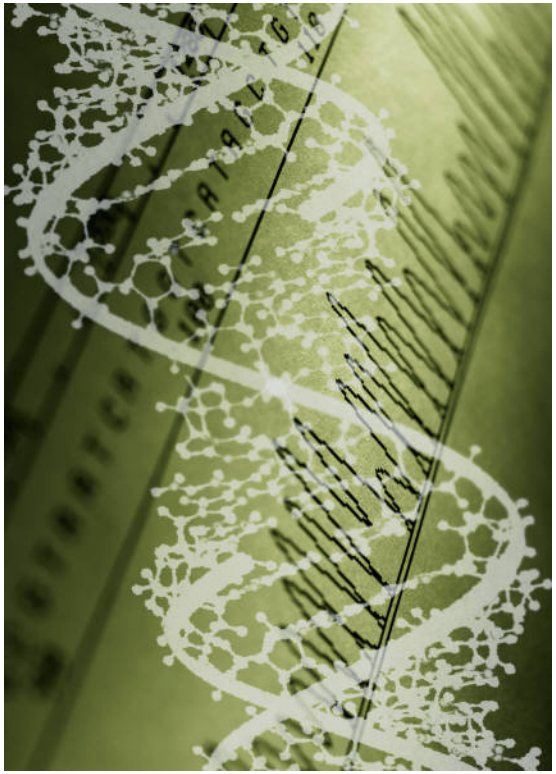
Learning Your Identity and Disease from Research Papers: Information Leaks in Genome-Wide Association Study

