BMI500
Genomics Exercises

Lee AD Cooper
Assistant Professor of Biomedical Informatics
Assistant Professor of Biomedical Engineering
Emory University / Georgia Institute of Technology

Agenda

- Introduction
- Background material - cancer
- Activity 1: Investigate *TP53* using CBioPortal, NCBI Entrez, UCSC Genome Browser
- Activity 2: Cluster Analysis of Gene Expression Data

Omic Databases

- Too many omics databases to count – primary categories are:
  - Databases housing genomic data
    - NCI Genomic Data Commons
    - NCBI Gene Omnibus Expression (GEO)
  - Databases defining –omes (genomes, proteomes, glycomes, etc.)
    - UCSC genome browser (genomes)
    - UniProt (proteome)
    - UniCarbKB (glycomics)
  - Databases that map genetic variation & disease
    - COSMIC (Cancer Alterations)
    - dbSNP and ClinVar (NCBI)
  - Databases defining molecular interactions and pathways
    - NDEx – The Network Data Exchange
    - MSigDB – Molecular Signatures DB
    - STRING – Protein interaction networks
  - Databases defining standards and ontologies
    - HUGO Gene Nomenclature Committee
    - Geneontology.org

Background Material - Cancer

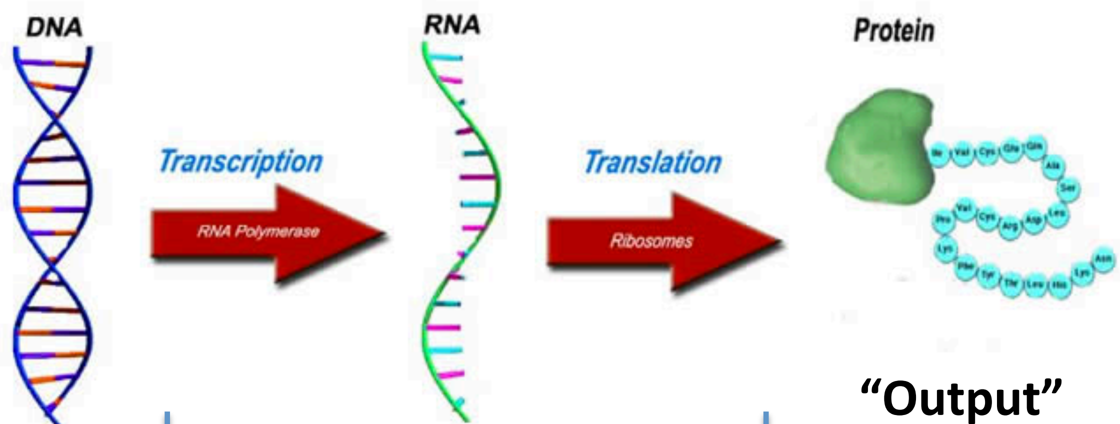
Biology 101 – Simplified (overly)

Instructions for making proteins

Requests for proteins

Proteins

Molecular Circuits



"Output"

"Source Code"

"Program in Execution"

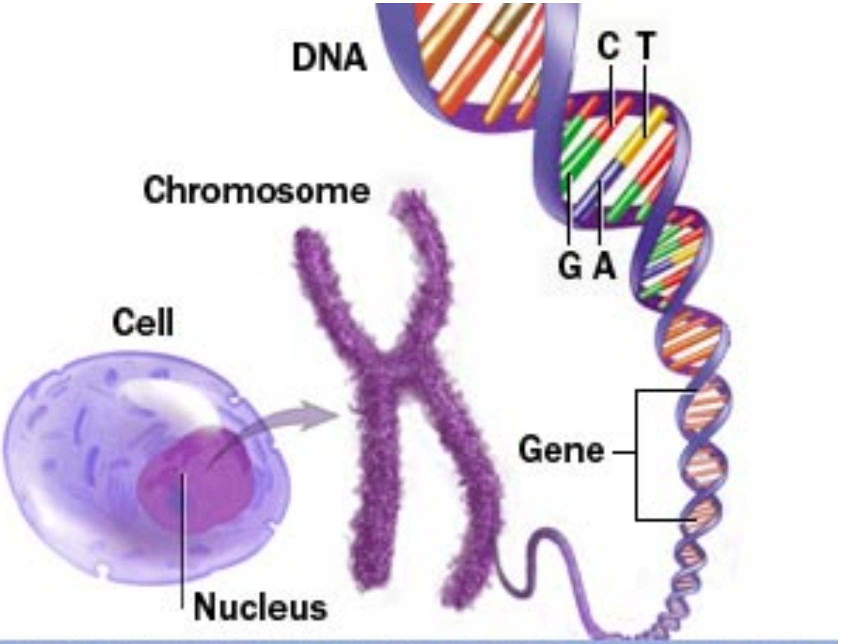
"Life" – "Cell Behavior"

Organization of DNA

- ***Gene*** – a specific region of DNA containing the instructions for making a protein (you have two copies of each)

- ***Chromosome*** – a collection of genes (you have 23 pairs, each has two arms)

- ***Genome*** – the whole thing



Cancer Simplified

- Cancer is a disease of the DNA
  - Corruption of the source code
- Leads to dysregulation of biological processes
  - Cell division
  - Programmed death
  - DNA maintenance and repair
- Sources of corruption
  - Mutations – flipping bits, truncations, bit shifts etc
  - Copy number – deletion or duplication of DNA
- Consequences
  - DNA → RNA → Protein → cell behavior

Activity 1:
Investigate *TP53* using CBioPortal, NCBI Entrez, UCSC Genome Browser

How do elephants avoid cancer?

