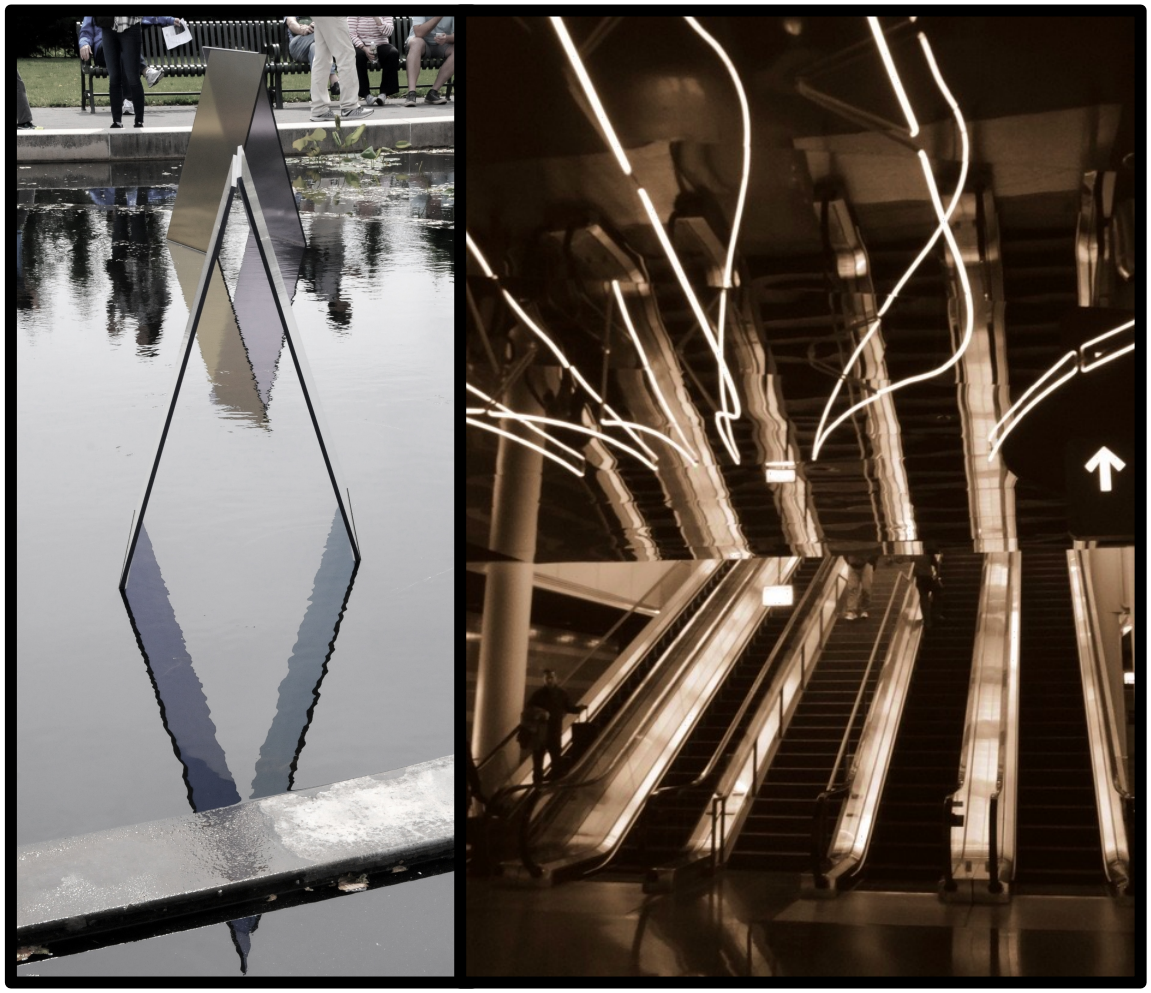
Genomic Privacy & the 2-sided Nature of RNA-seq: Quantifying the Leakage of Individual Information



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

Activity Patterns

- RNA Seq. gives rise to activity patterns of genes & regions in the genome

RNA-Seq Overview

Successive steps of Data Reduction

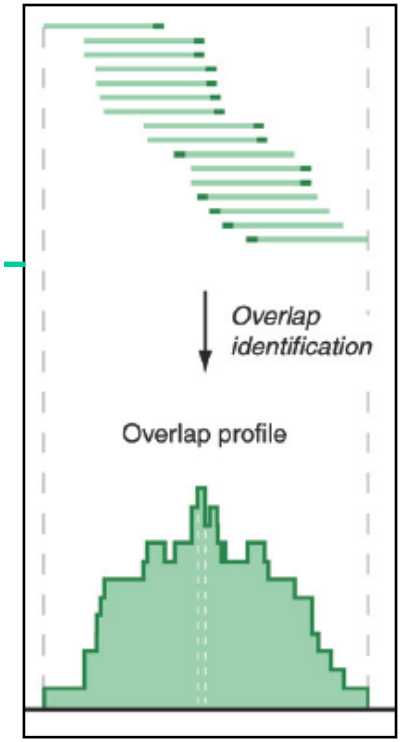
Fastq sequence files
~5-10 GB

```
ATACAAGCAAGTATAAGTTCGTATGCCGTCTT
GGAGGCTGGAGTTGGGGACGTATGCGGCATAG
TACCGATCGAGTCGACTGTAAACGTAGGCATA
ATTCTGACTGGTGTCATGCTGATGTACTTAAA
```

Index-building + Alignment to reference genome

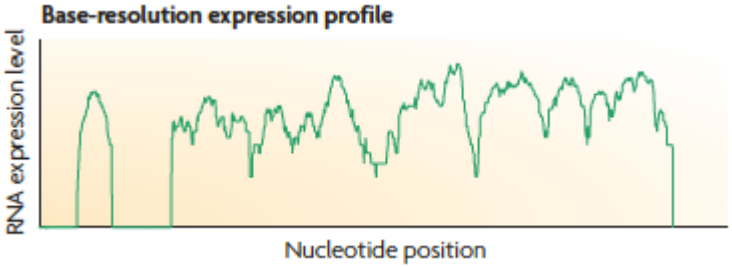
BAM files
~1-2-fold reduction

Conversion to signal track by overlapping reads



Reads => Signal

BigWig files
~25-fold reduction



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

Mapping to genes

Gene/Transcript expression matrix
~20-fold reduction

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

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: basic patterns and clusters of co-active genes across an organisms development
  - Evolutionary relationships: behavior preserved across a wide range of organisms
  - Tissue- and cell-type
  - Disease phenotypes: what genes go up in cancer?
- **Above are not tied to an individual's genotype. However, data is derived from an individual & tagged with an individual's genotype**
- Note, a few calculations aim to use explicitly genotype to derive general relations related to sequence variation & gene expression (eg allelic activity)

Importance of Leakage Quantification for Genomic Privacy

- The **overall dilemma of genomic privacy**
  - From sharing information, the individual is potentially harmed but society benefits in terms of medical research
  - How to balance risks v rewards?
- **Need to quantify leakage**
  - Cost Benefit Analysis: how helpful is identifiable data in genomic research v. potential harm from a breach
  - What is acceptable risk – ie what is acceptable data leakage?
- Also, **need 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)]

Genomic Privacy & the 2-sided Nature of RNA-seq: Quantifying the Leakage of Individual Information

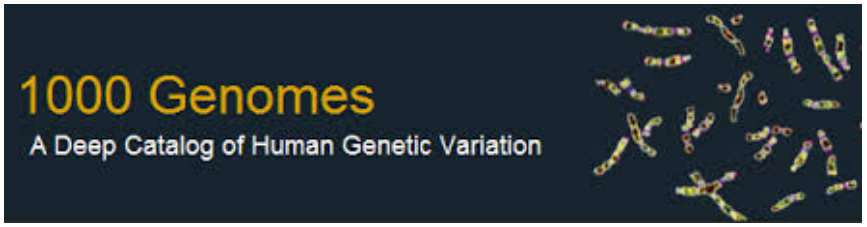
- **Intro on RNA-seq & the General Dilemma of Genomic Privacy**
  - RNA-seq Presents a tricky privacy issue since much of the sequencing is for general, non-individual specific results yet it's tagged with individual information
  - The need to quantify leaks
- **Quantifying RNA-seq Leakage …from Reads**
  - Almost as much as WGS
  - But can remove SNVs in reads w/ MRF
- **…from eQTLs**
  - Quantifying & removing variant info from expression levels + eQTLs using ICI & predictability
  - Instantiating a practical linking attack
- **…from Indels/SVs**
  - Another source of leakage in RNA-seq data & how to use these for a related linking attack