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

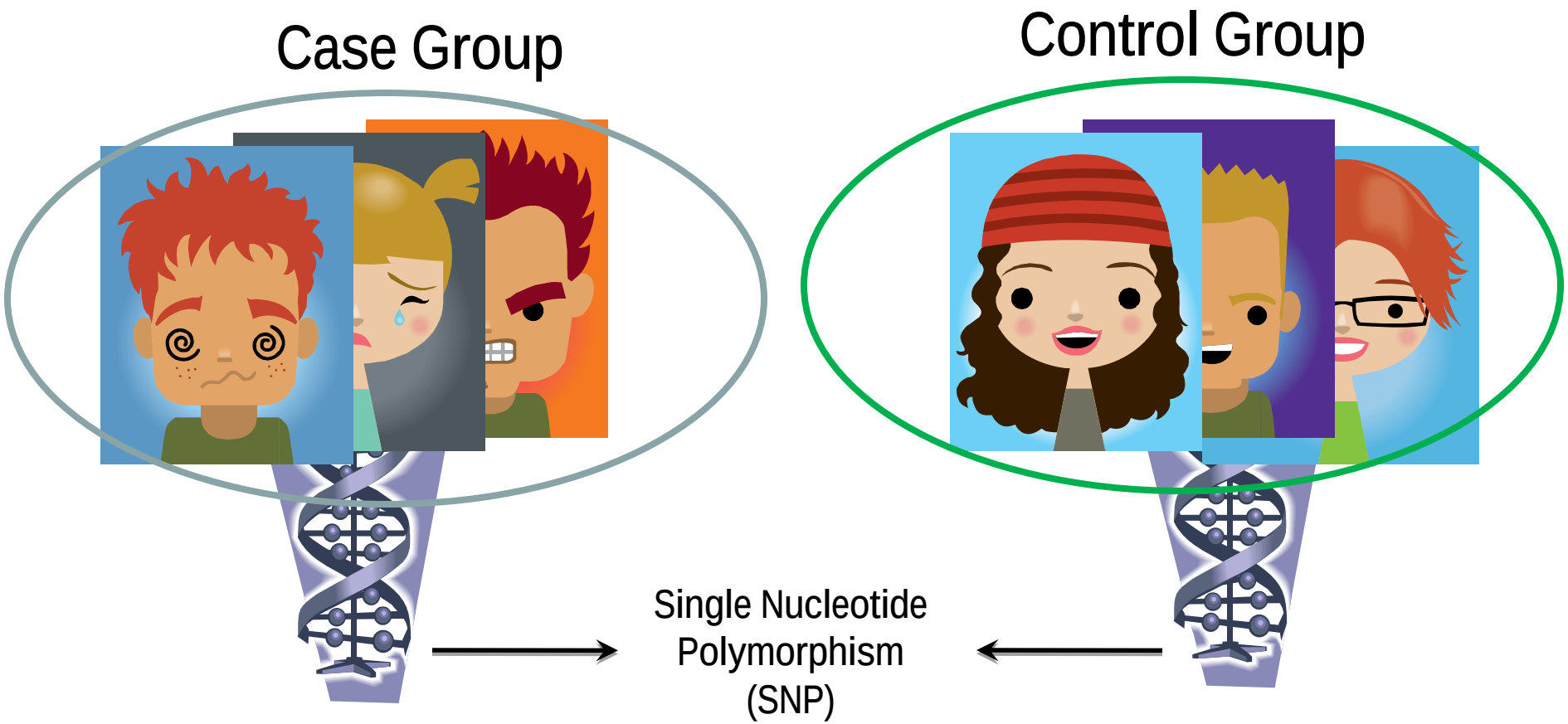
Indiana University at Bloomington

Genomic Revolutions

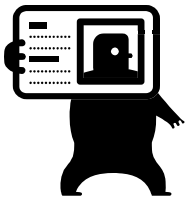


- Low-cost genotyping
- Revolutionary applications

Genome-Wide Association Study



Identification Risk



- Consequence of identifications

- Participant protection
  - De-identification
  - Aggregation
- Is this sufficient?

Attack on Aggregated Data

- Single-allele frequencies
  - Major: 0; Minor: 1
- Homer's attack
- NIH's Reactions

National Institutes of Health
The Nation's Medical Research Agency

The Rest of The Iceberg

- Other genome data
  - Test statistics
  - Linkage Disequilibrium (LD)
  - Haplotype sequences
- Other sources
  - Publications

Our Scary Findings

- ID from GWAS publications
  - Test statistics ⇨ Allele Frequencies
  - LD statistics ⇨ Statistical Identification
  - Pair-wise allele frequencies ⇨ SNP Sequences
- Work on real genome data
- Conclusion:

  Urgent needs to thoroughly study the problem

Why Doing This?

- Facilitate Dissemination of Genome Data SAFELY

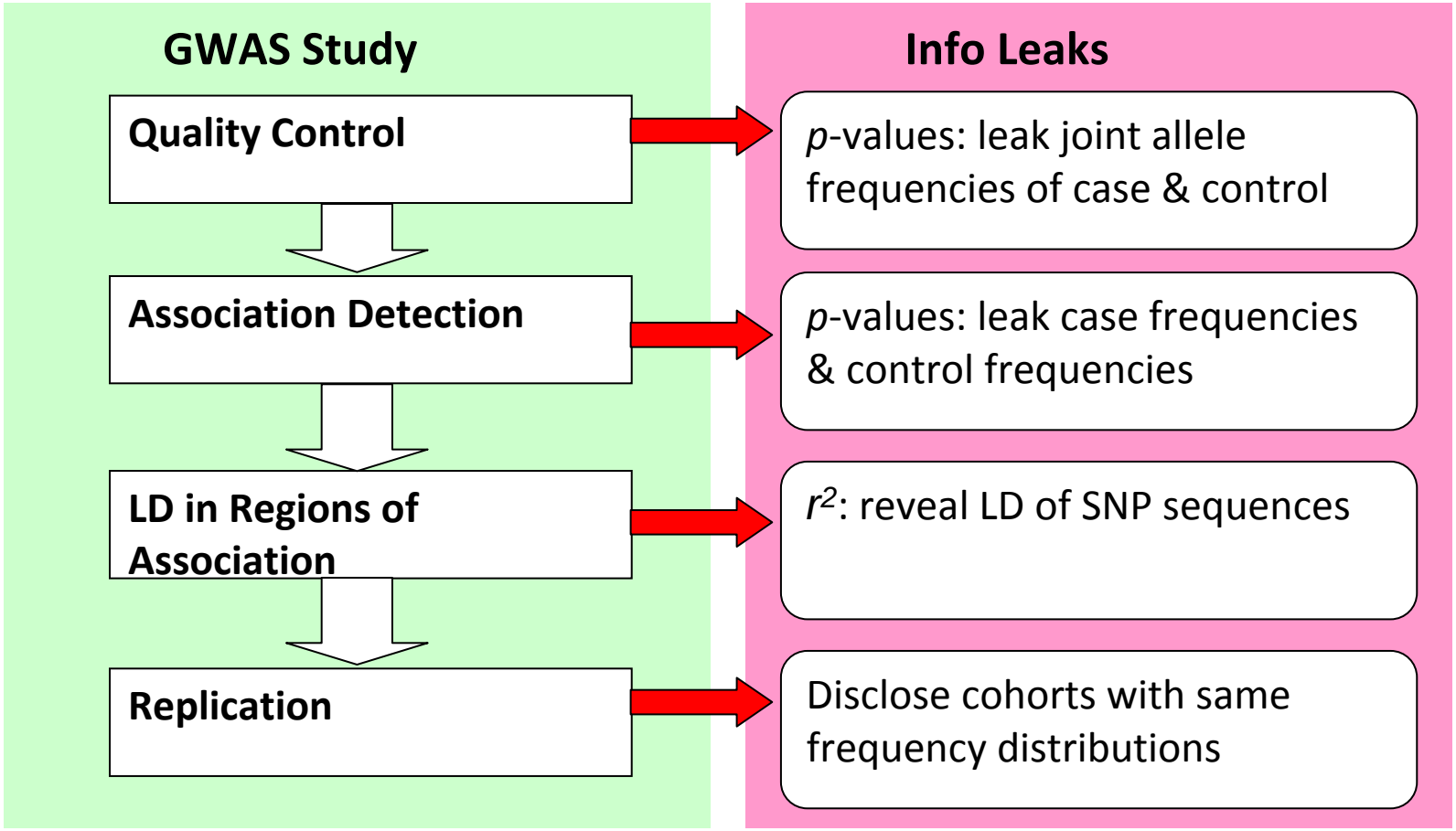
- A Lesson From the Internet:

