



# CANCER GENOMICS CLOUD

## SEVEN BRIDGES

TCGA Meeting Workshop

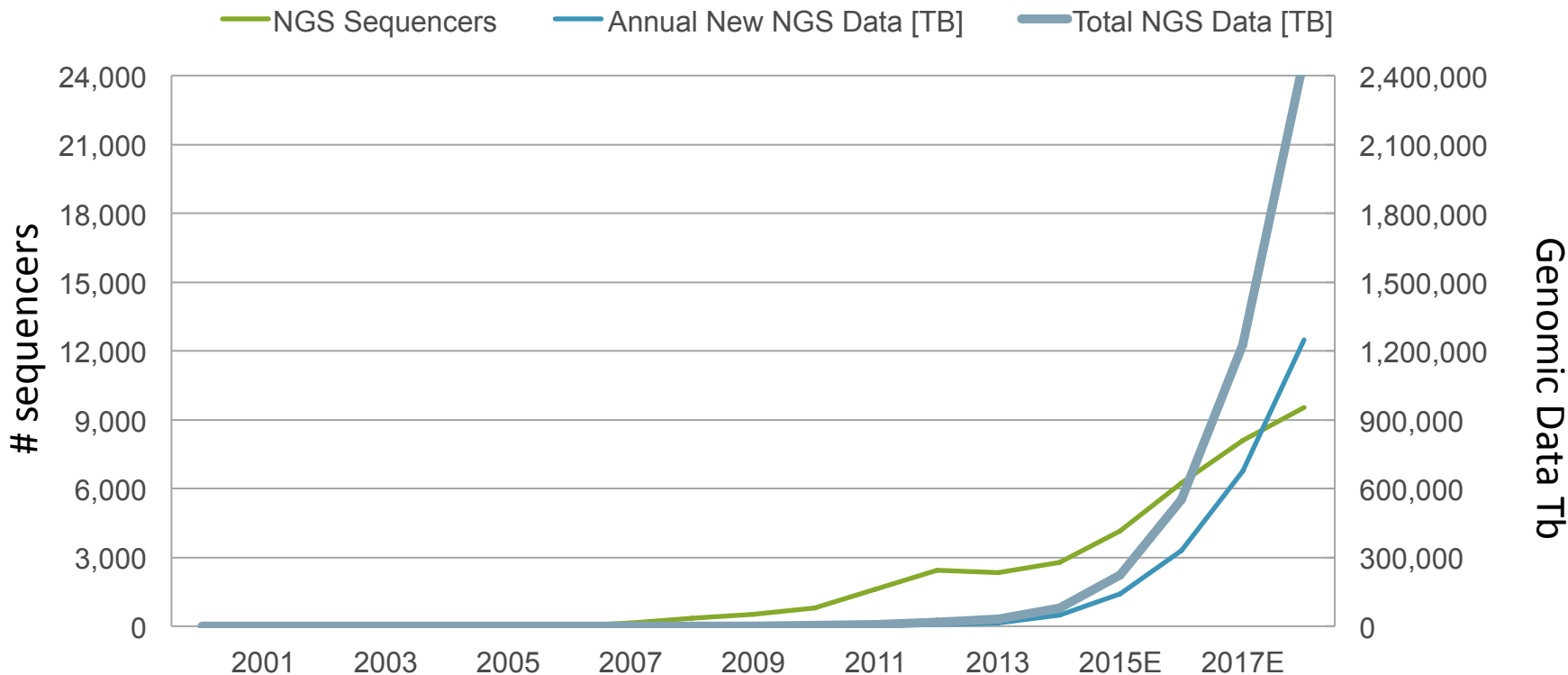
Deniz Kural, PhD

May 12, 2015

+ Brandi, Zeynep, Devin, Kate



# Amount of genomics data will exceed available resources



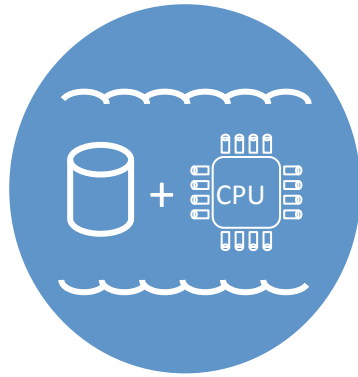
Between 2014-2018 production of new NGS data to exceed **2 Exabytes**

