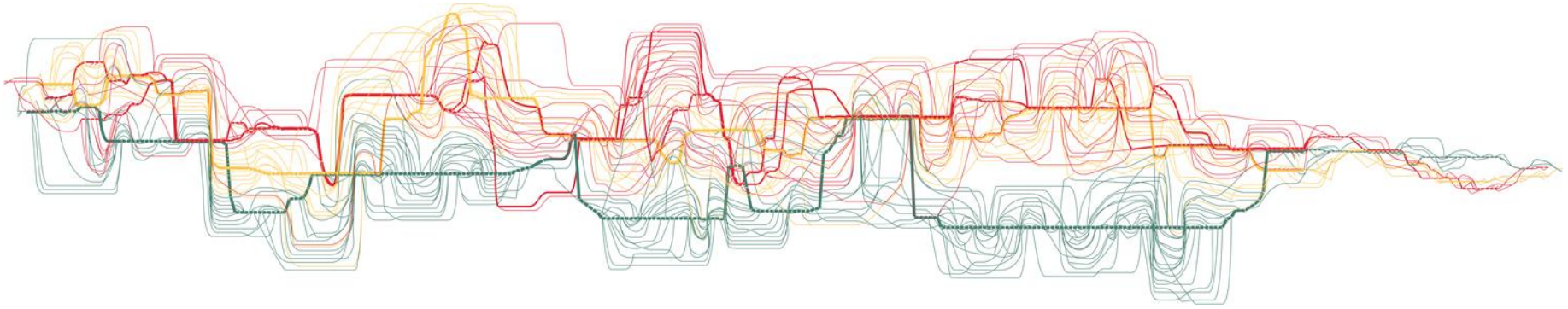
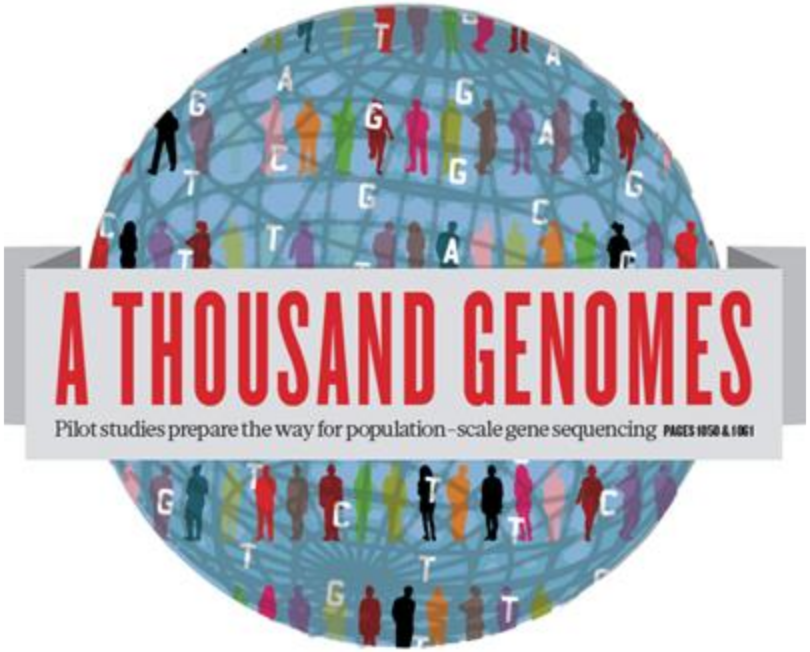
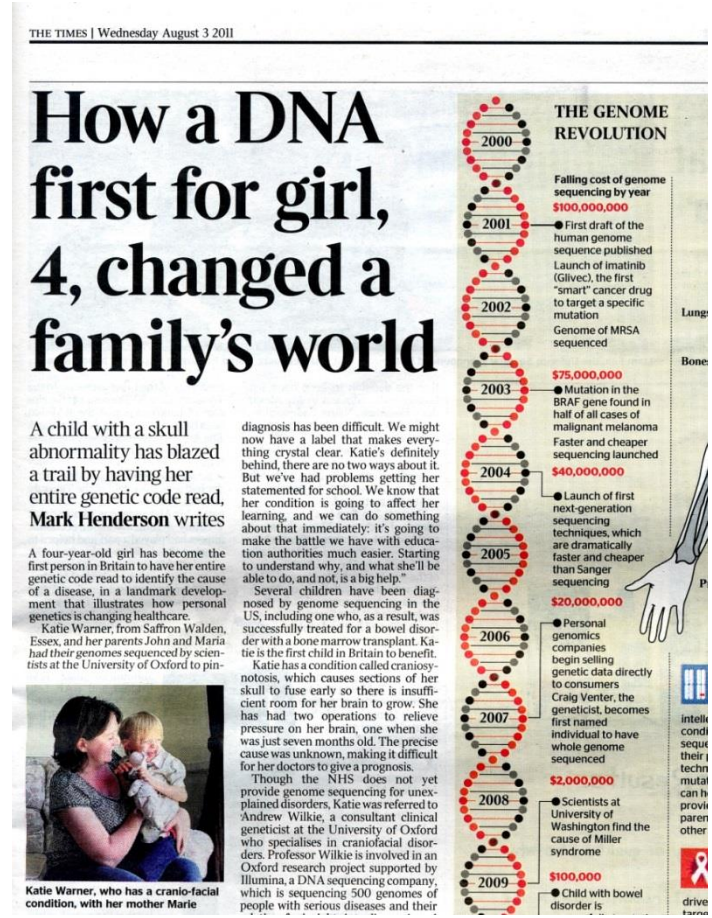
Making big data work for biomedicine

Gil McVean







THE TIMES | Wednesday August 3 2011

How a DNA first for girl, 4, changed a family's world

A child with a skull abnormality has blazed a trail by having her entire genetic code read, **Mark Henderson** writes

A four-year-old girl has become the first person in Britain to have her entire genetic code read to identify the cause of a disease, in a landmark development that illustrates how personal genetics is changing healthcare.

Katie Warner, from Saffron Walden, Essex, and her parents John and Maria had their genomes sequenced by scientists at the University of Oxford to pin-

Katie Warner, who has a cranio-facial condition, with her mother Marie

diagnosis has been difficult. We might now have a label that makes everything crystal clear. Katie's definitely behind, there are no two ways about it. But we've had problems getting her statemented for school. We know that her condition is going to affect her learning, and we can do something about that immediately: it's going to make the battle we have with education authorities much easier. Starting to understand why, and what she'll be able to do, and not, is a big help."

Several children have been diagnosed by genome sequencing in the US, including one who, as a result, was successfully treated for a bowel disorder with a bone marrow transplant. Katie is the first child in Britain to benefit.

Katie has a condition called craniosynotosis, which causes sections of her skull to fuse early so there is insufficient room for her brain to grow. She has had two operations to relieve pressure on her brain, one when she was just seven months old. The precise cause was unknown, making it difficult for her doctors to give a prognosis.

Though the NHS does not yet provide genome sequencing for unexplained disorders, Katie was referred to Andrew Wilkie, a consultant clinical geneticist at the University of Oxford who specialises in craniofacial disorders. Professor Wilkie is involved in an Oxford research project supported by Illumina, a DNA sequencing company, which is sequencing 500 genomes of people with serious diseases and their

THE GENOME REVOLUTION

Falling cost of genome sequencing by year

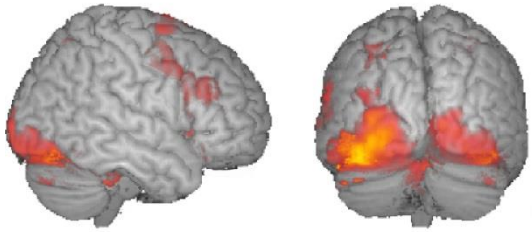
- 2000 — $100,000,000
- 2001 — First draft of the human genome sequence published
- Launch of imatinib (Glivec), the first "smart" cancer drug to target a specific mutation
- 2002 — Genome of MRSA sequenced
- $75,000,000
- 2003 — Mutation in the BRAF gene found in half of all cases of malignant melanoma
- Faster and cheaper sequencing launched
- 2004 — $40,000,000
- Launch of first next-generation sequencing techniques, which are dramatically faster and cheaper than Sanger sequencing
- 2005 — $20,000,000
- 2006 — Personal genomics companies begin selling genetic data directly to consumers
- 2007 — Craig Venter, the geneticist, becomes first named individual to have whole genome sequenced
- $2,000,000
- 2008 — Scientists at University of Washington find the cause of Miller syndrome
- 2009 — $100,000
- Child with bowel disorder is

Big data is...

- Large, population-scale data sets typically made possible by innovations in high throughput technology
  - Genome sequencing
  - Mobile and internet technology
  - Imaging and automated image processing

- Data sets that are large, high dimensional, semi-structured and highly heterogeneous
  - Large, requires distributed and cloud computing
  - Cannot be stored in standard database structures
  - Hard to summarise / visualise
  - Collected in many different ways by many different agents
  - Requiring new statistical and computational methods to analyse

