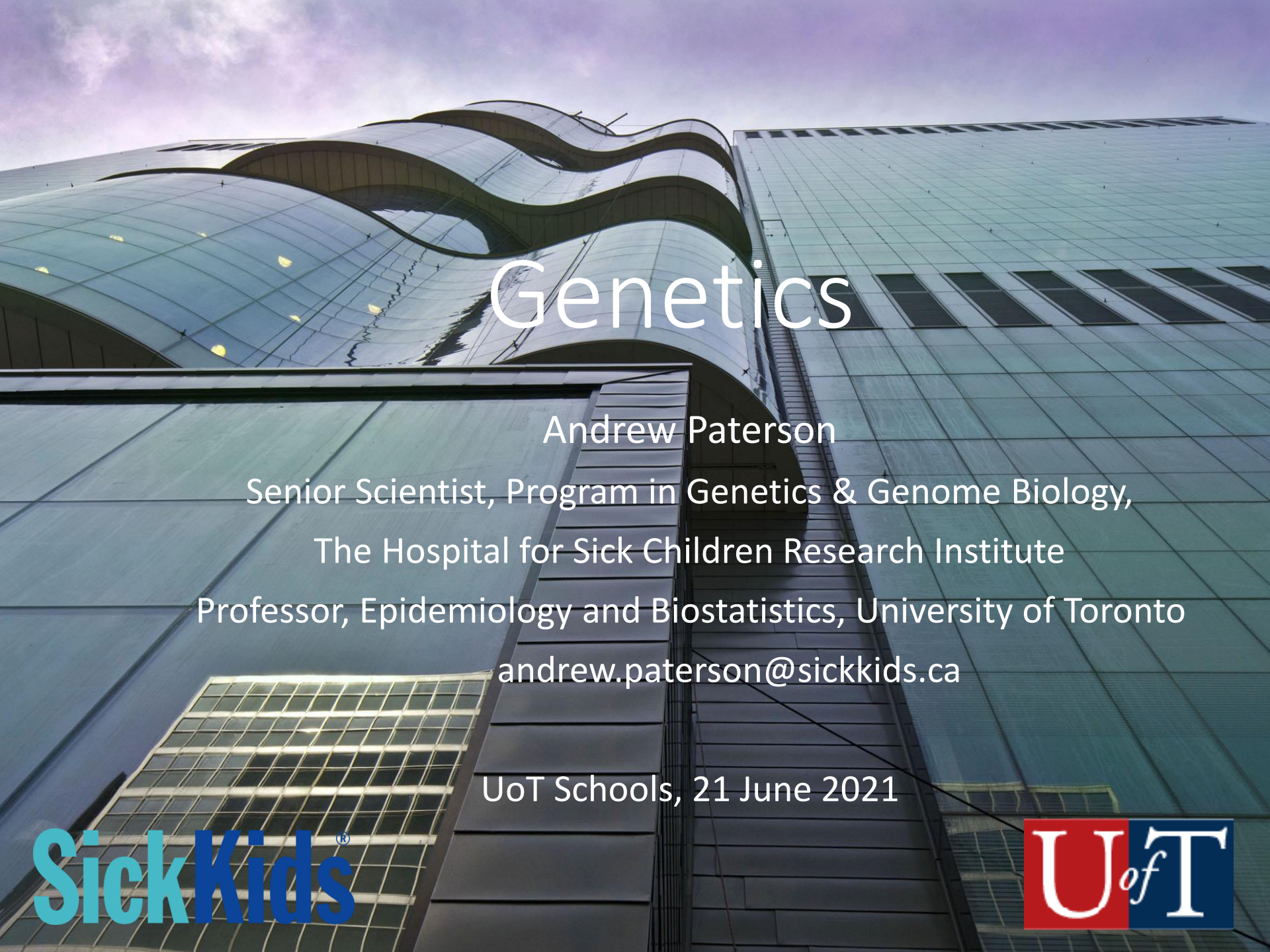


# Genetics

Andrew Paterson

Senior Scientist, Program in Genetics & Genome Biology,

The Hospital for Sick Children Research Institute

Professor, Epidemiology and Biostatistics, University of Toronto

andrew.paterson@sickkids.ca

UoT Schools, 21 June 2021

SickKids®

U of T

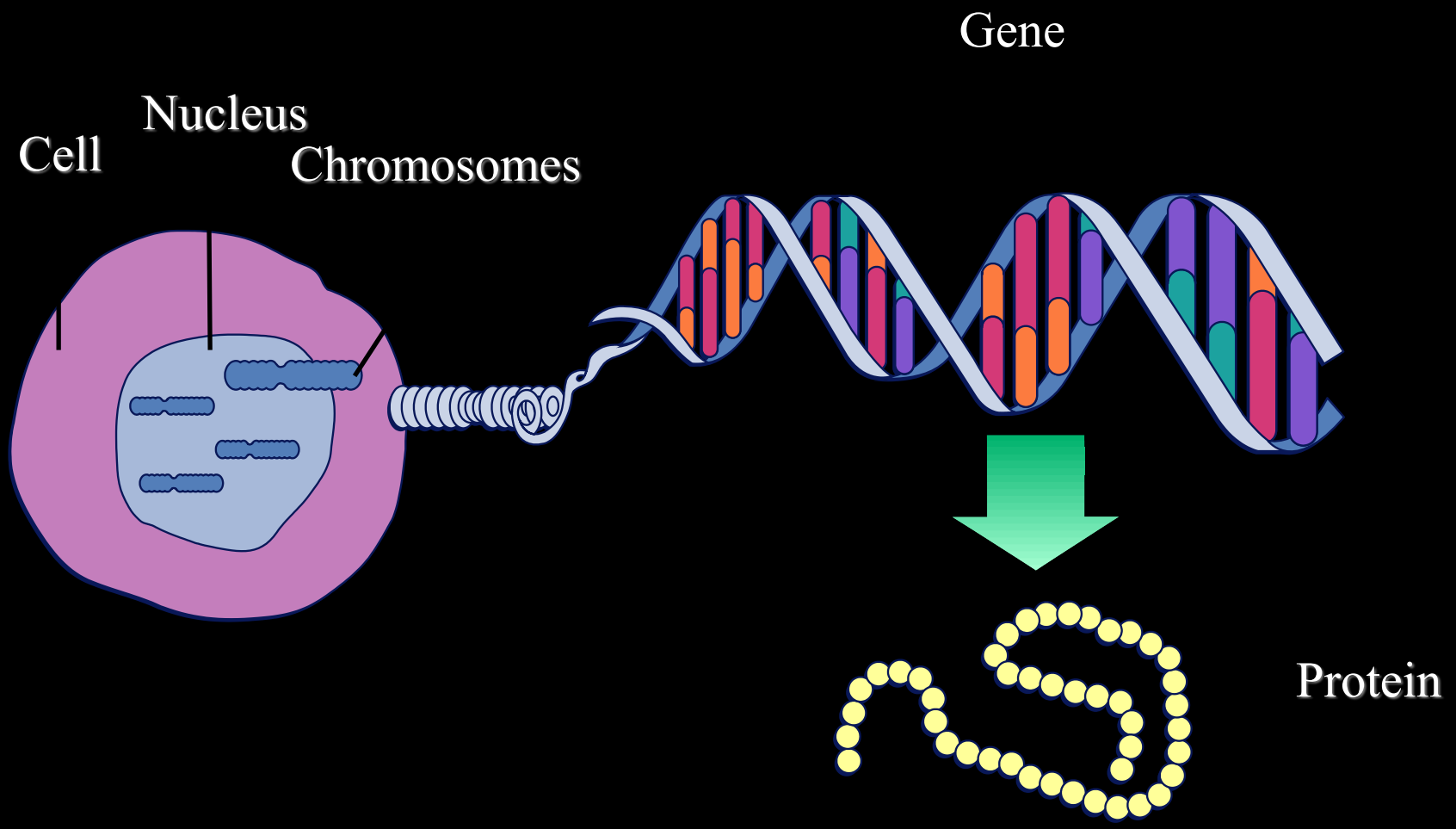
# General comments

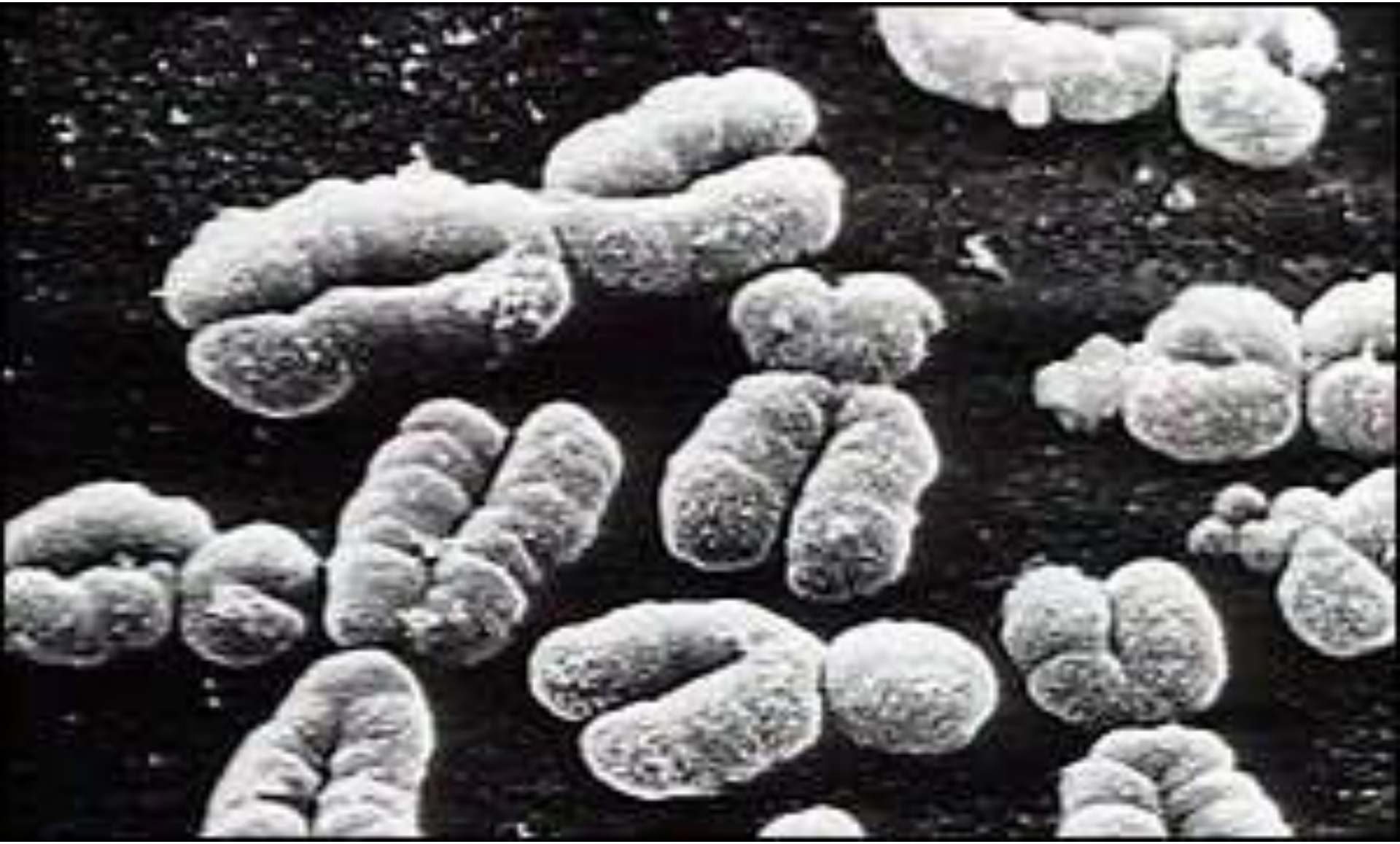
- There are no stupid questions
- Better to have a good understanding of a few basics rather … than superficial understanding of many things

# Human genome

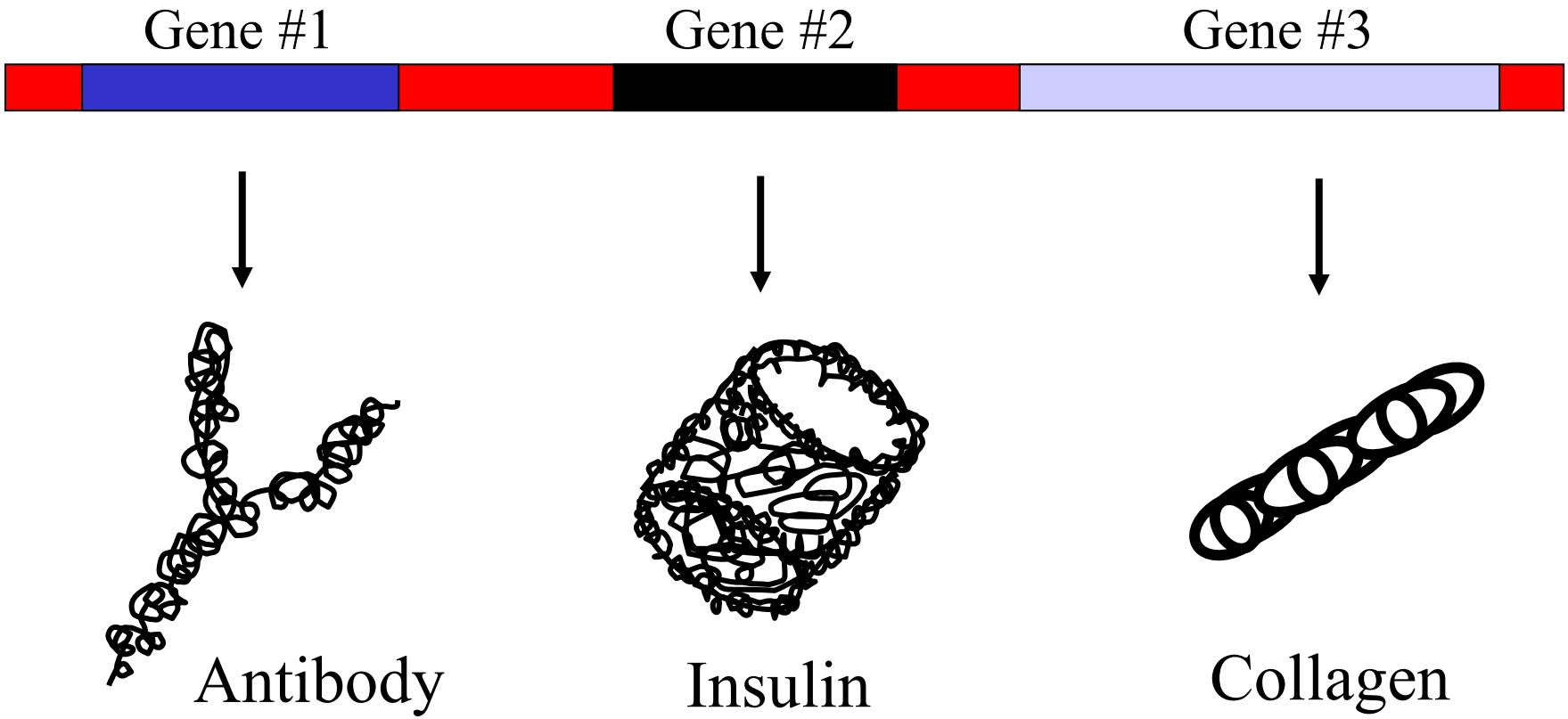
- Each person has 23 pairs of chromosomes
  - One copy of each chromosomes from mother and father
- Each chromosome contains a very long stretch of DNA made up of four different nucleotides (letters, A,T,C,G)
- Males have one X and one Y chromosome
  - Rare XXY
- Females have two X chromosomes
  - Rarely single X
- Thousands of genes on each chromosome
- About 20,000 genes in human genome

# Chromosomes, DNA and Genes

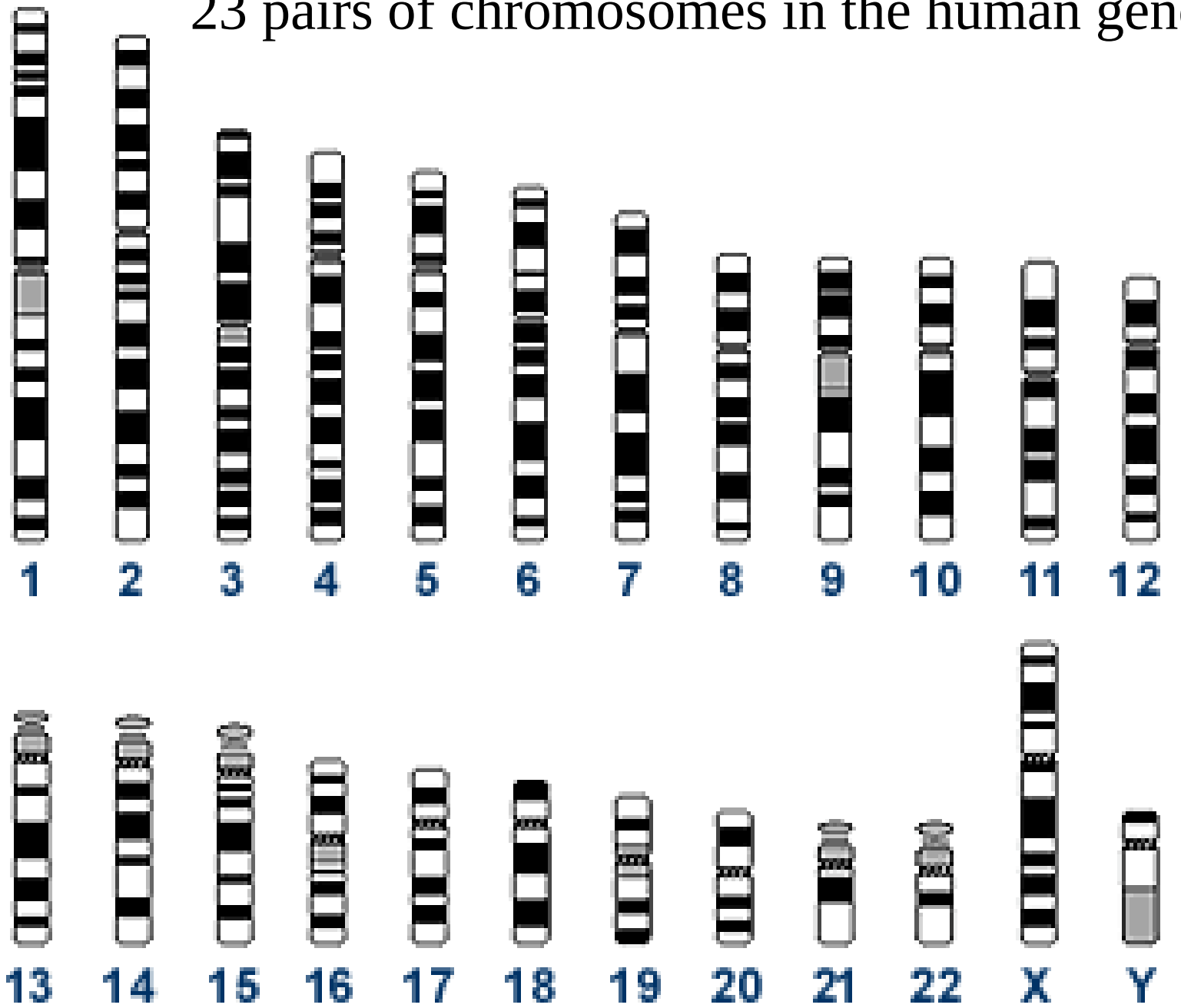




# A *Gene* is a segment of DNA that makes a protein...



# 23 pairs of chromosomes in the human genome



# Q: Do all humans really have two copies of each chromosome?

# Human genome

- 3,000,000,000 nucleotides (A,C,G,T) in the whole human genome
  - Paired, double helix
- 0.1% (3,000,000 nucleotides) where the DNA sequence differs between any two individuals
- Most of these variations are in ‘junk DNA’
  - regions between genes
- Minority of these variations change how products of genes (proteins) behave
- Genetic variations that influence protein behaviour may influence common human traits
- Which of these variations increases chance that a person has a particular trait

# Finding a gene is like finding the correct address of a house in Canada

- Chromosomes are like Provinces
- Bands on chromosomes are like Cities and Towns
- Genes are like Streets
- Positions in genes are like individual houses

- How do we find the right address?

# Man with giant kidneys to have major surgery to remove both

By Jack Guy, CNN

Updated 3:35 AM ET, Sun June 20, 2021


Warren Higgs will have major surgery in July.

**(CNN) —** A man from Windsor, southern England is set to have major surgery next month after his kidneys grew up to an estimated 40kg (88lb) due to polycystic kidney disease (PKD).

Warren Higgs has suffered multiple strokes and aneurysms due to his condition, he told CNN.

