

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

Lei's course, 14 Sept 2021

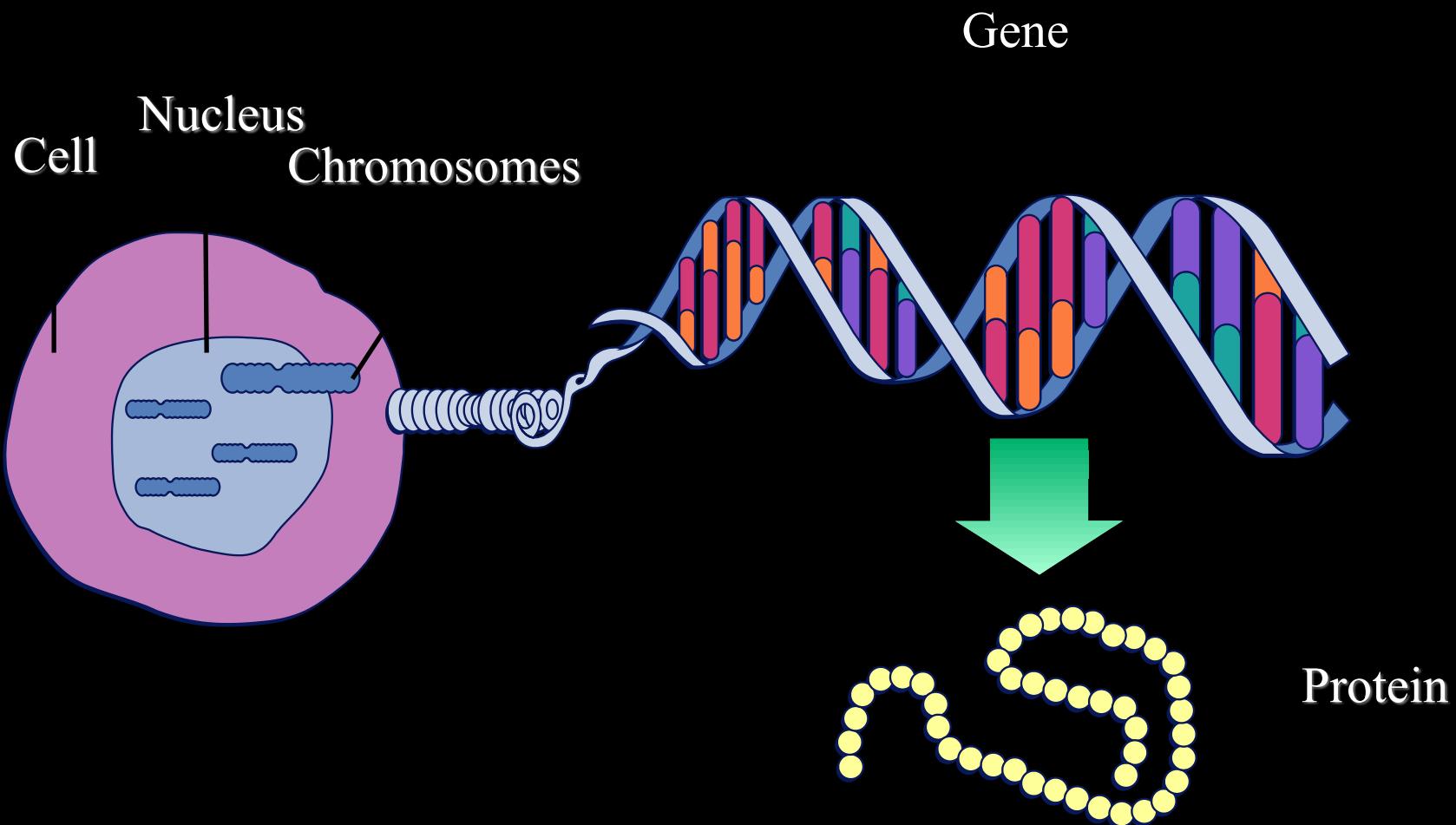
General comments

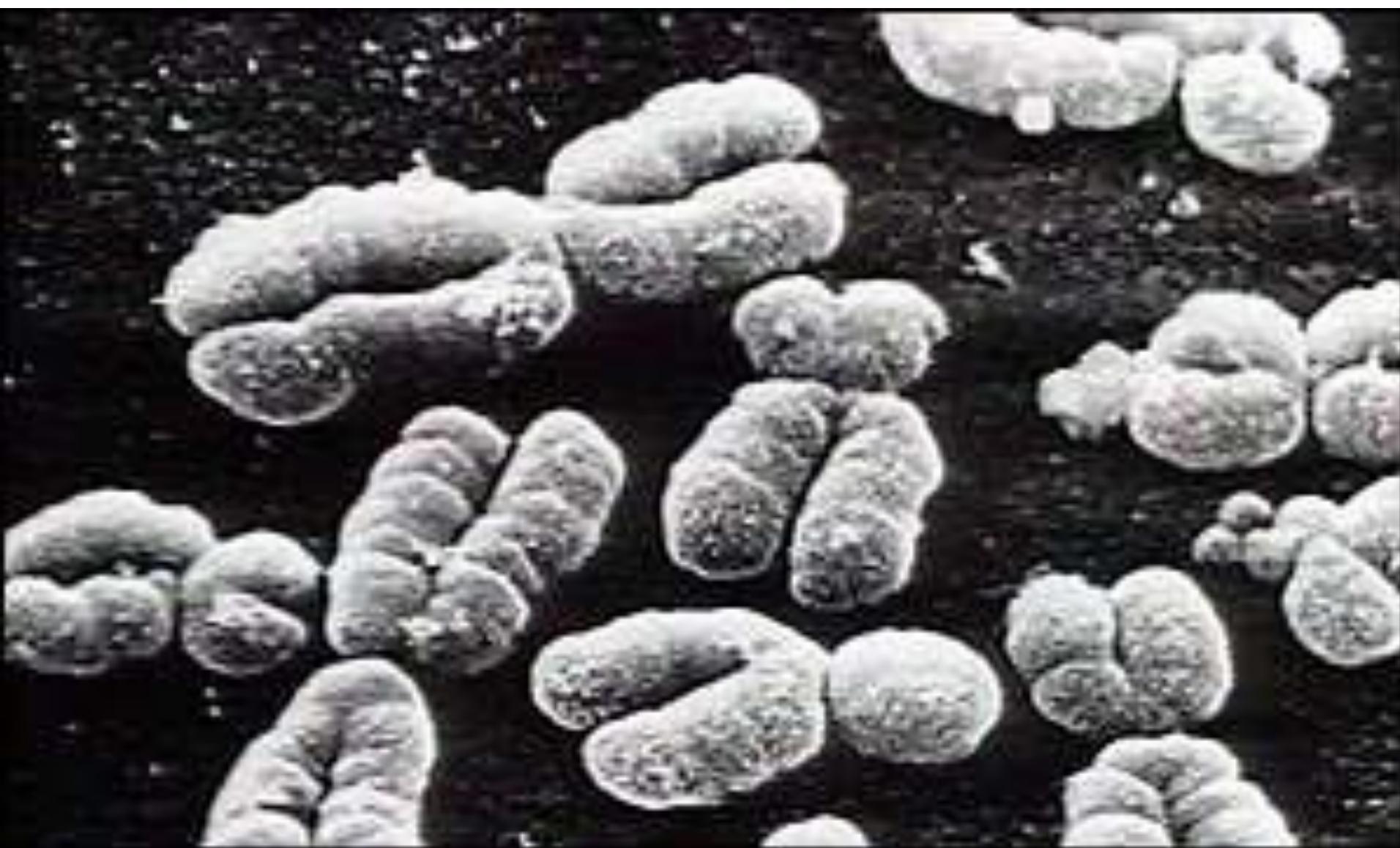
- There are no stupid questions
- Better to have a good understanding of a few basics rather ... than superficial understanding of many things

Human genome

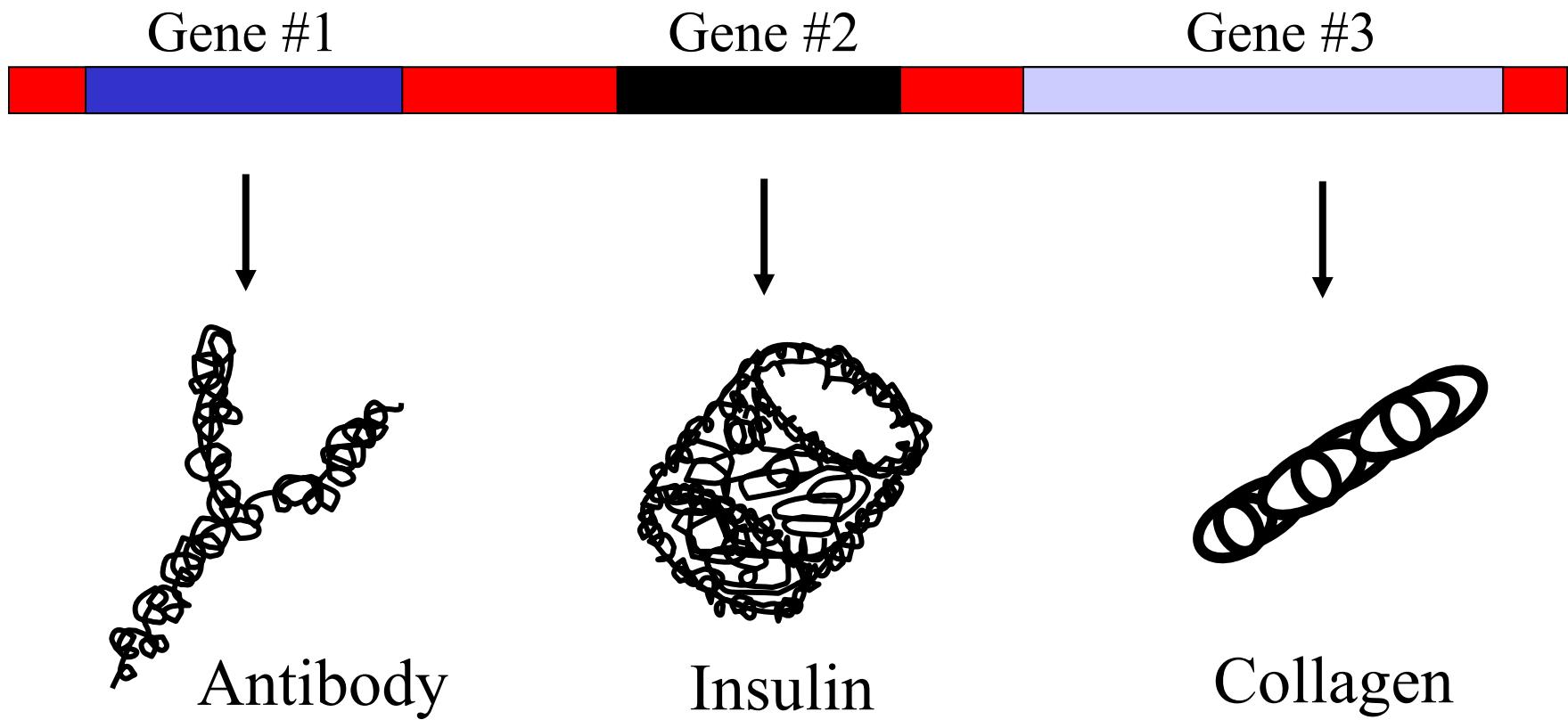
- Each person has 23 pairs of chromosomes
 - One copy of each chromosomes from mother and father
- Each chromosome contains a very long stretch of DNA made up of four different nucleotides (letters, A,T,C,G)
- Males have one X and one Y chromosome
 - Rare XXY
- Females have two X chromosomes
 - Rarely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

Chromosomes, DNA and Genes

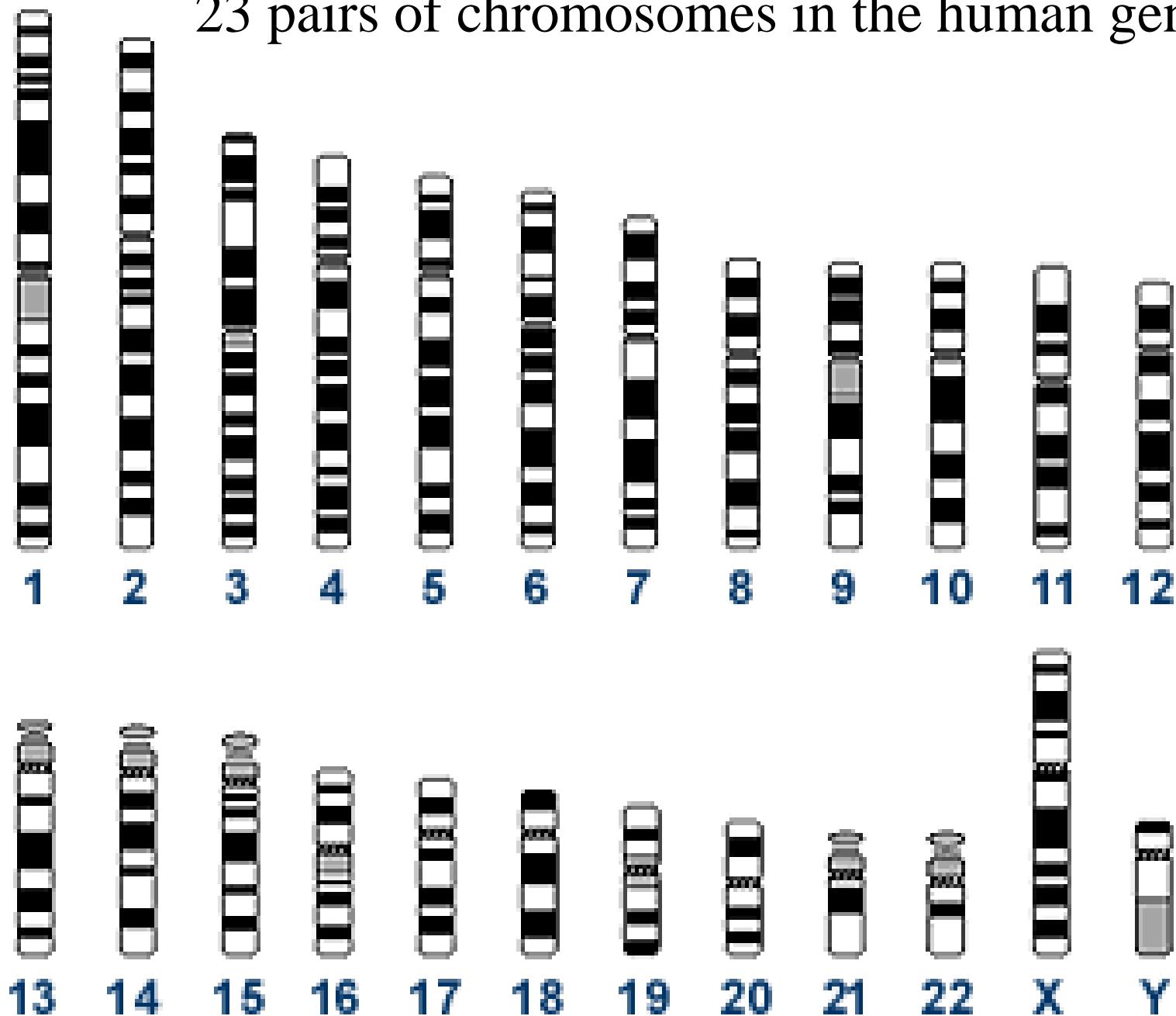




A Gene is a segment of DNA that makes a protein...



