

CAGI slide pack guidelines

I agree to my slides being used according to the following terms:

Madeleine Price Ball

- Authors: please include THIS slide at the beginning of your presentation to indicate that you agree with reuse of your slides in CAGI presentations, according to the terms below, and digitally sign by writing your name above.
- To disseminate the CAGI results, I encourage all CAGI participants to remix these CAGI slides for inclusion in their oral and poster presentations
- CAGI participants may use or publish any content of these slides only in compliance with the CAGI Data Use Agreement
- CAGI participants may not cite or publish any content of these slides except with the written permission of the originating author or in the primary CAGI publications
- CAGI participants must acknowledge the original authors of these slides and CAGI 2013. Include the acknowledgement banner across the bottom
- CAGI participants must include the CAGI credits slide in their presentations
- Slides with a red slash (no remix) may not be included in a presentation unless all attendees are registered CAGI participants who signed the data use agreement
- Slides with a red X over the entire slide may not be used in any circumstances
- Slides with a blue slash (embargoed) may not be used unless the embargo is explicitly lifted
- Slides without slash or X may be tweeted unless flagged with the “No tweet” icon.
- CAGI participants using these slides must explain the “No tweet” icon, which means that the slide content should not be disseminated outside the presenting conference hall. If the presentation venue permits pervasive tweeting, you may not include these slides in your talk
- Note for slide authors: we expect all CAGI participants to abide by these restrictions, and will make best effort to ensure they are followed. However, but bear in mind that we have limited means of enforcing them, and therefore the restrictions cannot be guaranteed.

About the Harvard Personal Genome Project Data & CAGI 4 Challenge

Madeleine P Ball, Abram Connelly, Preston W Estep, Nancy Ouyang,
Ward Vandewege, Alexander W Zaranek, and George M Church



