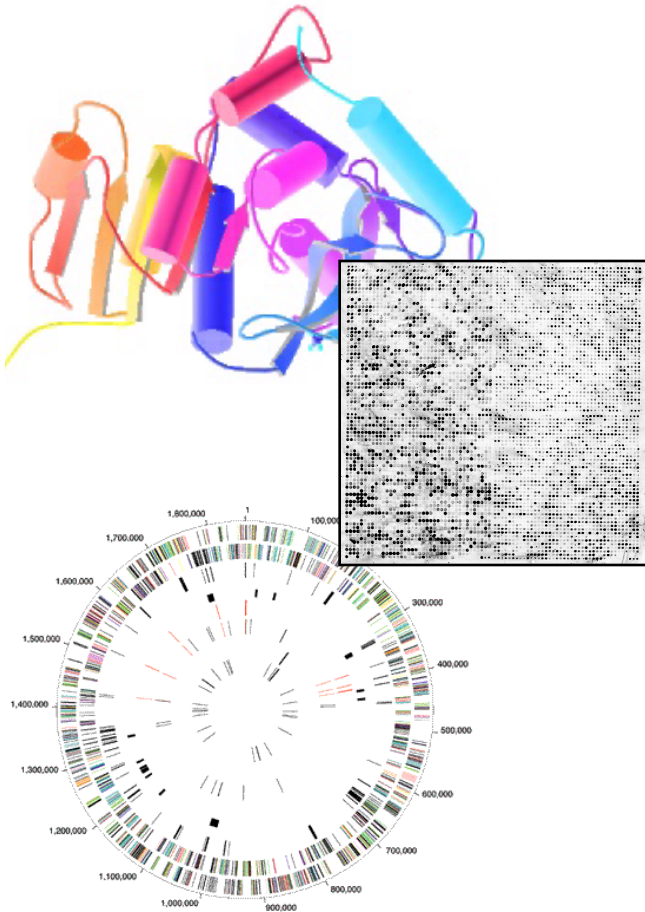


Biomed. Data Science:

# Privacy & Mining the “Data Exhaust”



Mark Gerstein, Yale University  
[gersteinlab.org/courses/452](http://gersteinlab.org/courses/452)  
(last edit in spring '19, pack #13)

# Privacy

# Genomics has similar "Big Data" Dilemma in the Rest of Society

- Sharing & "peer-production" is central to success of many new ventures, with the same risks as in genomics
  - **EG web search**: Large-scale mining essential



- We confront privacy risks every day we access the internet

**Specific  
Genomic  
Privacy  
Issues**

# From Personal Genomics to Genome Privacy



## Genome of an Individual

- Sequencing, analysis and interpretation
- Soon will become part of medical practice
- NCI: prevent, diagnose, and treat disease through personalized medicine

## Privacy risk

- **Identity tracing**
  - Link between unknown genome to a panel of individual through quasi-identifiers
- **Attribute Disclosure Attacks**
  - Known DNA sample to private data such as HIV status or drug abuse
- **Completion Techniques**
  - Impute sensitive information from partial genomic data (e.g. bipolar disorder risk)

# Genetic Data: Traditionally focus of genome privacy

## Participation Attacks

- Homer et al., 2008
  - Detecting an individual with known genotypes in a mixture
  - Allele frequency of a SNP in study vs. in population
  - If allele frequency of a SNP of an individual is more similar to the study than to the population → individual is in the mixture
  - Implications for GWAS
  - Shaped NIH GWAS sharing policy until '18

Snp	Allele Frequency ( $Y_{ij}$ )			Distance Measure	Interpretation at the given SNP	
	0.0	0.25	0.50			0.75
j					$=  1.0 - 0.25  -  1.0 - 0.75 $ $= 0.75 - 0.25$ $= 0.50$	most likely to be in the Mixture
j+1					$=  0.50 - 0.250  -  0.50 - 0.75 $ $= 0.25 - 0.25$ $= 0.00$	equally likely to be in the Mixture and in the Reference Population
j+2					$=  0.00 - 0.25  -  0.00 - 0.75 $ $= 0.25 - 0.75$ $= -0.50$	most likely to be in the Reference Population

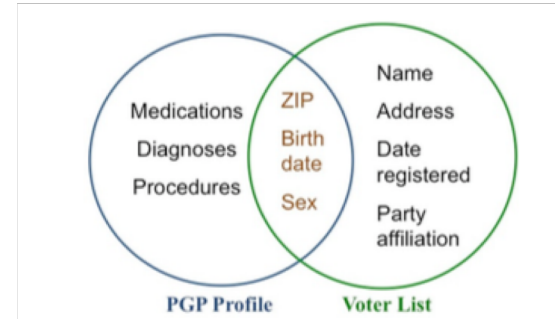
Person of Interest ( $Y_i$ )
Reference Population (Pop)
Mixture (M)

- Im et al., 2012
  - Regression coefficients of GWAS summary statistics
  - Created a statistics using the regression coefficients and compute the statistics for an individual
    - If individual falls within the reference distribution → no participation
    - If individual is in the tail → participation is inferred!

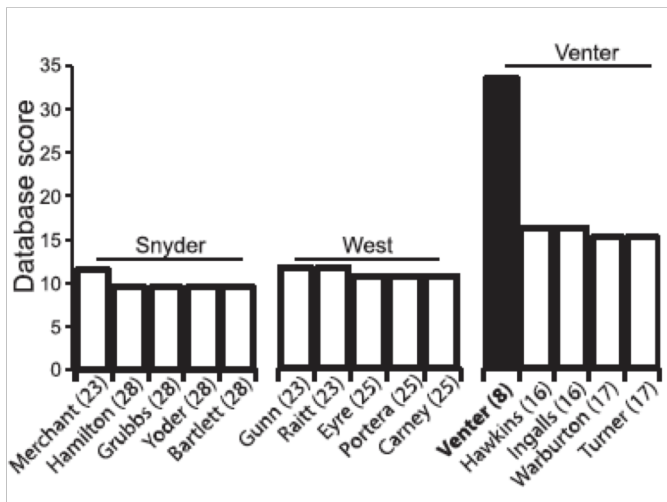
# Genetic Data: Traditionally focus of genome privacy

## Re-identification Attacks

- In 1997, Latanya Sweeney successfully identified then Massachusetts governor, William Weld to his medical records using publicly accessible records.
- significant impact on privacy centered policymaking including the health privacy legislation HIPAA
- publication of the experiment was rejected twenty times!
- In fact, a court ruling in *Southern Illinois v. Department of Public Health* barred her from publication and sharing of her



- Sweeney et al., 2013
  - Re-Identification by cross-referencing independent datasets



- Gymrek et al., 2013
  - Y-STRs with recreational genetic genealogy database
  - Majority of surnames are transferred from males in the family
  - STRs are unique to an individual, Y-STRs are unique to male individuals
  - Cross reference Y-STRs with genealogy database and find surname of the individual

