Characterizing genomic variation

Why are we so phenotypically different?

Classes of human genetic variation

L2Q1a

Common versus *rare*

Refers to the frequency of the *minor allele* in the human population:

- Common variants = minor allele frequency (MAF) >1% in the population. Also described as *polymorphisms*.
- Rare variants = MAF < 1%

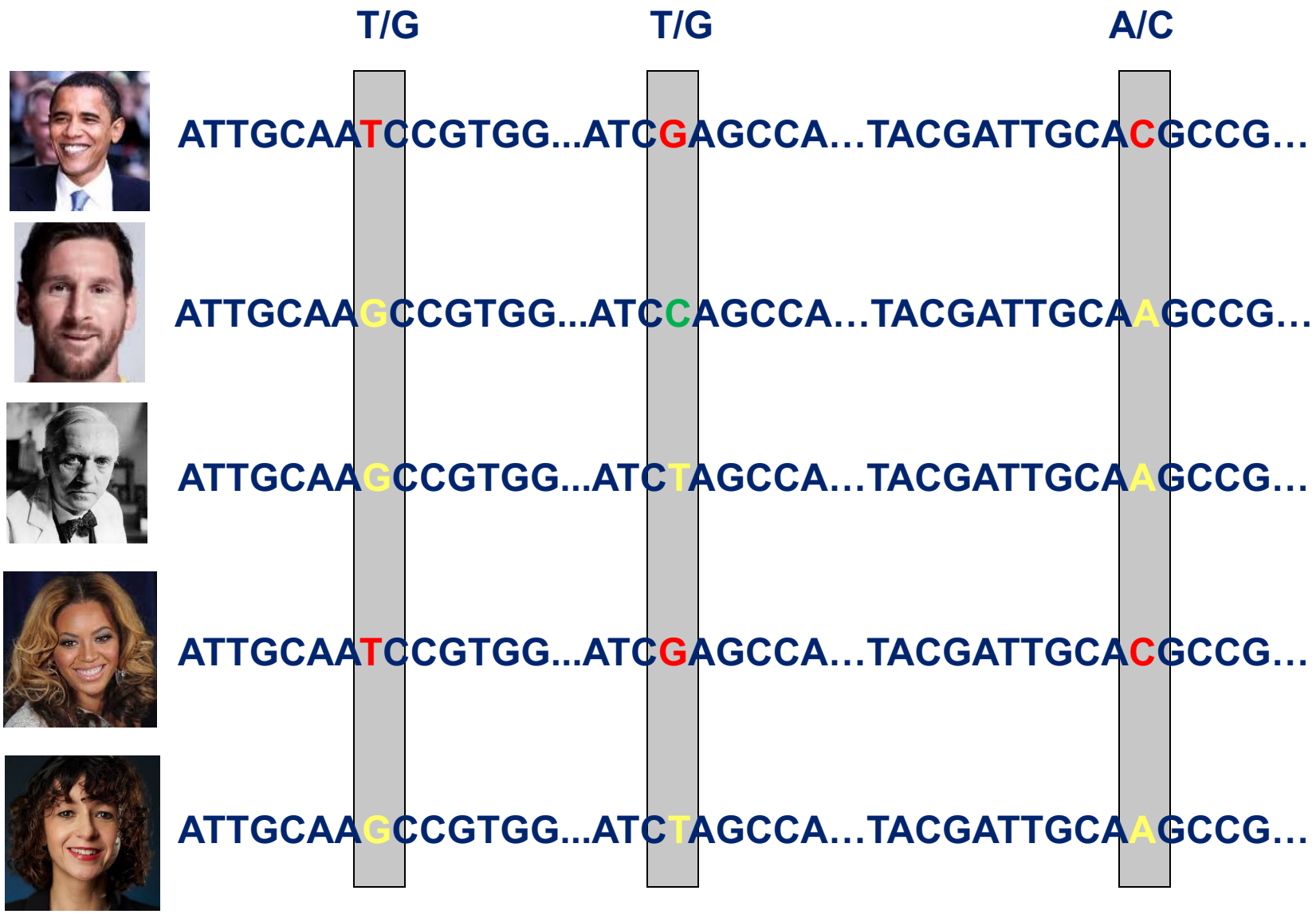
Neutrality:

- The vast majority of genetic variants are likely *neutral* = no contribution to phenotypic variation.
- Some may reach significant frequencies, but this is chance.

Two different nucleotide composition classes:

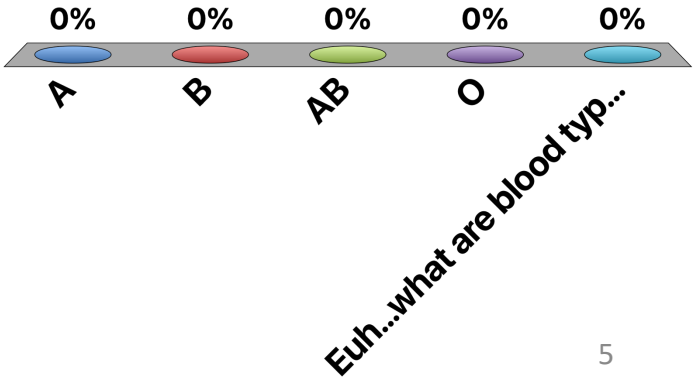
- Single nucleotide variants
- Structural variants

Single nucleotide variants

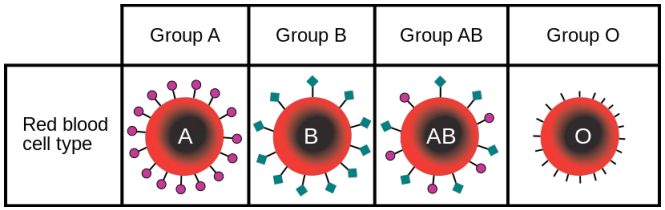


Blood types in the class

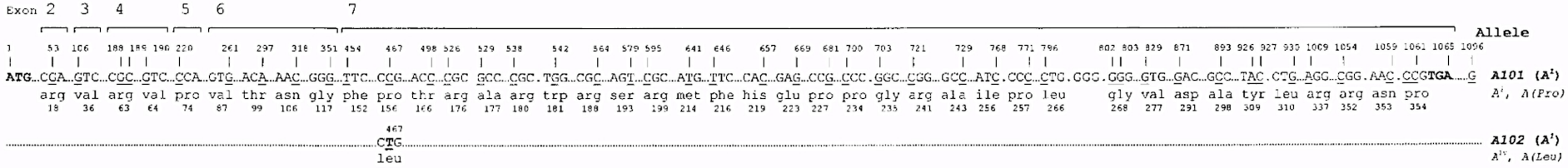
A. A
B. B
C. AB
D. O
E. Euh…what are blood types?

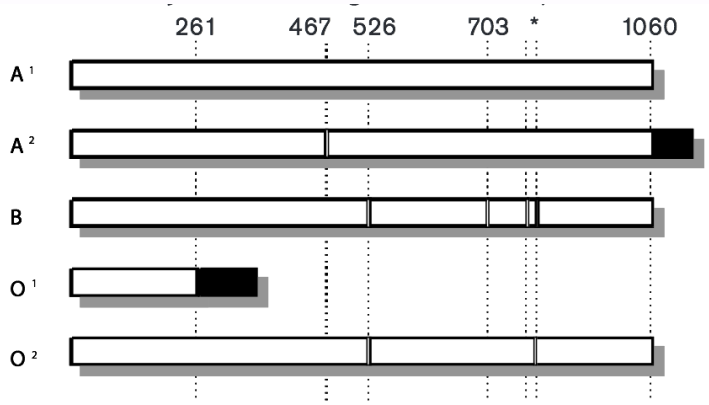


Concrete example: blood types (A, B, AB, & O)



Transferase A, alpha 1-3-N-acetylgalactosaminyltransferase
Transferase B, alpha 1-3-galactosyltransferase





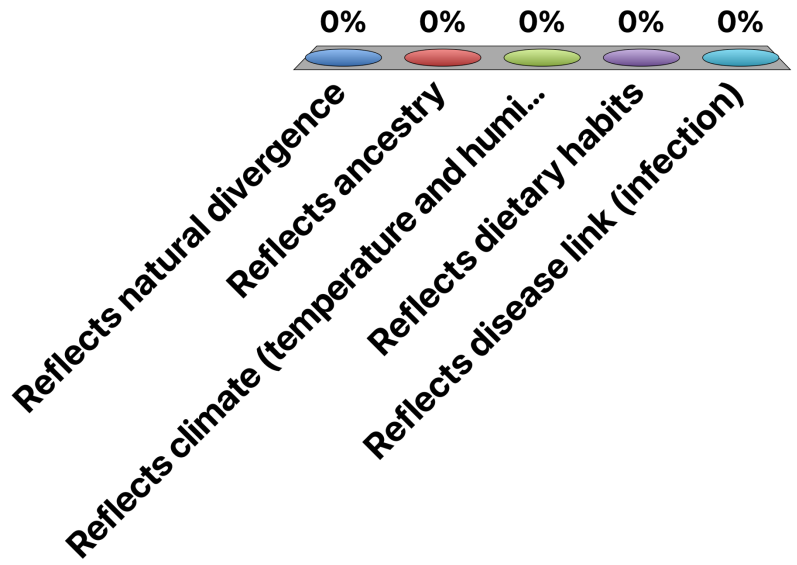
65 “A”alleles, 47 “B” alleles, 58 “O” alleles

L2Q1b Why so many alleles?

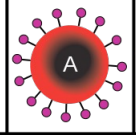
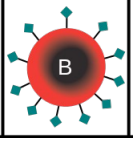
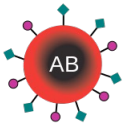
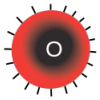
Yip, Annals of Human Genetics, 2002; Storry and Olsen, Immunohematology, 2009; Ridley, Genome, 1999

Why are these blood types maintained?

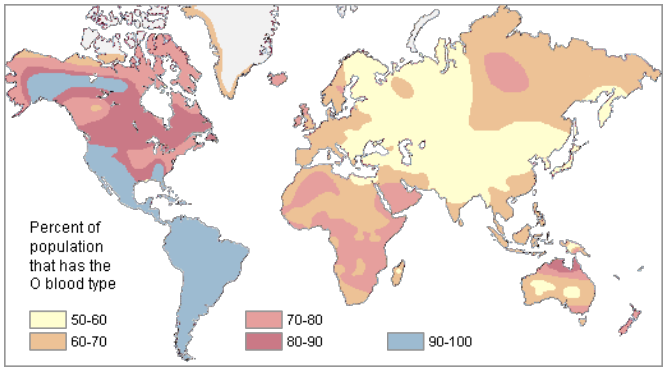
A. Reflects natural divergence
B. Reflects ancestry
C. Reflects climate (temperature and humidity)
D. Reflects dietary habits
E. Reflects disease link (infection)

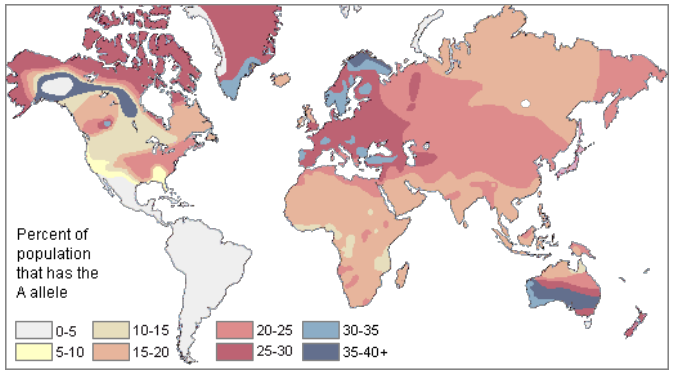


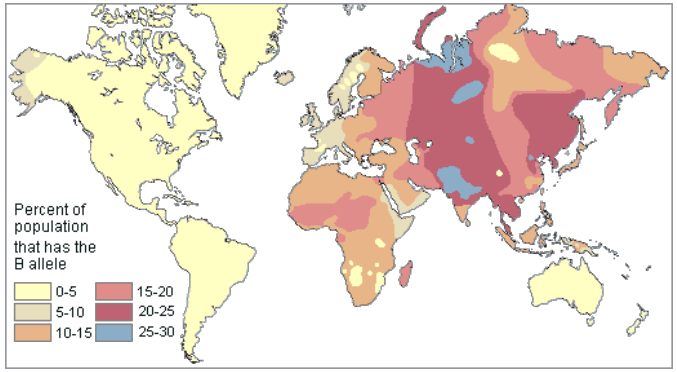
Concrete example: blood types (A, B, AB, & O)

	Group A	Group B	Group AB	Group O
Red blood cell type				

Transferase A, alpha 1-3-N-acetylgalactosaminyltransferase
Transferase B, alpha 1-3-galactosyltransferase





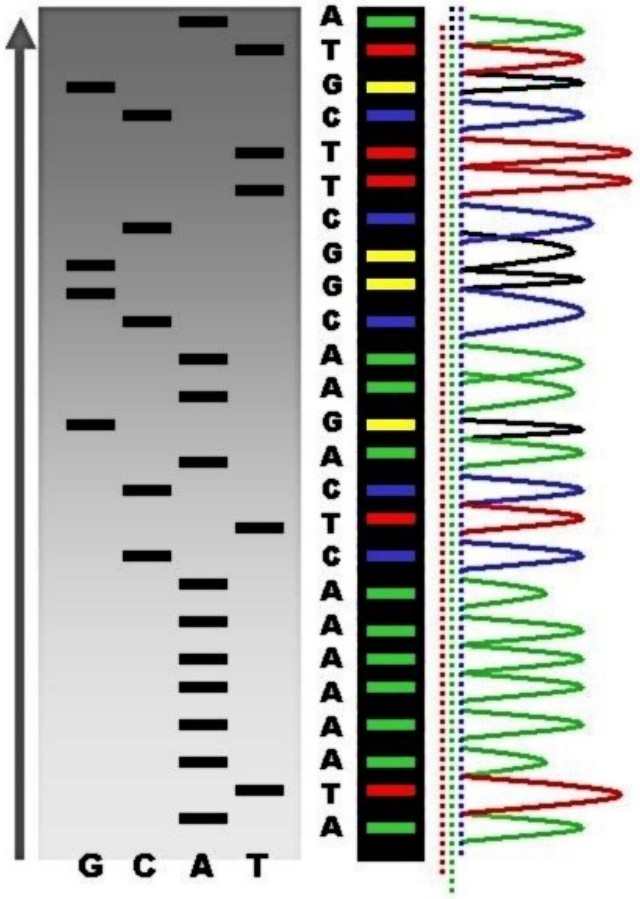


"The genome is a written record of our pathological past"

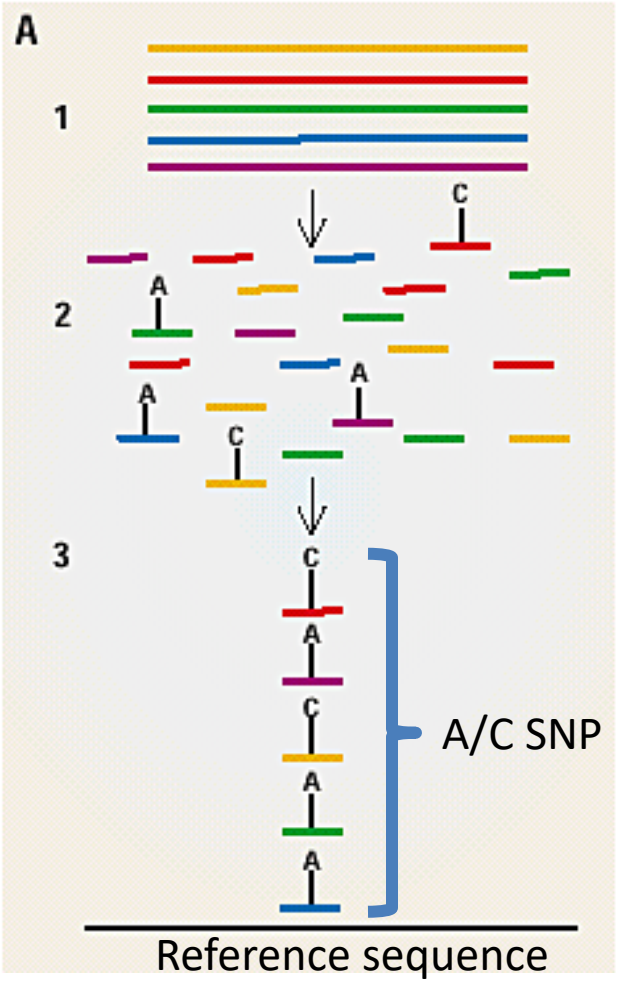
Yip, Annals of Human Genetics, 2002; Storry and Olsen, Immunohematology, 2009; Ridley, Genome, 1999

How are SNPs detected?

