Genomic Privacy

- Personalized genomic data generation is booming
- Main focus is on protecting variants
- "Detection of genome in a mixture"
  - Individuals give consent to participate but request anonymity
    - HAPMAP, Personal genome project, 1000 Genomes…
- Larger and more datasets leads to more realistic risks of linking attacks, that may be much more damaging than detection of genome in a mixture attacks

Identifying Personal Genomes by Surname Inference

Melissa Gymrek,[1,2,3,4] Amy L. McGuire,[5] David Golan,[6] Eran Halperin,[7,8,9] Yaniv Erlich[1]*

Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays

Nils Homer[1,2], Szabolcs Szelinger[1], Margot Redman[1], David Duggan[1], Waibhav Tembe[1], Jill Muehling[1], John V. Pearson[1], Dietrich A. Stephan[1], Stanley F. Nelson[2], David W. Craig[1]*

On Sharing Quantitative Trait GWAS Results in an Era of Multiple-omics Data and the Limits of Genomic Privacy

Hae Kyung Im,[1,*] Eric R. Gamazon,[2] Dan L. Nicolae,[2,3,4] and Nancy J. Cox[2,3,*]

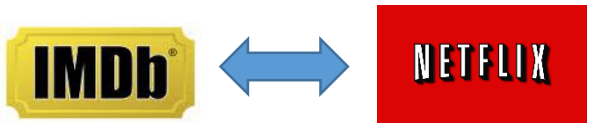
Identifying Participants in the Personal Genome Project by Name

Latanya Sweeney, Akua Abu, Julia Winn

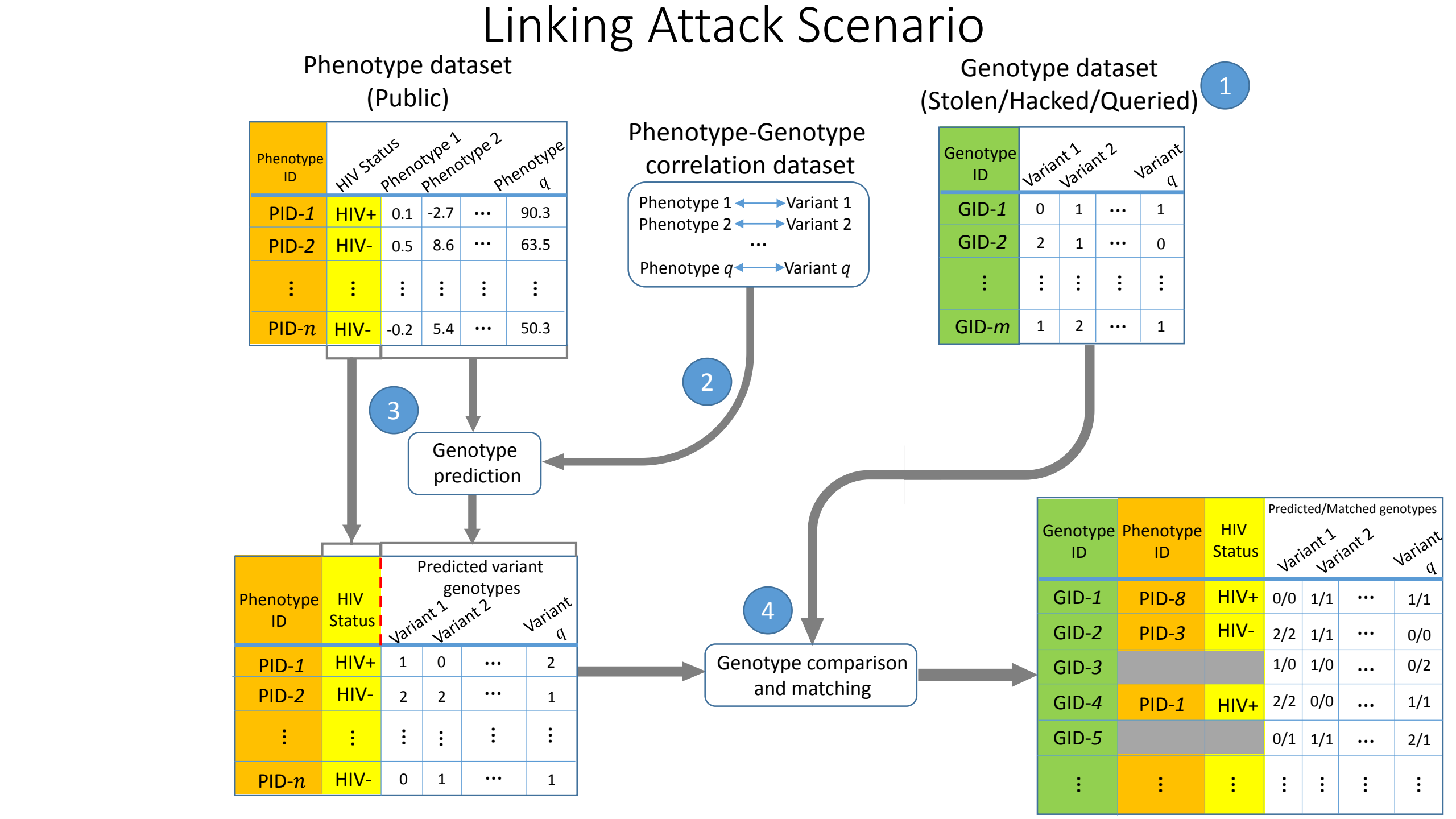
Harvard College
Cambridge, Massachusetts
latanya@fas.harvard.edu, aabu@college.harvard.edu, jwinn@post.harvard.edu

