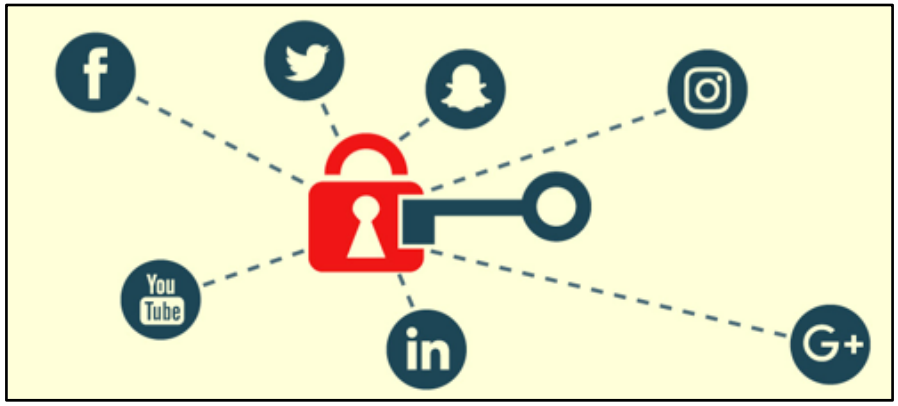
M Gerstein
Yale
(See last slide for more info.)

Slides freely "tweetable" (via **@MarkGerstein**) & downloadable from **Lectures.GersteinLab.org**

Privacy: Does Genomics has similar "Big Data" Dilemma as in the Rest of Society?



- We confront privacy risks every day we access the internet (e.g., social media, e-commerce).
- Sharing & "peer-production" is central to success of many new ventures, with analogous risks to genomics
  - **EG web search**: Large-scale mining essential

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?

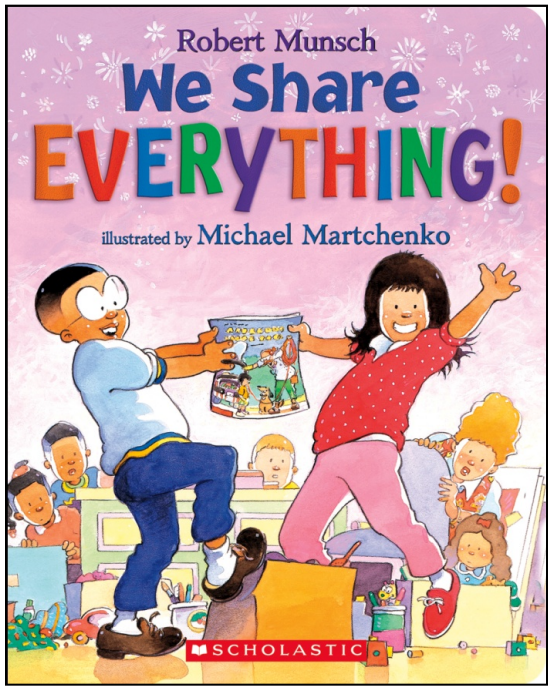
Once put on the web it can't be taken back

Ethically challenged history of genetics

Ownership of the data & what consent means (Hela)

Could your genetic data give rise to a product line?

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

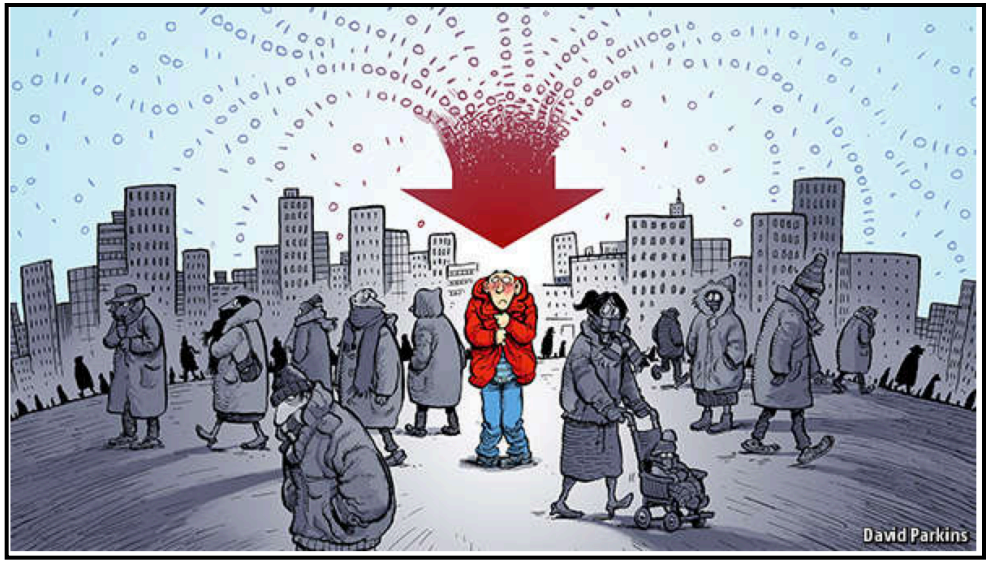


The Dilemma

- The individual (harmed?) v the collective (benefits)
  - But do sick patients care about their privacy?
- How to balance risks v rewards – Quantification

The Other Side of the Coin for Genomics: Why we should share

- Sharing helps **speed research**
  - Large-scale mining of this information is important for medical research
  - Statistical power
  - Privacy is cumbersome, particularly for big data



[Economist, 15 Aug '15]

[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)]

Current Social & Technical Solutions: The quandary where are now

- **Closed Data** Approach
  - Consents
  - "Protected" distribution via dbGAP
  - Local computes on secure computer
- Issues with Closed Data
  - Non-uniformity of consents & paperwork
    - Different, confusing int'l norms
  - Computer security is burdensome
  - Many schemes get "hacked" .
  - **Tricky aspects of high-dimensional data** (leakage & ease of creating quasi-identifiers)
- **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; Photo: futurism.com]

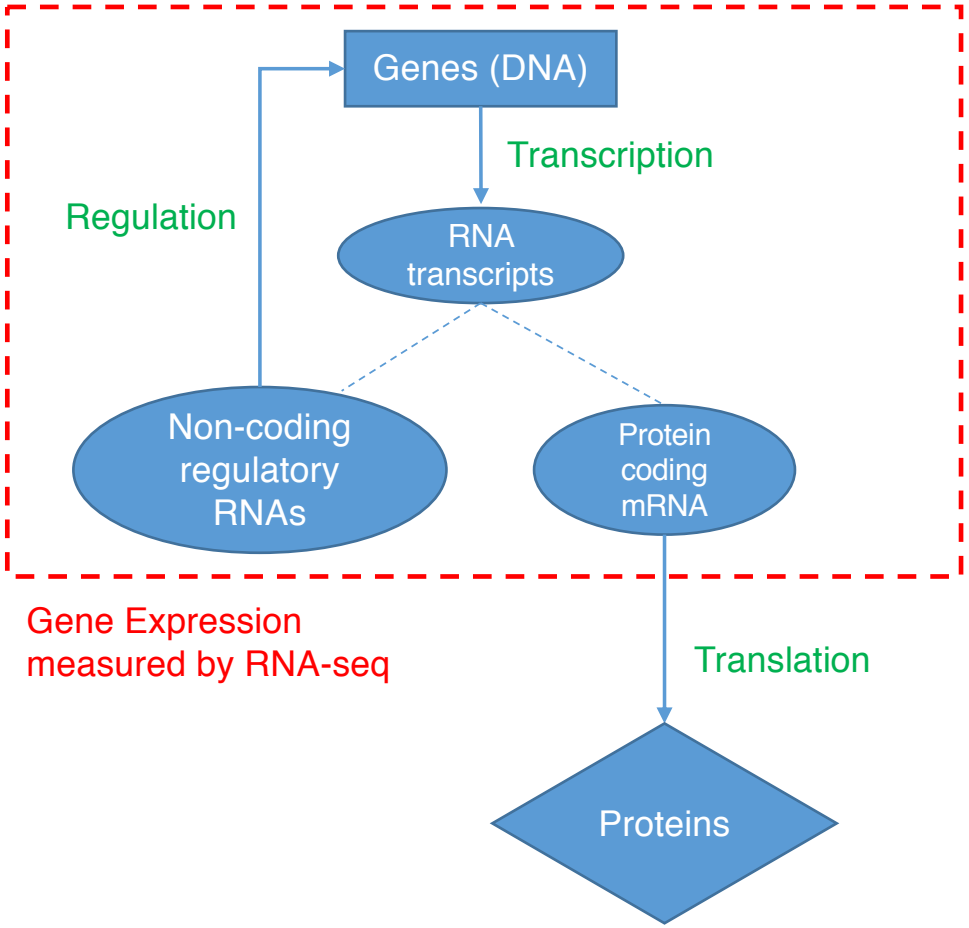
Quantification of sensitive information leakage from functional genomics data: Obvious v subtle leakages & practical file formats for addressing this

- Intro. to Genomic Privacy
  - The **dilemma**: The genome as fundamental, inherited info that's very private v. need for large-scale sharing & mining for med. research
- Privacy & Functional Genomics Data
  - **2-sided nature** of this data presents particularly tricky privacy issues
  - Overview of **types of the leakage**, from obvious to subtle
- Subtle Leakage #1: **eQTLs**
  - Quantifying & removing further variant info from expression levels w/ ICI & predictability.
  - Instantiating a practical linking attack w/ noisy quasi-identifiers
- Subtle Leakage #2: **Signal Profiles**
  - Manifest appreciable leakage from large & small deletions.
  - Linking attacks possible but additional complication of SV discovery in addition to genotyping
- Practical solutions & file formats
  - Using **pBAM** file format to remove obvious large-scale leakage
  - Small subtle leaks combatted by restricting large-scale access. Hence, developing secure, **blockchain-based logging** technology (response to the iDash challenge)

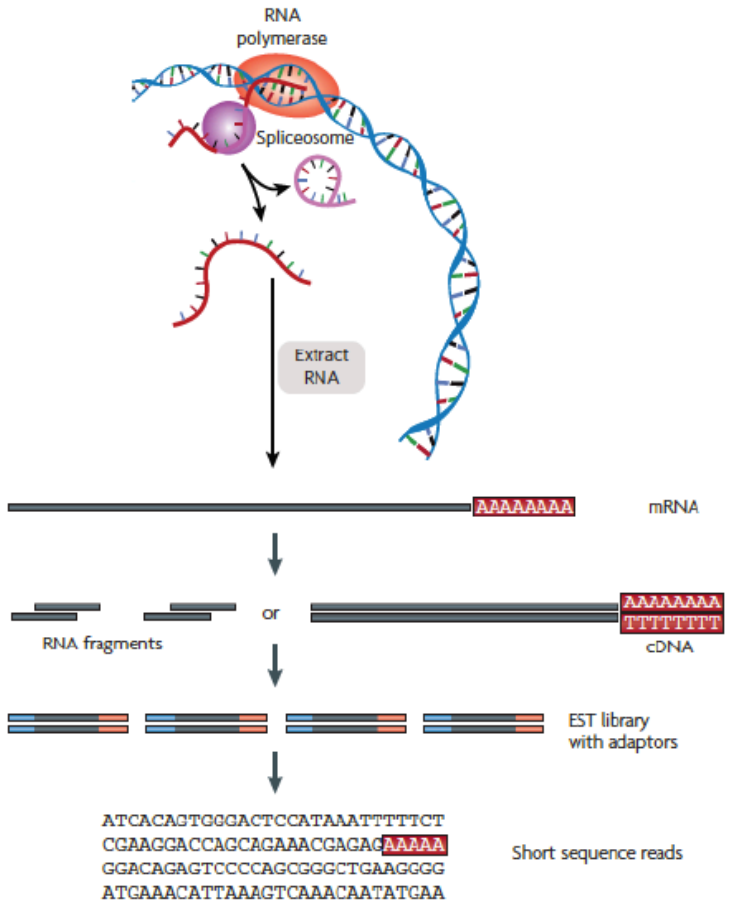
Quantification of sensitive information leakage from functional genomics data: Obvious v subtle leakages & practical file formats for addressing this

- Intro. to Genomic Privacy
  - The **dilemma**: The genome as fundamental, inherited info that's very private v. need for large-scale sharing & mining for med. research
- Privacy & Functional Genomics Data
  - **2-sided nature** of this data presents particularly tricky privacy issues
  - Overview of **types of the leakage**, from obvious to subtle
- Subtle Leakage #1: **eQTLs**
  - Quantifying & removing further variant info from expression levels w/ ICI & predictability.
  - Instantiating a practical linking attack w/ noisy quasi-identifiers
- Subtle Leakage #2: **Signal Profiles**
  - Manifest appreciable leakage from large & small deletions.
  - Linking attacks possible but additional complication of SV discovery in addition to genotyping
- Practical solutions & file formats
  - Using **pBAM** file format to remove obvious large-scale leakage
  - Small subtle leaks combatted by restricting large-scale access. Hence, developing secure, **blockchain-based logging** technology (response to the iDash challenge)

Transcriptome = Gene Activity of All Genes in the Genome, usually quantified by RNA-seq (RNA-seq is the most common type of functional genomics data)



Gene Expression measured by RNA-seq



[*NATURE 459: 927; NAT. REV. GEN. 10: 57*]

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

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)

Data Reduction in RNA-Seq: an Overview

Successive steps of Data Reduction

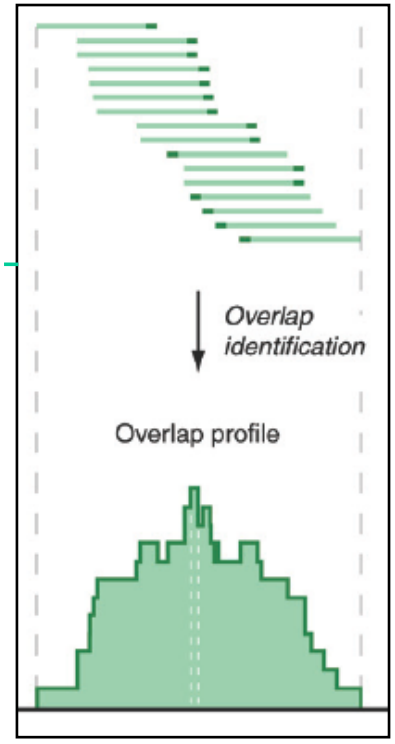
Fastq sequence files
~5-10 GB

```
ATACAAGCAAGTATAAGTTCGTATGCCGTCTT
GGAGGCTGGAGTTGGGGACGTATGCGGCATAG
TACCGATCGAGTCGACTGTAAACGTAGGCATA
ATTCTGACTGGTGTCATGCTGATGTACTTAAA
```