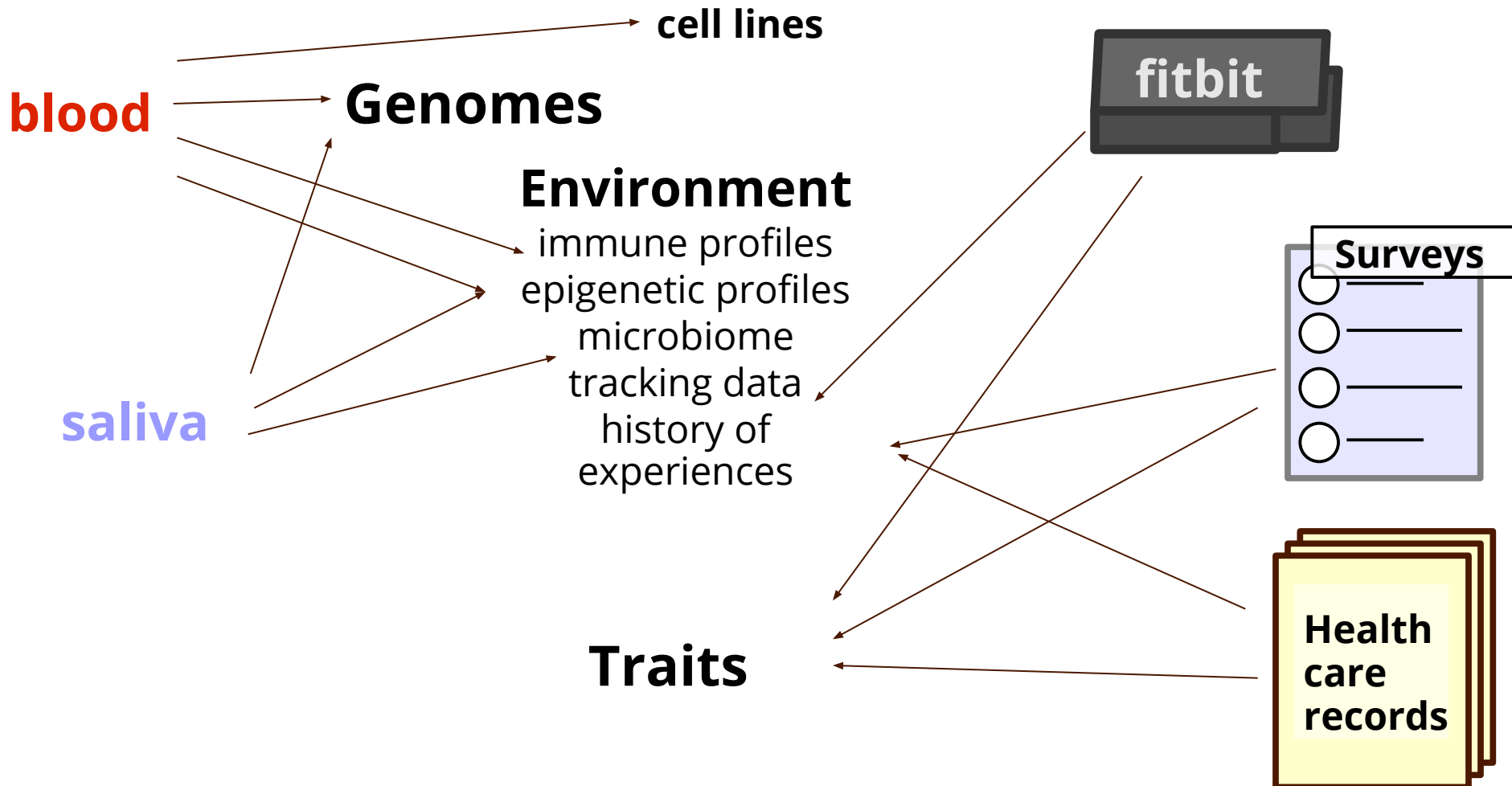
CAGI March 27 2016

I. About the Personal Genome Project

II. The CAGI 4 data & challenge

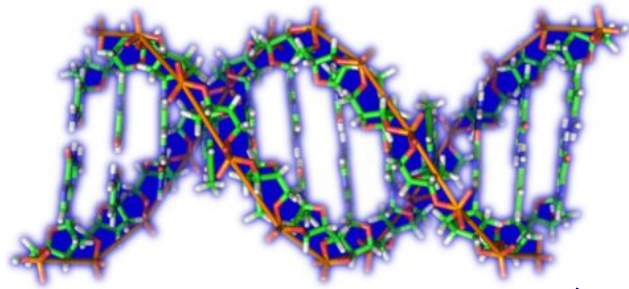
About the Harvard
Personal Genome Project

Understanding needs **more than genomes**

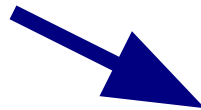


But **granular data** is generally a
a **privacy concern**.

Many approaches for **re-identification**,
especially combined with other data.



predictive



identifiable

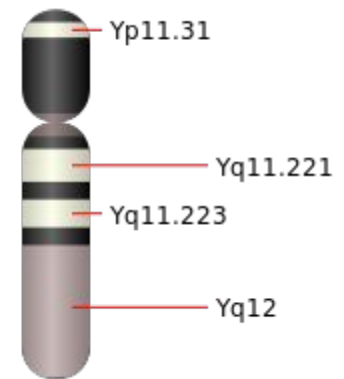


```
1000000000011
0011100011001
0010010011001
1000000000001
1011100111110
0100111110000
```

easy to digitize  **(and thus easy to share)**

Y chromosome = Surname

- Gymrek M, McGuire AL, Golan D, Halperin E, and Erlich Y
“Identifying Personal Genomes by Surname Inference”
Science 18 January 2013: Vol. 339 no. 6117 pp. 321-324
- How to:
 - Infer Y chromosome STR data
 - Compare to ancestry database
 - Get surname guesses
 - Compare to public records
 - Hint – study design gives time/region
- Re-identified 50 HapMap individuals



**Ancestry inference from genomes
leads to high identifiability.**

Data Sharing vs. Privacy

- Obfuscation is hard (maybe impossible)
- “Controlled access” databases:
tension between privacy and sharing
- But we want science to be **open**

Some people are willing to share,
despite the privacy risk!



