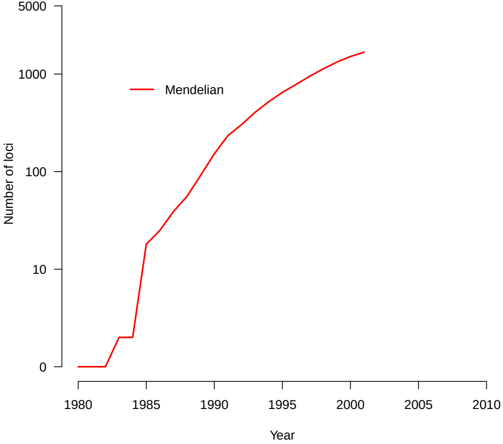
Genome-wide association studies

Jeff Barrett

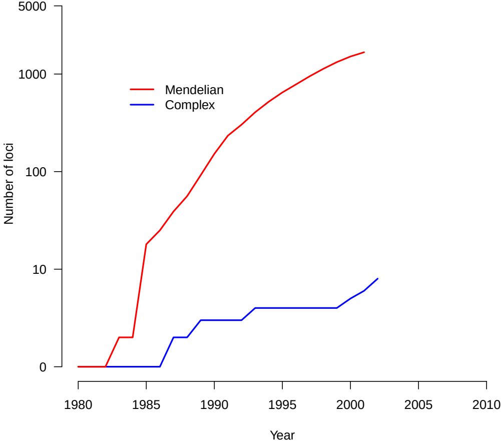
Boulder Workshop, 2011

Linkage mapping of Mendelian diseases accelerated. . .



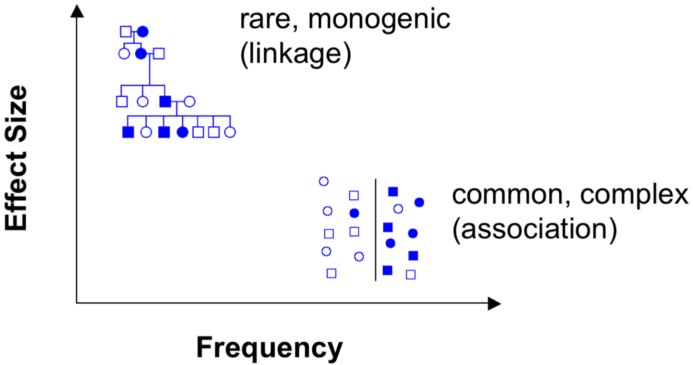
Adapted from Glazier *et al. Science.* 2002.

. . . but this success did not translate to complex disease

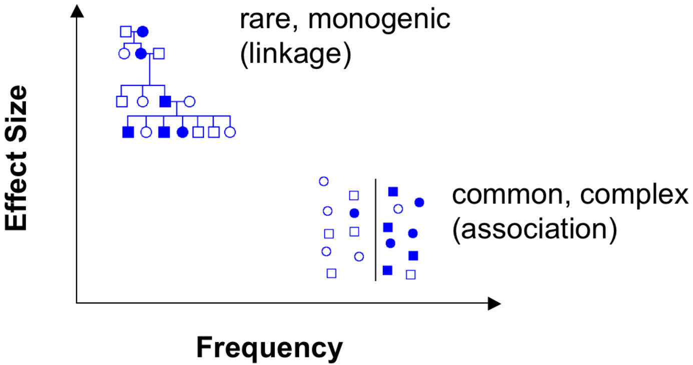


Adapted from Glazier *et al. Science.* 2002.

Different diseases require different methods



Different diseases require different methods



Challenge: find a genome-wide analysis well powered to find small effects

Genetic diversity

The two processes which increase genetic diversity in a population are mutation, which introduces novel variants into the population, and recombination, which re-shuffles the existing patterns of variation (haplotypes).

