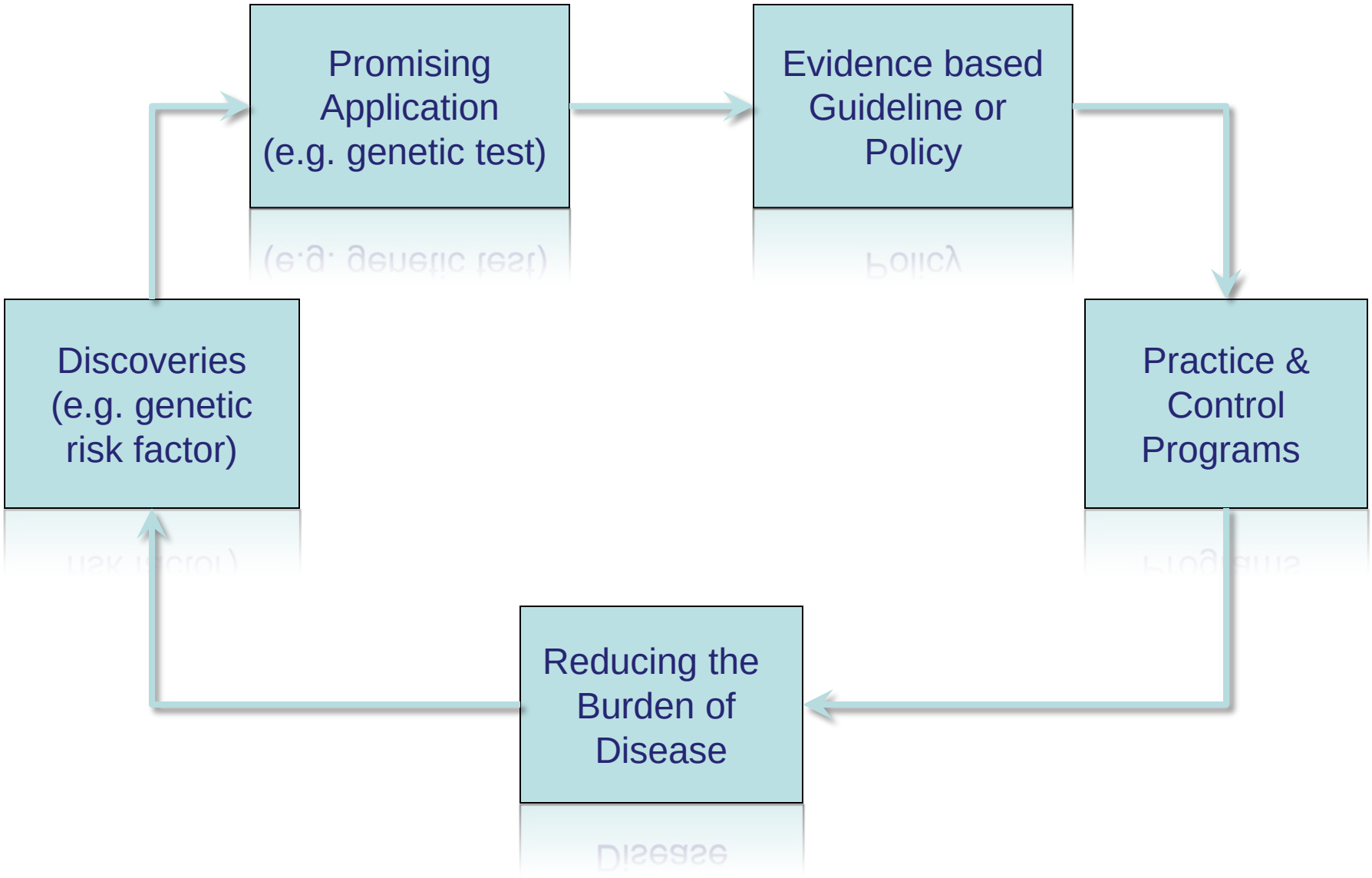
Personalized Medicine:

