Population genetics, linkage disequilibrium and GWAS

Jeff Barrett

Boulder Workshop, 2013

Overview

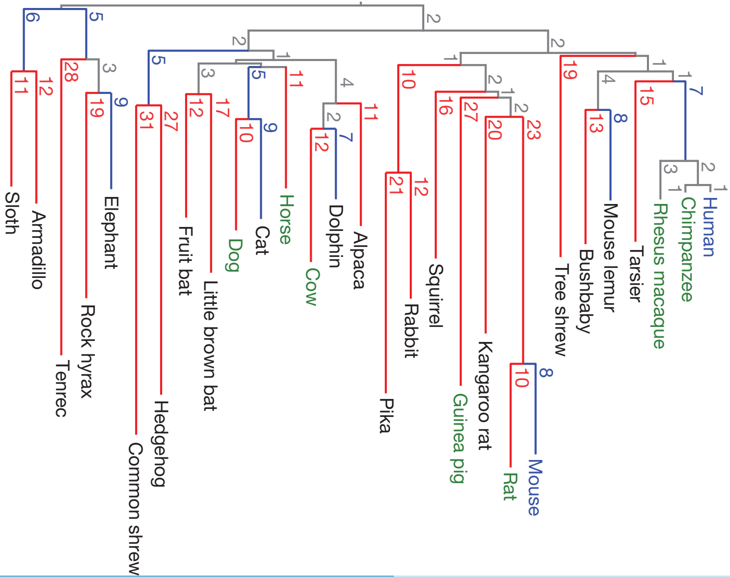
Human evolution

Genetic variation & linkage disequilibrium

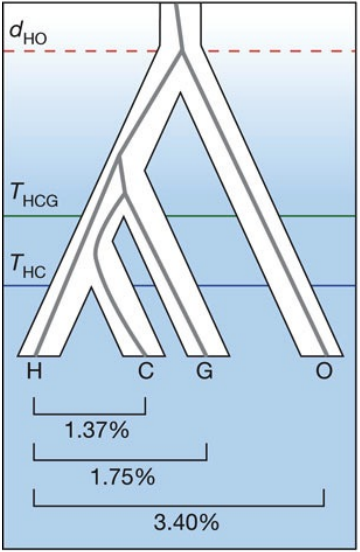
HapMap & tag SNPs

Genome-wide association studies & QC

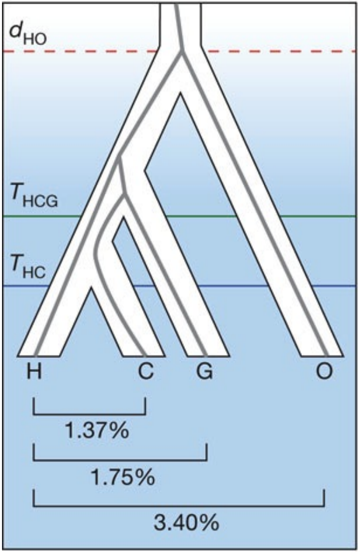
Humans are related to other species

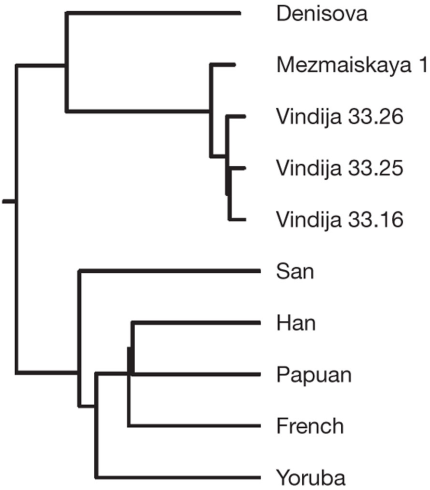


Our closer relatives

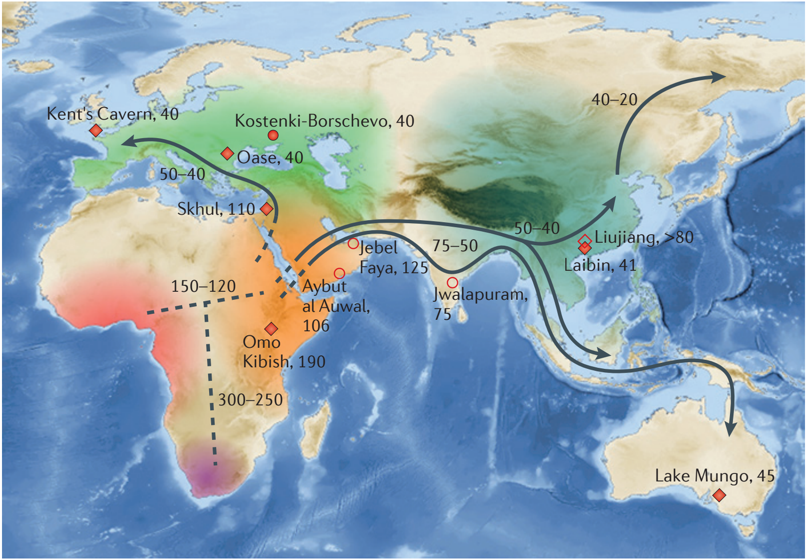


Our closer relatives





The history of human populations



Overview

Human evolution

Genetic variation & linkage disequilibrium

HapMap & tag SNPs

Genome-wide association studies & QC

Genetic diversity