Genomic Privacy & the 2-sided Nature of RNA-seq: Quantifying the Leakage of Individual Information

- Intro on RNA-seq & the General **Dilemma of Genomic Privacy**
  - RNA-seq Presents a tricky privacy issue since much of the sequencing is for general, non-individual specific results yet it's tagged with individual information
  - The need to quantify leaks
- **Quantifying RNA-seq Leakage …from Reads**
  - Almost as much as WGS
  - But can remove SNVs in reads w/ MRF
- **…from eQTLs**
  - Quantifying & removing variant info from expression levels + eQTLs using ICI & predictability
  - Instantiating a practical linking attack
- **…from Indels/SVs**
  - Another source of leakage in RNA-seq data & how to use these for a related linking attack

Representative Expression, Genotype, eQTL Datasets on Open Datasets

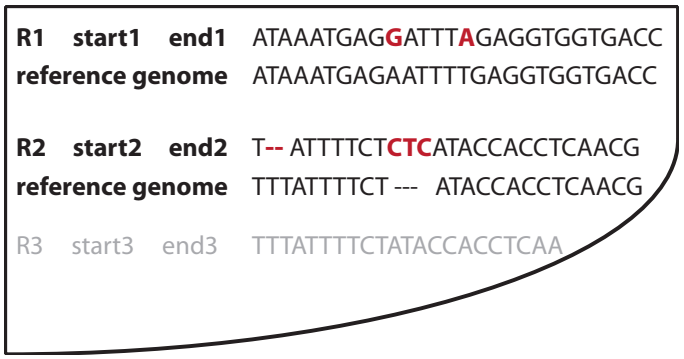
- Publically available genotypes (not controlled access) are available from the 1000 Genomes Project
- mRNA sequencing for 462 individuals from gEUVADIS and ENCODE
  - Publicly available quantification for protein coding genes
- Approximately 3,000 cis-eQTL (FDR<0.05)

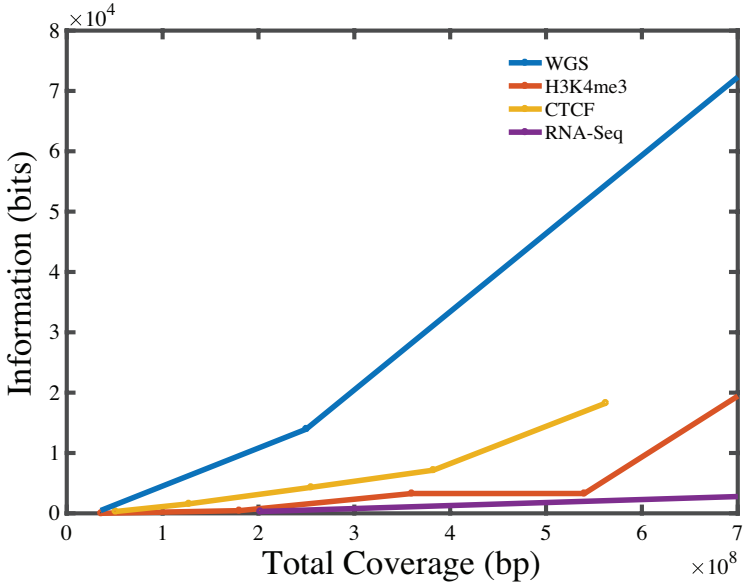


Variants directly in the reads

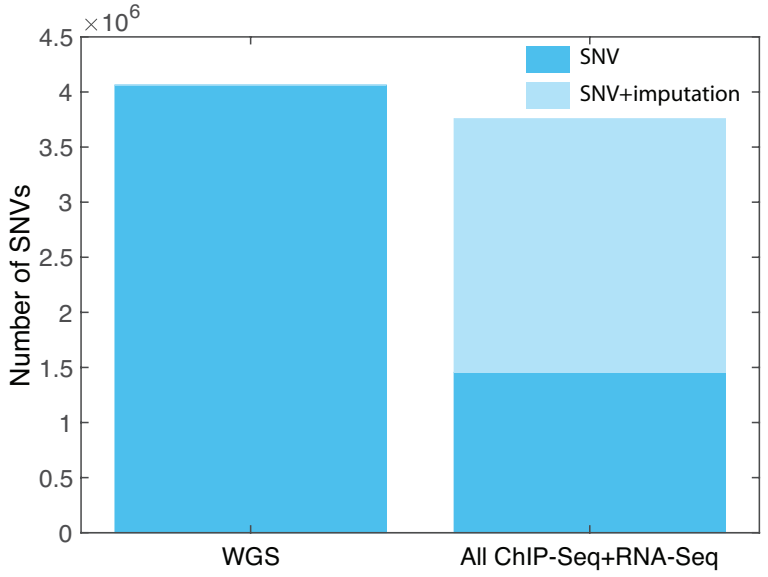
- Functional genomics data comes with a great deal of sequencing
  - NA12878 as case study - 1000 genomes variants are used as gold standard
- How much information, for example, do RNA-Seq reads (or ChIP-Seq) reads contain? Does that information enough to identify individuals?

Variants from RNA-Seq reads



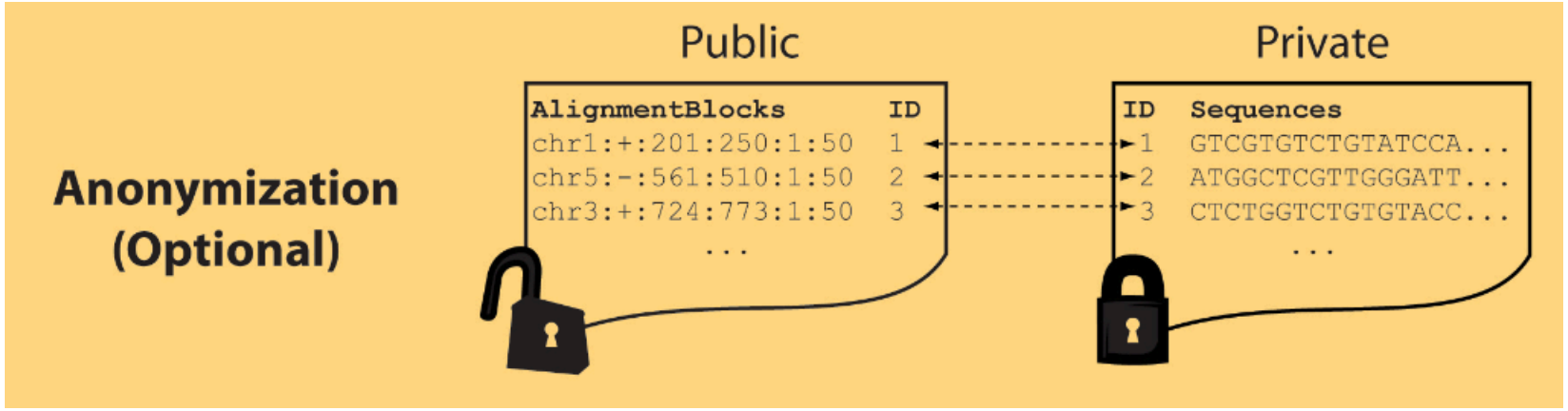


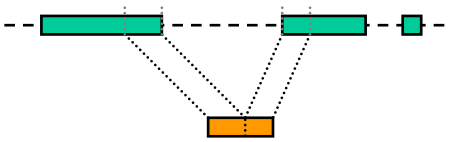
- It might seem like we don't infer much information from single ChIP-Seq and RNA-Seq experiments compared to WGS
  - However putting 10 different ChIP-Seq experiments and RNA-Seq together with imputation provides a great deal of information about the individual