According to a GoFundMe page, he had a severe stroke 15 years ago due to PKD. That incident

# Autosomal Dominant Polycystic Kidney Disease

- Affects ~1:500
- Mode of inheritance: autosomal dominant
  - 50% chance for offspring of affected individuals to inherit the variant
- Using 9 families with ADPKD, linkage to chr 16 in 1985
  - *PKD1* found in 1994
- 80% of cases due to variants in *PKD1*
- Typically causes kidney failure aged 50-60 years
  - Requires either dialysis or kidney transplant
- Diagnosis by ultrasound or CT/MRI of abdomen
- Recently new treatment that can delay development of kidney disease
  - Tolvaptan, $33,000 per year

- https://www.ncbi.nlm.nih.gov/books/NBK1246/

BY MICHAEL LOCCISANO/GETTY IMAGES

# Angelina Jolie

- Mother
  - diagnosed with breast cancer aged 46
  - diagnosed with ovarian cancer aged 49
  - died of cancer at age 56 (life expectancy ~80)
- Maternal grandmother
  - Died from ovarian cancer
- Maternal aunt:
  - died from breast cancer
- Because of strong family history of early onset breast and ovarian cancer was offered genetic testing:
  - DNA sequencing of *BRCA1* and *BRCA2*
  - Had a cancer-causing variant in *BRCA1*
- Age 37 years bilateral risk reducing mastectomy
- Age 39 years bilateral preventive salpingo-oophorectomy
- Described decision to undergo preventive surgery as a proactive measure for the sake of her six children
- “Angelina Jolie Effect”

# Familial breast cancer due to BRCA1

- Mode of inheritance: autosomal dominant
  - Sex-specific (females, predominantly)
- 23 families with multiple females affected with early onset breast cancer
- Disease was linked to chromosome 17q in 1990
  - Gene in which variations cause early onset breast cancer was found in 1994

- https://www.ncbi.nlm.nih.gov/books/NBK1247/

# Recommended reading

- **A brief history of human disease genetics**
  – *Nature* **577**, pages 179–189 (2020)

https://www.nature.com/articles/s41586-019-1879-7

# 3 Genetic Topics

- **1. Genetic Variants**
- 2. Genetic Linkage
- 3. Genetic Association

# Genetic markers / Polymorphisms

- Differences in the DNA sequence between individuals
- Are transmitted from parents to children
- Allow us to determine if a particular variation in gene is associated with a disease
- Allow us to ‘Find the correct address’

# Polymorphism

- ‘The condition or character of being polymorphous; the occurrence of something in several different forms’

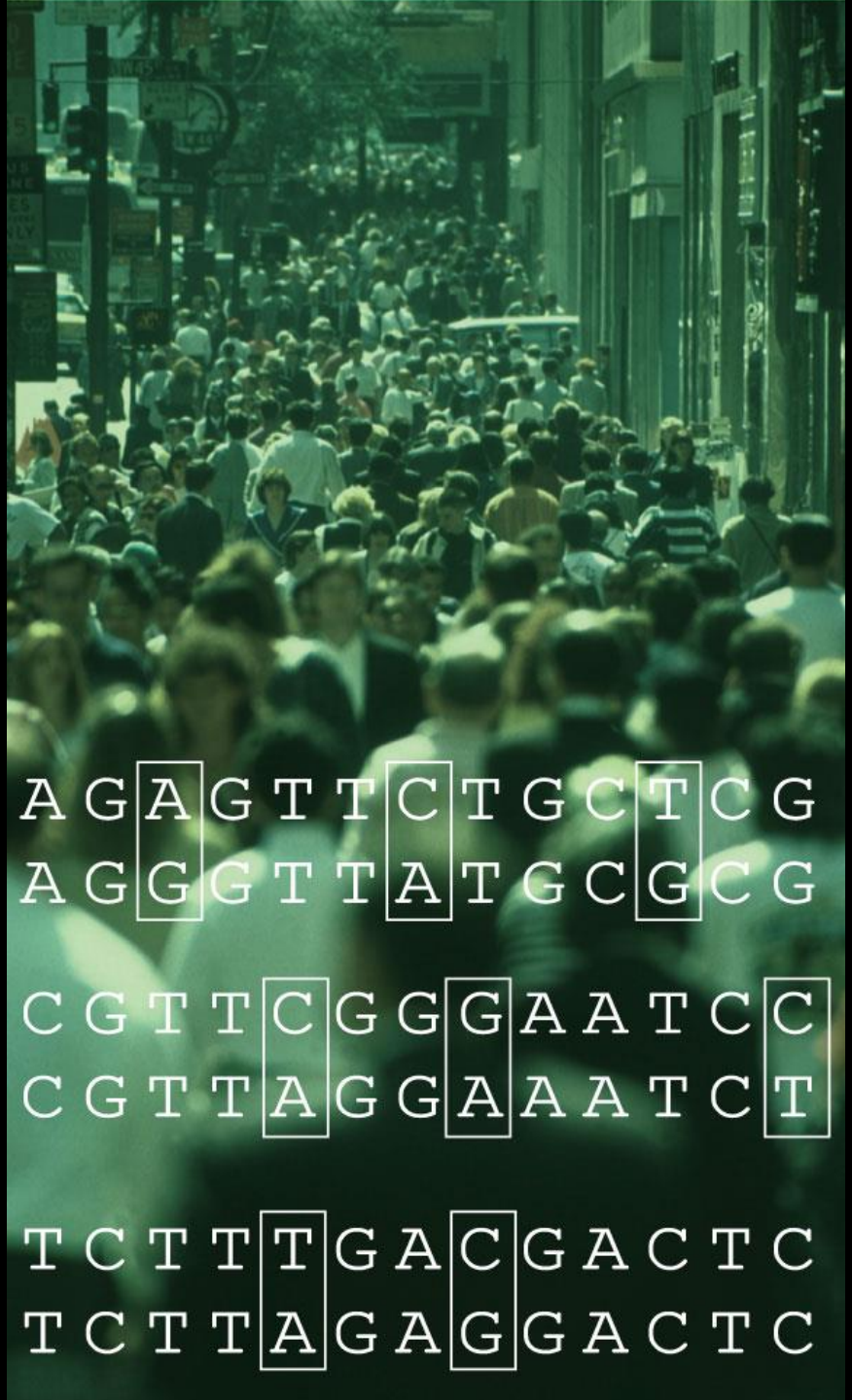
  – Oxford English Dictionary

# Polymorphism

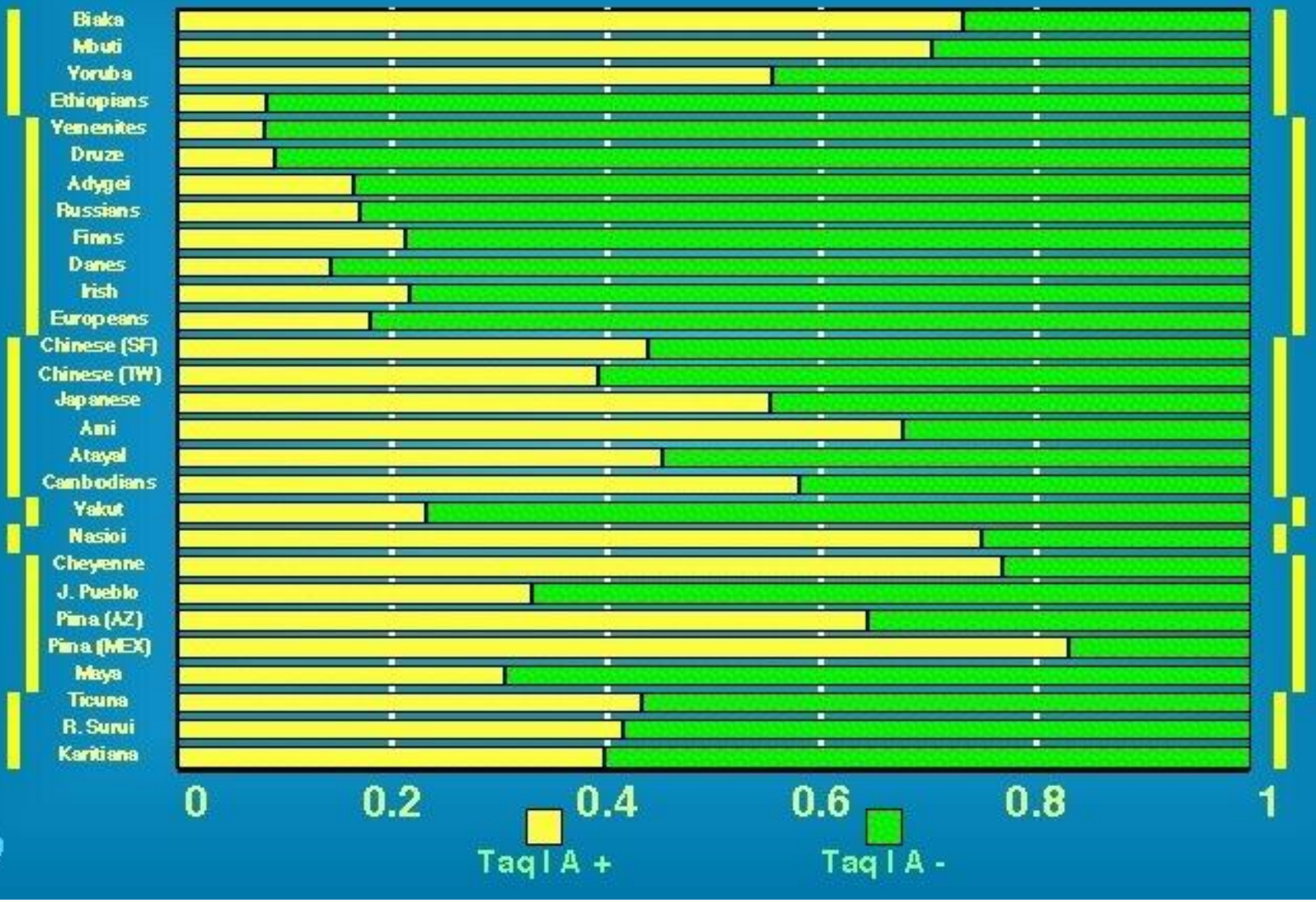
- Variation is the most interesting part of the human genome
- Phenotype
  - Sex/gender, hair colour, eye colour, height etc.
- Biochemical
  - blood groups, protein electrophoresis
- Genetic markers
  - DNA sequence differences between individuals

# ‘mutation’ vs ‘variant’

- Terminology has implications
  - If a disease is caused by a ‘mutation’
    - Does make the person a ‘mutant’?
  - Move away from stigmatizing people
    - ‘person with diabetes’ cf ‘diabetic’
  - ‘Variant’ has less stigma
    - We all have Millions of variants
    - Most likely do not contribute to disease/trait, but some do
- https://varnomen.hgvs.org/bg-material/basics/


AGAGTTCTGCTCG
AGGGTTATGCGCG
CGTTCGGGAATCC
CGTTAGGAAATCT
TCTTTGACGACTC
TCTTAGAGGACTC

# Frequencies of DRD2 Taq I A Alleles



# Single nucleotide polymorphisms (SNPs)

- Commonest polymorphism in human genome
- The most common class of variations causing Mendelian diseases
- Provide powerful approaches for genetic studies of common complex diseases
  - Frequency
  - Ease of automated genotyping at low cost
  - Relative old ‘age’ (low mutation rate)

# dbSNP

- Central database for *Homo sapiens* nucleotide (short genetic) variation (and other species – discontinued late 2017)

| **dbSNP Build #**<br>**date** | **118**<br>**Nov**<br>**2003** | **150**<br>**Feb**<br>**2017** | **151**<br>**Mar**<br>**2018** | **152**<br>**Dec**<br>**2018** | **154**<br>**June**<br>**2020** | **155**<br>**April**<br>**2021** |
|---|---|---|---|---|---|---|
| # Submissions (ss#) | 10.4 M | 907 M | 1,803 M | 1,998 M | 2,158 M | 3.3 B |
| # Unique positions (rs#) | 5.8 M | 325 M | 661 M | 695 M | 729 M | 1.1 B |

- http://www.ncbi.nlm.nih.gov/projects/SNP
- ftp://ftp.ncbi.nih.gov/pub/factsheets/Factsheet_SNP.pdf
- Population genetics: http://alfred.med.yale.edu/alfred/
- HGDP: http://hagsc.org/hgdp/files.html

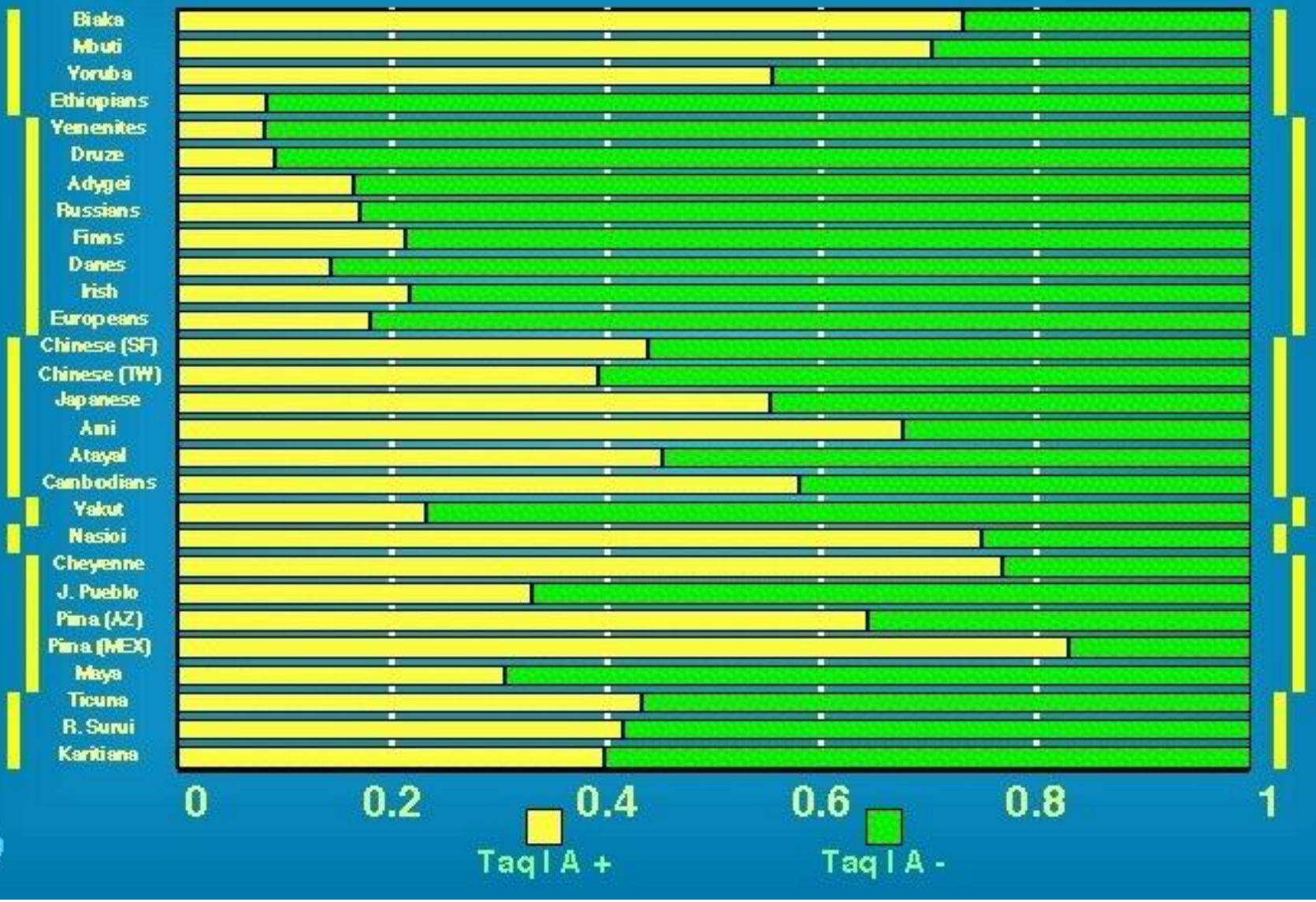
# Genotype and Allele frequencies

- Direct counting of unordered genotypes and alleles

| Genotypes | AA | AS | SS | Total |
|---|---|---|---|---|
| Counts | 189 | 89 | 9 | 287 |
| Genotype Frequencies | | | | |

| Alleles | A | S | Total |
|---|---|---|---|
| Counts | | | |
| Allele Frequencies | | | |

# Frequencies of DRD2 Taq I A Alleles



# Hardy-Weinberg Equilibrium (1908 CE)

- In a large random-mating population with no selection, mutation, migration, the allele frequencies and the genotype frequencies are constant from one generation to another, and there is a simple relationship between genotype and allele frequencies
- Many approaches in human genetics rely on the presence of Hardy-Weinberg Equilibrium

# Hardy-Weinberg Equilibrium

- At a biallelic marker, the frequencies of the two alleles (A and S) are *p* and *q*
  - *q=1-p*
- The expected genotype frequencies are:
  - AA: $p^2$
  - AS: $2pq$
  - SS: $q^2$
- The observed and expected genotype frequencies can be compared using a $Chi^2$ test

# Testing for departure from Hardy Weinberg Equilibrium

- Sickle cell anaemia
- Most common cause is due to autosomal recessive inheritance of mutations at Haemaglobin S (*HBB*)
- Highest frequency in Africa and descendants
- Genotype data from infants and adults for the Haemaglobin S in Tanzania
- Data from Allison 1956

# Infants

- *p(A)* = 0.81
- *q(S)* = 0.19

| **Group** | **AA** | **AS** | **SS** | **Total** |
|---|---|---|---|---|
| Observed | 189 (0.658) | 89 (0.310) | 9 (0.031) | 287 |
| Expected | | | | |

$\chi^2$ = 0.06, 1df, p value =1.0

# Adults

- *p(A)* = 0.80
- *q(S)* = 0.20

| Group | AA | AS | SS | Total |
|---|---|---|---|---|
| Observed | 400 (0.611) | 249 (0.381) | 5 (0.008) | 654 |
| Expected | | | | |

# Hardy-Weinberg Equilibrium (1908 CE)

- In a large random-mating population with no selection, mutation, migration, the allele frequencies and the genotype frequencies are constant from one generation to another, and there is a simple relationship between genotype and allele frequencies

# Detection of STR polymorphic markers

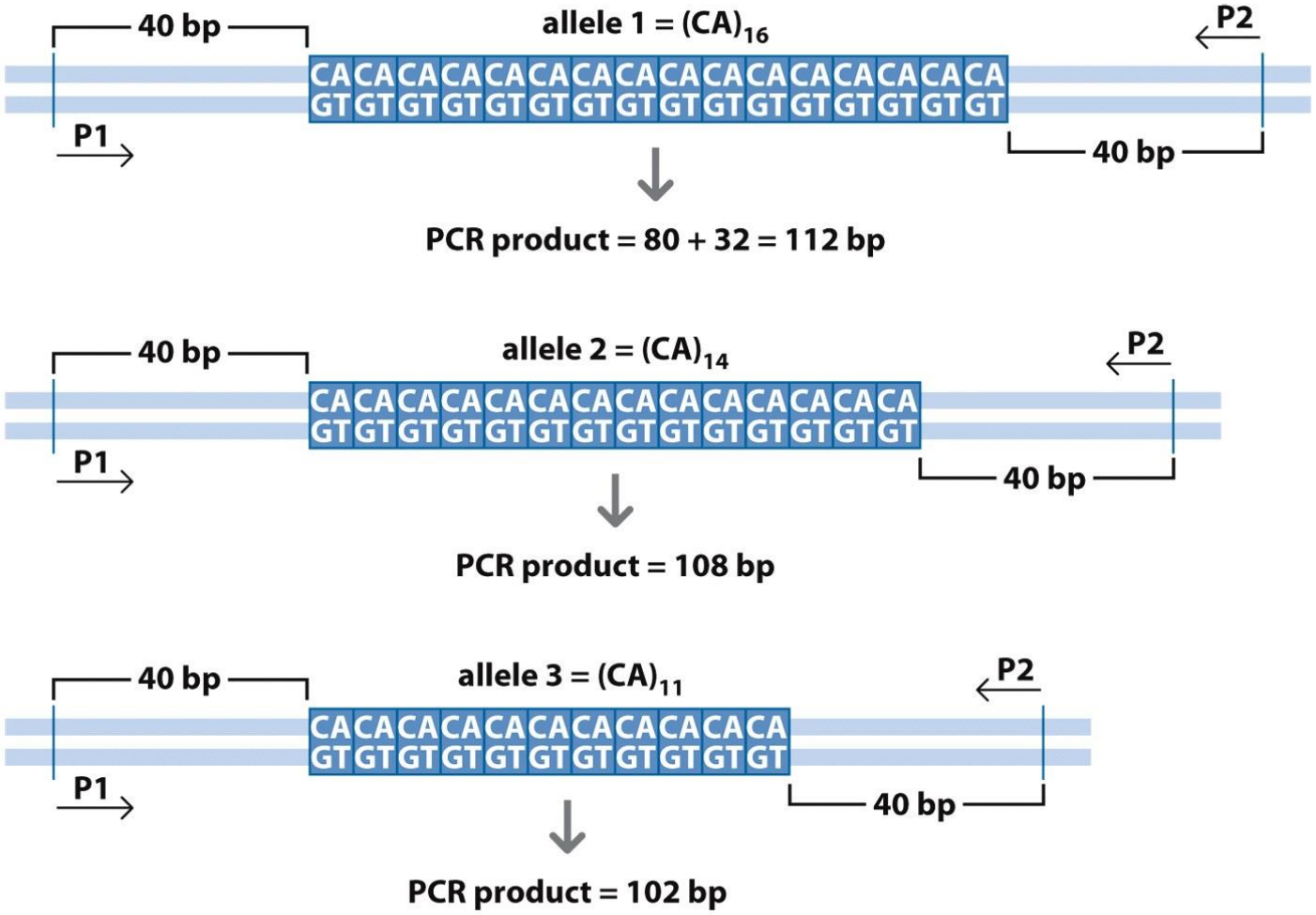



Figure 8.15 Human Molecular Genetics, 4ed. (© Garland Science)

Figure 8.15 from Strachan & Read, Human Molecular Genetics 4, Garland Science. PCR assay for a STR polymorphism. PCR products can be separated by size to determine alleles in an individual compared to other family members or control individuals.

# Forensic analysis

- Up to 15 unlinked Microsatellite markers
- Selected for
  - High heterozygosity
  - High quality genotype calling
  - Ability to multiplex (simultaneously PCR amplify all markers using a single reaction)
- Used in paternity, criminal, wrongful conviction
  - Probability of identity < $10^{-14}$
  - Probability of exclusion of paternity 0.99999
    - unless identical twin or clone
- http://www.cstl.nist.gov/biotech/strbase/

