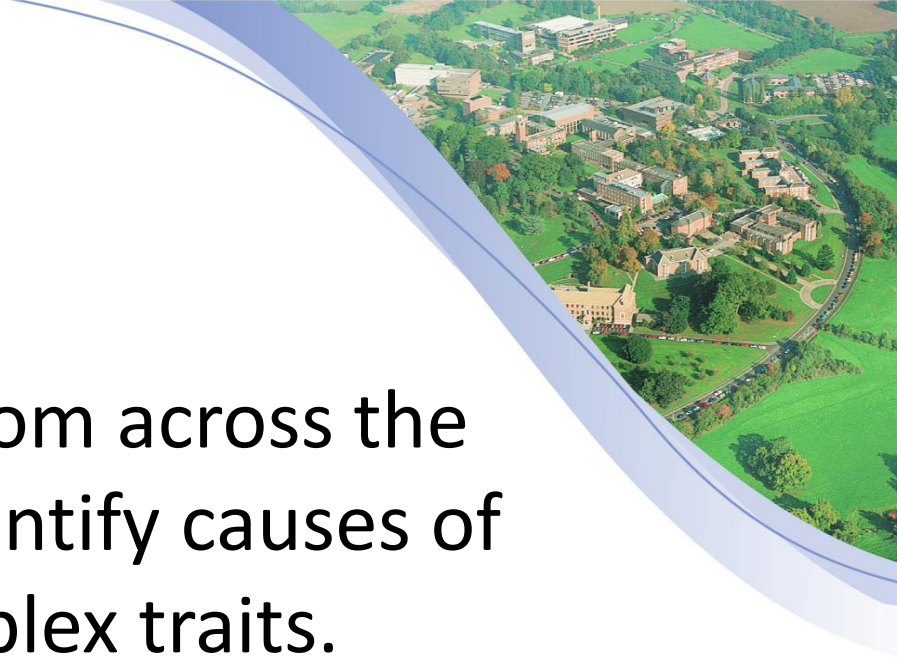
UNIVERSITY OF EXETER | MEDICAL SCHOOL

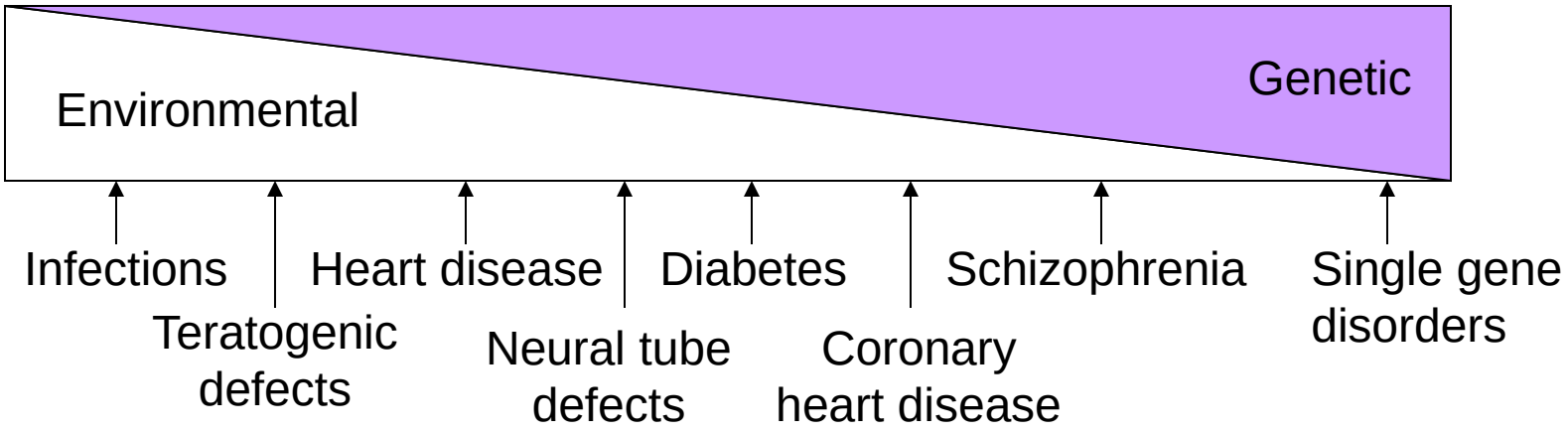


Using genetic data from across the human genome to identify causes of disease and complex traits.

Anna Murray
Senior Lecturer in Human Genetics
University of Exeter Medical School
Wellcome Trust Biomedical Informatics Hub
A.Murray@exeter.ac.uk

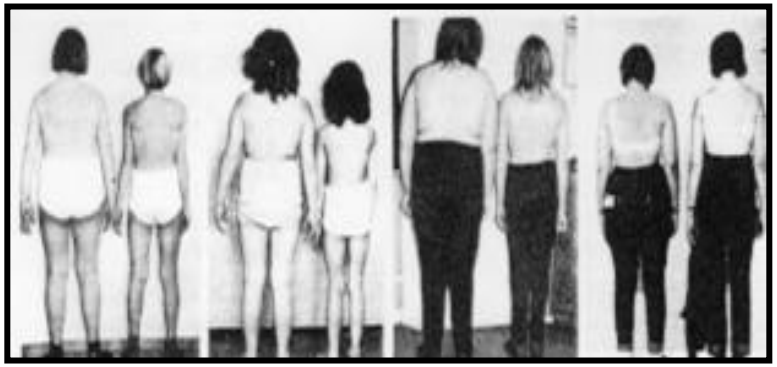
Many diseases and traits have a genetic cause

- Many common diseases, which have a huge impact on human health, have a genetic component and an environment one.
- Often the genes involved are numerous, each having a relatively small effect on the trait.



Most human traits are at least partly due to genetic variation

Dizygotic Twins



Monozygotic Twins

Twins separated at birth

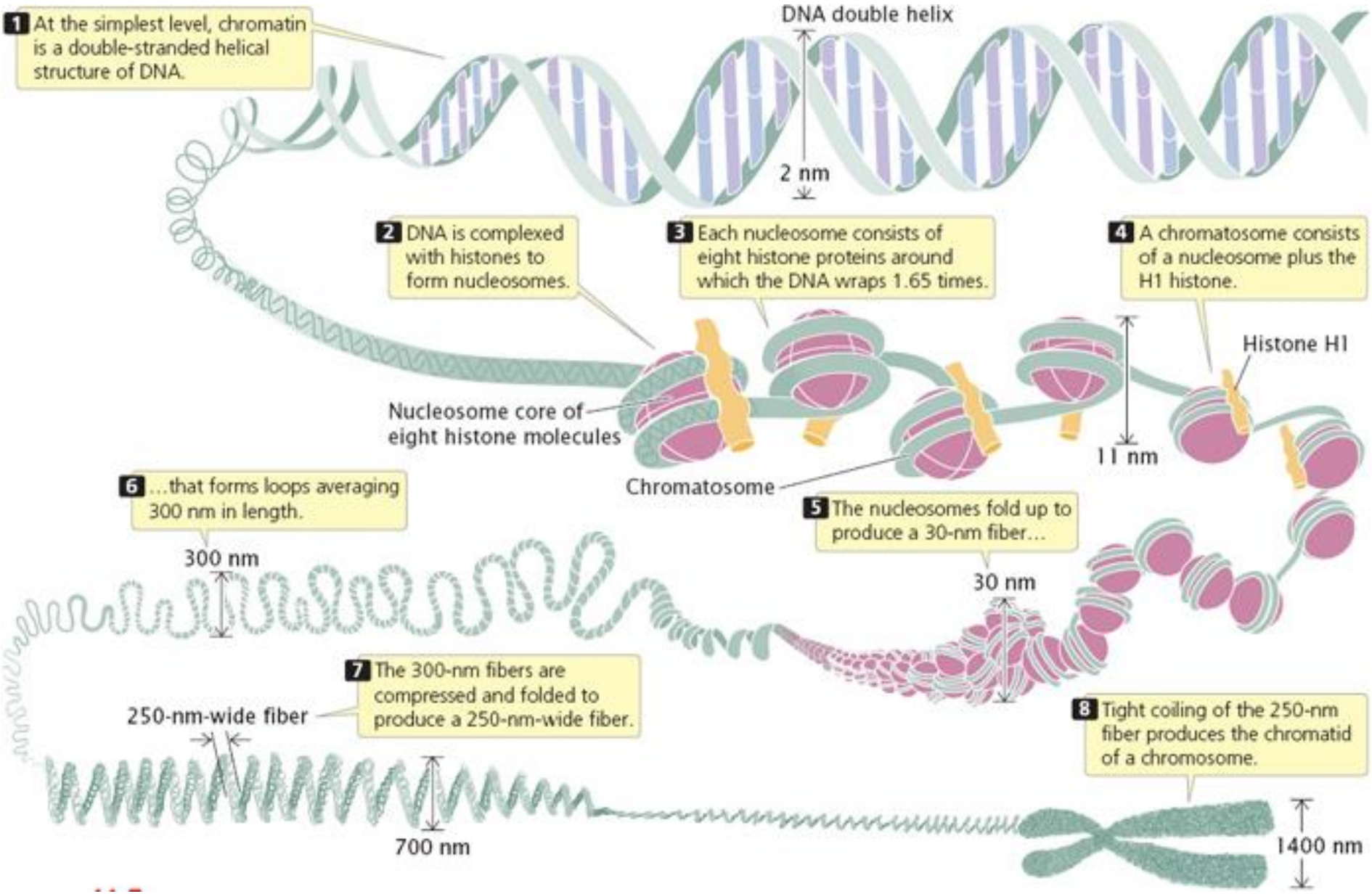
Borjeson, Acta Paed, 1976

- DZ twins share 50% genes and environment
- MZ twins share all their genes and environment

