Whiteboard

1

CS273A

The Human Genome Source Code

Gill Lecture 10: Pop Gen II, Genetic Disease

Mon, Wed 11:30 AM - 12:50, on Zoom*

Prof: Gill Bejerano

CA: Boyoung (Bo) Yoo

* Track class on Piazza

2

Announcements

3

Class Topics

good morning!

(0) Genome context:
cells, DNA, central dogma

(1) Genome content / genome function:
genes, gene regulation, epigenetics, repeats, SARS-CoV-2

(2) Genome sequencing:
technologies, assembly/analysis, technology dependence

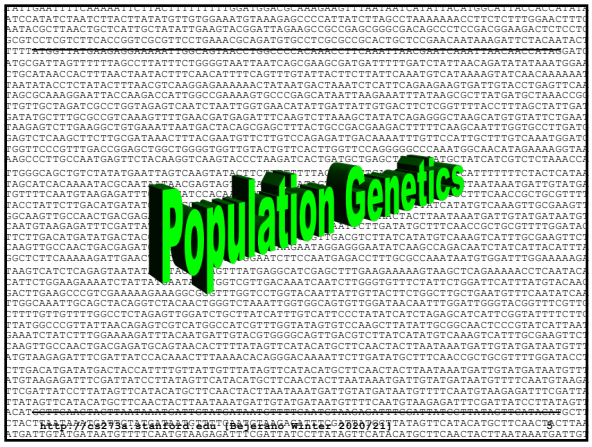
(3) Genome evolution:
evolution = mutation + selection, main forces of evolution:
Neutral evolution, Negative selection, Positive selection

(4) Population genomics:
Human migration, paternity testing, forensics, cryptogenomics

(5) Genomics of human disease:
personal genomics, GxE disease types, deep dive monogenics

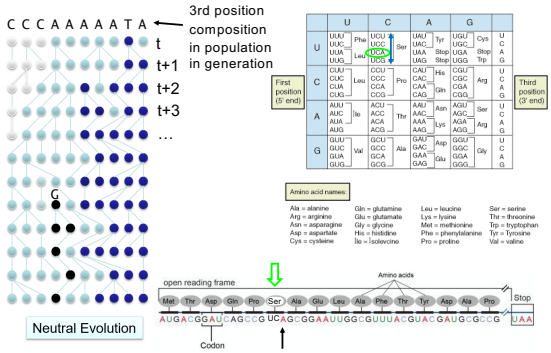
(6) Comparative Genomics :
Genomics of amazing animal adaptations, ultraconservation

4



5

If the 3rd position is under neutral evolution



6

Reminder

We used the facts that

- Every genome replication introduces a mutation
- Most of genome, and by extension most of these mutations are “silent”, evolving neutrally.

To identify/distinguish:

- Individuals, even monozygotic twins!
- Paternity, 1st – 3rd degree relations
- Ethnicity
- Birthplace to the resolution of villages (map of Europe)
- Ethnicities of person’s ancestors
- Human migration patterns (out of Africa)

7

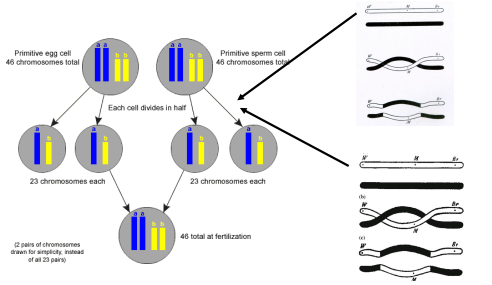
Ancestry Painting

- Danish
- French
- Spanish
- Mexican

Jessica Marie Alba was born in Pomona, California,[2] on April 28, 1981, to Catherine Louisa (*née* Jensen) and Mark David Alba. Her mother has Danish, Welsh, German, English and French ancestry, while her paternal grandparents, who were born in California, were both the children of Mexican immigrants.[11]

