Analysis tools and portals

Gabor Marth
Boston College
NIH Data Aggregation Meeting
Wednesday, June 5-6, 2012, Rockville, MD
Sequences in a central place, variant recalled or aggregated, metadata and phenotype harmonized ...

... what tools are needed make the data useful for the community?
What type of analyses can we do?

- Population genetic
  - Haplotype phasing
  - Single-variant allele frequencies
  - Variant burden
- Functional
  - Coding annotations
  - Disease databases
  - Non-coding annotations
  - Loss of Function analyses
- GWAS
  - Meta-analysis
  - De novo analysis
- Systems biology / higher order analyses
  - Network / pathway analysis

- Well-defined vs. open-ended
- Algorithmically easy vs. hard
- Tools mature vs. emerging
- Computation-heavy vs. light
Who are we trying to serve?

Statistician / tool developer
- How does my tool perform?
- What’s different about the non-responder?

Drug developer
- Do I get a bigger p-value from all the extra samples?

Biologist in small laboratory
- Is this variant associated with any known phenotype?

Medical consortium project analyst

Clinician
- I see a variant in this gene... should I alter the treatment of my patient?
How to make analysis accessible?

- Easy to install
- Easy to use
- Intuitive
- Fast
- Interactive
- Web-based
- Storage & hardware
- Informatics expertise
How can we provide the analyses?

- Raw data download
- Query portals, viewers, data slicers
- Static variant annotations, pre-computed resources
- Analysis environment with **central data**, facilities for users to add their **own data**, **tools installed** and **computational resources** to run the analysis
What static analyses make sense?

- Variants, variant allele frequencies, sample genotypes
- Phased haplotypes

- Variant consequence
  - Haplotype browsing
  - Phenotype browsing
What tasks require analysis services?

- Read mapping, and variant calling
- Imputation (using central haplotype resource)
- Custom functional annotations
  - Structural variants
  - Aggregate effect of phased variants

This category has the highest tool development cost because of the additional engineering required (Daniel McArthur)
One tool or multiple tools?

<table>
<thead>
<tr>
<th>Dataset</th>
<th>SNPs</th>
<th>FP metric</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tool 1</td>
<td>632,344</td>
<td>2.32</td>
<td></td>
</tr>
<tr>
<td>Tool 2</td>
<td>547,173</td>
<td>2.34</td>
<td></td>
</tr>
<tr>
<td>Tool 3</td>
<td>576,125</td>
<td>2.36</td>
<td></td>
</tr>
<tr>
<td>Tool 4</td>
<td>629,761</td>
<td>2.26</td>
<td></td>
</tr>
</tbody>
</table>

There are inherent advantages to having alternative tools available.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>SNPs</th>
<th>FP metric</th>
</tr>
</thead>
<tbody>
<tr>
<td>4 of 4</td>
<td>410,243</td>
<td>2.56</td>
</tr>
<tr>
<td>3 of 4</td>
<td>518,407</td>
<td>2.50</td>
</tr>
<tr>
<td>2 of 4</td>
<td>593,538</td>
<td>2.42</td>
</tr>
</tbody>
</table>
Centralized or distributed development?

- Tool development is iterative... once we get an answer, we want to ask a new question
- Often users are better served by light, flexible tools for customized analysis... a **tool ecosystem**
Who would develop the tools?

- Many current tools from large genome centers but the majority from smaller tool development groups.
- There is also a large and successful “cottage industry” of tool development, where small informatics groups can produce very sophisticated software, and respond nimbly to user needs.
How to move forward?

• Focus on the cloud
• Build an open environment for tool deployment to pull in the widest possible developer base
• Models and technologies exist (iPhone apps, etc.)