

# Mining TCGA Gene Expression Data

Kimberly J. Bussey, Ph.D.

Assistant Professor

Integrated Cancer Genomics Division

Translational Genomics Research Institute



Translational Genomics Research Institute

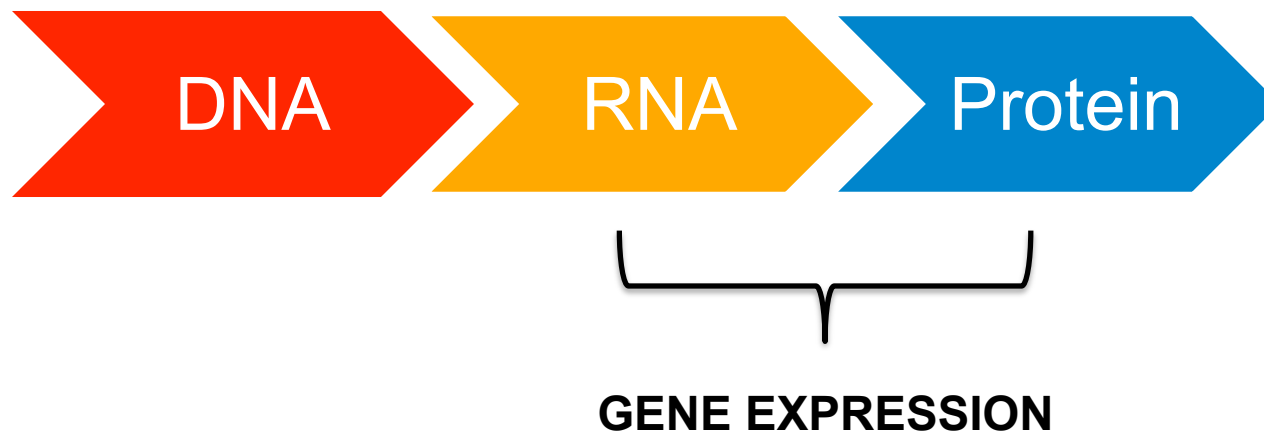


# Outline

- Biology
- Types of gene expression measurements
- Gene expression data in TCGA
- Tools for working with the data



# What do we mean by gene expression?



# Gene Expression Platforms - RNA

- Array-based
  - Affymetrix, Illumina, Agilent, etc.
  - Can be total mRNA, focused on exons or splice variants, or miRNA
  - Probes designed to specific sequences
- RNA-seq
  - Sequenced based
  - Either mRNA or miRNA

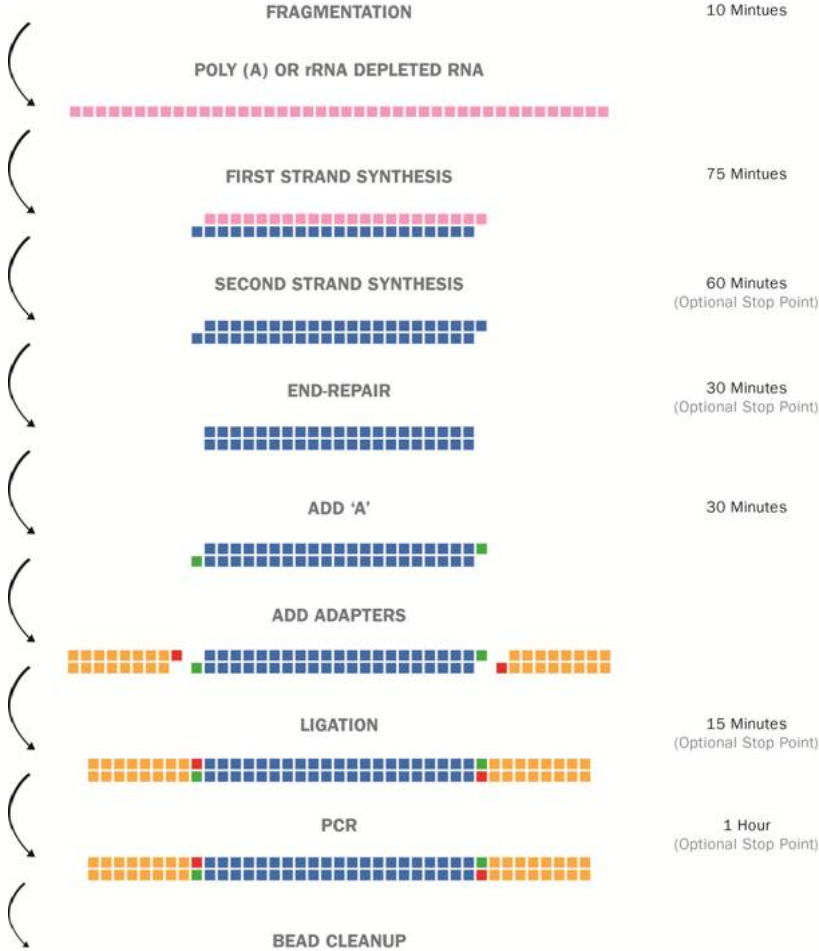


# RNA workflow

- RNA Extraction
  - Different protocols depending on what the target pool of RNA is: total RNA, mRNA, or miRNA
- Create cDNA or cRNA library
- For array-based method, hybridization and image analysis



