CS 364 COMPUTATIONAL BIOLOGY

Sara Mathieson
Haverford College

Outline

- Computational disease biology beyond GWAS
- Uses of DNA information
- DNA in forensics
- Non-invasive Prenatal Testing (NIPT)

Notes:

--Project meetings in lab next Thursday
--Presentations next Thursday
--Presentation instructions and final deliverables posted!

Project Presentation Notes

- Date: in-class **Thursday, Dec 11**
- Each person will have **10-12 minutes to present** (+ time for questions and transition)
- Email me your slides by **12pm on Dec 11!** (PDF only)
- I will have a laser pointer / slide advancer clicker

Project Presentation Notes

Your presentation should include

- Motivation and Scientific Question
- Data and Methods
- Results and Interpretation
- Conclusions and Future Work

Presentation Tips

- Speak loudly (to the back of the class)
- Avoid text-heavy slides, use images/diagrams
- Include citations for any figures you did not make
- Ask at least one question to another group

Submit by 12pm on Dec 20

- Lab Notebook (include references)
- All project code
- Presentation slides

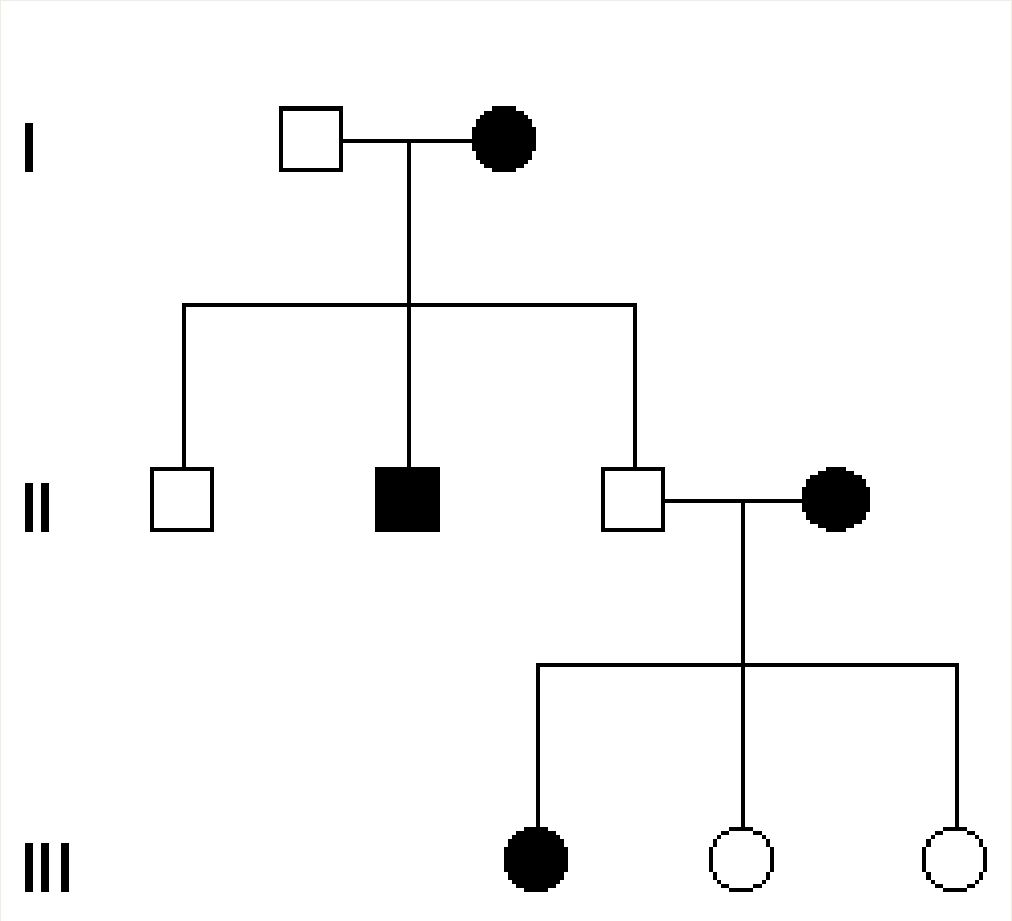
Think about reproducibility!