8

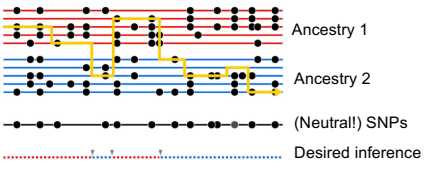
The “paint” comes from Meiotic Crossover

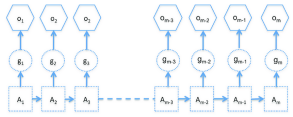


In humans:
1 crossover in ~100Mb

9

Ancestry Painting Computation

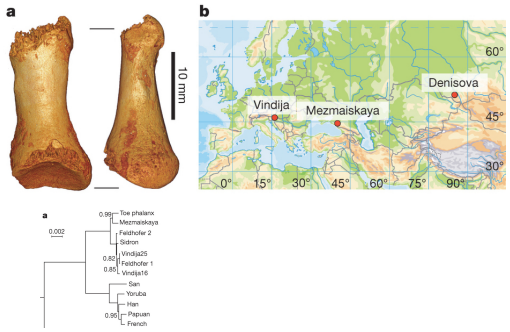




Graphic Illustration of Our Hidden Markov Model for Ancestry Decomposition There are three layers of information: the square boxes (A) represent the hidden ancestry at each site, the circles (g) represent hidden genotypes, and the hexagons (o) represent observed sequence reads.

10

Ancient DNA: The Neanderthal



11

The Neanderthal Genome

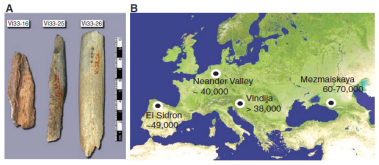


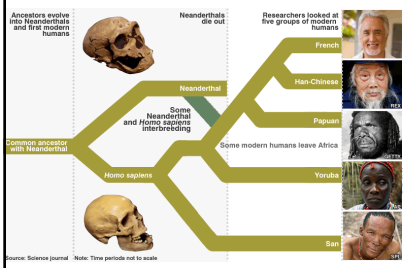
Fig. 1. Samples and sites from which DNA was retrieved. (**A**) The three bones from Vindija from which Neandertal DNA was sequenced. (**B**) Map showing the four archaeological sites from which bones were used and their approximate dates (years B.P.).

- From bones, compared genomes of three different Neanderthals with five genomes from modern humans from different areas of the world

- **Figure 1- R. E. Green et al., Science 328, 710-722 (2010)**

12

Neanderthal Genome *segments*



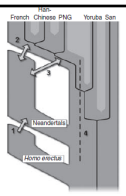


Fig. 6. Four possible scenarios of genetic mixture involving Neandertals. Scenario 1 represents gene flow into Neandertal from other archaic hominins, here collectively referred to as *Homo erectus*. This would manifest itself as segments of the Neandertal genome with unexpectedly high divergence from present-day humans. Scenario 2 represents gene flow between late Neandertals and early modern humans in Europe and/or western Asia. We see no evidence of this because Neandertals are equally distantly related to all non-Africans. However, such gene flow may have taken place without leaving traces in the present-day gene pool. Scenario 3 represents gene flow between Neandertals and the ancestors of all non-Africans. This is the most parsimonious explanation of our observation. Although we detect gene flow only from Neandertals into modern humans, gene flow in the reverse direction may also have occurred. Scenario 4 represents old substructure in Africa that persisted from the origin of Neandertals until the ancestors of non-Africans left Africa. This scenario is also compatible with the current data.

