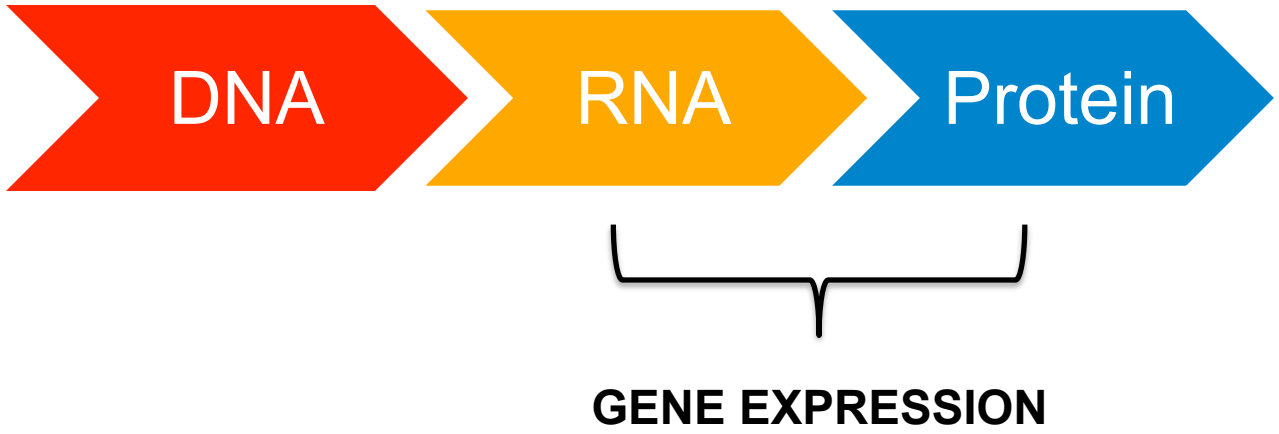
Mining TCGA Gene Expression Data

Kimberly J. Bussey, Ph.D.
Assistant Professor
Integrated Cancer Genomics Division
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?