# RNA-Seq



# Illumina Sequencing Technology Overview

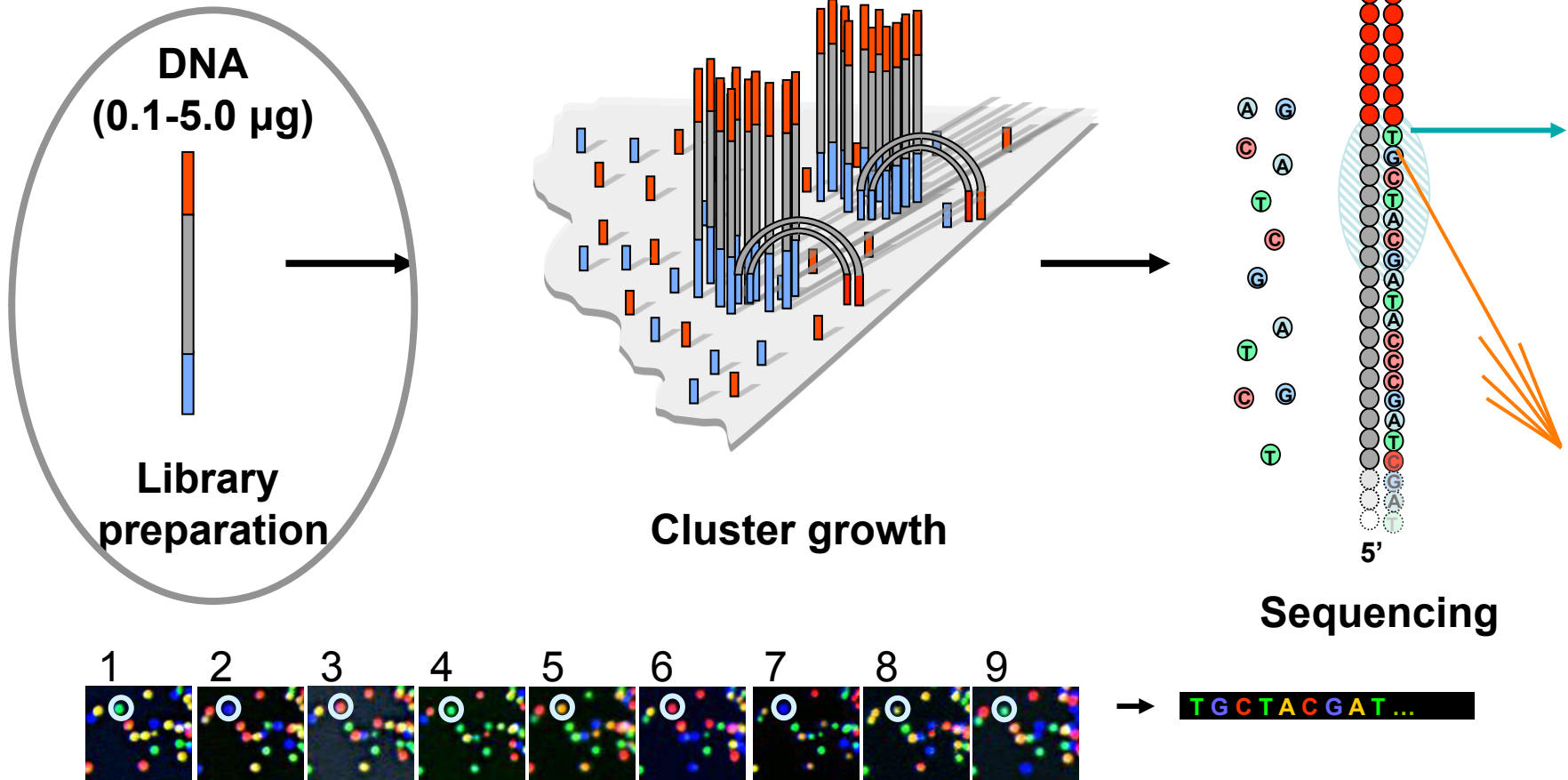


Image acquisition

Base calling

Translational Genomics Research Institute



# Gene Expression Platforms - Protein

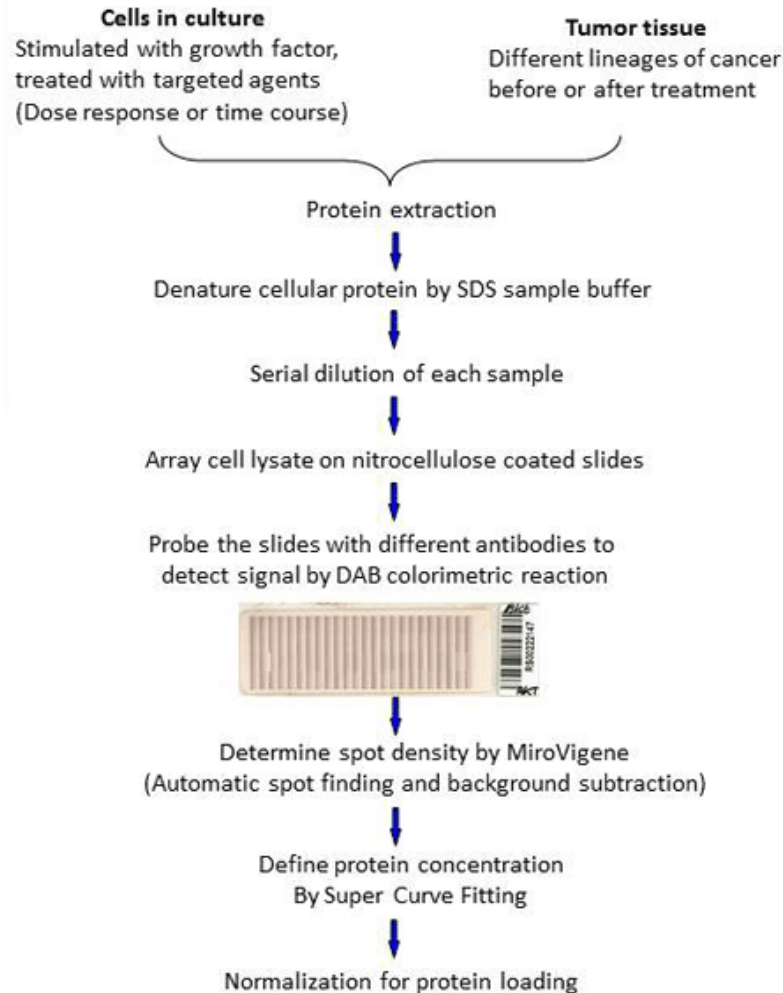
- Reverse Phase Protein Arrays
  - Serial dilution of protein lysate spotted array
  - Probed with antibody



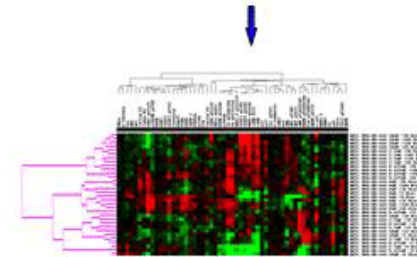


# Protein workflow

## The RPPA Process



**Data:**  
Supervised and Un-Supervised  
hierarchical clustering analysis  
(Set of 2 Heatmaps)



**AND**  
Excel file of protein expression  
and modification in a relative  
concentration

**Further analysis  
by Bioinformatics**

- 1) Identify on and off effects
- 2) Identify regulatory loops
- 3) Select drug candidates
- 4) Improve patient management



# TCGA data sets

- <https://wiki.nci.nih.gov/display/TCGA/TCGA+Data+Primer>
- Remember: what is public depends on the risk of being able to identify the subject
  - No BAM files, no FASTq without controlled access approval



# The Cancer Genome Atlas Data Portal

Understanding genomics to improve cancer care

[TCGA Home](#) | [Contact Us](#) | [For the Media](#)

- Home
- Query the Data
- Download Data
- Tools**
- About the Data
- Publication Guidelines

Home > Tools > Analytical Tools

## Analytical Tools

### The Cancer Imaging Archive (TCIA)

The Cancer Imaging Archive (TCIA) is a service provided by NCI that provides access to radiological imaging data sets in DICOM format from TCGA cases. TCIA supports imaging phenotype - genotype research, in addition to other imaging data sets for cancer imaging analysis.

### Cancer Genome Workbench (CGWB)

The Cancer Genome Workbench (CGWB) is an application developed by NCI that provides whole-genome and heatmap views of sample-level data.

### Integrative Genomics Viewer (IGV)

The Integrative Genomics Viewer (IGV) is a high-performance visualization tool created by the Broad Institute for interactive exploration of large, integrated datasets.