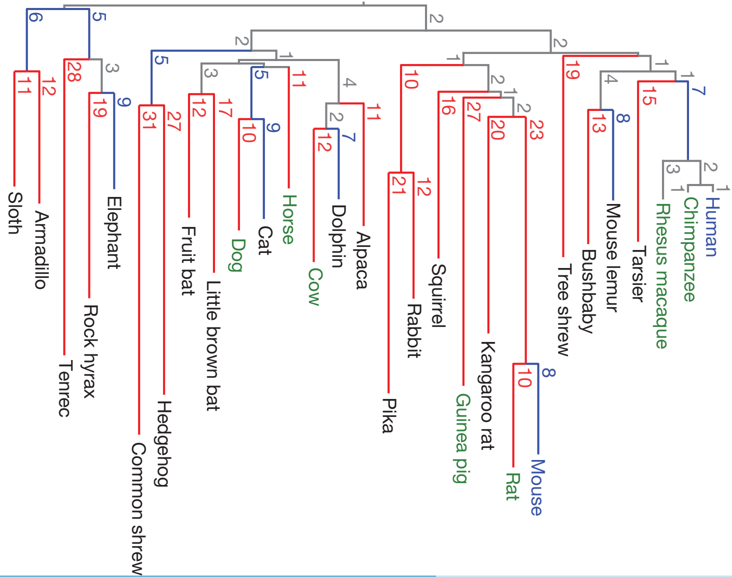
The two processes which increase genetic diversity in a population are mutation, which introduces novel variants into the population, and recombination, which re-shuffles the existing patterns of variation (haplotypes).

Genetic diversity

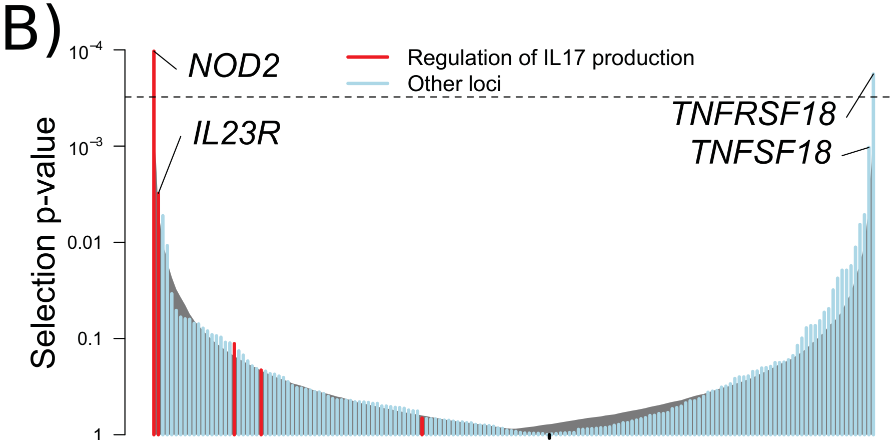
The two processes which increase genetic diversity in a population are mutation, which introduces novel variants into the population, and recombination, which re-shuffles the existing patterns of variation (haplotypes).

The fate of new mutations is 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.

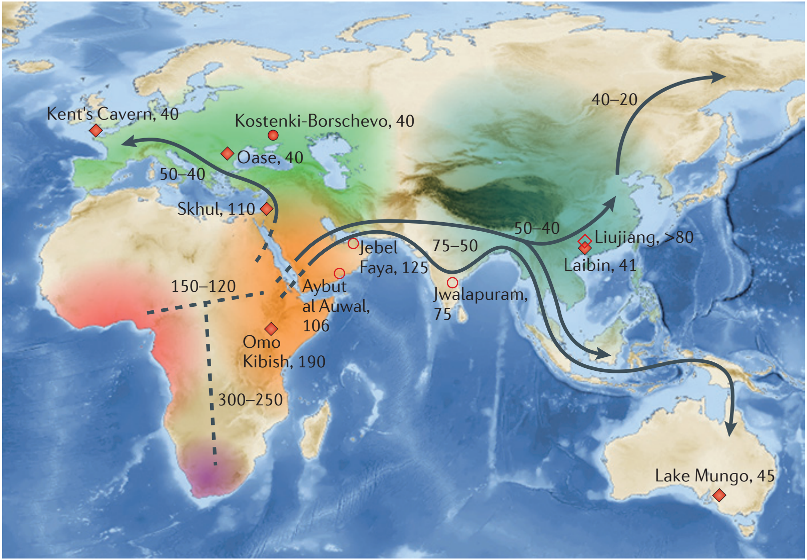
Weak selection pervades disease genetics



