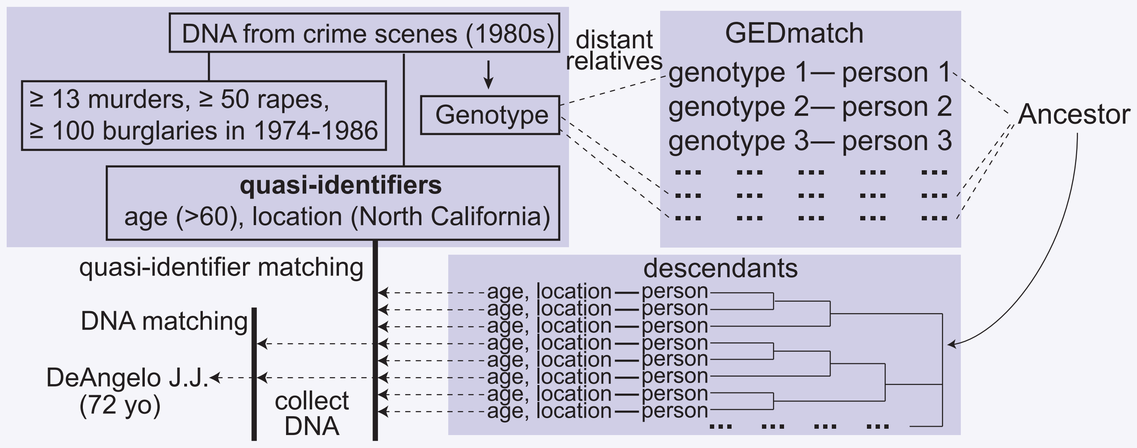
Privacy time bombs in omics data: latent risk manifests over time

Zhiqiang Hu, Steven E. Brenner
University of California, Berkeley, CA, USA
hu.zhiqiang@berkeley.edu; brenner@compbio.berkeley.edu

Consumer genetics and quasi-identifiers enable widespread re-identification from a genomes

Genomic data are unique fingerprints. Such data can be used to identify a person, as well as be used to infer private traits. Recent studies suggest the increasing consumer genetics data and the commonly attached quasi-identifiers pose challenges to privacy.

An example: how consumer genetics helped identify the Golden State killer

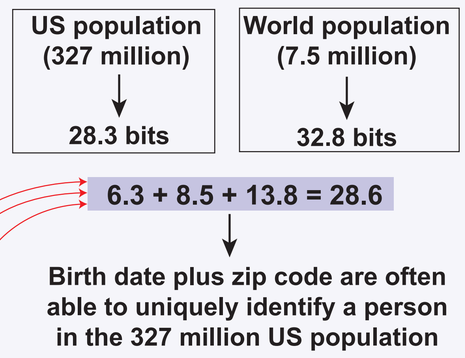


How do quasi-identifiers help re-identify a person from a population?

Table. Entropy and the contribution of quasi-identifiers.

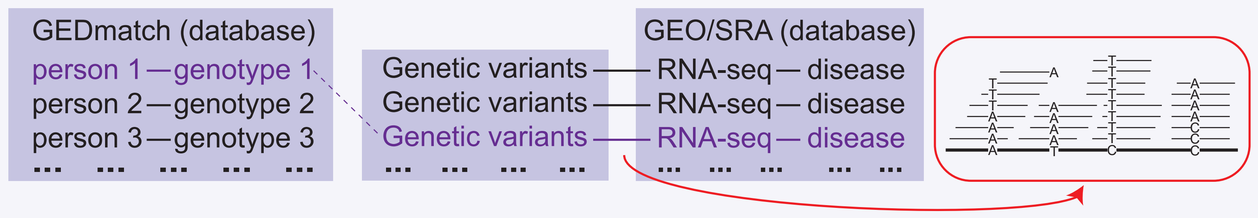
Quasi-identifier	Expected information content (bits)
Sex	1.0
Ethnic group	1.4
Eye color	1.4
Blood group (ABO and Rhesus systems)	2.2
State of residence	5.0
Height	5.0
Year of birth	6.3
Day and month of birth	8.5
Surname	12.9
Zip code	13.8

Table adopted from [1]

