Enhancing open data sharing for functional genomics experiments: Measures to quantify genomic information leakage & identify file formats for privacy preservation

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

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

Dilemma of Genomic Privacy + the 2-sided nature of functional genomics data: Analysis can be very General/Public or Individual/Private

- **General Dilemma of Genomic Privacy**
  - **Risks to the individual related to sharing variants**
  - **Yet benefit to medical research & society as a whole from data sharing**
- **Functional Genomics gives General quantifications** related to overall aspects of a state– ie gene activity as a function of: Developmental stage, Evolutionary relationships, Cell-type, Disease, &c.
- **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

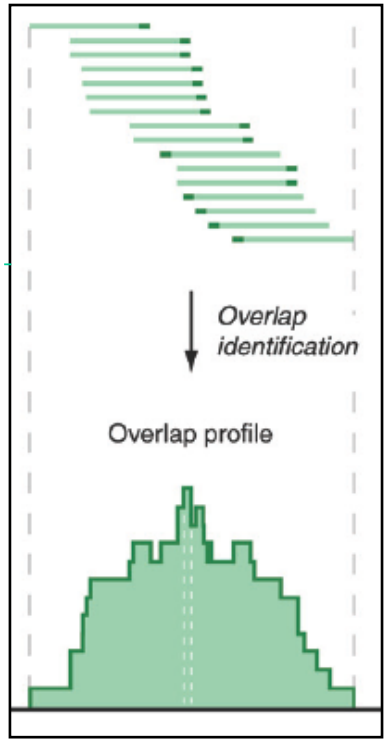
Fastq sequence files
~5-10 GB

```
ATACAAGCAAGTATAAGTTCGTATGCCGTCTT
GGAGGCTGGAGTTGGGGACGTATGCGGCATAG
TACCGATCGAGTCGACTGTAAACGTAGGCATA
ATTCTGACTGGTGTCATGCTGATGTACTTAAA
```

Index-building + Alignment to reference genome

BAM files
~1-2-fold reduction

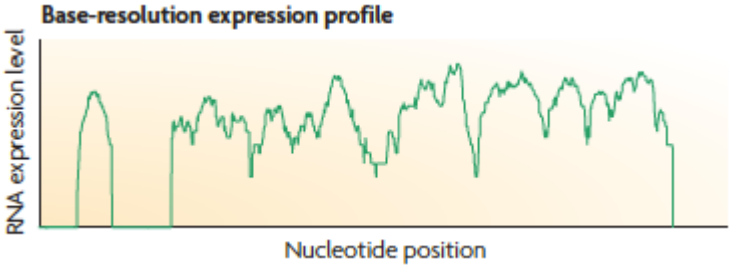
Conversion to signal track by overlapping reads



Reads => Signal

BigWig files
~25-fold reduction

Mapping
to genes



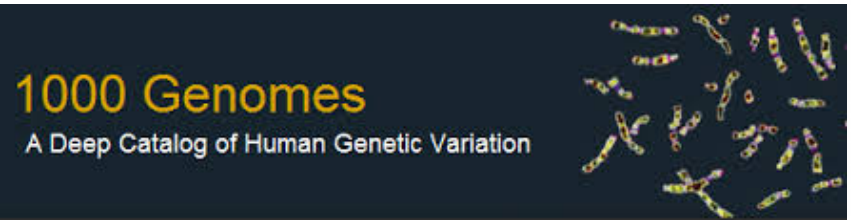
Quantitative information from RNA-seq signal:
average signals at exon level (RPKMs)

Gene/Transcript
expression matrix
~20-fold reduction

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

Representative Functional Genomics & Genotype 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



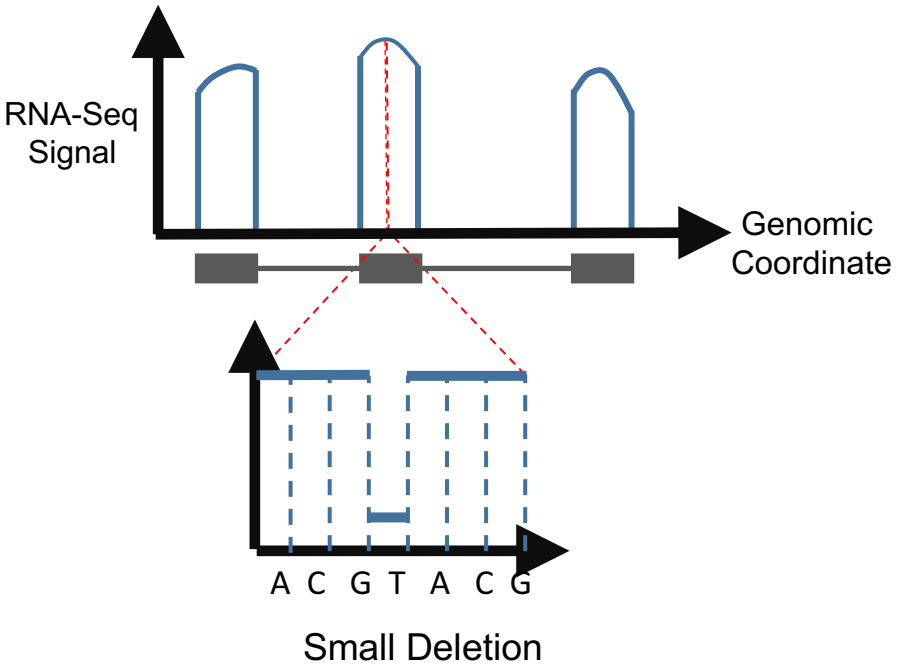
- (Functional) Genomic Privacy
  - The **dilemma**: The genome as fundamental, inherited info that's very private v. need for large-scale sharing for med. research
  - **2-sided nature** of functional genomics data presents particularly tricky privacy issues
- Subtle Leakage: **Signal Profiles**
  - Manifest appreciable leakage from large & small deletions
  - Quantified by ICI and π_{GW}
  - Linking attacks possible but additional complication of SV discovery in addition to genotyping

