Biological discovery and consumer genomics databases activate latent privacy risk in functional genomics data

Zhiqiang Hu, Steven E. Brenner
University of California, Berkeley, CA, USA

Summary

Research participants are typically assured that their personal data will be protected. We found that current functional genomics research data sharing practices may create privacy time bombs. We have demonstrated that some types of data, such as gene expression levels or DNase hypersensitive sites, can be accurately linked to a unique genome or genotype in consumer genealogy databases. Public research data typically had few privacy concerns at the time they were created and initially distributed. However, biological discoveries, new databases, and new techniques make it increasingly likely that shared research datasets could potentially compromise personal information. This poses unique challenges to the effective sharing of high-throughput molecular data.

Background

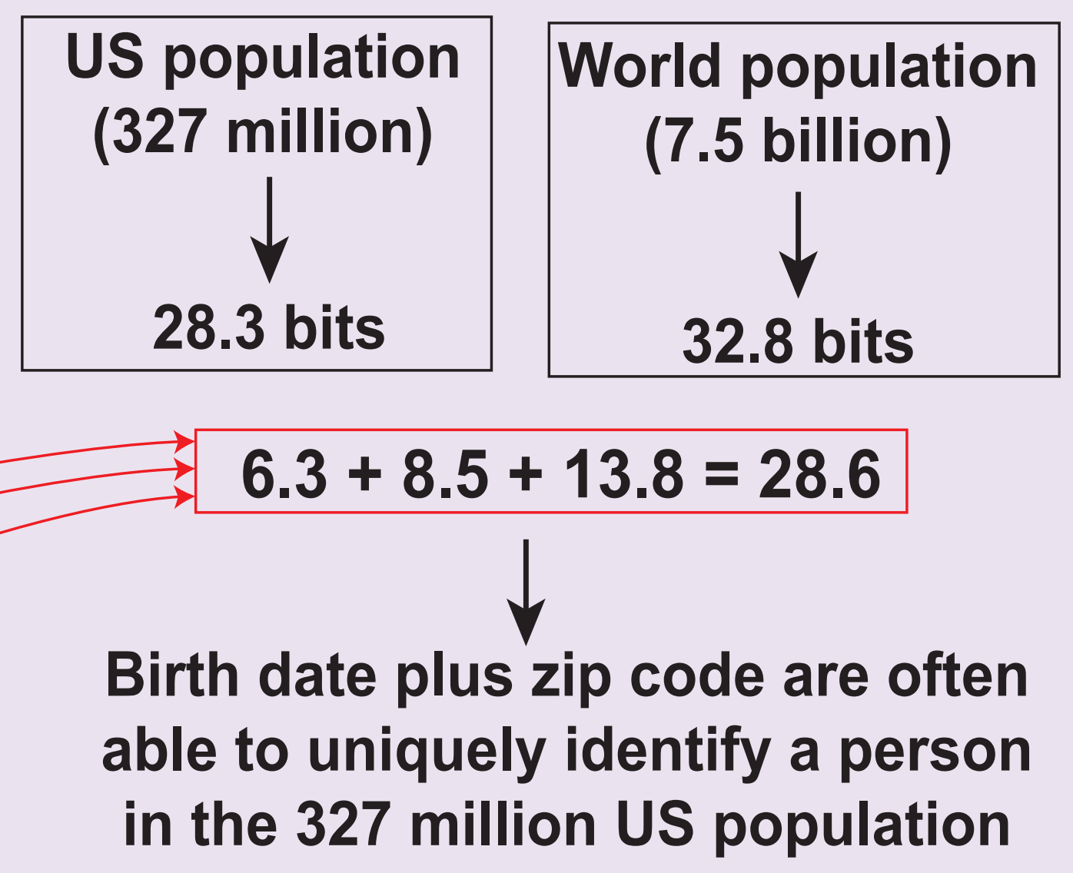
Consumer genetics and quasi-identifiers enable widespread re-identification from genomes

