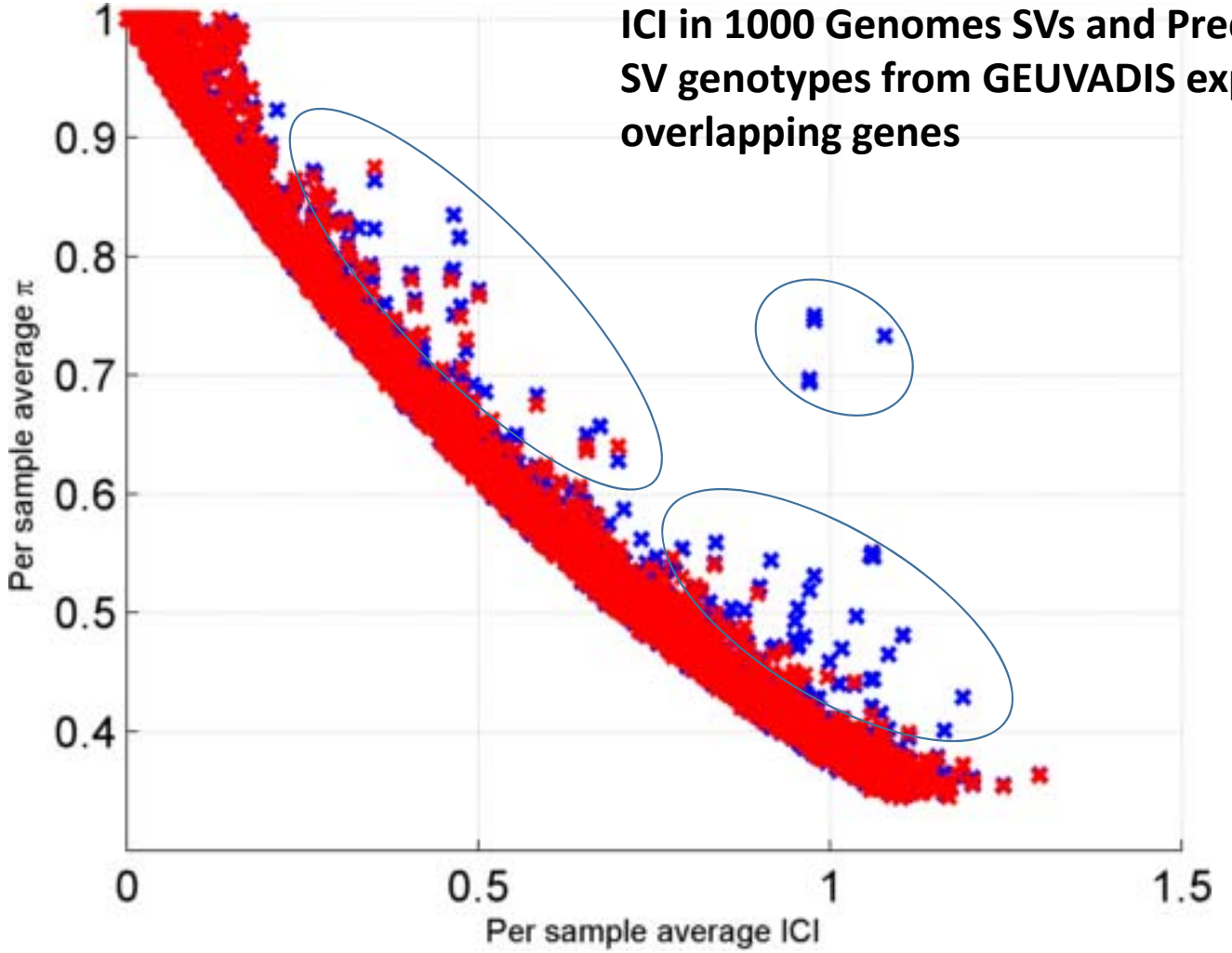

PrivaSeq2
Attacker Strikes Back

Focus on Rare Events

- Rare variants are valuable for identifying individuals, so attacker will want to use these
- Rare variant -> High ICI (☺ Happy attacker)
  - ***Example:*** A SNP genotype with <0.004 frequency has 6 bits of ICI. It can identify, on average, 1 individual among 256 individuals.
- Rare variant -> Low predictability (π) (☹ Sad attacker)
  - GWAS, eQTL studies are based on common variants and they were left out in PrivaSeq

π vs ICI for 1kG Structural Variants: Sample-wide

ICI in 1000 Genomes SVs and Predictability (π) of SV genotypes from GEUVADIS expression levels of overlapping genes