Light-weight formats to Hide Most of the Read Data (Signal Tracks)

- Some lightweight format clearly separate public & private info., aiding exchange
- Files become much smaller
- Distinction between formats to compute on and those to archive with – become sharper with big data





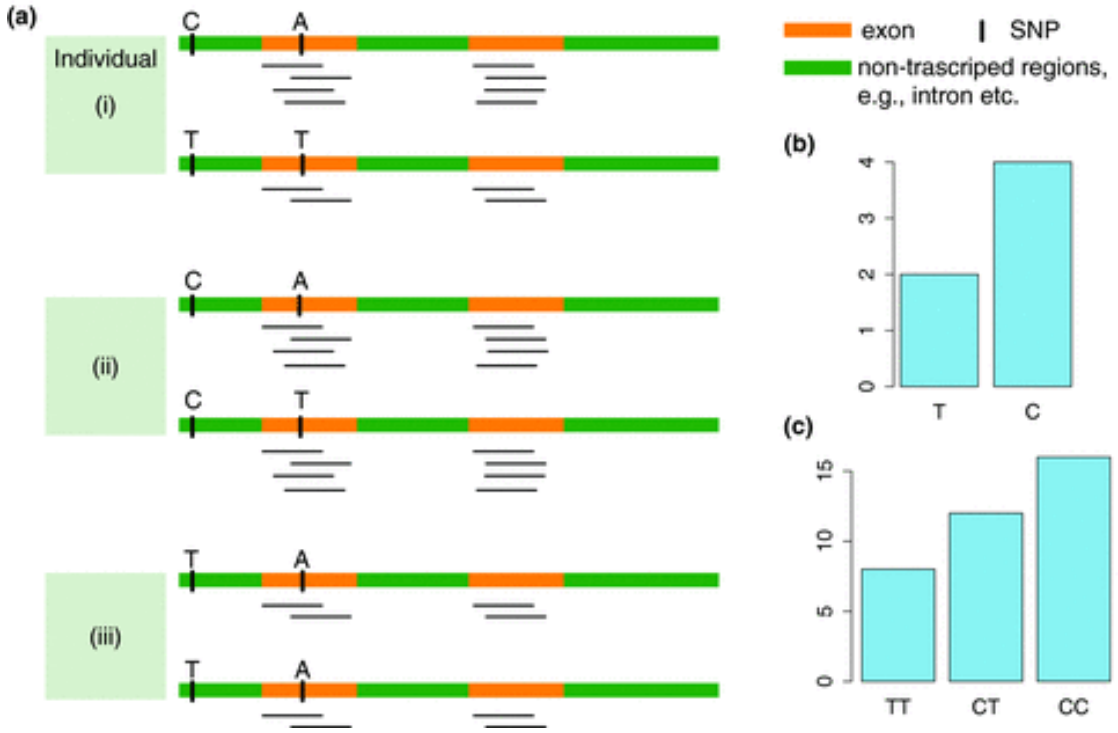
Mapping coordinates without variants (MRF)

Reads (linked via ID, 10X larger than mapping coord.)

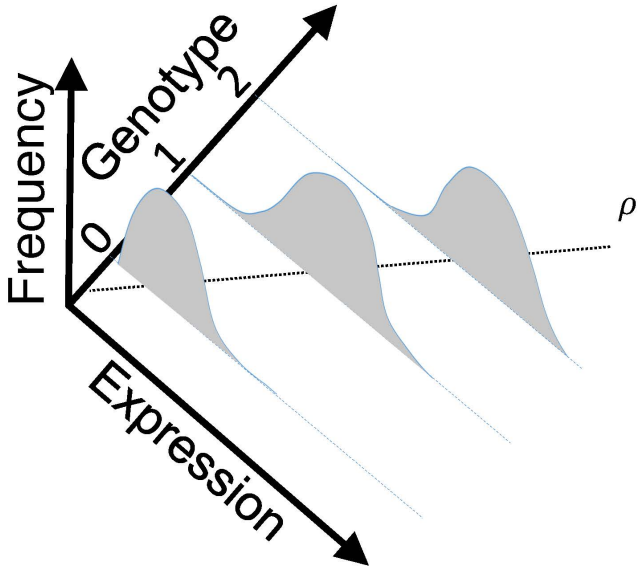
[*Bioinformatics* 27: 281]

Genomic Privacy & the 2-sided Nature of RNA-seq: Quantifying the Leakage of Individual Information

- Intro on RNA-seq & the General **Dilemma of Genomic Privacy**
  - RNA-seq Presents a tricky privacy issue since much of the sequencing is for general, non-individual specific results yet it's tagged with individual information
  - The need to quantify leaks
- **Quantifying RNA-seq Leakage …from Reads**
  - Almost as much as WGS
  - But can remove SNVs in reads w/ MRF
- **…from eQTLs**
  - Quantifying & removing variant info from expression levels + eQTLs using ICI & predictability
  - Instantiating a practical linking attack
- **…from Indels/SVs**
  - Another source of leakage in RNA-seq data & how to use these for a related linking attack



[*Biometrics 68(1) 1–11*]

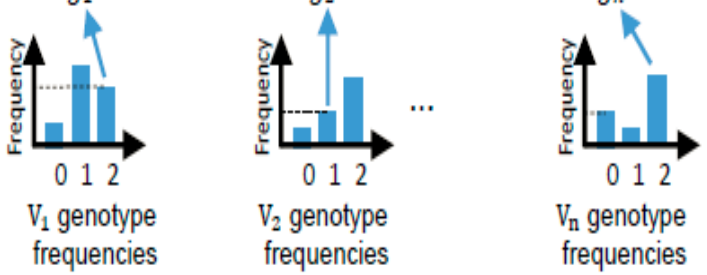


eQTL Mapping Using RNA-Seq Data

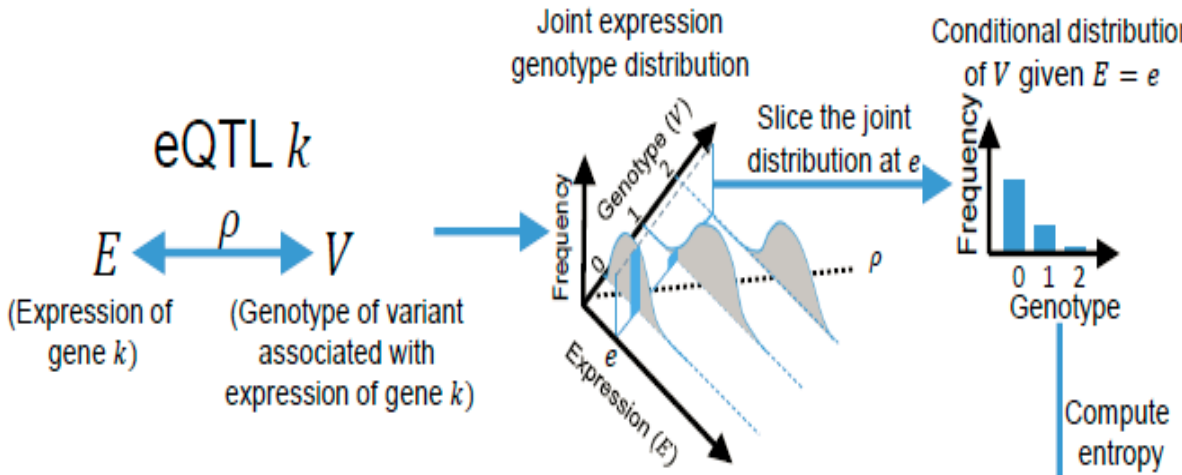
- eQTLs are genomic loci that contribute to variation in mRNA expression levels
- eQTLs provide insights on transcription regulation, and the molecular basis of phenotypic outcomes

Information Content and Predictability

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



- **Higher frequency: Lower ICI**
- **Lower frequency: Higher ICI**
- **Additive for multiple variants**