### cBio Cancer Genomics Portal

The cBioCancer Genomics Portal provides visualization, analysis and download of large-scale cancer genomics data sets. The portal is developed and maintained by the Computational Biology (cBio)

#### In This Section

[Tools](#)

[Analytical Tools](#)

[Annotations Manager](#)

[Biospecimen Metadata Browser](#)

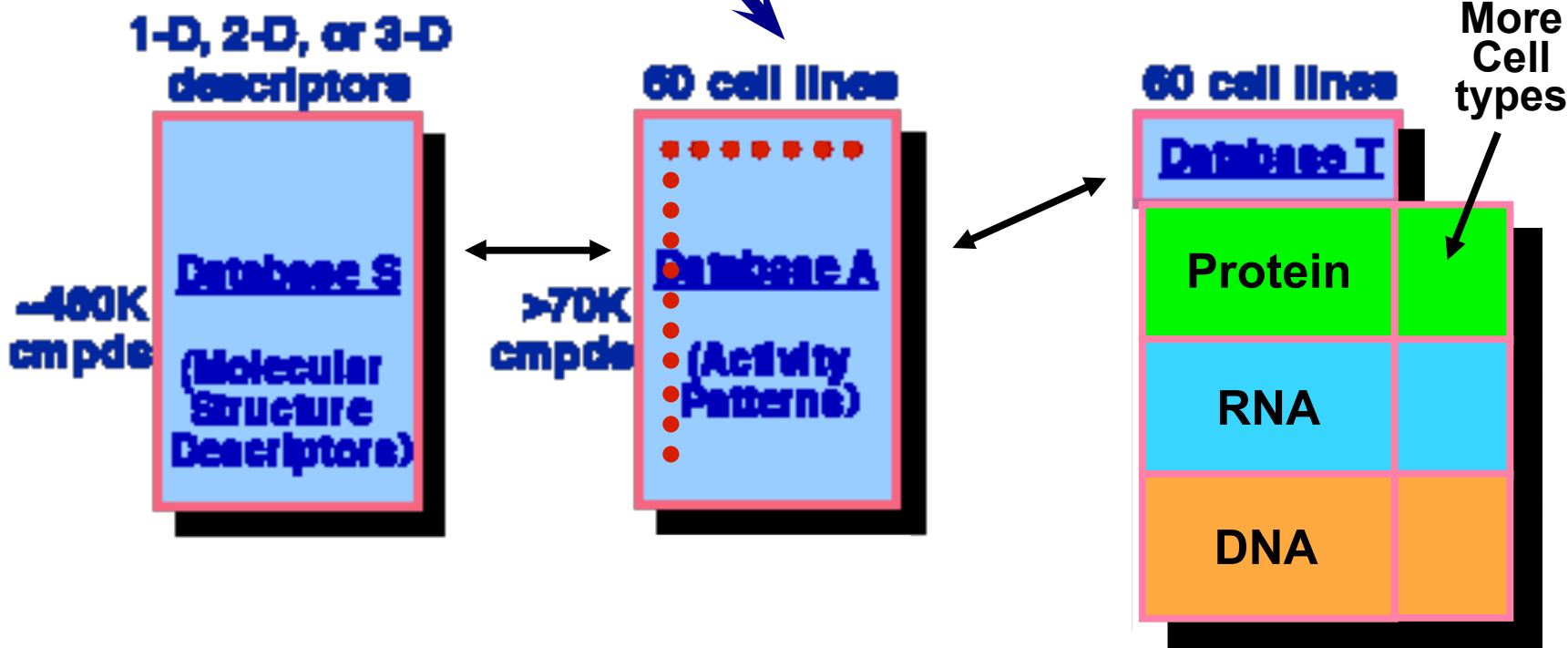
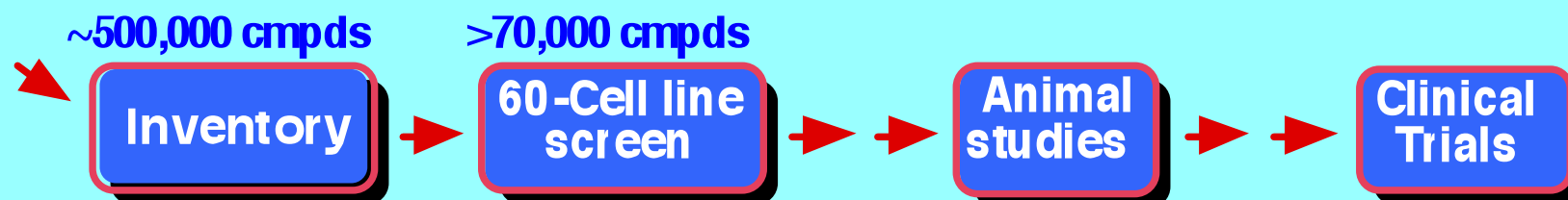
#### More TCGA Information