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

Adopted from Erlich Y, et al. 2014. *Nature Reviews Genetics 15:*409-421



US population (327 million) → 28.3 bits

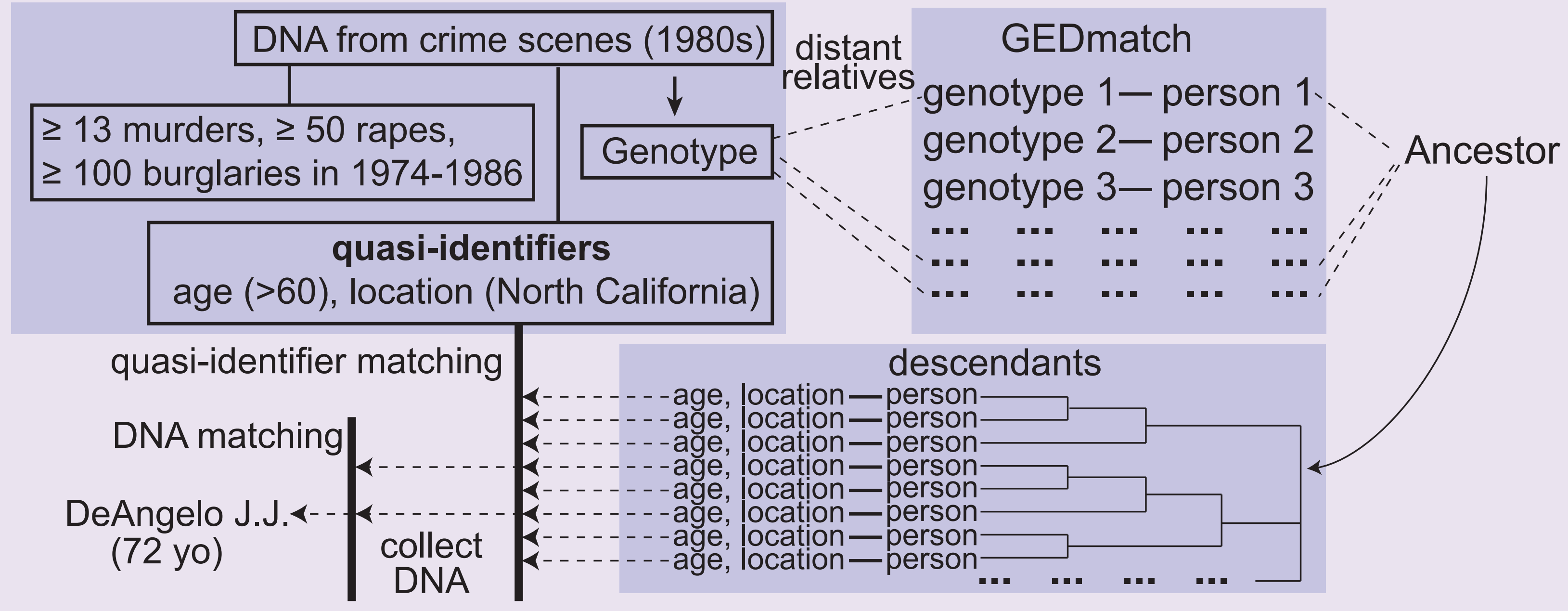
World population (7.5 billion) → 32.8 bits

6.3 + 8.5 + 13.8 = 28.6

Birth date plus zip code are often able to uniquely identify a person in the 327 million US population

Increasing consumer genetics data pose challenges to privacy

(An example: identifying the Golden State killer)