Build Protection Into the Core!

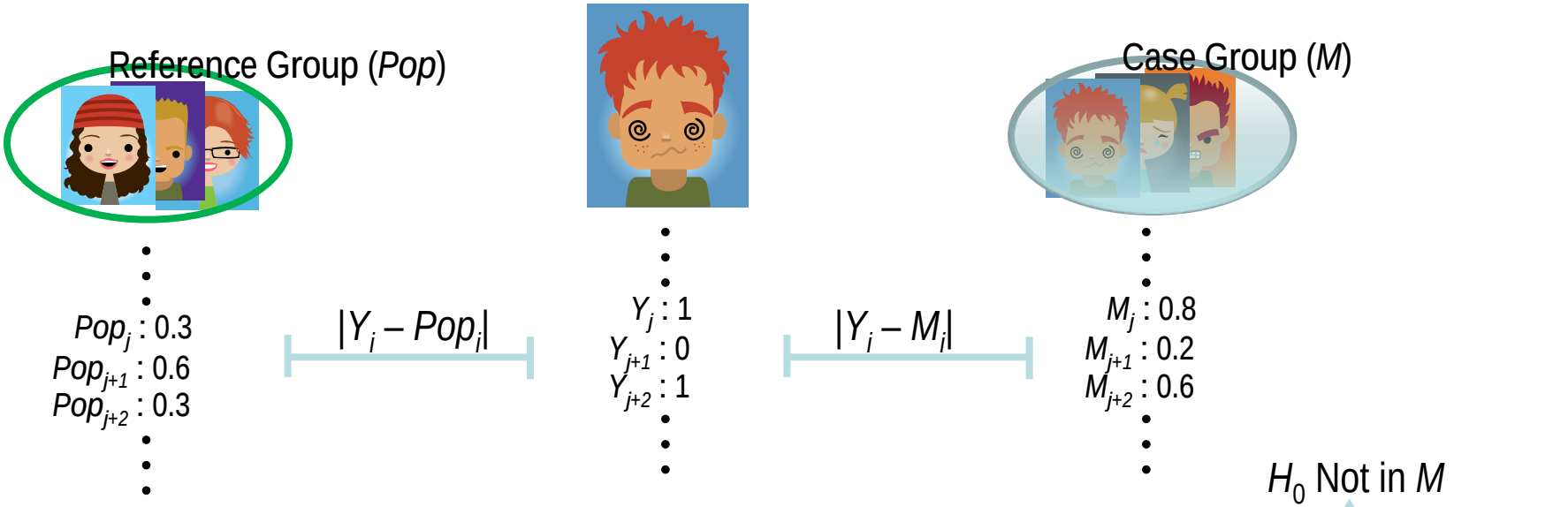
Terms

- Alleles
  - Single (0 1)
  - Pair-wise (00, 01, 10, 11)
- Genotype
  - Combinations of two sets of alleles
- Haplotype
  - SNP Sequence (phased genotype)
- Locus
  - Surrounding region of a SNP site