Vision vs. *Reality*

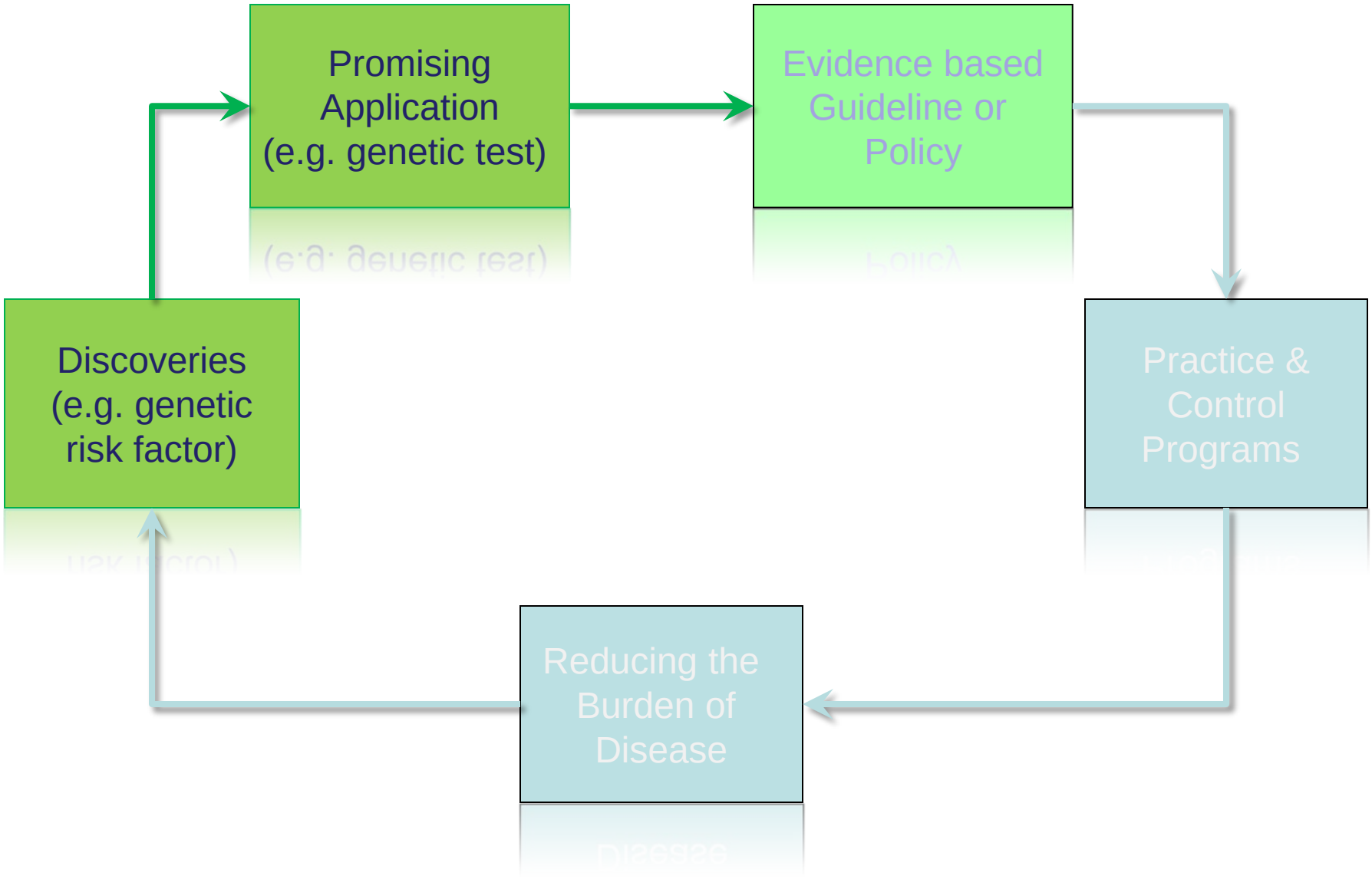
- Disruptive developments in science and technology
- Convergence of molecular biology, genetics, advanced technology, bioinformatics, broadband
- "Team science"
- Transformational changes in medicine
  - Molecular-based products and services
  - Shift towards prevention
  - Reclassification of disease
  - Integration and coordination
  - IT solutions; Interoperability
  - Consumer-centered
  - Premise that knowledge will change behavior
- Huge public & private investments in R&D
- Health as a national asset
- Ethical, legal and policy issues addressed in parallel with the science