Disease	*Concordance*	
	MZ	*DZ*
Manic depressive psychosis	67%	5%
Cleft lip and palate	38%	8%
Rheumatoid arthritis	34%	7%
Asthma	47%	24%
Coronary artery disease	19%	9%
Diabetes mellitus	56%	11%

UNIVERSITY OF EXETER | MEDICAL SCHOOL

DNA is tightly packaged into chromosomes within the nucleus of the cell



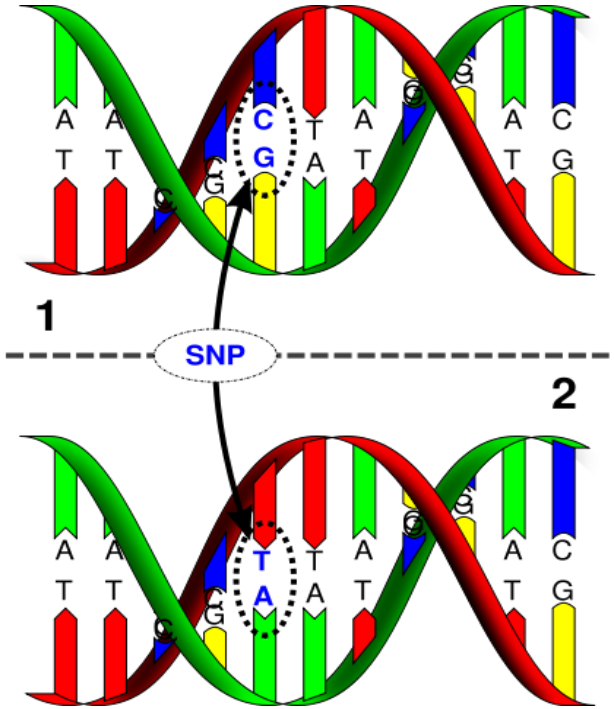
Genome facts

- The human genome is made up of 6 billion (6 000 000 000) bases of DNA, split into 24 chromosomes.
- This information would fill a stack of paperback books 61 m high
- The human body is made up of 100 trillion cells. Each cell has at least one nucleus, which houses the chromosomes.
- There is 1.8 m of DNA in each of our cells packed into a structure only 0.0001 cm across .
- If all the DNA in the 100 trillion cells of the human body was put end to end it would reach to the sun and back over 600 times
- The human genome contains about 20,000-25,000 genes.
- Between humans, our DNA differs by only 0.2 per cent, or 1 in 500 base (letters). (This takes into account that human cells have two copies of the genome.)
- Human DNA is 98 per cent identical to chimpanzees.

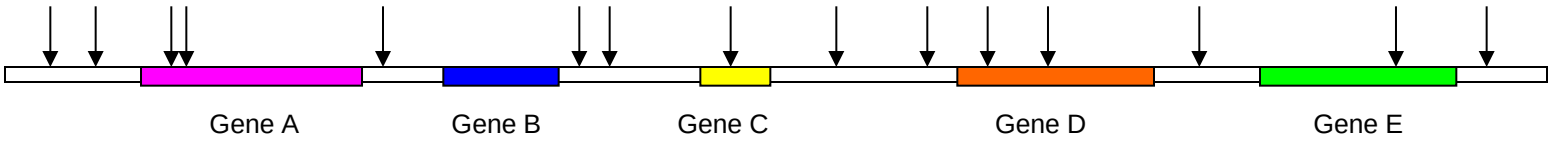
UNIVERSITY OF EXETER | MEDICAL SCHOOL

Genome-wide association studies

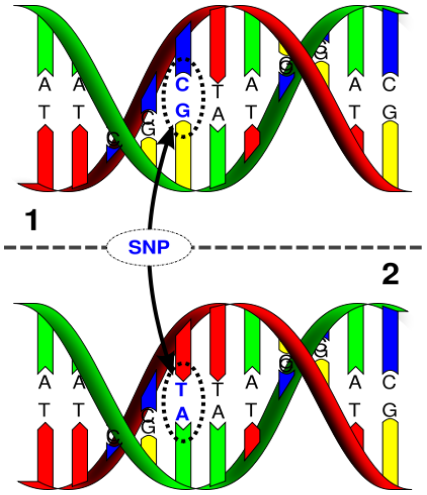
Single nucleotide polymorphism (SNP) (variation at a single base within the genome)

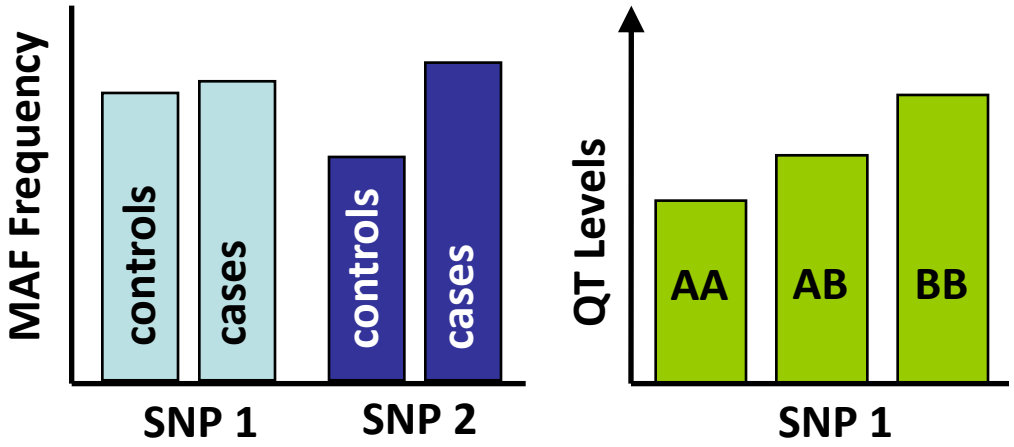


- Most common class of genomic variation
- Frequency of at least 1% in population
- Occur every 100-300 bases
- ~10 Million SNPs in human genome
- Occur both within genes and outside genes
- Predispose to, rather than cause, disease/trait