More information about The Cancer Genome Atlas program can be found by following the links below:

[TCGA website](#)

[TCGA Publications](#)

# Flow in one portion of NCI Drug Discovery Program



# Integration is the goal, but...

- You need to understand what the technology measured
- You need to know how that measurement was annotated
- Remember that not all identifiers are stable over time
- Excel does bad things to gene symbols and clone ids



# About Copy Number....



↑  
Normal



Cancer →



# Tools and Questions

- IGV: Visualize the output of FireHose
- UCSC Cancer Genome Viewer : Visualize and do subset analysis
- Regulome Explorer: Integrative analysis driven by statistical associations
- GenePattern: Suite of tools for many different types of data sets
- IPA, GeneGO, etc: Pathway enrichment analysis



# Biosig

Tcga - Mozilla Firefox

File Edit View History Bookmarks Tools Help

Analytical Tools - Data P... x Tcga x cBio Cancer Genomics Po... x Cancer Genomics Browse... x Home | Integrative Geno... x Cancer Genome Workben... x +

tcga.lbl.gov:8080/biosig/tcgadownload.do

tcga tools

Most Visited Getting Started Latest Headlines PubMed Home Entrez Gene BLAST: Basic Local Alig... UCSC Genome Browser ... GeneCards Homepage



TCGA Login

Search

BLCA  
BRCA  
CESC  
COAD  
DLBC  
ESCA  
GBM  
HNSC  
KICH  
KIRC  
KIRP  
LGG  
LIHC  
LUAD  
LUSC  
OV  
PAAD  
PRAD  
READ  
SARC  
SKCM  
STAD  
THCA  
UCEC

TumorType
<a href="#">BLCA</a>
<a href="#">BRCA</a>
<a href="#">CESC</a>
<a href="#">COAD</a>
<a href="#">DLBC</a>
<a href="#">ESCA</a>
<a href="#">GBM</a>
<a href="#">HNSC</a>
<a href="#">KICH</a>
<a href="#">KIRC</a>
<a href="#">KIRP</a>
<a href="#">LGG</a>
<a href="#">LIHC</a>
<a href="#">LUAD</a>
<a href="#">LUSC</a>
<a href="#">OV</a>
<a href="#">PAAD</a>
<a href="#">PRAD</a>
<a href="#">READ</a>
<a href="#">SARC</a>
<a href="#">SKCM</a>
<a href="#">STAD</a>
<a href="#">THCA</a>
<a href="#">UCEC</a>

About Privacy Copyright © 2005-2011 Lawrence Berkeley National Laboratory. All Rights Reserved.

Windows taskbar with icons for Internet Explorer, File Explorer, and PowerPoint. System tray shows network, volume, and clock (4:57 AM, 2/15/2013).



# Biosig

TCGA

Tissues for OV

LBNL Patient Id	Tissue	Tissue Type	Patient	Visualized	Processed	Thumbnail
LP1881	<a href="#">TCGA-13-0920-01A-01-BS1</a>	Frozen	TCGA-13-0920	true	false	
LP1881	<a href="#">TCGA-13-0920-01A-01-TS1</a>	Frozen	TCGA-13-0920	true	false	
LP1882	<a href="#">TCGA-13-0919-01A-01-BS1</a>	Frozen	TCGA-13-0919	true	false	
LP1882	<a href="#">TCGA-13-0919-01A-01-TS1</a>	Frozen	TCGA-13-0919	true	false	
LP1883	<a href="#">TCGA-24-1556-01A-01-BS1</a>	Frozen	TCGA-24-1556	true	false	
LP1883	<a href="#">TCGA-24-1556-01A-01-TS1</a>	Frozen	TCGA-24-1556	true	false	

Transferring data from tcga.lbl.gov...

4:58 AM  
2/15/2013



Translational Genomics Research Institute



# cBio

The screenshot shows the cBio Cancer Genomics Portal in a Mozilla Firefox browser. The address bar displays [www.cbioportal.org/public-portal/](http://www.cbioportal.org/public-portal/). The page header includes the Memorial Sloan-Kettering Cancer Center logo and the text "cBio Cancer Genomics Portal" with the tagline "Visualize, analyze, discover." A navigation menu contains links for HOME, TUTORIALS, NEWS, FAQ, DATA SETS, ABOUT, WEB API, R/MATLAB, and NETWORKS. The main content area features a description of the portal's purpose: "The cBio Cancer Genomics Portal provides visualization, analysis and download of large-scale cancer genomics data sets." It also mentions that the portal is developed and maintained by the Computational Biology Center at Memorial Sloan-Kettering Cancer Center, with a reference to *Cancer Discovery*, May 2012 2; 401. [Abstract].

