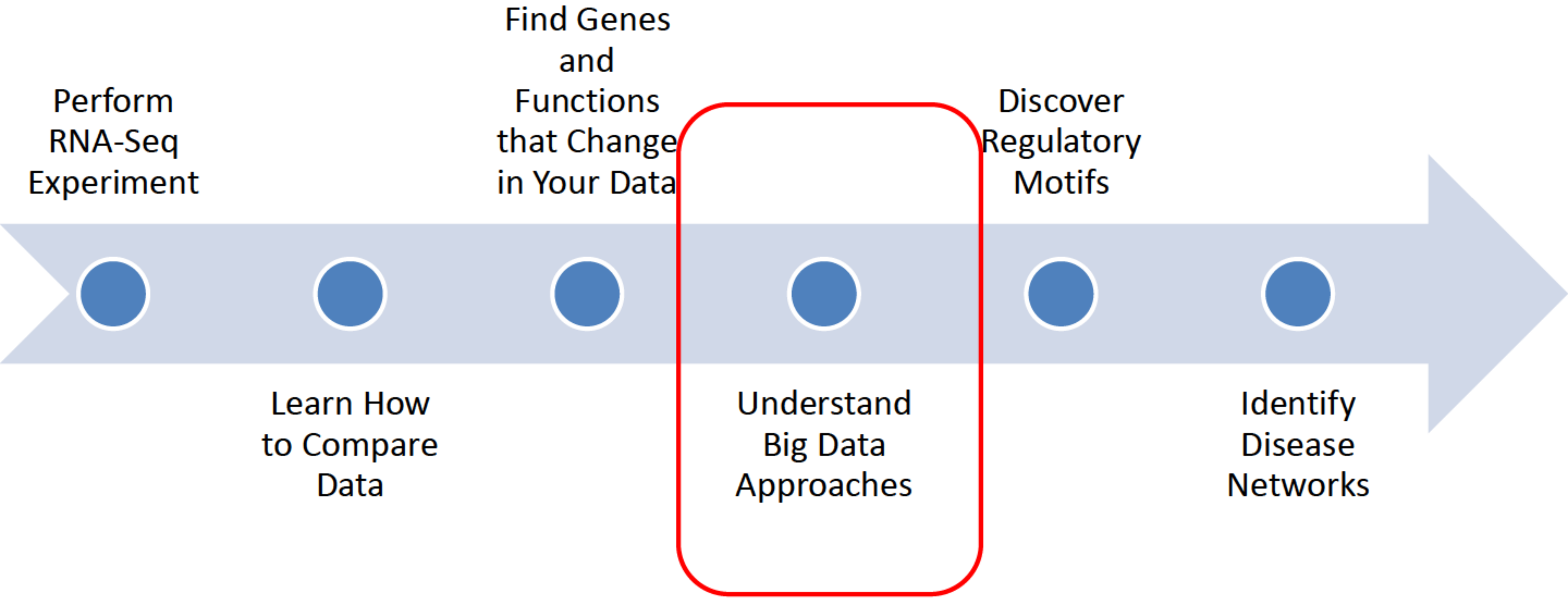


Perform RNA-Seq Experiment

Learn How to Compare Data

Find Genes and Functions that Change in Your Data

Understand Big Data Approaches

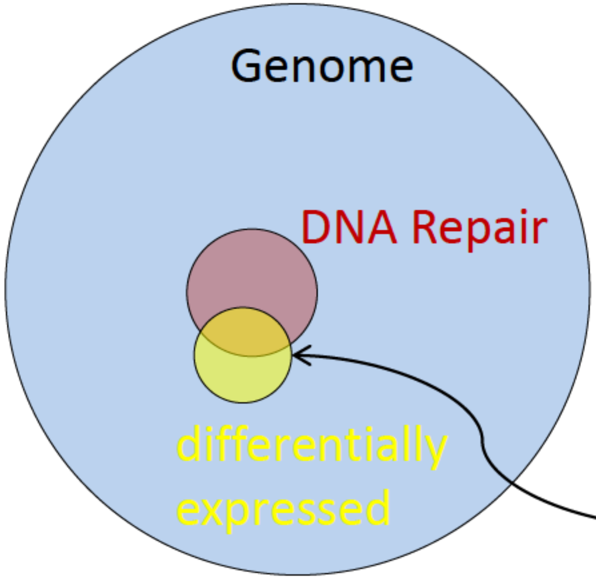
Discover Regulatory Motifs

Identify Disease Networks

Outline

- Statistical significance for gene annotations
- Big data
  - L1000 transcriptional assay
  - Chemical sensitivity dataset
  - PubChem
  - TCGA
  - Drug Repositioning

Statistical significance



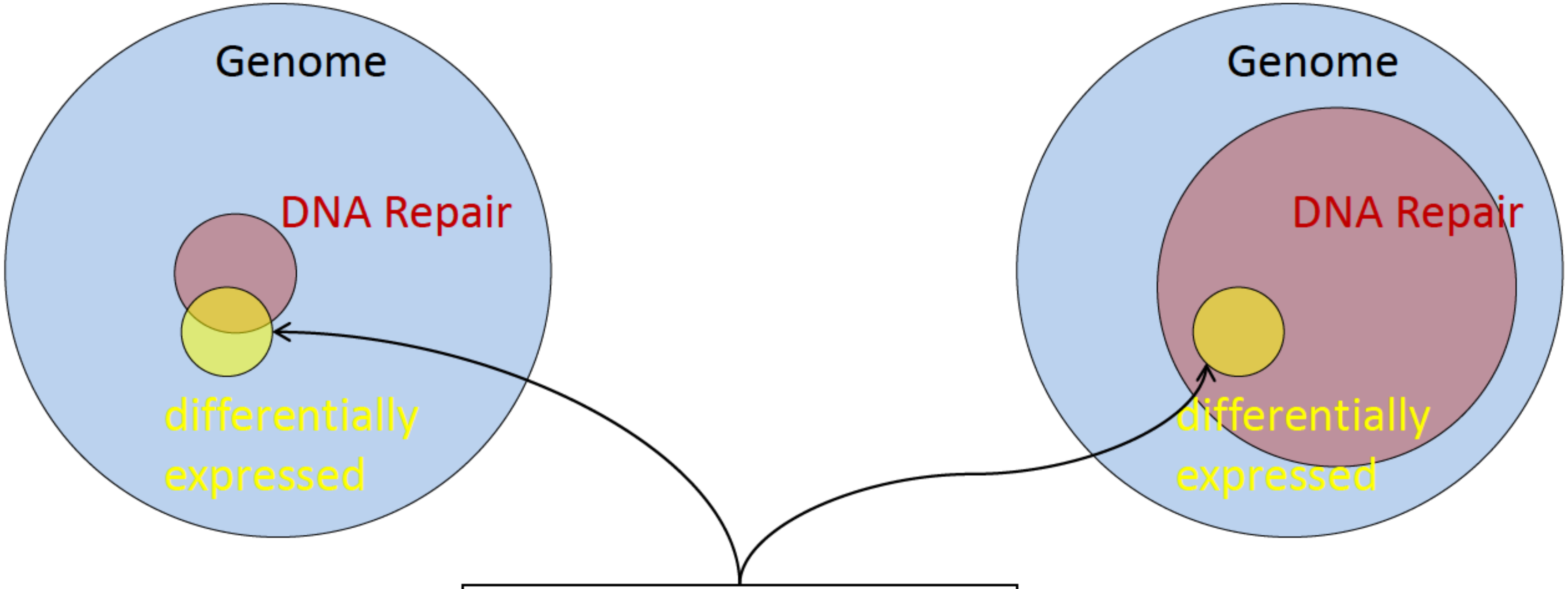
I found that ten of the upregulated genes in my dataset are annotated as“DNA Repair” ...

Is this overlap significant?

To answer this question we need a null model.

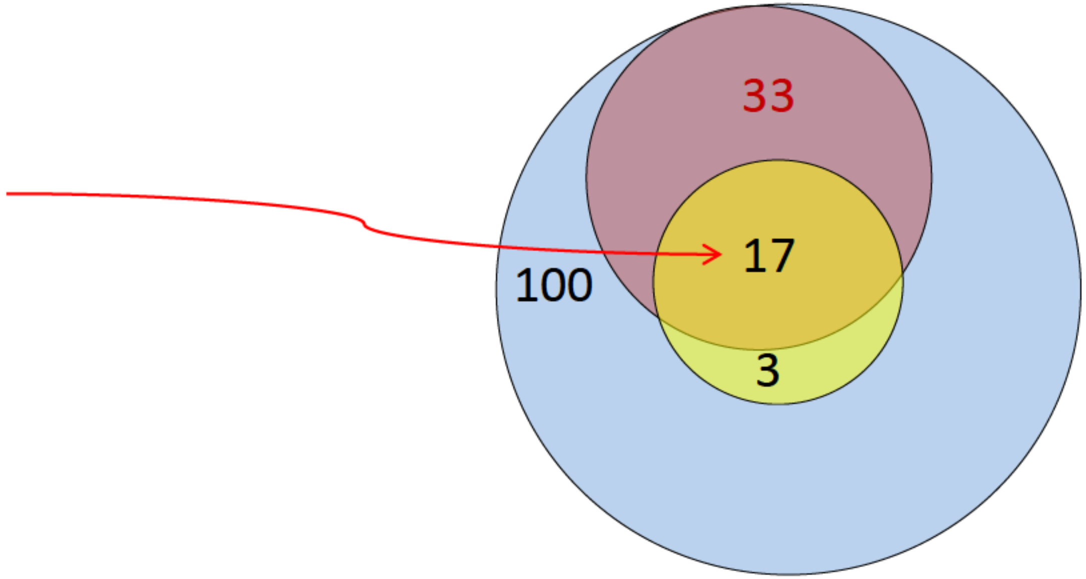
Statistical significance

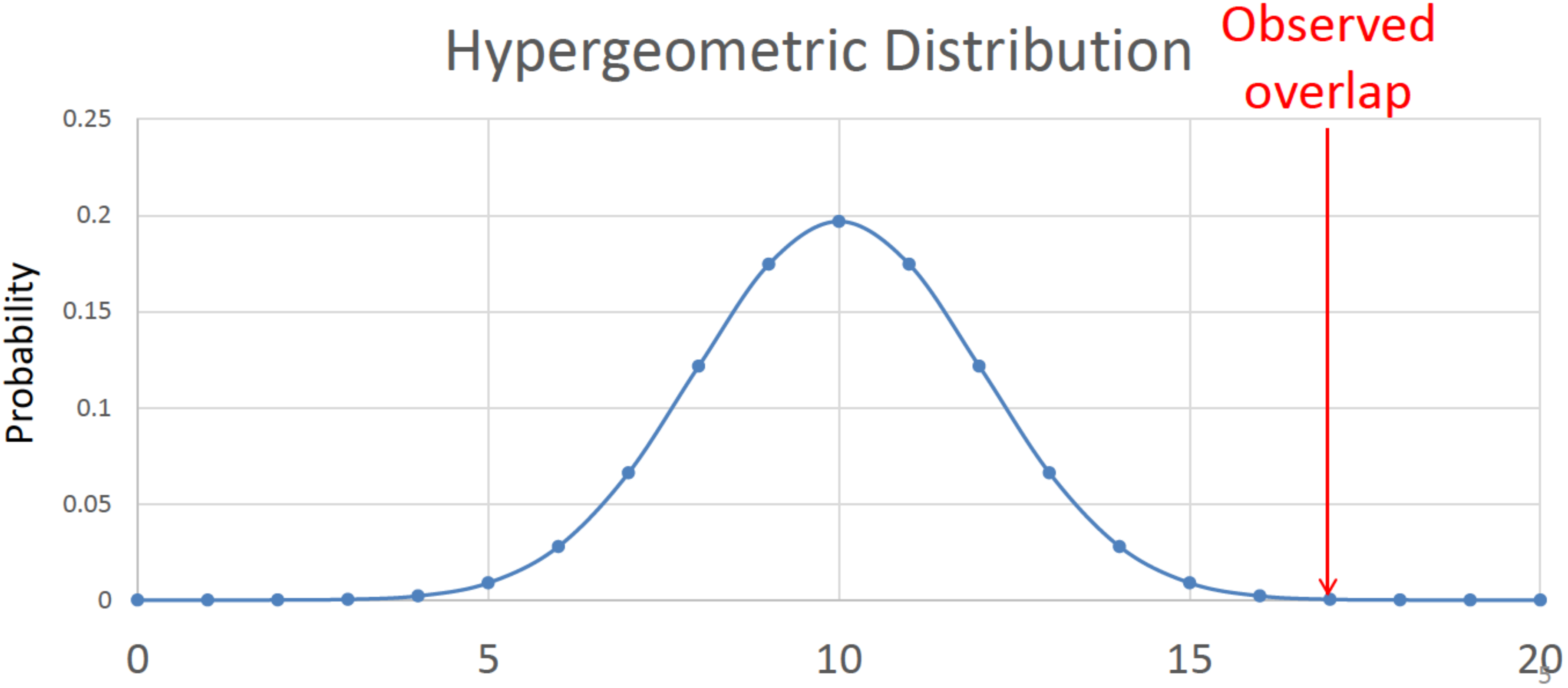
The significance depends on the size of the lists.



If the two lists had nothing in common, could we still get this degree of overlap?

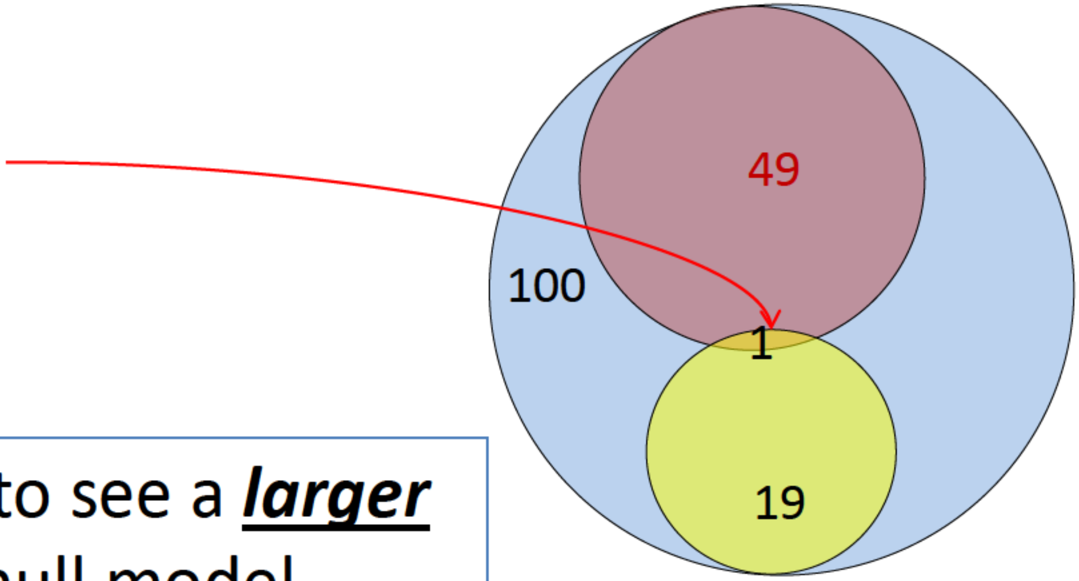
There are 17 overlapping genes.
Is that surprising?





Observed overlap

There is only one overlapping gene.
Is that surprising?

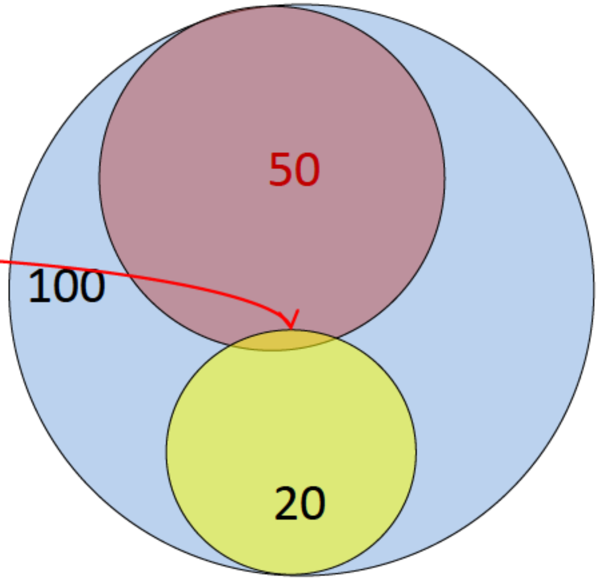


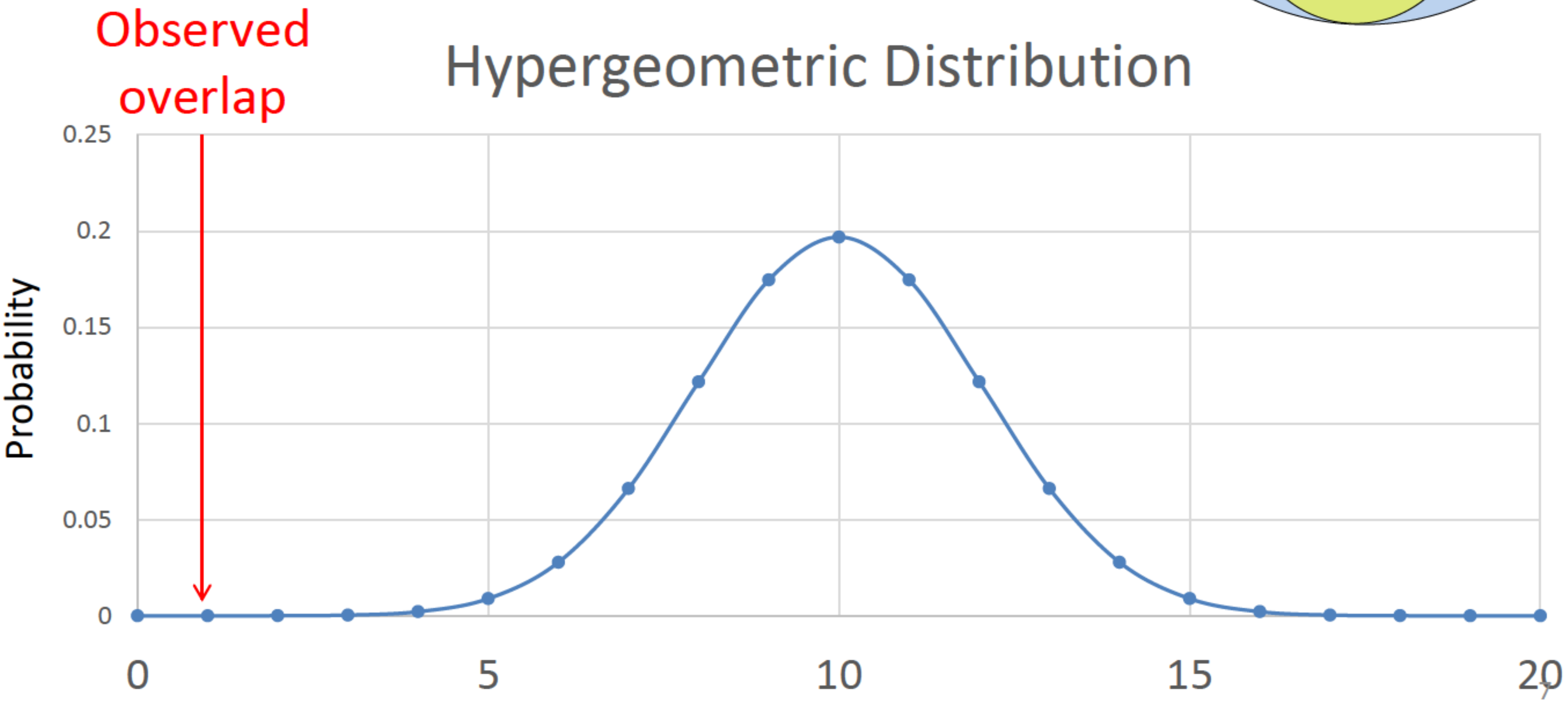
Yes! You would expect to see a ***larger*** overlap under the null model.

Are the yellow genes ***enriched*** for the red function?

No! Quite the opposite!