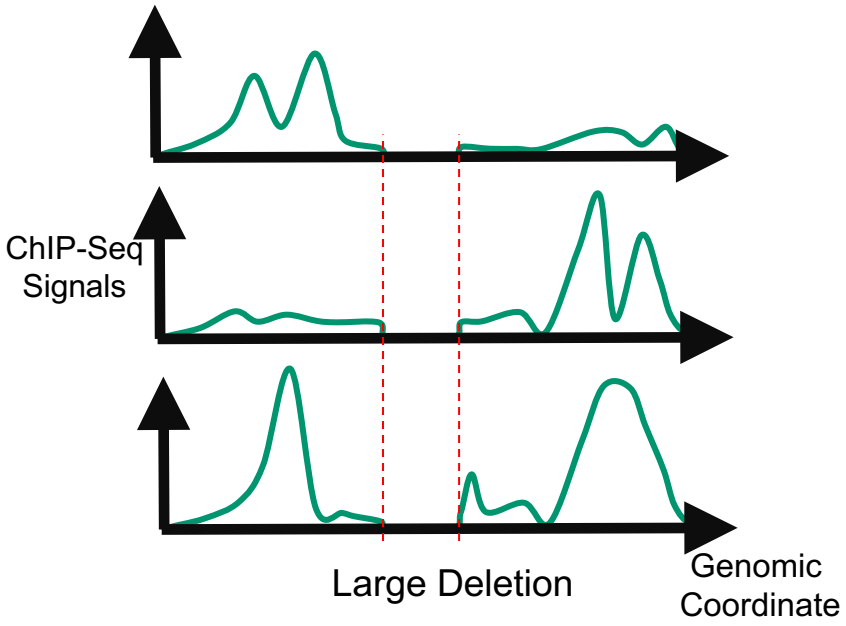
- **Practical solutions**
  - **FANCY Risk assessment:** Predicts the number of leaking variants from functional genomics data without the need for genotyping.
  - **pBAM file format:** Remove obvious variants from leakage +
    Can be used with existing softwares/pipelines, giving consistent quantifications

- (Functional) Genomic Privacy
  - The **dilemma**: The genome as fundamental, inherited info that's very private v. need for large-scale sharing for med. research
  - **2-sided nature** of functional genomics data presents particularly tricky privacy issues
- Subtle Leakage: **Signal Profiles**
  - Manifest appreciable leakage from large & small deletions
  - Quantified by ICI and π_{GW}
  - Linking attacks possible but additional complication of SV discovery in addition to genotyping

- **Practical solutions**
  - **FANCY Risk assessment:** Predicts the number of leaking variants from functional genomics data without the need for genotyping.
  - **pBAM file format:** Remove obvious variants from leakage +
    Can be used with existing softwares/pipelines, giving consistent quantifications