# Genetic Data: Traditionally focus of genome privacy

## Characterization Attacks

### About the James Watson Genotype Viewer

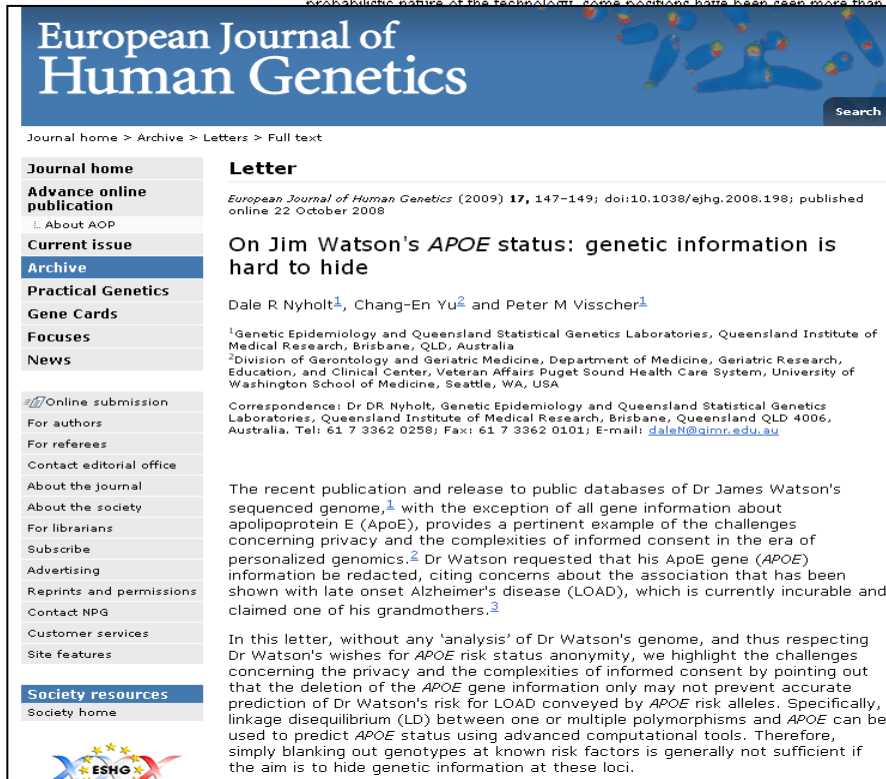
On May 31, 2007, Nobel Laureate James Watson received his personal genome sequence in a [ceremony at the Baylor College of Medicine](#). This genome sequence describes the six billion base pairs of DNA that James Watson received from his two parents, the unique combination of which are responsible for James Watson's genetic individuality. Dr. Watson is making his genome sequence available to the public in the hope that it will encourage the development of an era of "personalized medicine" when the information contained in our genomes is used to identify and prevent diseases to which we are genetically prone before they appear, and to create personalized medical therapies that have the maximum benefit and the minimum risk. This [simple browser](#) allows you to view the places where Watson's sequence is different from the "reference" human genome sequence, as well as to view the genes and some of the common diseases associated with them.

### What the Watson Sequence is

Dr. Watson's genome was sequenced at 6x coverage using [454 Life Sciences Technology](#). This means that each position on the genome was sequenced roughly six times. However, because of the probabilistic nature of the technology, some positions have been seen more than six times, and some less or not at all. The 454 technology produces short stretches of sequences called "reads" that are roughly 50,000 bases long, or 500 times the size of a 454 sequence. To interpret the Watson sequence, it was matched to the reference genome in 500 bp to gene-length pieces. The entire Watson sequence, with the exception of the ApoE gene, variants of which are associated with Alzheimer's disease, is available on the [NCBI Trace Repository](#), and will be available from many other web sites in the future.

These differences are called *variants* or *polymorphisms*. Because each of these differences involves only a single nucleotide change (SNP), they are called *single nucleotide polymorphisms*.

Each SNP found in Watson's sequence) are known as *alleles*. Each SNP has two possible alleles.



The screenshot shows the website for the European Journal of Human Genetics. The main article is titled "On Jim Watson's APOE status: genetic information is hard to hide" by Dale R Nyholt, Chang-En Yu, and Peter M Visscher. The article discusses the challenges of informed consent in the era of personalized genomics, specifically regarding the APOE gene and its association with late onset Alzheimer's disease (LOAD). The authors highlight the complexities of informed consent by pointing out that the deletion of the APOE gene information only may not prevent accurate prediction of Dr Watson's risk for LOAD conveyed by APOE risk alleles. They also mention that linkage disequilibrium (LD) between one or multiple polymorphisms and APOE can be used to predict APOE status using advanced computational tools. The article concludes by stating that simply blanking out genotypes at known risk factors is generally not sufficient if the aim is to hide genetic information at these loci.

Journal home > Archive > Letters > Full text

**Journal home** | **Letter**

**Advance online publication**  
L About AOP

**Current issue**

**Archive**

**Practical Genetics**

**Gene Cards**

**Focuses**

**News**

Online submission

For authors

For referees

Contact editorial office

About the journal

About the society

For librarians

Subscribe

Advertising

Reprints and permissions

Contact NPG

Customer services

Site features

**Society resources**  
Society home

ESHG



# Genetic Data: Now with a focus on forensic applications

LOCAL // CRIME

## Suspect in 1973 Palo Alto killing caught through DNA matchup

Science

REPORTS

Cite as: Y. Erlich *et al.*, *Science*  
10.1126/science.aau4832 (2018).

### Identity inference of genomic data using long-range familial searches

Yaniv Erlich<sup>1,2,3,4\*</sup>, Tal Shor<sup>1</sup>, Itsik Pe'er<sup>2,3</sup>, Shai Carmi<sup>5</sup>

<sup>1</sup>MyHeritage, Or Yehuda 6037606, Israel. <sup>2</sup>Department of Computer Science, Fu Foundation School of Engineering and Applied Sciences, Princeton University. <sup>3</sup>Department of Computational Biology and Bioinformatics (C2B2), Department of Systems Biology, Columbia University, New York, NY, USA. <sup>4</sup>Braun School of Public Health and Community Medicine, The Hebrew University of Jerusalem, Jerusalem, Israel. <sup>5</sup>Department of Computer Science, Princeton University, Princeton, NJ, USA.

\*Corresponding author. Email: erlichya@gmail.com

Cell

### Statistical Detection of Relatives Typed with Disjoint Forensic and Biomedical Loci

Graphical Abstract

GOAL:  
Test STR query profile against STR databases  
to determine whether it is a relative of a database profile



Authors

Jaehee Kim, Michael D. Edge,  
Bridget F.B. Algee-Hewitt, Jun Z. Li,  
Noah A. Rosenberg

# Tricky Privacy Considerations in Personal Genomics

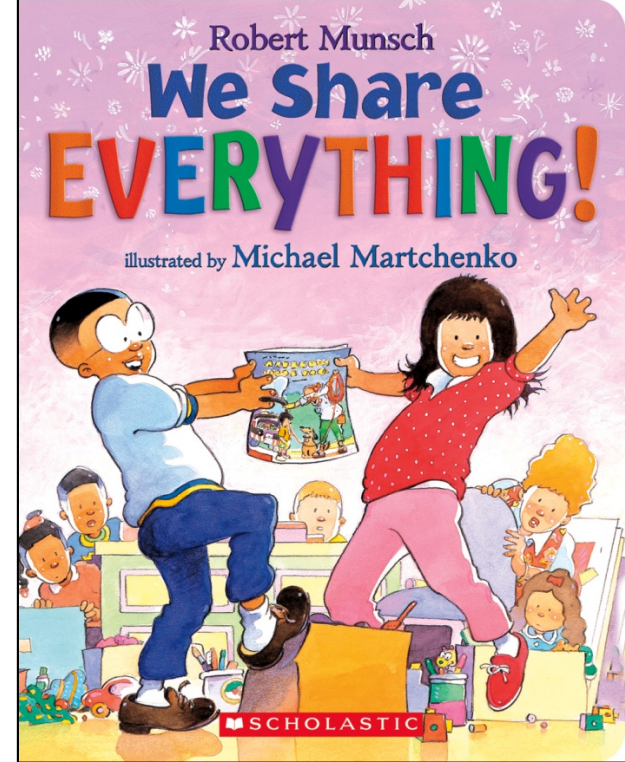
- **Genetic Exceptionalism :**  
The Genome is very fundamental data, potentially very revealing about one's identity & characteristics
- **Personal Genomic info. essentially meaningless currently but will it be in 20 yrs? 50 yrs?**
  - Genomic sequence very revealing about one's children. Is true consent possible?
  - Once put on the web it can't be taken back
- **Culture Clash:**  
Genomics historically has been a proponent of “open data” but not clear personal genomics fits this.
  - Clinical Medline has a very different culture.
- **Ethically challenged** history of genetics
  - Ownership of the data & what consent means (Hela)
    - Could your genetic data give rise to a product line?