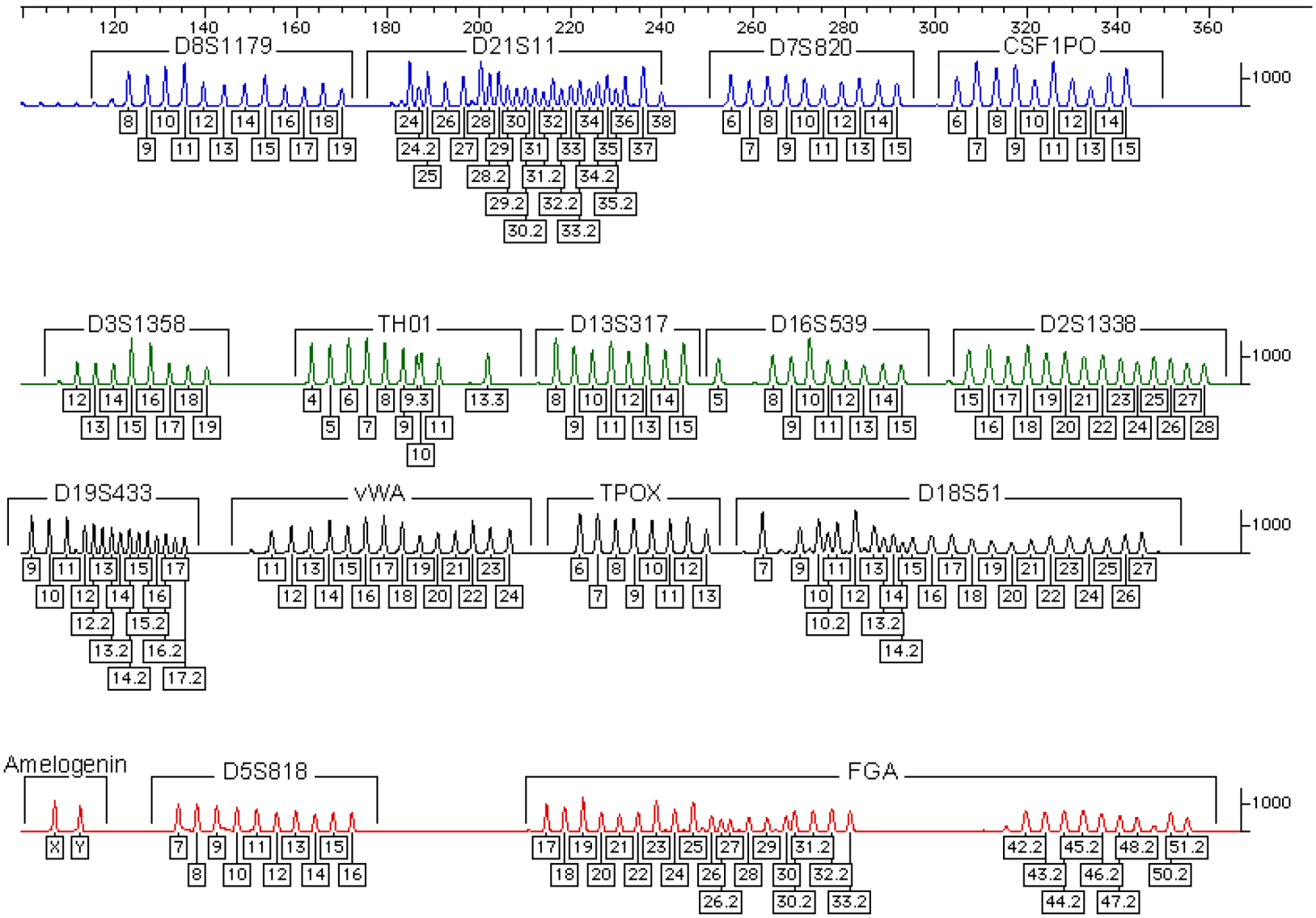
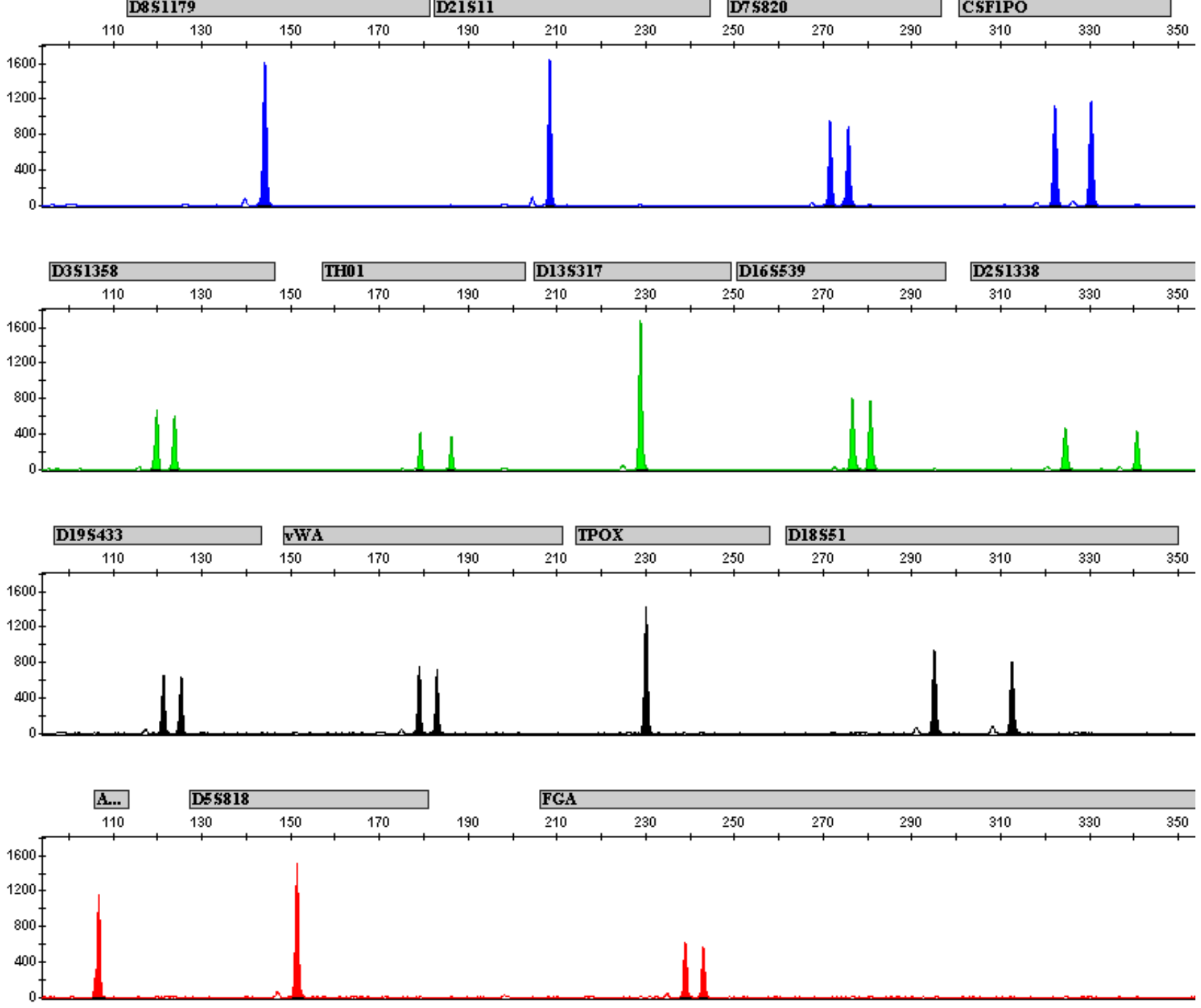



**Figure 5-1** Genotyper® software plot of the AmpF*ℓ*STR Identifiler Allelic Ladder, indicating the designation for each allele. These results were obtained on an ABI PRISM 310 Genetic Analyzer

Allelic ladder is a mix of all the alleles ever observed in the population:

Any single individual only usually has up to 2 alleles at each marker


D8S1179
D21S11
D7S820
CSF1PO
D3S1358
TH01
D13S317
D16S539
D2S1338
D19S433
vWA
TPOX
D18S51
A...
D5S818
FGA
110
130
150
170
190
210
230
250
270
290
310
330
350
1600
1200
800
400
0

# 3 Genetic Topics

- 1. Genetic Variants
- **2. Genetic Linkage**
- 3. Genetic Association

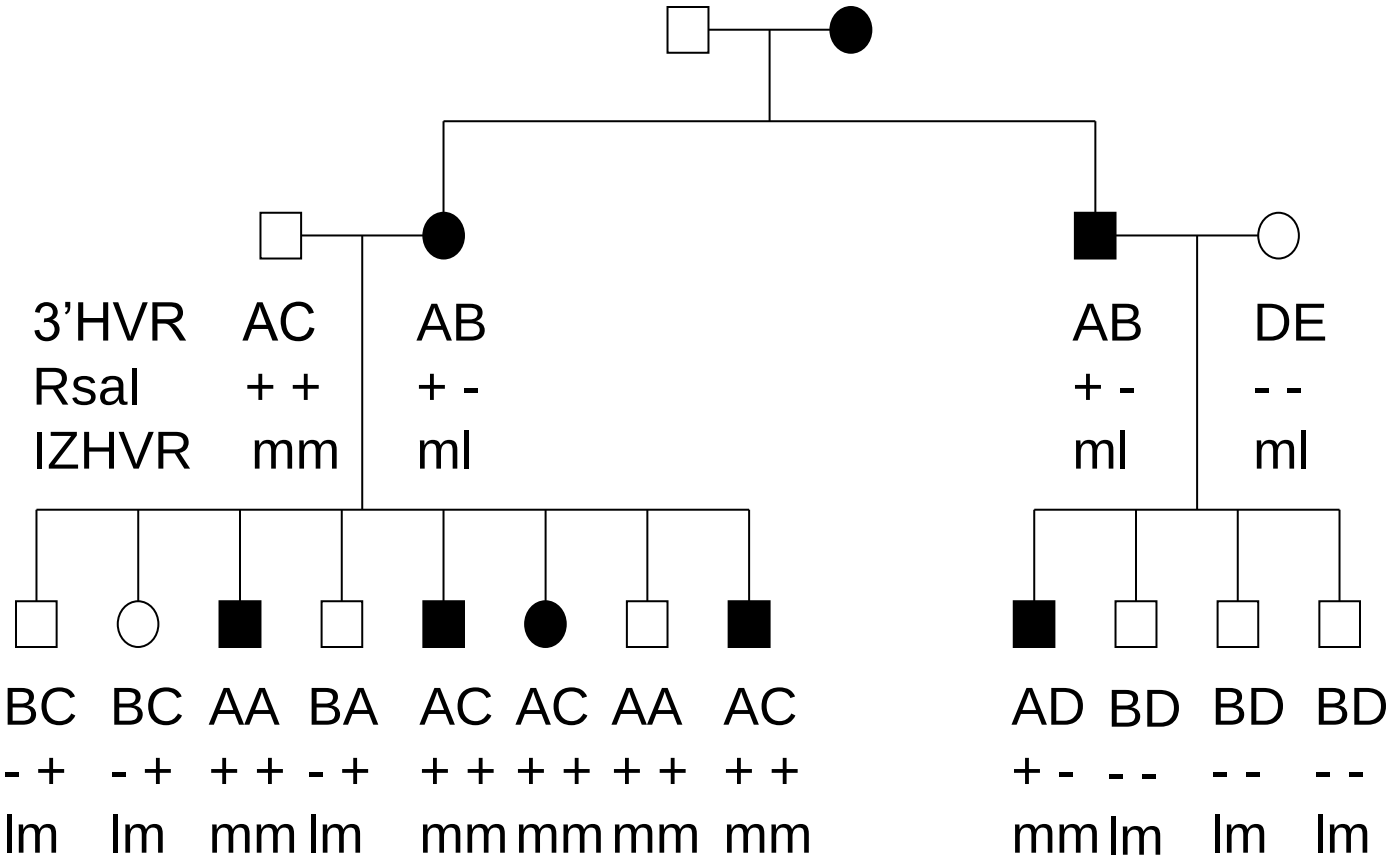
# Genetic linkage

- The cosegregation of a genetic marker with another marker or trait due to proximity on the same chromosome
- Meiotic recombination (crossing-over) in the germs cells of parents (sperm, ova) result in exchange of genetic material inherited from the grandparental chromosomes

# Genetic linkage is a fundamental concept

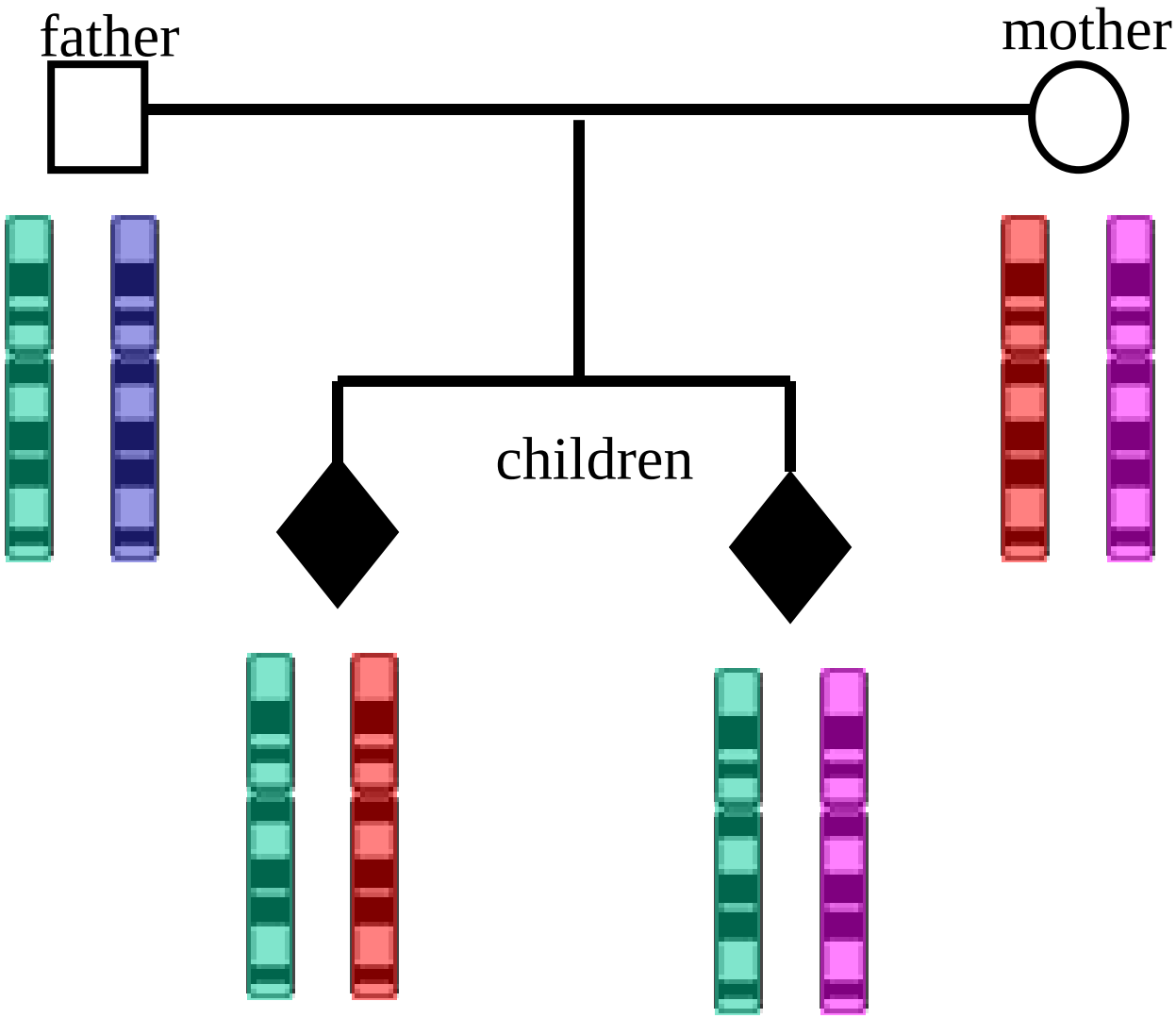
- Marker-trait linkage studies are the most popular approach for mapping
  - Mendelian disease genes
    - ADPKD ->PKD1
    - Early onset breast cancer -> BRCA1

# ADPKD

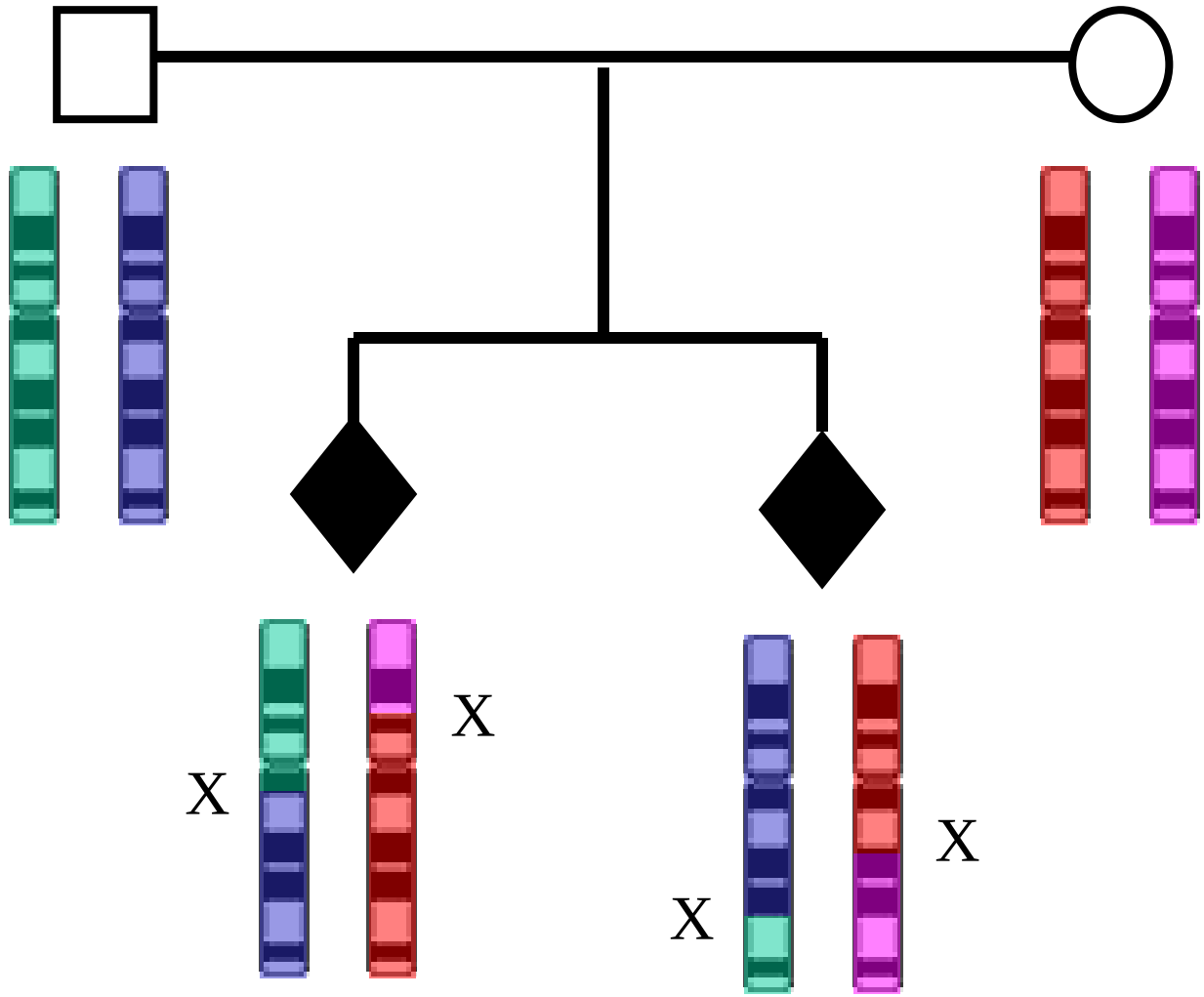



NDM-B pedigree (Reeders et al., Nature 317, 542-4; 1985)