NGS: Next Generation Sequencing

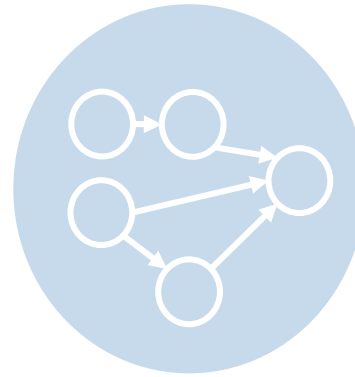
NGS sequencers include machines from Illumina, Life Technologies, and Pacific Biosciences. Human genome data based on estimates of whole human genomes sequenced

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

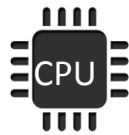
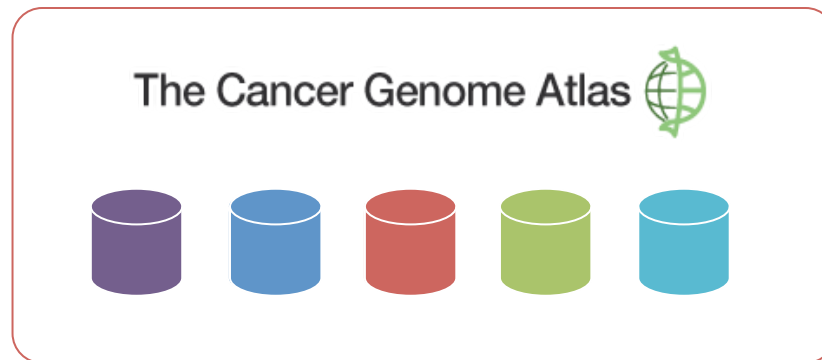


**CLOUD  
COMPUTATION**

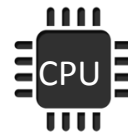


**PORTABLE  
WORKFLOWS**  
replace  
data transfers

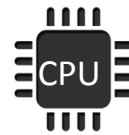
# Standard Model: Data comes to computation



Research Center 1



Research Center 2

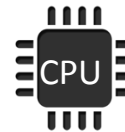
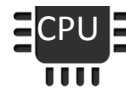
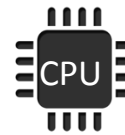
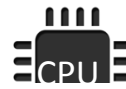


Research Center 3

# Cloud: Algorithms come to the data



The Cancer Genome Atlas 



ID	Name	Action	Status	Date
00001	John Doe	login	success	2013-01-01 10:00:00
00002	Jane Smith	login	failure	2013-01-01 10:05:00
00003	John Doe	logout	success	2013-01-01 10:15:00
00004	Jane Smith	login	success	2013-01-01 10:20:00
00005	John Doe	login	failure	2013-01-01 10:30:00
00006	Jane Smith	logout	success	2013-01-01 10:40:00
00007	John Doe	login	success	2013-01-01 10:50:00
00008	Jane Smith	login	failure	2013-01-01 11:00:00
00009	John Doe	logout	success	2013-01-01 11:10:00
00010	Jane Smith	login	success	2013-01-01 11:20:00