23 pairs of chromosomes in the human genome



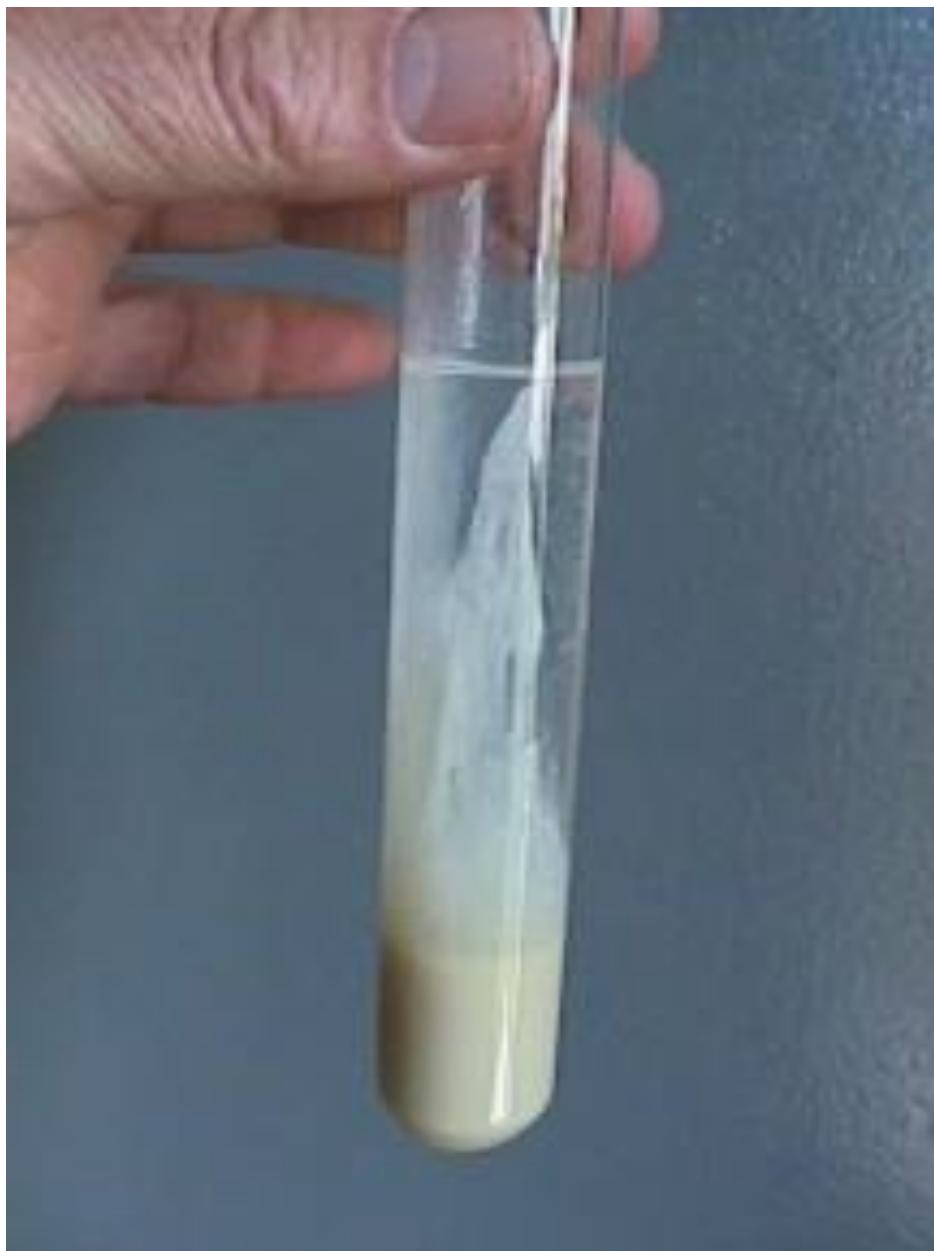
Q: Do all humans really have two copies of each chromosome?

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
 - *BRCA1*: 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/>



A G[A]G T T[C]T G C[T]C G
A G[G]G T T[A]T G C[G]C G

C G T T[C]G G[G]A A T C C
C G T T[A]G G[G]A A A T C T

T C T T[T]G A[C]G A C T C
T C T T[A]G A[G]G A C T C

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 #	118	150	151	152	154	155
date	Nov 2003	Feb 2017	Mar 2018	Dec 2018	June 2020	April 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

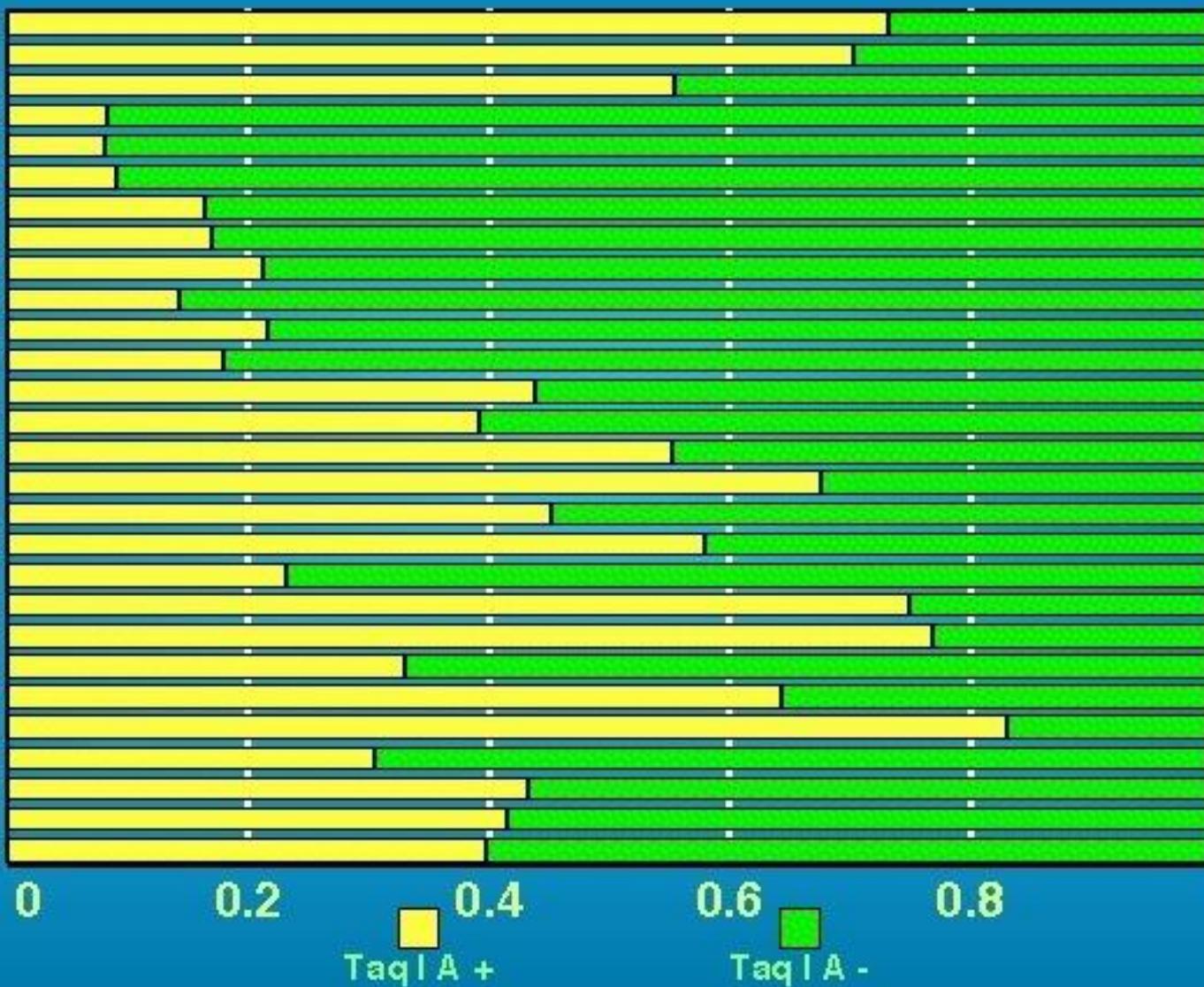
- Direct counting of unordered genotypes and alleles

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

Bialka
Mbuti
Yoruba
Ethiopians
Yemenites
Druze
Adygei
Russians
Fins
Danes
Irish
Europeans
Chinese (SF)
Chinese (TW)
Japanese
Aini
Atayal
Cambodians
Yakut
Nasioi
Cheyenne
J. Pueblo
Pima (AZ)
Pima (MEX)
Maya
Ticuna
R. Surui
Kantiana



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 Chi² test

Testing for departure from Hardy Weinberg Equilibrium

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

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

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

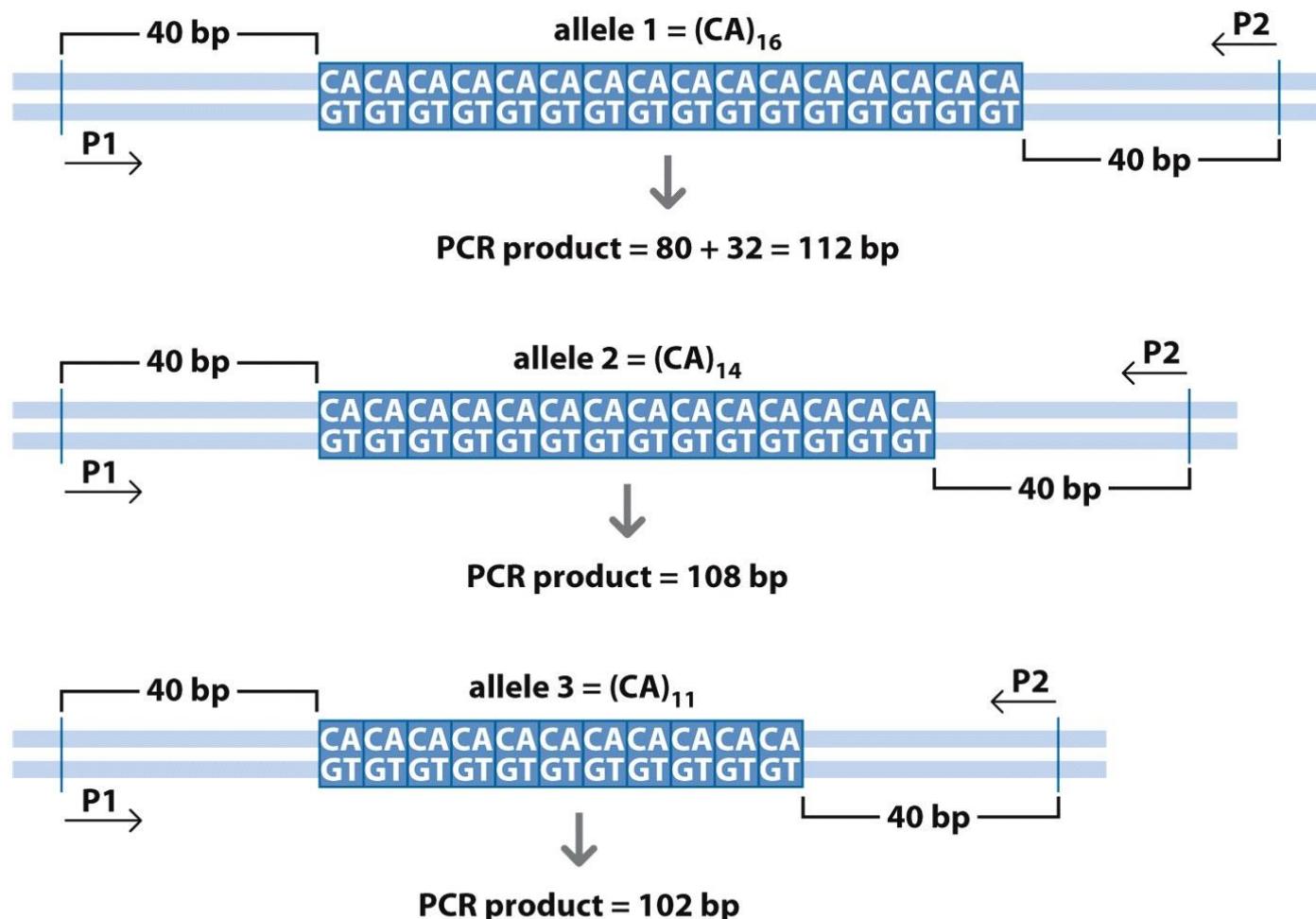


Figure 8.15 Human Molecular Genetics, 4ed. (© Garland Science)