Computational disease biology beyond GWAS

Pedigree Analysis

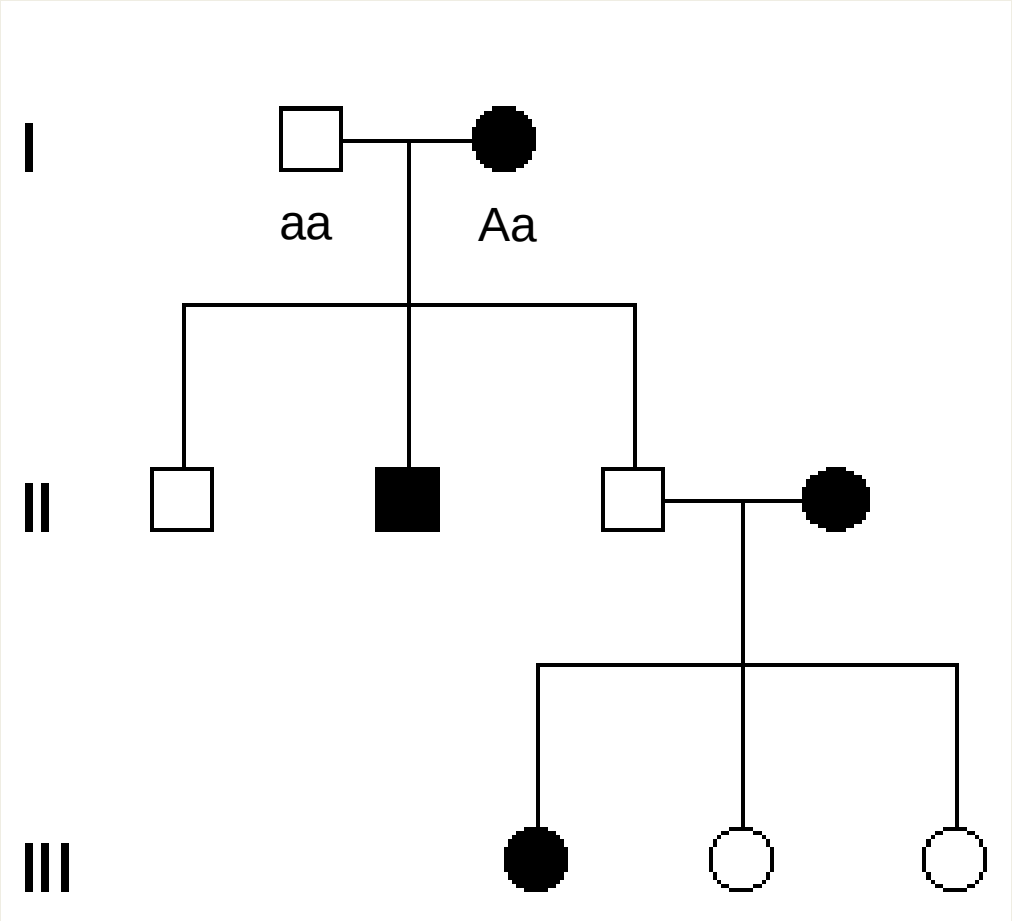
Beyond GWAS: pedigree analysis

Dominant



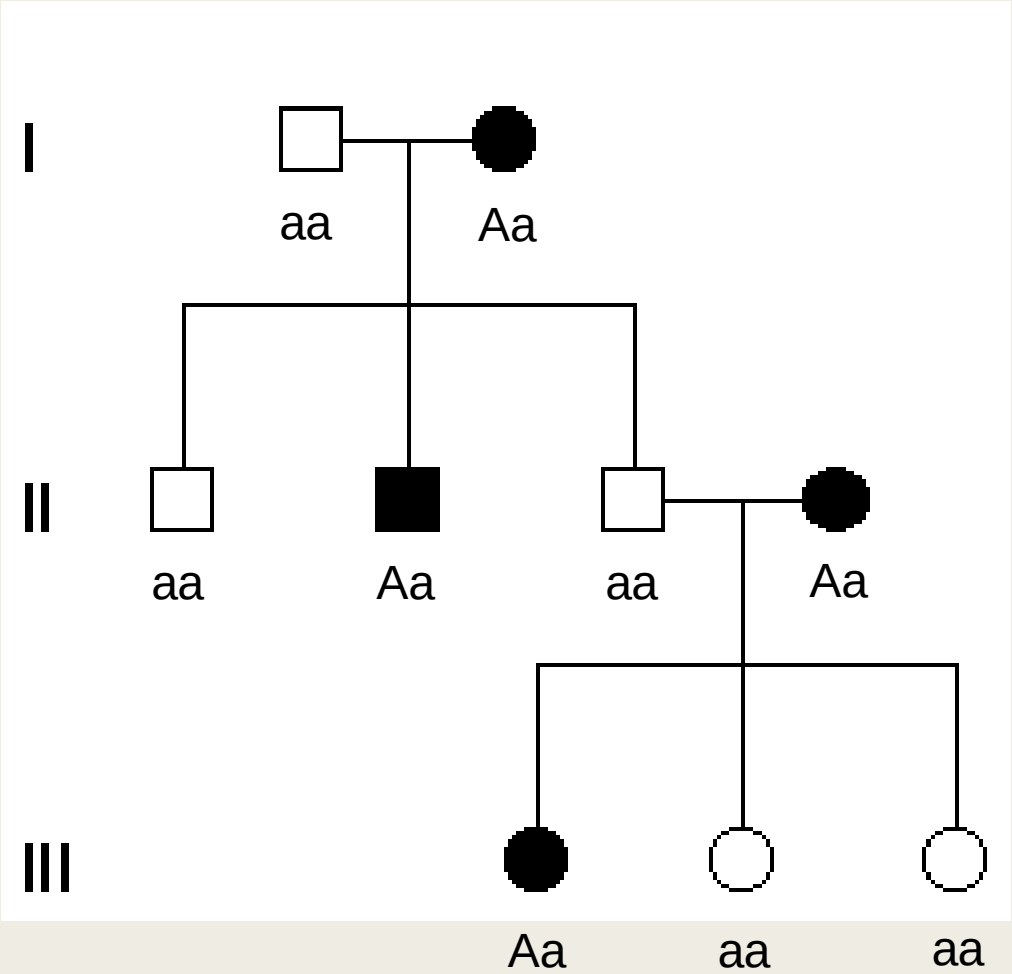
Beyond GWAS: pedigree analysis

Dominant



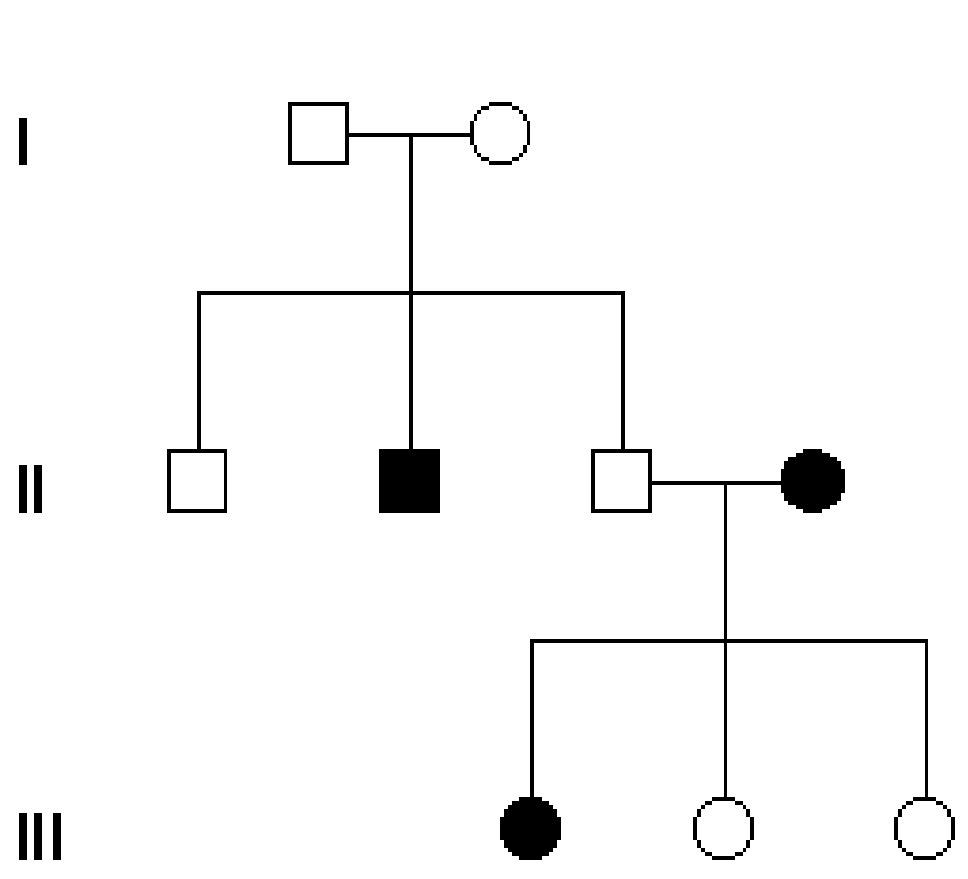
Beyond GWAS: pedigree analysis

Dominant



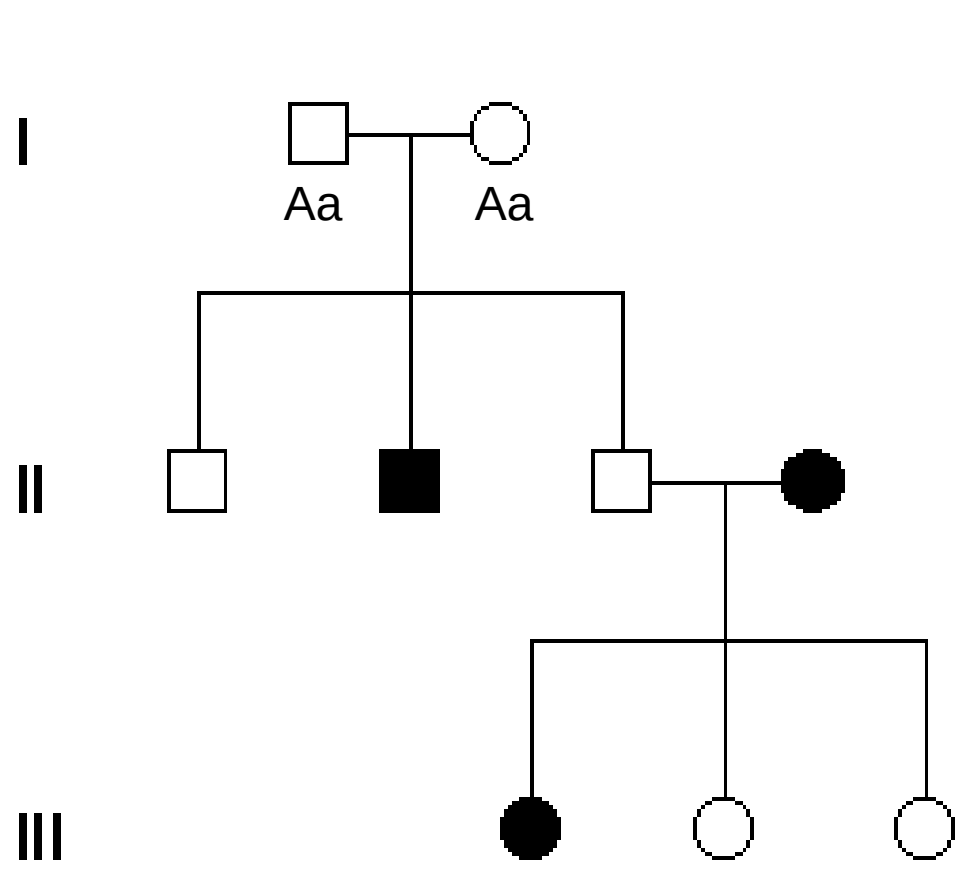
Beyond GWAS: pedigree analysis

Recessive



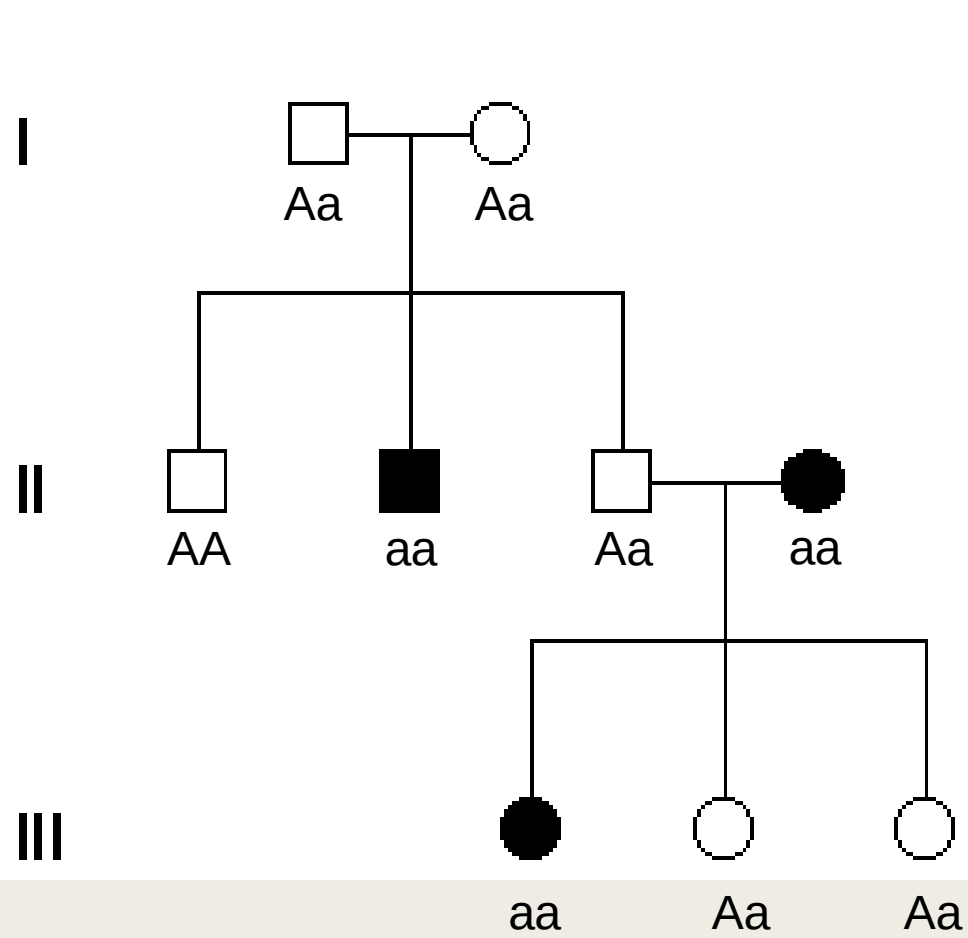
Beyond GWAS: pedigree analysis

Recessive



Beyond GWAS: pedigree analysis

Recessive



Ability to taste the chemical PTC (“bitter” taste)