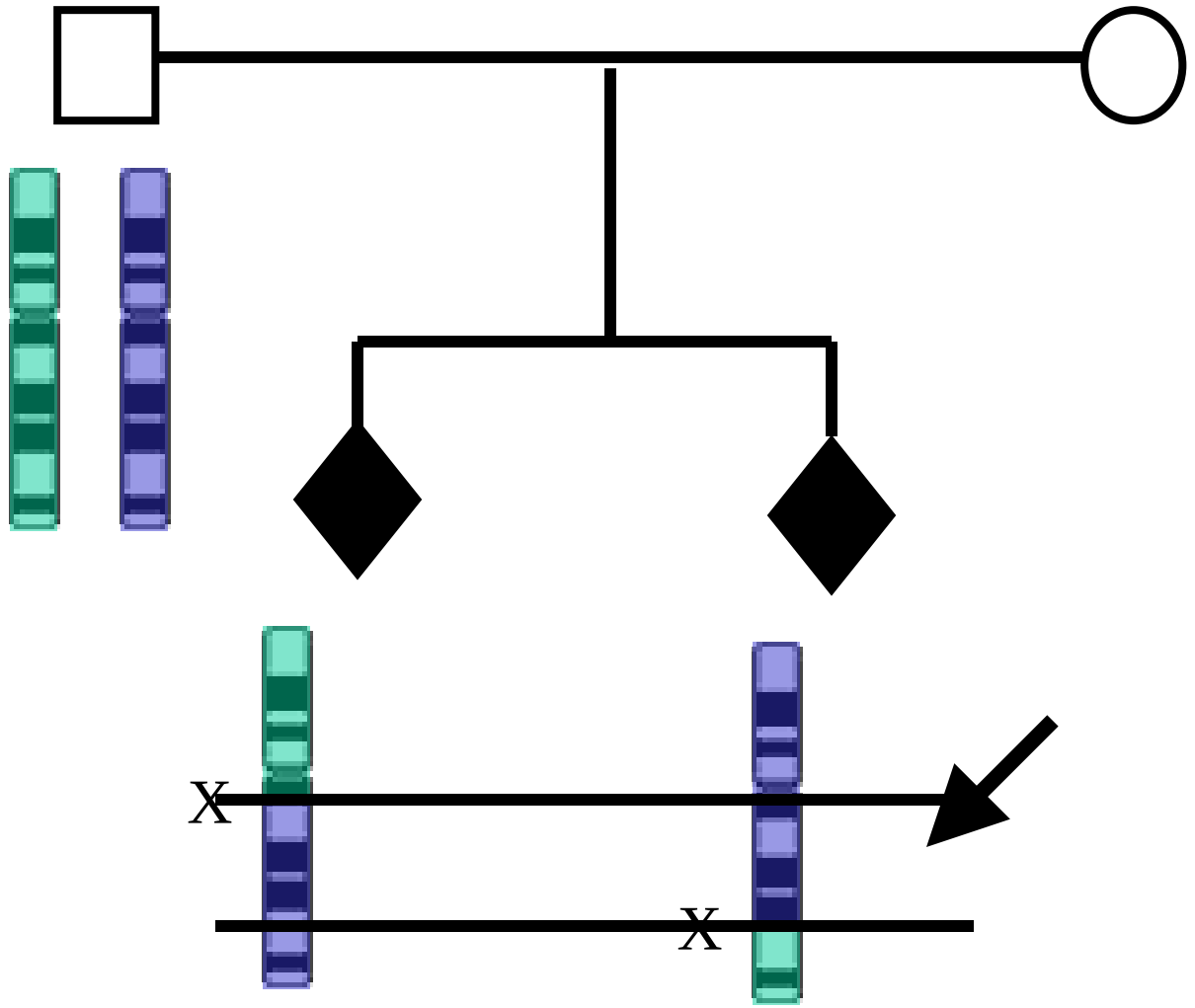
# Inheritance of chromosomes



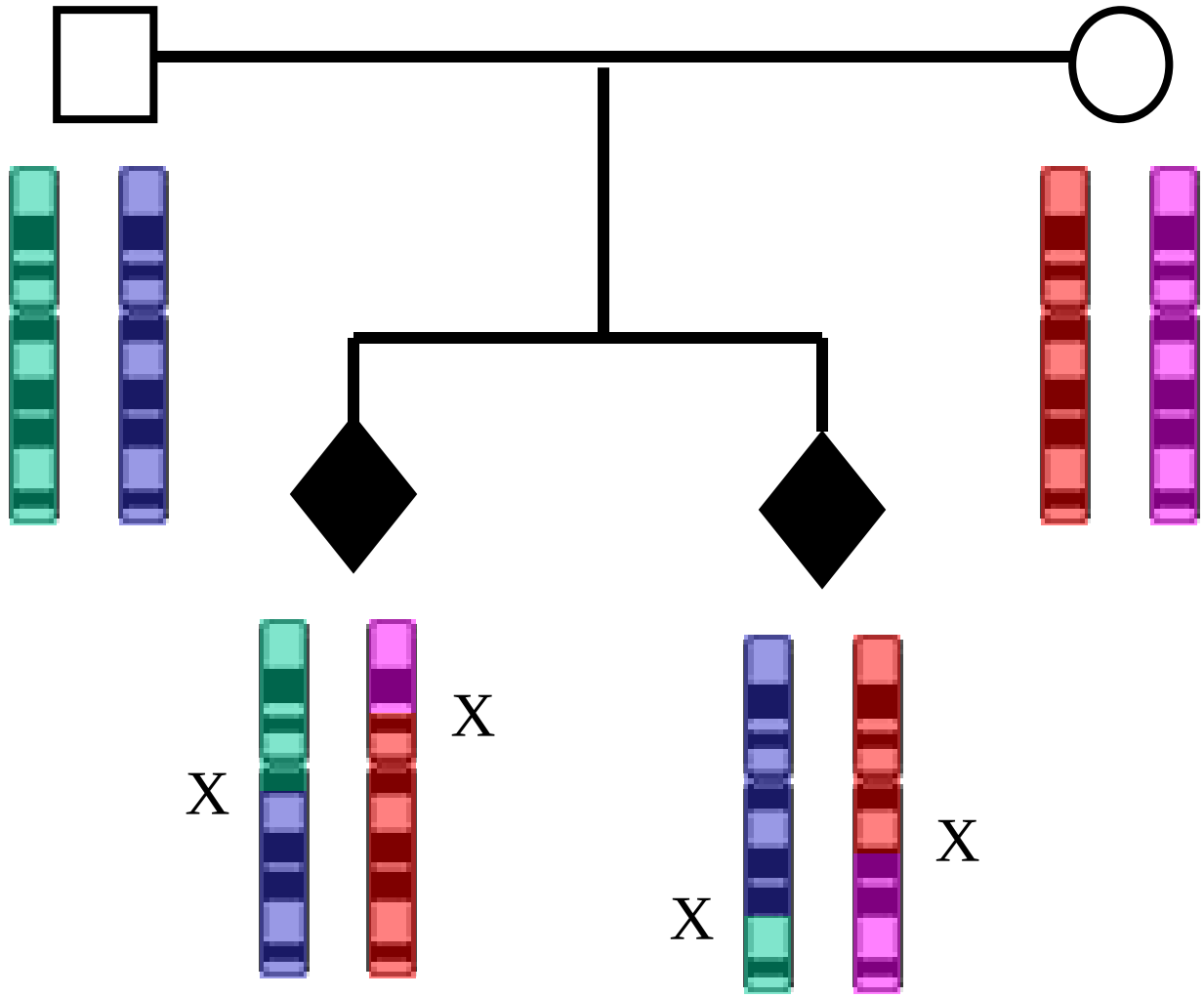
# Crossing-over (X) of chromosomes



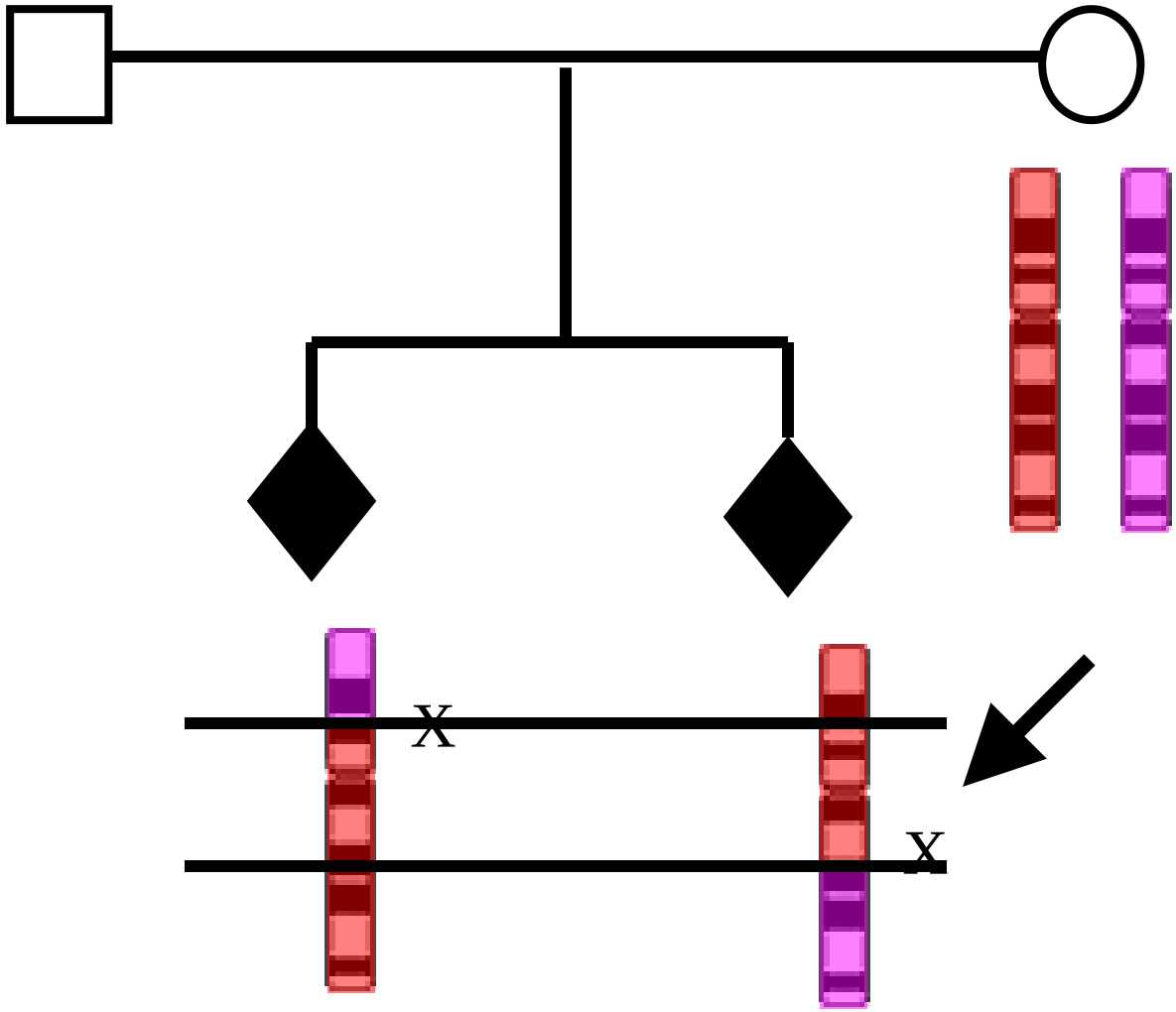
# Region of father's chromosome shared



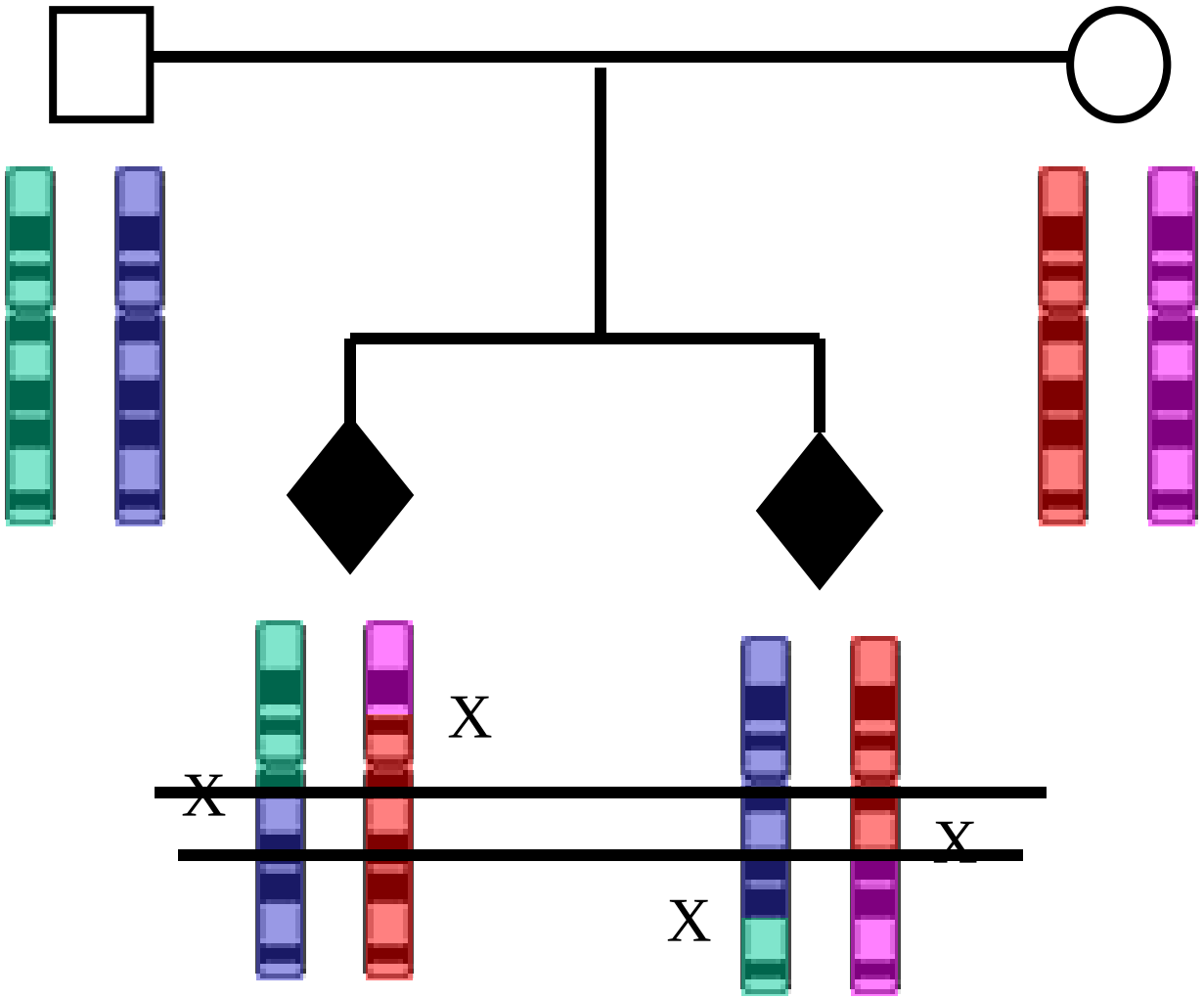
# Crossing-over (X) of chromosomes



# Region of mother's chromosome shared



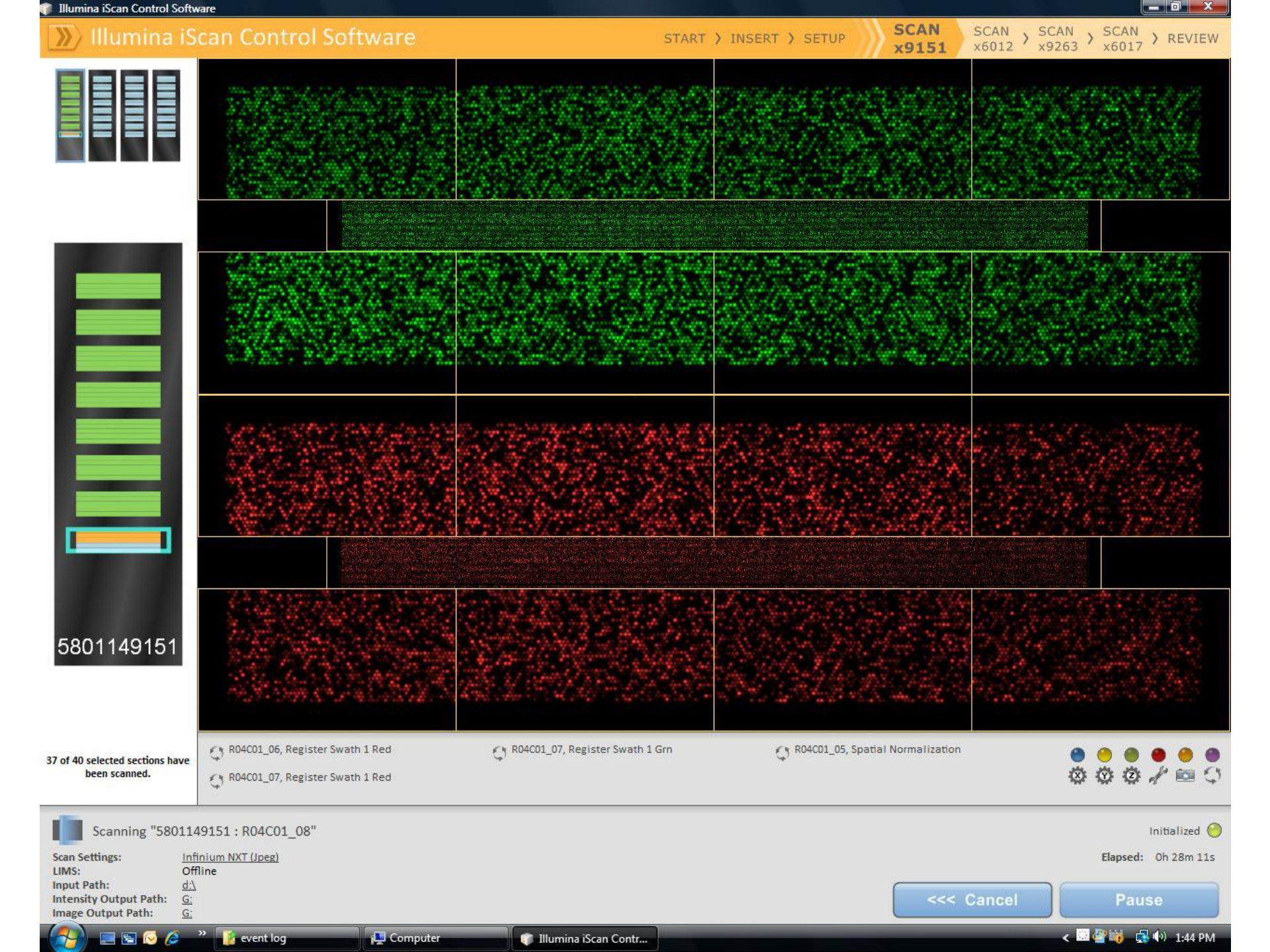
# Region of chromosome shared from both father and mother

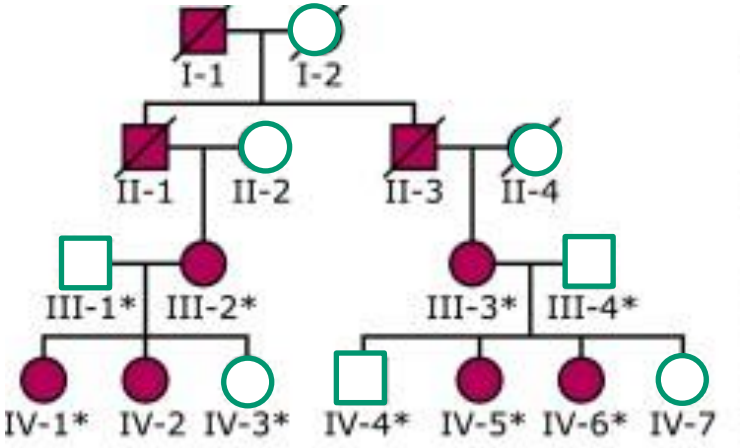


# Genome wide linkage study ‘genome scan’

- AIM: To map a new Mendelian disease to a specific location
- Genome-wide association (GWAS) SNP chips:
  - www.affymetrix.com (SNP 6.0 or Axiom)
  - www.illumina.com (HumanCore or Omni 1, 2.5 or 5)
  - Also allow for detection of CNVs


# The Omni family of microarrays.


Illumina iScan Control Software
Illumina iScan Control Software
START › INSERT › SETUP
SCAN x9151
SCAN x6012 › SCAN x9263 › SCAN x6017 › REVIEW
5801149151
37 of 40 selected sections have been scanned.
R04C01_06, Register Swath 1 Red
R04C01_07, Register Swath 1 Grn
R04C01_05, Spatial Normalization
R04C01_07, Register Swath 1 Red
Scanning "5801149151 : R04C01_08"
Initialized
Scan Settings: Infinium NXT (Jpeg)
LIMS: Offline
Input Path: d:\
Intensity Output Path: G:
Image Output Path: G:
Elapsed: 0h 28m 11s
<<< Cancel
Pause
event log
Computer
Illumina iScan Contr...
1:44 PM




Red = affected
white = unaffected
*= GWAS SNP (n=9)
Whole Exome Sequencing on 2 individuals

Linked regions:

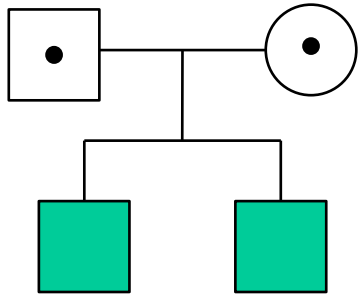
| Chromosome | Length (Mb) | Lod Score |
|---|---|---|
| 6 | 11.4 | 2.1 |
| 11 | 4.5 | 2.1 |
| 15 | 7.8 | 2.1 |
| | | |
| Total length | 23.7 Mb, <1% genome | |

105kb tandem duplication identified in RUNX2

Moffatt et al., Am J Hum Genet 2013; 92 (2), 252-258.
Metaphyseal dysplasia with maxillary hypoplasia and brachydactyly

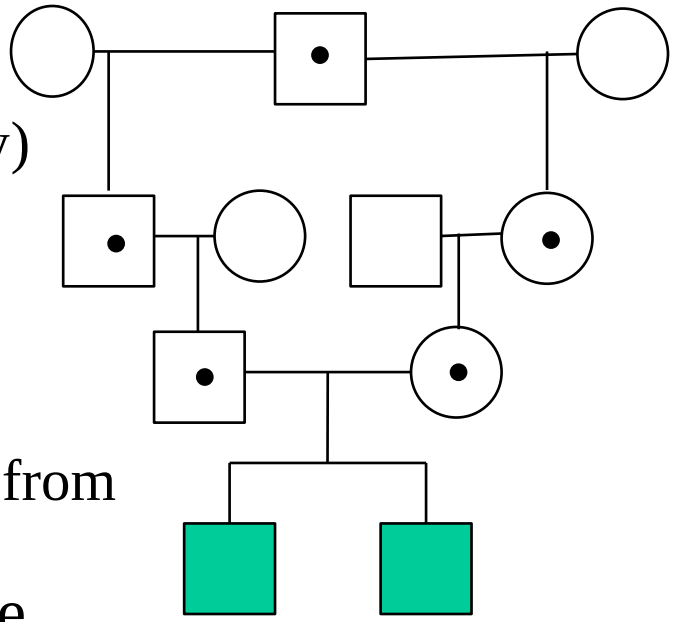
# Childhood cerebellar ataxia …

- Rare genetically heterogeneous disorder
  - Known loci excluded by Sanger sequencing
  - Assumed to be autosomal recessive
- Genome-wide linkage analysis using GWAS SNP chip
  - Max lod score=0.6 at ~¼ of the genome (637cM/ 479Mb)

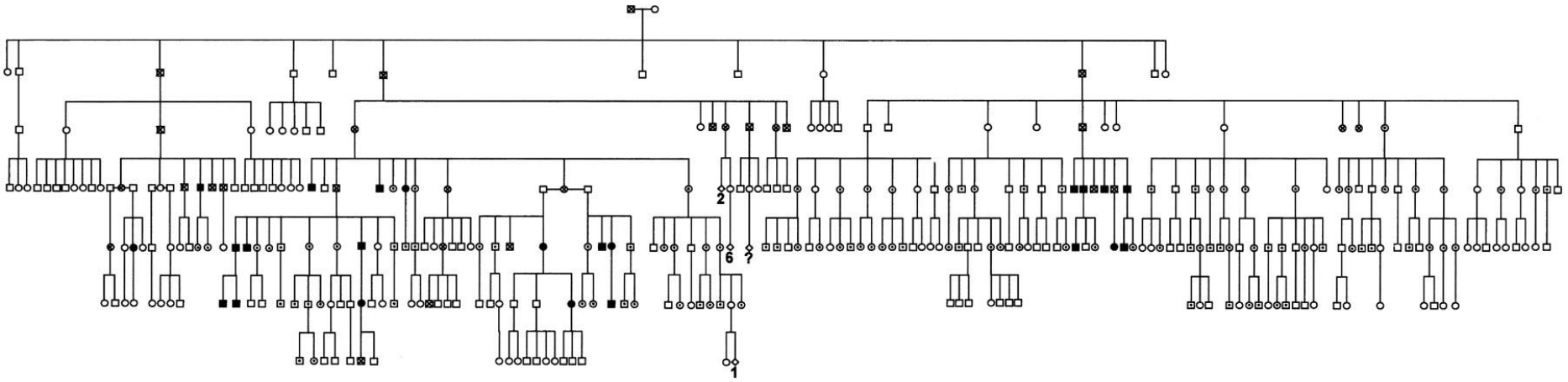


# … continued

- Parents from same village in Lebanon, not known to be related
  - Estimate inbreeding in parents
    - both close to first cousins
  - Parents also related ~ half-first cousins
  - Homozygosity linkage analysis (autozygosity)
- 4 ‘free’ informative meioses + 4 observed
  - lod score >1.2 at 2% of the genome
    - 78 cM/ 45 Mb
  - By generating NO MORE DATA, but just additional analysis, narrow the linked region from 25% to 2% of the genome
- Mutation in novel gene identified by exome sequencing within one of 4 homozygous regions with lod score >1: *PMPCA* p.Ala377Thr
  - https://academic.oup.com/brain/article/138/6/1505/2847363



# Quebec Platelet Disorder (QPD)



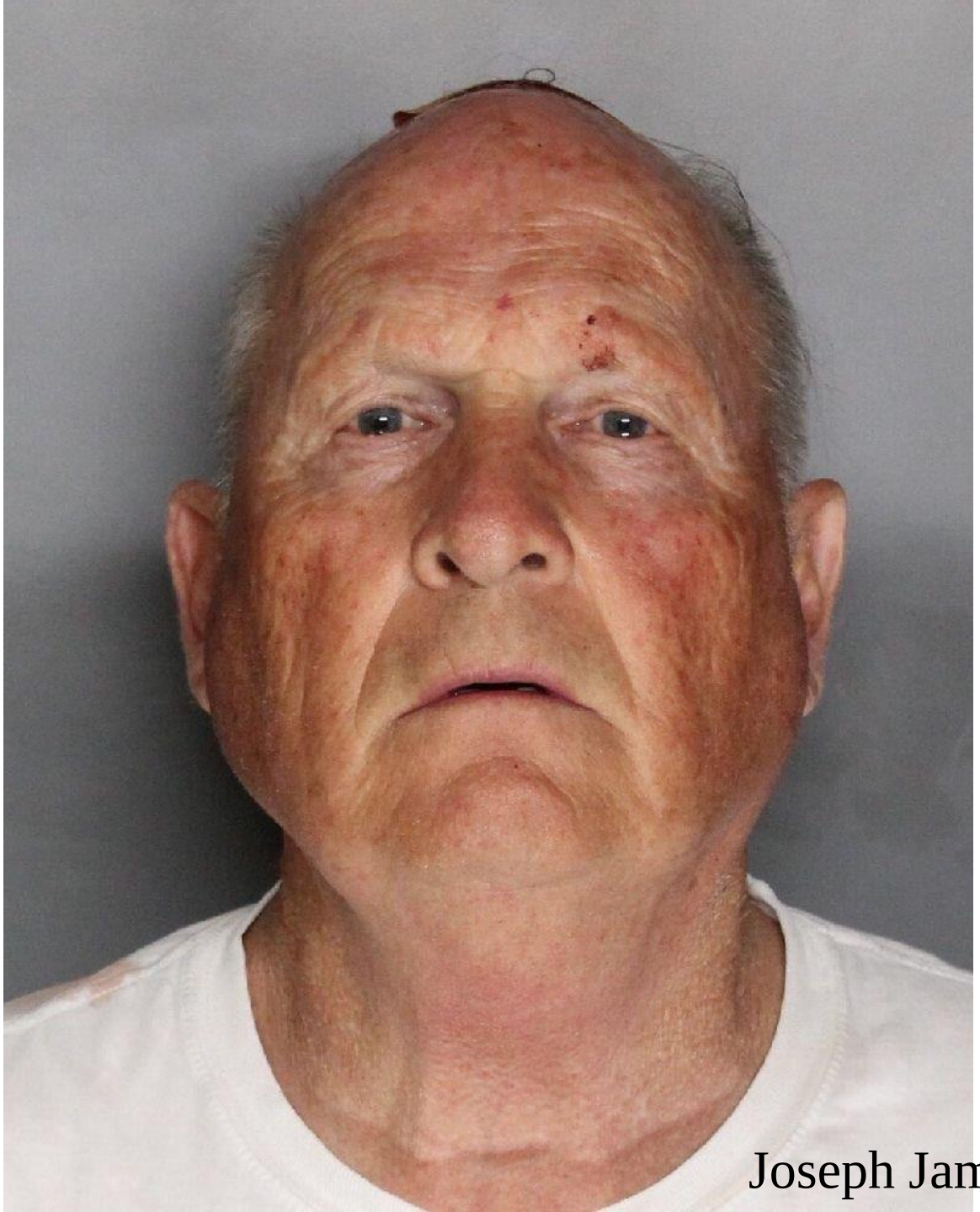
# Quebec Platelet Disorder

- Single large family with autosomal dominant fibrinolytic alpha-granule deficiency bleeding disorder
  - Massive increase in platelet urokinase plasminogen expression and storage (uPA=PLAU)
  - Genome-wide linkage analysis
    - lod score =11 (significance threshold>3) on chr 10q24
    - 2 Mb; including PLAU and 22 other genes
    - **<0.1% of the genome** (Diamandis et al. Blood 2009)
  - No candidate variants identified by Sanger sequencing of PLAU coding regions