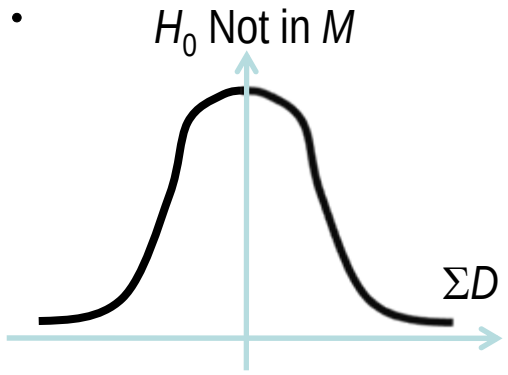
GWAS: Backgrounds



Homer's Attack



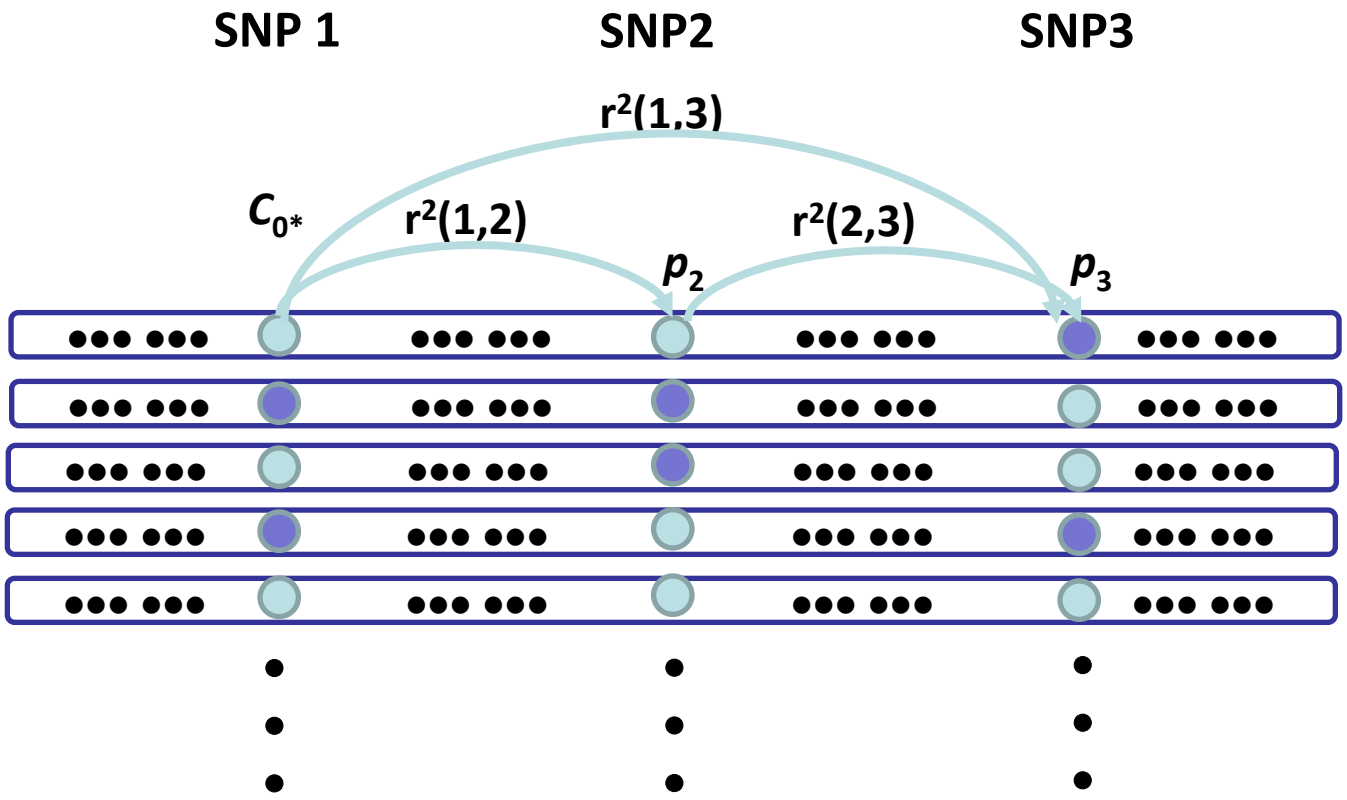
$$D(Y_i) = |Y_i - Pop_i| - |Y_i - M_i|$$



