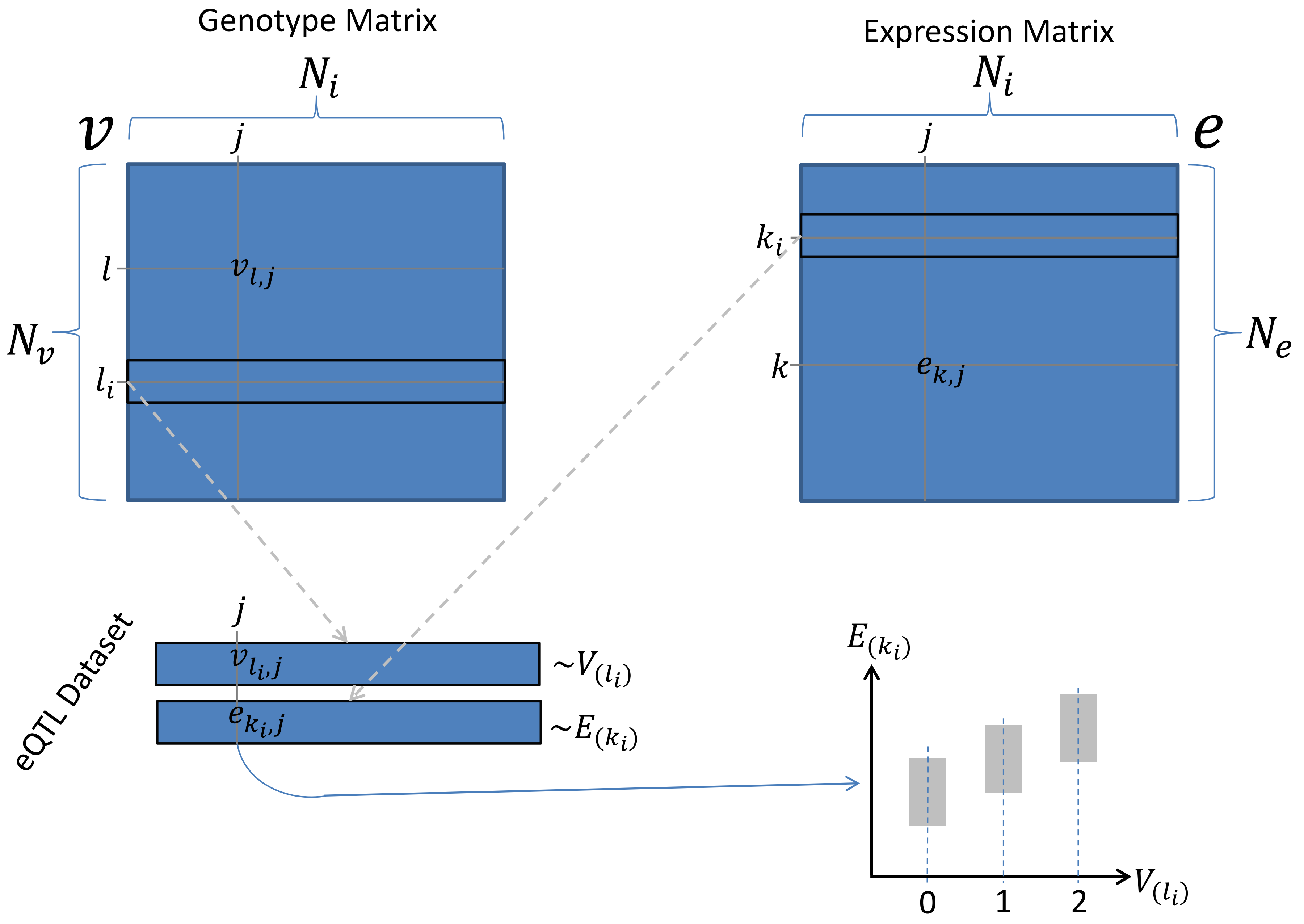
PrivaSeq Updates

January 26, 2015

Datasets [[Figure accompanies]]

- N_i individuals.
- Genotype data is stored in a $N_v \times N_i$ matrix denoted by v.
  - $v_{l,j}$ represents the genotype ($v_{l,j} \in \{0,1,2\}$) of l^{th} variant for j^{th} individual.
  - Let $\{v_{l,j}\}$ be a realization of the random variable $V_{(l)}$
- Expression data is stored in a $N_e \times N_i$ matrix denoted by e.
  - $e_{k,j}$ represents the expression of the k^{th} gene for j^{th} individual.
  - Let $\{e_{k,j}\}$ be a realization of the random variable $E_{(k)}$
- The eQTL dataset contains N_q eQTLs as a set of gene and variant RV pairs $\{(E_{(k_i)}, V_{(l_i)})\}, i < N_q, k_i < N_e, l_i < N_v$
  - There is significant correlation between $E_{(k_i)}$ and $V_{(l_i)}$
  - Correlation between $E_{(k_i)}$ and $V_{(l_i)}$: $\rho(E_{(k_i)}, V_{(l_i)})$.

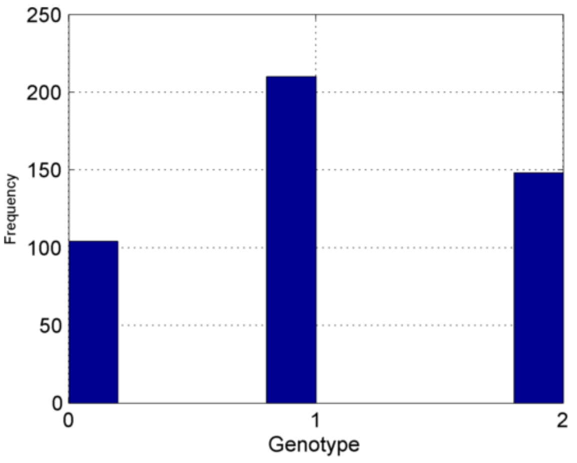
Datasets


Genotype Matrix
N_i
v
j
l
$v_{l,j}$
N_v
l_i
Expression Matrix
N_i
e
j
k_i
k
$e_{k,j}$
N_e
eQTL Dataset
j
$v_{l_i,j}$
$\sim V_{(l_i)}$
$e_{k_i,j}$
$\sim E_{(k_i)}$
$E_{(k_i)}$
0
1
2
$V_{(l_i)}$

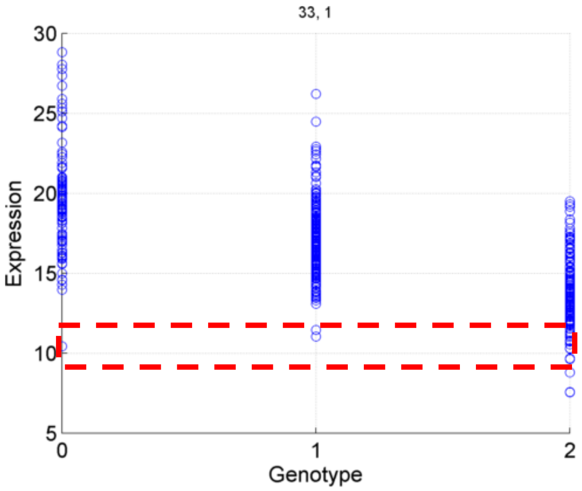
III Leakage: Attacker wants to identify an individual

- Main assumption: The attacker can predict the conditional probability distribution of the genotypes given the gene expression levels
  - $p\left(V_{(l_i)} \mid E_{(k_i)} = e\right)$

Prior: $p\left(V_{(l_i)}\right)$

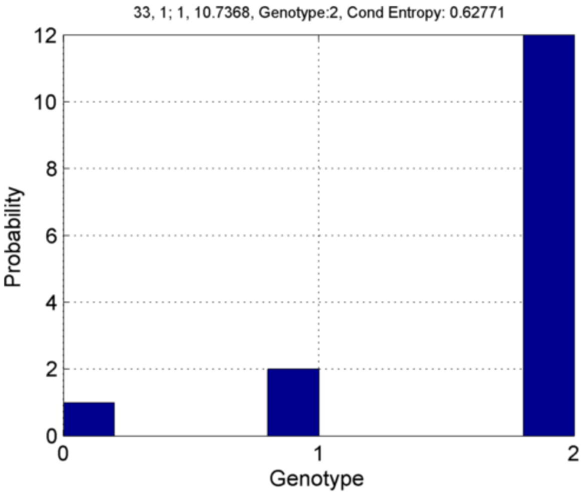


Joint: $p\left(V_{(l_i)}, E_{(k_i)}\right)$



Conditional (Posterior):

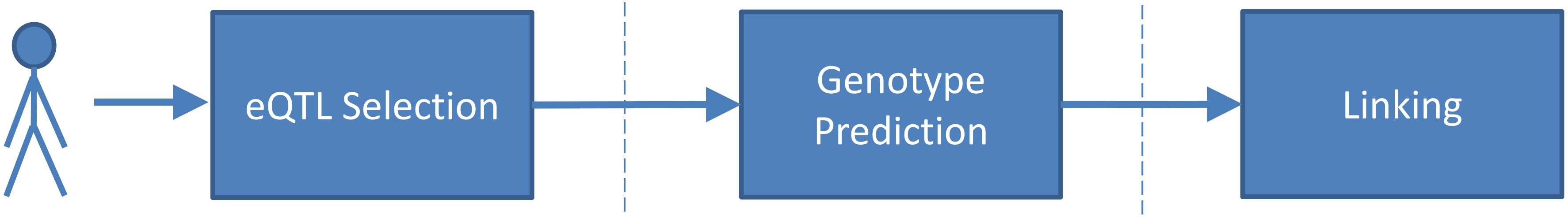
$p\left(V_{(l_i)} \mid E_{(k_i)} = 10.7\right)$