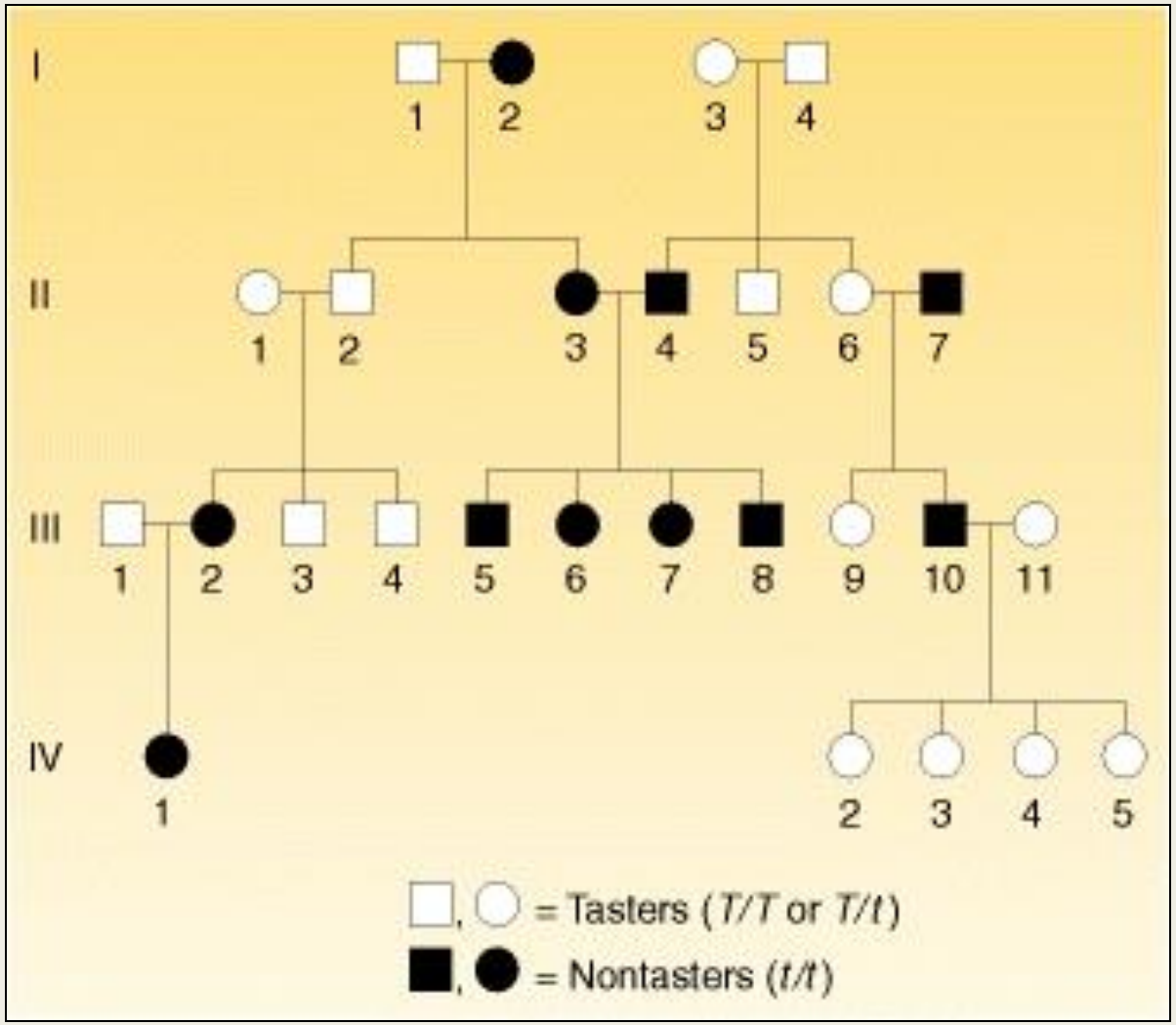


Image: *Human Pedigree Analysis* (1999)

Hemophilia in the Royal Family – X linked

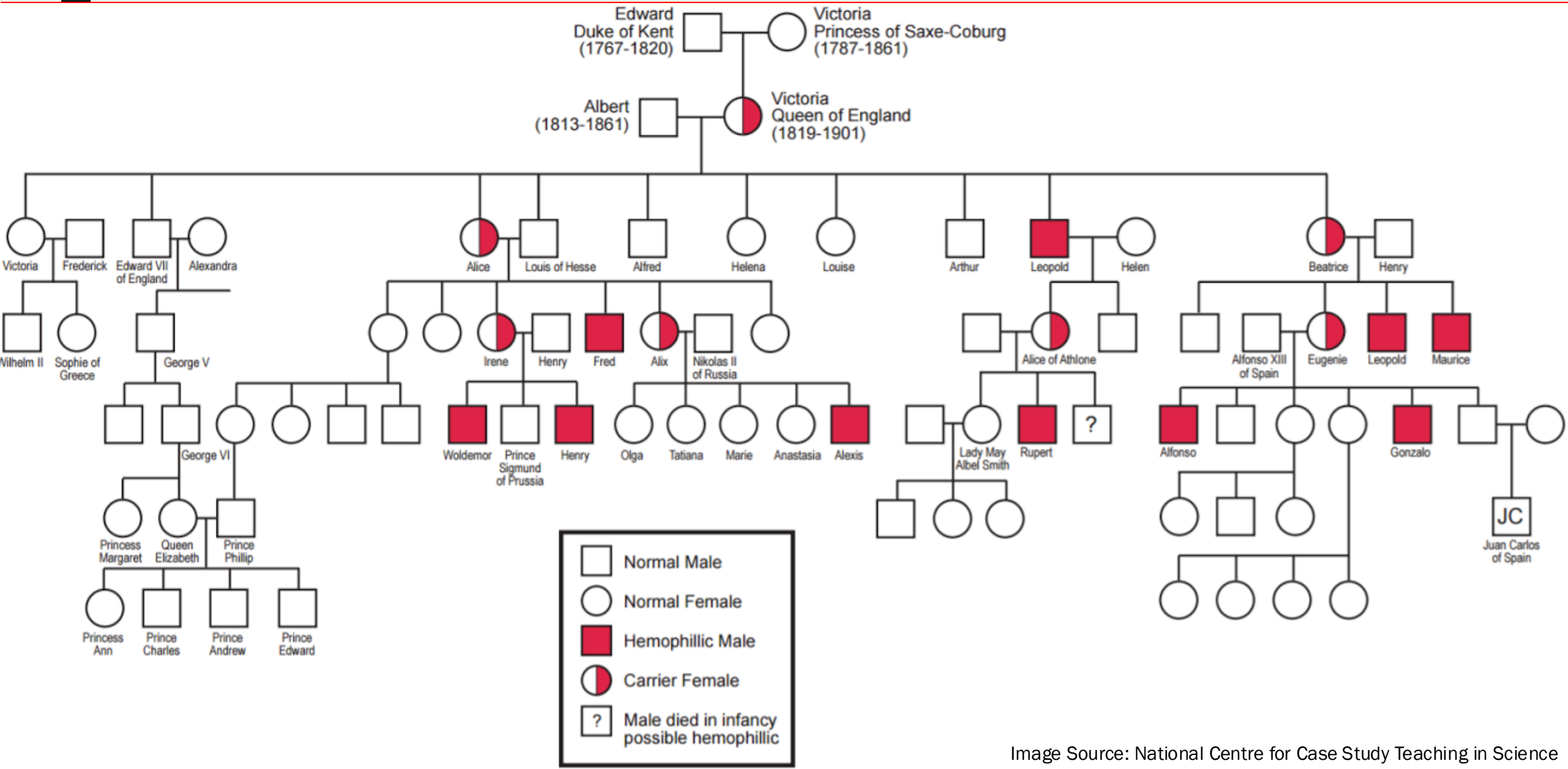
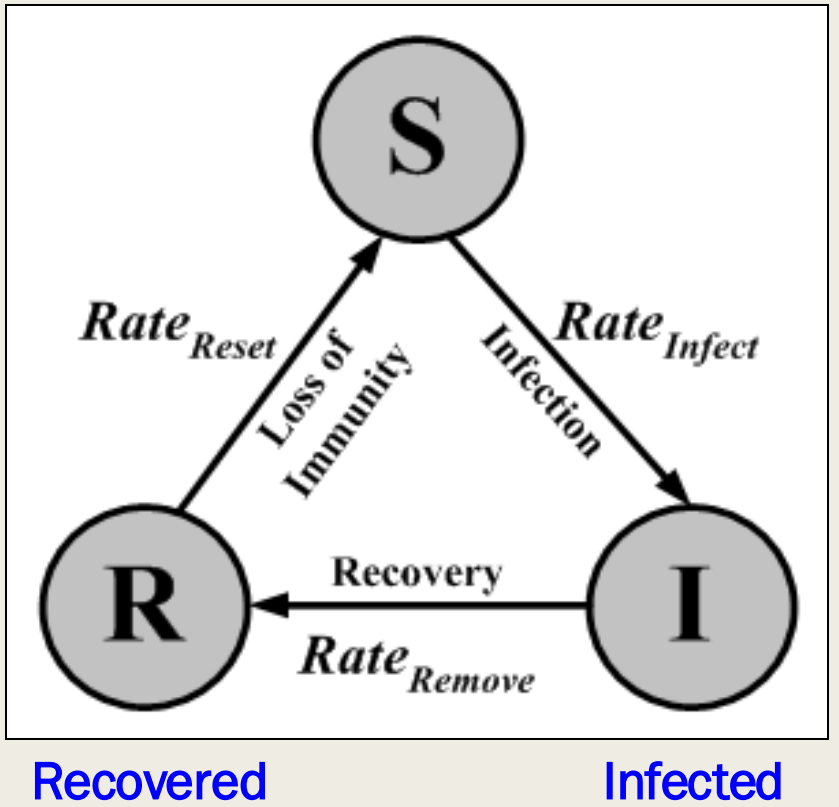