What we can do

- Reverse engineer test statistics
  - To find allele frequencies
- LD-based statistical identification
- Recover SNP sequences

Allele Frequency (Single)



Allele Frequencies (Pair-wise)

$$L < r^2 = \frac{(C_{00}C_{11} - C_{*0}C_{0*})^2}{C_{0*}C_{1*}C_{*0}C_{*1}} < U \quad (1)$$

$$C_{0*} = C_{00} + C_{01} \quad (2)$$

$$C_{1*} = C_{10} + C_{11} \quad (3)$$

$$C_{*0} = C_{00} + C_{10} \quad (4)$$

$$C_{*1} = C_{01} + C_{11} \quad (5)$$

- Catch: C_{00} not unique
  - Integer constraint
- Inaccurate r-squares
- Signs

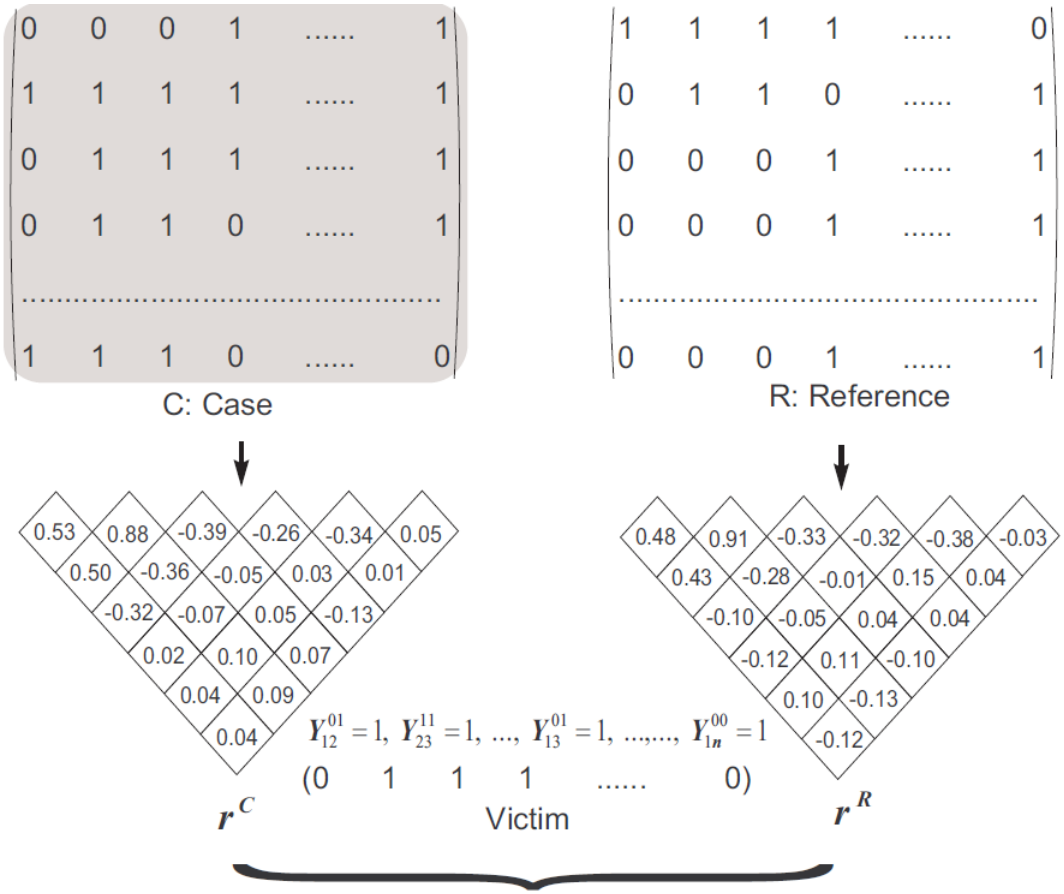
Homer-Style Attack Based On LD?