Initially: Sanger sequencing



Reduced representation shotgun sequencing followed by genomic alignment



From Rothberg et al. Nature Biotech, 2001

The SNP database - dbSNP

http://www.ncbi.nlm.nih.gov/projects/SNP/

Single nucleotide variants in four human genomes

	(*n*)	In dbSNP (%)
J. Craig Venter's genome	3,213,401	91.0
James D. Watson's genome	3,322,093	81.7
Asian genome	3,074,097	86.4
Yoruban genome	4,139,196	73.6

> (2)

High (1)

> (3)

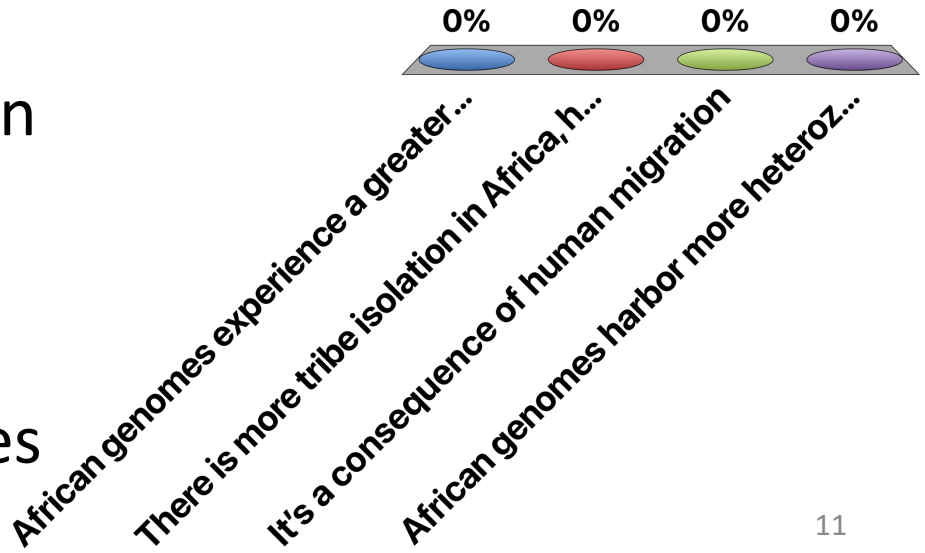
Some conclusions:

1) majority of these sites previously identified as variants in dbSNP → most human high-frequency SNPs (MAF > 10%) have been discovered
2) fewer novel SNPs in J. Craig Venter's genome: his genome contributed heavily to dbSNP
3) But why were more SNPs reported in the "Out of Africa" (Yoruban) genome?

Why were more SNPs reported in "Out of Africa" (Yoruban) genome?

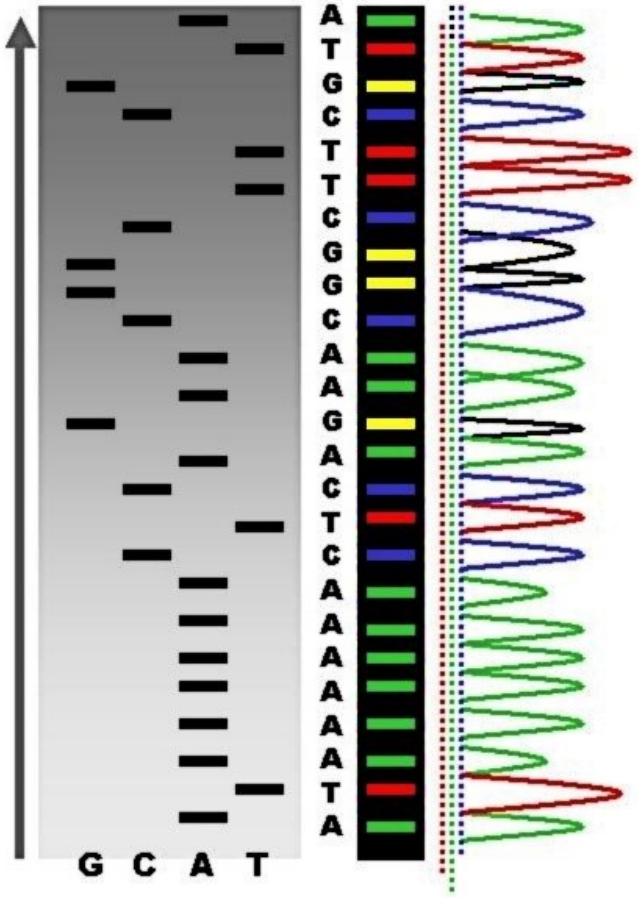
L2Q3

A. African genomes experience a greater mutation rate than e.g. Europeans ones
B. There is more tribe isolation in Africa, hence maintenance of genetic diversity
C. It's a consequence of human migration
D. African genomes harbor more heterozygous sites than European or Asian ones

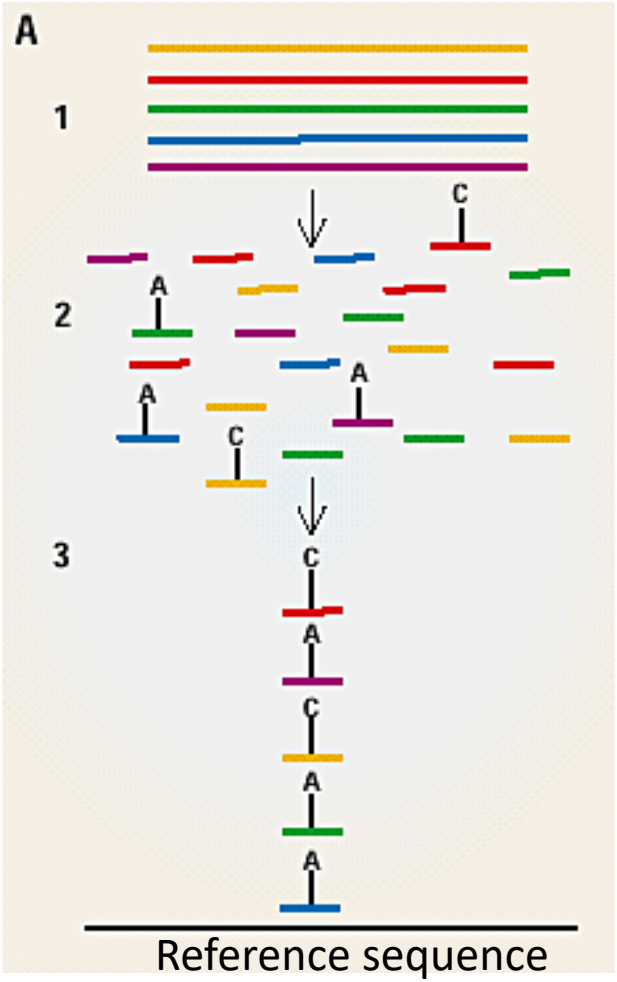


How are SNPs detected?

Initially: Sanger sequencing



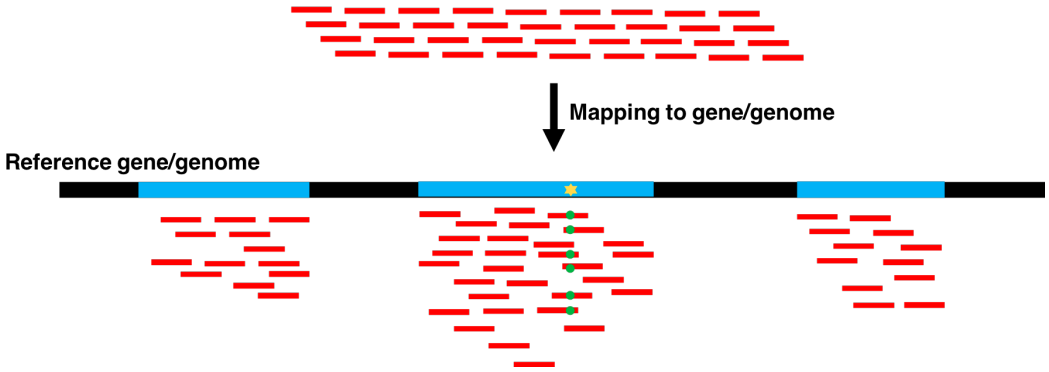
Reduced representation shotgun sequencing followed by genomic alignment



From Rothberg et al. Nature Biotech, 2001

How are SNPs detected?

L2Q2A



Short read alignment is the process of figuring out where in the genome a sequence is from.

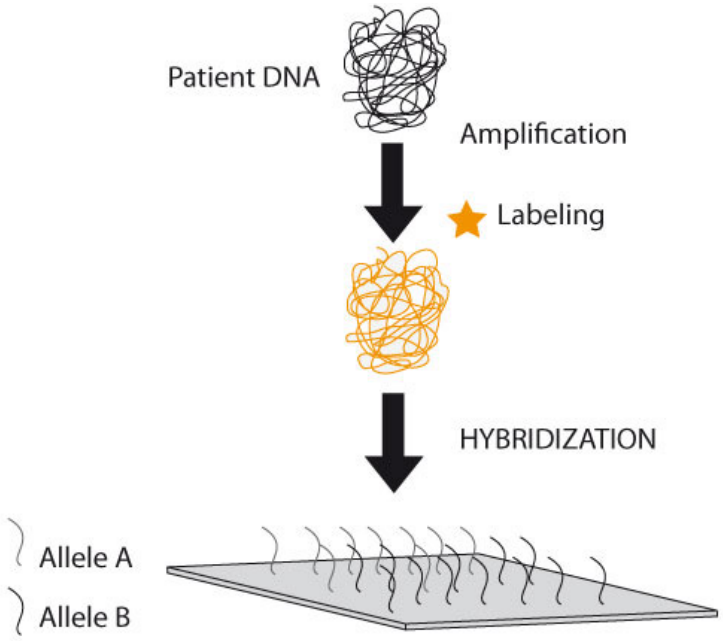
This is a huge computational challenge for several reasons:

1. The reference genome is really big. Searching big things is harder than searching small things.
2. You aren't always looking for *exact* matches in the reference genome–or, at least, probably not (true variation!)
3. Technical variation (bad reads)
4. Repeats, repeats, repeats…and oh, yes, …..repeats

How are SNPs detected?

Another strategy: High-density oligonucleotide arrays

- NGS sequencing is the preferred approach, but even though its cost has come down, sequencing the whole genome of large populations is still too expensive (e.g. 1,000 patients → 1k x 2k / genome = $$$$)
- The most used strategy is still high-density oligonucleotide arrays (only $50 / chip to get information on about 700k SNPs per genome)

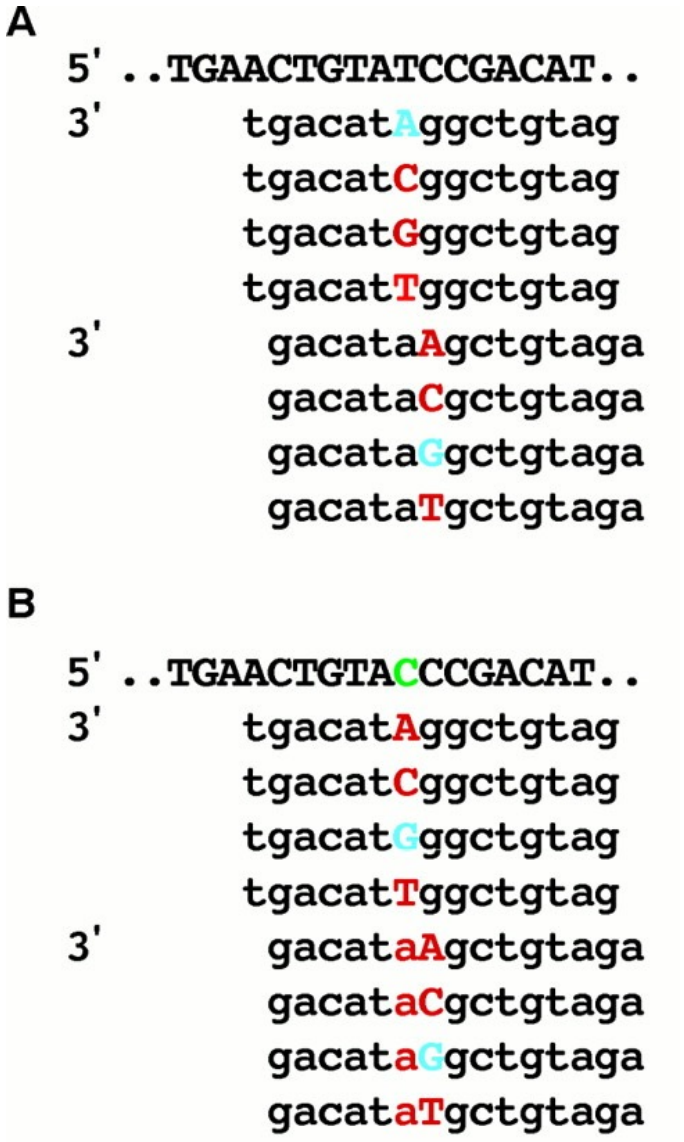


How are SNPs detected?

L2Q2B

Another strategy: High-density oligonucleotide arrays

Chee et al., Science, 1996



- Each position in the target sequence (uppercase) is queried by a set of four probes on the chip (lowercase), identical except at a single position, termed the *substitution position*, which is either A, C, G, or T (blue = complementarity, red = mismatch)
- Two sets of probes are shown, querying adjacent positions in the target

How are SNPs detected?

High-density oligonucleotide arrays

Chee et al., Science, 1996

A

```
5' ..TGAACTGTATCCGACAT..
3'       tgacatAggctgtag
         tgacatCggctgtag
         tgacatGggctgtag
         tgacatTggctgtag
3'        gacataAgctgtaga
          gacataCgctgtaga
          gacataGgctgtaga
          gacataTgctgtaga
```

B

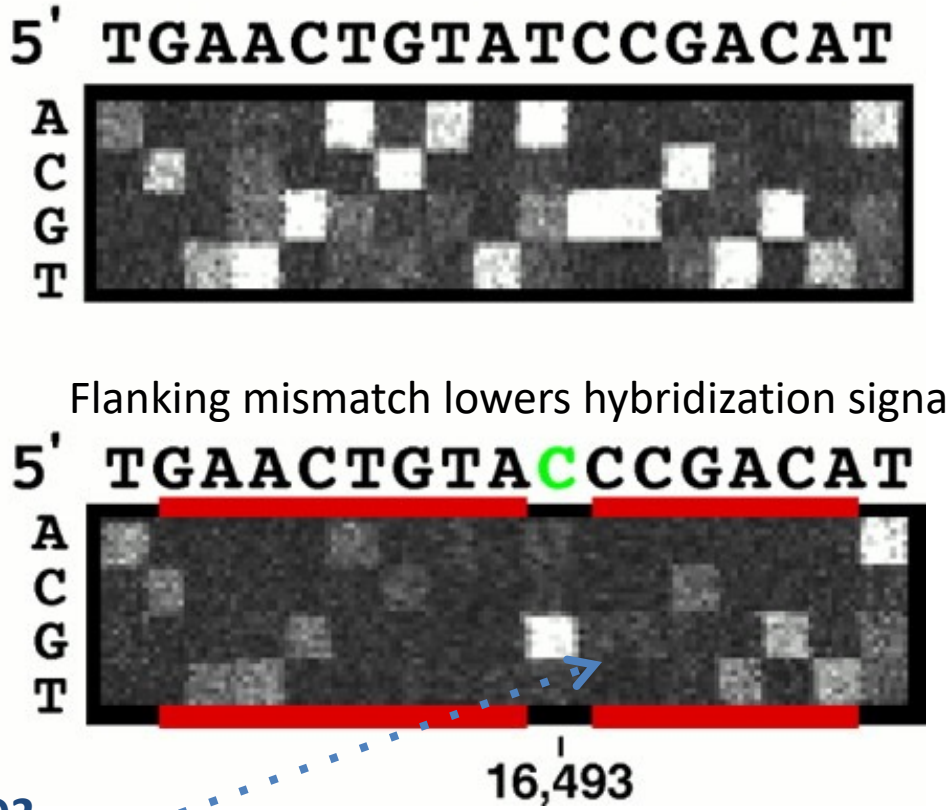
```
5' ..TGAACTGTACCCGACAT..
3'       tgacatAggctgtag
         tgacatCggctgtag
         tgacatGggctgtag
         tgacatTggctgtag
3'        gacataAgctgtaga
          gacataCgctgtaga
          gacataGgctgtaga
          gacataTgctgtaga
```

Reference

SNP

C

Simple 5' to 3' read-out



L2Q2

Because of the «A» mismatch, those probes yield a low hybridization signal around the mismatch as can be seen here (the same for those upstream or further downstream of the SNP)

The SNP database - dbSNP

http://www.ncbi.nlm.nih.gov/projects/SNP/

Single nucleotide variants in four human genomes

	(*n*)	In dbSNP (%)
J. Craig Venter's genome	3,213,401	91.0
James D. Watson's genome	3,322,093	81.7
Asian genome	3,074,097	86.4
Yoruban genome	4,139,196	73.6

> (2)

> (3)

High (1)

Some conclusions:

1) majority of these sites previously identified as variants in dbSNP → most human high-frequency SNPs (MAF > 10%) have been discovered
2) fewer novel SNPs in J. Craig Venter's genome: his genome contributed heavily to dbSNP
3) Yoruban genome → greater diversity (class discussion)

Single nucleotide variants

L2Q4

- Estimated that the human genome contains > 11 million SN*Ps* (~7 million with MAF > 5%, rest between 1-5%)
- Each individual between 3-4 million SNPs dependent on ethnicity
- Unknown how many rare or even novel ("*de novo*") SNPs

L2Q5

To know the genotype of a person, do we always need to sequence or can we target a subset of SNPs to infer the rest of the genetic information? ***(because designing e.g. an array targeting more than 11,000,000 SNPs would be really difficult)***

No! We can exploit the principle of linkage disequilibrium!

Single nucleotide variants

L2Q4

• Estimated that the human genome contains > 11 million SN*Ps* (~7 million with MAF > 5%, rest between 1-5%)

• Each individual between 3-4 million SNPs dependent on ethnicity

• Unknown how many rare or even novel ("*de novo*") SNVs

• SNP alleles in the same genomic interval are often correlated with one another → "*Linkage disequilibrium (LD)*" = Nonrandom association of alleles – varies in complex and unpredictable manner across the genome and between different populations.

• International HapMap Project → can we divide the genome into groups of highly correlated SNPs that are generally inherited together = "*LD bins*"

Number of tag SNPs required to capture common Phase II SNPs

L2Q5 To know the genotype of a person, do we always need to sequence or can we target a subset of SNPs to infer the rest of the genetic information?

