Medical Data Privacy and Ethics in the Age of Artificial Intelligence

Lecture 10: Genomic Data Sharing and Risks

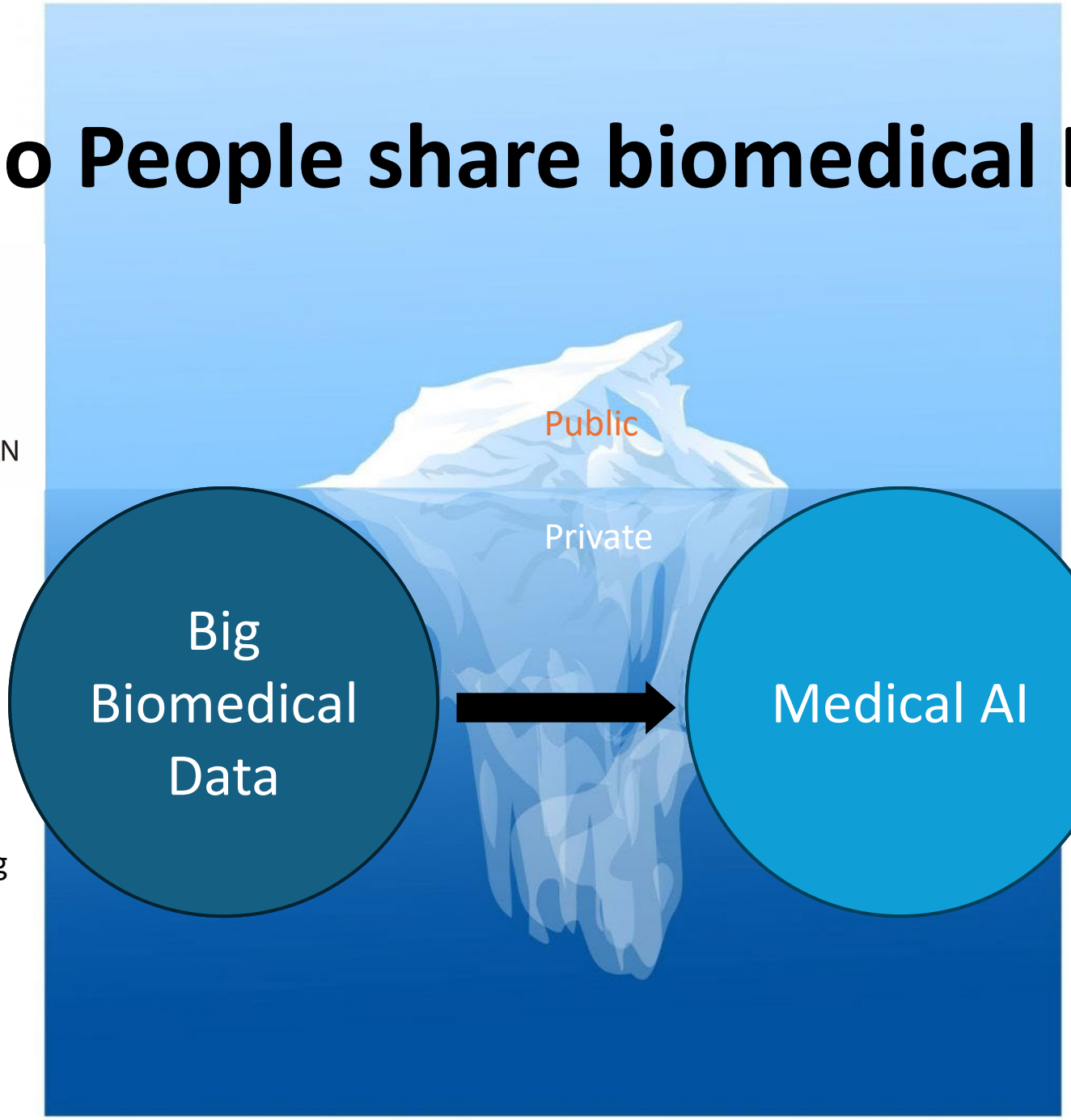
Zhiyu Wan, PhD (wanzhy@shanghaitech.edu.cn)
Assistant Professor of Biomedical Engineering
ShanghaiTech University