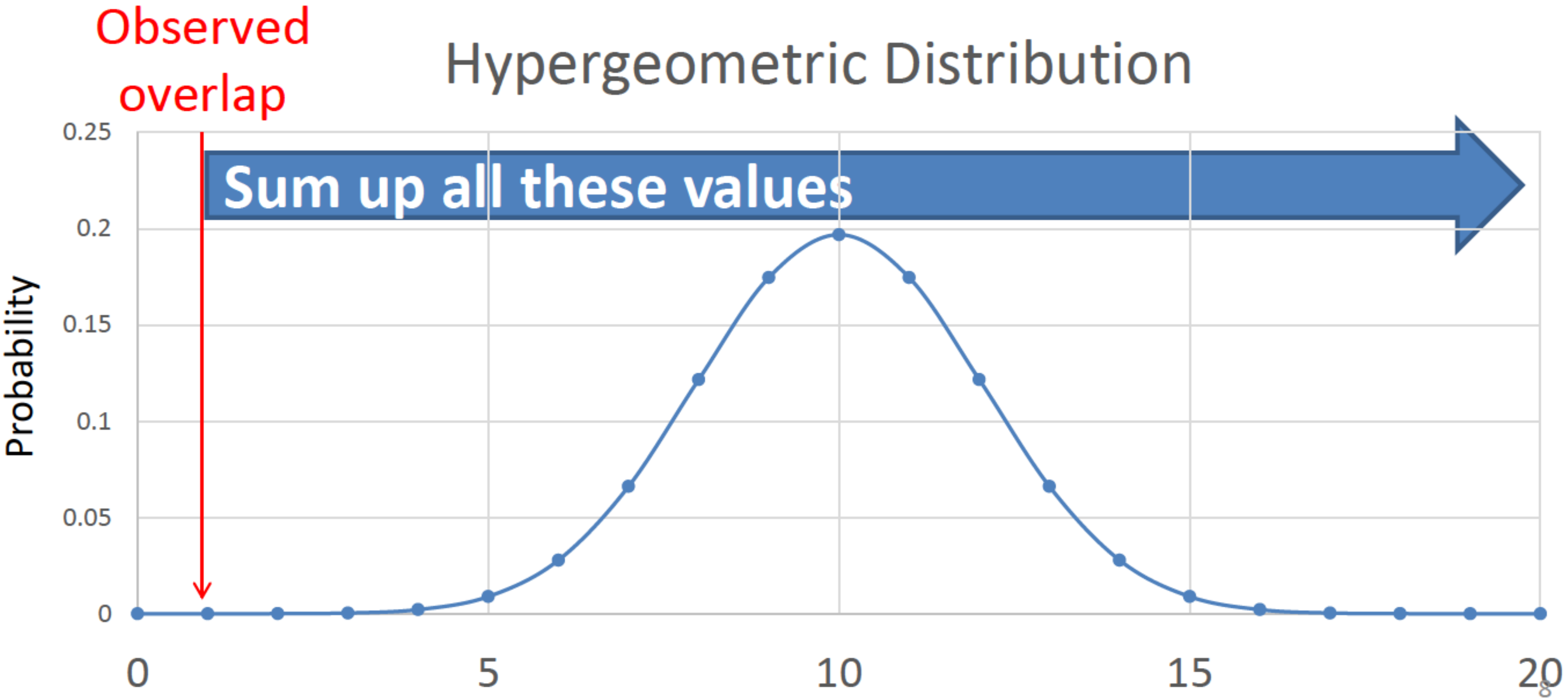
The Hypergeometric p-value is the probability of observing an exact overlap





The CDF helps us find enriched terms

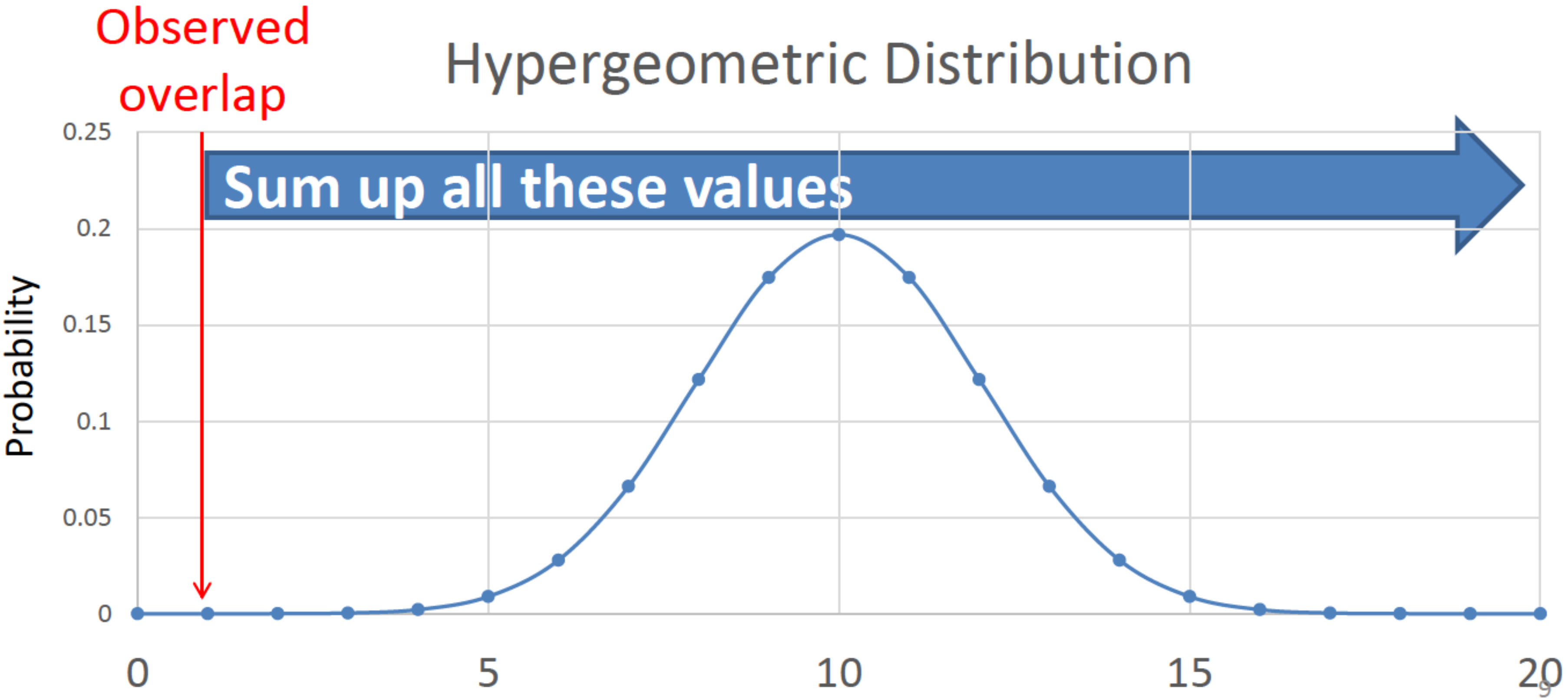
We want to compute the probability of observing ***at least*** this overlap under our null model.



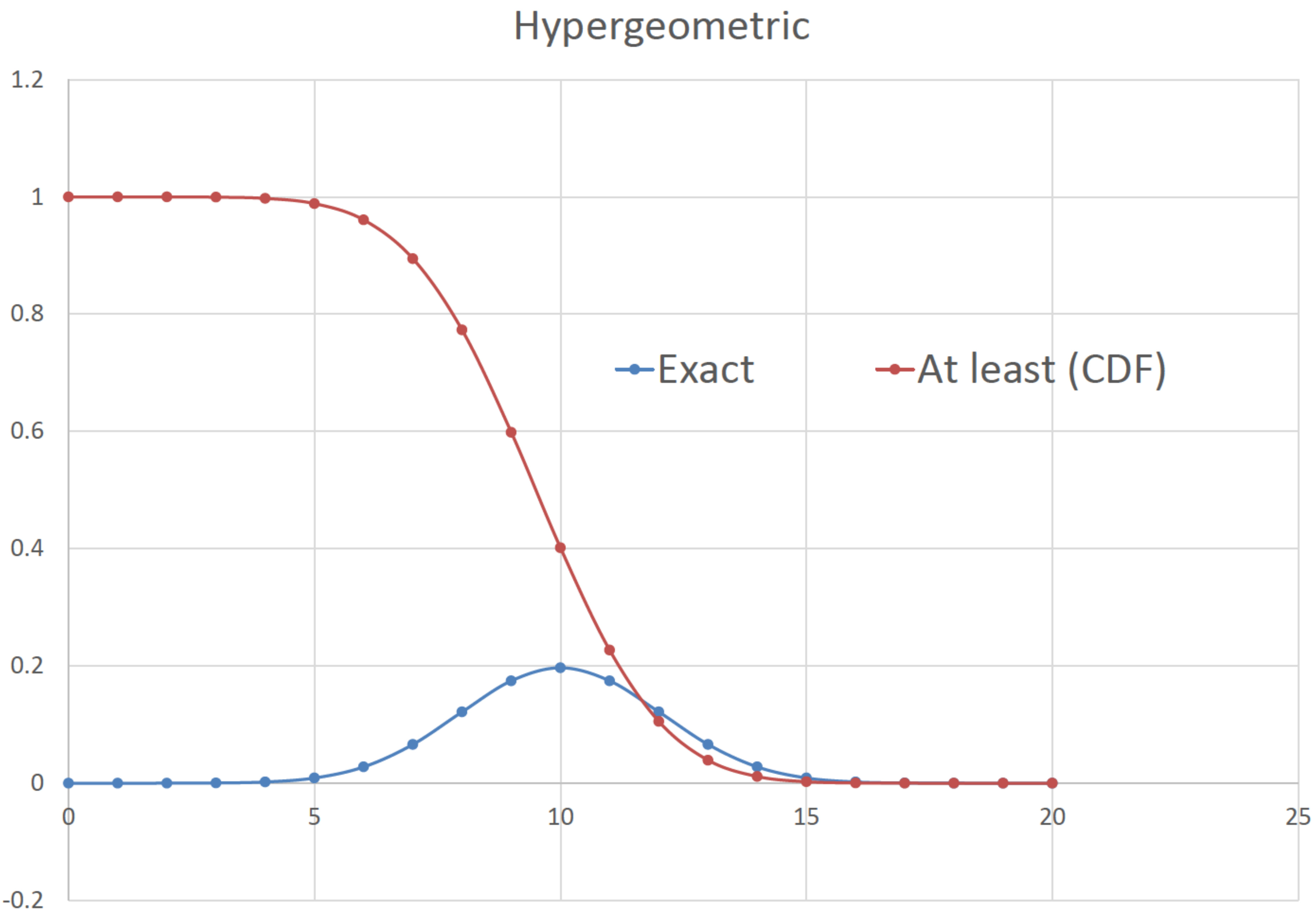
CDF=Cumulative distribution function

The CDF helps us find enriched terms

$$CDF(Overlap) = \sum_{n=overlap}^{\substack{Number\ of \\ genes\ in\ DNA\ Repair}} \frac{\binom{DNA\ repair}{n}\binom{Genome - DNA\ repair}{DiffExp - n}}{\binom{Genome}{DiffExp}}$$

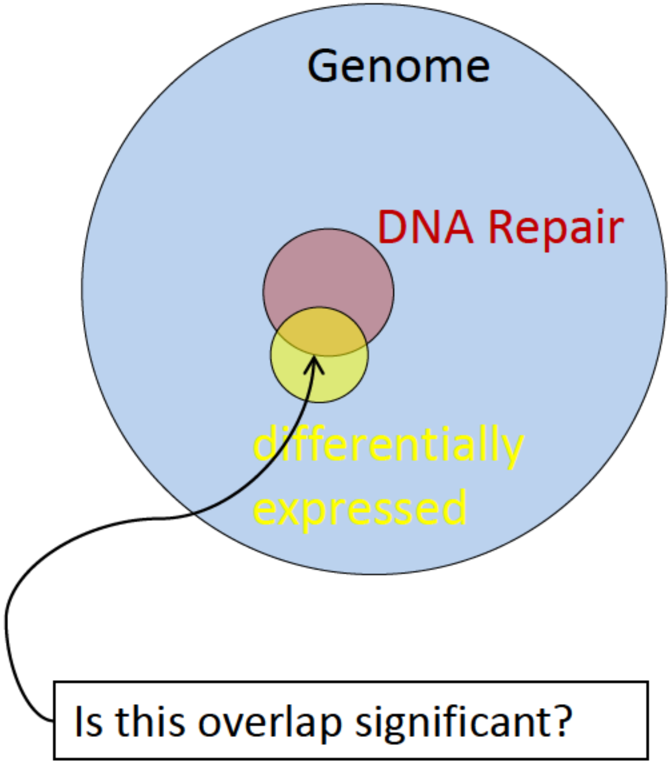


CDF=Cumulative distribution function


Hypergeometric
1.2
1
0.8
0.6
0.4
0.2
0
-0.2
0
5
10
15
20
25
Exact
At least (CDF)

Statistical significance

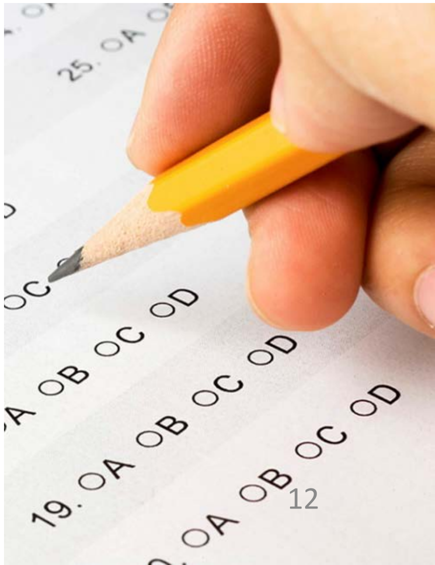
- We wish to test if a term is "enriched" in our data.
- But the hypergeometric gives the probability of getting **exactly** this amount of overlap for two randomly chosen sets of genes of the same size.
- Using the CDF, we can ask if we see ***more*** of a term than we would expect under the null model.



Is this overlap significant?

Testing Multiple Hypotheses

- Example: Filter GO terms using a p-value threshold of 0.01
- By definition, the null-hypothesis has a 1% probability of being correct ***for each test.***
- There are roughly 30,000 terms in GO.
- At this level, we expect roughly 300 false positives!



Multiple Hypotheses

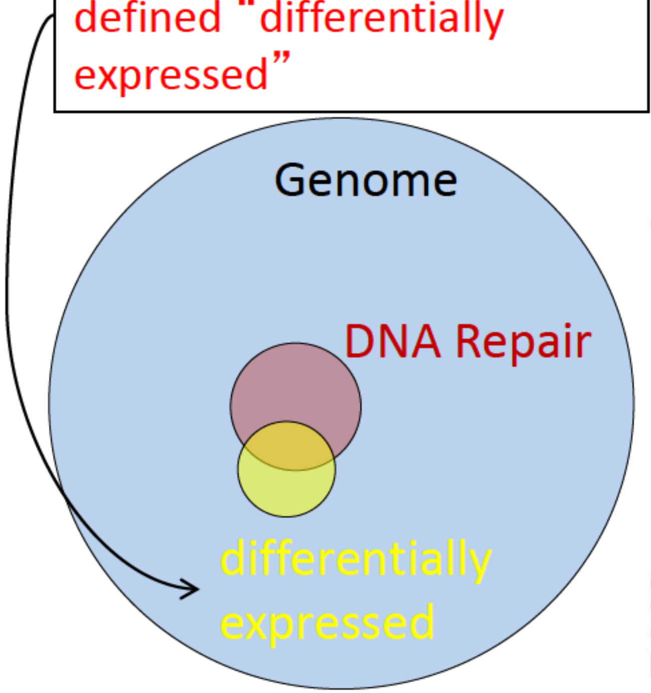
- A simple solution: require that the p-value be small enough to reduce the false positives to the desired level.
- This is called the Bonferroni correction.
- In our case, we would only accept terms with a

$$p \leq \frac{0.01}{30{,}000} = \frac{desired\ threshold}{number\ of\ tests}$$

- Since our tests are not all independent, this is very conservative, and will miss many true positives
- More sophisticated approaches exist, such as controlling the “false discovery rate”.

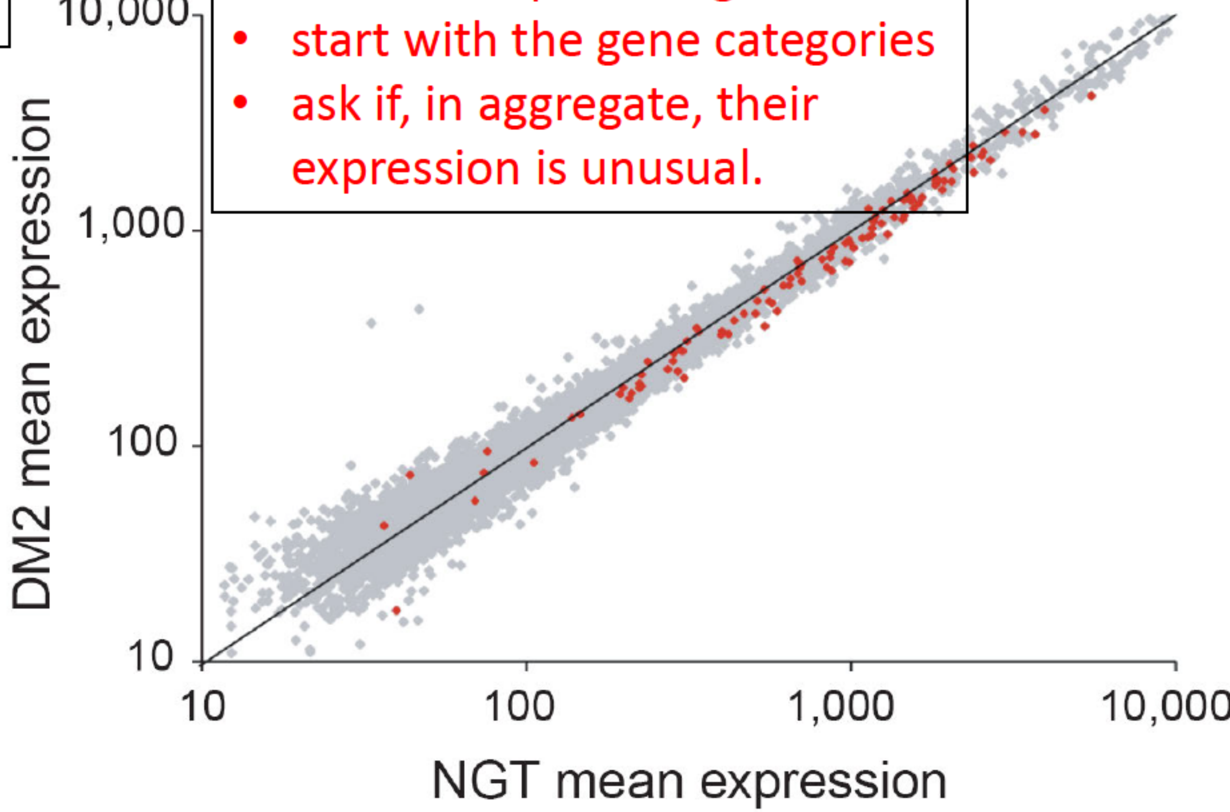
Aggregate score statistics

My results depend on how I defined "differentially expressed"



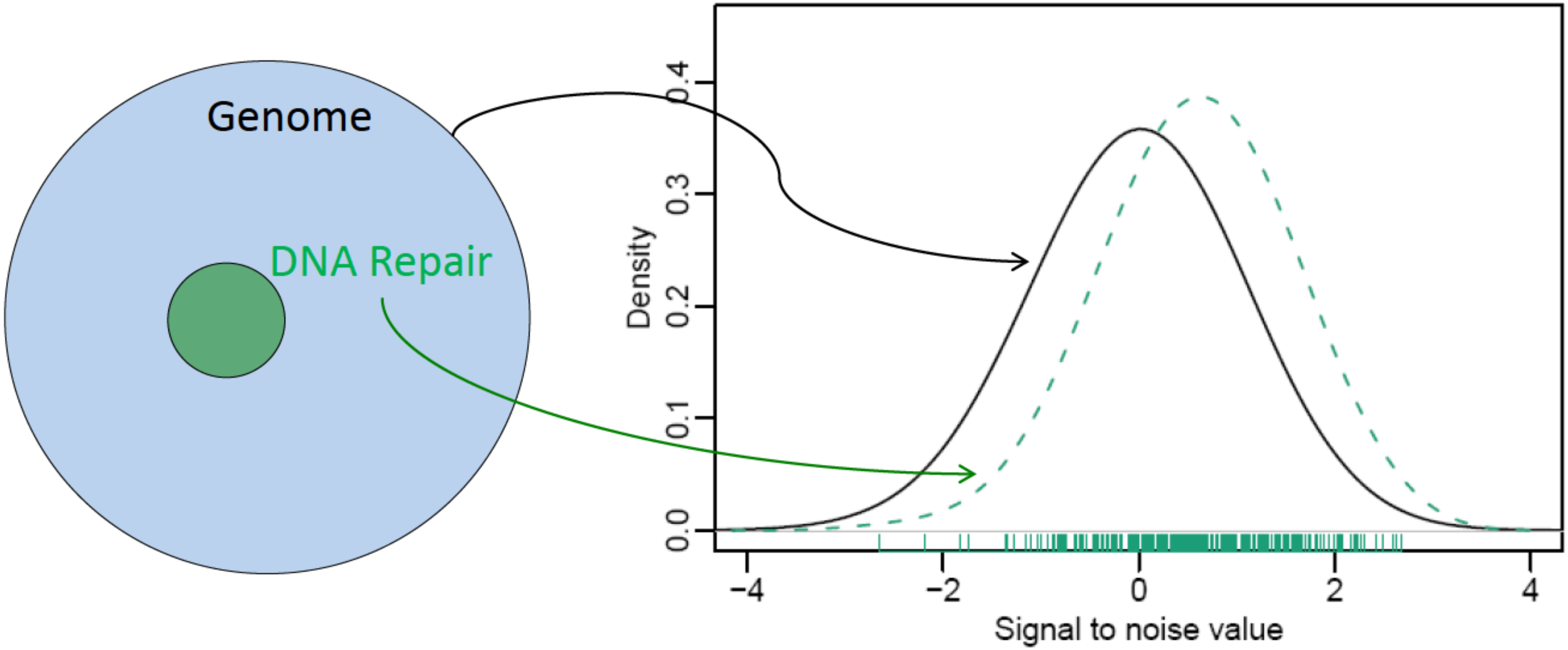
Instead of starting with differential expressed genes:

- start with the gene categories
- ask if, in aggregate, their expression is unusual.



