# Summer Camp '18 Event!

# 2 Sides of the Coin for RNA-seq: Ensuring Individual Privacy v. Allowing Easy Mining

Mark Gerstein, Yale

Slides freely downloadable from **Lectures.GersteinLab.org**
& "tweetable" (via **@markgerstein**). See last slide for more info.

# 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)

# 2 Sides of a Coin: Individual Privacy v. Easy RNA-Seq Mining

- **Introduction to Genomic Privacy**
  - The dilemma: The genome as fundamental, inherited info that's very private v need for large-scale mining for med. research
  - 2-sided nature of RNA-seq presents a particularly tricky privacy issue
- **Measuring Leakage from eQTLs**
  - Quantifying & removing further variant info from expression levels + eQTLs using ICI & predictability
- **Linking Attacks from eQTLs**
  - Instantiating a practical linking attack using extreme expression levels
- **Signal Profiles**
  - Appreciable leakage from large & small deletions evident in signal profiles
  - Linking attacks also possible but additional complication of SV discovery in addition to genotyping

# 2 Sides of a Coin: Individual Privacy v. Easy RNA-Seq Mining

- **Introduction to Genomic Privacy**
  - The dilemma: The genome as fundamental, inherited info that's very private v need for large-scale mining for med. research
  - 2-sided nature of RNA-seq presents a particularly tricky privacy issue
- **Measuring Leakage from eQTLs**
  - Quantifying & removing further variant info from expression levels + eQTLs using ICI & predictability
- **Linking Attacks from eQTLs**
  - Instantiating a practical linking attack using extreme expression levels
- **Signal Profiles**
  - Appreciable leakage from large & small deletions evident in signal profiles
  - Linking attacks also possible but additional complication of SV discovery in addition to genotyping

# 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

[Seringhaus & Gerstein ('09), *Hart. Courant* (Jun 5); Greenbaum & Gerstein ('11), *NY Times* (6 Oct)]

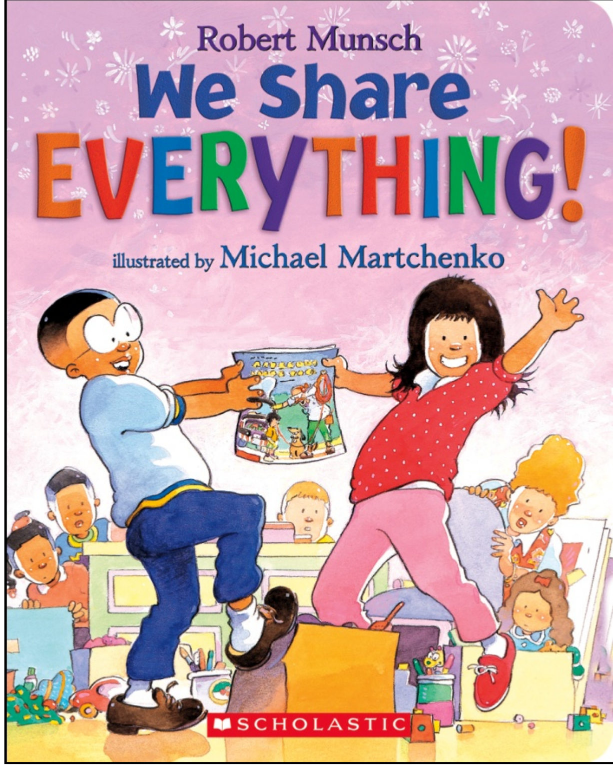
# Tricky Privacy Considerations in Personal Genomics

- **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?

[D Greenbaum & M Gerstein ('08). Am J. Bioethics; D Greenbaum & M Gerstein, Hartford Courant, 10 Jul. '08 ; SF Chronicle, 2 Nov. '08; Greenbaum et al. *PLOS CB* ('11) ; Greenbaum & Gerstein ('13), The Scientist; Photo from NY Times]

# 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



[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)]

[Economist, 15 Aug '15]

# The Dilemma

- 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

[Greenbuam et al ('04), Nat. Biotech; Greenbaum & Gerstein ('13), The Scientist]

# 2 Sides of a Coin: Individual Privacy v. Easy RNA-Seq Mining

- Introduction to Genomic Privacy
  - The dilemma: The genome as fundamental, inherited info that's very private v need for large-scale mining for med. research
  - 2-sided nature of RNA-seq presents a particularly tricky privacy issue
- Measuring Leakage from eQTLs
  - Quantifying & removing further variant info from expression levels + eQTLs using ICI & predictability
- Linking Attacks from eQTLs
  - Instantiating a practical linking attack using extreme expression levels
- Signal Profiles
  - Appreciable leakage from large & small deletions evident in signal profiles
  - Linking attacks also possible but additional complication of SV discovery in addition to genotyping

# RNA-Seq Overview

```
ATACAAGCAAGTATAAGTTCGTATGCCGTCTT
GGAGGCTGGAGTTGGGGACGTATGCGGCATAG
TACCGATCGAGTCGACTGTAAACGTAGGCATA
ATTCTGACTGGTGTCATGCTGATGTACTTAAA
```

Successive steps of Data Reduction

Fastq sequence files
~5-10 GB

Index-building + Alignment to reference genome

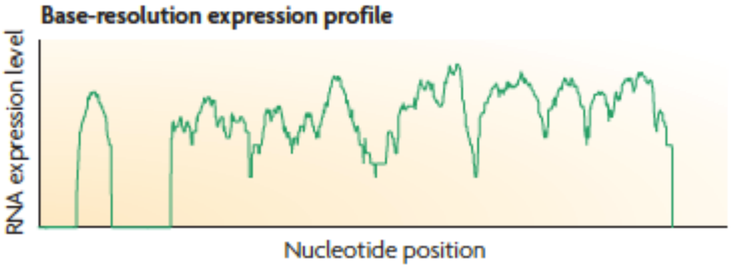
BAM files
~1-2-fold reduction

Conversion to signal track by overlapping reads

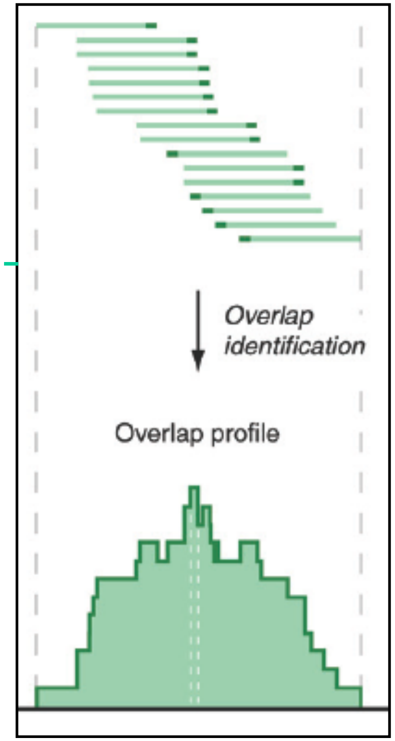
BigWig files
~25-fold reduction

Mapping to genes

Gene/Transcript expression matrix
~20-fold reduction



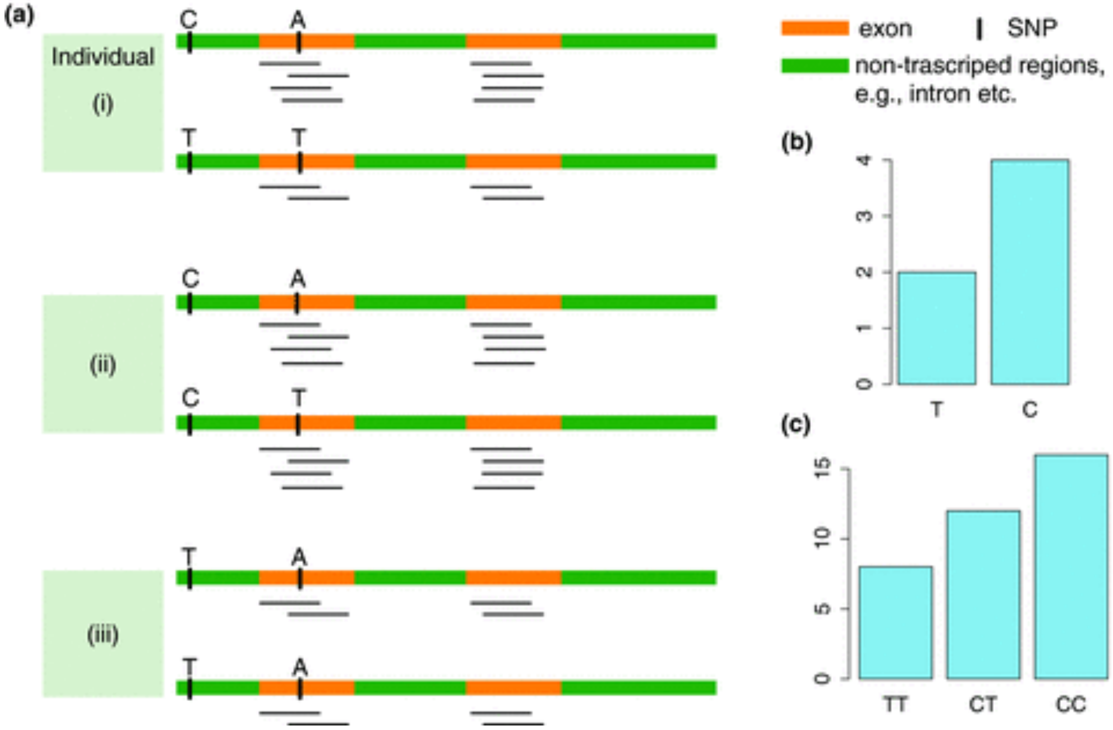
Quantitative information from RNA-seq signal: average signals at exon level (RPKMs)



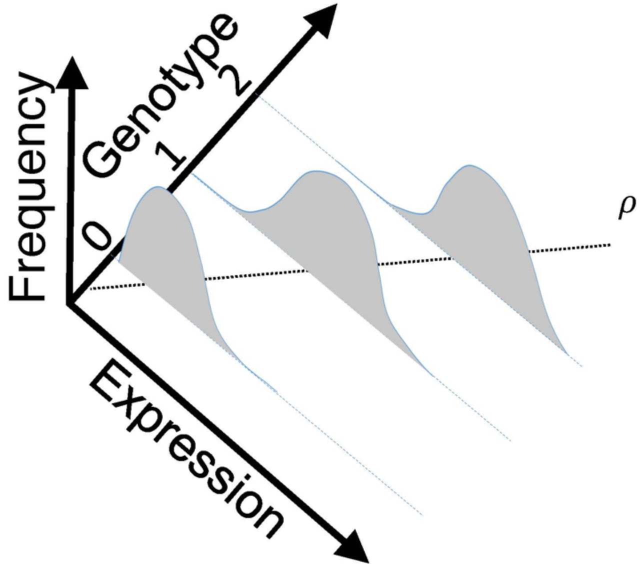
Reads => Signal

[*NAT. REV.* 10: 57; *PLOS CB* 4:e1000158; *PNAS* 4:107: 5254 ]

