Project B)

Standards for Reusable Genomic Resources

Jason Bobe
Executive Director at the Personal Genome Project

This session will explore ideal structures for the creation of reusable genomic resources.

An example of a resource that we want to be able to reuse, as individuals and as researchers, is a whole genome dataset. We want to move away from research models where genomic datasets are born into silos only to languish there under the lock and key of some narrow, pre-defined purpose and/or a territorial principal investigator. To reach our translational goals, we want genomic datasets to have some legs: we want data liquidity and long-term utility.

Data liquidity will be achieved when everyone has in their control a digital copy of their whole genome and they are free to move it from one context to another. This moment in history is near at hand, ushered here by the inexorable decline in costs of whole genome sequencing. Genome liquidity will arrive from multiple paths, including (1) individuals obtaining genomic data from research studies that, as part of study protocol, return genomic data to participants for use in other contexts and (2) individuals obtain genomic data from elsewhere (e.g. healthcare, commercial DTC, DIY) and donate it to research.

It is not necessarily true that genomic data generated in one context will be acceptable for use in another. In the PGP, we have both returned genomic datasets to participants and accepted donations of genomic datasets that participants have acquired from third parties. So far, we've learned that:

- Provenance matters: To state the obvious, meaningful differences exist between datasets generated from different platforms and protocols. Having protocols that allow for variants to be confirmed is helpful. For example, confirming through an independently collected and evaluated saliva specimen that a donated genomic dataset does in fact belong to person who signed the consent form.
- Curation is needed: Genomic data formats are variable and not always easily comparable. Even "premium" genome sequencing services available today can generate idiosyncratic data which needs to be re-organized prior to re-use.
- Variant evaluation: We have found that cross-referencing variants with published literature generates questionable findings, e.g. literature that predicts early onset, severe disorder in apparently healthy middle aged person.

We want a data commons that is able to reduce the activation energy required for a researcher to trust that the genomic datasets it contains is sufficiently vetted to jump in and invest her energy in making a reproducible discovery. Two important aspects of buiding a commons of reusable genomic resource are:

Transparency of Data Generation: A genomic dataset is generated from a variety of technologies, software tools and practices strung together in some protocol. The details of the protocol may play a major role in the “re-usability” of the genomic dataset in other contexts, such as a research study, healthcare setting, web application, or shared repository of data. This protocol should be transparent.

Governance: Having well-structured data with sufficient meta-data is only part of the challenge. Informed consent, data licenses and policies governing bi-directional communication between participants and investigators will also impact the long-term utility of a genomic dataset. The PGP has organized one governance regime (“open consent” and “CC0”) that attempts to maximize use and minimize restrictions that has been extended and standardized in the PLC.

In this session, we will explore these two structures in detail and attempt to define the components in an ideal genomic data commons.

Over to You

B- Standards for Reusable Genomic Resources

G	Organization	Last name	First
B	Duke University	Angrist	Misha
B-Lead	Personal Genome Project	Bobe	Jason
B	Common Causes Policy Advise	Bloemen	Sophia
B	Legal Pathways	Bovenberg	Jasper
B	University of Washington	Dorfman	Lizzie
B	Life Science Discovery Fund	Hertle	Mark
B	Stanford	Kasowski	Maya
B	Lilly	Krohn	Thomas
B	The Methodist Hospital Research Institute	Li	Fuhai
B	Coalition for Networked Information	Lynch	Clifford
B-Anchor	Sage Bionetworks	Mecham	Brig
B	SPARC	Rossini	Carolina
B	University of Wisconsin	Saha	Kris
B	Gladstone Institute	Salomonis	Nathan
B	UCSF	Sittler	Taylor
B	Notre Dame	Siwo	Geoffrey
B-Anchor	Sage Bionetworks	Wilbanks	John

Creating Reusable Genomic Resources
Jason Bobe, PersonalGenomes.org

Project Overview

Our goal is establish a set of requirements for creating re-usable genomic resources to grow the commons.

"re-usable" is defined as a genomic dataset generated in one research context can be used by others

Move away from data silos created by:
(1) Territorial investigators
(2) Governane Issues: privacy, consent restrictions, etc
(3) Lack of technical infrastructure for sharing

Potential alignment with existing Commons' approaches

List examples from other efforts that could be applied to this project.
Examples around Governance, Incentives, Platform, etc.

Policy:
Open access mandate NIH/Wellcome Trust by funders, e.g. 50% compliance now.

Technical Facilitation:
Blue Button Plus: An example of data liquidity has worked,
we could piggy back by adding an area for genomic files.

Standard Setting:
NIST "Genome in a Bottle"

Unmet needs and issues

Potential commons approaches needed but not yet built and remaining issues to be highlighted

Data liquidity: address barriers that prevent researchers from sharing data w/ participants and commons.

Governance: Expand PLC

Facilitation: Technical tools for making easy for data transfer, e.g. +1 button for sharing

Technical: Metadata standards for genomic datasets and provenance.

Communication Issues: How to manage longitudinal return of findings of data donated to a commons.

Project B

1-year vision for the future of this project

What is the first thing to do to align participants, researchers and funders.

Playbook for open clinical studies.

Use the PGP/NIST to generate basic standards. Piggyback on NIST "genome in bottle" effort and use that opportunity to specify a minimum metadata standards.

We want a template to give funders for making sure their investments in research are compliant with the commons and then giving them a compliance

Alignment:

Make sure that existing open data efforts, like Sage and the Personal Genome Project, are interoperable.

Project B