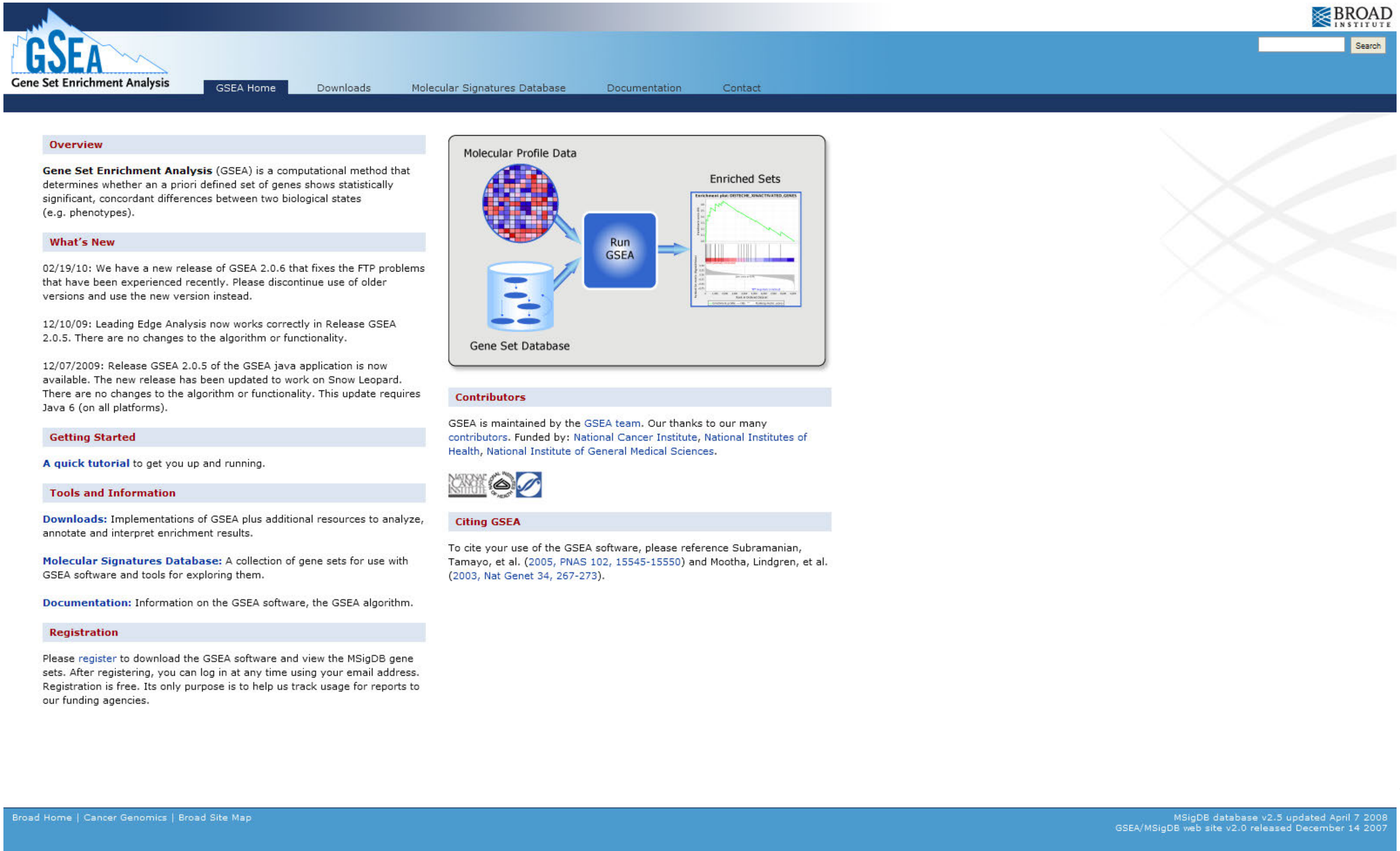
Mootha *et al.* (2003). *Nature Genetics* **34**, 267 – 273. doi:10.1038/ng1180

Aggregate score statistics



Aggregate score statistics

http://www.broadinstitute.org/gsea/



Overview

Gene Set Enrichment Analysis (GSEA) is a computational method that determines whether an a priori defined set of genes shows statistically significant, concordant differences between two biological states (e.g. phenotypes).

What's New

02/19/10: We have a new release of GSEA 2.0.6 that fixes the FTP problems that have been experienced recently. Please discontinue use of older versions and use the new version instead.

12/10/09: Leading Edge Analysis now works correctly in Release GSEA 2.0.5. There are no changes to the algorithm or functionality.

12/07/2009: Release GSEA 2.0.5 of the GSEA java application is now available. The new release has been updated to work on Snow Leopard. There are no changes to the algorithm or functionality. This update requires Java 6 (on all platforms).

Getting Started

A quick tutorial to get you up and running.

Tools and Information

Downloads: Implementations of GSEA plus additional resources to analyze, annotate and interpret enrichment results.

Molecular Signatures Database: A collection of gene sets for use with GSEA software and tools for exploring them.

Documentation: Information on the GSEA software, the GSEA algorithm.

Registration

Please register to download the GSEA software and view the MSigDB gene sets. After registering, you can log in at any time using your email address. Registration is free. Its only purpose is to help us track usage for reports to our funding agencies.

Contributors

GSEA is maintained by the GSEA team. Our thanks to our many contributors. Funded by: National Cancer Institute, National Institutes of Health, National Institute of General Medical Sciences.

Citing GSEA

To cite your use of the GSEA software, please reference Subramanian, Tamayo, et al. (2005, PNAS 102, 15545-15550) and Mootha, Lindgren, et al. (2003, Nat Genet 34, 267-273).

Broad Home | Cancer Genomics | Broad Site Map

MSigDB database v2.5 updated April 7 2008
GSEA/MSigDB web site v2.0 released December 14 2007

Learning Objectives

- To understand types and sources of biological "big data" and how they are used

Big Data Creates an Opportunity

Transcription

GEO

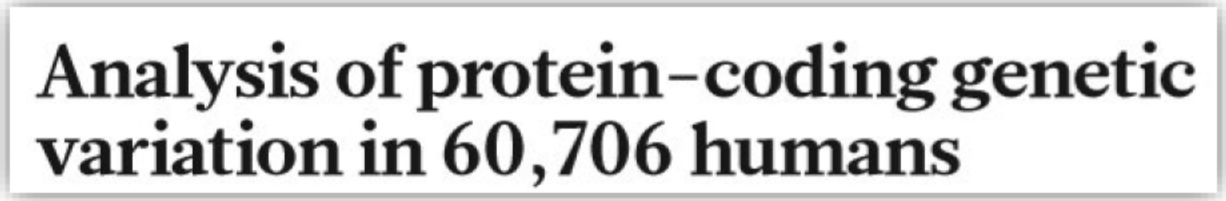
Gene Expression Omnibus

>2.3 million samples so far

PubChem

>94 million compounds

Genomics



Analysis of protein-coding genetic variation in 60,706 humans

The Exome Aggregation Consortium

ExAC

NIH THE CANCER GENOME ATLAS
National Cancer Institute
National Human Genome Research Institute

2.5 Petabytes

33 types of tumors

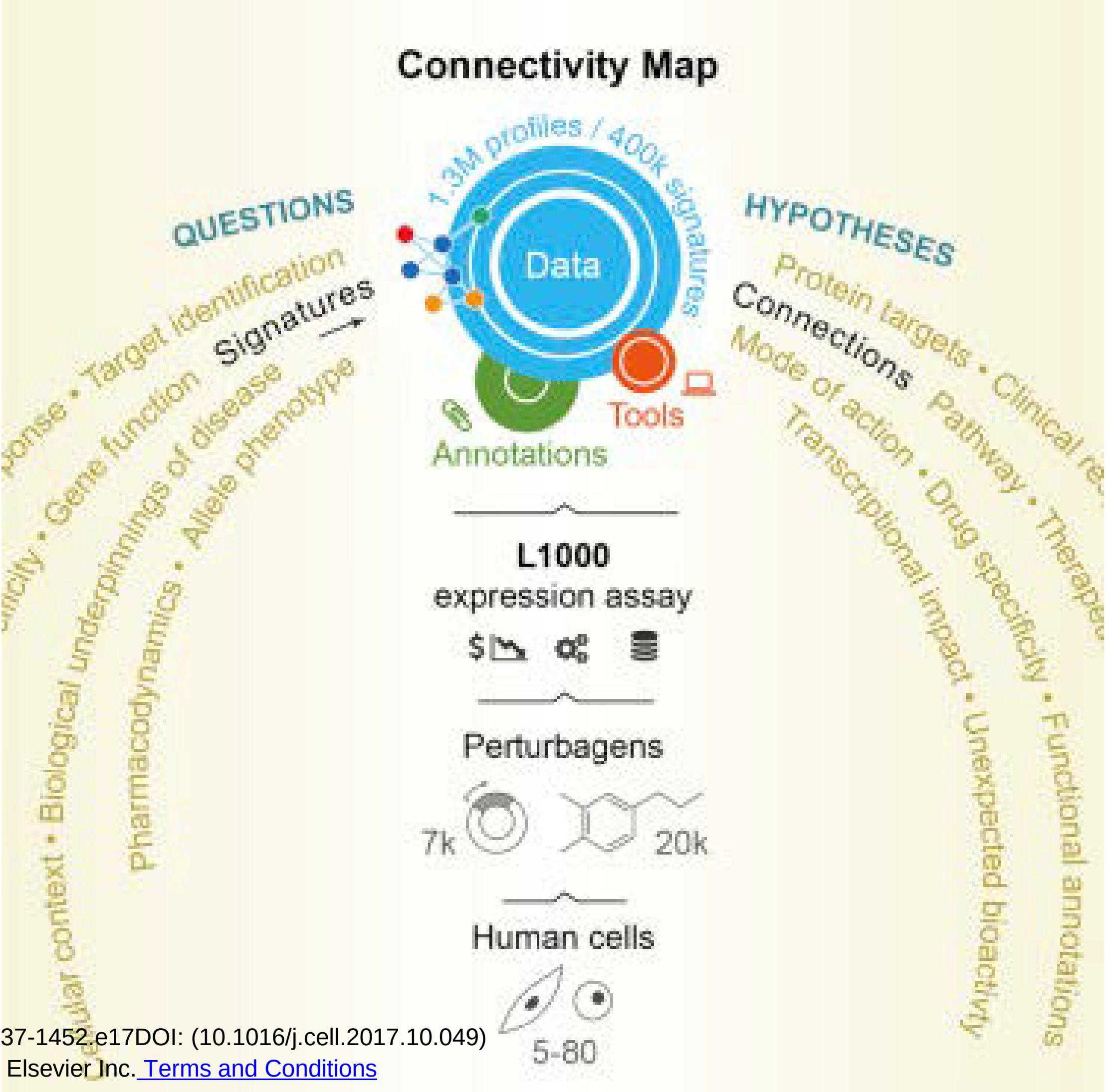
11,000 patients

7 data types

Example 1

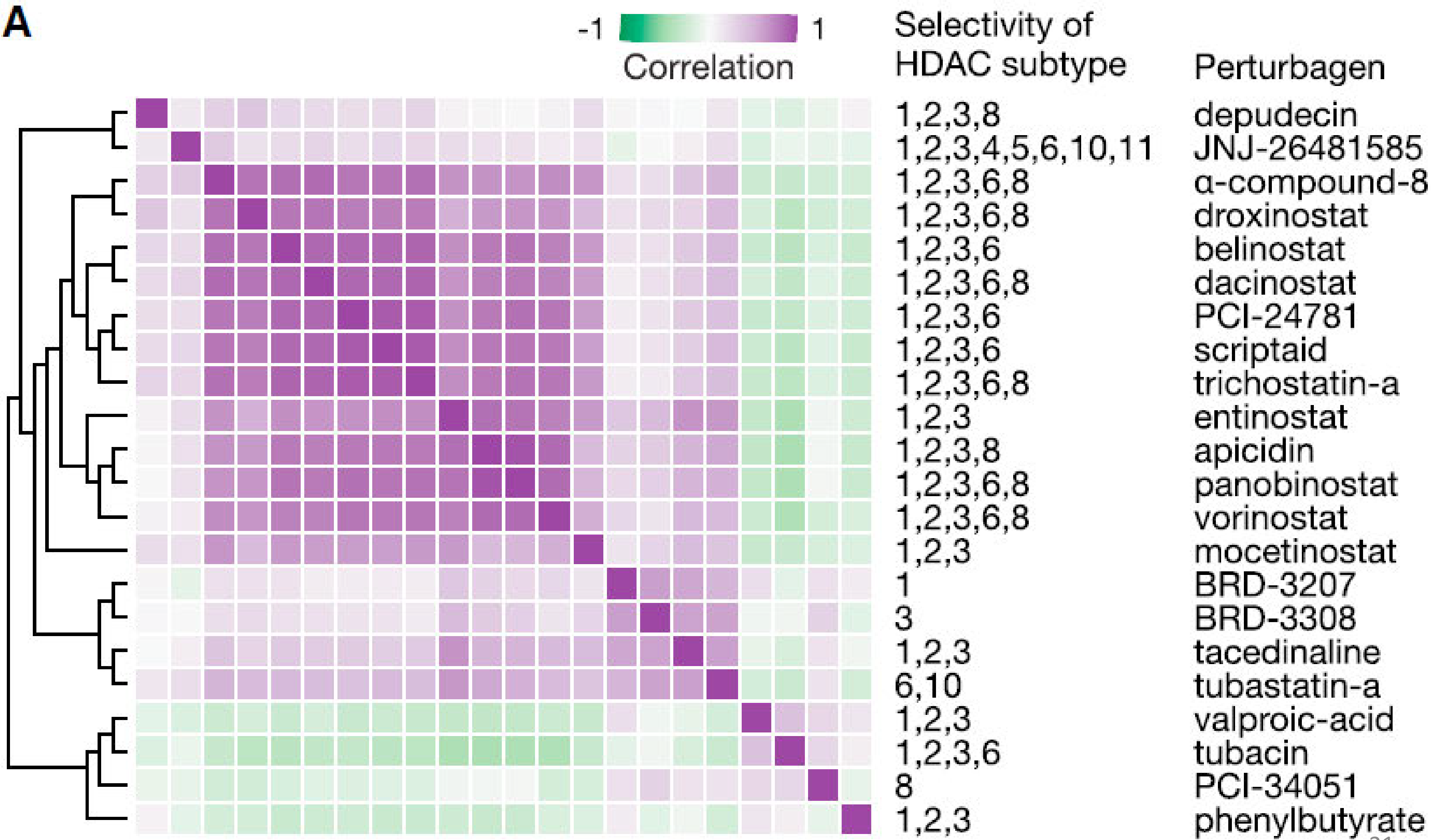
L1000: A VERY LARGE TRANSCRIPTIONAL DATASET

A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles


Connectivity Map
1.3M profiles / 400k signatures
Data
Annotations
Tools
QUESTIONS
ponse • Target identification
icity • Gene function
Signatures
llular context • Biological underpinnings of disease
Pharmacodynamics • Allele phenotype
HYPOTHESES
Protein targets • Clinical re
Connections
Pathway • Therape
Mode of action • Drug specificity • Functional annotations
Transcriptional impact • Unexpected bioactivity
L1000
expression assay
$
Perturbagens
7k
20k
Human cells
5-80

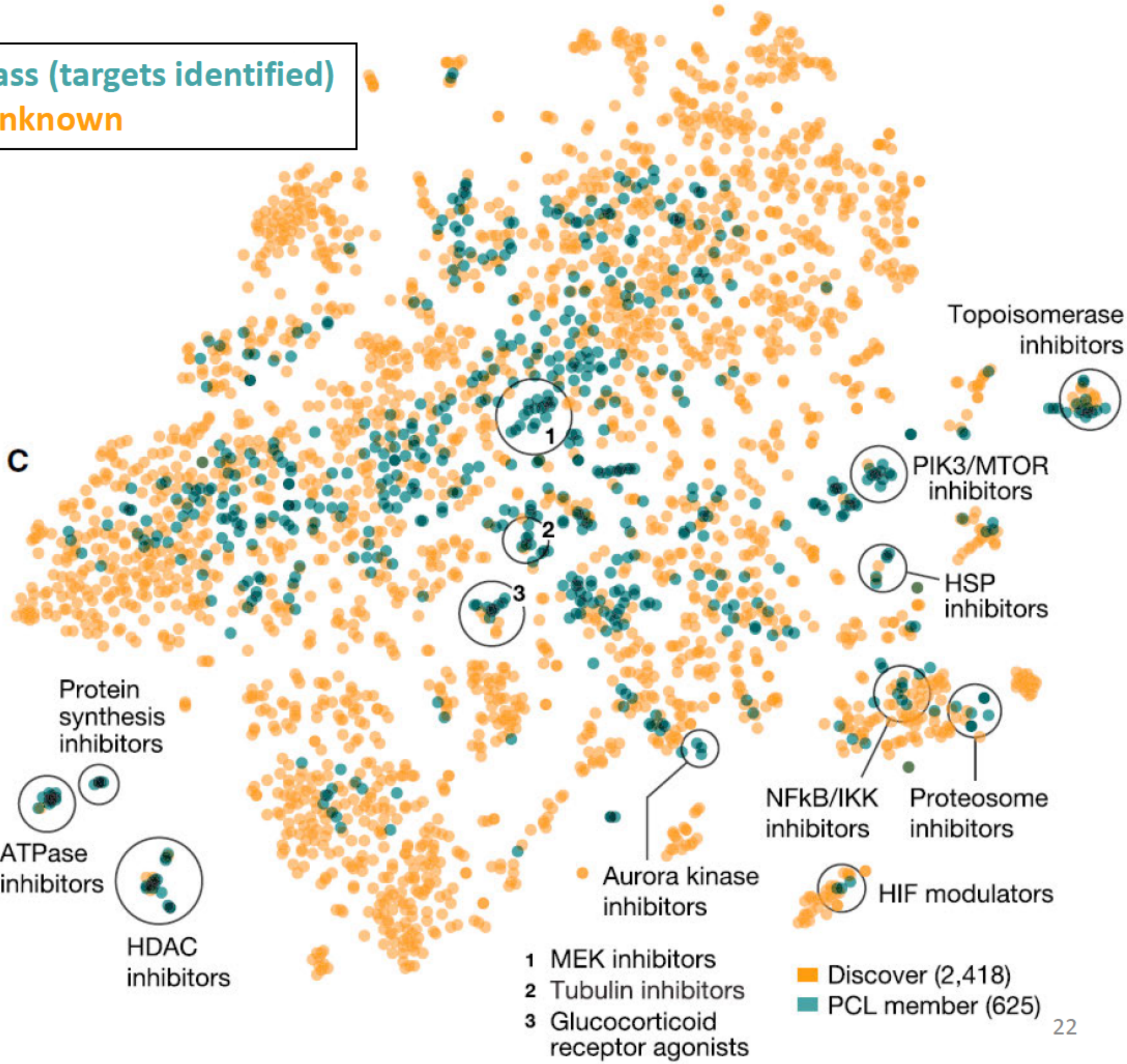
Cell 2017 171, 1437-1452.e17DOI: (10.1016/j.cell.2017.10.049)

Clustering Transcriptional Results



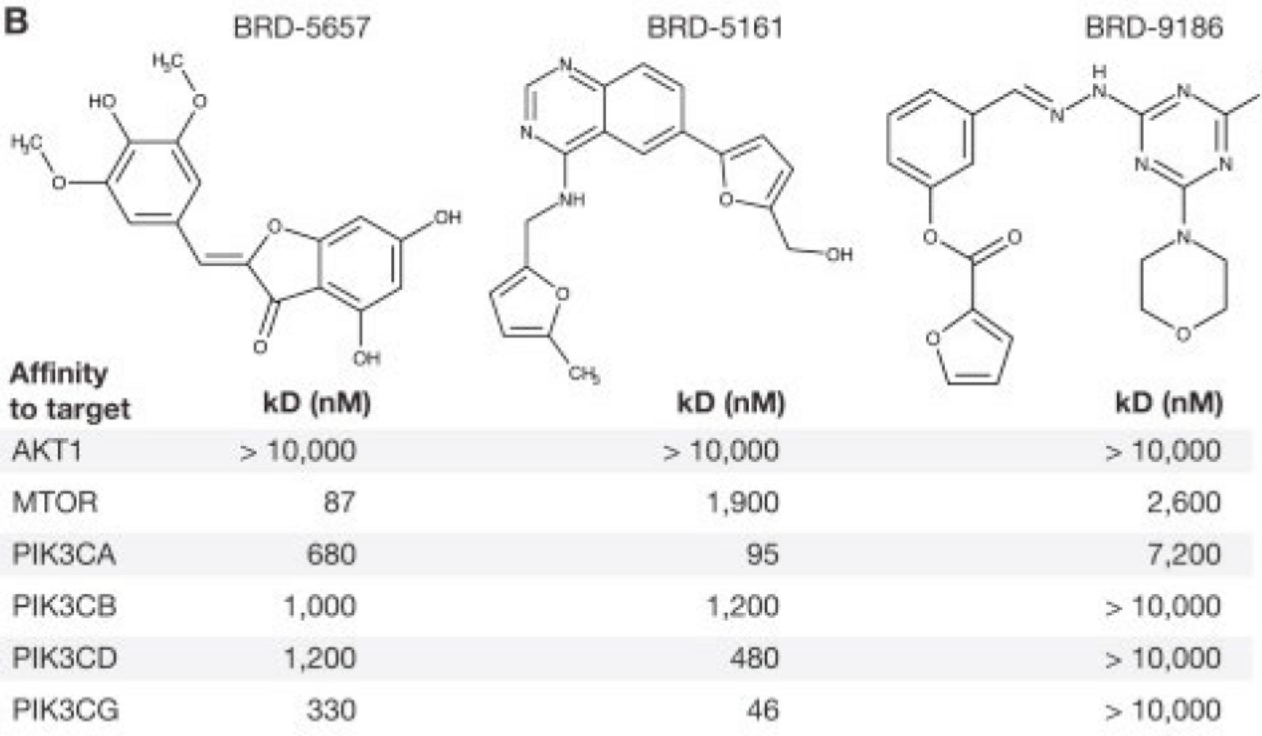
PCL=perturbagen class (targets identified)

Discovery = target unknown



PCL=perturbagen class (targets identified)

Discovery = target unknown



Affinity to target	BRD-5657 kD (nM)	BRD-5161 kD (nM)	BRD-9186 kD (nM)
AKT1	> 10,000	> 10,000	> 10,000
MTOR	87	1,900	2,600
PIK3CA	680	95	7,200
PIK3CB	1,000	1,200	> 10,000
PIK3CD	1,200	480	> 10,000
PIK3CG	330	46	> 10,000