III Leakage: 3 Level Process

- The attacker uses the conditional probability distribution to perform estimation
- Thus, identification is a 3-level process:
  1. Selection of the eQTLs to be used
     - Entropy of the conditional distribution of genotypes
     - Allele Frequency
  2. The prediction of the genotypes for the selected eQTLs
     - Probabilistic:
       - Attacker assigns genotype randomly with respect to conditional distribution
     - Maximum *a posteriori* estimate:
       - Attacker assigns the genotype that has the highest value
  3. Linking of the Predicted Genotypes
     - Identify nearest genotype: Given the predicted genotypes, identify the individual that matches best to the predicted genotypes

III Leakage (linking attack): 3 Level Process



eQTL Selection

- Entropy of conditional distribution
- Allele Frequency

Genotype Prediction

- Maximum a-posteriori genotype
- Maximum a-priori genotype

Linking

- Individual with exact matching of the predicted genotypes
- Individual whose genotypes have the closest match to predicted genotypes

III Leakage: Step 1: SNP Selection

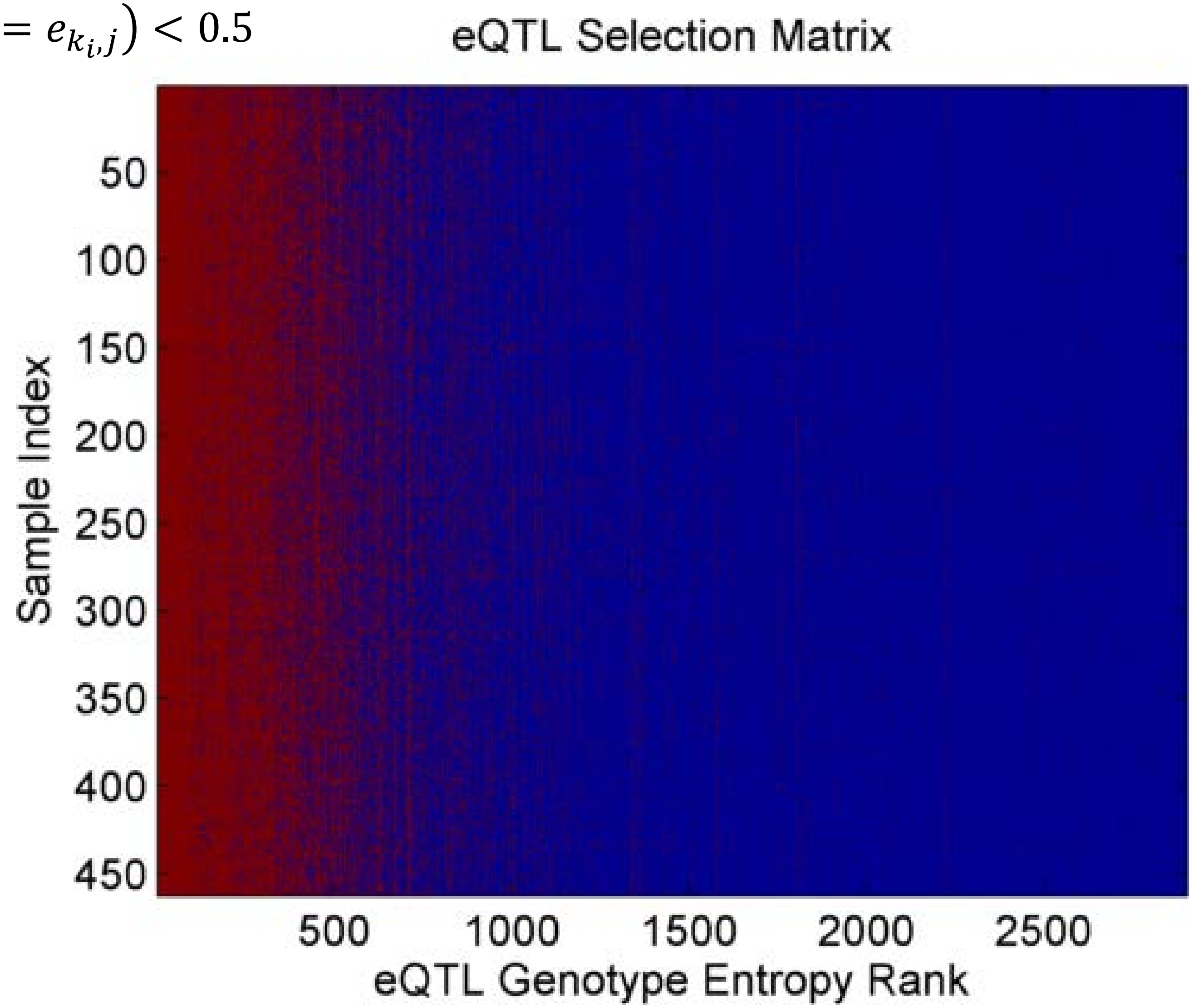
- Attacker goes over all the eQTLs and evaluates whether he will use the eQTL for prediction
- Different criteria:
  - Entropy of the conditional distribution
  - Allele frequency of the SNP
  - Strength of the reported association, i.e., absolute value of the correlation coefficient
- For an individual j, eQTL i is used in prediction if

$$H\left(V_{(l_i)} \mid E_{(k_i)} = e_{k_i,j}\right) < \gamma$$

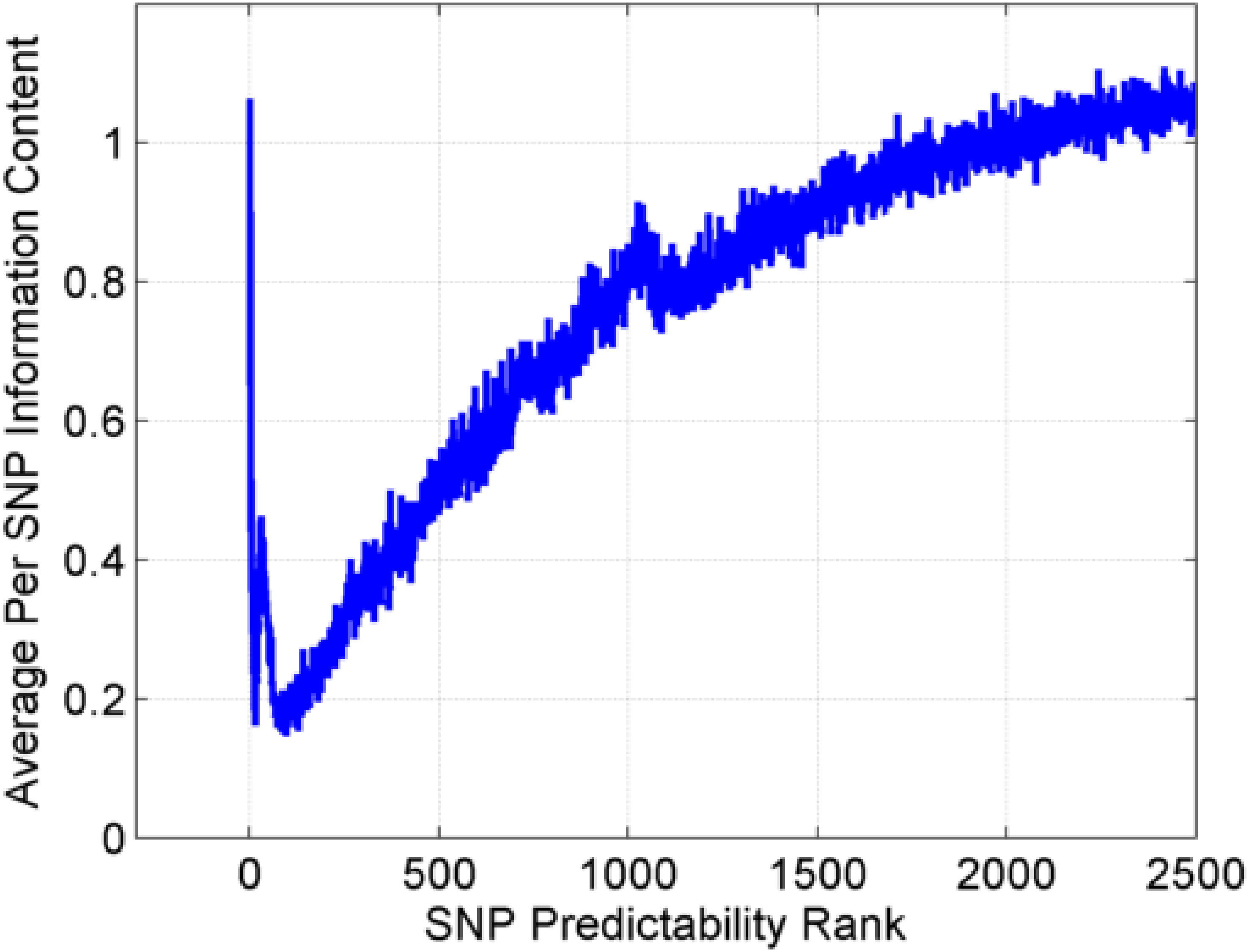
- Attacker gathers all the eQTLs that satisfy above for prediction

