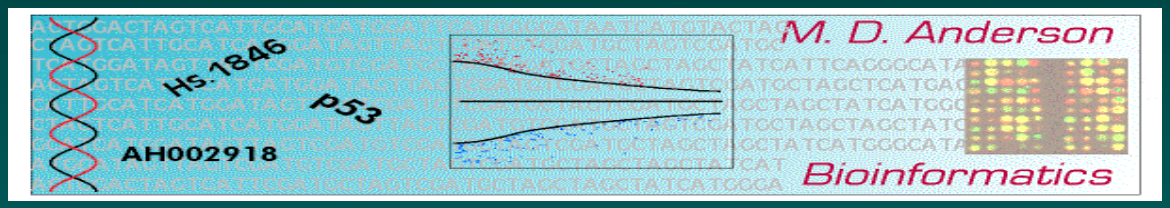
Exciting Big Data: Homework!

Keith A. Baggerly
Bioinformatics and Computational Biology
UT M. D. Anderson Cancer Center
kabagg@mdanderson.org

SISBID, July 18, 2016



Big Data is Revolutionizing Oncology

Looking at tons of measurements is changing how we model and understand the underlying biological processes.

Lots of this data is *public*.

We want you to play with it.

Since I work in the cancer field, that's where I'll draw most of my examples from.

Some Available Big Data Sources

The Cancer Genome Atlas, TCGA
TumorPortal
The Cancer Genomics Hub, CGHUB
The Cancer Cell Line Encyclopedia, CCLE
Genotype-Tissue Expression, GTEx
Project Achilles

GEO A public repository endorsed by journals for data deposition.

Now, what can we learn from this?

Putting data in context...

Oncology 101

Cancer is a genetic disease.

Every time a cell duplicates, it replicates its entire genome.

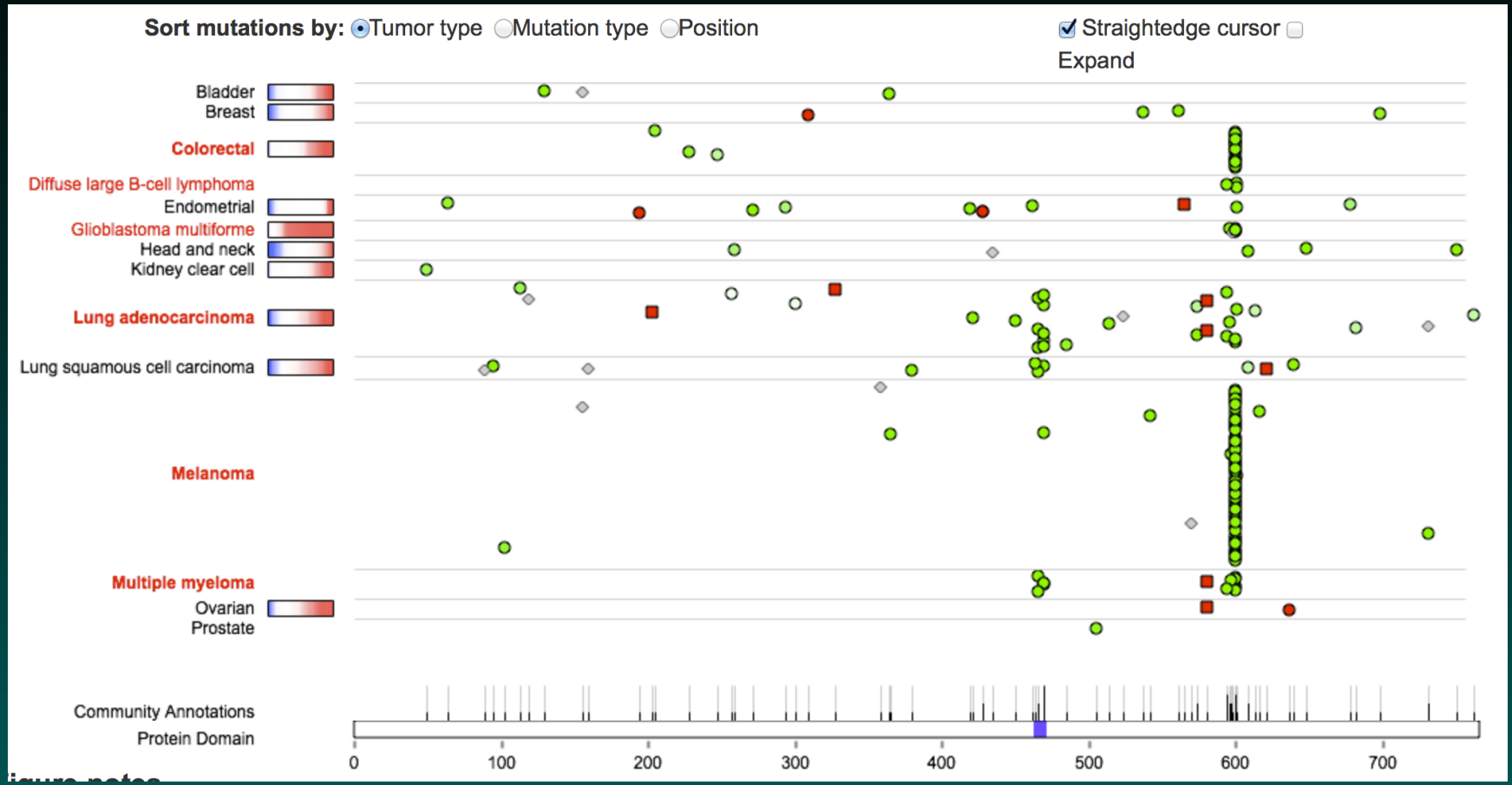
Replication is not perfect.

