Second Multi-IC Symposium Working Group 4:

Data Sharing in NIH GWA Studies

Sean Coady, NHLBI
Daniela Gerhard, NCI
Teri Manolio, NHGRI
Jim Ostell, NLM
Rebekah Rasooly, NIDDK
Andy Singleton, NIA

Data Sharing in NIH GWA Studies

- Available databases for submitting and obtaining data
- Challenges in receiving and coding forms/ protocols/individual data
- Quality control of submitted phenotype/ exposure data
- Approaches to facilitating access to DNA samples
- Types of genotyping data to be distributed
- Integrating next generation of genome-wide data on backbone of GWA data

Available Databases for Data Sharing

- caBIG/CGEMS (https://caintegrator.nci.nih.gov/cgems/)
  - Open access: Allele frequencies, associations
  - Controlled access: Individual gt/pt data
- dbGaP (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gap)
  - Open access: protocols, association findings
  - Controlled access: Individual gt/pt data
- NIA Genetics Initiative (http://www.niageneticsdata.org/)
- NIA Genetics of Alzheimer Disease Data Storage Site (http://zork.wustl.edu/nia/)
- NIDA Center for Genetic Studies (http://zork.wustl.edu/nida/)
- NIMH Human Genetics Initiative (http://zork.wustl.edu/nimh/)
- NINDS Human Genetic Resource Center (http://ccr.coriell.org/ninds/pd/pd.html)

Challenges in Receiving and Coding Forms, Protocols, and Individual Data

- Need for completely re-entering hard-copy forms or PDFs (ARIC, Framingham)
- Effort and expertise needed to recode forms and key them to protocols
- Maintaining data security (data transmission, storage, de-identification)
- Challenges of different levels of consent for different groups of participants
- Accommodating participants’ decisions to withdraw

Quality Control of Submitted Data: Focus on GAIN and GEI Experience

- Data submissions and NCBI-generated summaries from GAIN
  - Difficulties with IRBs in submitting actual data
  - High rates of missing data
  - Fewer elements or participants than promised
- Investigator-generated summaries from GEI
- Subsequent data submissions from GEI studies considered for funding
- Use of data summaries in peer review
- Quality control of genotyping data

Types of Genotyping Data to be Provided

- Methods: description of calling algorithms
- Genotype quality data for each SNP
  - Quality scores and thresholds
  - Concordance rates and call rates
  - Hardy-Weinberg equilibrium statistics
  - Q-Q plots for population heterogeneity
- Allele and genotype calls for each SNP for each participant
  - Intensity files and cluster plots for each SNP
- Calls that don’t reach quality thresholds
- Some assessment of sample quality

Integrating Next Generation of Genome-Wide Data on GWA Backbone

- Dealing with different releases of chips
- Copy number variants
- Sequence variants
- Expression
- DNA methylation

Approaches to Facilitating Access to DNA Samples

- Need standard, user-friendly approaches to identifying availability of samples and procedures for obtaining them
- Probably extends to other biospecimens such as lymphocytes, transformed cells
- Importance of assessing quality of DNA specimens prior to award– GAIN and GEI experience
- Removing identifiers from samples