III Leakage: Step 1: SNP Selection

$H\left(V_{(l_i)} \mid E_{(k_i)} = e_{k_i,j}\right) < 0.5$



III Leakage: Step 1: SNP Selection

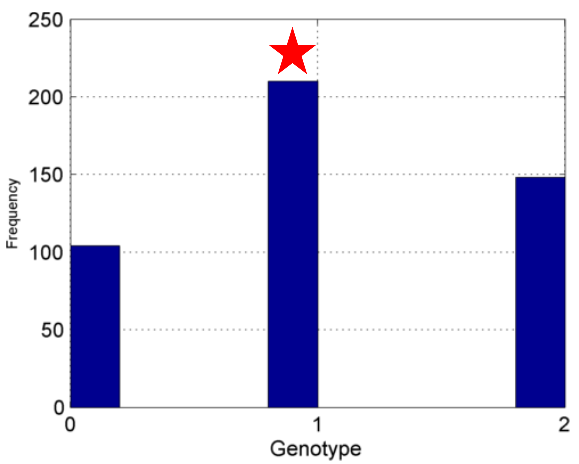


III Leakage: Step 2: MAP and max a-priori Genotype Prediction

- Select the genotype that has the highest a-posteriori (a-priori) probability

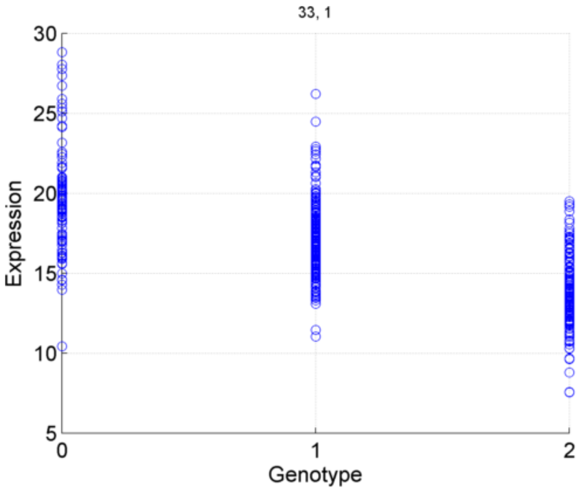
- Take j^th individual:

Prior: $p\left(V_{(l_i)}\right)$



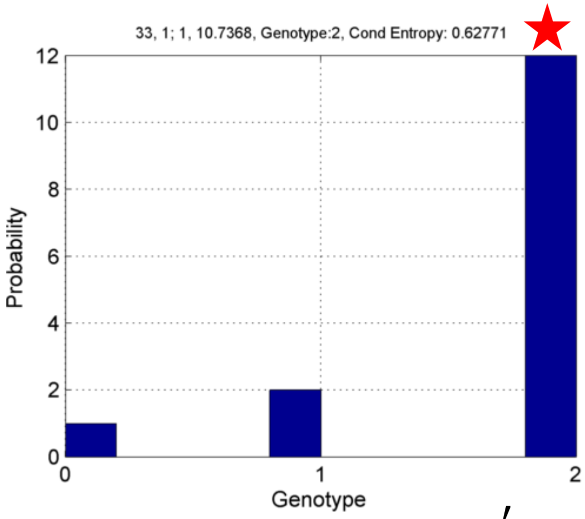
$v'_{l_i,j} = 1$

Joint: $p\left(V_{(l_i)}, E_{(k_i)}\right)$



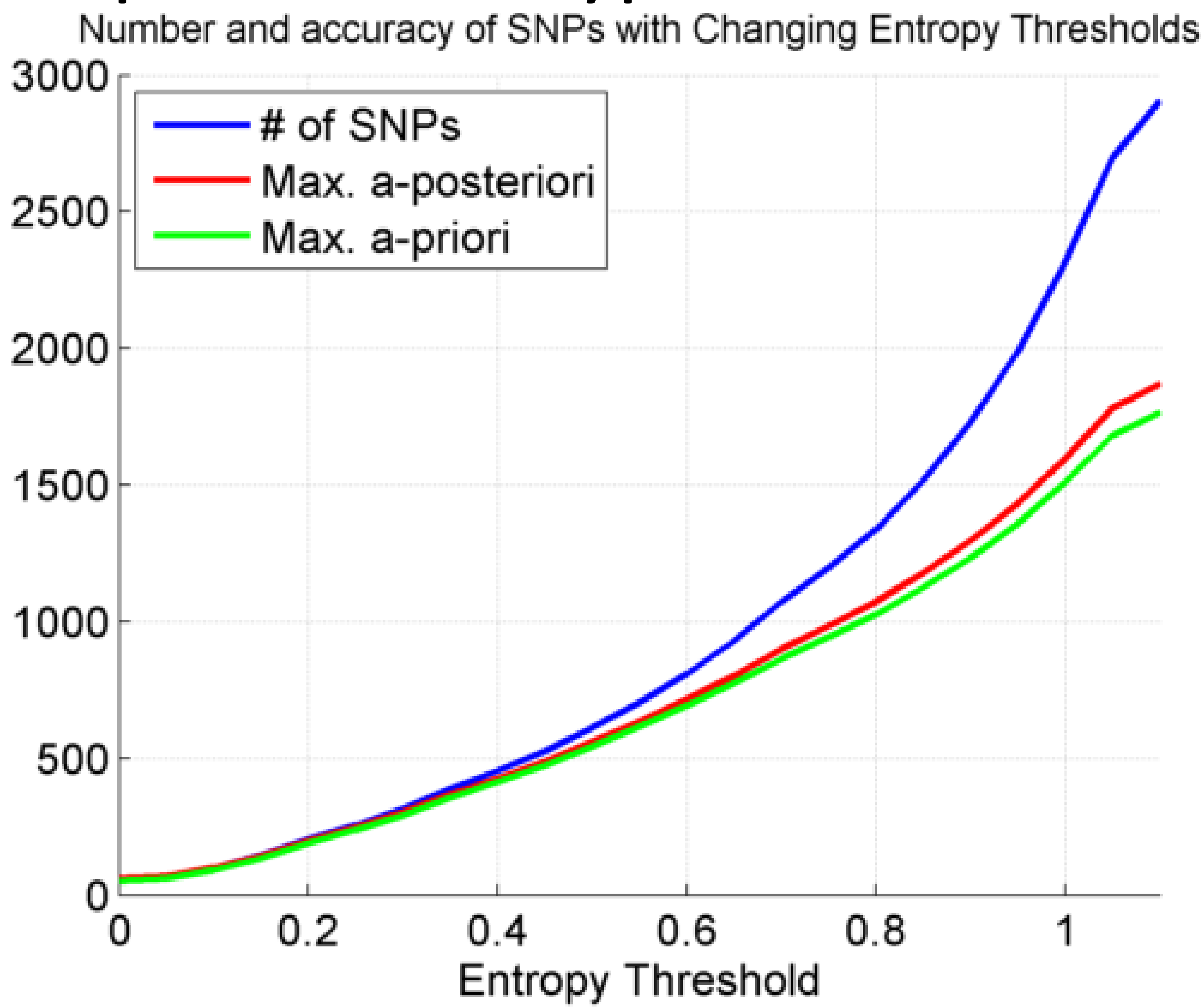
Conditional (Posterior):

