Protecting Kinship Privacy in Genomic Databases

Students

Emre Hilmi Songur
Farukhzhon Barotov

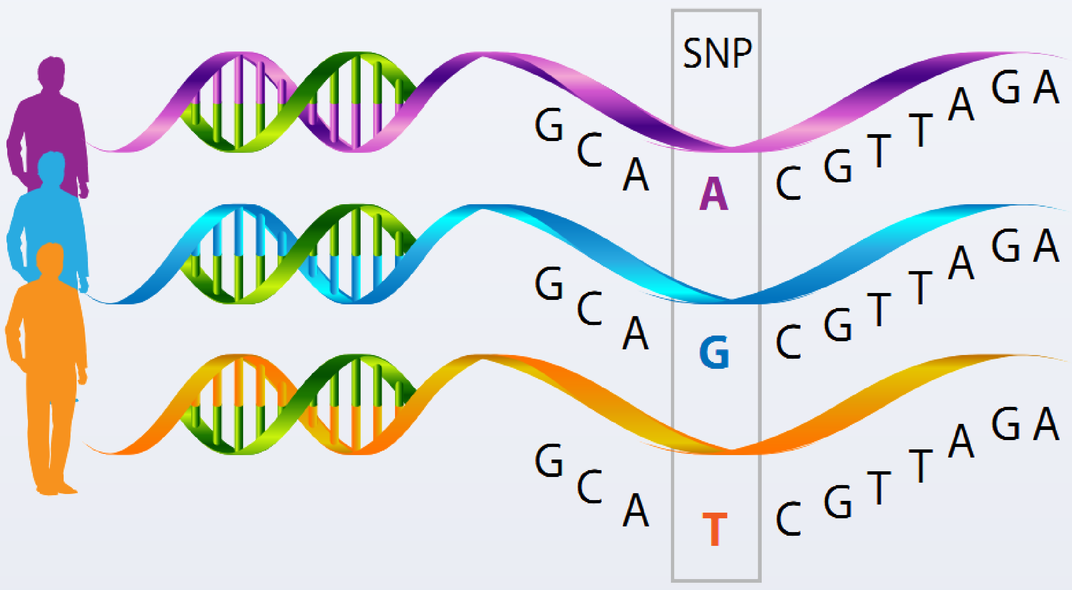
Faculty Member

Öznur Taştan

Sabancı Üniversitesi

PURE
PROGRAM FOR UNDERGRADUATE RESEARCH

FROM SNPs TO GENOMIC DATABASES



What Makes You Different?

- The human genome sequence is **99.6%** identical in all people.
- SNP (**Single Nucleotide Polymorphism**) is DNA sequence variation occuring when a single nucleotide differs between two or more genomes.
- SNPs make up **0.4%** of the genome.

What is Kinship?

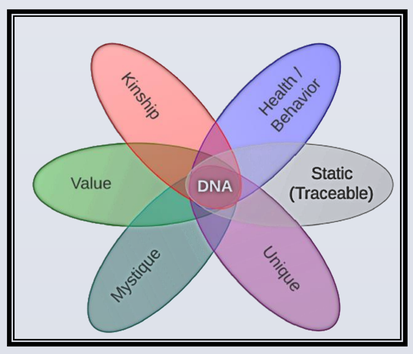
Relationship	Kinship Coefficient
Individual-self	1/2
Siblings	1/4
Parent-offspring	1/4
Grandparent-grandchild	1/8
Uncle/aunt-nephew/niece	1/8
First Cousins	1/16
Half-siblings	1/8

What is KING Coefficient?