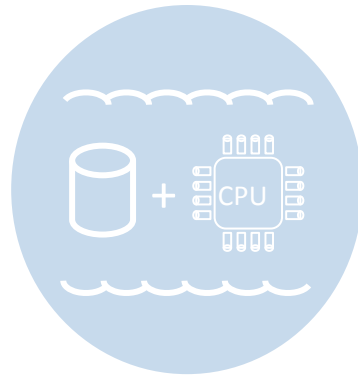
**Workflow** Research Center 1

**Workflow** Research Center 2

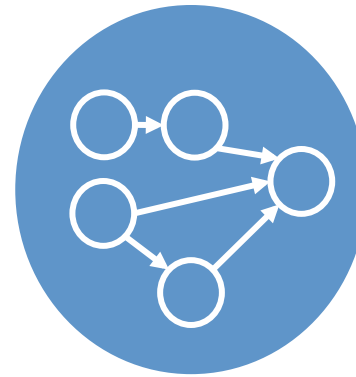
**Workflow** Research Center 3

Other awardees of the NCI Cancer Cloud pilots: Institute for Systems Biology, The Broad Institute

# Millions of genomes will shift how computation is done

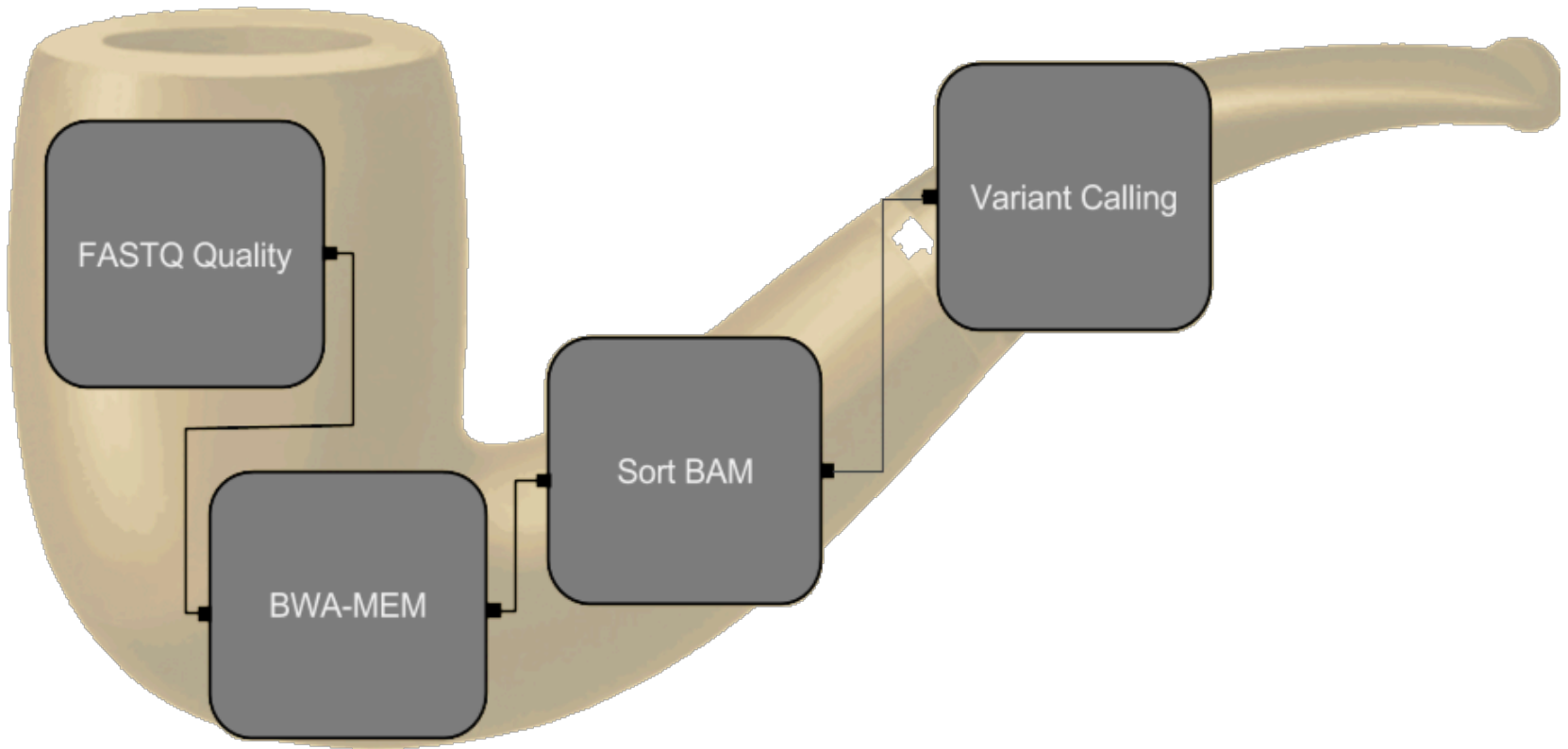


**CLOUD  
COMPUTE**
























**PORTABLE  
WORKFLOWS**  
replace  
data transfers

# Ceci n'est pas une pipeline!?



# Replicating computations is a hard problem

	Scripts	Machine Images	App containers + Pipeline-Language
Easy distribution			
Simple to run			
Compatible across environments			
100% reproducible			
Modular			
Automated optimization			
Distributed computing			



# 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

Public files

My files

## Import from...

My computer

Cluster or workstation

FTP or HTTP server

## Projects

My First Project

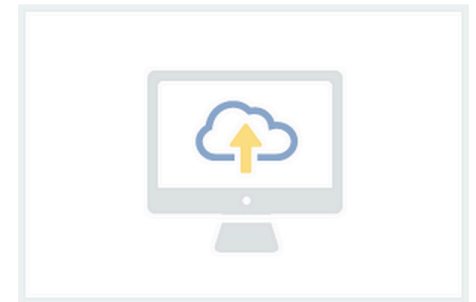
NCI demo

RNASeqDemo

## 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 uploader on Mac OS X

need it 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.

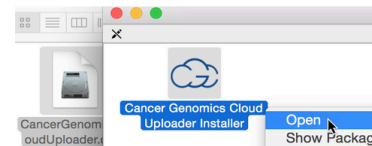