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

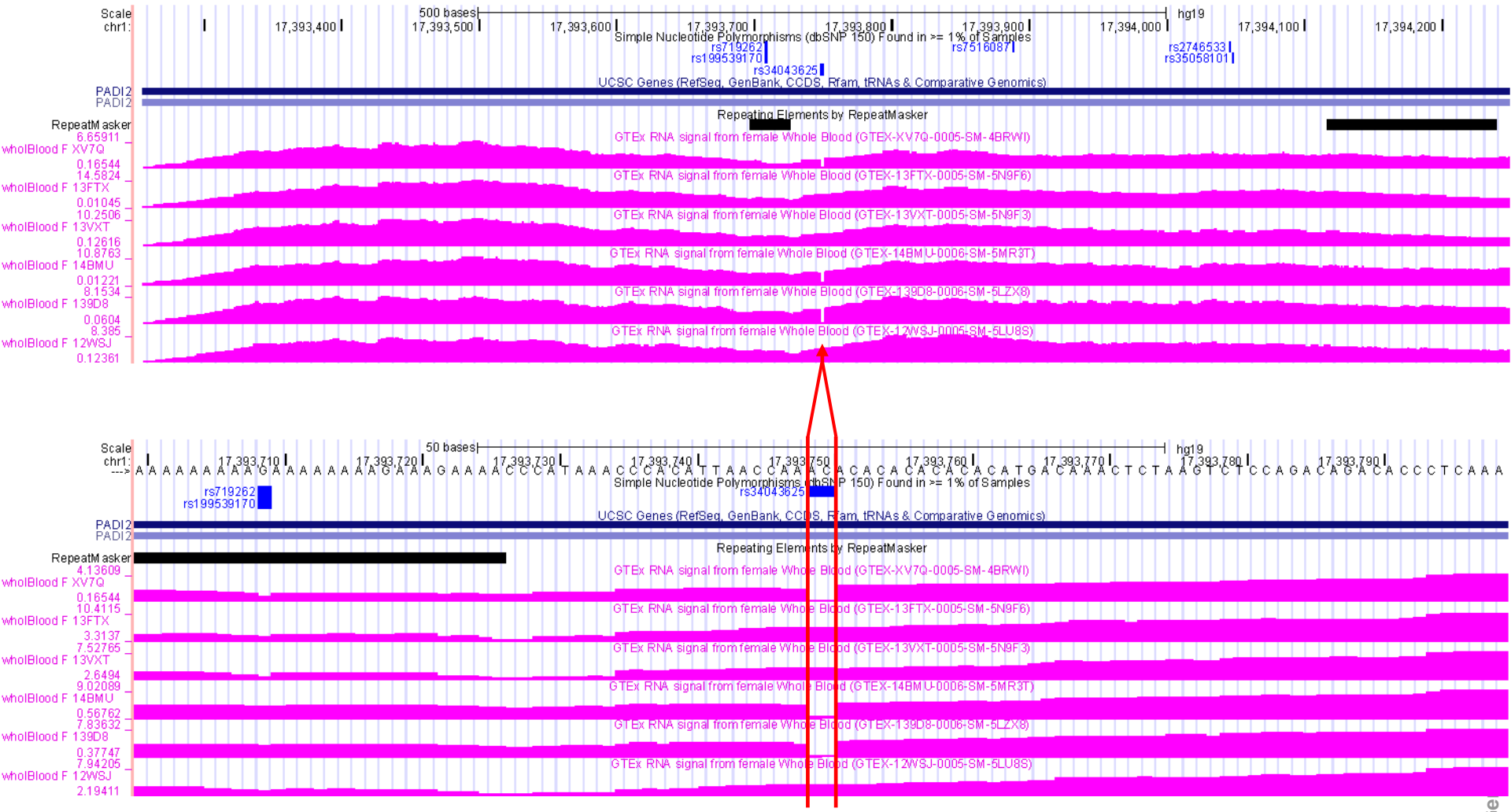




RNA-seq also shows large deletions

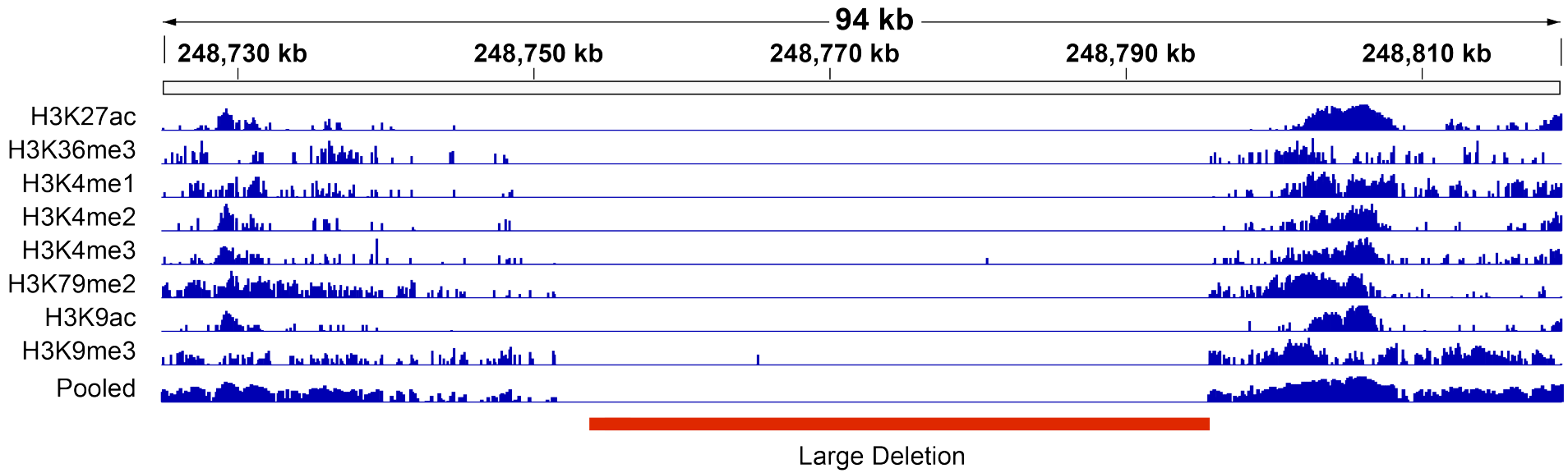
[Harmanci & Gerstein, *Nat. Comm.* ('18)]

Example of Small Deletion Evident in Signal Profile



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

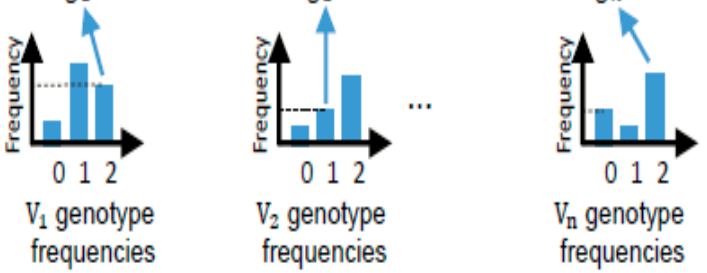
Example of Large Deletion Evident in Signal Profile



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

Metrics for quantifying leakage

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



- **ICI**: For a given variant genotype, measures how much information it supplies for pinpointing an individual in a population. Higher weight to genotypes that have low population frequency.