# eQTL Mapping Using RNA-Seq Data



[Biometrics 68(1) 1–11]



- eQTLs are genomic loci that contribute to variation in mRNA expression levels
- eQTLs provide insights on transcription regulation, and the molecular basis of phenotypic outcomes
- eQTL mapping can be done with RNA-Seq data

# 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)

1000 Genomes
A Deep Catalog of Human Genetic Variation

GEUVADIS
CONSORTIUM

ENCODE

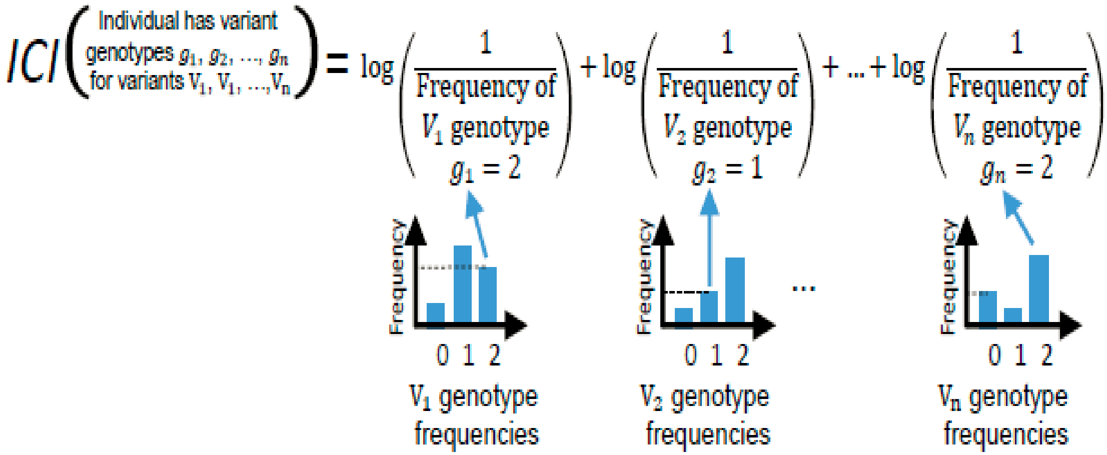
# Strawman Hybrid Social & Tech Proposed Solution?

- 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

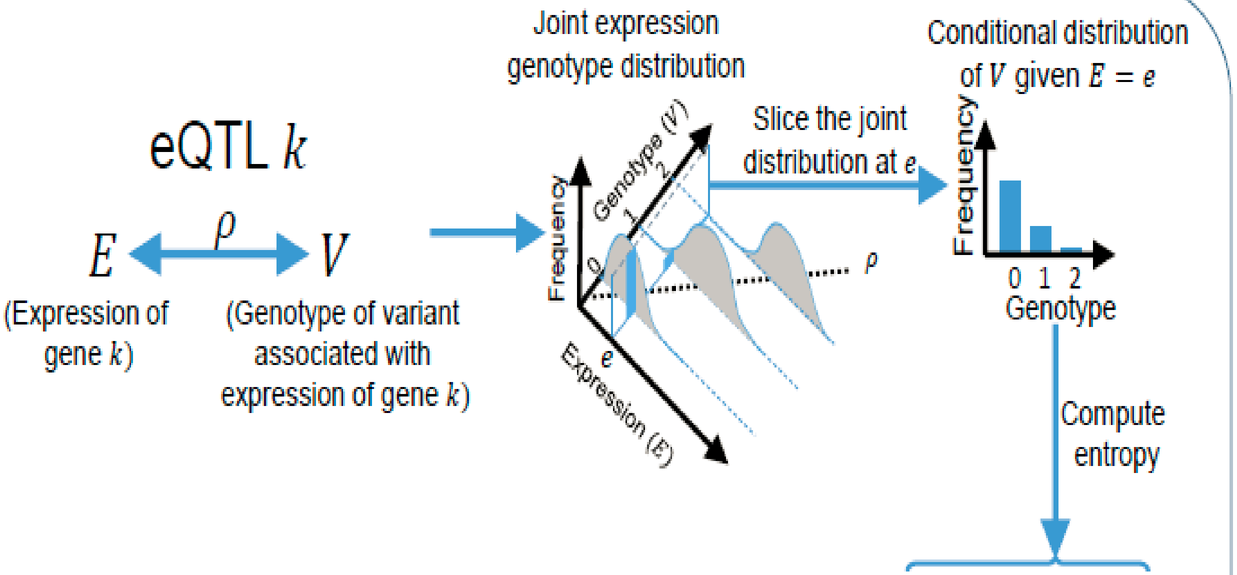
[D Greenbaum, M Gerstein ('11). Am J Bioeth 11:39. Greenbaum & Gerstein, The Scientist ('13)]

# Information Content and Predictability



$$ICI\left(\begin{array}{c}\text{Individual has variant}\\ \text{genotypes } g_1, g_2, \ldots, g_n\\ \text{for variants } V_1, V_1, \ldots, V_n\end{array}\right) = \log\left(\frac{1}{\begin{array}{c}\text{Frequency of}\\ V_1 \text{ genotype}\\ g_1 = 2\end{array}}\right) + \log\left(\frac{1}{\begin{array}{c}\text{Frequency of}\\ V_2 \text{ genotype}\\ g_2 = 1\end{array}}\right) + \ldots + \log\left(\frac{1}{\begin{array}{c}\text{Frequency of}\\ V_n \text{ genotype}\\ g_n = 2\end{array}}\right)$$

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



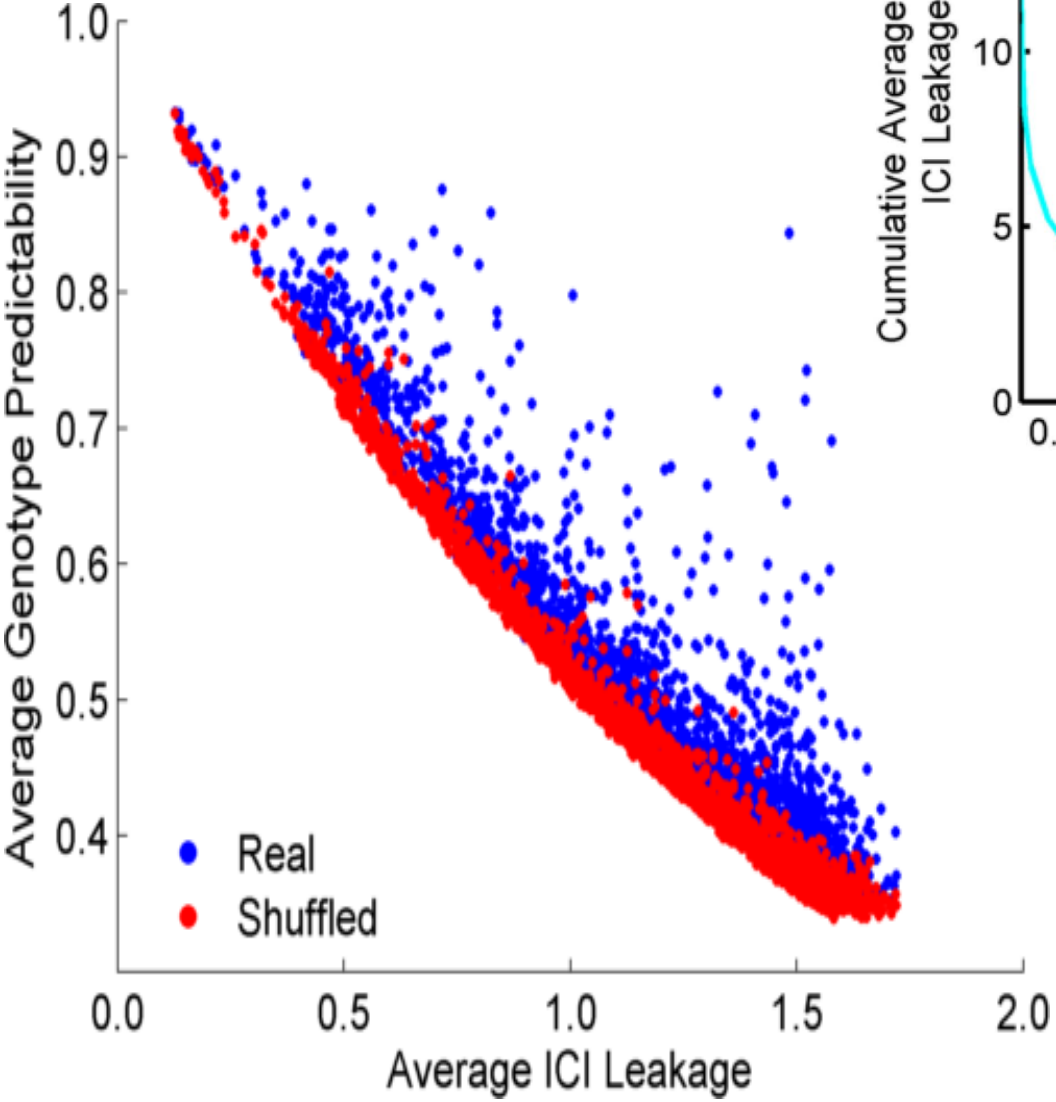
eQTL k
E
ρ
V
(Expression of gene k)
(Genotype of variant associated with expression of gene k)
Joint expression genotype distribution
Frequency
Genotype (V)
0
1
2
e
Expression (E)
ρ
Slice the joint distribution at e
Conditional distribution of V given E = e
Frequency
0 1 2
Genotype
Compute entropy

