Learning Your Identity and Disease from Research Papers: Information Leaks in Genome Wide Association Study

Rui Wang, Yong Li, XiaoFeng Wang, Haixu Tang, Xiaoyong Zhou

Presentation Overview

- Brief Introduction: Genomes, SNP, GWAS
- Privacy Implications of GWAS
- Authors' Attacks
- Defense
- Implementation
- Conclusion

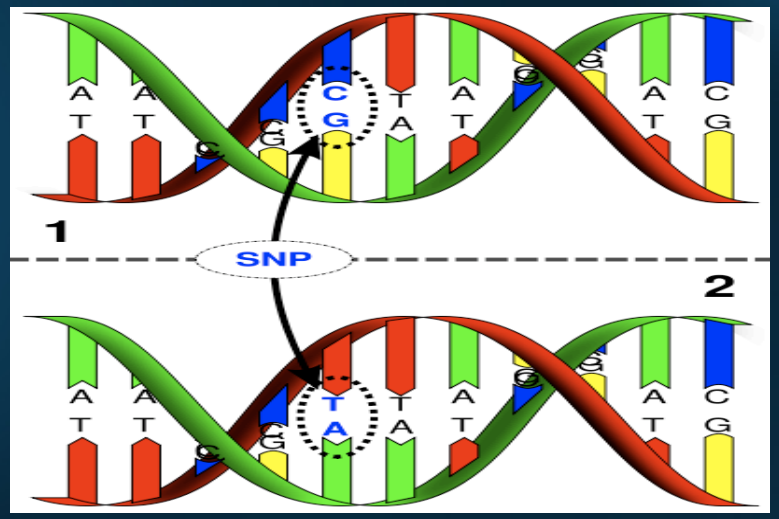
Genome

- Complete set of genes in a single organism
- Entirety of an organism's hereditary information
- Human Genome Project (HGP)[1] produced a reference sequence of the human genome

[1] http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml

Single Nucleotide Polymorphisms (SNP)

- DNA sequence variations that occur when a single nucleotide (A,T,C,or G) in the genome sequence is altered[1]



[1] http://www.ornl.org/sci/techresources/Human_Genome/faq/snps.shtml
[2] Picture: http://science.marshall.edu/murraye/341/Images/416px-Dna-SNP_svg.png

Single Nucleotide Polymorphisms (SNP)

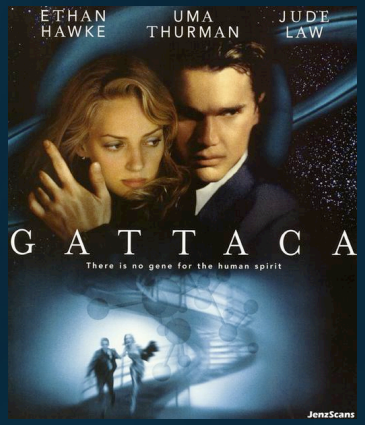
- Variation must occur in 1% population to be considered a SNP
- SNP contains a *major allele* (0) and a *minor allele* (1)
- Large amount of information
  - Individual frequency (1 or 0)
  - or SNP pairs of allele (00, 01, 10, 11)

Genome-wide Association Studies (GWAS)

- GWAS developed to leverage genome data to discover:
  - Genetic variations (SNPs)
  - Common diseases
- Data widely available
  - HapMap (http://hapmap.ncbi.nlm.nih.gov/)
- Individuals' disease susceptibility

Privacy Implications for GWAS DBs

- Privacy enforced through individuals' consent
- Individuals' disease susceptibility
  - Insurance
  - Profiling
  - Dating … or perhaps "Dataing"



Existing Database Attacks of GWAS

- Homer's Attack
  - Individual's blood compared to a target population
  - If distribution of risk alleles match, individual ID'd
- Subverting database anonymization
  - By analyzing the remaining data, feature information can be used to ID the individual
  - Ex: Blonde hair, blue eyes
- Database connections

Paper Framework

- Preexisting attacks
- Novel identification attacks on GWAS statistics
  - Smaller reference populations
- Implementation of attacks
- Study of the attack countermeasures
- Attack results and evaluations

Attack 1: From Statistics to Allele Frequencies

Attack I

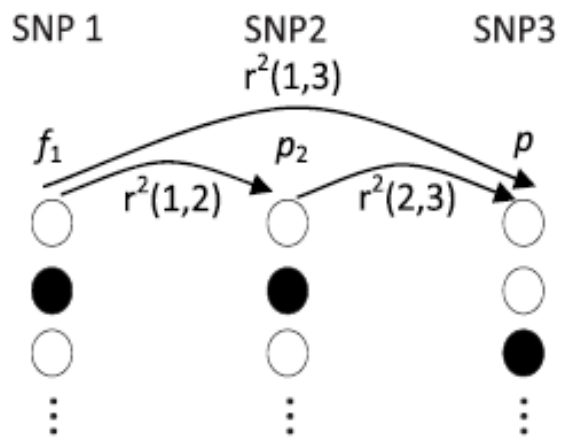


Figure 1: Recover allele frequencies.