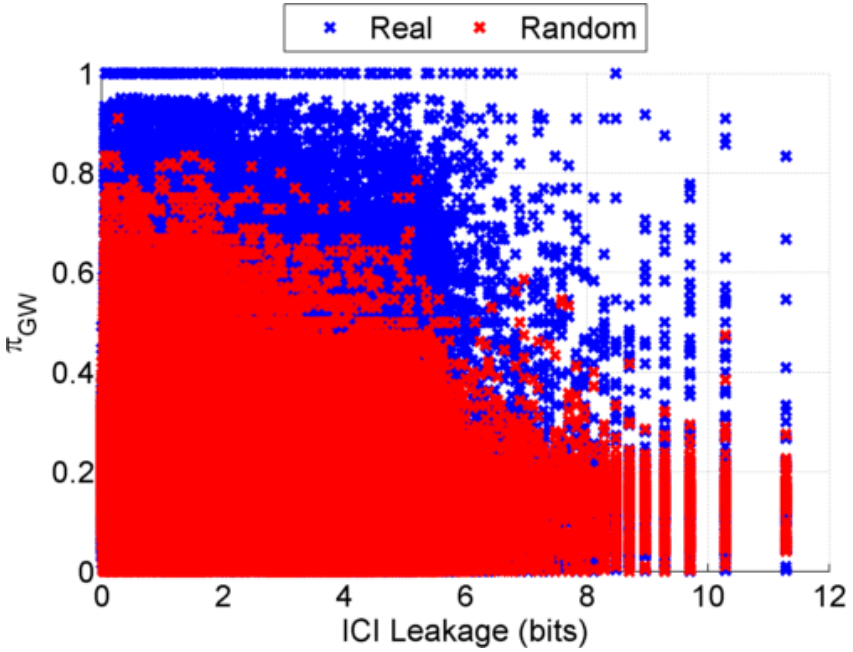
- π_{GW} :Predictability is defined as the conditional probability of a deletion genotype *g* given 5 features computed from a functional genomics signal profile (eg avg. signal, multi-mapability, signal shape, &c…)

$$\pi_{\mathrm{GW}}\left(G_{[i,j]} = g, \mathbf{S}_{[i,j]}\right) = P_{\mathrm{GW}}\left(G_{[i,j]} = g \,\middle|\, \begin{array}{l}\log_2\left(\bar{s}_{[i,j]}\right),\\ \log_2\left(\bar{m}_{[i,j]}\right),\\ \log_2\left(\rho_{[i,j]}\right),\\ \log_2\left(\eta_{[i,j]}\right),\\ \log_2\left(\tau_{[i,j]}\right)\end{array}\right).$$

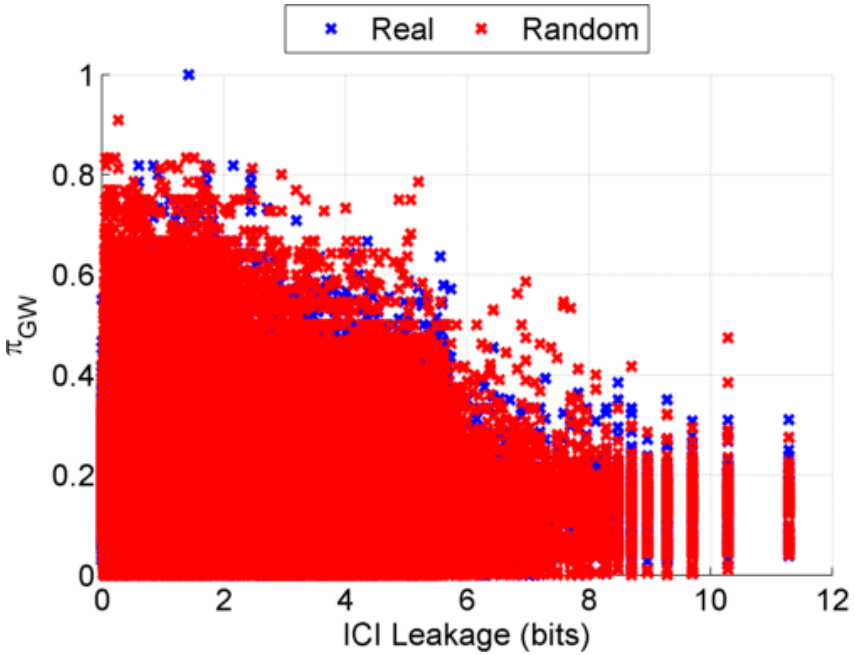
[Harmanci & Gerstein, *Nat. Comm.* ('18)]