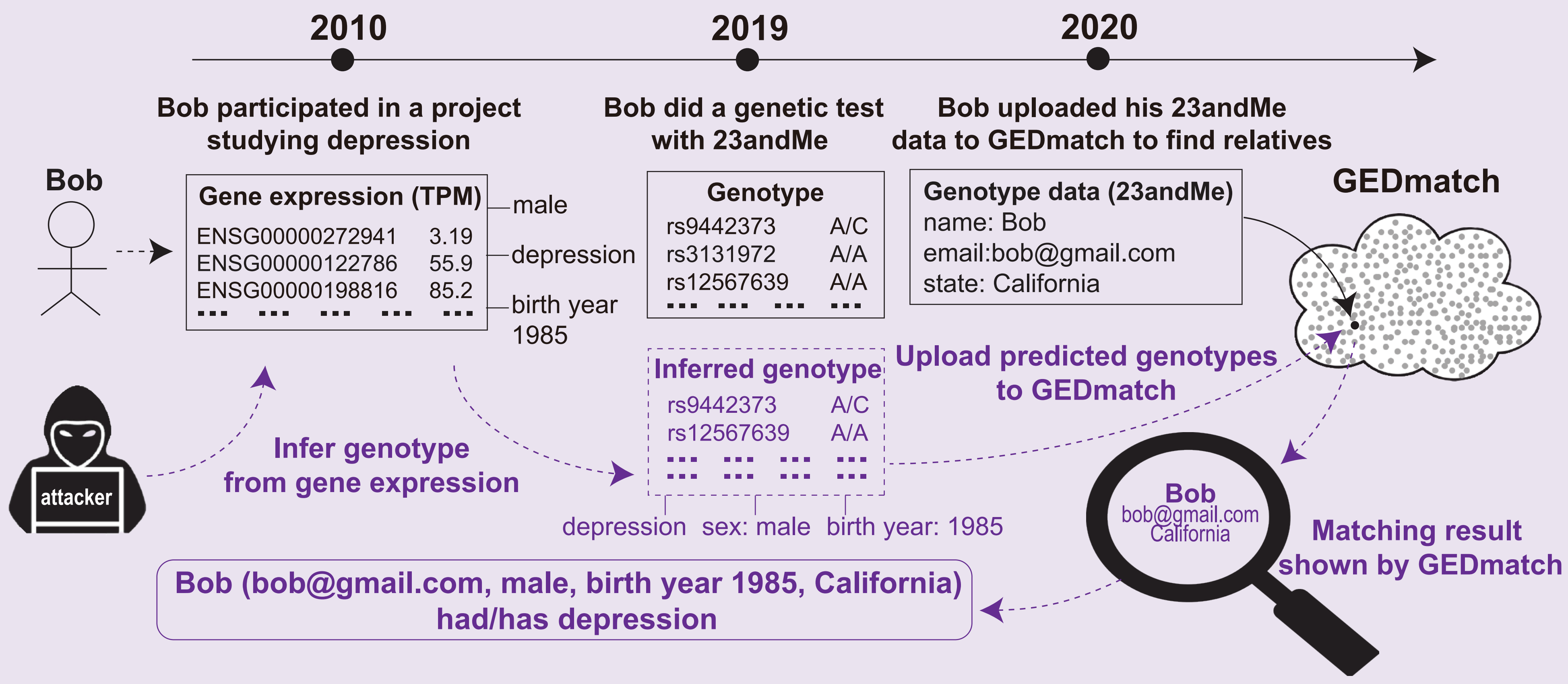
DNA from crime scenes (1980s)

≥ 13 murders, ≥ 50 rapes, ≥ 100 burglaries in 1974-1986

Genotype

quasi-identifiers: age (>60), location (North California)

Gene expression profiles (*without reads*) can be linked to genetic datasets, enabling re-identification and revealing medical conditions



2010: Bob participated in a project studying depression

2019: Bob did a genetic test with 23andMe

2020: Bob uploaded his 23andMe data to GEDmatch to find relatives

Gene expression (TPM)	
ENSG00000272941	3.19
ENSG00000122786	55.9
ENSG00000198816	85.2
...	...

male; depression; birth year 1985

Genotype	
rs9442373	A/C
rs3131972	A/A
rs12567639	A/A
...	...

Genotype data (23andMe)
name: Bob
email:bob@gmail.com
state: California

Inferred genotype	
rs9442373	A/C
rs12567639	A/A
...	...