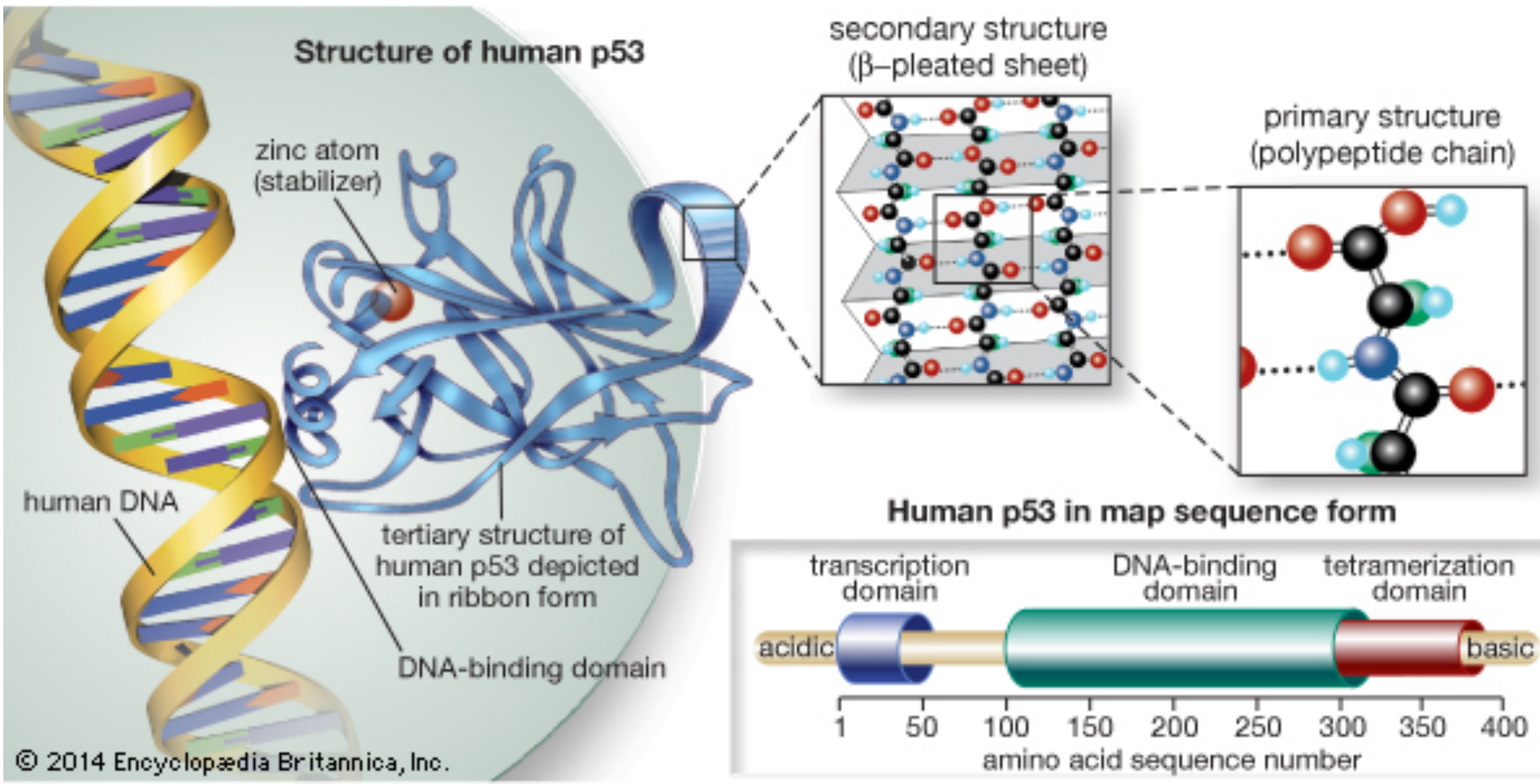
Theo Allofs/Minden Pictures/FLPA

Multiple copies of a tumour-suppressor gene help elephants avoid cancer.

Peto's Paradox

- Richard Peto - Oxford University
- *Why no relationship between body size and cancer?*
- Cancer is caused by alterations to genetic code
  - UV, carcinogen exposure
- Bigger animal → more cells → more opportunities for "bad luck"
- (Age – also more accumulation of events)