Robust De-anonymization of Large Sparse Datasets

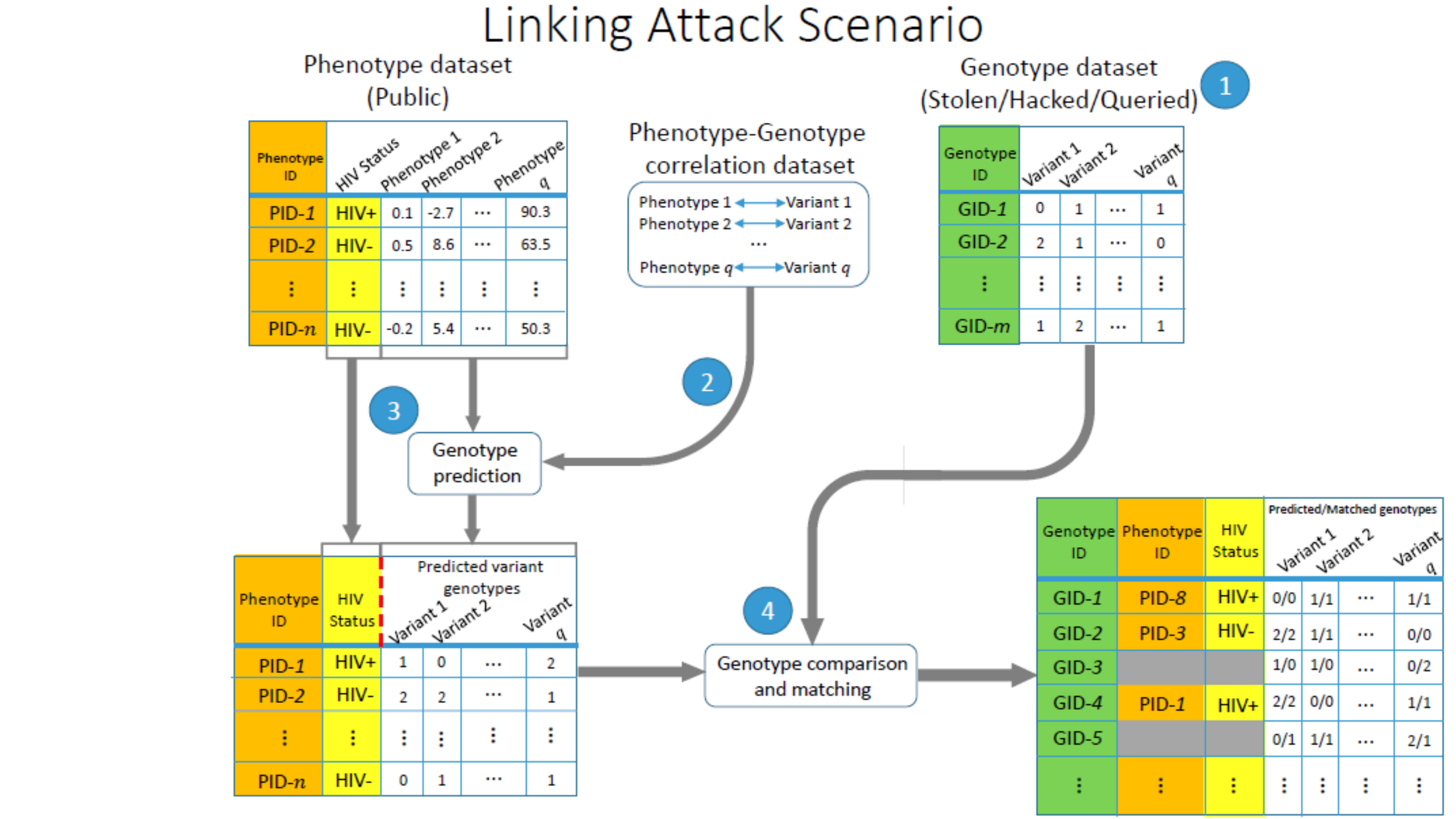
Arvind Narayanan and Vitaly Shmatikov
The University of Texas at Austin