depression sex: male birth year: 1985

Infer genotype from gene expression

Upload predicted genotypes to GEDmatch

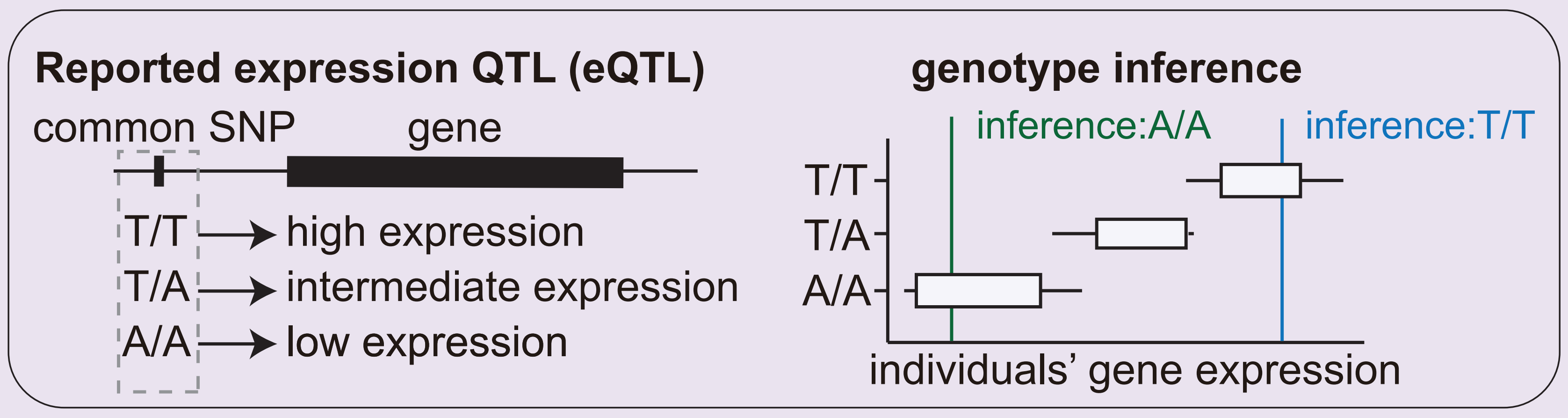
Matching result shown by GEDmatch

Bob (bob@gmail.com, male, birth year 1985, California) had/has depression

Consider Bob, who participated in a depression research study in 2010, where the researchers generated RNA-seq data from his blood sample. The researchers carefully considered his privacy and only released the gene expression levels, along with his depression status, sex, age, height, weight, education level, marital status, and income. In 2019, Bob purchased a genetic test from 23andMe. Later he uploaded his data to GEDmatch database to find potential relatives.

An attacker intends to identify individuals with depression. He obtains expression data from the publicly available depression study. He infers genotypes from the expression data, and uploads the inferred genotype data to GEDmatch, where he identifies matched DNA data. The attacker could now readily re-identify Bob if he submitted enough identifying information to GEDmatch (or answers the attacker's email). But even if Bob is circumspect, the combination of quasi-identifiers from the research study linked to quasi-identifiers from GEDmatch may be enough to uniquely identify him. Bob's depression status may be revealed.

How can genotypes be inferred from gene expression data?



Reported expression QTL (eQTL)