Medical big data sources

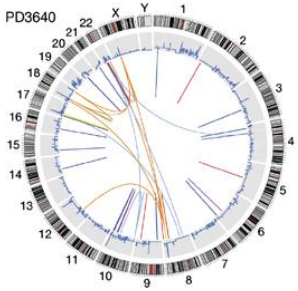


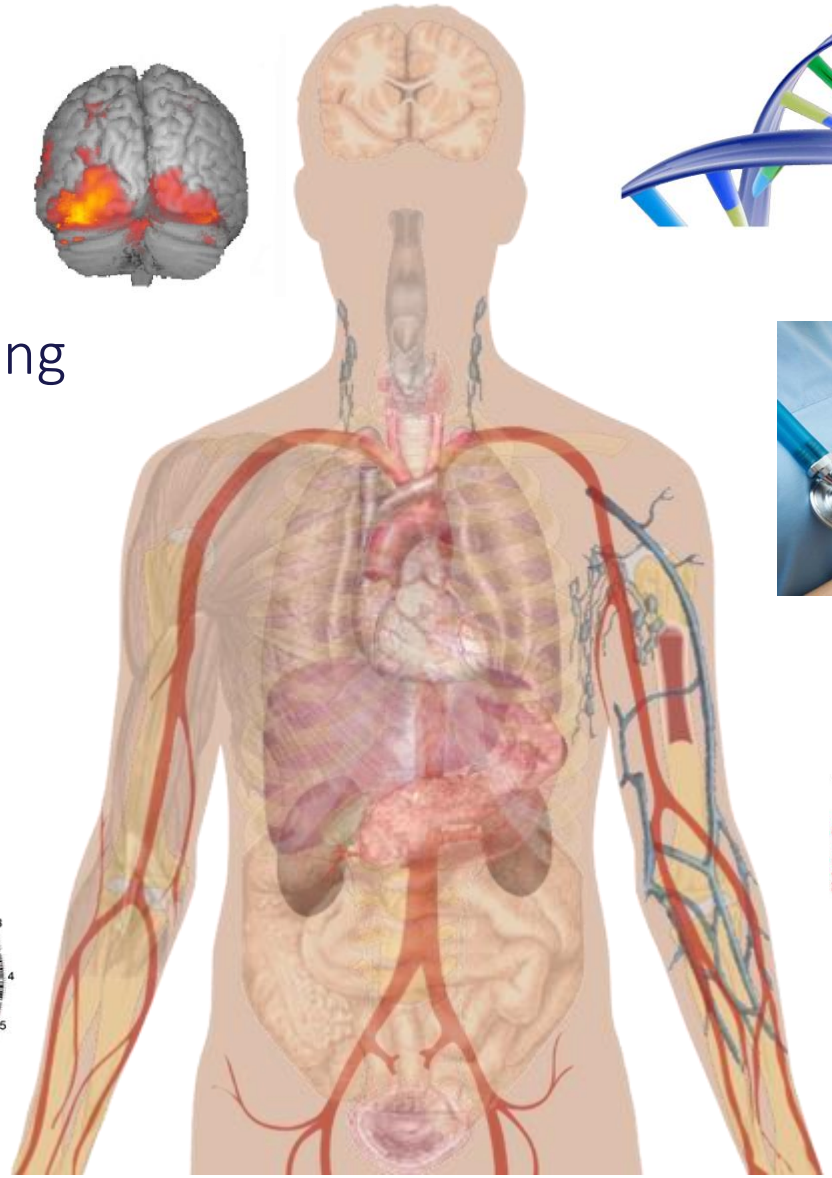
Imaging

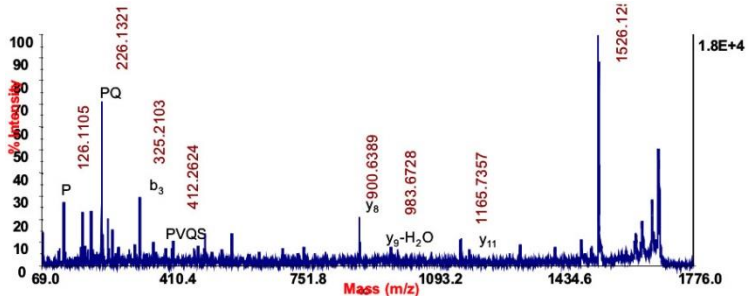
Germ-line genome

Electronic medical records

Tumour profile

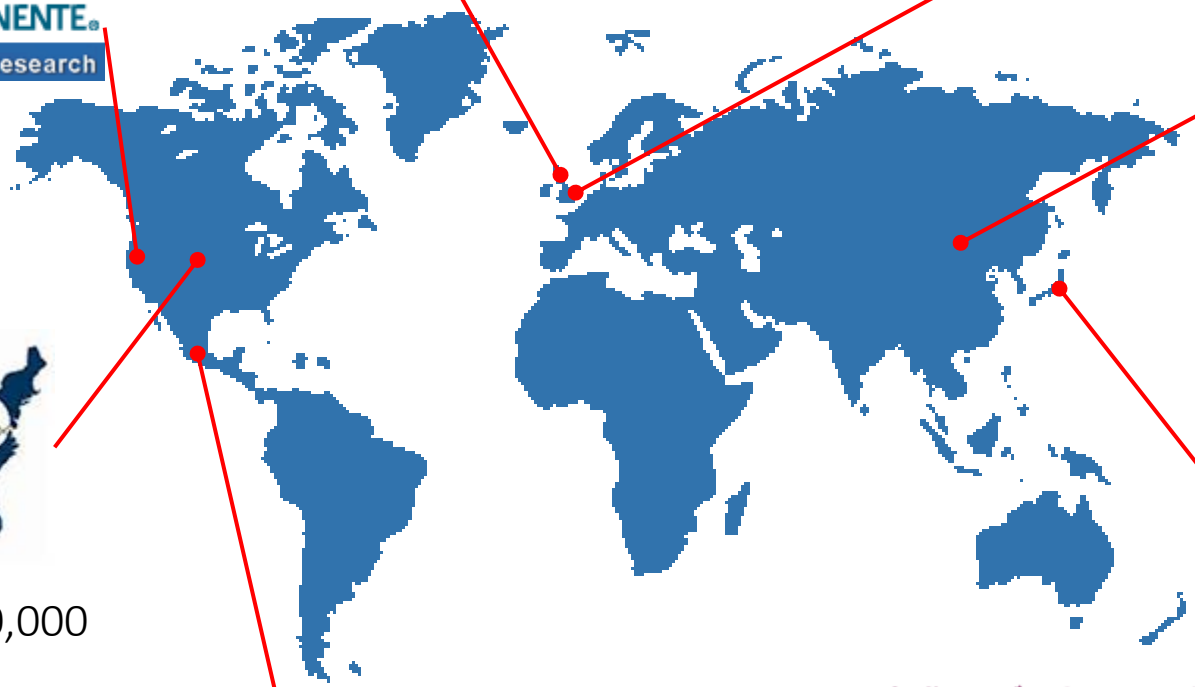






Biochemistry

Population-scale medical cohorts with genomic data



Genomics england

100,000

biobank uk

Improving the health of future generations

500,000

KAISER PERMANENTE

The Center for Health Research

500,000

CHINA KADOORIE BIOBANK

中国慢性病前瞻性研究

500,000

MVP

100,000

東北メディカル・メガバンク機構

TOHOKU MEDICAL MEGABANK ORGANIZATION

200,000

150,000

The Oxford Big Data Institute

Germ-line genomics

Infectious disease surveillance

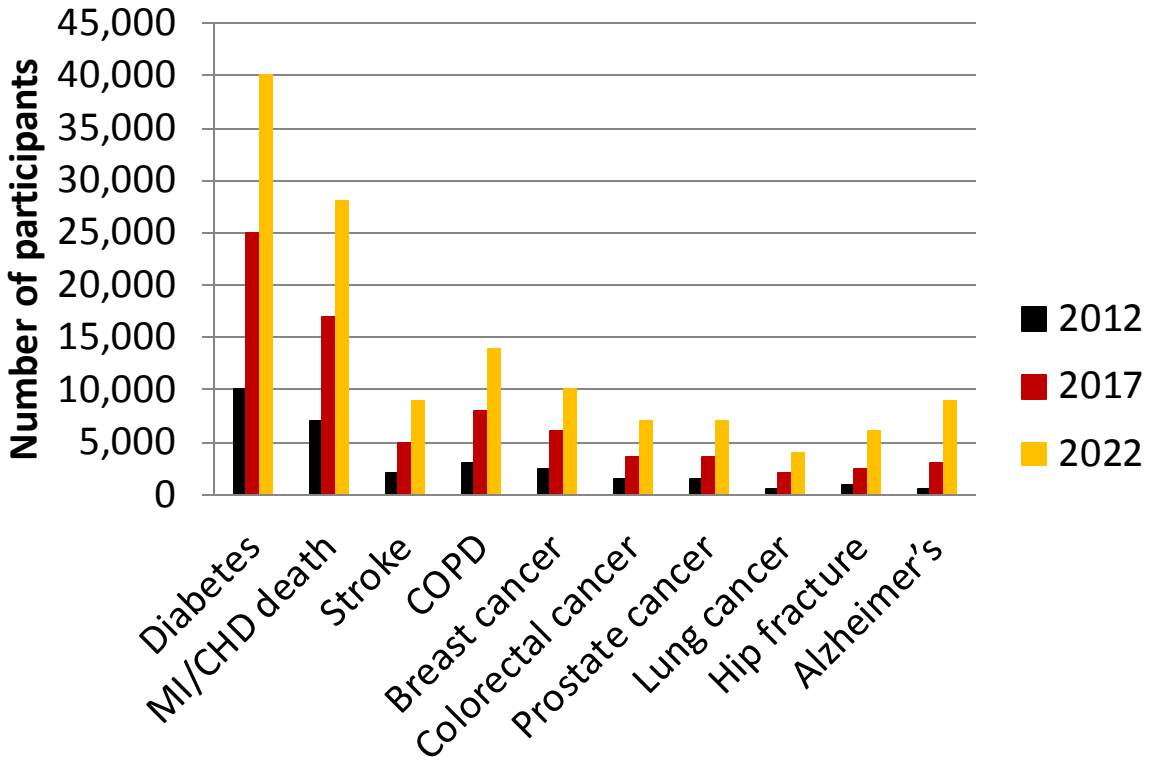
Electronic Medical Record linkage

Deep phenotypes in Large cohorts

biobank uk

Improving the health of future generations

Big data and epidemiology



biobank uk
Improving the health of future generations

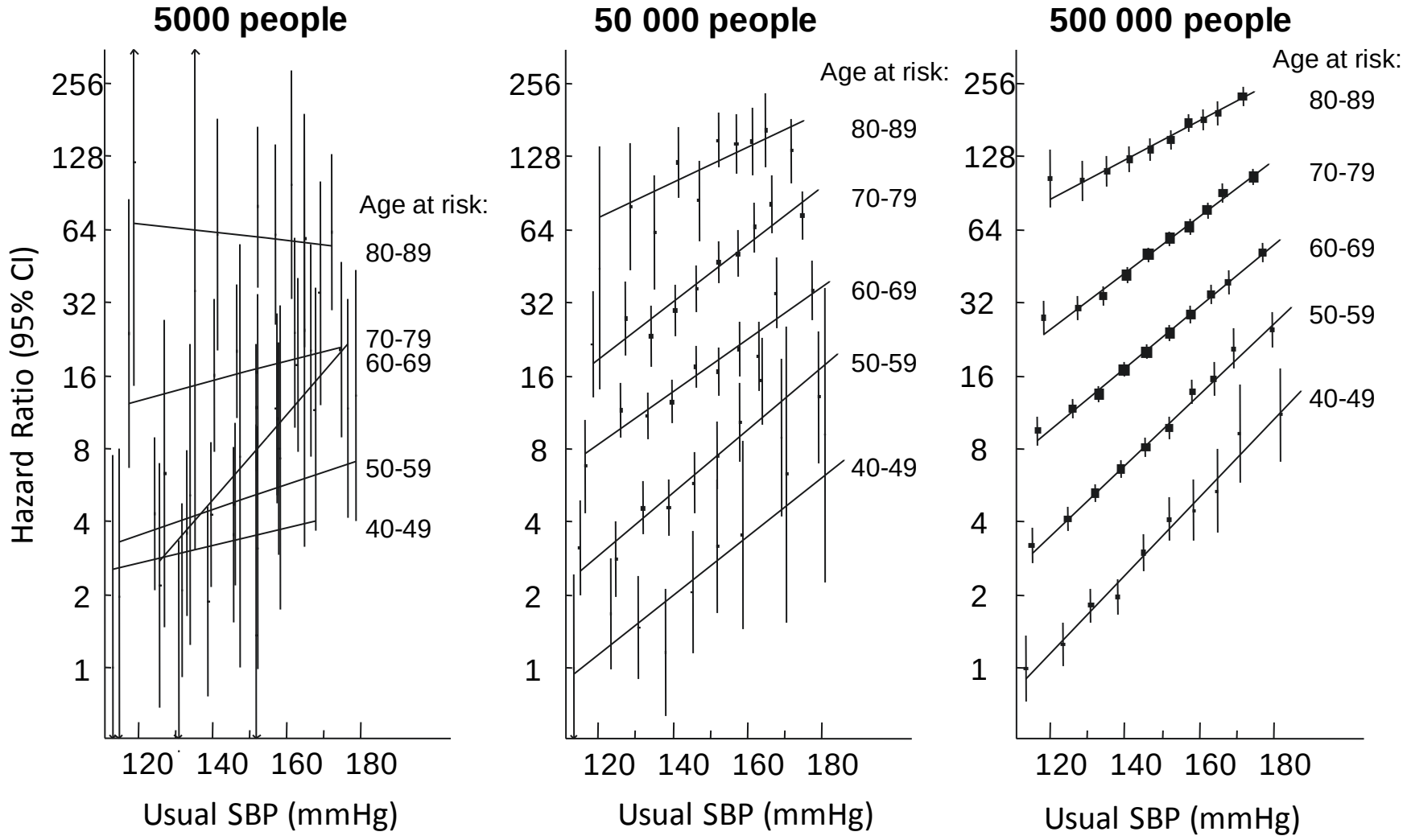
wellcome trust
Northwest Regional Development Agency
DH Department of Health
The Scottish Government
Llywodraeth Cynulliad Cymru Welsh Assembly Government
British Heart Foundation