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/>

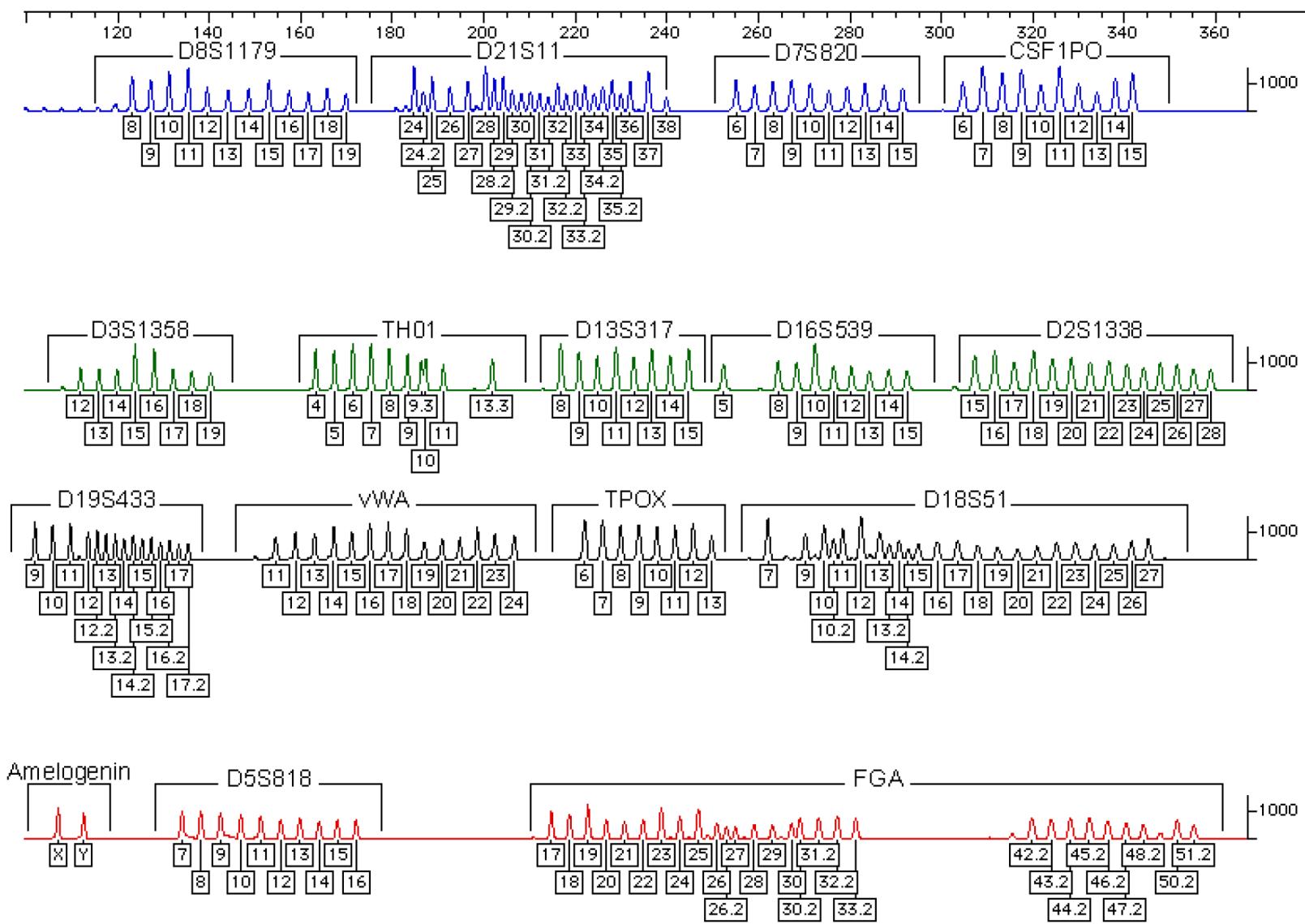
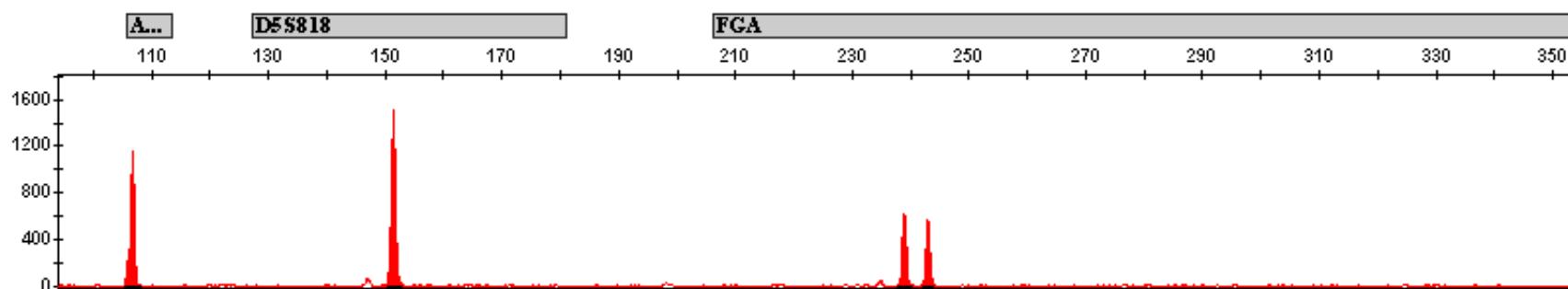
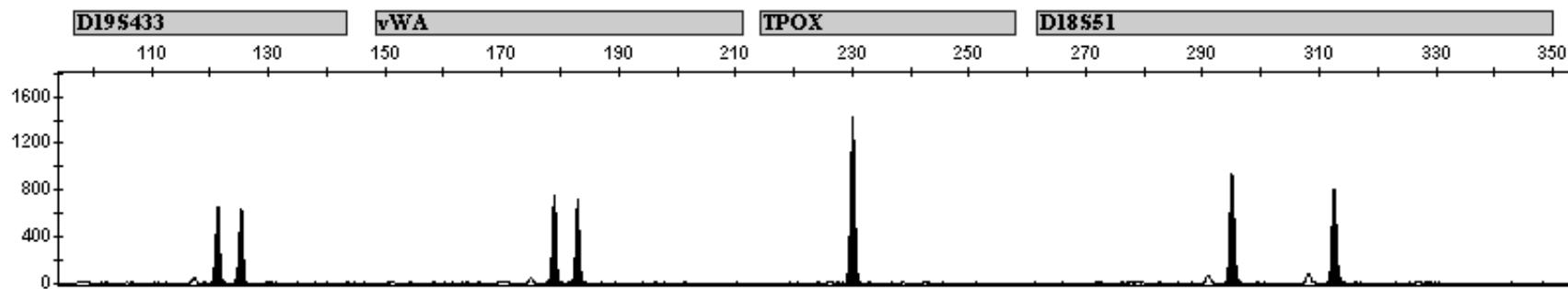
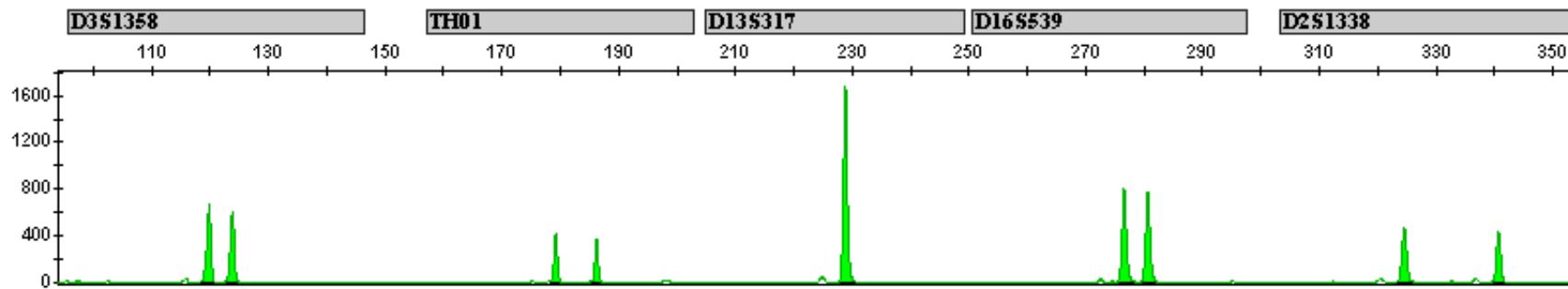
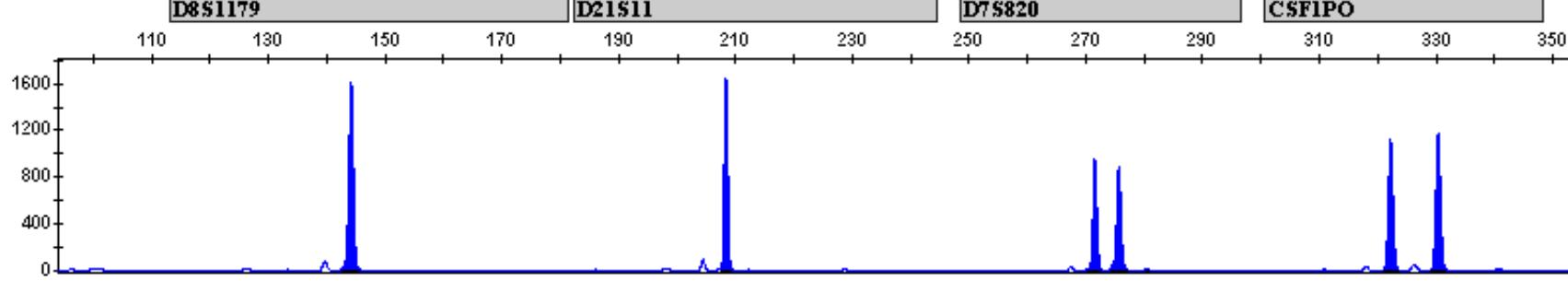


Figure 5-1 Genotyper® software plot of the AmpFℓSTR Identifiler Allelic Ladder, indicating the designation for each allele. These results were obtained on an ABI PRISM 310 Genetic Analyzer

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



3 Genetic Topics

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

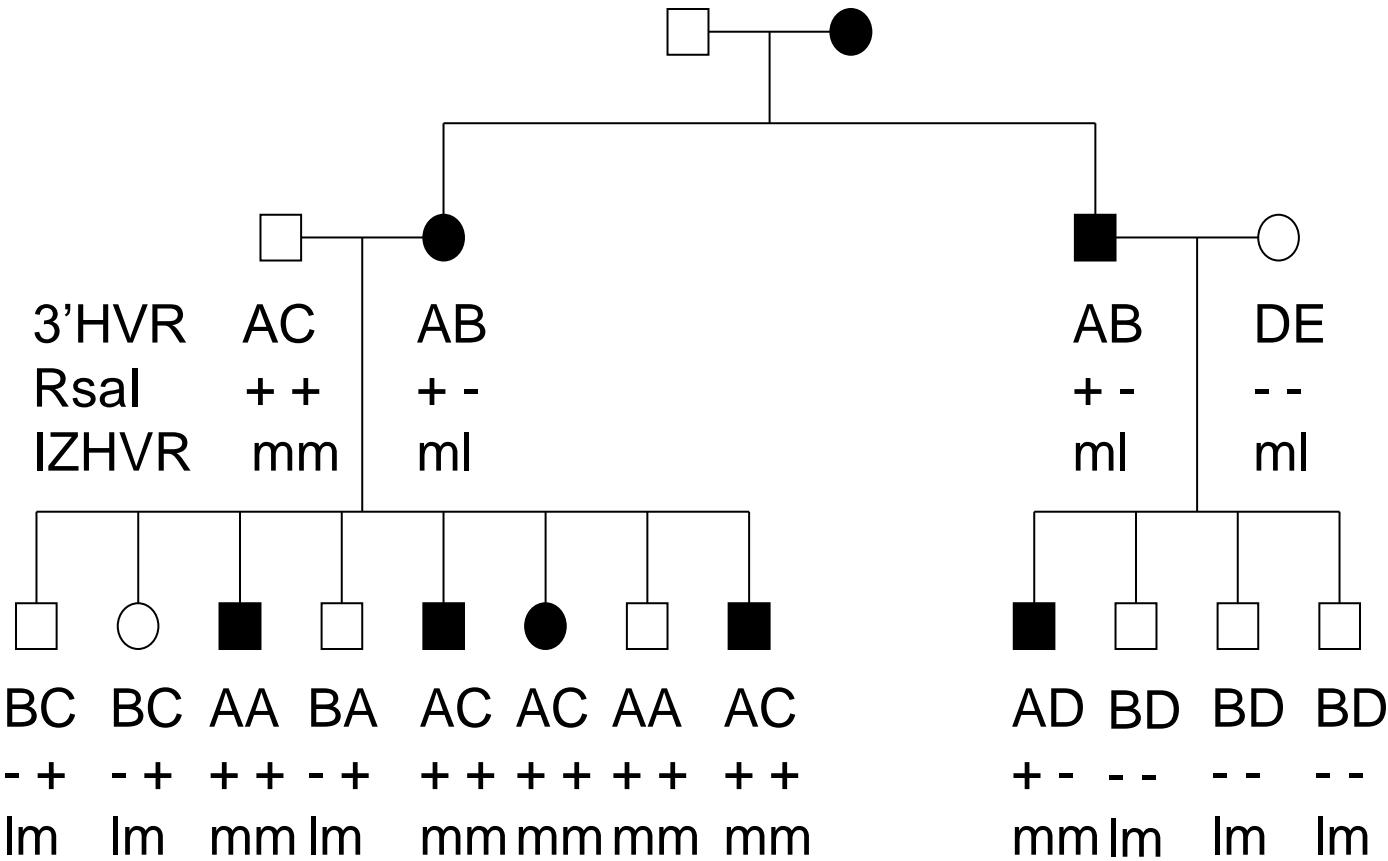
Genetic linkage

- 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

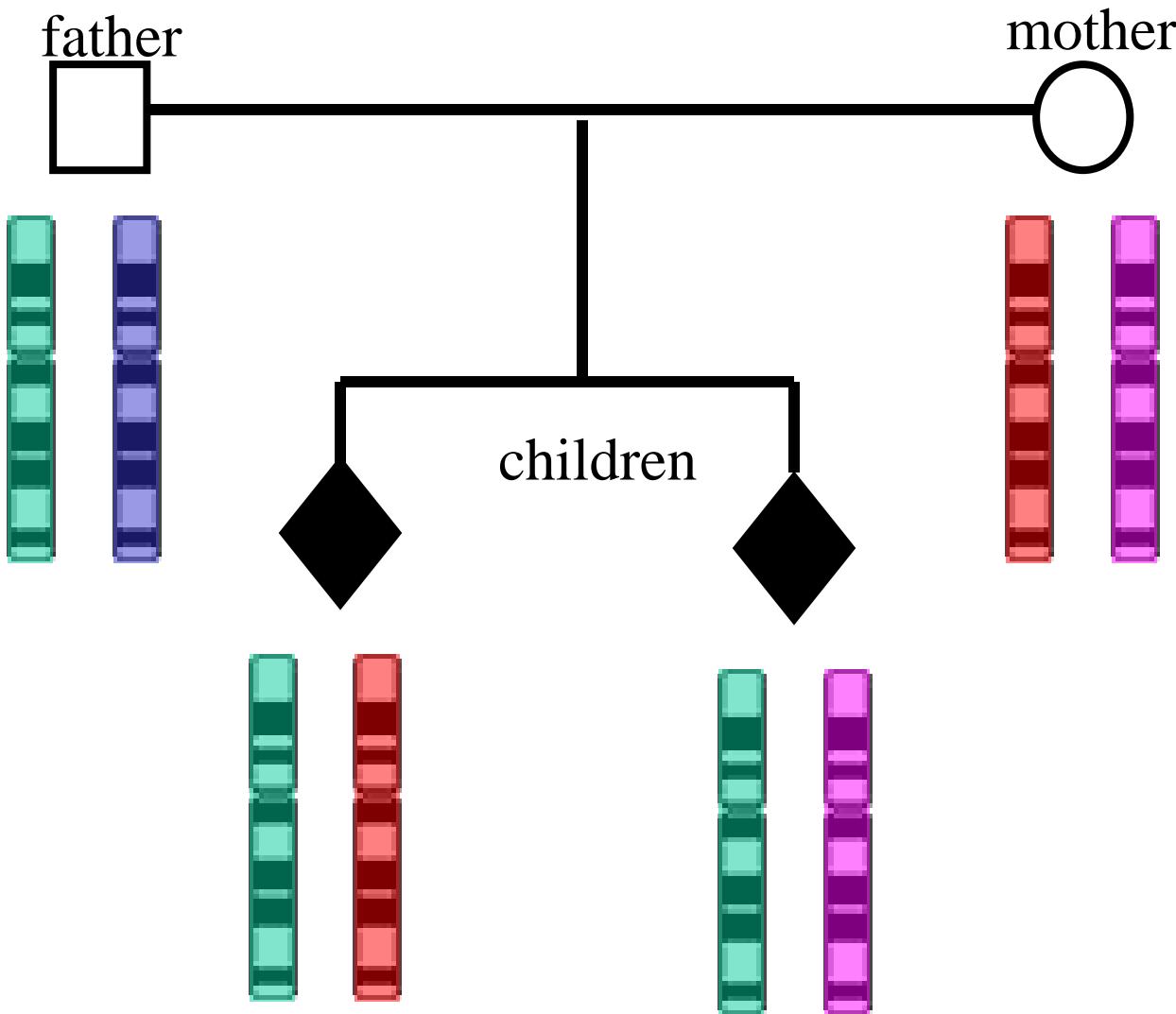
- Marker-trait linkage studies are the most popular approach for mapping
 - Mendelian disease genes
 - ADPKD -> PKD1
 - Early onset breast cancer -> BRCA1