Linking Attack Scenario



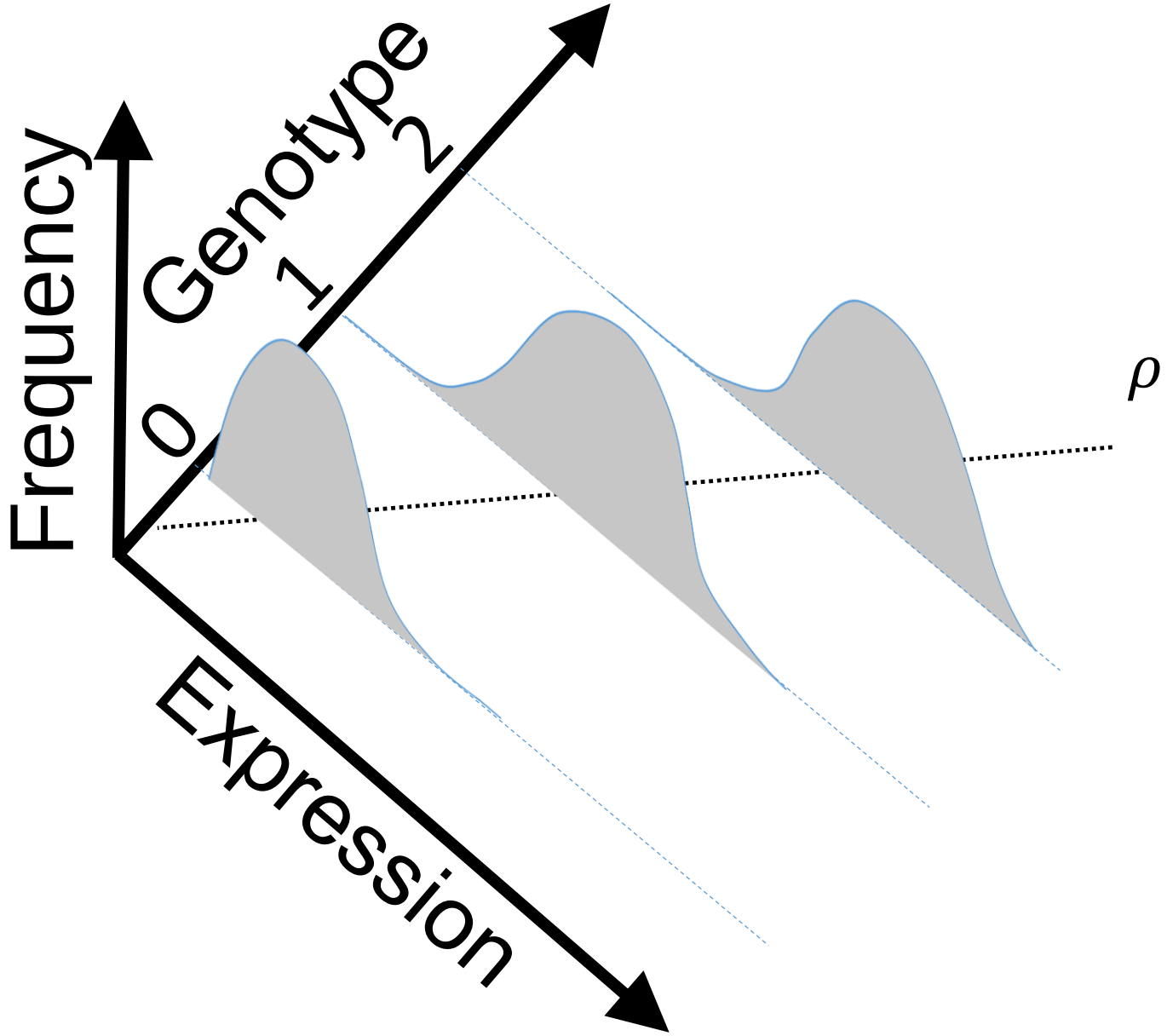
Linking Attack Scenario



Representative Expression, Genotype, eQTL Datasets

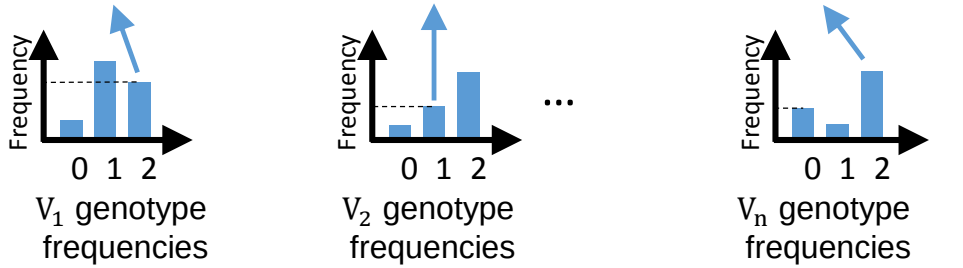
- mRNA sequencing for 462 individuals
  - Publicly availableQuantification for protein coding genes
- Approximately 3,000 cis-eQTL (FDR<0.05)
- Genotypes are available from the 1000 Genomes Project