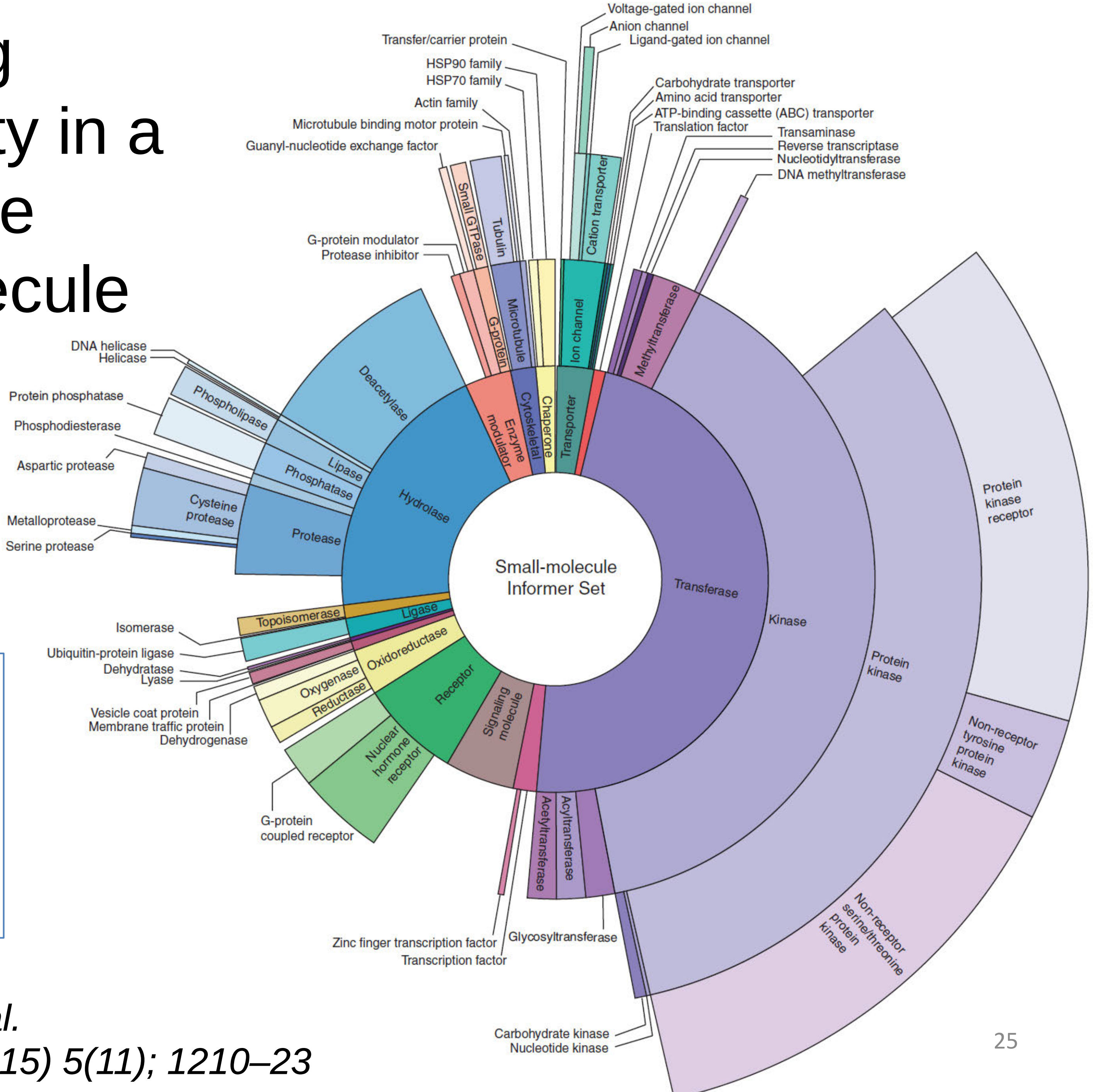
PIK3/MTOR inhibitors

Example 2

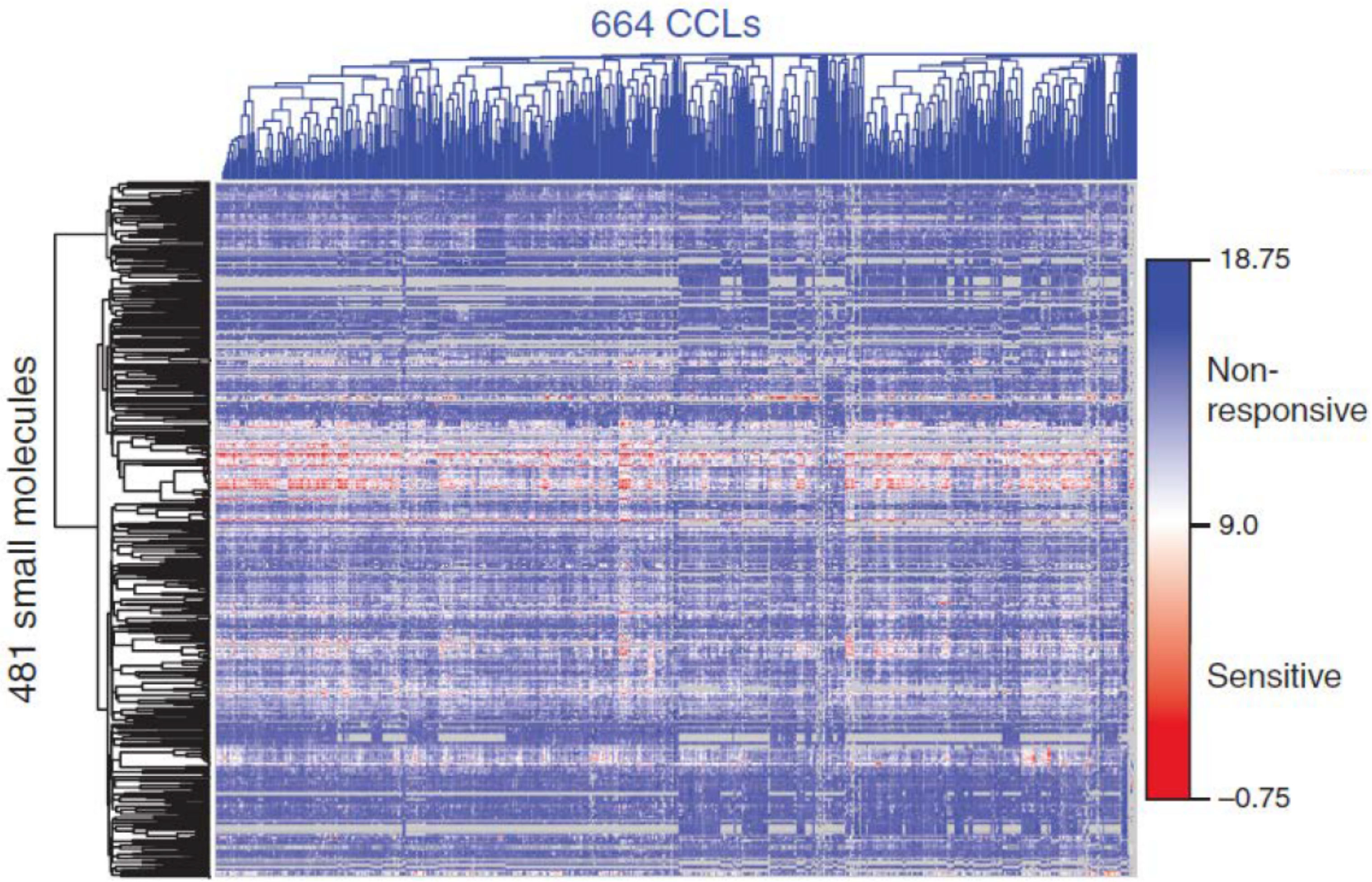
A VERY LARGE SENSITIVITY ASSAY

Harnessing Connectivity in a Large-Scale Small-Molecule Sensitivity Dataset.

481 compounds, including FDA-approved drugs, clinical candidates, and small-molecule probes



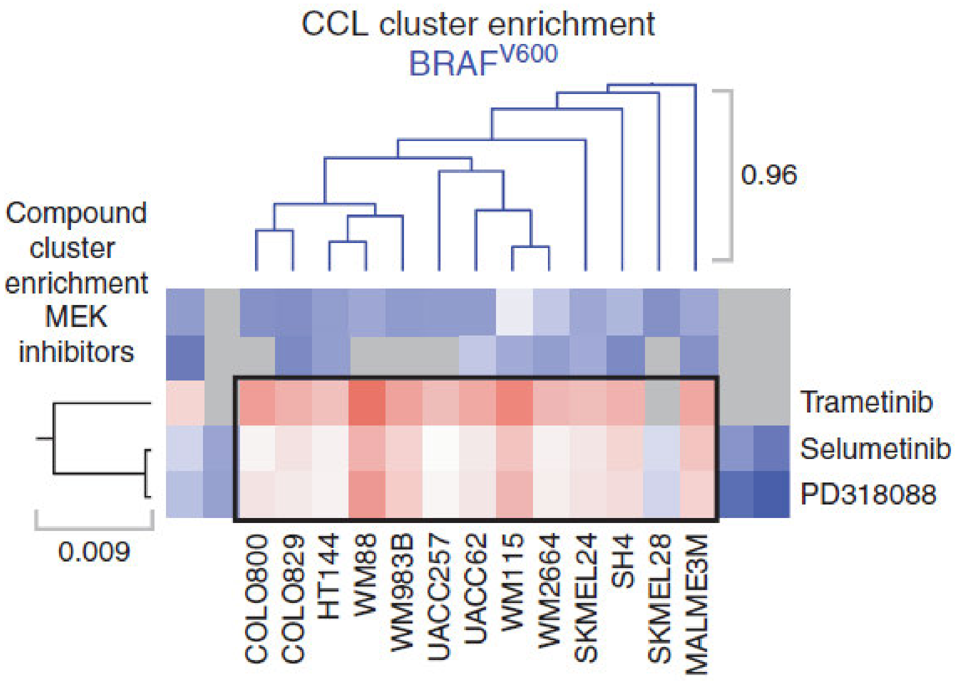
Seashore-Ludlow *et al.*
Cancer Discovery (2015) 5(11); 1210–23

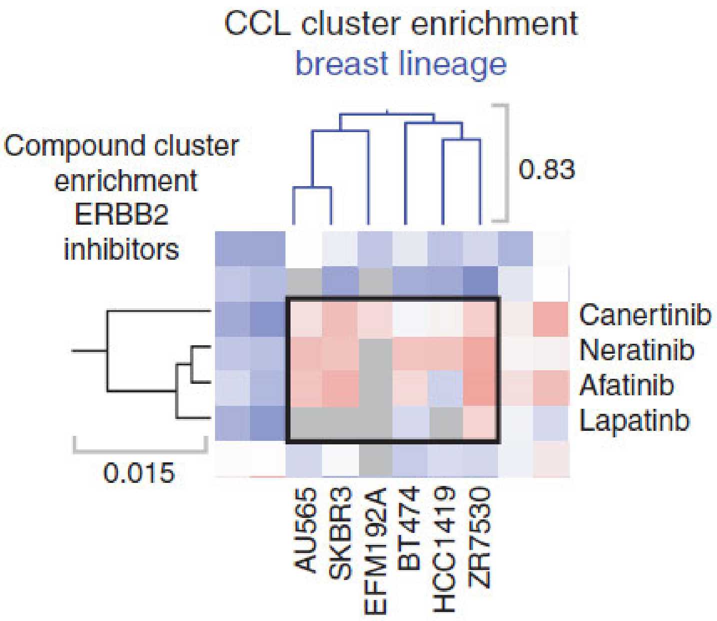


The 481 compounds were tested at 16 concentrations in duplicate against 664 cancer cell lines.

DOI: 10.1158/2159-8290.CD-15-0235 NOVEMBER 2015 CANCER DISCOVERY

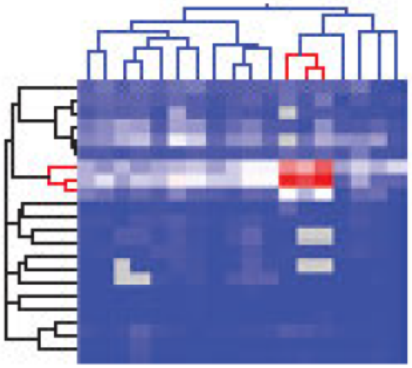
Cluster often represent common sensitivity to a mechanism

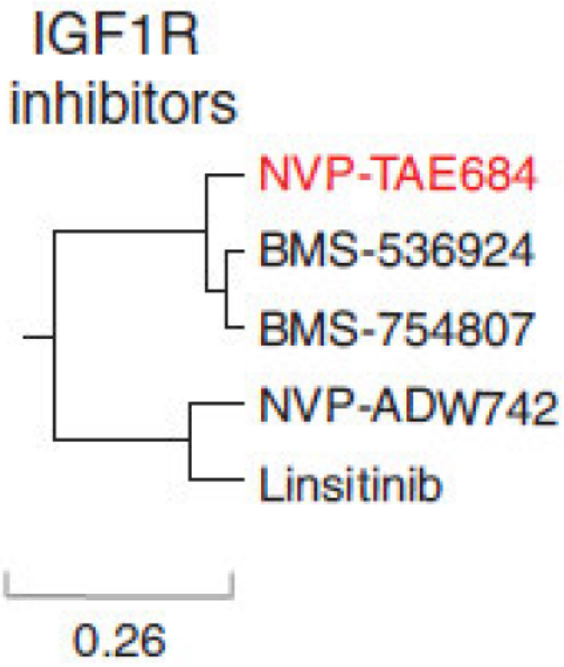


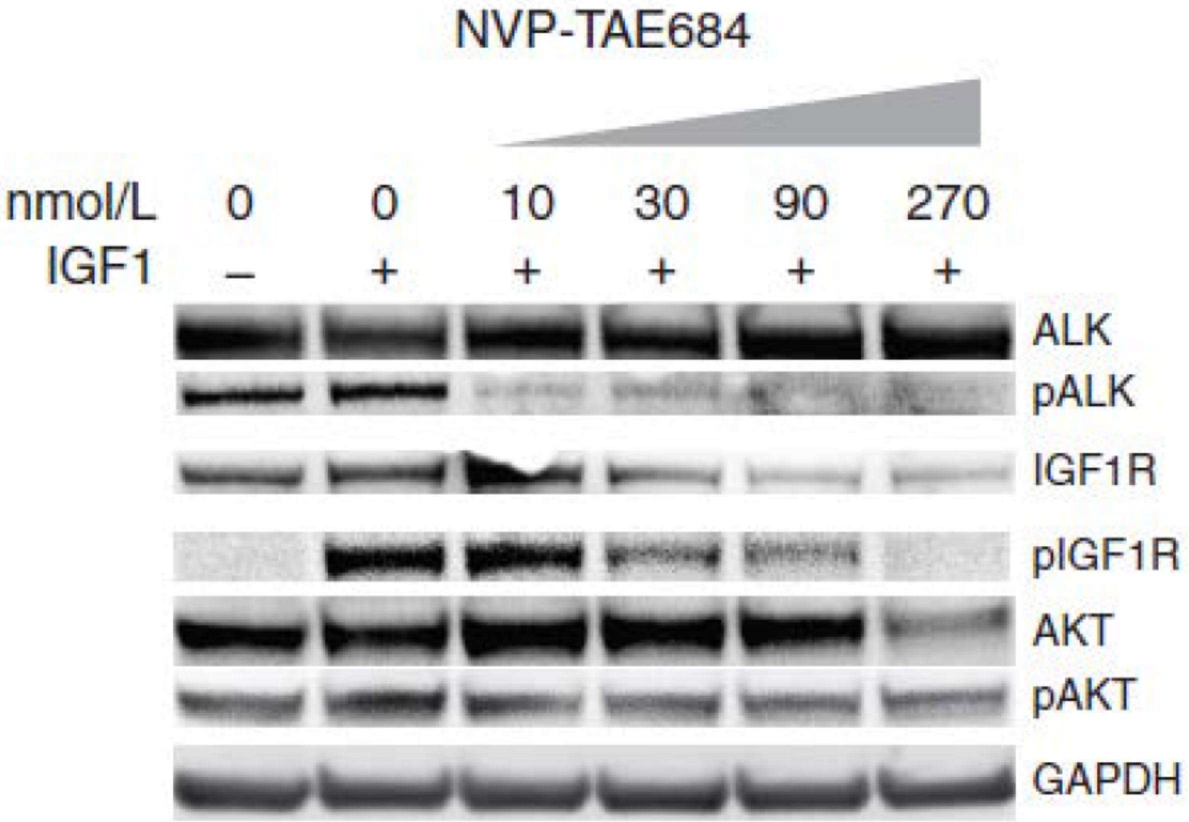


DOI: 10.1158/2159-8290.CD-15-0235

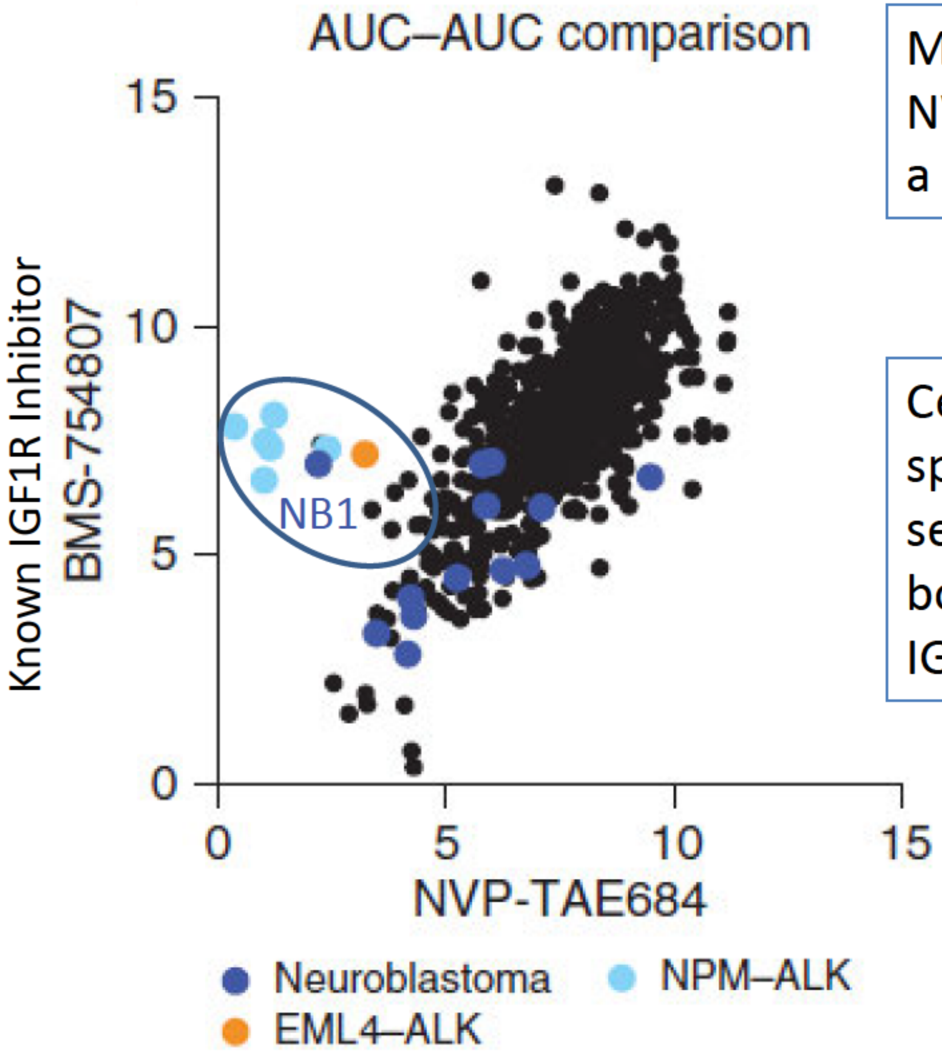
Discovery of new way to target neuroblastoma?







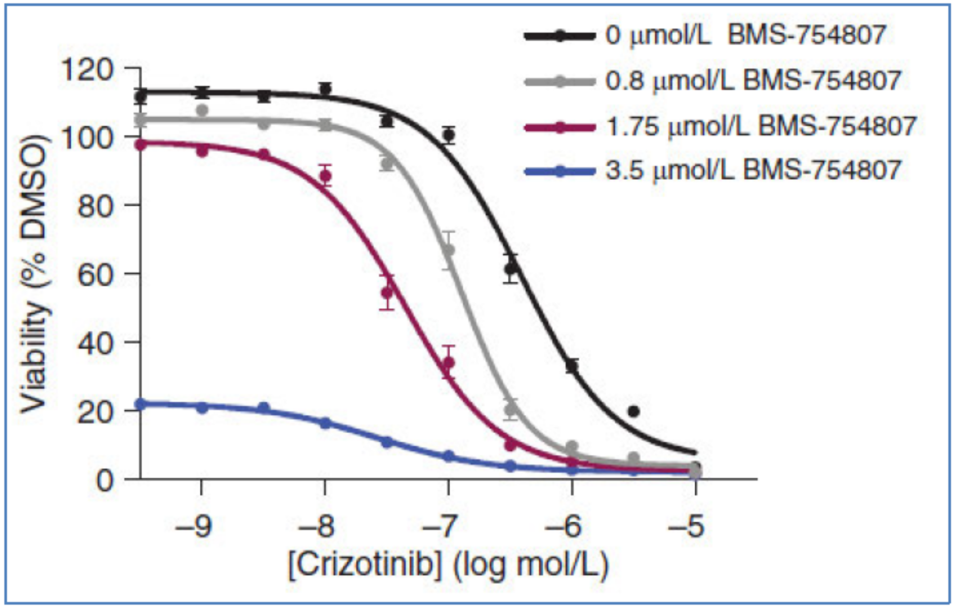
Discovery of new way to target neuroblastoma?