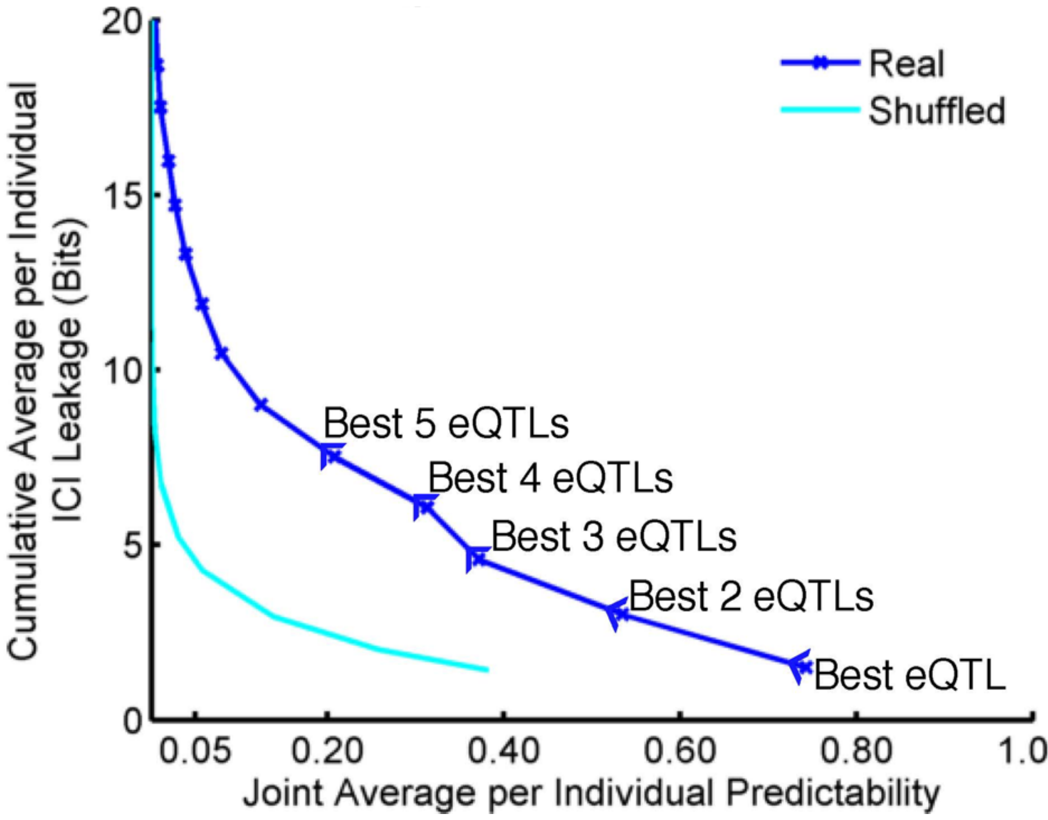
$$\pi\left(V \middle| E = e\right) = e^{-H\left(V \middle| E = e\right)}$$

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

[Harmanci et al. Nat. Meth. 2016]

[Harmanciet al. Nat. Meth. ('16]



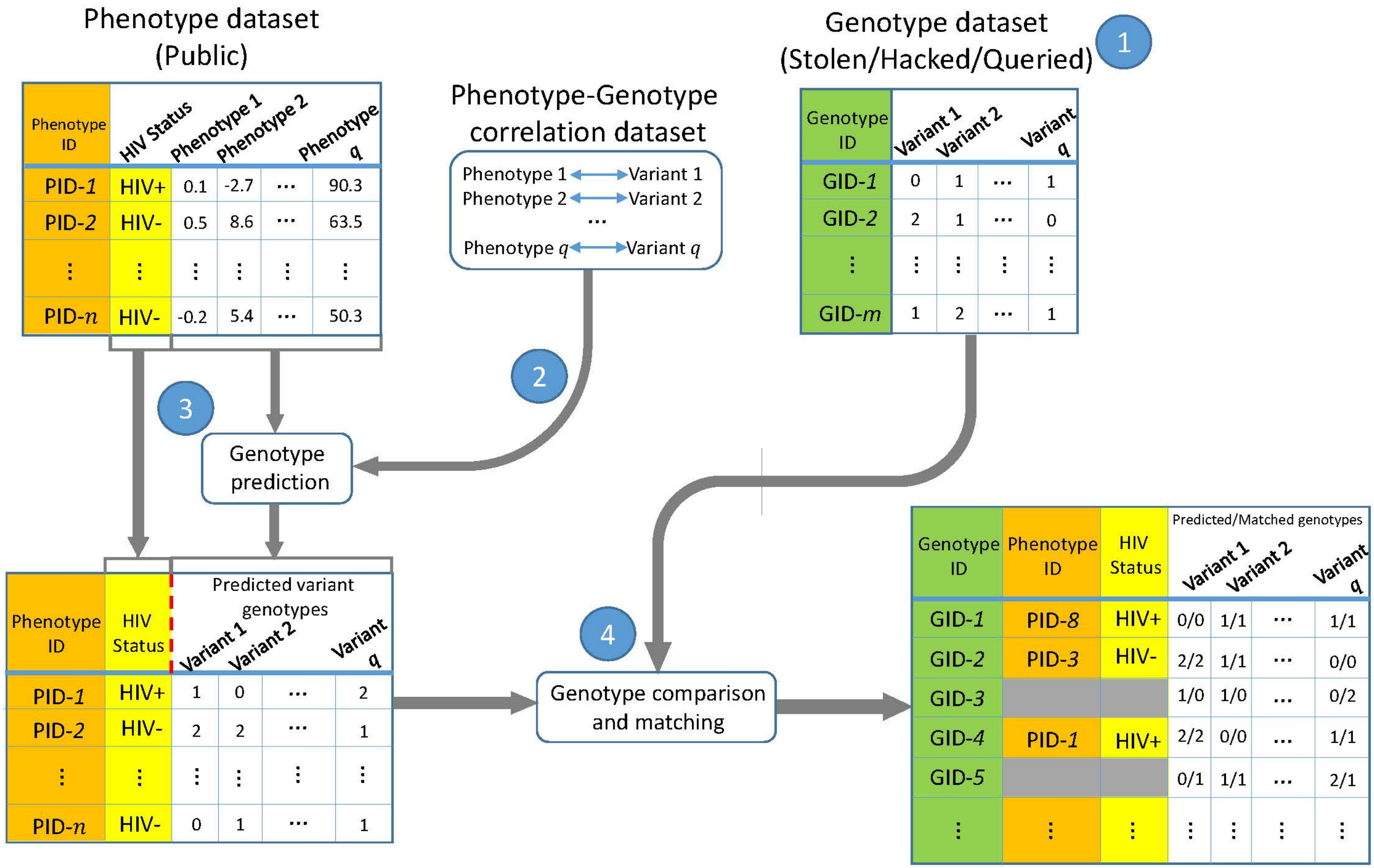


# ICI Leakage versus Genotype Predictability

# 2 Sides of a Coin: Individual Privacy v. Easy RNA-Seq Mining

- **Introduction to Genomic Privacy**
  - The dilemma: The genome as fundamental, inherited info that's very private v need for large-scale mining for med. research
  - 2-sided nature of RNA-seq presents a particularly tricky privacy issue
- **Measuring Leakage from eQTLs**
  - Quantifying & removing further variant info from expression levels + eQTLs using ICI & predictability
- **Linking Attacks from eQTLs**
  - Instantiating a practical linking attack using extreme expression levels
- **Signal Profiles**
  - Appreciable leakage from large & small deletions evident in signal profiles
  - Linking attacks also possible but additional complication of SV discovery in addition to genotyping

# Linking Attack Scenario



Phenotype dataset (Public)

| Phenotype ID | HIV Status | Phenotype 1 | Phenotype 2 | | Phenotype q |
|---|---|---|---|---|---|
| PID-1 | HIV+ | 0.1 | -2.7 | ⋯ | 90.3 |
| PID-2 | HIV- | 0.5 | 8.6 | ⋯ | 63.5 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| PID-$n$ | HIV- | -0.2 | 5.4 | ⋯ | 50.3 |

Phenotype-Genotype correlation dataset

Phenotype 1 ⟷ Variant 1
Phenotype 2 ⟷ Variant 2
⋯
Phenotype $q$ ⟷ Variant $q$

Genotype dataset (Stolen/Hacked/Queried)

| Genotype ID | Variant 1 | Variant 2 | | Variant q |
|---|---|---|---|---|
| GID-1 | 0 | 1 | ⋯ | 1 |
| GID-2 | 2 | 1 | ⋯ | 0 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| GID-$m$ | 1 | 2 | ⋯ | 1 |

Genotype prediction

| Phenotype ID | HIV Status | Predicted variant genotypes | | | |
|---|---|---|---|---|---|
| | | Variant 1 | Variant 2 | | Variant q |
| PID-1 | HIV+ | 1 | 0 | ⋯ | 2 |
| PID-2 | HIV- | 2 | 2 | ⋯ | 1 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| PID-$n$ | HIV- | 0 | 1 | ⋯ | 1 |

Genotype comparison and matching

| Genotype ID | Phenotype ID | HIV Status | Predicted/Matched genotypes | | | |
|---|---|---|---|---|---|---|
| | | | Variant 1 | Variant 2 | | Variant q |
| GID-1 | PID-8 | HIV+ | 0/0 | 1/1 | ⋯ | 1/1 |
| GID-2 | PID-3 | HIV- | 2/2 | 1/1 | ⋯ | 0/0 |
| GID-3 | | | 1/0 | 1/0 | ⋯ | 0/2 |
| GID-4 | PID-1 | HIV+ | 2/2 | 0/0 | ⋯ | 1/1 |
| GID-5 | | | 0/1 | 1/1 | ⋯ | 2/1 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |

[Harmanciet al. Nat. Meth. (’16)]