April 2, 2025

Learning Objectives of This Lecture

- Know the benefits of sharing genomic data
  - Advancing research and scientific knowledge
  - Help curing diseases
  - Genome-wide association studies
  - Basic research and discovery
  - Reproducibility
  - Genealogical search
- Know the risks of sharing genomic data
  - Re-identification attacks
  - Membership inference attacks
  - Reconstruction attacks
  - Familial attacks

Why do People share biomedical Data?



CHINA KADOORIE BIOBANK
中国慢性病前瞻性研究
512,000

BIOBANK JAPAN
270,000

华大基因 BGI
3,000,000
Non-invasive prenatal testing

Vanderbilt BioVU
278,000

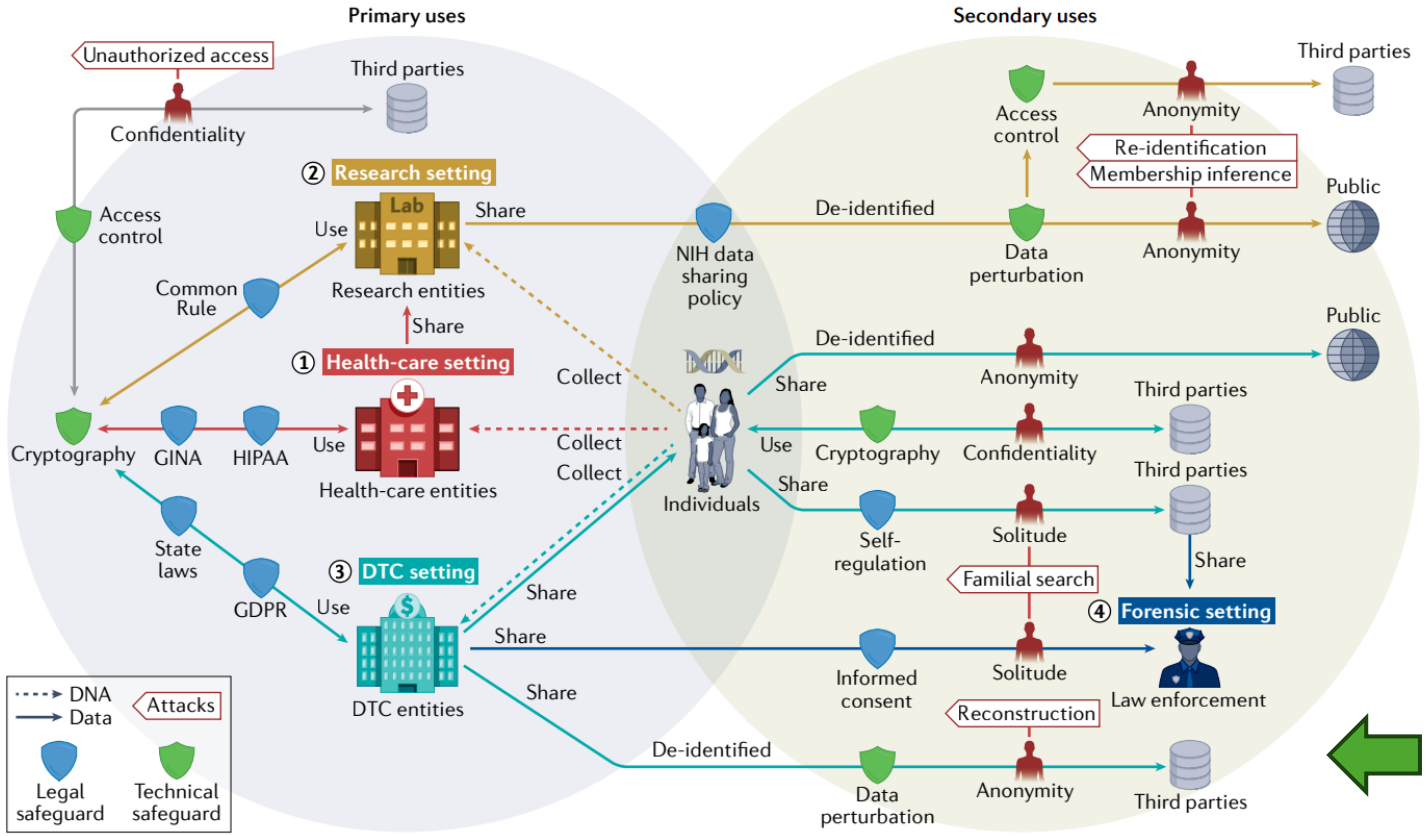
uk biobank