Association studies



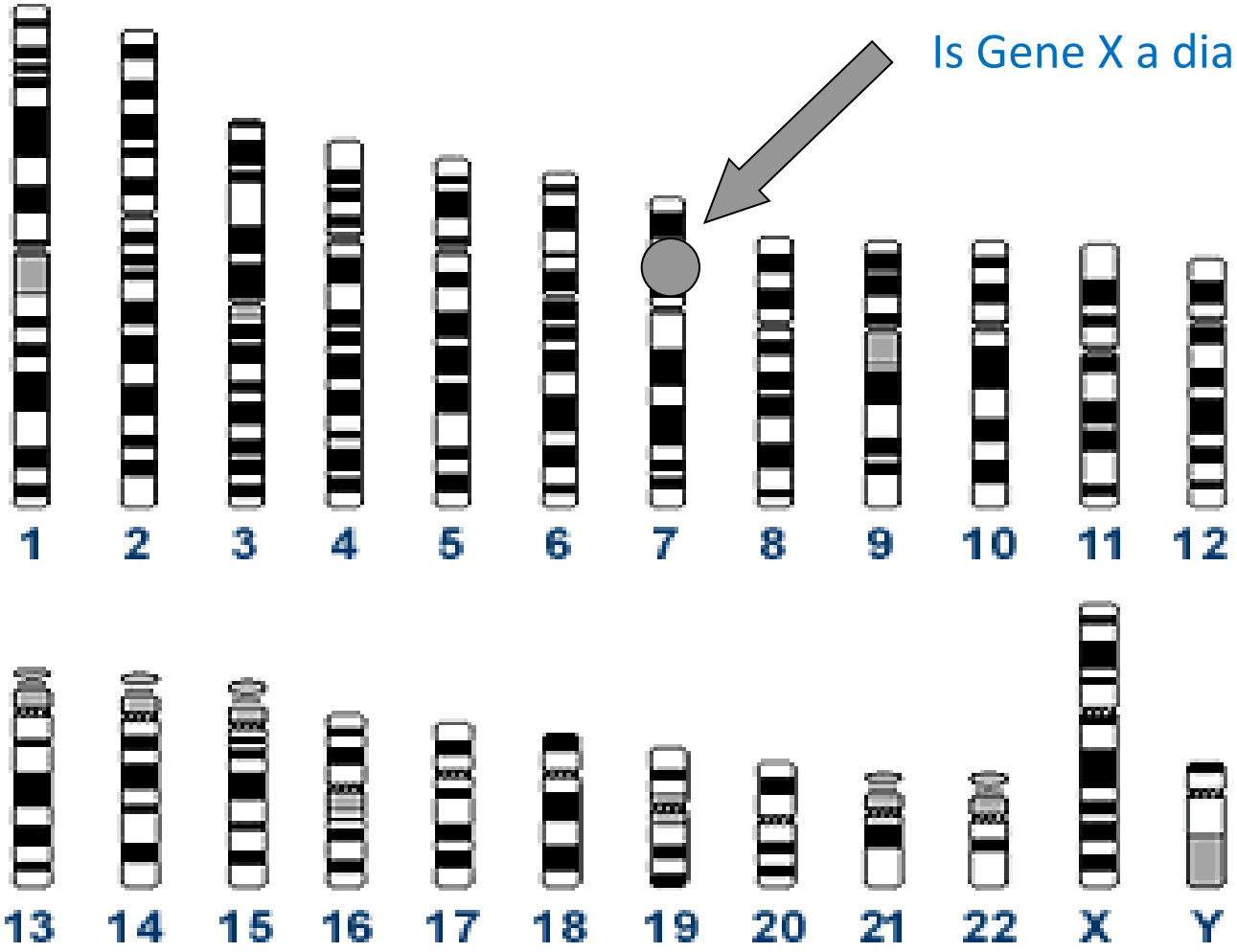


Genotype DNA samples at SNPs

Compare allele frequencies / quantitative trait levels

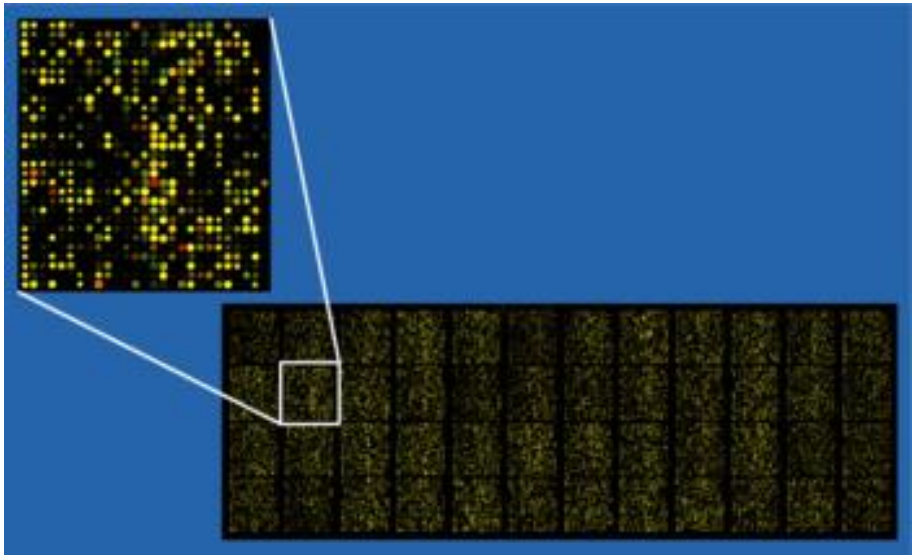
Before 2007

Association studies of candidate genes



If answer no – 24,999 more to look at

DNA Microarrays (gene chip)– used to test thousands of genetic variants simultaneously

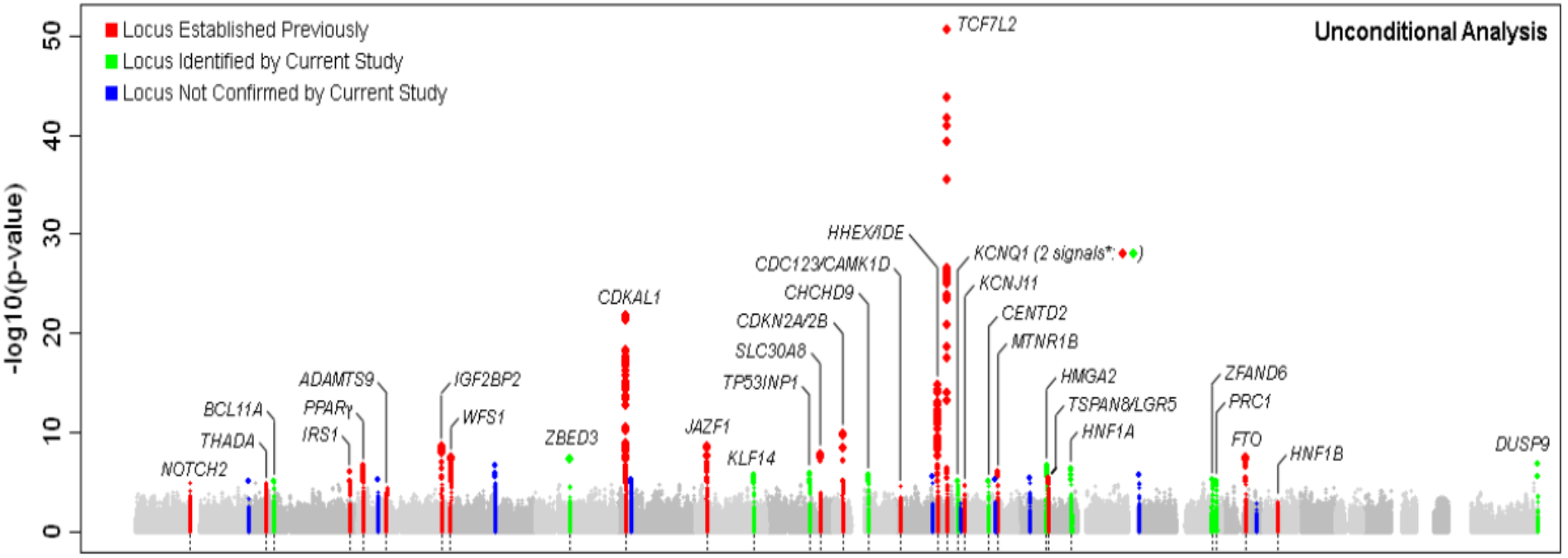


Thousands of short DNA sequences (probes) are spotted onto each chip and DNA target hybridised to it. The bound DNA is detected with a fluorescent signal. Multiple DNA sequences can be analysed simultaneously, but cost is high.

The Genome Wide Association Study era: 2007-
Study all variants in one go – involves large datasets and dealing with massive multiple hypothesis testing burden

Genotype DNA samples at 2.5million variants

~ 40 common loci have been robustly associated with type 2 diabetes susceptibility

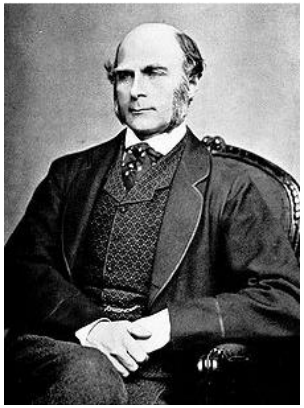


Discovery samples: 8000 cases vs 40,000 controls

Replication samples: 35,000 cases vs 60,000 controls

DIAGRAM+ Consortium: Voight *et al* Nature Genetics (2010)

The genetics of complex traits has been studied for over 100 years

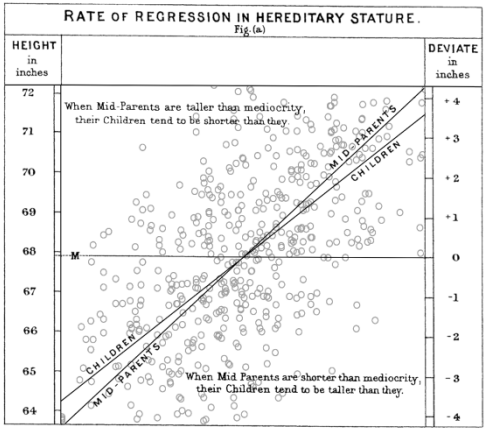


ANTHROPOLOGICAL MISCELLANEA.

Regression *towards* Mediocrity *in* Hereditary Stature.

By Francis Galton, F.R.S., &c.