Focus on Rare Events

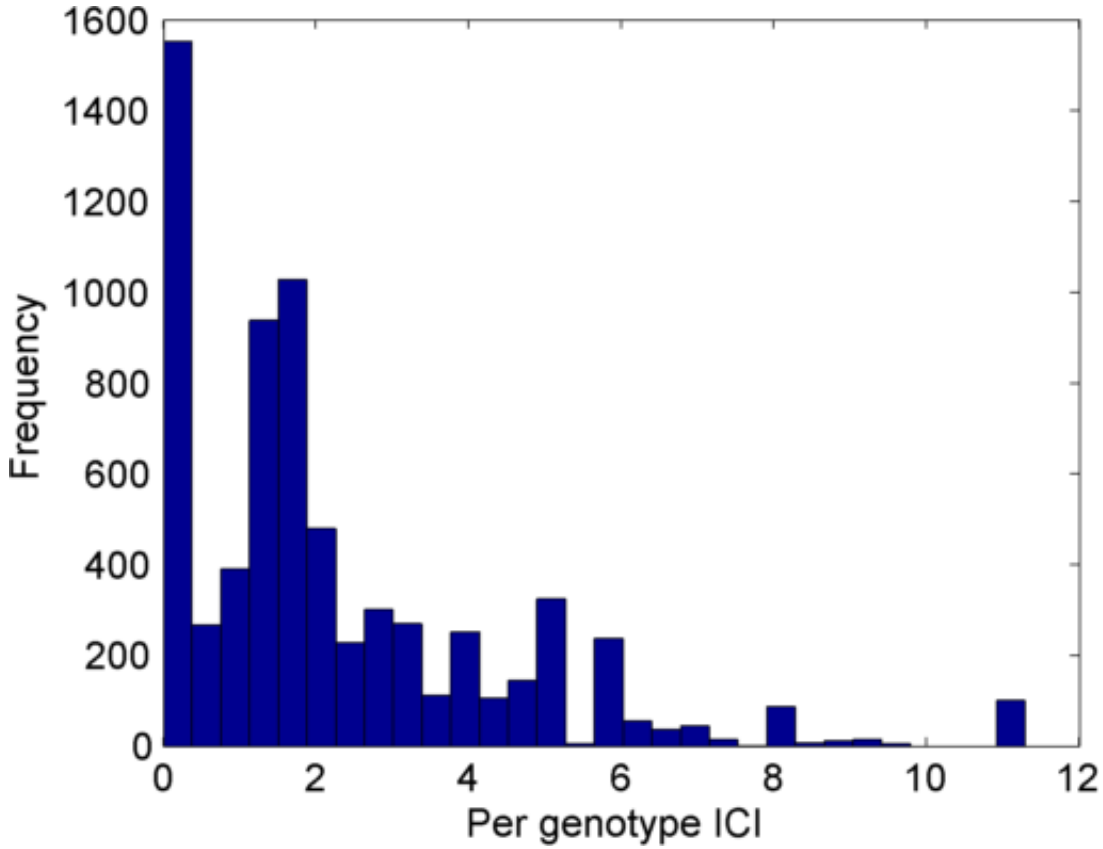
- To get around the low predictability, attacker focuses on two aspects:

1. Any variant can also be predicted in a genome-wide.
   - Up until now, the attacker used the population-wide predictability
     - Given the phenotypes for a sample of individuals, we predicted genotypes using extreme phenotypes
   - Focusing on one individual, given his/her genome-wide phenotype (RNA-seq signal, ChIP-Seq signal), can the attacker predict variants?
   - Will need to re-define π for the genomewide predictability: π_{GW}
   - More specifically, can we predict rare variants?
     - Maximal ICI
2. Attacker focuses on variants with high impact to ensure we do prediction correctly:
   - CNVs!