- Healthcare delivery focused on "sick care"
  - Standardization for quality improvement
- Fragmented, lack of coordination
- Costs growing and unsustainable
  - Pressures of expensive new technologies
  - Aging population in search of new services
  - Millions of Americans under- or uninsured
  - Employer-based system tenuous
  - No evidence of healthier citizenry
- Inefficient use of information
  - Lack of IT investment, connectivity
- Evidence base for medicine inadequate
  - Continuing debate about role of cost-effectiveness
- Huge provider knowledge gaps re genomics
- Complicated regulatory framework
- Reimbursement hurdles and uncertainties
- Powerful stakeholders in current system resist change

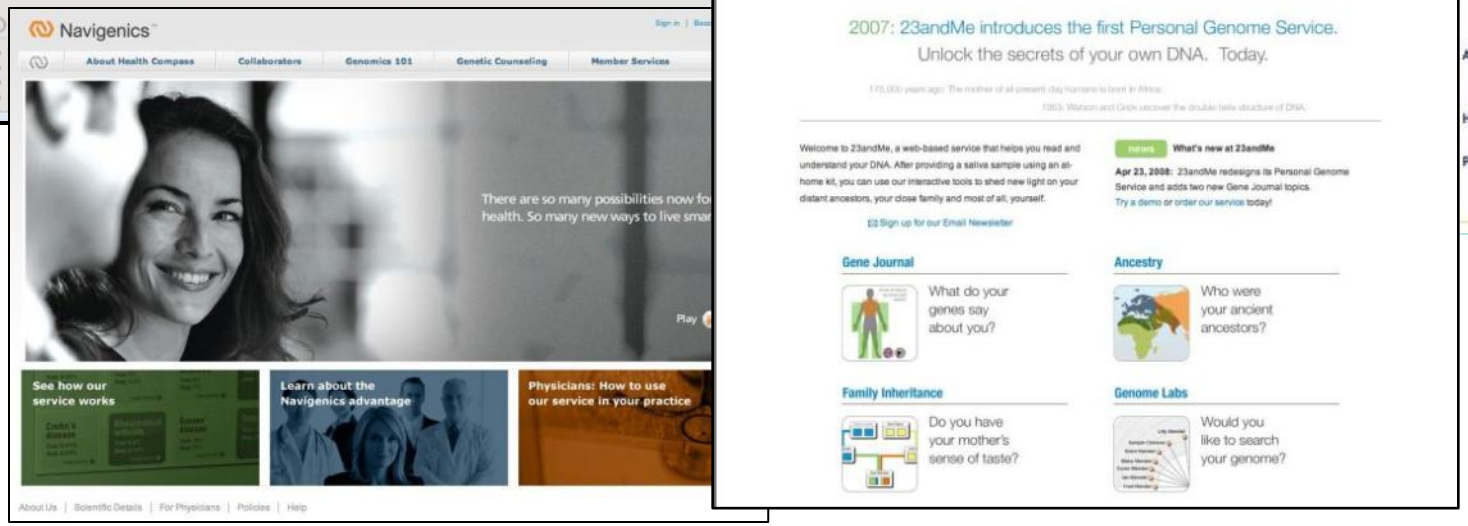
Only starting



Only starting



Where do we stand?



Where do we stand?

Disease Programs	Loci identified	Disease gene(s) isolated	Drug discovery and development
Cardiovascular			
Atherosclerosis (PAOD)	✓	✓	✓
Heart attack (Myocardial Infarction, MI)	✓	✓	✓
Hypertension	✓	✓	
Stroke	✓	✓	✓
Inflammatory			
Ankylosing spondilytis			
Asthma	✓	✓	✓
Atopy/Allergy	✓		
Chronic obstructive pulmonary disease	✓		
Inflammatory bowel disease			
Insulin-dependent diabetes			
Osteoarthritis	✓	✓	
Psoriasis	✓		
Rheumatoid arthritis	✓		
Central Nervous System			
Alzheimer's disease	✓		
Anxiety	✓		
Attention deficit and hyperactivity disorder			
Autism			
Bipolar disease			
Depression			
Dyslexia			
Familial essential tremor	✓		
Migraine	✓		
Multiple sclerosis	✓		
Narcolepsy	✓		
Parkinson's disease	✓		
Restless Leg Syndrome			
Schizophrenia	✓	✓	✓
Cancer			
Benign prostatic hypertrophy	✓		
Breast cancer			
Colon cancer			
Lung cancer			
Melanoma			
Prostate cancer	✓	✓	
Renal cancer			
Testicular cancer			
Thyroid cancer			
Eye			