- Why? Single AF: n LD: $n(n-1)/2$
- But how?
- Validity of the test statistic

$$D(Y_i) = |Y_i - Pop_i| - |Y_i - M_i|$$

$$r^2 = \frac{(C_{00}C_{11} - C_{10}C_{01})^2}{C_{0*}C_{1*}C_{*0}C_{*1}}$$

Our Statistical Attack



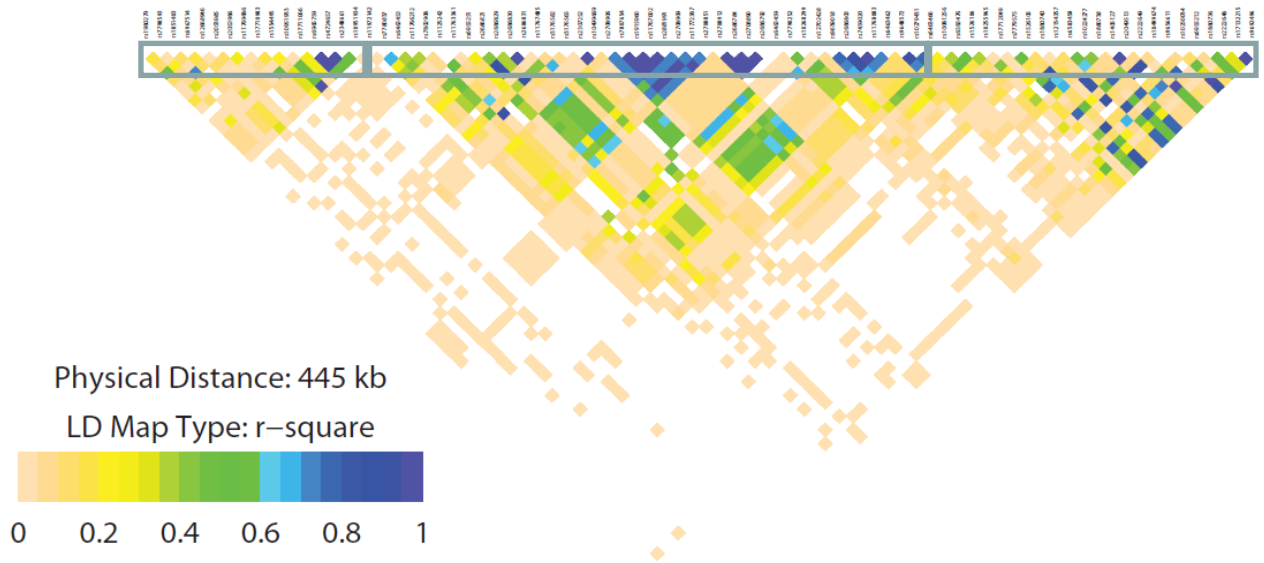
- We have to use signed *r*
- Distribution of T_r?
  - Markov model
- Reference?

$$T_{ij} = |(Y_{ij}^{00} + Y_{ij}^{11}) - (r_{ij}^{R} + 1)/2| - |(Y_{ij}^{00} + Y_{ij}^{11}) - (r_{ij}^{C} + 1)/2| = (r_{ij}^{C} - r_{ij}^{R})(Y_{ij}^{00} + Y_{ij}^{11} - Y_{ij}^{01} - Y_{ij}^{10})$$

$$T_r = \sum_{1 \le i \le j \le N} T_{ij}$$

Recover SNP Sequences

ID	Sequence
1	0 1 1 0 1 1 1 0 0 1 0 1 0 0 1 1 1 0 1 … … … … … 1 0 0 1 1 1 0 1 1 1 1 0 0 0 1 1 1 0 0 1 0
2	1 1 1 0 1 0 1 0 1 0 0 0 1 1 0 0 1 1 0 … … … … … 0 0 0 0 0 0 0 1 1 0 1 0 0 1 0 0 0 0 0 0 1
3	0 1 1 1 0 0 0 1 0 1 0 0 1 1 1 0 1 0 1 … … … … … 0 0 1 1 1 0 0 0 1 0 1 0 1 1 0 0 0 1 1 1 1
…	… … …
…	… … …
n-1	0 0 1 0 1 1 1 1 0 0 0 0 0 1 1 0 0 1 1 … … … … … 0 1 1 0 1 0 1 0 1 1 0 0 1 0 1 0 0 1 1 0 1
n	1 0 0 1 0 1 0 0 1 0 1 0 1 1 0 1 0 1 0 … … … … … 1 0 1 1 0 0 0 0 1 1 1 1 0 0 1 0 0 1 1 1 1