Threshold	CEU
$r^2 \geq 0.5$	290,969
$r^2 \geq 0.8$	552,853
$r^2 = 1.0$	1,024,665

Based on genotyping over 3.1 million SNPs in 270 individuals from 4 geographically diverse populations (Frazer et al., Nature, 2007)

SNPs with correlation $r^2 = 1.0$ are "statistically indistinguishable"
SNPs with $r^2 >= 0.8$ can be considered to be of high enough similarity [i.e., SNP 1 has 80% power to predict the genotypes of SNP 2 and vice versa]

Thus, by targeting only one “representative SNP”, we can get information on the nature of other SNPs and dependent on the implemented r^2, either with 100% confidence ($r^2 = 1$) or with less if a lower r^2 is used…..(i.e. called imputing missing genetic information), so it is a trade-off.

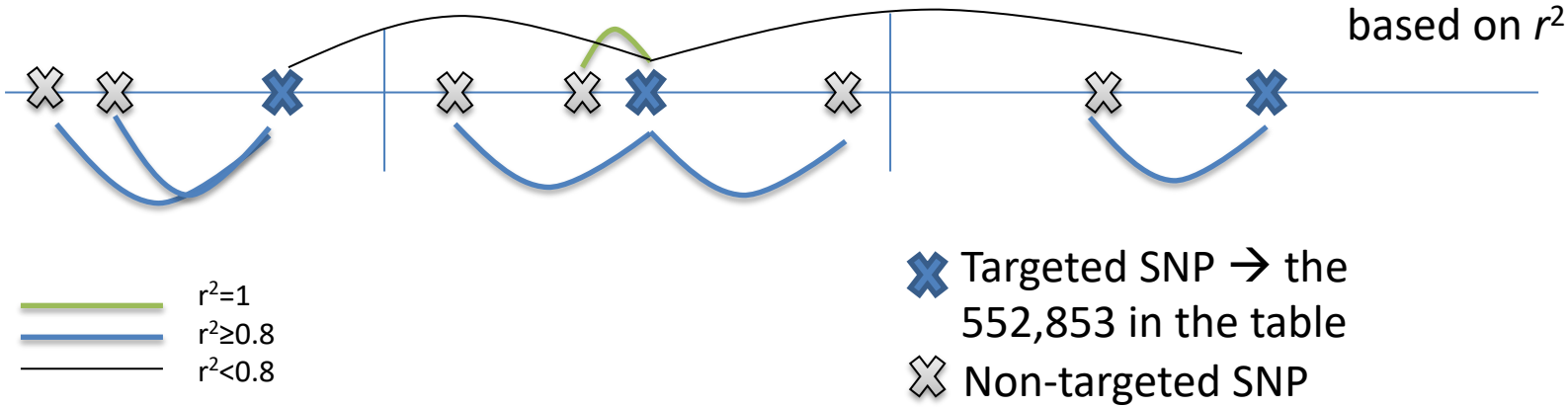
L2Q5

To know the genotype of a person, do we always need to sequence or can we target a subset of SNPs to infer the rest of the genetic information?

Threshold	CEU
$r^2 \geq 0.5$	290,969
$r^2 \geq 0.8$	552,853
$r^2 = 1.0$	1,024,665

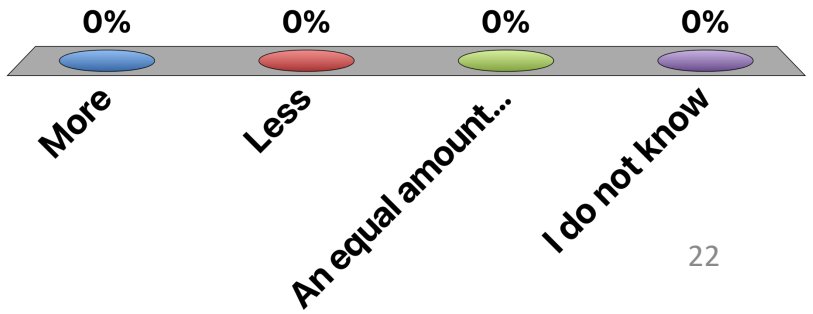
Based on genotyping over 3.1 million SNPs in 270 individuals from 4 geographically diverse populations (Frazer et al., Nature, 2007)

L2Q5



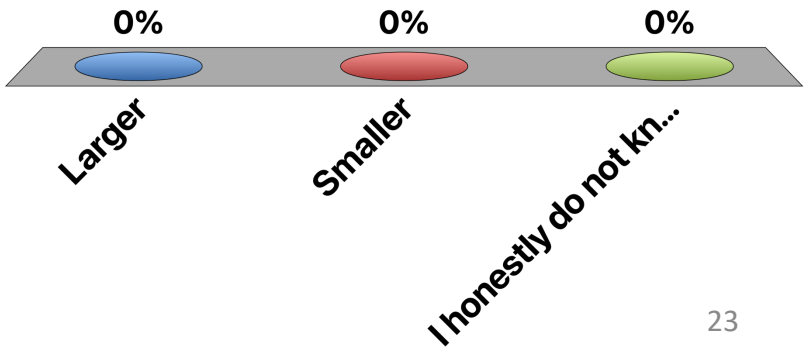
In the Yoruban genome, do we need to probe …(fill in)….. SNPs than the CEU one to achieve the same threshold of >=0.8?

A. More
B. Less
C. An equal amount of
D. I do not know



Is the average LD bin size in the Yoruban genome larger or smaller than in the European genome?

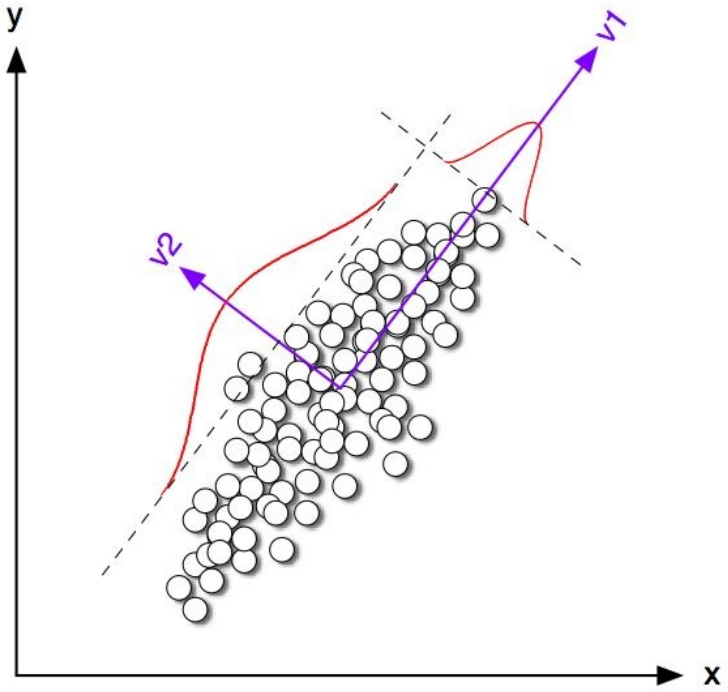
A. Larger
B. Smaller
C. I honestly do not know



Population Stratification

Subdivision of a population into different ethnic groups with potentially different marker allele frequencies and thus different disease prevalence

From Sven Bergmann, UNIL

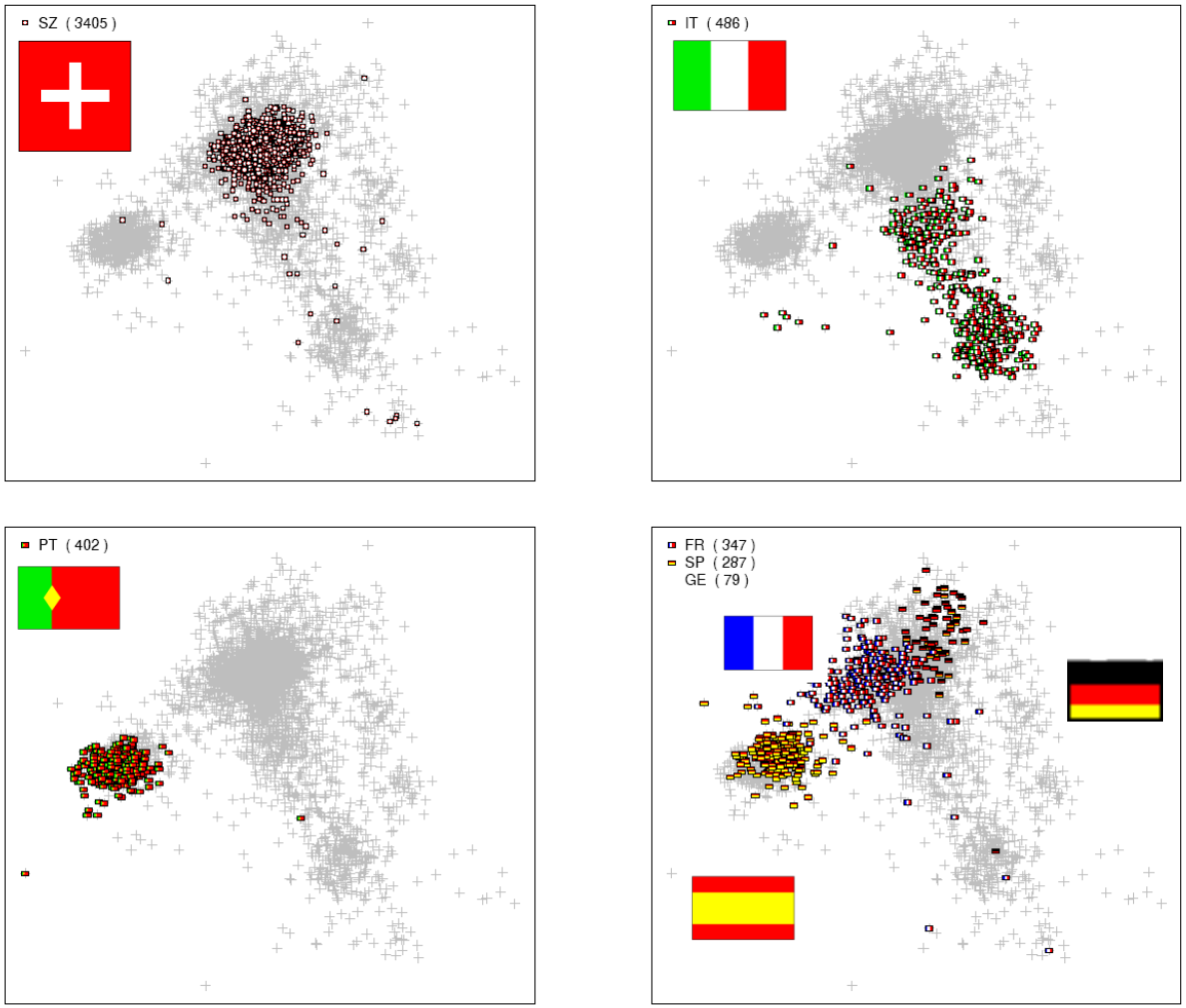


- Circles depict the SNP profile of human individuals
- The farther circles are from each other, the more genetically the respective people diverge
- Here two components can be identified that explain a large part of genetic variation

Principle Component Analysis reveals SNP vectors explaining largest variation in the data

Population Stratification

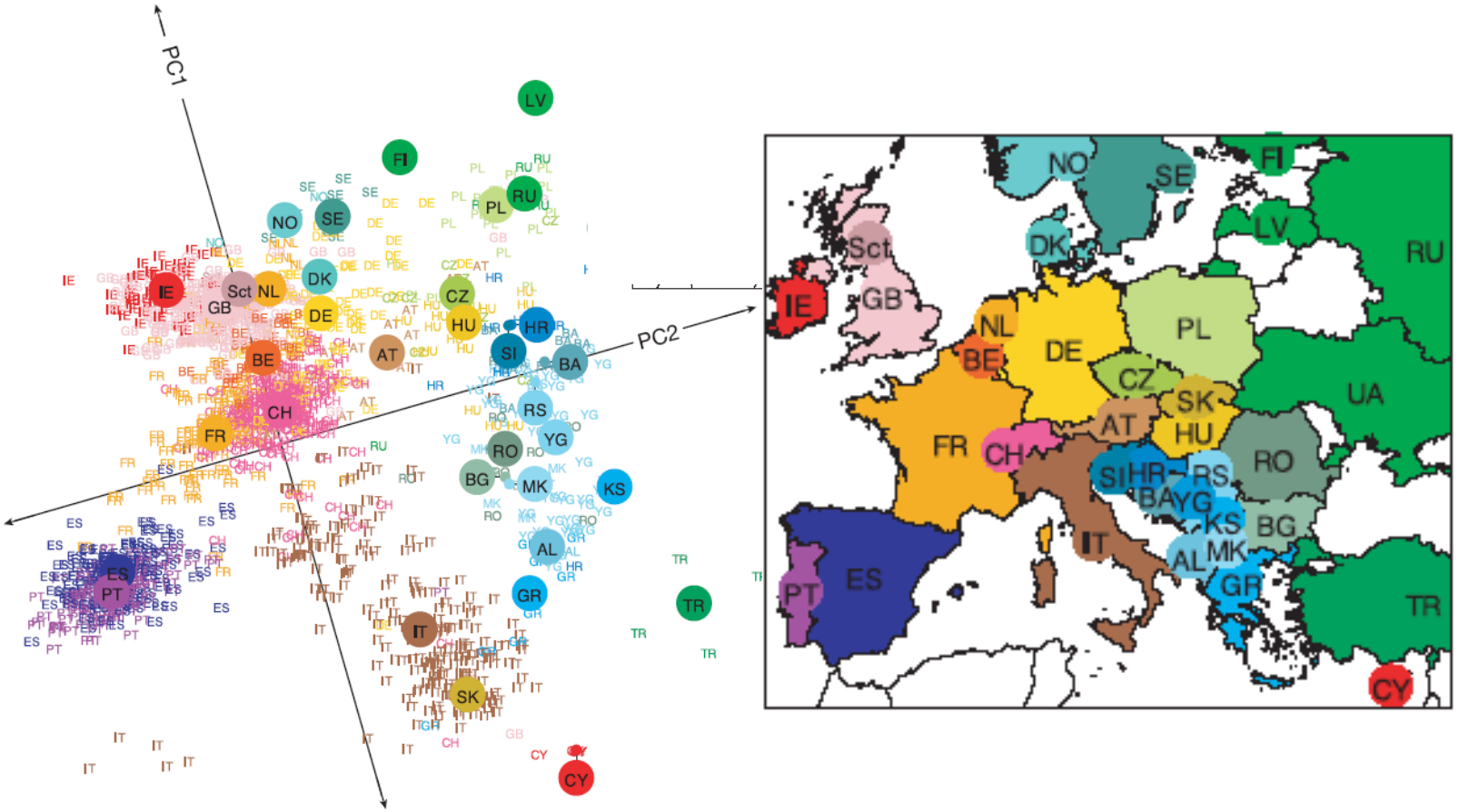
Ethnic groups cluster according to geographic distances



From Sven Bergmann, UNIL

Population Stratification

Principle component analysis of European human genomic variation



Genes mirror geography within Europe

nature Vol 456|6 November 2008|doi:10.1038/nature07331

John Novembre[1,2], Toby Johnson[4,5,6], Katarzyna Bryc[7], Zoltán Kutalik[4,6], Adam R. Boyko[7], Adam Auton[7], Amit Indap[7], Karen S. King[8], Sven Bergmann[4,6], Matthew R. Nelson[8], Matthew Stephens[2,3] & Carlos D. Bustamante[7]

Using genomics to elucidate human history:

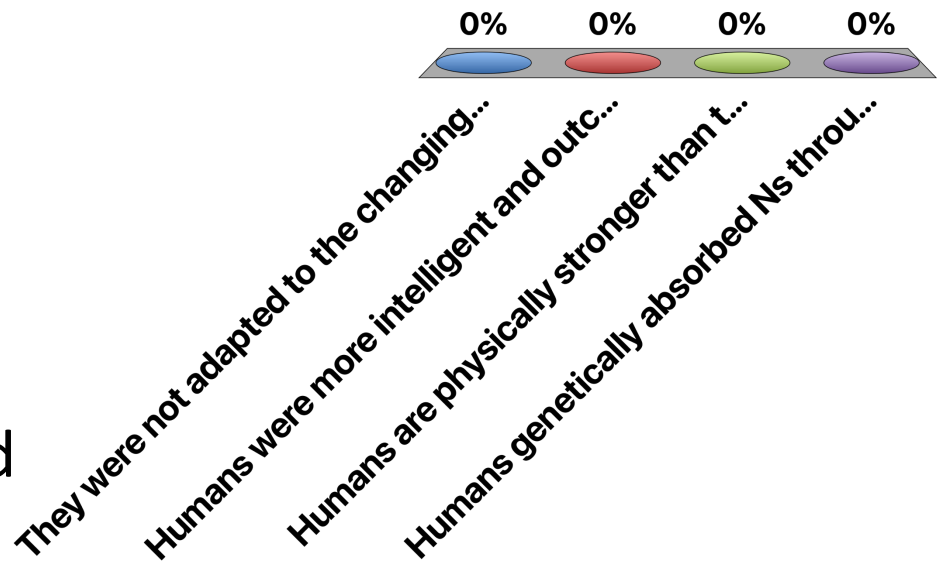
Why did Neanderthals (Ns) disappear in favor of modern humans?

A. They were not adapted to the changing climate

B. Humans were more intelligent and outcompeted Ns in food collection / people care

C. Humans are physically stronger than the “small” hobbit-like Ns and likely slaughtered most Ns

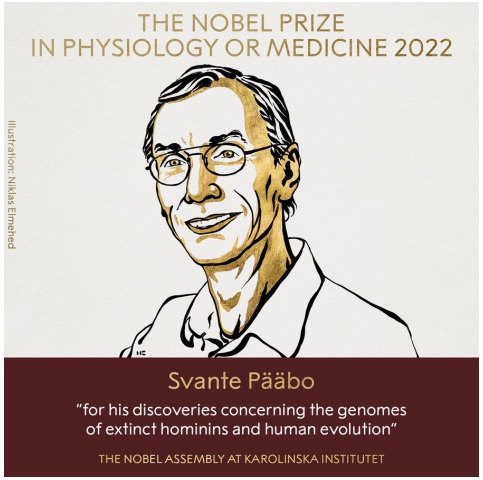
D. Humans genetically absorbed Ns through interbreeding



Using genomics to elucidate human history

Comparing the human versus the Neanderthal genome

- Europeans and Asians (not Africans!) have all inherited 1–4% of their genome from Neanderthals (“we are all a bit Neanderthal”): **sign of “admixture” (interbreeding)**
- Does it matter?