13

Denisovan – Another human relative

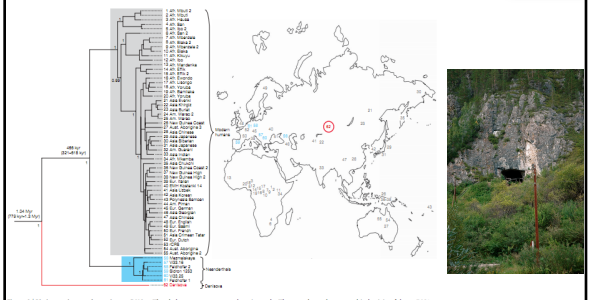


Figure 3 | Phylogenetic tree of complete mtDNAs. The phylogeny was estimated with a Bayesian approach under a GTR+I+Γ model using 54 present-day and one Pleistocene modern human mtDNA (grey), 6 Neanderthals (blue) and the Denisova hominin (red). The tree is rooted with a chimpanzee and a bonobo mtDNA. Posterior probabilities are given for each major node. The map shows the geographical origin of the mtDNAs (24, 25, 32, 44 are in the Americas). Note that two partial mtDNAs sequenced from Teshik Tash and Okladikov Cave in Central Asia fall together with the complete Neanderthal mtDNAs in phylogenies" (not shown).

Vol 464 | 8 April 2010 doi:10.1038/nature08976

14

Coalescent Theory

- The coalescent is a probability model for the tree underlying a sample of homologous DNA sequences drawn from a within-species population.
- The focus of interest can be the underlying genealogical tree, mutation or recombination rates, or demographic parameters such as historic population sizes or migration rates.

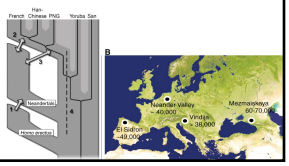
15

Human population migrations

- Out of Africa, Replacement
  - Single mother of all humans (Eve) ~190,000yr
  - Single father of all humans (Adam) ~340,000yr
  - Humans out of Africa ~50000 years ago replaced others (e.g., Neandertals)

Recent most likely migration & mating pattern estimates. They continue to be revised.

- Multiregional Evolution
  - Generally debunked, however,
  - ~5% of human genome in Europeans, Asians is Neanderthal, Denisova



16

Discoveries Continue

HUMAN GENETICS

Insights into human genetic variation and population history from 929 diverse genomes

RATIONALE: We sequenced 929 genomes from 54 geographically, linguistically, and culturally diverse human populations to an average of 35× coverage and analyzed the variation among them. We also physically resolved the haplotype phase of 26 of these genomes using linked-read sequencing.

Populations in central and southern Africa, the Americas, and Oceania each harbor tens to hundreds of thousands of private, common genetic variants. Most of these variants arose as new mutations rather than through archaic introgression, except in Oceanian populations, where many private variants derive from Denisovan admixture. Although some reach high frequencies, no variants are fixed between major geographical regions.

The low diversity among the Neanderthal haplotypes segregating in present-day populations indicates that, while more than one Neanderthal individual must have contributed genetic material to modern humans, there was likely only one major episode of admixture. By contrast, Denisovan haplotype diversity reflects a more complex history involving more than one episode of admixture.

ON OUR WEBSITE

Read the full article at http://dx.doi.org/10.1126/science.aay5012

17

Class Topics

Negative selection

(0) Genome context:
cells, DNA, central dogma

(1) Genome content / genome function:
genes, gene regulation, epigenetics, repeats, SARS-CoV-2

(2) Genome sequencing:
technologies, assembly/analysis, technology dependence

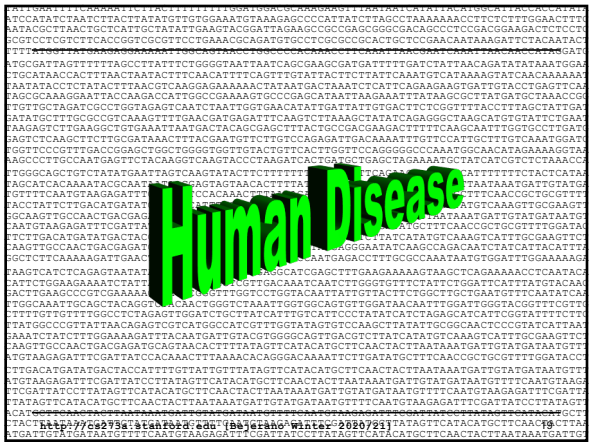
(3) Genome evolution:
evolution = mutation + selection, main forces of evolution:
Neutral evolution, Negative selection, Positive selection