A "Mutations of interest (4 of 35)" table is displayed:

Gene	Protein Change	Type
TP53	Q331*	NS
PPP2R1A	S256F	MS
FAT1	E314*	NS
EPHA7	H408Q	MS

Below the table is a "Query" section with a "Download Data" button. It includes a "Select Cancer Study" dropdown menu set to "All Cancer Studies", a "Select Data Type Priority" section with radio buttons for "Mutation and CNA" (selected), "Only Mutation", and "Only CNA", and an "Enter Gene Set" section with a text input field and a "Download" button. The text input field contains "Enter HUGO Gene Symbols or Gene Aliases".

On the right side, a "Data Sets" section states: "The Portal contains data for 7848 tumor samples from 26 cancer studies. [Details.]" Below this is a pie chart showing the distribution of samples across 26 cancer studies, with the largest slice representing 849 samples. Below the pie chart is an "Example Queries" section with the text "RAS/RAF alterations in colorectal cancer".

The browser's taskbar at the bottom shows various application icons and the system clock displaying 5:01 AM on 2/15/2013.

# Signatures/Gene Sets - MSigdb

GSEA | MSigDB - Mozilla Firefox

File Edit View History Bookmarks Tools Help

Analytical Tools - D... Patient cBio Cancer Genomi... Cancer Genomics Br... Home | Integrative ... Cancer Genome Wor... GSEA | MSigDB

www.broadinstitute.org/gsea/msigdb/index.jsp

Most Visited Getting Started Latest Headlines PubMed Home Entrez Gene BLAST: Basic Local Alig... UCSC Genome Browser ... GeneCards Homepage

login register BROAD INSTITUTE

GSEA Gene Set Enrichment Analysis

GSEA Home Downloads Molecular Signatures Database Documentation Contact

MSigDB Home About Collections Browse Gene Sets Search Gene Sets Investigate Gene Sets View Gene Families Help

**MSigDB**  
Molecular Signatures Database

## Molecular Signatures Database v3.1

**Overview**

The Molecular Signatures Database (MSigDB) is a collection of annotated gene sets for use with GSEA software. From this web site, you can

- ▶ **Search** for gene sets by keyword.
- ▶ **Browse** gene sets by name or collection.
- ▶ **Examine** a gene set and its annotations. See, for example, the **ANGIOGENESIS** gene set page.
- ▶ **Download** gene sets.
- ▶ **Investigate** gene sets:
  - ▶ **Compute overlaps** between your gene set and gene sets in MSigDB.
  - ▶ **Categorize** members of a gene set by gene families.
  - ▶ **View the expression profile** of a gene set in any of the three provided public expression compendia.

**Collections**

The MSigDB gene sets are divided into 6 major collections:

- c1 positional gene sets** for each human chromosome and cytogenetic band.
- c2 curated gene sets** from online pathway databases, publications in PubMed, and knowledge of domain experts.
- c3 motif gene sets** based on conserved cis-regulatory motifs from a comparative analysis of the human, mouse, rat, and dog genomes.
- c4 computational gene sets** defined by mining large collections of cancer-oriented microarray

Windows taskbar: 5:30 AM 2/15/2013

# MSigdb

The MSigDB gene sets are divided into 6 major collections:

**c1** **positional gene sets** for each human chromosome and cytogenetic band.

**c2** **curated gene sets** from online pathway databases, publications in PubMed, and knowledge of domain experts.

**c3** **motif gene sets** based on conserved cis-regulatory motifs from a comparative analysis of the human, mouse, rat, and dog genomes.

**c4** **computational gene sets** defined by mining large collections of cancer-oriented microarray data.

**c5** **GO gene sets** consist of genes annotated by the same GO terms.

**c6** **oncogenic signatures** defined directly from microarray gene expression data from cancer gene perturbations.



# UCSC Cancer Genomics Browser

The screenshot displays the UCSC Cancer Genomics Browser interface. The browser window title is "Cancer Genomics Browser at UCSC - Mozilla Firefox". The address bar shows the URL "https://genome-cancer.soe.ucsc.edu". The page header includes the UC Santa Cruz logo and the text "Cancer Genomics Browser". The main content area features a heatmap titled "TCGA lung adenocarcinoma (LUAD) copy number gistic2 estimate • N=205". The heatmap shows copy number data across chromosomes 1 to 22, X, and Y. Below the heatmap, there are columns for clinical data: Subgroup, smoking history, M (TNM), sample type, tumor stage, and Sample name. The y-axis of the heatmap ranges from -3,882 to 4,387. The browser's taskbar at the bottom shows various application icons and the system clock indicating 5:36 AM on 2/15/2013.