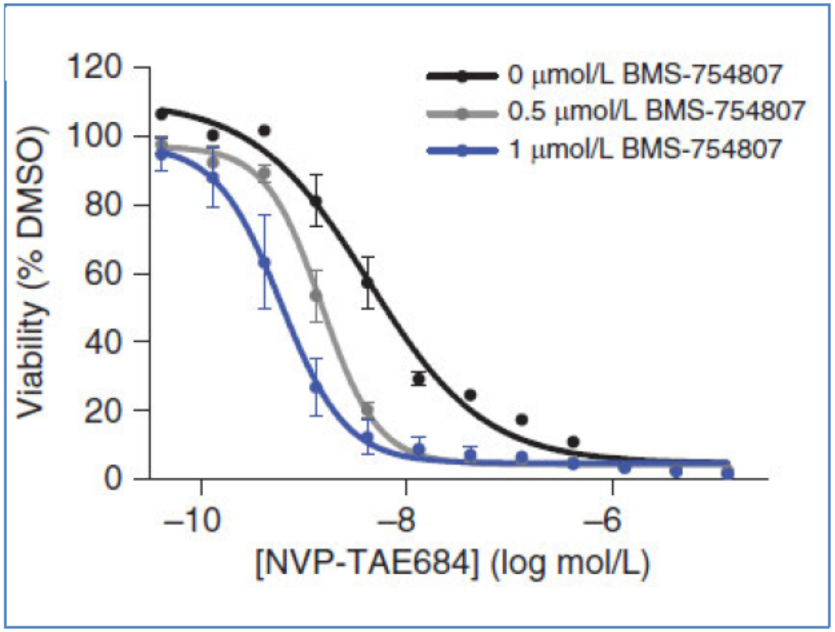
Most cells show a similar response to NVP-TAE684 and BMS-754807, suggesting a similar mechanism

Cells with high levels of ALK (NB1) or specific ALK genomic changes are more sensitive to NVP-TAE684, which targets both IGF1R and ALK than they are to the IGF1R inhibitor BMS-754807

NB1 responds to a combination of ALK and IGF1R inhibitors

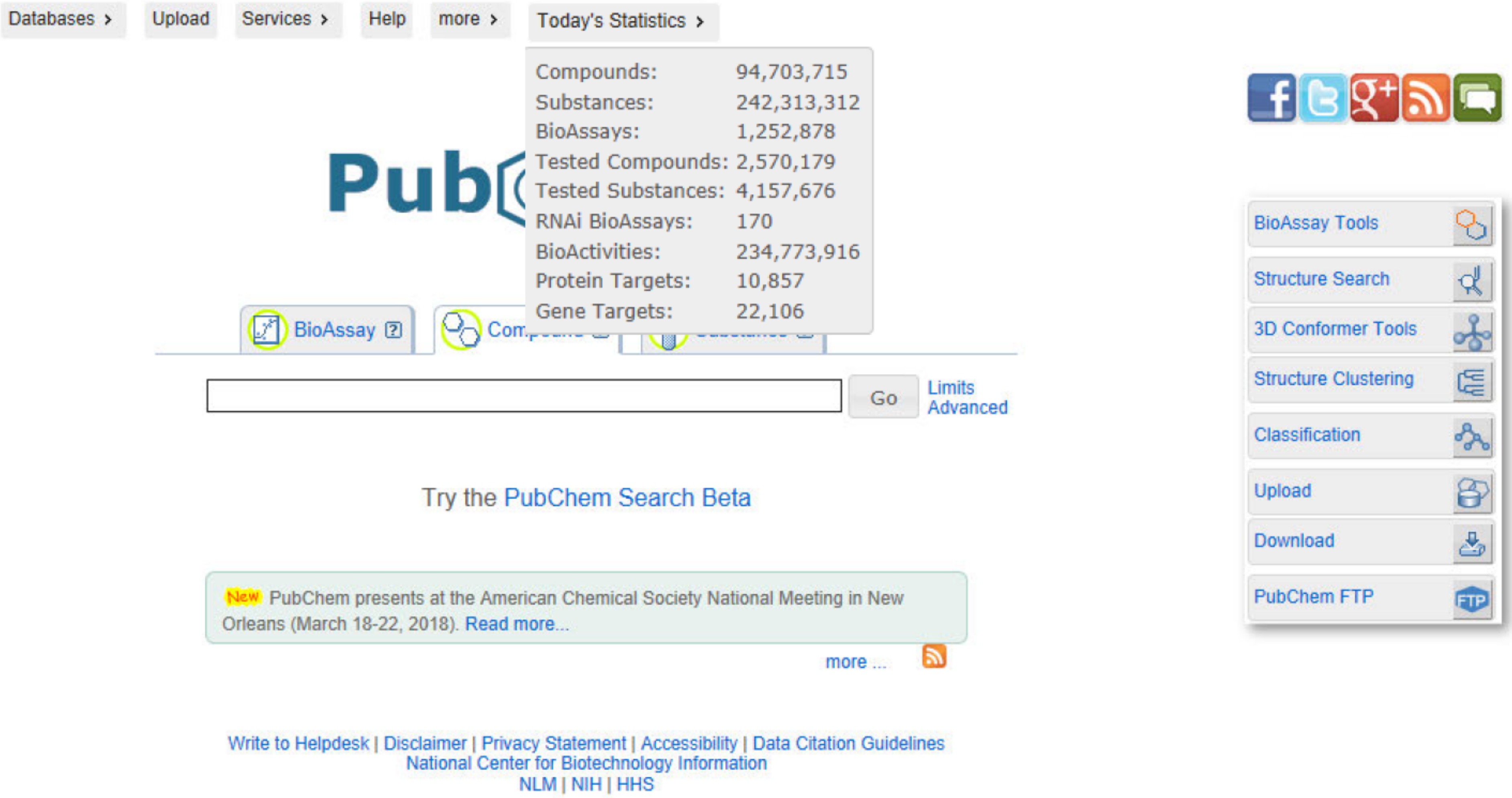


Known ALK Inhibitor



Example 3

PUBCHEM: A DATABASE OF CHEMICAL COMPOUNDS

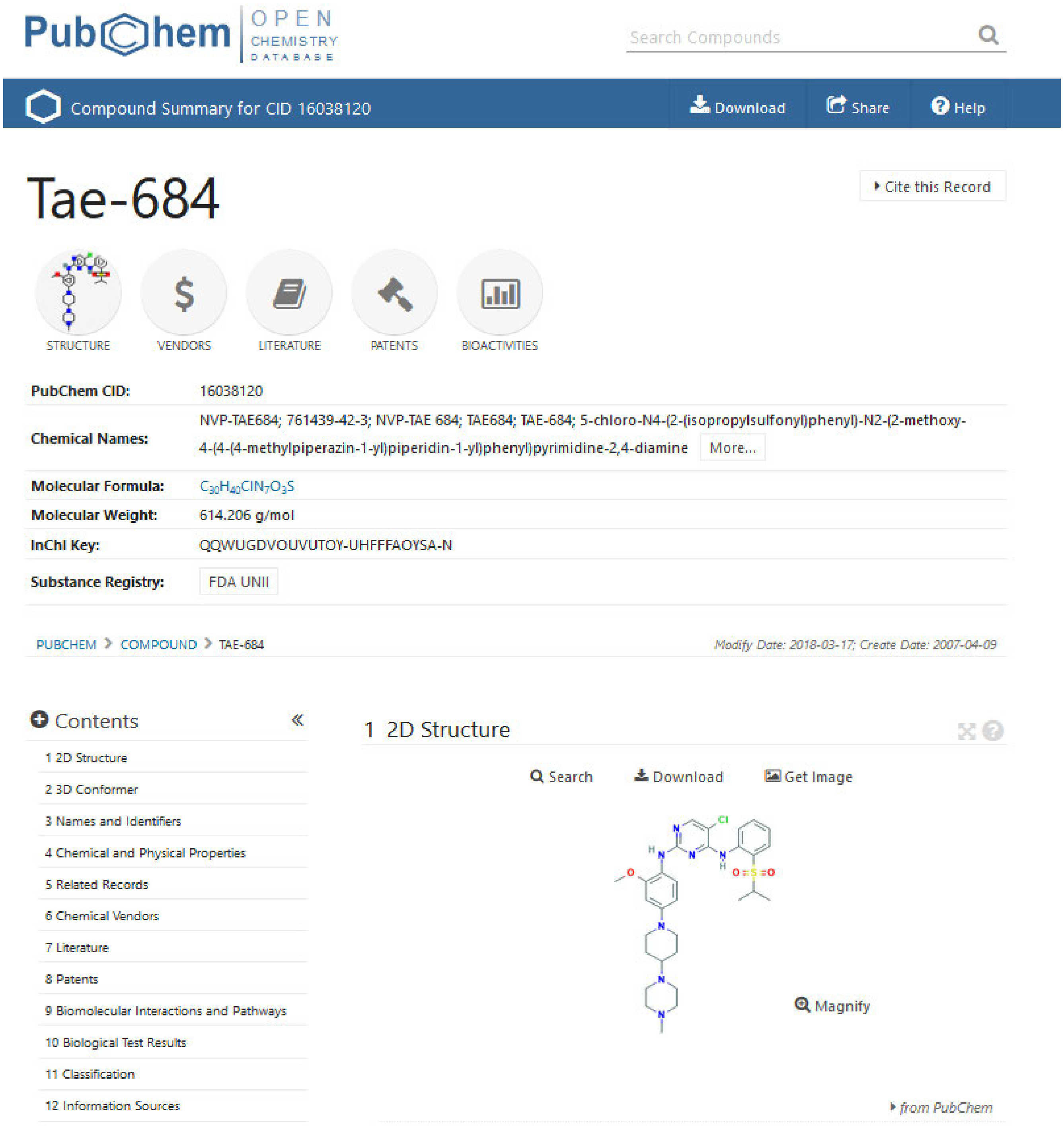


Compounds:	94,703,715
Substances:	242,313,312
BioAssays:	1,252,878
Tested Compounds:	2,570,179
Tested Substances:	4,157,676
RNAi BioAssays:	170
BioActivities:	234,773,916
Protein Targets:	10,857
Gene Targets:	22,106

Try the PubChem Search Beta

New PubChem presents at the American Chemical Society National Meeting in New Orleans (March 18-22, 2018). Read more...

more ...
BioAssay Tools
Structure Search
3D Conformer Tools
Structure Clustering
Classification
Upload
Download
PubChem FTP
Write to Helpdesk | Disclaimer | Privacy Statement | Accessibility | Data Citation Guidelines
National Center for Biotechnology Information
NLM | NIH | HHS



PubChem | OPEN CHEMISTRY DATABASE

Search Compounds

Compound Summary for CID 16038120

Download Share Help

Tae-684

Cite this Record

STRUCTURE VENDORS LITERATURE PATENTS BIOACTIVITIES

PubChem CID:	16038120
Chemical Names:	NVP-TAE684; 761439-42-3; NVP-TAE 684; TAE684; TAE-684; 5-chloro-N4-(2-(isopropylsulfonyl)phenyl)-N2-(2-methoxy-4-(4-(4-methylpiperazin-1-yl)piperidin-1-yl)phenyl)pyrimidine-2,4-diamine More...
Molecular Formula:	$C_{30}H_{40}ClN_7O_3S$
Molecular Weight:	614.206 g/mol
InChI Key:	QQWUGDVOUVUTOY-UHFFFAOYSA-N
Substance Registry:	FDA UNII

PUBCHEM > COMPOUND > TAE-684

Modify Date: 2018-03-17; Create Date: 2007-04-09

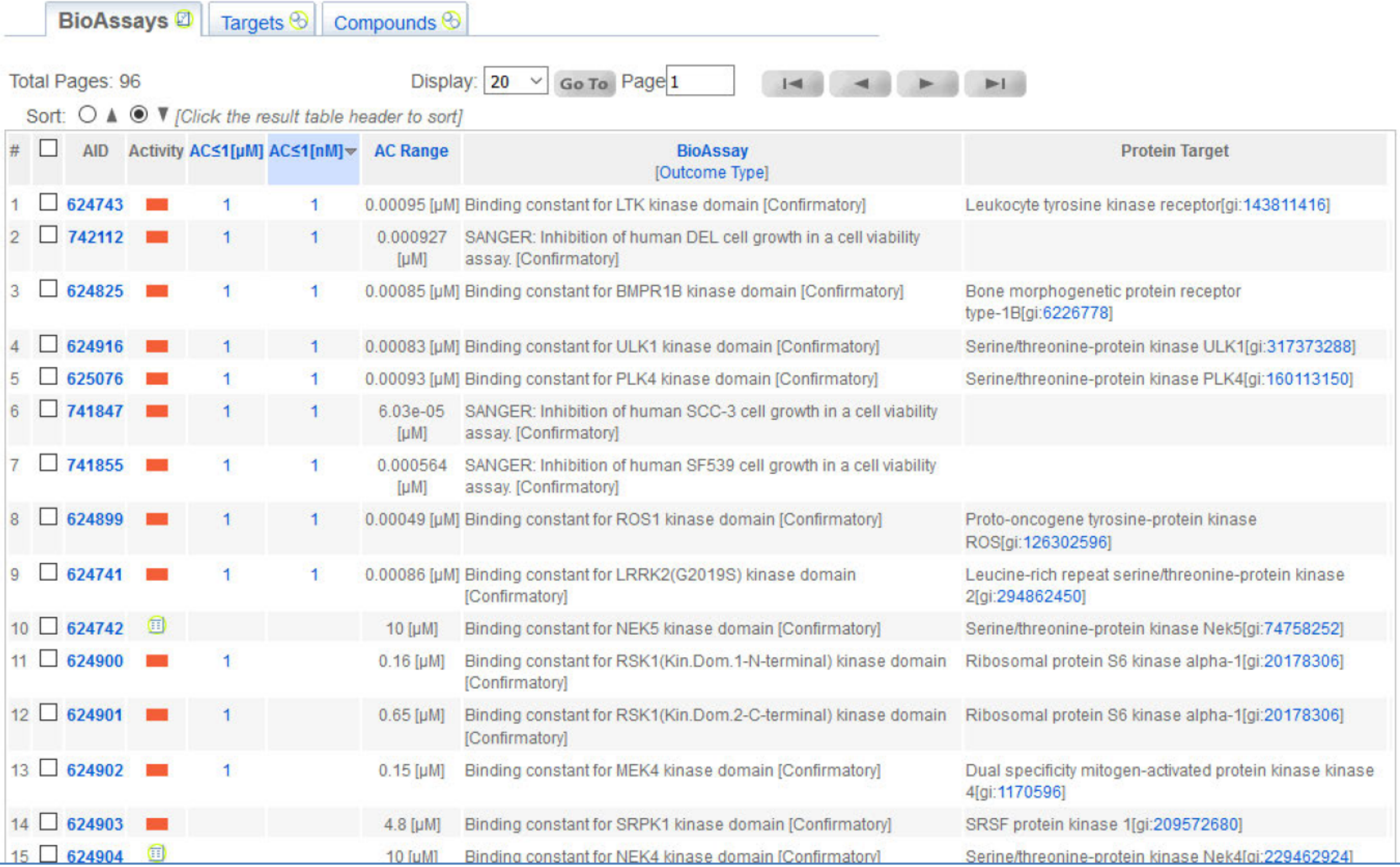
Contents «

1 2D Structure

Search Download Get Image

Magnify

from PubChem



Total Pages: 96 Display: 20 Go To Page 1

Sort: ○ ▲ ● ▼ *[Click the result table header to sort]*

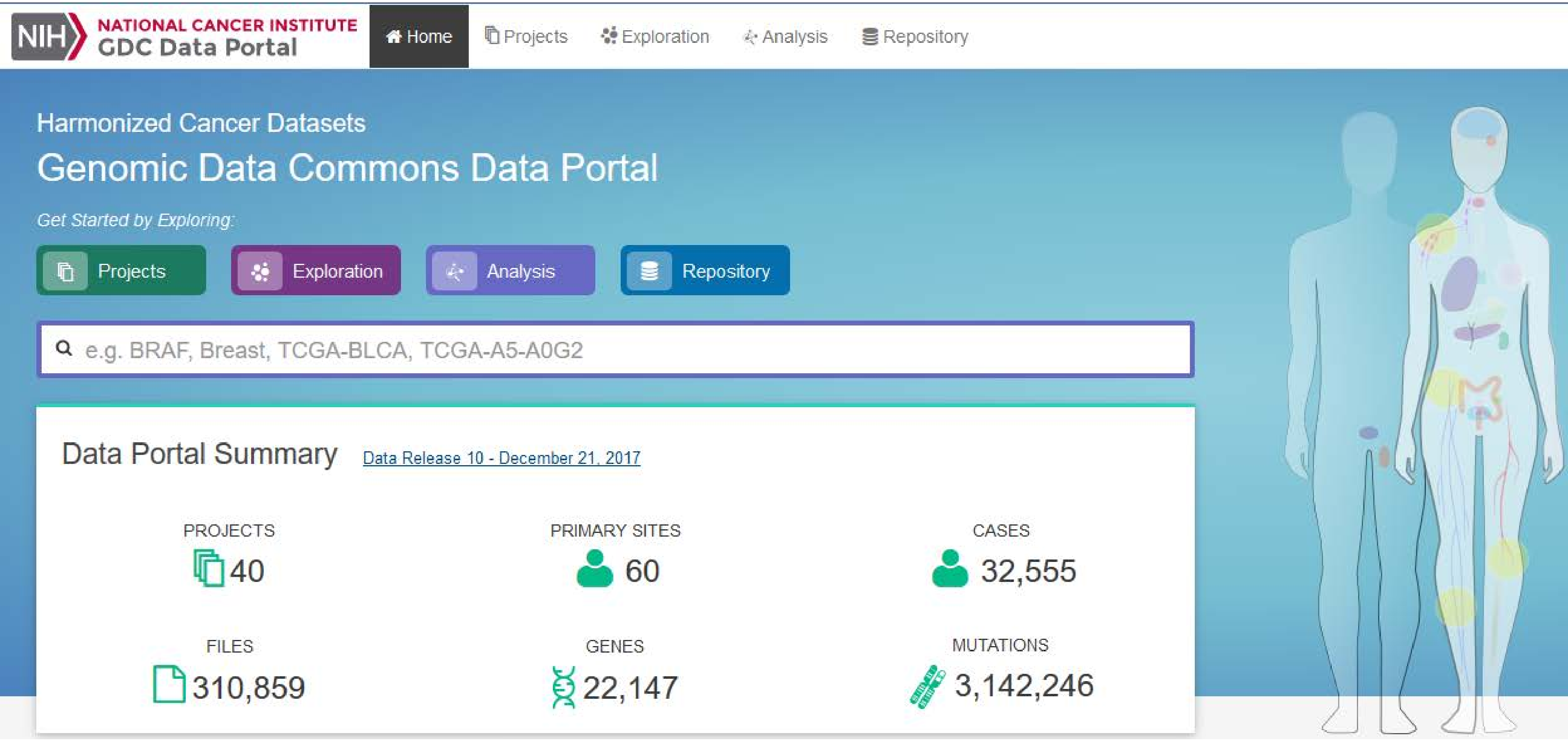
#	☐	AID	Activity	AC≤1[μM]	AC≤1[nM]	AC Range	BioAssay [Outcome Type]	Protein Target
1	☐	624743	🟥	1	1	0.00095 [μM]	Binding constant for LTK kinase domain [Confirmatory]	Leukocyte tyrosine kinase receptor[gi:143811416]
2	☐	742112	🟥	1	1	0.000927 [μM]	SANGER: Inhibition of human DEL cell growth in a cell viability assay. [Confirmatory]	
3	☐	624825	🟥	1	1	0.00085 [μM]	Binding constant for BMPR1B kinase domain [Confirmatory]	Bone morphogenetic protein receptor type-1B[gi:6226778]
4	☐	624916	🟥	1	1	0.00083 [μM]	Binding constant for ULK1 kinase domain [Confirmatory]	Serine/threonine-protein kinase ULK1[gi:317373288]
5	☐	625076	🟥	1	1	0.00093 [μM]	Binding constant for PLK4 kinase domain [Confirmatory]	Serine/threonine-protein kinase PLK4[gi:160113150]
6	☐	741847	🟥	1	1	6.03e-05 [μM]	SANGER: Inhibition of human SCC-3 cell growth in a cell viability assay. [Confirmatory]	
7	☐	741855	🟥	1	1	0.000564 [μM]	SANGER: Inhibition of human SF539 cell growth in a cell viability assay. [Confirmatory]	
8	☐	624899	🟥	1	1	0.00049 [μM]	Binding constant for ROS1 kinase domain [Confirmatory]	Proto-oncogene tyrosine-protein kinase ROS[gi:126302596]
9	☐	624741	🟥	1	1	0.00086 [μM]	Binding constant for LRRK2(G2019S) kinase domain [Confirmatory]	Leucine-rich repeat serine/threonine-protein kinase 2[gi:294862450]
10	☐	624742	▤			10 [μM]	Binding constant for NEK5 kinase domain [Confirmatory]	Serine/threonine-protein kinase Nek5[gi:74758252]
11	☐	624900	🟥	1		0.16 [μM]	Binding constant for RSK1(Kin.Dom.1-N-terminal) kinase domain [Confirmatory]	Ribosomal protein S6 kinase alpha-1[gi:20178306]
12	☐	624901	🟥	1		0.65 [μM]	Binding constant for RSK1(Kin.Dom.2-C-terminal) kinase domain [Confirmatory]	Ribosomal protein S6 kinase alpha-1[gi:20178306]
13	☐	624902	🟥	1		0.15 [μM]	Binding constant for MEK4 kinase domain [Confirmatory]	Dual specificity mitogen-activated protein kinase kinase 4[gi:1170596]
14	☐	624903	🟥			4.8 [μM]	Binding constant for SRPK1 kinase domain [Confirmatory]	SRSF protein kinase 1[gi:209572680]
15	☐	624904	▤			10 [μM]	Binding constant for NEK4 kinase domain [Confirmatory]	Serine/threonine-protein kinase Nek4[gi:229462924]

Targets with Kd or IC50 below 10nM:

ABL1 ALK BMPR1B DCLK1 EGFR FER FES FLT3 GAK IGF1R INSR INSRR LRRK2 LTK NUAK2 PLK4 PTK2 PTK2B ROS1 STK33 TNK1 TNK2 ULK1 ULK2 YES1