Most changes are irrelevant, but some can confer growth or survival advantages.

Some genes are particularly important.

These fall into two categories: *oncogenes* and *tumor suppressors*.

Looking at BRAF (an Oncogene)



Mutation sites are not random!

Mutations differ by tissue type

Looking at TP53 (a Tumor Suppressor)

Mutation patterns strikingly different; not as focused.

Oncology 102

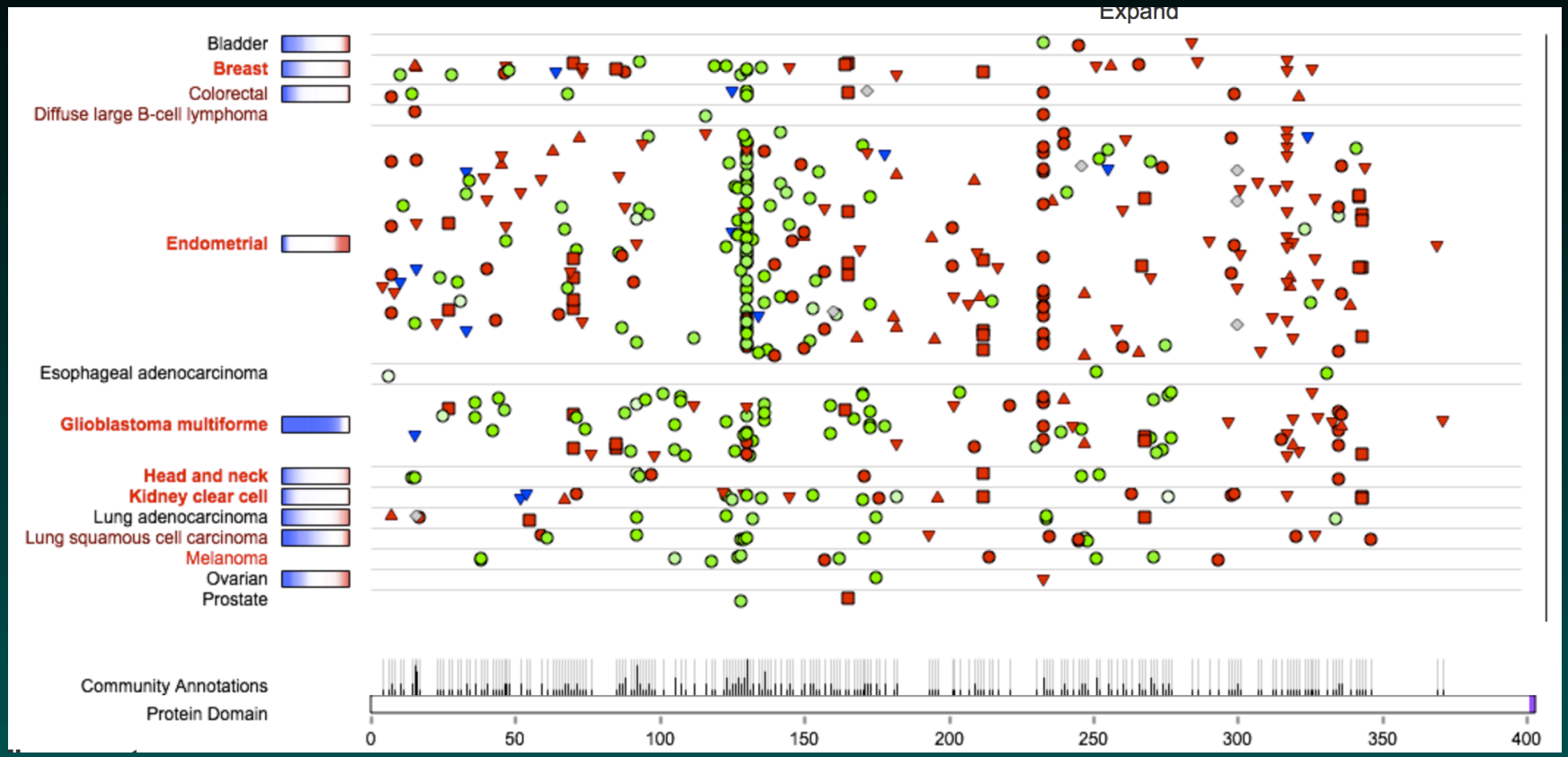
Changes can happen through many mechanisms; sequence alteration is the most direct.