Information Leakage from SV Deletions

a) Before Anonymization



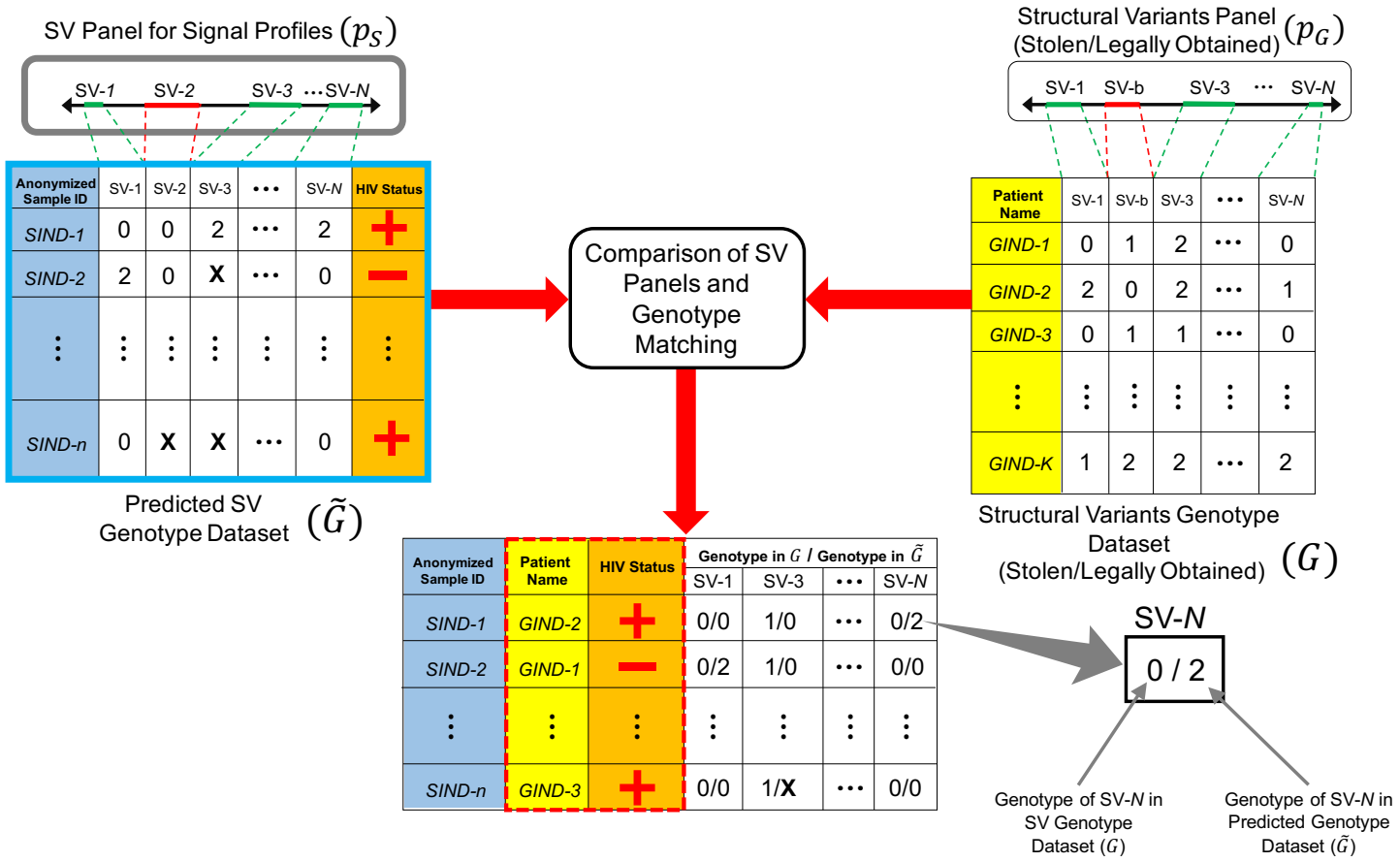
b) After Anonymization



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

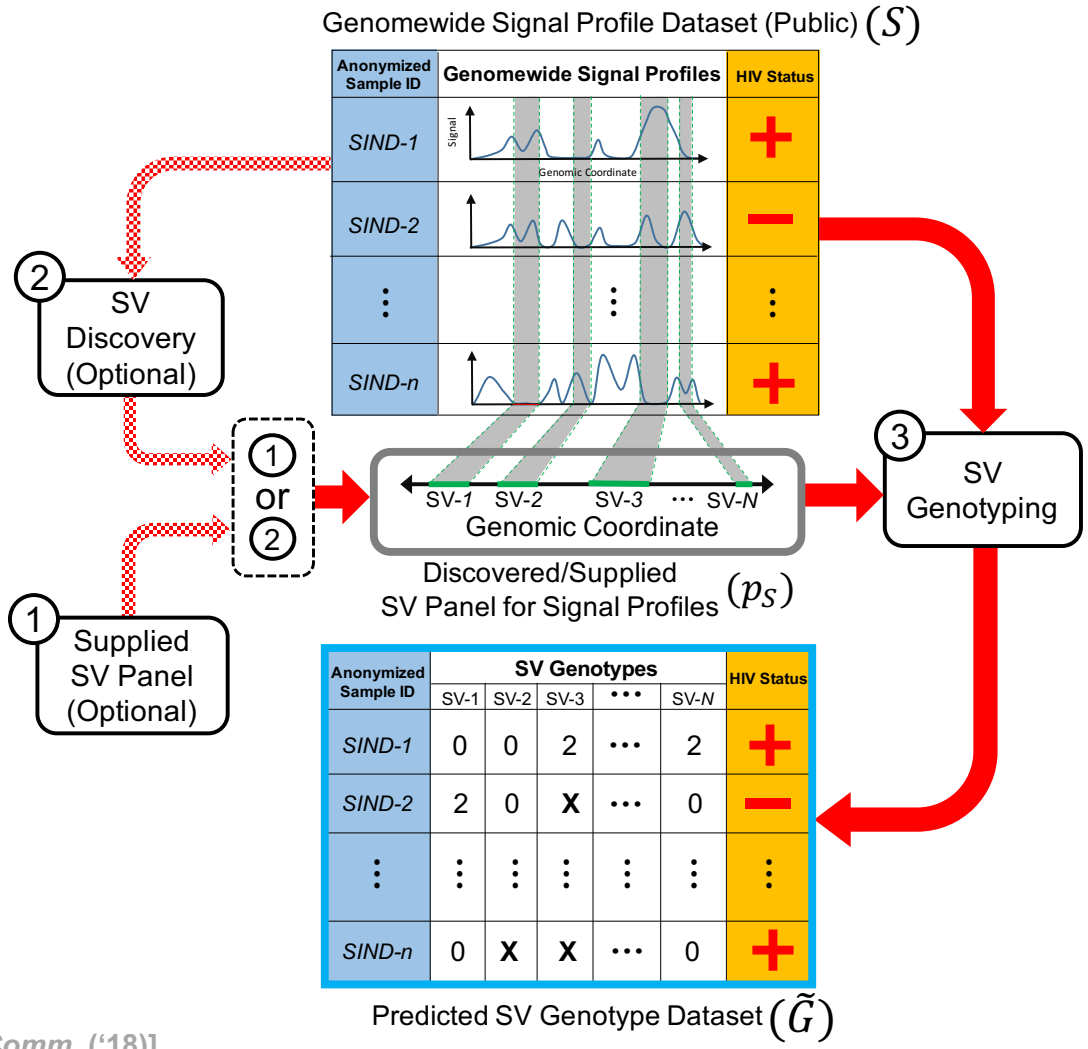
[Harmanci & Gerstein, *Nat. Comm.* ('18)]