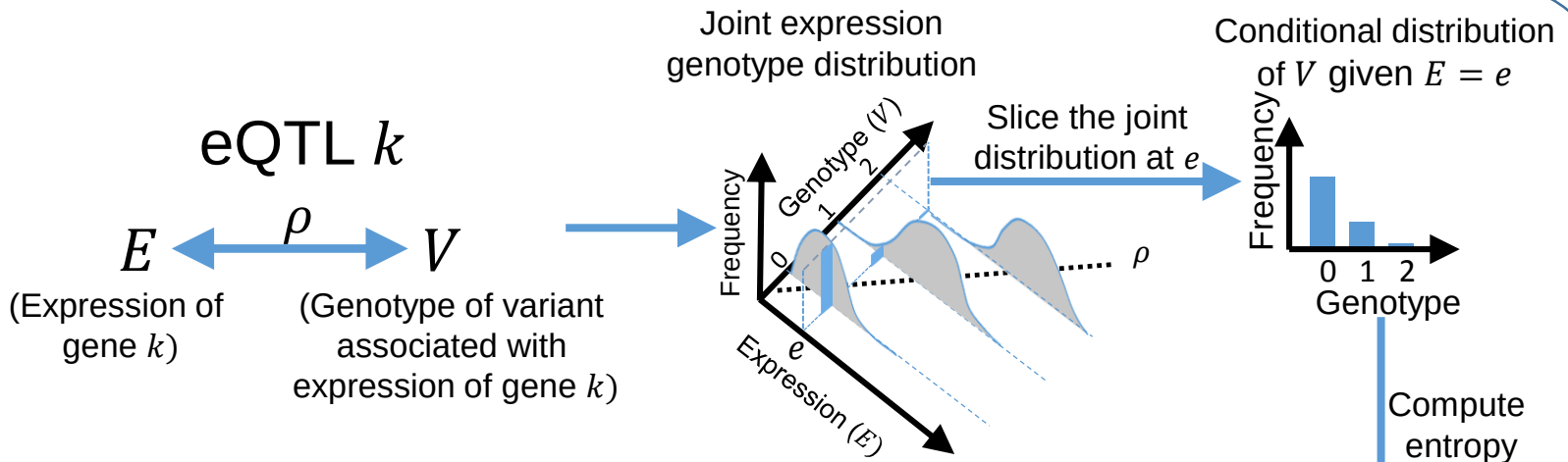
Expression and Genotype Distribution



Information Content and Predictability

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

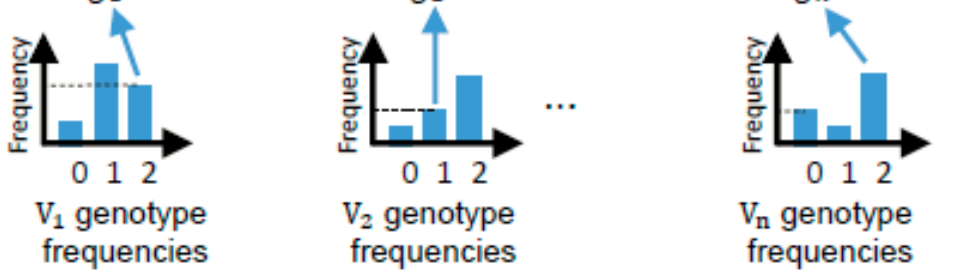




$$\pi\left(V \mid E = e\right) = e^{-H\left(V \mid E = e\right)}$$

Information Content and Predictability

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

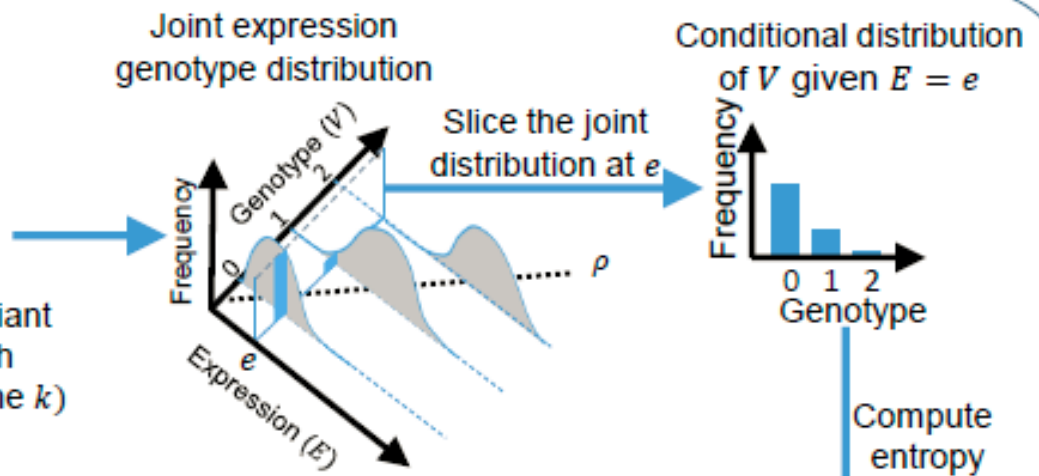


eQTL k

E ←ρ→ V

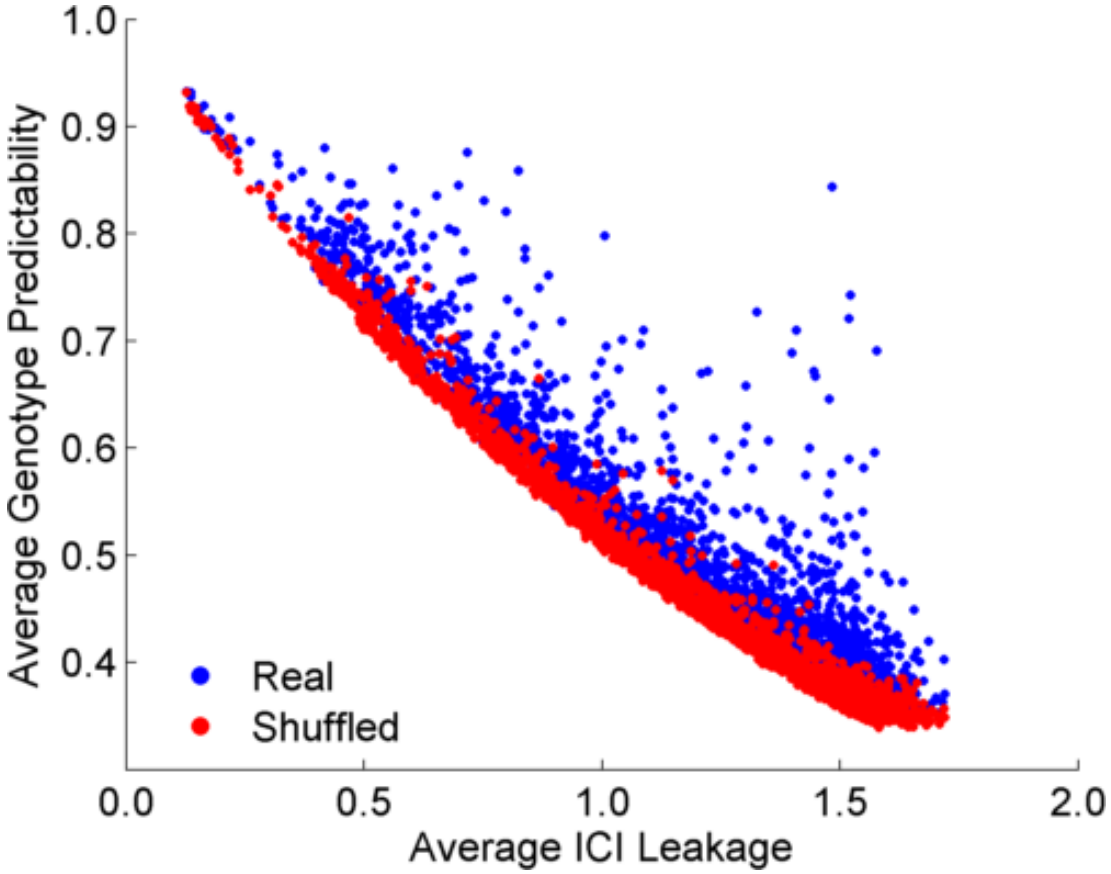
(Expression of gene k)

(Genotype of variant associated with expression of gene k)