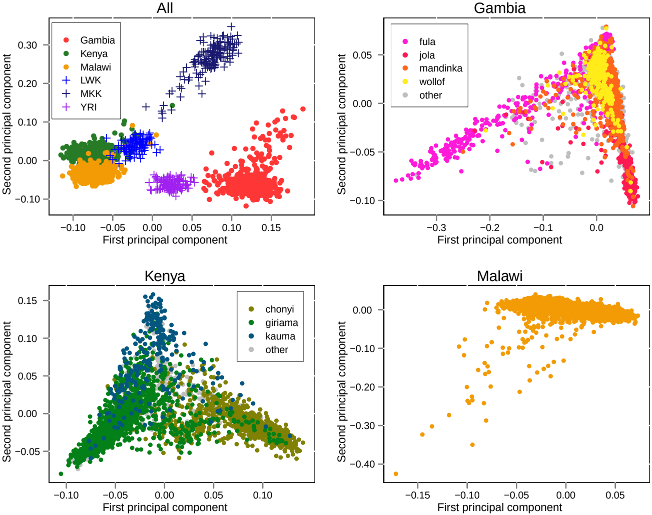
Weak selection pervades disease genetics



Population history and diversity

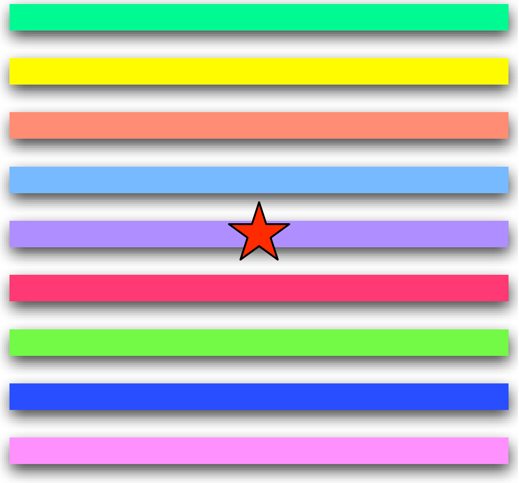


Population history and diversity

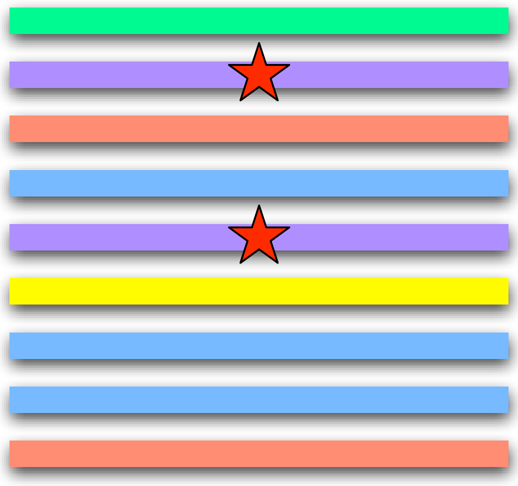


Mutation and recombination in a population

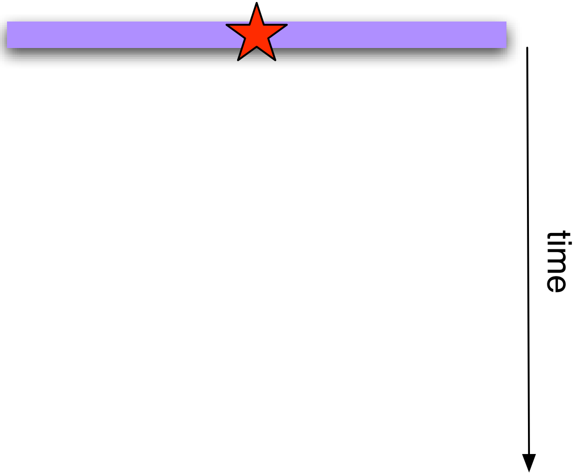
Mutation and recombination in a population