Enhanced phenotyping

- Web-based diet questionnaires on 300,000
  - cognitive assessments planned
- Repeat assessments on 20,000
  - further repeat every few years
- Wrist-worn accelerometers on 100,000
- Standard panel of laboratory assays + genotyping on 500,000
- Imaging visit in 100,000
  - to include whole body and brain MRI, carotid ultrasound, bone (DEXA)
- On-going linkage to additional data sources
  - hospital, primary care, disease register
  - environment

biobank uk
Improving the health of future generations

The value of large numbers: Ischaemic heart disease and systolic blood pressure

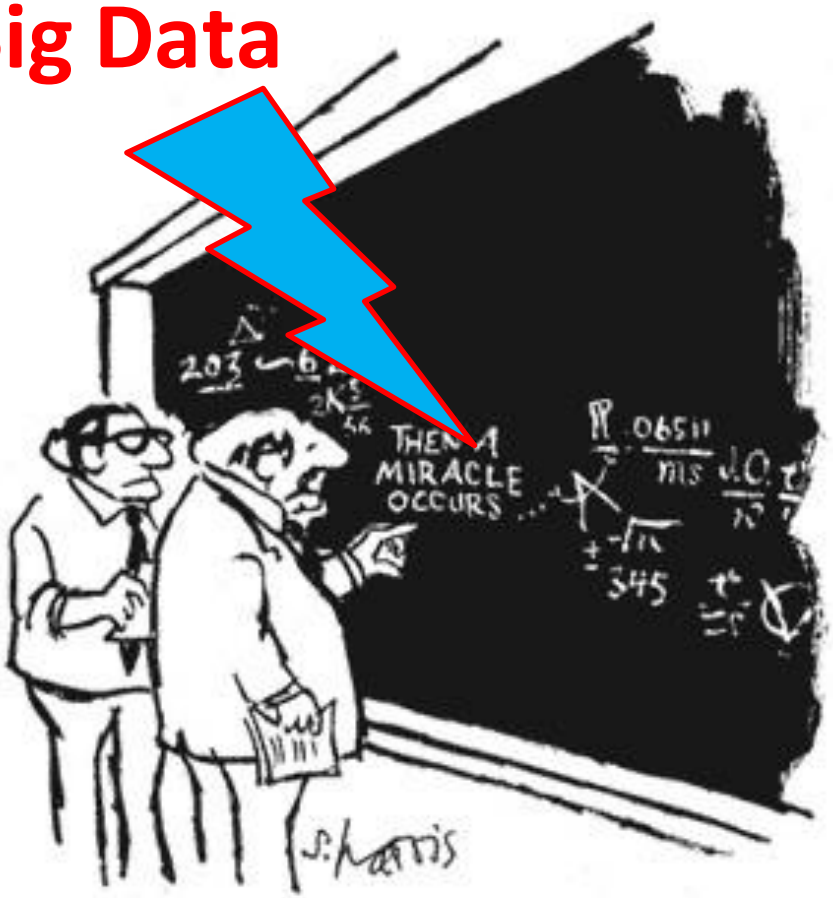


Courtesy of Prospective Studies Collaboration, unpublished

The promise of biomedical big data

- Better diagnosis
- Better treatment choice
- Improved target discovery
- Improved target validation
- Better outcomes

Big Data



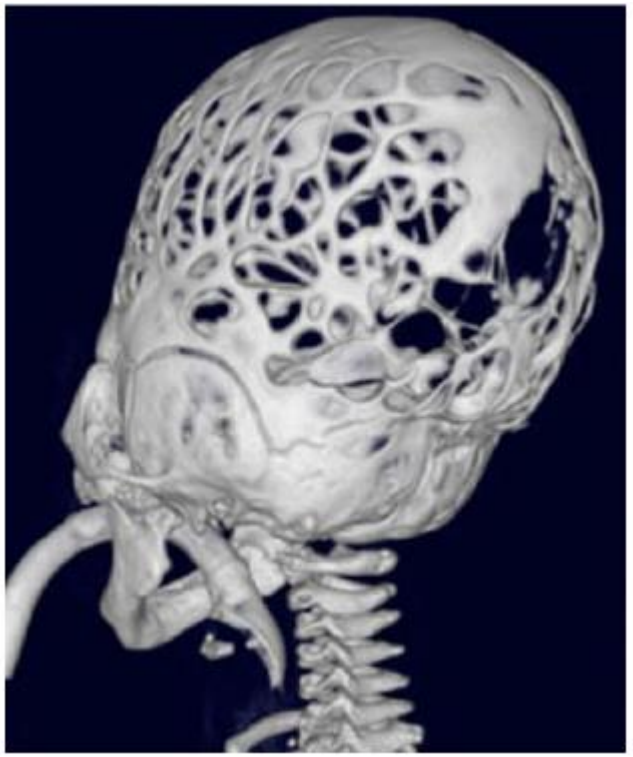
"I THINK YOU SHOULD BE MORE EXPLICIT HERE IN STEP TWO."

Putting genomics at the heart of Big Data

- Genetics has a direct and unequivocal relationship to phenotype
  - Germ-line gives exposure from birth
  - Finding the gene immediately informs about the patient
- Genomic data is accurate and easy to collect on a population scale
  - SNP genotype data -> GWAS on 100,000s
  - Sequencing at a population scale
- Genomic data can be used to probe causal relationships between biomarkers and disease
  - Natural variation mimics pharmaceutical interventions
- Genomic data provides a fingerprint for monitoring infectious disease control programmes
  - Can established transmission networks at micro and macro scale

Big data in the clinic

Genome sequencing as a clinical tool



The value of whole genome sequencing

- Whole-genome sequence is the only data type that can detect all types of information relevant to pathology in a single go:

Data type	Large-scale structural changes	Balanced translocations	Distant consanguinity	Uniparental disomy	Novel / known coding variants	Novel /known non-coding variants
Targeted gene sequencing	No	No	No	No	Yes	No
SNP arrays	Yes	No	Yes	Yes	No	No
Array CGH	Yes	No	No	No	No	No
Exome	Partial	No	Partial	Partial	Yes	No
Whole genome	Yes	Yes	Yes	Yes	Yes	Yes

WGS500 – initiated in 2011

- Collaboration

The Wellcome Trust Centre for Human Genetics | Nuffield Department of Clinical Medicine | University of Oxford

Oxford Biomedical Research Centre

Oxford University Hospitals NHS Trust

illumina

- Sequence 500 genomes at 30x

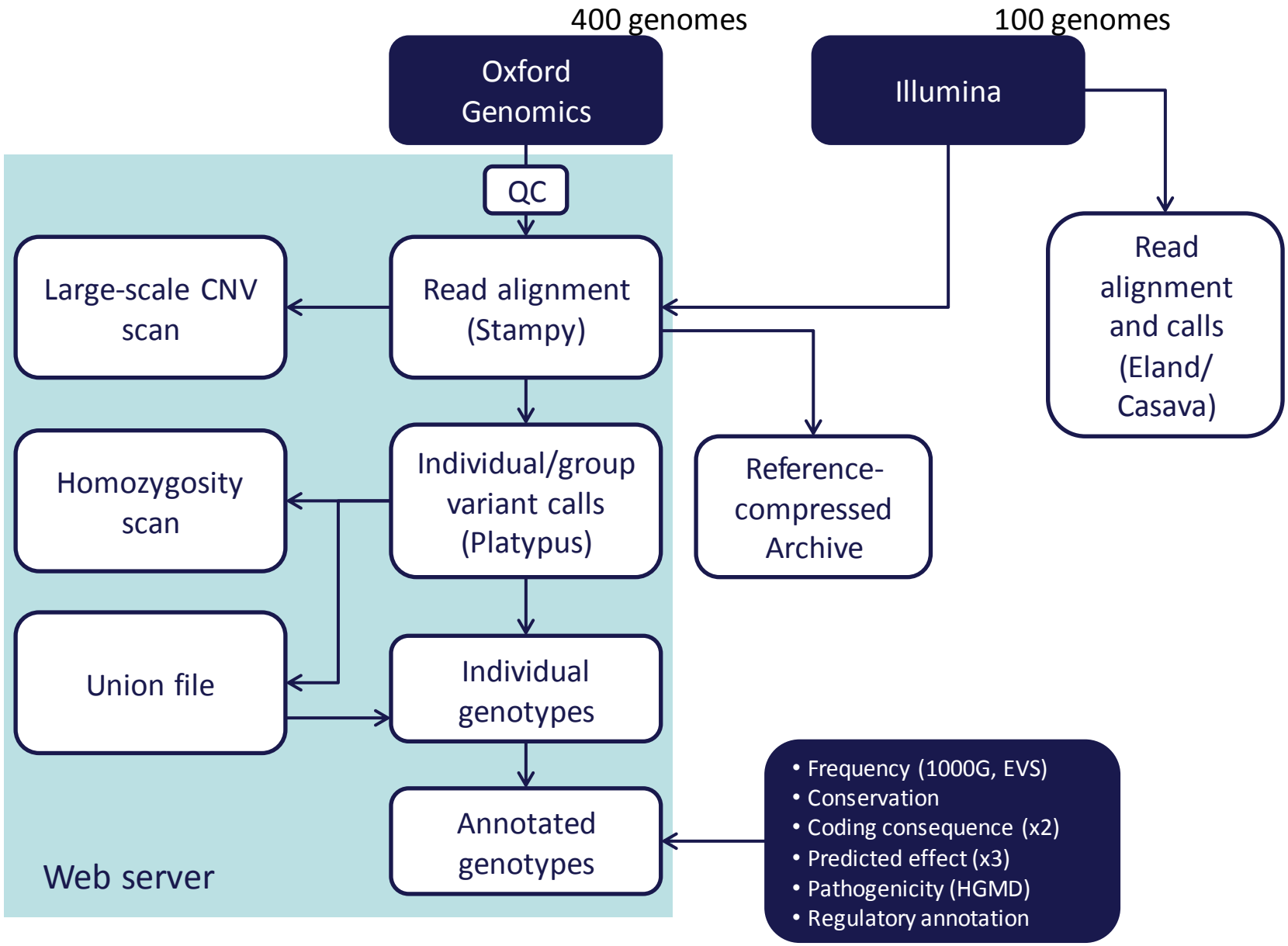
- Diverse set of diseases (>40 phenotypes)
  - Mendelian disorders (without mutation in screened genes)
  - Sporadic (extreme) immune phenotypes
  - Cancers

- Diverse set of experimental designs
  - Familial: Linkage information, trios, quartets
  - Cancer: Tumour-normal, metastases, multiple-mets, ..