"...stature is not a simple element, but a sum of accumulated lengths and thicknesses of more than a hundred of bodily parts... The beautiful regularity in the statures of a population is due to the number of variable elements the stature is the sum'

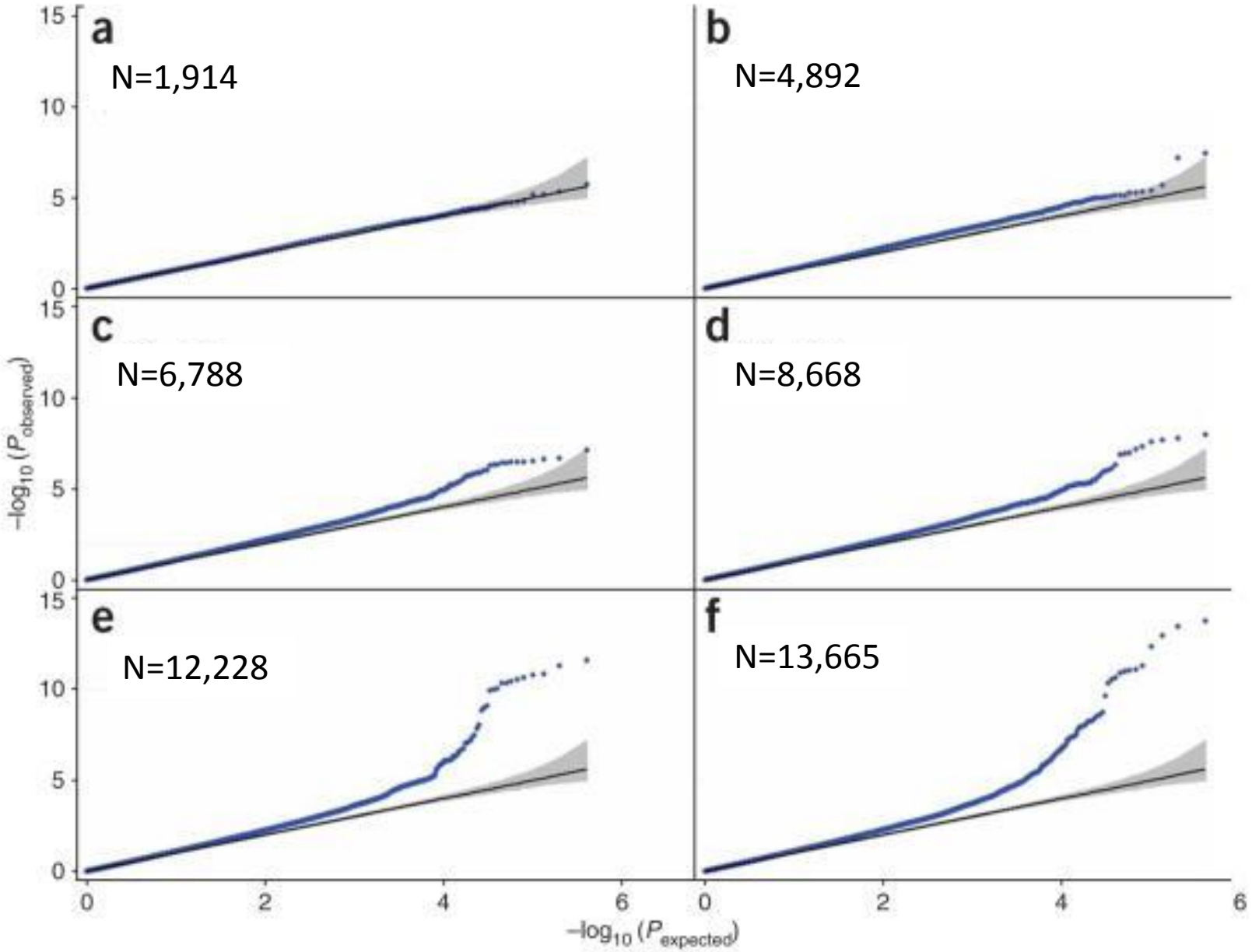


XV.—**The Correlation between Relatives on the Supposition of Mendelian Inheritance.** By **R. A. Fisher**, B.A. *Communicated by* Professor J. Arthur Thomson. (With Four Figures in Text.)

(MS. received June 15, 1918. Read July 8, 1918. Issued separately October 1, 1918.)

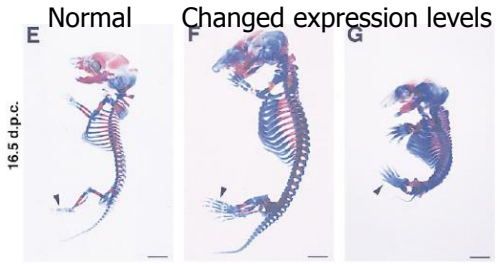
The simplest hypothesis, and the one which we shall examine, is that such features as stature are determined by a large number of Mendelian factors, and that the large variance among children of the same parents is due to the segregation of those factors in respect to which the parents are heterozygous.

Increasing sample size has a dramatic effect on our ability to detect significant associations



Common loci influencing adult height variation - discovery timeline

Zhou *et al*. 1995



Tsumaki *et al*. 1995

Ligon *et al.* 2005

Thomas *et al.* 2005

HMGA2

Weedon *et al.*

5,000 + 19,000

GDF5-UQCC

Sanna *et al.*

6,500 + 29,000

~44 new loci

Lettre *et al.*
15,000 + 10,000

Gudbjartsson *et al.*
27,000 + 5,000

Weedon *et al.*
14,000 + 16,000

GIANT

N=185,000!

180 loci

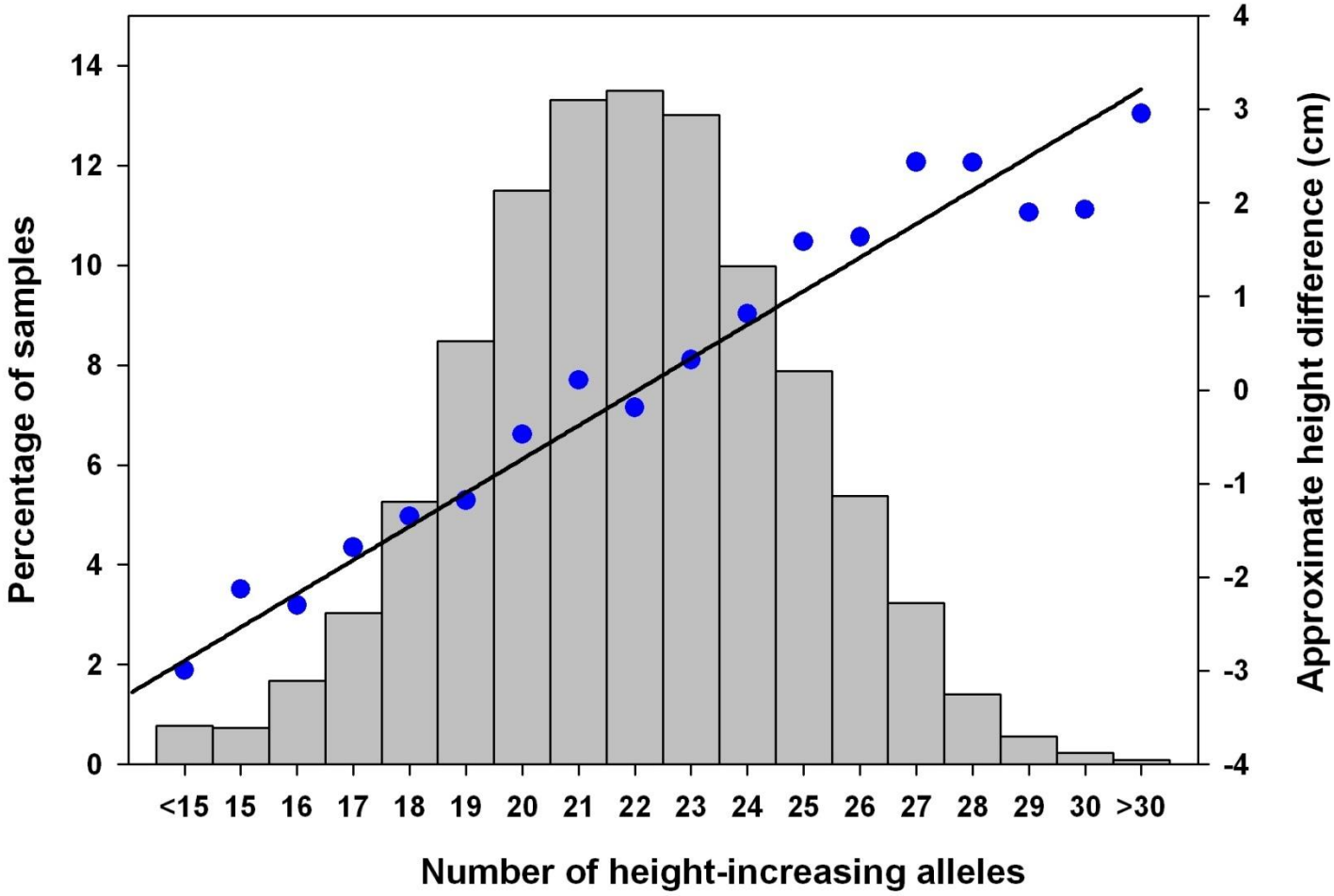
Oct 07

Feb 08

May 08

Nature, 2010

The heritability of polygenic diseases and traits is due to a very large number of variants each explaining only a small proportion of the population variation: height as an example