$$\phi_{ij} = \frac{2n_{11} - 4(n_{02} + n_{20}) - n_{*1} + n_{1*}}{4n_{1*}}.$$

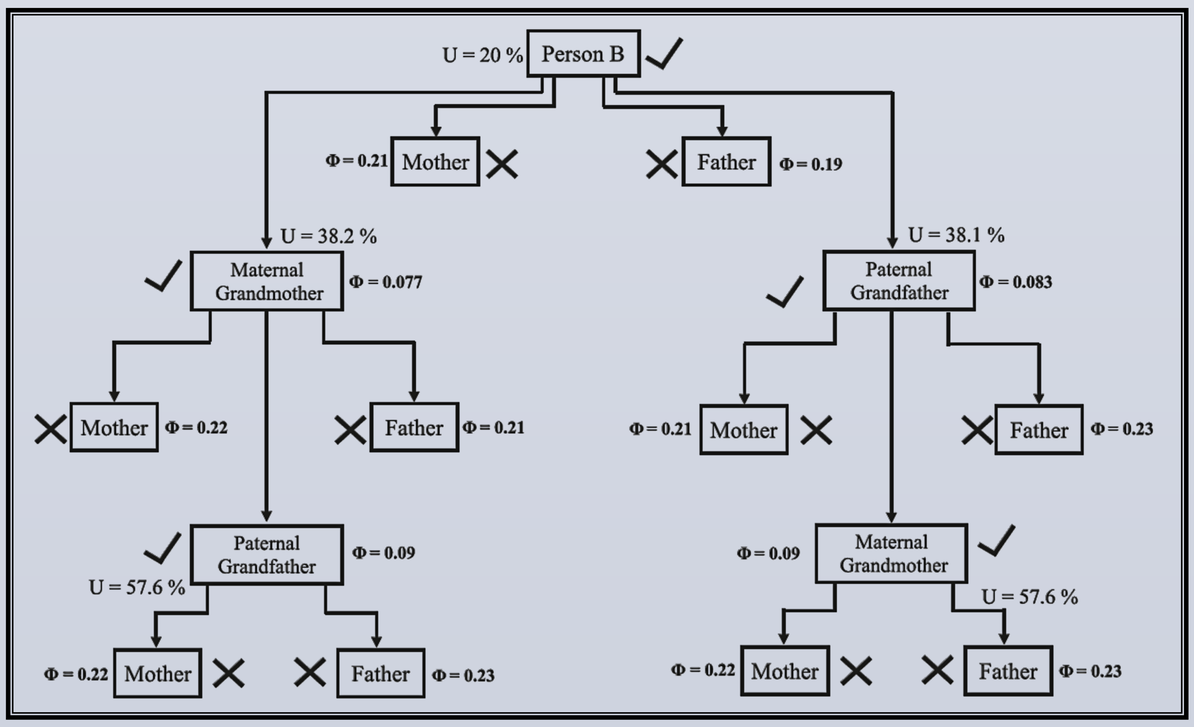
Here, n_{11} is the number of genomic positions that are heterozygous in both individuals, n_{02} is the number of SNPs where the first individual (i) is homozygous dominant and the second individual (j) is homozygous recessive. n_{20} denotes the positions where j is homozygous dominant and i is homozygous recessive. n_{1*} and n_{*1} are the number of SNPs that are heterozygous for individual i and for individual j, respectively.

Genomic Data has:



PURPOSE AND MOTIVATION

- A researcher found out that he had a half-sibling from genomic database:
  - «With genetic testing, I gave my parents the gift of divorce»
- The law enforcement recently tracked (and identified) the Golden State Killer
  - (by using a relative's genomic data in a database)
- Real-life examples led us build a privacy-protection system for **genomic databases**.



METHODOLOGY AND CALCULATIONS

We have a .gds file that contains a matrix of 262178 SNP positions belonging to 1412 people

	Individual 1	Individual 2	Individual 3	Individual n
SNP Position 1	1	0	1	2
SNP Position 2	2	1	1	0
SNP Position 3	0	0	2	1
SNP Position 4	2	1	0	1
SNP Position 5	1	0	2	2
SNP Position 6	0	2	2	1
SNP Position m	1	0	0	0

And this is how we convert the organic bases into 0, 1, 2 whilst creating this .gds file:

SNP Position	Alleles in one Individual	Alleles in the Reference Genome	How Many of the Alleles Changed?
788822	AG	AG	0
788825	GC	CC	1
788829	AA	GG	2

❑ **We focused on the correlation between the missing SNP position and all other SNP positions in the database.**

❑ **We considered the result of the SNP position with the highest correlation coefficient. Max{0,1,2} has been taken and injected into the masked .vcf file of the individual.**

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

CONCLUSION

❖ **Since we have the fully revealed file of the firstly added individual,**
❖ **Since we unmasked the masked individual by the calculated values**

❖ **We calculated KING before & after and compared, trying to see whether the kinship has been revealed.**

REFERENCES

- Ayday, E., & Humbert, M. (2017). Inference Attacks against Kin Genomic Privacy. *IEEE Security & Privacy*, *15*(5), 29-37. doi:10.1109/msp.2017.3681052
- Kale, G., Ayday, E., & Tastan, O. (2017). A utility maximizing and privacy preserving approach for protecting kinship in genomic databases. *Bioinformatics*, *34*(2), 181-189. doi:10.1093/bioinformatics/btx568
- Lange, Kenneth (2003). Mathematical and statistical methods for genetic analysis. Springer. pp. 81–83
- Manichaikul, A., Mychaleckyj, J. C., Rich, S. S., Daly, K., Sale, M., & Chen, W. (2010). Robust relationship inference in genome-wide association studies. *Bioinformatics*, *26*(22), 2867-2873. doi:10.1093/bioinformatics/btq559
- Naveed,M. et al. (2015) Privacy in the genomic era. ACM Comput. Surv., 48, 6.
- https://www.vox.com/2014/9/9/6107039/23andme-ancestry-dna-testing