(4) Population genomics:
Human migration, paternity testing, forensics, cryptogenomics

(5) Genomics of human disease:
personal genomics, GxE disease types, deep dive monogenics

(6) Comparative Genomics :
Genomics of amazing animal adaptations, ultraconservation

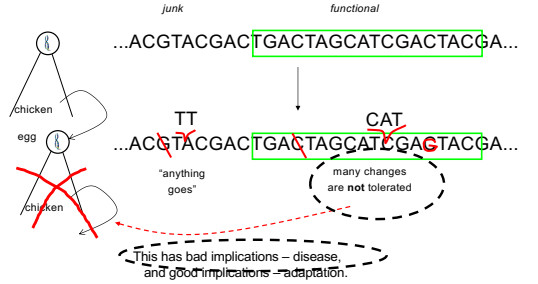
18



19

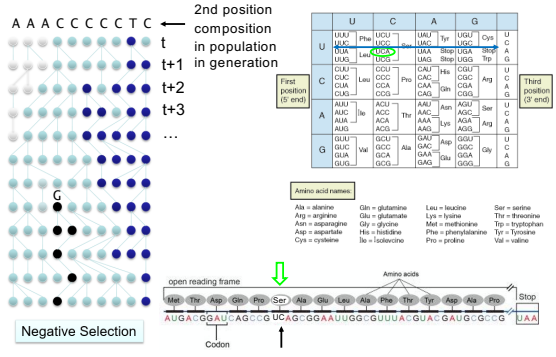
Evolution = Mutation + Selection

Mistakes can happen during DNA replication. Mistakes are oblivious to DNA segment function. But then selection kicks in.



20

If the 2nd position is under negative selection



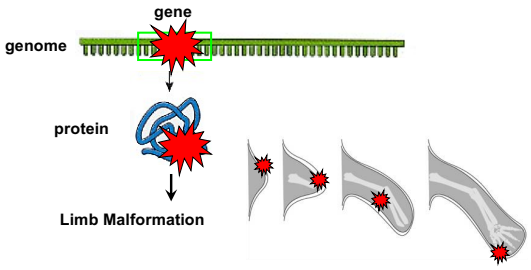
21

Genetic Disease

- Definition: “Any of the diseases and disorders that are caused by mutations in one or more genes”
- Genetic disease
  - Genes + Environment
    - Genes can be any functional genomic unit(s)
      - Viable chromosomal abnormalities (Down, XXY etc)
      - Genes and/or gene regulatory elements

22

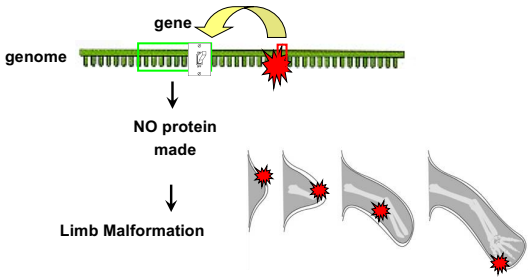
Developmental Defects



Hundreds of genes already implicated in limb malformations.

23

Genetic Causes



More and more cases are being discovered.