ADPKD

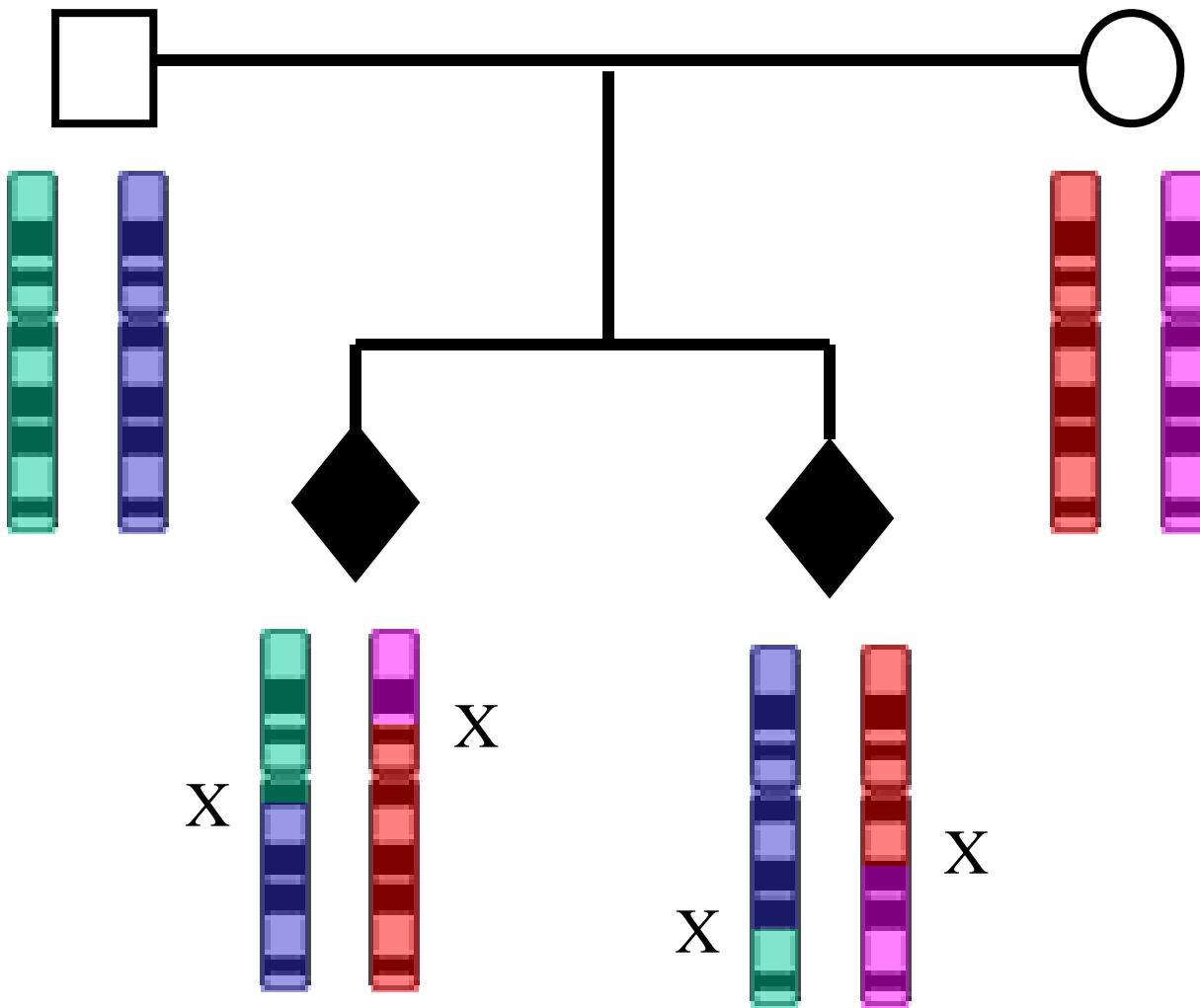


NDM-B pedigree (Reeders et al., Nature 317, 542-4; 1985)