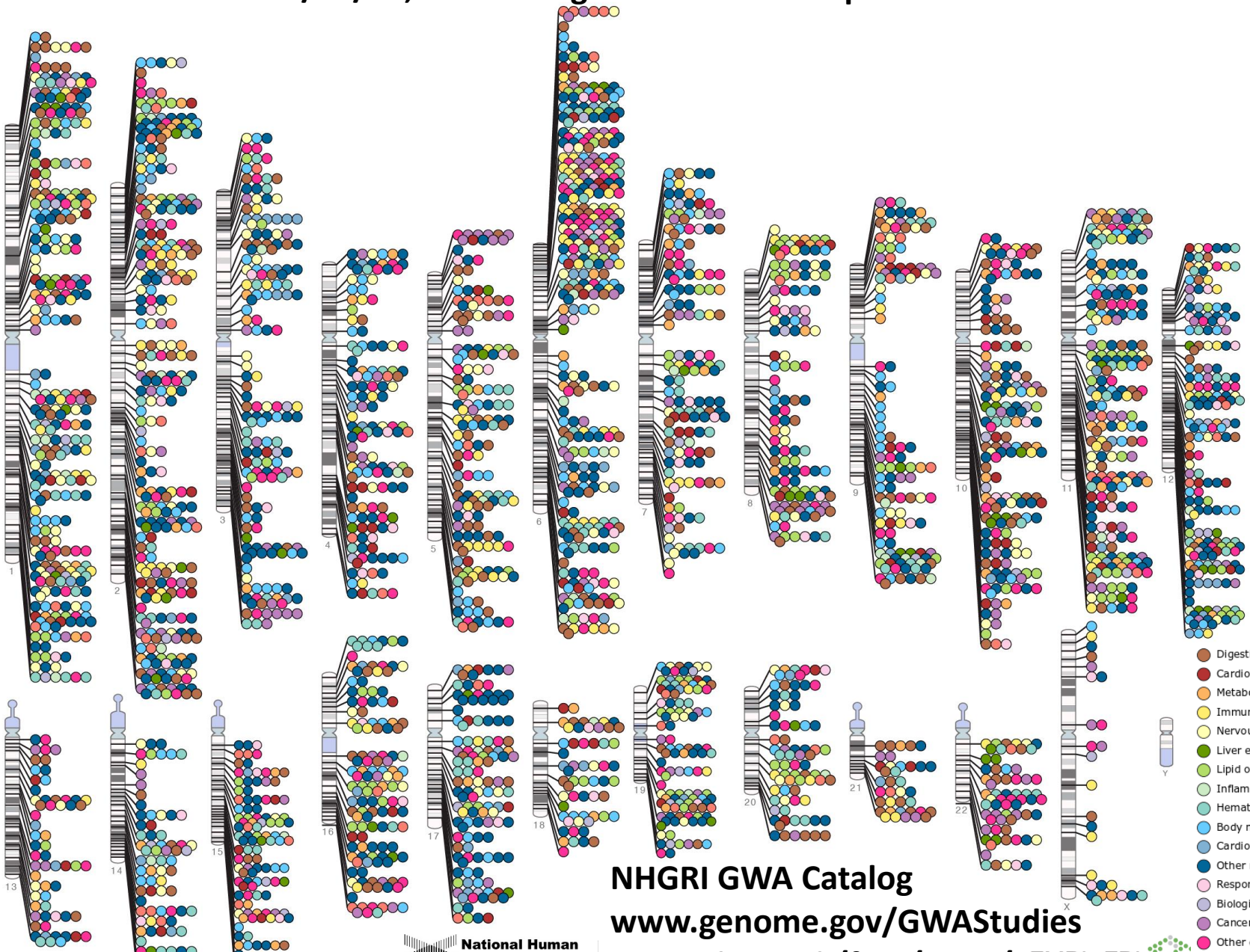


Weedon et al, Nature Genetics, 2008

UNIVERSITY OF EXETER | MEDICAL SCHOOL

Published Genome-Wide Associations in December 2012

As of 18/06/13, the catalogue includes 1636 publications and 10838 SNPs.



NHGRI GWA Catalog

www.genome.gov/GWAStudies

www.ebi.ac.uk/fgpt/gwas/

National Human Genome Research Institute

EMBL-EBI

Genome sequencing

In 2007 Sanger sequencing was used to produce the first complete genome sequence of an individual - Craig Venter.

The Diploid Genome Sequence of an Individual Human

Samuel Levy1*, Granger Sutton1, Pauline C. Ng1, Lars Feuk2, Aaron L. Halpern1, Brian P. Walenz1, Nelson Axelrod1, Jiaqi Huang1, Ewen F. Kirkness1, Gennady Denisov1, Yuan Lin1, Jeffrey R. MacDonald2, Andy Wing Chun Pang2,Mary Shago2, Timothy B. Stockwell1, Alexia Tsiamouri1, Vineet Bafna3, Vikas Bansal3, Saul A. Kravitz1, Dana A. Busam1, Karen Y. Beeson1, Tina C. McIntosh1, Karin A. Remington1, Josep F. Abril4, John Gill1, Jon Borman1, Yu-Hui Rogers1, Marvin E. Frazier1, Stephen W. Scherer2, Robert L. Strausberg1, J. Craig Venter1

PLoS Biol (2007) 5(10): e254

http://biology.plosjournals.org/archive/1545-7885/5/10/pdf/10.1371_journal.pbio.0050254-S.pdf

- a human genome contains ~6 billion bases and 20,000-25,000 genes
- ~0.5% variation between homologous chromosomes
- 3.2 million of the variants were single nucleotide polymorphisms
- but, nearly 1 million were much larger insertion/deletions and copy number variations accounting for ~74% of the variant DNA sequence.

In July 2007 next generation sequencing was used to generate the complete genomic sequence of James Watson

- It took 2 months to complete and cost $2 million
- This compares to the $3 billion for the human genome project, which took 10 years to complete.
- It is predicted that a $10,000 genome is a realistic target for the near future (currently ~$20,000)

The complete genome of an individual by massively parallel DNA sequencing
Nature. 2008 Apr 17;452(7189):872-6
http://jimwatsonsequence.cshl.edu/cgi-perl/gbrowse/jwsequence/

16 June 2013 Last updated at 19:50

Gene mutation means paracyclist has no fat under skin

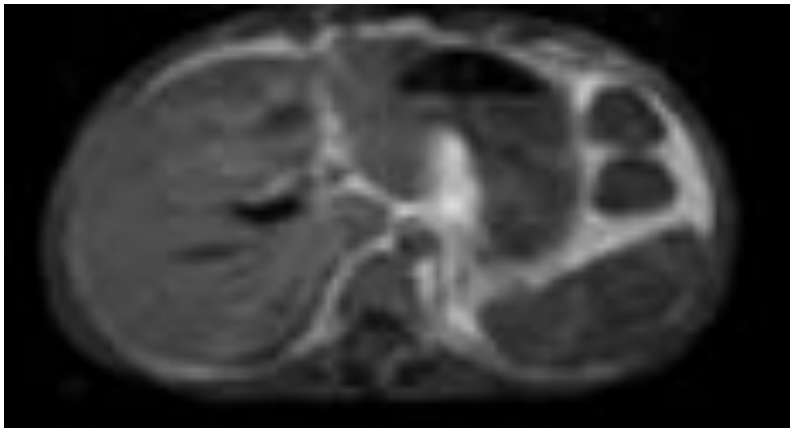
By Philippa Roxby
Health reporter, BBC News

Tom Staniford's ambition is to be a Paralympic cycling champion in Rio

Baffled doctors are nothing new to 23-year-old budding Paralympic cyclist Tom Staniford, from Exeter.

Related Stories

MDP syndrome is a rare novel multi-system disorder that presents with lipodystrophy



NATURE GENETICS | LETTER

An in-frame deletion at the polymerase active site of POLD1 causes a multisystem disorder with lipodystrophy

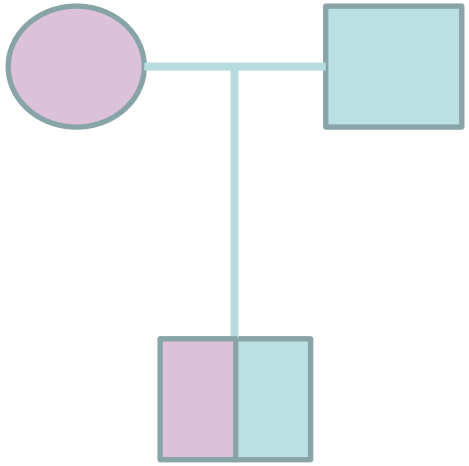
Michael N Weedon, Sian Ellard, Marc J Prindle, Richard Caswell, Hana Lango Allen, Richard Oram, Koumudi Godbole, Chittaranjan S Yajnik, Paolo Sbraccia, Giuseppe Novelli, Peter Turnpenny, Emma McCann, Kim Jee Goh, Yukai Wang, Jonathan Fulford, Laura J McCulloch, David B Savage, Stephen O'Rahilly, Katarina Kos, Lawrence A Loeb, Robert K Semple & Andrew T Hattersley