Cancer Genomics Browser  
UCSC Genome Browser  
News  
Tutorial  
User Guide  
FAQ  
About Us  
Publications  
Jobs  
Contact Us  
Conditions of Use

## UCSC Cancer Genomics Browser

*A tool to visualize and host functional genomic data*

Welcome to the UCSC Cancer Genomics Browser. The browser is a suite of web-based tools to visualize, integrate and analyze cancer genomics and its associated clinical data. It is developed and maintained by the UCSC Cancer Genomics Group, led by David Haussler and Josh Stuart, working closely with the UCSC Human Genome Browser team at the Center for Biomolecular Science and Engineering ([CBSE](#)) at the University of California Santa Cruz ([UCSC](#)). If you have

# Cancer Genome Workbench

Cancer Genome Workbench - Cancer Insights - Mozilla Firefox

File Edit View History Bookmarks Tools Help

Analytical Tools - D... Patient cBio Cancer Genomi... Cancer Genomics Br... Home | Integrative ... Cancer Genome Wor... GSEA | MSigDB

https://cgwb.nci.nih.gov

Most Visited Getting Started Latest Headlines PubMed Home Entrez Gene BLAST: Basic Local Alig... UCSC Genome Browser ... GeneCards Homepage

CGWB Home | Tools | Datadump | Help | Bambino | Log In | Heatmap

CGWB is developed by The National Cancer Institute's Center for Biomedical Informatics and Information Technology (CBIT)

CGWB (The Cancer Genome Workbench) hosts mutation, copy number, expression, and methylation data from a number of projects, including TCGA, TARGET, COSMIC, GSK, NCI60. It has tools for visualizing sample-level genomic and transcription alterations in various cancers. The three main viewers in CGWB are:

- ★ Integrated track: a sample-level view of genomic alterations from multiple data sources in a customized version of the UCSC genome browser.
- ★ Heatmap view: an interactive, high-level graphical view of gene expression and copy number data alongside the associated clinical features.
- ★ Bambino: an alignment viewer for next-generation sequencing data in SAM/BAM format. Also includes a command line SNP and indel caller.

[More about CGWB](#)

hg18  
hg19 Enter Genomic Coords or Gene Name: chr17:7512444-7531641 Go Tissue Browse

[Click here to reset](#) the browser user interface settings to their defaults. [Click here to add](#) a custom track

TCGA TARGET TSP COSMIC Johns\_Hopkins LPG GSK\_Cell\_Lines Miscellaneous

Acute Myeloid Leukemia-LAML Brain cancer (glioblastoma multiforme)-GBM Ovarian Cystadenocarcinoma-OV  
Breast Carcinoma-BRCA Colon Adenocarcinoma-COAD Renal Clear Cell Carcinoma-KIRC Lung Adenocarcinoma-LUAD  
Lung Squamous Cell-LUSC Uterine Corpus Endometrioid Carcinoma-UCEC Rectum Adednocarcinoma-READ  
Bladder Urothelial Carcinoma-BLCA Cervical Squamous Cell Carcinoma-CESC Head and Neck squamous cell carcinoma-HNSC  
Kidney renal paillary cell carcinoma-KIRP Lower Grade Glioma-LGG Liver hepatocellular carcinoma-LIHC  
Prostate adenocarcinoma-PRAD Stomach adenocarcinoma-STAD Thyroid carcinoma-THCA

Current Project: **Serous cystadenocarcinoma (OV) in The Cancer Genome Atlas (TCGA)**

List of Genes ordered by Frequency of Samples with Somatic Mutations (with percentage of mutated samples in parentheses):  
TP53 (86%) CYP3A4 (13%) DST (8%) LTF (8%) NOS3 (8%) PRPF4B (8%) PRKDC (8%) ABCA3 (4%) AXL (4%) CDK4 (4%) DNABJ1 (4%) ERBB2 (4%) EGFR (4%) GRB10 (4%) ING1 (4%) MAP3K14 (4%) MET (4%) MELK (4%) NLK (4%) NF1 (4%) PML (4%) PRKCI (4%) RB1 (4%) RAF1 (4%) SNX17 (4%) TRIP11 (4%)