### 1. Download

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

Cancer Genomics Cloud Uploader  
Mac OS X

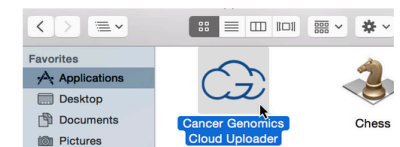
### 2. Install

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



### 3. Run

Once the installer has finished, you can start the client from your Applications folder.



# Metadata & finding the right datasets

The screenshot displays the CancerGenomicsCloud.org interface. At the top, there are navigation tabs for 'Projects', 'Data', and 'Tools & Pipelines'. Below this, a 'Demo' section contains tabs for 'Files', 'Pipelines', 'Tasks', and 'Settings'. A search bar and filters for 'File type' and 'Sample ID' are visible. A modal window titled 'Edit metadata for 1 selected files' is open, showing the following fields:

- File type: fastq
- Sample ID: Heart\_2
- Library ID: ERR315384
- Lane/slide: (empty)
- Paired End: 1
- Chunk number: (empty)
- Sequencing technology: Illumina
- Quality scale: illumina18

Buttons for 'Revert all' and 'Save' are at the bottom of the modal. In the background, a table of files is visible with columns for 'Size', 'File type', 'Task ID', and 'Sample ID'. A '+ Add Files' button is also present.