Index-building + Alignment to reference genome

BAM files
~1-2-fold reduction

Conversion to signal track by overlapping reads

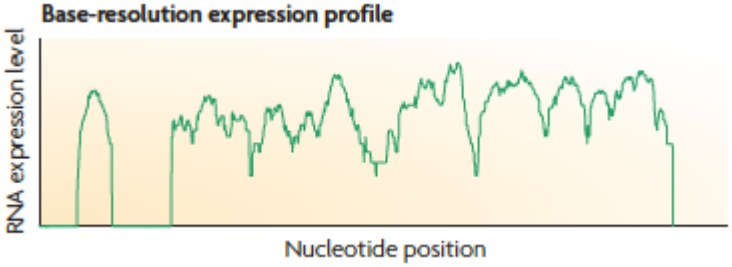


Reads => Signal

BigWig files
~25-fold reduction

Mapping to genes

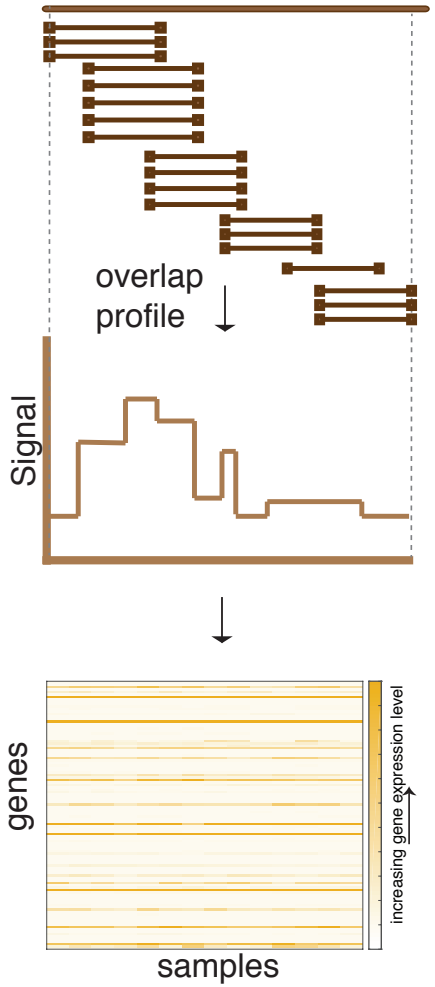
Gene/Transcript expression matrix
~20-fold reduction



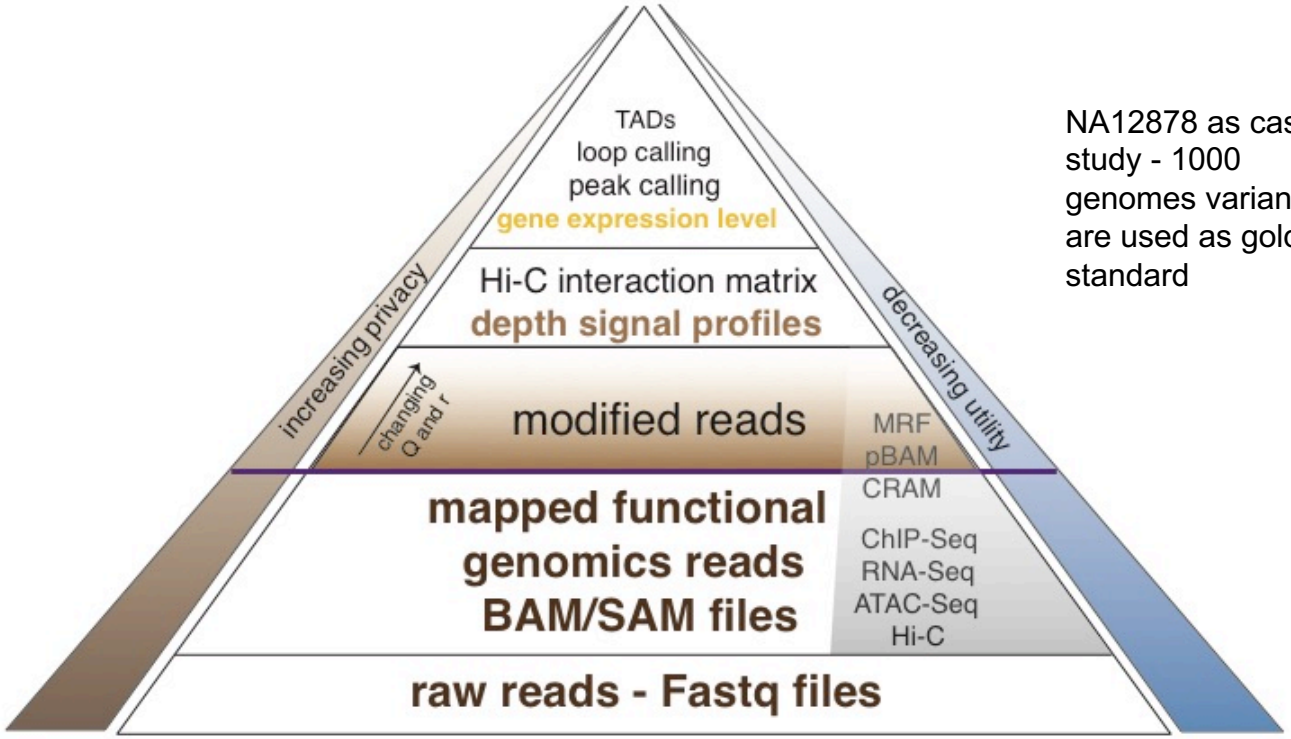
Quantitative information from RNA-seq signal:
average signals at exon level (RPKMs)

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

Functional genomics data comes with a great deal of sequencing; We can quantify amount of leakage at every step of the data summarization process.



[*Gursoy et al,* bioRxiv 345074]

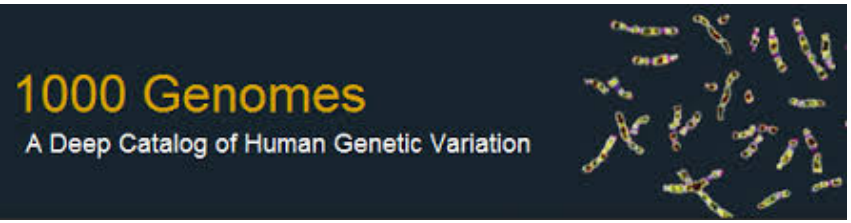


NA12878 as case study - 1000 genomes variants are used as gold standard

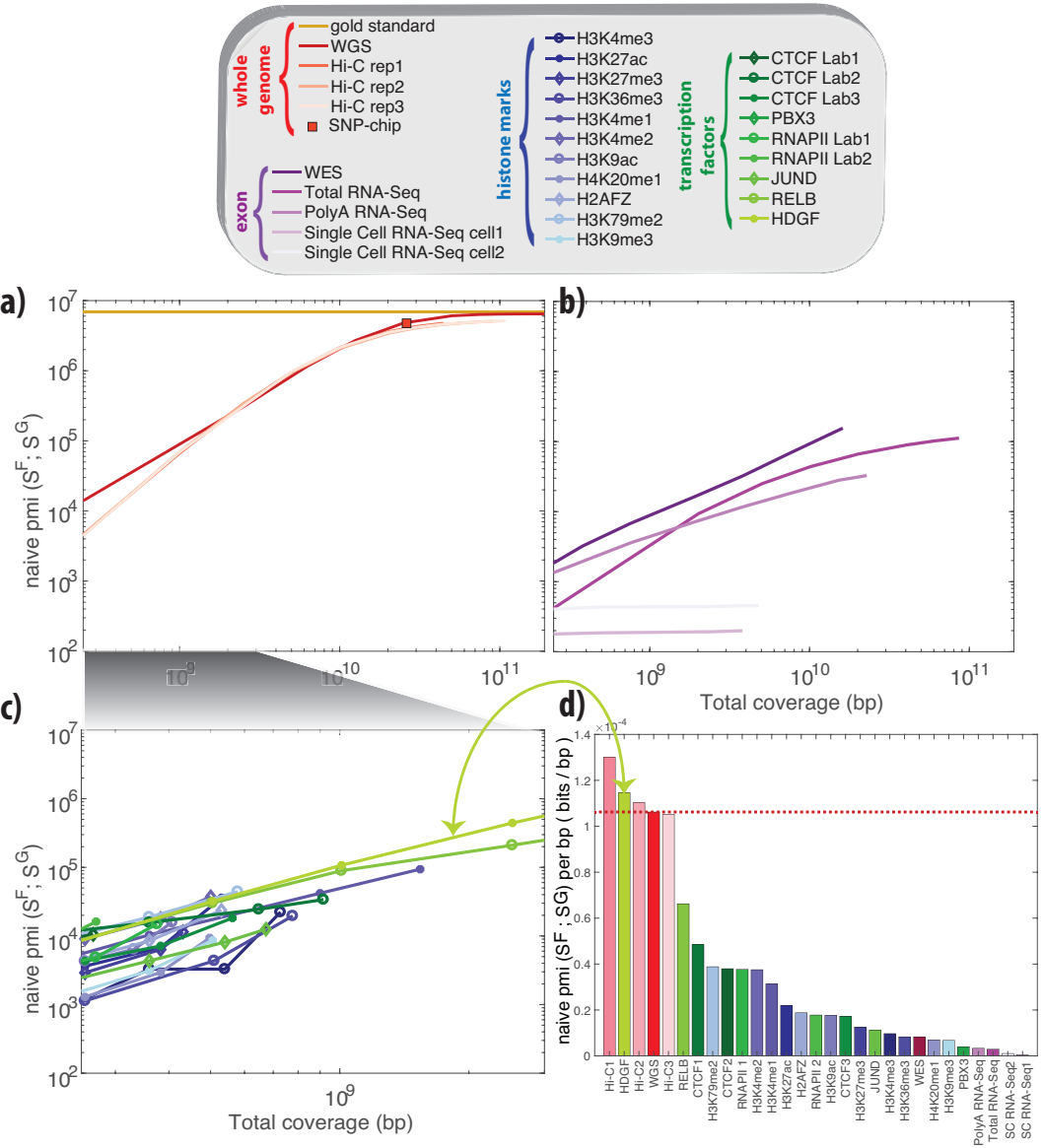
Leakage Source	Leaking Variants	# of potential variants	Average leakage per variant (bits)	Maximum leakage per variant (bits)	# of accessible variants	Total leakage (bits)
Raw reads	Exonic variants	2,682,417	0.10 ± 0.28	9.88 ± 2.12	246,893	24,689
Modified reads Q = {indels}	Exonic SNVs	2,607,969	0.09 ± 0.27	9.95 ± 2.02	231,031	207,92
Modified reads Q = {mismatches}	Exonic indels	51,408	0.33 ± 0.47	7.64 ± 2.42	15,862	5234
Signal profiles	Exonic deletions	48,019	0.29 ± 0.45	7.97 ± 2.42	1,067	298
Gene expression quantification	eQTLs	3,175	1.19 ± 0.36	4.00 ± 1.92	158	188

Representative Functional Genomics, Genotype, eQTL Datasets

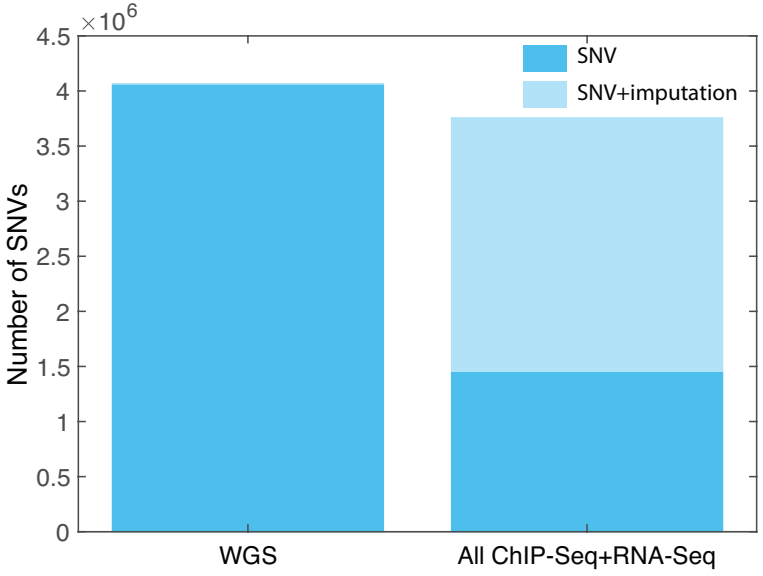
- Genotypes are available from the 1000 Genomes Project
- mRNA sequencing for 462 individuals from gEUVADIS and ENCODE
  - Publicly available quantification for protein coding genes
- Functional genomics data (ChIP-Seq, RNA-Seq, Hi-C) available from ENCODE
- Approximately 3,000 cis-eQTL (FDR<0.05)



- **How much information, for example, do RNA-Seq reads (or ChIP-Seq) reads contain? Does that information enough to identify individuals?**



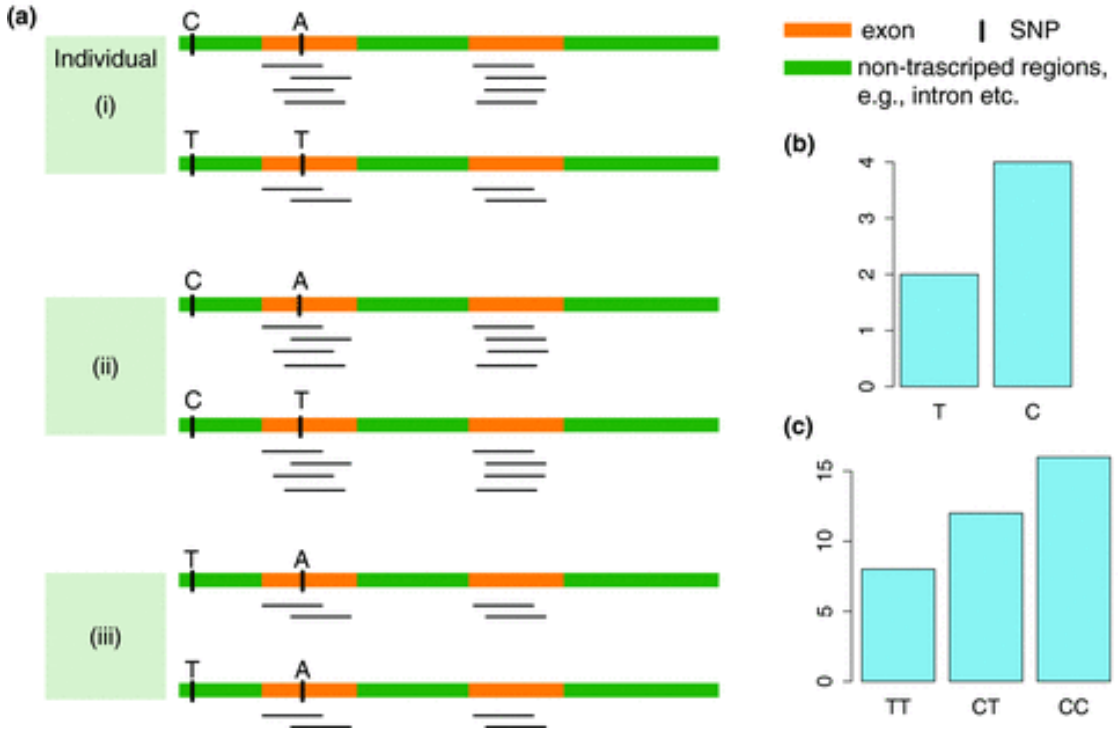
- **It might seem like we don't infer much information from single ChIP-Seq and RNA-Seq experiments compared to WGS**
  - **However putting 10 different ChIP-Seq experiments and RNA-Seq together with imputation provides a great deal of information about the individual**