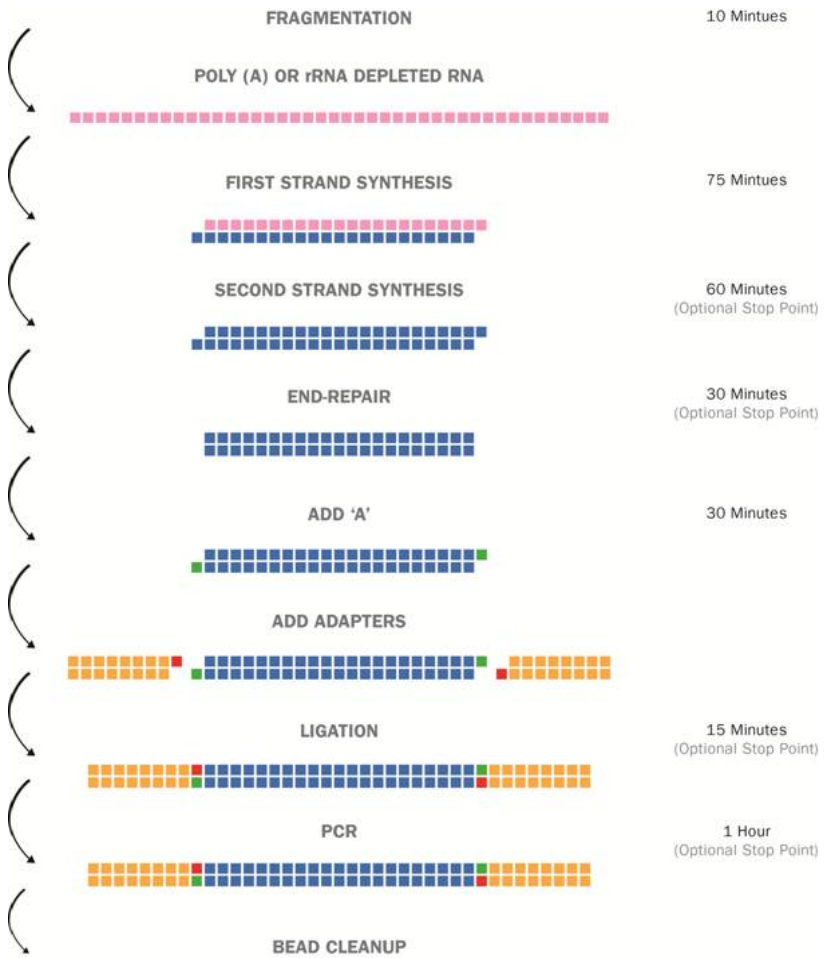
Gene Expression Platforms - RNA

- 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

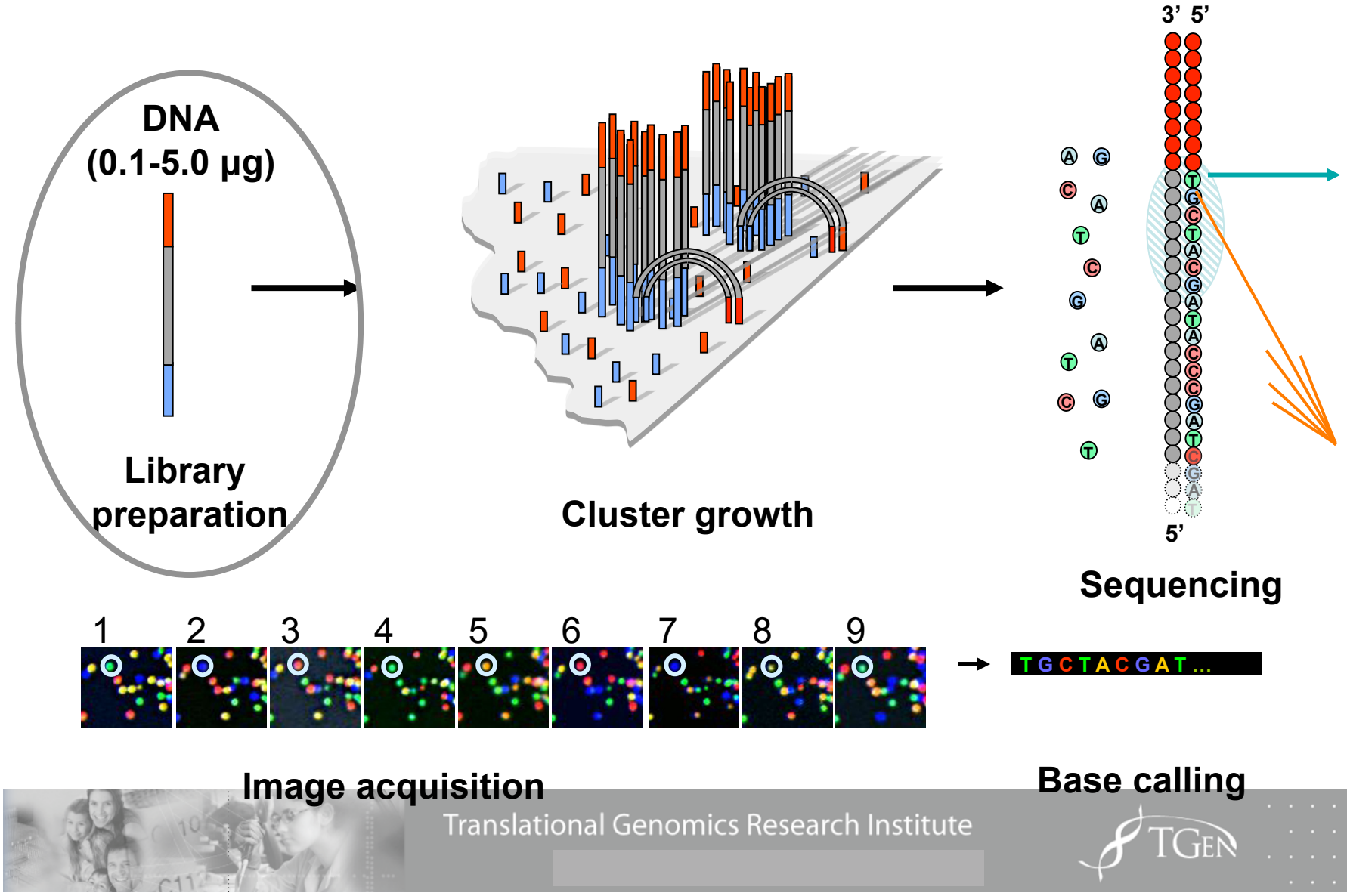
RNA workflow

- 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


FRAGMENTATION
10 Mintues
POLY (A) OR rRNA DEPLETED RNA
FIRST STRAND SYNTHESIS
75 Mintues
SECOND STRAND SYNTHESIS
60 Minutes
(Optional Stop Point)
END-REPAIR
30 Minutes
(Optional Stop Point)
ADD 'A'
30 Minutes
ADD ADAPTERS
LIGATION
15 Minutes
(Optional Stop Point)
PCR
1 Hour
(Optional Stop Point)
BEAD CLEANUP

Illumina Sequencing Technology Overview



Gene Expression Platforms - Protein

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

Protein workflow

The RPPA Process

Cells in culture
Stimulated with growth factor, treated with targeted agents (Dose response or time course)

Tumor tissue
Different lineages of cancer before or after treatment

Protein extraction

Denature cellular protein by SDS sample buffer

Serial dilution of each sample

Array cell lysate on nitrocellulose coated slides

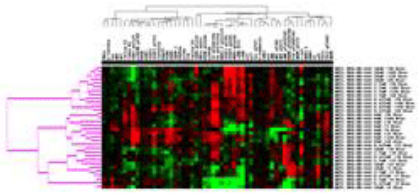
Probe the slides with different antibodies to detect signal by DAB colorimetric reaction

Determine spot density by MiroVigene (Automatic spot finding and background subtraction)