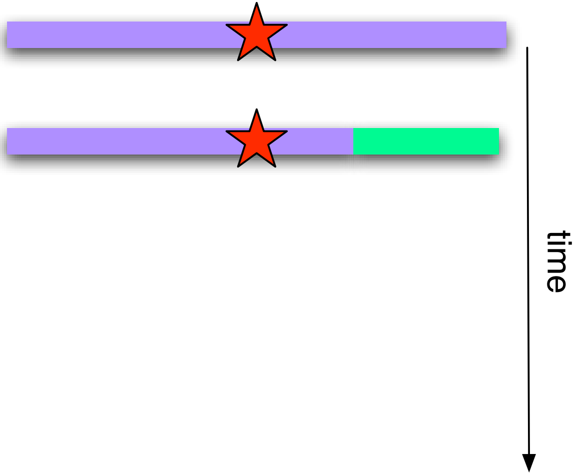
Mutation and recombination in a population



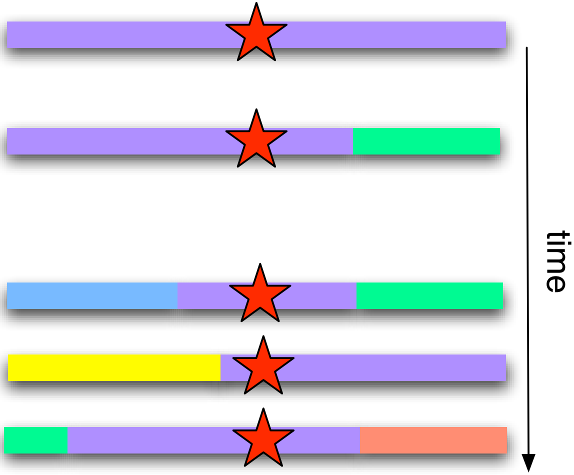
Mutation and recombination in a population



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 (called linkage disequilibrium or LD)would never be broken down and would extend a great distance along chromosomes.

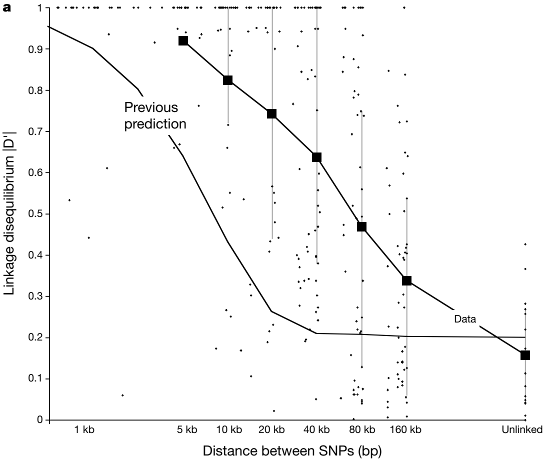
Consequences of mutation and recombination

- 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 (called linkage disequilibrium or LD)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

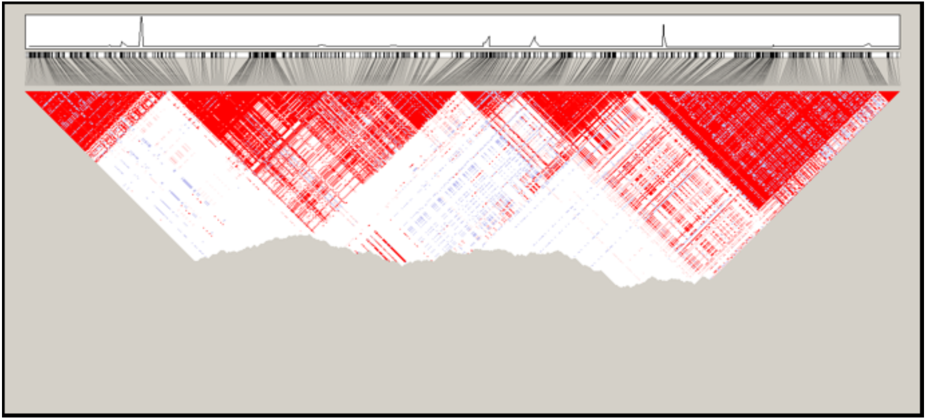
- Genetic variants are correlated because they occur on a particular haplotype background, and segregate in populations on that background.
- In the absence of recombination this correlation (called linkage disequilibrium or LD)would never be broken down and would extend a great distance along chromosomes.
- Recombination breaks down this correlation over many successive generations, leaving a narrower and narrower window of correlation.
- Under certain assumptions (neutral evolution, random mating, homogenous recombination), we can model exactly how far this correlation should extend.

Theoretical vs. empirical patterns of LD



Reich et al, *Nature*, 2001.