- How likely one SNP can be used to infer some of the subjects other SNPs

Attack I

- Allow for a range of acceptable boundaries by using inequalities:

$$L < r^2 < U$$

- Result is *positive* (true) if the signs hold, or *negative* (false) otherwise

- If false, then infers that the sign's may need to be recovered (switched)

Attack II: A Statistic Attack

Attack II

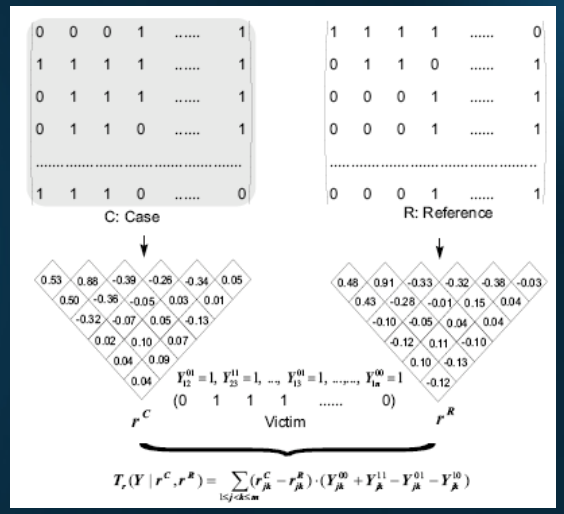
- Establish a reference group
  - SNP sequences from group of individuals
  - Same genetic background of the case group
- Derived from HMAP studies
- High confidence when results in *linkage distribution* (LD)
  - Combinations of alleles or genetic markers occur more or less frequently in a population than would be expected from a random formation

Attack II

- Assumes a null hypothesis that the victim is not in the case group
- T_r is the statistic designed to make the presence of an victim in the case group *valid*
- Given a positive result of T_r, an individual's SNP can be distinguished from the group therefore identifying the individual

Attack II

- Since single allele correlations are *not* normally completely independent, cannot assume null hypothesis



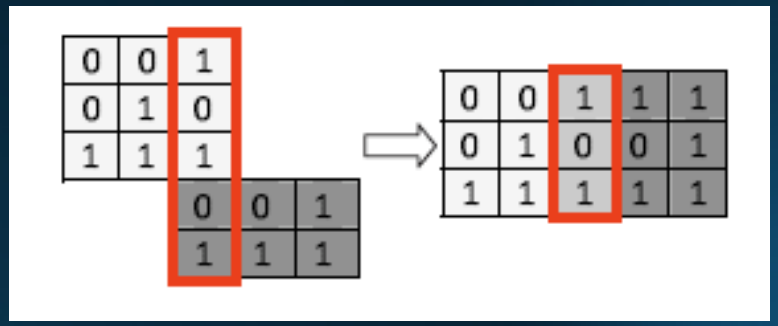
- Result is the similarity between the case group's r^2 and the victim's r^2

Attack III: Integer Programming Attack

Attack III

- Given allele frequencies for the surrounding regions of a SNP site (*locus*)
- Haplotypes
  - Specific combination of alleles across multiple neighboring SNP sites in a locus
  - Each individual has two haplotypes inherited from the parents
  - Population level - some haplotypes are more common than others.

Attack III



- “Divide and Conquer”
- Instead of computing every block derived from haplotypes merge haplotypes based on strong correlation between two SNPs

Defense

- Low-precision statistics
  - Downgrade the *linkage distribution* (LD)
  - Limiting the accuracy in comparing the victim's LD
  - Using allele frequencies still restored over 50% of pairwise frequencies and all the signs
- Thresholds
  - Publish less data → less informative
  - Sufficient information for recovering signs, attack still works
- Noise
  - Mitigates attack, but data becomes less useful

Implementation

- (1) Infer allele frequencies for individual SNPs and SNP from statistics (GWAS)
- (2) Propagate the marker SNP frequencies to other SNPs by using r^2
- Result:
  - Recovered all SNP frequencies
  - Half of pairwise frequencies
  - Most of the signs for r