24

Modularity: Disease Implications

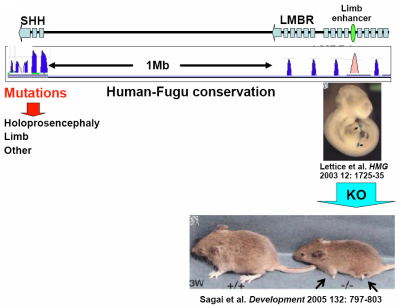
Sall1

GENE MUTATIONS

SYNDROMIC PATIENTS: multiple developmental abnormalities

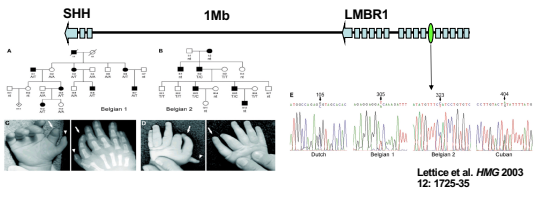
25

Critical regulatory sequences



26

Cis-regulatory contribution to Human Disease



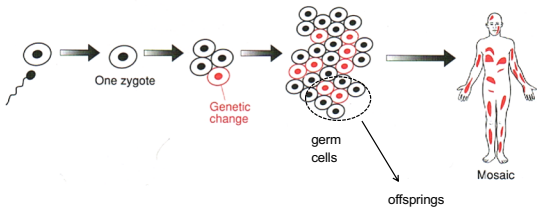
27

Genetic Disease

- Genetic disease
  - Genes + Environment
    - "Genes" can be any functional genomic unit(s) (Coding, splicing, RNA genes, promoters and enhancers)
    - Affected by any type of mutations (SNP to SV)
      - SVs will often affect multiple functional units
    - Effect can be loss OR gain of "gene" function
    - Effect can be dominant, recessive or sex linked
    - It can happen at any point in the individual's life time

28

Mosaicism



Mosaicism can be:

- **somatic** (ie in most body cells) or
- **gonadal** (confined solely to the gonads).

29

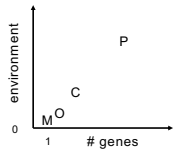
Genetic Disease

- Genetic disease
  - Genes + Environment
    - Environment is "everything else" (diet, life style, uv etc)
- Types of genetic diseases
  - Cancer
  - Polygenic
  - Monogenic, digenic, oligogenic

30

Disease "G x E"

Monogenic
Oligogenic
Cancer
Polygenic



Disease frequency in the population vs.
Rate of advance of our understanding to date

31

Variant composition

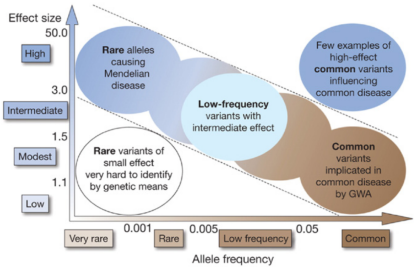
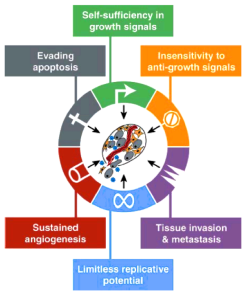


Fig. 1. Feasibility of identifying genetic risk variants by risk allele frequency and strengths of the genetic effects (odds ratio). From Manolio T, et al., Nature 2009;461:747–753, with permission.

32

Fundamental Changes in Cancer Cell Physiology



Exploitation of natural pathways for cellular growth

- Growth Signals (e.g. TGF family)
- Angiogenesis
- Tissue Invasion & Metastasis

Evasion of anti-cancer control mechanisms

- Apoptosis (e.g. p53)
- Antigrowth signals (e.g. pRb)
- Cell Senescence

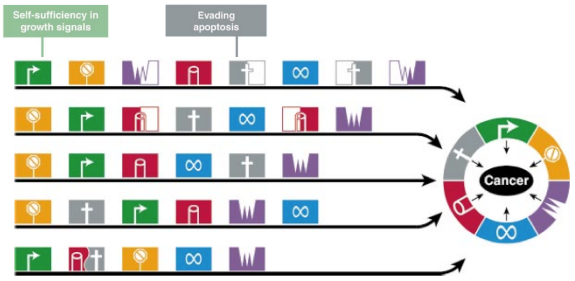
Acceleration of Cellular Evolution Via Genome Instability

- DNA Repair
- DNA Polymerase

Hanahan and Weinberg. 2000. The hallmarks of cancer. Cell 100: 57-70.