TP53: the Guardian of the Genome

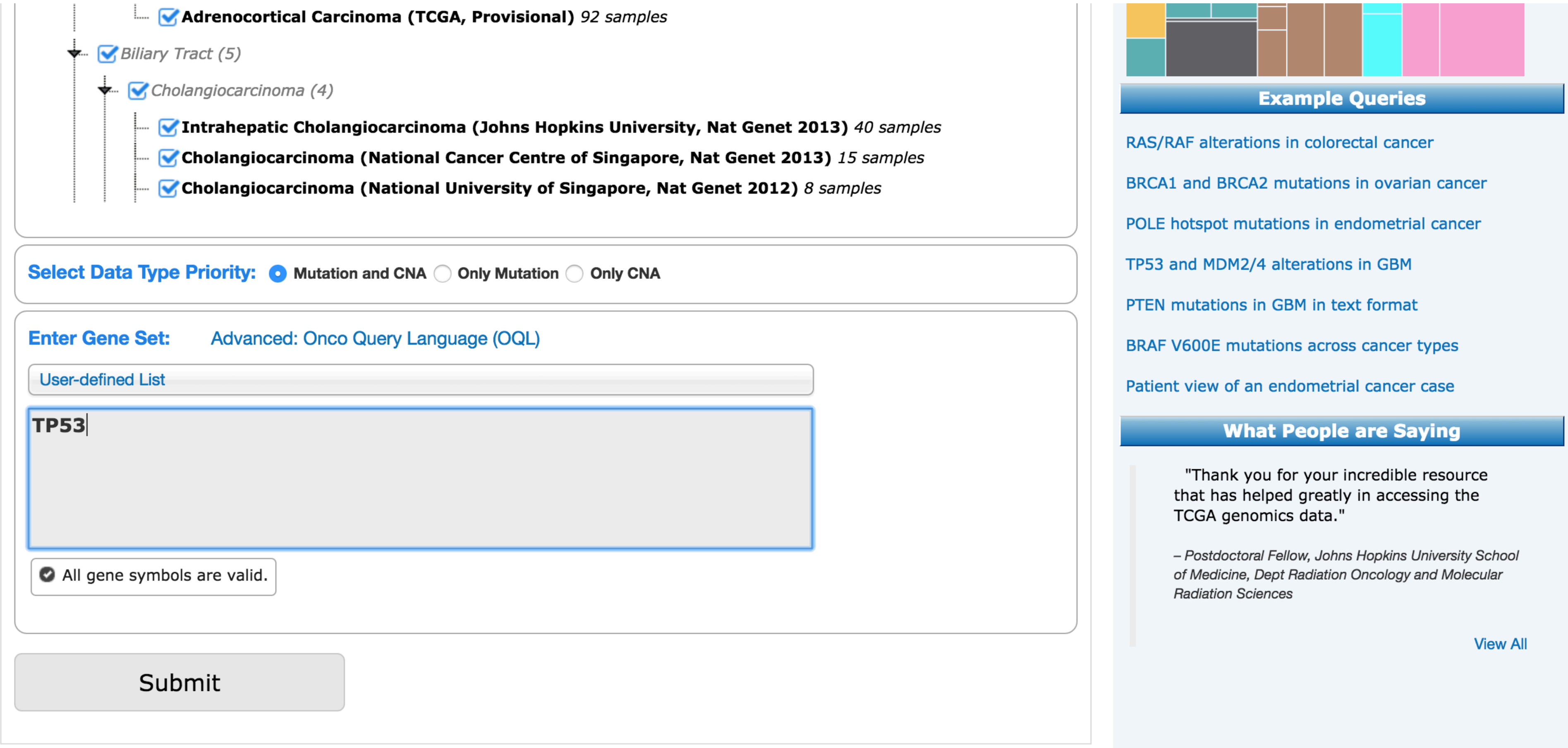


Cell Cycle Checkpoint

- TP53 is the most frequently mutated gene in human cancers

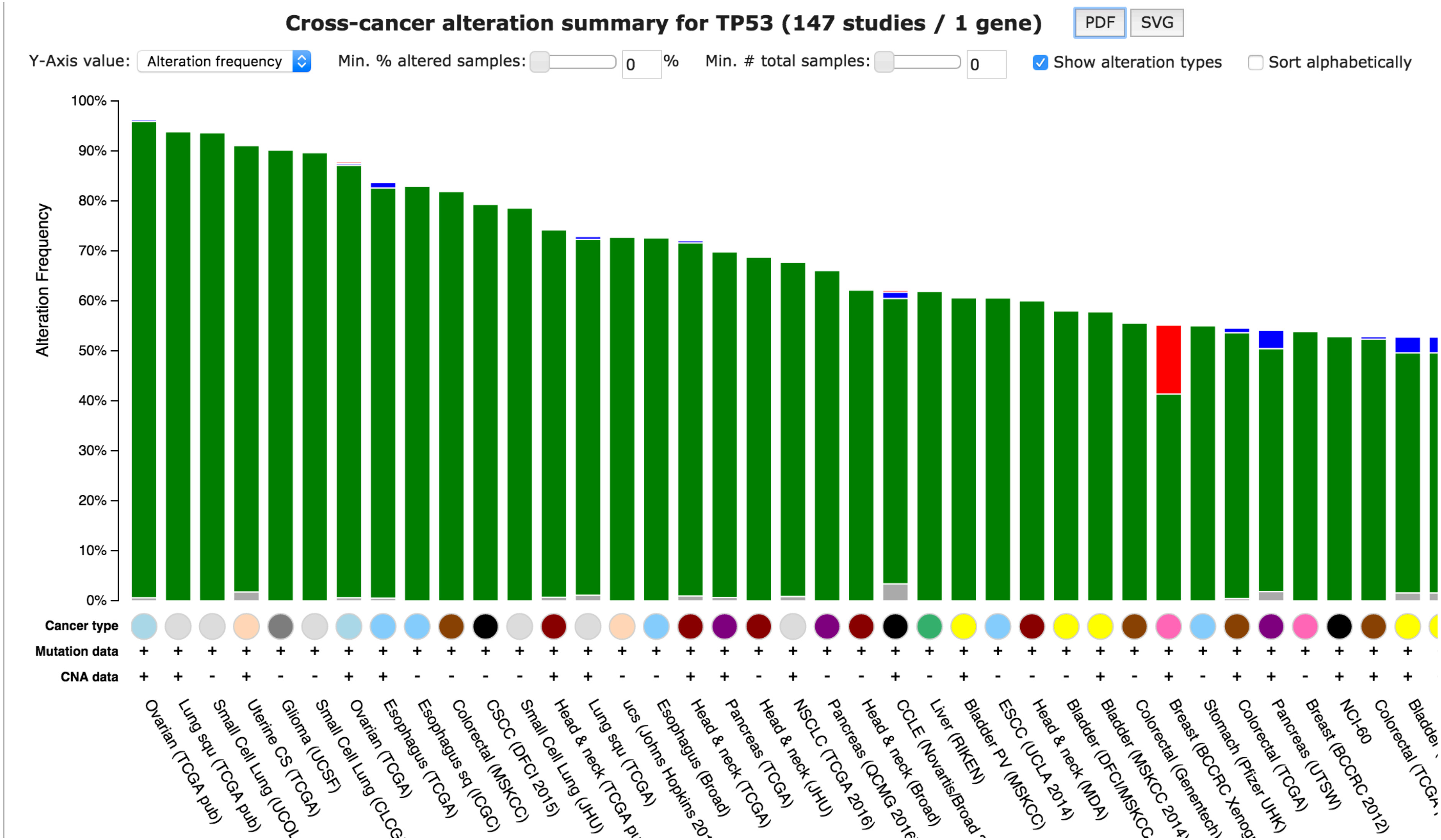
- Navigate to http://www.cbioportal.org

Select “All” cancer datasets

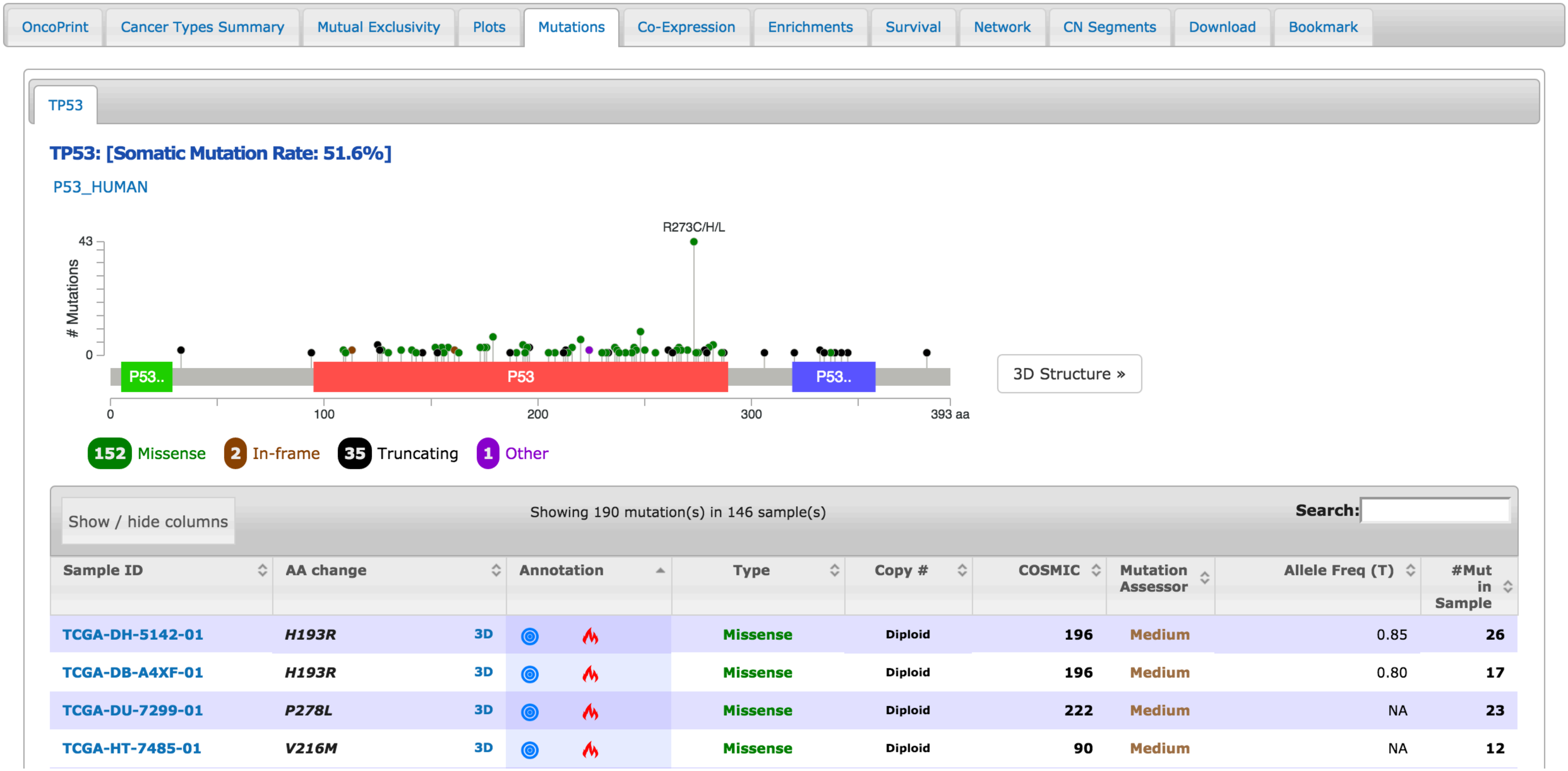


Adrenocortical Carcinoma (TCGA, Provisional) 92 samples
Biliary Tract (5)
Cholangiocarcinoma (4)
Intrahepatic Cholangiocarcinoma (Johns Hopkins University, Nat Genet 2013) 40 samples
Cholangiocarcinoma (National Cancer Centre of Singapore, Nat Genet 2013) 15 samples
Cholangiocarcinoma (National University of Singapore, Nat Genet 2012) 8 samples
Select Data Type Priority:
Mutation and CNA
Only Mutation
Only CNA
Enter Gene Set:
Advanced: Onco Query Language (OQL)
User-defined List
TP53
All gene symbols are valid.
Submit
Example Queries
RAS/RAF alterations in colorectal cancer
BRCA1 and BRCA2 mutations in ovarian cancer
POLE hotspot mutations in endometrial cancer
TP53 and MDM2/4 alterations in GBM
PTEN mutations in GBM in text format
BRAF V600E mutations across cancer types
Patient view of an endometrial cancer case
What People are Saying
"Thank you for your incredible resource that has helped greatly in accessing the TCGA genomics data."
– Postdoctoral Fellow, Johns Hopkins University School of Medicine, Dept Radiation Oncology and Molecular Radiation Sciences
View All

Enter “TP53”



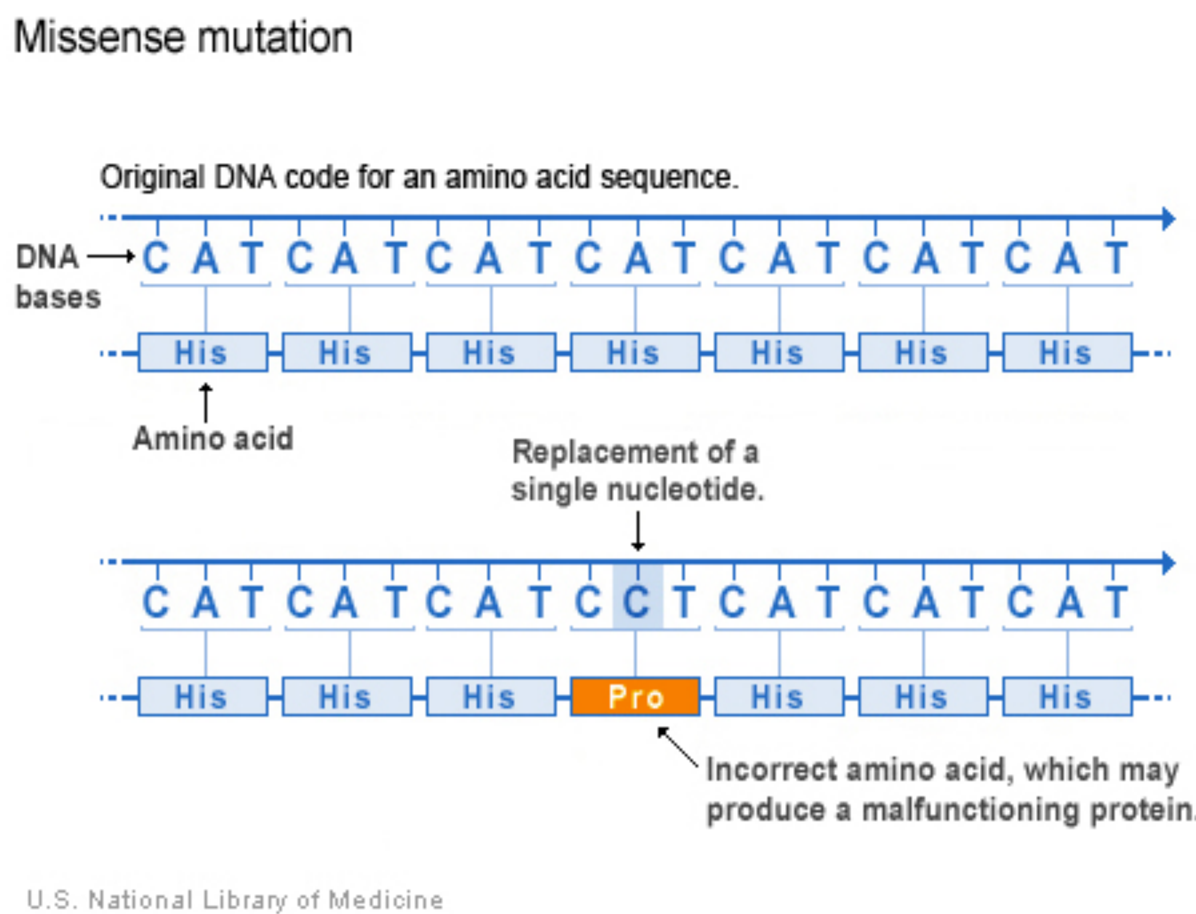
Select “Glioma”