Another type of Linking Attack: Linking based on SV Genotyping



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

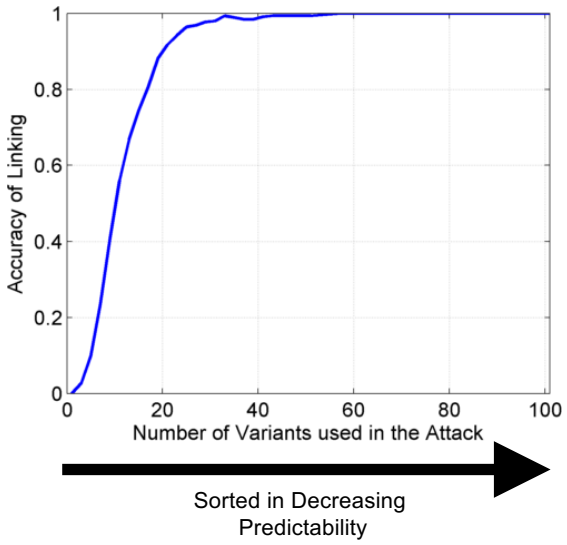
Another type of Linking Attack: First Doing SV Genotyping



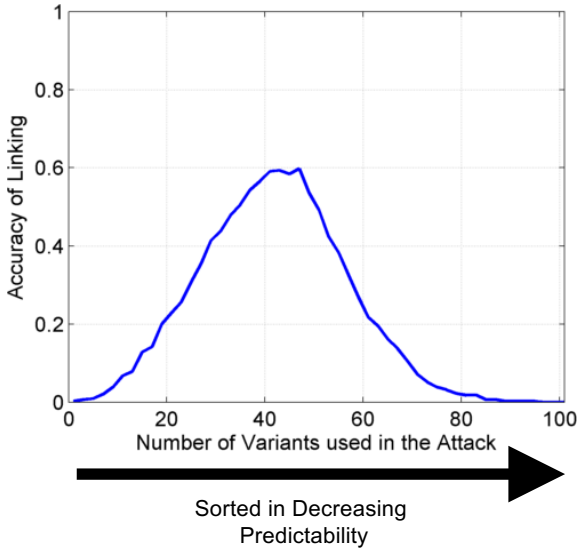
[Harmanci & Gerstein, *Nat. Comm.* ('18)]

Linking Attack Based on SV Deletions in gEUVADIS Dataset

c) Genotyping
(1kG MAF>0.01)



d) Discovery + Genotyping



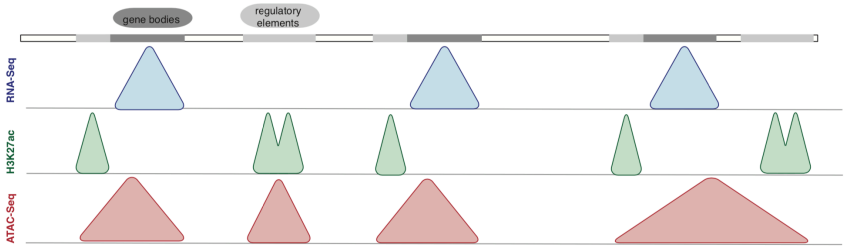
[Harmanci & Gerstein, *Nat. Comm.* ('18)]

- (Functional) Genomic Privacy
  - The **dilemma**: The genome as fundamental, inherited info that's very private v. need for large-scale sharing for med. research
  - **2-sided nature** of functional genomics data presents particularly tricky privacy issues
- Subtle Leakage: **Signal Profiles**
  - Manifest appreciable leakage from large & small deletions
  - Quantified by ICI and π_{GW}
  - Linking attacks possible but additional complication of SV discovery in addition to genotyping

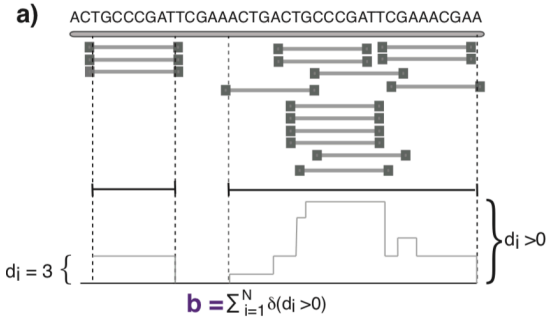
- **Practical solutions**
  - **FANCY Risk assessment:** Predicts the number of leaking variants from functional genomics data without the need for genotyping.
  - **pBAM file format:** Remove obvious variants from leakage +
    Can be used with existing softwares/pipelines, giving consistent quantifications

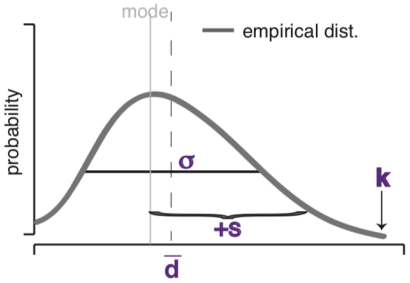
Number of leaking variants: Can we predict without genotyping?