Image Source: National Centre for Case Study Teaching in Science

Infectious disease modeling

SIR models for infectious disease

- Recent applications:
  - *H1N1, “swine flu”, 2009*
  - *Ebola, 2015*
  - *Covid, 2020*



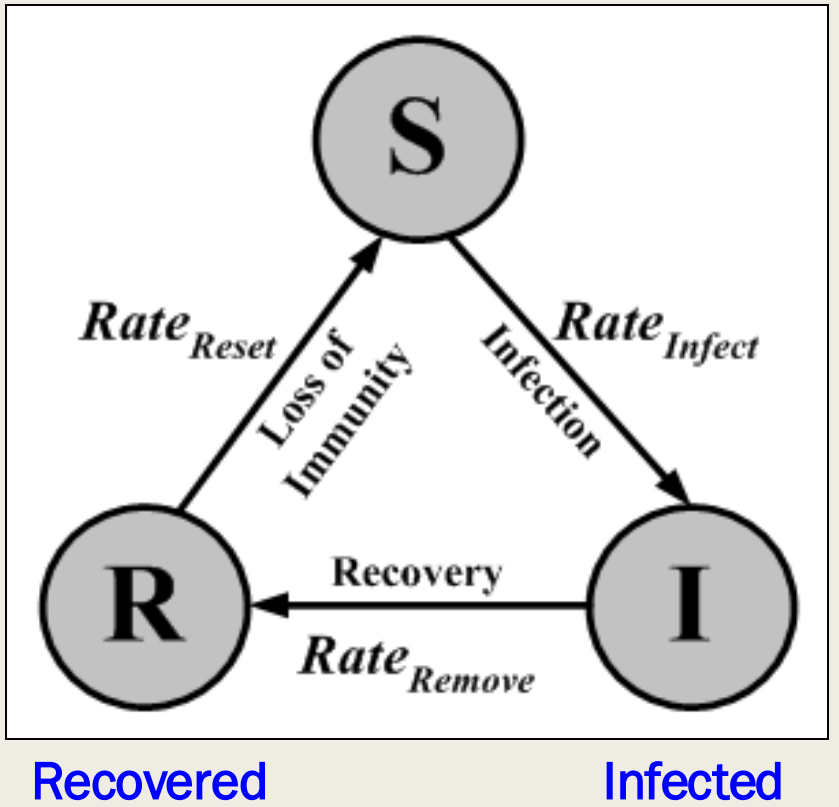
“Influence of Local Information on Social Simulations in Small-World Network Models” (2005)

SIR models for infectious disease

- Recent applications:
  - *H1N1, "swine flu", 2009*
  - *Ebola, 2015*
  - *Covid, 2020*

$$\frac{dS}{dt} = -\frac{\beta IS}{N},$$
$$\frac{dI}{dt} = \frac{\beta IS}{N} - \gamma I,$$
$$\frac{dR}{dt} = \gamma I.$$

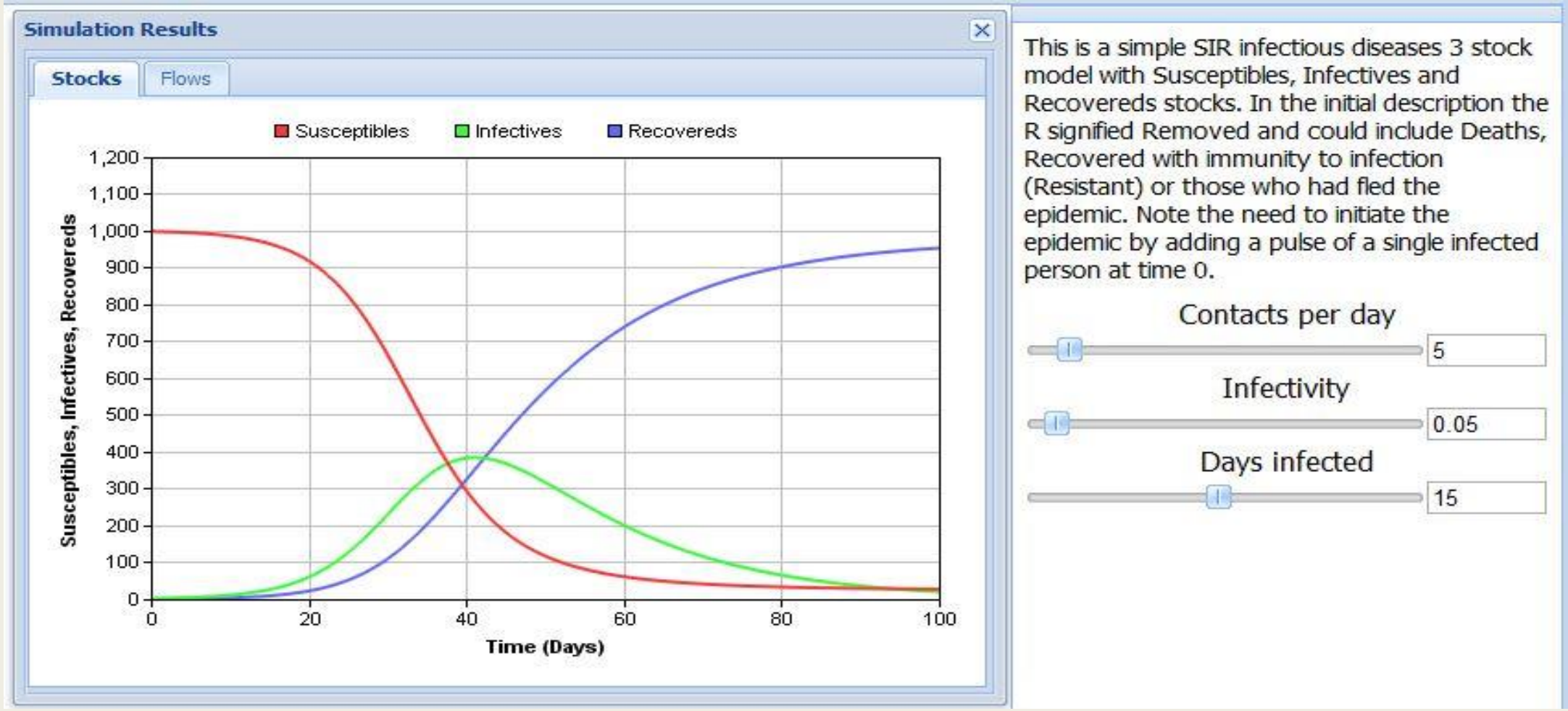
Modeled through differential equations



"Influence of Local Information on Social Simulations in Small-World Network Models" (2005)

