Thoughts related to integration of genomic information with clinical phenotypes & issues related to data privacy

Mark Gerstein, Yale

Thoughts related to integration of genomic information with clinical phenotypes & issues related to data privacy

- Charge: For NHGRI Computational Genomics & Data Science Workshop
  - Challenges in enabling new clinical insights
  - Helping NHGRI with planning for extramural computational and data science portfolio

- Overall recommendation
  - For future genomics, we **need to be able to do data sharing to enable large-scale data mining but still maintain individual privacy**
  - **We need to develop enabling technologies to allow this**

Background:
The Conundrum of Genomic Privacy

- The Genome is very fundamental data, potentially very revealing about one's identity & characteristics
- **Identification Risk**:
  Find that someone participated in a study [Craig, Erlich]
- **Characterization Risk**:
  Finding that you have a particular trait from studying your identified genome [eg Watson ApoE status]
- Genetics has a problematic ethical history & might need to be particularly careful to keep public trust
  - A single bad event (eg **HELA** like) could cause great damage to genomics research

- Sharing helps research
  - **Large-scale mining** of this information is essential for medical research
  - For statistical association, **1M is better than 1K**
  - **Privacy is cumbersome**, particularly for big data
- Sharing is important for **reproducible research**
- Data sharing & doing research on identifiable individuals is useful for **education**
  - More interesting to study the genome of someone you know

[Yale Law Roundtable ('10). Comp. in Sci. & Eng. 12:8; D Greenbaum & M Gerstein ('09). Am. J. Bioethics; D Greenbaum & M Gerstein ('10). SF Chronicle, May 2, Page E-4; Greenbaum et al. *PLOS CB* ('11)]

[Klitzman & Sweeney ('11), J Genet Couns 20:98l; Greenbaum & Gerstein ('09), New Sci. (Sep 23)]

Genomics has similar "Big Data" Privacy Issues in the Rest of Society… or does it ?

- Sharing & "peer-production" central to success of many new ventures, with the similar risks as in genomics
  - **EG web search**: Large-scale mining essential, but we confront privacy risks every day we access the internet
- Or is the genome exceptional?
  - We can't change it
  - The individual (harmed?) v the collective (benefits)
    - (Do sick patients care about their privacy?)
- **Different cultures** of genomics (open-data) & clinical medicine (private data)
- Genome is inherited so privacy risk is multi-generational
  - **What you can mine from a genome now is not clear, but what about in 20 years**

[Seringhaus & Gerstein ('09), *Hart. Courant* (Jun 5); Greenbaum & Gerstein ('11), *NY Times* (6 Oct)]

Current Social & Technical Solutions

- Notion of **Consent**
- "Protected" distribution of data (**dbGAP**)
- Local computes on secure computer

- Practical Issues
  - Non-uniformity of consents & paperwork
    - Different international norms, leading to confusion
  - Encryption & computer security creates **burdensome** requirements on data sharing & large scale analysis
  - Many schemes get "**hacked**"

[Greenbuam et al ('04), Nat. Biotech; Greenbaum & Gerstein ('13), The Scientist]

How to Surmount the Dilemma

- One approach is **just using open data**
  - Maybe we need a few "test pilots" (ala PGP)?
    - Sports stars & celebrities?
  - Some public data & data donation is helpful but is this a realistic solution for an unbiased sample of ~1M