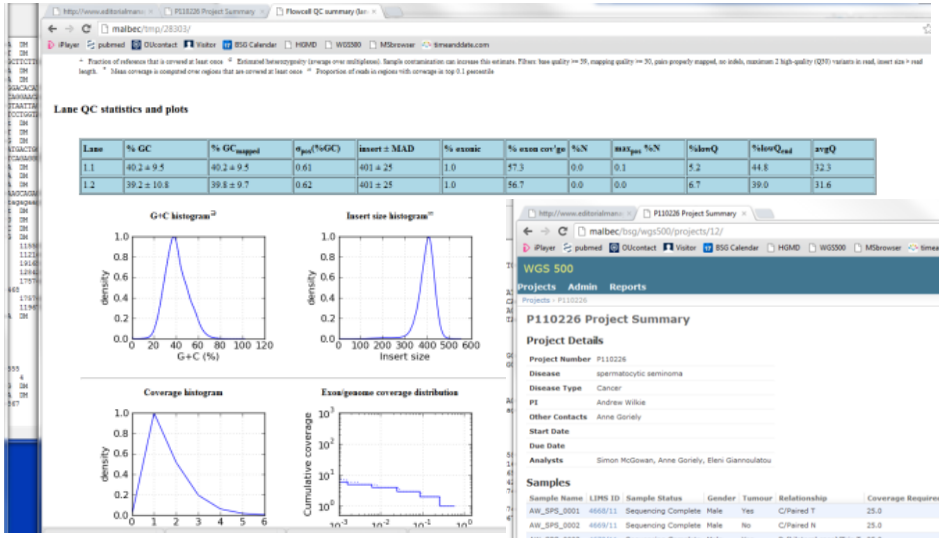
- Substantial follow-up (screening and functional) to establish candidacy

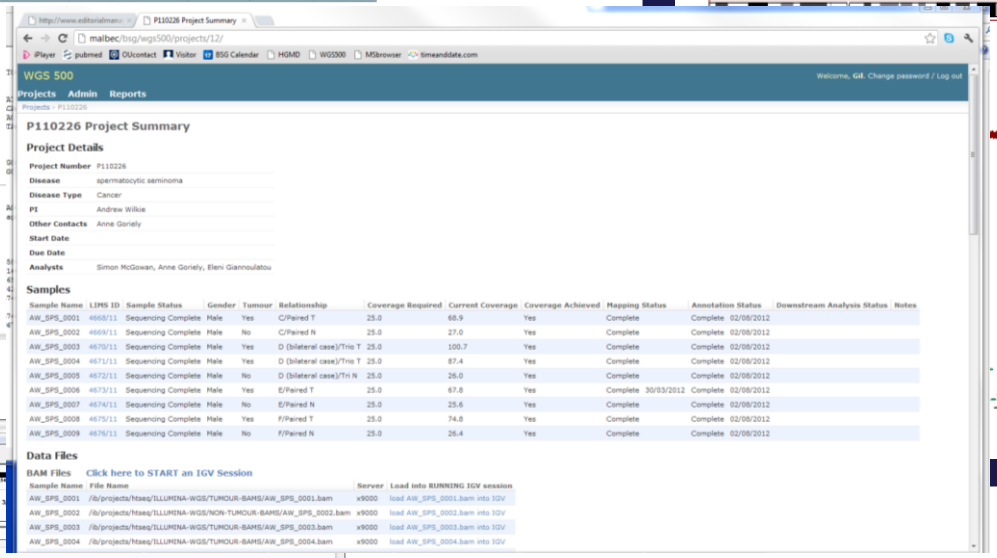
What is big data about this study?

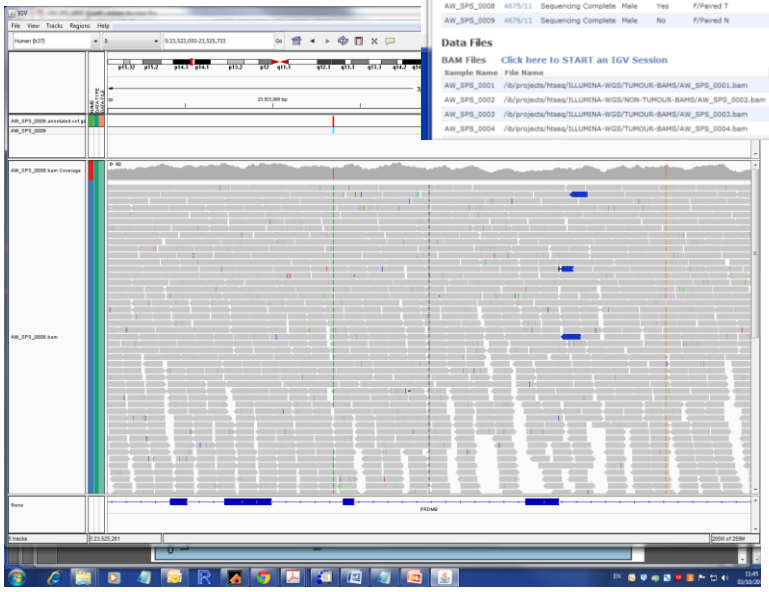
1. Infrastructure



2. Standards for data exchange and alignment







WGS500 Project Report on SPS June 2012

1. Disease: Spermatocytic Seminoma

PI: Andrew Wilkie (awilkie@hammer.ox.ac.uk)
Analysts: Anne Goriely (anne.goriely@imm.ox.ac.uk) and Eleni Giannoulatou (eleni.giannoulatou@imm.ox.ac.uk)
Last update: 02/03/2012

2. Experimental design and Strategy for identifying candidates

Spermatocytic seminoma (SPS) is a rare germ cell tumour that is slow growing and occurs specifically during adulthood in older males. We originally sequenced 9 samples, including 5 frozen tumour samples (SPS1, SPS6, SPS8) (one of them being a bilateral case (SPS3, SPS4)) and their 4 matched controls (2 from blood genomic DNA (SPS2, SPS5) and 2 from DNA extracted from normal tissue adjacent to the tumours (SPS7, SPS9)). However, it was noted that one of the presumed 'bilateral' sample SPS3 had many more discordant calls than its SPS4 counterpart, suggesting a sampling mismatched. Typing of a few SNPs for SPS3 suggest that the sampling error originate upstream of the WGS process. As a result, the 'trio' has been re-genotyped as a pair (SPS4/SPS5) and SPS3 is therefore an unmatched sample.

3. Use of heterogeneous external data

OMIM
Online Mendelian Inheritance in Man
Johns Hopkins University

ENCODE

HGMD
The Human Gene Mutation Database
Cardiff

DECIPHER
GRCh37

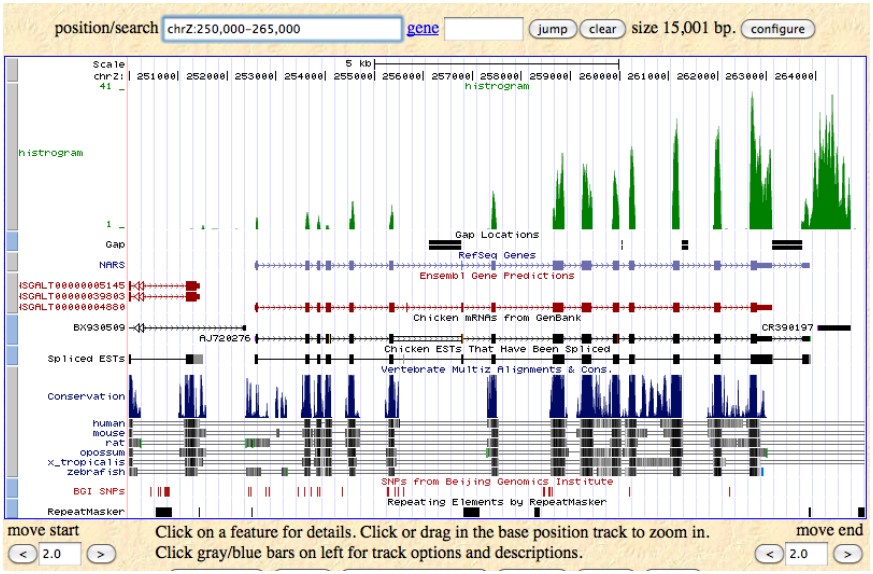
UK 10K
RARE GENETIC VARIANTS IN HEALTH AND DISEASE

PHAST
PHYLOGENETIC ANALYSIS WITH SPACE/TIME MODELS

Mutation Taster

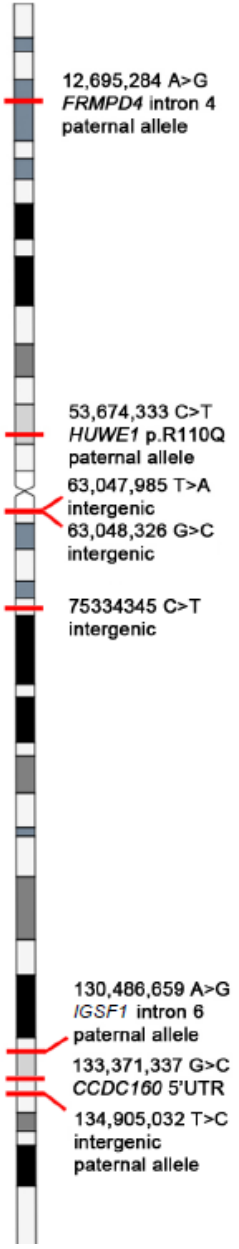
A THOUSAND GENOMES
Pilot studies prepare the way for population-scale gene sequencing

LOVD 3
Leiden Open Variation Database

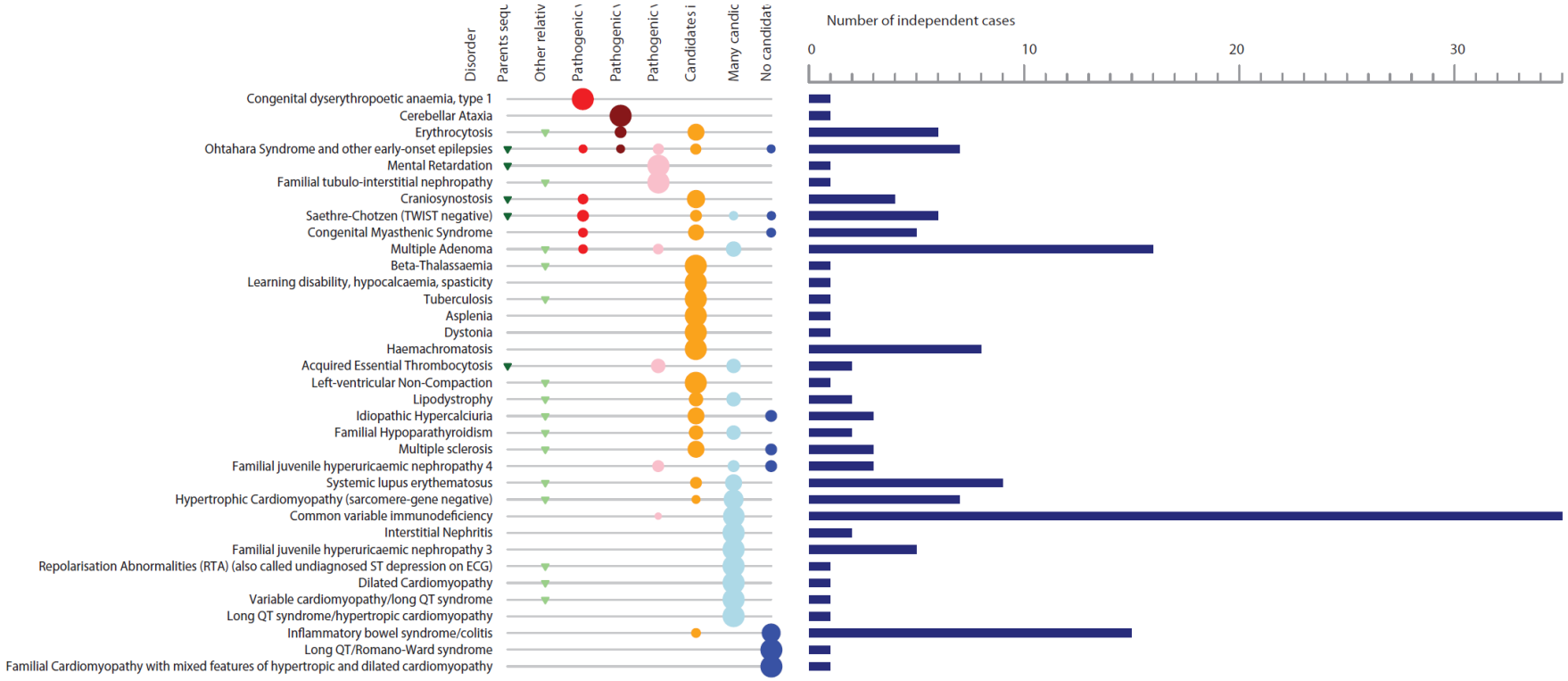


What did we find?