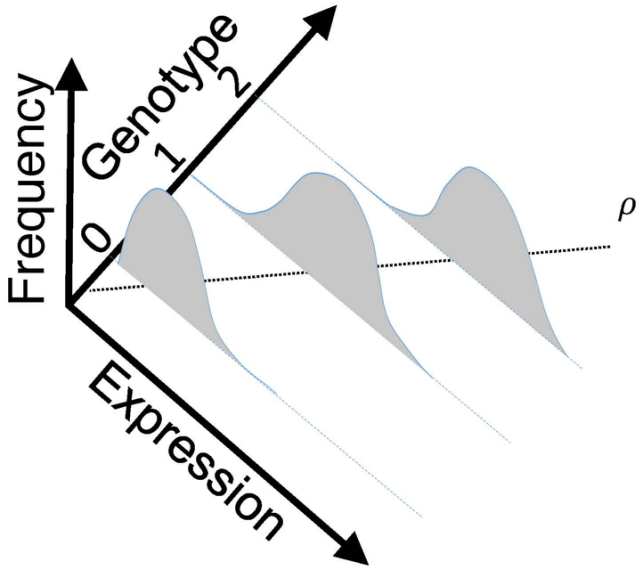
[*Gursoy et al*, bioRxiv 345074]

Quantification of sensitive information leakage from functional genomics data: Obvious v subtle leakages & practical file formats for addressing this

- Intro. to Genomic Privacy
  - The **dilemma**: The genome as fundamental, inherited info that's very private v. need for large-scale sharing & mining for med. research
- Privacy & Functional Genomics Data
  - **2-sided nature** of this data presents particularly tricky privacy issues
  - Overview of **types of the leakage**, from obvious to subtle
- Subtle Leakage #1: **eQTLs**
  - Quantifying & removing further variant info from expression levels w/ ICI & predictability.
  - Instantiating a practical linking attack w/ noisy quasi-identifiers
- Subtle Leakage #2: **Signal Profiles**
  - Manifest appreciable leakage from large & small deletions.
  - Linking attacks possible but additional complication of SV discovery in addition to genotyping
- Practical solutions & file formats
  - Using **pBAM** file format to remove obvious large-scale leakage
  - Small subtle leaks combatted by restricting large-scale access. Hence, developing secure, **blockchain-based logging** technology (response to the iDash challenge)



[*Biometrics 68(1) 1–11*]

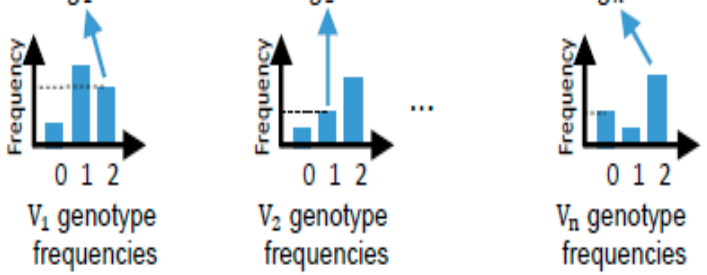


eQTL Mapping Using RNA-Seq Data

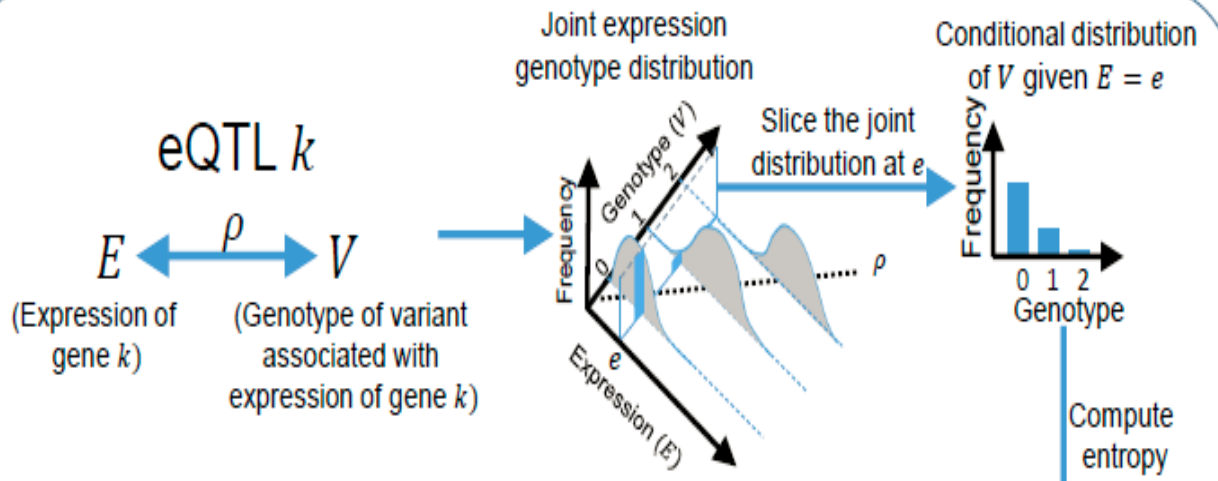
- eQTLs are genomic loci that contribute to variation in mRNA expression levels
- eQTLs provide insights on transcription regulation, and the molecular basis of phenotypic outcomes
- eQTL mapping can be done with RNA-Seq data

Information Content and Predictability

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



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

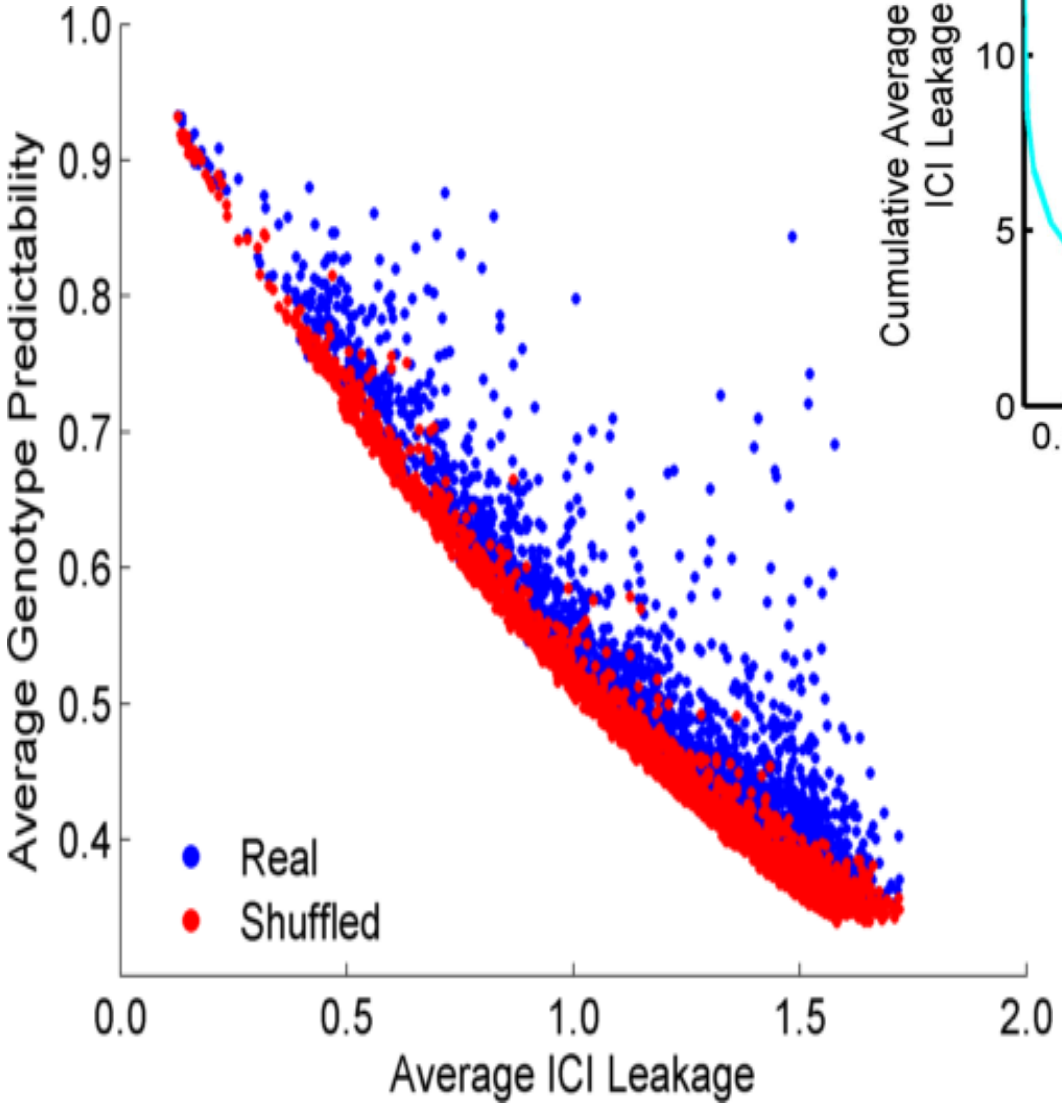


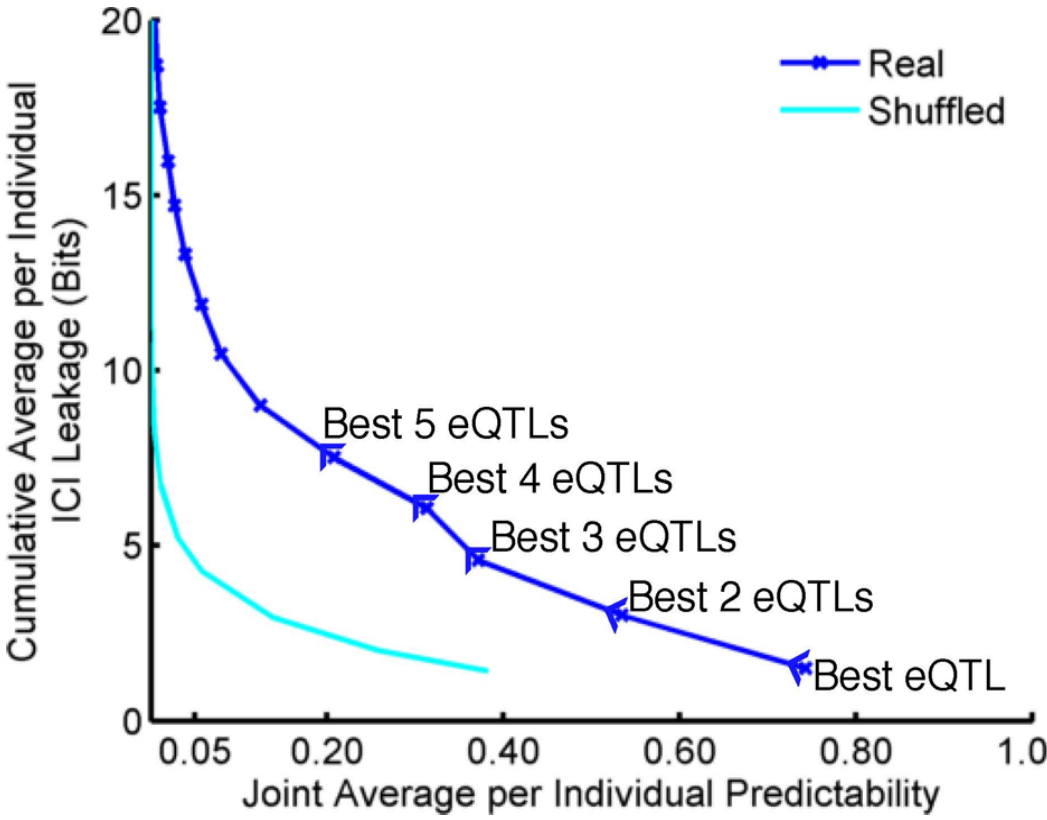
$$\pi(V|E=e) = e^{-H(V|E=e)}$$

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

[Harmanci et al. Nat. Meth. 2016]