# 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 |
| … | … | … | … |
| … | … | … | … |

**Names available for many users!**

| 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

NETFLIX

IMDb

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

- IMDB users are public

- NetFLIX and IMdB moves are public

# Linking Attacks: Case of Netflix Prize

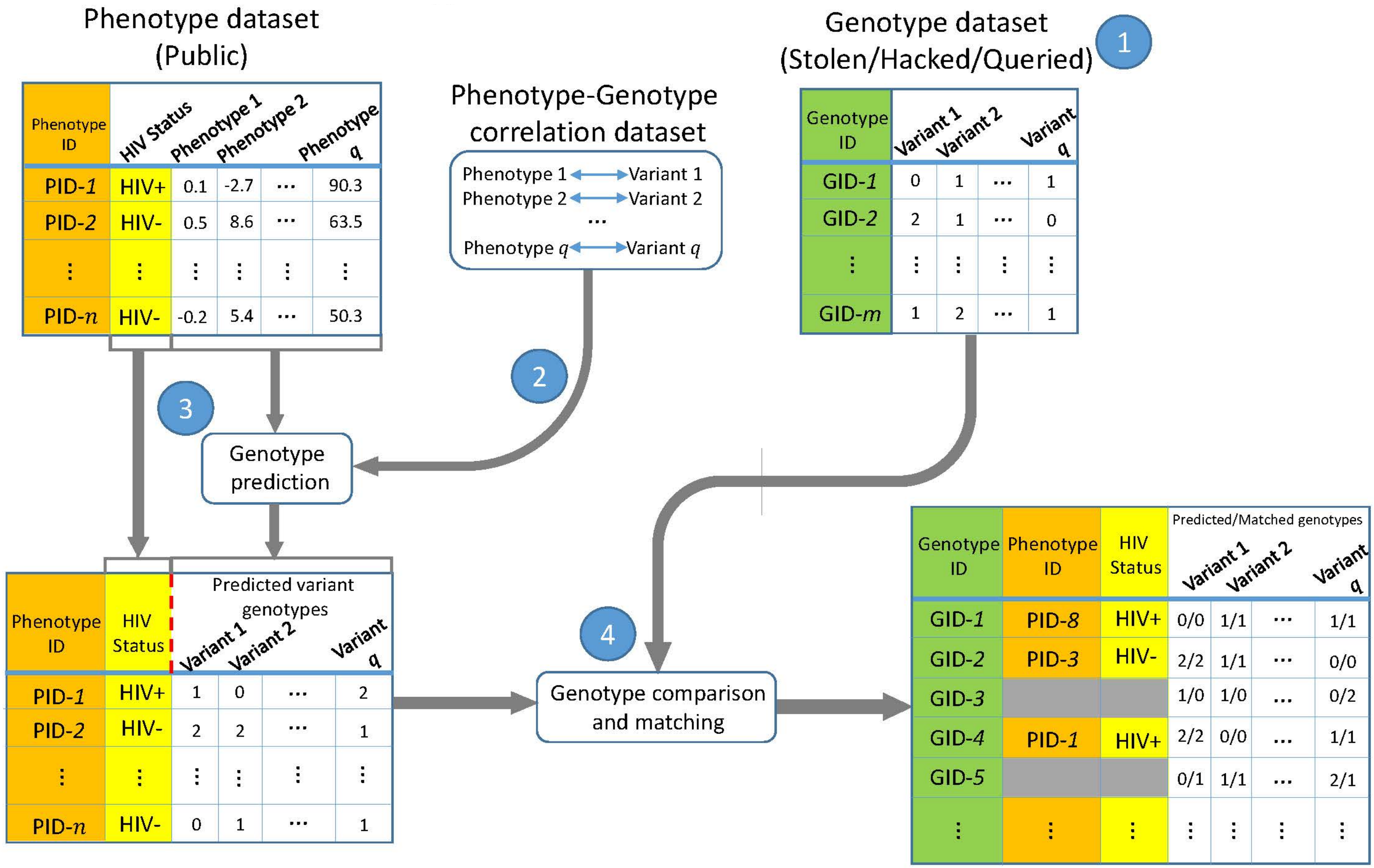
Names available for many users!

| 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

# Linking Attack Scenario



Phenotype dataset (Public)

| Phenotype ID | HIV Status | Phenotype 1 | Phenotype 2 | | Phenotype *q* |
|---|---|---|---|---|---|
| PID-*1* | HIV+ | 0.1 | -2.7 | ⋯ | 90.3 |
| PID-*2* | HIV- | 0.5 | 8.6 | ⋯ | 63.5 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| PID-*n* | HIV- | -0.2 | 5.4 | ⋯ | 50.3 |

Phenotype-Genotype correlation dataset

Phenotype 1 ⟷ Variant 1
Phenotype 2 ⟷ Variant 2
⋯
Phenotype *q* ⟷ Variant *q*

Genotype dataset (Stolen/Hacked/Queried)

| Genotype ID | Variant 1 | Variant 2 | | Variant *q* |
|---|---|---|---|---|
| GID-*1* | 0 | 1 | ⋯ | 1 |
| GID-*2* | 2 | 1 | ⋯ | 0 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| GID-*m* | 1 | 2 | ⋯ | 1 |

Genotype prediction

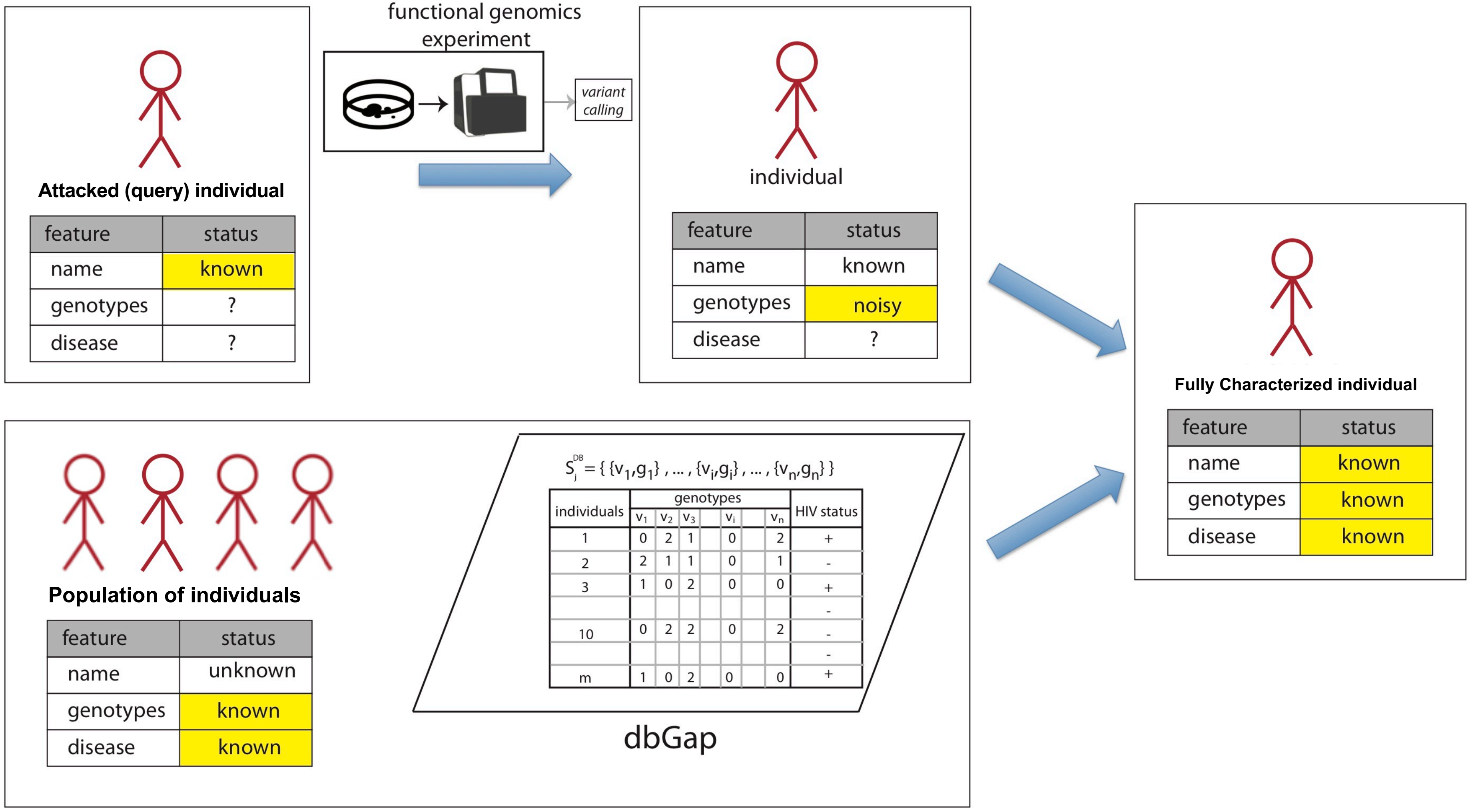
| Phenotype ID | HIV Status | Predicted variant genotypes | | | |
|---|---|---|---|---|---|
| | | Variant 1 | Variant 2 | | Variant *q* |
| PID-*1* | HIV+ | 1 | 0 | ⋯ | 2 |
| PID-*2* | HIV- | 2 | 2 | ⋯ | 1 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| PID-*n* | HIV- | 0 | 1 | ⋯ | 1 |

Genotype comparison and matching

| Genotype ID | Phenotype ID | HIV Status | Predicted/Matched genotypes | | | |
|---|---|---|---|---|---|---|
| | | | Variant 1 | Variant 2 | | Variant *q* |
| GID-*1* | PID-*8* | HIV+ | 0/0 | 1/1 | ⋯ | 1/1 |
| GID-*2* | PID-*3* | HIV- | 2/2 | 1/1 | ⋯ | 0/0 |
| GID-*3* | | | 1/0 | 1/0 | ⋯ | 0/2 |
| GID-*4* | PID-*1* | HIV+ | 2/2 | 0/0 | ⋯ | 1/1 |
| GID-*5* | | | 0/1 | 1/1 | ⋯ | 2/1 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |

[Harmanciet al. Nat. Meth. ('16)]

# Summary of a Linking Attack



**Attacked (query) individual**

| feature | status |
|---|---|
| name | known |
| genotypes | ? |
| disease | ? |

individual

| feature | status |
|---|---|
| name | known |
| genotypes | noisy |
| disease | ? |

**Population of individuals**

| feature | status |
|---|---|
| name | unknown |
| genotypes | known |
| disease | known |