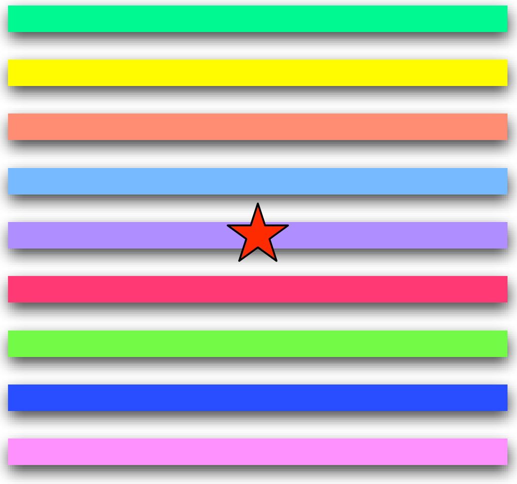
Genetic diversity

The two processes which increase genetic diversity in a population are mutation, which introduces novel variants into the population, and recombination, which re-shuffles the existing patterns of variation (haplotypes).

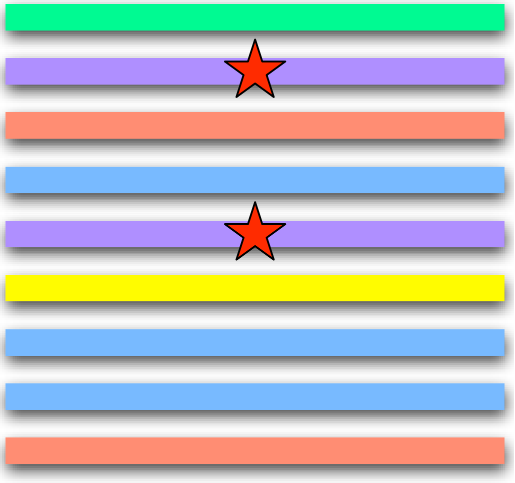
The fate of new mutations is also affected by drift, selection, and population history. Understanding the patterns left behind in genetic variation because of these forces is key to designing disease studies.

Mutation and recombination in a population

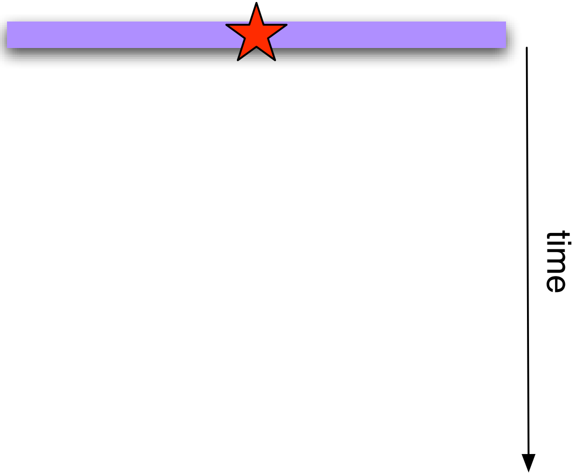
Mutation and recombination in a population



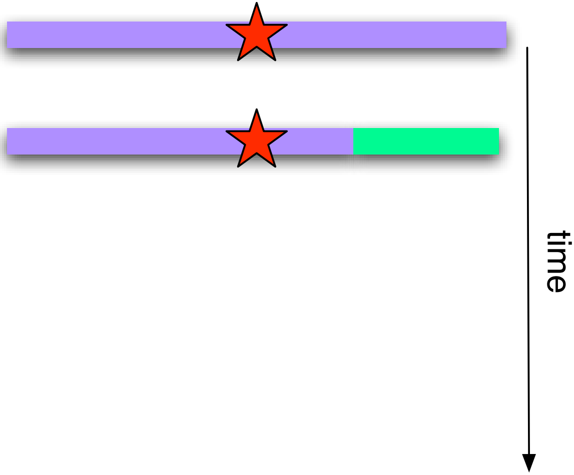
Mutation and recombination in a population



