Deriving Drug Discovery Value from Large-Scale Genetic Bioresources

Designing Cohorts to Maximize Discovery Capabilities

March 22, 2016

Joe Vockley, PhD

Inova Translational Medicine Institute

Importance of Study Design

Studies At ITMI

- Study design integrates physicians, genomic scientists and bioinformatic scientists.
- Standardize design across studies
- Complete data sets: Trio WGS, expression, DNA Methylation, miRNA
- Ancestral component
- Clinical, family history, and survey data (including nutrition, behavior and environmental data)

Improvements

- Randomness in study design inhibits ability to compare data.
- Study design needs to include a plan for biobanking through data analysis utilizing best practices.
- Centrally developed standard protocols that must be followed to obtain external funding?

Biobanking

Banking Samples at ITMI

- Version controlled SOPs
- Minimize warm and cold ischemic time
- Minimize time to stabilization
- Correct and constant storage temperature
- SNP chip upon deposition
- Barcoding
- Document metadata

- The value of banked specimens is enhanced by longitudinal samples with matched clinical and outcomes data.

Improvements

- Fund community hospitals for broad spectrum biobanking?
- Standardized “opt out” consent for all hospitals.
- National standards for biobanking required to obtain funding.
- QC check of SOPs and samples.
- Education of physicians, hospital administration and patients on importance of biobanking.

Data Generation

ITMI Data

- Data generated in HTP environment
- Optimize and stabilize protocols for biomolecule isolation
- Protocol stabilization for data generation
- Standardized data processing
- Record metadata

Improvements

- Most groups/companies that utilize Illumina sequencing technology modify the data generation protocols.
- Few groups/companies properly document or disclose modifications.
- These modifications significantly change the data, making it difficult to integrate and interpret.
- 150K biomarker manuscripts: 100 Biomarkers
- Biomarkers can’t validate or be reproduced

Data Integration

Hybrid Cloud at ITMI

- Cloudera-Hadoop Hybrid Cloud.
- Multi-omic integration.
- Integrate -omic data with pathway, ancestral, familial, clinical and network data.
- Individual data sets that do not achieve statistical significance can be combined to support each other in the context of pathways or genomic networks

Improvements

- Nextgen genomic data processing and analysis fundamentally changes the data.
- Few fully disclose how data are changed during analysis.
- Data sets are incomplete making integration and interpretation impossible.
- Data deposited into the public domain does not contain sufficient metadata and standardization to facilitate target discovery.

ITMI's Data Environment

- Standardization of all processes from biobanking to data analysis with complete metadata.
- Cloudera-Hadoop Hybrid Cloud – System of open source components that enables multiple simultaneous analytic workloads.
- SGI UV 2000 with 1000 cores, 16 TBytes RAM, 5TBytes Flash and 17 TFlops (10^{12} floating point operations/second) of analytics across multiple threads integrated with Amazon elastic analytic (EC2) space.
- Multi-PByte onsite storage integrated with Amazon S3 and Glacier storage.
- Phenotypic Database, Genotypic Database, MAF Database.