$S_j^{DB} = \{ \{v_1,g_1\} , \ldots , \{v_i,g_i\} , \ldots , \{v_n,g_n\} \}$

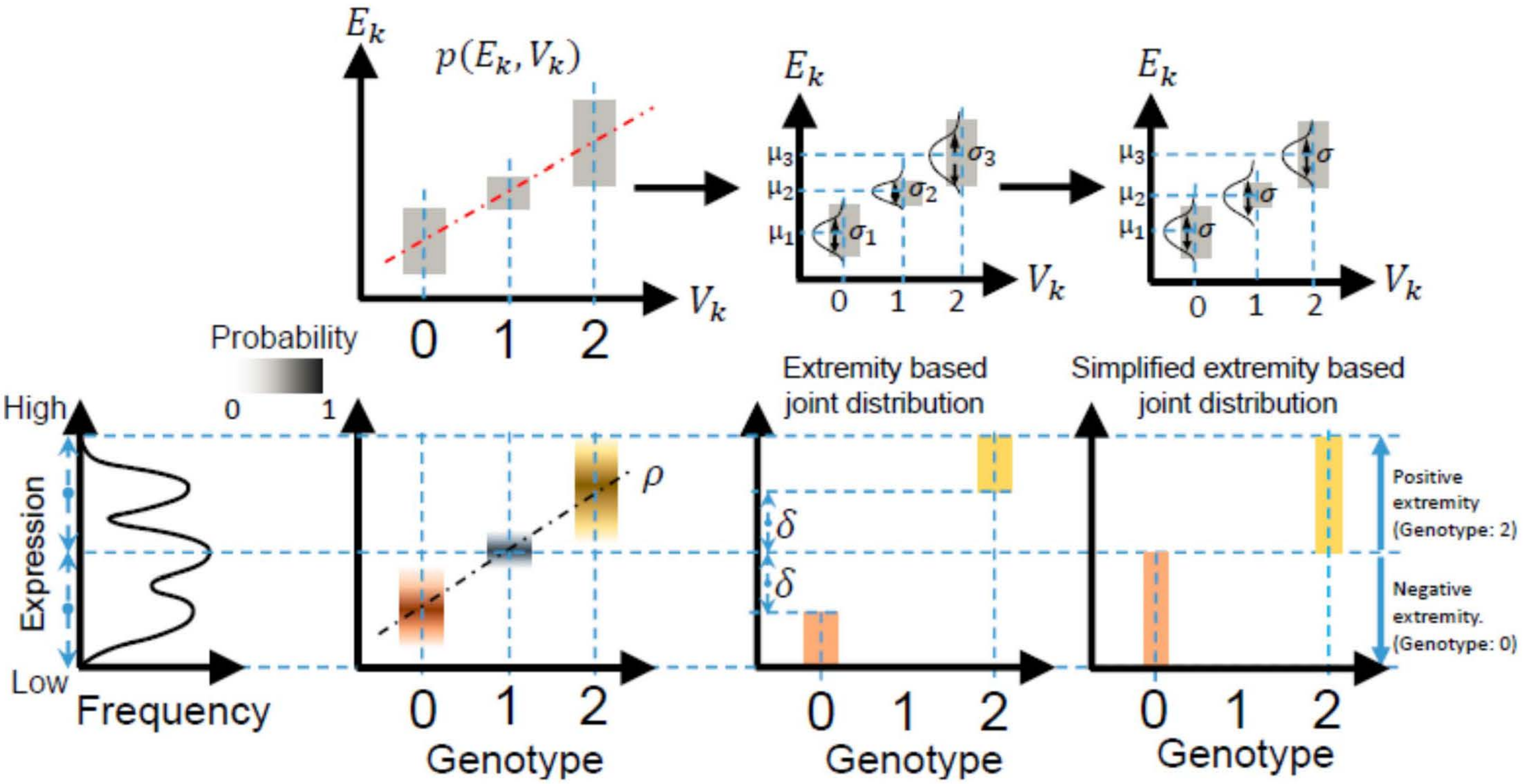
| individuals | genotypes | | | | | | | HIV status |
|---|---|---|---|---|---|---|---|---|
| | $v_1$ | $v_2$ | $v_3$ | | $v_i$ | | $v_n$ | |
| 1 | 0 | 2 | 1 | | 0 | | 2 | + |
| 2 | 2 | 1 | 1 | | 0 | | 1 | - |
| 3 | 1 | 0 | 2 | | 0 | | 0 | + |
| | | | | | | | | - |
| 10 | 0 | 2 | 2 | | 0 | | 2 | - |
| | | | | | | | | - |
| m | 1 | 0 | 2 | | 0 | | 0 | + |

dbGap

**Fully Characterized individual**

| feature | status |
|---|---|
| name | known |
| genotypes | known |
| disease | known |

# Levels of Expression-Genotype Model Simplifications for Genotype Prediction

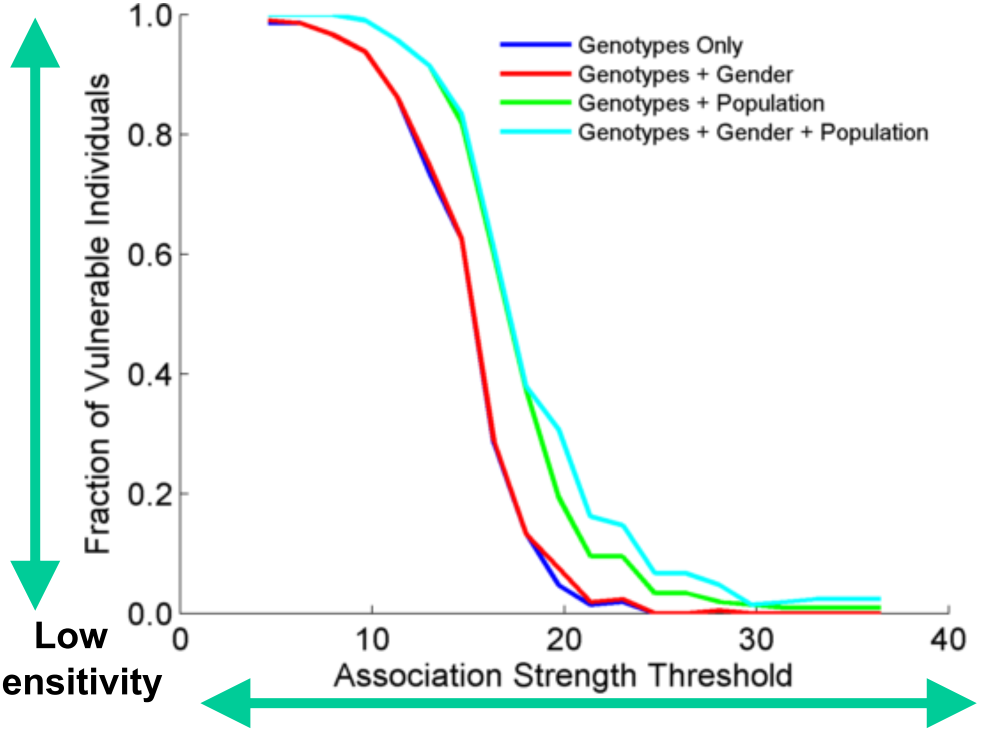


[Harmanci et al. *Nat. Meth.* (16)]

# Success in Linking Attack with Extremity based Genotype Prediction

**200 individuals eQTL Discovery**
**200 individuals in Linking Attack**

**High Sensitivity**



**Low Sensitivity**

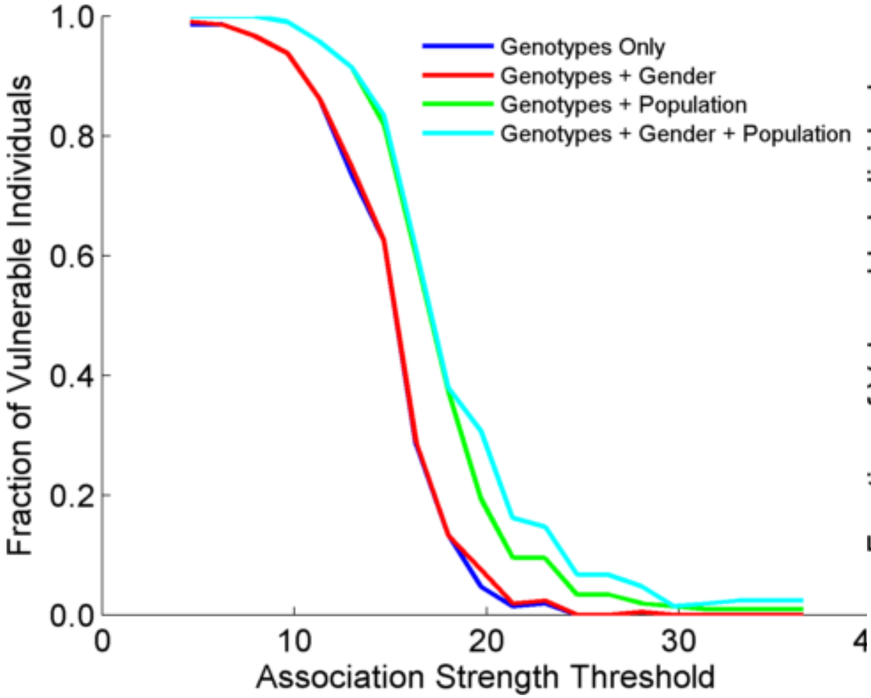
**High Number Of eQTLs**

**Low Number Of eQTLs**

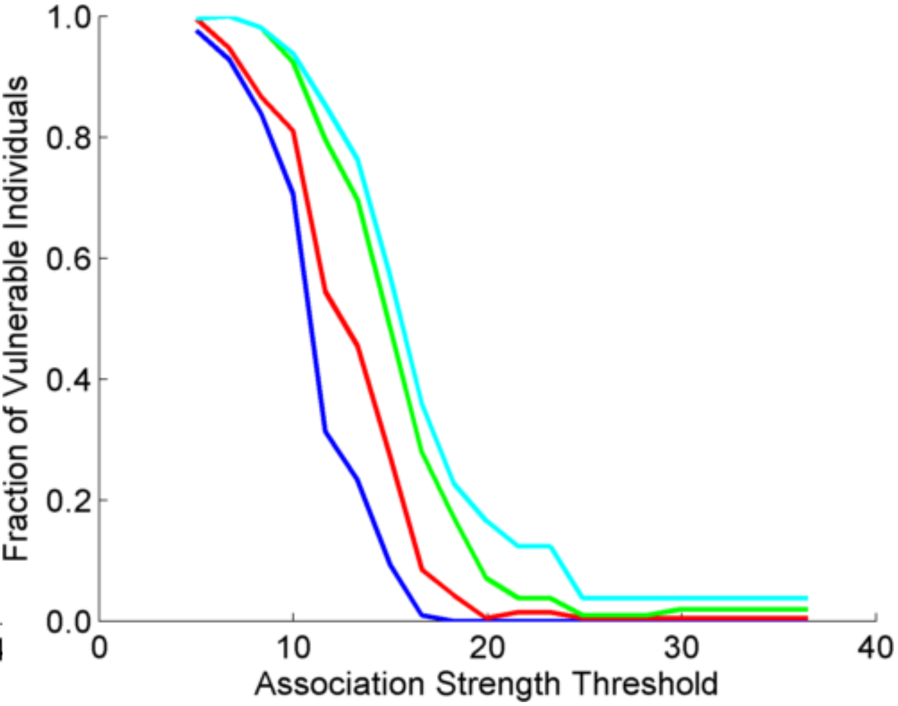
[Harmanci et al. *Nat. Meth.* (16)]