Define protein concentration By Super Curve Fitting

Normalization for protein loading

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



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

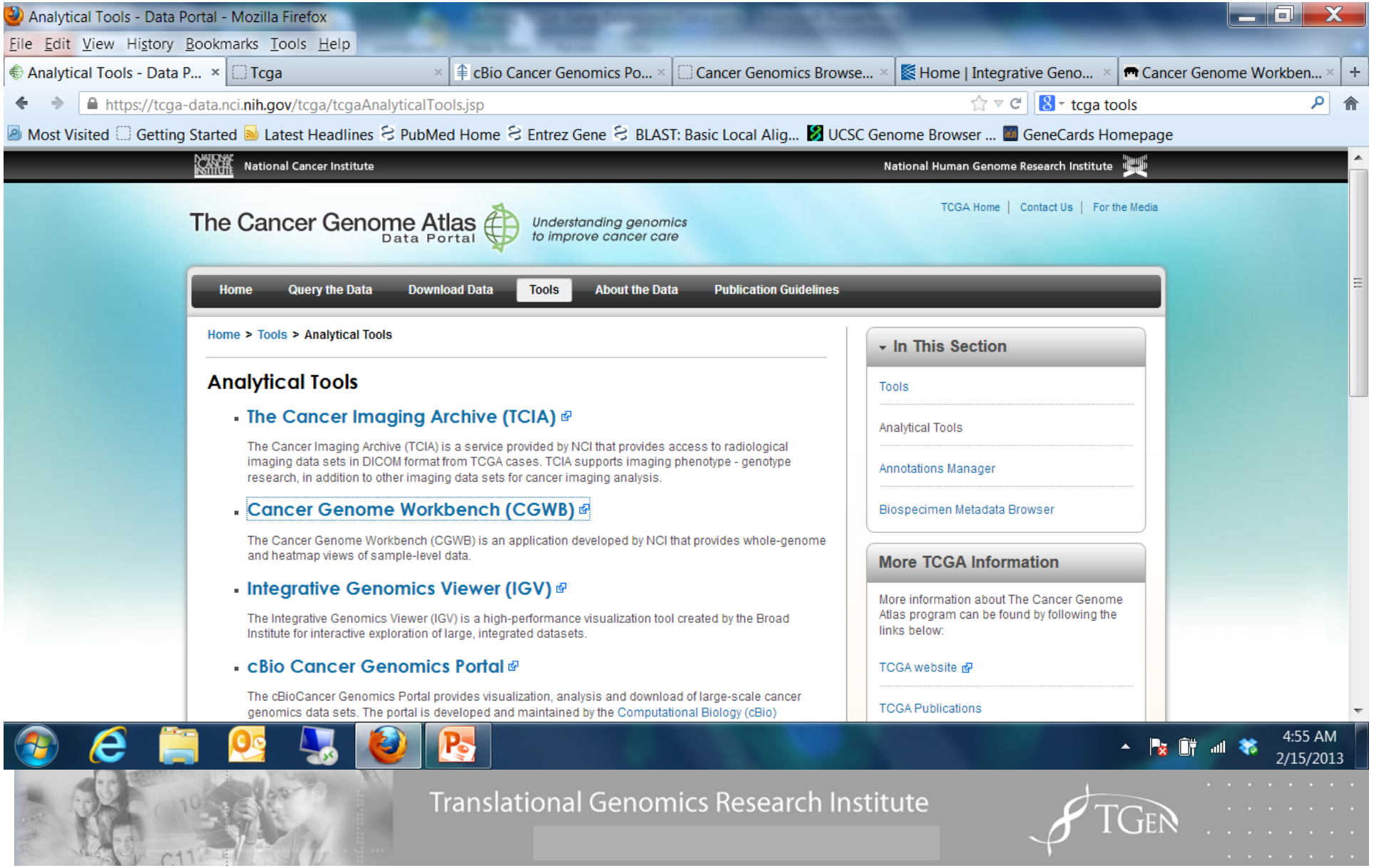
Further analysis by Bioinformatics

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

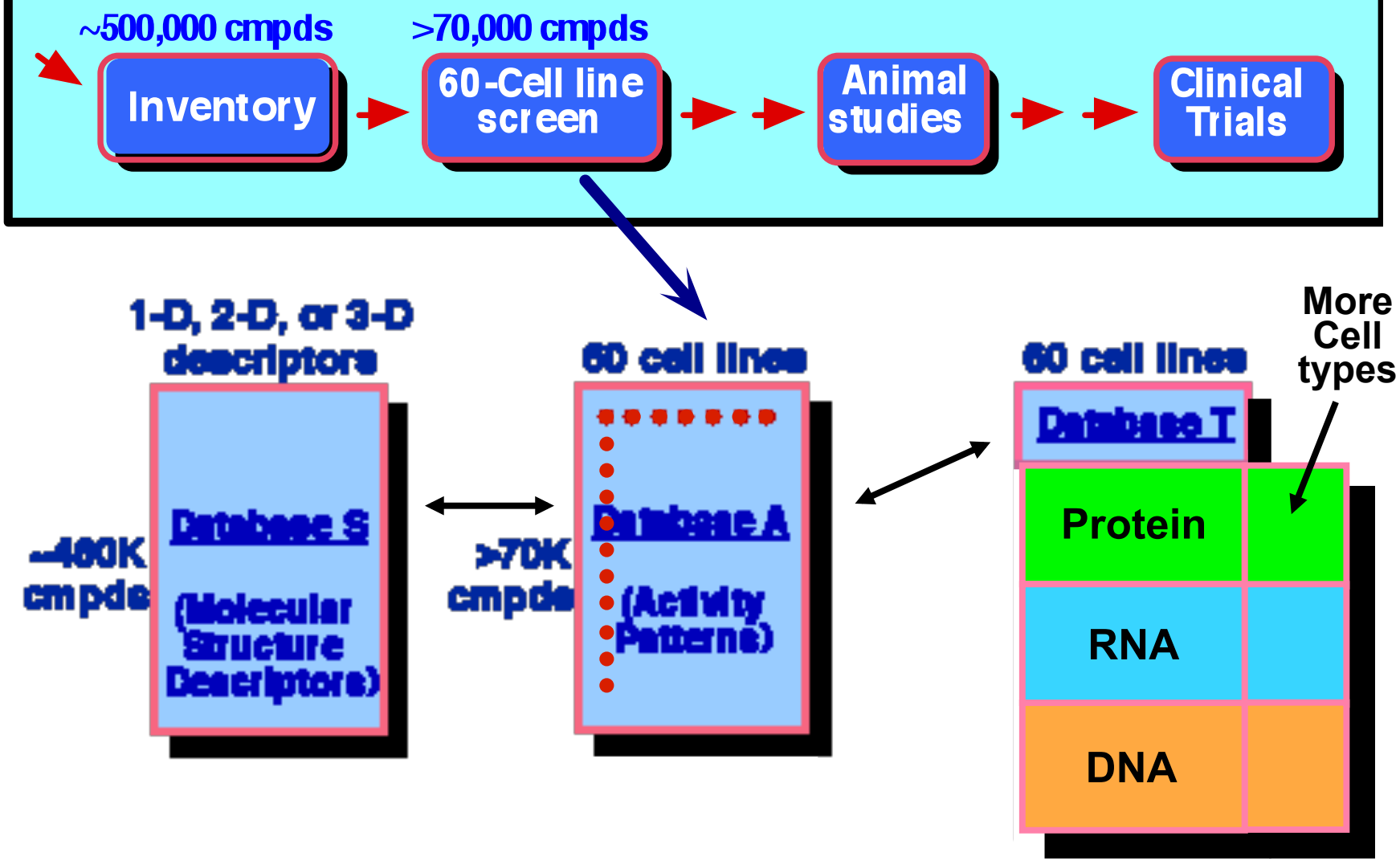
TCGA data sets

- 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


Analytical Tools - Data Portal - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Analytical Tools - Data P...
Tcga
cBio Cancer Genomics Po...
Cancer Genomics Browse...
Home | Integrative Geno...
Cancer Genome Workben...
https://tcga-data.nci.nih.gov/tcga/tcgaAnalyticalTools.jsp
tcga tools
Most Visited
Getting Started
Latest Headlines
PubMed Home
Entrez Gene
BLAST: Basic Local Alig...
UCSC Genome Browser ...
GeneCards Homepage
National Cancer Institute
National Human Genome Research Institute
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
4:55 AM
2/15/2013
Translational Genomics Research Institute
TGEN