# Sharing

# The Other Side of the Coin: Why we should share

- Sharing helps **speed research**
  - Large-scale mining of this information is important for medical research
  - Privacy is cumbersome, particularly for big data
- Sharing is important for **reproducible research**
- Sharing is useful for **education**
  - More fun to study a known person's genome
    - Eg Zimmer's Game of Genomes in STAT



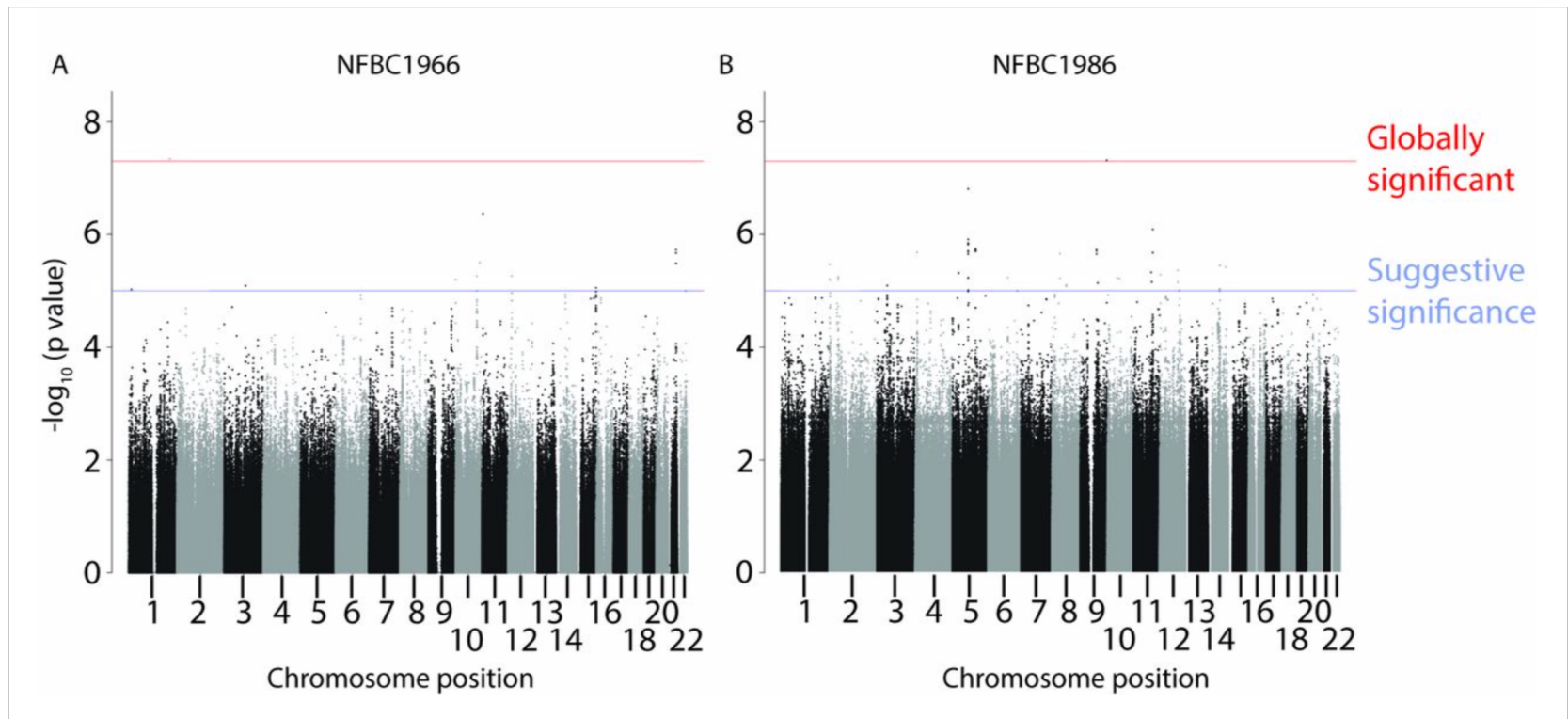
[Yale Law Roundtable ('10). *Comp. in Sci. & Eng.* 12:8; D Greenbaum & M Gerstein ('09). *Am. J. Bioethics*; D Greenbaum & M Gerstein ('10). *SF Chronicle*, May 2, Page E-4; Greenbaum et al. *PLOS CB* ('11)]

CARL ZIMMER'S  
**GAME OF GENOMES**  
SEASON 1



# Statistical power

- The genomic characterization of millions of individuals is useful for medical research
- Having a larger number of studied individuals is assured to boost statistical power, hence discoveries
- Expected to have genomes from increasingly more individuals will be sequenced going forward



# Reproducibility

“The idea is: An article about computational science in a scientific publication is not the scholarship itself, it is merely advertising of the scholarship. The actual scholarship is the complete ... set of instructions [and data] which generated the figures.”

David Donoho, 1998

# FAIR



## Data should be findable

- both for human researchers as well as computers
- requires unique and persistent identifiers for the data.

## Data ought to be accessible.

- good data stewardship
- long term electronic storage,
- legally in terms of licensing and access conditions that easily provide for authentication of authorized

# FAIR

Find

Access

Interoperate

Re-use

Data



Data needs to be **interoperable**

- human readers can clearly
  - understand the connection of the data to the main text.
  - appreciate the nature of the data from the presentation of the data
- easily able to digested by computational systems, e.g., in a standard that allows for straightforward data manipulation.