Using genomics to elucidate human history

Comparing the human versus the Neanderthal genome

- Europeans and Asians (not Africans!) have all inherited 1–4% of their genome from Neanderthals ("we are all a bit Neanderthal"): **sign of "admixture" (interbreeding)**
- Does it matter?

Article | Published: 30 September 2020

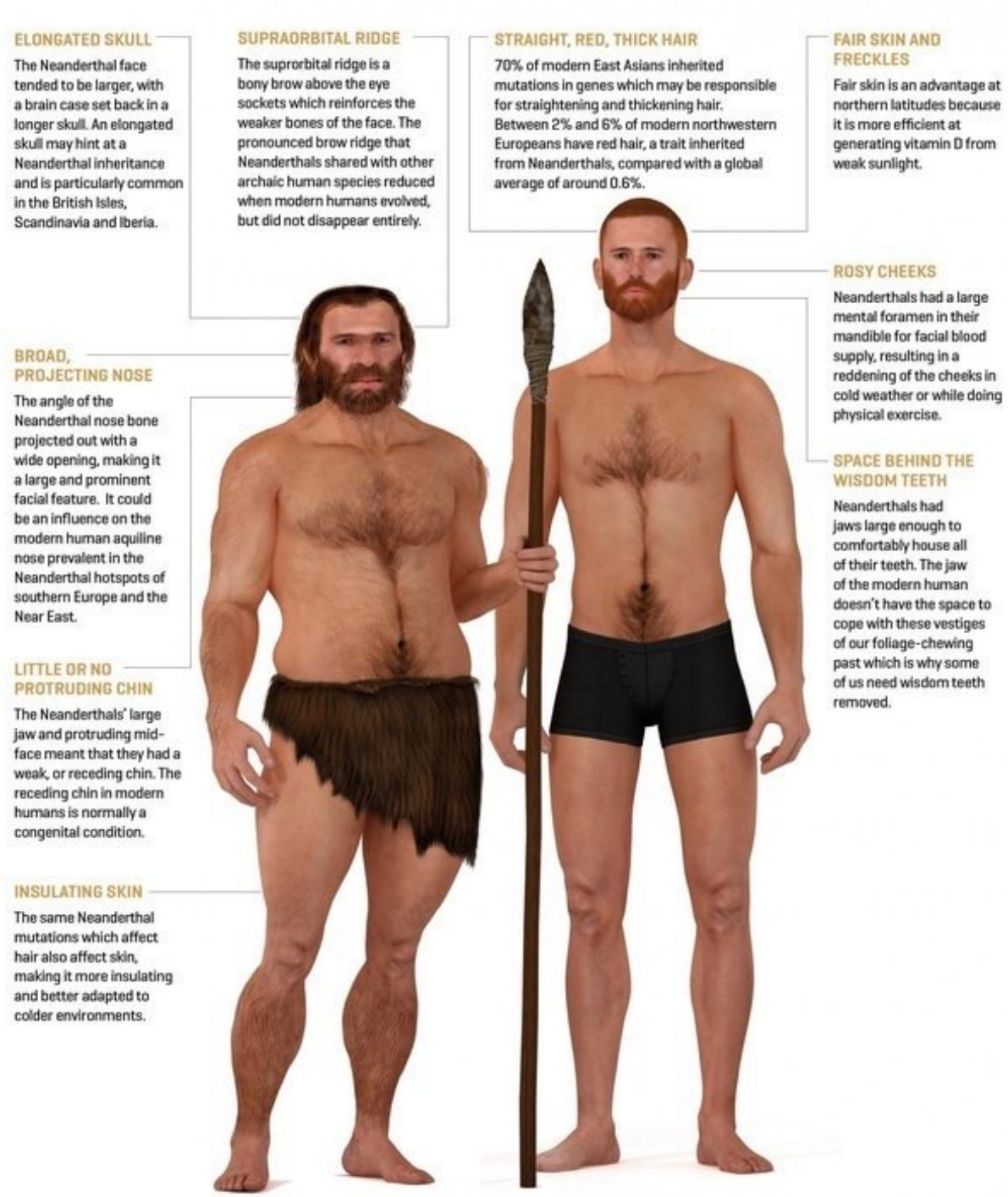
The major genetic risk factor for severe COVID-19 is inherited from Neanderthals

Hugo Zeberg & Svante Pääbo

Nature **587**, 610–612 (2020) | Cite this article

Abstract

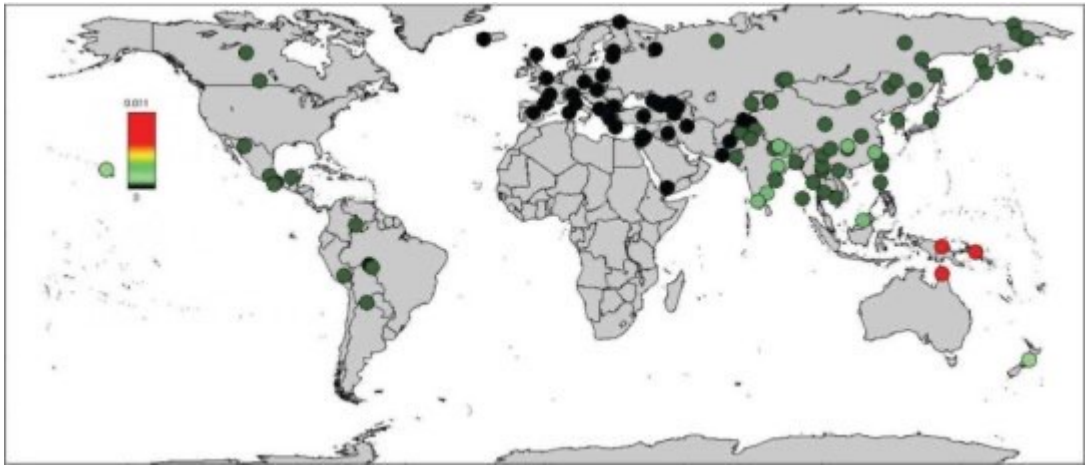
A recent genetic association study[1] identified a gene cluster on chromosome 3 as a risk locus for respiratory failure after infection with severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). A separate study (COVID-19 Host Genetics Initiative)[2] comprising 3,199 hospitalized patients with coronavirus disease 2019 (COVID-19) and control individuals showed that this cluster is the major genetic risk factor for severe symptoms after SARS-CoV-2 infection and hospitalization. Here we show that the risk is conferred by a genomic segment of around 50 kilobases in size that is inherited from Neanderthals and is carried by around 50% of people in south Asia and around 16% of people in Europe.


ELONGATED SKULL
The Neanderthal face tended to be larger, with a brain case set back in a longer skull. An elongated skull may hint at a Neanderthal inheritance and is particularly common in the British Isles, Scandinavia and Iberia.
SUPRAORBITAL RIDGE
The suprorbital ridge is a bony brow above the eye sockets which reinforces the weaker bones of the face. The pronounced brow ridge that Neanderthals shared with other archaic human species reduced when modern humans evolved, but did not disappear entirely.
STRAIGHT, RED, THICK HAIR
70% of modern East Asians inherited mutations in genes which may be responsible for straightening and thickening hair. Between 2% and 6% of modern northwestern Europeans have red hair, a trait inherited from Neanderthals, compared with a global average of around 0.6%.
FAIR SKIN AND FRECKLES
Fair skin is an advantage at northern latitudes because it is more efficient at generating vitamin D from weak sunlight.
ROSY CHEEKS
Neanderthals had a large mental foramen in their mandible for facial blood supply, resulting in a reddening of the cheeks in cold weather or while doing physical exercise.
SPACE BEHIND THE WISDOM TEETH
Neanderthals had jaws large enough to comfortably house all of their teeth. The jaw of the modern human doesn't have the space to cope with these vestiges of our foliage-chewing past which is why some of us need wisdom teeth removed.
BROAD, PROJECTING NOSE
The angle of the Neanderthal nose bone projected out with a wide opening, making it a large and prominent facial feature. It could be an influence on the modern human aquiline nose prevalent in the Neanderthal hotspots of southern Europe and the Near East.
LITTLE OR NO PROTRUDING CHIN
The Neanderthals' large jaw and protruding mid-face meant that they had a weak, or receding chin. The receding chin in modern humans is normally a congenital condition.
INSULATING SKIN
The same Neanderthal mutations which affect hair also affect skin, making it more insulating and better adapted to colder environments.

But many mysteries remain….

Denisovan: The unknown ancestor!

Photo credit: Thilo Parg

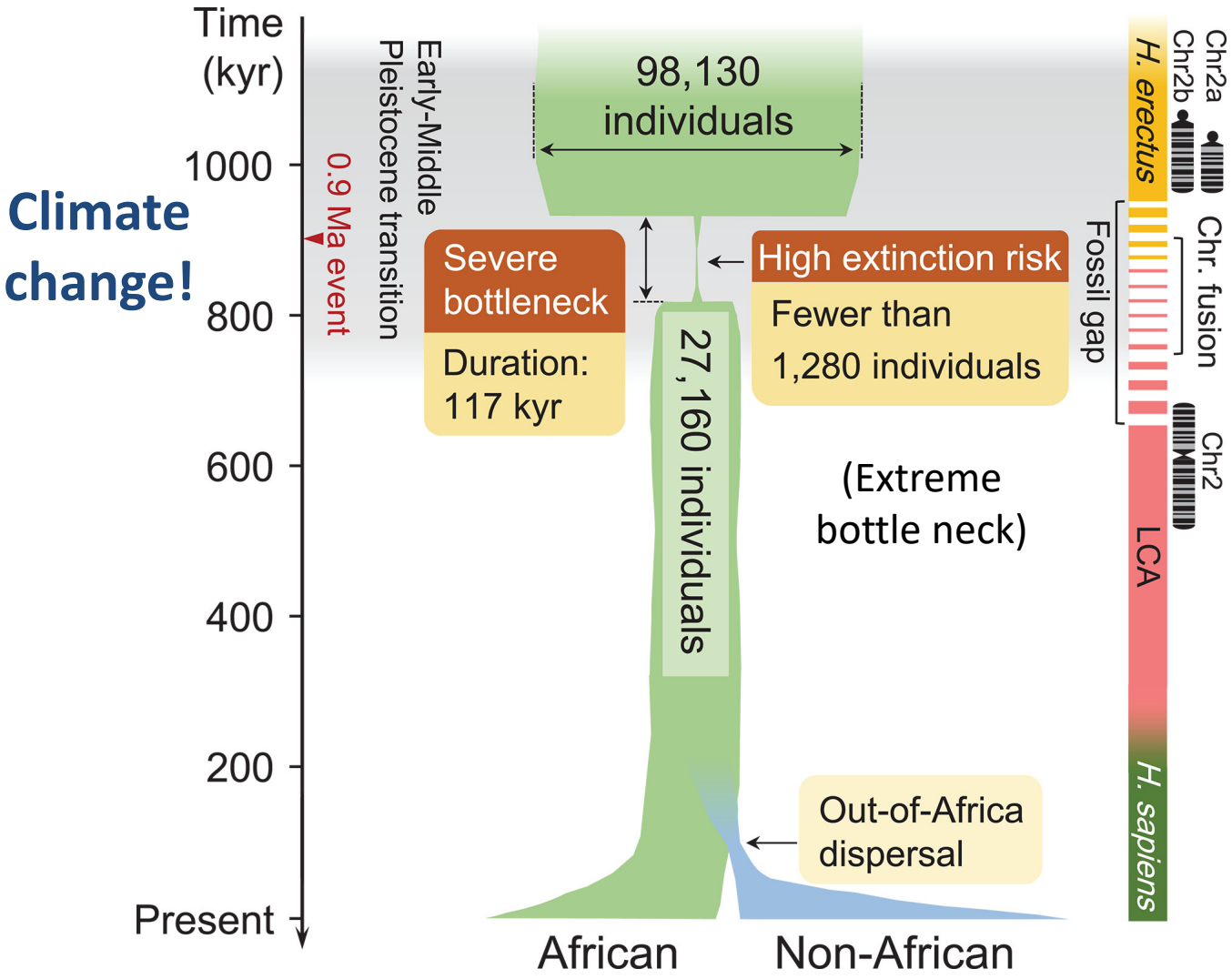


"Lord of the Rings–type world" in which multiple human species coexisted and mated

This map shows the proportion of the genome inferred to be Denisovan in ancestry in diverse non-Africans.

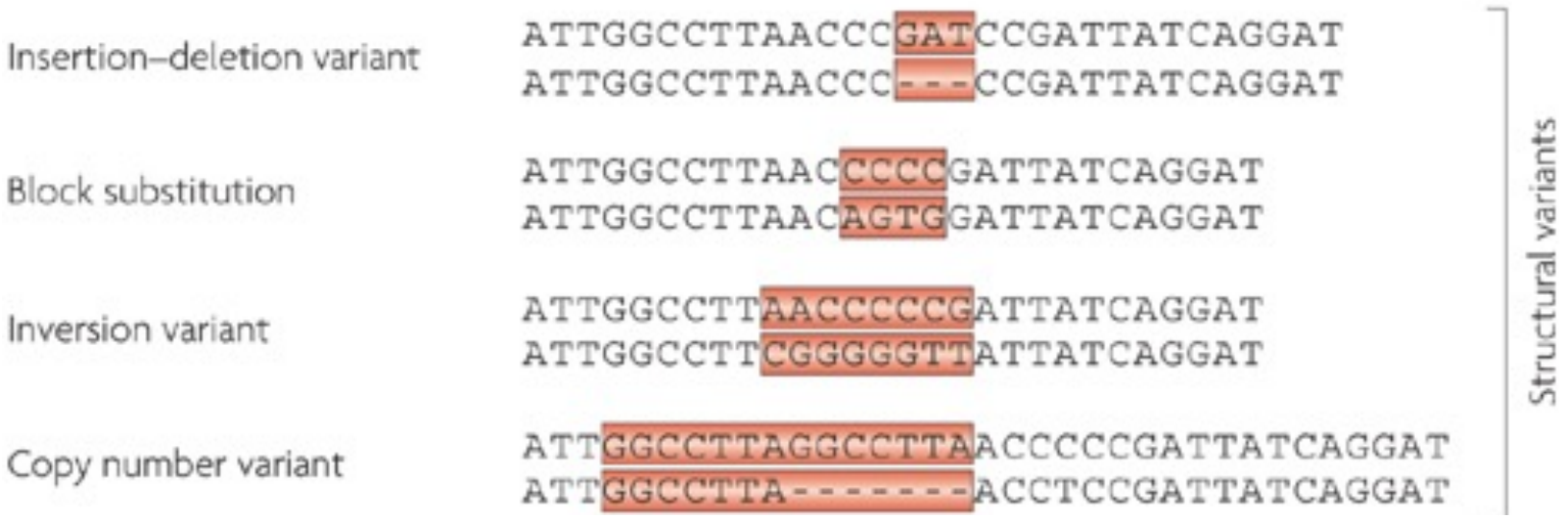
But many mysteries remain….

Homo sapiens was almost not meant to be



Hu et al., Science, 2023

Structural variants (SVs)



(Frazer et al., Nature Rev. Genetic., 2010)

SVs are more difficult to detect than SNPs!

Structural variants in the Venter genome		
	(*n*)	**length (bp)**
Block substitutions	53,823	2–206
Indels (heterozygous)	851,575	1–82,711
Inversions	90	7–670,345
Copy number variants	62	8,855–1,925,949

A map of human genome variation from population-scale sequencing, Nature, 2010

A global reference for human genetic variation (>2,500 genomes), Nature 2015

- Millions of variants are added to the variant databases
- Involved --> a dozen different algorithms
- Still a difficult problem
- Engineers!

L2Q7a

Some insights about a "typical genome":

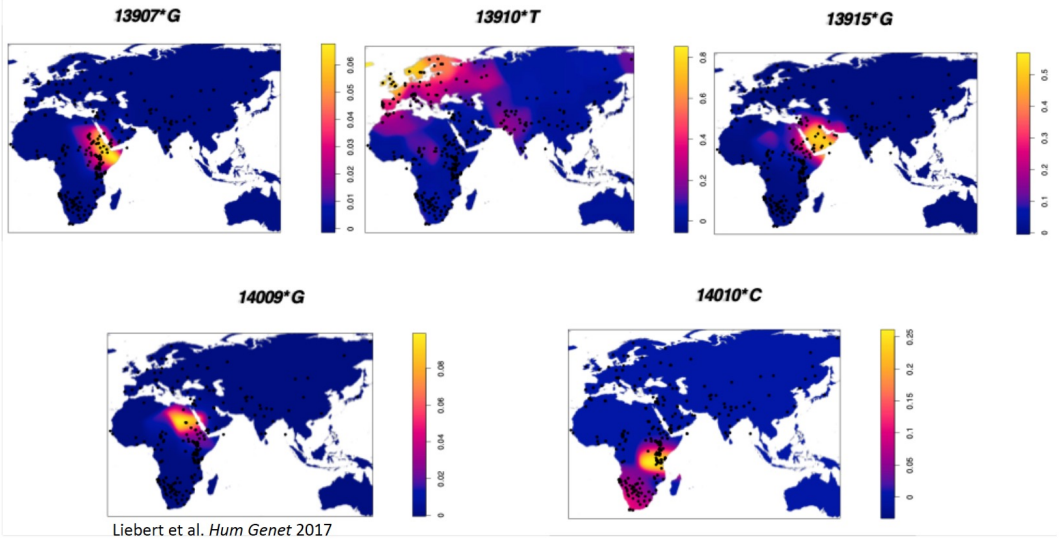
- 149–182 sites with protein truncating variants
- 10,000 to 12,000 sites with peptide sequence-altering variants
- 459,000 to 565,000 variant sites overlapping known regulatory regions (untranslated regions (UTRs))

Are we still evolving?