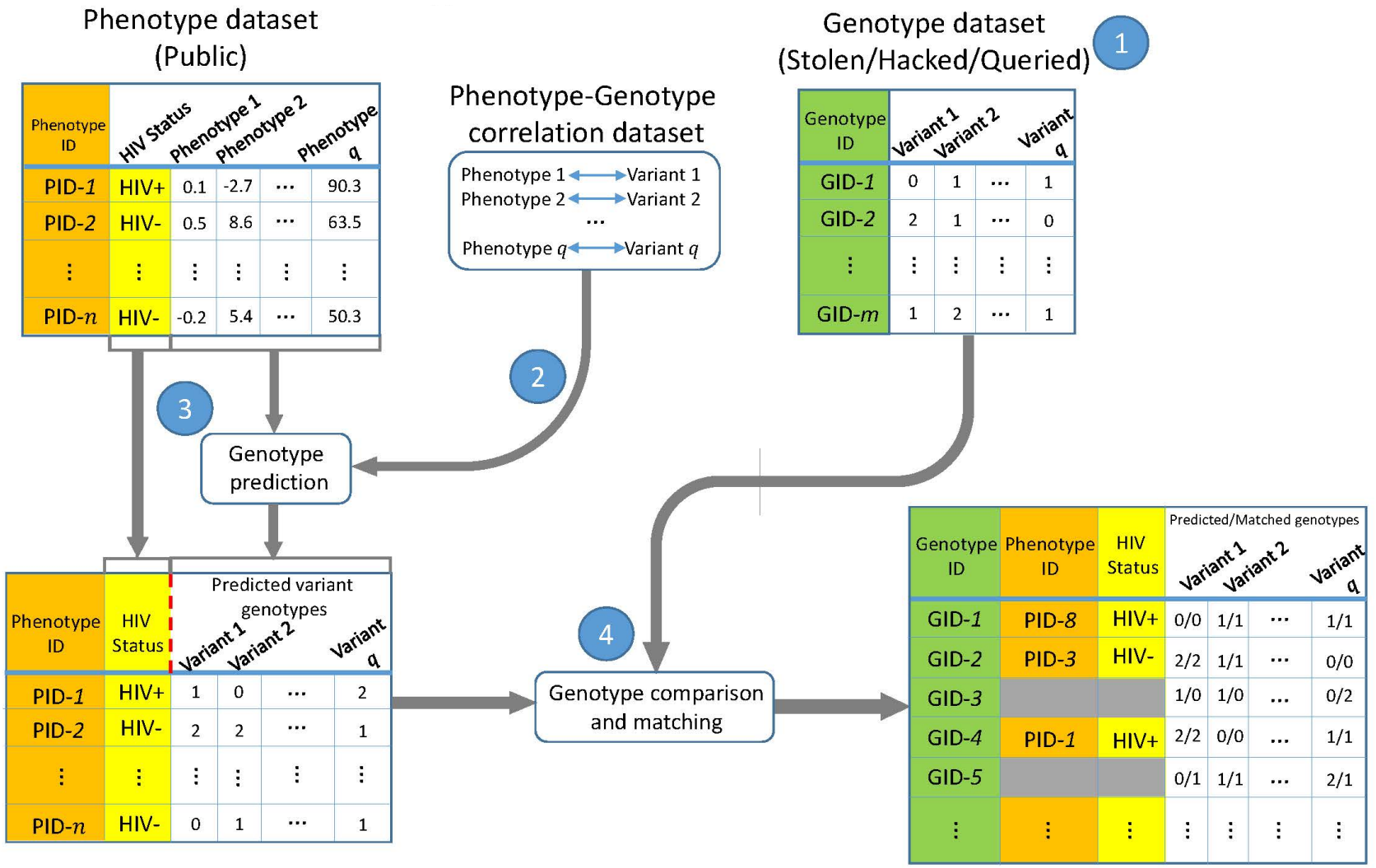
[Harmanciet al. Nat. Meth. ('16]





ICI Leakage versus Genotype Predictability

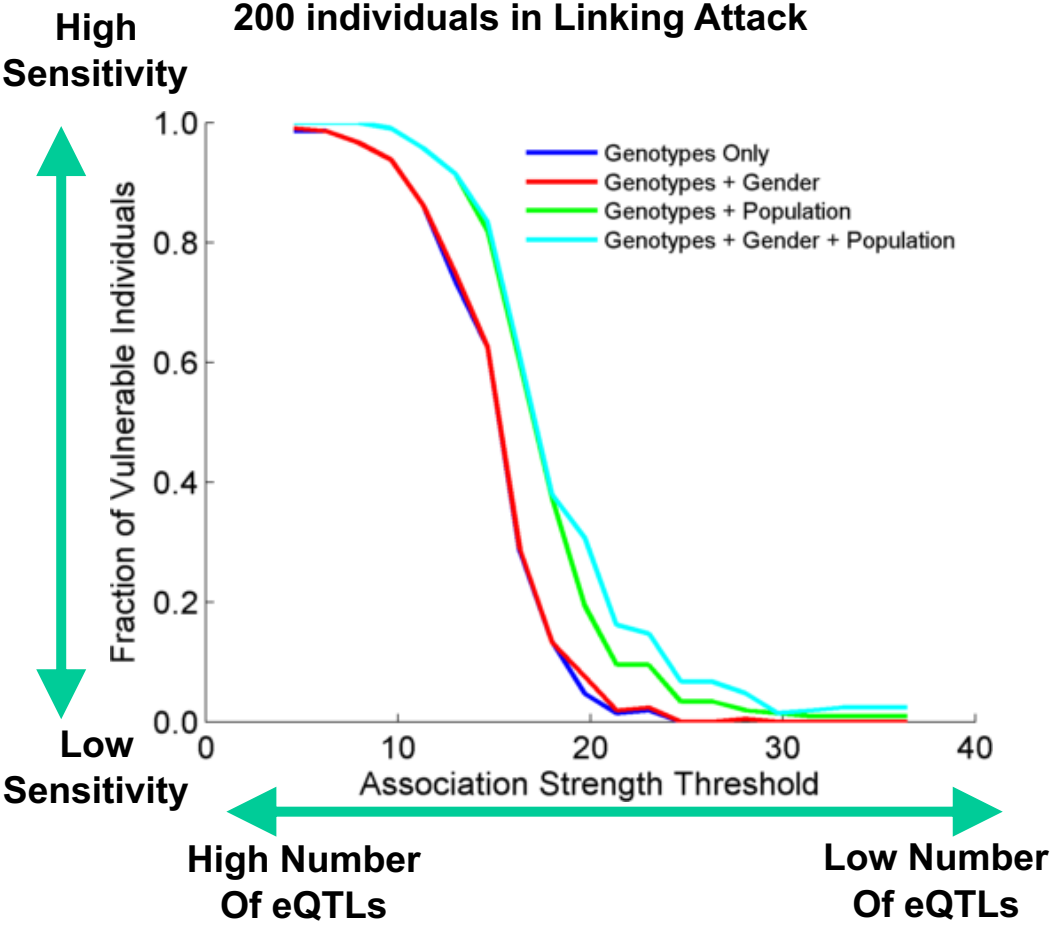
Linking Attack Scenario



[Harmanciet al. Nat. Meth. (‘16)]

Success in Linking Attack with Extremity based Genotype Prediction

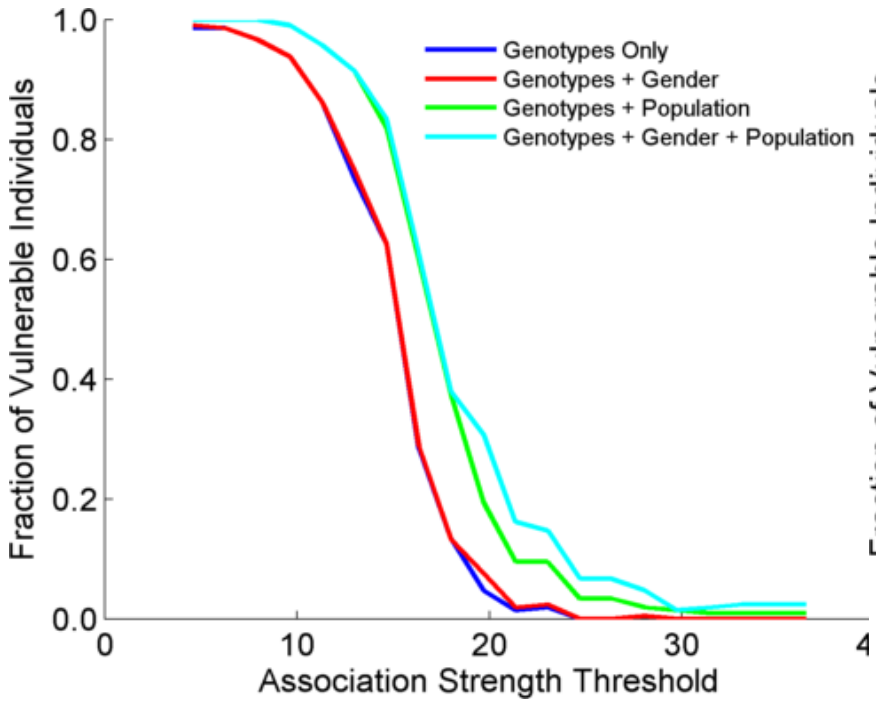
200 individuals eQTL Discovery
200 individuals in Linking Attack



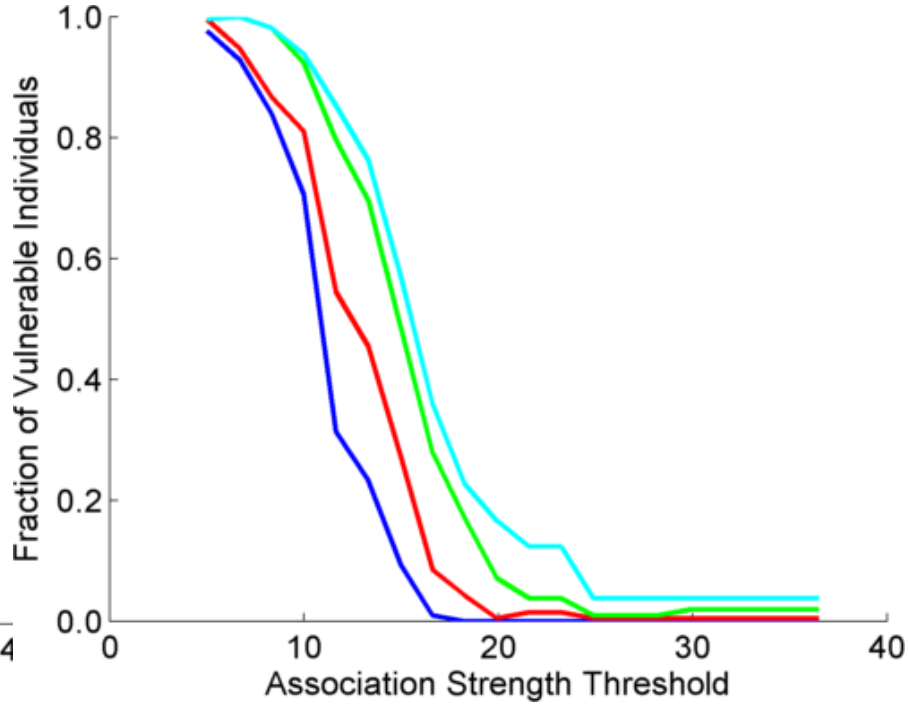
[Harmanci et al. *Nat. Meth.* (16)]

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

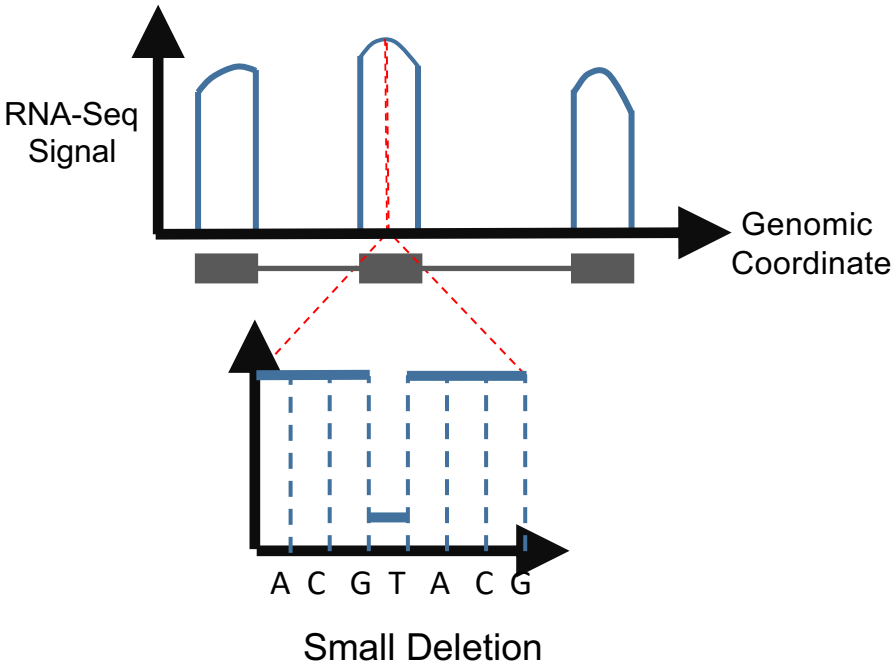


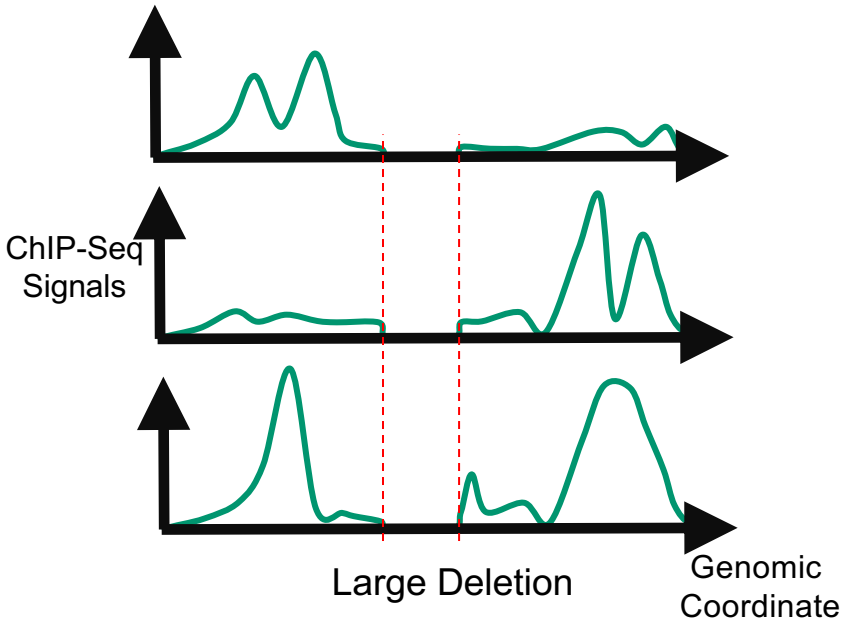
[Harmanci et al. *Nat. Meth.* (16)]

Quantification of sensitive information leakage from functional genomics data: Obvious v subtle leakages & practical file formats for addressing this

- Intro. to Genomic Privacy
  - The **dilemma**: The genome as fundamental, inherited info that's very private v. need for large-scale sharing & mining for med. research
- Privacy & Functional Genomics Data
  - **2-sided nature** of this data presents particularly tricky privacy issues
  - Overview of **types of the leakage**, from obvious to subtle
- Subtle Leakage #1: **eQTLs**
  - Quantifying & removing further variant info from expression levels w/ ICI & predictability.
  - Instantiating a practical linking attack w/ noisy quasi-identifiers
- Subtle Leakage #2: **Signal Profiles**
  - Manifest appreciable leakage from large & small deletions.
  - Linking attacks possible but additional complication of SV discovery in addition to genotyping
- Practical solutions & file formats
  - Using **pBAM** file format to remove obvious large-scale leakage
  - Small subtle leaks combatted by restricting large-scale access. Hence, developing secure, **blockchain-based logging** technology (response to the iDash challenge)