500,000

All of Us℠ The Precision Medicine Initiative®
THE FUTURE OF HEALTH BEGINS WITH YOU
1,000,000

23andMe®
12,000,000
(5,000,000 share)

Groups working on genomic privacy

Genomic Data Protection Methods



nature reviews genetics

Explore content · About the journal · Publish with us

nature › nature reviews genetics › review articles › article

Review Article | Published: 04 March 2022

Sociotechnical safeguards for genomic data privacy

Zhiyu Wan, James W. Hazel, Ellen Wright Clayton, Yevgeniy Vorobeychik, Murat Kantarcioglu & Bradley A. Malin

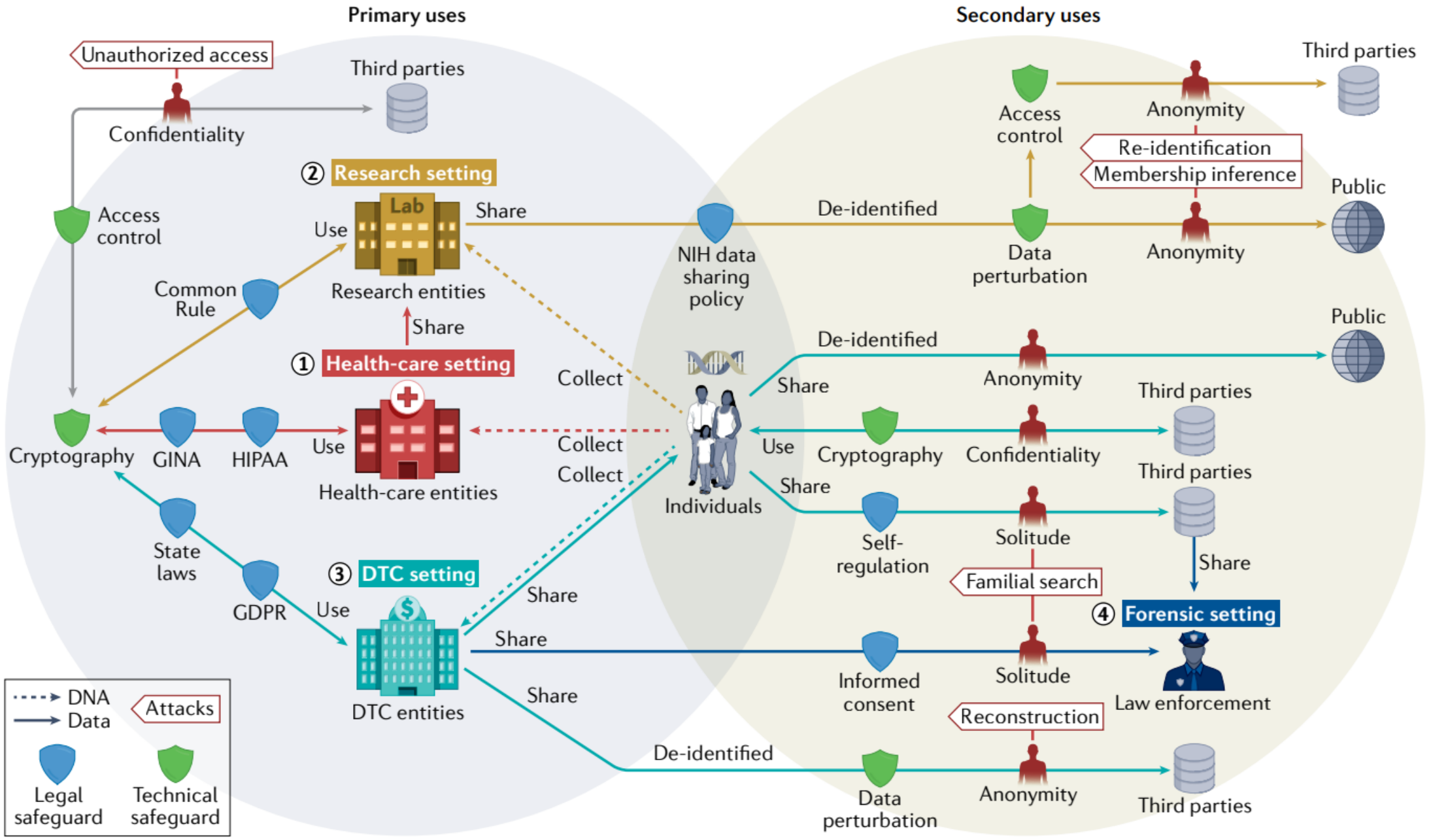
Nature Reviews Genetics **23**, 429–445 (2022) | Cite this article