Are we still evolving?

Convergent evolution of lactase persistence

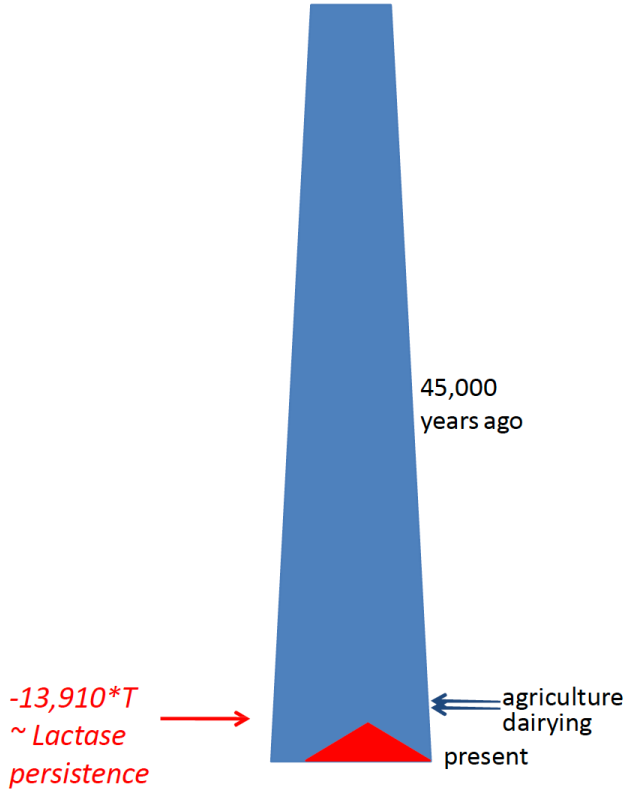
Yes! L2Q7b



Liebert et al. *Hum Genet* 2017

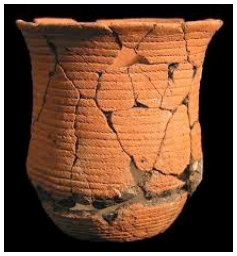
How old is lactase persistence?

Central Europe → about 5,000 years ago



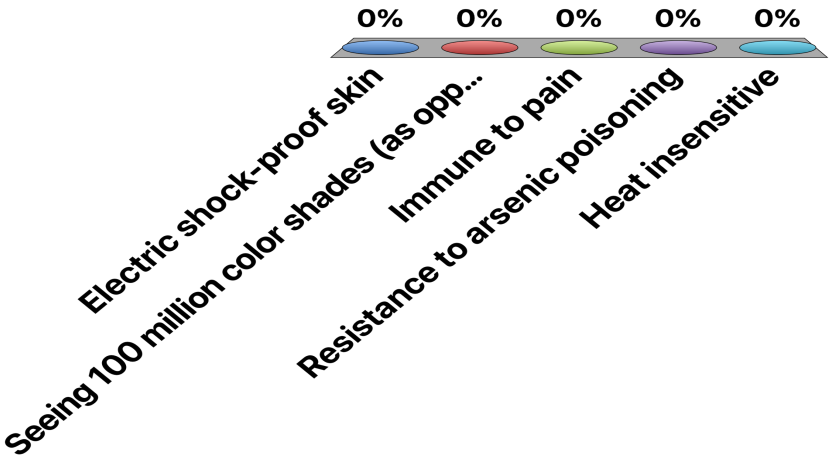
How old dairying?

Near east → 8-9,000 years ago



Some X-men mutation fun: which of the following X-men does not exist?

A. Electric shock-proof skin
B. Seeing 100 million color shades (as opposed to "only" 1 million for the average human)
C. Immune to pain
D. Resistance to arsenic poisoning
E. Heat insensitive

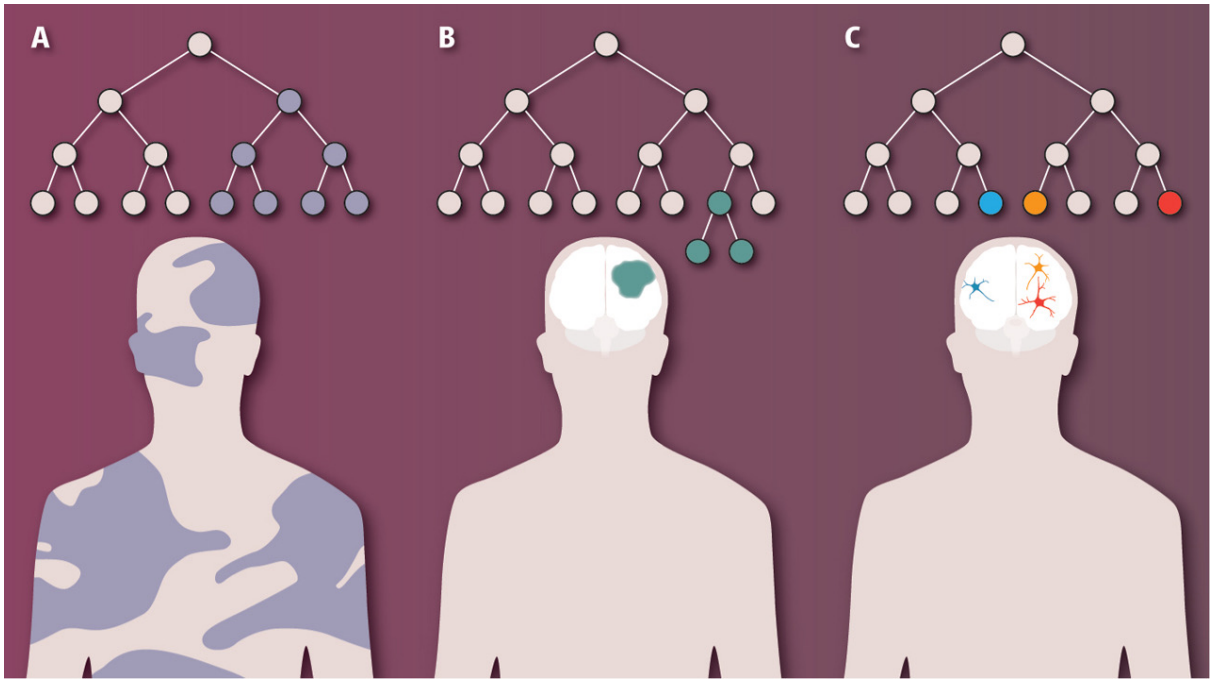


From a person's "meta-genome" to a cell-specific genome

L2Q8a

So far:

- Each genome has largely been assumed to be the same throughout one's body = "the meta-genome"
- This now appears to be a false assumption:



Deletions, duplications, and other mutations may arise at different places in a developmental lineage.

(A) Mutations early in development → large-scale somatic mosaicism in the body.

(B) Mutations that cause cells to proliferate may lead to detectable somatic mosaicism, even if they arise later in development.

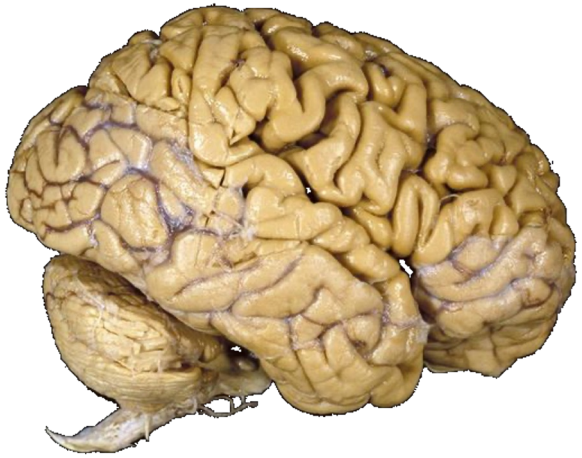
(C) Mutations that arise late in development may be unique events in individual cells.

Macosco and McCarroll, Science, 2013

From a person's "meta-genome" to a cell-specific genome

A striking example:

Post-mortem brains



Single cell genome sequencing

- Up to 41% of the individual frontal cortex neurons had at least one Mb-scale de novo copy number variation (no two alike)
- Deletions are twice as common as duplications
- A subset of neurons have highly aberrant genomes marked by multiple alterations

L2Q8a

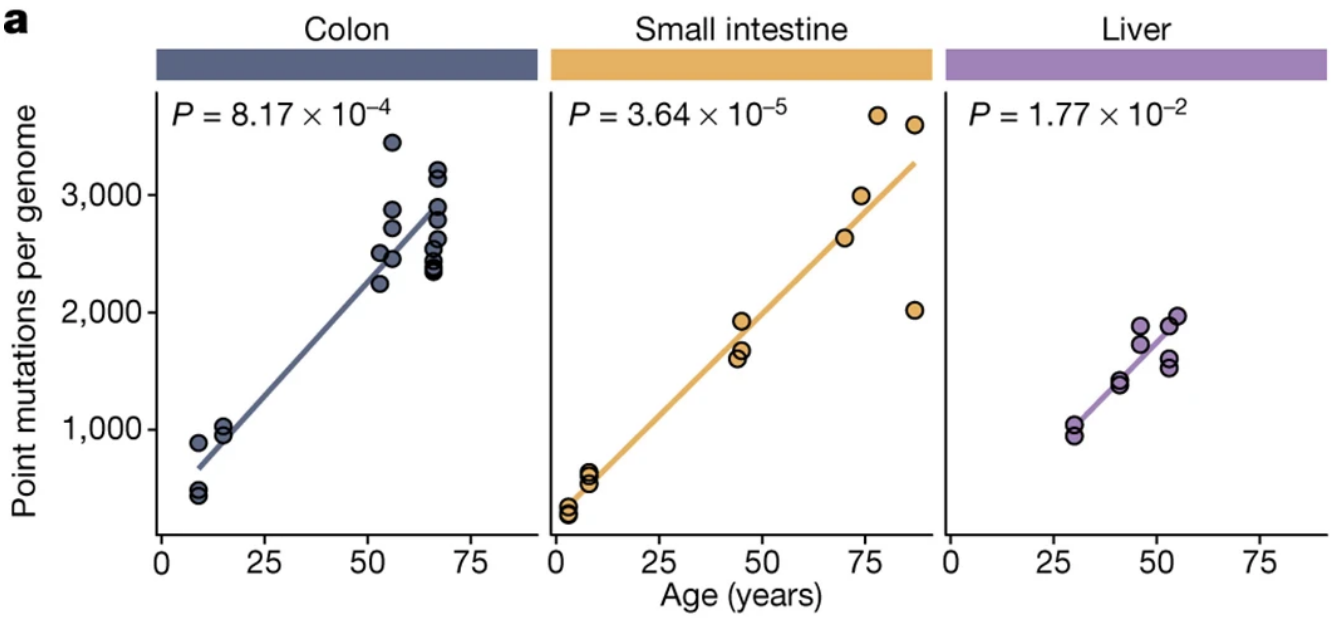
These genetic differences may have an important impact on brain cell function, and they may even shape our personalities, academic abilities, and susceptibilities to neurological diseases.

McConnell et al., Science, 2013

From a person's "meta-genome" to a cell-specific genome

Clock-like accumulation of somatic mutations during ageing

L2Q8b



2,000-3,000 somatic mutations per genome at the age of 65 years

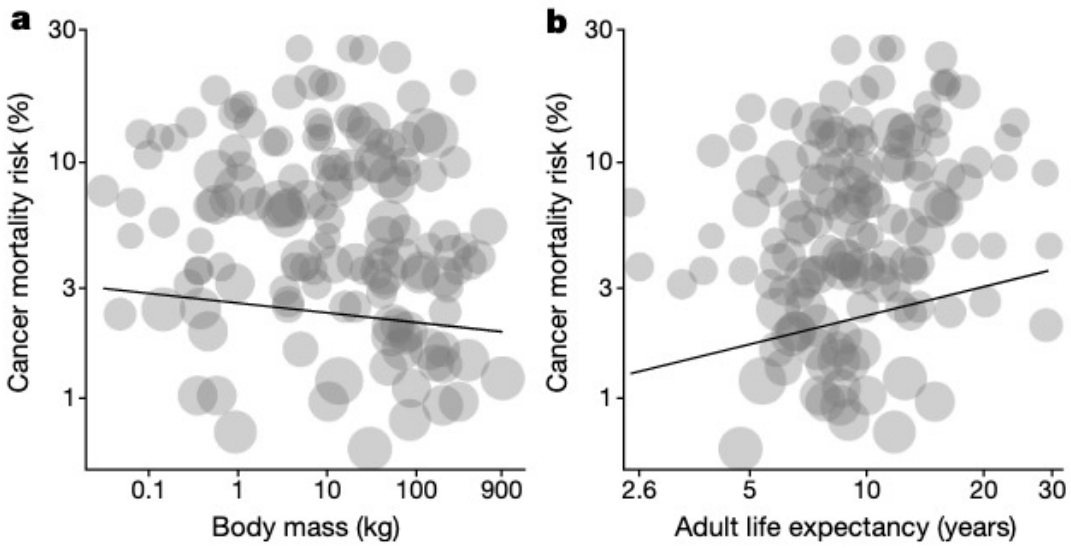
Blokzijl et al. *Nature* 2016

The probability of acquiring a "bad" mutation increases linearly with age

Contributed by Sebastian Waszak (EPFL)

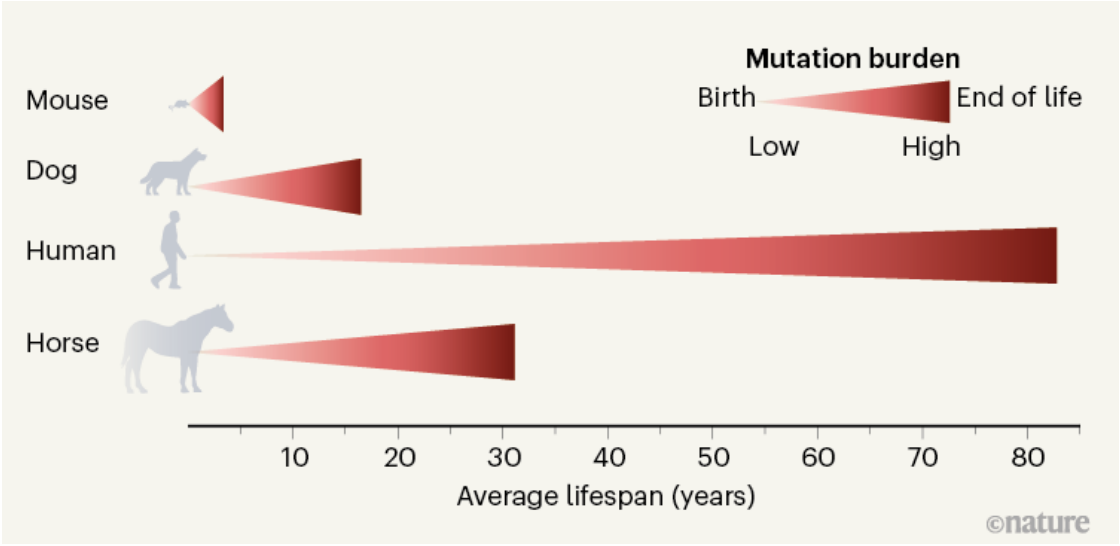
L2Q8b

But….Peto's paradox*



Vincze et al., Nature, 2022
Cancer risk across mammals
(110,148 individuals, 191 species)

No / Poor cancer association with body size and longevity
(but carnivorous mammals (raw meat) are an exception)

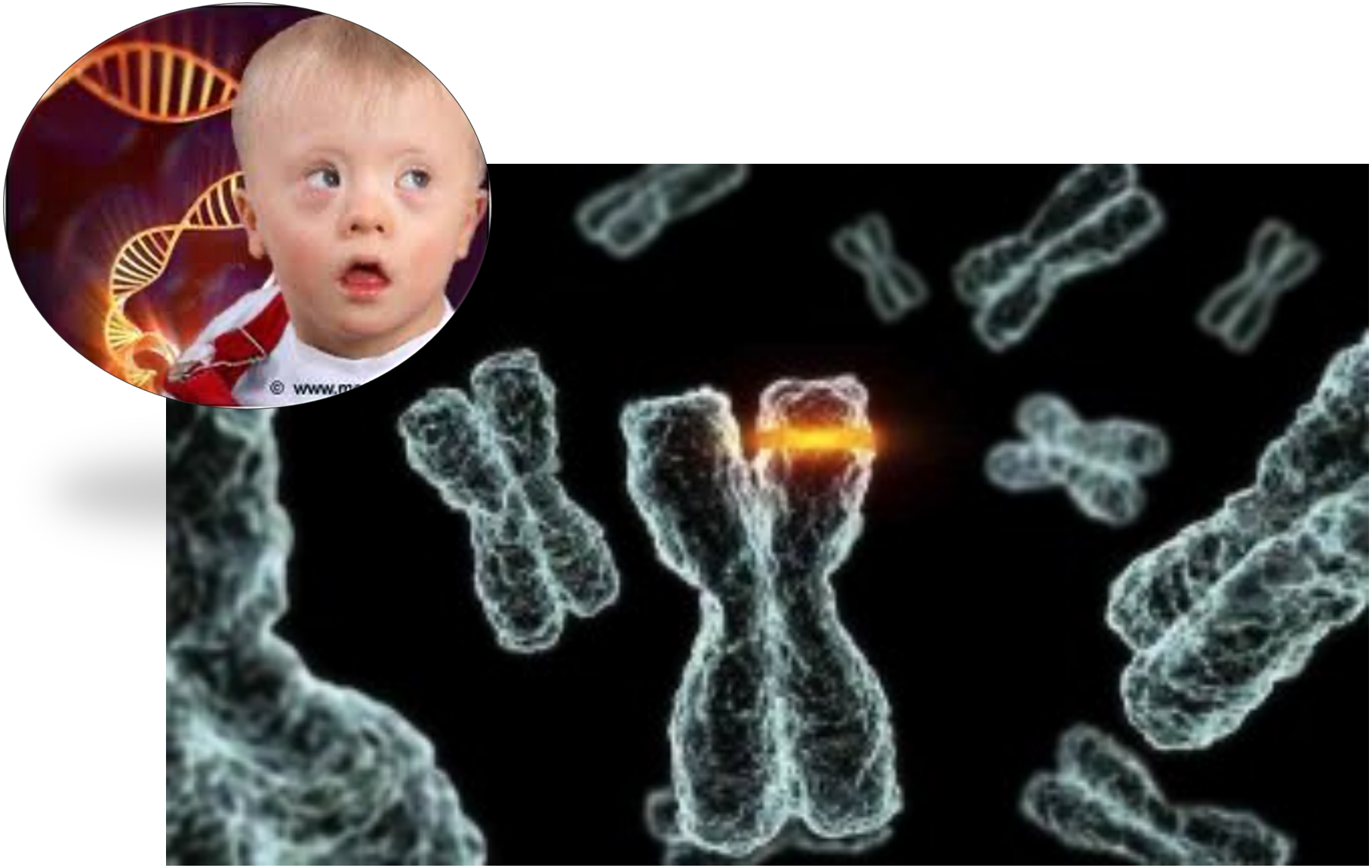


Cagan et al., Nature, 2022

Longer-lived animals have a low(er) mutation rate, which brings their risk of cancer mortality down to levels similar to those of shorter-lived species