$p\left(V_{(l_i)} \mid E_{(k_i)} = e_{k_i,j=}10.7\right)$

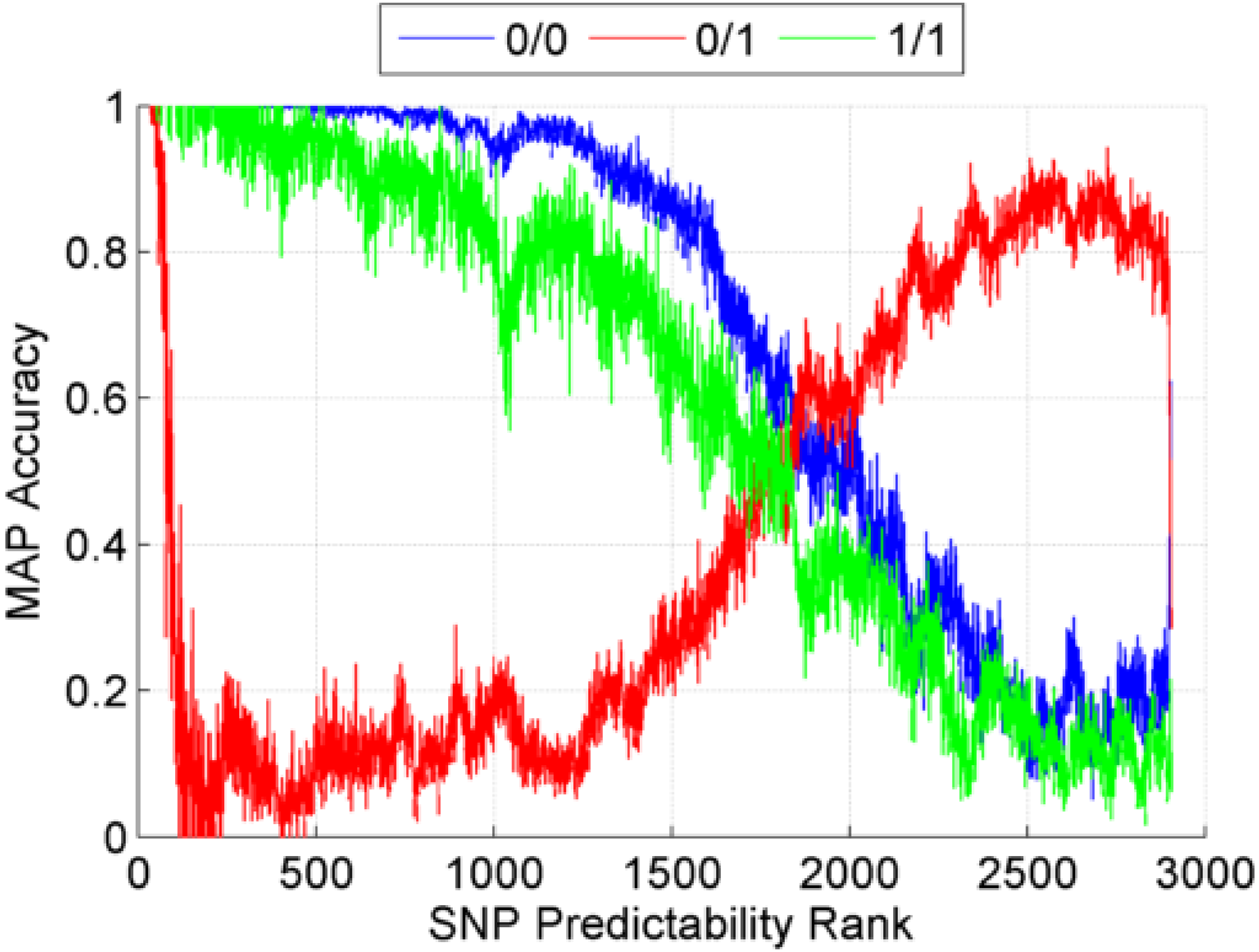


$v'_{l_i,j} = 2$

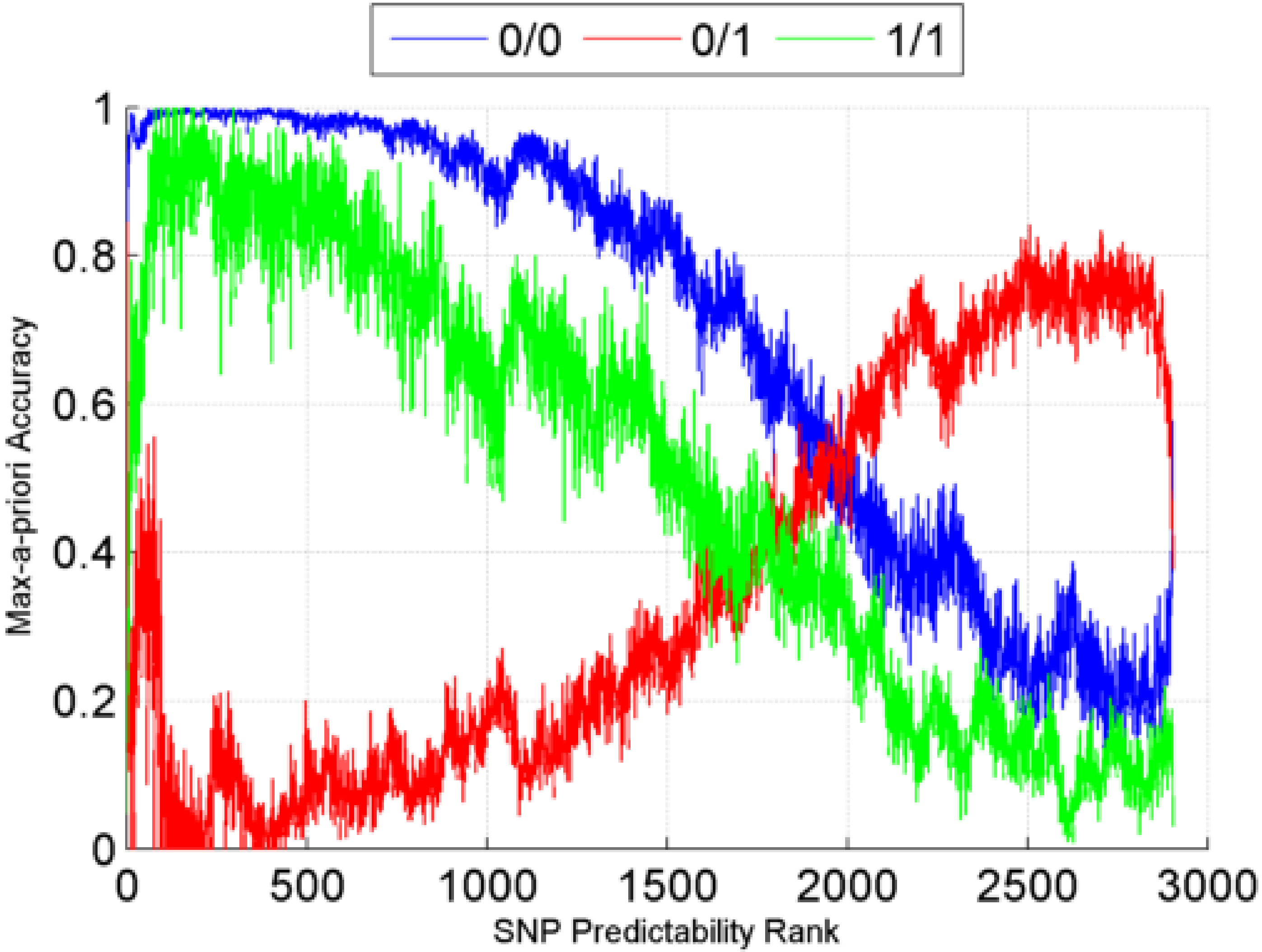
III Leakage: Step 2: MAP and max a-priori Genotype Prediction



III Leakage: Step 2: MAP and max a-priori Genotype Prediction



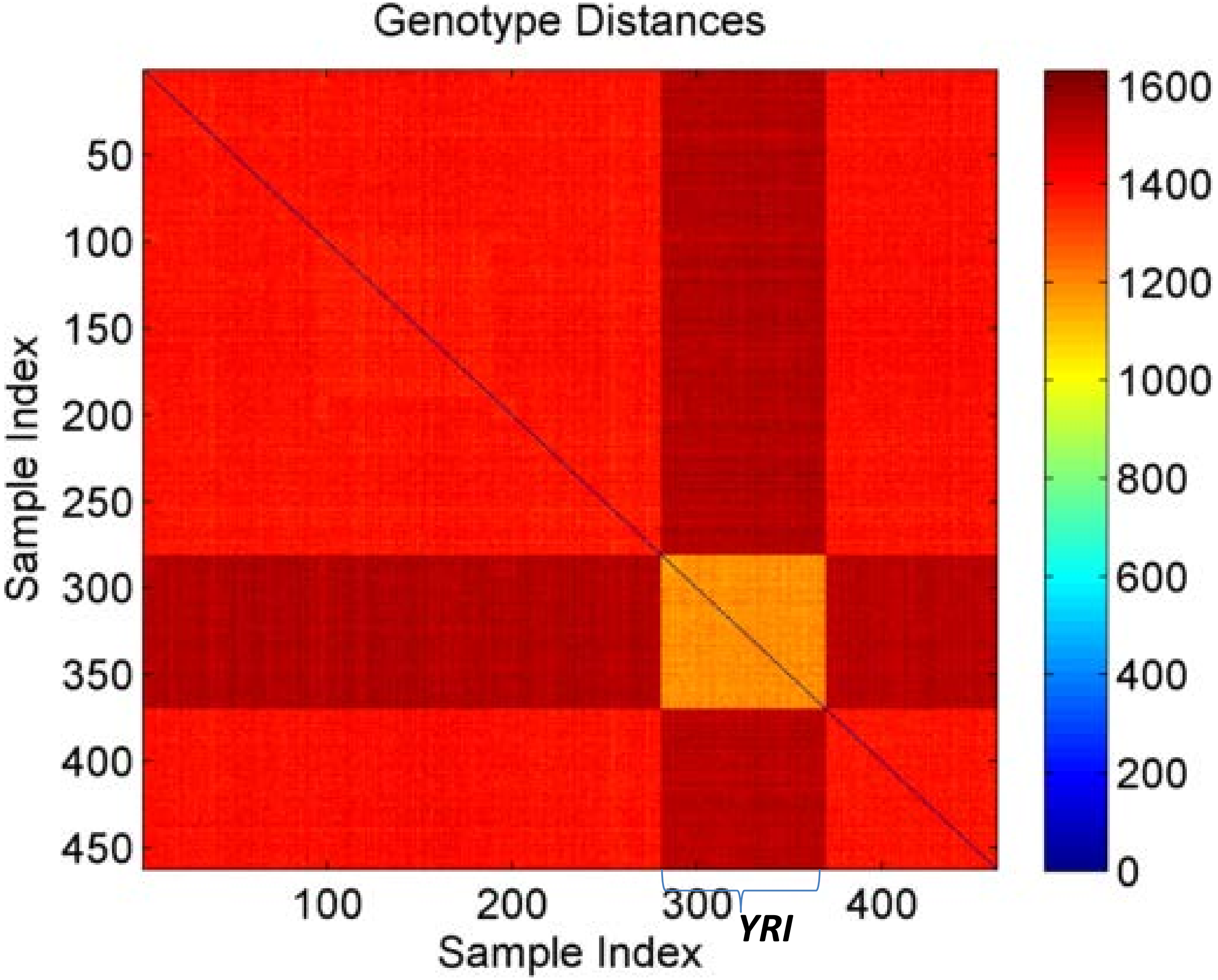
III Leakage: Step 2: MAP and max a-priori Genotype Prediction