Flow in one portion of NCI Drug Discovery Program



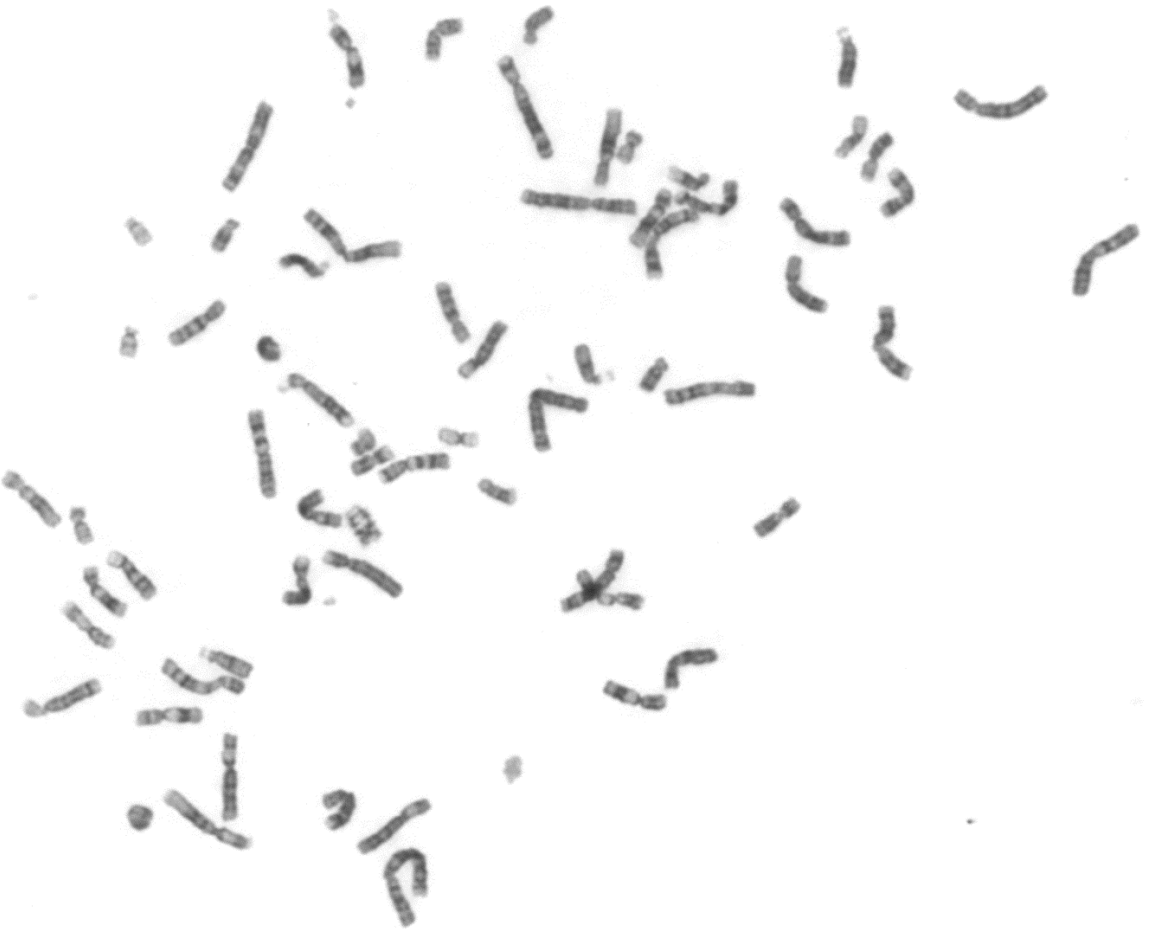
Integration is the goal, but…

- 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

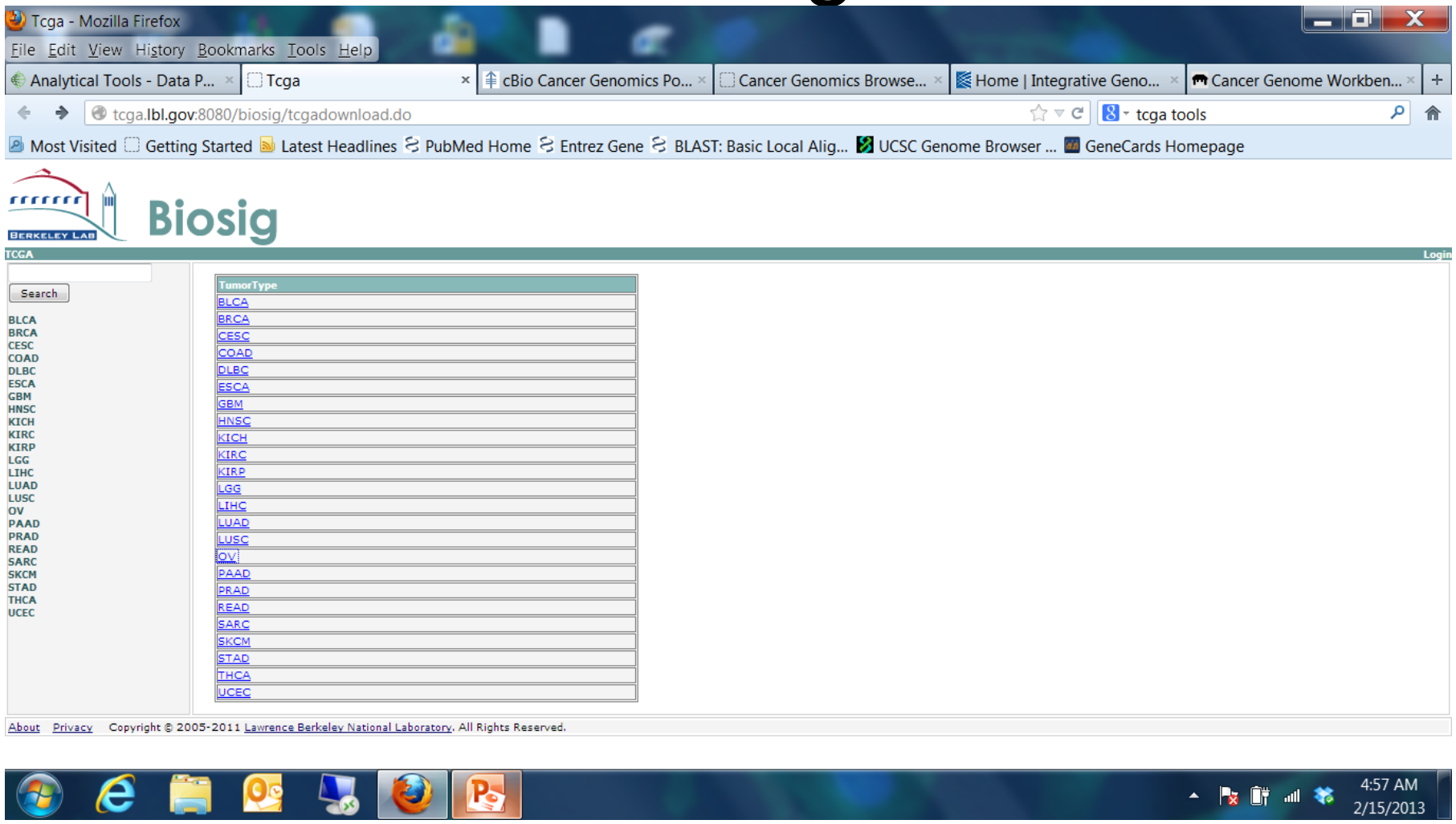
Cancer →



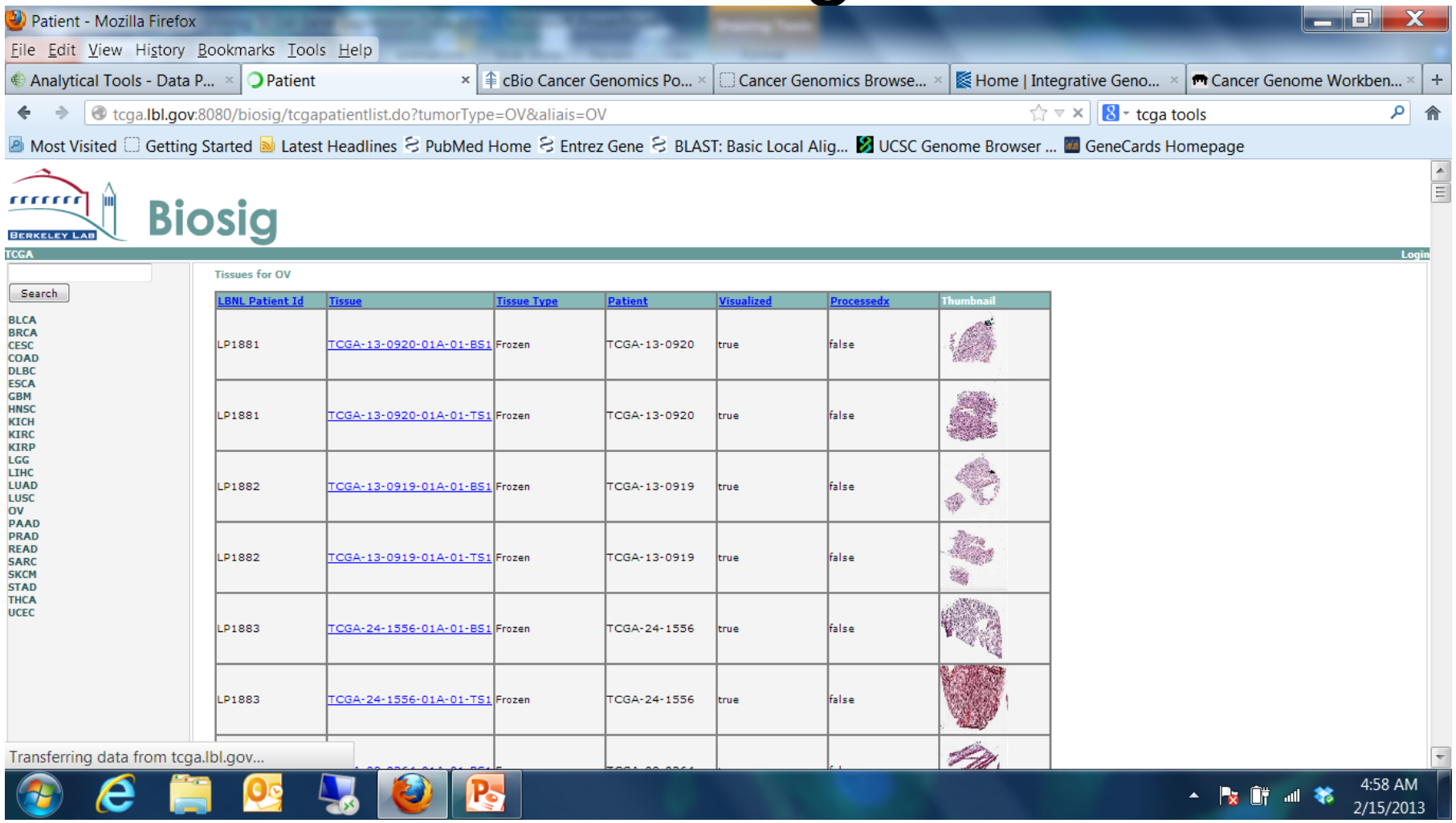
Tools and Questions

- 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

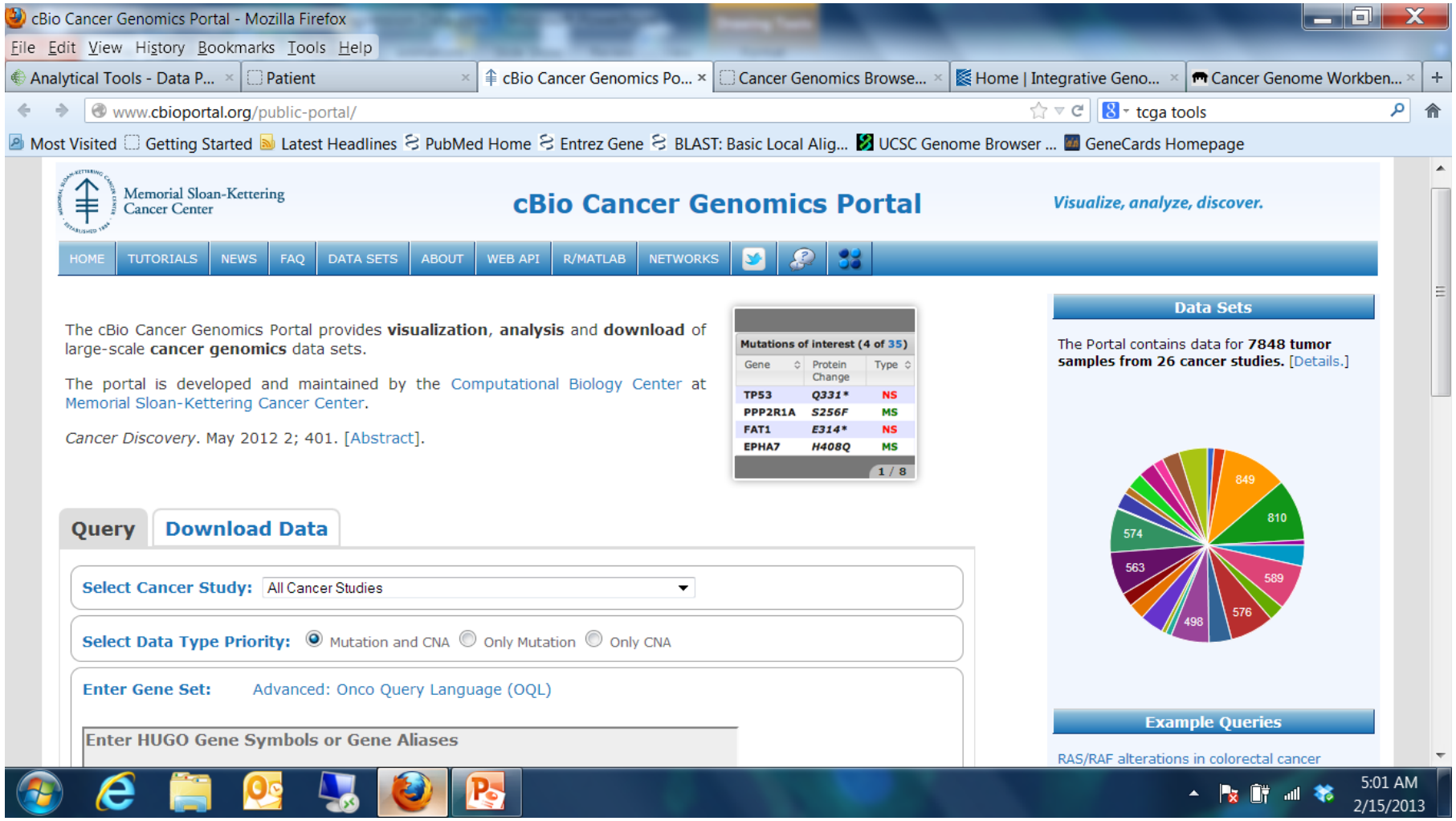
Biosig


Tcga - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Analytical Tools - Data P...
Tcga
cBio Cancer Genomics Po...
Cancer Genomics Browse...
Home | Integrative Geno...
Cancer Genome Workben...