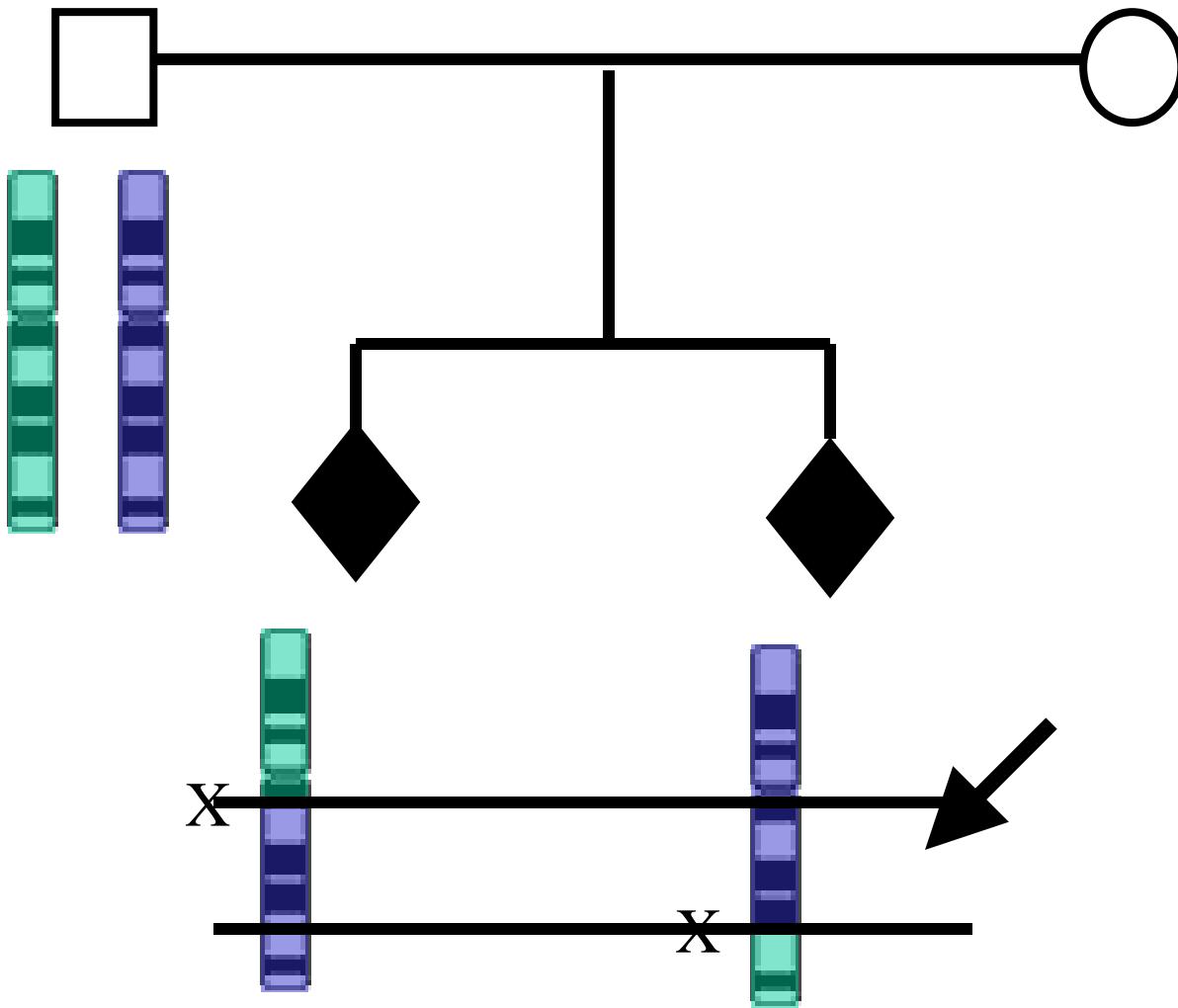
Inheritance of chromosomes



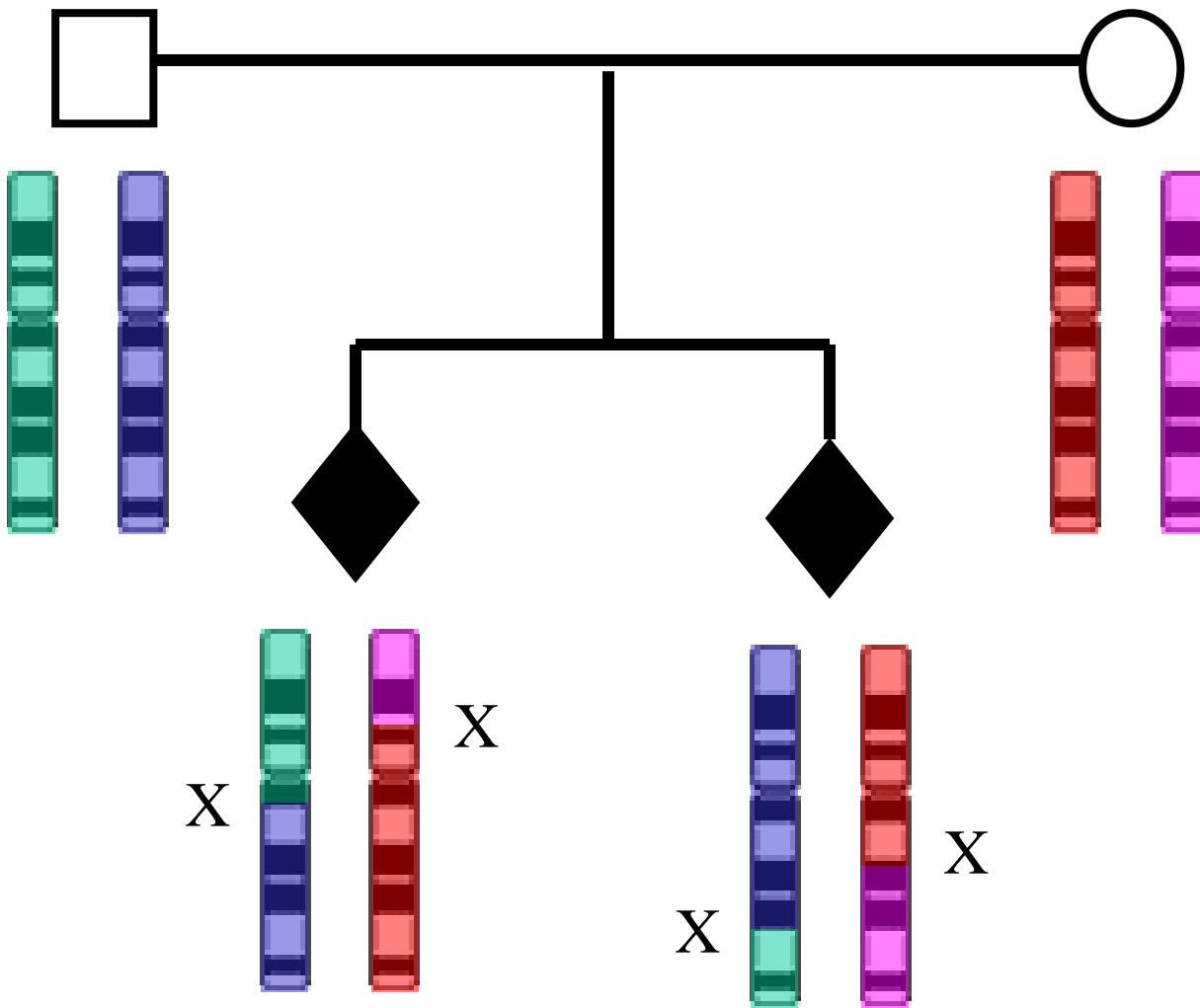
Crossing-over (X) of chromosomes



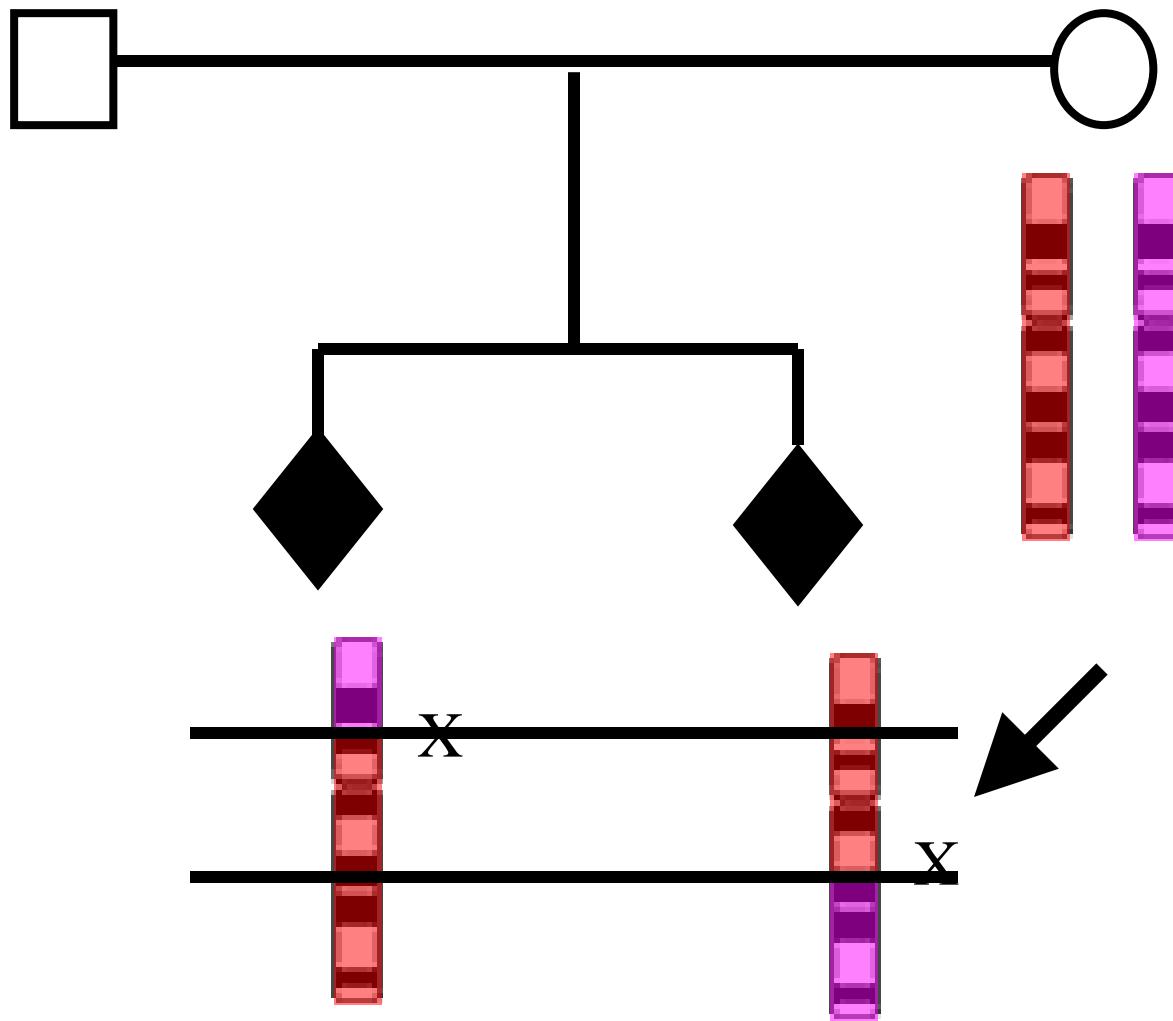
Region of father's chromosome shared



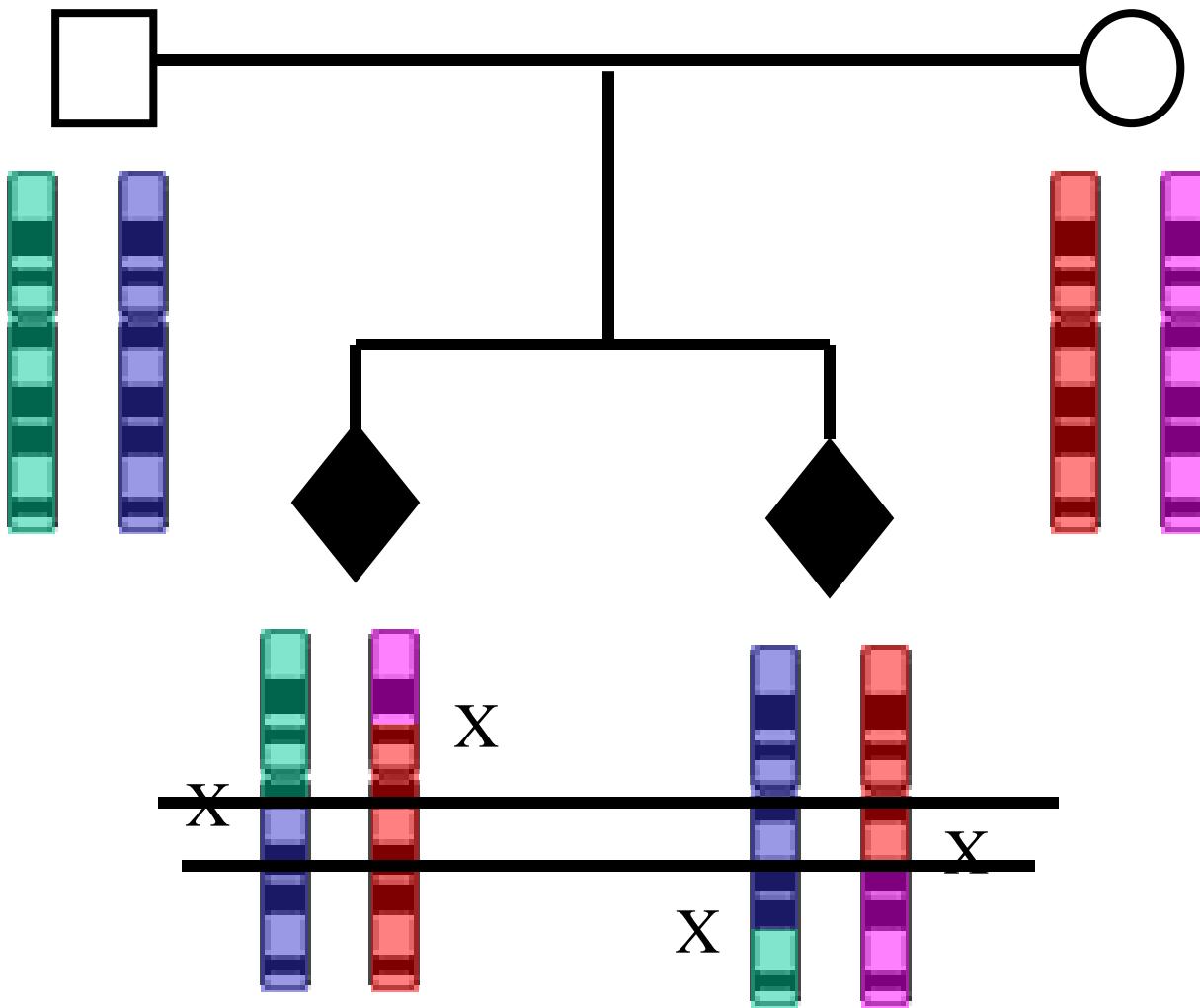
Crossing-over (X) of chromosomes



Region of mother's chromosome shared

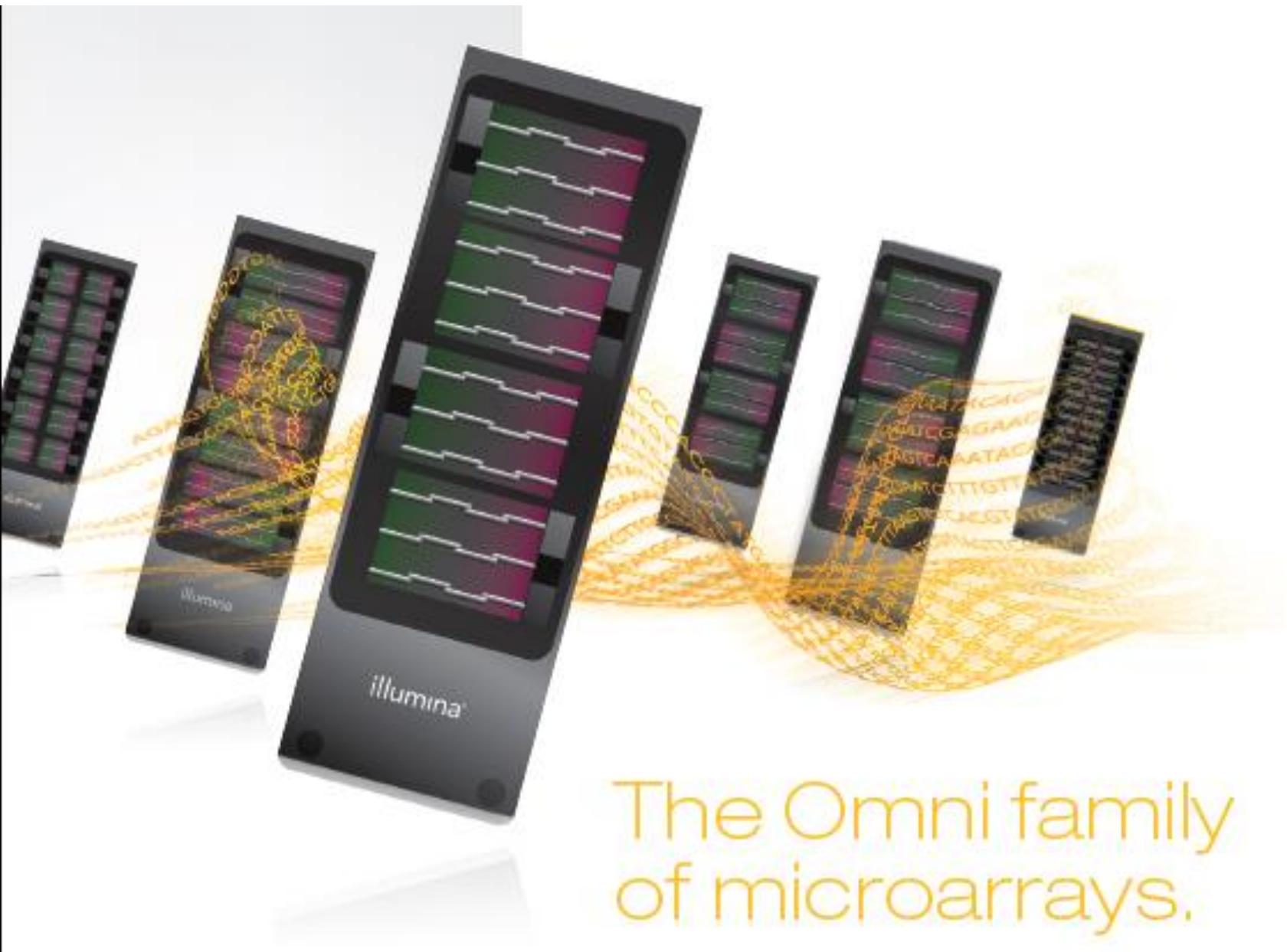


Region of chromosome shared from both father and mother



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

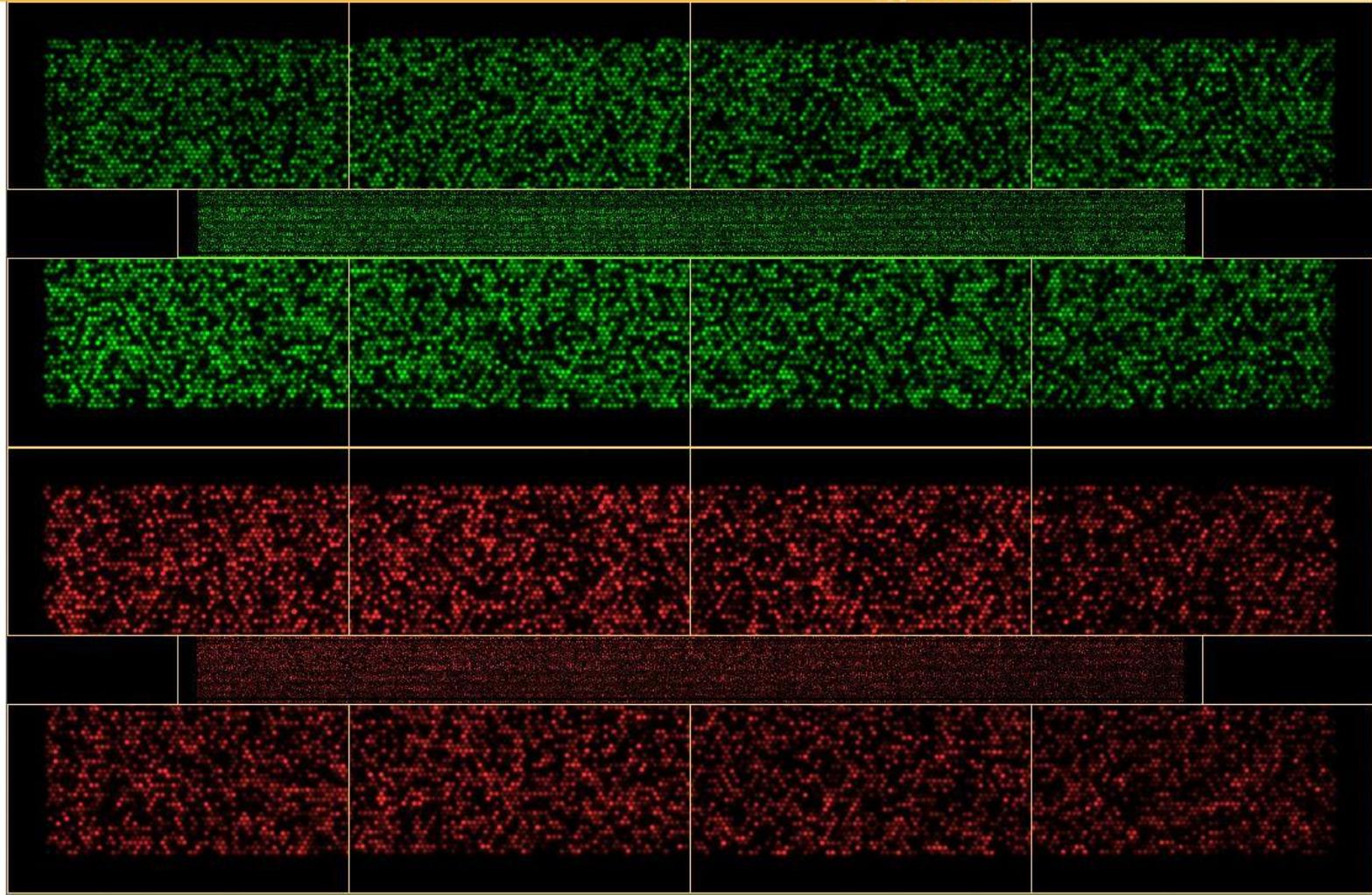


The Omni family
of microarrays.



5801149151

37 of 40 selected sections have been scanned.