Example 4

TCGA: MULTI-OMIC TUMOR DATA



NIH NATIONAL CANCER INSTITUTE GDC Data Portal

Home | Projects | Exploration | Analysis | Repository

Harmonized Cancer Datasets

Genomic Data Commons Data Portal

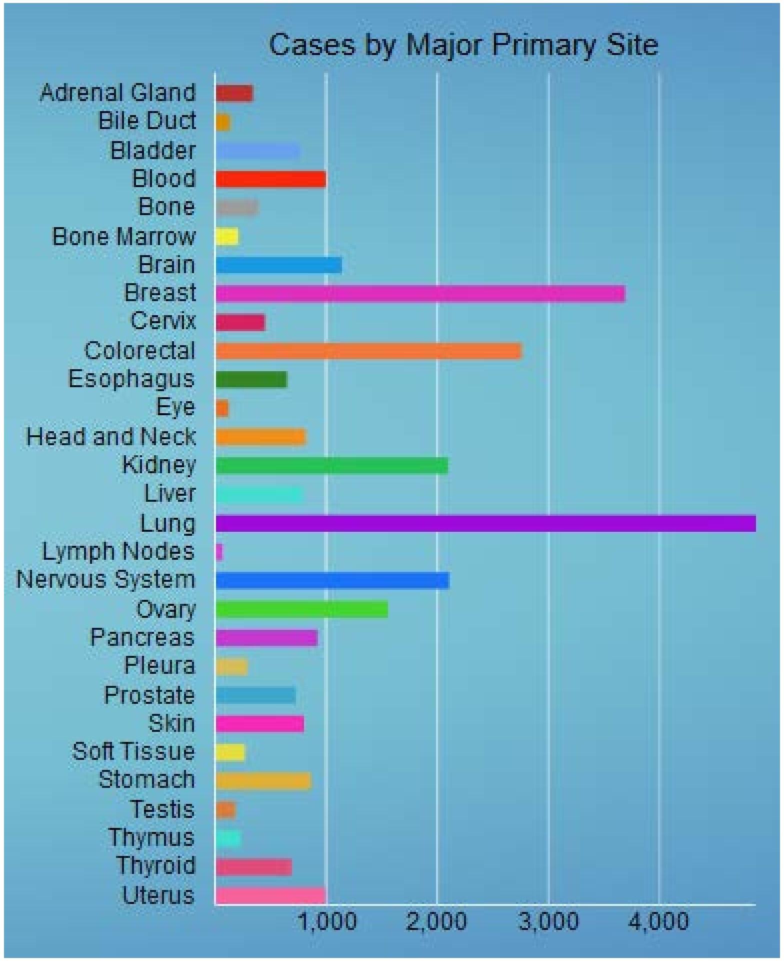
Get Started by Exploring:

Projects | Exploration | Analysis | Repository

e.g. BRAF, Breast, TCGA-BLCA, TCGA-A5-A0G2

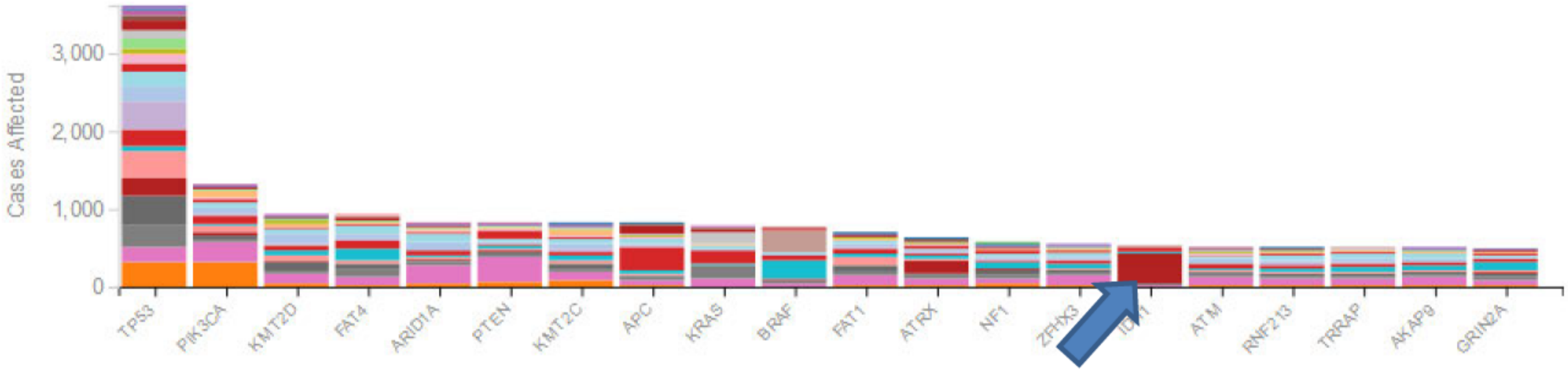
Data Portal Summary Data Release 10 - December 21, 2017

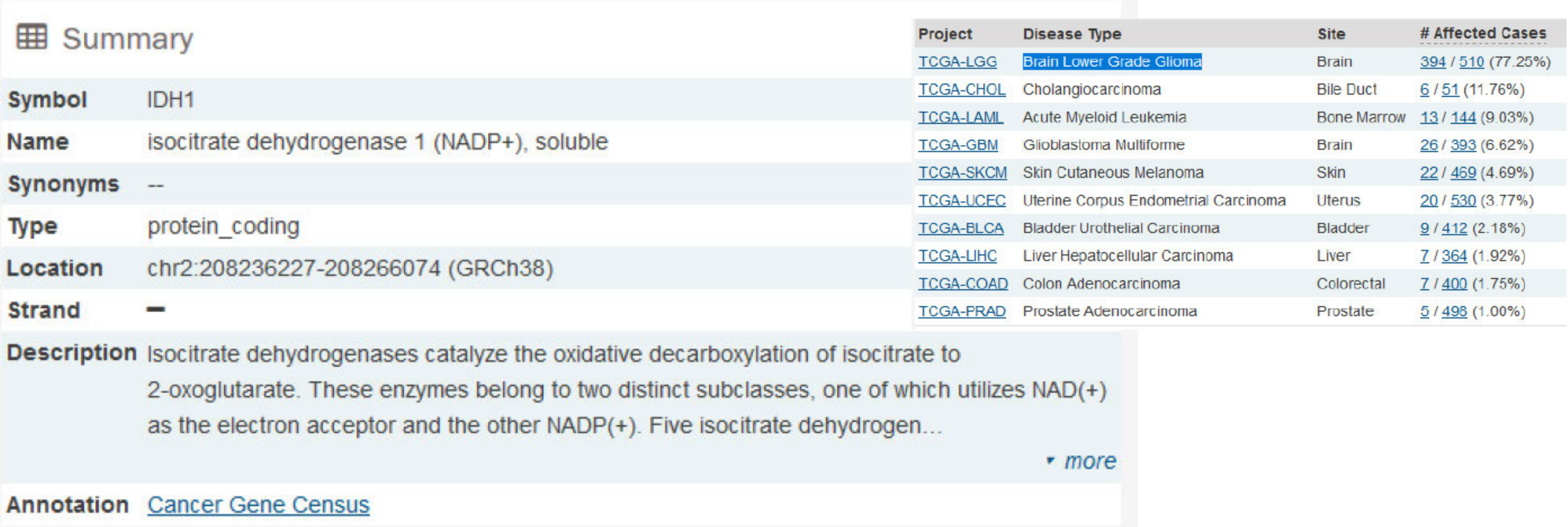
PROJECTS 40	PRIMARY SITES 60	CASES 32,555
FILES 310,859	GENES 22,147	MUTATIONS 3,142,246



What questions might you ask using these sequencing data?

Top Mutated Cancer Genes

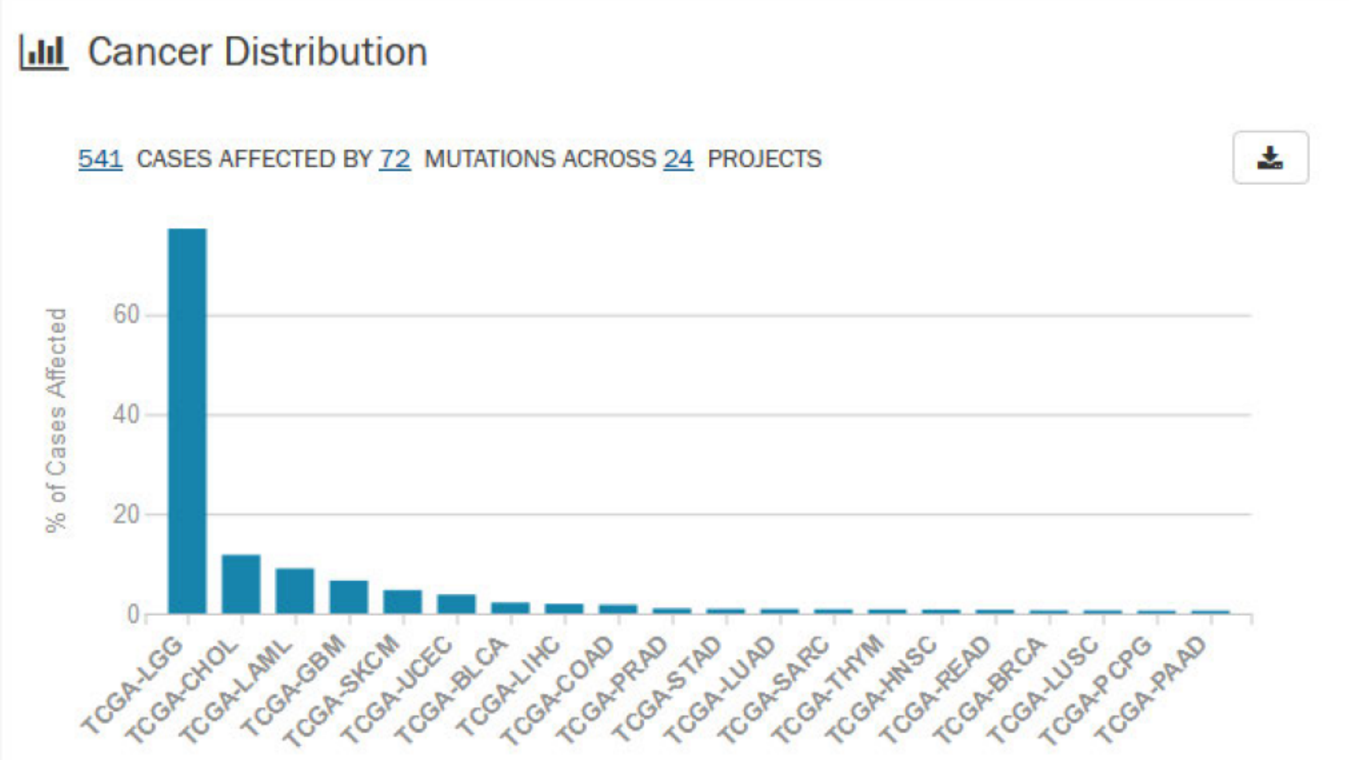




Summary

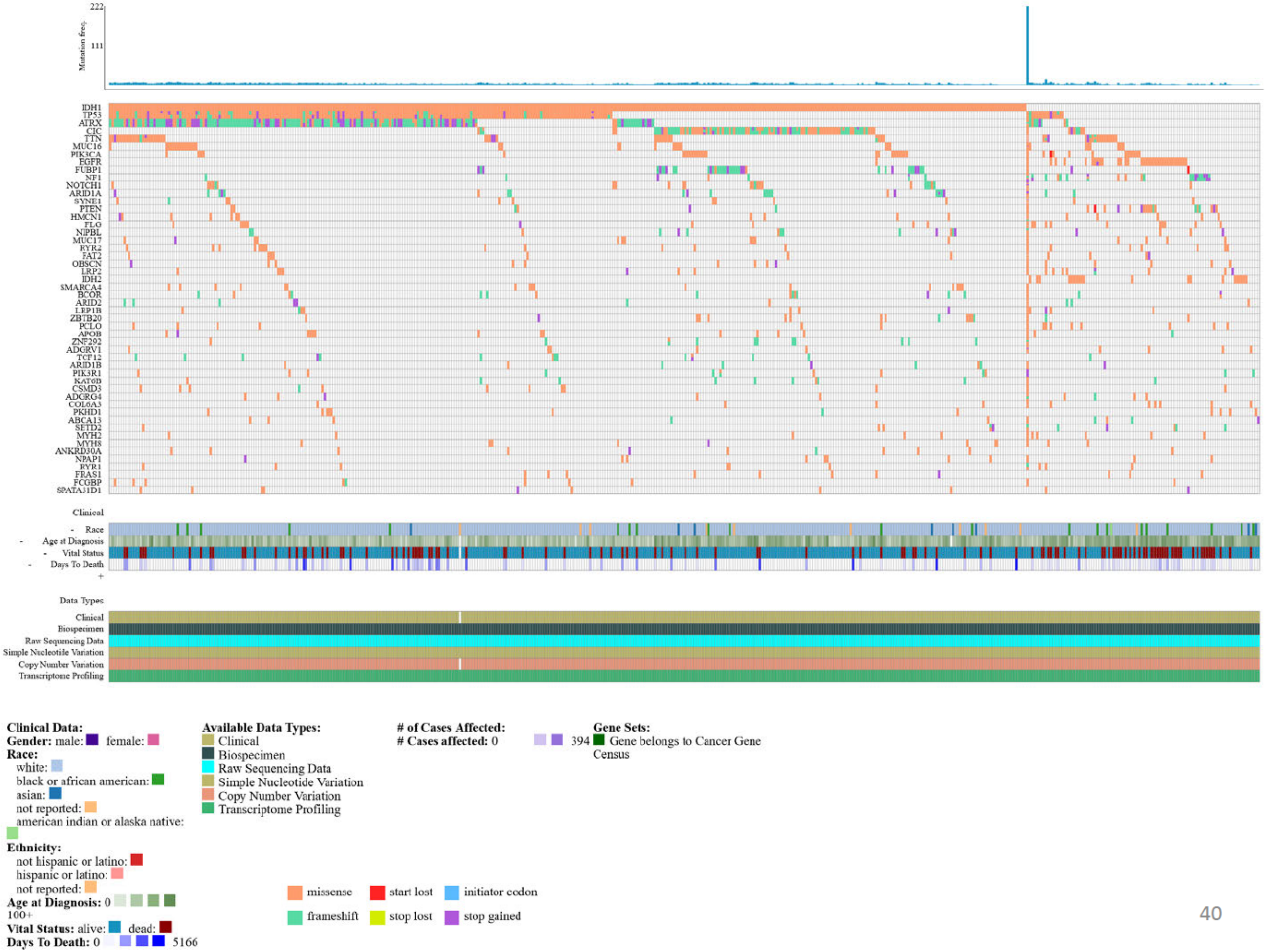
Symbol	IDH1
Name	isocitrate dehydrogenase 1 (NADP+), soluble
Synonyms	--
Type	protein_coding
Location	chr2:208236227-208266074 (GRCh38)
Strand	━
Description	Isocitrate dehydrogenases catalyze the oxidative decarboxylation of isocitrate to 2-oxoglutarate. These enzymes belong to two distinct subclasses, one of which utilizes NAD(+) as the electron acceptor and the other NADP(+). Five isocitrate dehydrogen... ▾ *more*
Annotation	Cancer Gene Census

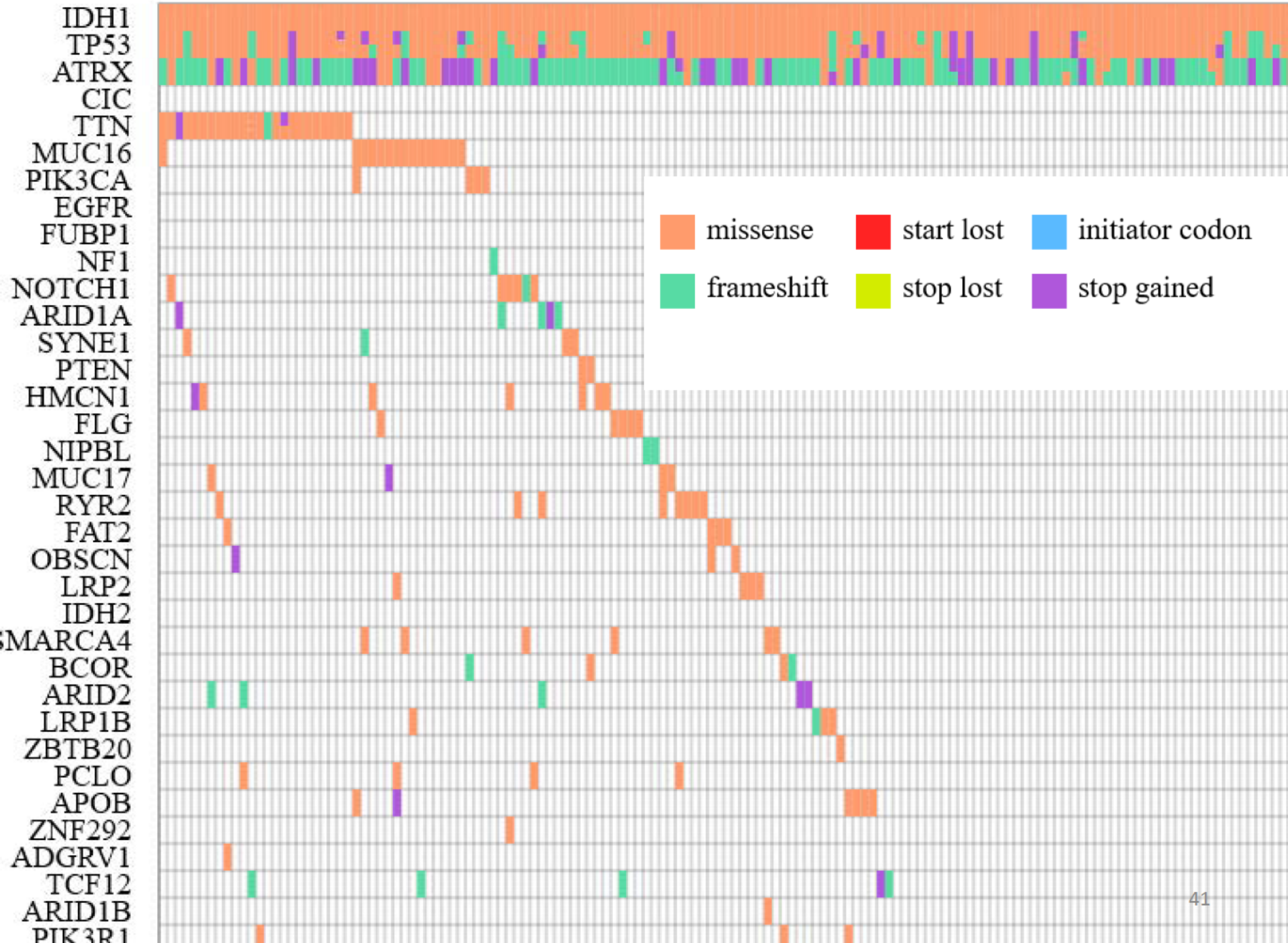
Project	Disease Type	Site	# Affected Cases
TCGA-LGG	Brain Lower Grade Glioma	Brain	394 / 510 (77.25%)
TCGA-CHOL	Cholangiocarcinoma	Bile Duct	6 / 51 (11.76%)
TCGA-LAML	Acute Myeloid Leukemia	Bone Marrow	13 / 144 (9.03%)
TCGA-GBM	Glioblastoma Multiforme	Brain	26 / 393 (6.62%)
TCGA-SKCM	Skin Cutaneous Melanoma	Skin	22 / 469 (4.69%)
TCGA-UCEC	Uterine Corpus Endometrial Carcinoma	Uterus	20 / 530 (3.77%)
TCGA-BLCA	Bladder Urothelial Carcinoma	Bladder	9 / 412 (2.18%)
TCGA-LIHC	Liver Hepatocellular Carcinoma	Liver	7 / 364 (1.92%)
TCGA-COAD	Colon Adenocarcinoma	Colorectal	7 / 400 (1.75%)
TCGA-PRAD	Prostate Adenocarcinoma	Prostate	5 / 498 (1.00%)