# 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**

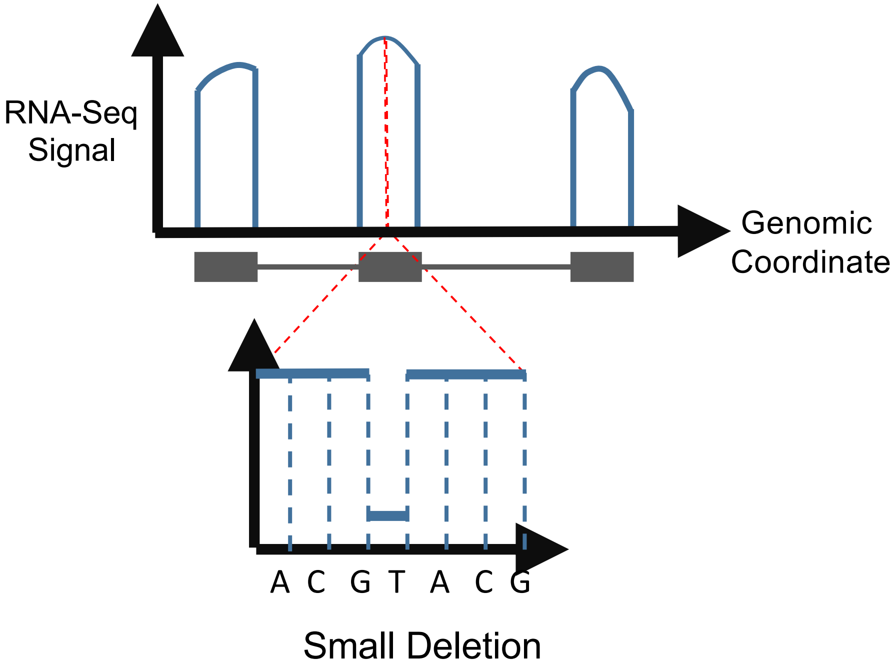


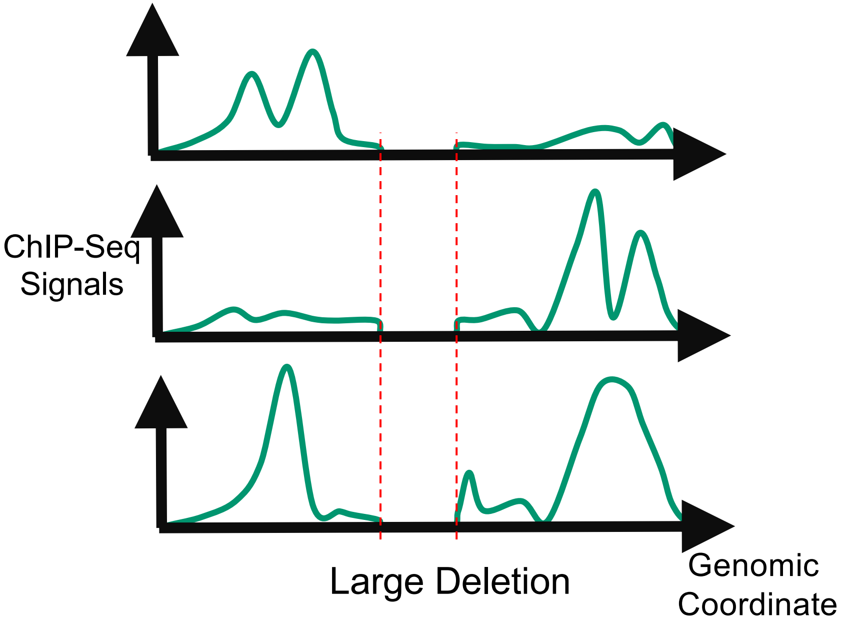
[Harmanci et al. *Nat. Meth.* (16)]

# 2 Sides of a Coin: Individual Privacy v. Easy RNA-Seq Mining

- **Introduction to Genomic Privacy**
  - The dilemma: The genome as fundamental, inherited info that's very private v need for large-scale mining for med. research
  - 2-sided nature of RNA-seq presents a particularly tricky privacy issue
- **Measuring Leakage from eQTLs**
  - Quantifying & removing further variant info from expression levels + eQTLs using ICI & predictability
- **Linking Attacks from eQTLs**
  - Instantiating a practical linking attack using extreme expression levels
- **Signal Profiles**
  - Appreciable leakage from large & small deletions evident in signal profiles
  - Linking attacks also possible but additional complication of SV discovery in addition to genotyping

# Detection & Genotyping of small & large SV deletions from signal profiles

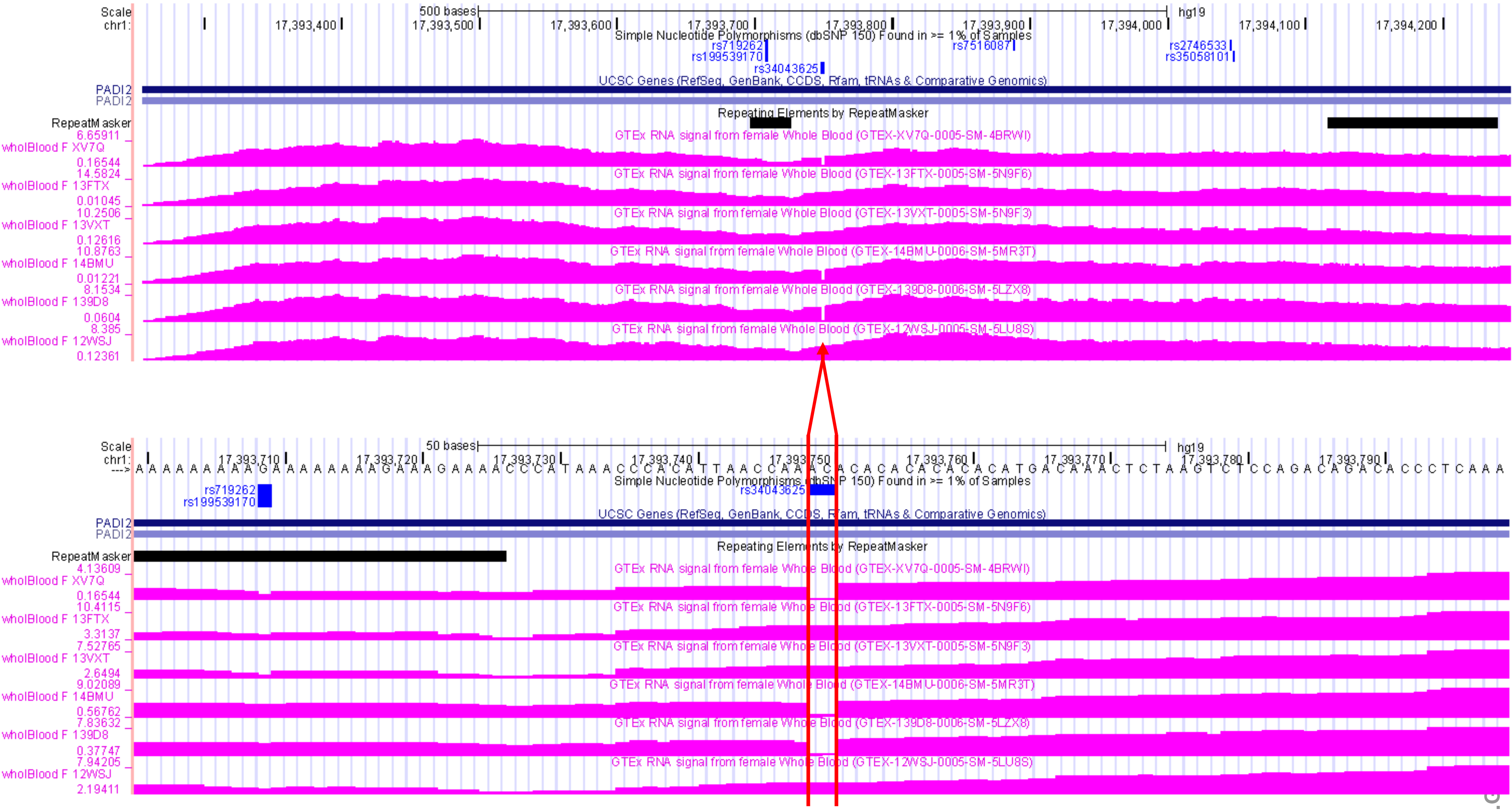




RNA-seq also shows large deletions

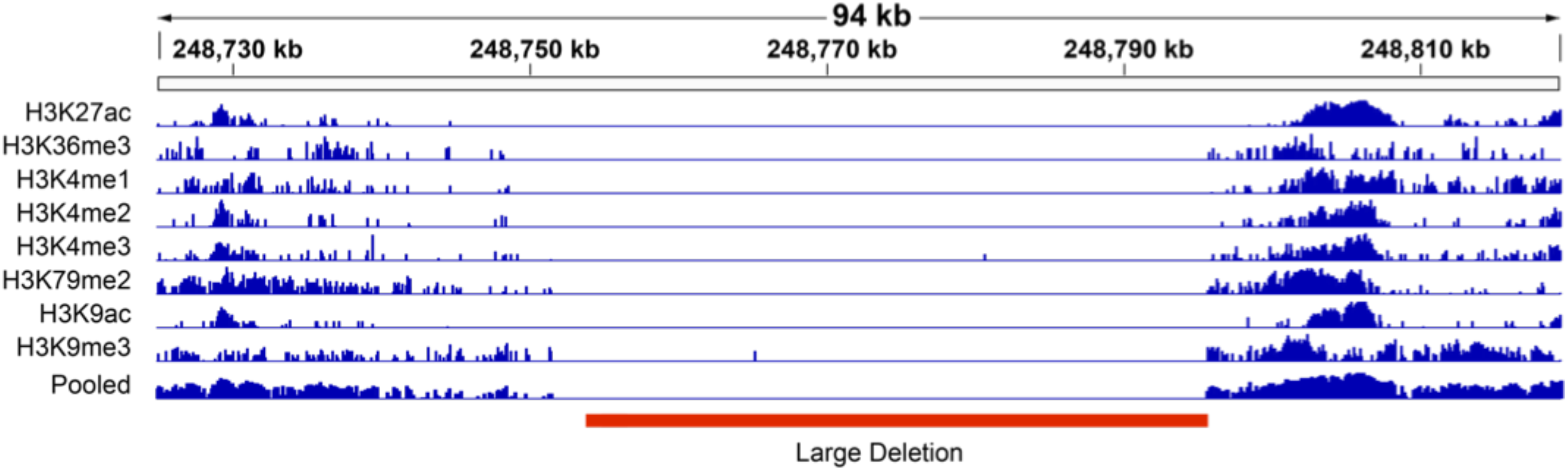
[Harmanci & Gerstein, *Nat. Comm.* ('18)]