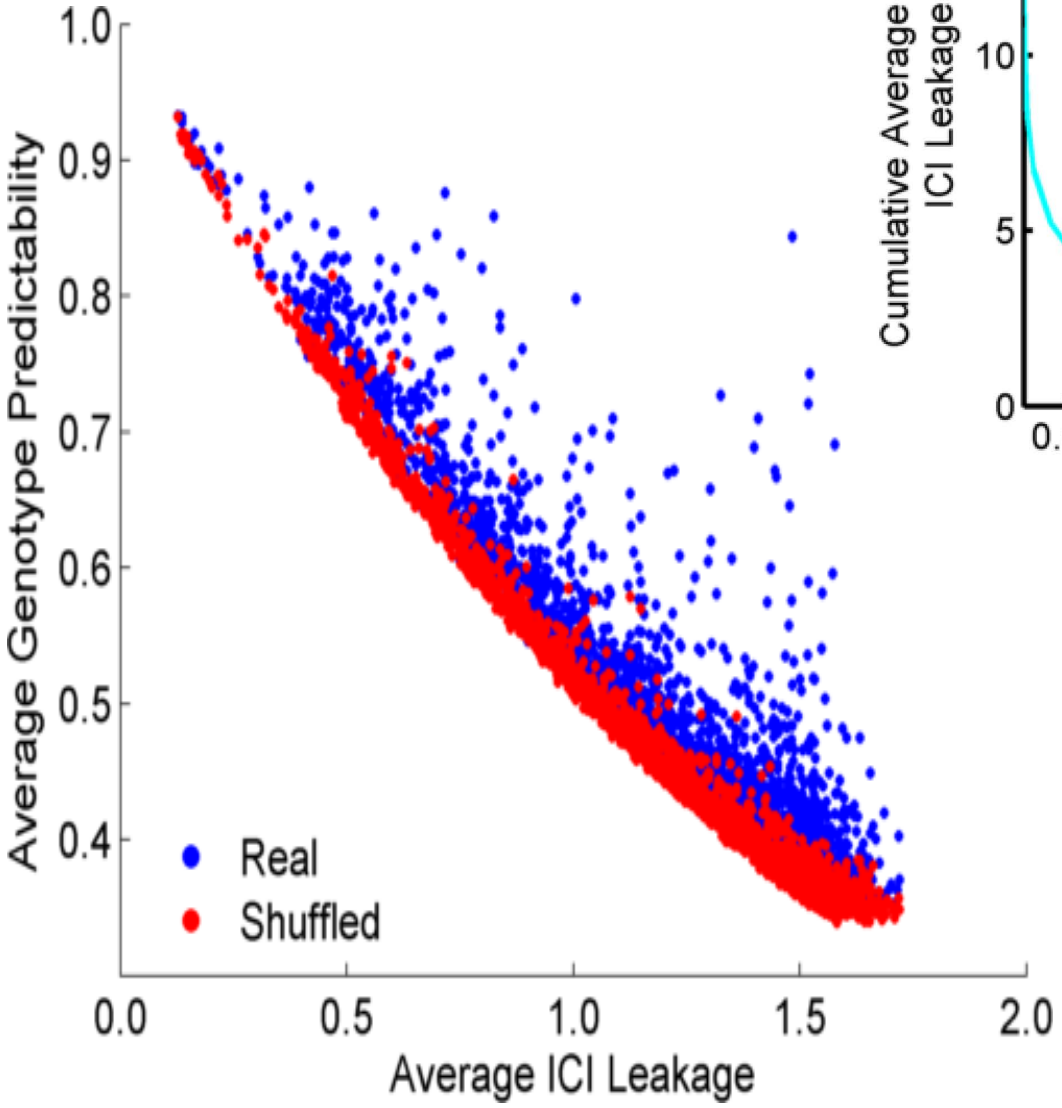


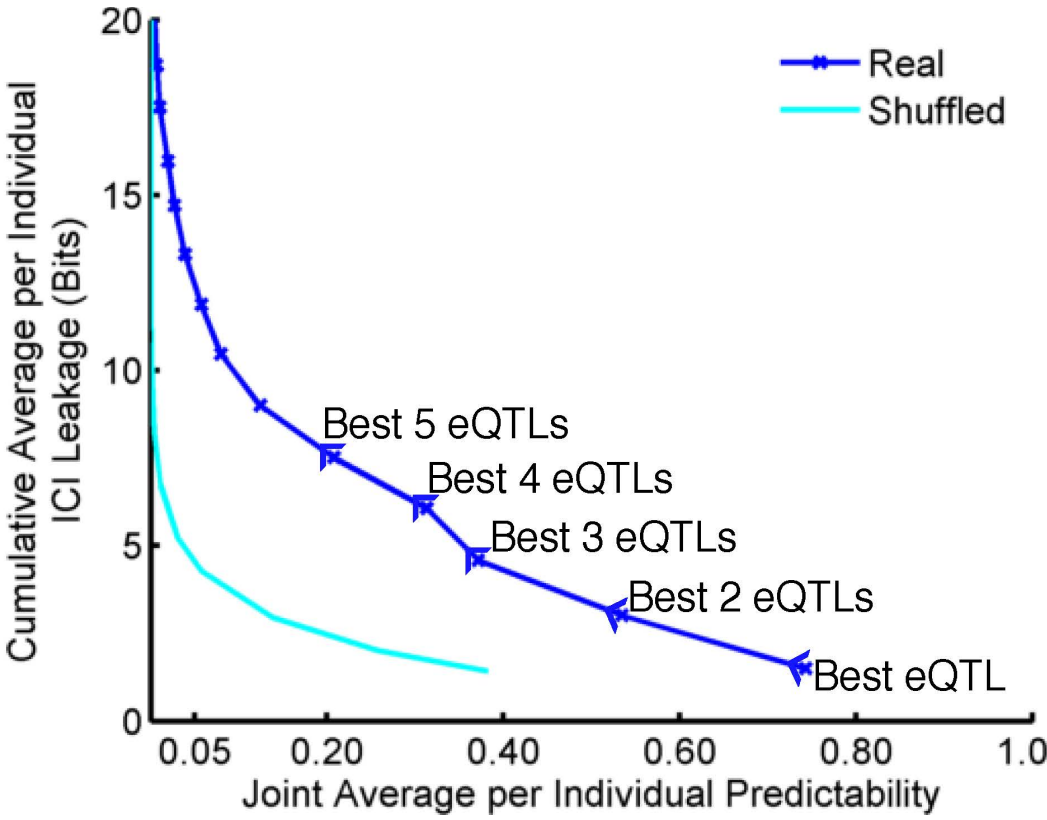
$$\pi\left(V \middle| E = e\right) = e^{-H\left(V \middle| E = e\right)}$$

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

[Harmanci et al. Nat. Meth. 2016]