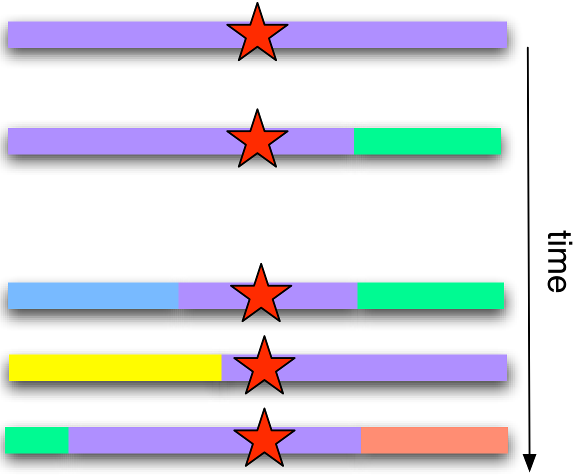
Mutation and recombination in a population



Mutation and recombination in a population



Mutation and recombination in a population



Consequences of mutation and recombination

- Genetic variants are correlated because they occur on a particular haplotype background, and segregate in populations on that background.

Consequences of mutation and recombination

- Genetic variants are correlated because they occur on a particular haplotype background, and segregate in populations on that background.
- In the absence of recombination this correlation would never be broken down and would extend a great distance along chromosomes.

Consequences of mutation and recombination

- Genetic variants are correlated because they occur on a particular haplotype background, and segregate in populations on that background.
- In the absence of recombination this correlation would never be broken down and would extend a great distance along chromosomes.
- Recombination breaks down this correlation over many successive generations, leaving a narrower and narrower window of correlation.

Consequences of mutation and recombination

- Genetic variants are correlated because they occur on a particular haplotype background, and segregate in populations on that background.
- In the absence of recombination this correlation would never be broken down and would extend a great distance along chromosomes.
- Recombination breaks down this correlation over many successive generations, leaving a narrower and narrower window of correlation.
- This correlation (or linkage disequilibrium, LD) enables GWAS to capture most common variation in a population without genotyping every marker.

Quantifying LD

		SNP 1	
		p	1-p
SNP 2	q	pq	q(1-p)
	1-q	p(1-q)	(1-p)(1-q)

Quantifying LD

		SNP 1	
		p	1-p
SNP 2	q	π_{11}	π_{12}
	1-q	π_{21}	π_{22}

Quantifying LD

		SNP 1	
		p	1-p
SNP 2	q	π_{11}	π_{12}
	1-q	π_{21}	π_{22}

$$D = \pi_{11} - pq$$

Quantifying LD

		SNP 1	
		p	1-p
SNP 2	q	π_{11}	π_{12}
	1-q	π_{21}	π_{22}

$$D = \pi_{11} - pq$$

$$D' = D/D_{\text{max}}$$

Quantifying LD

		SNP 1	
		p	1-p
SNP 2	q	π_{11}	π_{12}
	1-q	π_{21}	π_{22}

$$D = \pi_{11} - pq$$

$$D' = D/D\text{max}$$

$$r^2 = D/p(1-p)q(1-q)$$

A haplotype map of the human genome

27 October 2005 | www.nature.com/nature
THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE
nature
THE HAPMAP PROJECT
Chapter and verse on human genetic variation

Project details (Phase I/II)

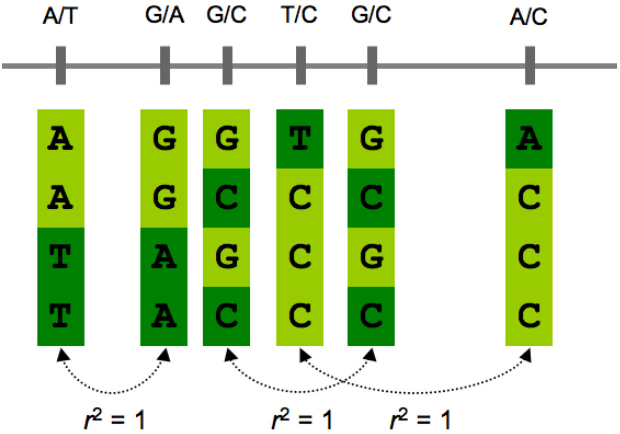
Samples:

- 90 Yoruba (30 parent-parent-offspring trios) from Ibadan, Nigeria (YRI)
- 90 CEPH samples (30 trios) of European descent from Utah (CEU)
- 45 Han Chinese from Beijing (CHB)
- 45 Japanese from Tokyo (JPT)

SNPs: Original goal was 1 SNP every 5kb, but as genotyping costs dropped, eventual catalogue included approximately 4 million polymorphic SNPs scattered across the genome.