23k Accesses | **53** Citations | **106** Altmetric | Metrics

Abstract

Recent developments in a variety of sectors, including health care, research and the direct-to-consumer industry, have led to a dramatic increase in the amount of genomic data that are collected, used and shared. This state of affairs raises new and challenging concerns for personal privacy, both legally and technically. This Review appraises existing and emerging threats to genomic data privacy and discusses how well current legal frameworks and technical safeguards mitigate these concerns. It concludes with a discussion of remaining and emerging challenges and illustrates possible solutions that can balance protecting privacy and realizing the benefits that result from the sharing of genetic information.

Wan Z, Hazel JW, Clayton EW, Vorobeychik Y, Kantarcioglu M, Malin BA. Sociotechnical safeguards for genomic data privacy. *Nature Reviews Genetics*. 2022 Jul;23(7):429-45.



Wan Z, Hazel JW, Clayton EW, Vorobeychik Y, Kantarcioglu M, Malin BA. Sociotechnical safeguards for genomic data privacy. *Nature Reviews Genetics*. 2022 Jul;23(7):429-45.

Advancing research and scientific knowledge

- The foundation of biomedical and healthcare studies relies on the data collected.
- **Genome sequencing technologies** can help identify genetic variations in humans that cause or influence diseases ranging from Huntington disease to cancer.
- Functional genomics assays, such as RNA-Seq, can help us understand genetic activity that is different in disease and health.
- It is extremely difficult and expensive to collect all types of genomic data at one site or institution.
- Therefore, data sharing across institutions and labs is essential for data integration.

Responsible Genomic Data Sharing Challenges and Approaches Ch.1

Help curing diseases

- **Rare disease**
  - In the US, a disease is characterized as rare if it affects fewer than 200,000 Americans at any given time.
  - There are 30 million people in the United States and 350 million people in the world currently suffering from a rare disease.
  - There are currently more than 6000 rare diseases identified and cataloged in the world.
  - 80% of rare diseases have been determined to have a genetic origin.
  - There is great value in collecting and sharing genetic data on a worldwide scale in the context of rare diseases.
  - By means of pharmacogenomics, one can look at how genetic variation affects the response of a patient to a drug.

Help curing diseases (Cont.)