File Name	Size	File type	Task ID	Sample ID
ERR315384.Aligned.out.bam	2.3 GB	bam	553	Heart_2
:R315384.Aligned.out.bam.bai	0.2 KB	bam_index	553	Heart_2
:R315384.Log.final.out	1.7 KB	text	553	Heart_2
:R315384.SJ.out.tab	7.6 MB	text	553	Heart_2
:R315384.Unmapped.out.mate1	5.3 MB	fastq	553	Heart_2
:R315384.Unmapped.out.mate2	5.3 MB	fastq	553	Heart_2
:R315384_1.fastq.gz	7.7 MB	fastq	-	Heart_2
:R315384_2.fastq.gz	2.3 MB	fastq	-	Heart_2
:R315389_1.fastq.gz	457.0 MB	fastq	-	Heart_1
:R315389_2.fastq.gz	470.4 MB	fastq	-	Heart_1

# Versioning & Reproducibility of pipelines


Files Pipelines Tasks Settings

ne List

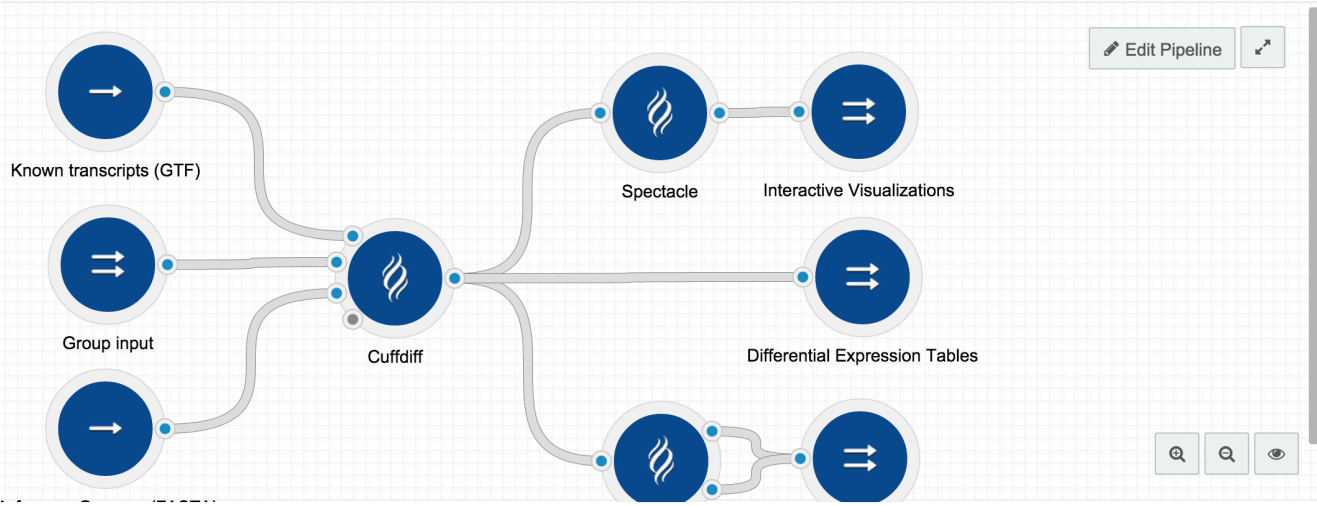
Revision 2 ▾ | Delete Clone

## q Differential Expression - Cuffdiff (with Visualization)

Jemo. Last updated by [cgcdemo](#) on May 10, 2015.

cant differences in transcript expression between groups of samples using Cuffdiff. 

Change Log



Known transcripts (GTF)


Group input

Cuffdiff

Spectacle

Interactive Visualizations


Differential Expression Tables

Edit Pipeline 

Search Search Eye

Cuffdiff (2.0.2) ▾

- Time series
- Upper\_quartile normalisation
- Hits normalisation Compatible hits
- Multi read correct
- Min alignment count
- False discovery rate
- Library type
- Fragment length mean
- Fragment length standard dev...
- Maximum MLE iterations
- Poisson dispersion
- Maximum bundle frags
- Number of fragment count dra...
- Number of fragment assign dr...
- Maximum fragment multihits
- Minimum outlier p
- Minimum replicates for js test
- No effective length correction