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

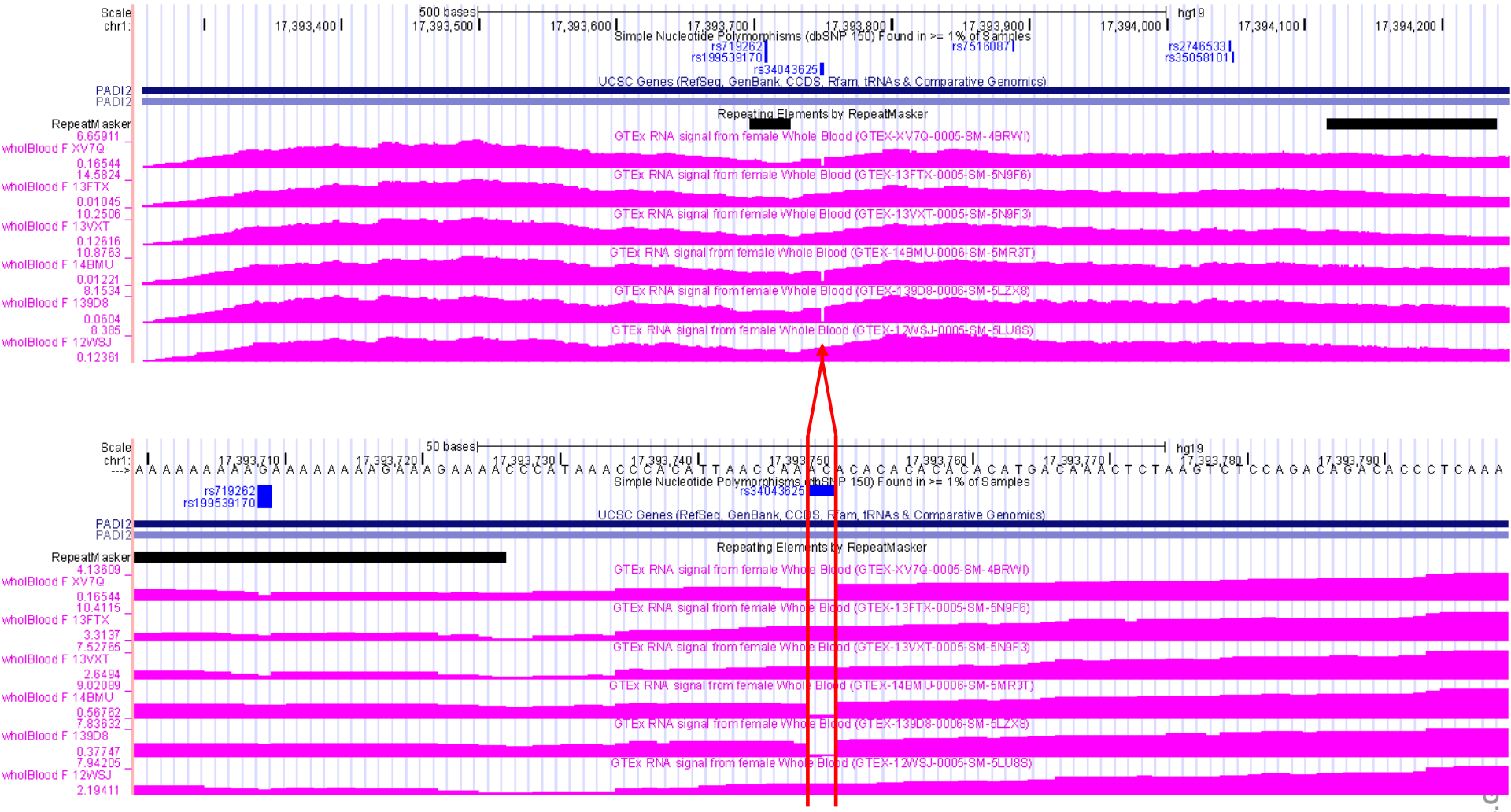




RNA-seq also shows large deletions

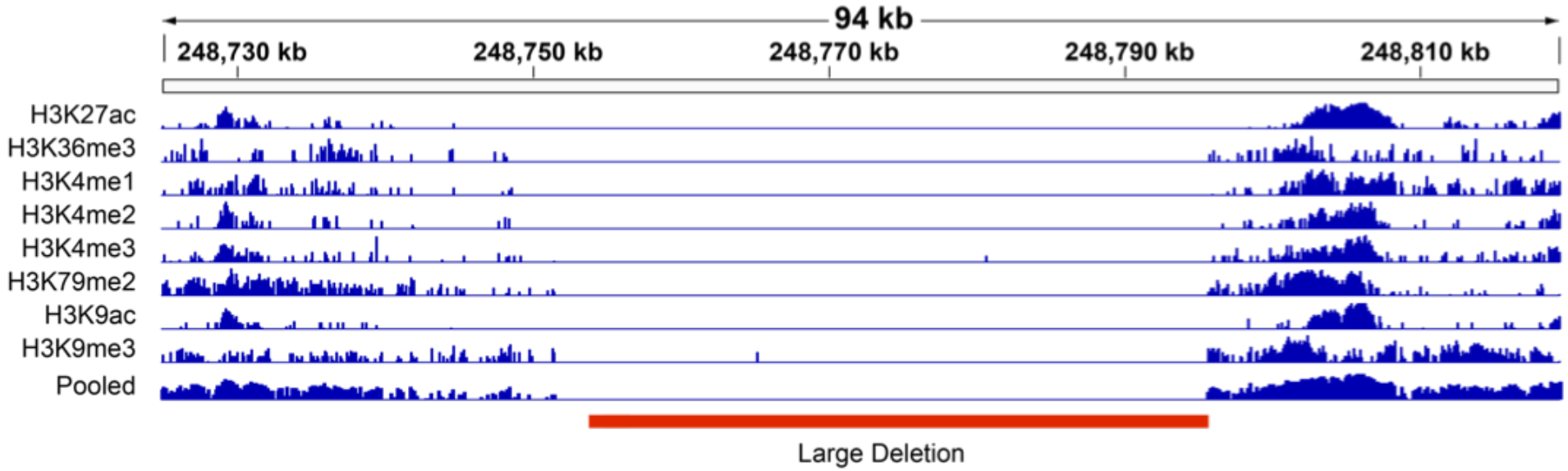
[Harmanci & Gerstein, *Nat. Comm.* ('18)]

Example of Small Deletion Evident in Signal Profile



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

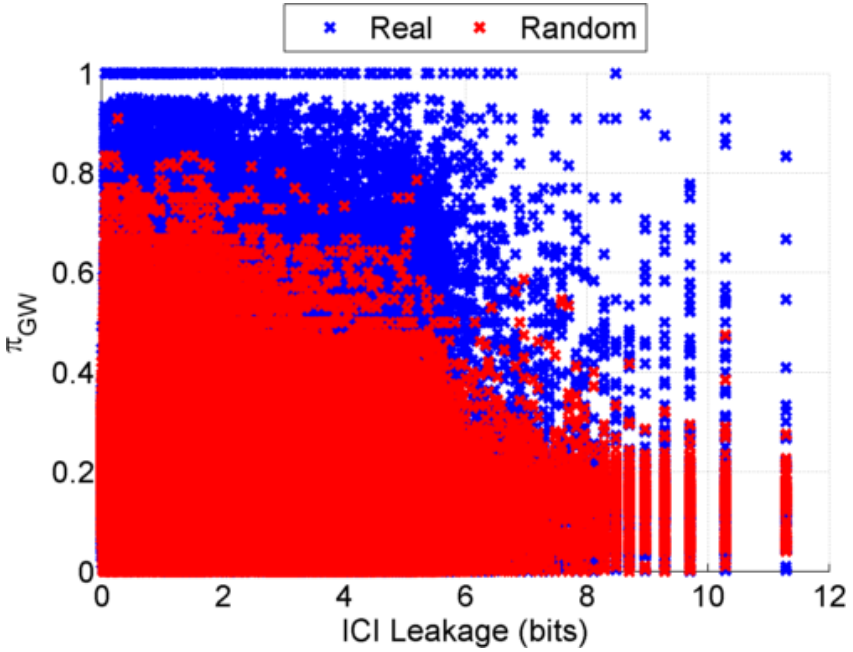
Example of Large Deletion Evident in Signal Profile



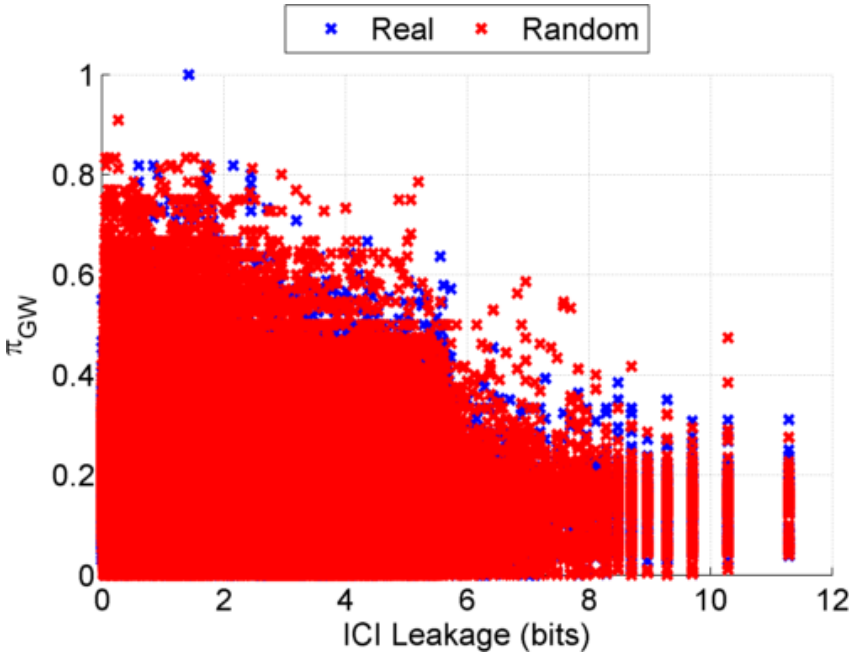
[Harmanci & Gerstein, *Nat. Comm.* ('18)]