Affiliations | **Contributions** | **Corresponding author**

Nature Genetics (2013) | doi:10.1038/ng.2670
Received 27 March 2013 | Accepted 22 May 2013 | Published online 16 June 2013

UNIVERSITY OF EXETER | MEDICAL SCHOOL

de novo mutations occur in every generation



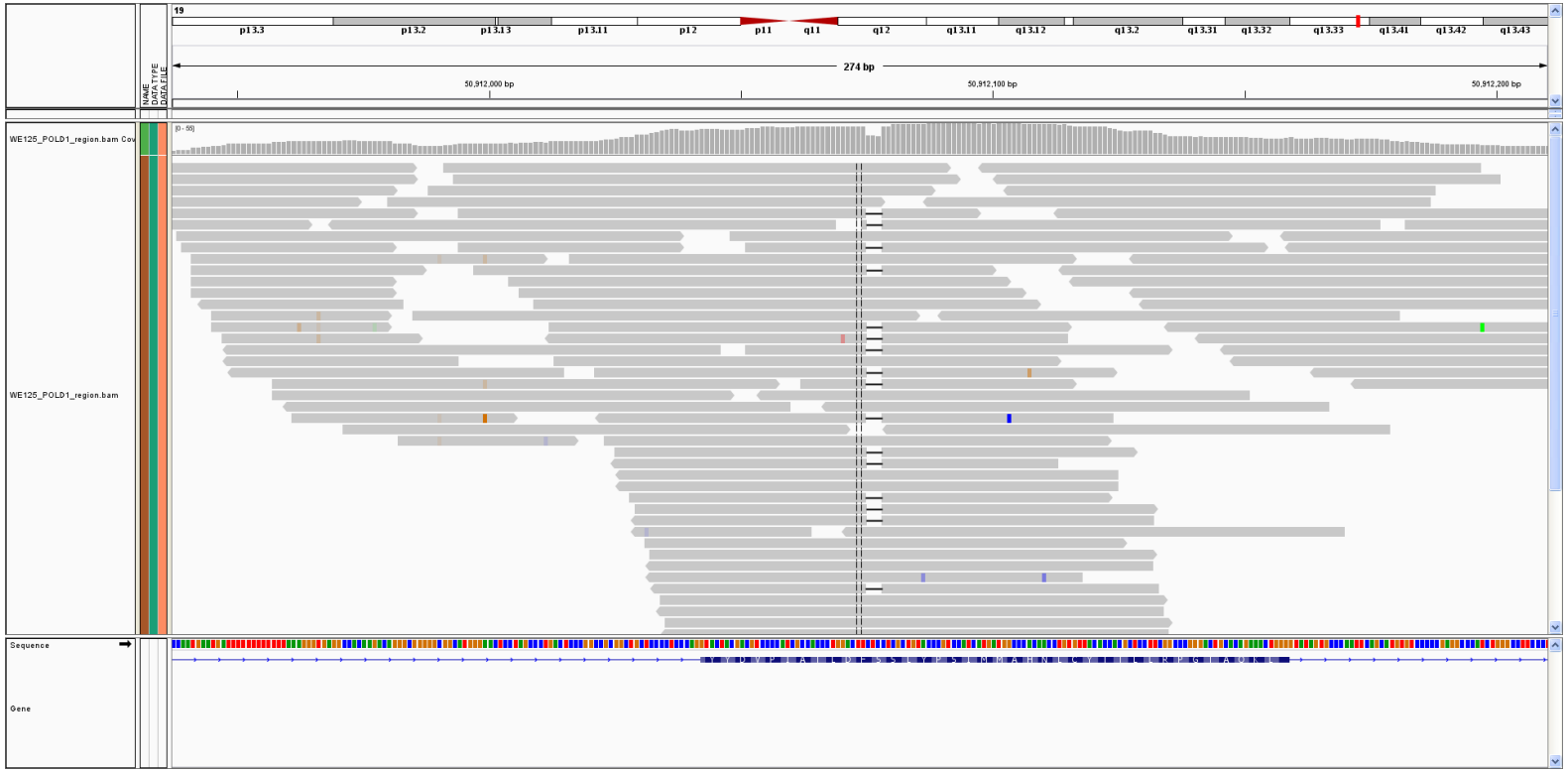
50% of genes come from Dad
and 50% from Mum

74 *de novo* single nucleotide changes
3 *de novo* insertion/deletions
0.02 *de novo* copy number variants
1 out of 20,563 protein-coding genes are hit by a *de novo* mutation per generation

UNIVERSITY OF EXETER | MEDICAL SCHOOL

We sequenced the protein-coding genome of two patients and their unaffected parents to look for *de novo* mutations

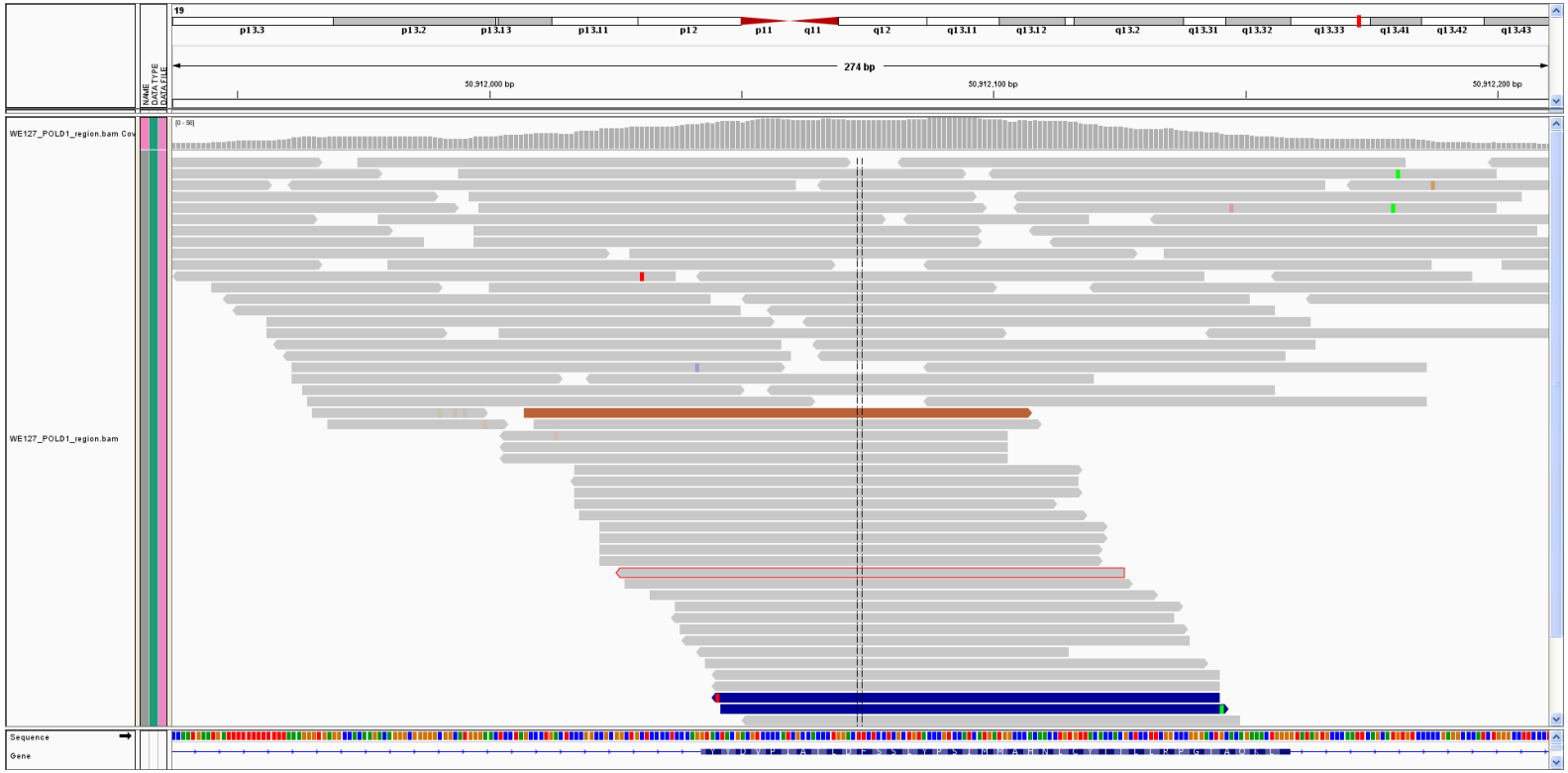
Patient 1 - reads over POLD1 region



UNIVERSITY OF EXETER | MEDICAL SCHOOL

We sequenced the protein-coding genome of two patients and their unaffected parents to look for *de novo* mutations