- Contingency table problem
  - Studied for decades
  - Very difficult
- Divide-and-Conquer
  1. Construct each haplotype block
  2. Connect different blocks

Simple Defense

- Low-precision statistics
  - Correlation among SNPs
- Thresholds
  - How to determine them?
- Noises
  - Consistency check
  - Maximum-likelihood approximation

Evaluations

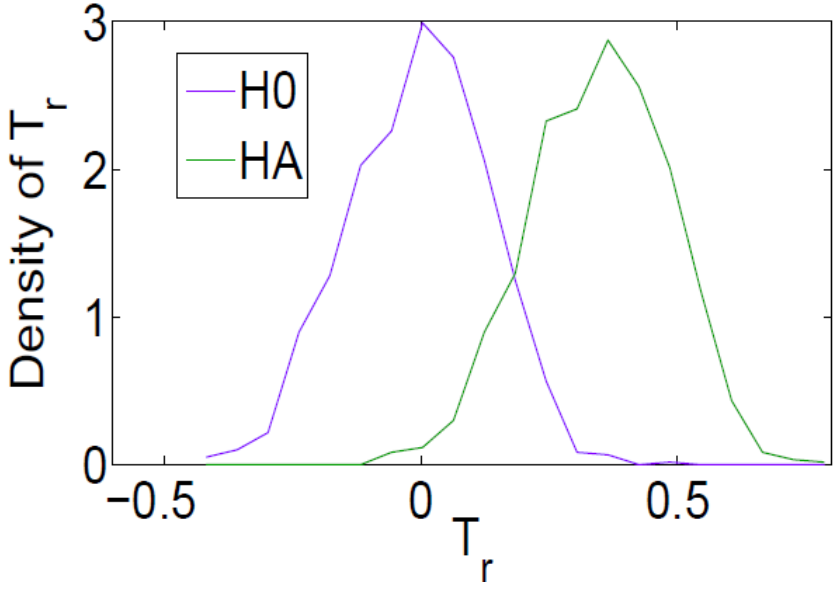
- Data: the HapMap project
- Locus: FGFR2
  - 174 SNPs
  - Used in a real GWAS study
- Population
  - Africa backgrounds
  - 200: half cases and half controls

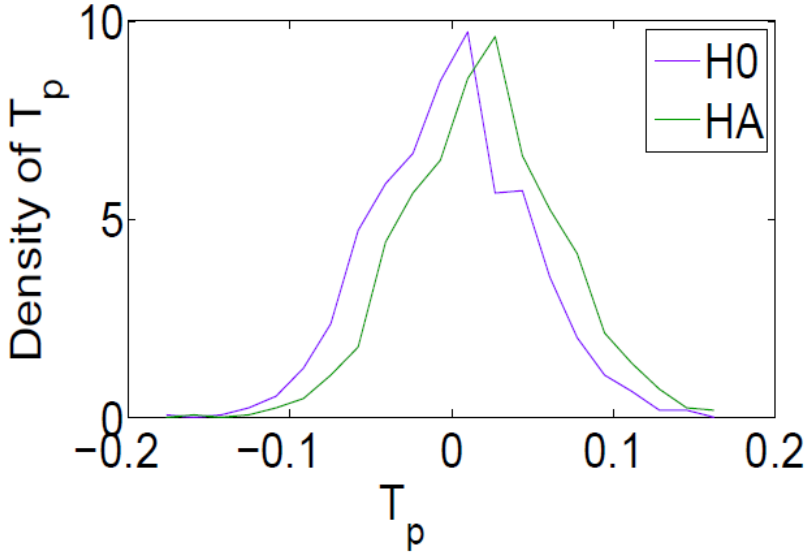
Allele Frequencies and Signs

Statistics Precision		Recovered Information %		
r^2	p-value	single SNP frequency	pair-wise frequency	sign of r
0.1	0.1	12.1	1.8	6.7
0.1	0.00001	40.6	11.7	31.7
0.01	*	100	50.1	98.7
0.001	*	100	90.4	100
0.0001	*	100	95.1	100