Information Leakage from SV Deletions

a) Before Anonymization



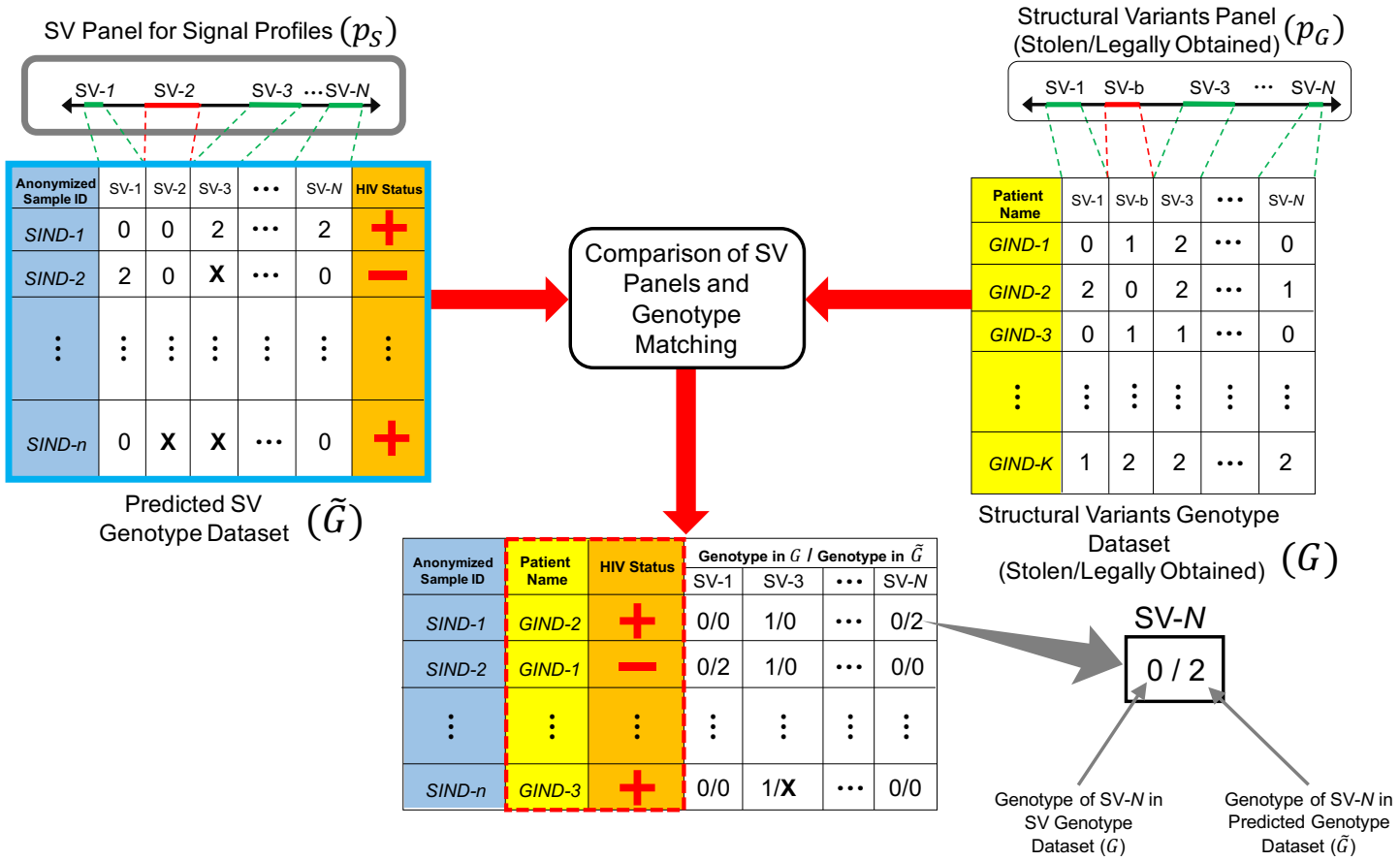
b) After Anonymization



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

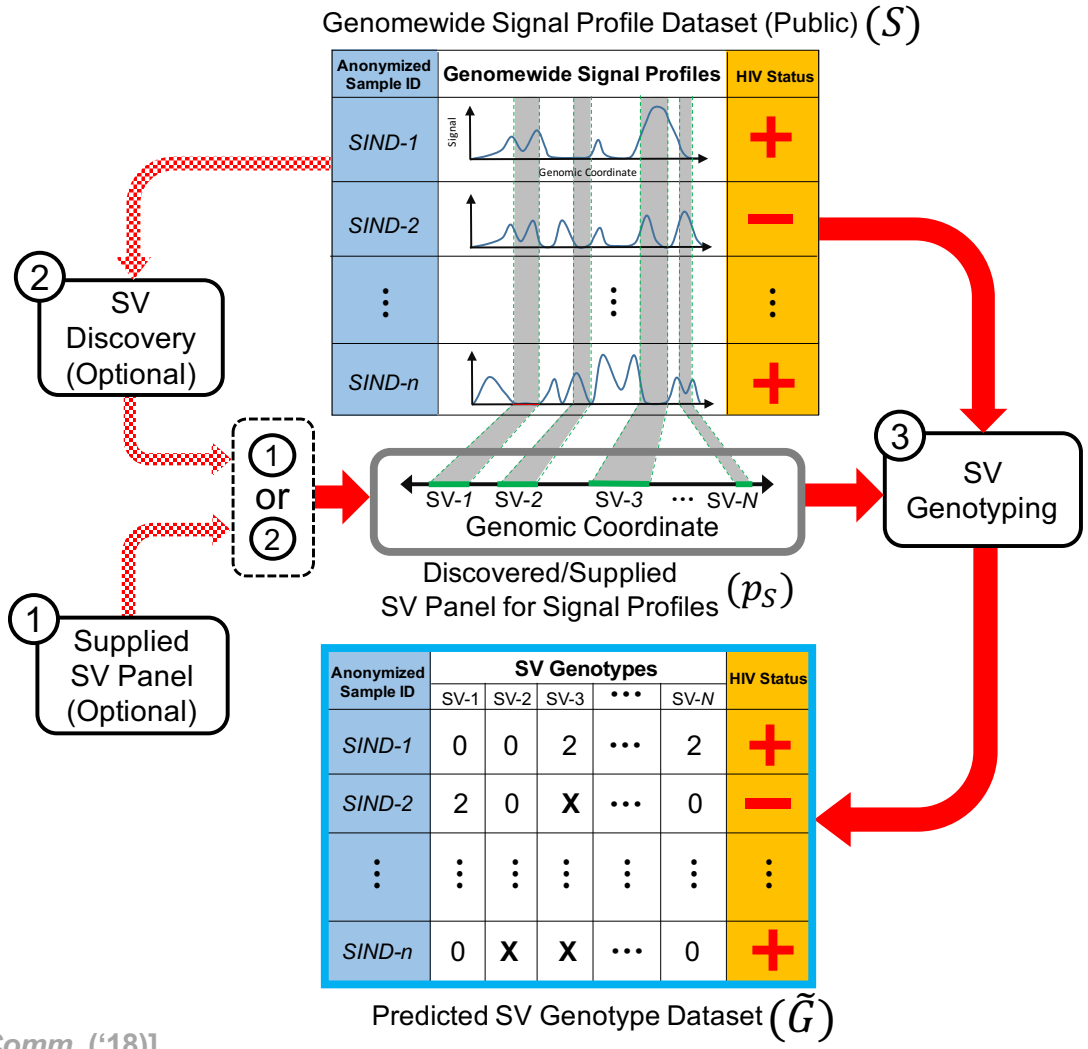
[Harmanci & Gerstein, *Nat. Comm.* ('18)]

Another type of Linking Attack: Linking based on SV Genotyping



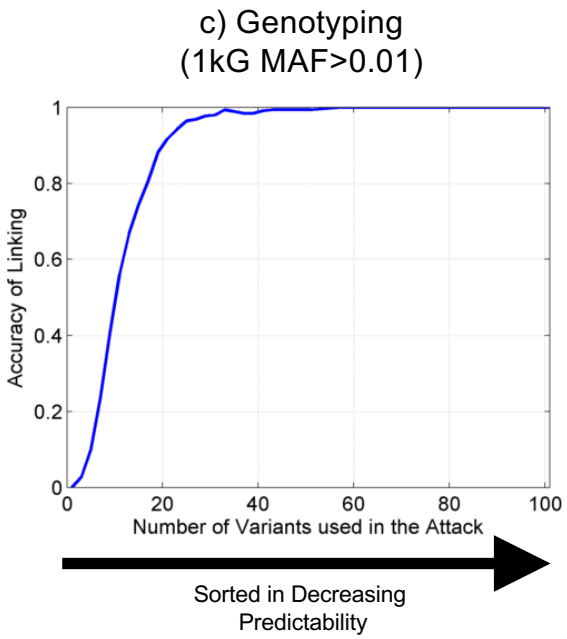
[Harmanci & Gerstein, *Nat. Comm.* ('18)]

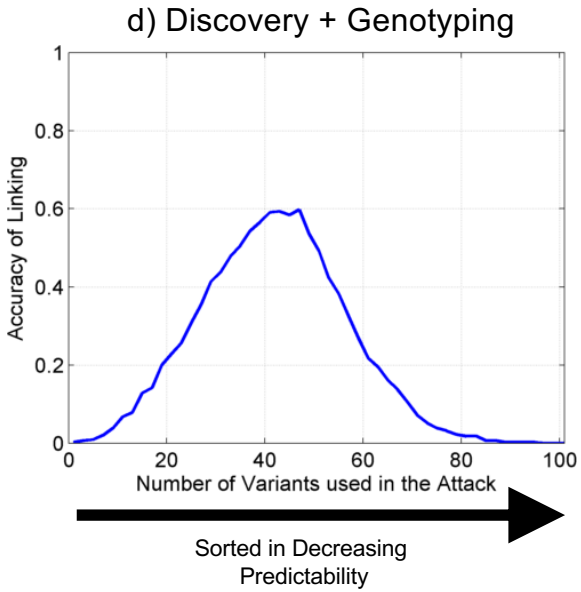
Another type of Linking Attack: First Doing SV Genotyping



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

Linking Attack Based on SV Deletions in gEUVADIS Dataset



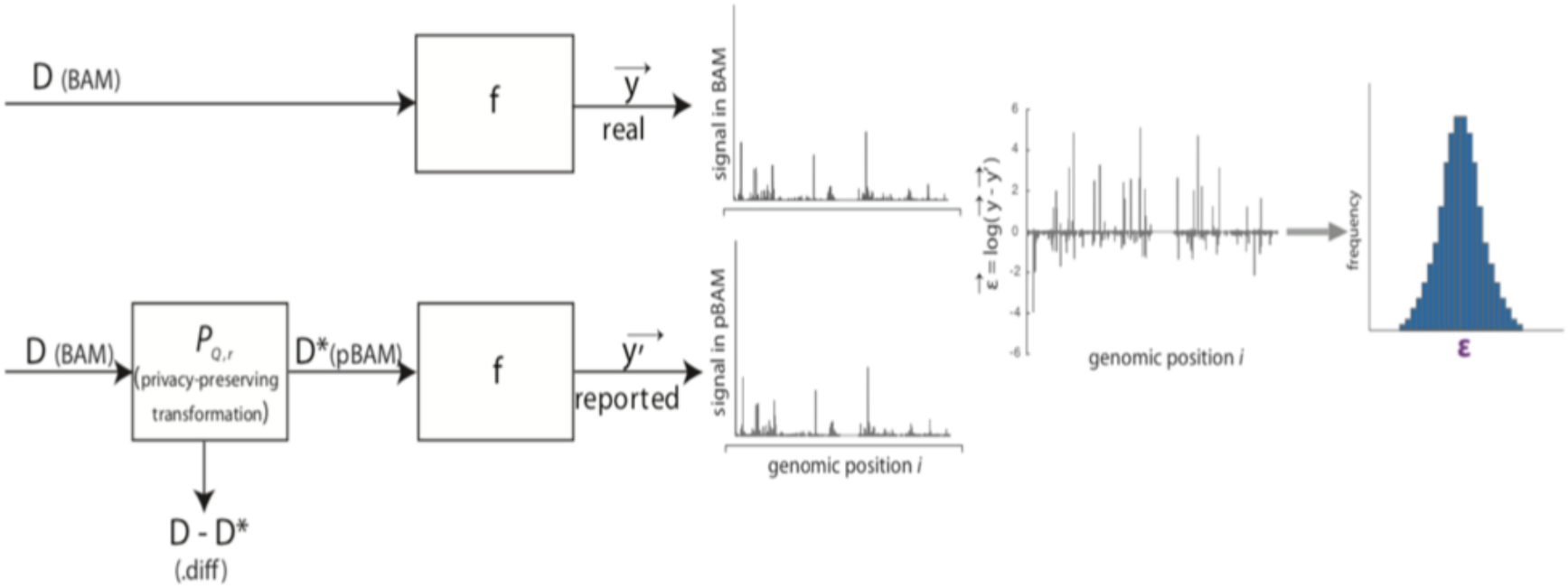


[Harmanci & Gerstein, *Nat. Comm.* ('18)]

Quantification of sensitive information leakage from functional genomics data: Obvious v subtle leakages & practical file formats for addressing this

- Intro. to Genomic Privacy
  - The **dilemma**: The genome as fundamental, inherited info that's very private v. need for large-scale sharing & mining for med. research
- Privacy & Functional Genomics Data
  - **2-sided nature** of this data presents particularly tricky privacy issues
  - Overview of **types of the leakage**, from obvious to subtle
- Subtle Leakage #1: **eQTLs**
  - Quantifying & removing further variant info from expression levels w/ ICI & predictability.
  - Instantiating a practical linking attack w/ noisy quasi-identifiers
- Subtle Leakage #2: **Signal Profiles**
  - Manifest appreciable leakage from large & small deletions.
  - Linking attacks possible but additional complication of SV discovery in addition to genotyping
- Practical solutions & file formats
  - Using **pBAM** file format to remove obvious large-scale leakage
  - Small subtle leaks combatted by restricting large-scale access. Hence, developing secure, **blockchain-based logging** technology (response to the iDash challenge)

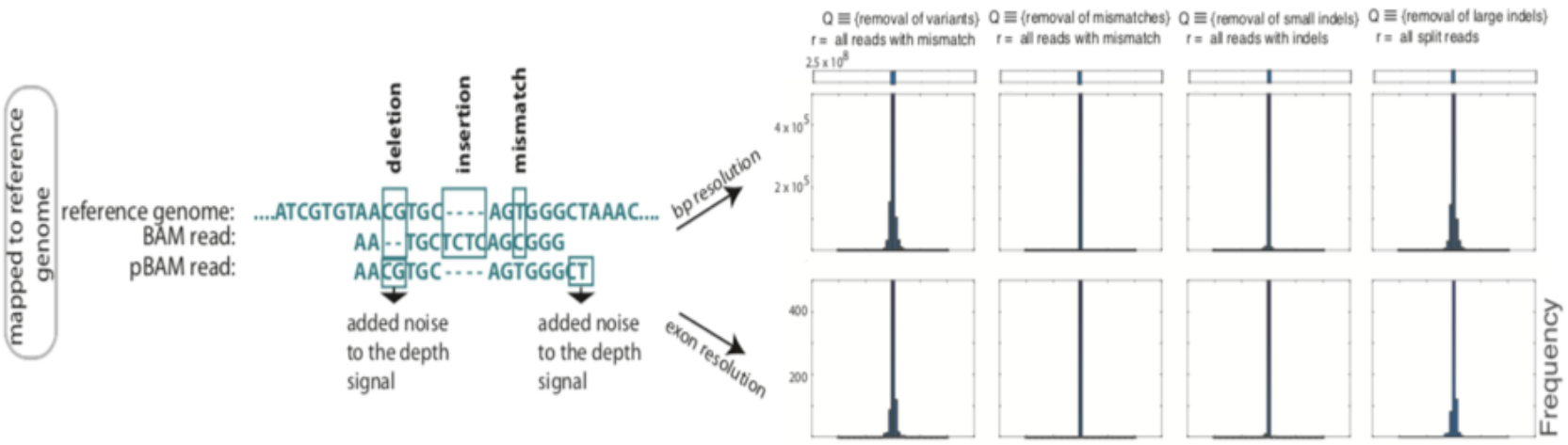
Privacy-aware Binary Alignment Mapping (pBAM)



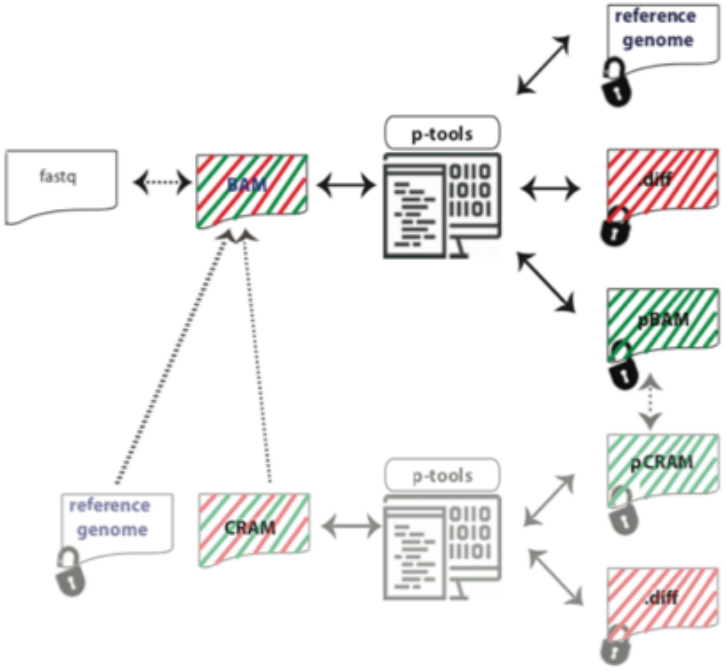
- No need to know the sequence of mapped reads to aggregate them
- A manipulation on Binary Alignment Files (BAM)
  - Find leaky fields/tags
  - Suppression
  - Generalization
- Goal:
  - Accurate gene/transcript expression quantification
  - Works with the pipelines / SAMtools

[Gursoy et al, BMC Med. Inform. (in press)]

pBAMs are high in utility and can be converted BAM



- **Works well with**
  - **STAR signal tracks,**
  - **RSEM gene expression and quantification**
  - **MACS2 for ChIP-Seq peak calling.**
- **The original BAM does not need be stored.**
  - **a smaller file called .diff**



[Gursoy et al, BMC Med. Inform. (in press)]

Quantification of sensitive information leakage from functional genomics data: Obvious v subtle leakages & practical file formats for addressing this

- Intro. to Genomic Privacy
  - The **dilemma**: The genome as fundamental, inherited info that's very private v. need for large-scale sharing & mining for med. research
- Privacy & Functional Genomics Data
  - **2-sided nature** of this data presents particularly tricky privacy issues
  - Overview of **types of the leakage**, from obvious to subtle
- Subtle Leakage #1: **eQTLs**
  - Quantifying & removing further variant info from expression levels w/ ICI & predictability.
  - Instantiating a practical linking attack w/ noisy quasi-identifiers
- Subtle Leakage #2: **Signal Profiles**
  - Manifest appreciable leakage from large & small deletions.
  - Linking attacks possible but additional complication of SV discovery in addition to genotyping
- Practical solutions & file formats
  - Using **pBAM** file format to remove obvious large-scale leakage
  - Small subtle leaks combatted by restricting large-scale access. Hence, developing secure, **blockchain-based logging** technology (response to the iDash challenge)

Total RNA Sequencing

Data Subtype	Cancer Types Applicable	Data Type Name	Level 1	Level 2	Level 3	Important Metadata
mRNA Sequencing sequence	Applicable to some tumor types	TotalRNASeqV2	mRNA sequence for each participant's tumor sample File type: binary alignment	n/a	n/a	Experimental protocol, including primer information, is contained in the metadata **.xml file** associated with each .bam file

Protected Data and Raw Data

Due to the nature of our donor consent agreement, raw data and attributes which might be used to identify the donors are not publicly available on the GTEx Portal. You may apply for access to the data through dbGaP.