Hethcote H (2000). *"The Mathematics of Infectious Diseases"*

SIR models for infectious disease



Cancer biology

Evolution of a cancerous tumor

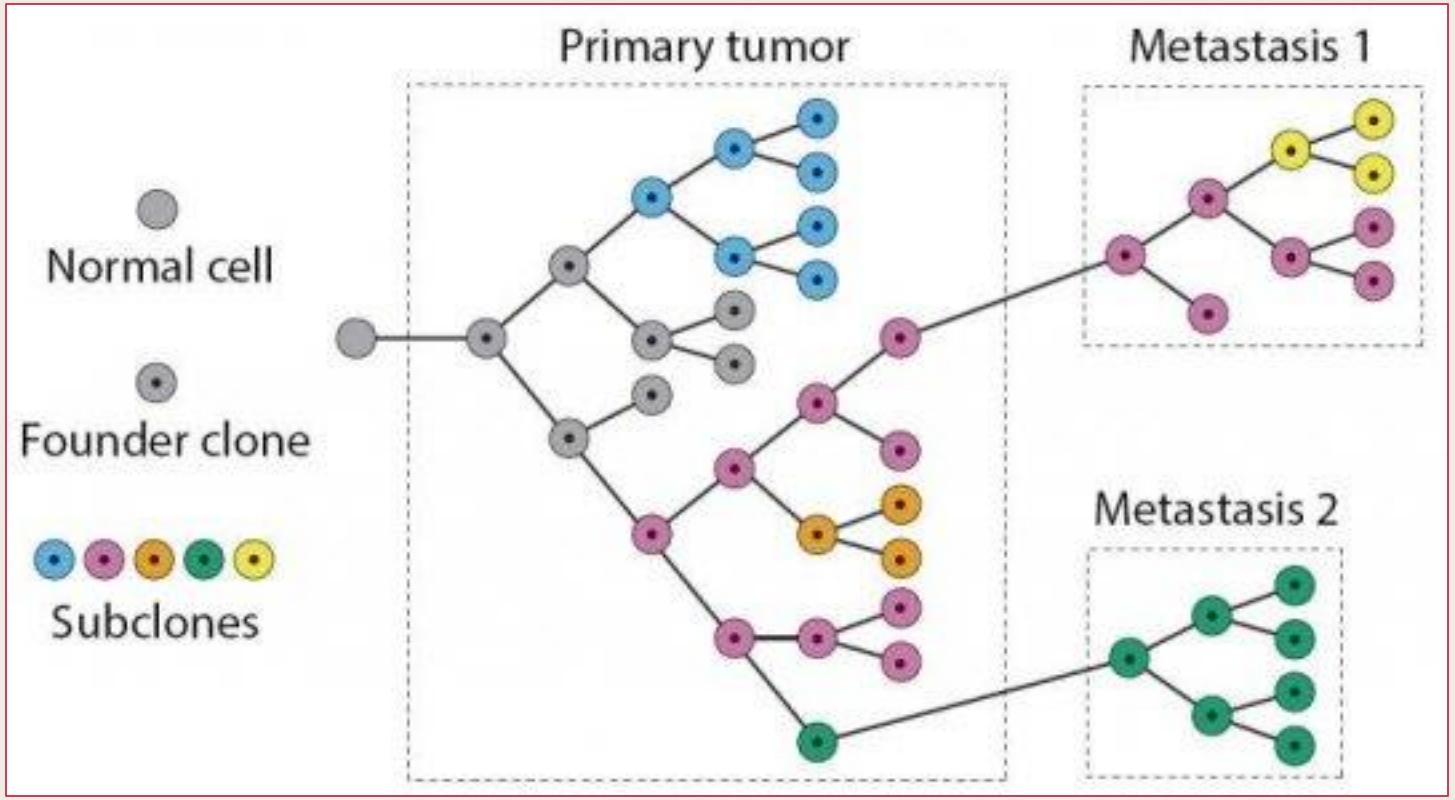
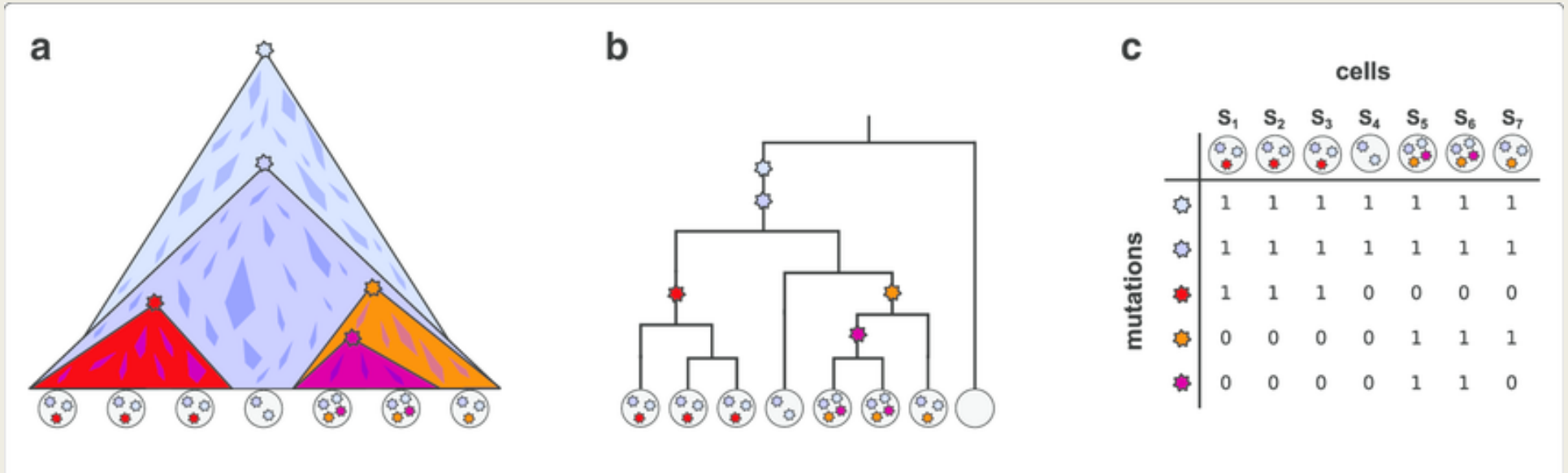


Image: ScienceDaily

Phylogenetic analysis of cancer cells

- Cancerous tumors often contain many different types of cells
- Once one mutation happens that causes the initial issue, mutations accumulate
- We can try to reconstruct the “ancestral” state to figure out what first went wrong



“Tree inference for single-cell data”, Genome Biology, 2016

Uses of DNA information

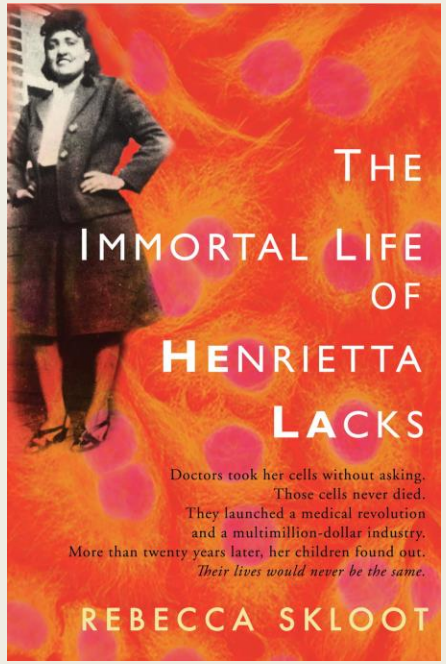
Big question: who owns DNA information?

- Early example: 1951, cells from Henrietta Lacks were immortalized (HeLa cells)

- These cells have been used extensively in research
  - *20 tons of cells grown*
  - *Involved in 11,000 patents*

- In 1990, court ruled that discarded tissue/cells are not the person's property and can be commercialized

Further reading:



Recently: companies offer DTC genetic testing

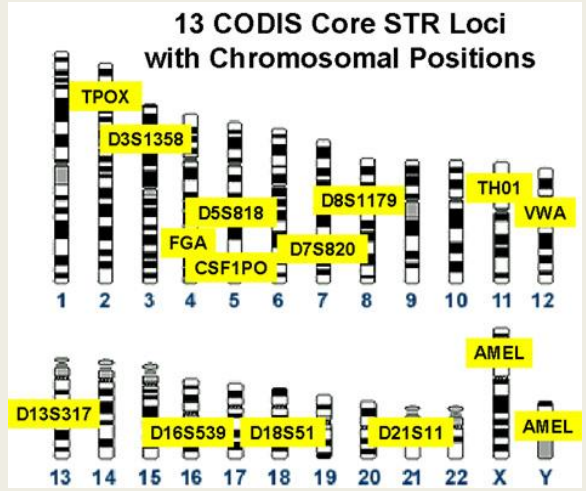
- 23andMe now has SNP data from 14 million individuals (2024)
- They offer both ancestry and health related information
- They ask users to answer survey questions (health history, physical traits, etc)
- Partnering with drug companies, as well as creating therapies in-house

Forensic uses of DNA

- Current technology uses short tandem repeats (STRs) for identification
- Repeats are typically 2-6bp long
- DNA from a crime scene is taken and then matched against an existing database of previously collected samples
- Even if existing samples are not linked to people, matches can indicate the crime was committed by the same unknown person

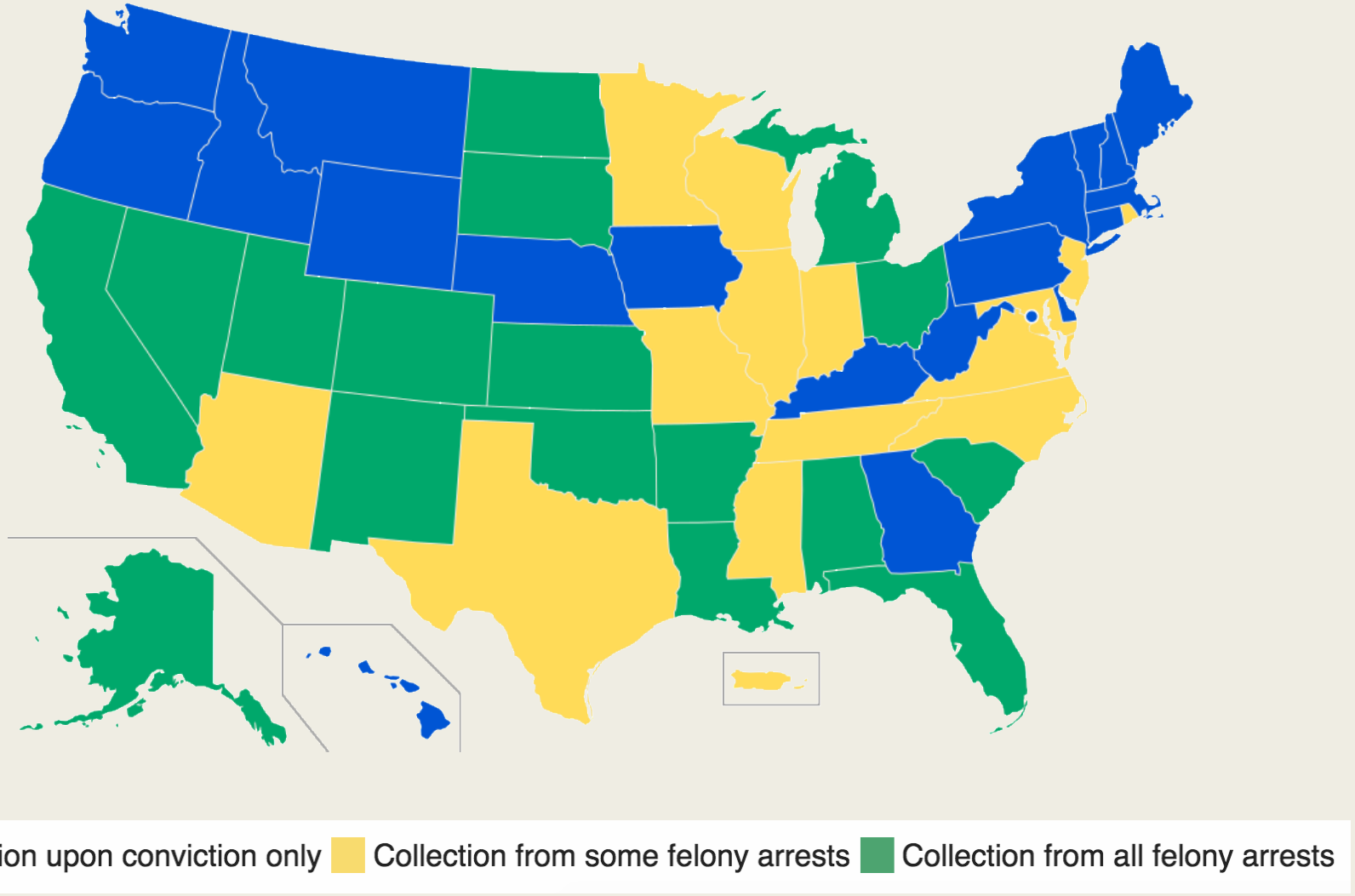
Combined DNA Index System (CODIS)