ation 

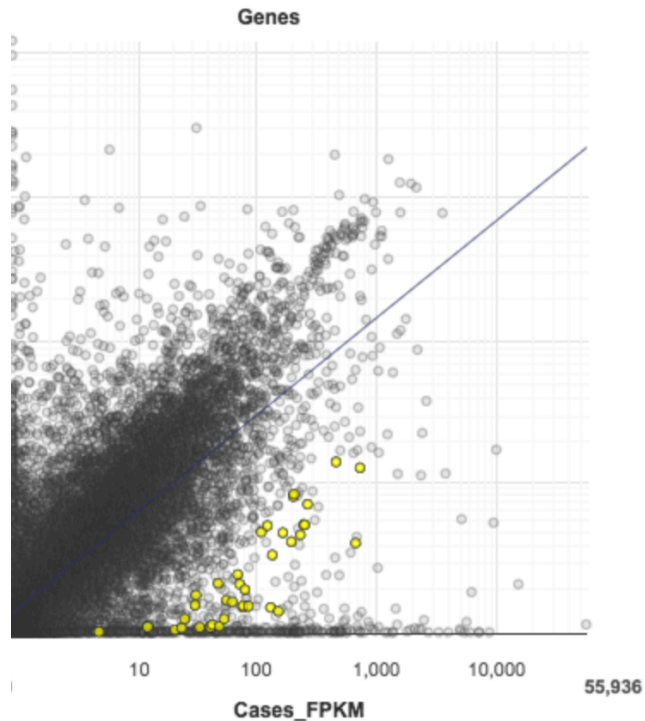
# Monitor and optimize computational resource usage.



# Visual interface for comparative analyses

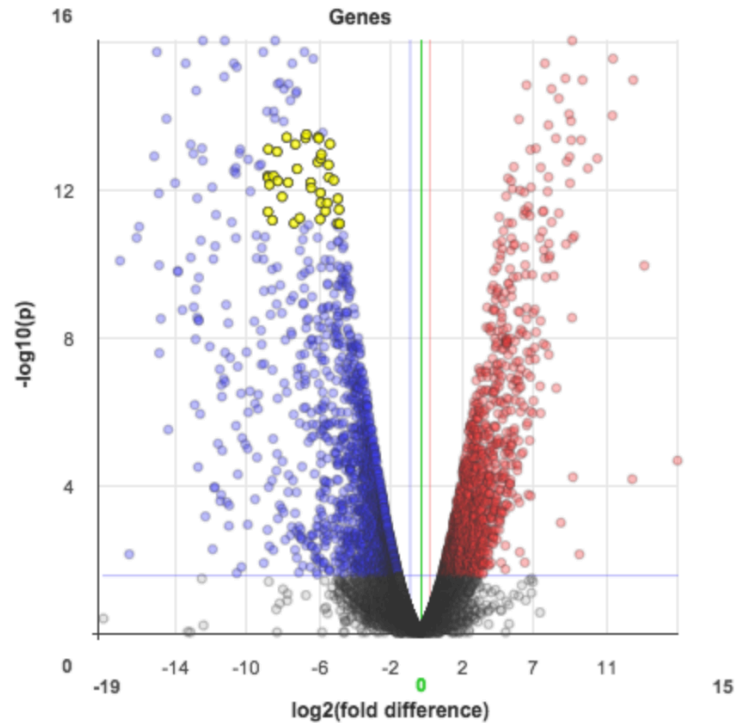
Select Groups

Cases vs Controls



Select Groups

Cases vs Controls



Find Gene

Enter gene name

Gene names of interest

Enter gene names, separat  
new line

Selected Genes

[cgn](#)

[onecut2](#)

[amdhd1](#)