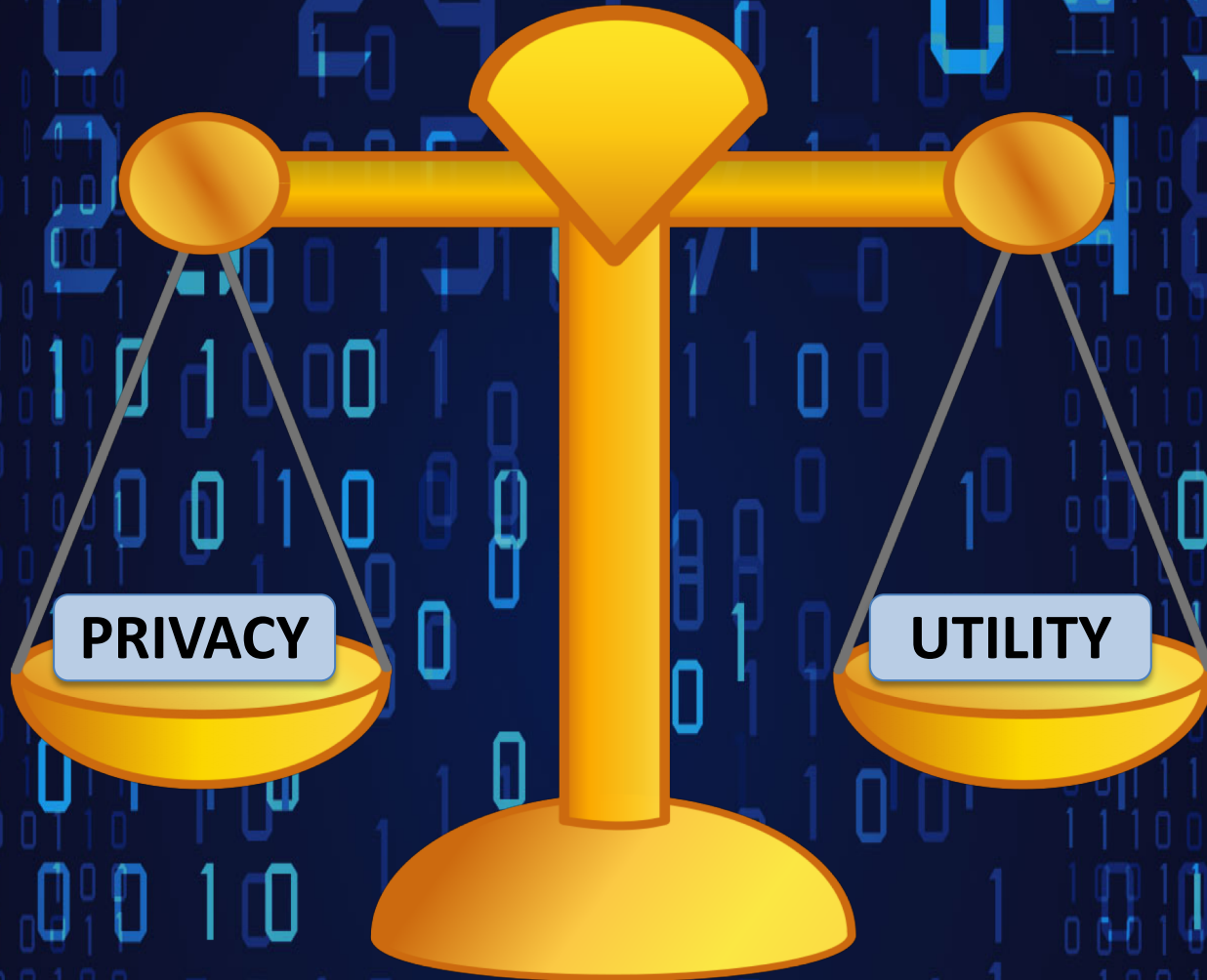
Data needs to be **reusable**

- both humans and machines should be able to either apply the data to follow-up research or additional computational analysis



# Balancing Privacy & Sharing

Goal is to find the balance





## The Dilemma

[Economist, 15 Aug '15]

- The individual (harmed?) v the collective (benefits)
  - But do sick patients care about their privacy?
- How to balance risks v rewards - Quantification
  - What is acceptable risk?  
Can we quantify leakage?
    - Ex: photos of eye color
  - Cost Benefit Analysis

# Current Social & Technical Solutions

## • **Closed Data** Approach

- Consents
- “Protected” distribution via dbGAP
- Local computes on secure computer

## • Issues with Closed Data

- Non-uniformity of consents & paperwork
  - Different international norms, leading to confusion
- Encryption & computer security creates burdensome requirements on data sharing & large scale analysis
- Many schemes get “hacked”

## • **Open Data**

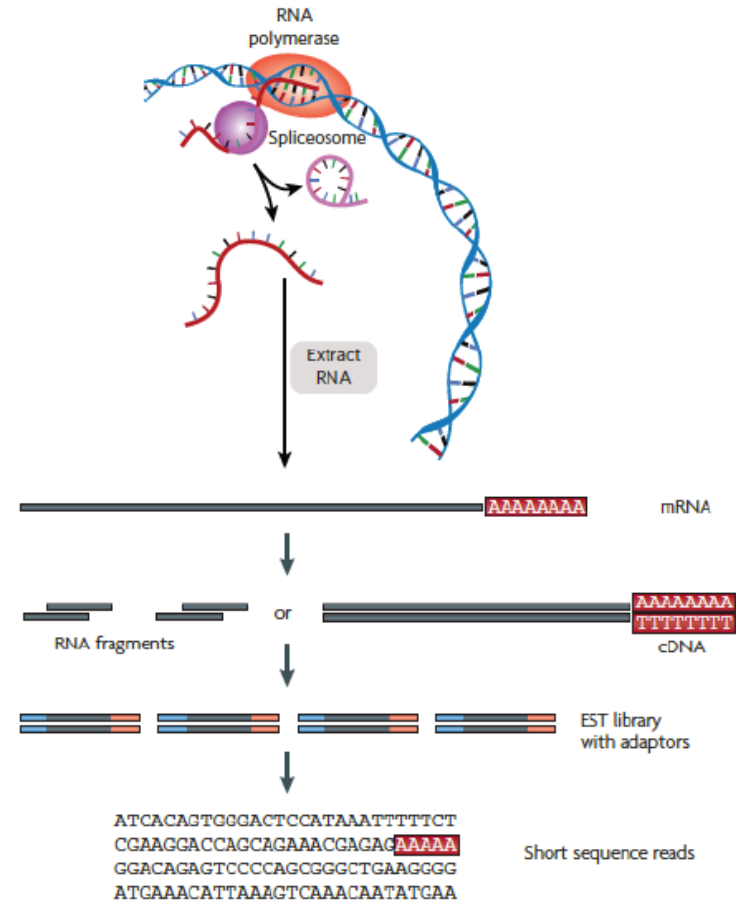
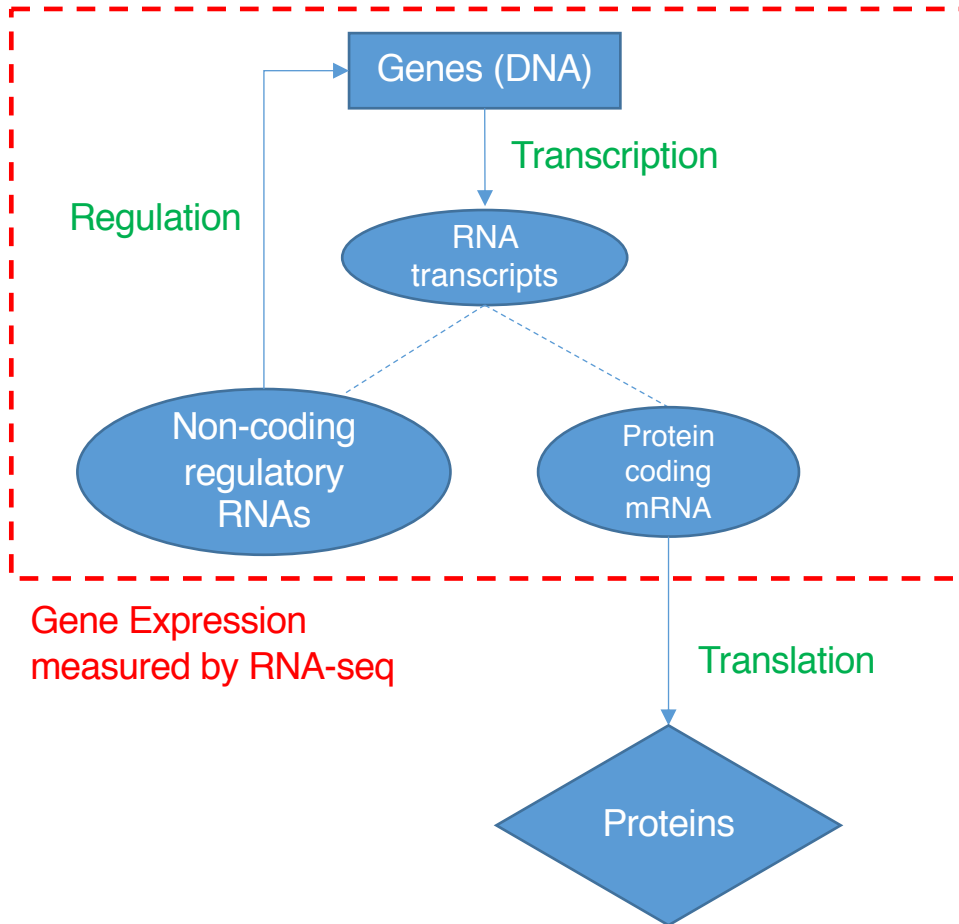
- Genomic “test pilots” (ala PGP)?
  - Sports stars & celebrities?
- Some public data & data donation is helpful but is this a realistic solution for an unbiased sample of ~1M

