Trans-NIH Workshop

Establishing a Central Resource of Data from Genome Sequencing Projects

"Data Aggregation"

Eric Green, M.D., Ph.D.
Director, NHGRI

Major Challenge: Data Access

Victims of Our Own Success:

- ❑ Data production (genome sequencing)
- ❑ Core value: Open data release/access
- ❑ 'Data rich' but 'analysis poor'

Current systems need streamlining

'Research Commons' – certified users can access all data sets at once

'Central Data Servers' – provide results and summaries but not underlying data

Other Challenges

Approaches to variant calling

Harmonizing phenotype and environment data across studies

Computing on very large data sets (component of 'Big Data' problem)

Workshop Planning

Michael Boehnke and Wylie Burke, Co-Chairs
David Altshuler and Paul Flicek

NIH staff (10 ICs):

Lisa Brooks, Adam Felsenfeld, Teri Manolio

David Cabrera, Hemin Chin, Peter Good, Emily Harris, Allison Lea, Thomas Lehner, Nicole Lockhart, Catherine McKeon, Alan Michelson, Erin Ramos, Laura Rodriguez, Winifred Rossi, Kenna Shaw, Stephen Sherry, Heidi Sofia

Logistics: Nicholas Clemm, Saundra Bromberg

Workshop Participants

47 non-NIH staff, 40 NIH staff

Expertise areas:
- ➤ ELSI and policy
- ➤ Genomic and functional data
- ➤ Disease studies, GWAS, & cohorts
- ➤ Database, computation, & data analysis
- ➤ Drug development
- ➤ Scientific publication

Workshop Goals

Discuss the scientific questions that analyses across these data sets could address

Discuss the challenges to obtaining and analyzing across many data sets

Discuss the options for dealing with these challenges (including costs and tradeoffs)

Recommend steps to address these challenges

Questions to Consider

How should we deal with data already collected (retrospective) and with data to be collected (prospective)?

What options are high priority?

What are the cost-benefit tradeoffs?

What can be done now, and what would require changes in policies or other work?

NOTE: 90% solutions would still be progress!

Questions for NIH Leadership?