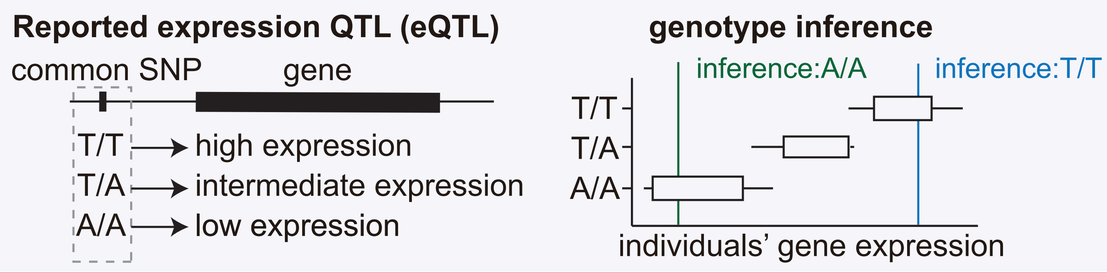


Gene expression profiles can be linked to genetic datasets, enabling re-identification and revealing medical conditions

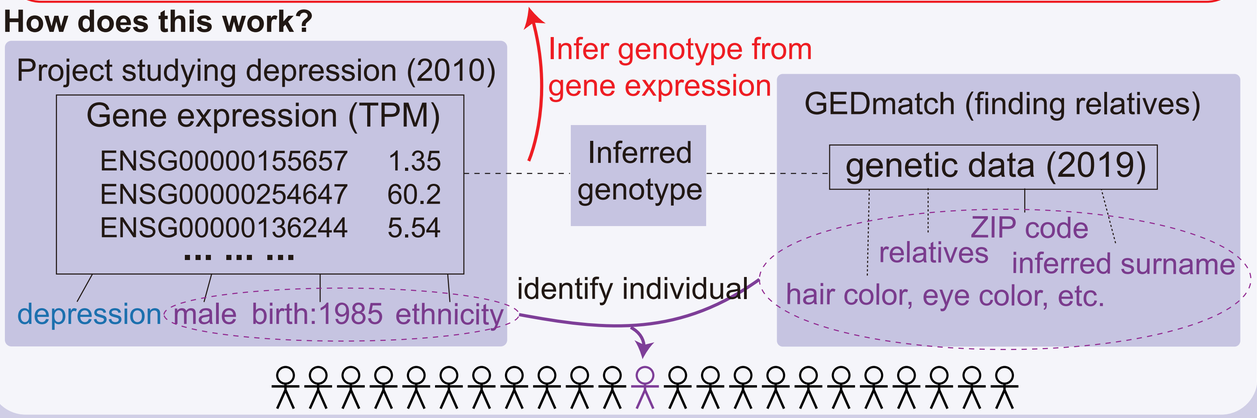
RNA-seq reads contain genetic variants, thus can be readily linked to public genotypes



Gene expression levels (without reads) can be used to infer genetic variants via expression QTLs, thus can be also linked to public genotypes



How does this work?



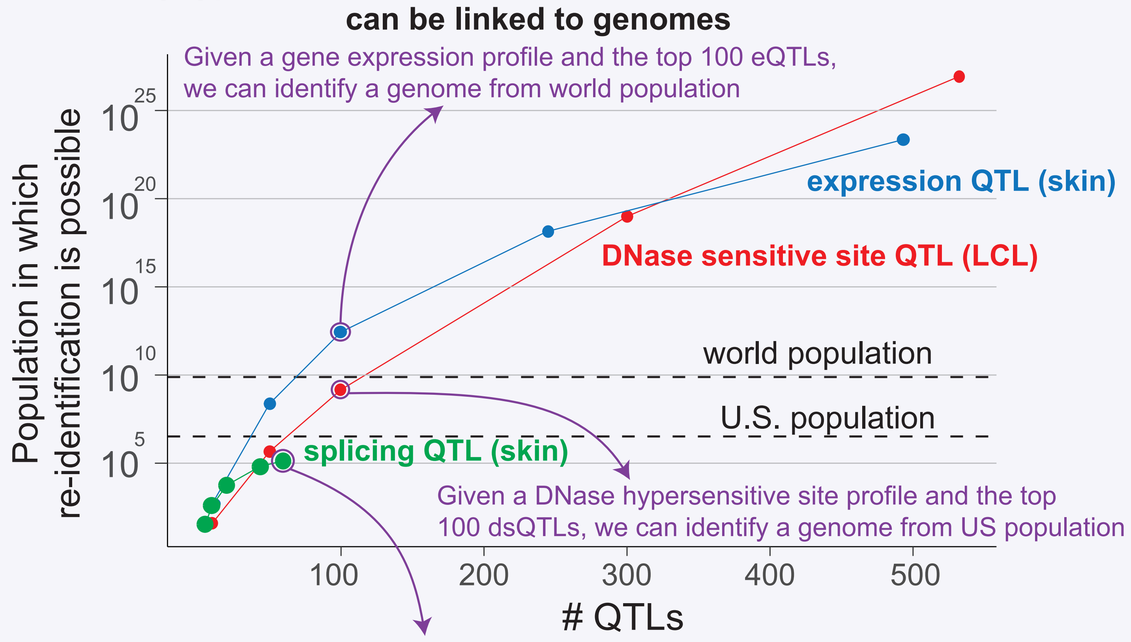
Many types of omics data have numerous QTL usable for re-identification

