TCGA Meeting Workshop
Deniz Kural, PhD
May 12, 2015

+ Brandi, Zeynep, Devin, Kate
Amount of genomics data will exceed available resources

Between 2014-2018 production of new NGS data to exceed 2 Exabytes

NGS: Next Generation Sequencing
NGS sequencers include machines from Illumina, Life Technologies, and Pacific Biosciences. Human genome data based on estimates of whole human genomes sequenced
Sources: Financial reports of Illumina, Life Technologies, Pacific Biosciences; revenue guidances; JP Morgan; The Economist; Seven Bridges Analysis.
Large-scale cancer genomics will shift how computation is done

CLOUD COMPUTATION

PORTABLE WORKFLOWS
replace
data transfers
Standard Model: Data comes to computation

The Cancer Genome Atlas

Research Center 1

Research Center 2

Research Center 3
Cloud: Algorithms come to the data

Other awardees of the NCI Cancer Cloud pilots: Institute for Systems Biology, The Broad Institute
Millions of genomes will shift how computation is done

**CLOUD COMPUTE**

**PORTABLE WORKFLOWS**
replace data transfers
Ceci n’est pas une pipeline!?
Replicating computations is a hard problem

<table>
<thead>
<tr>
<th>Feature</th>
<th>Scripts</th>
<th>Machine Images</th>
<th>App containers + Pipeline-Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>Easy distribution</td>
<td>🔄</td>
<td>🔄</td>
<td>🔄</td>
</tr>
<tr>
<td>Simple to run</td>
<td>🔄</td>
<td>🔄</td>
<td>🔄</td>
</tr>
<tr>
<td>Compatible across environments</td>
<td>🔄</td>
<td>🔄</td>
<td>🔄</td>
</tr>
<tr>
<td>100% reproducible</td>
<td>🔄</td>
<td>🔄</td>
<td>🔄</td>
</tr>
<tr>
<td>Modular</td>
<td>🔄</td>
<td>🔄</td>
<td>🔄</td>
</tr>
<tr>
<td>Automated optimization</td>
<td>🔄</td>
<td>🔄</td>
<td>🔄</td>
</tr>
<tr>
<td>Distributed computing</td>
<td>🔄</td>
<td>🔄</td>
<td>🔄</td>
</tr>
</tbody>
</table>
PORTABLE WORKFLOWS will replace data transfers enabled by open standards that...

......are independent of the computation environment
...ensure 100% reproducibility and auditability
...are platform independent
Interaction Methods

Programmatic Access:
- API
  - Includes Data API
- SDK
  - Upload & Execute own code

Graphical Analysis Tools:
- Data-Mining Tools
- Genome Browser
- QC tools
Upload additional datasets, analyze in tandem with TCGA

How to upload files from your computer

We offer a standalone uploading client as a convenient way to upload your datasets from your laptop or desktop computer to Cancer Genomics Cloud.

The Cancer Genomics Cloud Uploader is a flexible, fast and secure client that installs on your local computer, can be started and stopped at your convenience and accommodates to a wide range of network topologies.

Installing the uploder on Mac OS X

For Windows or Linux?

Note: Cancer Genomics Cloud Uploader works on OS X 10.4 or newer. If you have an older version of OS X, please use the command-line uploader instead.

1. Download
Click the button below to download the installer. Double-Click the downloaded file to open the archive.

2. Install
Right- or control-click the installer icon and select "Open", then "Open" again in the dialog that pops up.

   Cancer Genomics Cloud Uploader
   Mac OS X

3. Run
Once the installer has finished, you can start the client from your Applications folder.
Metadata & finding the right datasets
Versioning & Reproducibility of pipelines

q Differential Expression - Cuffdiff (with Visualization)

Detect differences in transcript expression between groups of samples using Cuffdiff.
Monitor and optimize computational resource usage.
Visual interface for comparative analyses
Integrated Genome Browser with Annotations
Please help us develop the tools useful to you

• www.CancerGenomicsCloud.org
• ~$1M in funding to support computation and private data storage in evaluation period.
• Please share feedback!

• .. Or if you forget www.tcga.ninja