T/T → high expression

T/A → intermediate expression

A/A → low expression

genotype inference

Harmanci A, et al. 2016. *Nature Methods 13*:251-256

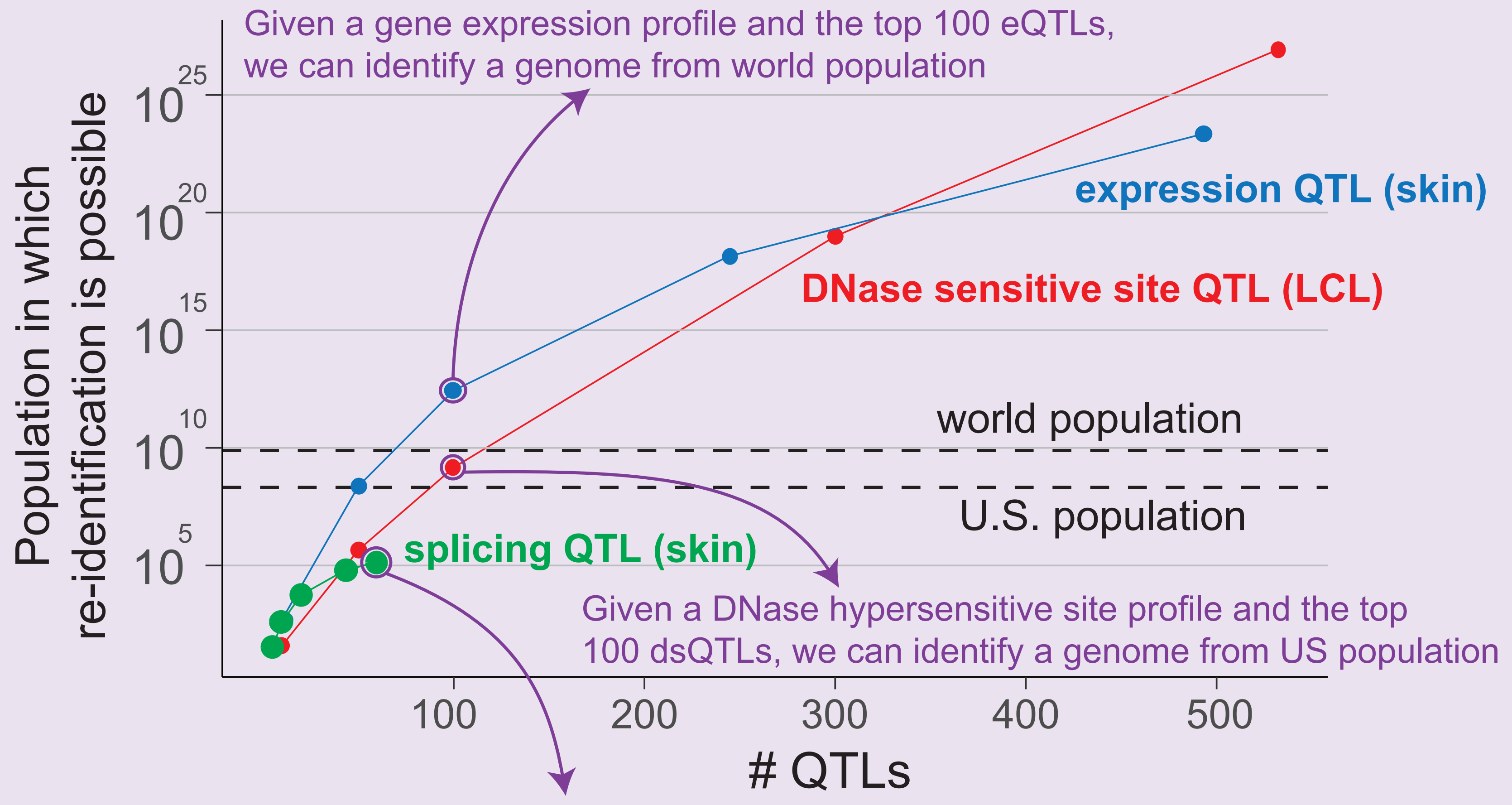
Genotypes can be predicted from gene expression values via eQTLs.

For example, an eQTL can be a common SNP located in an enhancer region upstream a gene that impacts expression. Given the eQTL, one can predict likely expression levels. Harmanci and colleagues showed how this process can be inverted to compromise privacy. In this example, a high expression value would lead to the prediction that the eQTL genotype is likely to be T/T.

Contact: hu.zhiqiang@berkeley.edu; brenner@compbio.berkeley.edu

Fund: NIH (U01 EB023686, U41 HG0077346), Tata Consultancy Services

Many types of omics data can be uniquely matched to a genome from a genome library, the size of world population