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
Biosig
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
About
Privacy
Copyright © 2005-2011 Lawrence Berkeley National Laboratory. All Rights Reserved.
4:57 AM
2/15/2013

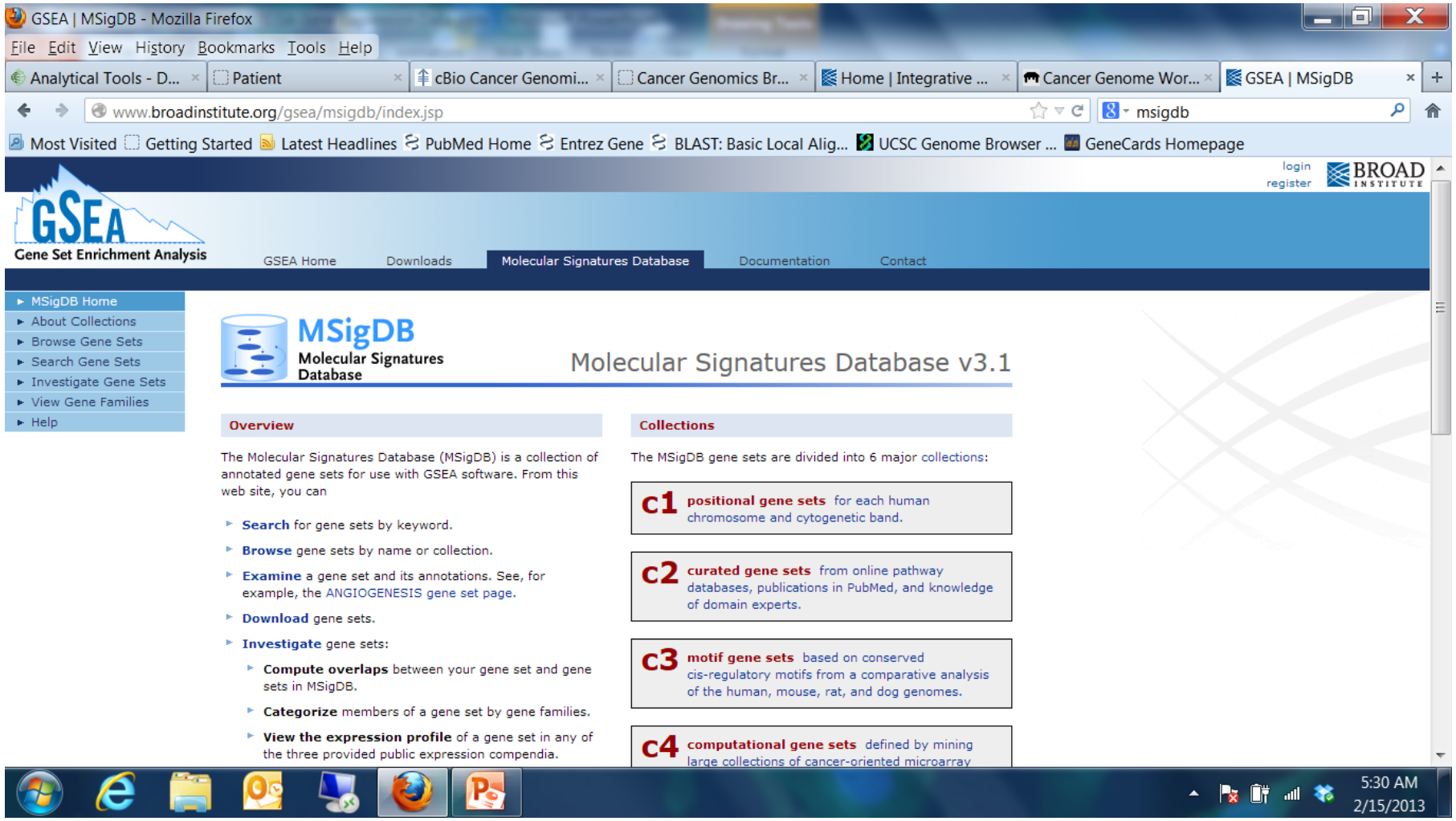
Biosig


Patient - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Analytical Tools - Data P...
Patient
cBio Cancer Genomics Po...
Cancer Genomics Browse...
Home | Integrative Geno...
Cancer Genome Workben...
tcga.lbl.gov:8080/biosig/tcgapatientlist.do?tumorType=OV&aliais=OV
tcga tools
Most Visited
Getting Started
Latest Headlines
PubMed Home
Entrez Gene
BLAST: Basic Local Alig...
UCSC Genome Browser ...
GeneCards Homepage
Biosig
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
Tissues for OV
LBNL Patient Id
Tissue
Tissue Type
Patient
Visualized
Processedx
Thumbnail
LP1881
TCGA-13-0920-01A-01-BS1
Frozen
TCGA-13-0920
true
false
LP1881
TCGA-13-0920-01A-01-TS1
Frozen
TCGA-13-0920
true
false
LP1882
TCGA-13-0919-01A-01-BS1
Frozen
TCGA-13-0919
true
false
LP1882
TCGA-13-0919-01A-01-TS1
Frozen
TCGA-13-0919
true
false
LP1883
TCGA-24-1556-01A-01-BS1
Frozen
TCGA-24-1556
true
false
LP1883
TCGA-24-1556-01A-01-TS1
Frozen
TCGA-24-1556
true
false
Transferring data from tcga.lbl.gov...
4:58 AM
2/15/2013

