Establishing a Central Resource of Data from Genome Sequencing Projects

Scope of the NIH Data Sets

June 5/6 2012
### Inventory of NIH Projects

**Projected through end of ~2012**

http://www.genome.gov/27545796

<table>
<thead>
<tr>
<th>Institute or Center</th>
<th>Project</th>
<th>Whole Exome (yes/no)</th>
<th>Whole Genome (yes/no)</th>
<th>Depth of Coverage (3X, 20X, etc)</th>
<th>Sequencing Platform</th>
<th>Date Sequencing Initiated</th>
<th>Date Sequencing (to be) Completed</th>
<th>Date Data to be Shared</th>
<th>Venue for Data Sharing</th>
<th>Number of Individuals (to be) Sequenced</th>
<th>Criteria for Selection of Individuals</th>
<th>Primary Phenotype</th>
<th>Age Range (yr)</th>
<th>Race/Ethnic Comp</th>
</tr>
</thead>
<tbody>
<tr>
<td>NCI-CCR</td>
<td>DRCT Whole Exome</td>
<td>Yes</td>
<td>No</td>
<td>&gt;100</td>
<td>SOLiD</td>
<td>Jan-12</td>
<td>Mar-12</td>
<td>Upon Publication</td>
<td>dbGaP</td>
<td>25</td>
<td>Desmoplastic Small Round Cell Tumor</td>
<td>Child</td>
<td>Any</td>
<td></td>
</tr>
<tr>
<td>NCI-CCR</td>
<td>EWS Whole Genome</td>
<td>No</td>
<td>Yes</td>
<td>&gt;40</td>
<td>Complete Genomics</td>
<td>May-11</td>
<td>Sep-11</td>
<td>Upon Publication</td>
<td>dbGaP</td>
<td>6</td>
<td>Ewing’s Sarcoma</td>
<td>Child</td>
<td>Any</td>
<td></td>
</tr>
<tr>
<td>NCI-CCR</td>
<td>Genome structure of DLBCL</td>
<td>Yes</td>
<td>Unknown</td>
<td>Complete genomics</td>
<td>Mar-11</td>
<td>Jul-11</td>
<td>2012</td>
<td>dbGAP</td>
<td>2</td>
<td>One patient with ABC DLBCL and One with GCB DLBCL</td>
<td>Development of DLBCL</td>
<td>38-67</td>
<td>Caucasian</td>
<td></td>
</tr>
<tr>
<td>NCI-CCR</td>
<td>GIST Whole Exome</td>
<td>Yes</td>
<td>No</td>
<td>&gt;100</td>
<td>SOLiD</td>
<td>Mar-11</td>
<td>Apr-11</td>
<td>Upon Publication</td>
<td>dbGaP</td>
<td>3</td>
<td>GIST with Family Member</td>
<td>Gastrointestinal Stromal Tumor</td>
<td>Any</td>
<td></td>
</tr>
<tr>
<td>NCI-CCR</td>
<td>HBL Whole Exome</td>
<td>Yes</td>
<td>No</td>
<td>&gt;100</td>
<td>SOLiD</td>
<td>Nov-11</td>
<td>Jan-12</td>
<td>Upon Publication</td>
<td>dbGaP</td>
<td>25</td>
<td>Hepatoblastoma</td>
<td>Child</td>
<td>Any</td>
<td></td>
</tr>
</tbody>
</table>
Inventory High Level Numbers:

By end of year:

• Projects: ~190
• Samples: ~68,800 (~18K as WGS?)
Growing.....

Current samples in dbGaP
Data Are Organized into Projects

Projects by sample size

~50K samples
Projects by Disease Area
“High-Value” Samples?

Out of ~68,800 total

~26,000 samples have no data use limitations. (~2600 are completely public…)

~15,900 samples - participant recontact is permitted (could *in principal* be re-phenotyped)

~Physical samples actually available – 25K?
Other Things We may Want To Know

- What phenotype data?
- What exposure data?
- What population?
- What age?
These data are all at NCBI/dbGaP, right? So what’s the problem?

~200 projects; ~400 Consent Groups; Go through multiple DAC’s to access; Inconsistent metadata, etc.
And, Not All the Data are in dbGaP

• Some of the NIH-funded data are in CGHub
• Disease-specific databases will proliferate
• More data outside US
Summary

• A lot of data: easy to imagine 100K+ samples and 1000 Tbases just from NIH by end of 2013. *What number of samples should we plan for? (1M?)*

• If we do nothing, also easy to imagine these being divided into many projects

• Large numbers of samples are concentrated in a few studies. If “high value” (eg., no restrictions, recontactable, etc.) then could be useful for aggregate analyses in short term. But far from optimal, and not scalable
Many Thanks To

Nicholas Clemm – Inventory digest and slides

Steve Sherry, Eugene Yaschenko, Martin Shumway – dbGaP summary data

Teri Manolio, Ian Marpuri – Inventory

Lisa Brooks – general
end
These data are (mostly) Available via dbGaP