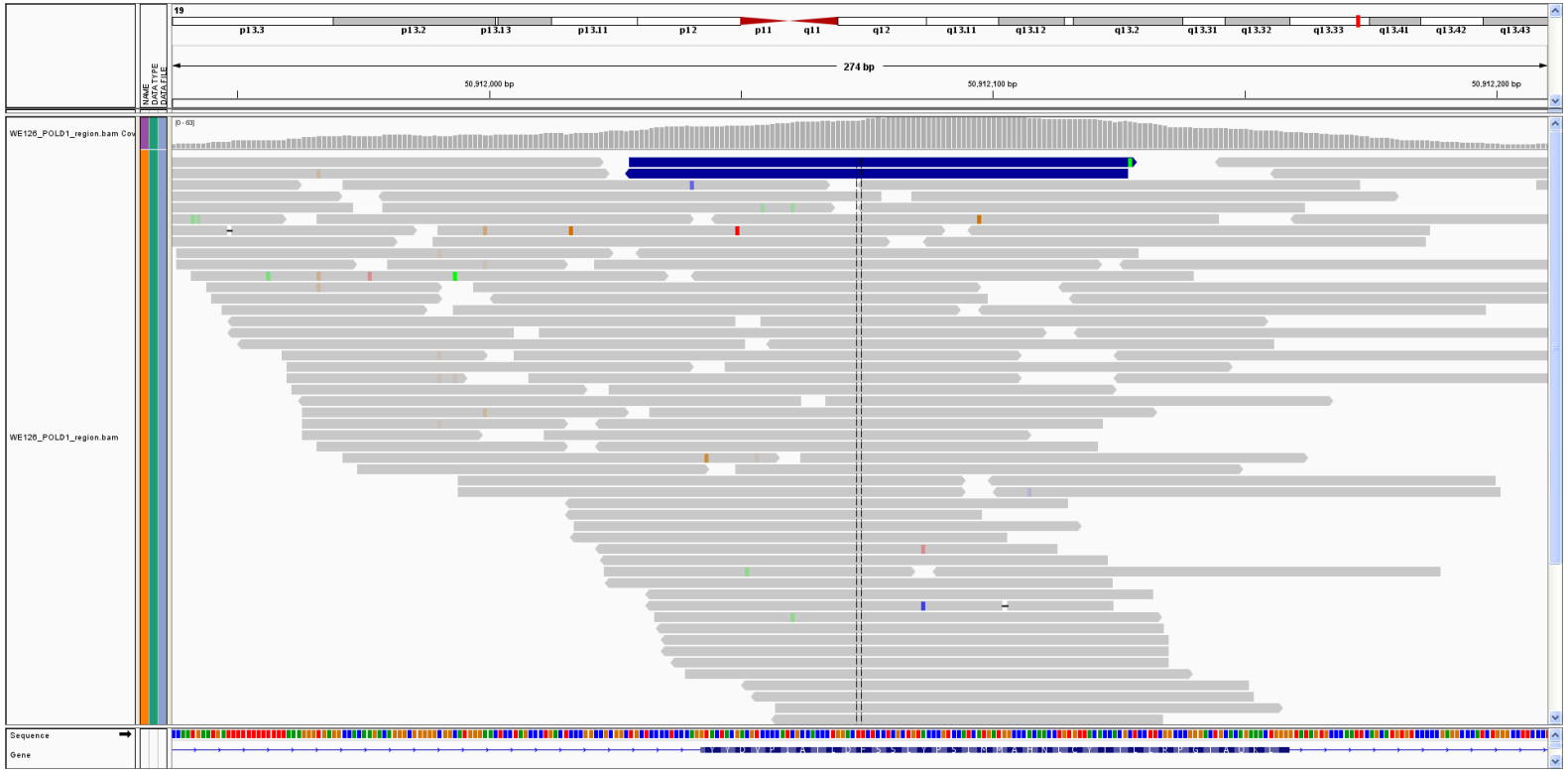
Patient 1's unaffected mother



UNIVERSITY OF EXETER | MEDICAL SCHOOL

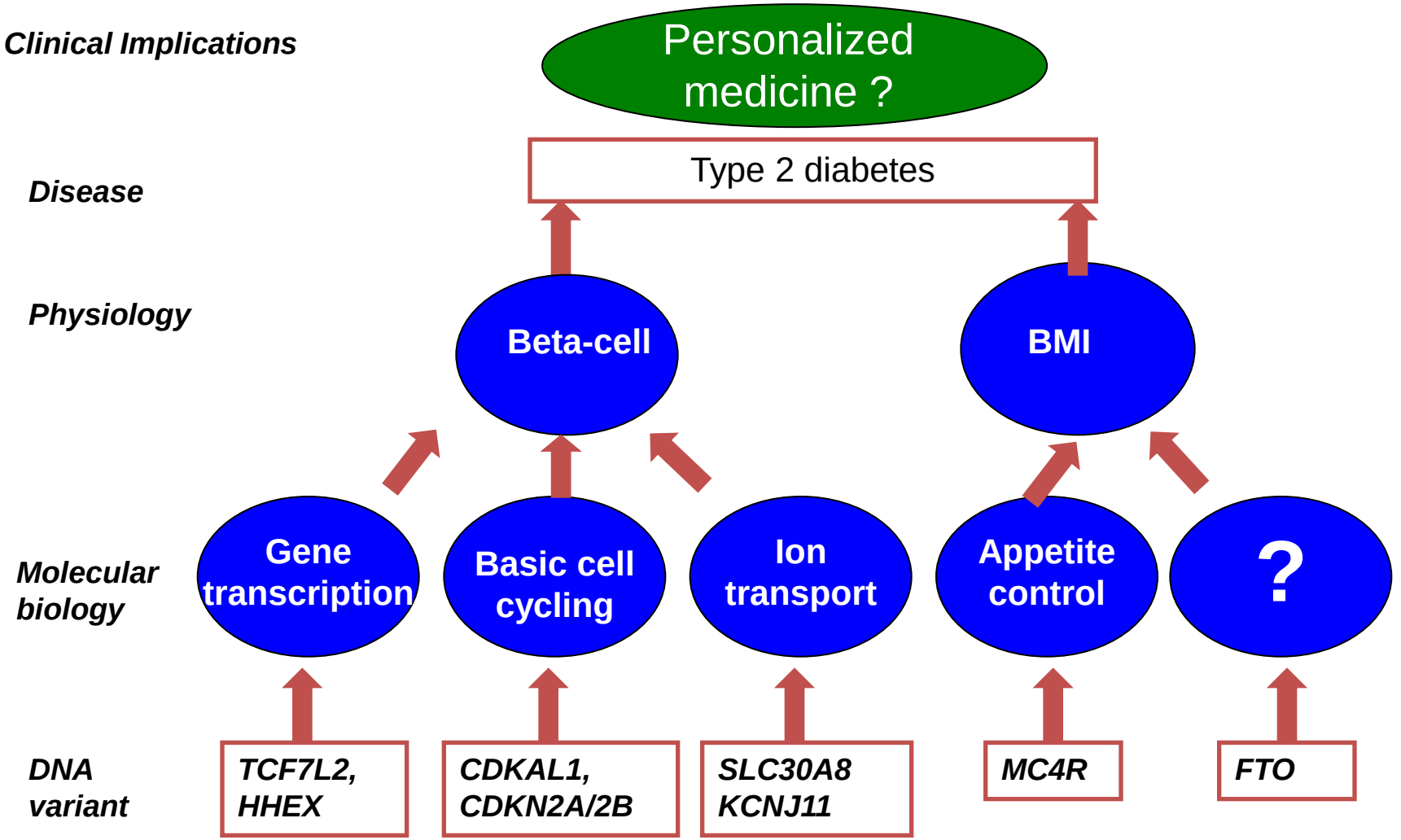
We sequenced the protein-coding genome of two patients and their unaffected parents to look for *de novo* mutations

Patient 1's unaffected father



UNIVERSITY OF EXETER | MEDICAL SCHOOL

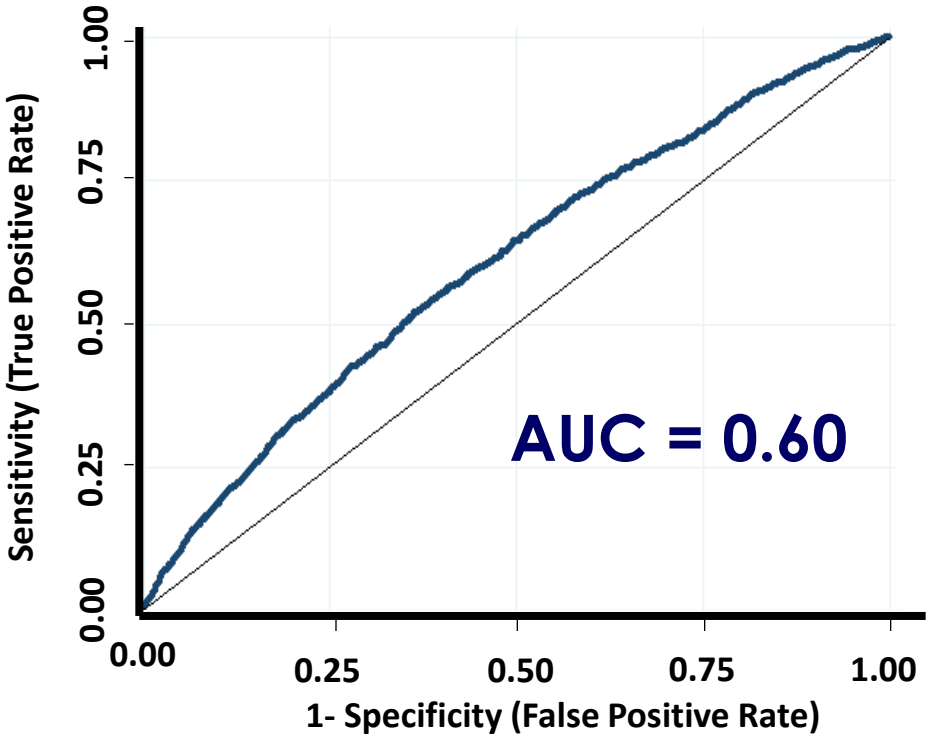
What's the point ?
Example: Dissection of the biology of Type 2 diabetes