Heterogeneous recombination drives observed LD patterns



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}

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

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

Quantifying LD

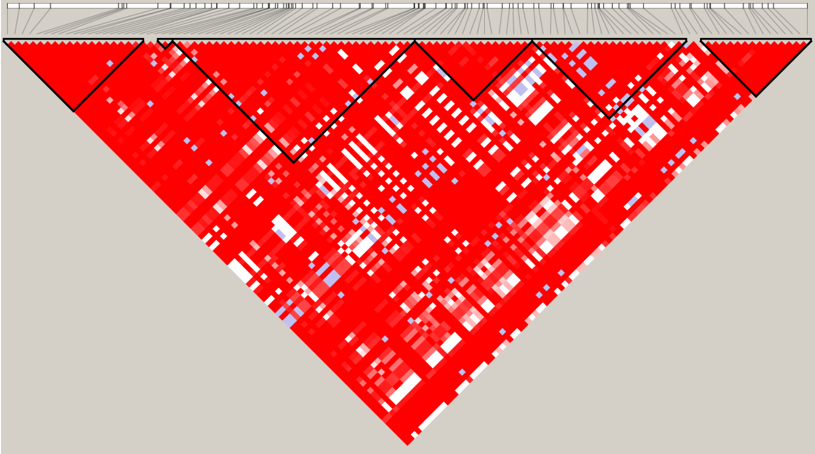
		SNP 1	
		p	1-p
SNP 2	q	π_{11}	π_{12}
	1-q	π_{21}	π_{22}

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

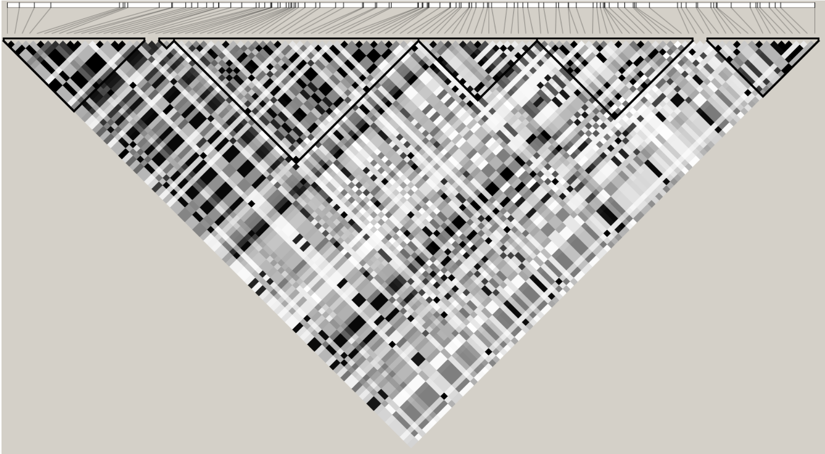
$$D' = D/D_{\text{max}}$$

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

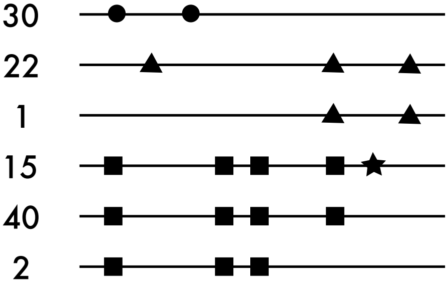
D' for common SNPs in a region of 100kb



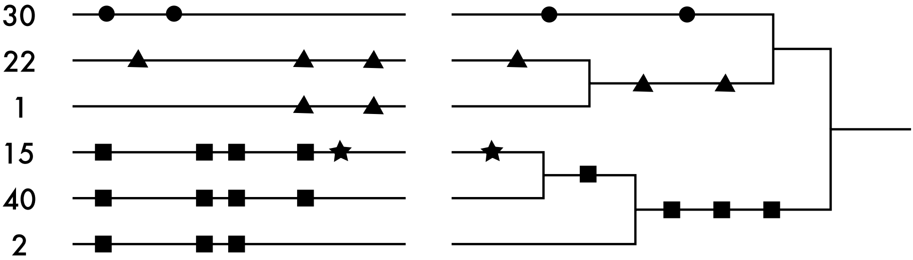
r^2 for common SNPs in a region of 100kb



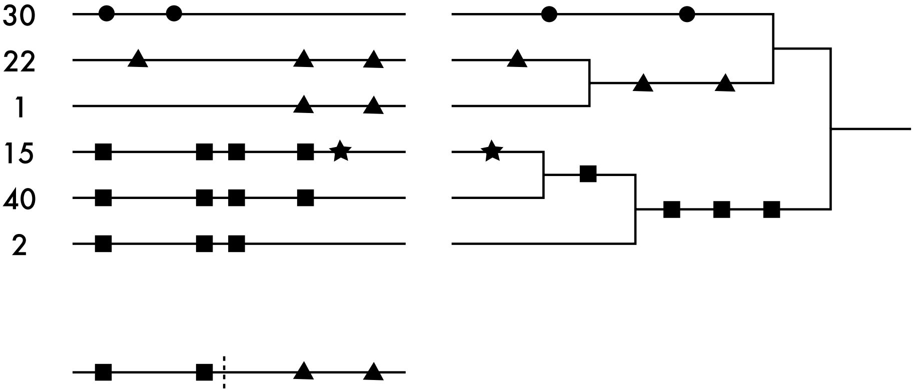
D' and r^2 in a haplotypic context



D' and r^2 in a haplotypic context



D' and r^2 in a haplotypic context



Overview

Human evolution

Genetic variation & linkage disequilibrium

HapMap & tag SNPs

Genome-wide association studies & QC

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)