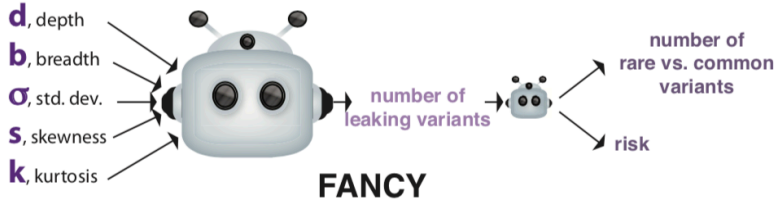
- Genotyping is costly & requires re-parameterization for different assays
- Est. number of leaking variants is important before data release
  - Different assays sequence different regions.
  - May not leak enough individually, but can be combined



- Waterman-Lander statistics can inform us about genotyping capacity
  - Hypothesis: The extended statistics from reads of functional genomics experiments can be used as features for a supervised regression model
    - **Noisy features → Gaussian Process Learning**

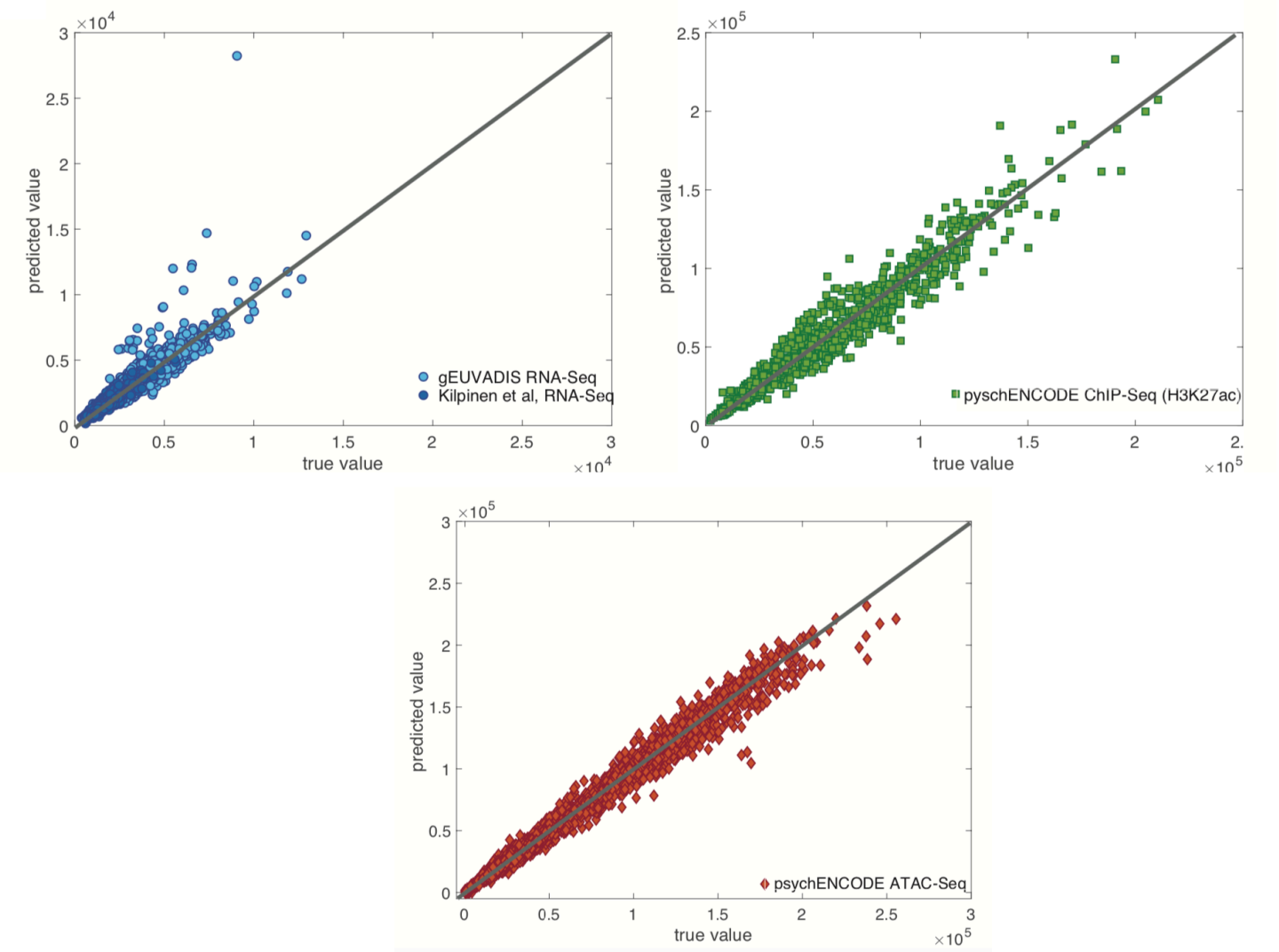






[*Gursoy et al,* bioRxiv 775338]