# Example of Small Deletion Evident in Signal Profile


Scale
chr1:
500 bases
hg19
17,393,400
17,393,500
17,393,600
17,393,700
17,393,800
17,393,900
17,394,000
17,394,100
17,394,200
Simple Nucleotide Polymorphisms (dbSNP 150) Found in >= 1% of Samples
rs719262
rs199539170
rs34043625
rs7516087
rs2746533
rs35058101
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
PADI2
PADI2
RepeatMasker
Repeating Elements by RepeatMasker
6.65911
wholBlood F XV7Q
0.16544
GTEx RNA signal from female Whole Blood (GTEX-XV7Q-0005-SM-4BRWI)
14.5824
wholBlood F 13FTX
0.01045
GTEx RNA signal from female Whole Blood (GTEX-13FTX-0005-SM-5N9F6)
10.2506
wholBlood F 13VXT
0.12616
GTEx RNA signal from female Whole Blood (GTEX-13VXT-0005-SM-5N9F3)
10.8763
wholBlood F 14BMU
0.01221
GTEx RNA signal from female Whole Blood (GTEX-14BMU-0006-SM-5MR3T)
8.1534
wholBlood F 139D8
0.0604
GTEx RNA signal from female Whole Blood (GTEX-139D8-0006-SM-5LZX8)
8.385
wholBlood F 12WSJ
0.12361
GTEx RNA signal from female Whole Blood (GTEX-12WSJ-0005-SM-5LU8S)
Scale
chr1:
50 bases
hg19
17,393,710
17,393,720
17,393,730
17,393,740
17,393,750
17,393,760
17,393,770
17,393,780
17,393,790
---> A A A A A A A A A G A A A A A A A A G A A A G A A A A C C C A T A A A C C C A C C A T T A A C C A A A C A C A C A C A C A C A C A T G A C A A A C T C T A A G T C T C C A G A C A G A C A C C C T C A A A
Simple Nucleotide Polymorphisms (dbSNP 150) Found in >= 1% of Samples
rs719262
rs199539170
rs34043625
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
PADI2
PADI2
RepeatMasker
Repeating Elements by RepeatMasker
4.13609
wholBlood F XV7Q
0.16544
GTEx RNA signal from female Whole Blood (GTEX-XV7Q-0005-SM-4BRWI)
10.4115
wholBlood F 13FTX
3.3137
GTEx RNA signal from female Whole Blood (GTEX-13FTX-0005-SM-5N9F6)
7.52765
wholBlood F 13VXT
2.6494
GTEx RNA signal from female Whole Blood (GTEX-13VXT-0005-SM-5N9F3)
9.02089
wholBlood F 14BMU
0.56762
GTEx RNA signal from female Whole Blood (GTEX-14BMU-0006-SM-5MR3T)
7.83632
wholBlood F 139D8
0.37747
GTEx RNA signal from female Whole Blood (GTEX-139D8-0006-SM-5LZX8)
7.94205
wholBlood F 12WSJ
2.19411
GTEx RNA signal from female Whole Blood (GTEX-12WSJ-0005-SM-5LU8S)

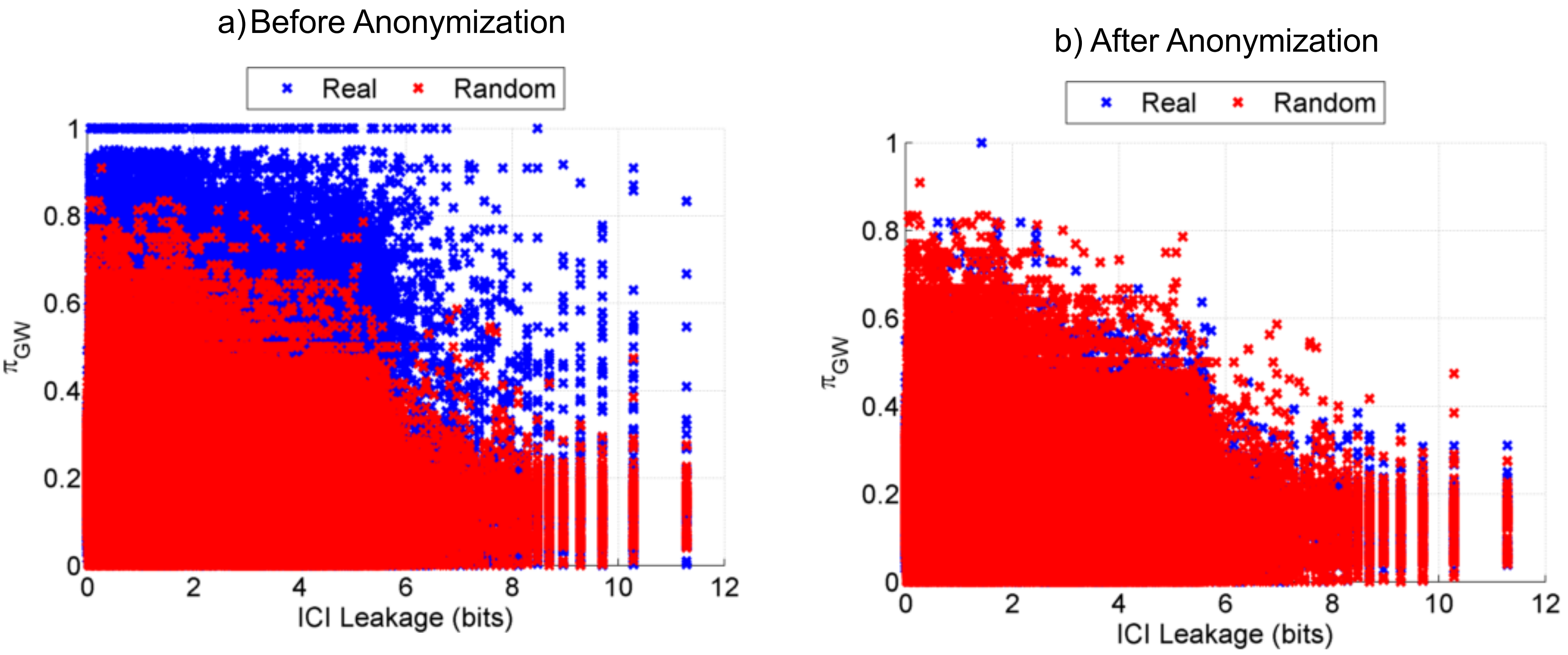
[Harmanci & Gerstein, *Nat. Comm.* ('18)]

# Example of Large Deletion Evident in Signal Profile



[Harmanci & Gerstein, *Nat. Comm.* ('18)]

# Information Leakage from SV Deletions

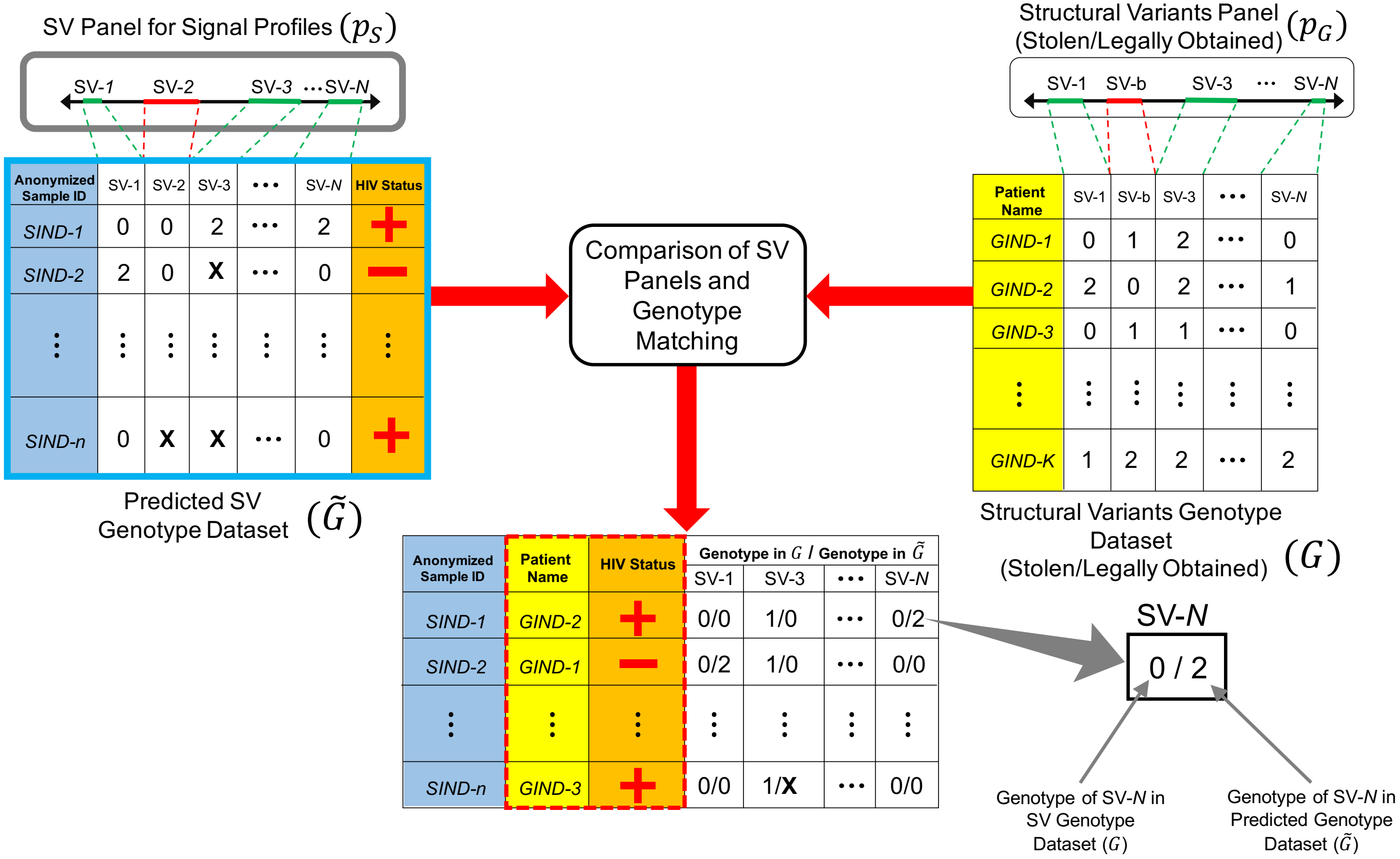

a) Before Anonymization

b) After Anonymization

Simple anonymization procedure (filling in deletion by value at endpoints) has dramatic effect

[Harmanci & Gerstein, *Nat. Comm.* ('18)]

# Another type of Linking Attack: Linking based on SV Genotyping



SV Panel for Signal Profiles $(p_S)$

| Anonymized Sample ID | SV-1 | SV-2 | SV-3 | ··· | SV-N | HIV Status |
|---|---|---|---|---|---|---|
| SIND-1 | 0 | 0 | 2 | ··· | 2 | + |
| SIND-2 | 2 | 0 | **X** | ··· | 0 | − |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| SIND-n | 0 | **X** | **X** | ··· | 0 | + |

Predicted SV Genotype Dataset $(\tilde{G})$

Structural Variants Panel (Stolen/Legally Obtained) $(p_G)$

| Patient Name | SV-1 | SV-b | SV-3 | ··· | SV-N |
|---|---|---|---|---|---|
| GIND-1 | 0 | 1 | 2 | ··· | 0 |
| GIND-2 | 2 | 0 | 2 | ··· | 1 |
| GIND-3 | 0 | 1 | 1 | ··· | 0 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| GIND-K | 1 | 2 | 2 | ··· | 2 |