# Quebec Platelet Disorder

- ~78 kb tandem duplication including PLAU
  - Likely due to de novo enhancer of PLAU expression in megakaryocytes
    - Paterson et al., Blood 2010; Hayward et al., Plos ONE 2017
- Specific anti-fibrinolytic therapy: tranexamic acid
- Created simple molecular diagnostic: extend pedigree & phenotyping
- **26 years** from initial description to molecular characterization

# More families



Joseph James DeAngelo

The New York Times

# How a Genealogy Site Led to the Front Door of the Golden State Killer Suspect


A Sacramento County sheriff's deputy carried bags of evidence from the home of the suspect in the Golden State Killer case on Thursday. Jim Wilson/The New York Times

By Thomas Fuller

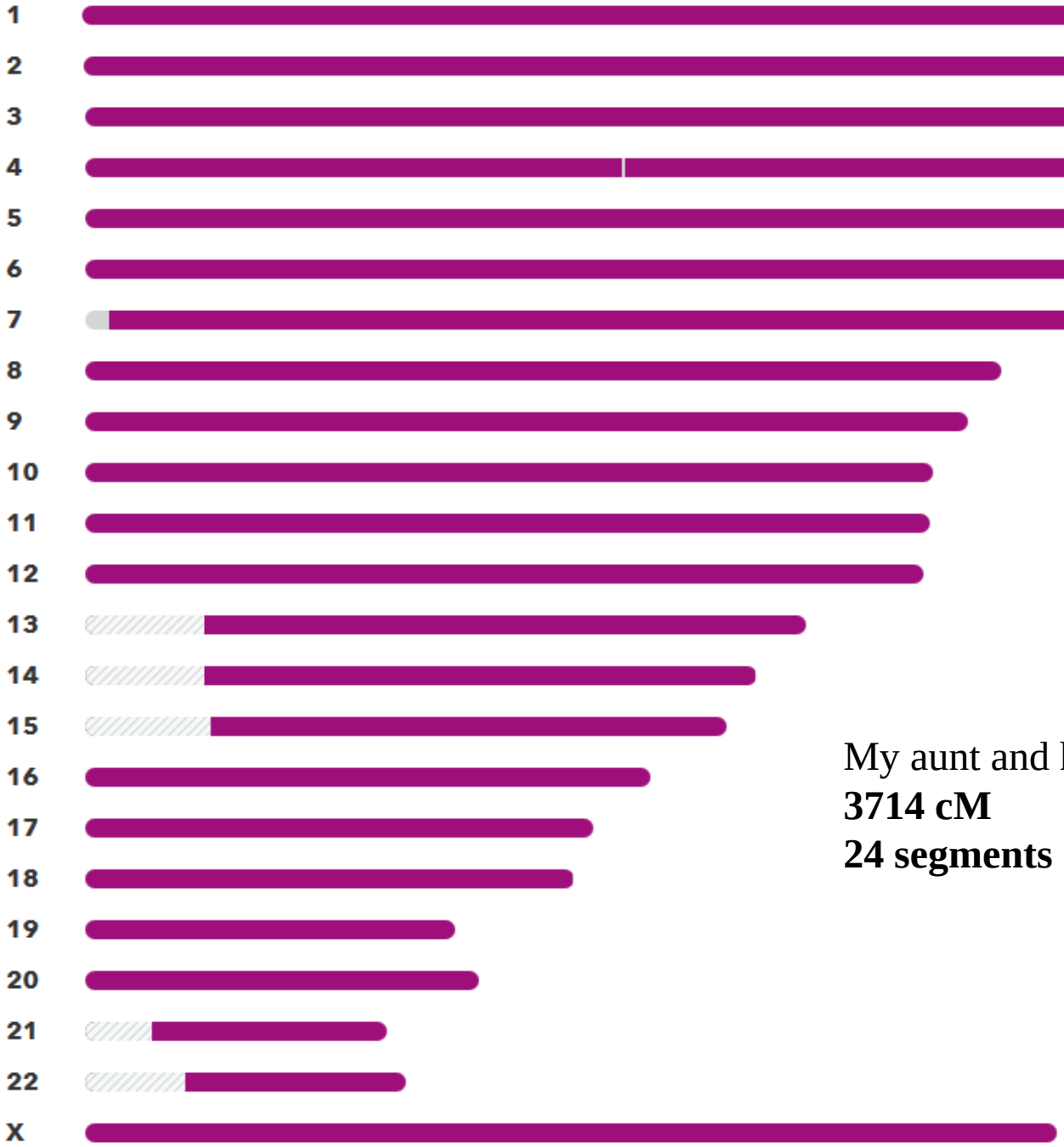
April 26, 2018

# Golden State Killer

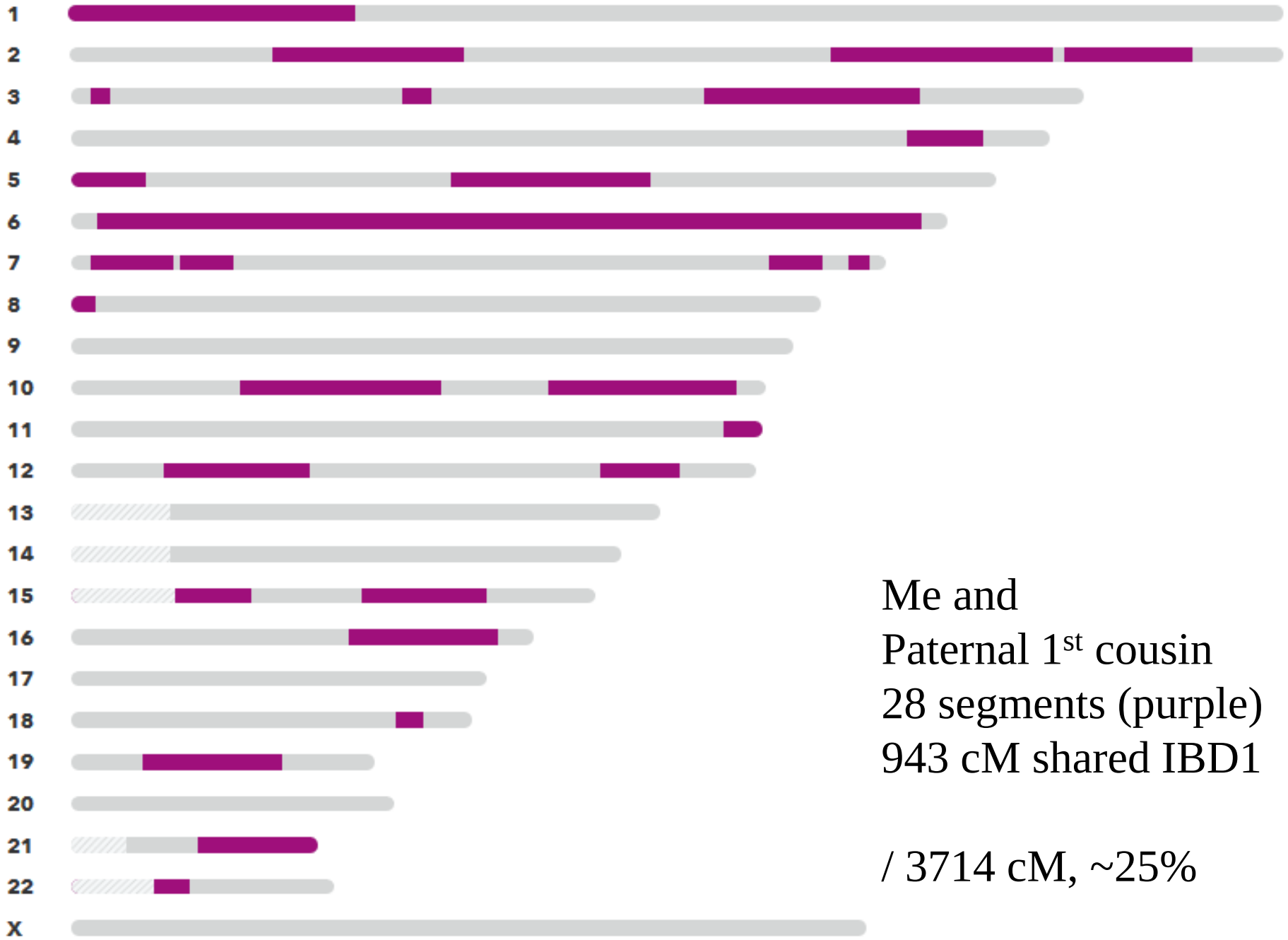
- Single suspect for >50 rapes and >12 murders in California from 1974 to 1986
  - Connected crimes by forensic perp DNA left at scenes
    - But no match in FBI database
      - 15 unlinked microsatellite markers
      - -> cold case
  - Genotyped perp DNA with GWAS array (2018)
    - ~ 1M SNPs
      - Uploaded to GEDMatch
        - » Genealogy db: 650k people who have downloaded Direct To Consumer (e.g. 23andMe; ancestry.com) genetic data
          - » Upload to GEDMatch
      - Found multiple 3rd cousins match perp for few long shared segments
        - » Built family tree, recruited additional relatives
        - » Charge suspect; plead guilty to multiple charges; life sentence

# > 20 additional serious cold cases solved by similar approach

- Estimated that 60% of European ancestry Americans will have similar close relatives now in DTC
  - will continue to increase
    - 23andMe have >10 M customers
    - Assume people deposit data in GEDMatch
    - doi:10.1126/science.aav7021

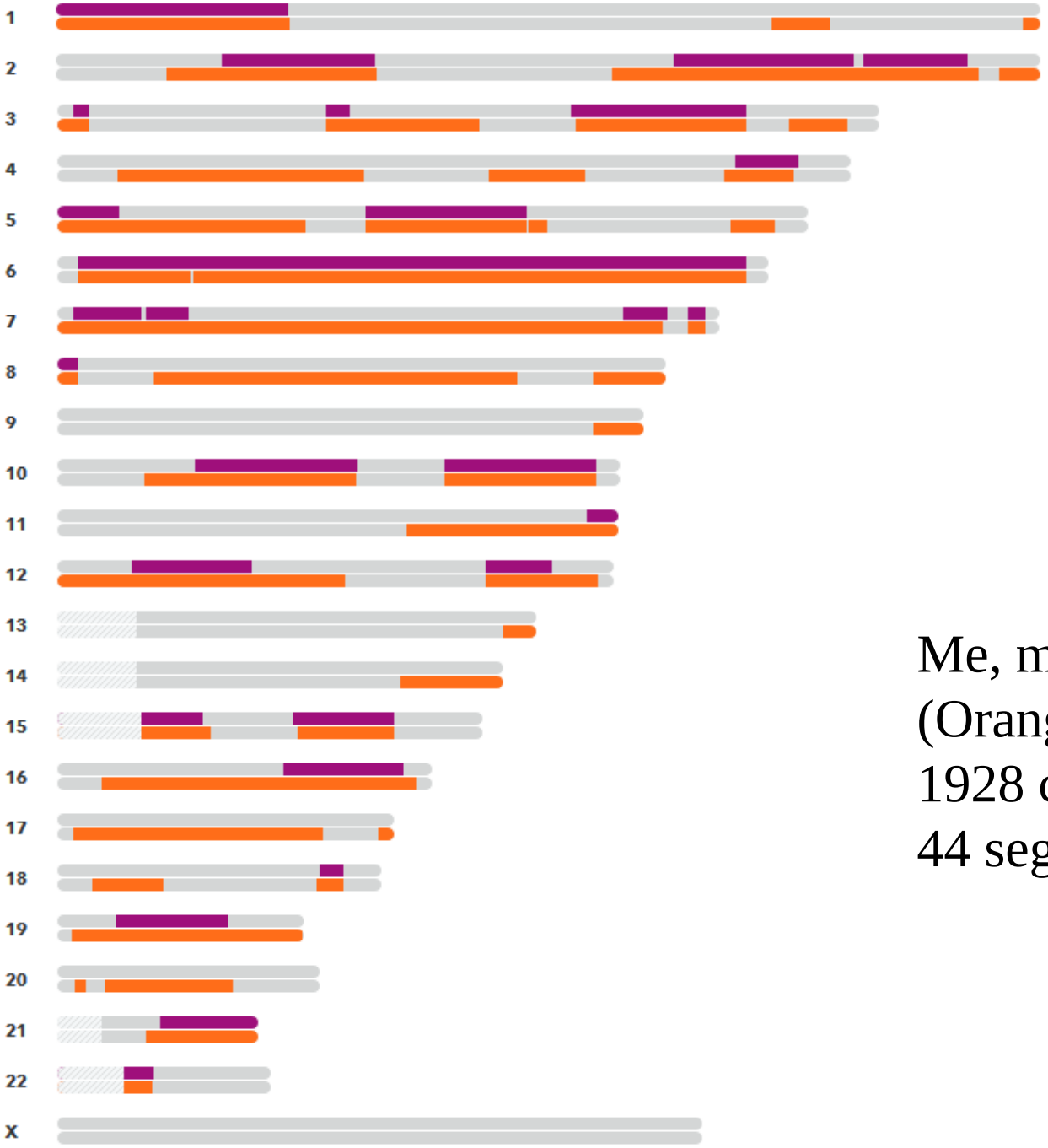



My aunt and her son (my cousin)
**3714 cM**
**24 segments**




Me and
Paternal 1st cousin
28 segments (purple)
943 cM shared IBD1

/ 3714 cM, ~25%




Me, my paternal aunt: (Orange), my cousin (purple)
1928 cM
44 segments

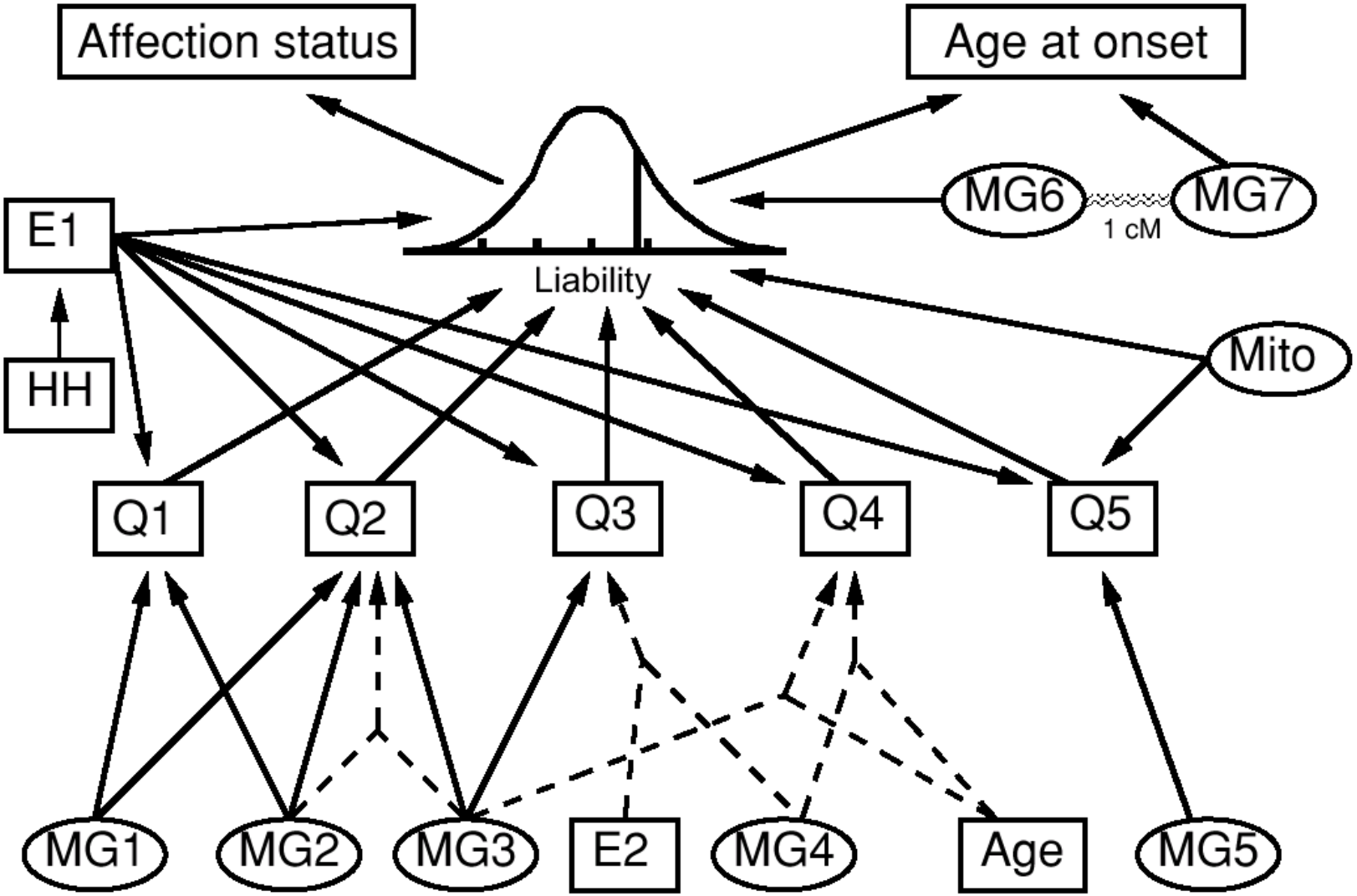

Me and 2nd to 3rd cousin

Purple:
6 segments sharing 1 haplotype identical by descent

159 cM / 3,714 cM
~4%

# 3 Genetic Topics

- 1. Genetic Variants
- 2. Genetic Linkage
- **3. Genetic Association**


Affection status
Age at onset
MG6
MG7
1 cM
E1
Liability
HH
Mito
Q1
Q2
Q3
Q4
Q5
MG1
MG2
MG3
E2
MG4
Age
MG5

# Phenotype/Trait

- Phenotypes
  - binary (discrete)
    - Affected vs. unaffected
- Traits
  - continuous, quantitative (e.g. height, weight, cholesterol, blood pressure etc.)
  - classically mean and SD

# Evidence for genetic influence

- Clustering of phenotypes in groups
  - Ethnicity
  - Families
  - Extent of clustering depends on relationship
- Correlation of traits between related individuals
  - Correlation differs by relationship
    - MZ twin: distant
  - Correlation between cryptically (distant) relationships in population samples
    - ‘chip’ heritability (Visscher, Yang)

# Common complex diseases

- E.g. most forms of common cancers, diabetes (type 1 and 2), rheumatoid arthritis, schizophrenia, autism, inflammatory bowel disease, hypertension etc.
- Family
  - No clear mode of inheritance in most families
  - Sibling recurrence risk
- Twin
- adoption
- ? segregation analysis

# Twin studies

- Two main types of twins
  - Monozygotic (Identical) MZ
  - Dizygotic (Fraternal) DZ
- Population based
- Ascertainment biases
- Same sex (MM,FF) and opposite sex (MF) DZ twins
- Compare to sibs

# Why find QTLs?

- Public Health
  - Use to identify/screen those at risk
    - Number of loci
    - Effect size (Odds Ratio, Hazard Ratio, Population Attributable Fraction, etc.), genetic model, genotype and allele frequencies
    - Interactions with each other Genes (epistasis) and Environment
    - Genetic risk scores (across all loci p<threshold)
- Biology
  - pathways, opportunities for intervention/treatment

# Future uses of genetic information

- Identify those at increased and decreased risk of disease
  - Genetic risk scores
    - Either loci that meet genome-wide significance, or also include those with less stringent significance criteria
- Screening and preventative strategies can be tested in those at risk
  - Behaviour modification may be required only in those at risk
- Pharmacogenetics
  - Predict response and adverse effects to particular medicines
- Understanding biology
  - Novel therapeutics
  - Insights into mechanisms and pathways, perhaps even environmental factors

# Association studies

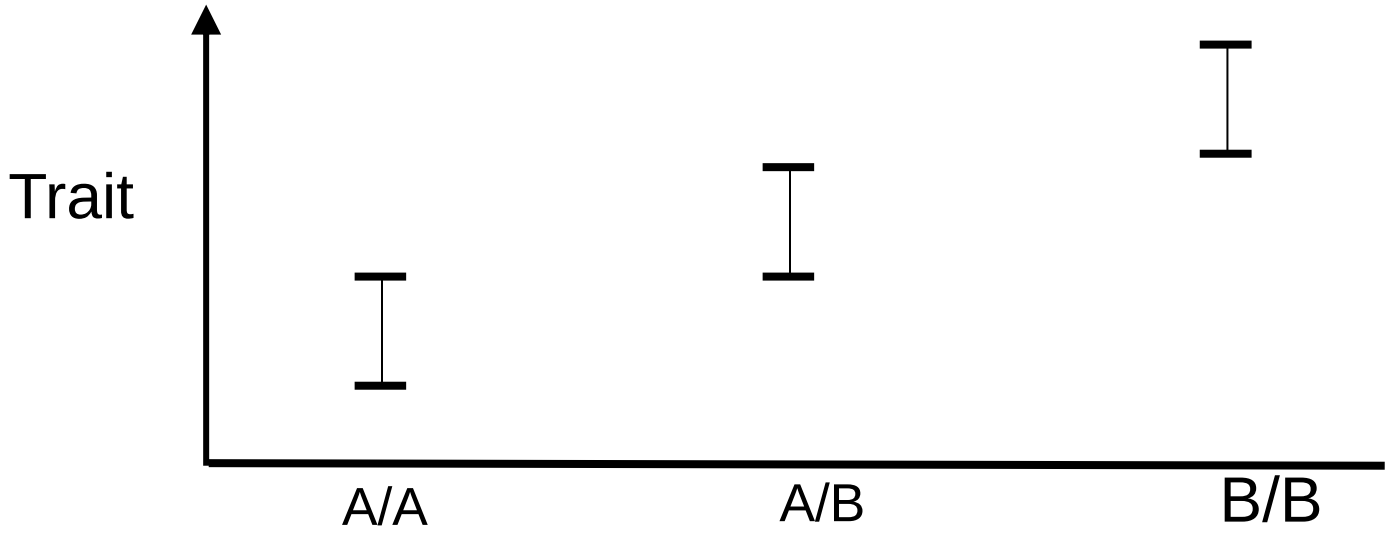
- Case-control
  - Compare genotype and/or allele frequencies

# Matching cases and controls

- Match for
  - Age
  - Sex
  - Ethnicity
  - Environmental exposures
- Difficulty is recruiting appropriate controls
  - Random digit dialing, voters lists

| | A/A | A/B | B/B |
|---|---|---|---|
| Cases | a | b | c |
| Controls | d | e | f |

Cochran-Armitage trend test (1df chisq)