R04C01_06, Register Swath 1 Red
R04C01_07, Register Swath 1 Grn
R04C01_07, Register Swath 1 Red

R04C01_07, Register Swath 1 Grn

R04C01_05, Spatial Normalization



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

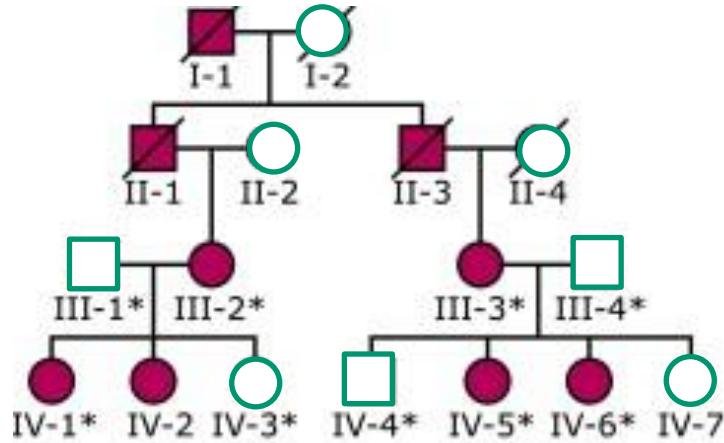


>> event log

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



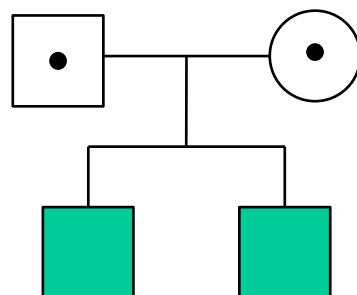
105kb tandem
duplication identified
in RUNX2

Moffatt et al., Am J Hum Genet 2013; 92 (2), 252-258.

Metaphyseal dysplasia with maxillary hypoplasia and brachydactyly

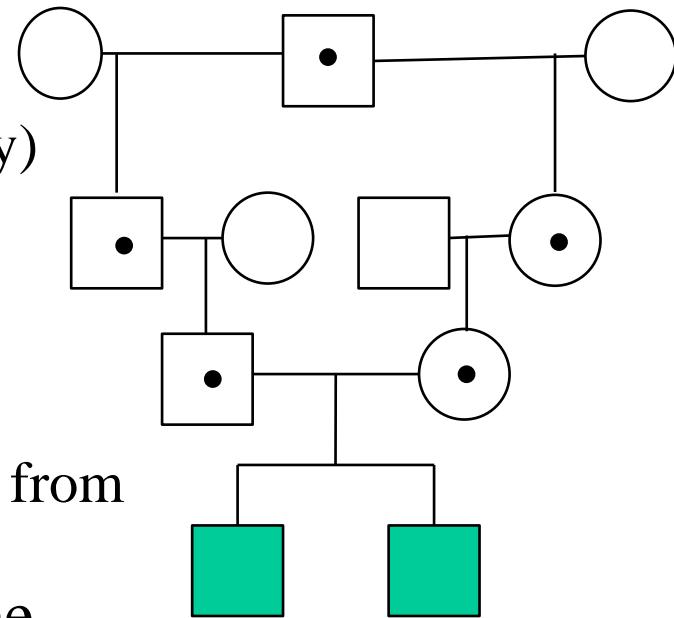
Childhood cerebellar ataxia ...

- 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 ~ $\frac{1}{4}$ of the genome (637cM/479Mb)

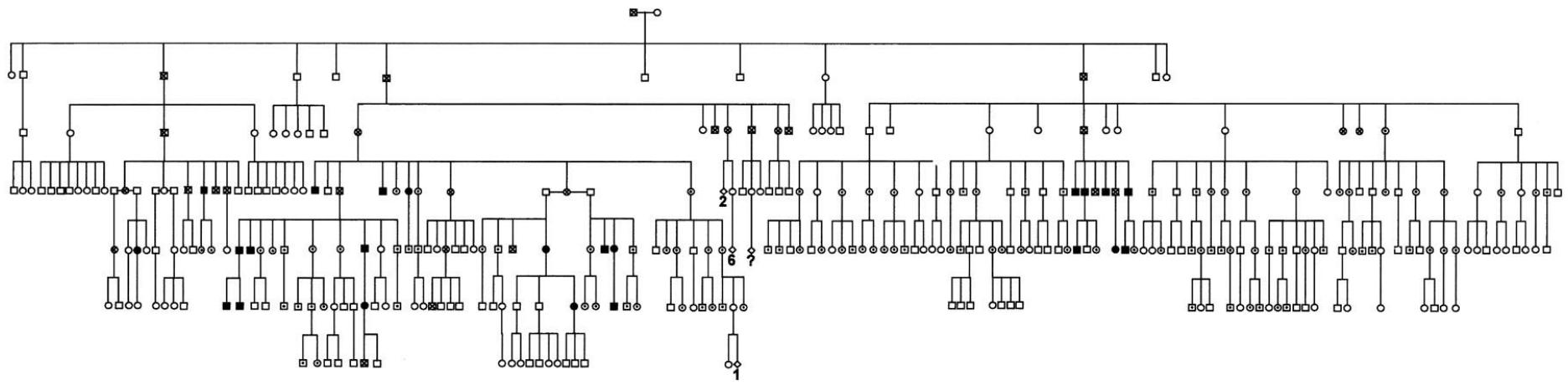


... continued

- 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

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

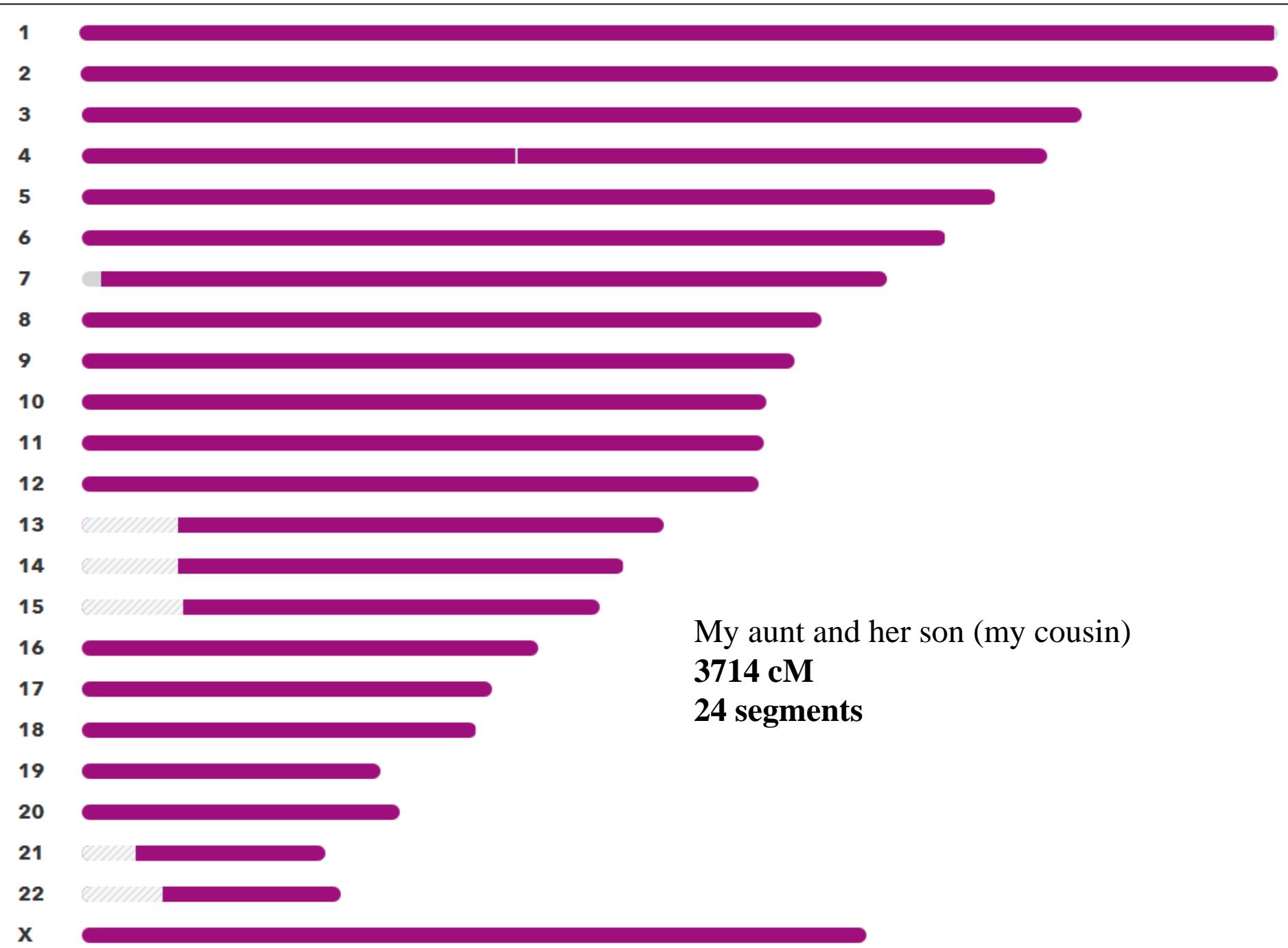
April 26, 2018

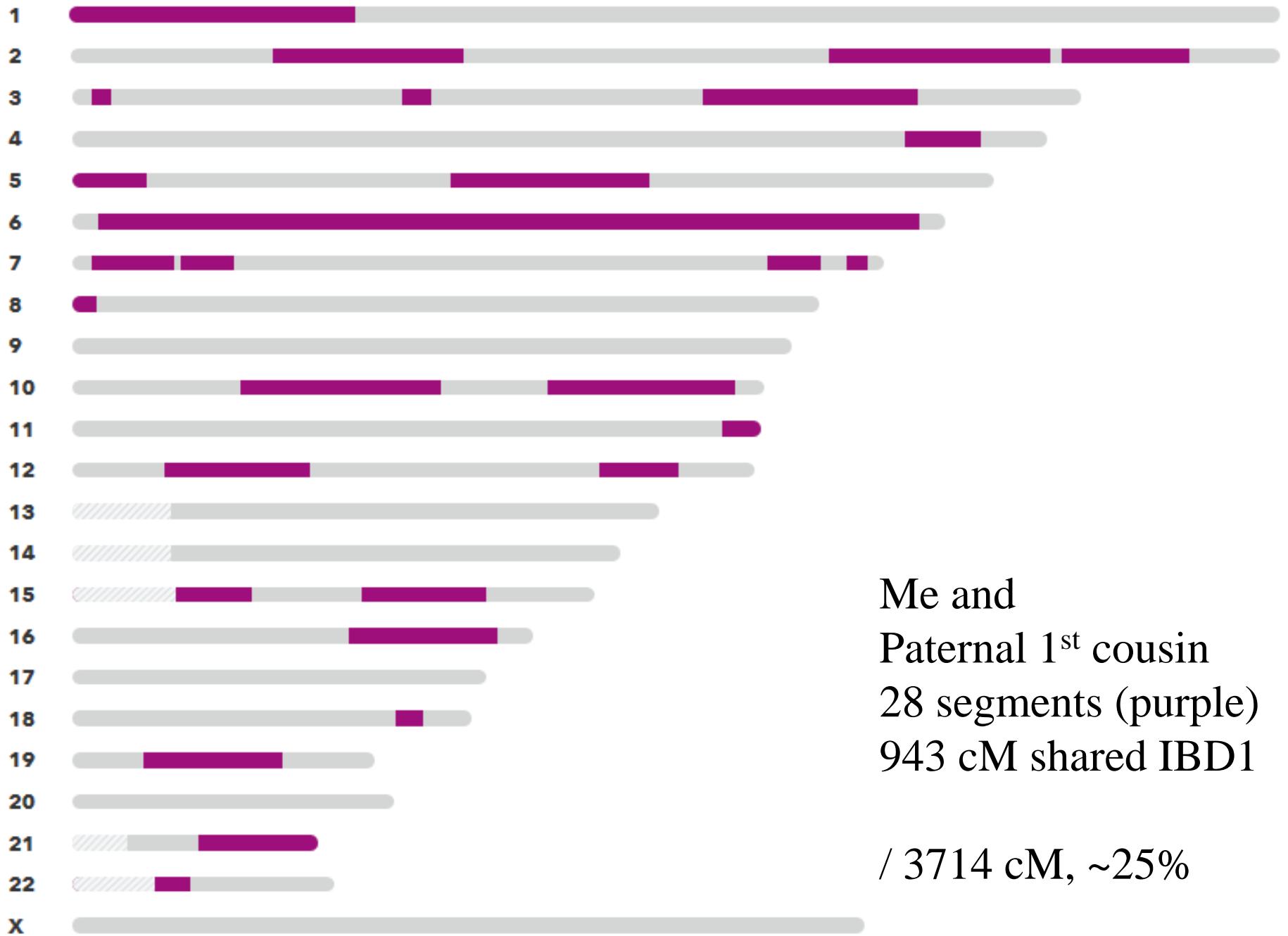
Golden State Killer

- 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

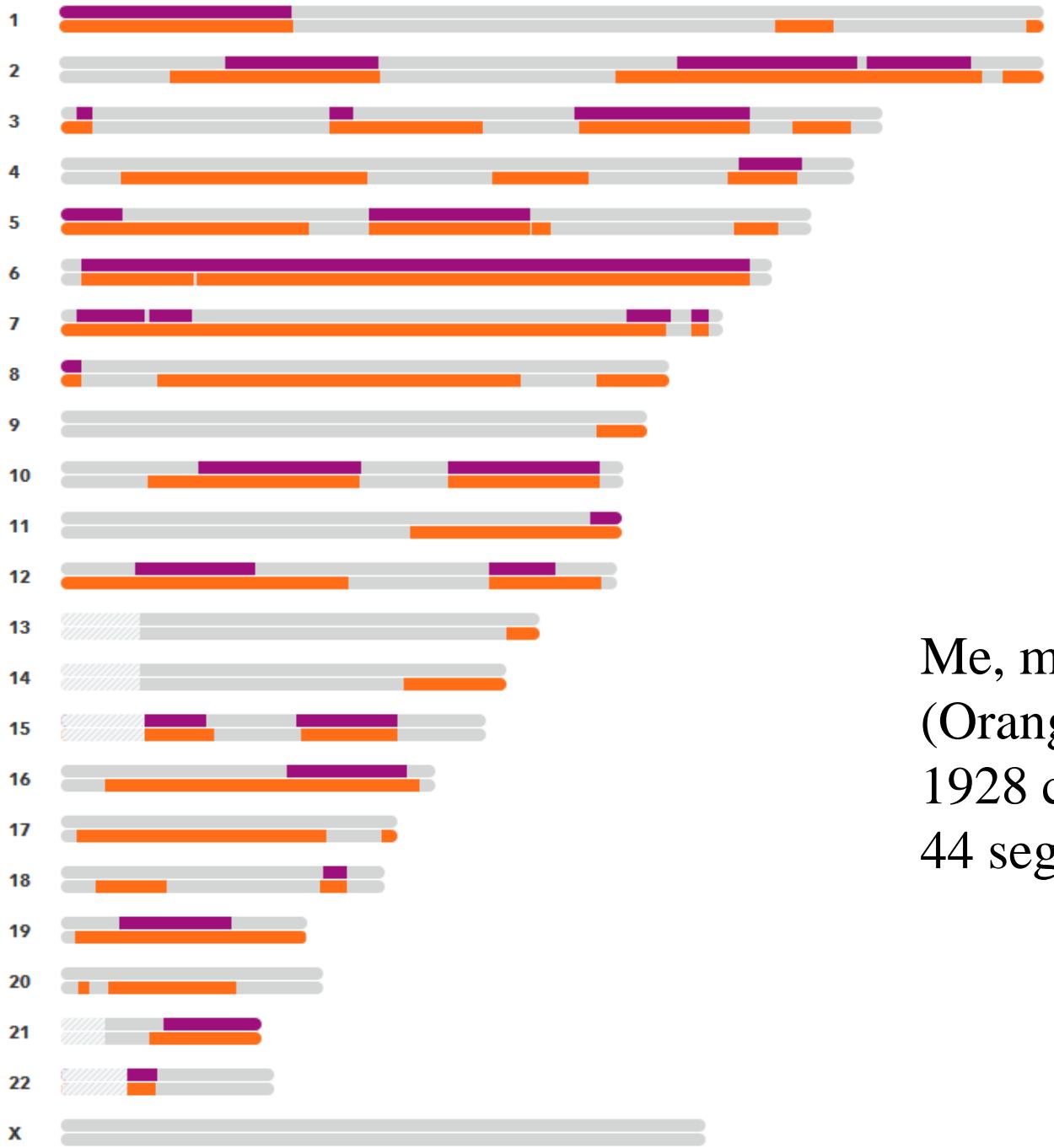
- Estimated that 60% of European ancestry Americans will have similar close relatives now in DTC
 - will continue to increase
 - 23andMe have >10 M customers
 - Assume people deposit data in GEDMatch
 - doi:10.1126/science.aav7021