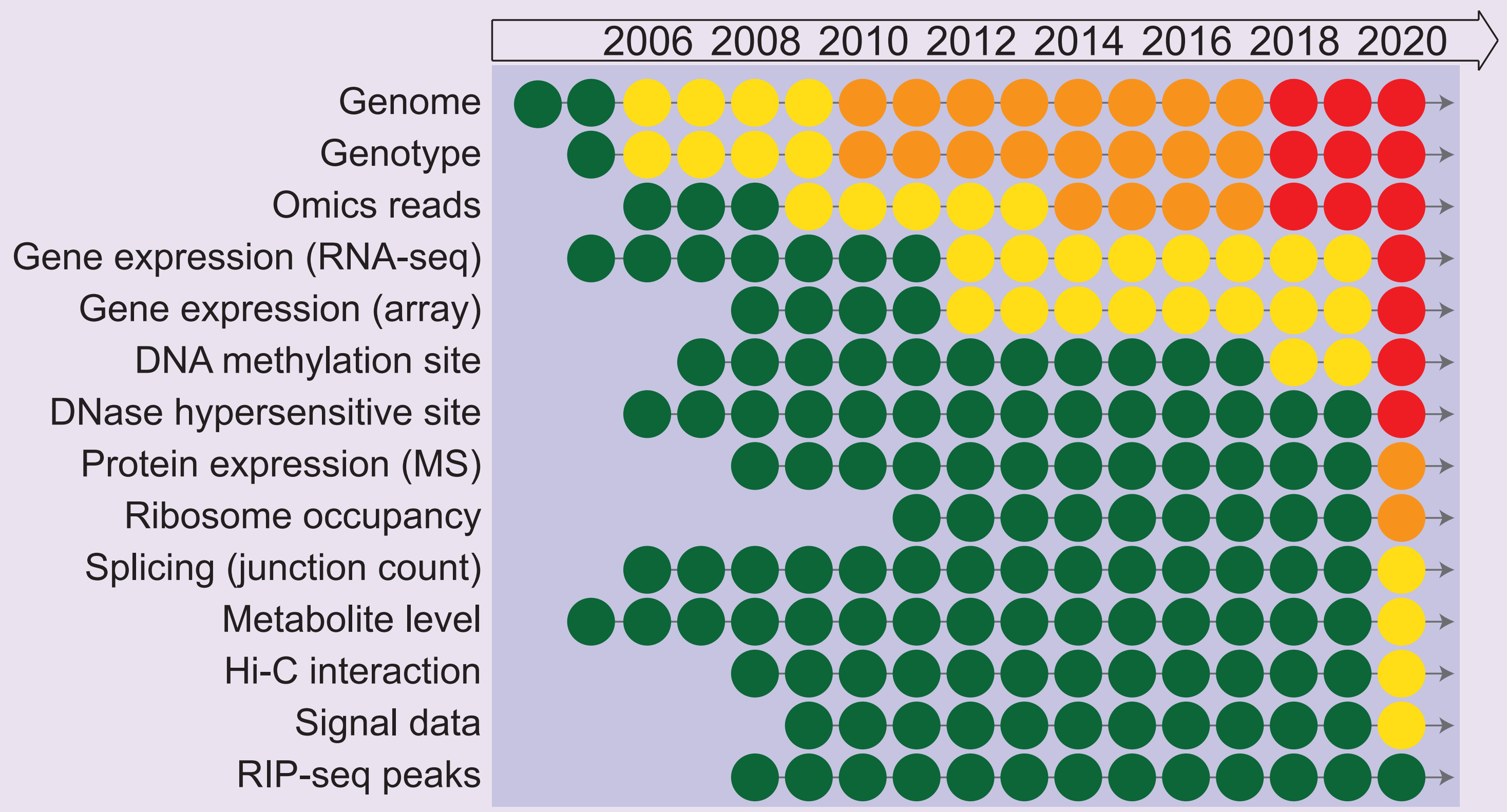


Given a gene expression profile and the top 100 eQTLs, we can identify a genome from world population

Given a DNase hypersensitive site profile and the top 100 dsQTLs, we can identify a genome from US population

Given a splicing profile and public sQTLs, we can identify a genome from a set of >100,000 genomes.

Hidden privacy risks in functional genomics data manifest over time, due to new data, new discovery and new techniques



High-throughput functional genomics data previously considered unlikely to compromise privacy may pose risks today or in the future.

For example, gene expression data were long considered safe to share without restriction. Concerns were raised the ability to link expression data to genomes via eQTLs were initially reported. But the risk at that time had been considered low, both because the linking ability was estimated to be limited, and because only a small number of genomes were readily available to match against. Our results here show the linking ability is now potentially high and the risks can be activated by consumer genealogy databases. Therefore, sharing gene expression level data today has potentially significant privacy issues.

It is impossible to know what additional risks may accrue in these and other biological research data over time.