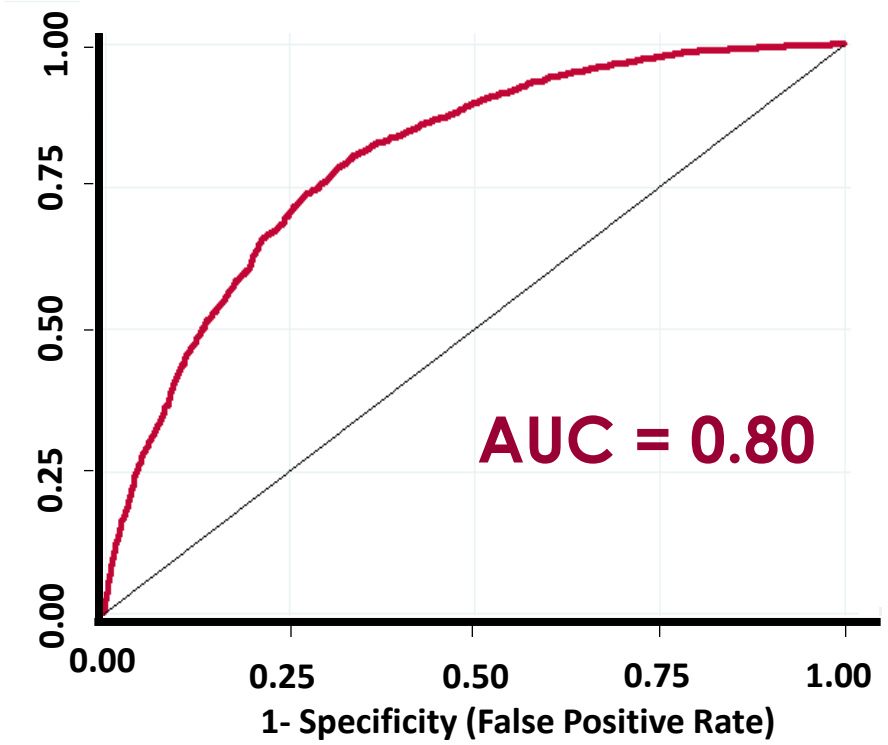
UNIVERSITY OF EXETER | MEDICAL SCHOOL

Genetic variants for T2D not particularly useful for prediction yet, but may be more so in future

Gene variants alone

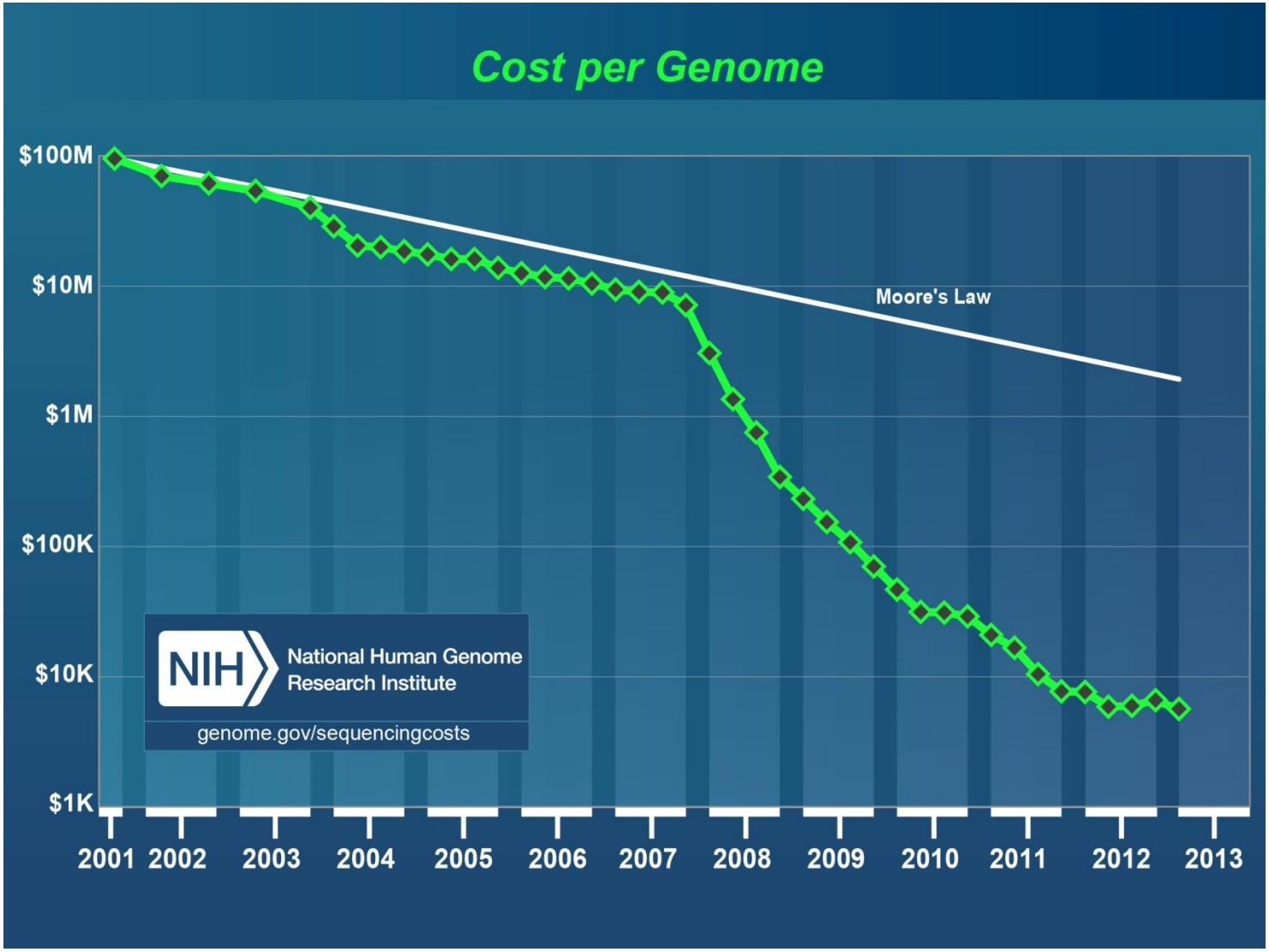


Gene variants, BMI, age & gender



BMI, age & gender
AUC = 0.78

The cost of genome sequencing is falling dramatically



Francis S. Collins

(National Human Genome Research Institute):

where to begin?

The real question is, "What wouldn't we do?" At the National Human Genome Research Institute, we'd be like kids in a candy shop—there are so many exciting possibilities from which to choose. Bearing in mind our mission of using genomic research to improve human health, we'd probably take most of our current annual spending on DNA sequencing, about $120 million, and devote it to sequencing 100,000 human samples for $100 million. About 75,000 of those samples would come from obtaining the complete genome sequences of 2,500 affected individuals for each of 30 common, complex diseases, such as asthma, arthritis, diabetes, various types of cancer, heart disease, stroke, Alzheimer's disease and depression. This would enable us to systematically find both the common and the rare genetic variations that contribute to the risk of developing these diseases. The remaining genomes to be sequenced would be those of 25,000 people who have made it to the age of 100 in relatively good health and retaining the capacity for independent function. The aim of that endeavor would be to see what's special about the genomes of healthy centenarians, and then to use that information to explore the genetics of good health and longevity in all humans.

(posted 2 January 2007)

10 December 2012 Last updated at 08:42

Fergus Walsh
Medical correspondent
More from Fergus

DNA mapping for cancer patients

COMMENTS (224)

Up to 100,000 patients with cancer and rare diseases in England are to have their entire genetic code sequenced.

The Prime Minister will announce £100m has been set aside for the project over the next three to five years.

The aim is to give doctors a better understanding of patients' genetic make-up, condition and treatment needs, and help develop new cancer treatments.

The human genome contains three billion pairs of code