Statistical Powers

- 20 times more powerful than Homer's test (T_p)





Recover Haplotypes

- Linear equation solving: *rref*
- Integer Programming: *bintprog*
- 100 individuals, 10 blocks, 174 SNPs
- System: 2.80GHz Core 2 Duo, 3GB memory
- Fully restored within 12 hours

Discussion

- Genotypes vs. Haplotypes

- Defense
  - Differential privacy

Conclusion

- New attacks and new understanding

- Many open research problems

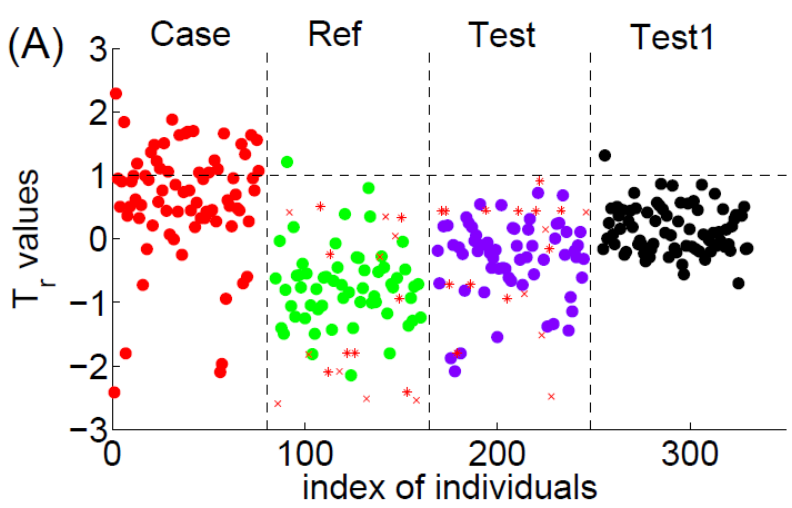
Contacts

- Dr. XiaoFeng Wang
- 812-856-1862
- Web: www.informatics.indiana.edu/xw7
- System Security Lab: sysseclab.informatics.indiana.edu

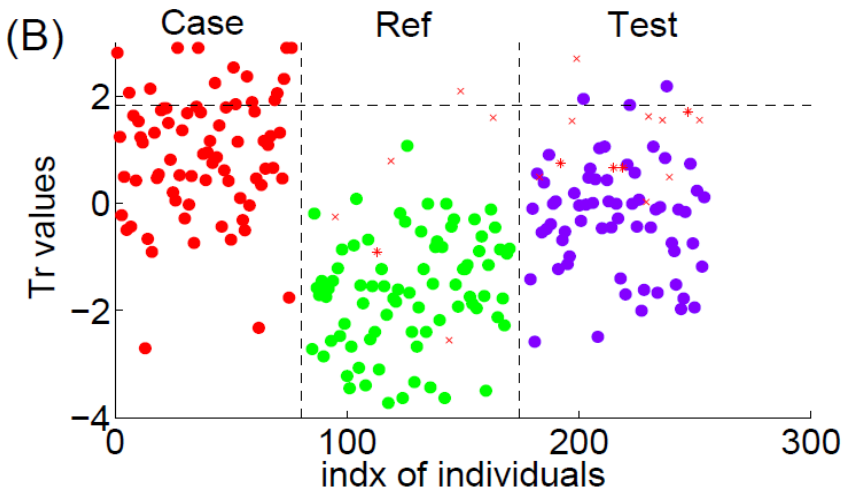
- Dr. Haixu Tang
- 812-856-1859
- Web: www.informatics.indiana.edu/hatang

References

- Good: from the same population
- Bad: from different populations



good reference



average reference

More In-depth Studies

- Larger populations:

N	50	100	200	400	800	1600
power (%)	99.9	85.7	67.2	40.4	36.2	18.1

- Low-precision statistics (200 cases, 200 references)

Precision of r^C	0.5	0.2	0.1	0.01	0.001
% power π left	12	74	85	100	100