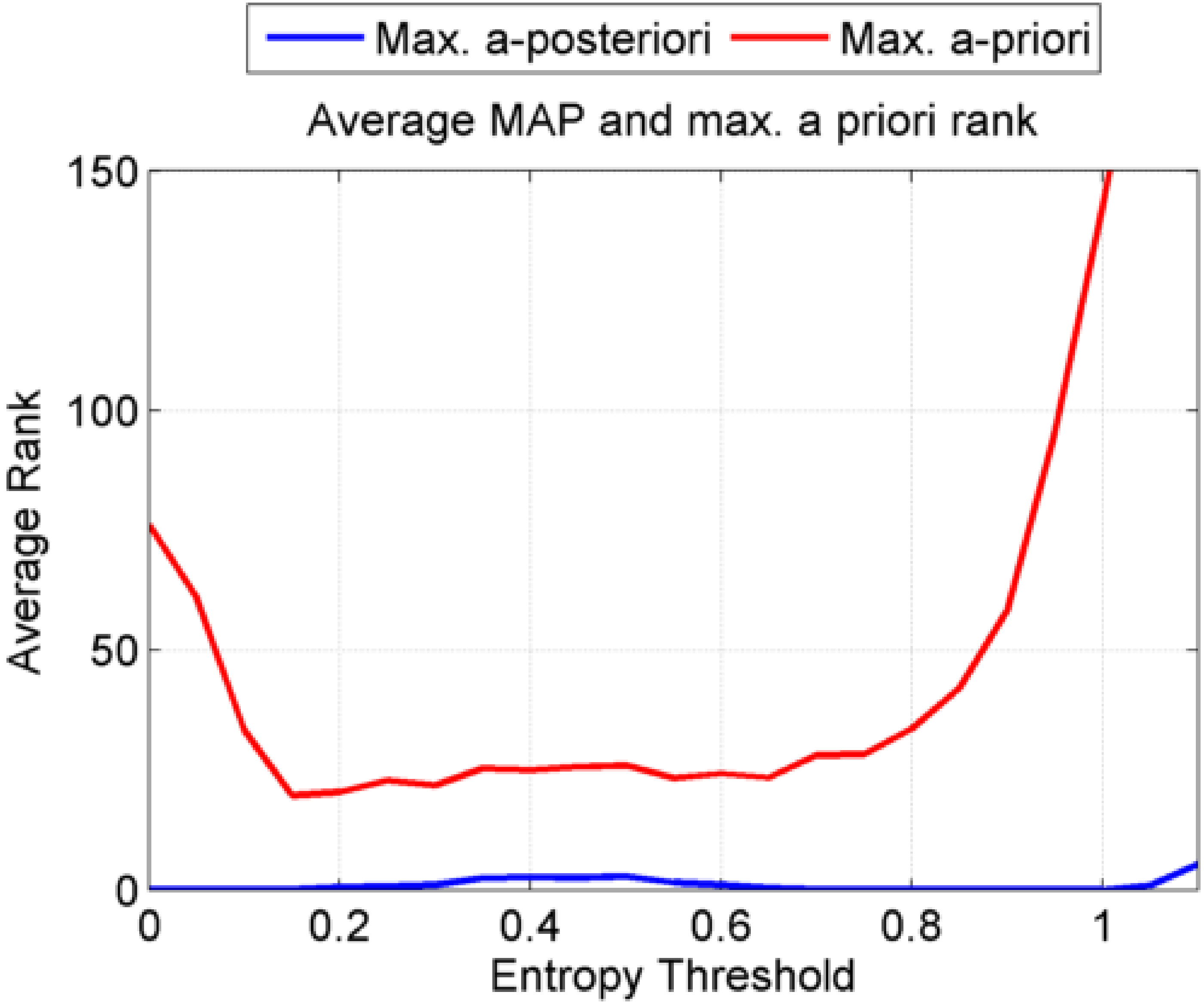
III Leakage: Step 3: Linking

- Using the predicted genotypes, select the genotype to match the predicted individual to.
    - Perfect matching: Attacker tries to match all the predicted genotypes
    - Nearest neighbor matching: Attacker identifies the individual with genotypes that matches closest to the predicted genotypes
        - Given set of predicted genotypes for individual j, $\{v'_{l,j}\}$;
        - $pred_j = \underset{a}{\operatorname{argmax}}\{\sum_b I(v'_{b,j}, v_{b,a})\}$
        - If $pred_j =$ j; j is vulnerable