Reported molecular QTLs.

QTL type	QTL number	Year
expression QTLs	3,124,346	2017 [2]
splicing QTLs	16,483	2015 [3]
DNA methylation QTLs	2,907,234	2018 [4]
protein expression QTLs	16,602	2018 [5]

QTL type	QTL number	Year
metabolite QTLs	145	2014 [6]
histone modification QTLs	315	2015 [7]
ribosome occupancy QTLs	939	2015 [8]
DNase sensitive site QTLs	8,902	2012 [9]

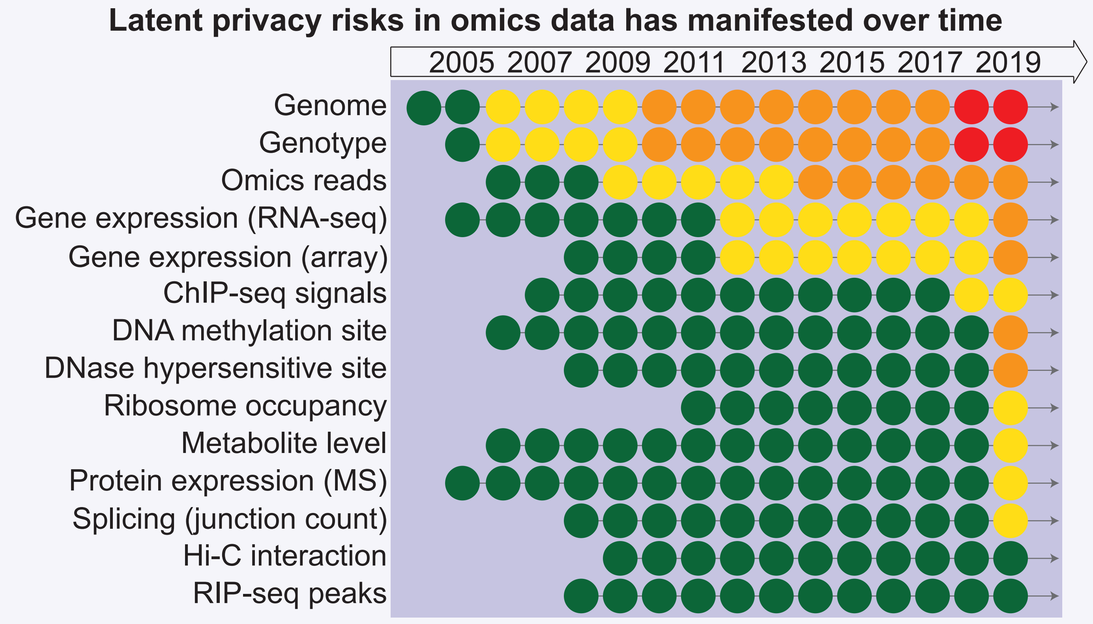
Many types of omics data, whose QTLs are less abundant, can be linked to genomes



Hidden privacy risks exist in omics data (even high-level summary data), which will only manifest over time

- more people will make genetic data public
- more omics data and more types of omics data will be available
- more QTLs will be detected, due to accumulated biological knowledge and sequencing
- more phenotypes will be able to be inferred directly from genotypes
- more powerful linking strategies will be developed, due to improved algorithms and compute resources