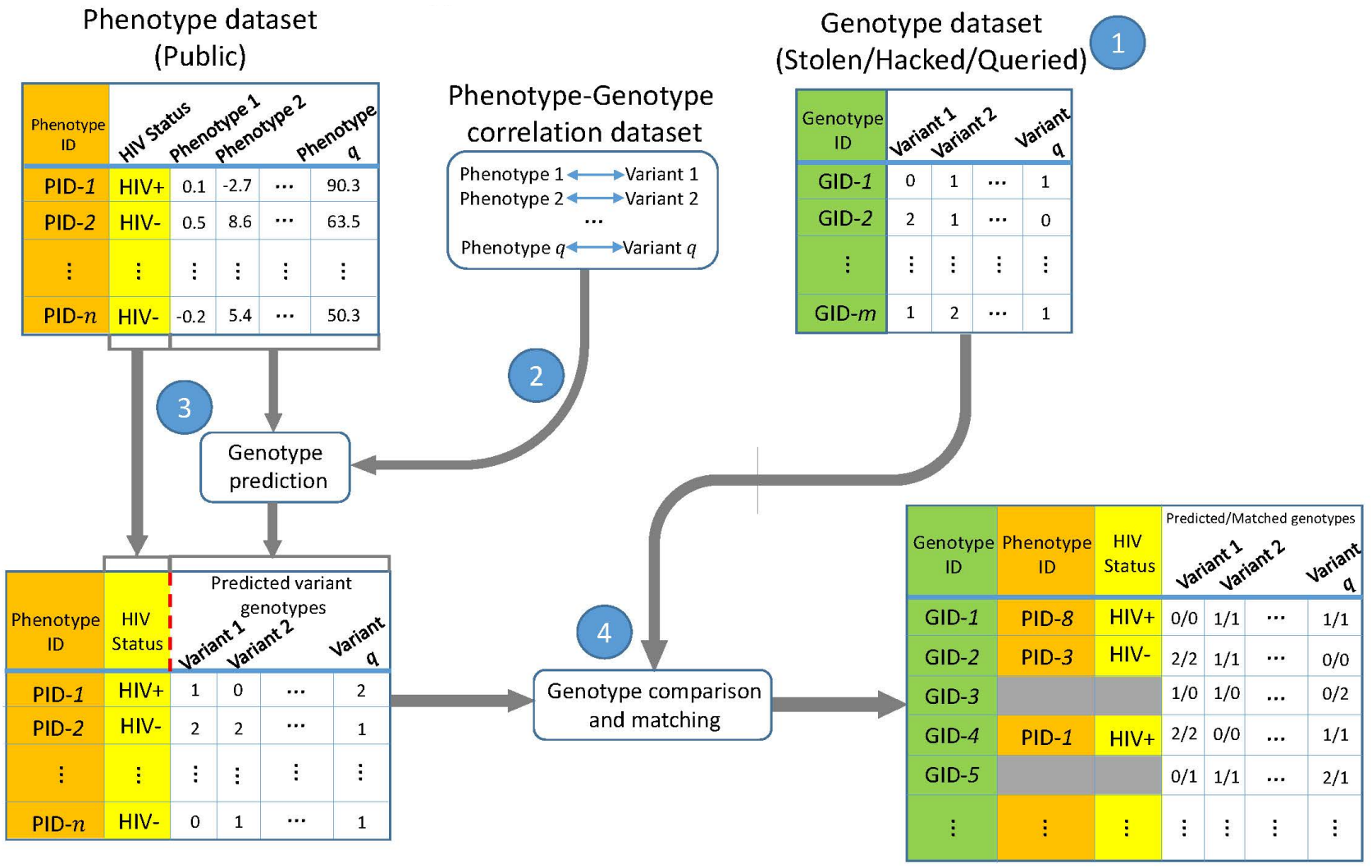
[Harmanciet al. Nat. Meth. (‘16]





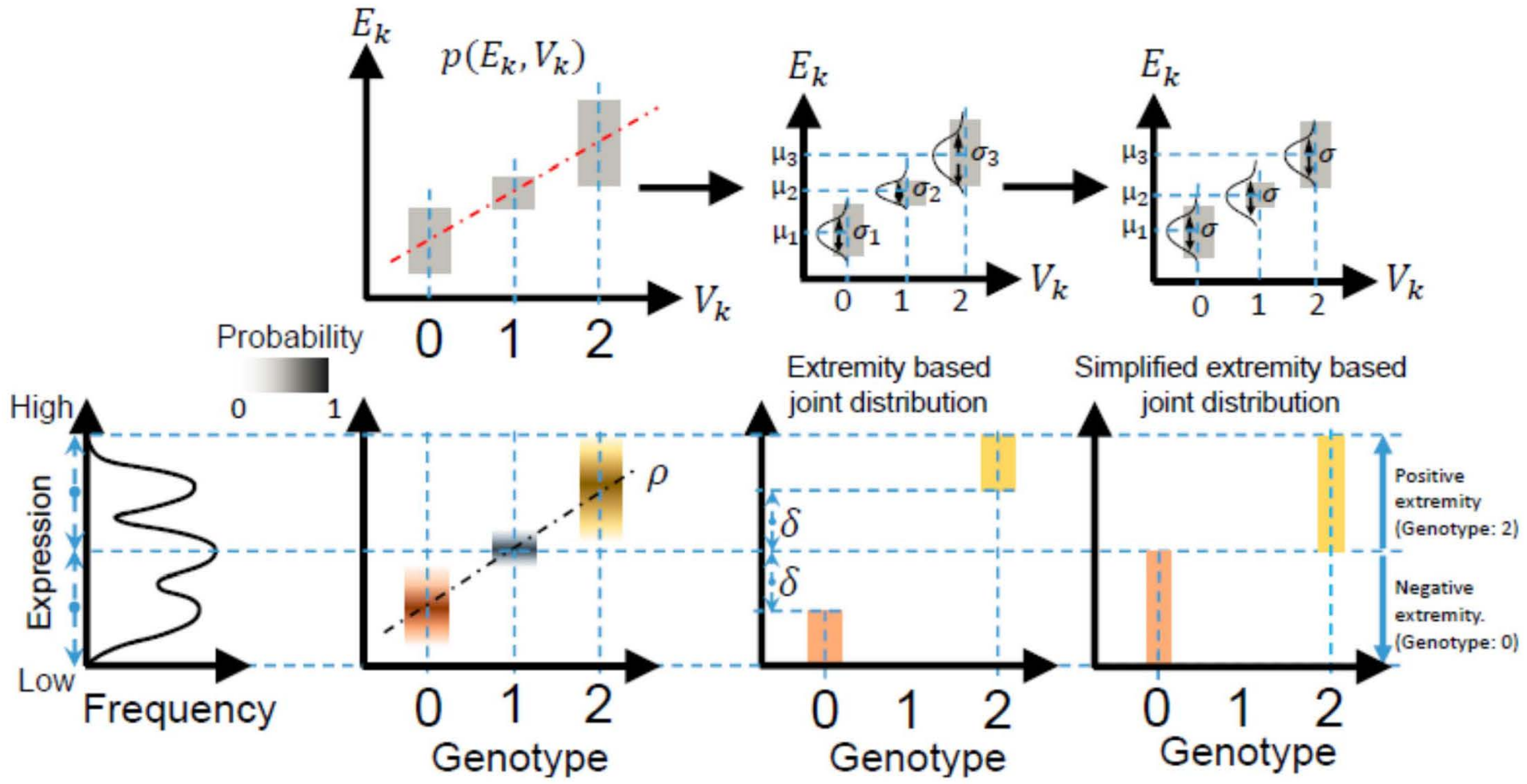
ICI Leakage versus Genotype Predictability

Linking Attack Scenario



[Harmanciet al. Nat. Meth. (in revision)]

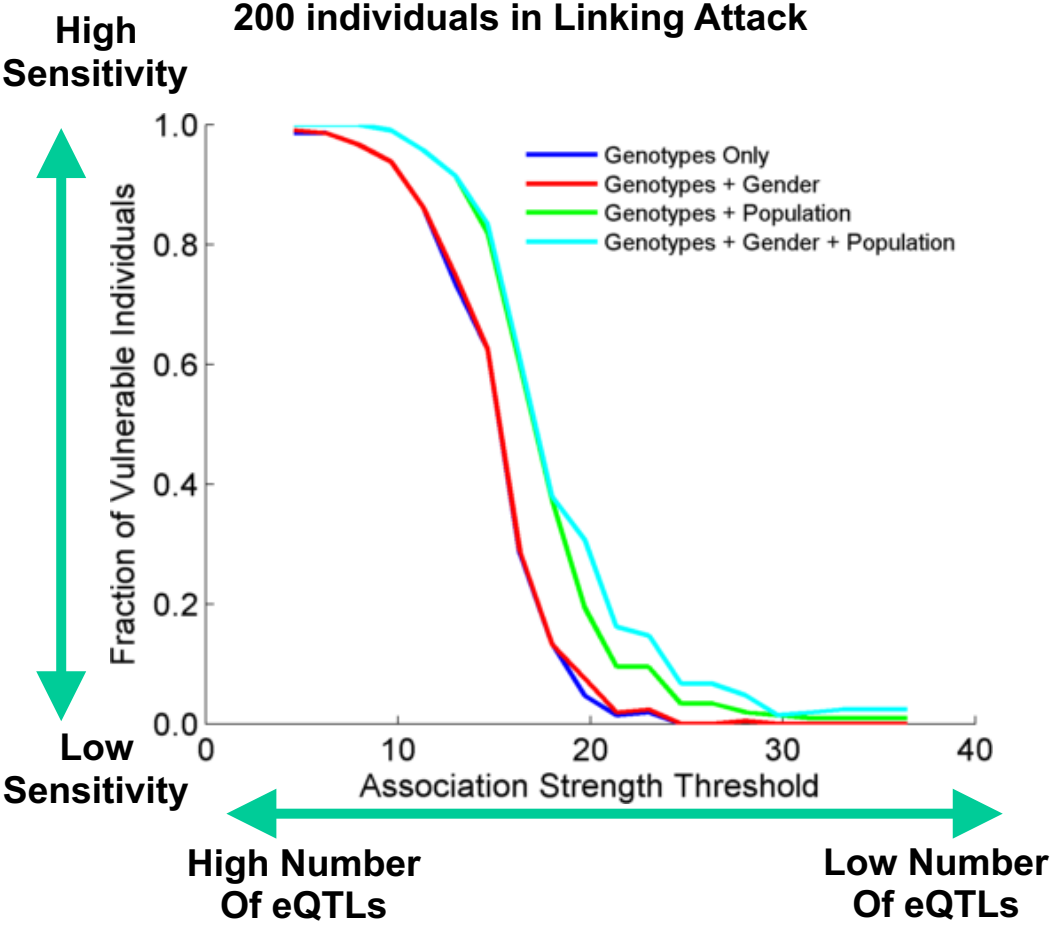
Levels of Expression-Genotype Model Simplifications for Genotype Prediction