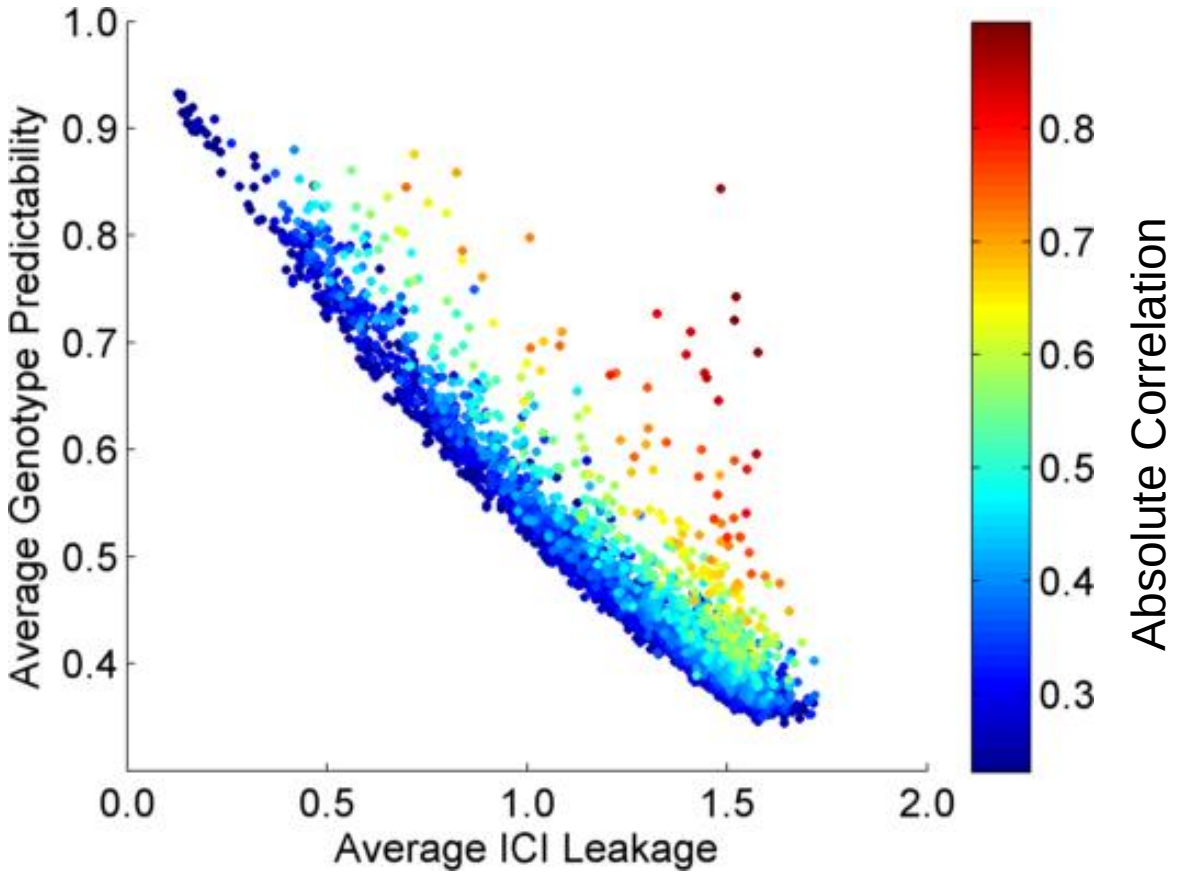


$$\pi\left(V \middle| E = e\right) = e^{-H\left(V \middle| E = e\right)}$$

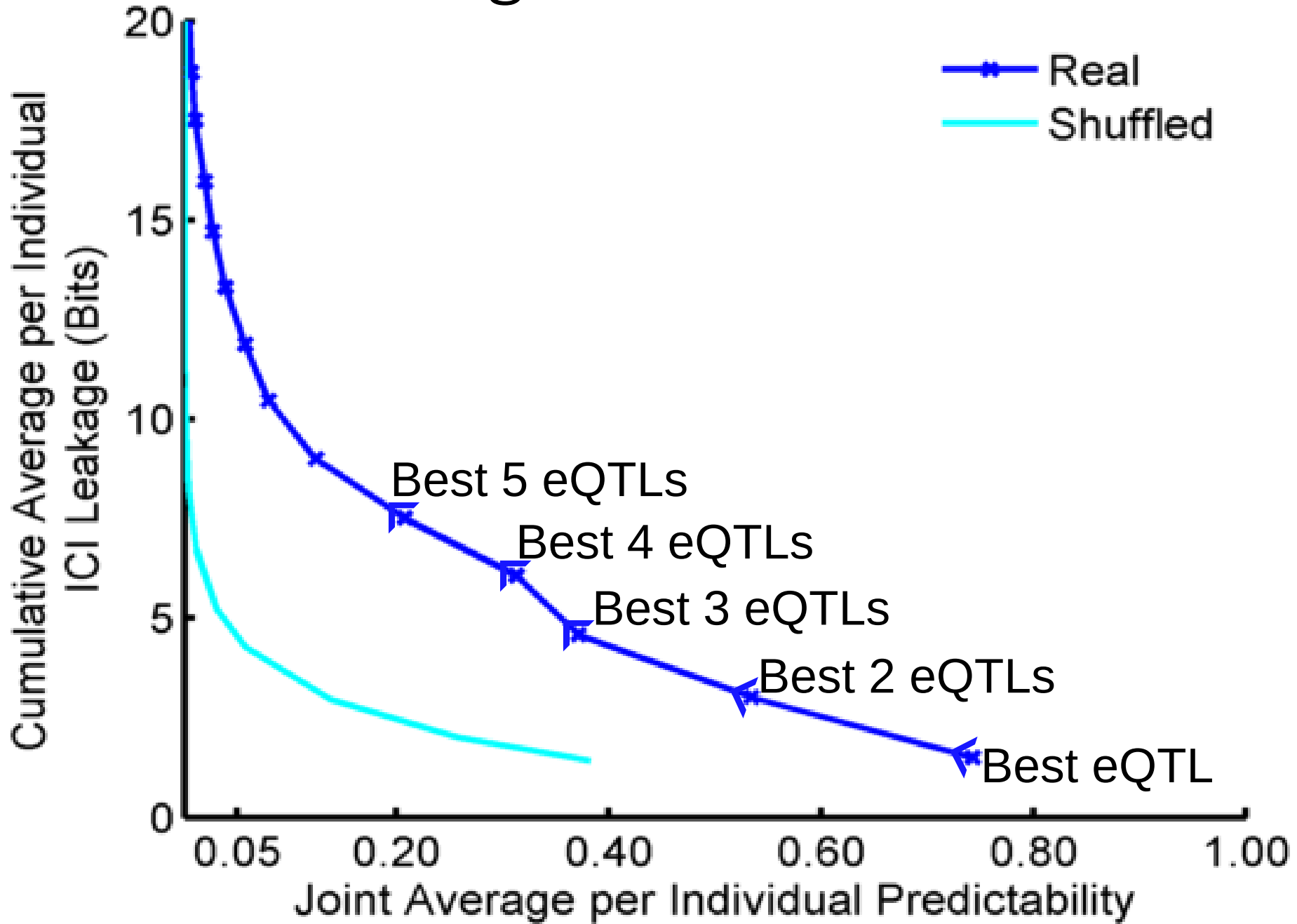
Per eQTL and ICI Cumulative Leakage versus Genotype Predictability



