Genetic Alliance Privacy and Progress Webinar Series

Technology – Is a Deeper Dive Needed

Debra JH Mathews, PhD, MA
8 October 2013
1. Culture of openness
2. WGS at $100-1000
3. Social media/networks
   • Evolving notions of privacy
   • Control over data
4. Identifiability
Cost per Genome

Moore's Law

NIH National Human Genome Research Institute

genome.gov/sequencingcosts
We’re crowd-sourcing health discovery by helping anyone create group health studies.

Ready to discover your family story?
Simply start with yourself and we’ll do the rest. Find your ancestors’ stories. Discover your identity.

Learn valuable health & ancestry information.

- Reports on 240+ health conditions and traits
- Testing for 40+ inherited conditions
- Discover your ancestry composition
- Updates on your DNA as science advances

$99

Order Now

Live better, together
making healthcare better for everyone through sharing, support, and research

learn  connect  track
Identifying Personal Genomes by Surname Inference

Melissa Gymrek,1,2,3,4 Amy L. McGuire,5 David Golan,6 Eran Halperin,7,8,9 Yaniv Erlich1,*

Sharing sequencing data sets without identifiers has become a common practice in genomics. Here, we report that surnames can be recovered from personal genomes by profiling short tandem repeats on the Y chromosome (Y-STRs) and querying recreational genetic genealogy databases. We show that a combination of a surname with other types of metadata, such as age and state, can be used to triangulate the identity of the target. A key feature of this technique is that it entirely relies on free, publicly accessible Internet resources. We quantitatively analyze the probability of identification for U.S. males. We further demonstrate the feasibility of this technique by tracing back with high probability the identities of multiple participants in public sequencing projects.

Surnames are paternally inherited in most human societies, resulting in their cosegregation with Y-chromosome haplotypes (1–5). Based on this observation, multiple genetic genealogy companies offer services to reunite distant patrilineal relatives by genotyping a few dozen highly polymorphic short tandem repeats across the Y chromosome (Y-STRs). The association between surnames and haplotypes can be confounded by nonpaternity events, mutations, and adoption of the same surname by multiple founders (5). The genetic genealogy community addresses these barriers with massive databases that list the test results of Y-STR haplotypes along with their corresponding surnames. Currently, there are at least eight databases and numerous surname project Web sites that allow individual users to look up the surnames of other users.

By combining other pieces of demographic information, such as date and place of birth, they fully exposed the identity of their biological fathers. Lunshof et al. (10) were the first to speculate that this technique could expose the full identity of participants in sequencing projects. Gitschier (11) empirically approached this hypothesis by testing 30 Y-STR haplotypes of CEU participants in these databases and reported that potential surnames can be detected. [CEU participants are multigenerational families of northern and western European ancestry in Utah who had originally had their samples collected by CEPH (Centre d’Etude du Polymorphism Humain) and were later reconsented to participate in the HapMap project.] However, these surnames could match thousands of individuals, and the study did not pursue full re-identification at a single-person resolution.

Our goal was to quantitatively approach the question of how readily surname inference might be possible in a more general population, apply this approach to personal genome data sets, and demonstrate end-to-end identification of individuals with only public information. We show that full identities of personal genomes can be exposed via surname inference from recreational genomics data sets.
1. Culture of openness
2. WGS at $100-1000
3. Social media/networks
   • Evolving notions of privacy
   • Control over data
4. Identifiability
Ethics
Genetics & Privacy

Privacy: “a condition or state in which cognitive access to personal information is restricted” (Powers, 1997)

- Why do we care about genetic privacy?
  - This information can be used both for good and bad purposes
  - Social and economic well-being
  - Psychological interests
  - Relational interests
  - Autonomy interests
Genetics & Privacy

Ethics

- **Training**
  - Make sure people touching the data appreciate and respect the power they have and behave appropriately/ethically

- **Consent**
  - Scrap identifiability as a determinant of consent
  - Make clear that privacy cannot be promised

- **Misuse**
  - Beef up tracking and governance of data use
  - Penalize data misuse
Control
- Shift to more participatory models of research that allow ongoing engagement of and control by, research participants.
Why control?
- Informed consent was devised as a way to protect/foster autonomy
- We now have the tools necessary to develop truly participatory models
- Norms of privacy have shifted, as have expectations of control
- The research community needs to earn the trust of the public
- Agency may offset the risks
- It’s the right thing to do