ANOVA or non-parametric-test (Kruskal-Wallis), linear regression


feature

# The International HapMap Project

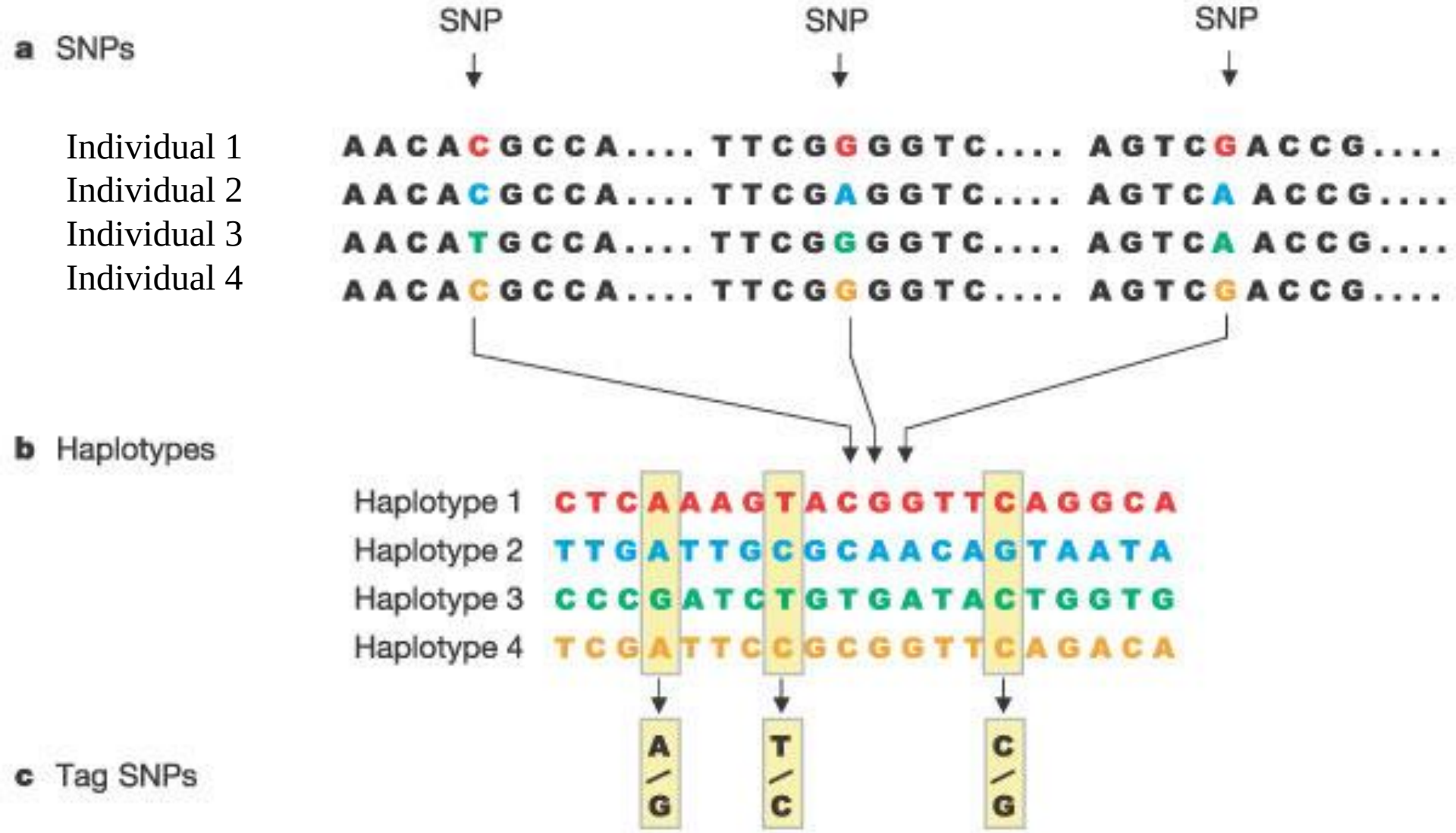
**The International HapMap Consortium***

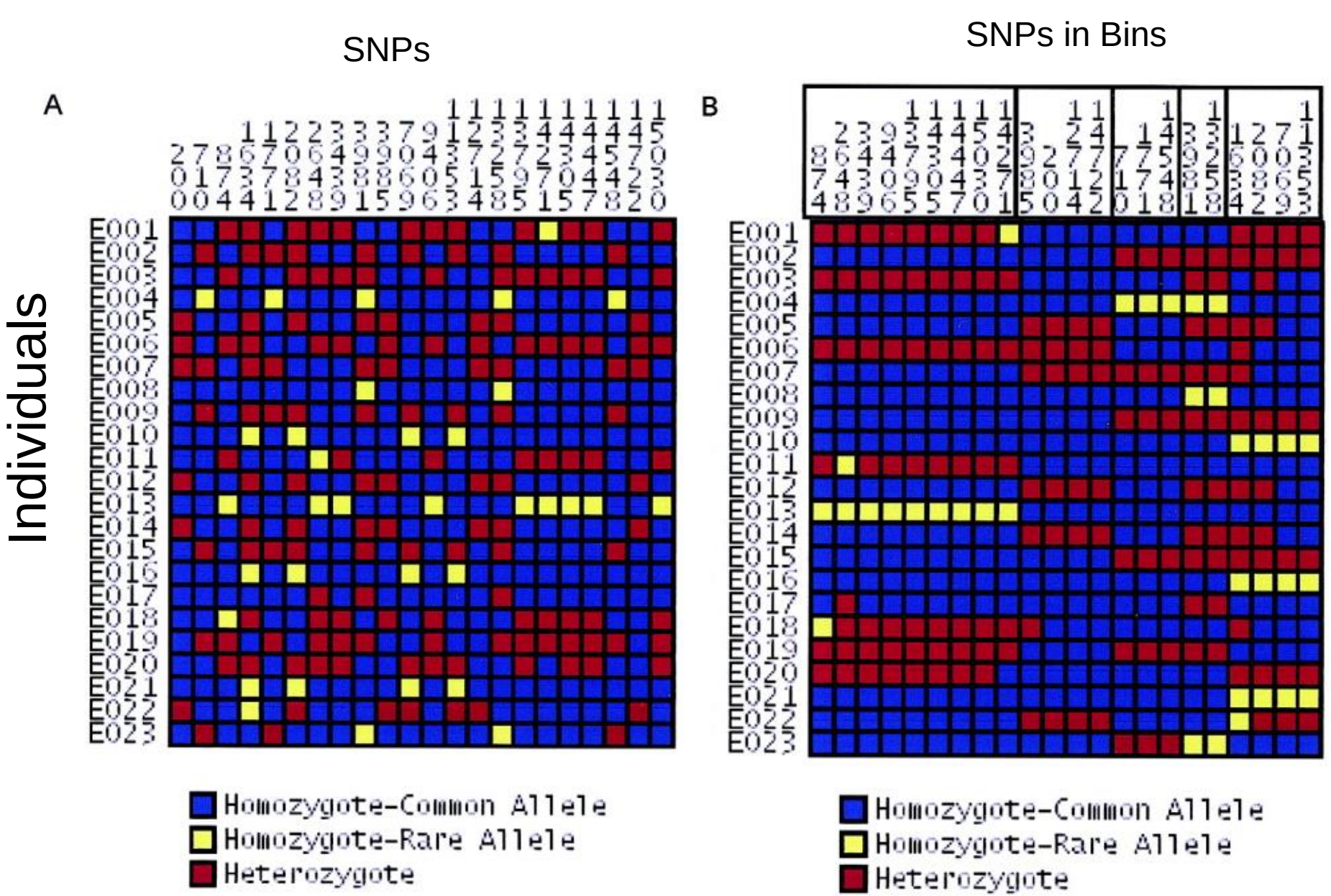
*Lists of participants and affiliations appear at the end of the paper

**The goal of the International HapMap Project is to determine the common patterns of DNA sequence variation in the human genome and to make this information freely available in the public domain. An international consortium is developing a map of these patterns across the genome by determining the genotypes of one million or more sequence variants, their frequencies and the degree of association between them, in DNA samples from populations with ancestry from parts of Africa, Asia and Europe. The HapMap will allow the discovery of sequence variants that affect common disease, will facilitate development of diagnostic tools, and will enhance our ability to choose targets for therapeutic intervention.**

# Haplotype map (HapMap)

- Recognition that genotypes at markers that are physically close are correlated (associated) in a population
- ~ 15M common variations in human genome
- Majority of common genetic variation can be assayed by a smaller set of markers (~500K)
- 28th data release (Aug 2010): 4.0M SNPs pass QC
- http://hapmap.ncbi.nlm.nih.gov/
- *Nature* 426, 789-796. 2003
  - *Nature* 437, 1299-1320. 2005
  - *Nature* 449, 851-861. 2007


a SNPs
SNP
SNP
SNP
Individual 1
Individual 2
Individual 3
Individual 4
AACACGCCA.... TTCGGGGTC.... AGTCGACCG....
AACACGCCA.... TTCGAGGTC.... AGTCA ACCG....
AACATGCCA.... TTCGGGGTC.... AGTCA ACCG....
AACACGCCA.... TTCGGGGTC.... AGTCGACCG....
b Haplotypes
Haplotype 1 CTCAAAGTACGGTTCAGGCA
Haplotype 2 TTGATTGCGCAACAGTAATA
Haplotype 3 CCCGATCTGTGATACTGGTG
Haplotype 4 TCGATTCCGCGGTTCAGACA
c Tag SNPs
A/G
T/C
C/G


SNPs
SNPs in Bins
A
B
Individuals
Homozygote-Common Allele
Homozygote-Rare Allele
Heterozygote
Homozygote-Common Allele
Homozygote-Rare Allele
Heterozygote

# Genome-wide association analysis

- Allows analysis of the majority of common variation in the human genome for association with a disease or trait
- Large-scale projects, e.g.
  - Wellcome Trust Case-Control Consortium
  - 3K controls, 2K cases each of 7 different diseases
  - 500K SNP chip (Affymetrix)
  - Multiple testing: p<5 x 10-8 considered ‘significant’
    - www.wtccc.org.uk, Nature 2007
- Off-the-shelf panels (~1M SNPs)
  - Affymetrix SNP6.0 / Axiom platform
  - Illumina Omni-express, -2.5M, -5M, and HumanCore (Exome)

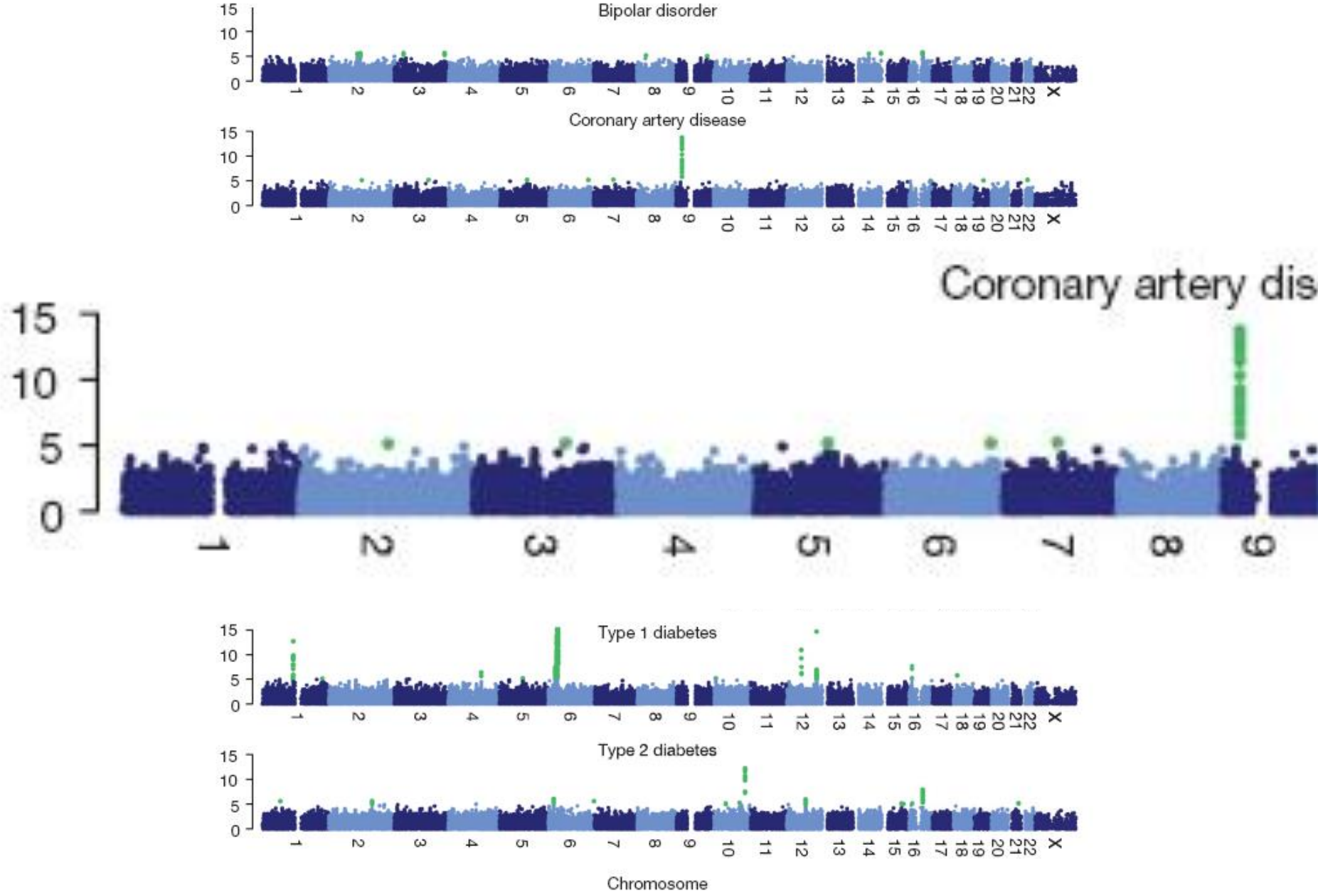



**Figure 4 | Genome-wide scan for seven diseases.** For each of seven diseases $-\log_{10}$ of the trend test *P* value for quality-control-positive SNPs, excluding those in each disease that were excluded for having poor clustering after visual inspection, are plotted against position on each chromosome. Chromosomes are shown in alternating colours for clarity, with *P* values $<1 \times 10^{-5}$ highlighted in green. All panels are truncated at $-\log_{10}(P\text{value}) = 15$, although some markers (for example, in the MHC in T1D and RA) exceed this significance threshold.

# Region plots

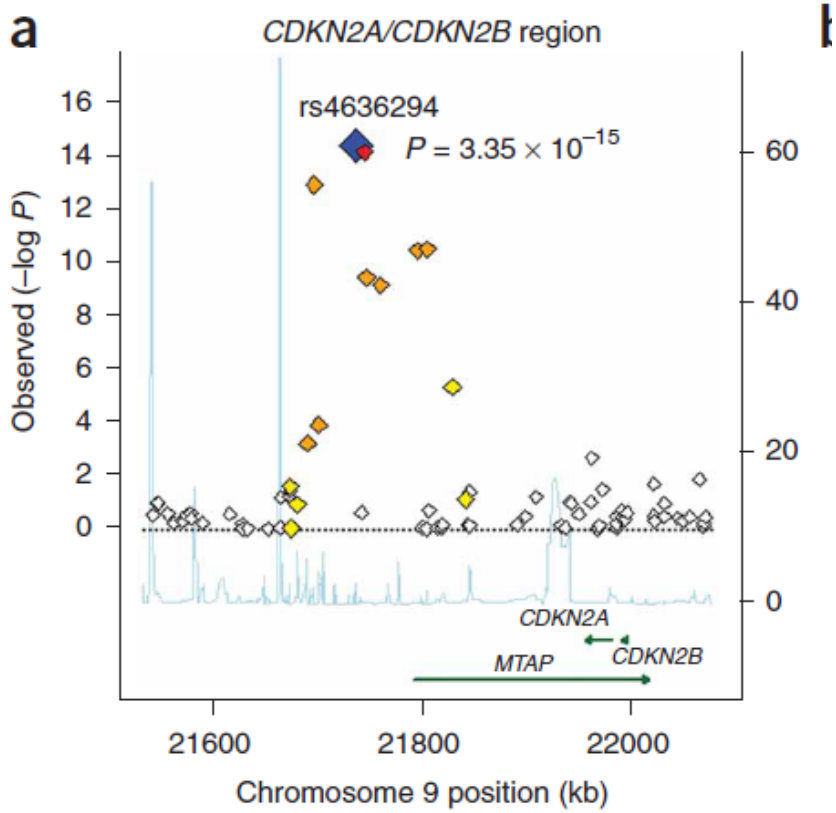



**Figure 2** Regional plots. (**a**) Chromosome 9p21. (**b**) Chromosome 22q13. Meta analysis $\log_{10}$ *P* values are plotted as a function of genomic position (build 36). The *P* values for the lead SNPs are denoted by large blue (combined discovery and replication) diamonds. Proxies are indicated with diamonds of smaller size, with colors assigned based on the pairwise $r^2$ values with the lead SNP in the HapMap CEU sample: red ($r^2 > 0.8$), orange ($0.5 < r^2 < 0.8$) or yellow ($0.2 < r^2 < 0.5$). White indicates either no LD with the lead SNP ($r^2 < 0.2$) or loci where such information was not available. Recombination rate estimates (HapMap Phase II) are given in light blue, RefSeq genes (NCBI) in green.

# rs17696736, 12q24 (C12orf30) associated with type 1 diabetes

Case-control:

| Allele | Cases | Controls | OR (95%CI) | P value |
|---|---|---|---|---|
| A | 5,981 (51.5%) | 7,083 (56.5%) | 1.00 (ref) | 1.7 x 10-13 |
| G | 5,637 (48.5%) | 5,451 (43.5%) | 1.22 (1.15-1.28) | |

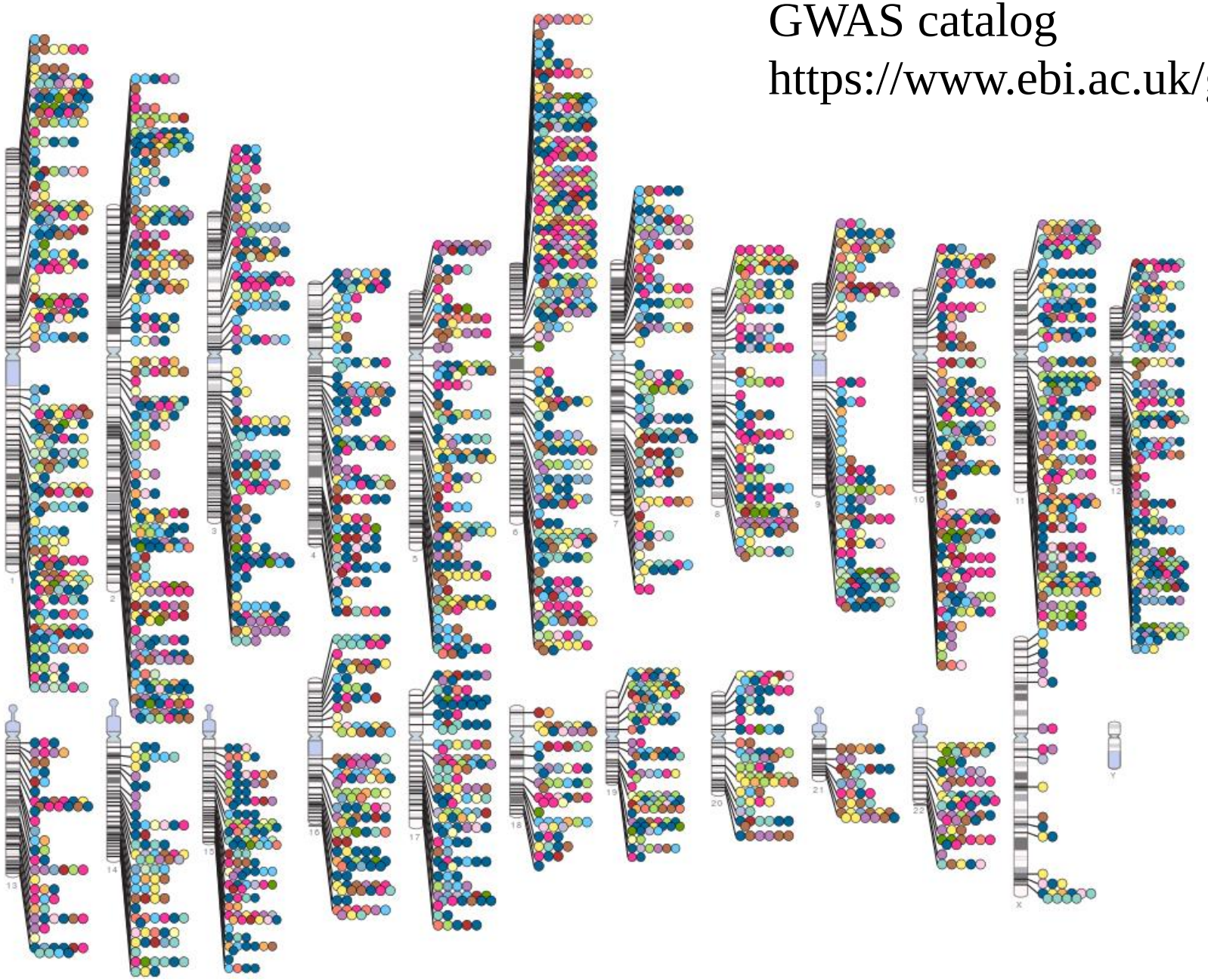
| Genotype | Cases | Controls | OR (95%CI) | P value |
|---|---|---|---|---|
| A/A | 1,545 (26.6%) | 1,984 (31.7%) | 1.00 (ref) | 9.5 x 10-13 |
| A/G | 2,891 (49.8%) | 3,115 (49.7%) | 1.17 (1.08-1.28) | |
| G/G | 1, 373 (23.6%) | 1,168 (18.6%) | 1.48 (1.34-1.65) | |

Todd et al., 2007 Nature Genetics

# NHGRI-EBI GWAS catalogue

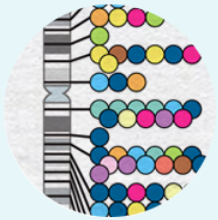
- https://www.ebi.ac.uk/gwas/
- List of top SNPs (typically P<10-8) for any disease or trait in humans
- Searchable by disease, trait, gene, SNP, chromosomal region
- As of 2021-06-08, the GWAS Catalog contains 5,106 publications and 258,738 associations

GWAS catalog
https://www.ebi.ac.uk/gwas/



# Find a trait/disease in GWAS catalog

- Something a trait/disease that you are interested in
- Each tell us about tomorrow (2 mins max)
  - Include why the trait is important to you
- Two options
  - 1.
    - https://www.ebi.ac.uk/gwas/docs/file-downloads
    - Download ‘all associations v1’
      - ~135 Mb tsv
      - Open in excel
      - Browse
      - Column H is ‘phenotype/trait’
  - 2. Use search box online



# GWAS Catalog

The NHGRI-EBI Catalog of human genome-wide association studies

Search the catalog

Examples: breast carcinoma, rs7329174, Yao, 2q37.1, HBS1L, 6:16000000-25000000





## Downloading the GWAS Catalog

| Description | Download Link | Format | Column header descriptions |
|---|---|---|---|
| All associations **v1.0** | Click to download | tab separated file | Click to view |
| All associations **v1.0.2 - with added ontology annotations, GWAS Catalog study accession numbers and genotyping technology** | Click to download | tab separated file | Click to view |