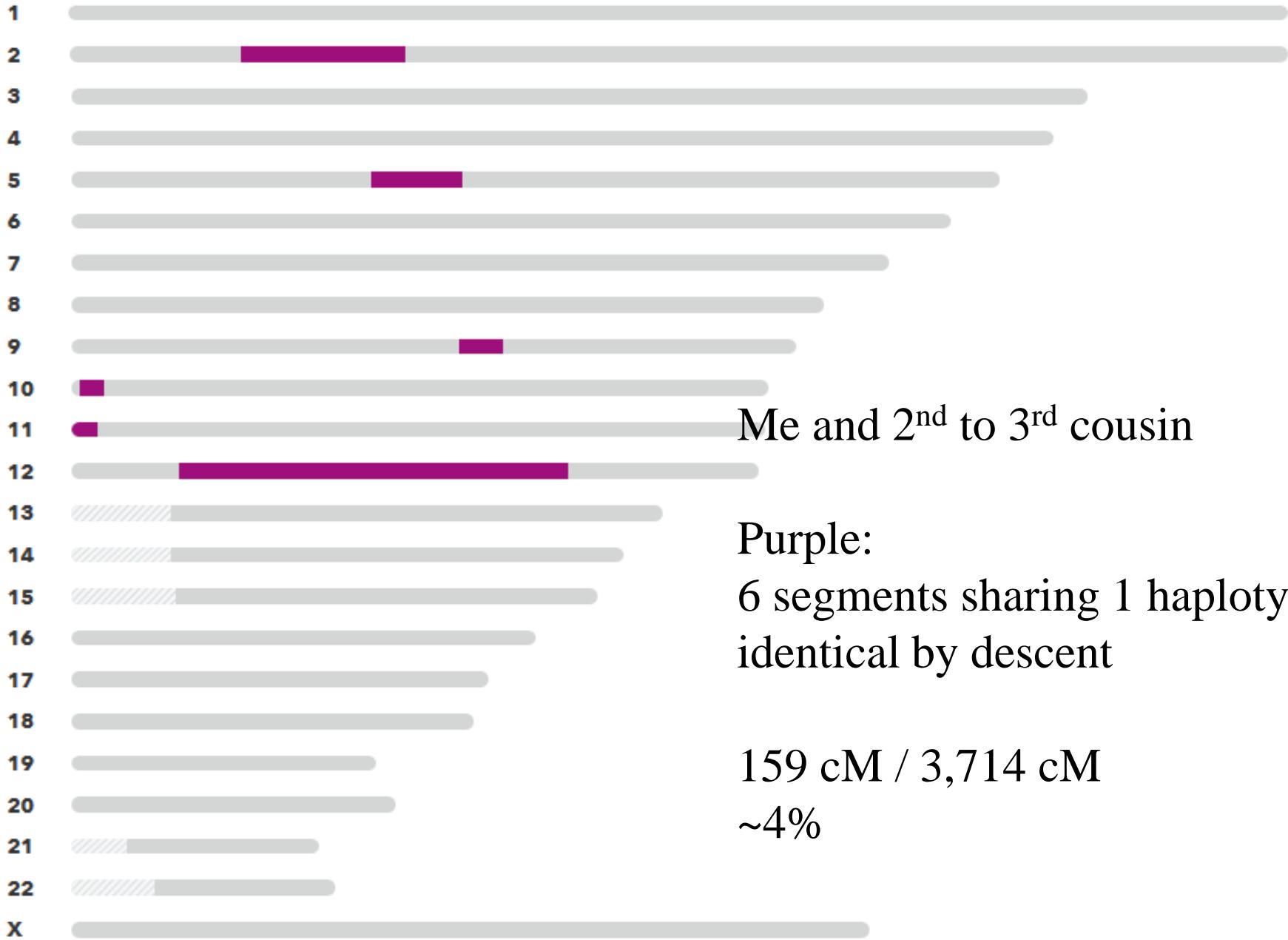


Me and
Paternal 1st cousin
28 segments (purple)
943 cM shared IBD1

/ 3714 cM, ~25%



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

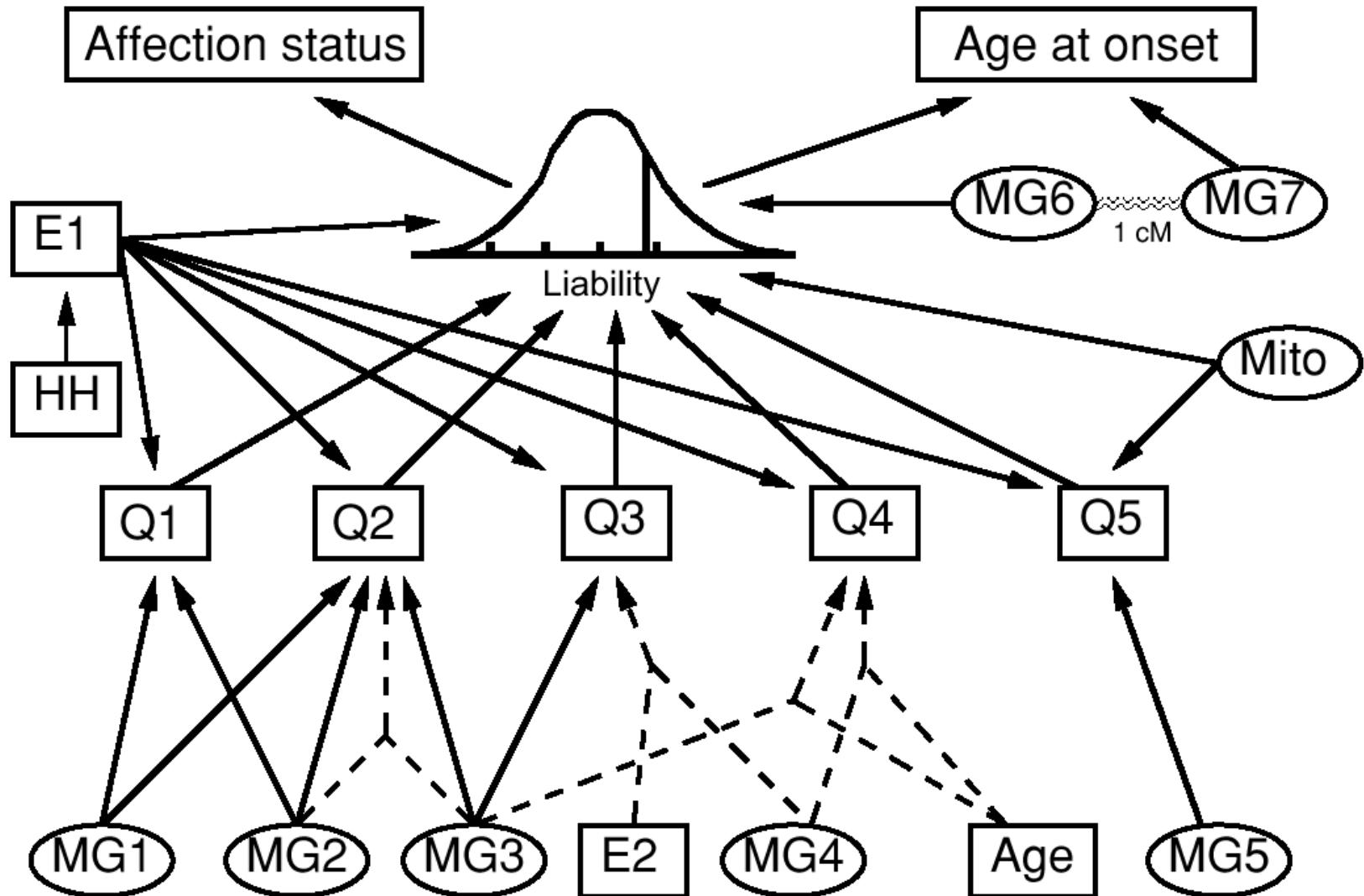


Purple:
6 segments sharing 1 haplotype
identical by descent

159 cM / 3,714 cM
~4%

3 Genetic Topics

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



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 $p < \text{threshold}$)
- 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

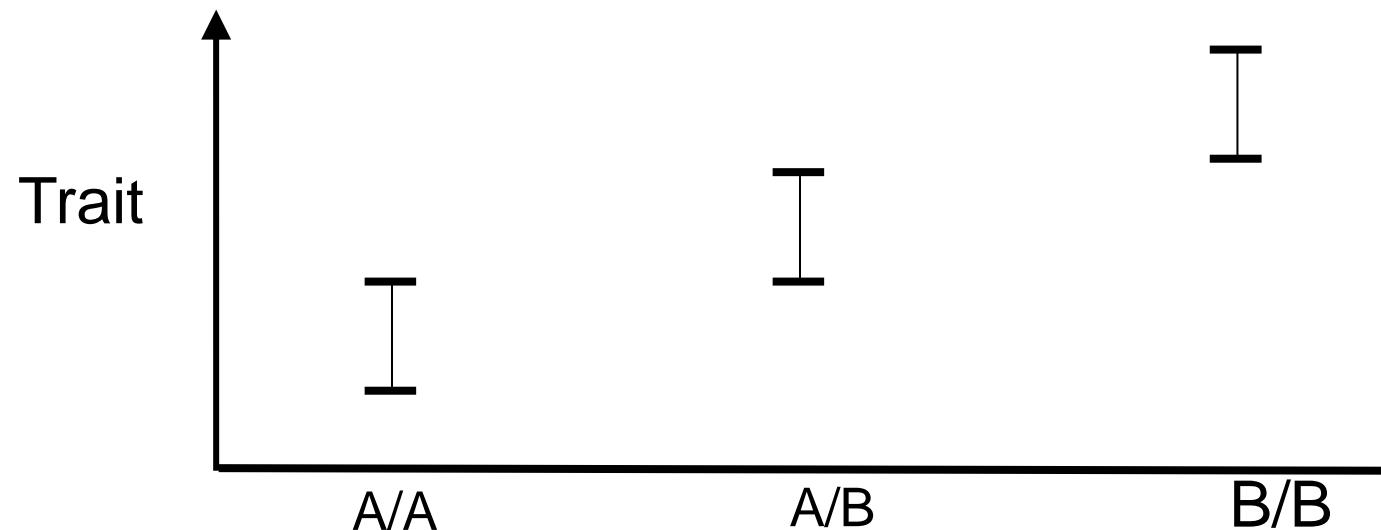
- Case-control
 - Compare genotype and/or allele frequencies

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

International HapMap Project



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
↓
Individual 1 AACAC**C**GCCA.... TTTCG**G**GGGTc.... AGT**T**C**G**ACCG....
Individual 2 AACAC**C**GCCA.... TTTCG**A**GGGTc.... AGT**T**CA**C**CCG....
Individual 3 AACAC**T**GCCA.... TTTCG**G**GGGTc.... AGT**T**CA**A**CCG....
Individual 4 AACAC**G**GCCA.... TTTCG**G**GGGTc.... AGT**T**CG**A**CCG....