Biomarker	Application
Her-2/neu receptor	Select Herceptin (trastuzumab) for breast cancer
BRCA1/2	Breast and ovarian cancer inherited risk, prophylactic tamoxifen and surgery
Transcriptional profile – 21 genes	Avoid use of chemotherapy in breast CA patients with low risk of recurrence
CYP2D6/CYP2D19	Guide prescribing/ adjust dose of ~25% of commonly used drugs
VKOR/CYP2C9	Dosing of warfarin

Where do we stand?

Table 1 Examples of genetic and genomic testing in personalized medicine

Pre-symptomatic risk assessment

- *BRCA1/2* testing for breast cancer[a]
- Lynch syndrome testing for hereditary colon cancer[b]
- Long QT interval[c,d]
- Spinal Muscular Atrophy[e]

Diagnosis

- Beta thalassemia[f]
- Fusion genes and rearrangements including *BCR-ABL*, *E2A-PBX1*, *TEL-AML1*, and *MLL* in pediatric leukemia[g]
- Gene expression profiles define subtypes of breast cancer[h]
- Human Papilloma Virus detection[i]
- Hepatitis C detection[j]
- PCR detection of micro-organisms (bacteria, fungi)[k]

Prognosis

- Fragile X syndrome (number of trinucleotide repeats predicts severity)[l]
- Gene expression signatures and prognosis in breast cancer[m]
- Gene expression analysis and lymphoma prognosis[n]

Treatment and pharmacogenomics

- Therapies for targeted gene mutations in cancer[o]
  - *EGFR* point mutations in lung cancer and glioblastoma and cetuximab, gefitinib, erlotinib, panitumumab, lapatinib treatment
  - *KIT*, *PDGFR* mutations in sarcoma, glioma, liver and renal cancer, melanoma and imatinib, nilotinib, sunitinib, sorafenib treatment
  - *BRAF* mutations in melanoma treated by RAF inhibitors
  - *BCR-ABL* translocation in chronic myelogenous leukemia treated by imatinib
  - *KRAS* wild-type status correlated with resistance to EGFR inhibition
  - PARP inhibitors in *BRCA* mutant breast, ovarian, prostate and pancreatic cancer
  - Herceptin (Trastuzumab) in HER2 + breast cancer
- Pharmacogenomic applications[p]
  - *CYP 2C19**2* variant (rs4244285) associated with diminished clopidogrel response[q]
  - Rs2395029 testing for HLA-B*5701 allele, correlated with hypersensitivity to abacavir treatment for HIV+ patients[r]

[a] Robson and Offit (2007), [b] EGAPP (2009a), [c] Napolitano et al. (2005), [d] Lehnart et al. (2007), [e] Prior et al. (2008), [f] Galanello and Origa (2010), [g] Carroll et al. (2003), [h] Sorlie et al. (2001), [i] Nicol et al. (2010), [j] Pham et al. (2010), [k] Tsalik et al. (2010), [l] Sherman et al. (2005), [m] Kim and Paik (2010), [n] Rosenwald et al. (2002), [o] Macconaill and Garraway (2010), [p] U.S. Food and Drug Administration (2011), [q] Shuldiner et al. (2009), [r] Colombo et al. (2008)


© Francis Collins, 2008

Where do we stand?

Examples of going from individual genome-wide analyses to treatment are accumulating (e.g., Nic Volker, the Beery twins, Mike Snyder, John Lauerman).