5:40 AM  
2/15/2013

Translational Genomics Research Institute



# IGV

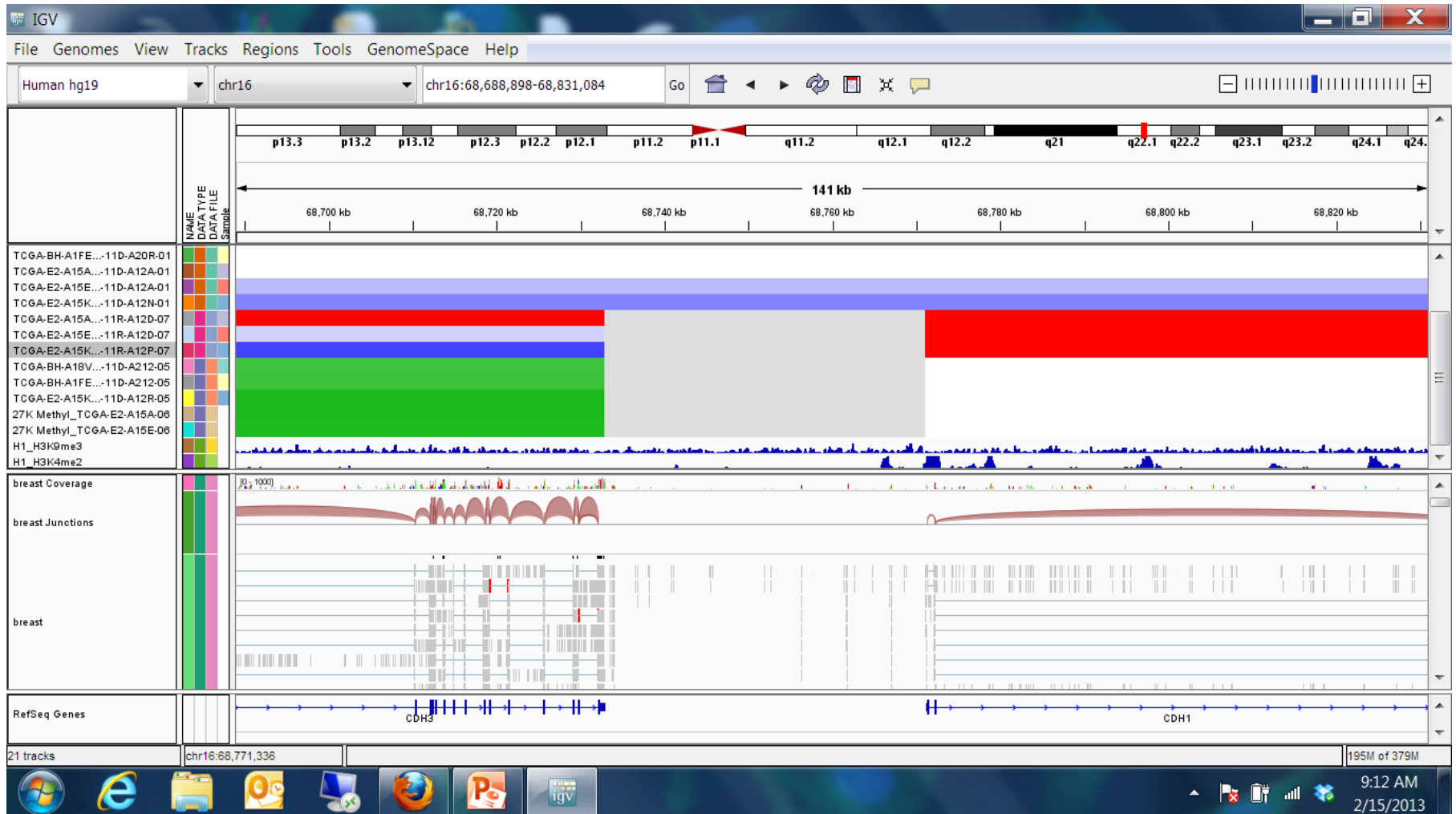
The screenshot shows the IGV website in a Mozilla Firefox browser. The browser's address bar displays [www.broadinstitute.org/software/igv/](http://www.broadinstitute.org/software/igv/). The page features a navigation menu on the left with links for Home, Downloads, Documents, Hosted Genomes, FAQ, IGV User Guide, File Formats, Release Notes, Credits, and Contact. A search bar is also present. The main content area includes a large banner for the Integrative Genomics Viewer, a 'What's New' section with news items from December 18, 2012, and April 20, 2012, and a 'Citing IGV' section with a citation for the 2012 paper by Helga Thorvaldsdóttir et al. The Windows taskbar at the bottom shows the system time as 5:46 AM on 2/15/2013.



Translational Genomics Research Institute



# IGV



Translational Genomics Research Institute





# Take Home Points

- Gene expression data can be RNA or protein-based measurements
- Different tools are good for showing you different relationships
- Interpretation of the results requires an understanding of what was actually measured



# QUESTIONS?



Translational Genomics Research Institute