DNA makes RNA makes Protein

Some other types of changes involve

copy number (DNA),
gene expression (mRNA), and
methylation (DNA/mRNA regulation)

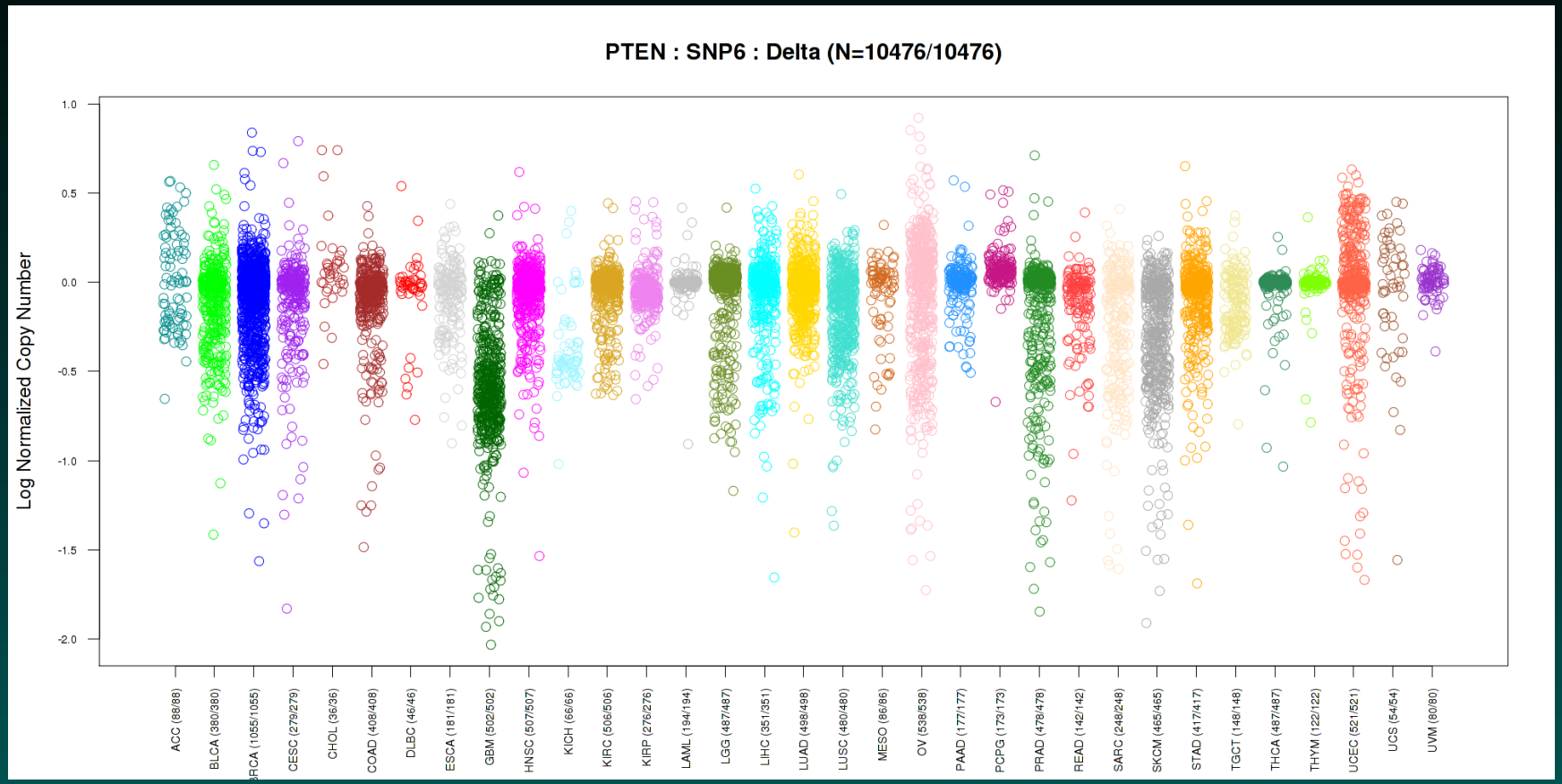
Exploring PTEN mutation



The dominant mutation in UCEC; but note CN in GBM

Mutated in UCEC