# Strawman Hybrid **Social** & **Tech** Proposed Solution?

- Fundamentally, researchers have to keep genetic secrets.
  - **Need for an (international) legal framework**
  - Genetic Licensure & training for individuals (similar to medical license, drivers license)
- Technology to make things easier
  - Cloud computing & enclaves (eg solution of Genomics England)
- Technological barriers shouldn't create a social incentive for “hacking”
- **Quantifying Leakage & allowing a small amounts of it**
- **Careful separation & coupling of private & public data**
  - **Lightweight, freely accessible secondary datasets coupled to underlying variants**
  - Selection of stub & "test pilot" datasets for benchmarking
  - Develop programs on public stubs on your laptop, then move the program to the cloud for private production run

**Privacy &  
Functional  
Genomics Data –  
the Data Exhaust**

# Transcriptome = Gene Activity of All Genes in the Genome, usually quantified by RNA-seq



Expression of genes is quantified by transcription:  
RNA-Seq measures mRNA transcript amounts

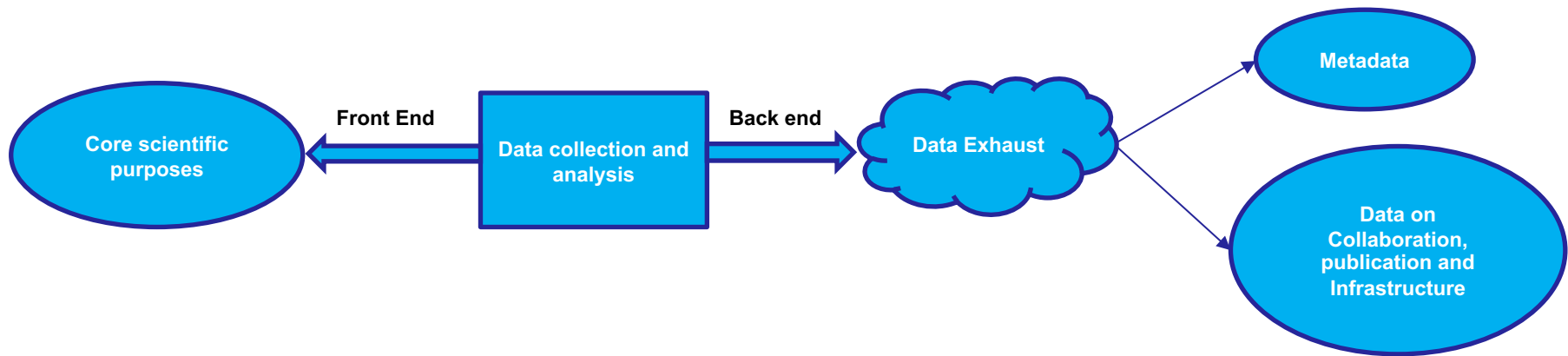
# Some Core Science Qs Addressed by RNA-seq

- Gene activity as a function of:
  - **Developmental** stage: basic patterns of co-active genes across development
  - **Cell-type** & Tissue: relationship to specialized functions
  - **Evolutionary** relationships: behavior preserved across a wide range of organisms; patterns in model organisms in relation to those in humans
  - **Individual**, across the human population
  - **Disease** phenotypes: disruption of patterns in disease
- Some overarching Qs:
  - Are there core patterns of gene activity ?**
  - How do they vary across individual ?**
  - Are they disrupted by disease?**



Studying large-scale transcriptome data  
also produces

## Data Exhaust



- Data Exhaust = Exploitable byproducts of big data collection and analysis
- Creative use of Data is key to Data Science !

[PHOTO: RELAXNEWS; from <http://www.lapresse.ca>]

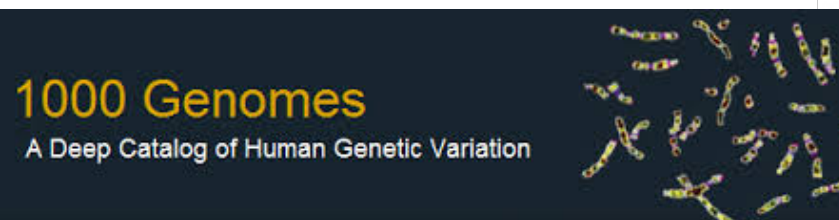
## 2-sided nature of functional genomics data: Analysis can be very **General/Public** or **Individual/Private**



- **General quantifications** related to overall aspects of a condition – ie gene activity as a function of:
  - Developmental stage, Evolutionary relationships, Cell-type, Disease
- **Above are not tied to an individual's genotype. However, data is derived from individuals & tagged with their genotypes**
- (Note, a few calculations aim to use explicitly genotype to derive general relations related to sequence variation & gene expression - eg allelic activity)

# Representative Functional Genomics, Genotype, eQTL Datasets

- Genotypes are available from the 1000 Genomes Project
- mRNA sequencing for 462 individuals from gEUVADIS and ENCODE
  - Publicly available quantification for protein coding genes
- Functional genomics data (ChIP-Seq, RNA-Seq, Hi-C) available from ENCODE
- Approximately 3,000 cis-eQTL (FDR<0.05)



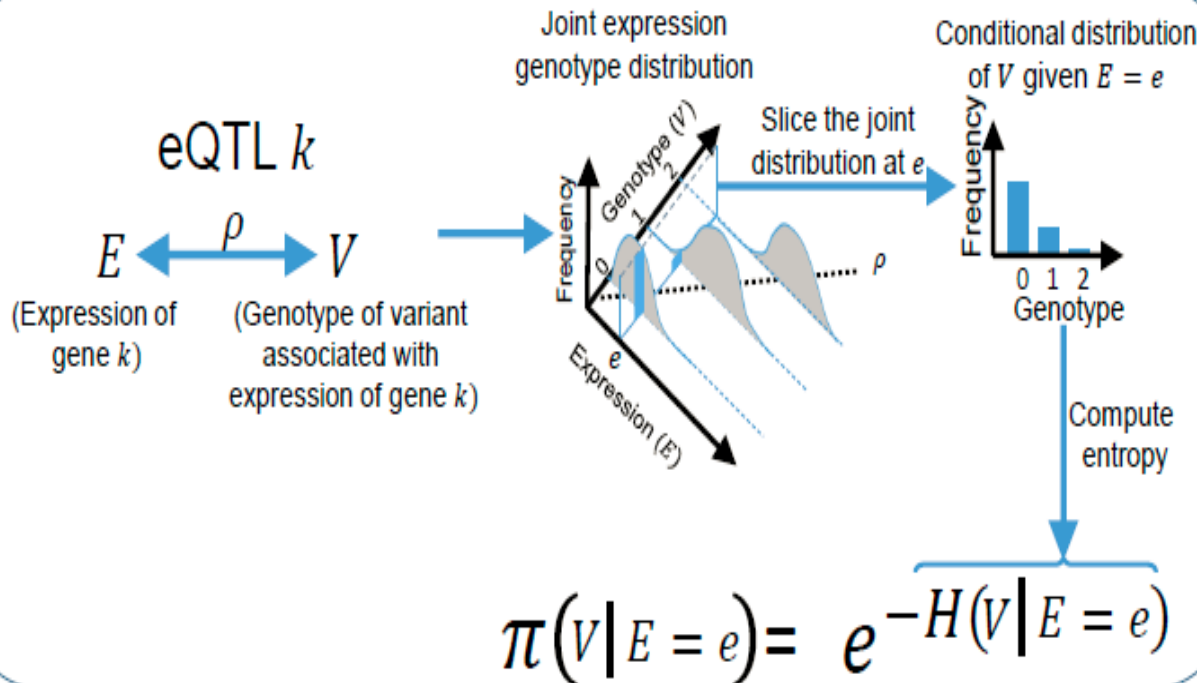
# Information Content and Predictability

$$ICI \left( \begin{array}{l} \text{Individual has variant} \\ \text{genotypes } g_1, g_2, \dots, g_n \\ \text{for variants } V_1, V_2, \dots, V_n \end{array} \right) = \log \left( \frac{1}{\text{Frequency of } V_1 \text{ genotype}} \right) + \log \left( \frac{1}{\text{Frequency of } V_2 \text{ genotype}} \right) + \dots + \log \left( \frac{1}{\text{Frequency of } V_n \text{ genotype}} \right)$$