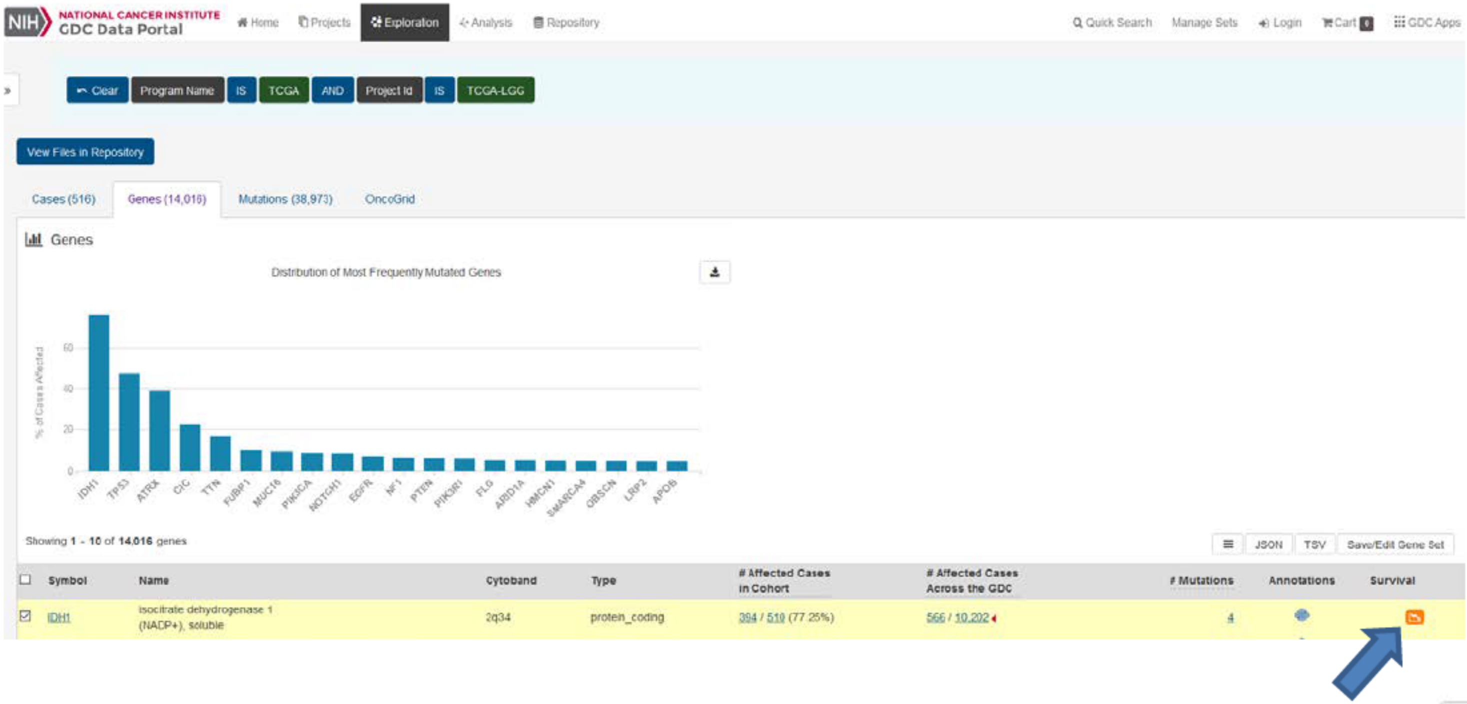


Cancer Distribution

541 CASES AFFECTED BY 72 MUTATIONS ACROSS 24 PROJECTS

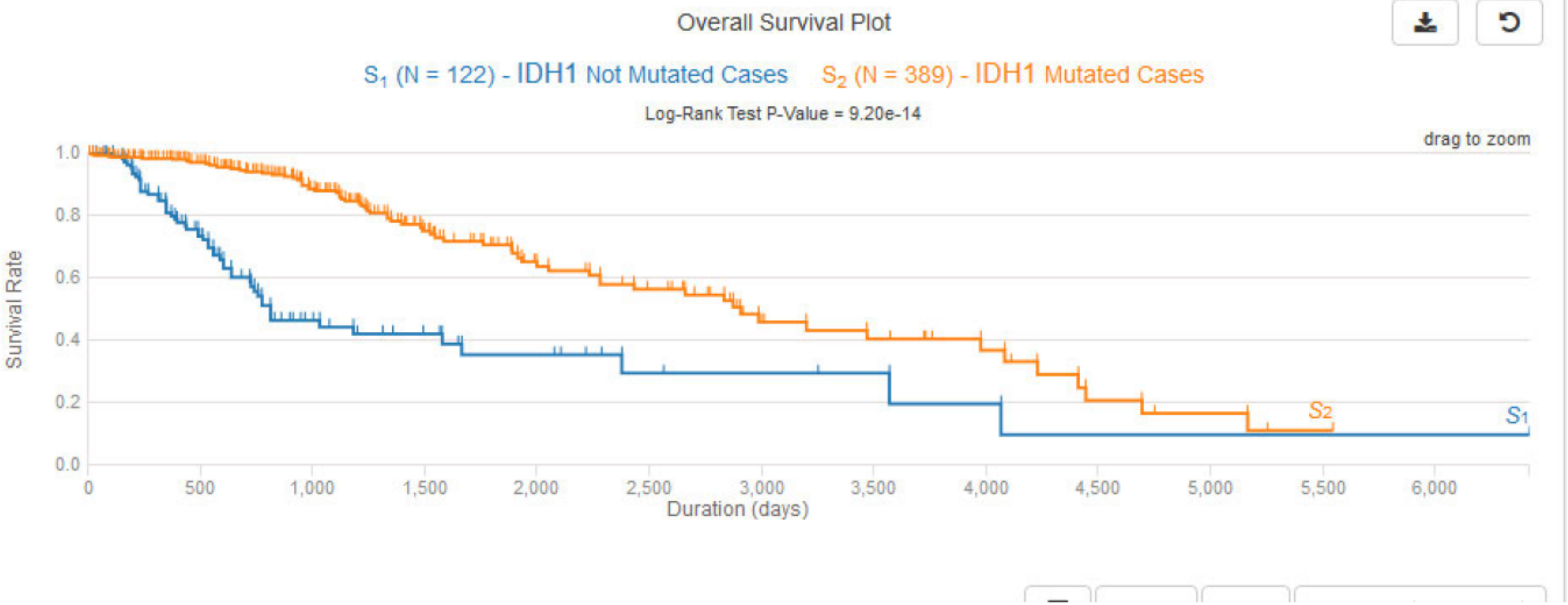

222
111
Mutation freq.
IDH1
TP53
ATRX
CIC
TTN
MUC16
PIK3CA
EGFR
FUBP1
NF1
NOTCH1
ARID1A
SYNE1
PTEN
HMCN1
FLG
NIPBL
MUC17
RYR2
FAT2
OBSCN
LRP2
IDH2
SMARCA4
BCOR
ARID2
LRP1B
ZBTB20
PCLO
APOB
ZNF292
ADGRV1
TCF12
ARID1B
PIK3R1
KAT6B
CSMD3
ADGRG4
COL6A3
PKHD1
ABCA13
SETD2
MYH2
MYH8
ANKRD30A
NPAP1
RYR1
FRAS1
FCGBP
SPATA31D1
Clinical
- Race
- Age at Diagnosis
- Vital Status
- Days To Death
+
Data Types
Clinical
Biospecimen
Raw Sequencing Data
Simple Nucleotide Variation
Copy Number Variation
Transcriptome Profiling
Clinical Data:
Gender: male: female:
Race:
white:
black or african american:
asian:
not reported:
american indian or alaska native:
Ethnicity:
not hispanic or latino:
hispanic or latino:
not reported:
Age at Diagnosis: 0
100+
Vital Status: alive: dead:
Days To Death: 0 5166
Available Data Types:
Clinical
Biospecimen
Raw Sequencing Data
Simple Nucleotide Variation
Copy Number Variation
Transcriptome Profiling
# of Cases Affected:
# Cases affected: 0
394
Gene Sets:
Gene belongs to Cancer Gene Census
missense
frameshift
start lost
stop lost
initiator codon
stop gained

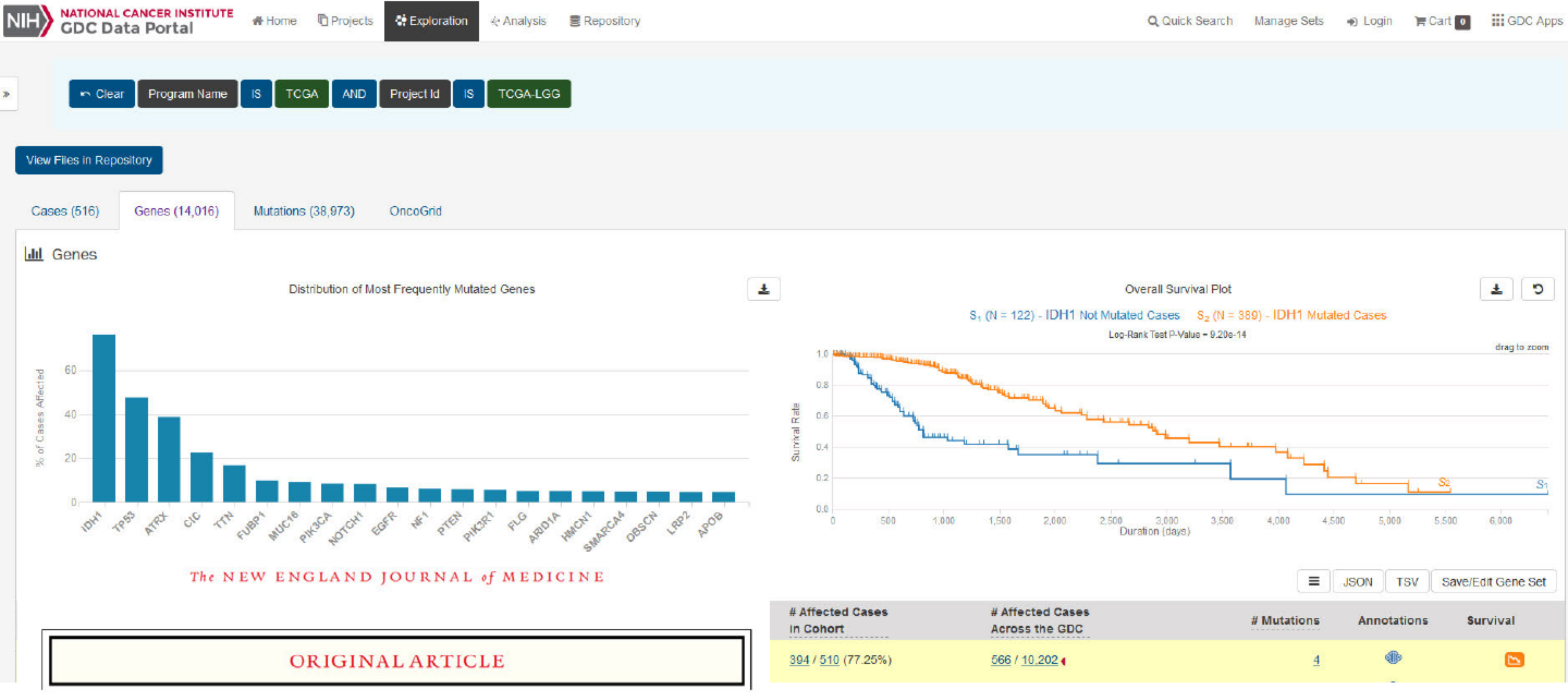

IDH1
TP53
ATRX
CIC
TTN
MUC16
PIK3CA
EGFR
FUBP1
NF1
NOTCH1
ARID1A
SYNE1
PTEN
HMCN1
FLG
NIPBL
MUC17
RYR2
FAT2
OBSCN
LRP2
IDH2
SMARCA4
BCOR
ARID2
LRP1B
ZBTB20
PCLO
APOB
ZNF292
ADGRV1
TCF12
ARID1B
PIK3R1
missense
start lost
initiator codon
frameshift
stop lost
stop gained



Showing **1** - **10** of **14,016** genes

□	Symbol	Name	Cytoband	Type	# Affected Cases in Cohort	# Affected Cases Across the GDC	# Mutations	Annotations	Survival
☑	IDH1	isocitrate dehydrogenase 1 (NADP+), soluble	2q34	protein_coding	394 / 510 (77.25%)	566 / 10,202	4	●	▣





# Affected Cases in Cohort	# Affected Cases Across the GDC	# Mutations	Annotations	Survival
394 / 510 (77.25%)	566 / 10,202	4		

The NEW ENGLAND JOURNAL *of* MEDICINE

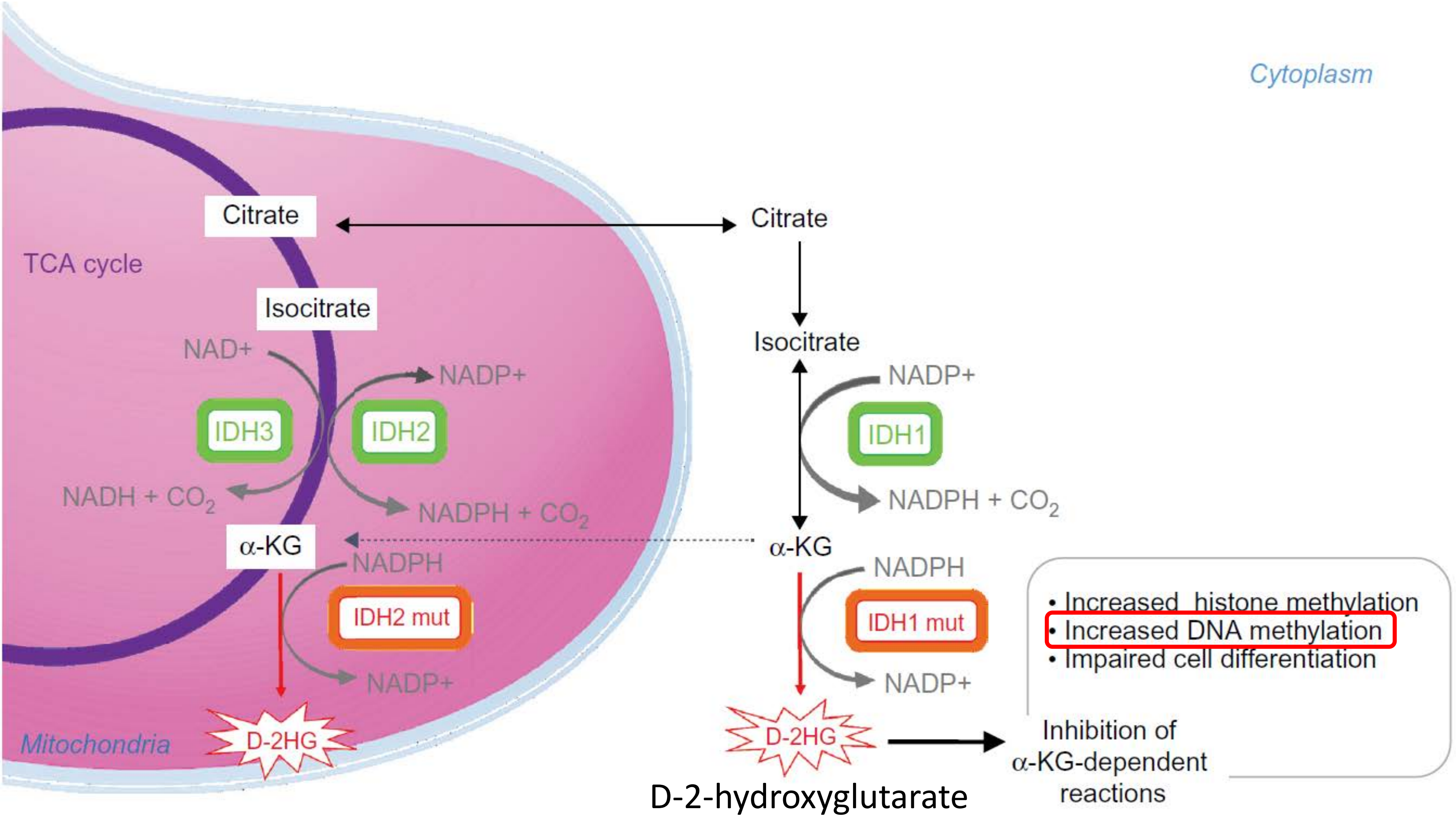
ORIGINAL ARTICLE

IDH1 and *IDH2* Mutations in Gliomas

N ENGL J MED 360;8 NEJM.ORG FEBRUARY 19, 2009

RESULTS

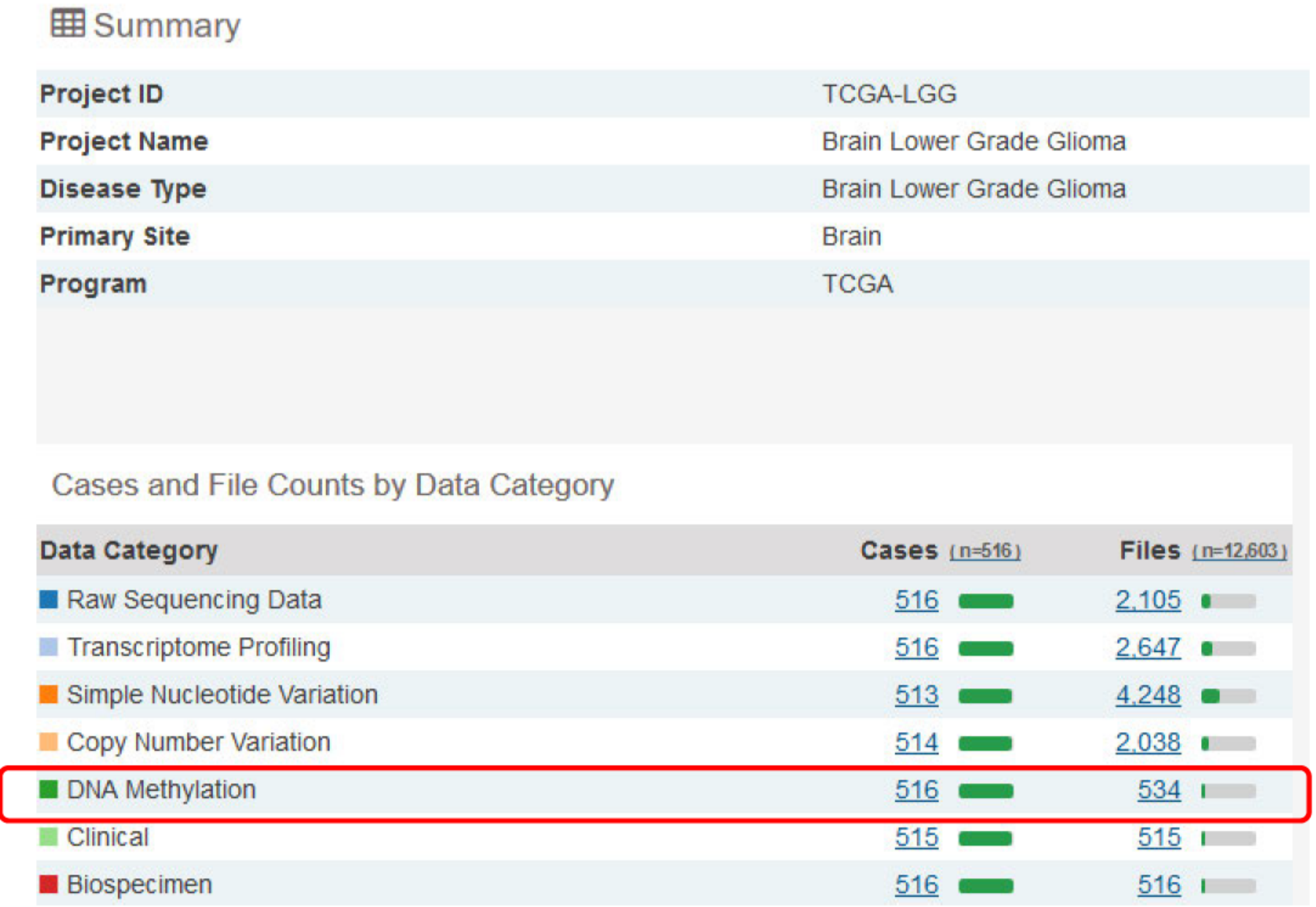
We identified mutations that affected amino acid 132 of *IDH1* in more than 70% of WHO grade II and III astrocytomas and oligodendrogliomas and in glioblastomas that developed from these lower-grade lesions. Tumors without mutations in *IDH1* often had mutations affecting the analogous amino acid (R172) of the *IDH2* gene. Tumors with *IDH1* or *IDH2* mutations had distinctive genetic and clinical characteristics, and patients with such tumors had a better outcome than those with wild-type *IDH* genes. Each of four tested *IDH1* and *IDH2* mutations reduced the enzymatic activity of the encoded protein.



D-2-hydroxyglutarate

J Blood Med. 2016; 7: 171–180 doi: 10.2147/JBM.S70716
IDH1 and IDH2 mutations as novel therapeutic targets: current perspectives

Many Types of Data Available

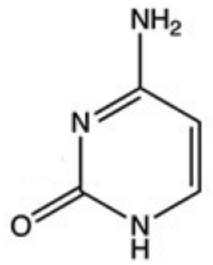


Summary

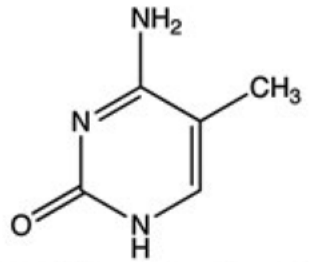
Project ID	TCGA-LGG
Project Name	Brain Lower Grade Glioma
Disease Type	Brain Lower Grade Glioma
Primary Site	Brain
Program	TCGA

Cases and File Counts by Data Category

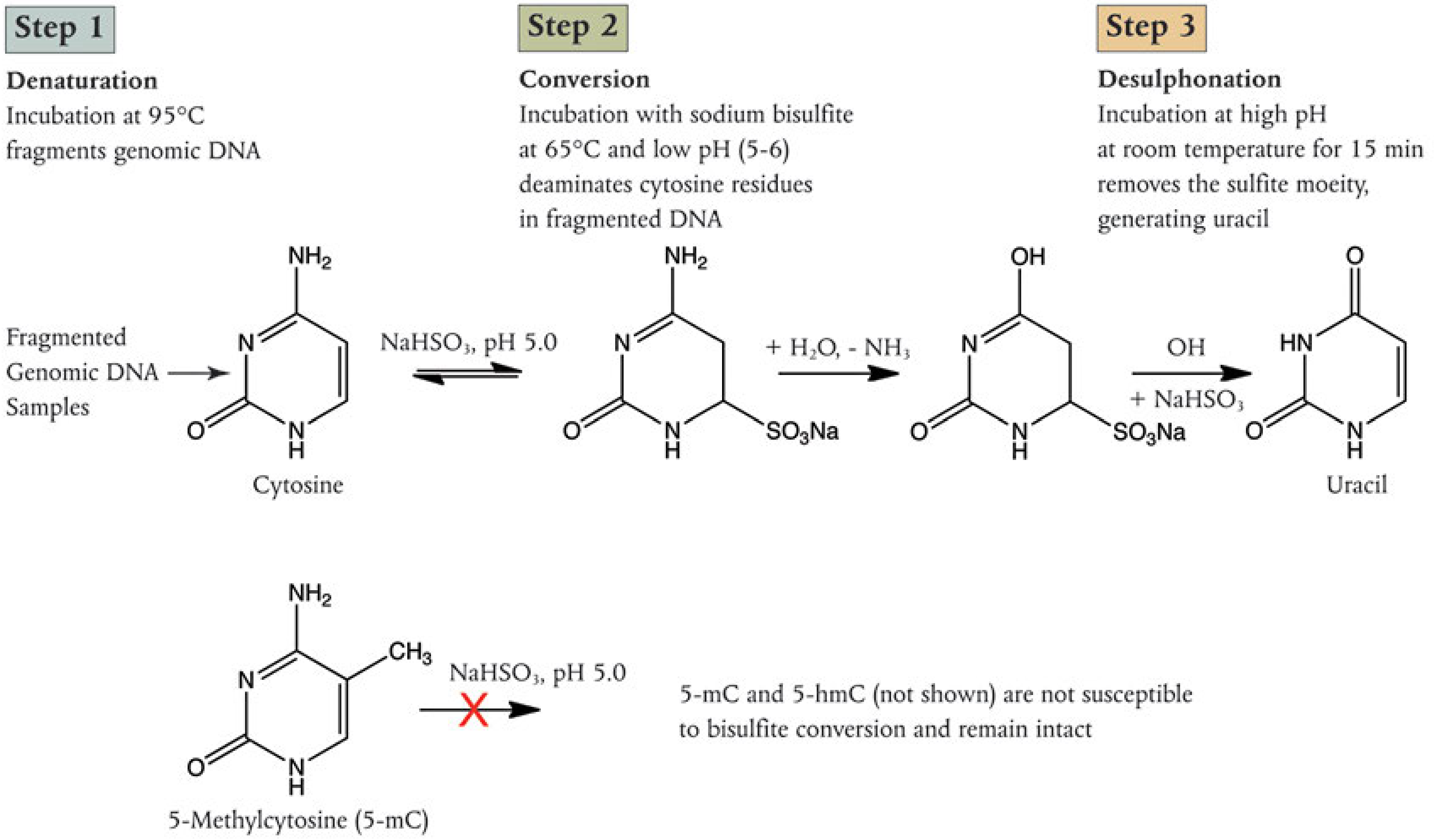
Data Category	Cases (n=516)	Files (n=12,603)
🟦 Raw Sequencing Data	516	2,105
🟦 Transcriptome Profiling	516	2,647
🟧 Simple Nucleotide Variation	513	4,248
🟧 Copy Number Variation	514	2,038
🟩 DNA Methylation	516	534
🟩 Clinical	515	515
🟥 Biospecimen	516	516