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

Success in Linking Attack with Extremity based Genotype Prediction

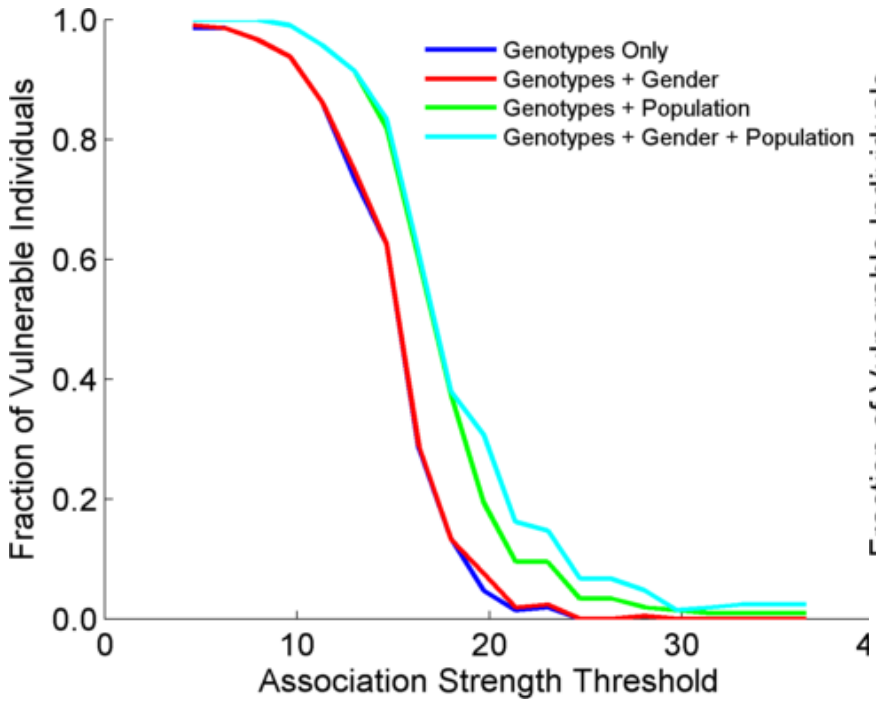
200 individuals eQTL Discovery
200 individuals in Linking Attack



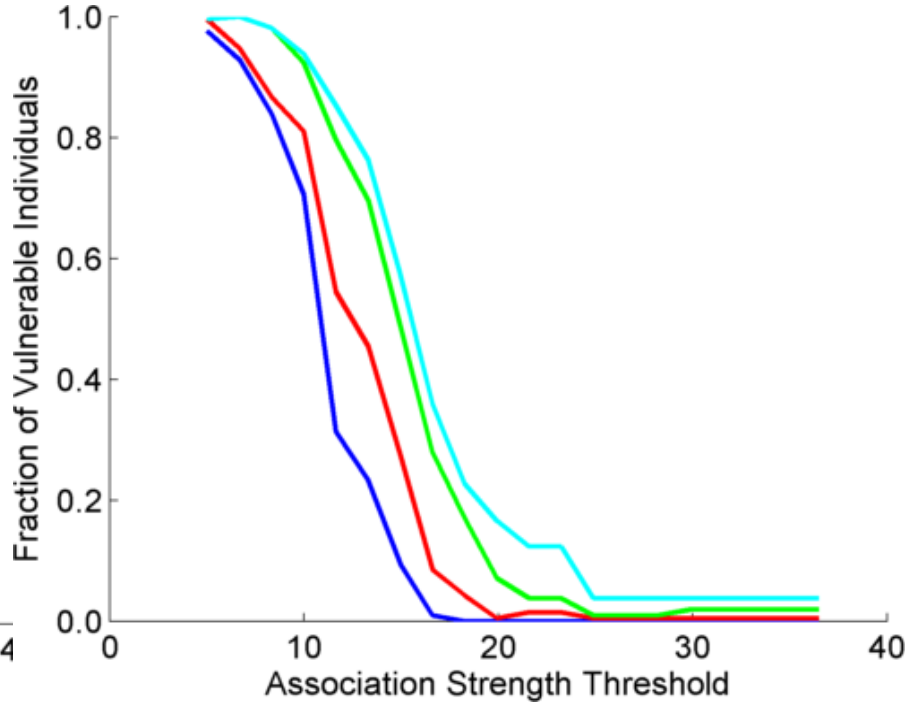
[Harmanci et al. *Nat. Meth.* (16)]