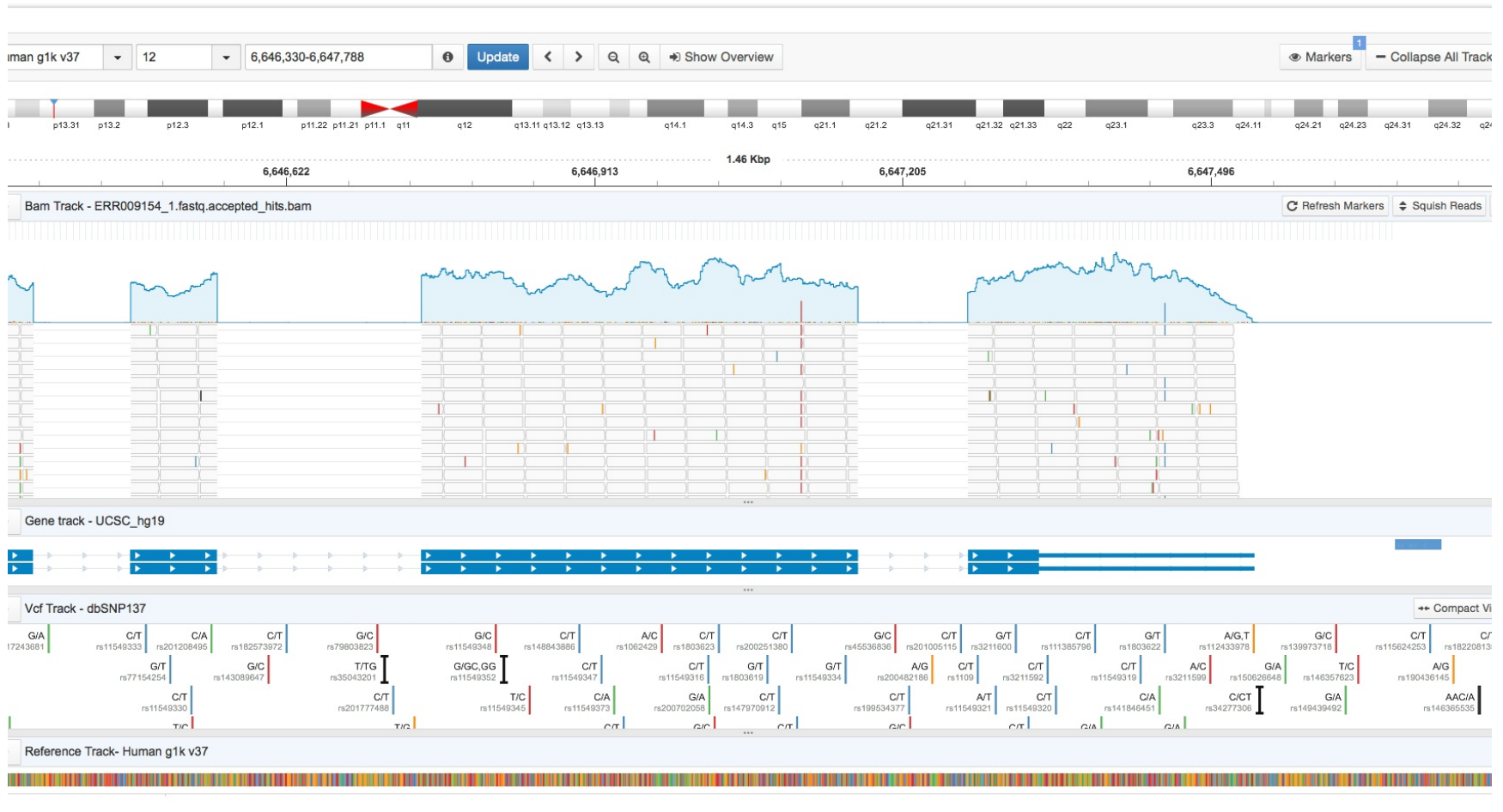
[rnf128](#)

[hal](#)

[c2orf72](#)

Export Selected (38)

# Integrated Genome Browser with Annotations





# Please help us develop the tools useful to you

- [www.CancerGenomicsCloud.org](http://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](http://www.tcga.ninja)