33

Many Paths Lead to Cancer Self-Sufficiency



~90% of cancer cases are thought to be sporadic.

Hanahan, Douglas, and Ra Weinberg. 2000. The hallmarks of cancer. Cell 100: 57-70.

34

Cancer Mutations

The relation between sporadic and inherited forms of the same tumour.

The target tissue contains *n* cells and the chance of one cell suffering a loss of function mutation in the tumour suppressor gene is μ.

35

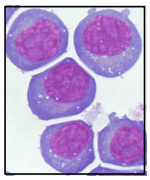
Inherited vs. Somatic Mutations

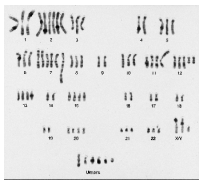
- Retinoblastoma is an aggressive, often childhood, eye tumor.
- 60-70% of all cases are sporadic, 30-40% inherited.
- The relevant gene, RB, encodes a protein that normally functions as a suppressor of cell cycle progression and is considered a classic tumor suppressor gene.
- Children who inherit one mutant copy of the RB gene are at nearly 100% risk to develop retinoblastoma, because the probability that their one remaining functional RB gene will sustain a mutation in at least one retinal cell is nearly assured.
- In contrast, children who inherit two functional copies of the RB gene must experience two mutations at the RB locus in the same retinal cell in order to develop retinoblastoma; this is a very rare event.

36

Cancer: A Disease of the Genome

Cancer genomes have a really messed up genome.



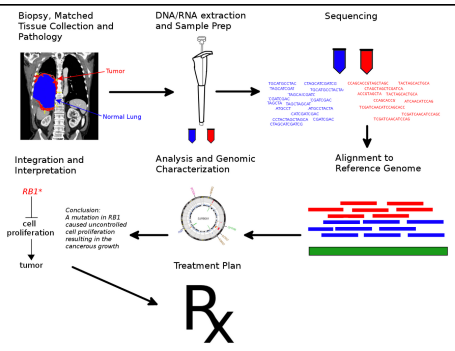


Challenge in Treating Cancer:

- Every tumor is different
- Every cancer patient is different
- By the time we sequence it, every cancer genome is (often dramatically) different

37

Cancer Genomics Goal



38

Cell Free DNA / Liquid Biopsy

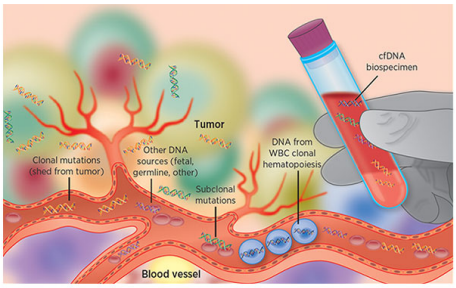
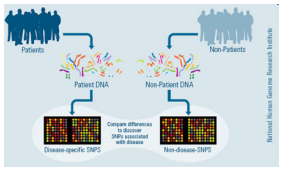
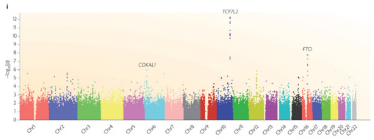


Figure originally published in Clin Cancer Res; Published OnlineFirst May 10, 2018.