Select "Mutations"

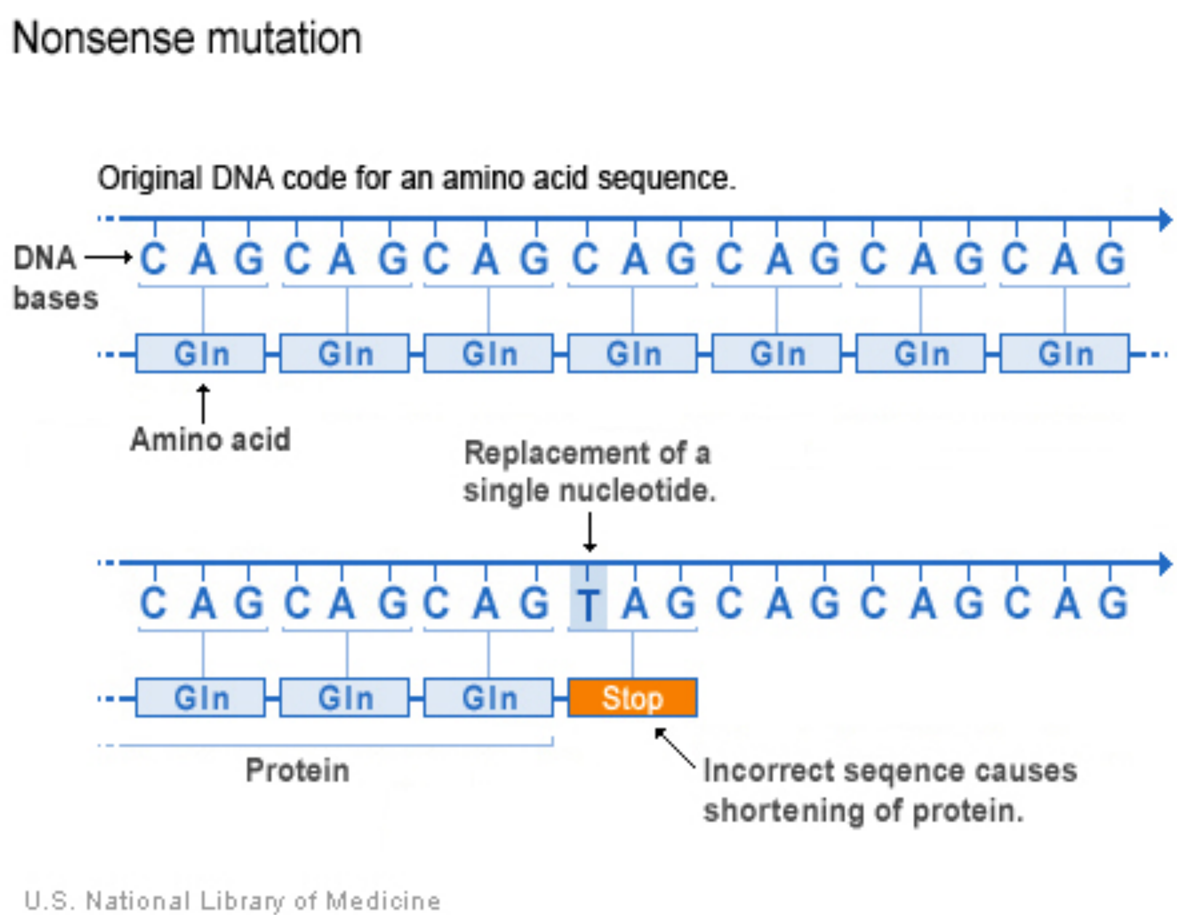
Missense Mutation

- Single base substitution
- Frequently "activating" but not always



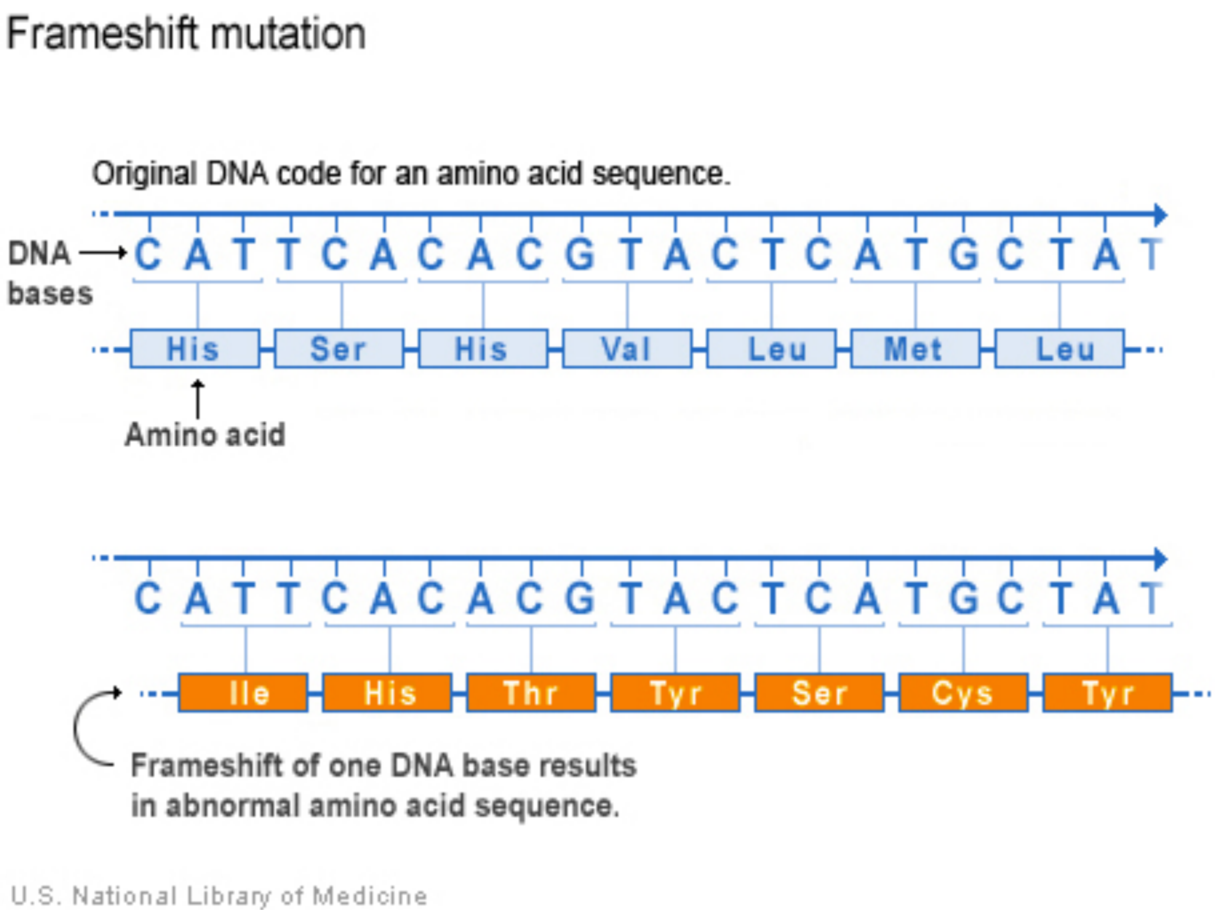
Nonsense Mutation

- Single base substitution
- Premature truncation signal



Frameshift Mutation

- Shift of reading frame by insertion or deletion of bases



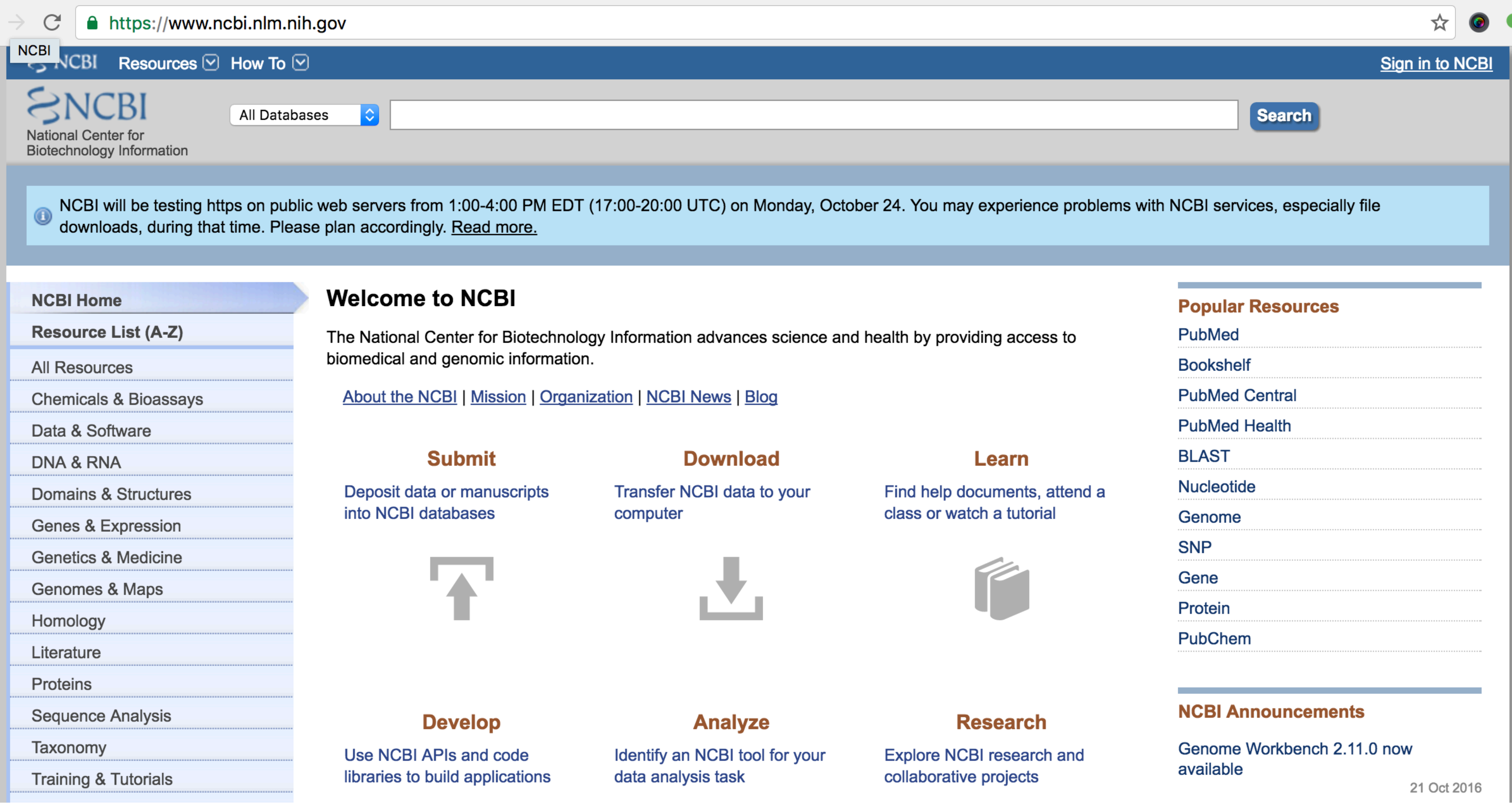
NCBI - Entrez

- National Center for Biotechnology Information

  https://www.ncbi.nlm.nih.gov

- National Library of Medicine / National Institutes of Health

- Clearinghouse for biomedical information, databases and services



Select “Gene” and search “TP53”

Using Sequence Viewer…

- Zoom out to see who's in the neighborhood
- Why so many versions?
- Select a sequence variant – what do the green, blue and red represent?
- Accession numbers for sequences, proteins
- Zoom in until you see sequence
- What are ClinVar, dbSNP?

Other NCBI Information

- How many *TP53*-related articles are in PubMed?
- What medical conditions are associated with *TP53*?
- Find the accession number for a protein that *TP53* interacts with
- What are some biological processes associated with *TP53*?
- What are some locations where *TP53* is found (localized) within a cell?