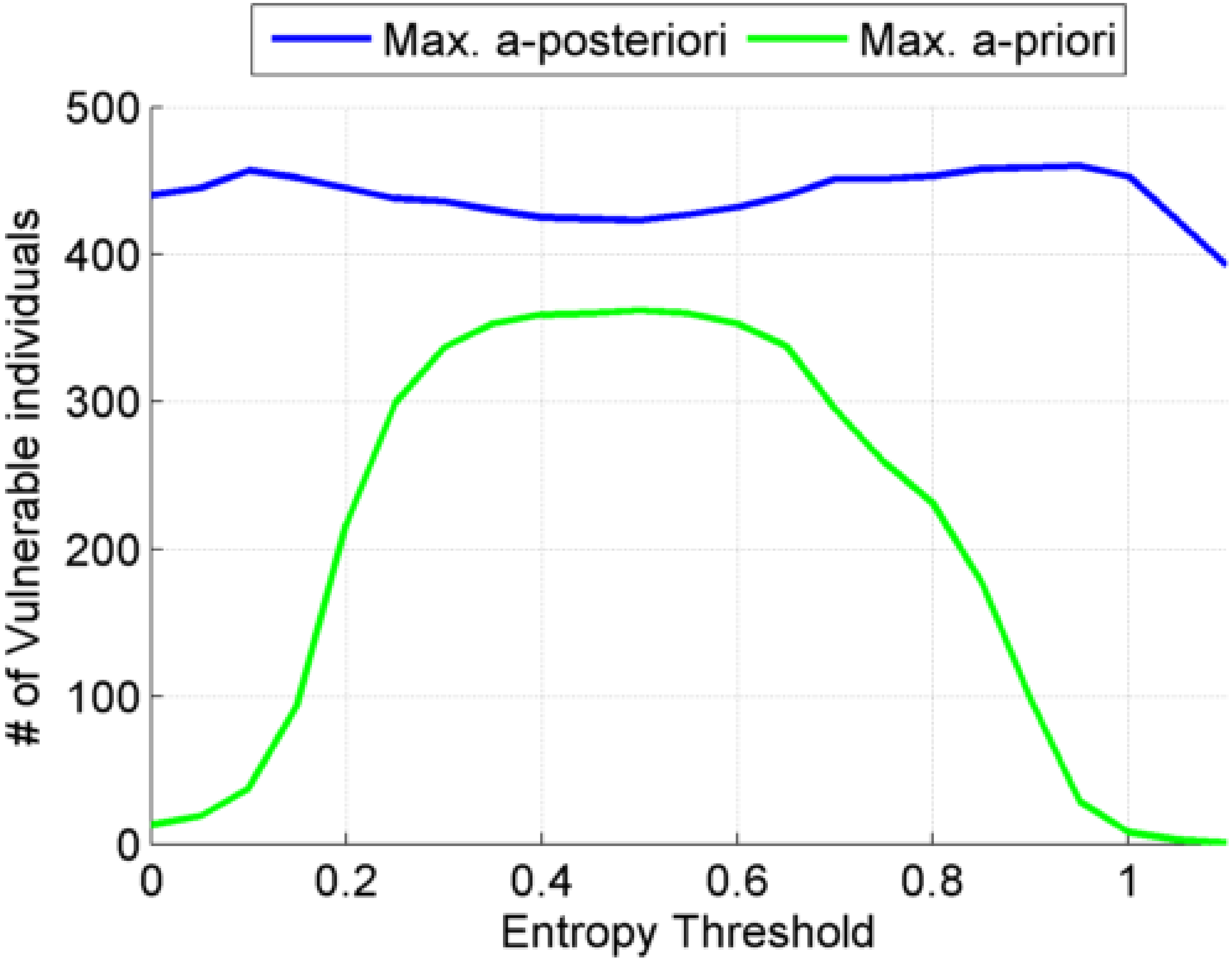
III Leakage: Step 3: Linking



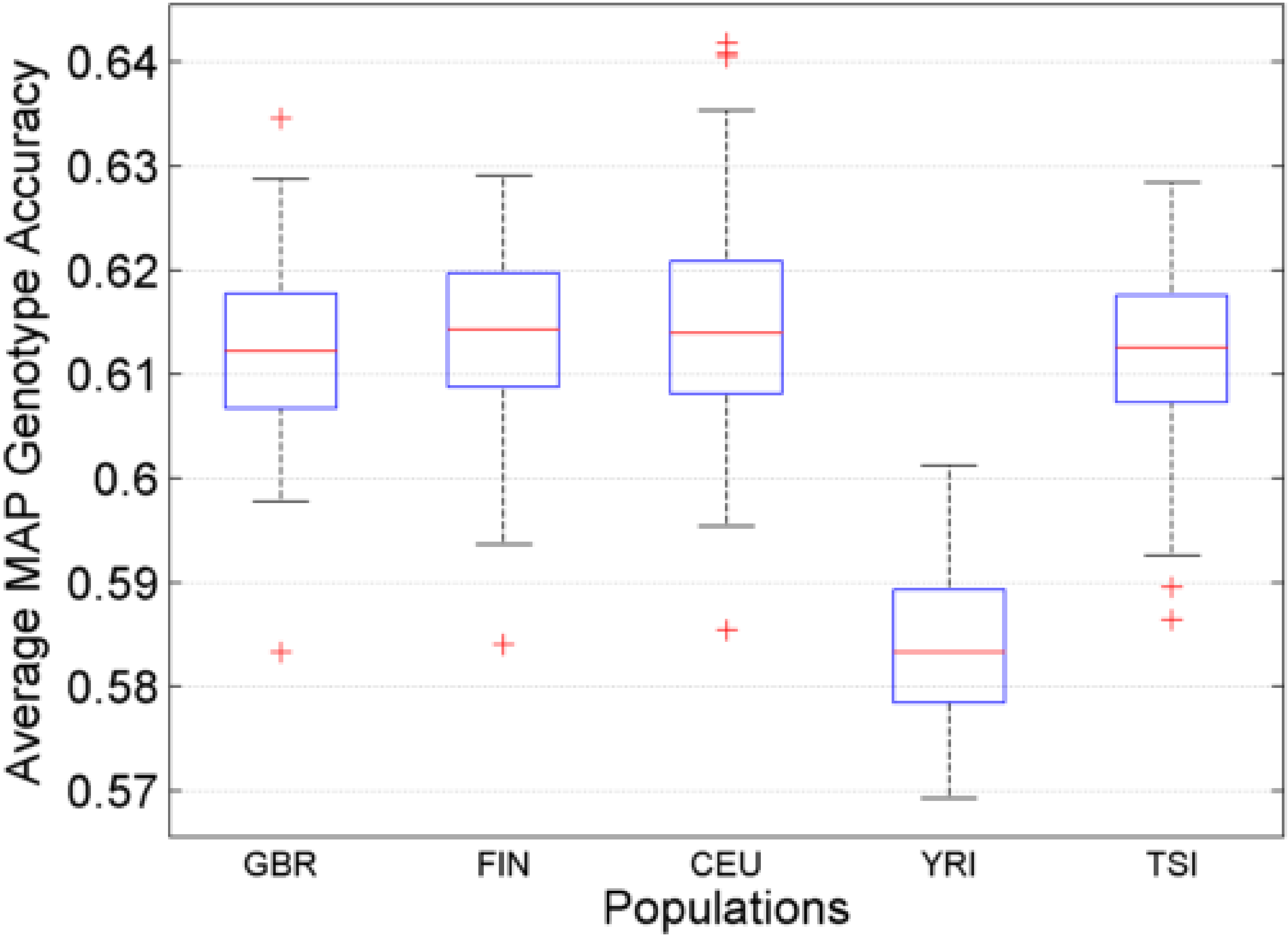
III Leakage: Step 3: Linking



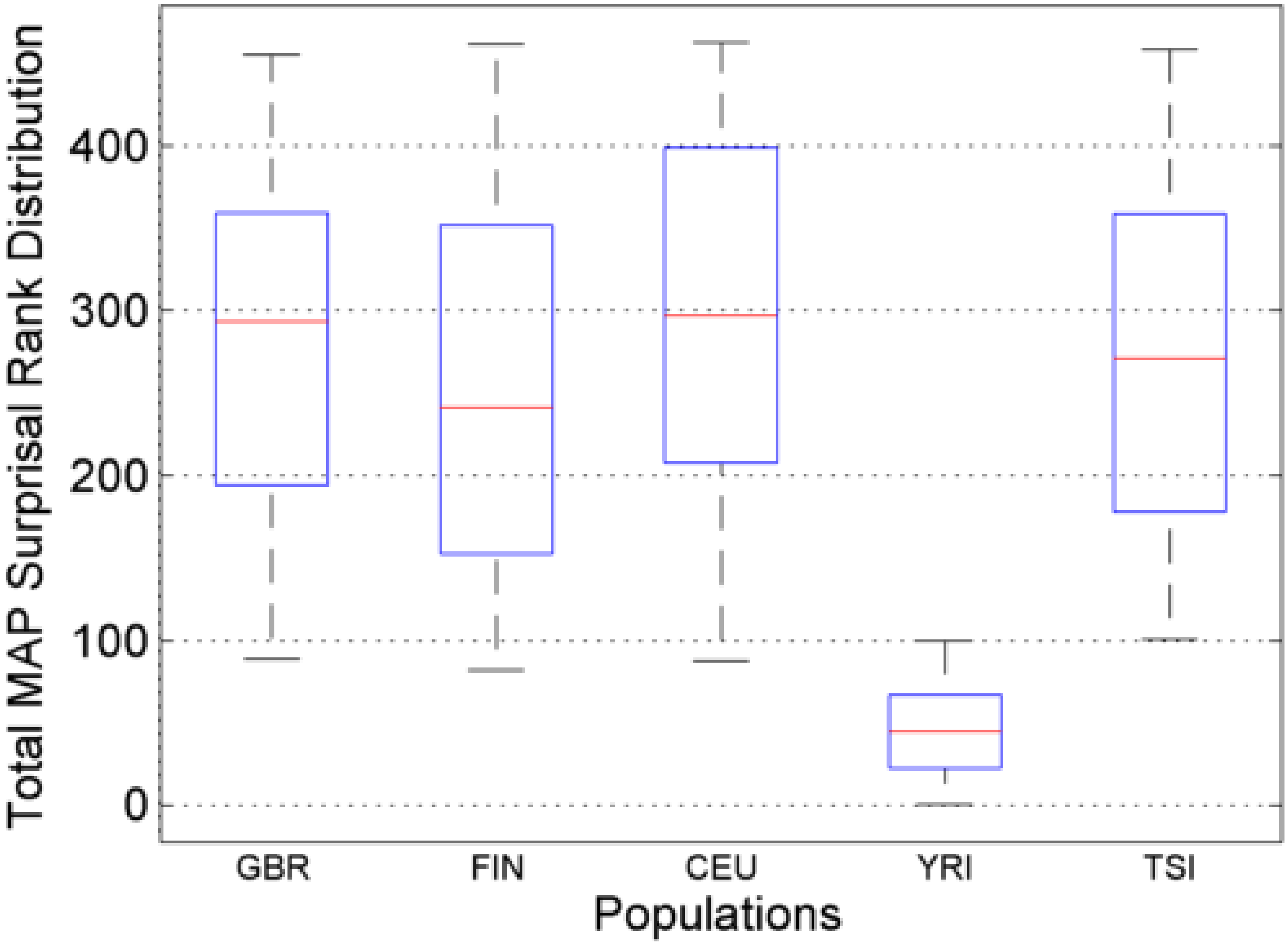
III Leakage: Step 3: Linking



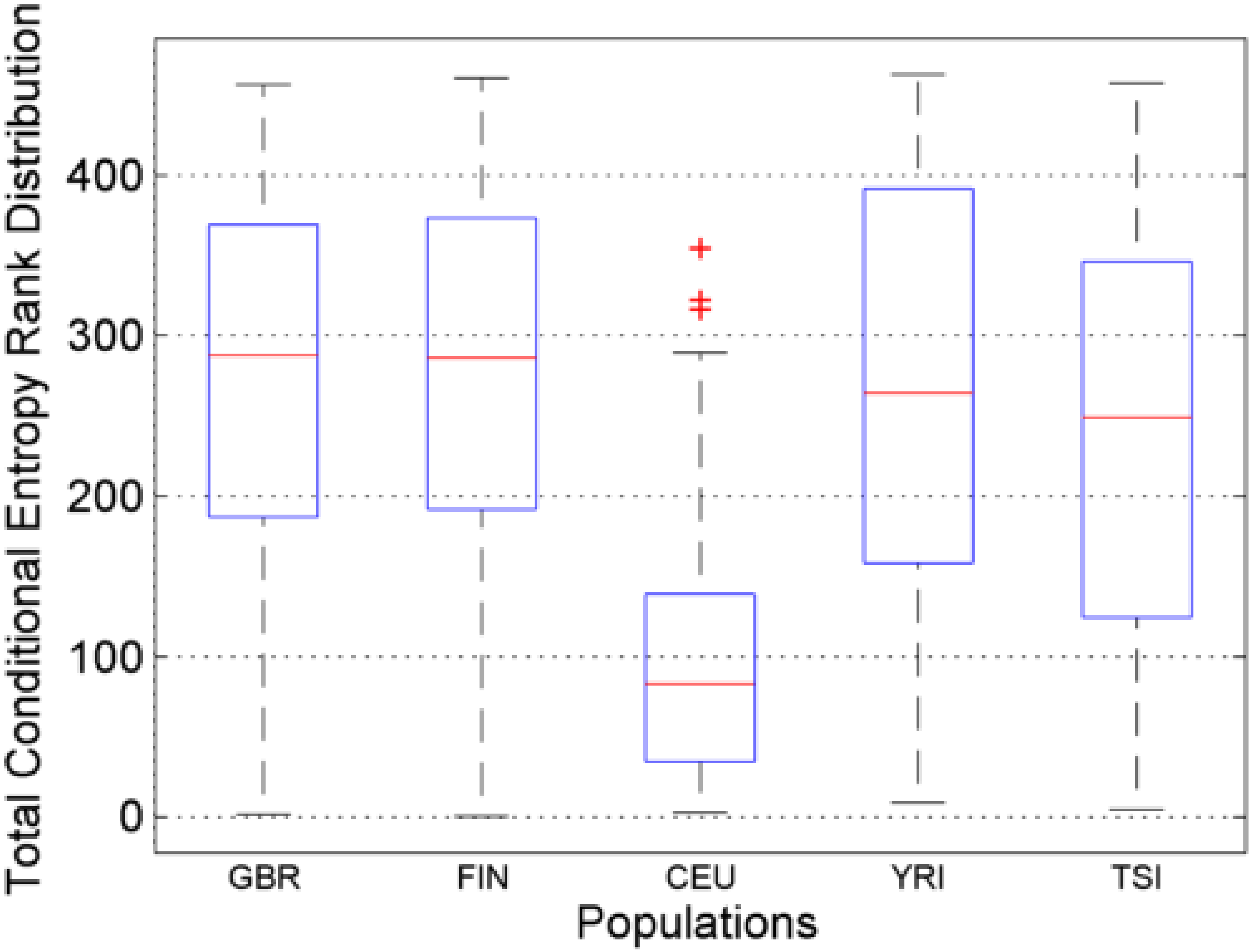
III Leakage: Step 3: Linking: Average MAP Genotype Accuracy