Structural Variants Genotype Dataset (Stolen/Legally Obtained) $(G)$

Comparison of SV Panels and Genotype Matching

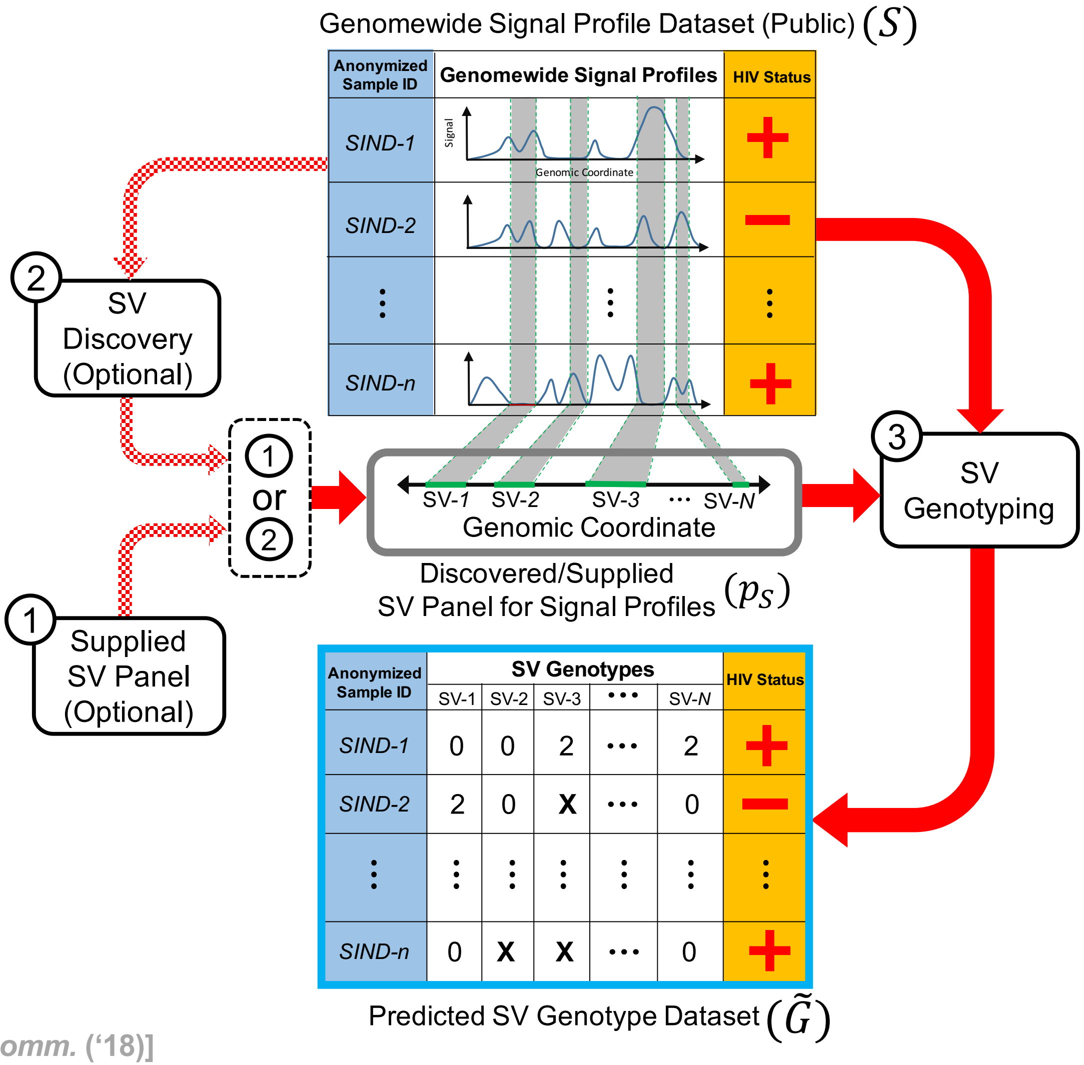
| Anonymized Sample ID | Patient Name | HIV Status | Genotype in $G$ / Genotype in $\tilde{G}$ | | | |
|---|---|---|---|---|---|---|
| | | | SV-1 | SV-3 | ··· | SV-N |
| SIND-1 | GIND-2 | + | 0/0 | 1/0 | ··· | 0/2 |
| SIND-2 | GIND-1 | − | 0/2 | 1/0 | ··· | 0/0 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| SIND-n | GIND-3 | + | 0/0 | 1/**X** | ··· | 0/0 |

Genotype of SV-N in SV Genotype Dataset $(G)$

Genotype of SV-N in Predicted Genotype Dataset $(\tilde{G})$

[Harmanci & Gerstein, *Nat. Comm.* ('18)]

# Another type of Linking Attack: First Doing SV Genotyping



Genomewide Signal Profile Dataset (Public) $(S)$

| Anonymized Sample ID | Genomewide Signal Profiles | HIV Status |
|---|---|---|
| *SIND-1* | | + |
| *SIND-2* | | − |
| ⋮ | ⋮ | ⋮ |
| *SIND-n* | | + |

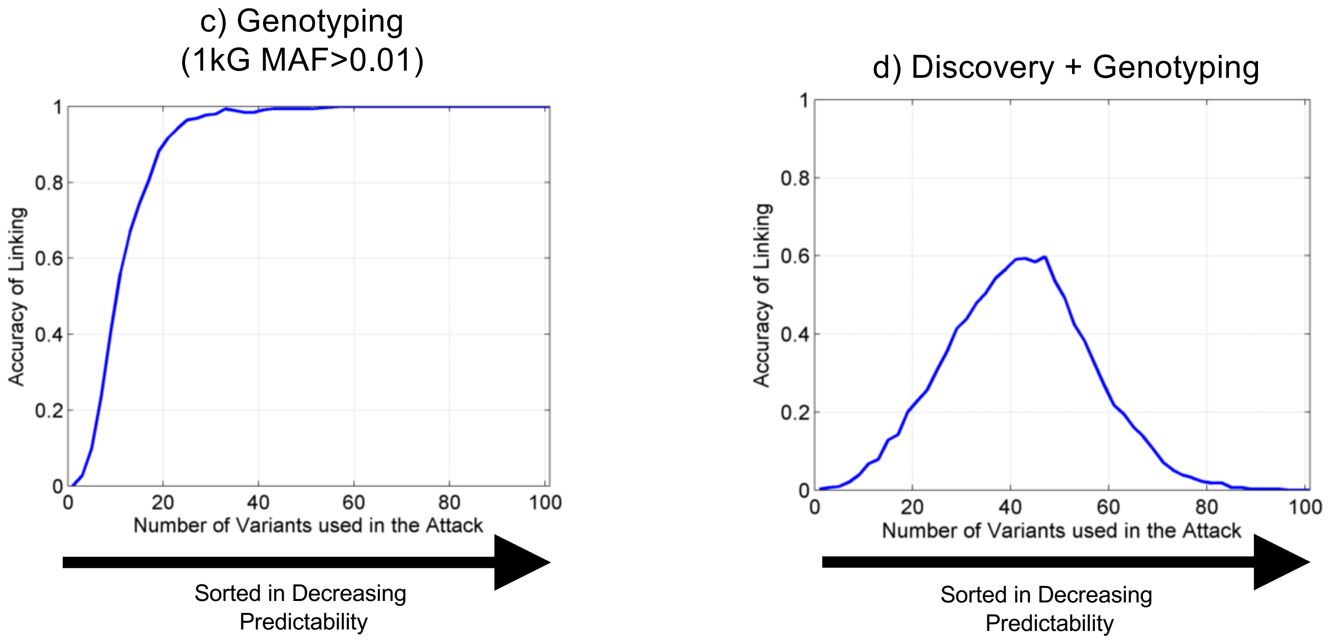
Discovered/Supplied SV Panel for Signal Profiles $(p_S)$

| Anonymized Sample ID | SV Genotypes | | | | | HIV Status |
|---|---|---|---|---|---|---|
| | SV-1 | SV-2 | SV-3 | ••• | SV-*N* | |
| *SIND-1* | 0 | 0 | 2 | ••• | 2 | + |
| *SIND-2* | 2 | 0 | X | ••• | 0 | − |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| *SIND-n* | 0 | X | X | ••• | 0 | + |

Predicted SV Genotype Dataset $(\tilde{G})$

[Harmanci & Gerstein, *Nat. Comm.* ('18)]

# Linking Attack Based on SV Deletions in gEUVADIS Dataset



[Harmanci & Gerstein, *Nat. Comm.* ('18)]

# 2 Sides of a Coin: Individual Privacy v. Easy RNA-Seq Mining

- **Introduction to Genomic Privacy**
  - The dilemma: The genome as fundamental, inherited info that's very private v need for large-scale mining for med. research
  - 2-sided nature of RNA-seq presents a particularly tricky privacy issue
- **Measuring Leakage from eQTLs**
  - Quantifying & removing further variant info from expression levels + eQTLs using ICI & predictability
- **Linking Attacks from eQTLs**
  - Instantiating a practical linking attack using extreme expression levels
- **Signal Profiles**
  - Appreciable leakage from large & small deletions evident in signal profiles
  - Linking attacks also possible but additional complication of SV discovery in addition to genotyping

# 2 Sides of a Coin: Individual Privacy v. Easy RNA-Seq Mining

- **Introduction to Genomic Privacy**
  - The dilemma: The genome as fundamental, inherited info that’s very private v need for large-scale mining for med. research
  - 2-sided nature of RNA-seq presents a particularly tricky privacy issue
- **Measuring Leakage from eQTLs**
  - Quantifying & removing further variant info from expression levels + eQTLs using ICI & predictability
- **Linking Attacks from eQTLs**
  - Instantiating a practical linking attack using extreme expression levels
- **Signal Profiles**
  - Appreciable leakage from large & small deletions evident in signal profiles
  - Linking attacks also possible but additional complication of SV discovery in addition to genotyping

## Acknowledgements

papers.gersteinlab.org/subject/privacy –

D **Greenbaum**

**PrivaSeq**.gersteinlab.org

**PrivaSig**.gersteinlab.org

A **Harmanci**

# Info about content in this slide pack

- General PERMISSIONS
  - This Presentation is copyright Mark Gerstein, Yale University, 2017.
  - Please read permissions statement at www.**gersteinlab.org/misc/permissions.html** .
  - Feel free to use slides & images in the talk with PROPER acknowledgement (via citation to relevant papers or link to gersteinlab.org).
  - Paper references in the talk were mostly from Papers.GersteinLab.org.

- PHOTOS & IMAGES. For thoughts on the source and permissions of many of the photos and clipped images in this presentation see http://streams.gerstein.info .
  - In particular, many of the images have particular EXIF tags, such as kwpotppt , that can be easily queried from flickr, viz: http://www.flickr.com/photos/mbgmbg/tags/kwpotppt