- U.S. National DNA database, maintained by the FBI
- Each profile has 13 STR (short tandem repeat) loci
- 13 million offender profiles, 3 million arrestee profiles
- Aided over 390,000 investigations
- False matches are rare, but occur more frequently in closely related individuals



By Chemical Science & Technology Laboratory, National Institute of Standards and Technology – Wikipedia

U.S. arrestee collection laws as of 2017



By Majora – Wikipedia

Beyond database hits: reconstructing appearance with DNA

"Current DNA-based appearance prediction includes group-specific traits such as eye colour, hair colour and age with categorical prediction accuracies suitable for practical applications, and additional group-specific traits such as skin colour, hair morphology or baldness may follow. Individual-specific DNA-based facial morphology prediction would be most appreciated for finding unknown persons, but is currently beyond our level of genetic knowledge."

“IrisPlex” eye color prediction from DNA alone

Review Article

Improving human forensics through advances in genetics, genomics and molecular biology

Manfred Kayser ✉ & Peter de Knijff

Nature Reviews Genetics **12**, 179–192 (2011) | Published: 18 February 2011

2018: police capture criminal by using a fake DNA profile

- Uploaded a DNA sample from 1980 to the site GEDmatch
- Several relatives matched, which led the investigators to an address
- Joseph James DeAngelo, then 72, convicted as the “Golden State Killer”

Joseph James DeAngelo was arrested in Citrus Heights on April 24.

“Everything else up to this time had failed,” Holes said. “For 44 years, law enforcement has been trying to solve this case. No other case has had more resources poured into it in the history of California. I was just stunned.”

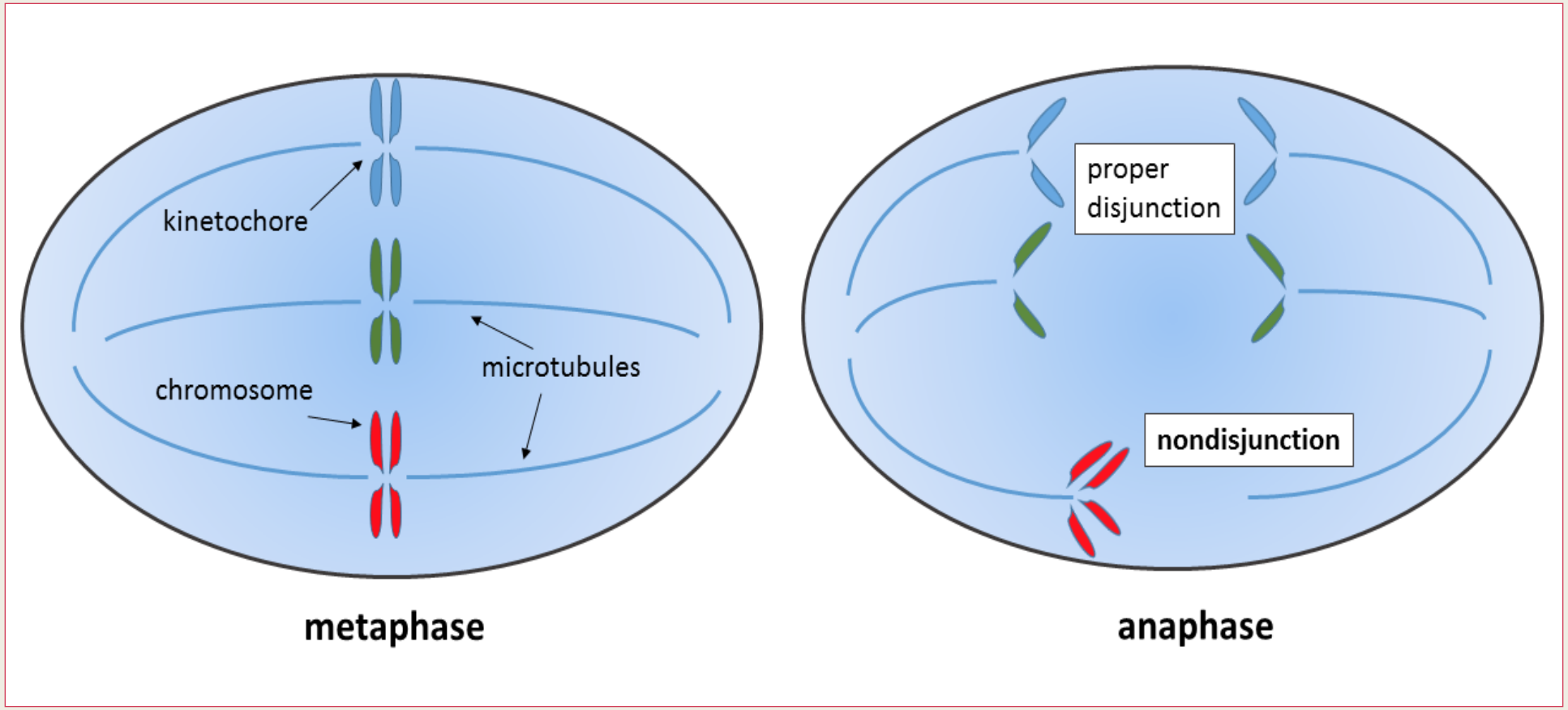
Washington Post

Discussion Questions in groups

Prenatal Testing Background

Causes of aneuploidy

abnormal number of chromosomes (usually 1 or 3 instead of 2)



"Mitotic nondisjunction" by Wpeissner - https://commons.wikimedia.org

Motivation: aneuploidy testing

Autosomal chromosomes:

- Trisomy 13: Patau syndrome
- Trisomy 18: Edwards syndrome
- Trisomy 21: Down syndrome

Motivation: aneuploidy testing

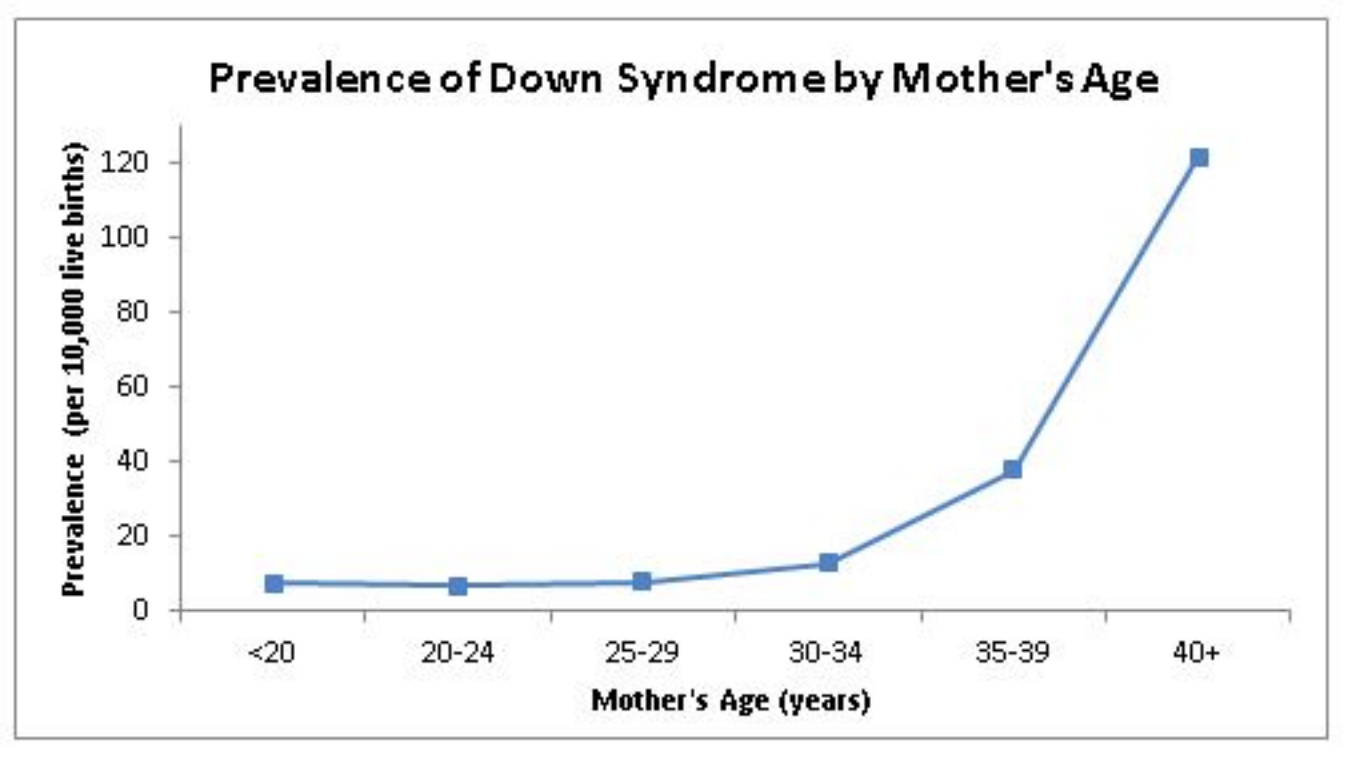
Autosomal chromosomes:

- Trisomy 13: Patau syndrome
- Trisomy 18: Edwards syndrome
- Trisomy 21: Down syndrome

Sex chromosomes:

- X0 (one X chromosome): Turner's syndrome
- XYY or XXX: normal male or female phenotype
- XXY: Klinefelter Syndrome

Older women are more at risk



Current testing procedures

- cFTS (combined First-Trimester Screening), 12-14 weeks
  - *Looks at biomarkers and other data*

- Amniocentesis (invasive)
  - *16-22 weeks, uses karyotyping to determine anueploidy*

- 1997: proof of fetal DNA in maternal blood (Y-chromosome)

- ≈2011: companies offer Non-Invasive Prenatal Testing (NIPT)
  - *Verinata, Harmony, NIFTY*
  - *All using next-generation sequencing*

Basic Procedure

1) Sample maternal blood

- contains cell-free fetal DNA (cffDNA)

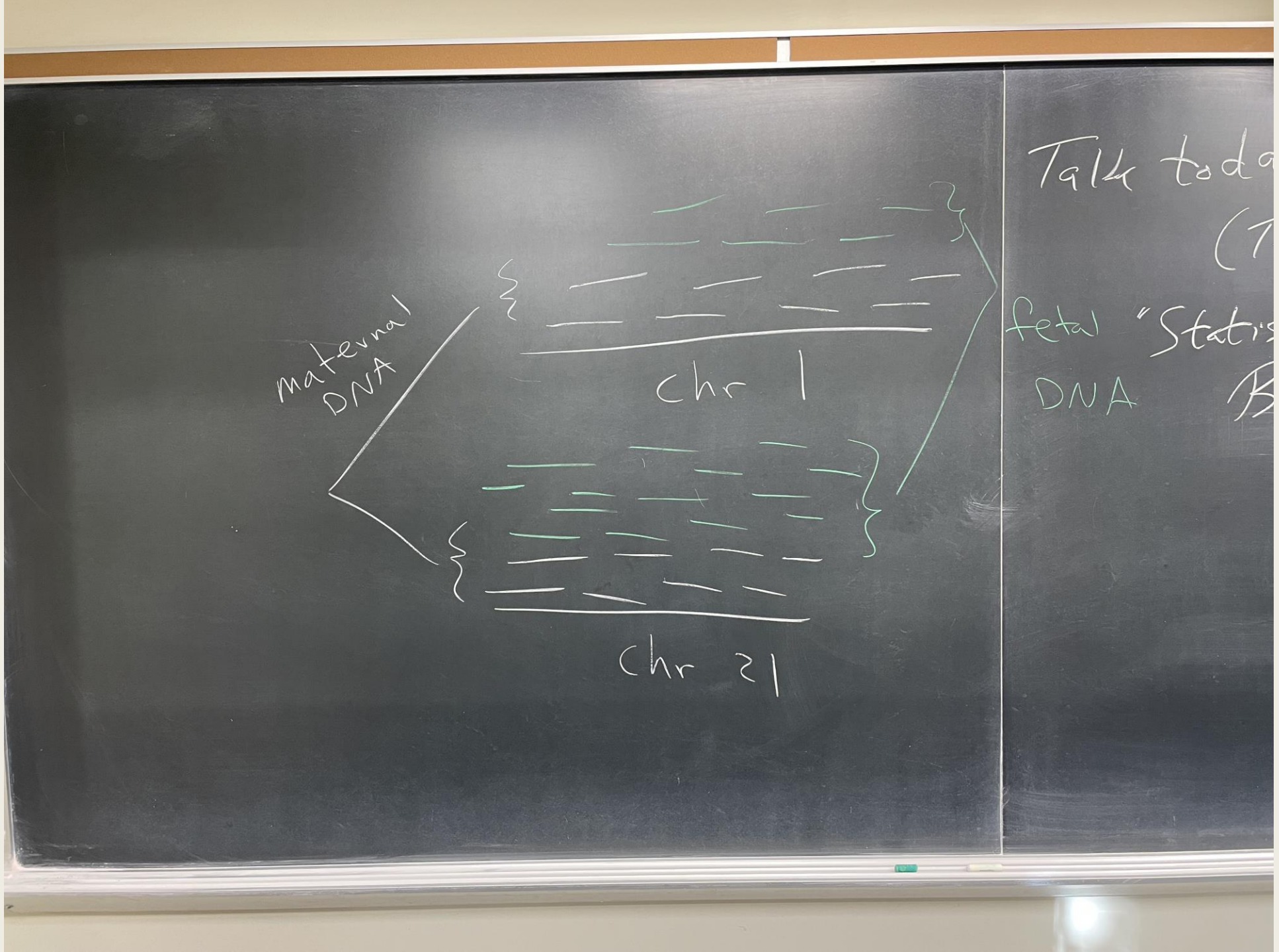
2) Low-coverage sequencing (0.1x - 4x)

3) Read alignment/mapping to human genome (using BWA or similar)

- save reads that map uniquely

4) Compute coverage for each chromosome

5) *t*-test for aneuploidy (coverage differences)


maternal DNA
Chr 1
Chr 21
Talk toda
(T
fetal "Statis
DNA

Direct to Consumer (DTC) testing

verinata

Continuous innovation

Increased safety and peace of mind for your patients

Swift acceptance of the verifi® prenatal test has made a world of difference to high-risk patients across the country:

- SAFE—Routine blood draw, **just one tube** (7-10 ml)
- ACCURATE—Directly analyzes cell-free fetal DNA with our proprietary SAFeR™ algorithm
- EASY—Test as early as 10 weeks, no limitations in reference to patient ethnicity, BMI, ART, or egg donor cases
- FAST—Results reported in 3-6 business days after sample receipt

The basic verifi® test detects:

- T21 (Down syndrome)
- T18 (Edwards syndrome)
- T13 (Patau syndrome)

Trisomy 21, 18, 13

Now a wider option is available for sex chromosomes at no extra charge:

- Monosomy X (MX; Turner syndrome)
- XXX (Triple X)
- XXY (Klinefelter syndrome)
- XYY (Jacobs syndrome)
- Fetal sex (XX or XY)—aids in stratifying the risk for X-linked disorders such as hemophilia, Duchenne muscular dystrophy, or cases of ambiguous genitalia, such as congenital adrenal hyperplasia