cBio


cBio Cancer Genomics Portal - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Analytical Tools - Data P...
Patient
cBio Cancer Genomics Po...
Cancer Genomics Browse...
Home | Integrative Geno...
Cancer Genome Workben...
www.cbioportal.org/public-portal/
tcga tools
Most Visited
Getting Started
Latest Headlines
PubMed Home
Entrez Gene
BLAST: Basic Local Alig...
UCSC Genome Browser ...
GeneCards Homepage
Memorial Sloan-Kettering Cancer Center
cBio Cancer Genomics Portal
Visualize, analyze, discover.
HOME
TUTORIALS
NEWS
FAQ
DATA SETS
ABOUT
WEB API
R/MATLAB
NETWORKS
The cBio Cancer Genomics Portal provides visualization, analysis and download of large-scale cancer genomics data sets.
The portal is developed and maintained by the Computational Biology Center at Memorial Sloan-Kettering Cancer Center.
Cancer Discovery. May 2012 2; 401. [Abstract].
Mutations of interest (4 of 35)
Gene
Protein Change
Type
TP53 Q331* NS
PPP2R1A S256F MS
FAT1 E314* NS
EPHA7 H408Q MS
1 / 8
Data Sets
The Portal contains data for 7848 tumor samples from 26 cancer studies. [Details.]
849
810
574
563
589
576
498
Query
Download Data
Select Cancer Study: All Cancer Studies
Select Data Type Priority: Mutation and CNA Only Mutation Only CNA
Enter Gene Set: Advanced: Onco Query Language (OQL)
Enter HUGO Gene Symbols or Gene Aliases
Example Queries
RAS/RAF alterations in colorectal cancer
5:01 AM
2/15/2013

Signatures/Gene Sets - MSigdb



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.

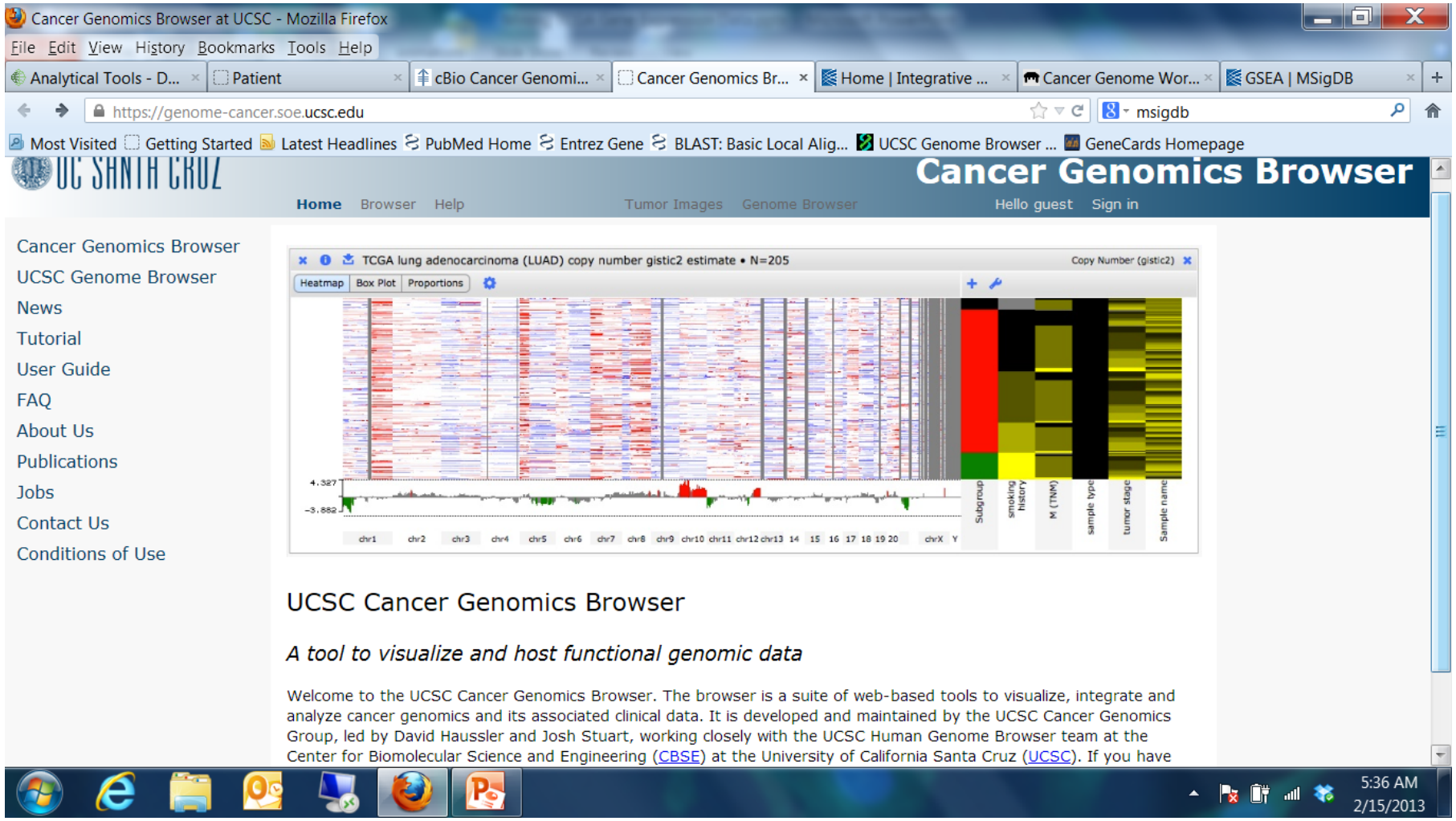
c3 **motif gene sets** based on conserved cis-regulatory motifs from a comparative analysis of the human, mouse, rat, and dog genomes.