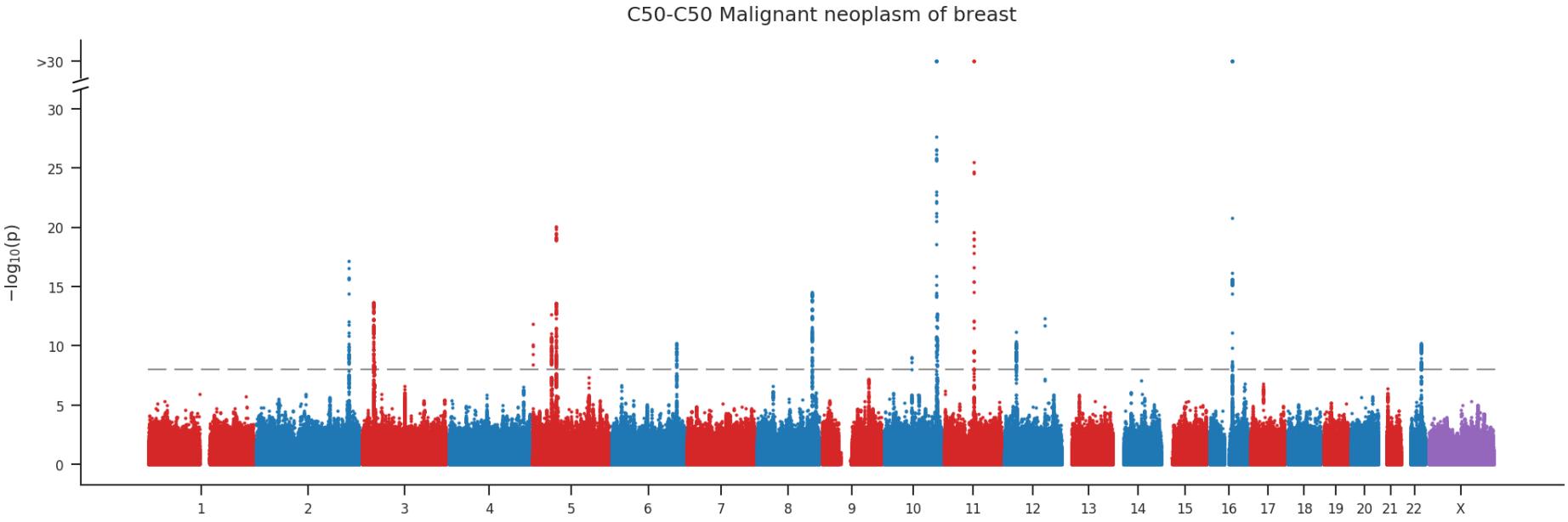
# UK biobank

- Aged 40-69 years at recruitment, n~ 500K
- Rich phenotypic and health-related information
  - Self-report demographics, diet, exercise
  - Physical and cognitive measures
  - Medical records & cancer registries
  - Blood and urine biomarkers
  - Imaging of brain, heart, bones, carotid arteries, abdominal fat in ~100K
- GWAS genotyping ~805K markers
- ~96M variants imputed
- Tools for efficient GWAS & PheWAS
  - standing height
- Raw data publicly available to all scientists (small cost)

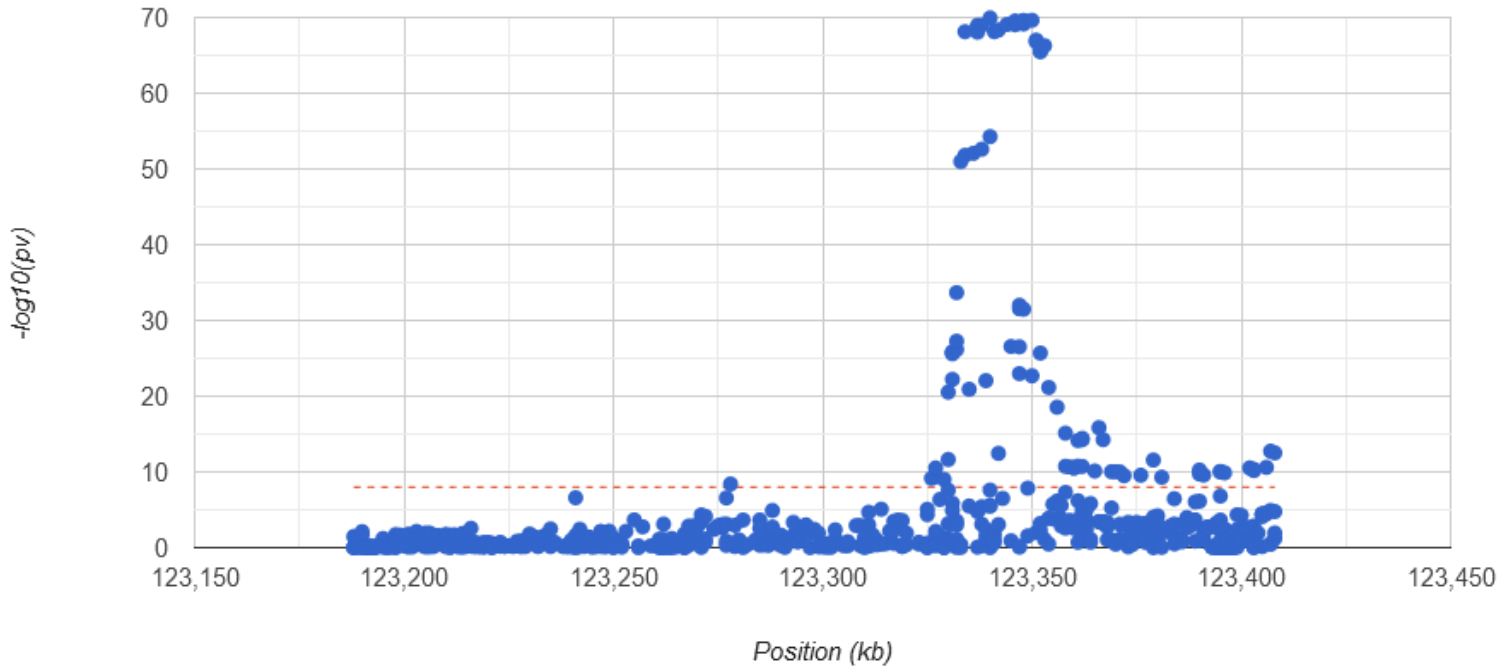
UKBB summary results

http://geneatlas.roslin.ed.ac.uk/

9.1 M variants that pass QC
Females:
10,478 cases
235,016 controls

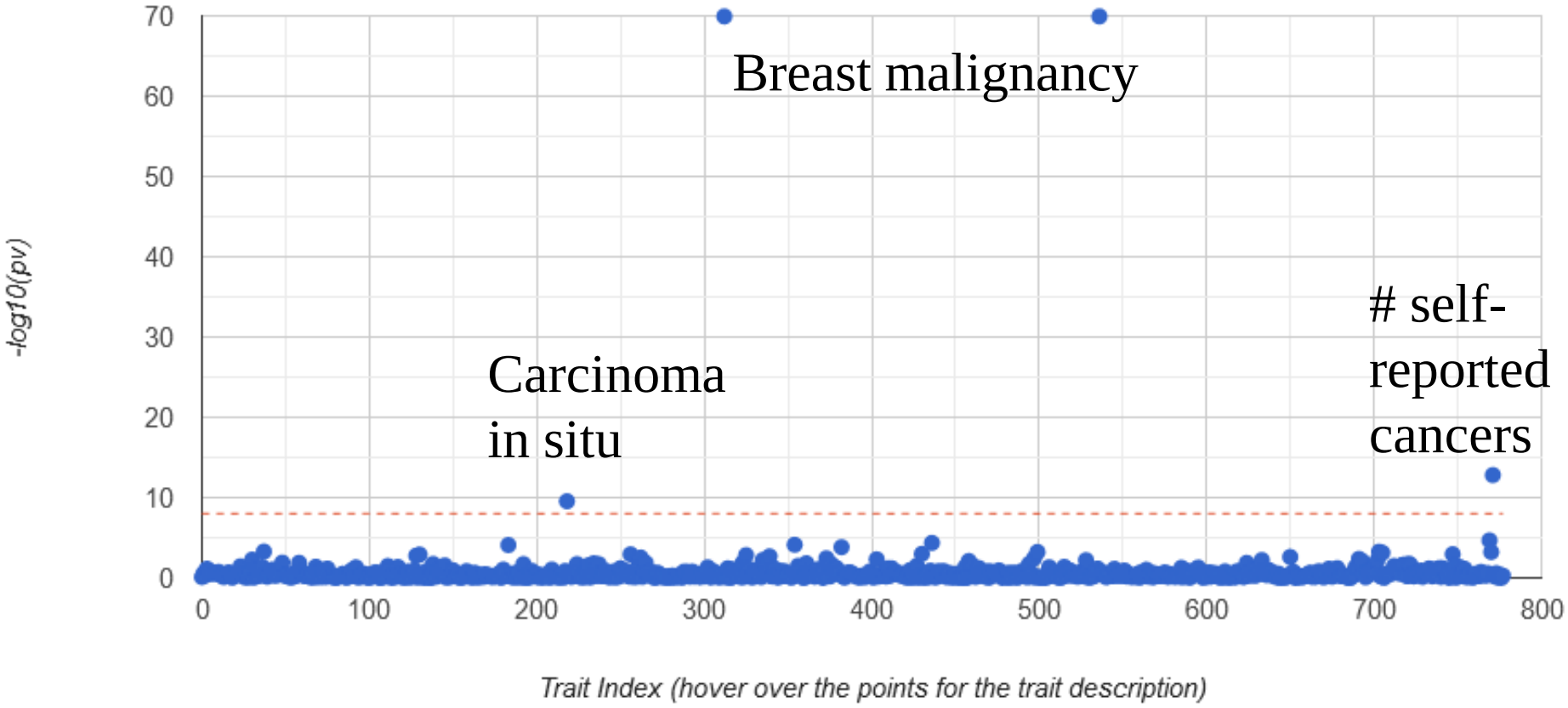


C50-C50 Malignant neoplasm of breast



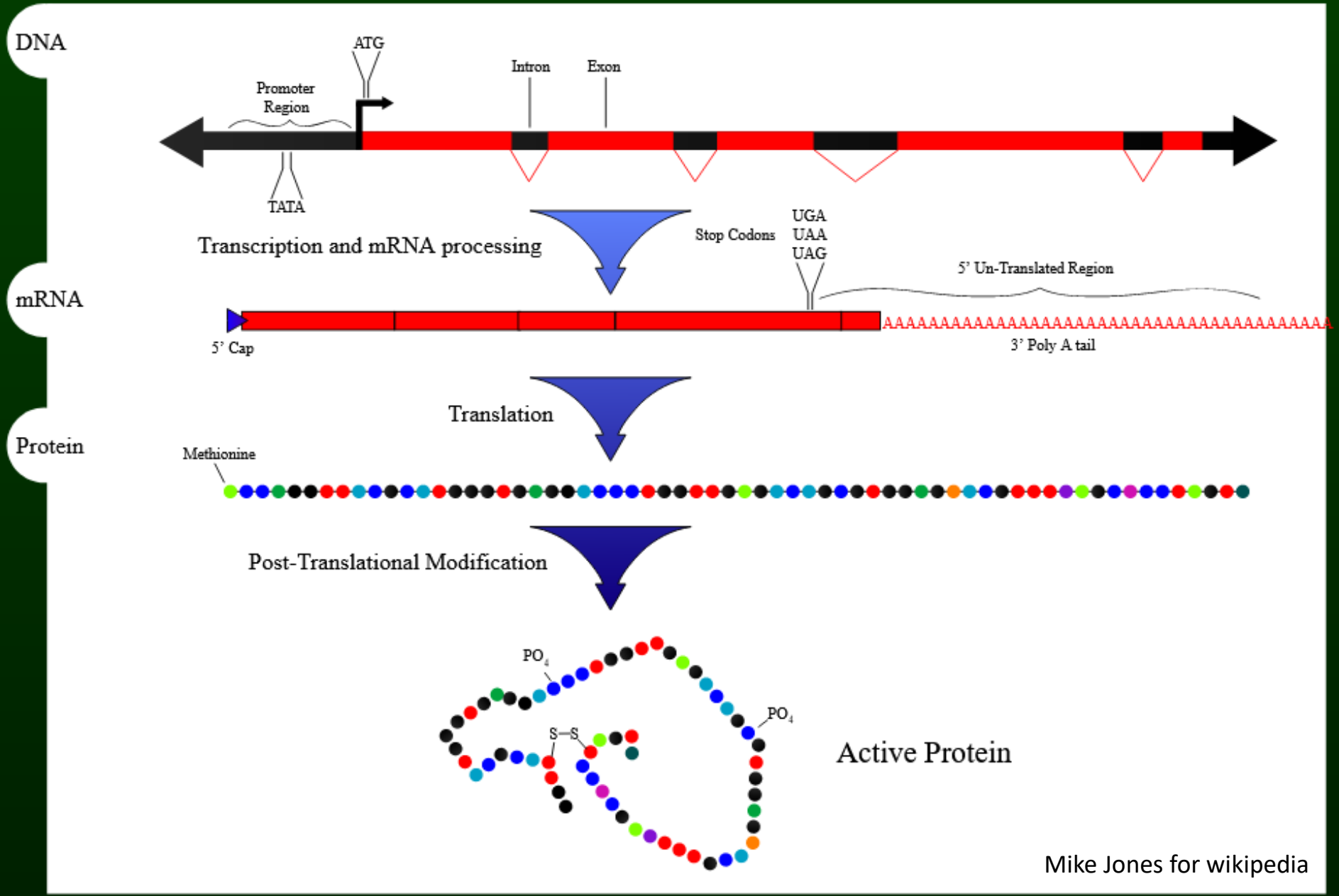

Chr 11, around FGFR2

PheWAS of top SNP for malignant breast cancer (rs11599804, FGFR2
Across 700 phenotypes



# ERAP2 details

# Central Dogma of Molecular Biology : Eukaryotic Model




Mike Jones for wikipedia

https://www.ncbi.nlm.nih.gov/gene/64167

Search

# ERAP2 endoplasmic reticulum aminopeptidase 2 [ *Homo sapiens* (human) ]

Download Datasets

Gene ID: 64167, updated on 11-Jun-2021

## Summary

**Official Symbol** ERAP2 provided by HGNC
**Official Full Name** endoplasmic reticulum aminopeptidase 2 provided by HGNC
**Primary source** HGNC:HGNC:29499
**See related** Ensembl:ENSG00000164308 MIM:609497
**Gene type** protein coding
**RefSeq status** REVIEWED
**Organism** Homo sapiens
**Lineage** Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo
**Also known as** LRAP; L-RAP
**Summary** This gene encodes a zinc metalloaminopeptidase of the M1 protease family that resides in the endoplasmic reticulum and functions in N-terminal trimming antigenic epitopes for presentation by major histocompatibility complex (MHC) class I molecules. Certain mutations in this gene are associated with the inflammatory arthritis syndrome ankylosing spondylitis and pre-eclampsia. This gene is located adjacent to a closely related aminopeptidase gene on chromosome 5. [provided by RefSeq, Jul 2016]
**Expression** Broad expression in lymph node (RPKM 22.9), spleen (RPKM 19.0) and 23 other tissues See more
**Orthologs** all

NEW **Try the new Gene table**
**Try the new Transcript table**

## Genomic context

**Location:** 5q15

See ERAP2 in Genome Data Viewer

**Exon count:** 19

| Annotation release | Status | Assembly | Chr | Location |
|---|---|---|---|---|
| 109.20210514 | current | GRCh38.p13 (GCF_000001405.39) | 5 | NC_000005.10 (96875939..96919716) |
| 105.20201022 | previous assembly | GRCh37.p13 (GCF_000001405.25) | 5 | NC_000005.9 (96211643..96255420) |

Chromosome 5 - NC_000005.10

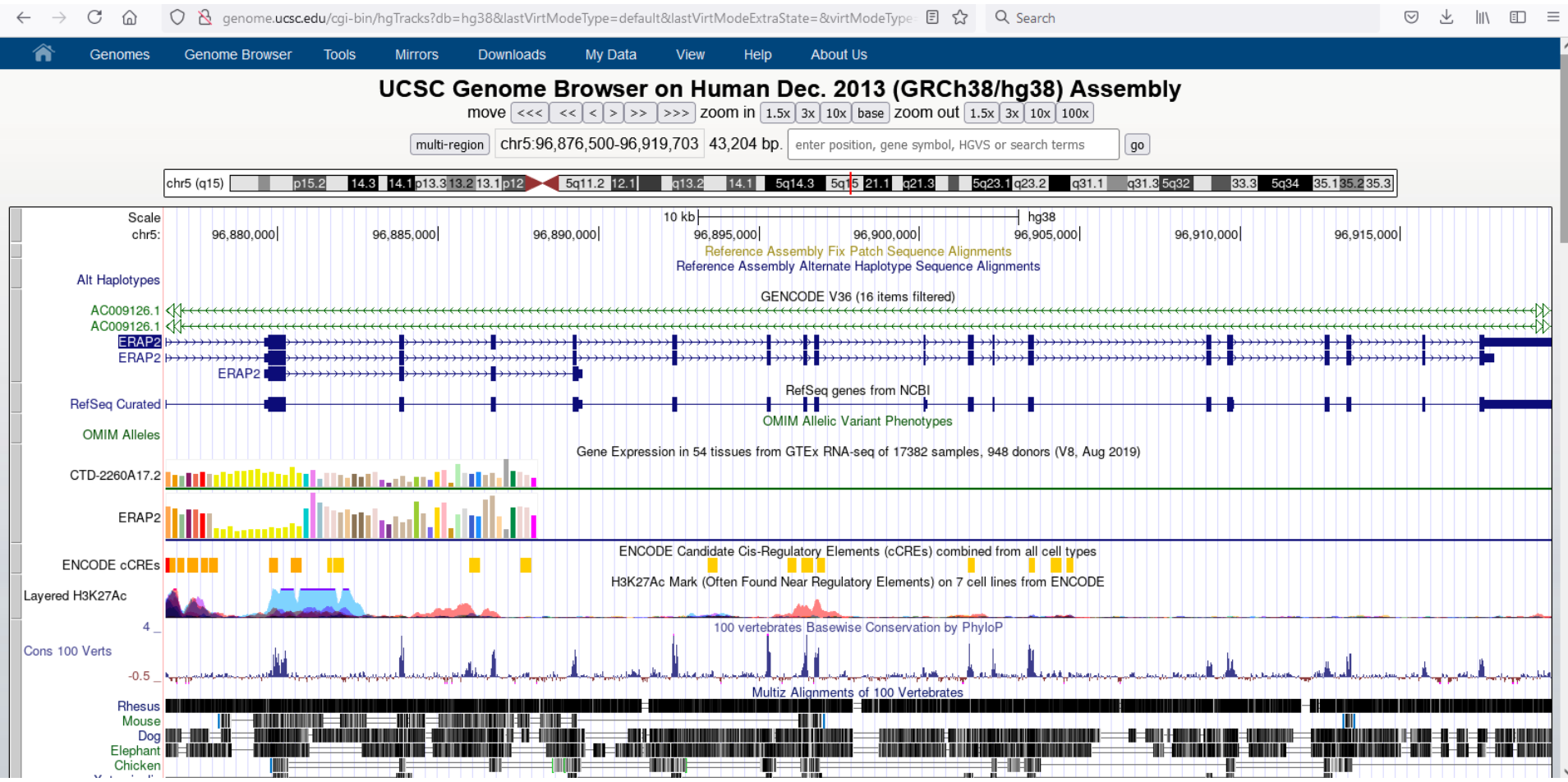
Table of contents



Genome Browsers



Related information


genome.ucsc.edu/cgi-bin/hgTracks?db=hg38&lastVirtModeType=default&lastVirtModeExtraState=&virtModeType
Search
Genomes
Genome Browser
Tools
Mirrors
Downloads
My Data
View
Help
About Us
UCSC Genome Browser on Human Dec. 2013 (GRCh38/hg38) Assembly
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x 100x
multi-region
chr5:96,876,500-96,919,703
43,204 bp.
enter position, gene symbol, HGVS or search terms
go
chr5 (q15)
Scale
chr5:
10 kb
hg38
96,880,000
96,885,000
96,890,000
96,895,000
96,900,000
96,905,000
96,910,000
96,915,000
Reference Assembly Fix Patch Sequence Alignments
Reference Assembly Alternate Haplotype Sequence Alignments
Alt Haplotypes
GENCODE V36 (16 items filtered)
AC009126.1
AC009126.1
ERAP2
ERAP2
ERAP2
RefSeq Curated
RefSeq genes from NCBI
OMIM Alleles
OMIM Allelic Variant Phenotypes
Gene Expression in 54 tissues from GTEx RNA-seq of 17382 samples, 948 donors (V8, Aug 2019)
CTD-2260A17.2
ERAP2
ENCODE cCREs
ENCODE Candidate Cis-Regulatory Elements (cCREs) combined from all cell types
Layered H3K27Ac
H3K27Ac Mark (Often Found Near Regulatory Elements) on 7 cell lines from ENCODE
4
Cons 100 Verts
-0.5
100 vertebrates Basewise Conservation by PhyloP
Multiz Alignments of 100 Vertebrates
Rhesus
Mouse
Dog
Elephant
Chicken

https://gtexportal.org/home/gene/ERAP2

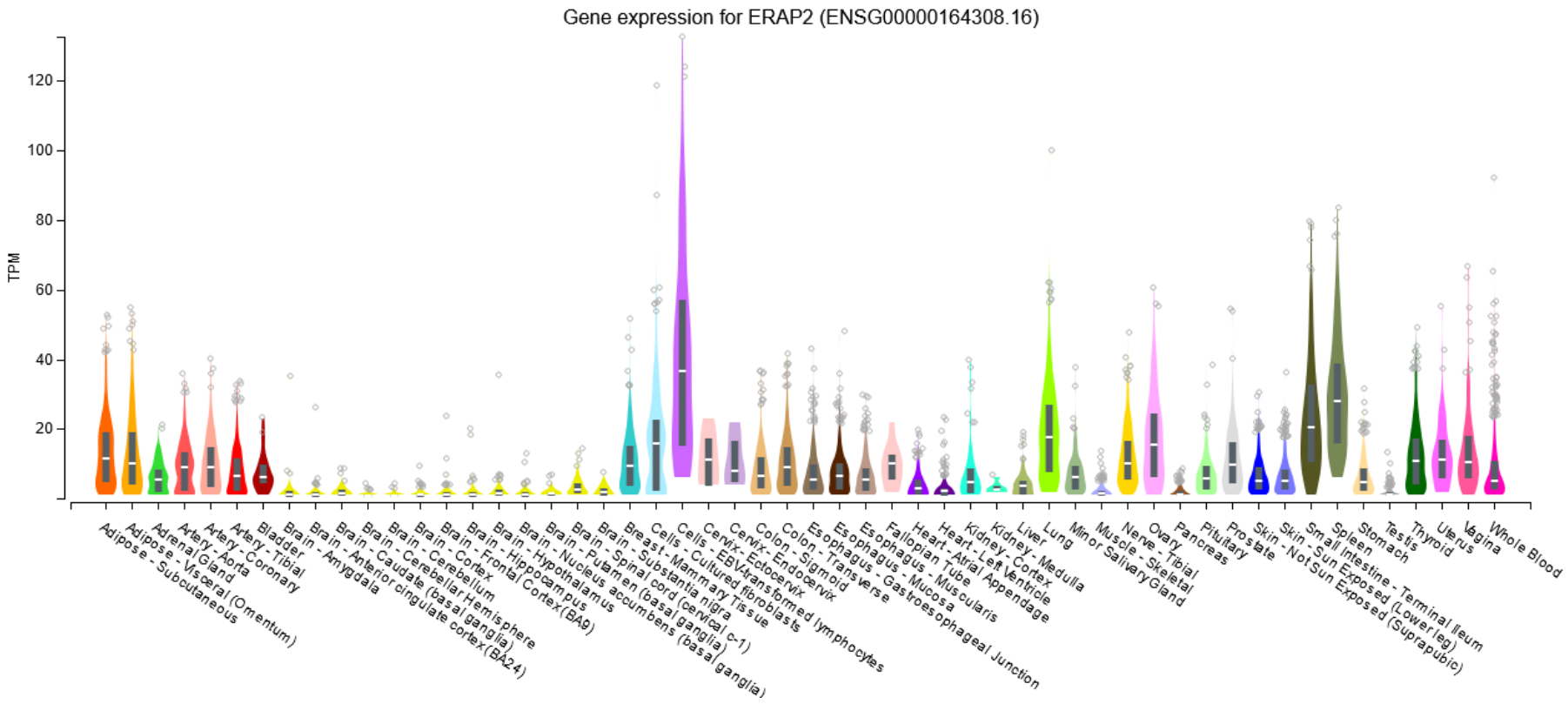
Search

## Gene expression for ERAP2 (ENSG00000164308.16)

Data Source: GTEx Analysis Release V8 (dbGaP Accession phs000424.v8.p2)