Cytosine



5-Methylcytosine (5-mC)



Step 1

Denaturation
Incubation at 95°C fragments genomic DNA

Step 2

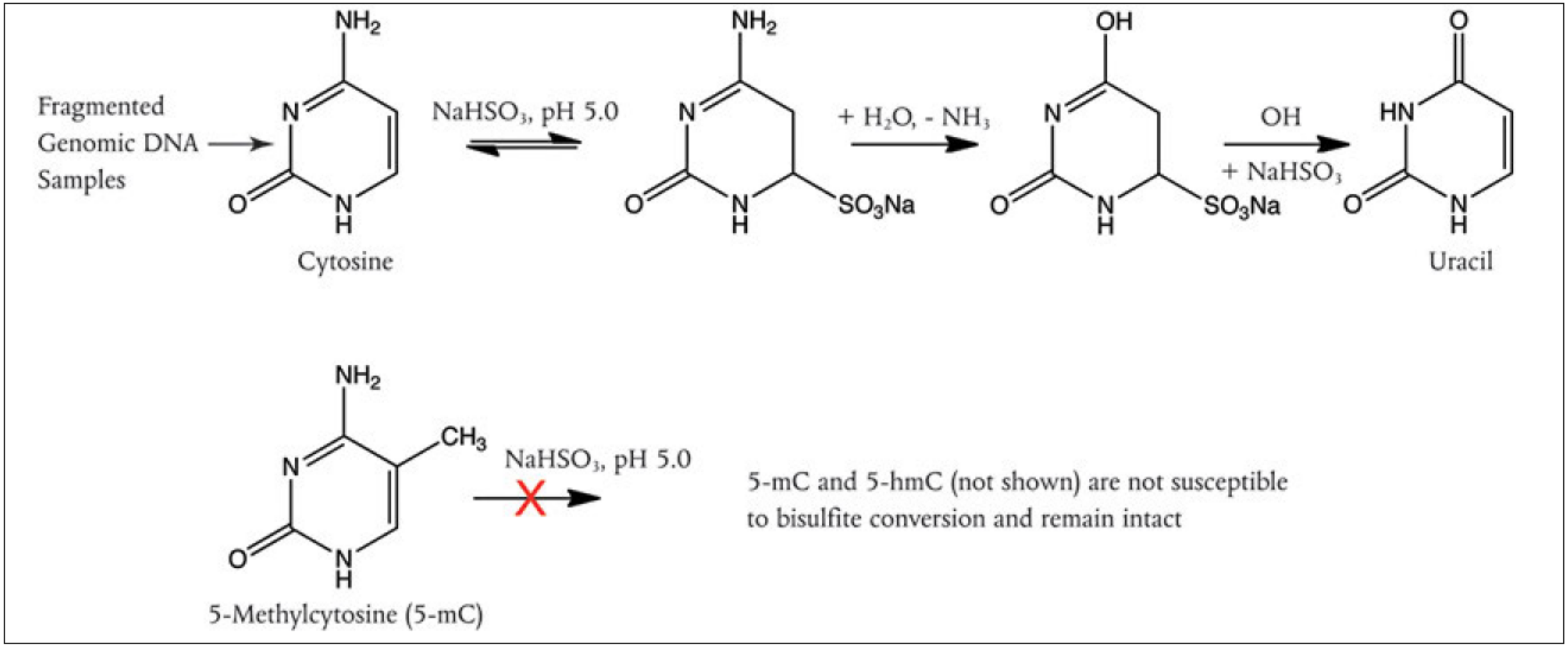
Conversion
Incubation with sodium bisulfite at 65°C and low pH (5-6) deaminates cytosine residues in fragmented DNA

Step 3

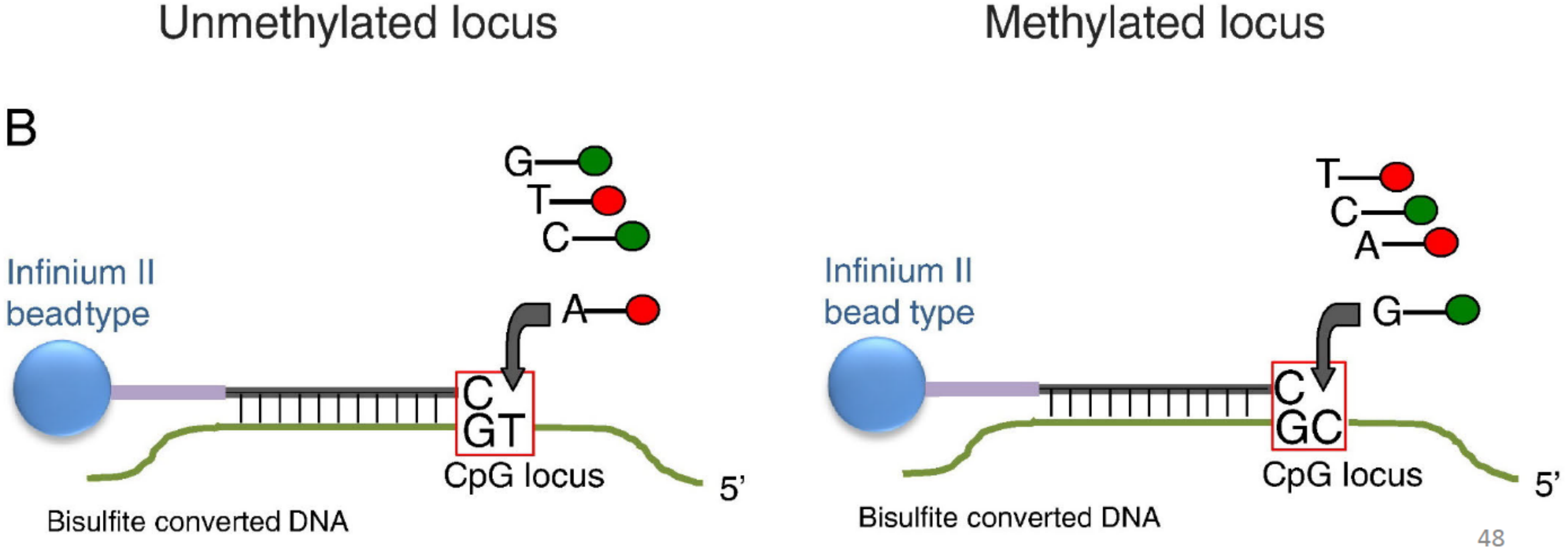
Desulphonation
Incubation at high pH at room temperature for 15 min removes the sulfite moeity, generating uracil

5-mC and 5-hmC (not shown) are not susceptible to bisulfite conversion and remain intact

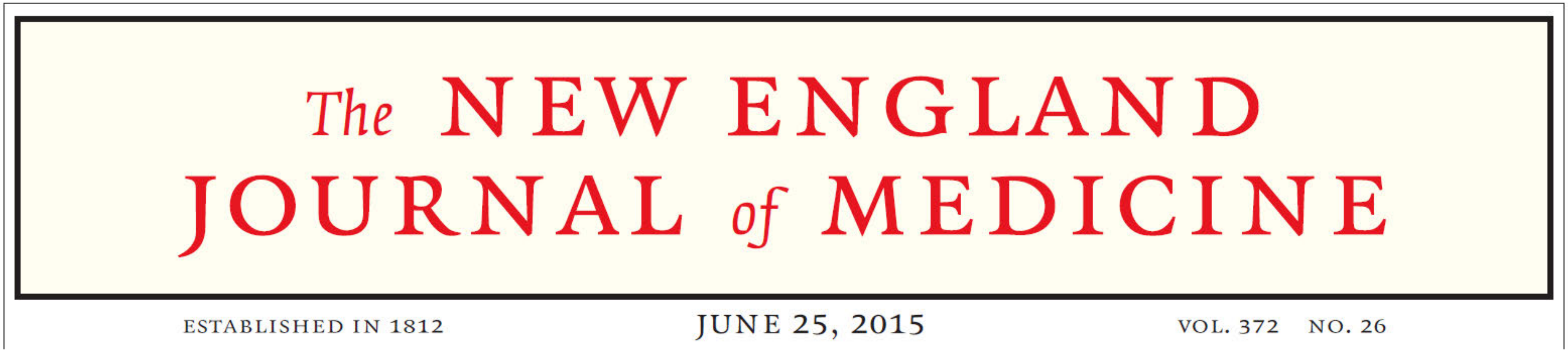
https://www.neb.com/-/media/nebus/page-images/newsized-brochure-images/epigenetics/bisulfite_conversion.jpg?la=en&device=modal



5-mC and 5-hmC (not shown) are not susceptible to bisulfite conversion and remain intact



https://doi.org/10.1016/j.ygeno.2011.07.007

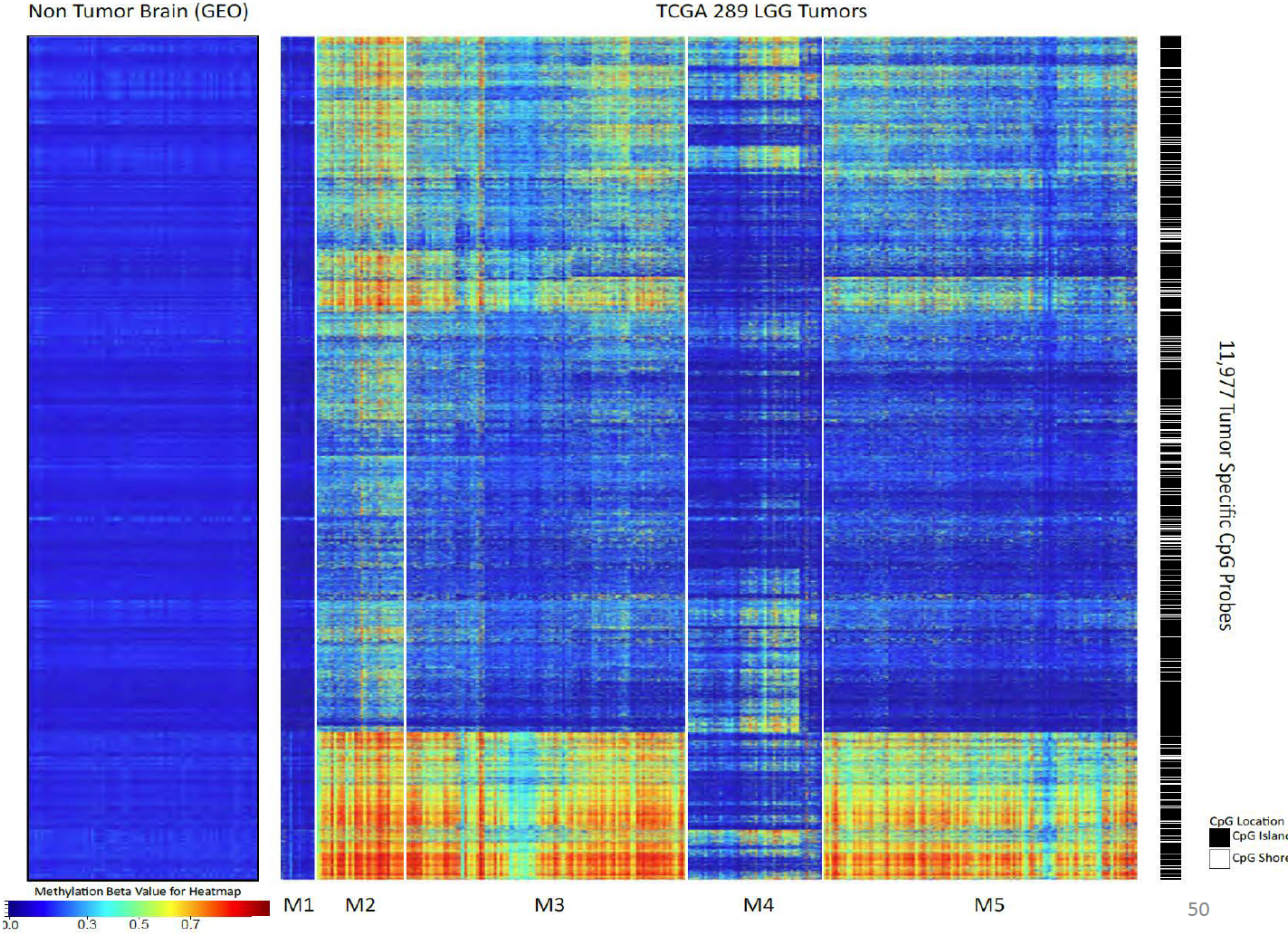


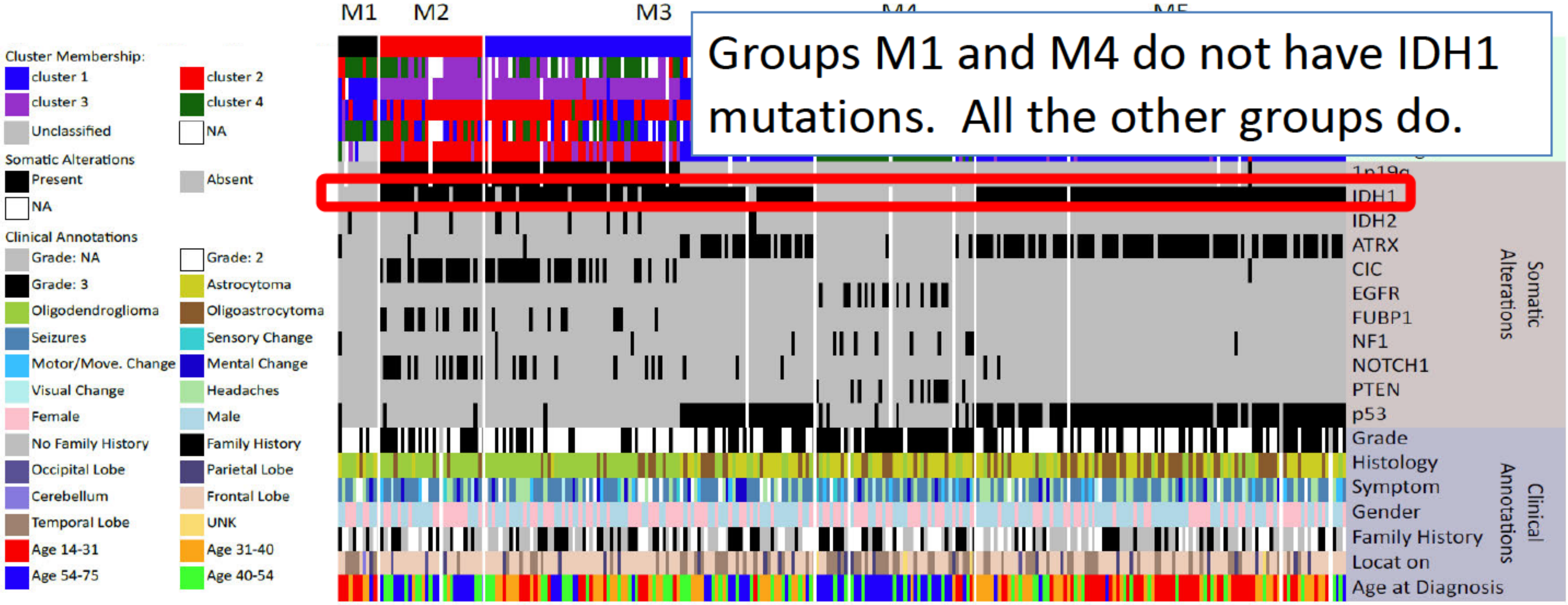
The NEW ENGLAND JOURNAL of MEDICINE

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

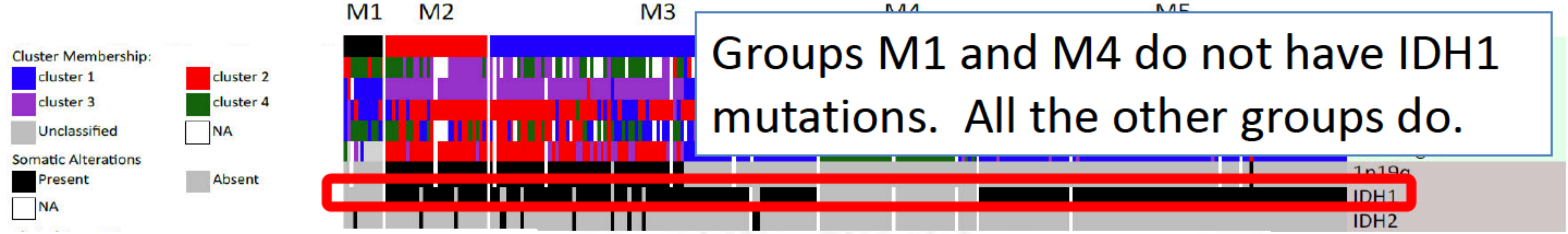
Comprehensive, Integrative Genomic Analysis of Diffuse Lower-Grade Gliomas

The Cancer Genome Atlas Research Network*

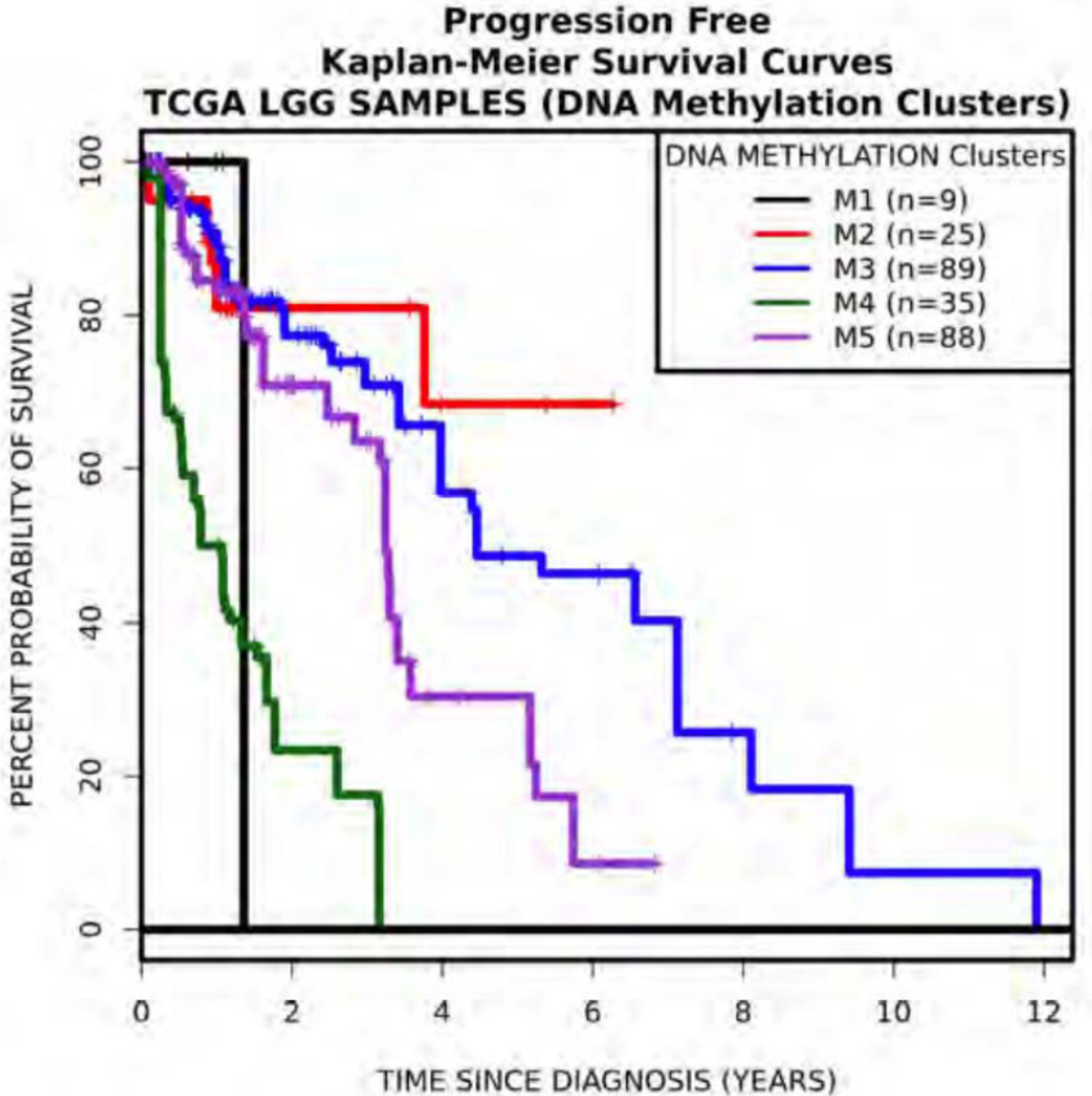

Non Tumor Brain (GEO)
TCGA 289 LGG Tumors
11,977 Tumor Specific CpG Probes
CpG Location
CpG Islan
CpG Shore
Methylation Beta Value for Heatmap
0.0
0.3
0.5
0.7
M1
M2
M3
M4
M5



Groups M1 and M4 do not have IDH1 mutations. All the other groups do.



Groups M1 and M4 do not have IDH1 mutations. All the other groups do.


Progression Free
Kaplan-Meier Survival Curves
TCGA LGG SAMPLES (DNA Methylation Clusters)
DNA METHYLATION Clusters
M1 (n=9)
M2 (n=25)
M3 (n=89)
M4 (n=35)
M5 (n=88)
PERCENT PROBABILITY OF SURVIVAL
TIME SINCE DIAGNOSIS (YEARS)