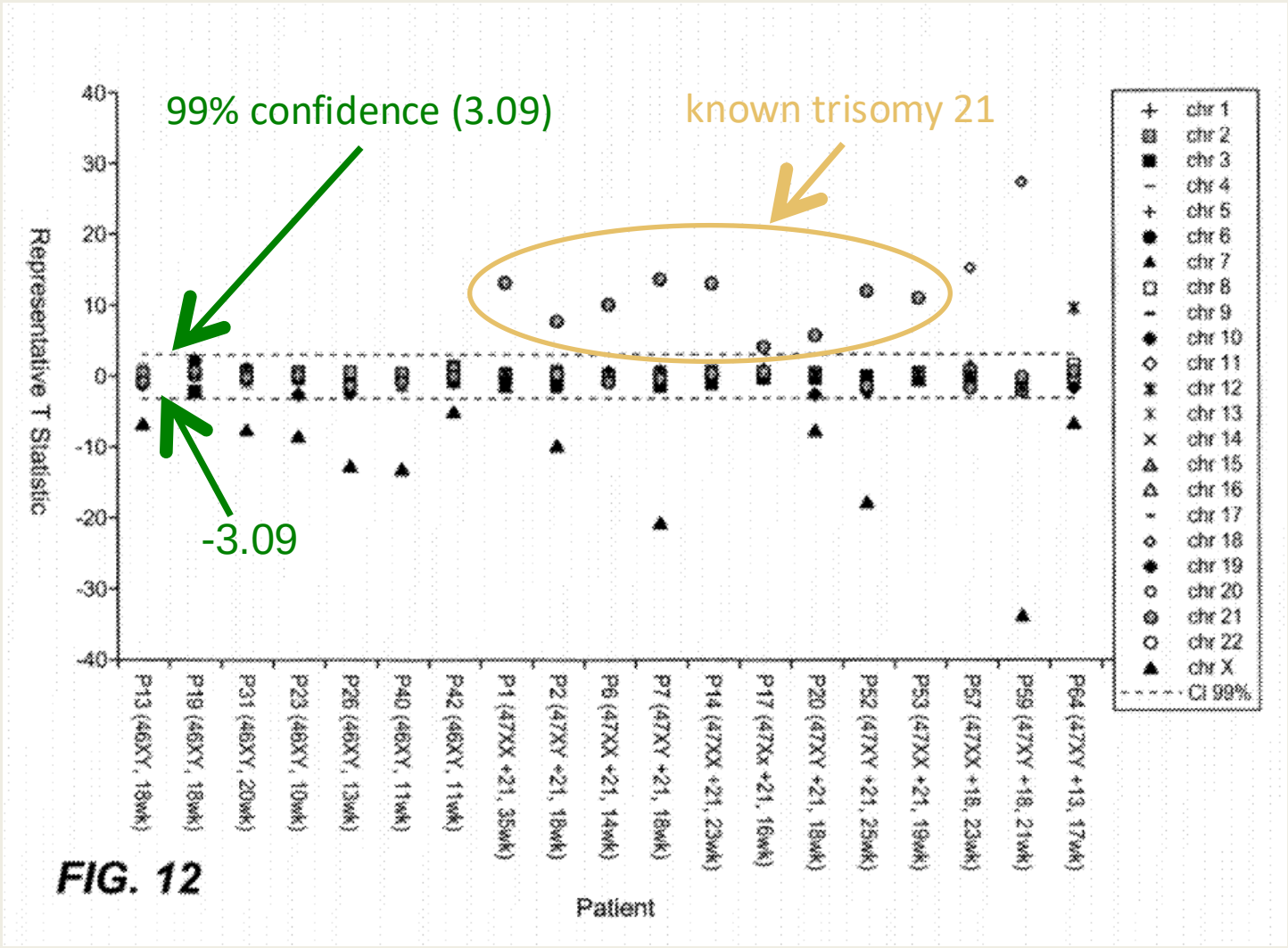

99% confidence (3.09)
known trisomy 21
-3.09
Representative T Statistic
40
30
20
10
0
-10
-20
-30
-40
chr 1
chr 2
chr 3
chr 4
chr 5
chr 6
chr 7
chr 8
chr 9
chr 10
chr 11
chr 12
chr 13
chr 14
chr 15
chr 16
chr 17
chr 18
chr 19
chr 20
chr 21
chr 22
chr X
CI 99%
P13 (46XY, 18wk)
P19 (46XY, 18wk)
P31 (46XY, 20wk)
P23 (46XY, 10wk)
P26 (46XY, 13wk)
P40 (46XY, 11wk)
P42 (46XY, 11wk)
P1 (47XX +21, 35wk)
P2 (47XY +21, 18wk)
P6 (47XX +21, 14wk)
P7 (47XY +21, 18wk)
P14 (47XX +21, 23wk)
P17 (47Xx +21, 16wk)
P20 (47XY +21, 18wk)
P52 (47XY +21, 25wk)
P53 (47XX +21, 19wk)
P57 (47XX +18, 23wk)
P59 (47XY +18, 21wk)
P64 (47XY +13, 17wk)
Patient

FIG. 12

verinata

DNA Sequencing versus Standard Prenatal Aneuploidy Screening

Diana W. Bianchi, M.D., R. Lamar Parker, M.D., Jeffrey Wentworth, M.D., Rajeevi Madankumar, M.D., Craig Saffer, M.D., Anita F. Das, Ph.D., Joseph A. Craig, M.D., Darya I. Chudova, Ph.D., Patricia L. Devers, M.S., C.G.C., Keith W. Jones, Ph.D., Kelly Oliver, B.S., Richard P. Rava, Ph.D., and Amy J. Sehnert, M.D. for the CARE Study Group

N Engl J Med 2014; 370:799-808 | February 27, 2014 | DOI: 10.1056/NEJMoa1311037

	false positive (sequencing)	false positive (standard)	false negative (sequencing)	num positives	PPV (sequencing)	PPV (standard)
T21	0.30%	3.60%	0	5	45.50%	4.20%
T18	0.20%	0.60%	0	2	40.00%	8.30%
T13			0	1		

Probability Considerations

Clinical Trials Example

- Disease affects 1/100 people: *P*(disease) = 0.01
- Test for the disease with 90% accuracy
  - *P(positive|disease) = 0.9*
  - *P(negative|healthy) = 0.9*

Clinical Trials Example

- Disease affects 1/100 people: $P(\text{disease}) = 0.01$
- Test for the disease with 90% accuracy
  - $P(\text{positive}|\text{disease}) = 0.9$
  - $P(\text{negative}|\text{healthy}) = 0.9$

$$P(\text{disease}|\text{positive}) = \frac{P(\text{positive}|\text{disease})\,P(\text{disease})}{P(\text{positive})}$$

$P(\text{disease}) = \frac{1}{100}$

$P(P|D) = \frac{9}{10}$

$P(N|H) = \frac{9}{10}$

$\underbrace{PPV}_{\text{positive predictive value}} = P(D|P)$

$= \dfrac{P(P|D)P(D)}{P(P|D)P(D) + P(H)P(P|H)}$ } $P(P)$ ↑ pos test

$= \dfrac{\frac{9}{10} \cdot \frac{1}{100}}{\frac{9}{10} \cdot \frac{1}{100} + \frac{99}{100}\left(\frac{10}{10} - \frac{9}{10}\right)}$

$P(P,H) + P(P,D)$

$= \dfrac{9}{9+99} = \dfrac{9}{108} = \dfrac{1}{12} = \boxed{8.25\%}$

lots of false positives!

Clinical Trials Example

- Disease affects 1/100 people: $P(\text{disease}) = 0.01$
- Test for the disease with 90% accuracy
  - $P(\text{positive}|\text{disease}) = 0.9$
  - $P(\text{negative}|\text{healthy}) = 0.9$

$$P(\text{disease}|\text{positive}) = \frac{P(\text{positive}|\text{disease})\,P(\text{disease})}{P(\text{positive})}$$

$$\approx 8.3\%$$

Bayesian Model

Input data are read counts for each chromosome (1,2,…,n):

$$q_1, q_2, \cdots, q_n = \vec{q}$$

$$\sum_{i=1}^{n} q_i = N$$

Bayesian Model

Input data are read counts for each chromosome (1,2,…,n):

$$q_1, q_2, \cdots, q_n = \vec{q}$$

$$\sum_{i=1}^{n} q_i = N$$

Goal:

$$\mathbb{P}(T_{21}|\vec{q}\,) = \frac{\mathbb{P}(\vec{q}\,|T_{21}) \cdot \mathbb{P}(T_{21})}{\mathbb{P}(\vec{q}\,)}$$

$$= \frac{\mathbb{P}(\vec{q}\,|T_{21}) \cdot \mathbb{P}(T_{21})}{\mathbb{P}(\vec{q}\,|T_{21}) \cdot \mathbb{P}(T_{21}) + \mathbb{P}(\vec{q}\,|T_{21}^{C}) \cdot \mathbb{P}(T_{21}^{C})}$$

Bayesian Model

Input data are read counts for each chromosome (1,2,…,n):

$$q_1, q_2, \cdots, q_n = \vec{q}$$

$$\sum_{i=1}^{n} q_i = N$$

Goal:

$$\mathbb{P}(T_{21}|\vec{q}\,) = \frac{\mathbb{P}(\vec{q}\,|T_{21}) \cdot \mathbb{P}(T_{21})}{\mathbb{P}(\vec{q}\,)}$$

$$= \frac{\mathbb{P}(\vec{q}\,|T_{21}) \cdot \mathbb{P}(T_{21})}{\mathbb{P}(\vec{q}\,|T_{21}) \cdot \mathbb{P}(T_{21}) + \mathbb{P}(\vec{q}\,|T_{21}^{C}) \cdot \mathbb{P}(T_{21}^{C})}$$

Prior:

$P(T_{21})$

Maternal Age	Trisomy 21	All Trisomies
20	1 in 1,667	1 in 526
21	1 in 1,429	1 in 526
22	1 in 1,429	1 in 500
23	1 in 1,429	1 in 500
24	1 in 1,250	1 in 476
25	1 in 1,250	1 in 476
26	1 in 1,176	1 in 476
27	1 in 1,111	1 in 455
28	1 in 1,053	1 in 435
29	1 in 1,000	1 in 417
30	1 in 952	1 in 384
31	1 in 909	1 in 384
32	1 in 769	1 in 323
33	1 in 625	1 in 286
34	1 in 500	1 in 238
35	1 in 385	1 in 192
36	1 in 294	1 in 156
37	1 in 227	1 in 127
38	1 in 175	1 in 102
39	1 in 137	1 in 83
40	1 in 106	1 in 66
41	1 in 82	1 in 53
42	1 in 64	1 in 42
43	1 in 50	1 in 33
44	1 in 38	1 in 26
45	1 in 30	1 in 21
46	1 in 23	1 in 16
47	1 in 18	1 in 13
48	1 in 14	1 in 10
49	1 in 11	1 in 8

Research paper

Refusing to provide a prenatal test: can it ever be ethical?

Rony E Duncan, research officer,[1] Bennett Foddy, PhD candidate,[2] and Martin B Delatycki, director[3]

"A couple in which the man carries the mutation for Huntington's disease request prenatal testing during their first pregnancy. Though they would not terminate an affected pregnancy, they would like the information. There is no treatment available that can change the course of the disease so the diagnosis will not result in medical benefit for the child."

Viewpoint 1: not to test
Viewpoint 2: to test

Discussion Questions

- Form small groups
- Discuss questions below (or anything else you find interesting/relevant about this topic)
- Choose a representative to mention an interesting part of your discussion to the class

1) Which side of the article (about Huntington's Disease) do you find most compelling?

2) Considering the potential of NIPT to resolve the entire fetal genome, what would be your recommendation about how to use this technology?