Data processing and normalization

SUBSET None Sex SCALE Log Linear TISSUE SORT MEDIAN SORT OUTLIERS On Off



https://gtexportal.org/home/gene/ERAP2

Significant Single-Tissue e

Data Source: GTEx Analysis Releas

Views QTLs of ERAP2 in the Locus

Copy CSV

Show 10 entries

| Gencode Id | Tissue | Actions |
|---|---|---|
| ENSG00000164308.16 | Adipose - Subcutaneous | eQTL violin plot, IGV Browser, Multi-tissue eQTL Plot |
| ENSG00000164308.16 | Adipose - Subcutaneous | eQTL violin plot, IGV Browser, Multi-tissue eQTL Plot |
| ENSG00000164308.16 | Whole Blood | eQTL violin plot, IGV Browser, Multi-tissue eQTL Plot |
| ENSG00000164308.16 | Whole Blood | eQTL violin plot, IGV Browser, Multi-tissue eQTL Plot |
| ENSG00000164308.16 | Muscle - Skeletal | eQTL violin plot, IGV Browser, Multi-tissue eQTL Plot |
| | | eQTL violin plot, IGV Browser, |

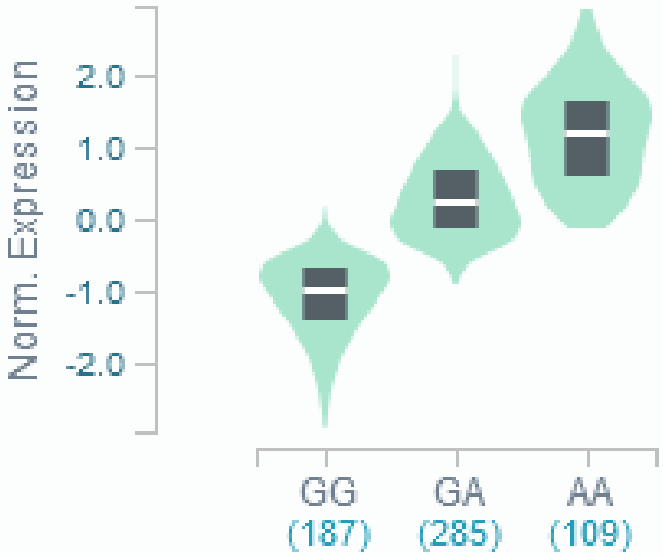
## eQTL Violin Plots

Clear All

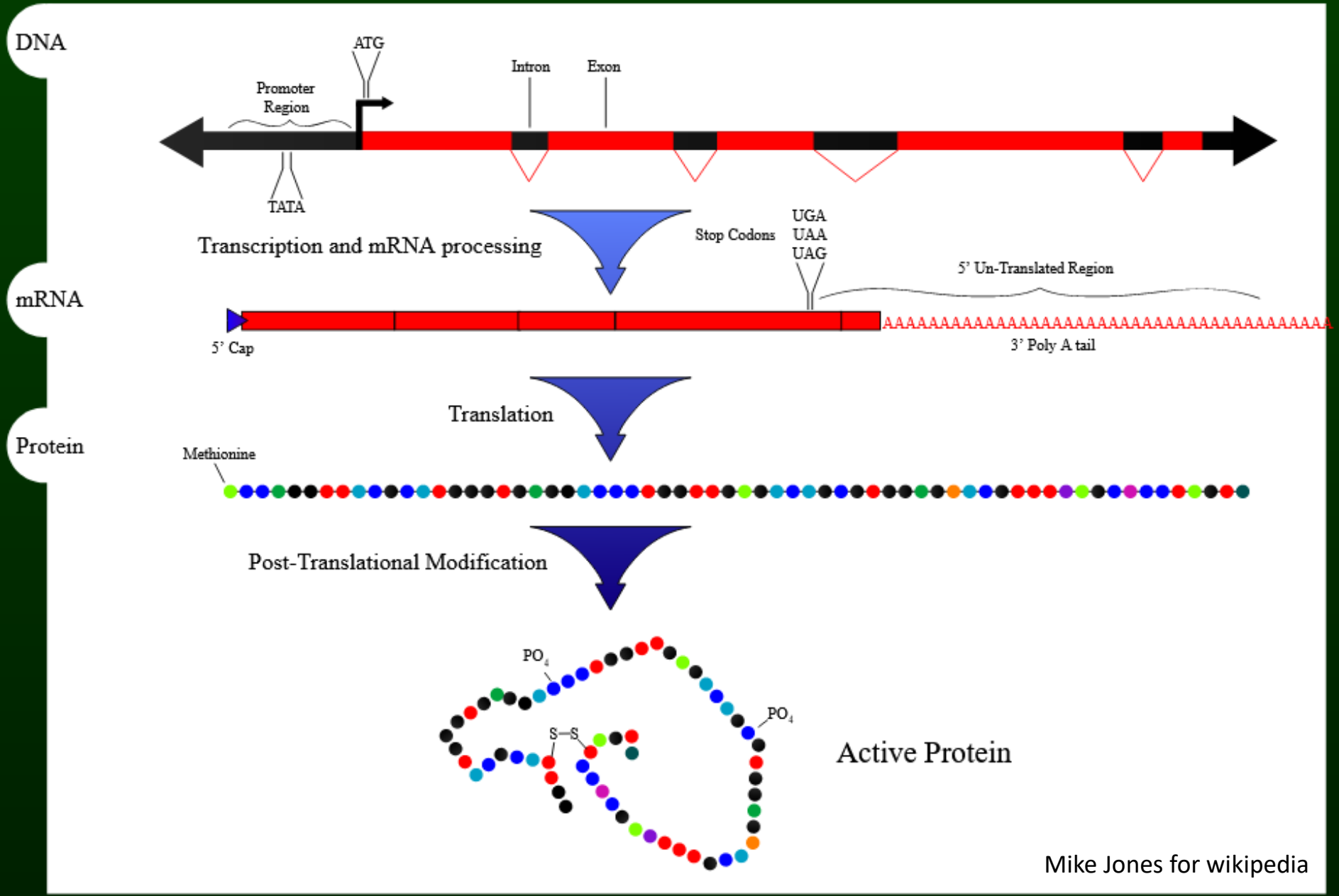
ERAP2

chr5_96916728_G_A_b38

Adipose - Subcutaneous



# Central Dogma of Molecular Biology : Eukaryotic Model




Mike Jones for wikipedia

Article | Published: 06 June 2018

# Genomic atlas of the human plasma proteome

Benjamin B. Sun, Joseph C. Maranville, [...]Adam S. Butterworth

*Nature* **558**, 73–79 (2018) | Cite this article



Does a GWAS of 1478 proteins in blood

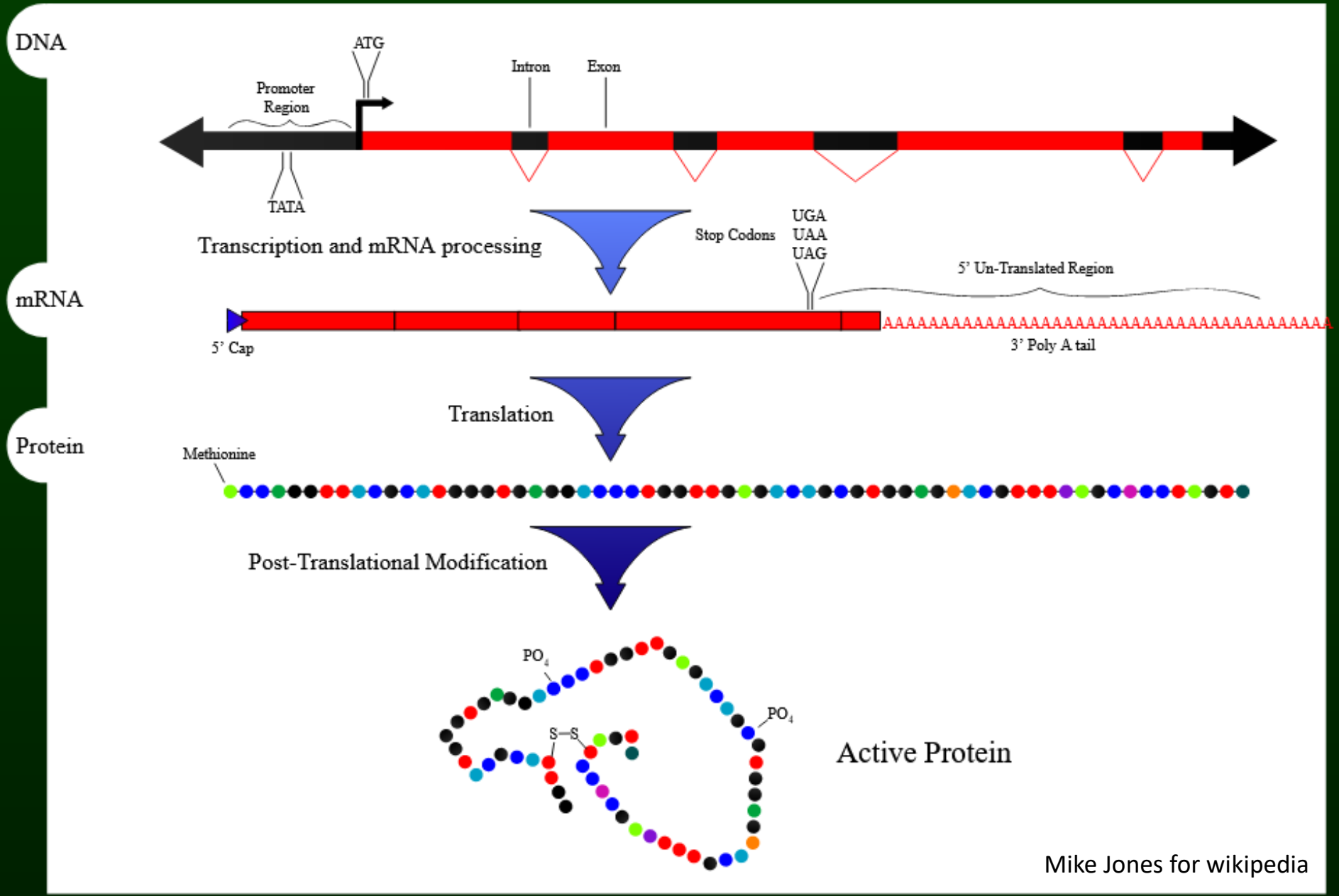
## Abstract

Although plasma proteins have important roles in biological processes and are the direct targets of many drugs, the genetic factors that control inter-individual variation in plasma protein levels are not well understood. Here we characterize the genetic architecture of the human plasma proteome in healthy blood donors from the INTERVAL study. We identify 1,927 genetic associations with 1,478 proteins, a fourfold increase on existing knowledge, including *trans* associations for 1,104 proteins. To understand the consequences of perturbations in plasma protein levels, we apply an integrated approach that links genetic variation with biological pathway, disease, and drug databases. We show that protein quantitative trait loci overlap with gene expression quantitative trait loci, as well as with disease-associated loci, and find evidence that protein biomarkers have causal roles in disease using Mendelian randomization analysis. By linking genetic factors to diseases via specific proteins, our analyses highlight potential therapeutic targets, opportunities for matching existing drugs with new disease indications, and potential safety concerns for drugs under development.

# One of their many results

- Rs2927608 associated with ERPA2 plasma protein level
  - Located in intron 18 of ERPA2
- Discovery cohort p=3 x 10-416
- Replication cohort, p=5 x 10-151
- Meta-analysis p=7 x 10-858

# Central Dogma of Molecular Biology : Eukaryotic Model




Mike Jones for wikipedia

https://www.ebi.ac.uk/gwas/search?query=ERAP2 80% Search

GWAS Catalog Home Diagram Submit Download Documentation About EMBL-EBI NIH National Human Genome Research Institute

Refine search results

- Publications 1
- Variants 11
- Genes 2

*Catalog stats*

- *Last data release on 2021-06-08*
- *5106 publications*
- *161014 SNPs*
- *258738 associations*
- *Genome assembly GRCh38.p13*
- *dbSNP Build 153*
- *Ensembl Build 100*

# Search results for *ERAP2*

## G ERAP2

**Description:** endoplasmic reticulum aminopeptidase 2
**Location:** 5:96875986-96919703 **Cytogenetic region:** 5q15 **Biotype:** protein coding

Associations 20 Studies 17

## P A genome-wide association study identifies a functional ERAP2 haplotype associated with birdshot chorioretinopathy.

Kuiper JJ et al. 2014 Hum Mol Genet PMID:24957906

Associations 3 Studies 1

## V rs1363907

**Location:** 5:96917099 **Cytogenetic region:**5q15 **Most severe consequence:** Intron variant **Mapped gene(s):** ERAP2,AC009126.1

Associations 6 Studies 6

## V rs2549794

**Location:** 5:96908845 **Cytogenetic region:**5q15 **Most severe consequence:** Intron variant **Mapped gene(s):** ERAP2,AC009126.1

Associations 1 Studies 1

## V rs4869313

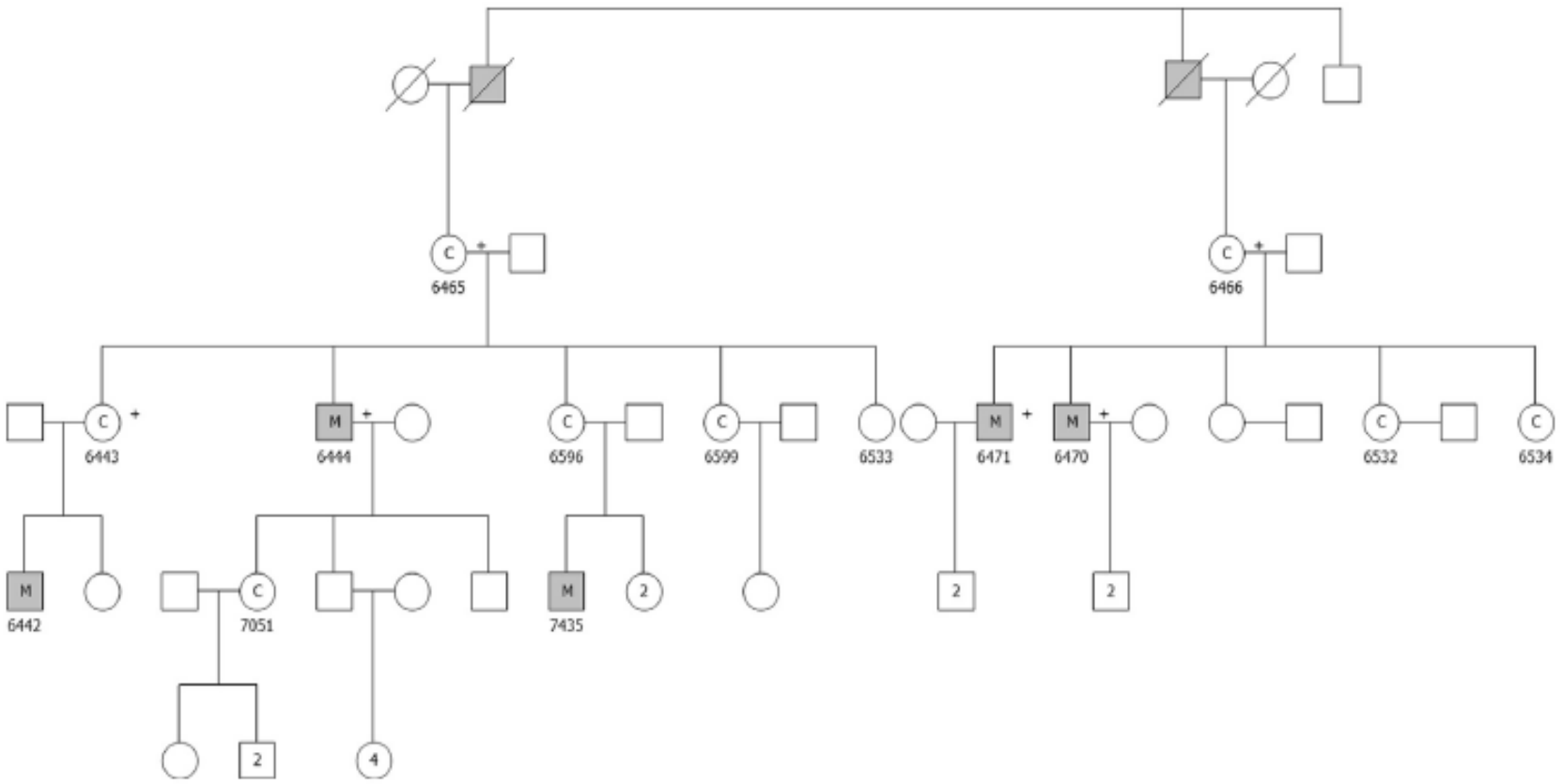



**Figure 1.** Pedigree of X-linked recessive family with unusual glomerulopathy, pathology, and genetic analysis. Individuals for whom DNA was available are represented with a 4-digit number. The proband is 6442. Affected individuals are shown in gray. Whole-exome sequencing was performed in individuals 6442 and 6444, revealing a shared rare *COL4A5* variant, c.T665G (p.Phe222Cys). Sanger sequencing of the *COL4A5* region was done in all individuals for whom DNA was available. Males hemizygous for the variant are denoted with an "M" (mutation), while females heterozygous for the variant are denoted with a "C" (carrier). Individuals for whom whole-genome genotyping was performed for linkage analysis are indicated with a "+".

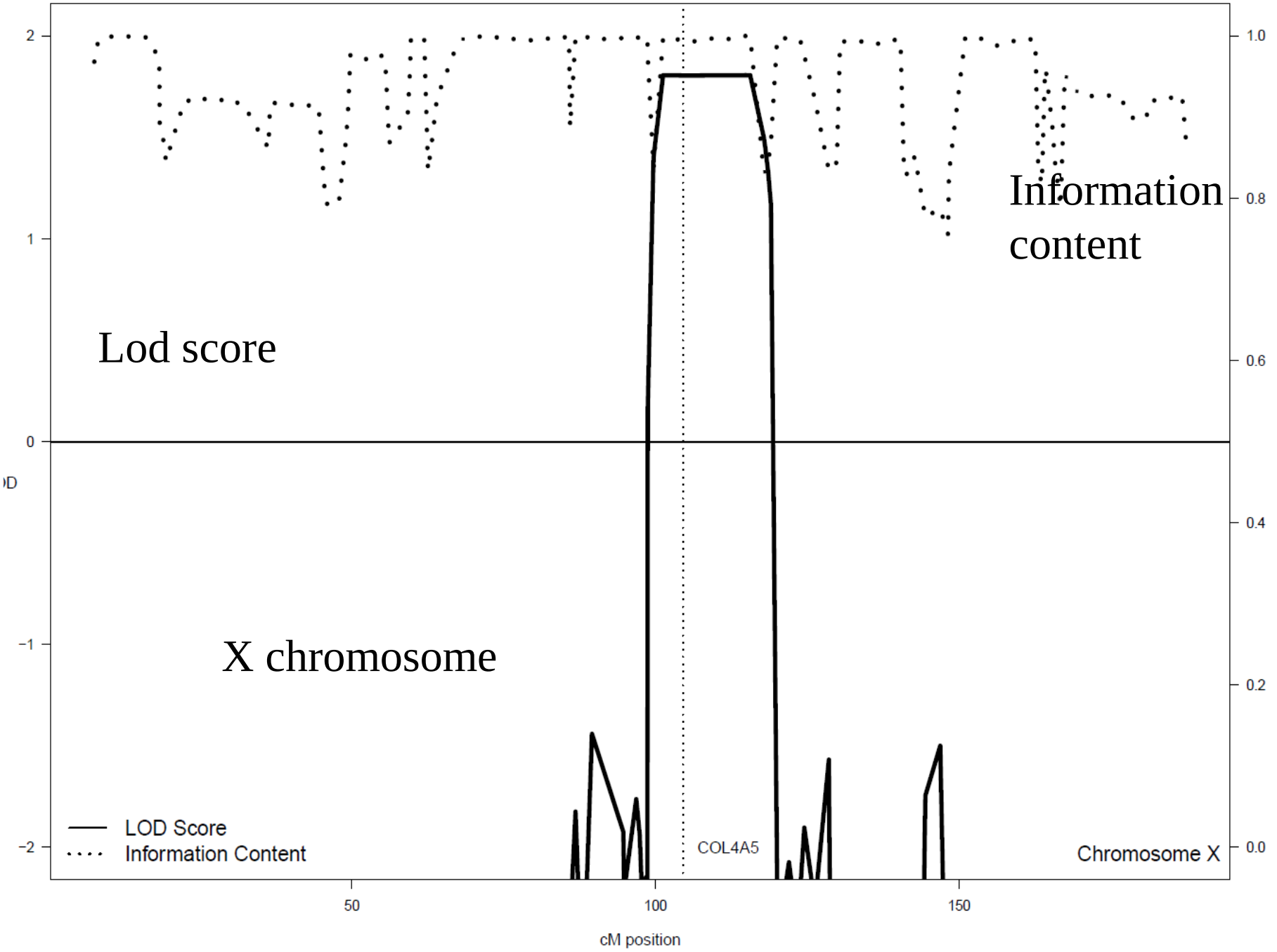



Figure S1: Linkage analysis. Analysis revealed a region on the X-chromosome from 90,000,000 to 118,700,000 (hg19 build 37) spanning 14.35 cM with a peak LOD score of 1.8.

# questions

- how can Ab have specific for different peptides (e.g. SARS-Cov2 spike protein, COVID-19) if encoded by the static genome?

- BRCA1, how many women with variants that increase risk have inherited them maternally vs paternally?

- what are implications for screening based on FHx. Ashkenazi Jewish founder mutation (variants), like Patrick mentioned

- poll