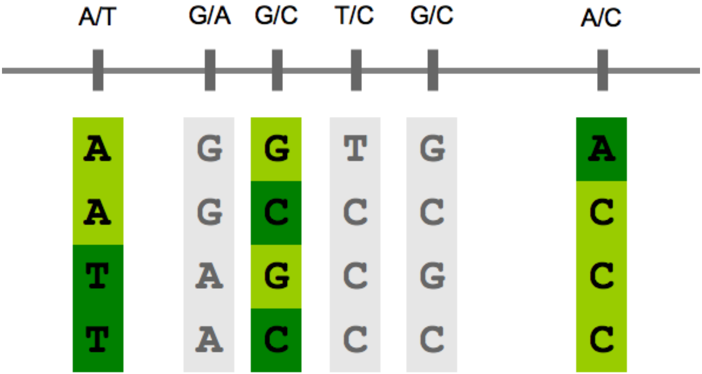
Panel	% $r^2 > 0.8$	mean max r^2
YRI	81	0.90
CEU	94	0.97
CHB+JPT	94	0.97

How can we use HapMap knowledge for disease studies?

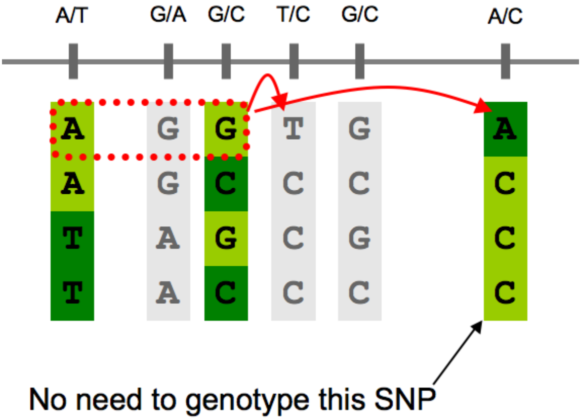


A/T	G/A	G/C	T/C	G/C	A/C
A	G	G	T	G	A
A	G	C	C	C	C
T	A	G	C	G	C
T	A	C	C	C	C

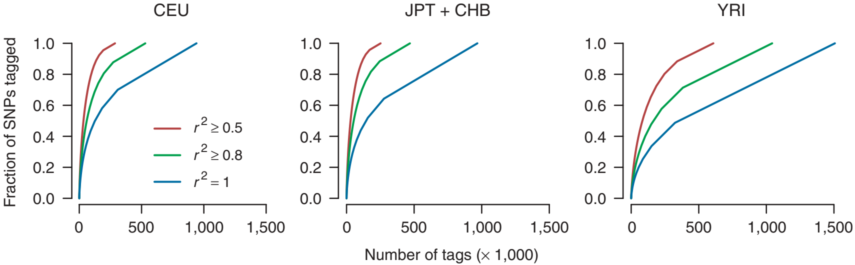
Gain efficiency by removing redundant SNPs



Haplotypes can yield additional gains in efficiency

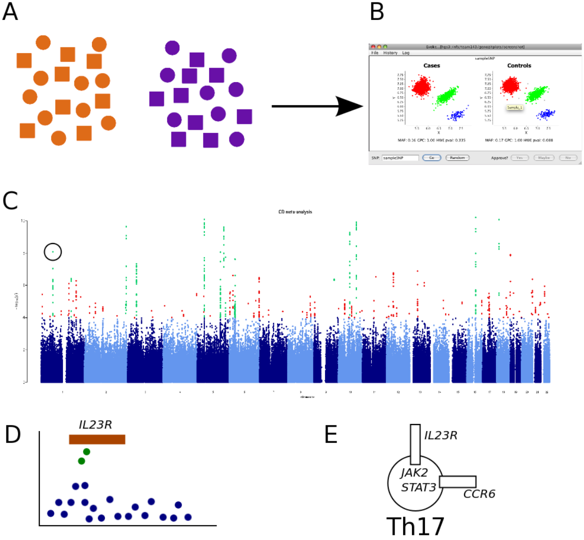


Cheap genotyping arrays allowed this idea to be implemented genome-wide



Barrett & Cardon. *Nature Genetics*, 2006.