- Sequenced child and both parents
- De novo mutation in HUWE1, a known mental retardation gene
- Skewed inactivation on the X chromosome towards chromosome with mutation
- Multiple additional de novo mutations
- No other HUWE1 cases in >100 additional cases screened



Over 25% of cases with clear diagnosis



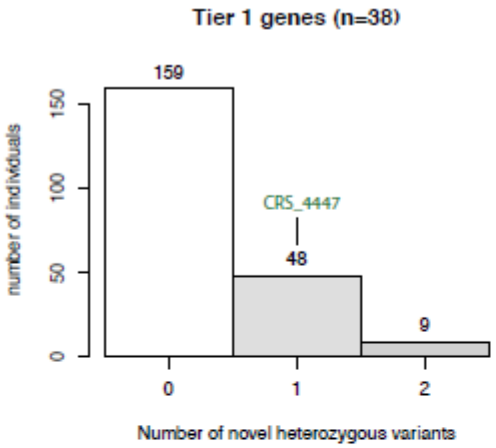
What limits success?

1. Ability to predict biological consequence

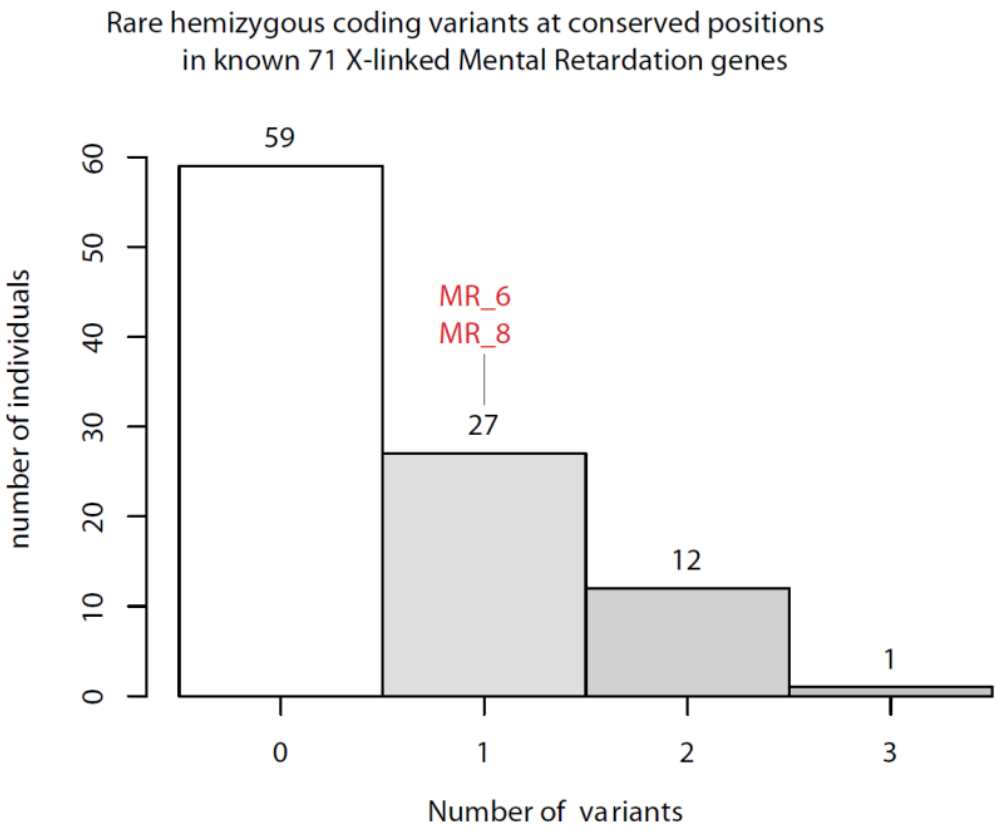
Known disease genes

Protein-protein interactions and related disorders

Same pathway



Do 40% of males have mental retardation?



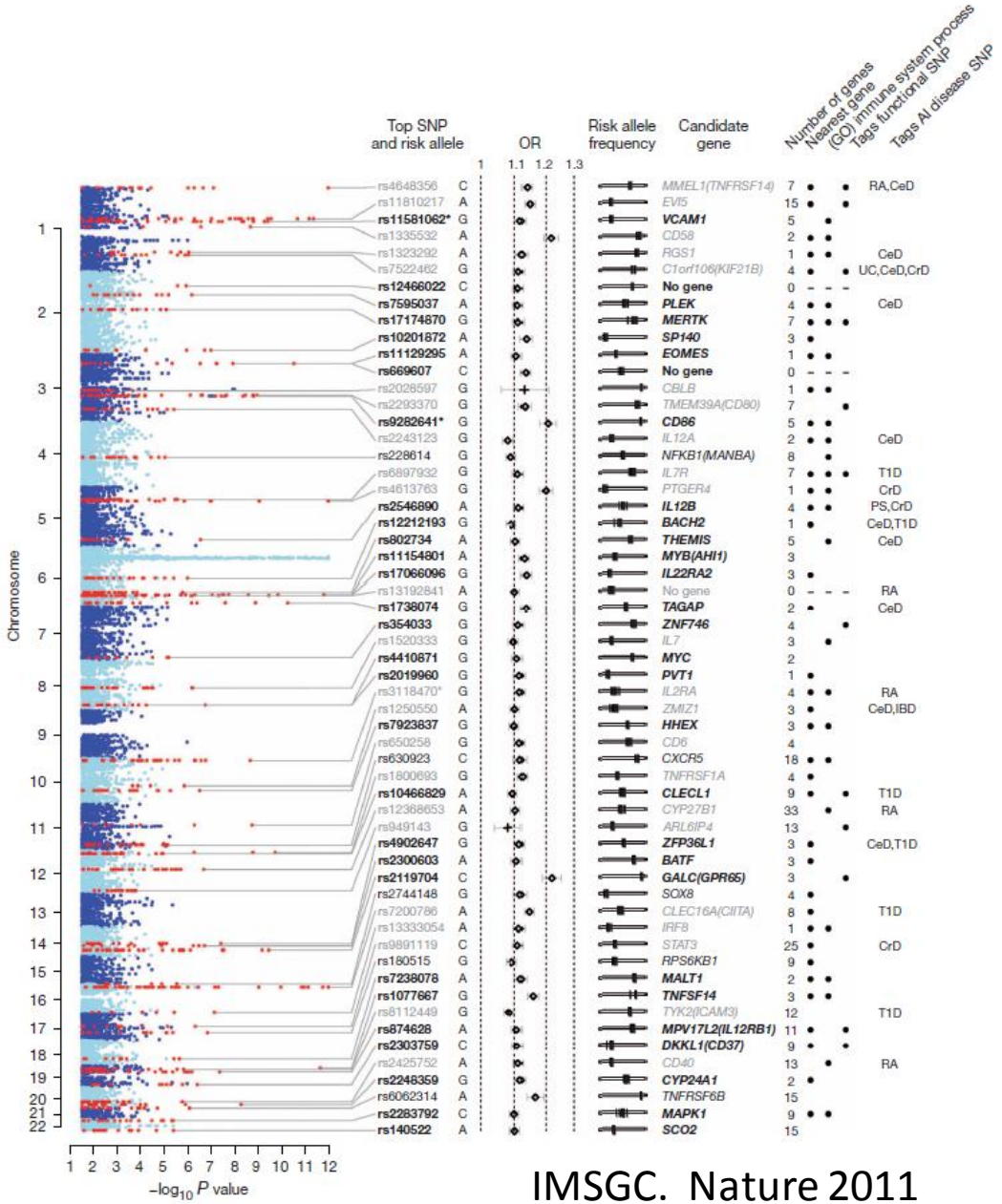
2. Sample size: The Genomics England 100k project

- 100,000 genomes sequenced in rare disease and cancer by 2017
- Linkage to medical records

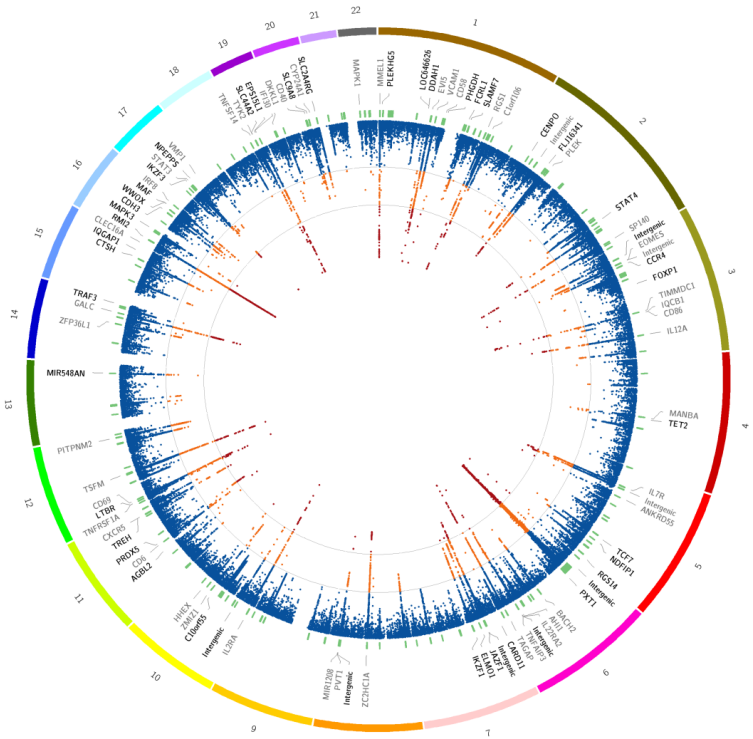
Global Alliance for Genomics and Health

What will large data sets deliver?