- **Cancer**
  - The inheritance and accumulation of mutations in the genome is the main cause of many cancer types.
  - The current situation is that such data are confined within the particular hospital database and not shared widely with the research community
  - The National Cancer Institute Genomic Data Commons (GDC), launched in 2016
  - There are various genetic indicators that predict the probability of effectiveness of immunotherapy treatment
  - The Pediatric Cancer Data Commons (PCDC)

Basic research and discovery

- **International HapMap project**, started in 2002
  - develop a haplotype map of the human genome to provide a resource for researchers to find disease associating genes and their response to drugs.
- **1000 Genomes Project**, first released in 2008
  - 2504 samples
  - finding most genetic variants with frequencies of at least 1% in the populations studied using some of the samples from the HapMap project
- **The Cancer Genome Atlas (TCGA)**, initiated in December 2005
  - to catalog the genomic changes underlying multiple cancer types
  - focused on three different cancer types: brain, lung, and ovarian
  - over $300 million in total funding

Basic research and discovery (cont.)

- **UK Biobank**, established in 2006
  - 500,000 volunteers between the ages of 40 and 69
- **ENCODE, the encyclopedia of DNA elements**, is a public research consortium supported by The National Human Genome Research Institute (NHGRI), started in 2003.
  - to identify the functional elements in the human and mouse genome beyond coding sequences.
- **Genotype-Tissue Expression (GTEx) project**, launched in 2010 by the NIH
  - to investigate how variation in the human genome affects tissue expression.

Reproducibility

- In a poll launched by Nature in 2016, 70% of 1500 scientists claimed that they failed to reproduce at least one other publication's experiment.
- Although access to data is not the only reason causing replication problems, in computational sciences, encouraging access to data will increase reproducibility.