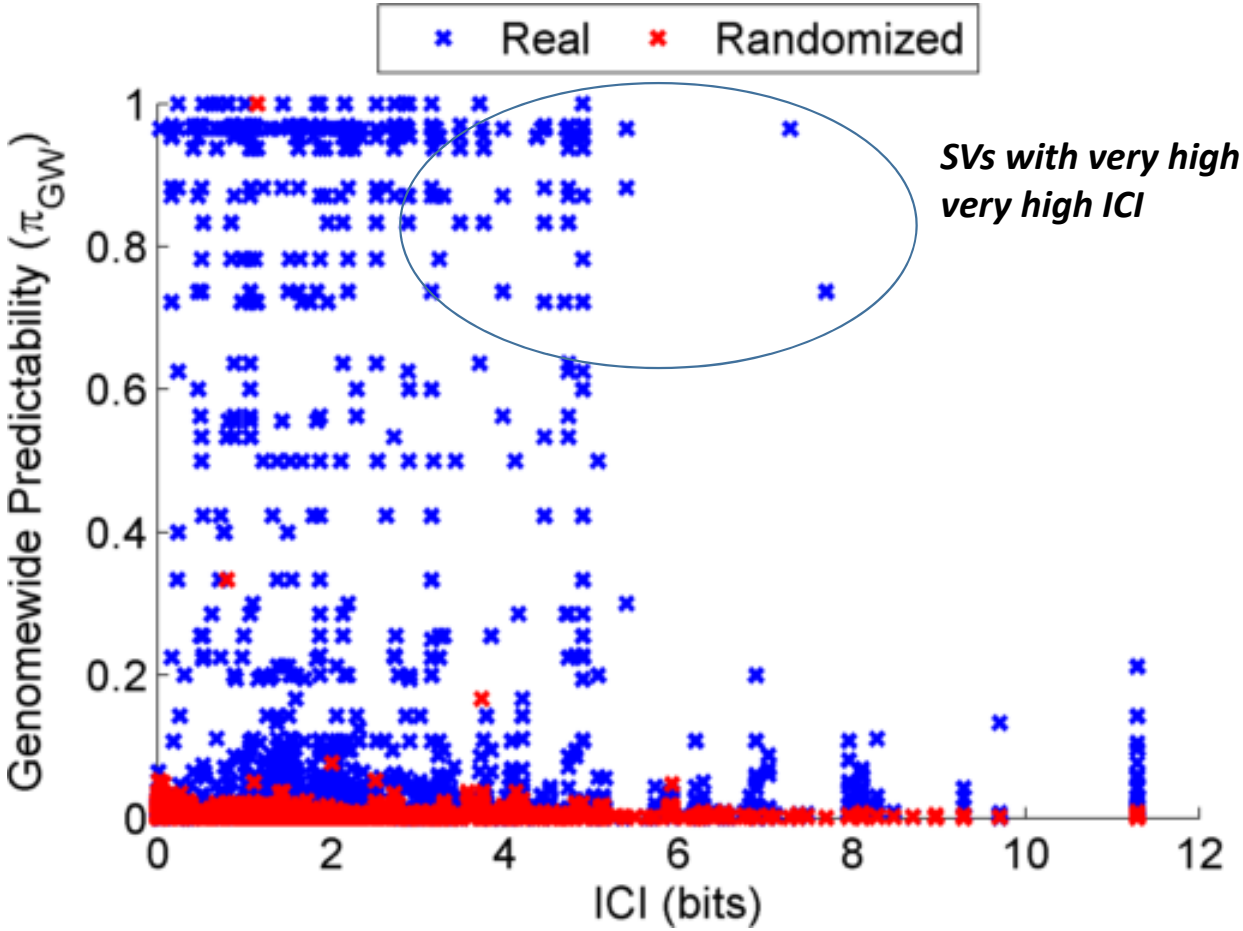
π_{GW}: Genomewide Predictability of a Genotype

- $\pi_{GW}(g_i = g, s_i = s, mapp_i = m) = P(g_i = g | s_i = s, mapp_i = m)$
  - g_i is the genotype RV for i^{th} variant and s_i is the expression RV for the overlapping gene and $mapp_i$ is the mappability RV of i^{th} variant
  - "Conditional probability of the genotype given signal level and mappability"
- Estimate the genomewide predictability for NA12878:
  - Divide the genome into 1000 bp windows
  - Pool H3K36me3, H3K27me3, H3K9me3, H3K79me2, and Control signal tracks for NA12878 from ENCODE2
  - Compute average signal and mappability in each window
  - Estimate π_{GW} for all the windows that are overlapping with SVs

Distribution of ICI for NA12878 SV genotypes



π_{GW} versus ICI for 1kG SVs: NA12878



SVs with very high GW predictability and very high ICI

Genome-wide Extremity Attack

- Attacker can adapt the extremity attack to exploit genomewide predictability of SVs, mainly the CNVs:
  - Any homozygous CNV will remove all the signal in the genomewide signal profile
- Sort the windows with respect to increasing signal levels
- Select a number of windows with smallest signal levels with good mappability
- Assign homozygous deletion to the CNV genotype of all the windows
- Compare and match the predicted CNV genotypes to the best matching 1000 Genomes individual (2504 individuals)

Genome-wide Extremity Attack for NA12878