Evaluations

- (1) Infer allele frequencies for individual SNPs and SNP from statistics (GWAS)
- (2) Propagate the marker SNP frequencies to other SNPs by using r^2

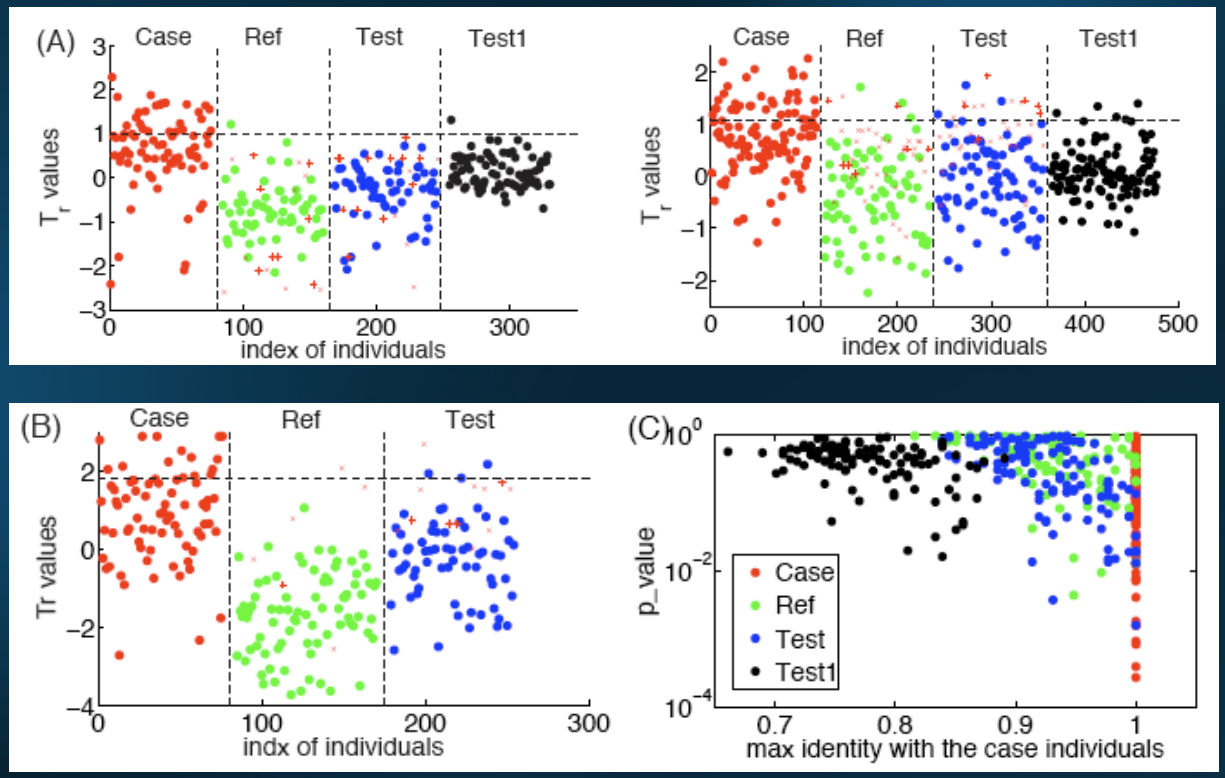
- Result:
  - Recovered all SNP frequencies
  - Half of pairwise frequencies
  - Most of the signs for r

Evaluations

- Using Markov model against GWASs
  - *Low-precision attacks*
    - 79% statistical power retained
  - *Threshold defense*
    - 85% statistical power retained
- Integer-programming attack
  - Run on 100 individuals
  - Within 12 hours successfully restored 174 SNPs for all 100 participants

Implementation

- Case = red dots, References (Ref) = green dots
- Tests: Test = blue dots, Test1 = black dots



Conclusion

- GWAS is a burgeoning field with a lot of attention placed upon the privacy, defense, and attacks of the studies' data
- This paper presents two new techniques that can lead to identification of victims in a GWAS
- Key: Form a **small set of statistics** routinely published in GWAS studies

Questions…

- …for the authors?

Attack I: Correlation and Recovery of SNP Alleles

- High r^2 value (0.93) =

$$r^2 = \frac{(C_{00}N - C_{*0}C_{0*})^2}{C_{0*}C_{1*}C_{*0}C_{*1}}$$

- Quick rundown…