Analysis Pipeline Introduction

David Levine
August 9, 2017
Session Outline

• Computing environments
• Analysis pipeline architecture
• Demo on local cluster
• Demo on AWS Batch
• Wrap up
Generally not powerful enough

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Samples</th>
<th>Variants</th>
</tr>
</thead>
<tbody>
<tr>
<td>workshop</td>
<td>1,126</td>
<td>25,760</td>
</tr>
<tr>
<td>freeze.1c</td>
<td>2,643</td>
<td>112,275,224</td>
</tr>
<tr>
<td>freeze.2a</td>
<td>9,109</td>
<td>140,980,783</td>
</tr>
<tr>
<td>freeze.3a</td>
<td>16,558</td>
<td>185,970,832</td>
</tr>
<tr>
<td>freeze.3a.phased</td>
<td>18,258</td>
<td>140,980,783</td>
</tr>
<tr>
<td>freeze.4</td>
<td>18,526</td>
<td>219,154,455</td>
</tr>
<tr>
<td>freeze.5</td>
<td>60,000</td>
<td>470,000,000</td>
</tr>
</tbody>
</table>

- Memory
- CPU
- Disk space
Single server OK for small data sets
Workshop: Server access via AWS*
TOPMed data sets belong on a cluster

- Local Cluster?
- Cloud Cluster?
GENESIS on a cluster

Local cluster
- SGE
- Python
- GENESIS
- R

AWS
- CfnCluster
- Python
- GENESIS
- R

AWS
- AWS Batch
- Python
- GENESIS
- R

Google
- Dataflow
- Python
- GENESIS
- R

Azure
- Batch
- Python
- GENESIS
- R

current

future
Calling GENESIS in other applications

- DNANexus
- GENESIS
- R

- NIH Data Commons
- GENESIS
- R

current  
possible future