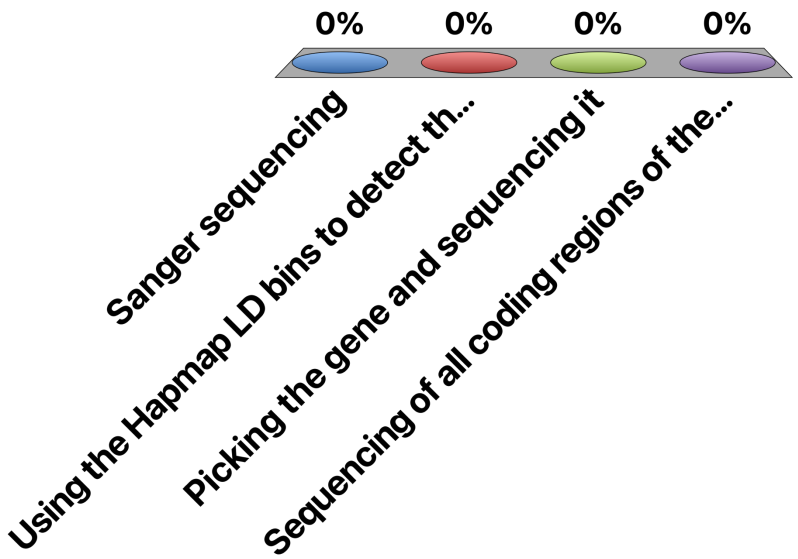
* https://wellcomecollection.org/works/p3pm7b2e

Contribution of variants to phenotypes / disease?



Rare diseases tend to be caused by rare variants: how to detect them?

A. Sanger sequencing
B. Using the Hapmap LD bins to detect the variant(s)
C. Picking the gene and sequencing it
D. Sequencing of all coding regions of the genome



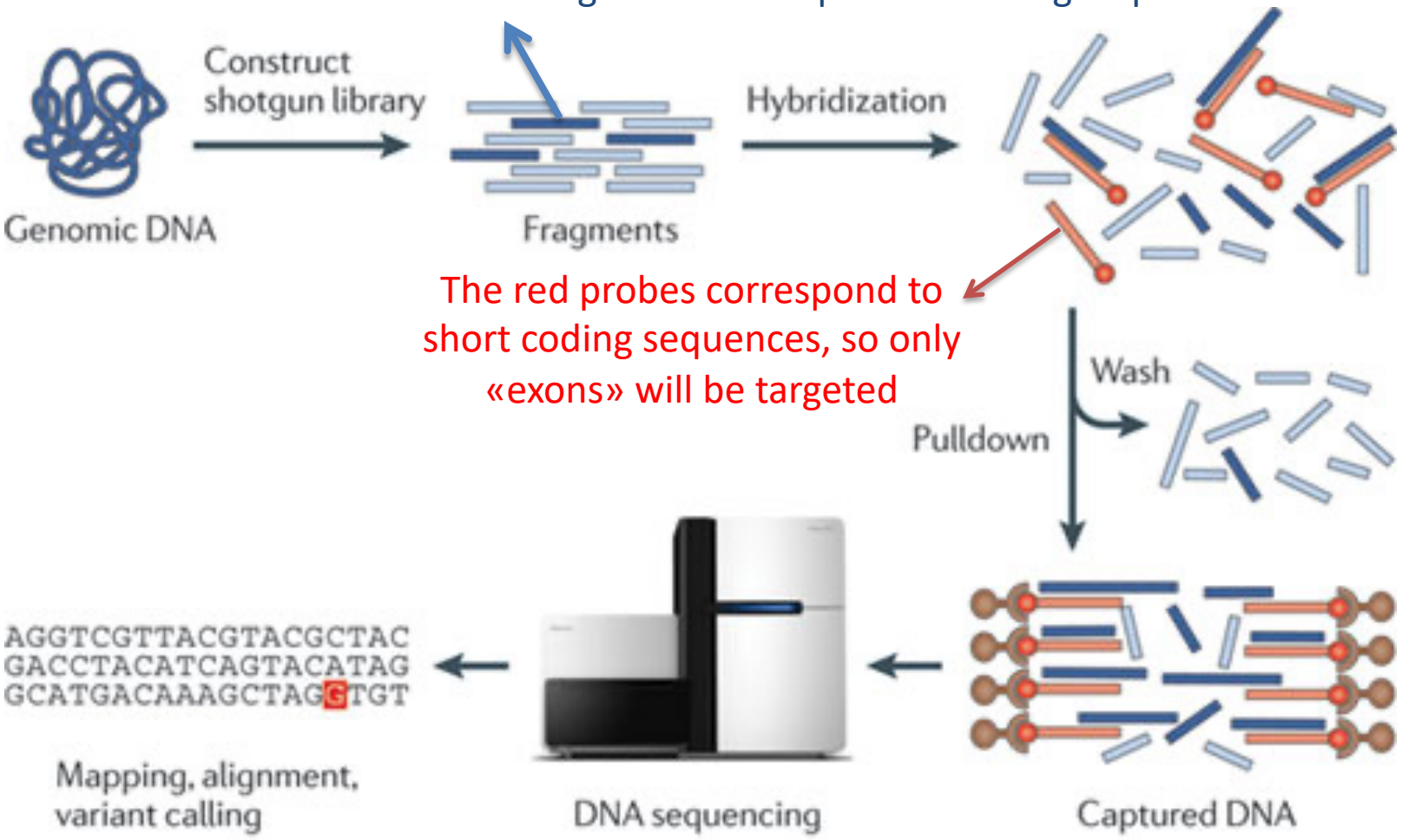
Targeting the very rare (<0.1%) variants

Exome sequencing as a tool for Mendelial disease gene discovery

Workflow

L2Q9

The dark blue fragments correspond to coding sequences



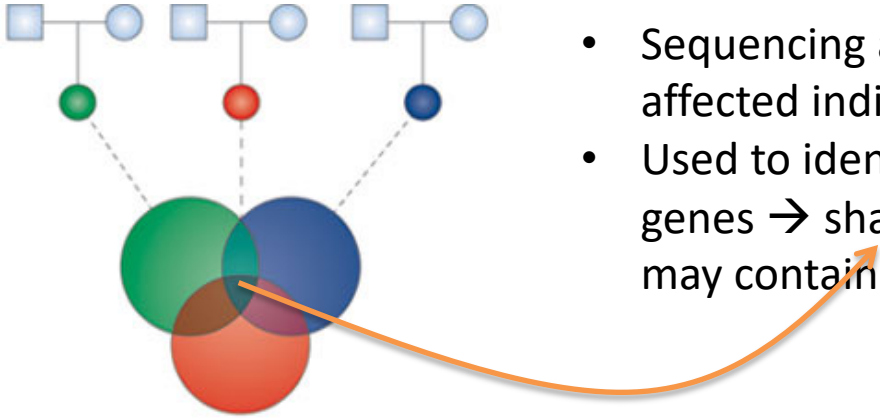
You map onto the reference human genome and you look for nucleotides that diverge from the reference sequence

Bamshad et al. Nat Rev Genet, 2011

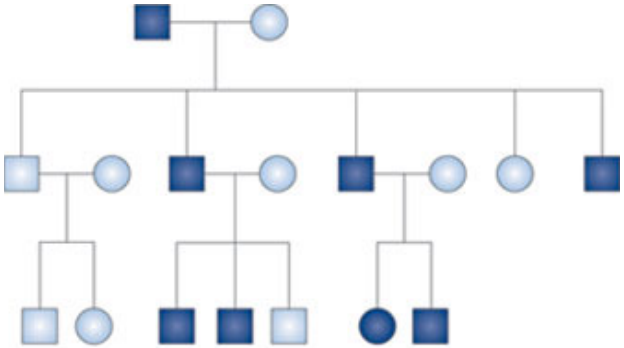
Targeting the very rare variants

Exome sequencing as a tool for Mendelian disease gene discovery

Disease gene identification strategies **L2Q10**



- Sequencing and filtering across multiple unrelated, affected individuals (3 colored circles).
- Used to identify novel variants in the same gene (or genes → shaded region shared by the 3 individuals may contain variants that cause the disease



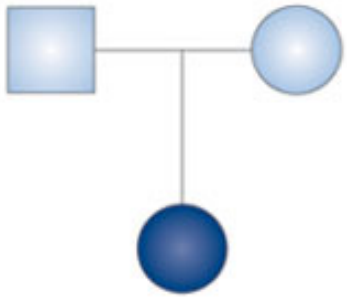
Sequencing and filtering among multiple affected individuals from within a pedigree (shaded circles and squares) to identify a gene (or genes) with a novel variant in a shared region of the genome

Bamshad et al. Nat Rev Genet, 2011

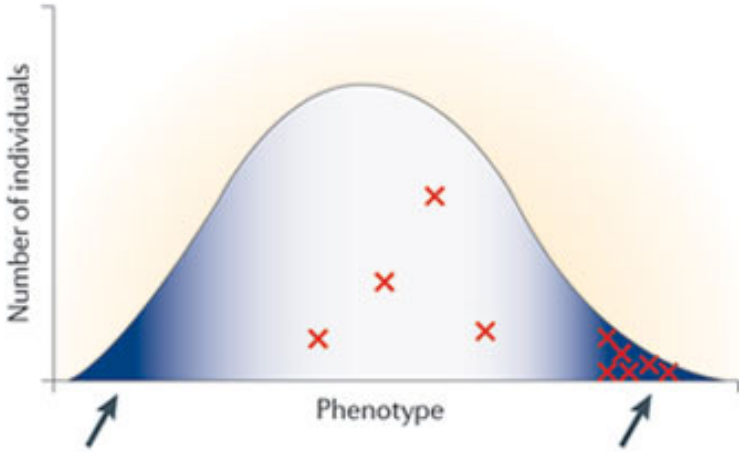
Targeting the very rare variants

Exome sequencing as a tool for Mendelial disease gene discovery

Disease gene identification strategies



Sequencing parent–child trios for identifying de novo mutations (i.e. which variants are in the child's genome that are not present in the genomes of the parents)

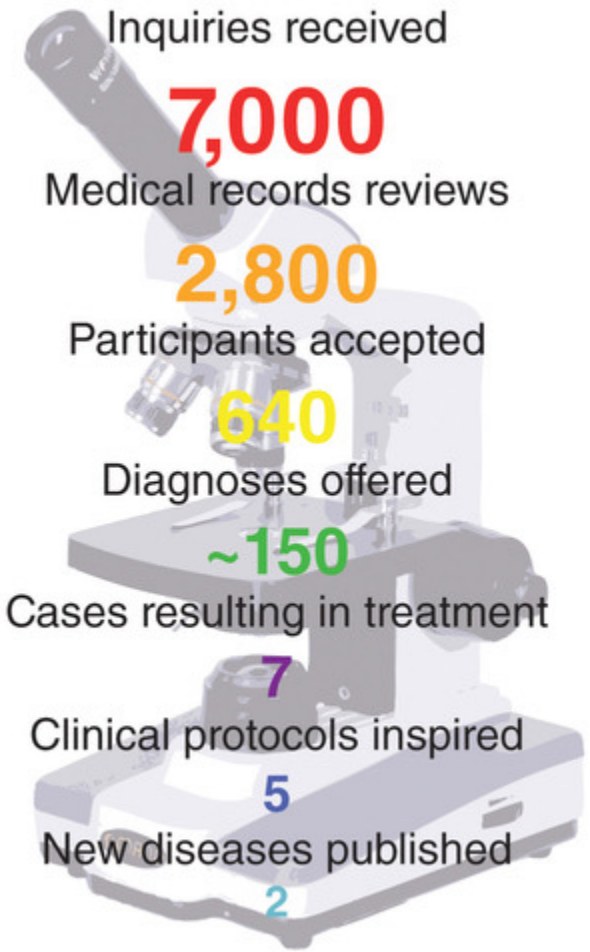


- Sampling and comparing the extremes of the distribution (arrows) for a quantitative phenotype.
- Individuals with rare variants in the same gene (red crosses) are concentrated in one extreme of the distribution.

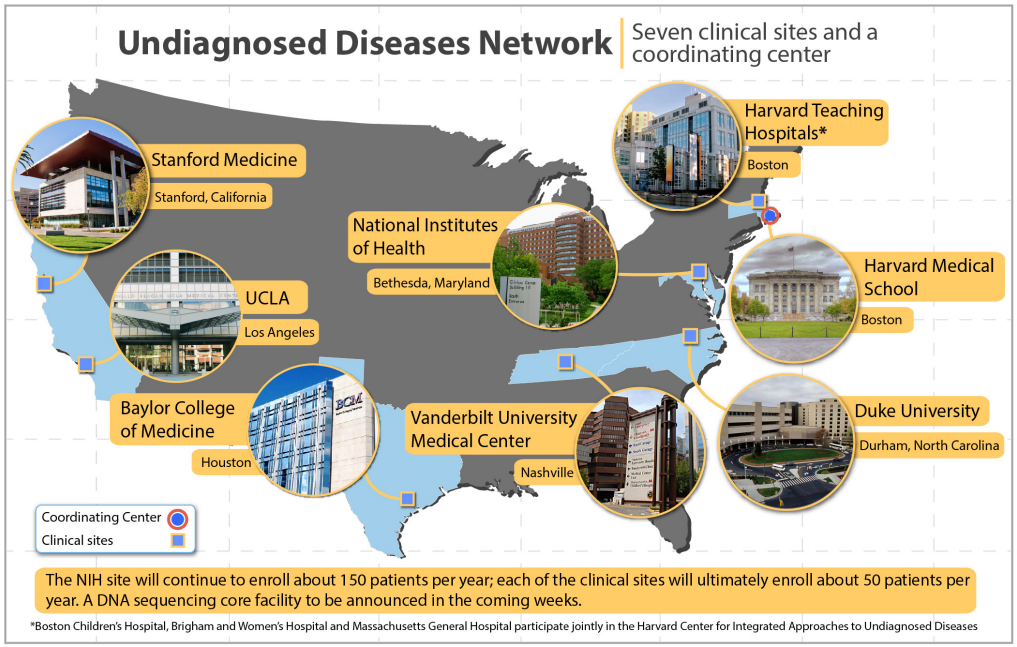
Bamshad et al. Nat Rev Genet, 2011

Targeting the very rare variants

Future of Mendelian disease gene identification is bright

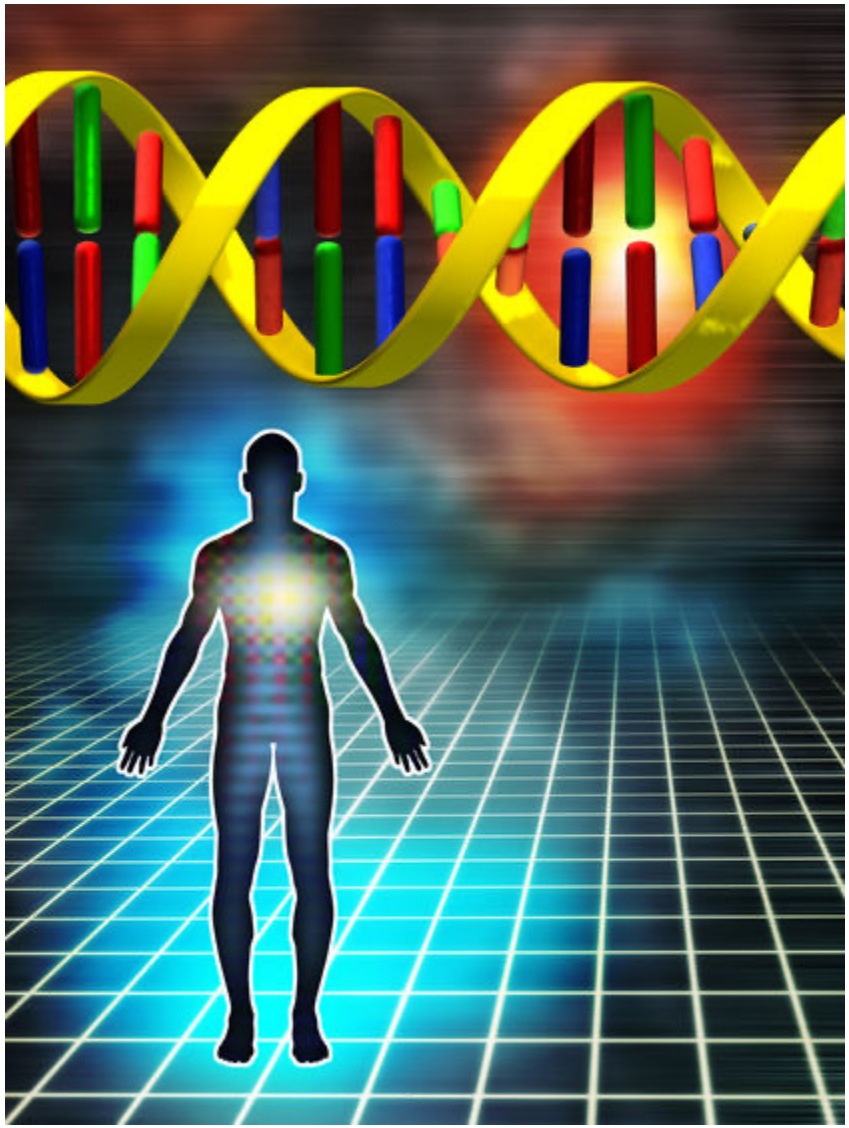


Dolgin, Nature Medicine, 2014



Hope is that genes will have been identified for the predicted 6,000-8,000 Mendelian diseases by 2030

Pre-disposition to complex disease



Where to look?