- As one accesses the protected data, privacy leakage/gb increases
- To prevent more than certain amount of leakage → log user access
- Can we securely store and query these logs?

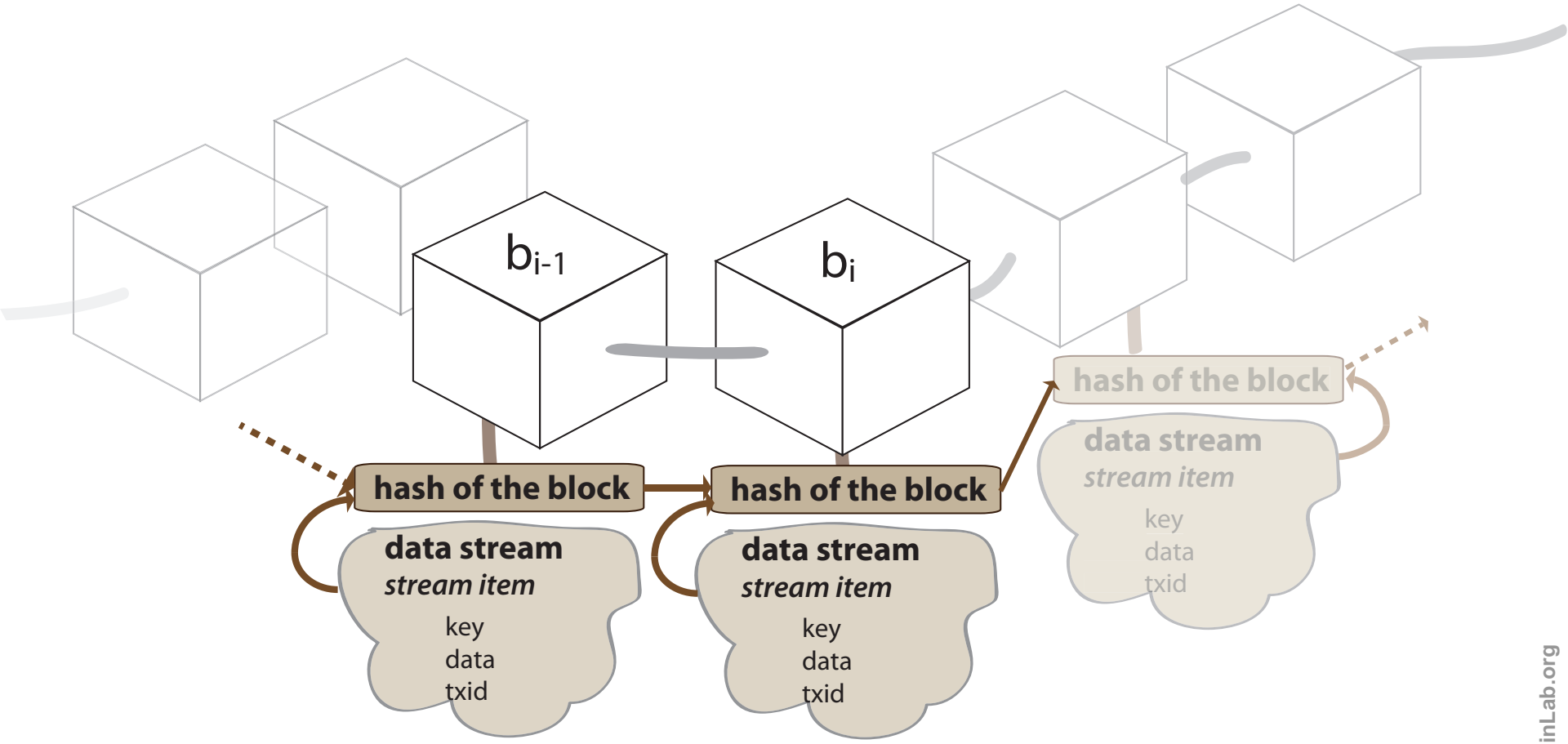
IDASH PRIVACY & SECURITY WORKSHOP 2018
- secure genome analysis competition

*NHGRI R13HG009072

Goal: Develop blockchain-based ledgering solutions to log and query the user activities of accessing genomic datasets across multiple sites

Find a way of storing it in the chain so you can access quickly!

Take advantage of data stream property of Multichain API



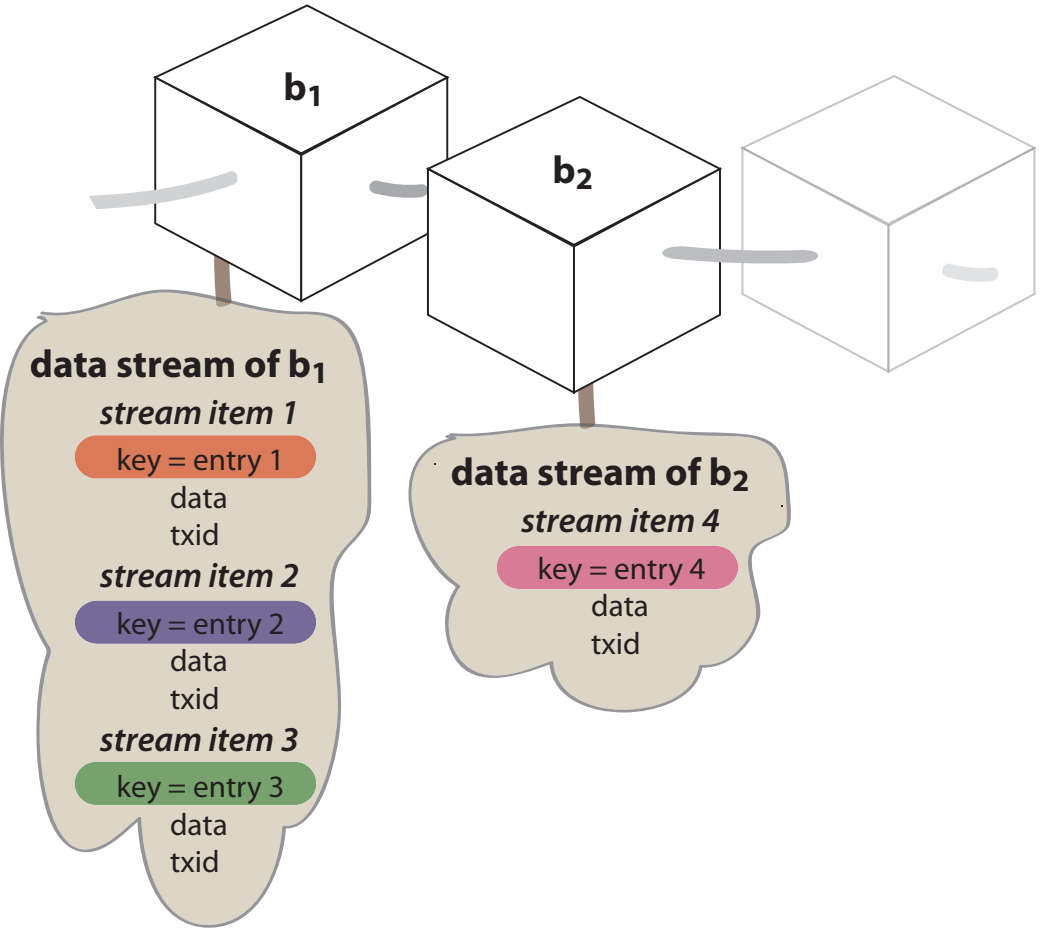
Every transaction appends a list of data items
key-value property

[Gursoy et al, BMC Med. Gen. (in press)]

Challenge Solution

	Timestamp	Node	ID	Ref ID	User	Activity	Resource
entry 1	1522000002801	1	1	1	1	REQ_RESOURCE	MOD_UCSC_Genome_Bioinformatics
entry 2	1522000008352	1	2	1	1	VIEW_RESOURCE	MOD_UCSC_Genome_Bioinformatics
entry 3	1522000016966	1	3	3	6	REQ_RESOURCE	MOD_FlyBase
entry 4	1522000019451	1	4	1	1	FILE_ACCESS	MOD_UCSC_Genome_Bioinformatics
						⋮	

LOG



- Save each entry as a key to a stream item
- For query
  - Download all the keys
  - Create a dataframe
  - Query locally on the dataframe

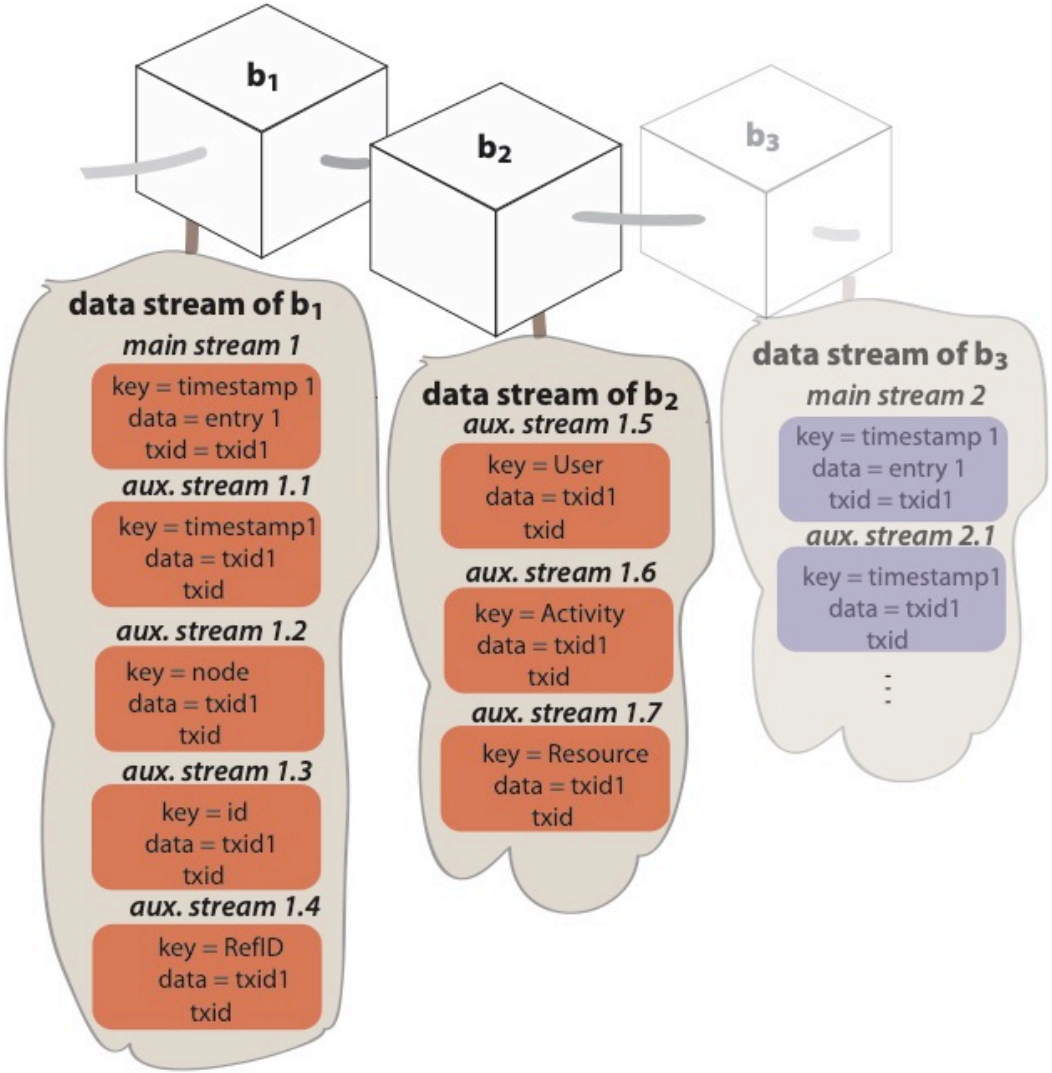
Quick but memory intensive for millions of entries

[Gursoy et al, BMC Med. Gen. (in press)]

Bigmem Solution

	Timestamp	Node	ID	Ref ID	User	Activity	Resource
entry 1	1522000002801	1	1	1	1	REQ_RESOURCE	MOD_UCSC_Genome_Bioinformatics
entry 2	1522000008352	1	2	1	1	VIEW_RESOURCE	MOD_UCSC_Genome_Bioinformatics
	⋮						