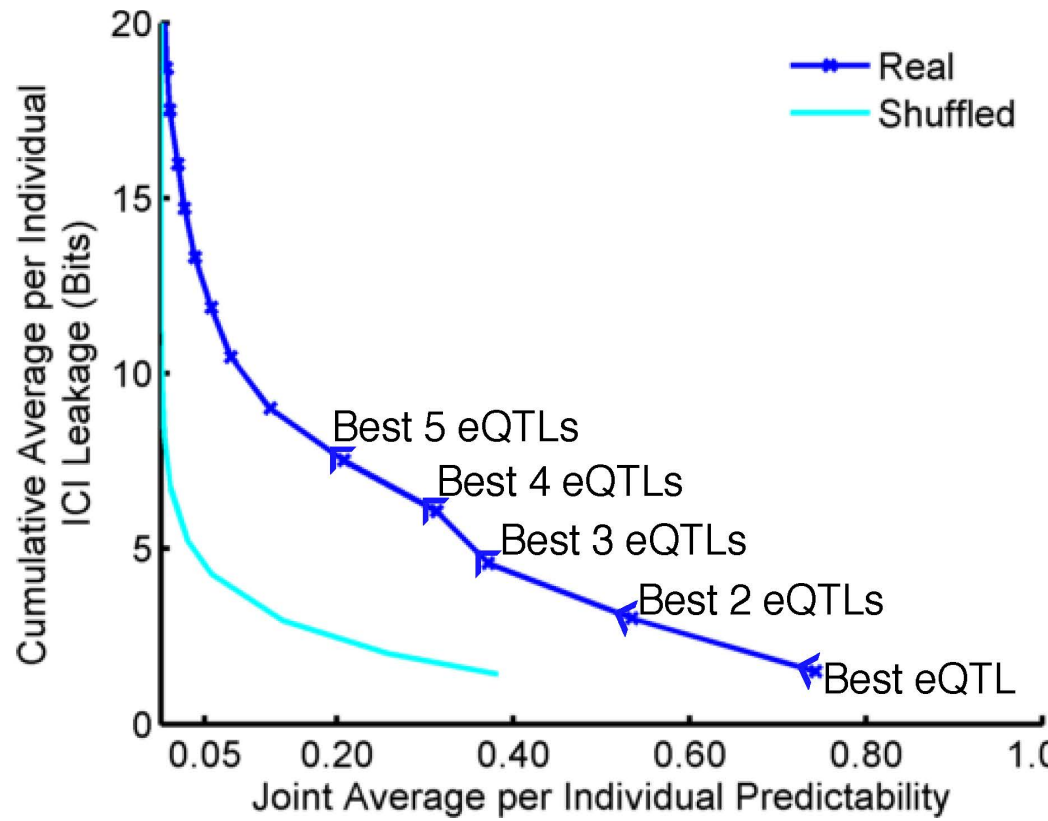
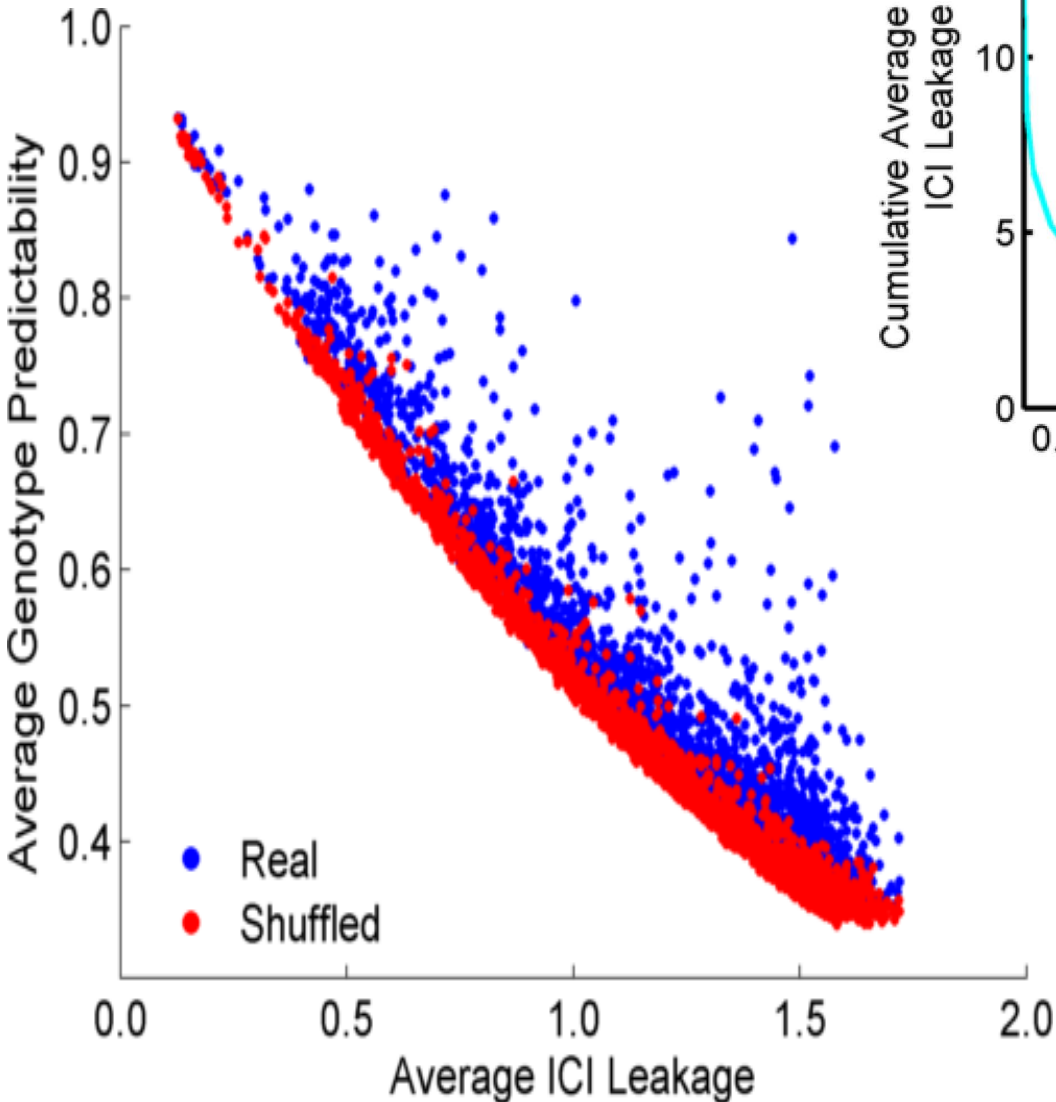
$g_1 = 2$                        $g_2 = 1$                        $g_n = 2$

$V_1$  genotype frequencies                       $V_2$  genotype frequencies                       $V_n$  genotype frequencies

- Naive measure of information (no LD, distant correlations, pop. struc., &c)
- Higher frequency: Lower ICI
- Additive for multiple variants