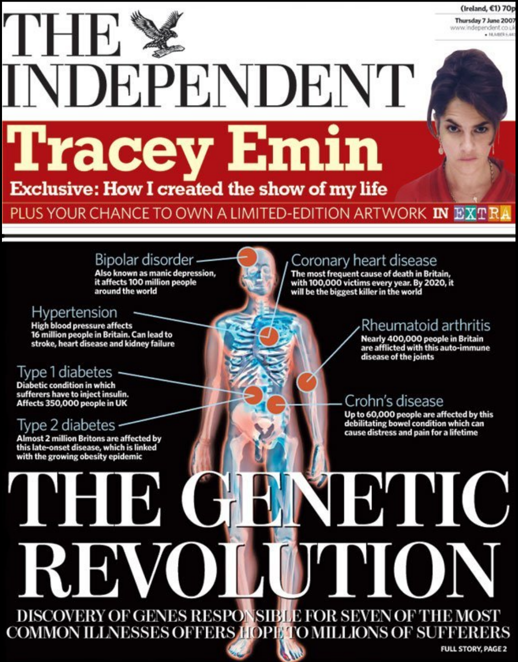
Genome wide association studies



MAF: 0.16 GPC: 1.00 HWE pval: 0.335

MAF: 0.17 GPC: 1.00 HWE pval: 0.088

SNP: sampleSNP



THE INDEPENDENT

(Ireland, €1) 70p

Thursday 7 June 2007
www.independent.co.uk
• NUMBER 6,4

Tracey Emin

Exclusive: How I created the show of my life

PLUS YOUR CHANCE TO OWN A LIMITED-EDITION ARTWORK IN EXTRA

Bipolar disorder
Also known as manic depression, it affects 100 million people around the world

Coronary heart disease
The most frequent cause of death in Britain, with 100,000 victims every year. By 2020, it will be the biggest killer in the world

Hypertension
High blood pressure affects 16 million people in Britain. Can lead to stroke, heart disease and kidney failure

Rheumatoid arthritis
Nearly 400,000 people in Britain are afflicted with this auto-immune disease of the joints

Type 1 diabetes
Diabetic condition in which sufferers have to inject insulin. Affects 350,000 people in UK

Crohn's disease
Up to 60,000 people are affected by this debilitating bowel condition which can cause distress and pain for a lifetime

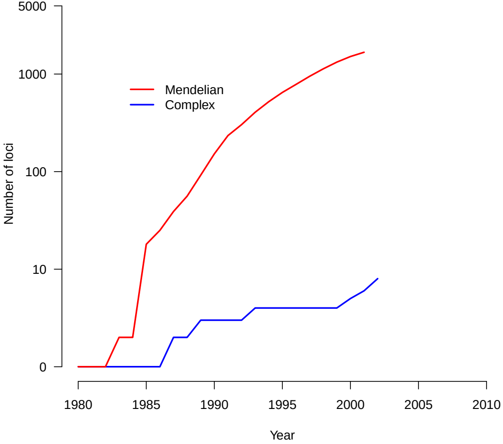
Type 2 diabetes
Almost 2 million Britons are affected by this late-onset disease, which is linked with the growing obesity epidemic