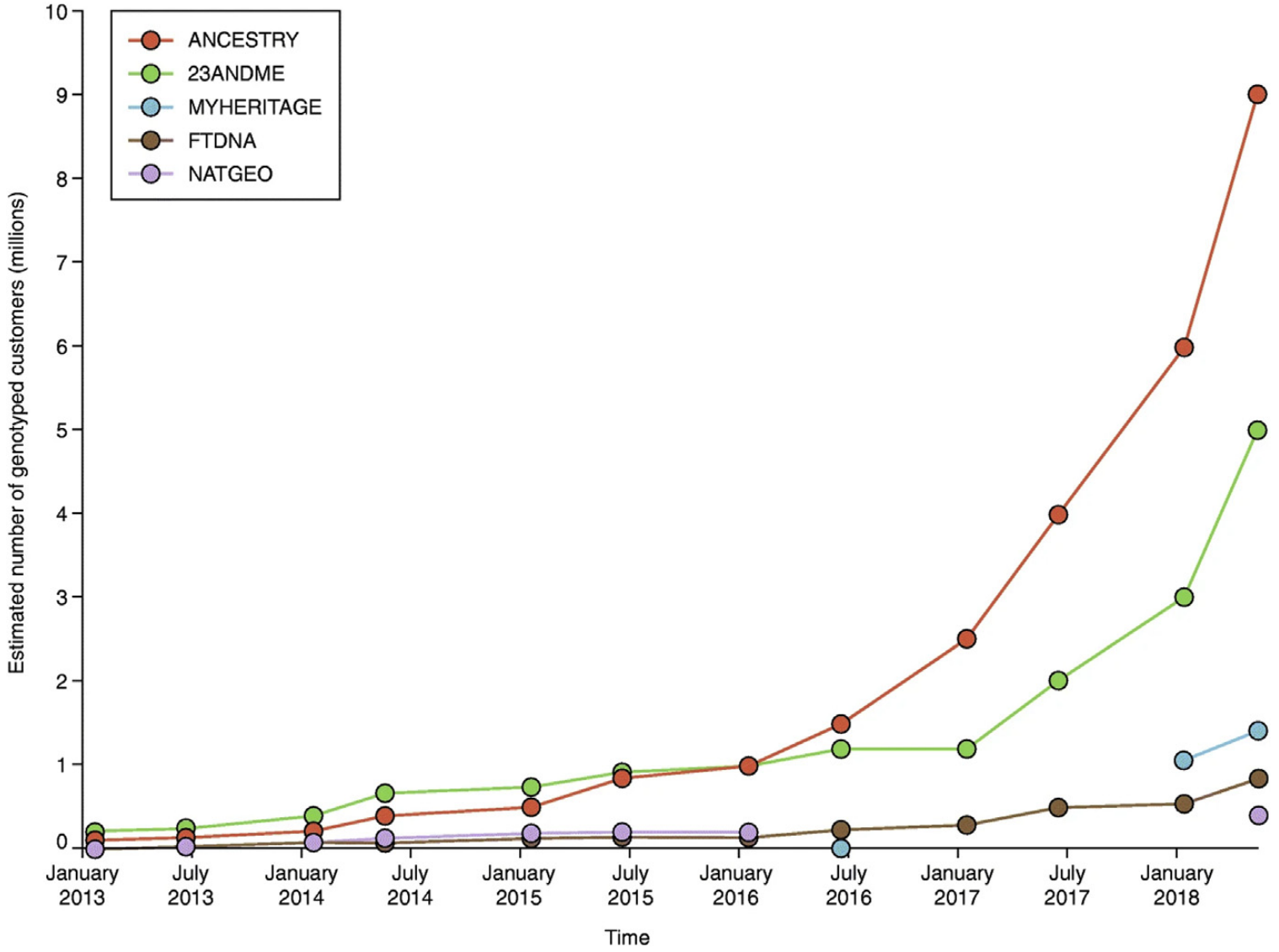
Public's views on genomic data sharing

- **Patients or individuals who are extremely sick** do not seem to think about who owns and controls their genomic data. They are interested instead on sharing their data quickly and with multiple institutions in case one can find novel ways of treating their medical condition.
- On the other hand, **healthy individuals** tend to think more about their data ownership, and the potential for leaking their identifying information.

Direct-to-consumer genetic companies

- **Family Tree DNA**, started in 2000
- **23andMe**, started in 2000
  - sell its kit at $99 starting December of 2012
  - Reaches 1-million customers in 2015.
  - In 2015, it gained FDA approval for a genetic test to predict Bloom syndrome
  - In 2017, it received FDA approval to sell a genetic risk test. The test provided information on 10 health conditions, most notably, Parkinson's disease and late-onset Alzheimer's disease.
  - As of early 2018, they had tested over 3 million customer samples,
- **AncestryDNA**, started in 2012
  - Provides ancestry data
  - low price, $59, and access to genealogy tools
  - Reaches 1-million customers served by the end of 2015
  - Reaches an astounding 7 million customers by the beginning of 2018

Responsible Genomic Data Sharing Challenges and Approaches Ch.4



Autosomal DNA database growth

File format

- Variant call format (VCF) is a tab-delimited text file format storing gene sequence variations.
  - It was originally developed to support the 1000 Genomes Project.

##fileformat=VCFv4.1
##FILTER=<ID=PASS,Description="All filters passed">
##fileDate=20150218
##reference=ftp://ftp.1000genomes.ebi.ac.uk//vol1/ftp/technical/reference/phase2_reference_assembly_sequence/hs37d5.fa.gz
##source=1000GenomesPhase3Pipeline
##contig=<ID=1,assembly=b37,length=249250621>
##...
##ALT=<ID=DEL,Description="Deletion">
##...
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##...
##INFO=<ID=AC,Number=A,Type=Integer,Description="Total number of alternate alleles in called genotypes">
##INFO=<ID=AF,Number=A,Type=Float,Description="Estimated allele frequency in the range (0,1)">
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of samples with data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total read depth; only low coverage data were counted towards the DP, exome data were not used">
##INFO=<ID=VT,Number=.,Type=String,Description="indicates what type of variant the line represents">
##...