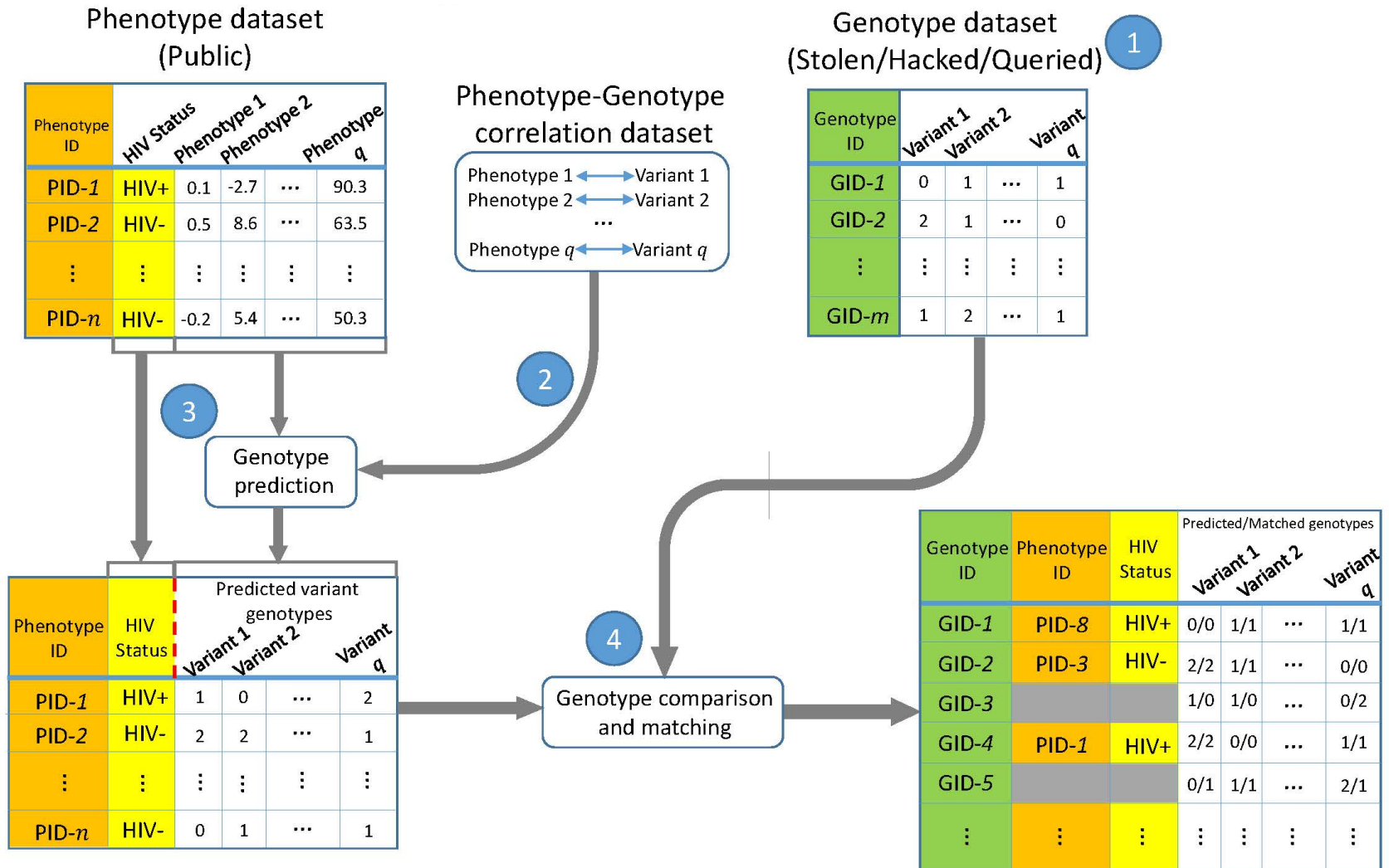
- Condition specific entropy
- Higher cond. entropy: Lower predictability
- Additive for multiple eQTLs



# ICI Leakage versus Genotype Predictability

# Linking Attacks

# Linking Attack Scenario



# Linking Attacks: Case of Netflix Prize



Names available for many users!

User (ID)	Movie (ID)	Date of Grade	Grade [1,2,3,4,5]
NTFLX-0	NTFLX-19	10/12/2008	1
NTFLX-1	NTFLX-116	4/23/2009	3
NTFLX-2	NTFLX-92	5/27/2010	2
NTFLX-1	NTFLX-666	6/6/2016	5
...	...	...	...
...	...	...	...

User (ID)	Movie (ID)	Date of Grade	Grade [0-10]
IMDB-0	IMDB-173	4/20/2009	5
IMDB-1	IMDB-18	10/18/2008	0
IMDB-2	IMDB-341	5/27/2010	-
...	...	...	...
...	...	...	...
...	...	...	...

- Many users are shared
- The grades of same users are correlated
- A user grades one movie around the same date in two databases

Anonymized Netflix Prize Training Dataset  
made available to contestants



# Linking Attacks: Case of Netflix Prize



User (ID)	Movie (ID)	Date of Grade	Grade [1,2,3,4,5]
NTFLX-0	NTFLX-19	10/12/2008	1
NTFLX-1	NTFLX-116	4/23/2009	3
NTFLX-2	NTFLX-92	5/27/2010	2
NTFLX-1	NTFLX-666	6/6/2016	5
...	...	...	...
...	...	...	...

User (ID)	Movie (ID)	Date of Grade	Grade [0-10]
IMDB-0	IMDB-173	4/20/2009	5
IMDB-1	IMDB-18	10/18/2008	0
IMDB-2	IMDB-341	5/27/2010	-
...	...	...	...
...	...	...	...
...	...	...	...

- Many users are shared
- The grades of same users are correlated
- A user grades one movie around the same date in two databases
- IMDB users are public
- NetFLIX and IMdB moves are public