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

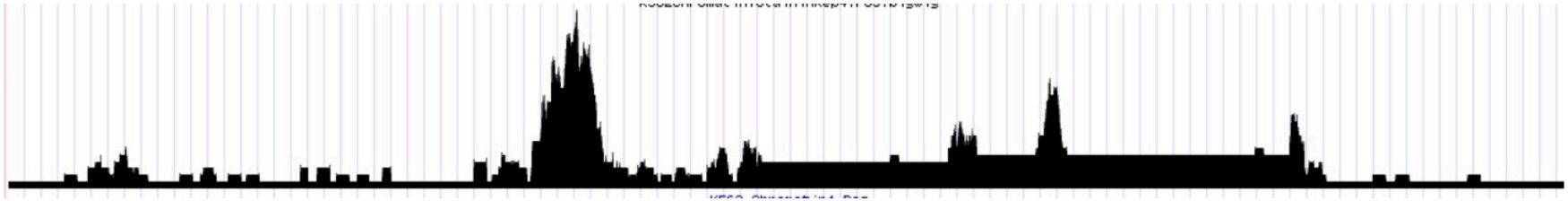


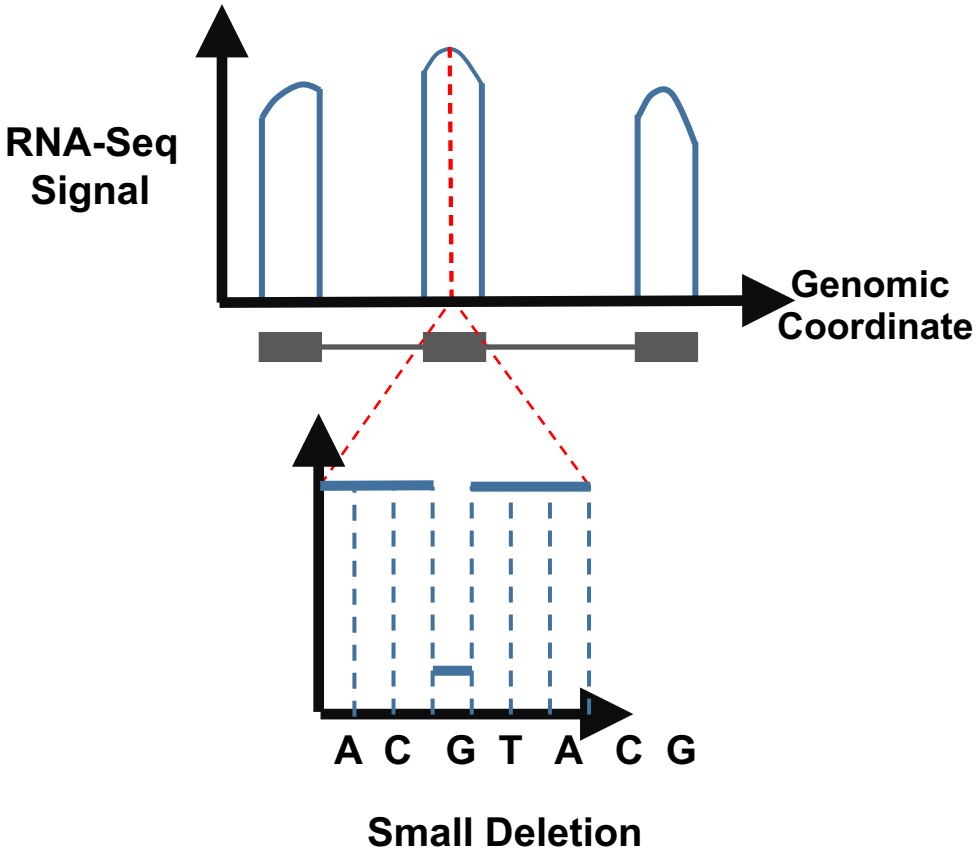
[Harmanci et al. *Nat. Meth.* (16)]

Genomic Privacy & the 2-sided Nature of RNA-seq: Quantifying the Leakage of Individual Information

- Intro on RNA-seq & the General **Dilemma of Genomic Privacy**
  - RNA-seq Presents a tricky privacy issue since much of the sequencing is for general, non-individual specific results yet it's tagged with individual information
  - The need to quantify leaks
- **Quantifying RNA-seq Leakage …from Reads**
  - Almost as much as WGS
  - But can remove SNVs in reads w/ MRF
- **…from eQTLs**
  - Quantifying & removing variant info from expression levels + eQTLs using ICI & predictability
  - Instantiating a practical linking attack
- **…from Indels/SVs**
  - Another source of leakage in RNA-seq data & how to use these for a related linking attack