1. A (growing) understanding of complex disease mechanisms



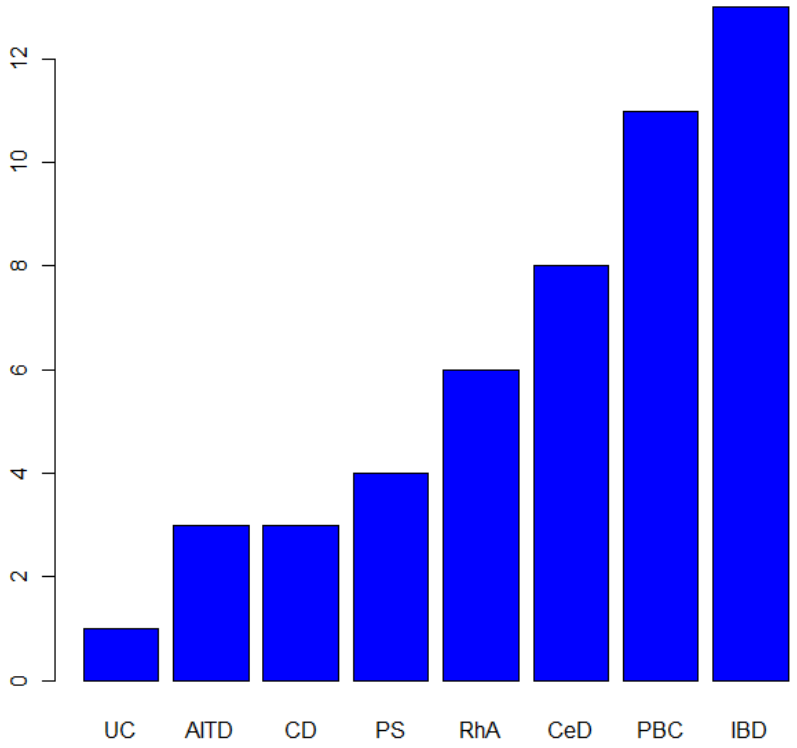
IMSGC. Nature 2011



IMSGC. Nature Genet. 2013

What have we learned?

- About 113 current risk loci for multiple sclerosis
- Explains c. 30% sibling recurrence risk (10% of that is HLA)
- C. 30% of loci overlap with other autoimmune diseases



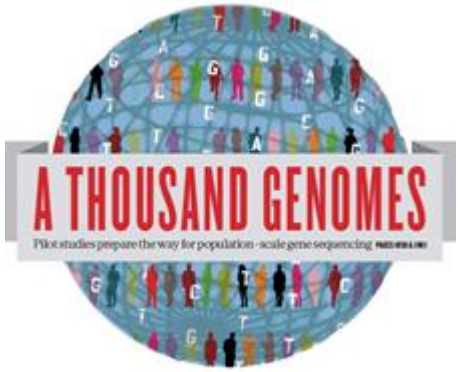
Can we map causal variants from GWAS?

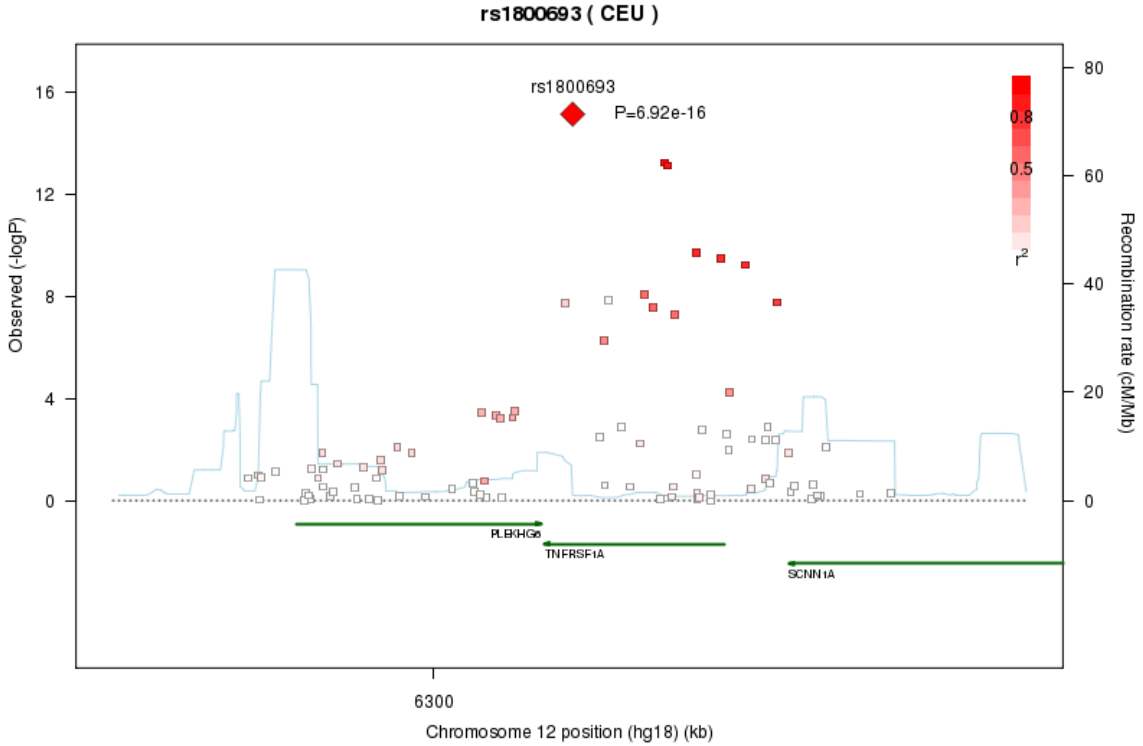
ARTICLES

nature genetics

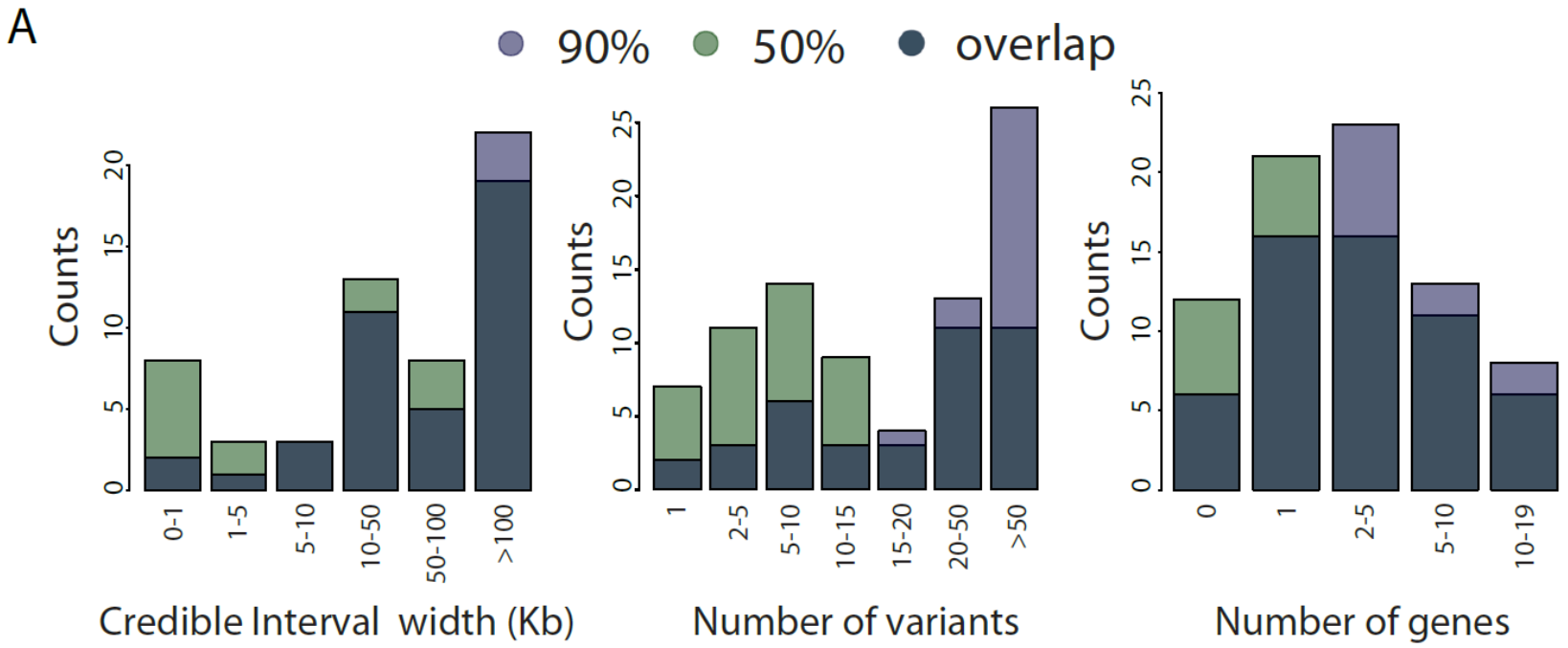
Bayesian refinement of association signals for 14 loci in 3 common diseases

The Wellcome Trust Case Control Consortium[1,2]

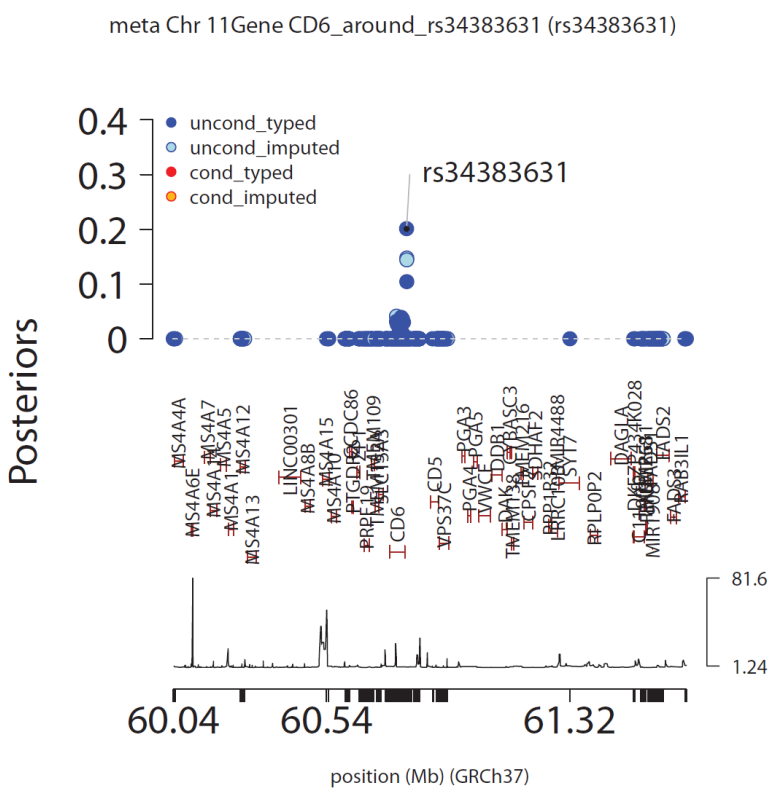


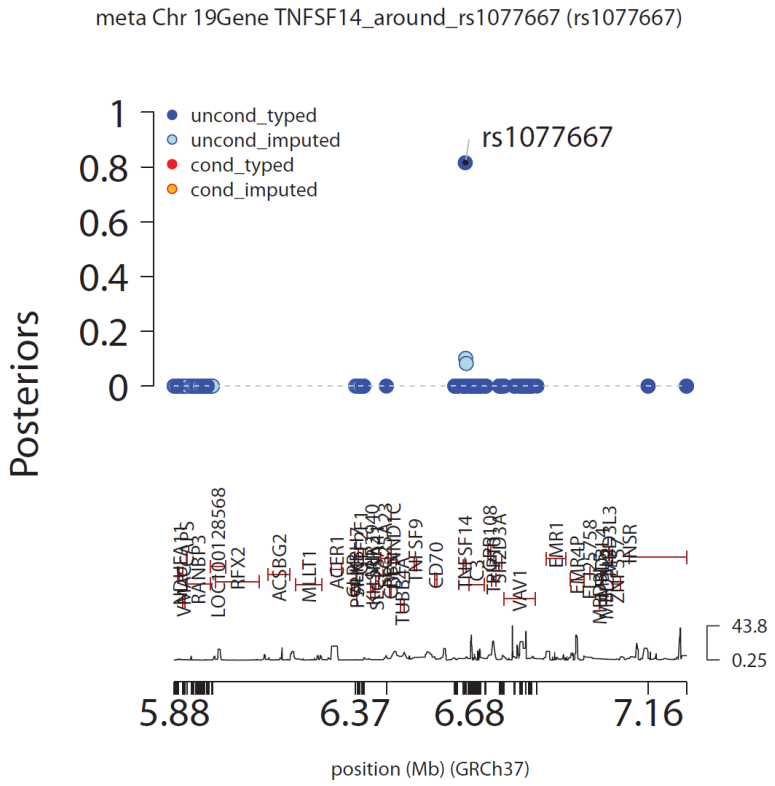


About 5% of signals fine-map to <5 variants



Occasional success at fine-mapping





IMSGC, 2013

But there is little indication of function….