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

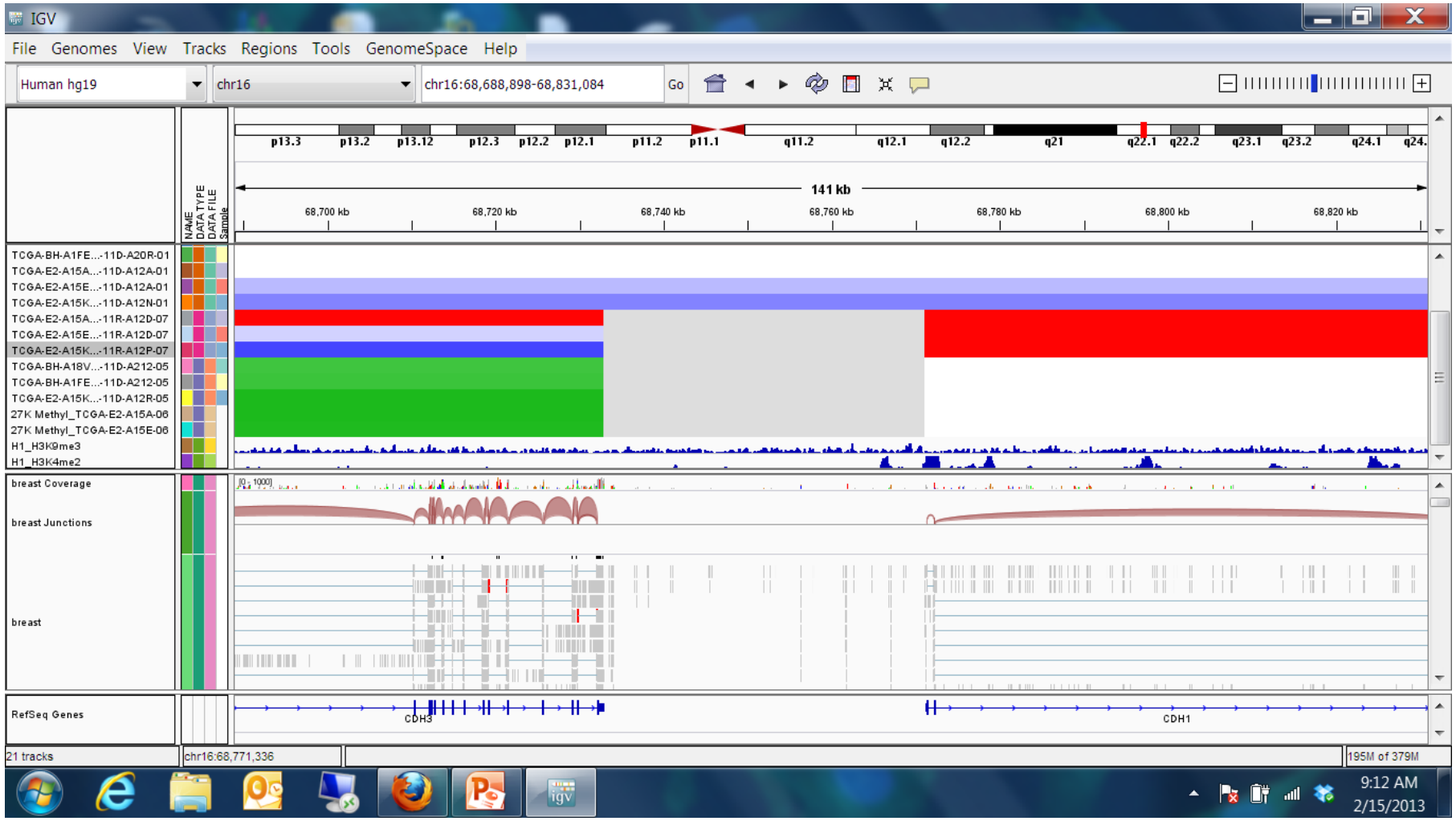


Cancer Genome Workbench

IGV


Home | Integrative Genomics Viewer - 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/software/igv/
Home | Integrative Genomics Viewer
Most Visited
Getting Started
Latest Headlines
PubMed Home
Entrez Gene
BLAST: Basic Local Alig...
UCSC Genome Browser ...
GeneCards Homepage
Integrative Genomics Viewer
Home
Downloads
Documents
Hosted Genomes
FAQ
IGV User Guide
File Formats
Release Notes
Credits
Contact
Search website
search
Broad Home
Cancer Program
BROAD INSTITUTE
© 2012 Broad Institute
Home
Integrative Genomics Viewer
What's New
December 18, 2012. IGV 2.2 has been released. See the release notes for more details.
April 20, 2012. IGV 2.1 has been released. See the release notes for more details.
April 19, 2012. See our new IGV paper in Briefings in Bioinformatics.
Citing IGV
To cite your use of IGV in your publication:
Helga Thorvaldsdóttir, James T. Robinson, Jill P. Mesirov. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Briefings in Bioinformatics 2012.
James T. Robinson, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S. Lander, Gad Getz, Jill P. Mesirov.
5:46 AM
2/15/2013

IGV


IGV
File Genomes View Tracks Regions Tools GenomeSpace Help
Human hg19
chr16
chr16:68,688,898-68,831,084
Go
p13.3 p13.2 p13.12 p12.3 p12.2 p12.1 p11.2 p11.1 q11.2 q12.1 q12.2 q21 q22.1 q22.2 q23.1 q23.2 q24.1 q24.
141 kb
68,700 kb 68,720 kb 68,740 kb 68,760 kb 68,780 kb 68,800 kb 68,820 kb
NAME DATA TYPE DATA FILE Sample
TCGA-BH-A1FE...-11D-A20R-01
TCGA-E2-A15A...-11D-A12A-01
TCGA-E2-A15E...-11D-A12A-01
TCGA-E2-A15K...-11D-A12N-01
TCGA-E2-A15A...-11R-A12D-07
TCGA-E2-A15E...-11R-A12D-07
TCGA-E2-A15K...-11R-A12P-07
TCGA-BH-A18V...-11D-A212-05
TCGA-BH-A1FE...-11D-A212-05
TCGA-E2-A15K...-11D-A12R-05
27K Methyl_TCGA-E2-A15A-06
27K Methyl_TCGA-E2-A15E-06
H1_H3K9me3
H1_H3K4me2
breast Coverage
breast Junctions
breast
RefSeq Genes
CDH3
CDH1
21 tracks
chr16:68,771,336
195M of 379M
9:12 AM
2/15/2013

Take Home Points

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

QUESTIONS?