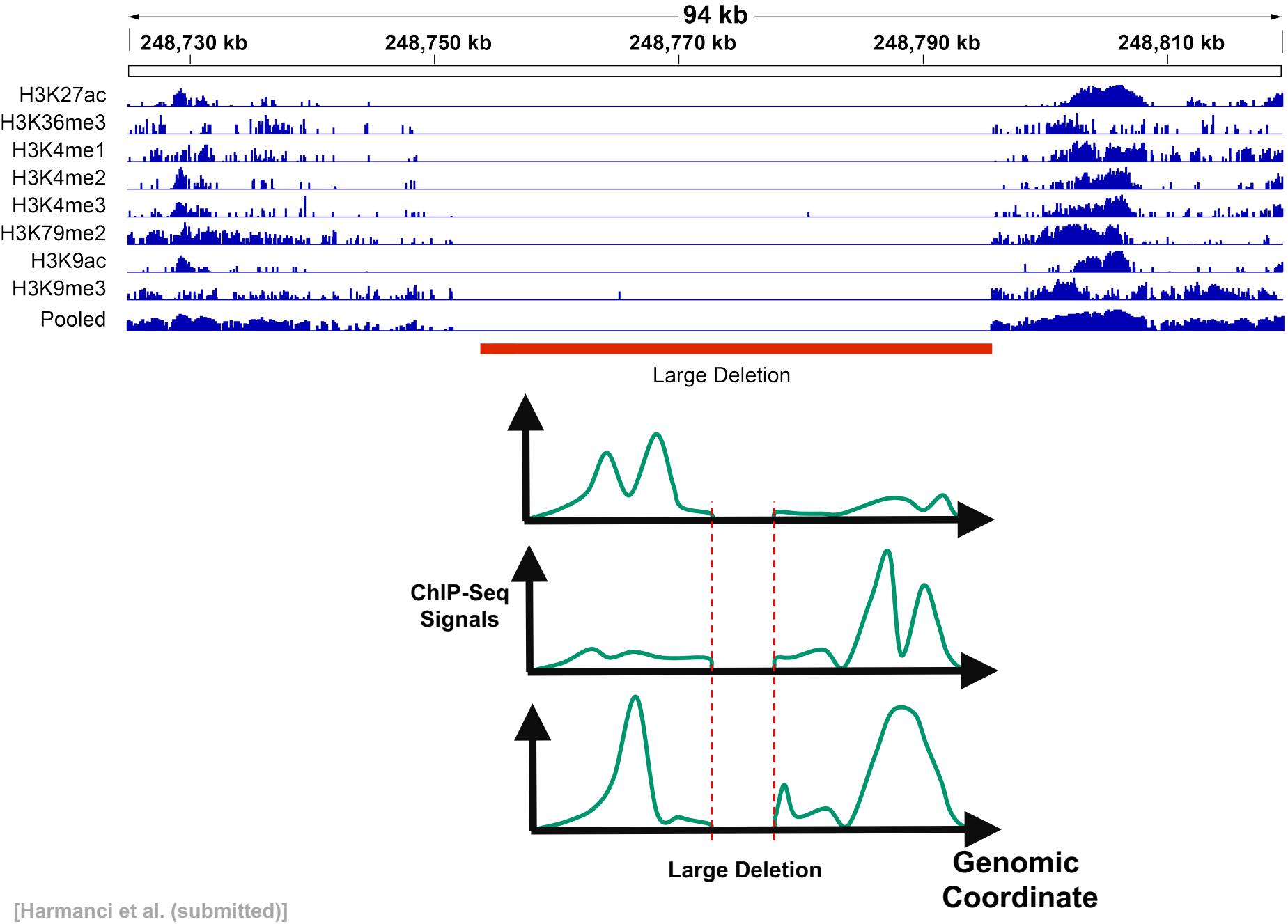
Small Deletions and RNA-seq Signal



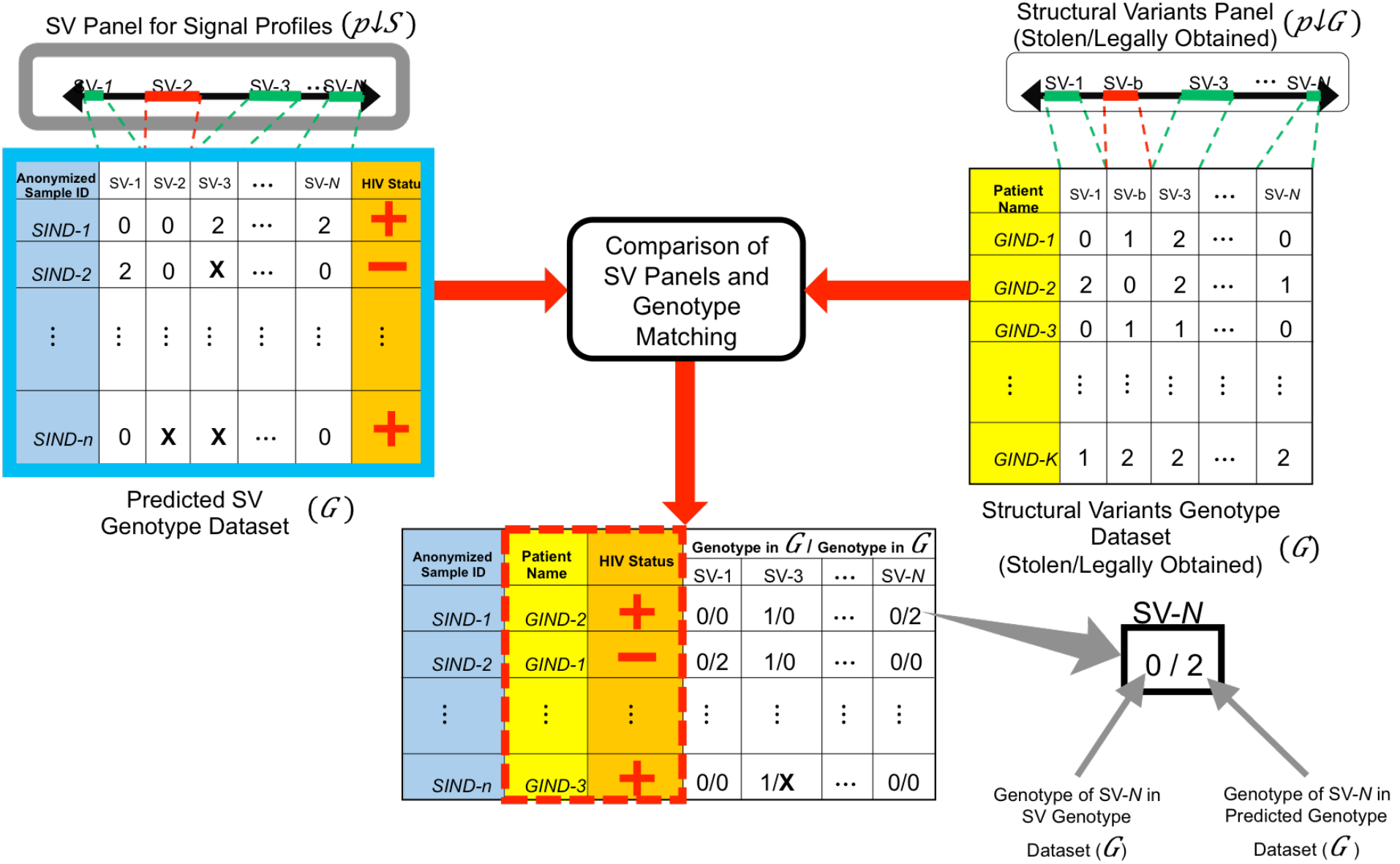


[Harmanci et al. (submitted)]

Structural Variants are Generally Detectable from Functional Genomics Data

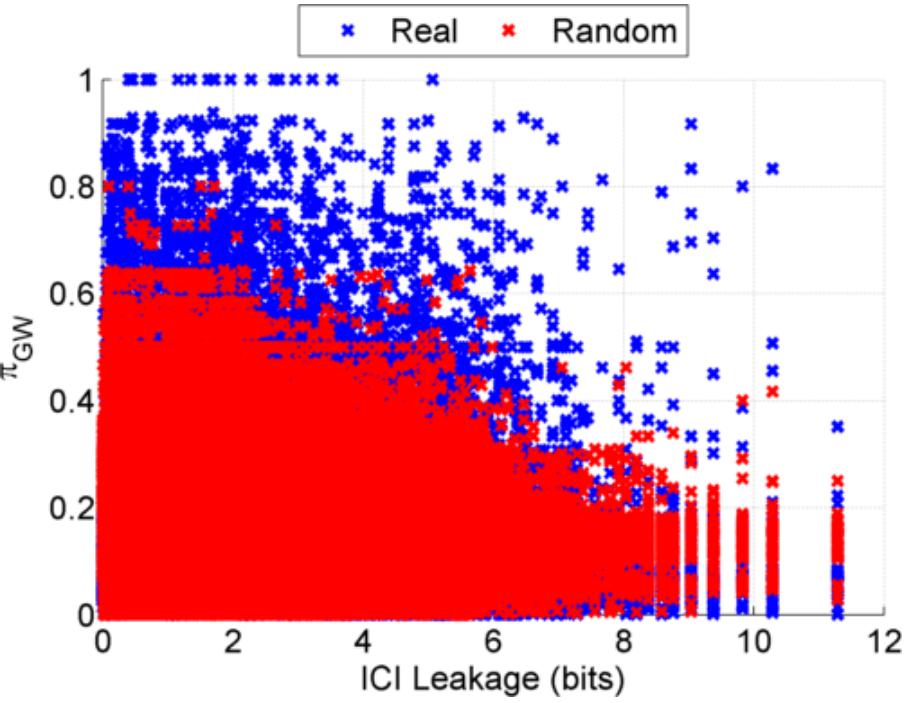


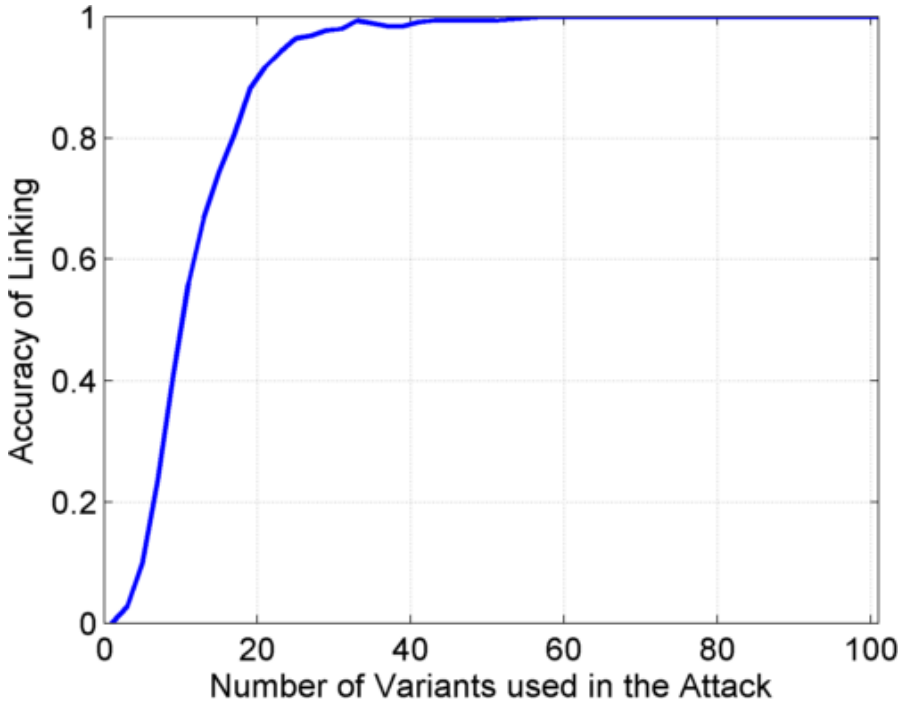
[Harmanci et al. (submitted)]



[Harmanci et al. (submitted)]

Predictability vs. Information Leakage & Accuracy of Linking





[Harmanci et al. (submitted)]

Genomic Privacy & the 2-sided Nature of RNA-seq: Quantifying the Leakage of Individual Information

- Intro on RNA-seq & the General **Dilemma of Genomic Privacy**
  - RNA-seq Presents a tricky privacy issue since much of the sequencing is for general, non-individual specific results yet it's tagged with individual information
  - The need to quantify leaks
- **Quantifying RNA-seq Leakage …from Reads**
  - Almost as much as WGS
  - But can remove SNVs in reads w/ MRF
- **…from eQTLs**
  - Quantifying & removing variant info from expression levels + eQTLs using ICI & predictability
  - Instantiating a practical linking attack
- **…from Indels/SVs**
  - Another source of leakage in RNA-seq data & how to use these for a related linking attack

Genomic Privacy & the 2-sided Nature of RNA-seq: Quantifying the Leakage of Individual Information

- **Intro on RNA-seq & the General Dilemma of Genomic Privacy**
  - RNA-seq Presents a tricky privacy issue since much of the sequencing is for general, non-individual specific results yet it's tagged with individual information
  - The need to quantify leaks
- **Quantifying RNA-seq Leakage …from Reads**
  - Almost as much as WGS
  - But can remove SNVs in reads w/ MRF
- **…from eQTLs**
  - Quantifying & removing variant info from expression levels + eQTLs using ICI & predictability
  - Instantiating a practical linking attack
- **…from Indels/SVs**
  - Another source of leakage in RNA-seq data & how to use these for a related linking attack

Acknowledgements

NIH BD2K Program

PrivaSeq.gersteinlab.org –
A **Harmanci**, G **Gürsoy**,
F Navarro, S Wagner, X Kong

Hiring Postdocs. See JOBS.gersteinlab.org !

Extra

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