39

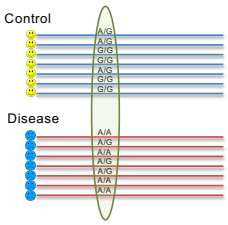
Polygenic Diseases





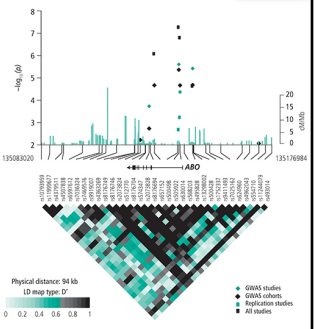
40

Genome Wide Association Studies (GWAS)



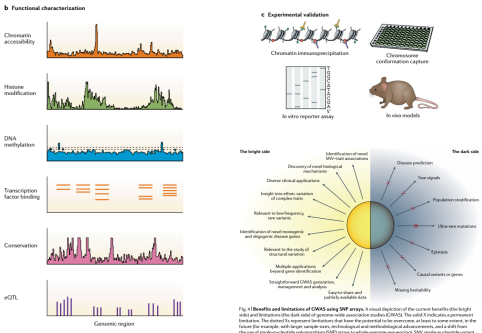
AA	0	4
AG	3	3
GG	4	0

p-value



41

Hunting down the functional variant/s



42

Neanderthal heritage

A major risk locus for COVID coincides with genomic variants some humans have inherited from Neanderthals.

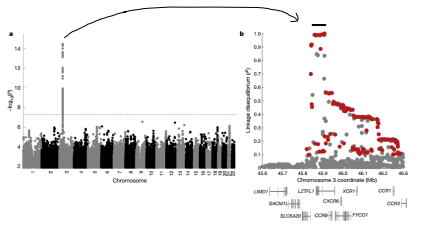
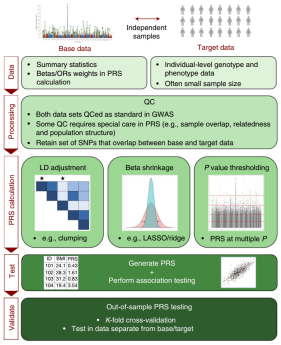


Fig. 1 | Genetic variants associated with severe COVID-19. a, Manhattan plot of a genome-wide association study of 1,299 hospitalized patients with COVID-19 and 897,488 population controls. The dashed line indicates genome-wide significance ($P = 5 \times 10^{-8}$). Data were modified from the COVID-19 Host Genetics Initiative[9] (https://www.covid19hg.org/). **b**, Linkage disequilibrium between the index risk variant (rs35044562) and genetic variants in the 1000 Genomes Project. Red circles indicate genetic variants for which the alleles are correlated to the risk variant ($r^2 > 0.1$) and the risk alleles match the Vindija 33.19 Neanderthal genome. The core Neanderthal haplotype ($r^2 > 0.98$) is indicated by a black bar. Some individuals carry longer Neanderthal-like haplotypes. The location of the genes in the region are indicated below using standard gene symbols. The *x* axis shows hg19 coordinates.

43

Polygenic Risk Scores (PRS)



44

Knowledge = technology + pre-conceptions

Our knowledge is strongly influenced by what our technology allows us to see. What it does not, we will with pre-conceptions. And those may often be wrong.

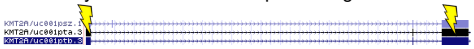
Even after half the human genome was sequenced, people thought there will be over 100,000 coding genes in it.
There are in fact ~20,000 (in par with the gene repertoire of much simpler animals).

Before the next generation sequencing revolution people thought the majority (99%) of human disease mutations will be in the coding regions.
80-90% of human disease GWAS associated SNPs are in fact non-coding. The majority is likely gene regulatory.

45

Monogenic Diseases

- Caused by 1 or 2 mutations in patient's genome



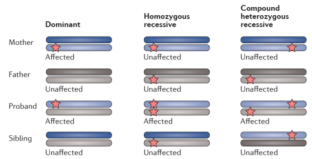
- Complicated phenotypes

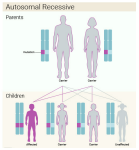
- ~3% of newborns (250 million) affected

46

Monogenic Diseases

- Definition
  - Loss of function
  - Gain of function
- Transmission modes
  - Dominant
    - Heterozygous
  - Recessive
    - Homozygous
    - Compound heterozygous
  - Hemizygous (X-linked male)
  - (Mitochondrial, and more)
- Penetrance





47