THE GENETIC REVOLUTION

DISCOVERY OF GENES RESPONSIBLE FOR SEVEN OF THE MOST COMMON ILLNESSES OFFERS HOPE TO MILLIONS OF SUFFERERS

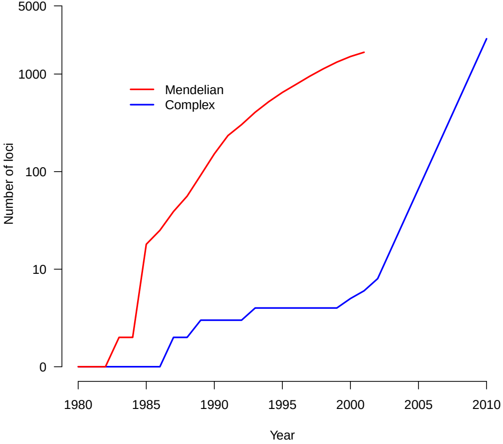
FULL STORY, PAGE 2

GWAS revolutionized complex disease genetics



Adapted from Glazier *et al. Science.* 2002, and NHGRI GWAS catalog.

GWAS revolutionized complex disease genetics



Adapted from Glazier *et al. Science.* 2002, and NHGRI GWAS catalog.

Expected challenges

Given that GWAS are feasible, what are the obstacles which stand in the way of finding genes?

- No common, single SNP main effects: all epistasis, or haplotypes, or rare variation or. . .
- Population structure
- Multiple testing corrections will drown out signal
- Computational burden
- Sample sizes too small to detect the effects
- SNP chips don't cover enough of the genome

Expected challenges

Given that GWAS are feasible, what are the obstacles which stand in the way of finding genes?

- No common, single SNP main effects: all epistasis, or haplotypes, or rare variation or. . .
- Population structure
- Multiple testing corrections will drown out signal
- Computational burden
- Sample sizes too small to detect the effects
- SNP chips don't cover enough of the genome

Expected challenges

Given that GWAS are feasible, what are the obstacles which stand in the way of finding genes?

- **Data quality control**
- No common, single SNP main effects: all epistasis, or haplotypes, or rare variation or. . .
- Population structure
- Multiple testing corrections will drown out signal
- Computational burden
- Sample sizes too small to detect the effects
- SNP chips don't cover enough of the genome

SNP quality control metrics

SNP QC for GWAS is straightforward, and generally similar to any other genotyping experiment. Commonly used QC checks include:

- Hardy-Weinberg equilibrium (expected ratios of three possible genotypes)
- Fraction of missing genotypes
- Allele frequency
- Frequency differences in separate control groups (if available)

SNP quality control metrics

SNP QC for GWAS is straightforward, and generally similar to any other genotyping experiment. Commonly used QC checks include:

- Hardy-Weinberg equilibrium (expected ratios of three possible genotypes)
- Fraction of missing genotypes
- Allele frequency
- Frequency differences in separate control groups (if available)

...but the crucial difference to all previous experiments is scale! The WTCCC had 8.5 billion genotypes, and datasets are growing all the time.

Sample quality control metrics

Collecting, processing and genotyping thousands of samples (often from many different clinicians, hospitals, countries. . .) is difficult.

- Duplicates
- Unexpected relatives
- Low quality DNA samples
- Sample mix-ups
- Samples with different ethnic ancestry

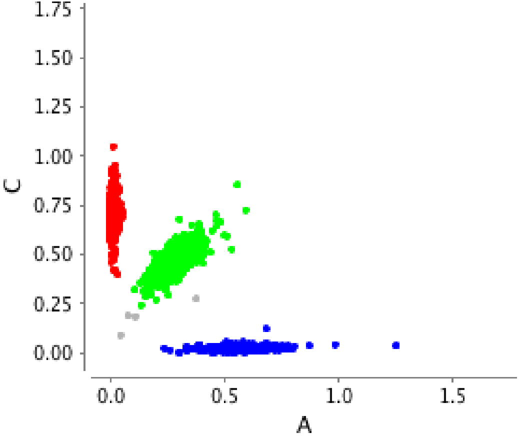
Sample quality control metrics

Collecting, processing and genotyping thousands of samples (often from many different clinicians, hospitals, countries. . .) is difficult.