The Personal Genome Project



Initiated by George Church in 2005 the PGP aims to advance genome understanding by **sequencing and publicly sharing genomes** along with health and other diverse data.

- **Participants can add data additional public data**
(health records, surveys, name)
- **Identification is an acknowledged risk**
(not required – but some choose to share their name)
- **Participants must demonstrate understanding**
(enrollment exam)

Highly educated cohort, enriched for academics, researchers, and citizens scientists.

More communities creating open data without privacy assurance

The Harvard PGP is no longer alone in this approach...

- **Harvard Personal Genome Project (2005)**
Whole genome sequencing, health data, surveys
- **OpenSNP (2011)**
Donated genetic data, activity trackers, surveys
- **PGP Canada (2012), PGP UK (2013), Genom Austria (2014)**
Whole genome sequencing & additional data
- **Open Humans (2015)**
Donated genetic data, genotyping, microbiome, viral profiling, activity trackers, GPS maps

The CAGI 4 PGP **data set & challenge**

Sample collection logistics

- **Blood collection**

Blood is the easiest sample for pure human gDNA.

Saliva is easier for collection but has >50% bacterial DNA.

That's okay for genotyping microarrays (e.g. 23andMe and AncestryDNA), the non-human sequences are ignored.

But for whole genome sequencing, it means >50% of reads are "wasted" on bacterial DNA.

- **Sample collection events in 2013 and 2014**

2013 and 2014 GET Conferences, and collection events in Boston/Cambridge, Mountain View, San Diego, St. Louis.

Who to collect samples from?

- **Only individuals who already shared trait data**

Getting individuals to respond can be a challenge.
This way we ensure this additional data already exists.

(Otherwise we risk sequencing the genome and never getting other data – a bad use of sequencing resources!)

- **Trait data immediately public**

Public trait data demonstrates commitment to public data.
(It's expensive for a participant to "change their mind"!)

Also consistently communicates "public data".
We need to ensure participants understand their genome and trait data will be public.

Genome sequencing

- **Sequencing is in bursts, based on available resources.**
- **These samples were sequenced by the research group at Complete Genomics, Inc.**

Samples were sequenced as part of Complete Genomics' "long fragment read" technology development.

Data which can become public enables publication of their research. (They're currently working on a publication for this data.)

Data release

- **Data return to participants**

PGP protocol: participants to have 30 day private access. If they withdraw during this time, the data stays private.

- **Participants also receive a "preliminary research report"**

Interpretation highlights some "high penetrance" effects published about some variants in the genome. Notably, these published claims are often incorrect and contradicted by allele frequency information (e.g. ExAC).

- **Genome publicly shared**

30 days later, the genome is (normally) automatically made public on the participant's public profile.

CAGI 4 challenge

- **Unlinked genome / trait data**

For the purposes of CAGI 4, the Harvard PGP temporarily changed the method of making genome data public.

Instead of releasing genomes directly to profiles, genome data was public – but not linked to profiles.

- **Match genomes to existing trait profiles**

A challenge to "predict traits from genomes" wasn't wanted because trait profiles could be used to create predictions.

So, the challenge was instead: match to existing trait profiles.

Acknowledgements



Harvard Personal Genome Project

- George M Church (PI)
- Jason Bobe
- Michael F Chou
- Abram Connelly
- Preston (Pete) W Estep
- Jeantine Lunshof
- Nancy Ouyang
- Celia Fulton Walden
- Ward Wandewege
- Alexander (Sasha) Wait Zaranek

Complete Genomics, Inc.

- Radoje Drmanac
- Brock Peters