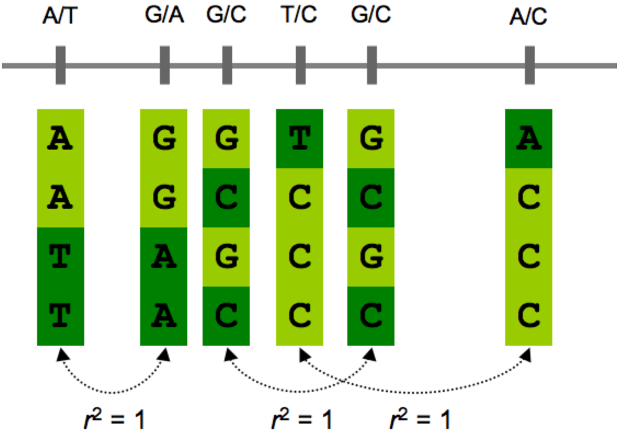
Samples:

- 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)

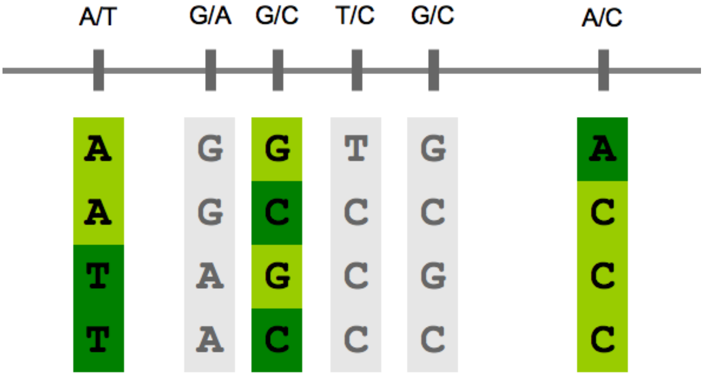
SNPs: Original goal was 1 SNP every 5kb, but as genotyping costs dropped, eventual catalogue included approximately 4 million polymorphic SNPs scattered across the genome.

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

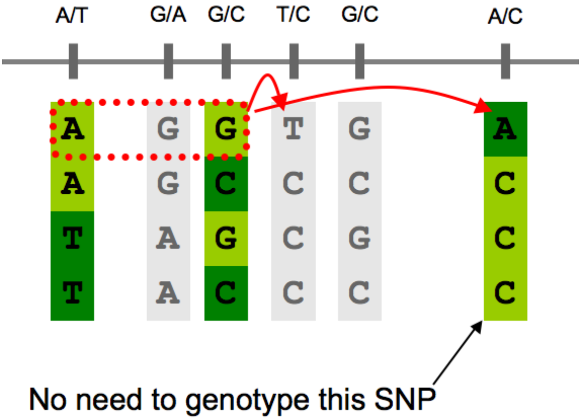
How can we use HapMap knowledge for disease studies?



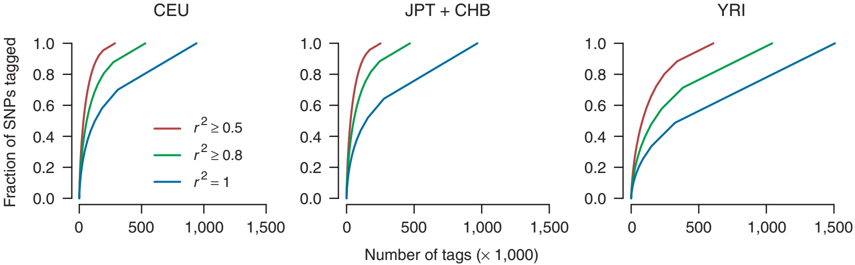
Gain efficiency by removing redundant SNPs



Haplotypes can yield additional gains in efficiency



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



Barrett & Cardon. *Nature Genetics*, 2006.