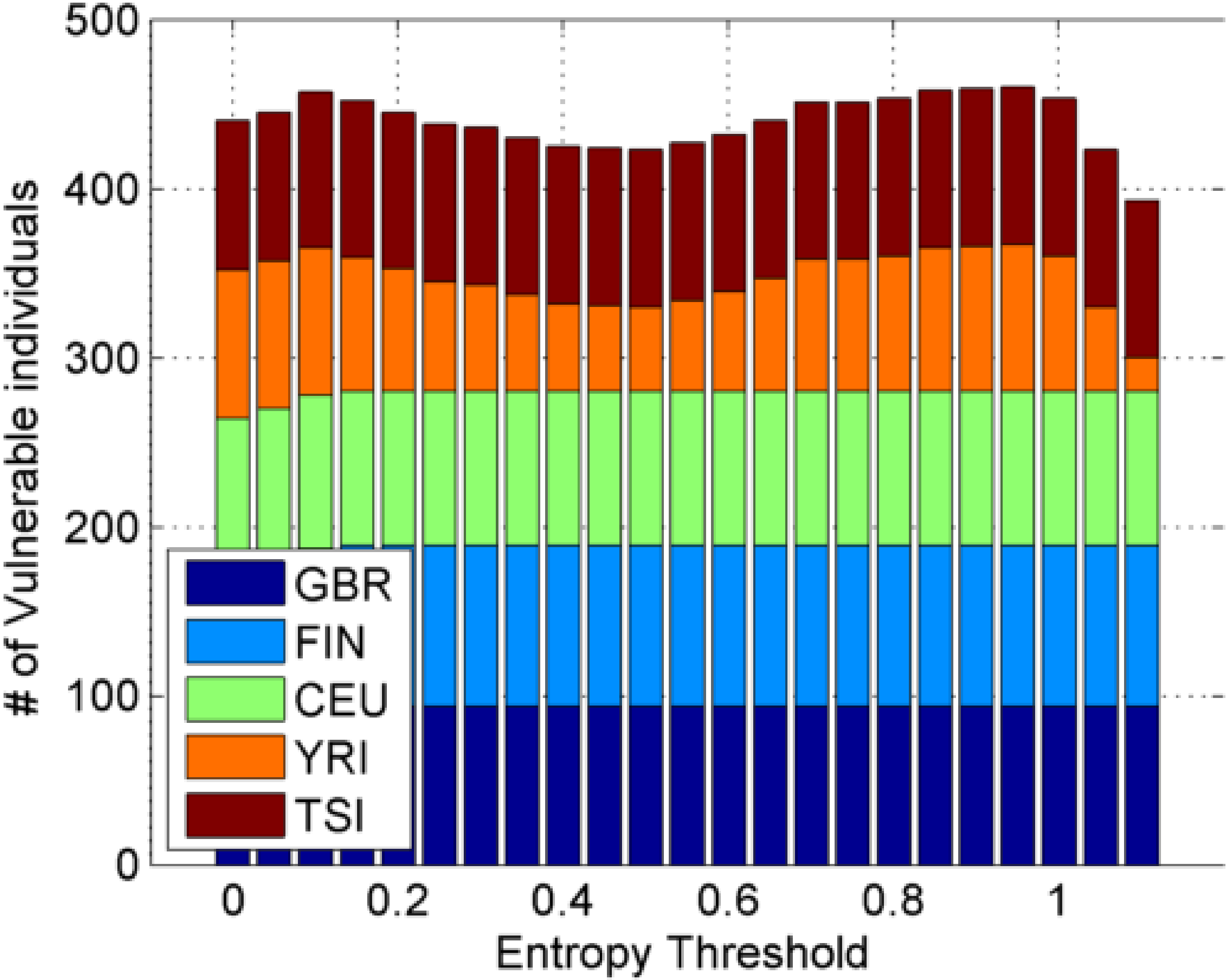
III Leakage: Step 3: Linking: Leaking III Rank



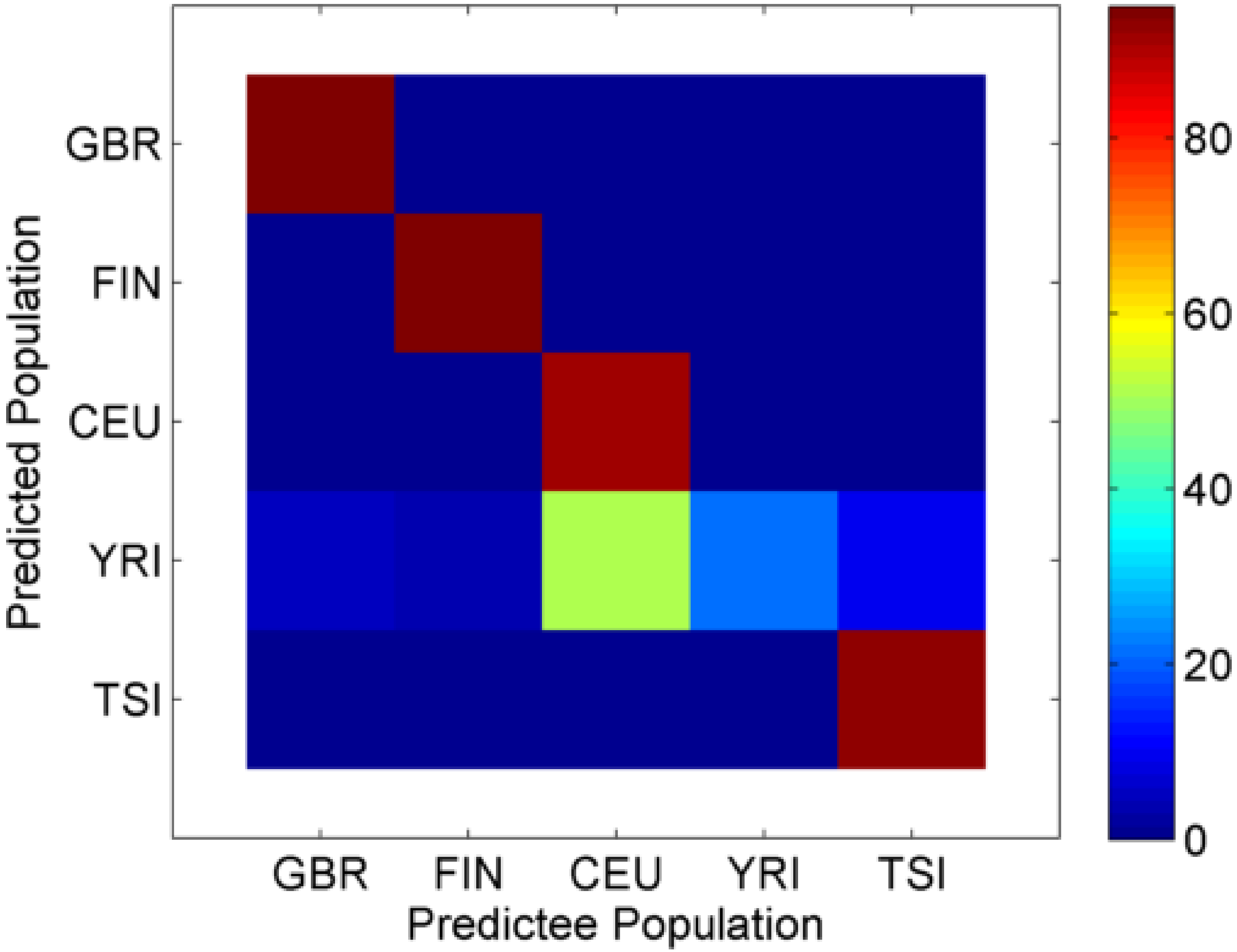
III Leakage: Step 3: Linking: Estimated Predictability