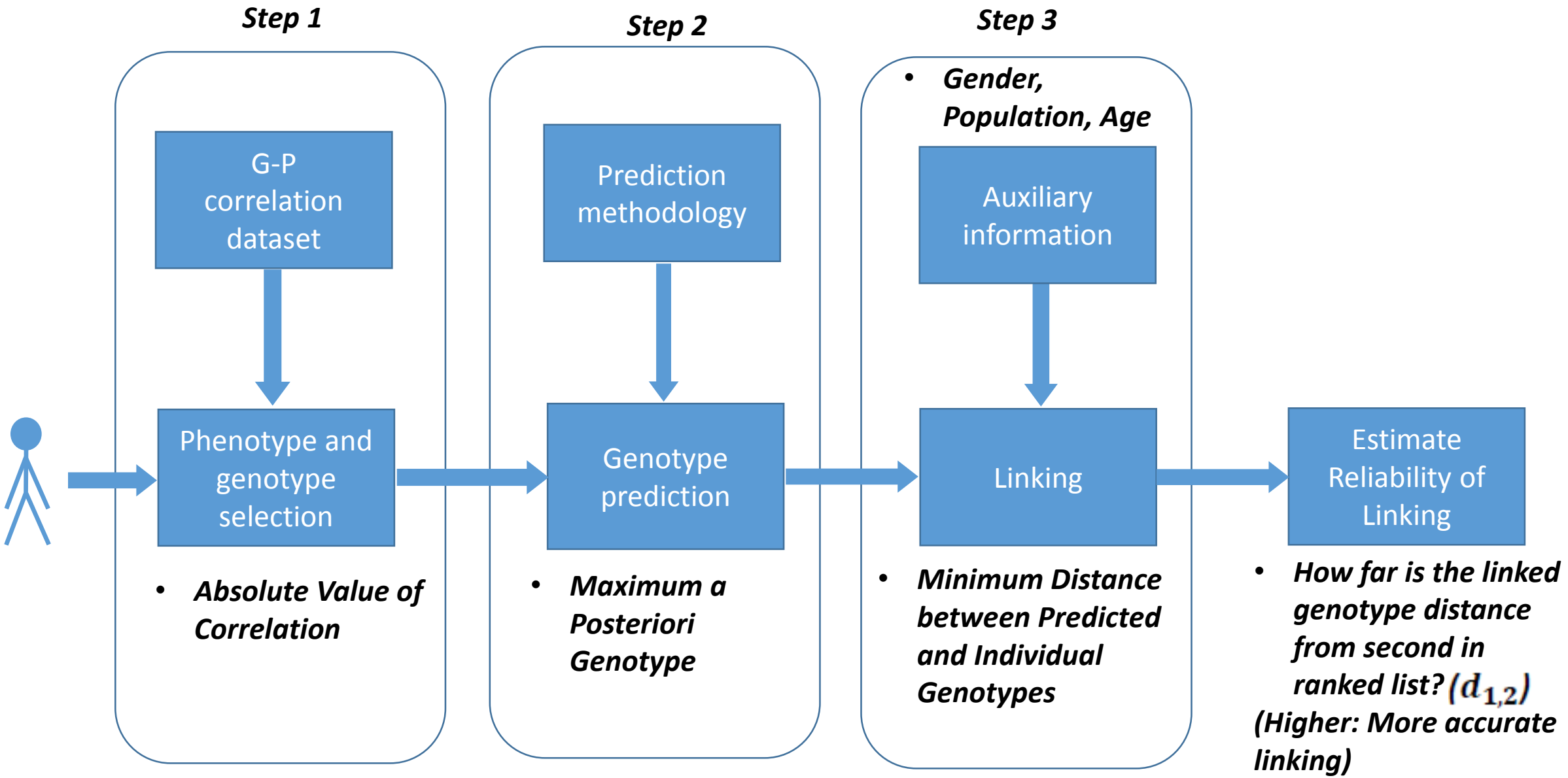
Colors by absolute correlation

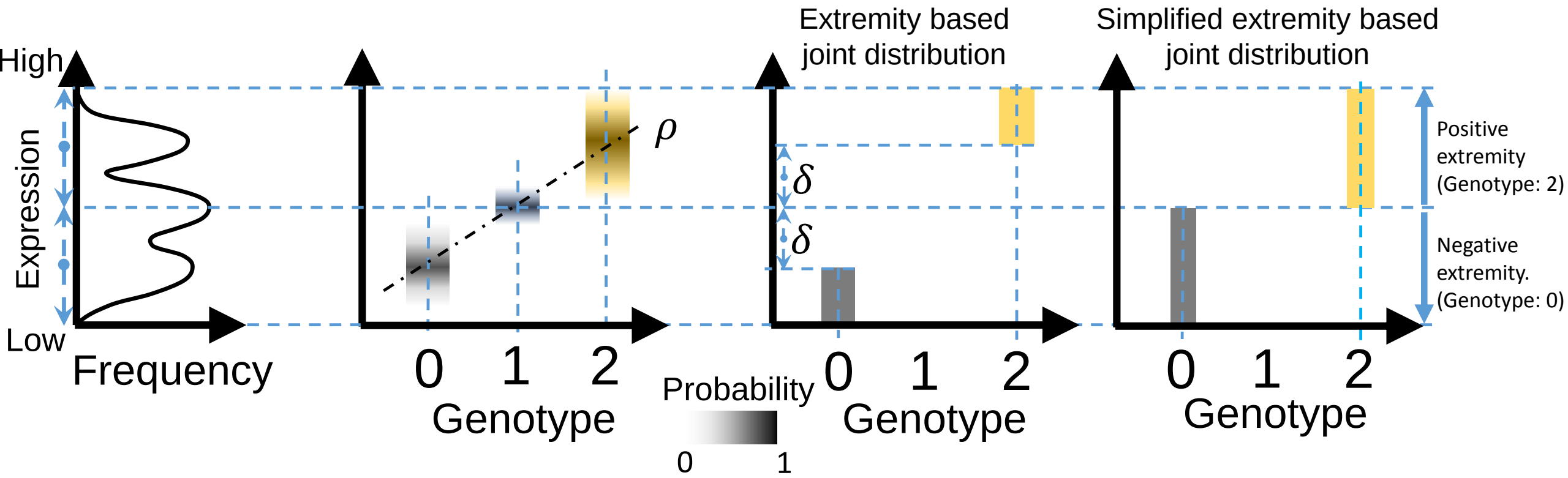


Cumulative Leakage versus Joint Predictability

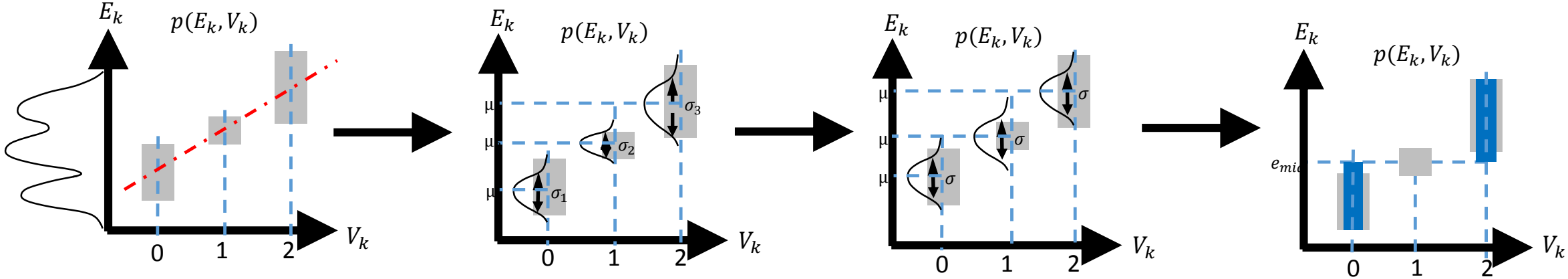