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	HG00096	HG00097	HG00099
22	16050075	rs587697622	A	G	100	PASS	AC=1;AF=0.000199681;AN=5008;NS=2504;DP=8012;VT=SNP	GT	0|0	0|0	0|0
22	16050115	rs587755077	G	A	100	PASS	AC=32;AF=0.00638978;AN=5008;NS=2504;DP=11468;VT=SNP	GT	0|0	0|0	0|0
22	16050213	rs587654921	C	T	100	PASS	AC=38;AF=0.00758786;AN=5008;NS=2504;DP=15092;VT=SNP	GT	0|0	0|0	0|0
22	16050319	rs587712275	C	T	100	PASS	AC=1;AF=0.000199681;AN=5008;NS=2504;DP=22609;VT=SNP	GT	0|0	0|0	0|0
22	16050527	rs587769434	C	A	100	PASS	AC=1;AF=0.000199681;AN=5008;NS=2504;DP=23591;VT=SNP	GT	0|0	0|0	0|0
22	16050568	rs587638893	C	A	100	PASS	AC=2;AF=0.000399361;AN=5008;NS=2504;DP=21258;VT=SNP	GT	0|0	0|0	0|0
22	16050607	rs587720402	G	A	100	PASS	AC=5;AF=0.000998403;AN=5008;NS=2504;DP=20274;VT=SNP	GT	0|0	0|0	0|0
22	16050627	rs587593704	G	T	100	PASS	AC=2;AF=0.000399361;AN=5008;NS=2504;DP=21022;VT=SNP	GT	0|0	0|0	0|0

A sample VCF file.

DTC format

- The human genetic data formats used by DTC companies are very similar. They are all tab-delimited text files.

#Genetic data is provided below as five TAB delimited columns. Each line							
#corresponds to a SNP. Column one provides the SNP identifier (rsID where							
#possible). Columns two and three contain the chromosome and basepair position							
#of the SNP using human reference build 37.1 coordinates. Columns four and five							
#contain the two alleles observed at this SNP (genotype). The genotype is reported							
#on the forward (+) strand with respect to the human reference.							
rsid	chromosome	position	allele1	allele2			
rs587697622	22	16050075	A	A			
rs587755077	22	16050115	G	G			
rs587654921	22	16050213	C	C			
rs587712275	22	16050319	C	C			
rs587769434	22	16050527	C	C			
rs587638893	22	16050568	C	C			
rs587720402	22	16050607	G	G			
rs587593704	22	16050627	G	G			

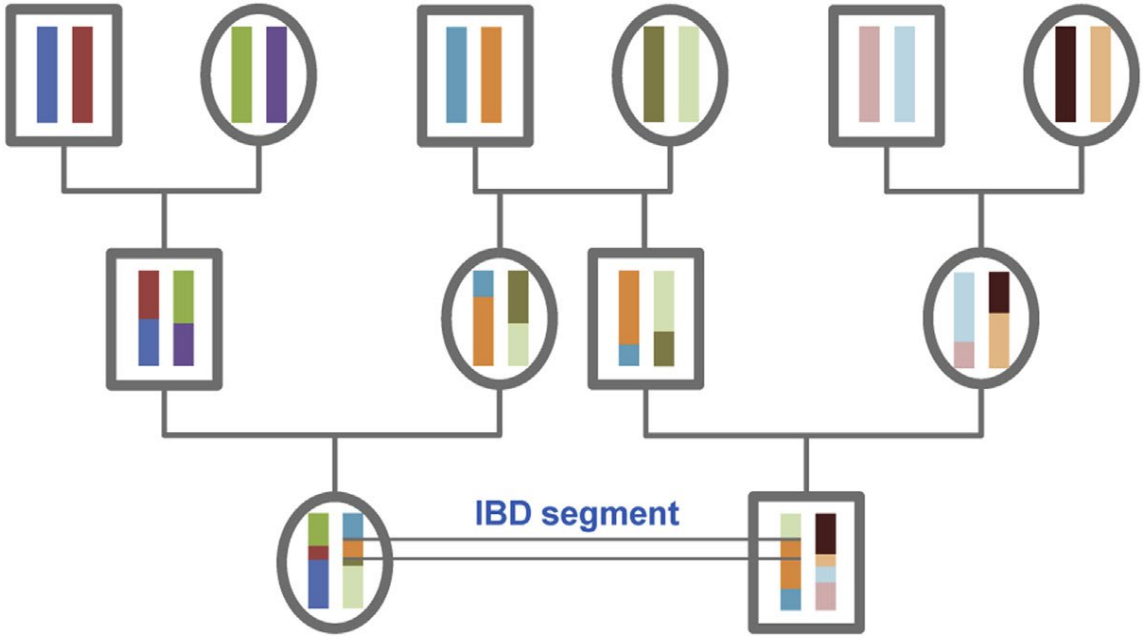
A sample DTC file.