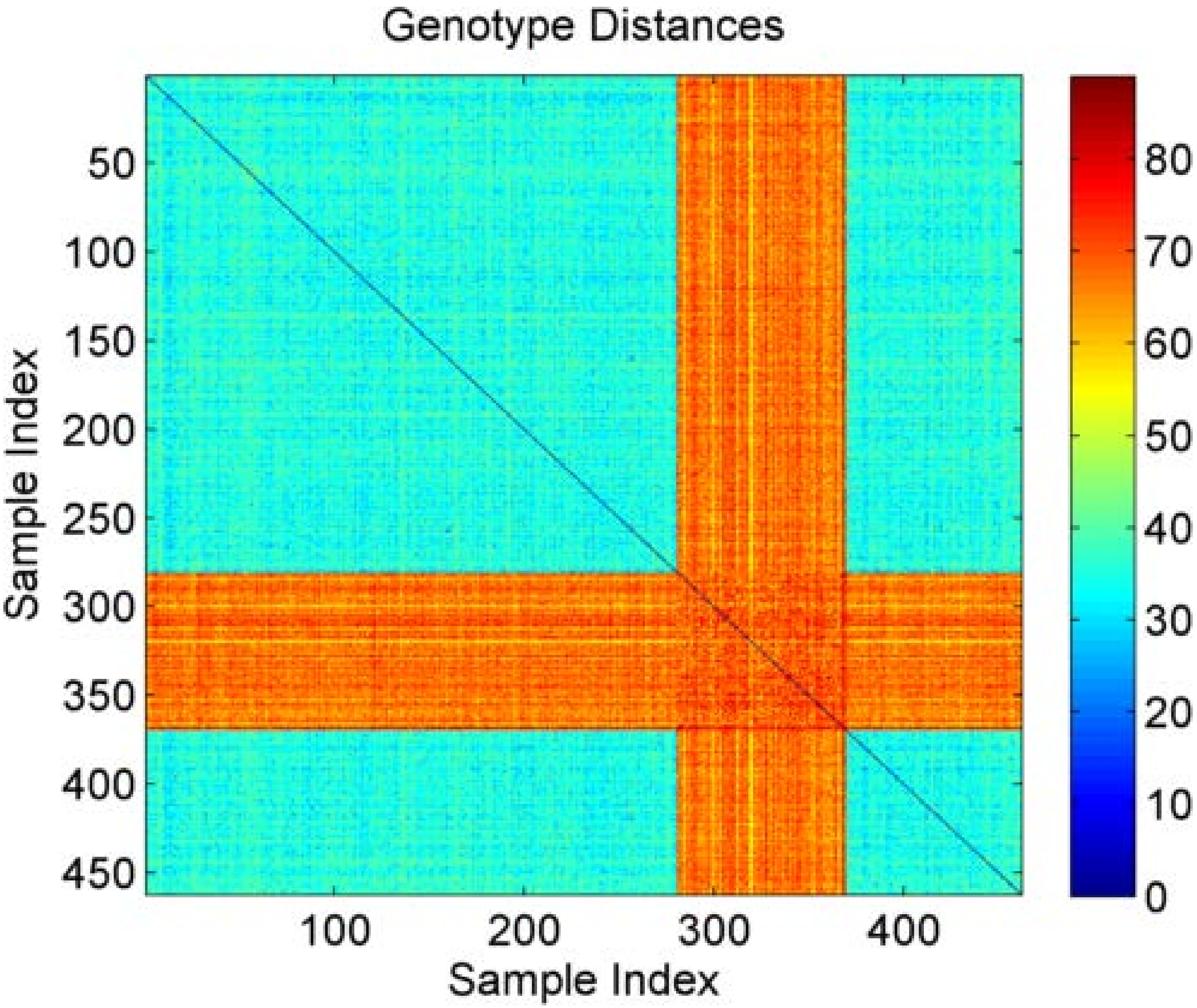
III Leakage: Step 3: Linking



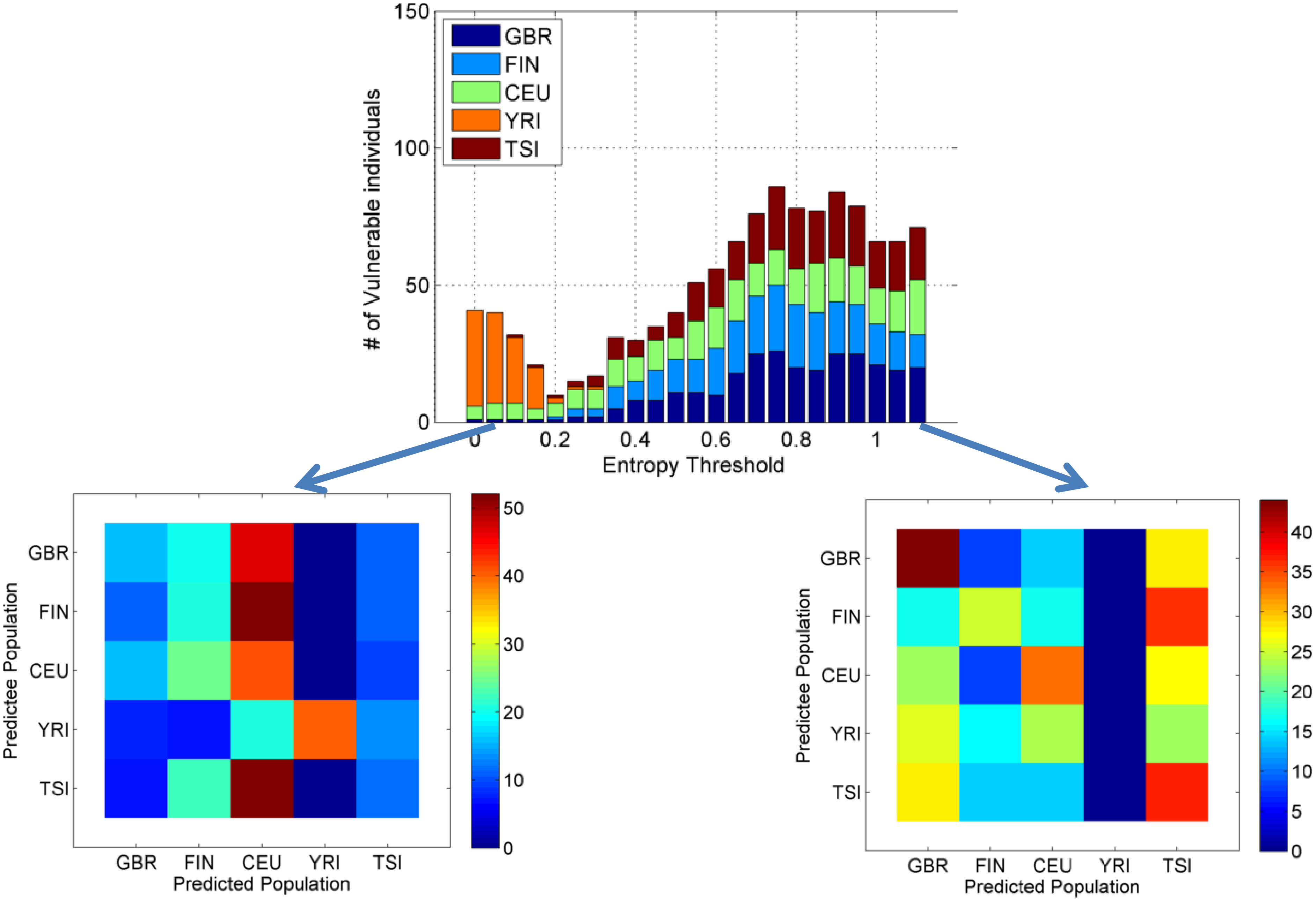
III Leakage: Step 3: Linking: Population Confusion



III Leakage: Step 3: Linking (YRI eQTLs)



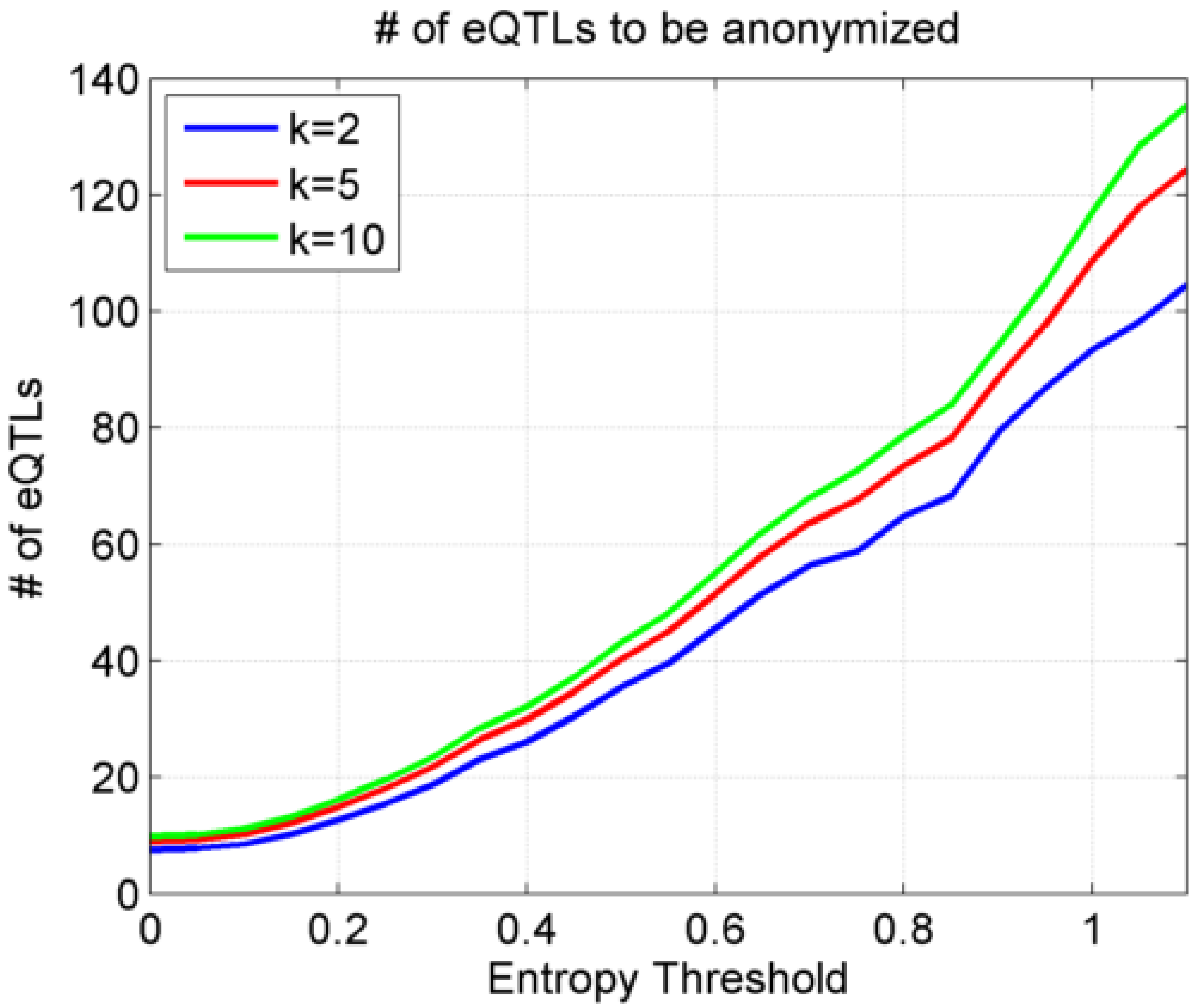
III Leakage: Step 3: Linking (YRI Only eQTLs)



Anonymization

- # of associations to anonymize depends on the number of associations between the top matching individual and second top matching individual.

Anonymization


# of eQTLs to be anonymized
k=2
k=5
k=10
# of eQTLs
Entropy Threshold