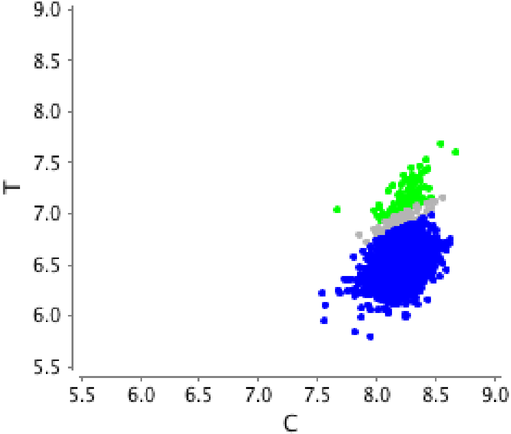
- Duplicates
- Unexpected relatives
- Low quality DNA samples
- Sample mix-ups
- Samples with different ethnic ancestry

But the good news is that simple analyses of genome-wide data can be very informative.

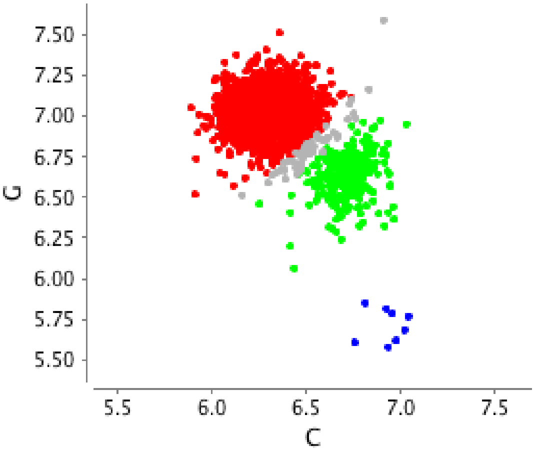
From intensity measurements to genotypes



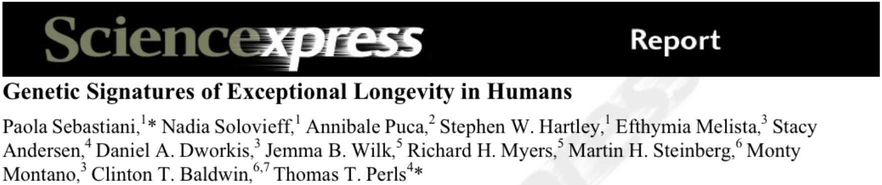
From intensity measurements to genotypes



From intensity measurements to genotypes



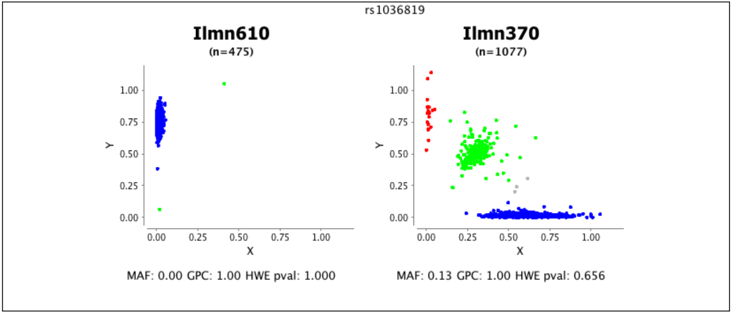
Clean data matters!



Sciencexpress Report

Genetic Signatures of Exceptional Longevity in Humans

Paola Sebastiani,[1]* Nadia Solovieff,[1] Annibale Puca,[2] Stephen W. Hartley,[1] Efthymia Melista,[3] Stacy Andersen,[4] Daniel A. Dworkis,[3] Jemma B. Wilk,[5] Richard H. Myers,[5] Martin H. Steinberg,[6] Monty Montano,[3] Clinton T. Baldwin,[6,7] Thomas T. Perls[4]*



GWAS resources

PLINK: analysis toolset
`http://pngu.mgh.harvard.edu/purcell/plink/`

Worked example: Data quality in case-control association studies, Anderson CA *et al. Nature Protocols* 5, 1564–1573 (2010).