Complex disease: Common versus rare

"Common disease – common variant hypothesis"

versus

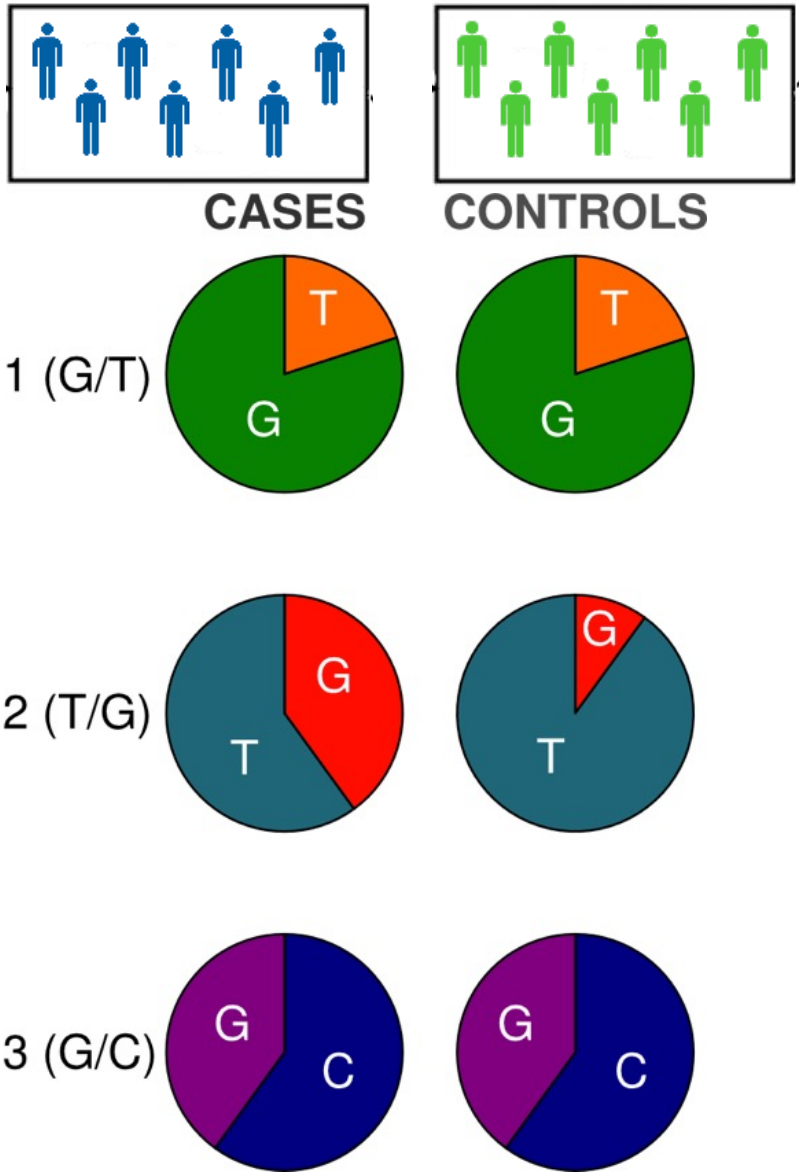
Common complex traits are the summation of low-frequency, high-penetrance variants

L2Q11

Common disease variants	Rare disease variants
• Discovery using GWAS	Discovery by targeted sequencing
• Need large populations to reach significance	Assessed by frequency increase in a population
• MAF > 5%	MAF > 0.1% to 2-3% (more frequent than rare familial mutations)
• OR b/w 1.2-1.5	OR > 2

OR = odds ratio or $\frac{p_1/(1-p_1)}{p_2/(1-p_2)} = \frac{p_1/q_1}{p_2/q_2} = \frac{p_1 q_2}{p_2 q_1},$ **L2Q12**

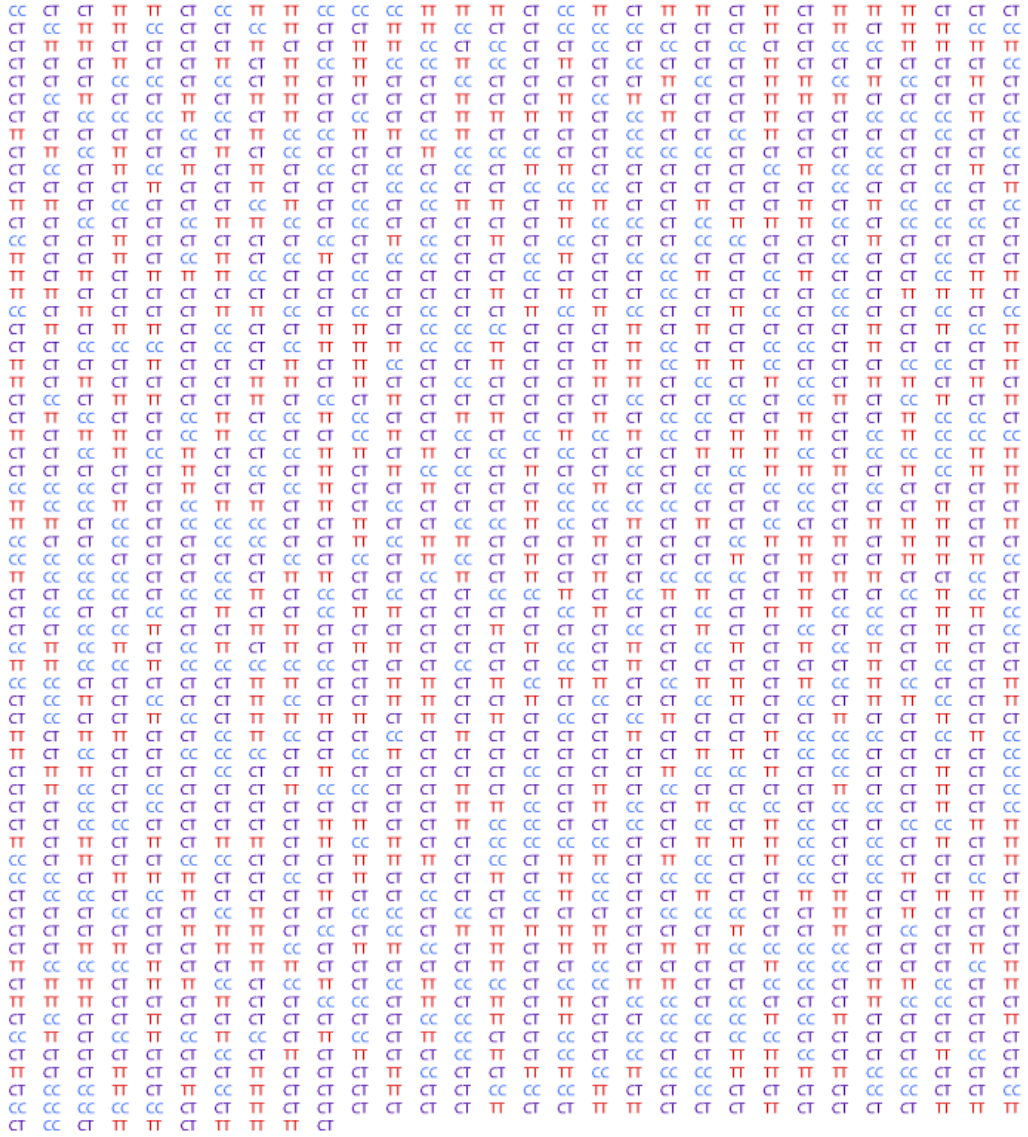
Genome-Wide Association Studies (GWAS) attempt to uncover the biology behind complex diseases using a case-control study design



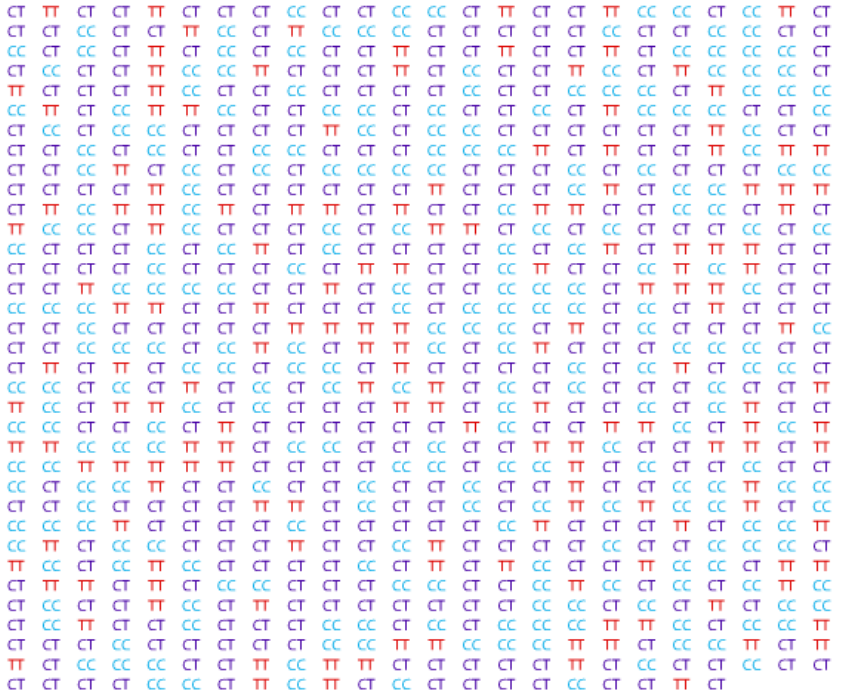
From Kathleen Bailey (U Chicago)

GWAS: a concrete example

controls



cases



Odds ratio for C allele:

1.35, $p = 6.3 \times 10^{-7}$

How significant is this?

For more info on how to calculate the P-value, see e.g. http://www.bmj.com/content/343/bmj.d2304 (beyond the scope of this class)

Whole genome association studies

P-value

the probability of seeing your data or more extreme data if the null hypothesis is *true*.

By chance, with 1,000,000 statistical tests:

- a threshold of p=0.05
  would show 50,000 "significant" associations
- a threshold of $p = 0.05/1{,}000{,}000$ (5×10^{-8})
  would show 0.05 "significant" associations

→ **Bonferroni correction**

Note: "Genome-wide" is a misnomer

- LD r^2>=0.8, thus not all variants are tagged **L2Q13**
- Rare variants not tagged at all

Common versus rare

"Common disease – common variant hypothesis"

versus

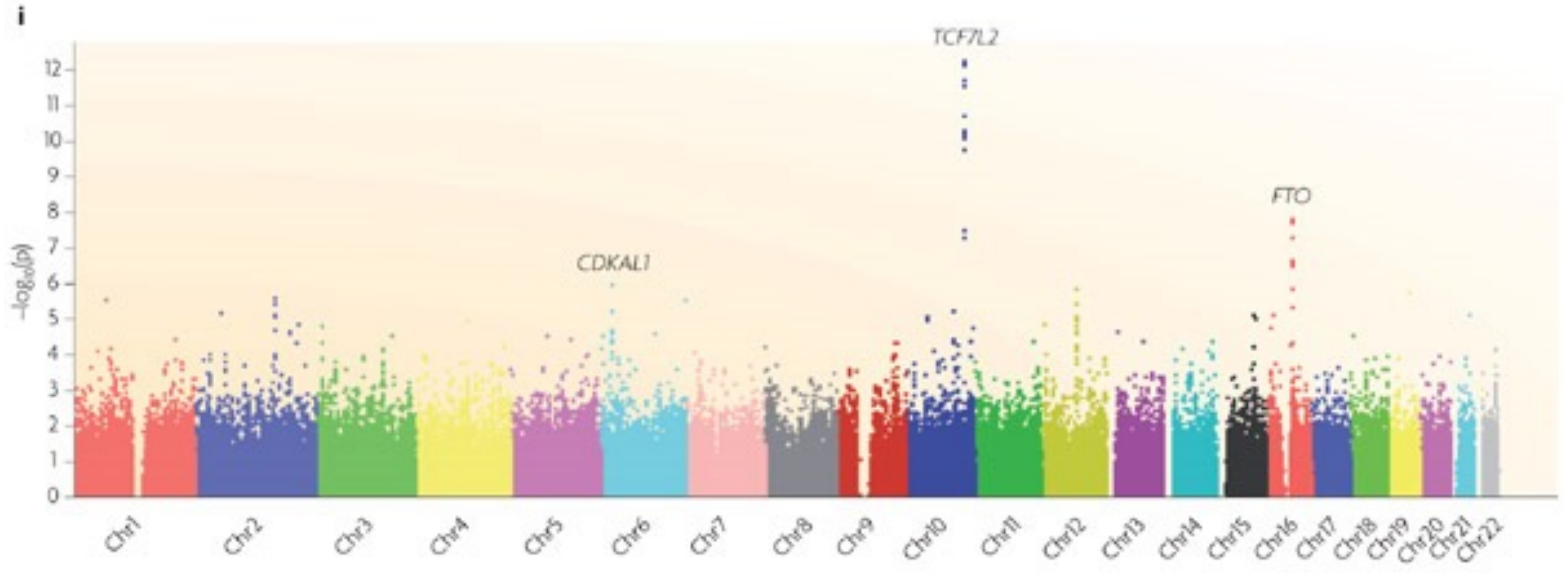
Common complex traits are the summation of low-frequency, high-penetrance variants

Common disease variants	Rare disease variants
• Discovery using GWAS	Discovery by targeted sequencing
• Need large populations to reach significance	Assessed by frequency increase in a population
• MAF > 5%	MAF > 0.1% to 2-3% (more frequent than rare familial mutations)
• OR b/w 1.2-1.5	OR > 2
• Difficult to identify causal variants	Causality more likely
• Medical intervention difficult	Intervention could be justified

Note: Variants that significantly associate with a trait / disease susceptibility → **Quantitative Trait Loci (QTLs)**

Whole Genome Association studies

Visualization



Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447, 661–678 (2007).

L2Q14

In **GWAS Manhattan plots**→ X-axis: genomic coordinates; Y-axis: the negative logarithm of the association *P*-value for each SNP displayed on the Y-axis.

Note that because the strongest associations have the smallest P-values (e.g., 10^{-15}), their negative logarithms will be the greatest (e.g., 15).

McCarthy et al., Nature Rev. Genet., 2008

Whole genome association studies

An avalanche of GWA studies

***HTRA1* Promoter Polymorphism in V Age-Related Macular Degeneration**

Andrew DeWan,[1] Mugen Liu,[2]* Stephen Hartman,[3]* Samuel Shao-Min Zhang,[2]* David T. Connie Zhao,[5] Pancy O. S. Tam,[4] Wai Man Chan,[4] Dennis S. C. Lam,[4] Michael Snyder,[1] Colin Barnstable,[2] Chi Pui Pang,[4] Josephine Hoh[1,2]†

www.sciencemag.org SCIENCE VOL 314 10 NOVEMBER 2006

A Genome-Wide Association Study Identifies *IL23R* as an Inflammatory Bowel Disease Gene

Richard H. Duerr,[1,2] Kent D. Taylor,[3,4] Steven R. Brant,[5,6] John D. Rioux,[7,8] Mark S. Silverberg,[9] Mark J. Daly,[9,10] A. Hillary Steinhart,[9] Clara Abraham,[11] Miguel Regueiro,[1] Anne Griffiths,[12] Themistocles Dassopoulos,[5] Alain Bitton,[13] Huiying Yang,[3,4] Stephan Targan,[4,14] Lisa Wu Datta,[5] Emily O. Kistner,[15] L. Philip Schumm,[15] Annette T. Lee,[16] Peter K. Gregersen,[16] M. Michael Barmada,[2] Jerome I. Rotter,[3,4] Dan L. Nicolae,[13,17] Judy H. Cho[18]*

www.sciencemag.org SCIENCE VOL 314 1 DECEMBER 2006

Corrected 4 September 2019. See full text.

RESEARCH

RESEARCH ARTICLE SUMMARY

HUMAN GENETICS

Large-scale GWAS reveals insights into the genetic architecture of same-sex sexual behavior

Andrea Ganna, Karin J. H. Verweij, Michel G. Nivard, Robert Maier, Robbee Wedow, Alexander S. Busch, Abdel Abdellaoui, Shengru Guo, J. Fah Sathirapongsasuti, 23andMe Research Team, Paul Lichtenstein, Sebastian Lundström, Niklas Långström, Adam Auton, Kathleen Mullan Harris, Gary W. Beecham, Eden R. Martin, Alan R. Sanders, John R. B. Perry, Benjamin M. Neale, Brendan P. Zietsch*

A genome-wide association study identifies novel risk loci for type 2 diabetes

Robert Sladek[1,2,4], Ghislain Rocheleau[1]*, Johan Rung[4]*, Christian Dina[5]*, Lishuang Shen[1], David Serre[1], Philippe Boutin[5], Daniel Vincent[4], Alexandre Belisle[4], Samy Hadjadj[6], Beverley Balkau[7], Barbara Heude[7], Guillaume Charpentier[8], Thomas J. Hudson[4,9], Alexandre Montpetit[4], Alexey V. Pshezhetsky[10], Marc Prentki[10,11], Barry I. Posner[2,12], David J. Balding[13], David Meyre[5], Constantin Polychronakos[1,3] & Philippe Froguel[5,14]

doi:10.1038/nature05616 nature

Published in final edited form as:
Mol Psychiatry. 2015 June ; 20(6): 786–792. doi:10.1038/mp.2014.130.