LOG



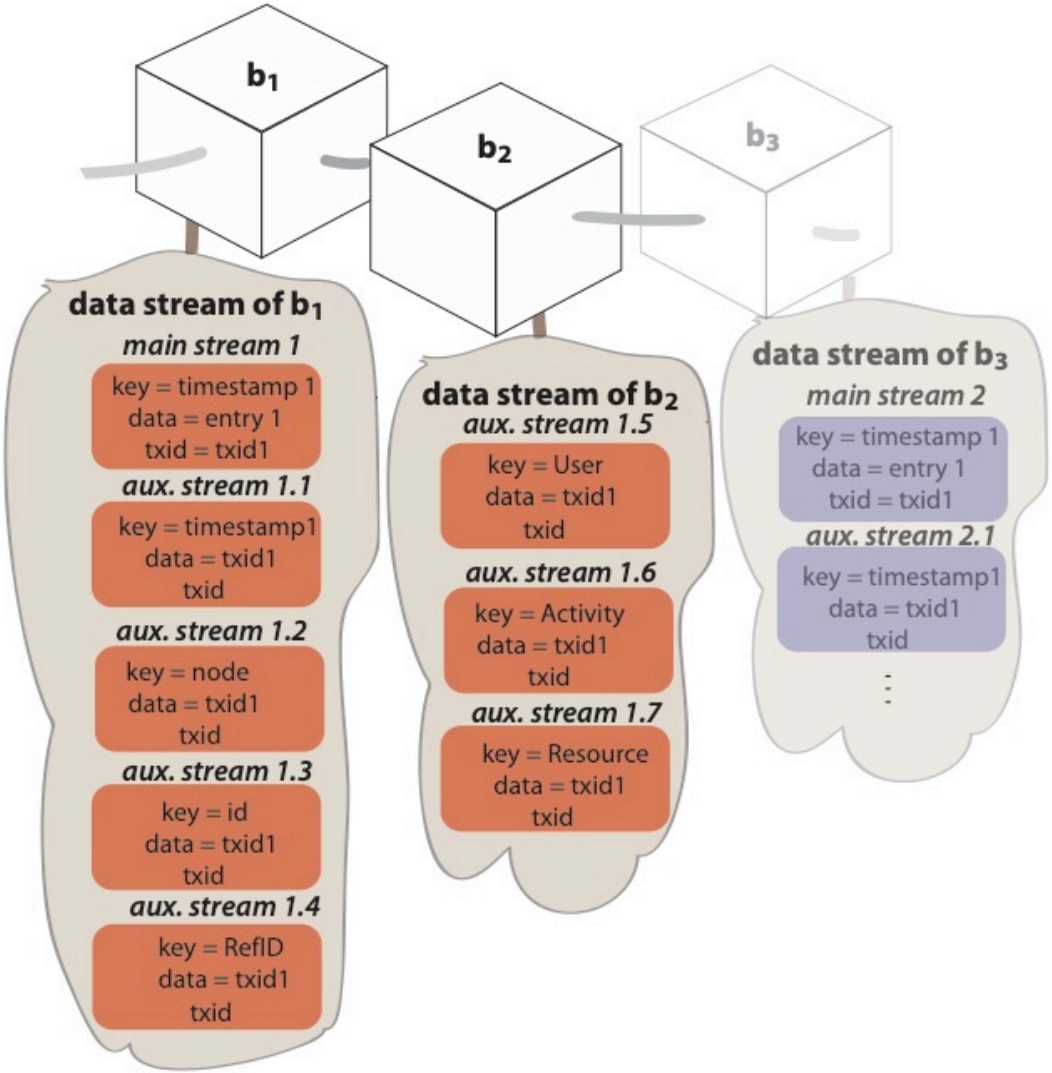
- Create a main stream item for each entry and 7 auxiliary stream items
- Save txid of main stream as data in auxiliary streams
- For query
  - Query on keys
  - Return txids of results
  - List main streams with above txids
  - Set intersection of lists

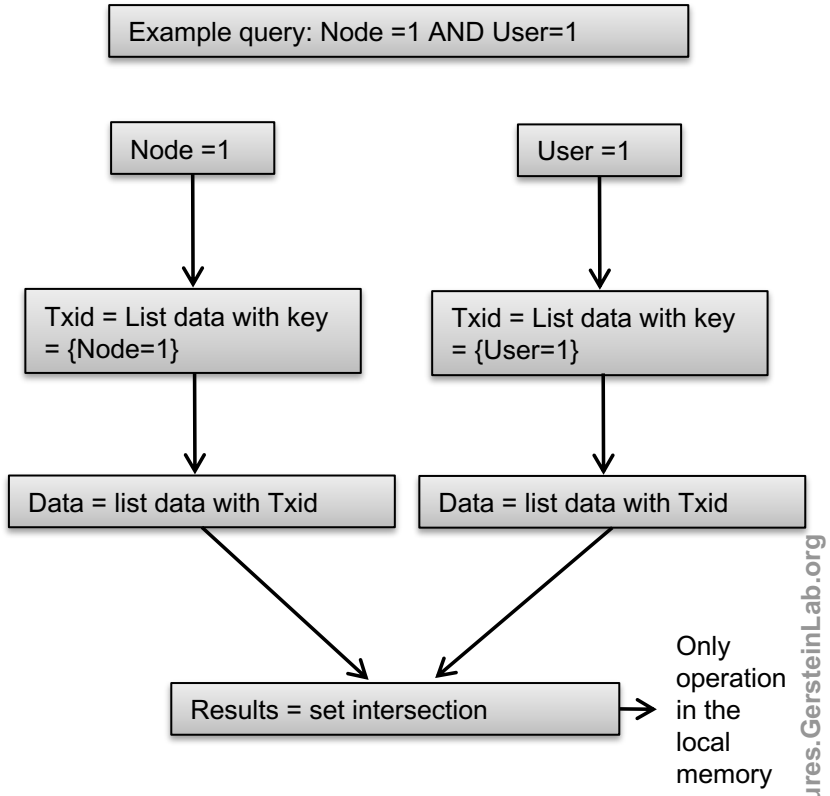
[Gursoy et al, BMC Med. Gen. (in press)]

Bigmem Solution

	Timestamp	Node	ID	Ref ID	User	Activity	Resource
entry 1	1522000002801	1	1	1	1	REQ_RESOURCE	MOD_UCSC_Genome_Bioinformatics
entry 2	1522000008352	1	2	1	1	VIEW_RESOURCE	MOD_UCSC_Genome_Bioinformatics
						⋮	

LOG

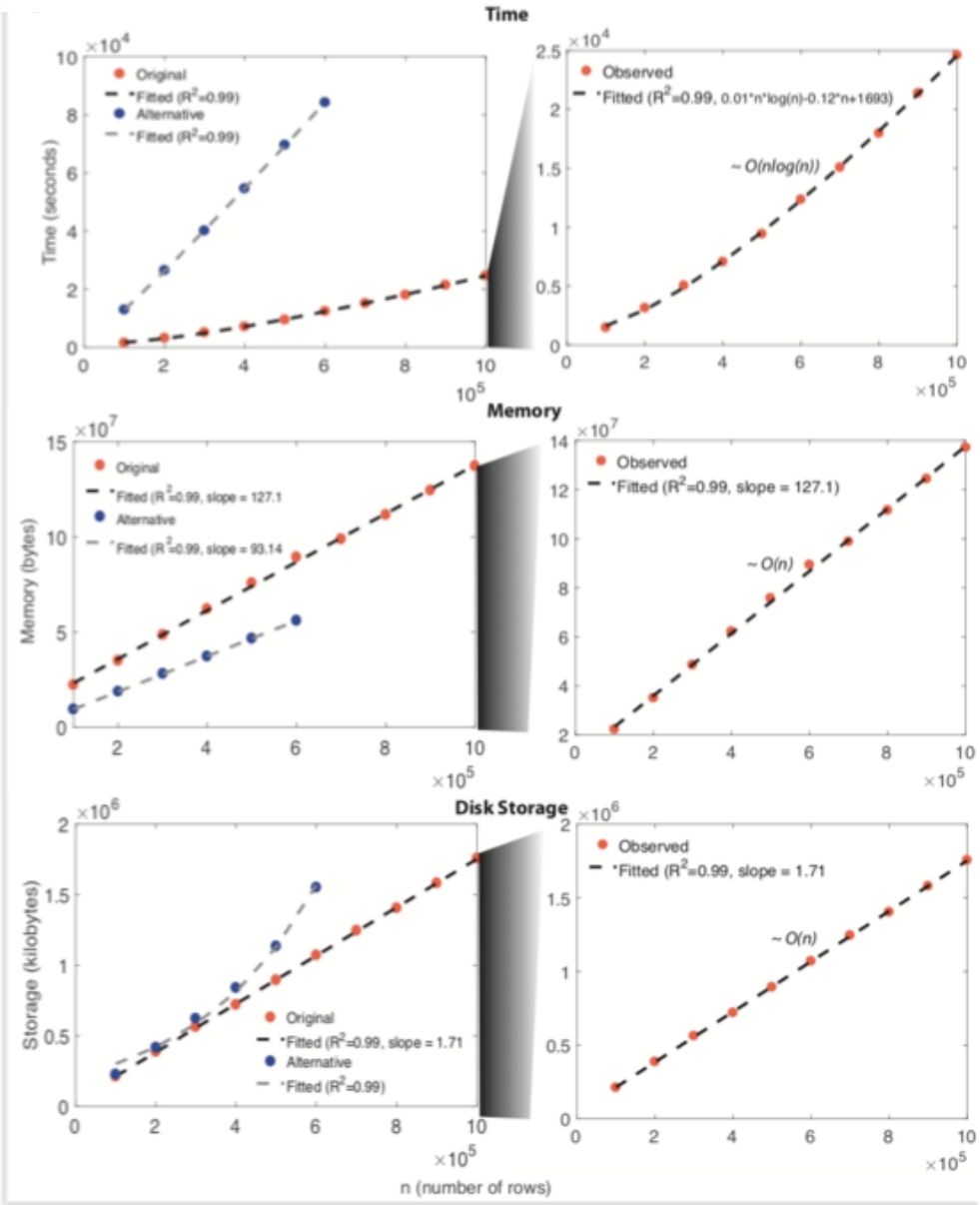




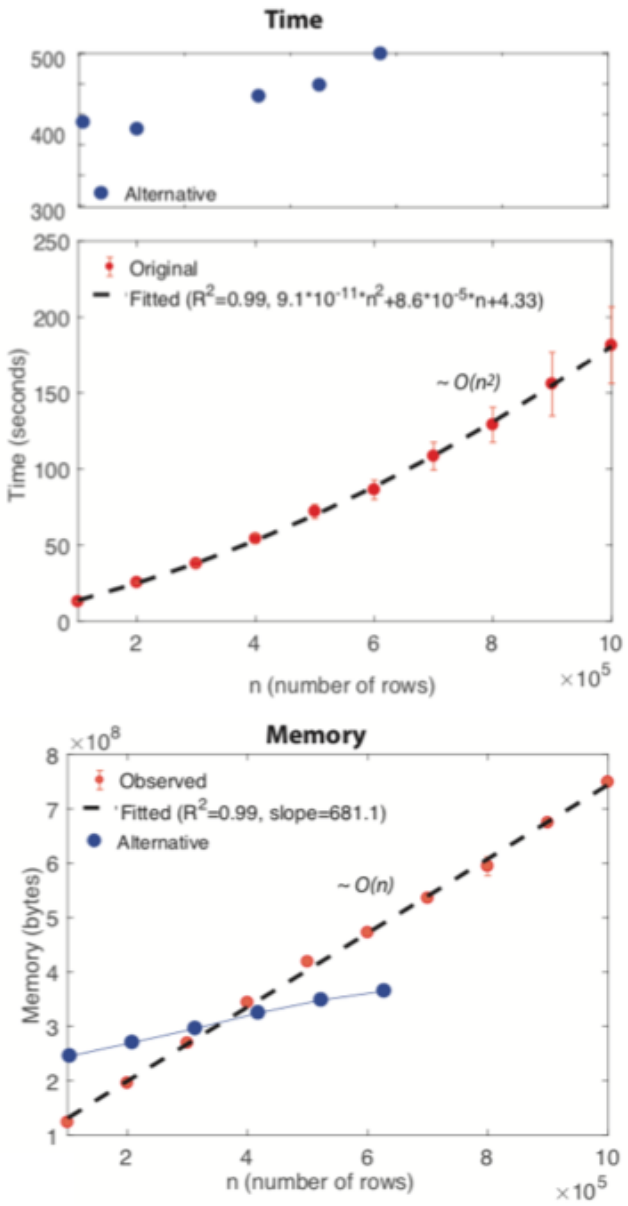
[Gursoy et al, BMC Med. Gen. (in press)]

Reasonable time/space efficiency

insertion



query



[Gursoy et al, BMC Med. Gen. (in press)]

Quantification of sensitive information leakage from functional genomics data: Obvious v subtle leakages & practical file formats for addressing this

- Intro. to Genomic Privacy
  - The **dilemma**: The genome as fundamental, inherited info that's very private v. need for large-scale sharing & mining for med. research
- Privacy & Functional Genomics Data
  - **2-sided nature** of this data presents particularly tricky privacy issues
  - Overview of **types of the leakage**, from obvious to subtle
- Subtle Leakage #1: **eQTLs**
  - Quantifying & removing further variant info from expression levels w/ ICI & predictability.
  - Instantiating a practical linking attack w/ noisy quasi-identifiers
- Subtle Leakage #2: **Signal Profiles**
  - Manifest appreciable leakage from large & small deletions.
  - Linking attacks possible but additional complication of SV discovery in addition to genotyping
- Practical solutions & file formats
  - Using **pBAM** file format to remove obvious large-scale leakage
  - Small subtle leaks combatted by restricting large-scale access. Hence, developing secure, **blockchain-based logging** technology (response to the iDash challenge)

Quantification of sensitive information leakage from functional genomics data: Obvious v subtle leakages & practical file formats for addressing this

- Intro. to Genomic Privacy
  - The **dilemma**: The genome as fundamental, inherited info that's very private v. need for large-scale sharing & mining for med. research
- Privacy & Functional Genomics Data
  - **2-sided nature** of this data presents particularly tricky privacy issues
  - Overview of **types of the leakage**, from obvious to subtle
- Subtle Leakage #1: **eQTLs**
  - Quantifying & removing further variant info from expression levels w/ ICI & predictability.
  - Instantiating a practical linking attack w/ noisy quasi-identifiers
- Subtle Leakage #2: **Signal Profiles**
  - Manifest appreciable leakage from large & small deletions.
  - Linking attacks possible but additional complication of SV discovery in addition to genotyping
- Practical solutions & file formats
  - Using **pBAM** file format to remove obvious large-scale leakage
  - Small subtle leaks combatted by restricting large-scale access. Hence, developing secure, **blockchain-based logging** technology (response to the iDash challenge)

Acknowledgements

A Harmanci,
D Greenbaum,
G Gürsoy,
R Bjornson, M Green,
S Strattan, O Jolanki,
F Navarro

papers.gersteinlab.org/subject/
privacy

PrivaSig.gersteinlab.org
PrivaSeq.gersteinlab.org
PrivaSeq3.gersteinlab.org

github.com/gersteinlab/
iDASH-blockchain

Also:

JOBS.gersteinlab.org

Extra

Info about content in this slide pack

- General PERMISSIONS
  - This Presentation is copyright Mark Gerstein, Yale University, 2019.
  - 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