Editorial, Molecular Systems Biology 9. January 2013

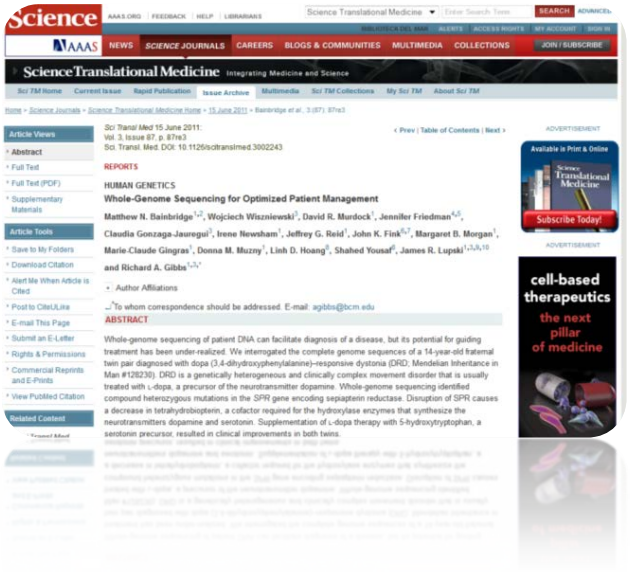
BRIEF REPORT

Making a definitive diagnosis: Successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease

Elizabeth A. Worthey, PhD[1,2], Alan N. Mayer, MD, PhD[2,3], Grant D. Syverson, MD[2], Daniel Helbling, BSc[1], Benedetta B. Bonacci, MSc[2], Brennan Decker, BSc[1], Jaime M. Serpe, BSc[2], Trivikram Dasu, PhD[2], Michael R. Tschannen, BSc[1], Regan L. Veith, MSc[2], Monica J. Basehore, PhD[4], Ulrich Broeckel, MD, PhD[1,2,3], Aoy Tomita-Mitchell, PhD[1,2,3], Marjorie J. Arca, MD[3,5], James T. Casper, MD[2,3], David A. Margolis, MD[2,3], David P. Bick, MD[1,2,3], Martin J. Hessner, PhD[1,2], John M. Routes, MD[2,3], James W. Verbsky, MD, PhD[2,3], Howard J. Jacob, PhD[1,2,3,6], and David P. Dimmock, MD[1,2,3]

Nic Volker and the XIAP gene (Worthey et all. Genetics In Medicine 2011)

Science Transl Med 87 June 2011



Science | Science Translational Medicine

Sci Transl Med 15 June 2011:
Vol. 3, Issue 87, p. 87re3
Sci. Transl. Med. DOI: 10.1126/scitranslmed.3002243

REPORTS

HUMAN GENETICS

Whole-Genome Sequencing for Optimized Patient Management

Matthew N. Bainbridge[1,*], Wojciech Wiszniewski[2], David R. Murdock[1], Jennifer Friedman[3,5], Claudia Gonzaga-Jauregui[2], Irene Newsham[1], Jeffrey G. Reid[1], John K. Fink[3,7], Margaret B. Morgan[1], Marie-Claude Gingras[1], Donna M. Muzny[1], Linh D. Hoang[2], Shahed Yousaf[2], James R. Lupski[1,3,8,10] and Richard A. Gibbs[1,3,*]

ABSTRACT

Whole-genome sequencing of patient DNA can facilitate diagnosis of a disease, but its potential for guiding treatment has been under-realized. We interrogated the complete genome sequences of a 14-year-old fraternal twin pair diagnosed with dopa (3,4-dihydroxyphenylalanine)-responsive dystonia (DRD; Mendelian Inheritance in Man #128230). DRD is a genetically heterogeneous and clinically complex movement disorder that is usually treated with L-dopa, a precursor of the neurotransmitter dopamine. Whole-genome sequencing identified compound heterozygous mutations in the *SPR* gene encoding sepiapterin reductase. Disruption of SPR causes a decrease in tetrahydrobiopterin, a cofactor required for the hydroxylase enzymes that synthesize the neurotransmitters dopamine and serotonin. Supplementation of L-dopa therapy with 5-hydroxytryptophan, a serotonin precursor, resulted in clinical improvements in both twins.

cell-based therapeutics the next pillar of medicine

Take home message:

Gene Discovery → **Closing the Gap** → **Population Health**

Lots to do that we MUST be doing!!!