Genetic background of extreme violent behavior

J Tiihonen[1,2,3,19], **M-R Rautiainen**[3,19], **HM Ollila**[3,4], **E Repo-Tiihonen**[2], **M Virkkunen**[5,6], **A Palotie**[7,8,9,10,11], **O Pietiläinen**[3], **K Kristiansson**[3], **M Joukamaa**[12], **H Lauerma**[3,13,14], **J Saarela**[15], **S Tyni**[16], **H Vartiainen**[16], **J Paananen**[17], **D Goldman**[18], and **T Paunio**[3,5,6]

Genome-wide association study using whole-genome sequencing rapidly identifies new genes influencing agronomic traits in rice

Kenji Yano[1], Eiji Yamamoto[2], Koichiro Aya[1], Hideyuki Takeuchi[1], Pei-ching Lo[1], Li Hu[1], Masanori Yamasaki[3], Shinya Yoshida[4], Hidemi Kitano[1], Ko Hirano[1] & Makoto Matsuoka[1]

More than 2,000 GWAS have been published identifying >15,000 SNPs associated with different traits

- Digestive system disease
- Cardiovascular disease
- Metabolic disease
- Immune system disease
- Nervous system disease
- Liver enzyme measurement
- Lipid or lipoprotein measurement
- Inflammatory marker measurement
- Hematological measurement
- Body weights and measures
- Cardiovascular measurement
- Other measurement
- Response to drug
- Biological process
- Cancer
- Other disease
- Other trait

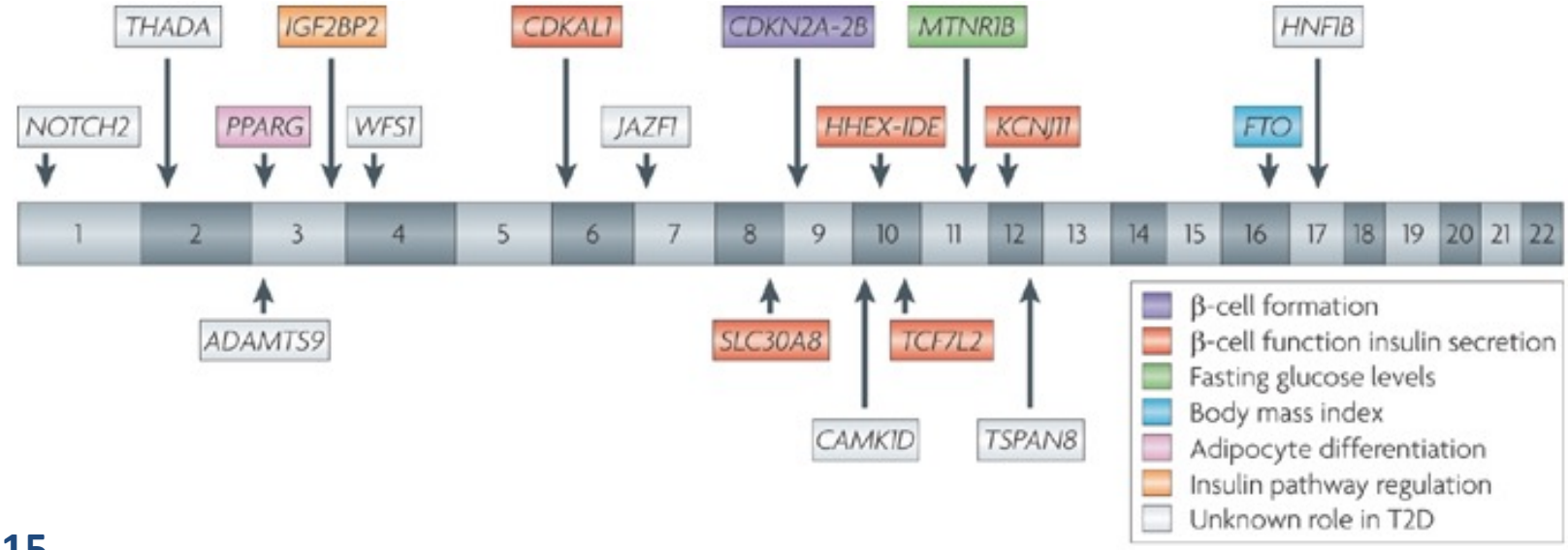
published GWAS at $p \leq 5 \times 10^{-8}$ through 2018

= GWAS locus

NHGRI-EBI GWAS Catalog, www.ebi.ac.uk/gwas

Whole genome association studies

Type 2 diabetes: an example



L2Q15

Frazer et al., Nat. Rev. Genet., 2010

- 18 genomic intervals with 4 containing previously implicated genes
- Genes with diverse functions in distinct biological processes are implicated
- Each patient may therefore have a distinct genetic reason why she/he developed type 2 diabetes → personalized medicine!!

Whole genome association studies

Longevity: an example

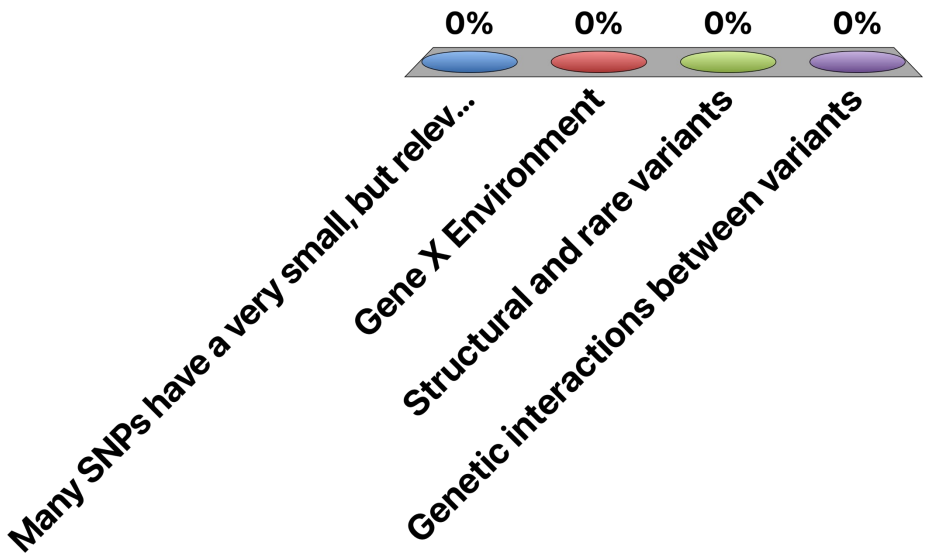
SNP	Chr	Gene	Distance (kb)	EA	EAF	*p* Value	OR	Discovery 95% CI
rs1416280	6	*GRIK2*	369	C	0.75	5.09×10^{-8}	1.24	1.15–1.34
rs9841144	3	*CADM2*	−236	A	0.79	9.66×10^{-7}	0.81	0.74–0.88
rs4611001	1	*RGS7*	−28	A	0.97	1.84×10^{-6}	1.79	1.41–2.27
rs11023737	11	*SOX6*	−28	A	0.32	3.64×10^{-6}	0.83	0.77–0.90
rs11753077	6	*MBOAT1*	−76	T	0.64	7.51×10^{-6}	1.17	1.09–1.26
rs10875746	12	*PFKM*	Intron	A	0.76	$7.8\ 3\times10^{-6}$	1.20	1.11–1.30
rs10007810	4	*LIMCH1*	Intron	A	0.23	8.80×10^{-6}	1.20	1.11–1.30
rs10457180	6	*FOXO3*	Intron	A	0.70	8.56×10^{-5}	0.87	0.81–0.93

L2Q16a

Around 20 loci by now associated with longevity but common variability accounts “only” for 25 % of human lifespan variability, What about the other 75%?

Where is the ”missing heritability”?

A. Many SNPs have a very small, but relevant effect
B. Gene X Environment
C. Structural and rare variants
D. Genetic interactions between variants

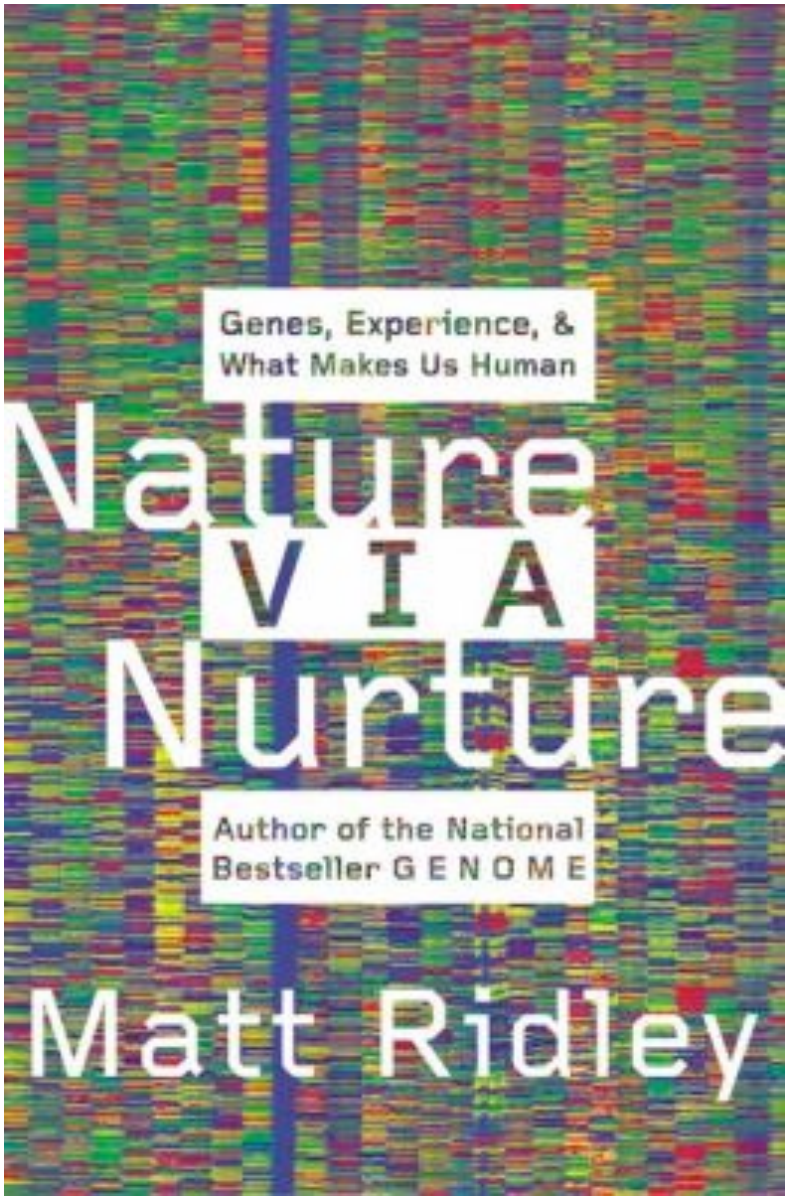


L2Q16b

“It takes two to tango”

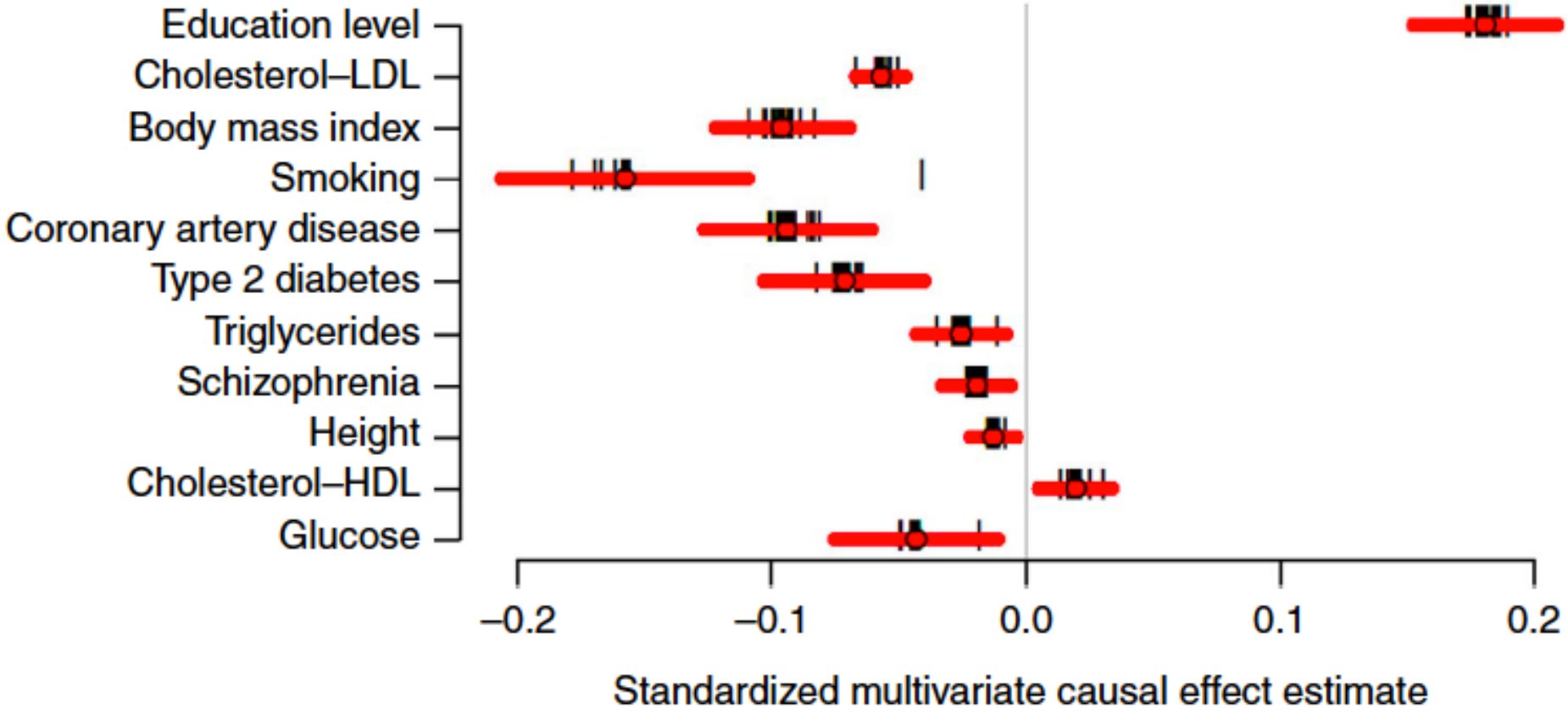
Phenotype = Genotype + Environment + GxE

(we are the product of our genes, of our environment, and of the interaction between these two)



Effect of environmental factors on longevity

L2Q16



Each year of education translates to +11 months lifespan lifespan

Courtesy of Prof. Zoltan Kutalik; McDaid et al., Nat. Comm,, 2017

The pace of the course is OK

A. True
B. False

1. Super Flexibility: Those who are affected by Marfan syndrome have a mutation in the protein Fibrillin-1, which affects the body's connective tissue, giving people with this mutation the ability to bend in all different directions. Some common signs of the disorder include long arms, legs, fingers, a curved spine, flat feet, and a tall, thin body. People with the condition may also experience serious health complications affecting the heart, eyes, blood vessels, and bones.

2. Super Speed: Mutations in the gene ACTN3 are associated with the ability to run fast. ACTN3 produces the protein alpha-actinin-3, which is responsible for the fast-twitch muscle fibers that allow us to run. Research has indicated that people with higher than average levels of this protein may be able to out-run the majority of us.

3. Resistance To Poisoning: An entire community of people in a small town in Argentina have inherited a gentic mutation that makes them resistant to arsenic poisoning. The village of 6,000 people are known to survive after consuming more than 80 times the amount of arsenic an average person would die from. Scientists believe this phenomenon is because the village's water supply has been laced with naturally forming arensic for thousands of years. People in this community are believed to have the gene AS3MT, which helps them flush out toxins much faster than the average person.

4. Resistance To Fat: The inability to gain weight may sound like a dream come true to many people, but the reality of your body being resistant to fat is far from glamourous. MDP syndrome, which affects 8 people in the world, prevents fat from being stored under the skin. Instead, it gets deposited in the bloodstream, leading to diabetes and other health complications.

5. Super Vision: A condition called tetrachromacy allows people to see almost 100 million different colors. In comparison, the average human can only differentiate between about 1 million. This is caused by a mutation in the opsin gene, which is responsible for producing visual pigments for color vision.

6. Shock-Proof Skin: Slavisa Pajkic, or "The Battery Man," has a gene mutation that results in no sweat or salivary glands. Since these glands are absent, his body is able to resist electricity. According to the video, Pajkic has been known to withstand voltages as high as 20,000 volts. To put that into perspective, most people would be severely injured and burned by 50 volts.

7. Immunity To Pain: Congenital insensitivity to pain (CIP) is a rare condition in which a person can't feel pain, even if they break a bone, get burned, or experience any other type of injury. CIP is caused by a mutation in the gene SCN11A. This mutation decreases the amount of sodium in the body's cells, which is key to alerting your brain of pain. CIP affects fewer than 100 people worldwide.

8. Super Strength: People who have a genetic mutation in the gene MSTN will pack on muscle quickly. MSTN's job is to produce myostatin, a protein that tells the body to stop creating muscle when there's already enough. Those who have this mutation typically have at least double the amount of muscle than the average human.