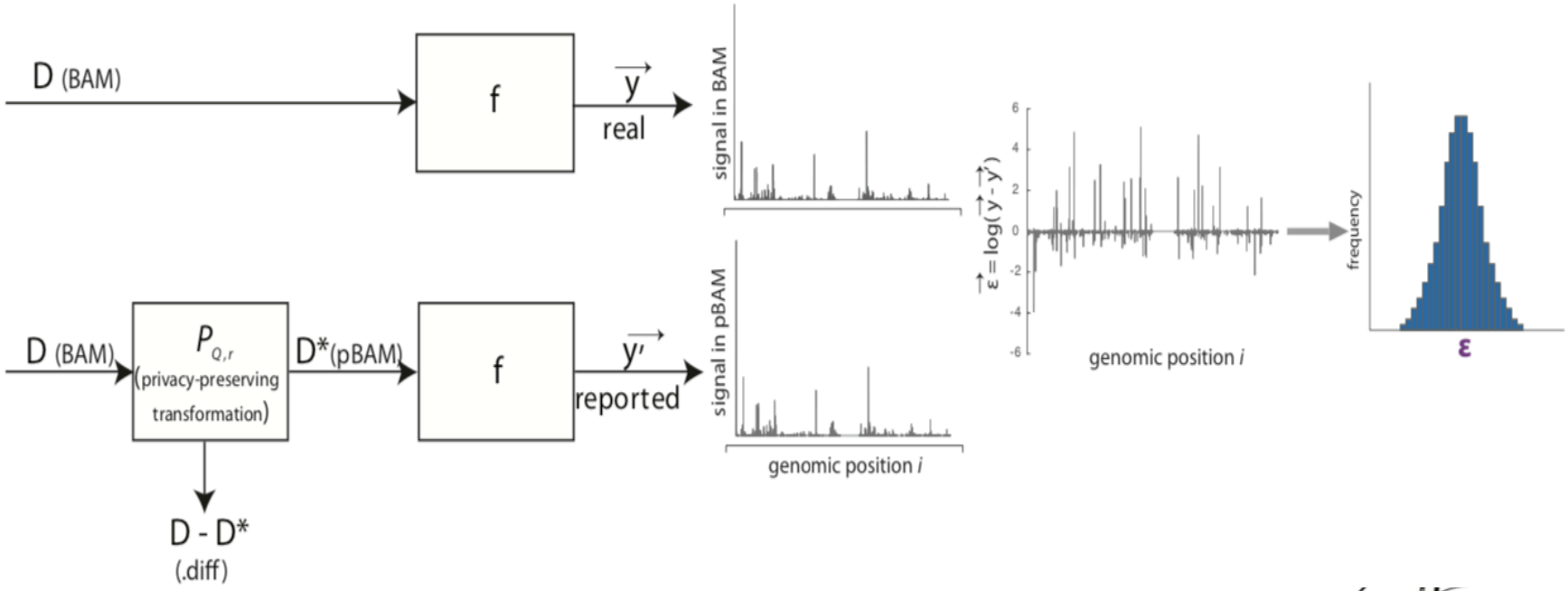
FANCY: Trained on gEUVADIS & psychENCODE and tested on an independent test dataset



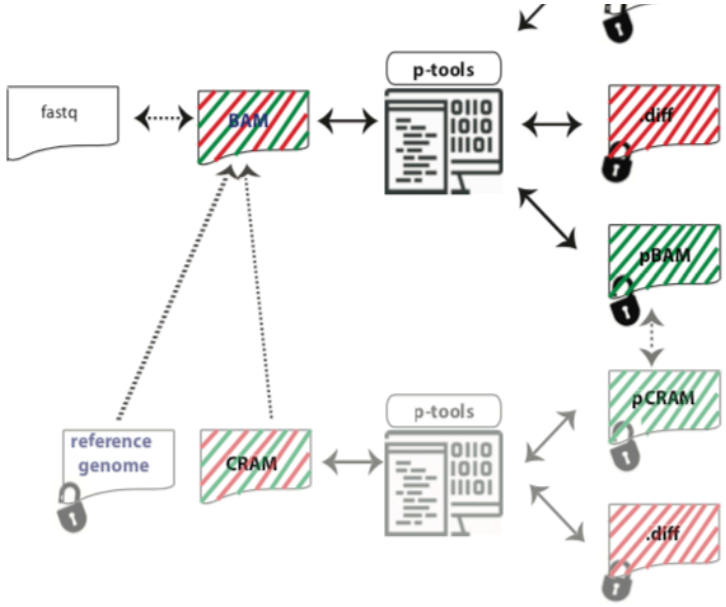
TEST Results

[*Gursoy et al,* bioRxiv 775338]

Privacy-aware Binary Alignment Mapping (pBAM)



- **A manipulation of 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)]

- (Functional) Genomic Privacy
  - The **dilemma**: The genome as fundamental, inherited info that's very private v. need for large-scale sharing for med. research
  - **2-sided nature** of functional genomics data presents particularly tricky privacy issues
- Subtle Leakage: **Signal Profiles**
  - Manifest appreciable leakage from large & small deletions
  - Quantified by ICI and π_{GW}
  - Linking attacks possible but additional complication of SV discovery in addition to genotyping

- **Practical solutions**
  - **FANCY Risk assessment:** Predicts the number of leaking variants from functional genomics data without the need for genotyping.
  - **pBAM file format:** Remove obvious variants from leakage +
    Can be used with existing softwares/pipelines, giving consistent quantifications

Enhancing open data sharing for functional genomics experiments:
Measures to quantify genomic information leakage & identify file formats for privacy preservation

- (Functional) Genomic Privacy
  - The **dilemma**: The genome as fundamental, inherited info that's very private v. need for large-scale sharing for med. research
  - **2-sided nature** of functional genomics data presents particularly tricky privacy issues
- Subtle Leakage: **Signal Profiles**
  - Manifest appreciable leakage from large & small deletions
  - Quantified by ICI and π_{GW}
  - Linking attacks possible but additional complication of SV discovery in addition to genotyping

- **Practical solutions**
  - **FANCY Risk assessment:** Predicts the number of leaking variants from functional genomics data without the need for genotyping.
  - **pBAM file format:** Remove obvious variants from leakage + Can be used with existing softwares/pipelines, giving consistent quantifications

Acknowledgements

G Gürsoy,

A Harmanci

C Brannon,

S Strattan,
O Jolanki,
F Navarro

PrivaSig.gersteinlab.org

PrivaSeq3.gersteinlab.org

github.com/gersteinlab/FANCY

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