Overview

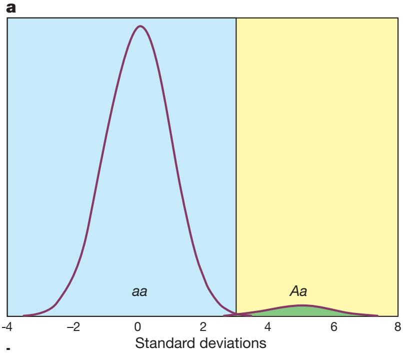
Human evolution

Genetic variation & linkage disequilibrium

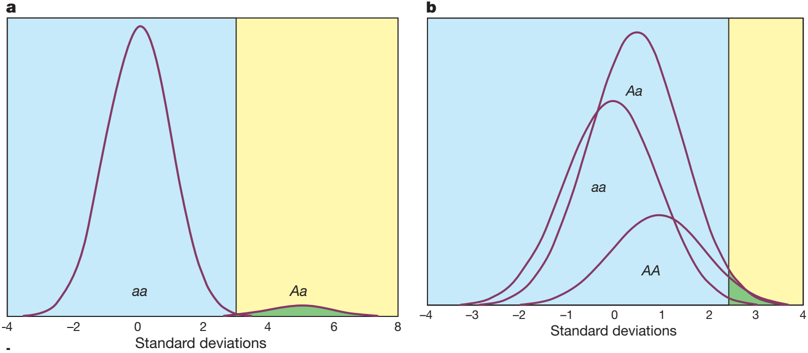
HapMap & tag SNPs

Genome-wide association studies & QC

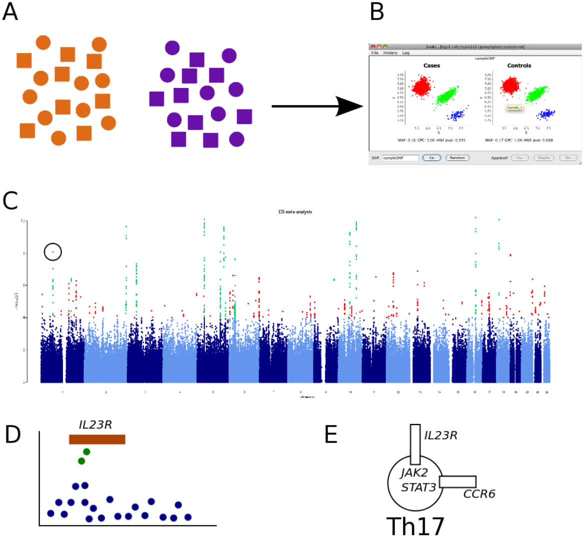
Two competing models to explain genetics of complex traits



Two competing models to explain genetics of complex traits



Genome wide association studies



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?

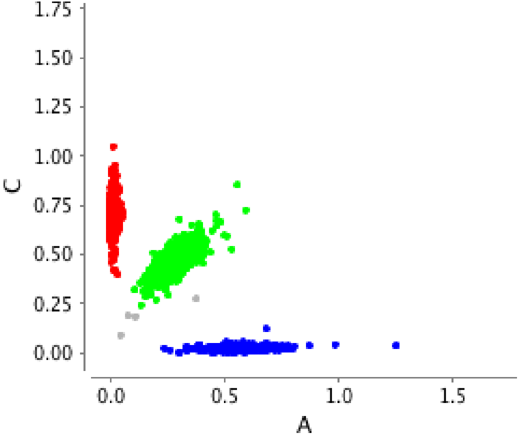
- No common, single SNP main effects: all epistasis, or haplotypes, or rare variation or. . .
- Population structure
- Multiple testing corrections will drown out signal
- Computational burden
- Sample sizes too small to detect the effects
- SNP chips don't cover enough of the genome

Expected challenges

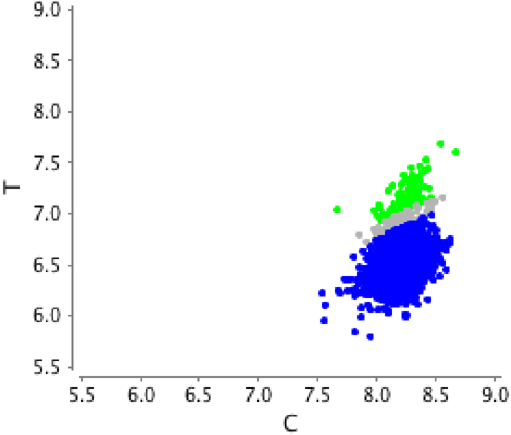
Given that GWAS are feasible, what are the obstacles which stand in the way of finding genes?

- **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

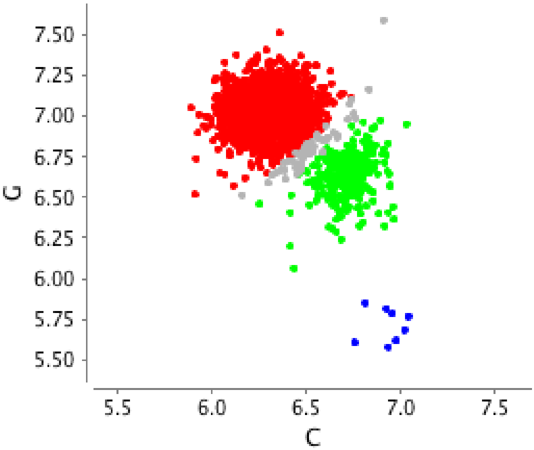
From intensity measurements to genotypes



From intensity measurements to genotypes



From intensity measurements to genotypes



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 largest meta-analyses involve 100 billion genotypes.