Table 3 The 22 variants from the 8 regions with consistent high resolution fine-mapping

Gene	SNP	Chr	Position[a]	Posterior	GERP	Functional Annotation[b]
TNFSF14	rs1077667	19	6668972	0.81	-3.89	intronic, TFBS / DNase1 peak, correlates with serum levels of TNFSF14
	rs2291668[c]	19	6669934	0.10	-9.78	intronic / synonymous, TFBS / DNAase1 peak
IL2RA	rs2104286	10	6099045	0.99	-0.47	intronic, correlates with soluble IL-2RA levels
TNFRSF1A	rs1800693	12	6440009	0.70	2.53	intronic, causes splicing defect and truncated soluble TNFRSF1A
	rs4149580[c]	12	6446990	0.10	2.06	intronic
IL12A	rs1014486	3	159691112	0.79	0.24	-
CD6	rs34383631	11	60793330	0.32	1.66	-
	rs4939490[c]	11	60793651	0.23	-0.53	-
	rs4939491[c]	11	60793722	0.23	-0.37	-
	rs4939489	11	60793648	0.16	3.25	-
TNFAIP3	rs632574	6	137959118	0.27	-1.15	-
	rs498549[c]	6	137984935	0.20	0.52	-
	rs651973	6	137996134	0.17	2.41	downstream of RP11-95M15.1 lincRNA gene
	rs536331	6	137993049	0.15	0.19	upstream of RP11-95M15.1 lincRNA gene
CD58	rs6677309	1	117080166	0.21	-1.18	intronic, TFBS / DNase1 peak
	rs35275493[c]	1	117095502	0.24	0.75	intronic (insertion)
	rs10754324[c]	1	117093035	0.22	0.32	intronic
	rs1335532	1	117100957	0.17	-1.32	intronic
STAT4	rs9967792	2	191974435	0.35	-3.96	intronic
	rs10197066[c]	2	191985459	0.21	0.05	intronic
	rs10804037	2	191991891	0.21	-0.36	intronic
	rs71301540[c]	2	192001443	0.20	0.08	intronic (deletion)

All listed variants have posterior ≥ 0.1 in regions where ≤ 5 variants explain the top 50% of the posterior and the top SNP from the frequentist analysis lives in the 90% confidence interval, ordered by maximum posterior.
Posterior denotes the posterior probability of any variant driving association. GERP denotes Genomic Evolutionary Rate Profiling.
[a]Position is based on human genome 19 and dbSNP 137.
[b]Functional data from VEP, eQTL browser, Fairfax et al. (2012), pubmed searches, 1000G. Dash indicates intergenic with no additional annotation.
Variants without annotation are intergenic and have no reported regulatory consequence.
[c]Imputed variant.

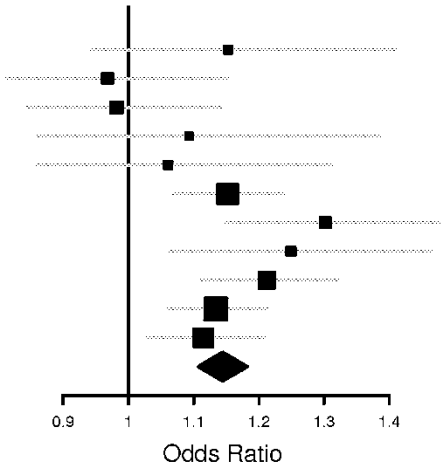
LETTER

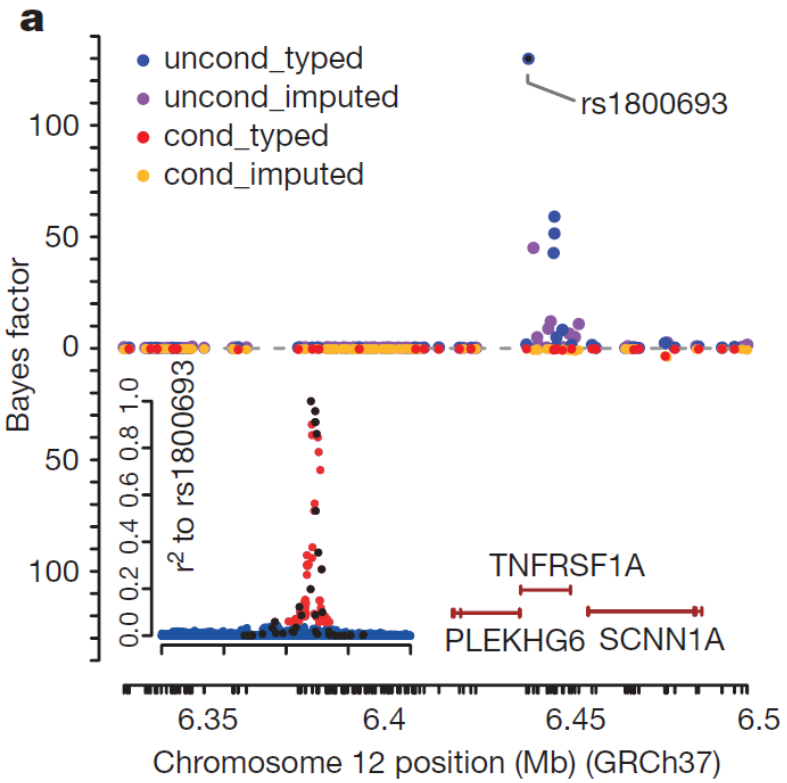
doi:10.1038/nature11307

TNF receptor 1 genetic risk mirrors outcome of anti-TNF therapy in multiple sclerosis

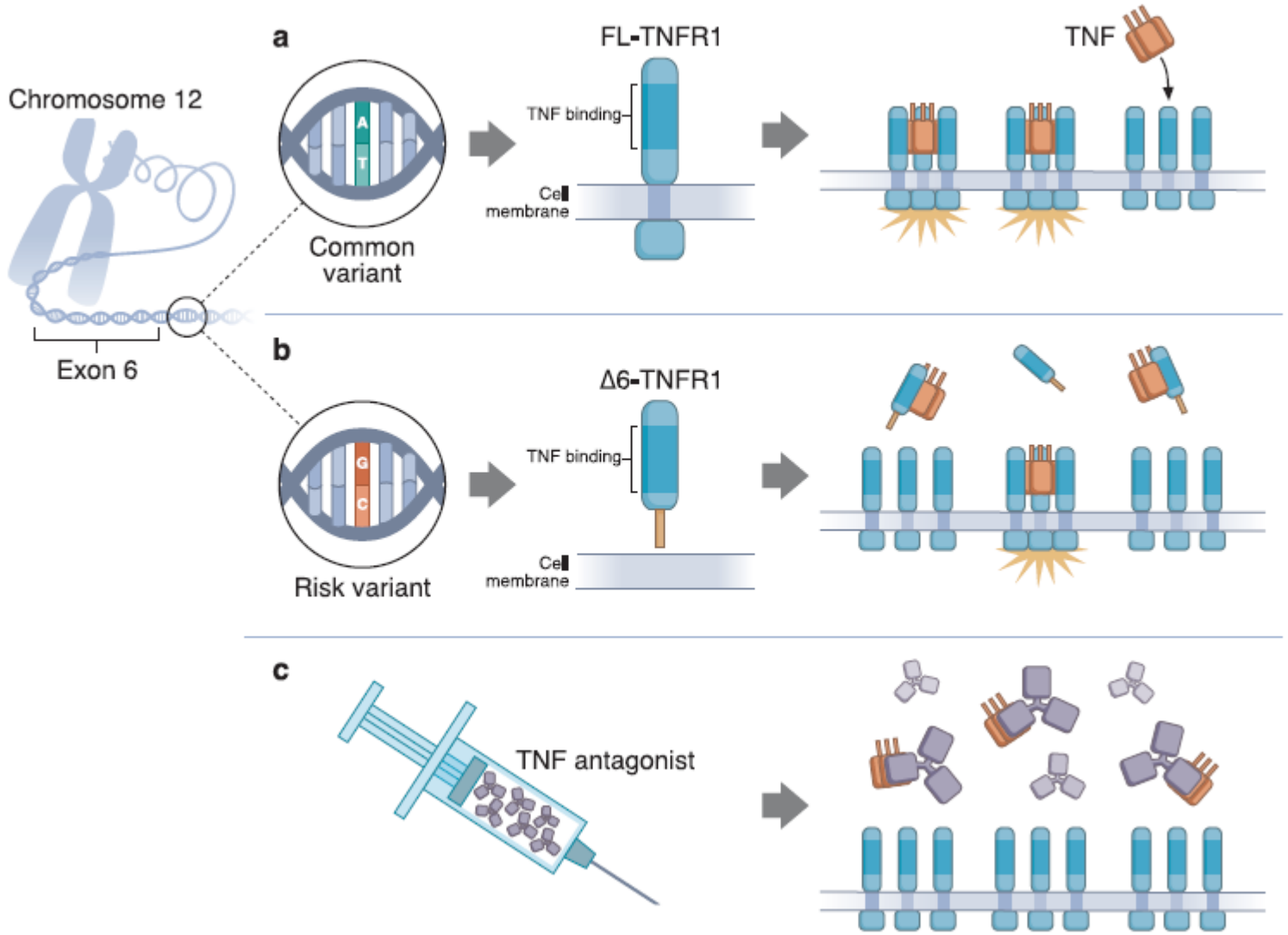
Adam P. Gregory[1]*, Calliope A. Dendrou[2]*, Kathrine E. Attfield[2], Aiden Haghikia[2,3], Dionysia K. Xifara[4], Falk Butter[5], Gereon Poschmann[6], Gurman Kaur[1], Lydia Lambert[2], Oliver A. Leach[2], Simone Prömel[2], Divya Punwani[1], James H. Felce[1], Simon J. Davis[1], Ralf Gold[3], Finn C. Nielsen[7], Richard M. Siegel[8], Matthias Mann[5], John I. Bell[9], Gil McVean[4] & Lars Fugger[1,2,10]

Stratum	OR	95% CI
AUSNZ	1.15	[0.94–1.41]
Belgium	0.97	[0.81–1.15]
Denmark	0.98	[0.84–1.14]
Finland	1.09	[0.86–1.39]
France	1.06	[0.86–1.31]
Germany	1.15	[1.07–1.24]
Italy	1.30	[1.15–1.48]
Norway	1.25	[1.06–1.46]
Sweden	1.21	[1.11–1.32]
UK	1.13	[1.06–1.21]
US	1.11	[1.03–1.21]
Summary	**1.14**	**[1.11–1.18]**





Gregory et al. (2012) Nature.