Compressed format

- PLINK format (BED)
  - PLINK is a genetic data toolset developed by Shaun Purcell et al. The goal of PLINK format is to handle large datasets and perform analysis on them in a computationally efficient manner.
- UK Biobank format (BGEN)
  - BGEN is a data format storing either typed or imputed genotype data with the imputed genotype probability (i.e., dosage value). It was developed by Gavin Band and Jonathan Marchini.
- GDS format
  - Xiuwen Zheng et al. propose an array-oriented data format to store genome-wide data, named Genomic Data Structure (GDS).

What is IBD

- Identical by descent (or identity by descent) (IBD) is a biological terminology proposed by Gustave Malecot.
- Originally, it serves as an indication of two homologous alleles descending from a common ancestor.
- Contemporarily, from the genome-scale perspective, IBD is redefined as two homologous chromosome segments being inherited from a common ancestor.



The average and the range of shared IBD segments per relationship

Relationship group (cluster)	Average shared IBD (percentage)	Range of shared IBD (percentage) [112]	Average shared IBD (centiMorgan)	Range of shared IBD (centiMorgan) [126]
Identical twin	100%	Not available	6800.00 [127]	Not available
Parent, child	50%	Variable	3400.00	Variable
Sibling	50%	Variable	2550.00 [125]	2209.00–3384.00
Grand parent, aunt (or uncle), niece (or nephew), grand child	25%	Variable	1700.00	1294.00–2230.00
Great grand parent, great aunt (or uncle), first cousin, great niece (or nephew), great grand child	12.5%	7.31%–13.80%	850.00	486.00–1761.00
second great grand parent, great grand aunt (or uncle), first cousin once removed, great grand niece (or nephew), second great grand child	6.25%	3.30%–8.51%	425.00	131.00–851.00
Third great grand parent, second great grand aunt (or uncle), first cousin twice removed, second cousin, second great grand niece (or nephew), third great grand child	3.125%	2.85%–5.04%	212.50	47.00–517.00

Genealogical search

- Genealogy search involves constructing a family tree, locating relatives, and sometimes providing historical records to the customer

Cousin relationship	Probability of having detectable shared DNA segment			
		In practice [129]		
	In Theory [128]	By 23andMe	By AncestryDNA	By FamilyTreeDNA
First cousin	100.00%	100.00%	100.00%	100.00%
Second cousin	100.00%	100.00%	100.00%	99.00%
Third cousin	97.70%	89.70%	98.00%	90.00%
Fourth cousin	69.30%	45.90%	71.00%	50.00%
Fifth cousin	30.20%	14.90%	32.00%	10.00%
Sixth cousin	10.10%	4.10%	11.00%	2.00%

Readings for the Next Week

- 1. Sankararaman S, Obozinski G, Jordan MI, Halperin E. **Genomic privacy and limits of individual detection in a pool**. *Nature genetics*. 2009 Sep;41(9):965-7.

- Optional
  - ❑ 2. Homer N, Szelinger S, Redman M, Duggan D, Tembe W, Muehling J, Pearson JV, Stephan DA, Nelson SF, Craig DW. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS genetics*. 2008 Aug 29;4(8):e1000167.
  - ❑ 3. 《*Responsible Genomic Data Sharing Challenges and Approaches*》Ch.3.

Feedback Survey

- One thing you learned or felt was valuable from today's class & reading

- Muddiest point: what, if anything, feels unclear, confusing or "muddy"

- https://www.wjx.cn/vm/hX0mIro.aspx