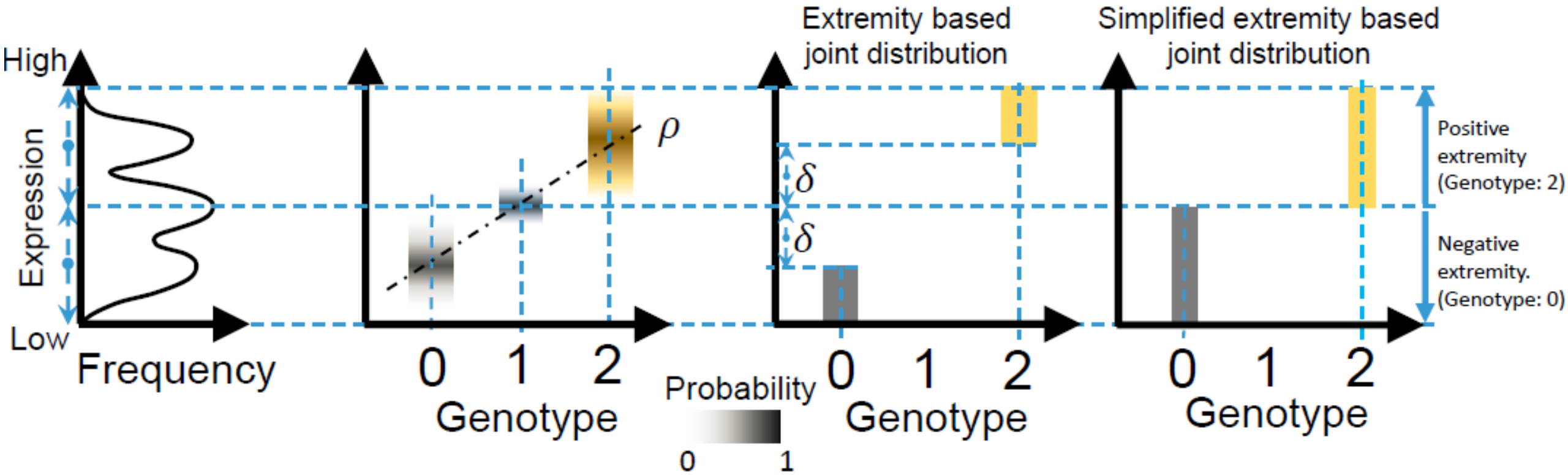
Steps in Instantiation of a (Mock) Linking Attack



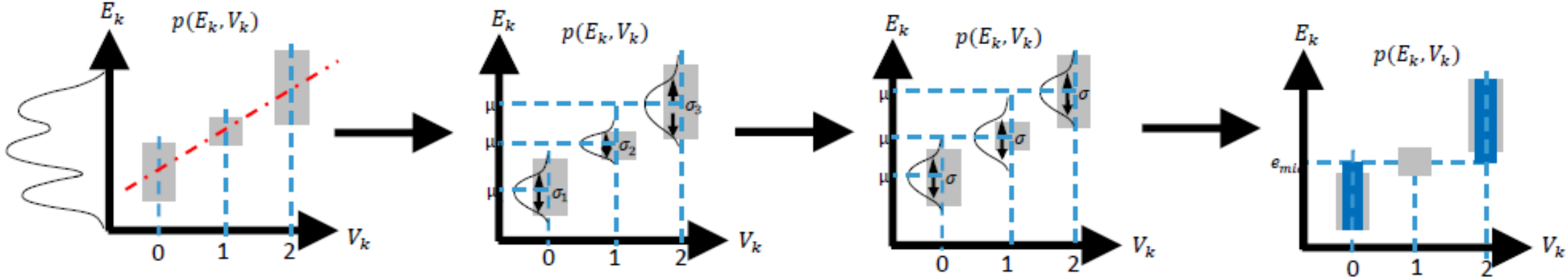


Levels of Expression-Genotype Model Simplifications:

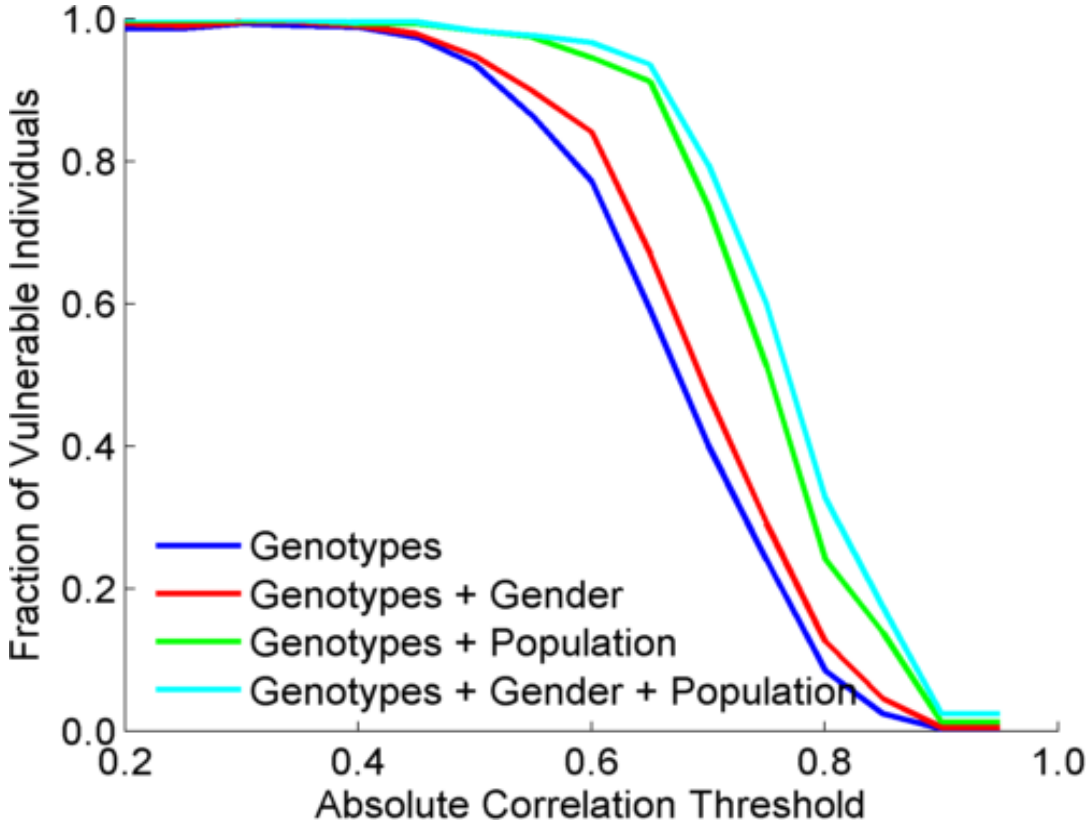




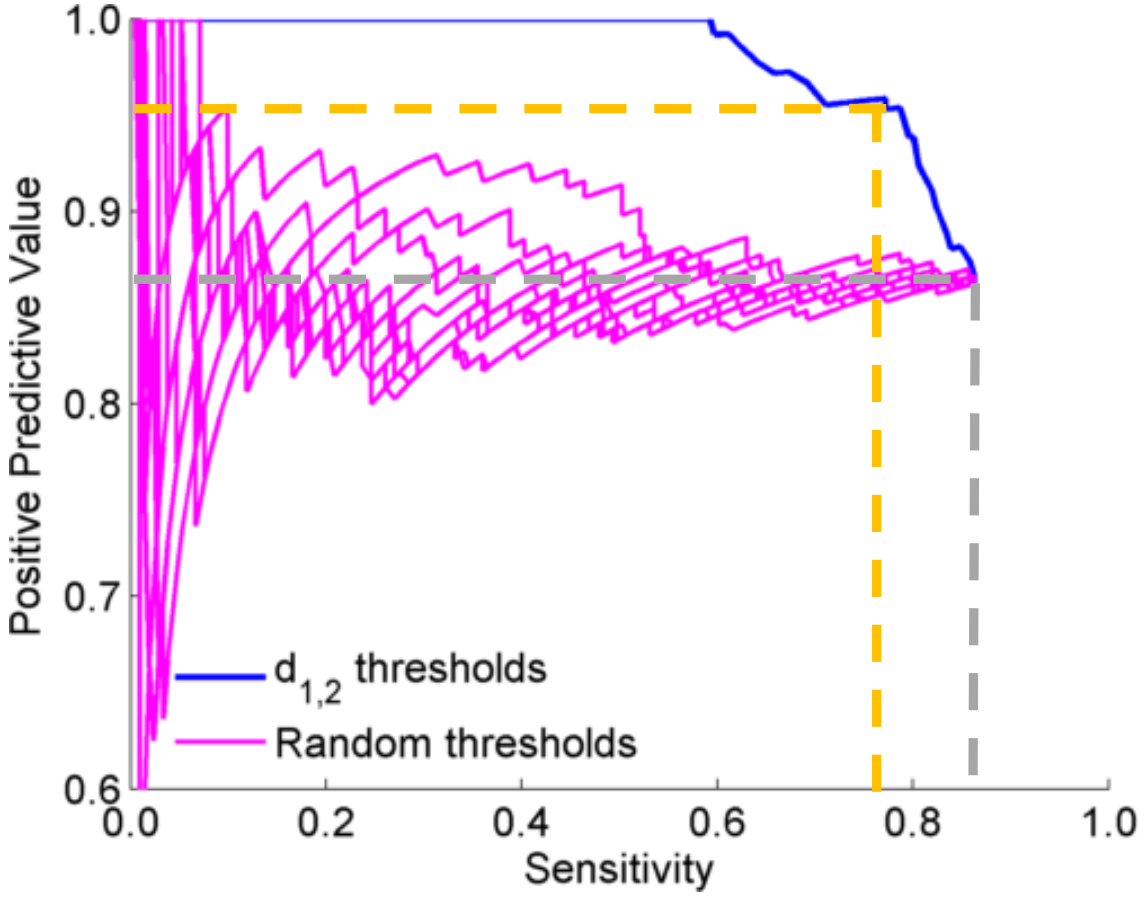
Levels of Expression-Genotype Model Simplifications:



Extremity based linking with homozygous genotypes



Attacker can estimate the reliability of linkings



Sensitivity: Fraction of individuals that are correctly linked

PPV: Fraction of selected individuals that are correctly linked

Risk Management Framework