Fugger et al. (2012)

2. Greater ability to validate therapeutic targets

A case study: Darapladib

- Epidemiological evidence shows that people with lower levels of a particular enzyme (Lp-PLA2) have reduced risk of heart disease.
- GSK developed a drug, darapladib, which inhibits Lp-PLA2
- Common variants around the gene PLA2G7 are modestly associated with Lp-PLA2 levels

…but not with heart disease

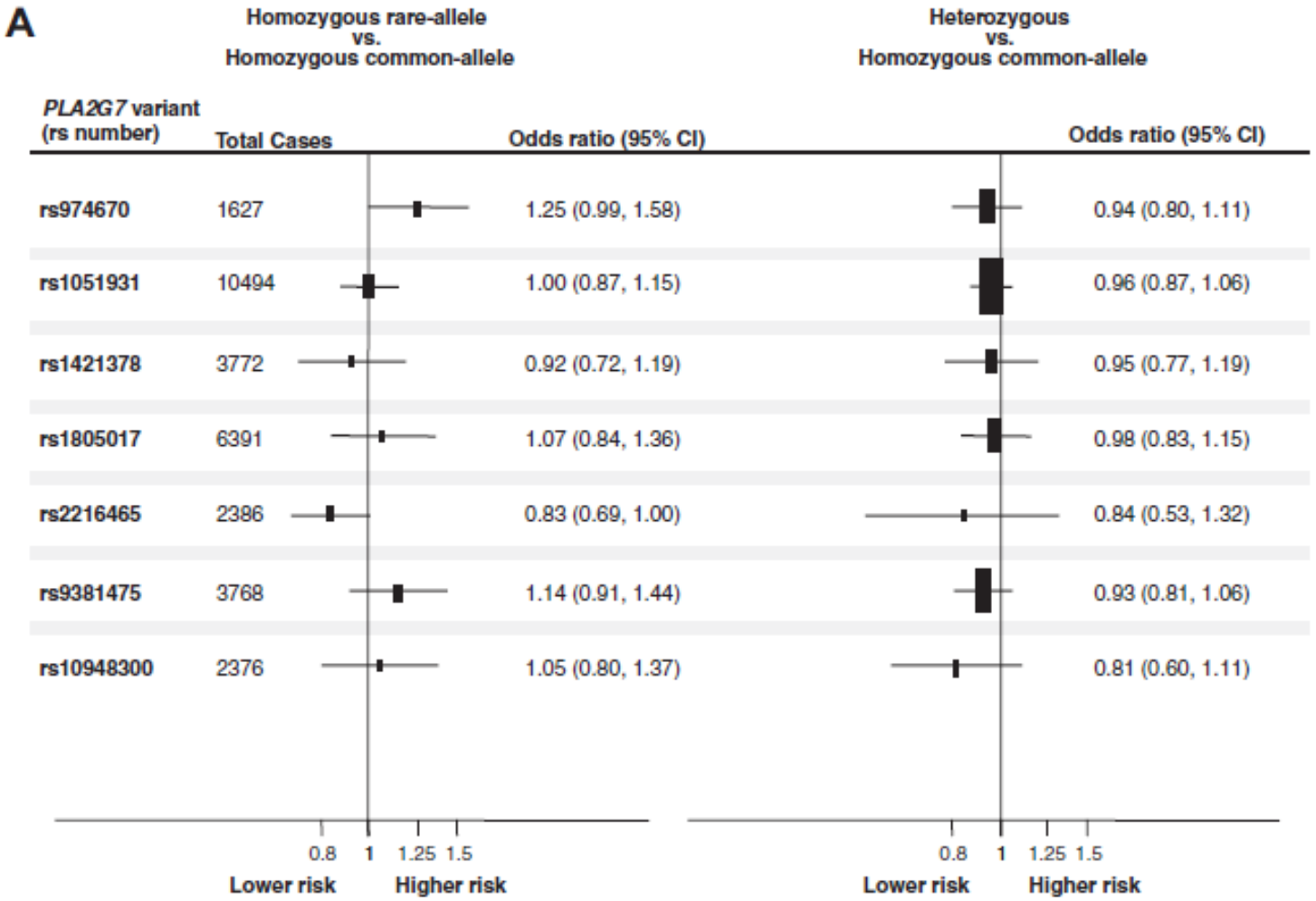
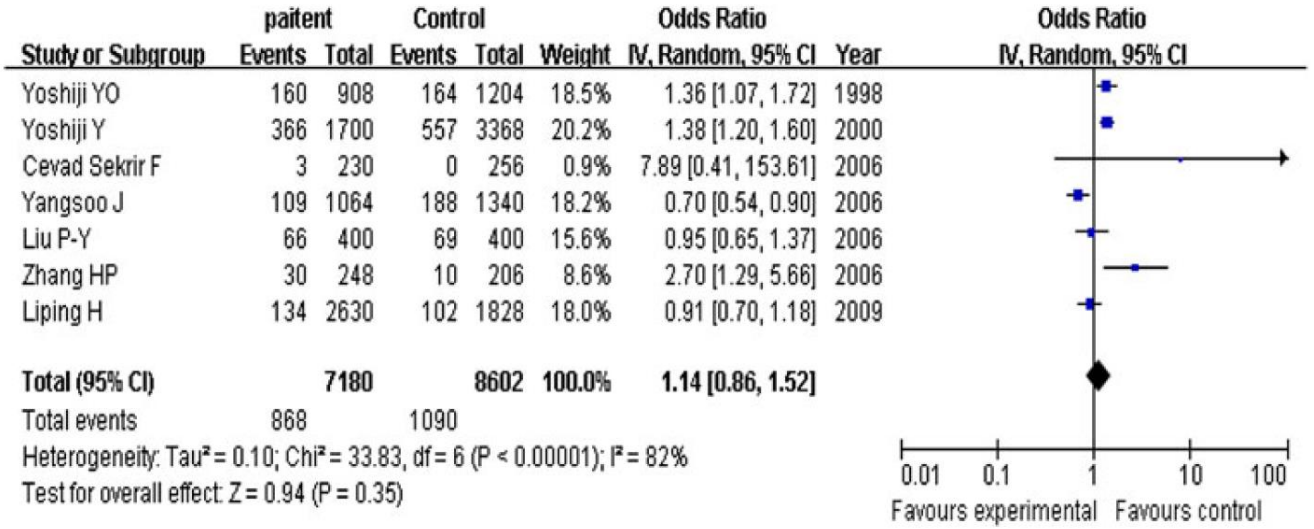


Figure 5. A, Relative odds of CHD associated with *PLA2G7* variants. Data are pooled from up to 10 studies (NPHS-II, EPIC-Norfolk, WH-II, HIFMECH, EAS, Athero*Gene*, LURIC, Cyprus, SAS, and WTCCC-CHD) including up to 10 494 CHD events. B, Effect of the

Casas et al. Circulation 2010

LOF variant in PLA2G7 creates “Lp-PLA2 human knockouts” but is still not associated with risk of CVD

Fig. 2 Results of association between the V279F polymorphism in PLA2 gene and coronary heart disease under the additive model



Study or Subgroup	paitent Events	Total	Control Events	Total	Weight	Odds Ratio IV, Random, 95% CI	Year
Yoshiji YO	160	908	164	1204	18.5%	1.36 [1.07, 1.72]	1998
Yoshiji Y	366	1700	557	3368	20.2%	1.38 [1.20, 1.60]	2000
Cevad Sekrir F	3	230	0	256	0.9%	7.89 [0.41, 153.61]	2006
Yangsoo J	109	1064	188	1340	18.2%	0.70 [0.54, 0.90]	2006
Liu P-Y	66	400	69	400	15.6%	0.95 [0.65, 1.37]	2006
Zhang HP	30	248	10	206	8.6%	2.70 [1.29, 5.66]	2006
Liping H	134	2630	102	1828	18.0%	0.91 [0.70, 1.18]	2009
Total (95% CI)		**7180**		**8602**	**100.0%**	**1.14 [0.86, 1.52]**	
Total events	868		1090				

Heterogeneity: $Tau^2 = 0.10$; $Chi^2 = 33.83$, df = 6 ($P < 0.00001$); $I^2 = 82\%$

Test for overall effect: $Z = 0.94$ ($P = 0.35$)

Zheng et al. (2011)

An unhappy ending

- GSK have recently completed “STABILITY” a large clinical trial (c. $800M) of darapladib.

- On Tuesday, November 12, 2013, GSK announced that the drug had failed to meet Phase III endpoints in a trial of 16,000 patients with acute coronary syndrome. An additional trial of 13,000 patients (SOLID-TIMI 52) is ongoing.

What does genetics tell us to date about MS treatments?

Compound	Drugs	Mode of action	Relevant genes	Other uses	Comments	eQTLs	GWAS
Teriflunomide	Aubagio	Blocks dihydroorotate dehydrogenase. Inihibts pyrimidine de novo synthesis, hence rapidly dividing cells including activated T cells. Also blocks NF-kB and tyrosine kinases at high dose	DHODH (mutations cause Miller syndrome)	Inactive form (leflunomide) used in severe RA and psoriatic arthritis (also pyrimidine synthesis inhibitor)	Poor efficacy	None	None
Interferon beta-1a	Avonex, Rebif, CinnoVex	Is an interferon type I. Binds to IFN-alpha receptor (IFNAR1/IFNAR2). Cytokine (activate NK cells, macrophages, upregulate antigen presentation). Produced by leukocytes	IFNAR1, IFNAR2 (activate JAK/STAT, Tyk2, etc.). IFN-beta actually 3 products from 3 genes, IFNB1, IFNB3 and IL6 (also called IFNB2). IL6 secreted by CD4 Th cells	Chemotherapy	30% MS patients unresponsive	IFNB1 - no eQTLs reported. IL6 - LPS stimulation specific eQTLs	None
Interferon beta-1b	Betaferon, Extavia	See above	See above	See above	See above	See above	See above
Glatiramer acetate	Copaxone	Random polymer of 4 AA found in MBP	MBP	Dry ARMD (Phase 1)	Doesn't seem to be effective	NA	NA
Fingolimod	Gilenya	Sphingosine 1-phosphate receptor modulator, which sequesters lymphocytes in lymph nodes, preventing them from contributing to an autoimmune reaction	S1PR1 / EDG1	Candidate for heart failure and arrythmia	Effective treatment	None reported	None
Alemtuzumab	Lemtrada	Monocloncal that binds to CD52 on mature lymphocytes	CD52 (indirect)	Used for CLL	Very serious side effects. Efficacy questioned.	NA	NA
Dimethyl fumarate	Tecfidera	Attached by glutathione, which leads to HO-1 induction (anti-inflammatory). Possible up-regulation of NRF2. HO-1 upregulate IL10 and IL-1R	HMOX1 encode HO-1. IL10, IL1R1, IL1R2	Psoriasis, sarcoidosis, others	Effective treatment	HMOX1 - none reported. IL10 - rs3024490 and rs1554286. None reported for IL1 or receptors.	None. Though IL20 is structurally related to IL10 and is a MS GWAS signal. Note Ils strongly clustered.
Natalizumab	Tysabri	Monocloncal against a4-integrin	ITGA4 (CD49D)		Low risk of PML caused by reactivation of JC virus	ITGA4 has both eQTLs and multiple exon QTLs reported	None

..at best, confusing

Big data, big challenges

- Biomedical big data needs many components:
  - High throughput measurement
  - Large cohorts
  - Ease of data access
  - Powerful analysis
  - Appropriate governance
  - Engagement with patients
  - …
- Medical data has many complexities, but genetics provides a useful instrument to help disentangle causal and indirect associations.
- Much to do to integrate genetic and functional data across many diseases, tissues, cell-types, stimulations, etc.

With thanks to

- The WGS500 project team
- The IMSGC
- Lars Fugger and his group
- UK Biobank
- Members of the WTCHG