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

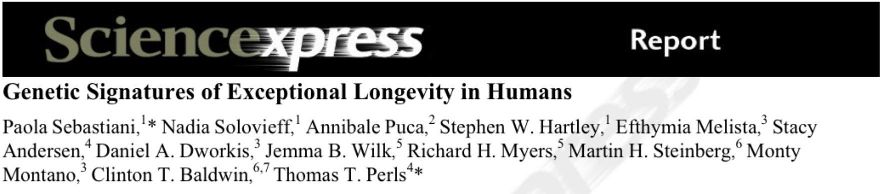
Sample quality control metrics

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.

Clean data matters!

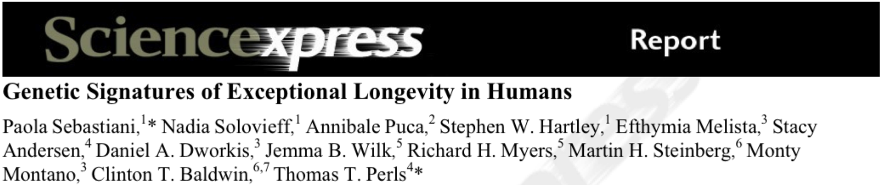


Sciencexpress **Report**

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]*

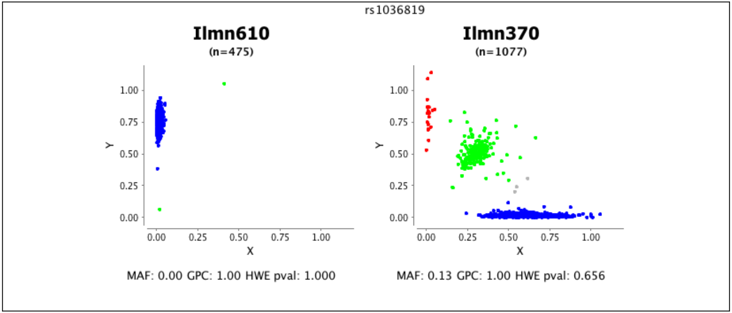
Clean data matters!



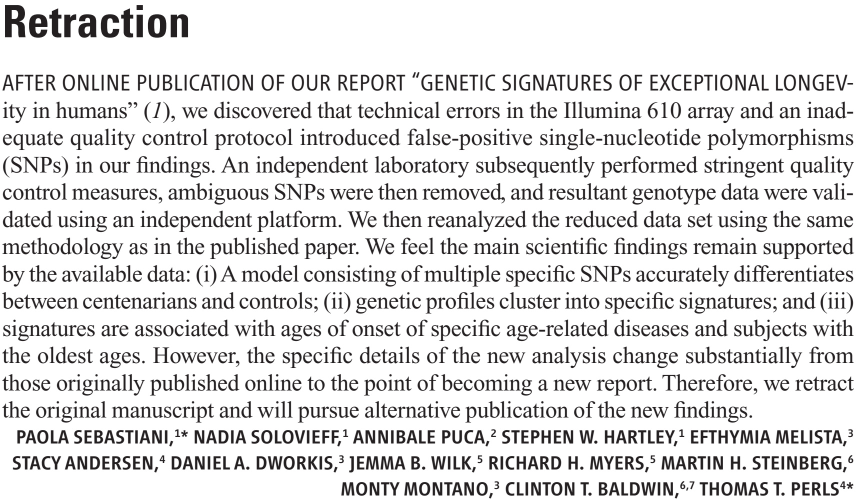
Sciencexpress Report

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]*



Clean data matters!



Retraction

AFTER ONLINE PUBLICATION OF OUR REPORT "GENETIC SIGNATURES OF EXCEPTIONAL LONGEVity in humans" (*1*), we discovered that technical errors in the Illumina 610 array and an inadequate quality control protocol introduced false-positive single-nucleotide polymorphisms (SNPs) in our findings. An independent laboratory subsequently performed stringent quality control measures, ambiguous SNPs were then removed, and resultant genotype data were validated using an independent platform. We then reanalyzed the reduced data set using the same methodology as in the published paper. We feel the main scientific findings remain supported by the available data: (i) A model consisting of multiple specific SNPs accurately differentiates between centenarians and controls; (ii) genetic profiles cluster into specific signatures; and (iii) signatures are associated with ages of onset of specific age-related diseases and subjects with the oldest ages. However, the specific details of the new analysis change substantially from those originally published online to the point of becoming a new report. Therefore, we retract the original manuscript and will pursue alternative publication of the new findings.

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).