Summary

- Used CBioPortal to investigate frequency and localization of mutations of TP53 in human cancers
- Discussed mutation types and visualized 3D protein structure
- Queried NCBI gene database to identify
  - mRNA isoforms
  - Medical conditions associated with *TP53*
  - *TP53* interacting partners
  - *TP53*-related publications

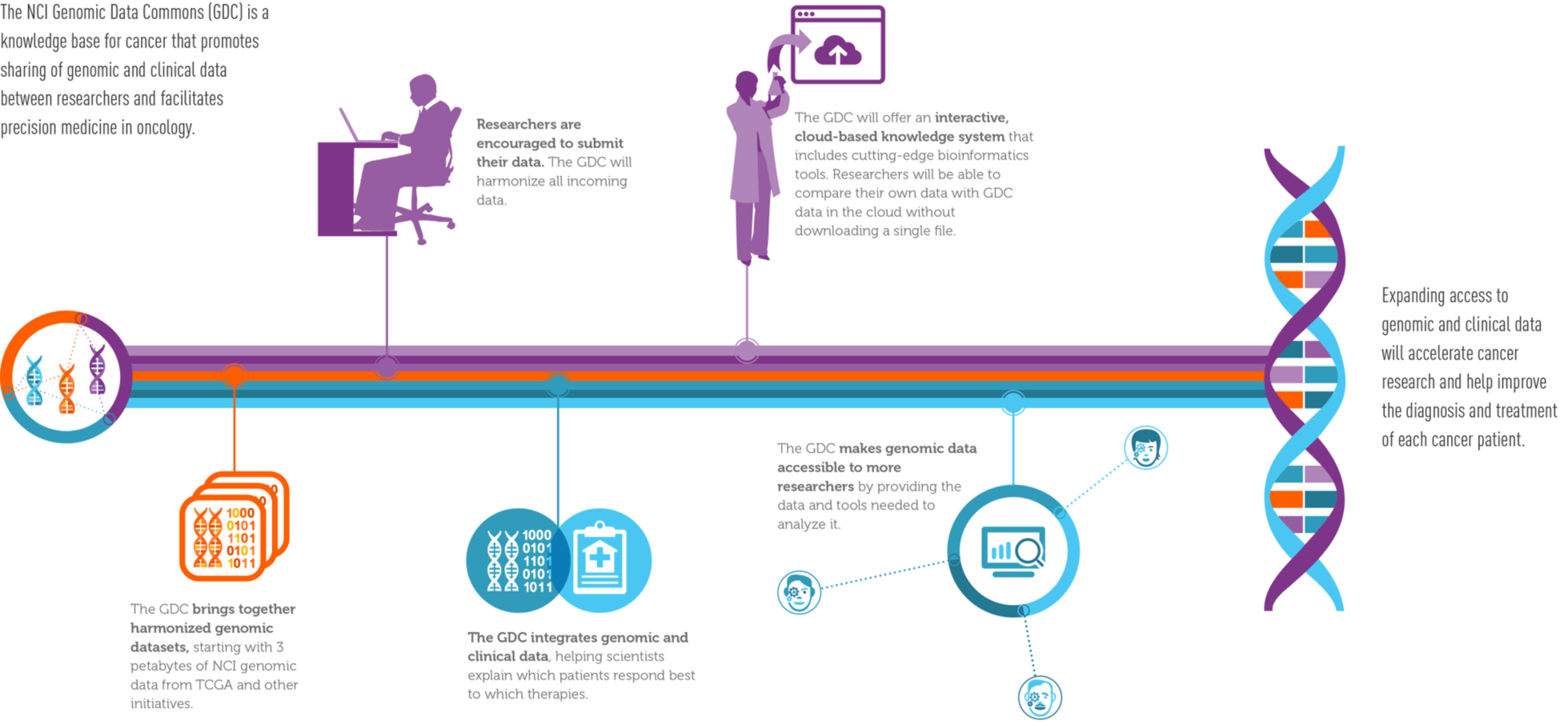
Activity 2:
Cluster Analysis of Gene Expression Data

Genomic Data Commons

- https://gdc.cancer.gov/
- A clearinghouse for cancer genomic and clinical data
- From summary spreadsheets to protected read-level data

NATIONAL CANCER INSTITUTE
GENOMIC DATA COMMONS

The NCI Genomic Data Commons (GDC) is a knowledge base for cancer that promotes sharing of genomic and clinical data between researchers and facilitates precision medicine in oncology.

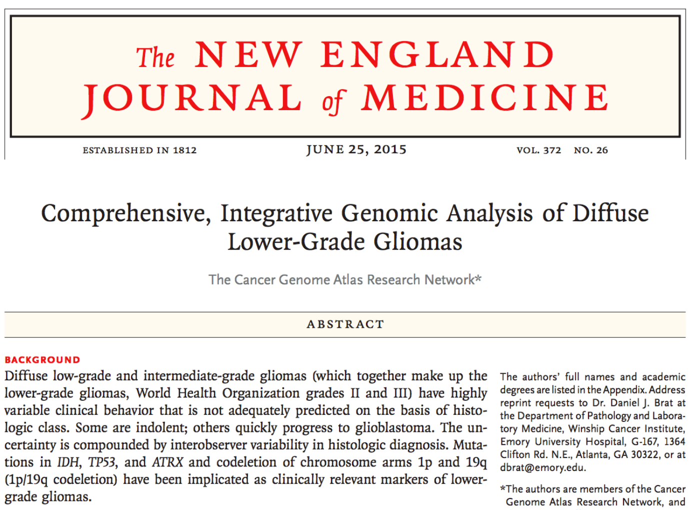


www.cancer.gov

Cancer moonshot – data sharing

Exercise summary

- We will cluster brain tumor patients by their mRNA expression profiles
- Exercise2.mat - goo.gl/E7OMZ6