# Linking Attacks: Case of Netflix Prize

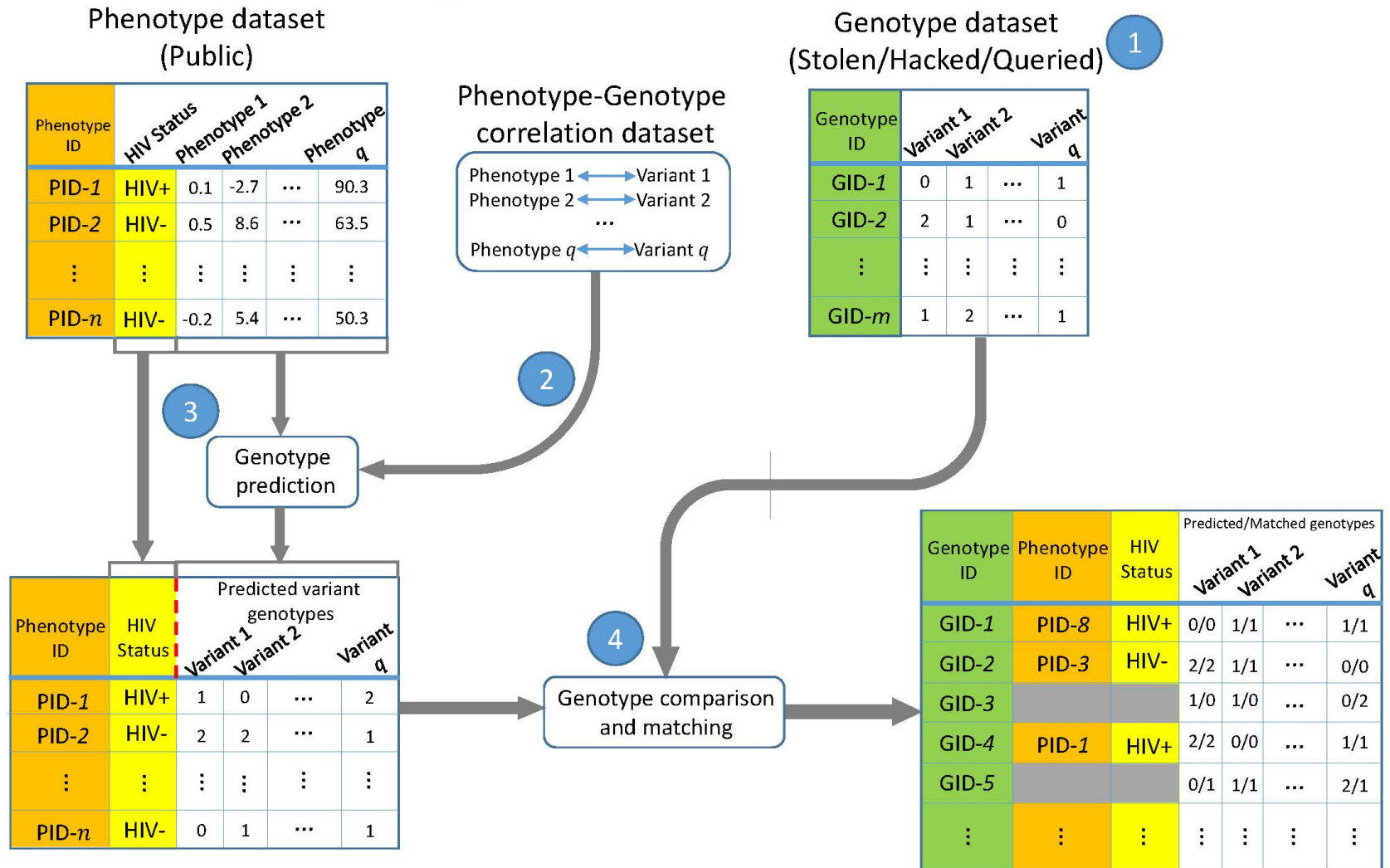


User (ID)	Movie (ID)	Date of Grade	Grade [1,2,3,4,5]
NTFLX-0	NTFLX-19	10/12/2008	1
NTFLX-1	NTFLX-116	4/23/2009	3
NTFLX-2	NTFLX-92	5/27/2010	2
NTFLX-1	<b>NTFLX-666</b>	6/6/2016	5
...	...	...	...
...	...	...	...

User (ID)	Movie (ID)	Date of Grade	Grade [0-10]
IMDB-0	IMDB-173	4/20/2009	5
IMDB-1	IMDB-18	10/18/2008	0
IMDB-2	IMDB-341	5/27/2010	-
...	...	...	...
...	...	...	...
...	...	...	...

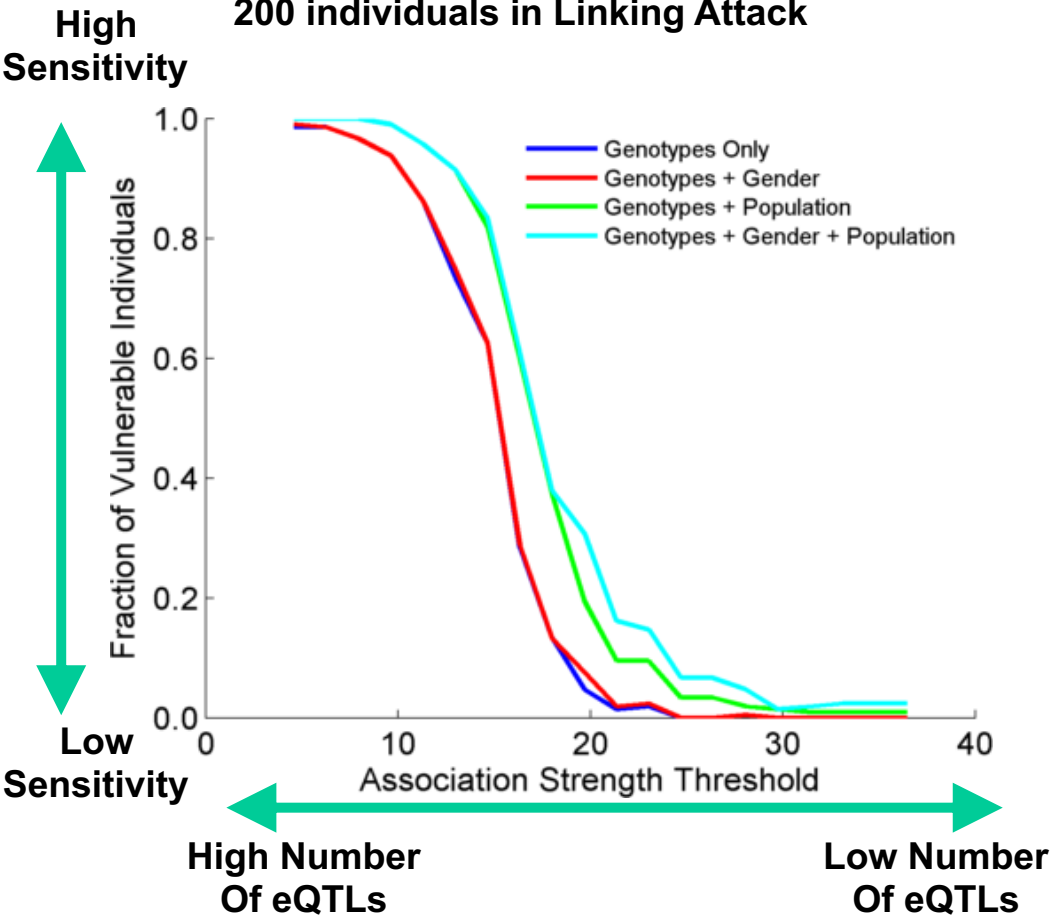
- Many users are shared
- The grades of same users are correlated
- A user grades one movie around the same date in two databases

# Linking Attack Scenario



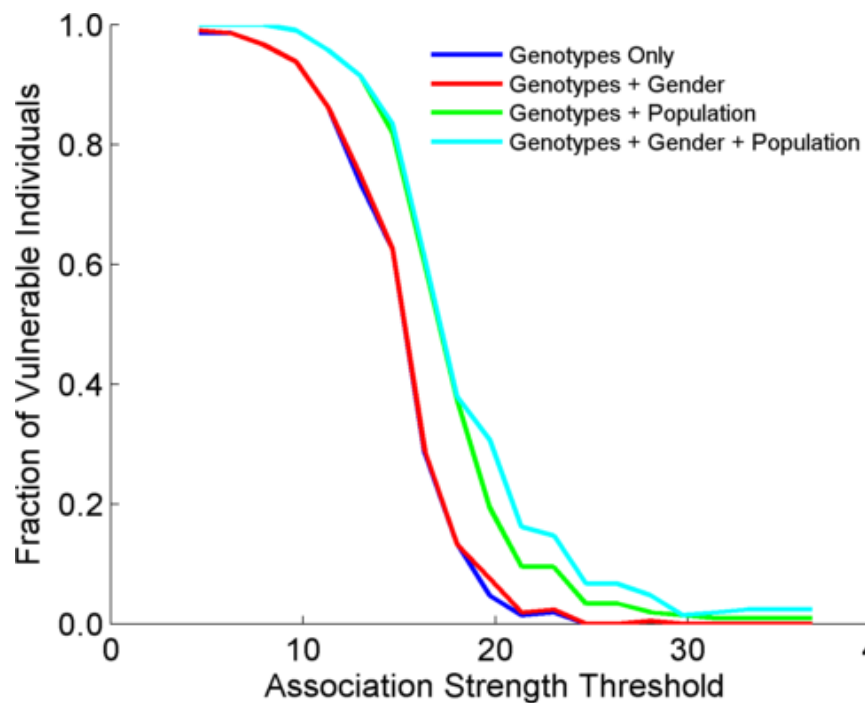
# Success in Linking Attack with Extremity based Genotype Prediction

200 individuals eQTL Discovery  
200 individuals in Linking Attack



# Success in Linking Attack with Extremity based Genotype Prediction

200 individuals eQTL Discovery  
200 individuals in Linking Attack



200 individuals eQTL Discovery  
100,200 individuals in Linking Attack