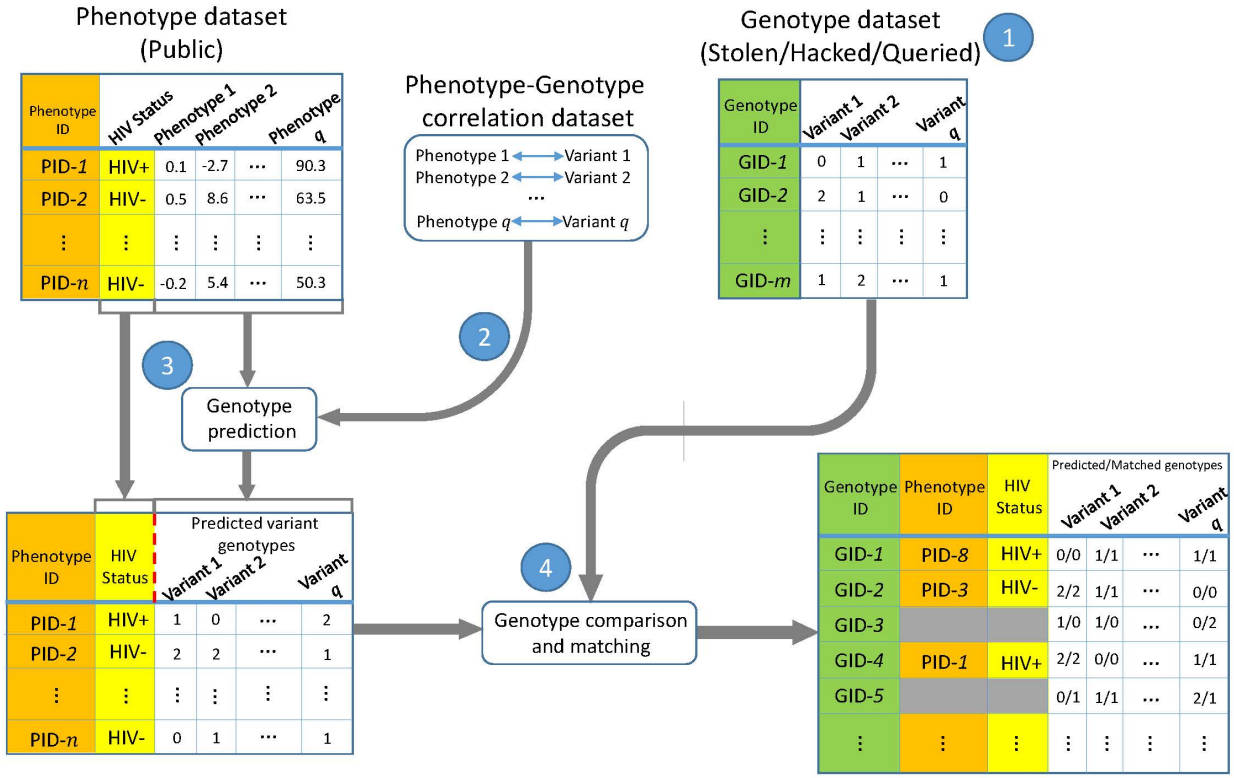
- **Hybrid Social & Tech Solution** (a strawman)
  - Fundamentally, researchers have to keep genetic secrets
  - Legal framework for this (not charge of this group)
    - Genetic licensure & training for individuals (similar to medical license, drivers license)
    - Barriers shouldn't create a incentive for “hacking”
  - Enabling Technologies
    - We should develop technologies for Quantifying Leakage & allowing a small amounts of it
    - We should develop technologies for the careful separation & coupling of private & public data

[D Greenbaum, M Gerstein (‘11). Am J Bioeth 11:39. Greenbaum & Gerstein, The Scientist ('13)]

More on Enabling Technologies

- **Making secure cloud computing easier**
  - Standard & open **workflow systems** cloud computing & enclaves (eg solution of Genomics England)
  - Homomorphic encryption, Differential privacy
- We should develop technologies for quantifying privacy
  - What is acceptable risk? What is acceptable data leakage? **Can we quantify leakage?**
    - Ex: photos of eye color
  - Cost Benefit Analysis: how helpful is identifiable data in genomic research v. potential harm from a breach?
- Separating but Linking Public & Private Data
  - Lightweight, freely accessible secondary datasets coupled to underlying variants (eg gene expression quantifications)
  - Selection of stub & "test pilot" datasets for benchmarking
  - Developing standards for **developing code on public stubs on your laptop**; then move programs to the cloud for private production runs

Example: Quantifying Information Leakage in RNA-seq Data & Providing guidelines on anonymizing public DBs



- Quantifying the **information from a rare SNP** (eg log[1/freq]) and information-theoretic predictability of an eQTL
- Showing how **a small but defined amount of leakage allows the genomic equivalent of the NETFLIX linking attack**
  - A small but defined number of strong eQTLs have enough information to statistically link a few SNPs (eg from a wine glass) to a record in an anonymized public gene expression database and then onto a private phenotype (eg HIV+)

[Harmanci & Gerstein, Nat. Meth. ('16)]

Further Thoughts on presenting genomic data to clinicians & more mainstream audience

- We need to simplify genomic data
- We need to make NHGRI products easier to use for a larger public
  - Particularly for noncoding regions of the genome
- Need to more clearly integrate human data with relevant data from model organisms