The NEW ENGLAND JOURNAL of MEDICINE

ESTABLISHED IN 1812 JUNE 25, 2015 VOL. 372 NO. 26

Comprehensive, Integrative Genomic Analysis of Diffuse Lower-Grade Gliomas

The Cancer Genome Atlas Research Network*

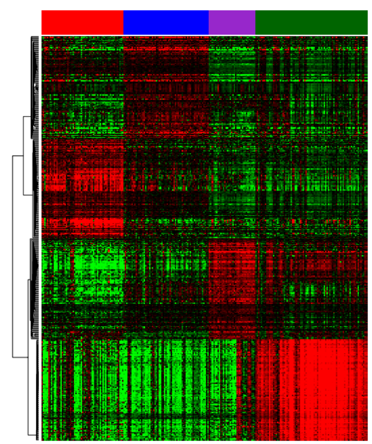
ABSTRACT

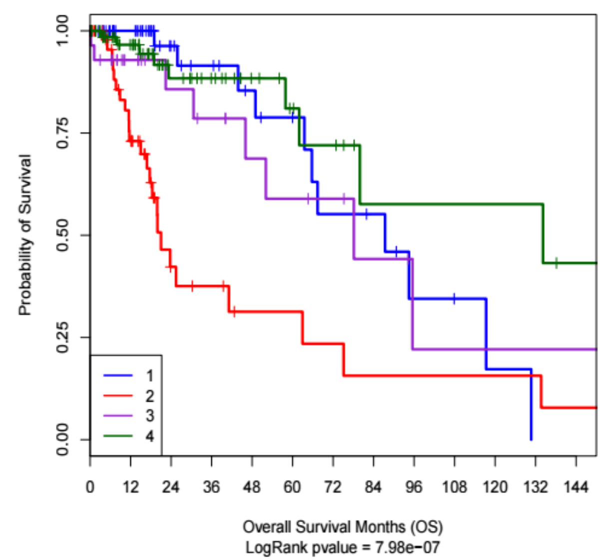
BACKGROUND

Diffuse low-grade and intermediate-grade gliomas (which together make up the lower-grade gliomas, World Health Organization grades II and III) have highly variable clinical behavior that is not adequately predicted on the basis of histologic class. Some are indolent; others quickly progress to glioblastoma. The uncertainty is compounded by interobserver variability in histologic diagnosis. Mutations in *IDH*, *TP53*, and *ATRX* and codeletion of chromosome arms 1p and 19q (1p/19q codeletion) have been implicated as clinically relevant markers of lower-grade gliomas.

The authors' full names and academic degrees are listed in the Appendix. Address reprint requests to Dr. Daniel J. Brat at the Department of Pathology and Laboratory Medicine, Winship Cancer Institute, Emory University Hospital, G-167, 1364 Clifton Rd. N.E., Atlanta, GA 30322, or at dbrat@emory.edu.

*The authors are members of the Cancer Genome Atlas Research Network, and





Exercise Summary

Filtering Steps

- Remove features with > 20% ‘0’ values
- Rank genes by standard deviation
- Select top 500 genes with highest SDs

Generate a clustered heatmap

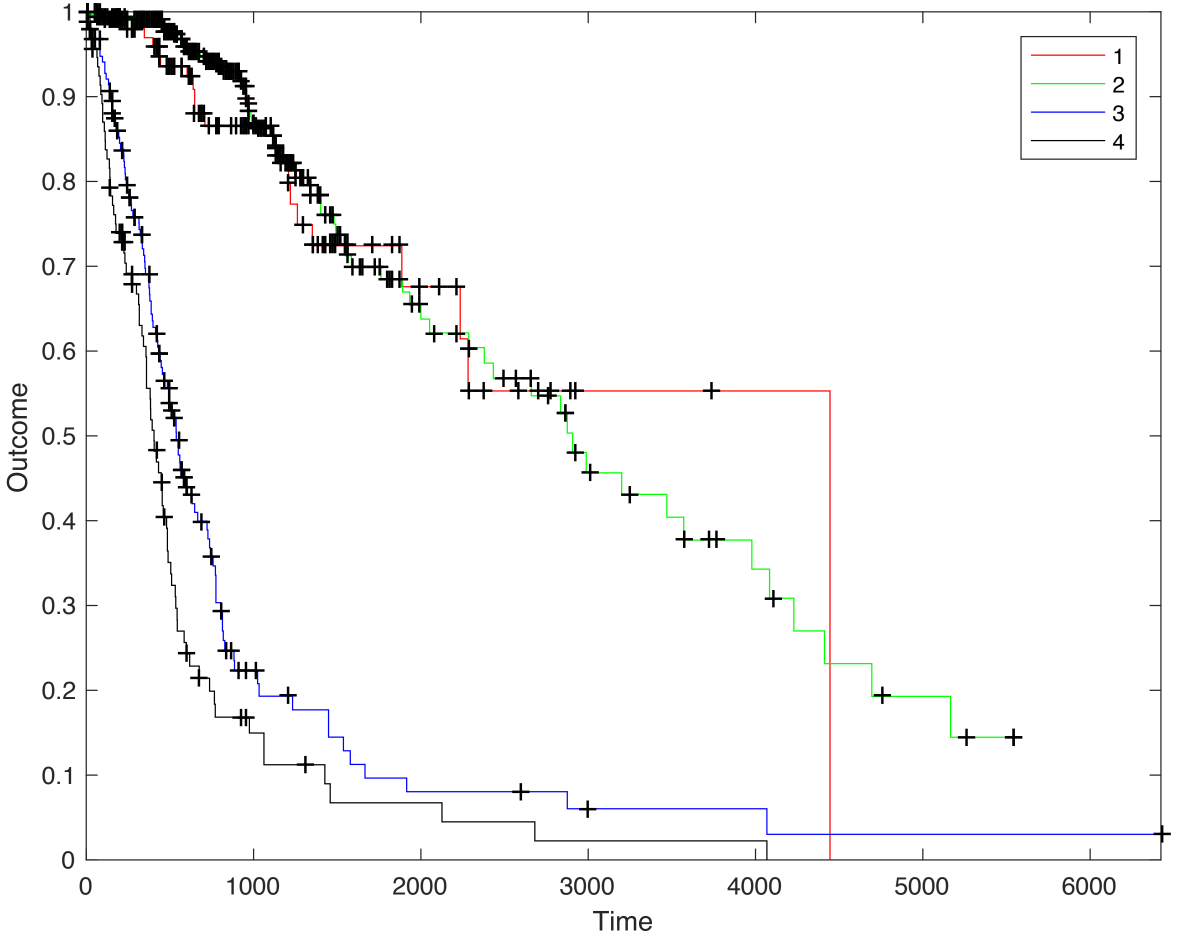
```
Subset = log(Subset + 1e-10);
Subset = (Subset - mean(Subset, 2) * ...
            ones(1, length(Survival))) ./ ...
              (std(Subset, [], 2) * ...
              ones(1, length(Survival)));

clustergram(Features, 'Colormap',redbluecmap);
```

Generate cluster labels

```
rng(1);
C = kmeans(Subset.', 4);
```

Survival Analysis - Kaplan Meier Plot

KMPlot(Survival, Censored, C, {'1', '2', '3', '4'});



Identify Key Genes

- Differential expression between clusters 1/2 and 3/4

```
Good = (C == 1 | C == 2);
Bad = (C == 3 | C == 4);
[~, p] = ttest2(Features(:, Good).', Features(:, Bad).');
Difference = median(Features(:, Bad), 2) - ...
   median(Features(:, Good), 2);
Score = -log(p) .* sign(Difference.');
[Score, Order] = sort(Score, 'descend');
cell2text(Symbols(Order(1:500)), 'Genes.txt');
```

Pathway Analysis

- Reactome.org