b Haplotypes

Haplotype 1 **CTCAAAAGTACGGTT**CAGGCA

Haplotype 2 **TTGATTGCGCAACAGTAATA**

Haplotype 3 **CCCGATCTGTGATAACTGGTG**

Haplotype 4 **TCGATTCCGGGGTT**CAGACA

↓
A / G

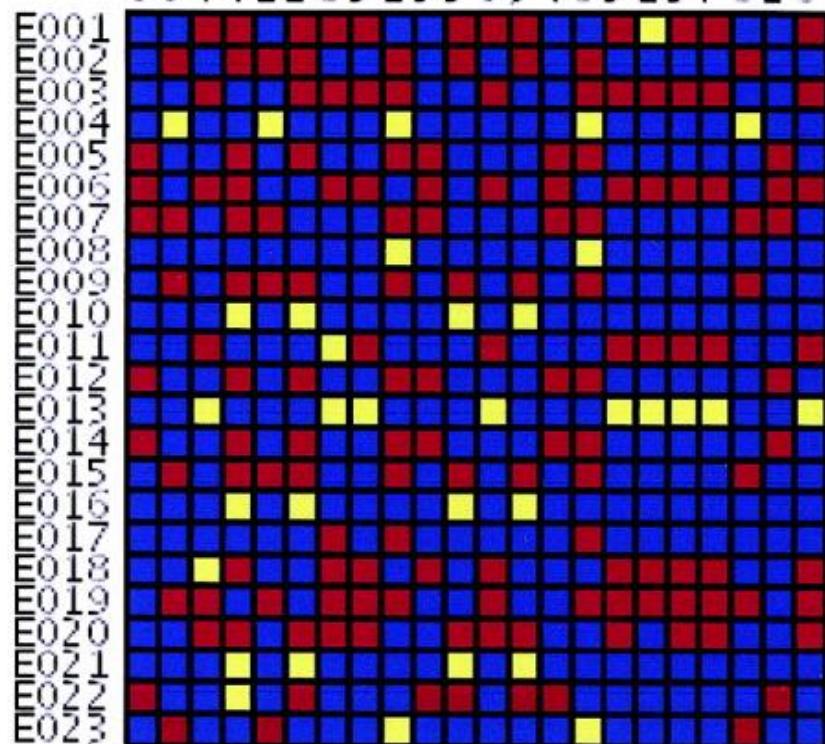
↓
T / C

↓
C / G

c Tag SNPs

SNPs

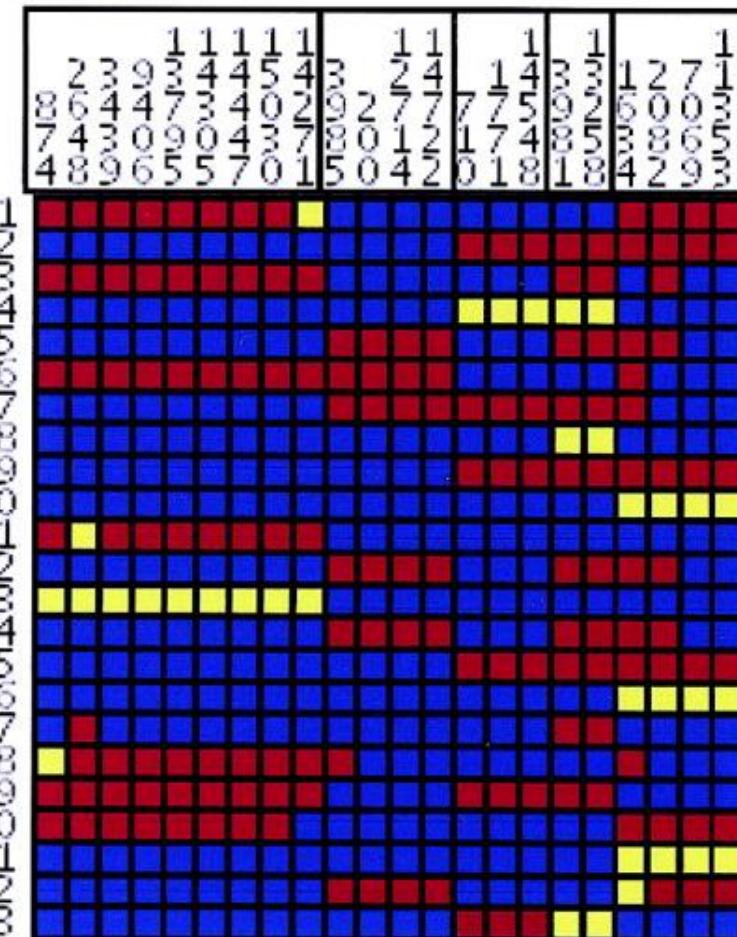
A



■ Homozygote-Common Allele
■ Homozygote-Rare Allele
■ Heterozygote

SNPs in Bins

B



- 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: $p < 5 \times 10^{-8}$ 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)

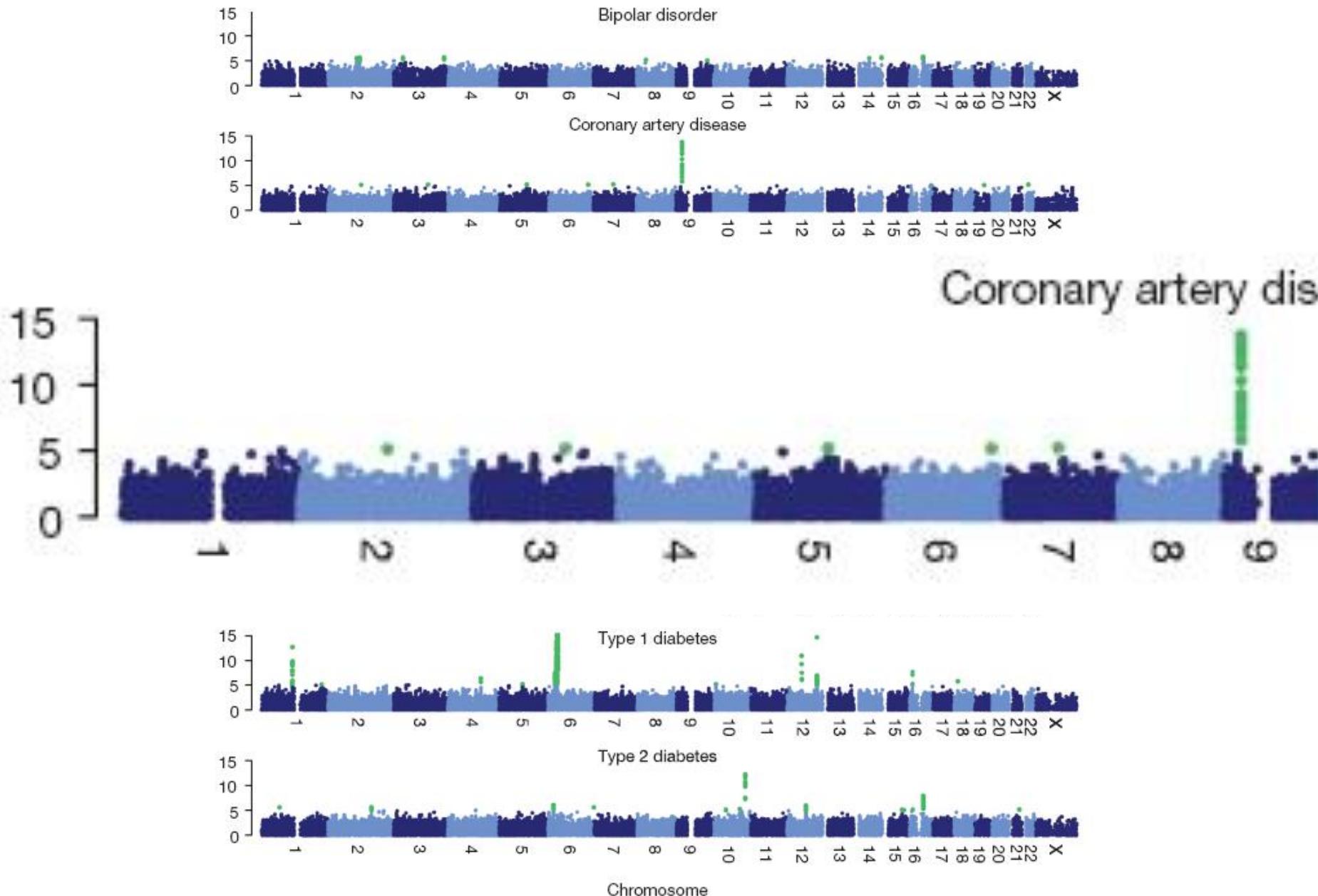


Figure 4 | Genome-wide scan for seven diseases. For each of seven diseases $-\log_{10}$ of the trend test P value for quality-control-positive SNPs, excluding those in each disease that were excluded for having poor clustering after visual inspection, are plotted against position on each chromosome.

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

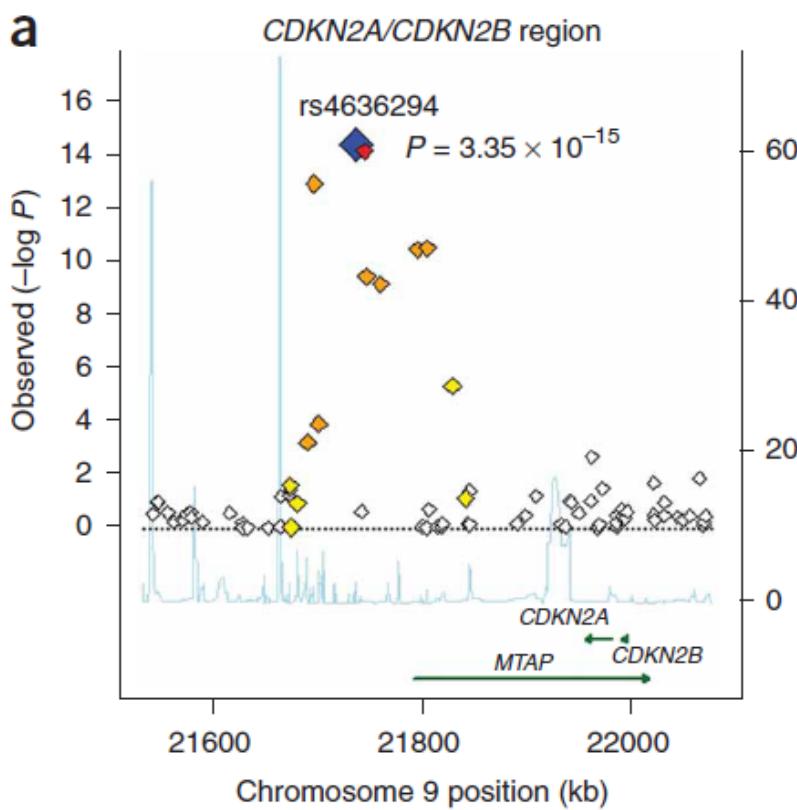


Figure 2 Regional plots. (a) Chromosome 9p21. (b) Chromosome 22q13. Meta analysis $\log_{10} P$ values are plotted as a function of genomic position (build 36). The P values for the lead SNPs are denoted by large blue (combined discovery and replication) diamonds. Proxies are indicated with diamonds 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)	

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/>



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/>

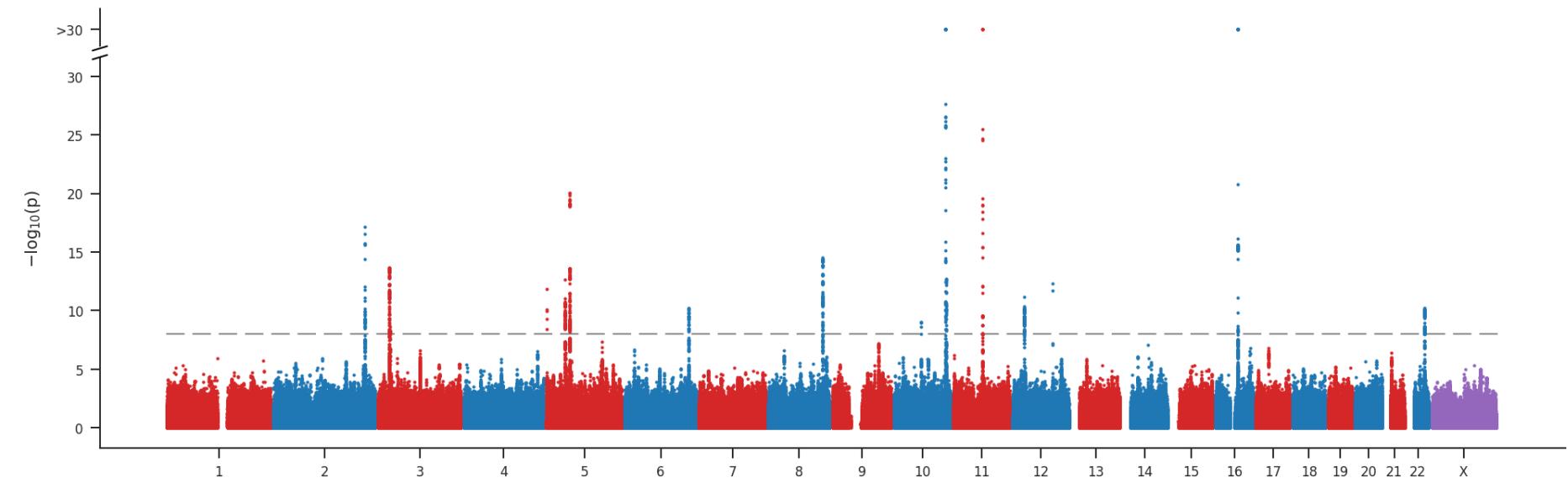
9.1 M variants that pass QC

Females:

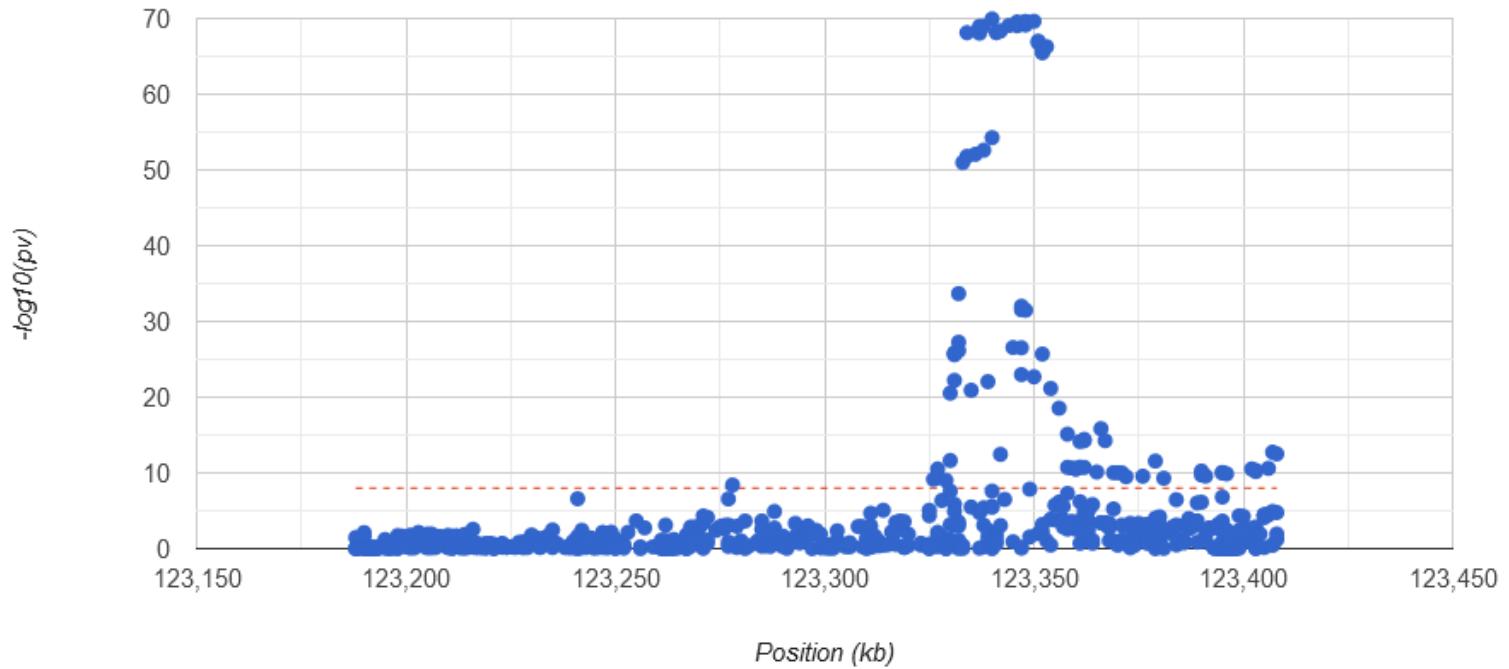
10,478 cases

235,016 controls

C50-C50 Malignant neoplasm of breast



C50-C50 Malignant neoplasm of breast



Chr 11, around FGFR2

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