Latent privacy risks in omics data has manifested over time



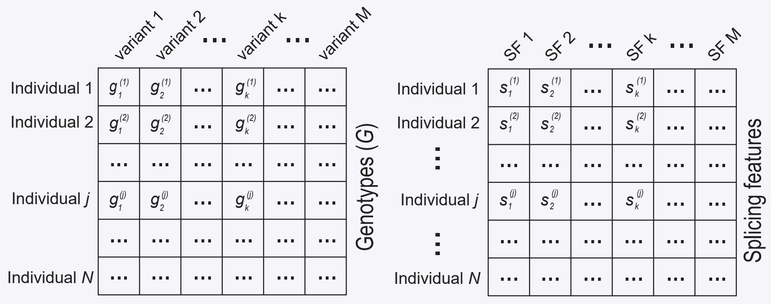
★ **The security of cryptographic methods is known to degrade over time. Approaches like MD5 and DES were state-of-the-art but now deprecated as compromised.**
The need to preserve individuals' privacy for their lifetimes (and beyond, for descendants) poses unique challenges to the effective sharing of omics data, as public data have ~100 times the impact of controlled access data.

Detailed linking strategy, applied to splicing data

Task description

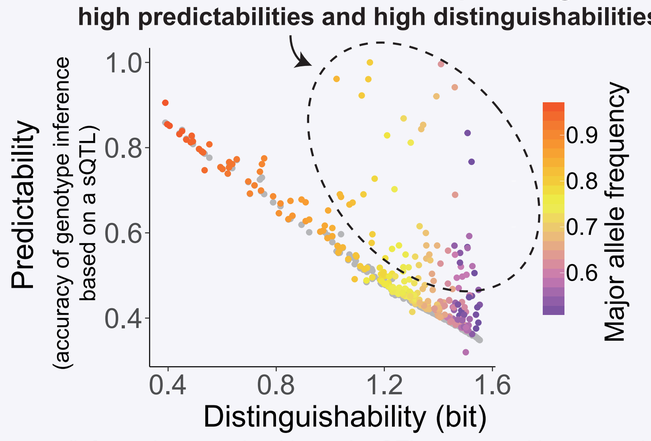
Given: (1) A pool of individual genotypes (G)
(2) A splicing feature ("SF", can be either PSI or relative isoform expression) profile of an unknown individual from G
(3) public sQTLs: variants associated with splicing features

Task: Identify the corresponding genome out of the genome pool.

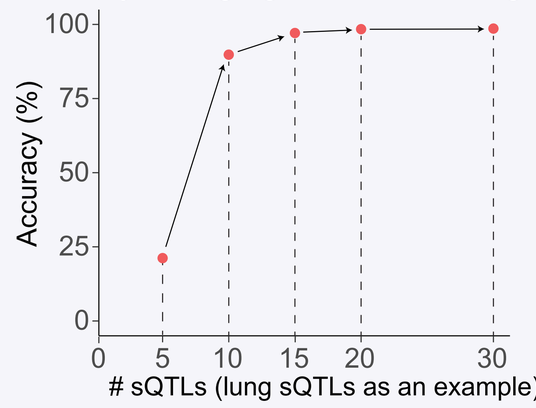


sQTLs are ~0.5% as abundant as eQTLs. Here we use GTEx data, which contains both genotypes and RNA-seq data, and the public sQTLs [3], to evaluate the feasibility of sQTL-based linking attack.

Some sQTLs are informative, providing both high predictabilities and high distinguishabilities.



Using a splicing profile from a sample and a very small number of sQTLs, we can identify the target genome out of ~200 genomes



References

1. Erlich Y, et al. 2014. Nature Review Genetics. doi:10.1038/nrg3723
2. GTEx Consortium. 2017. Nature. doi:10.1038/nature24277
3. GTEx Consortium. 2015. Science. doi:10.1126/science.1262110
4. Hannon E, et al. 2018. AJHG. doi:10.1016/j.ajhg.2018.09.007
5. Yao C, et al. 2018. Nature Communication. doi:10.1038/s41467-018-05512-x
6. Sin S, et al. 2014. Nature Genetics. doi:10.1038/ng.2982
7. Wasak SM, et al. 2015. Cell. doi:10.1016/j.cell.2015.08.001
8. Battle A, et al. 2015. Science. doi:10.1126/science.1260793
9. Degner JF, et al. 2012. Nature. doi:10.1038/nature10808

Acknowledgments

H.Z. was supported by an NIH/NIBMB grant U01 EB023686 to Mark Gerstein and by Tata Consultancy Services.
We thank Jingqi Chen for pre-processing of the GTEx data. We thank Gamze Gursoy and Arif Harmanci from Yale University for discussions.