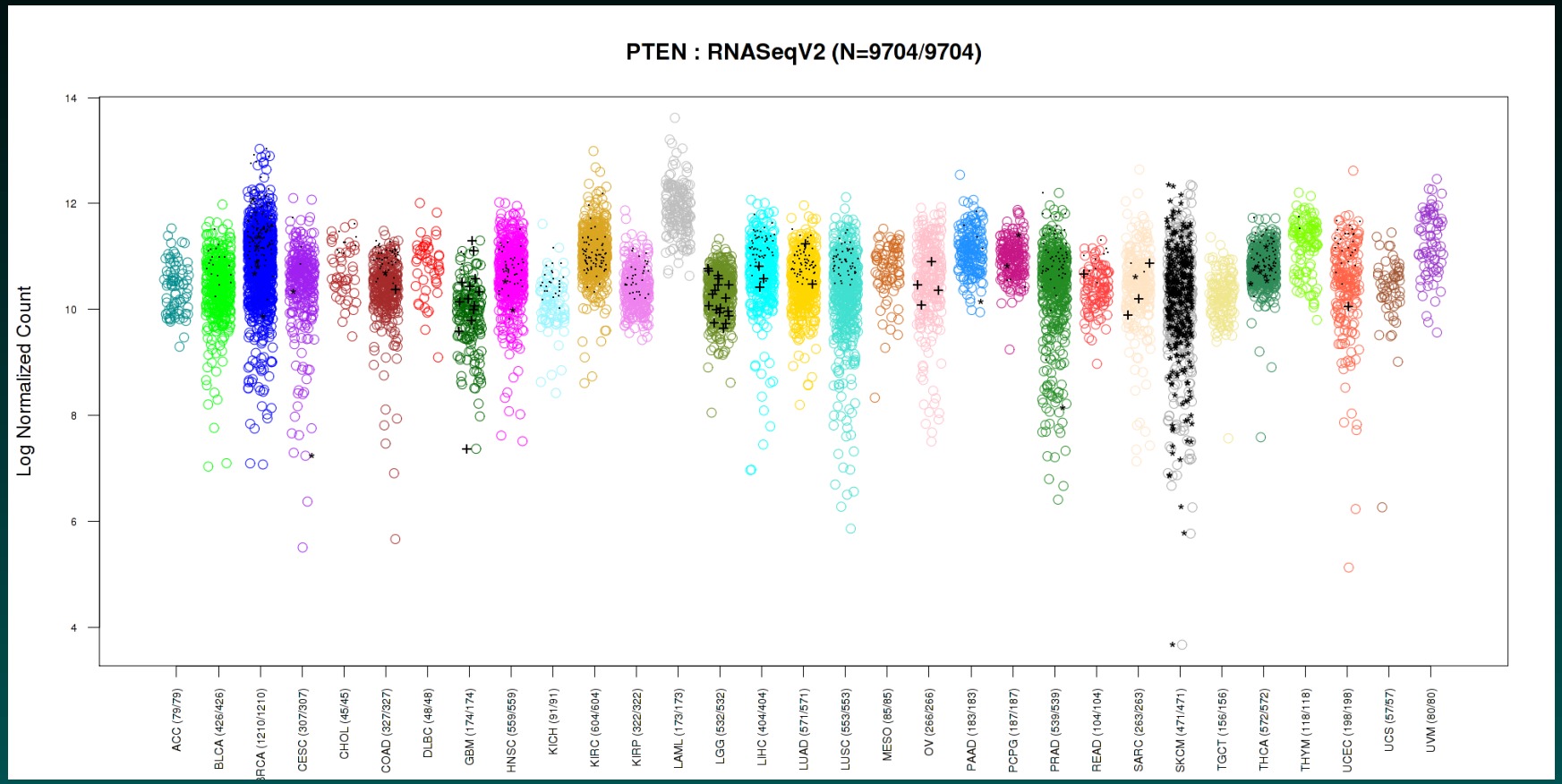
Exploring PTEN copy number



Loss in GBM, some in LGG, some in PRAD and SKCM

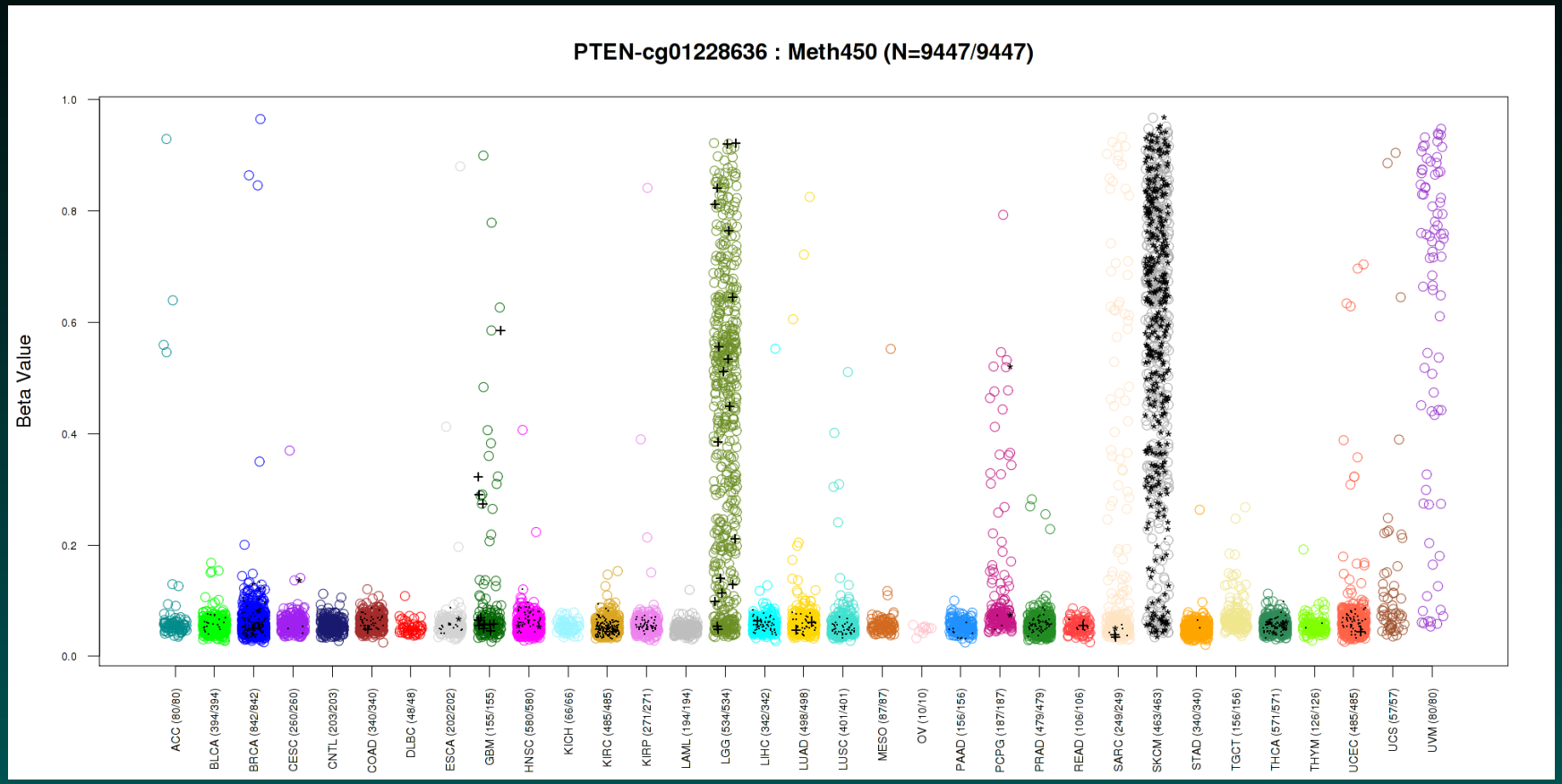
Lost in GBM

Exploring PTEN expression



The story with expression is less clear - what's normal?

Exploring PTEN methylation



Lots in LGG and SKCM, less in GBM

Methylated in LGG

Oncology 103

Cell safeguards prevent single errors from getting out of control, but once multiple hits accumulate, the system breaks.

Important relationships are often highlighted through co-occurrence or mutual exclusivity of changes

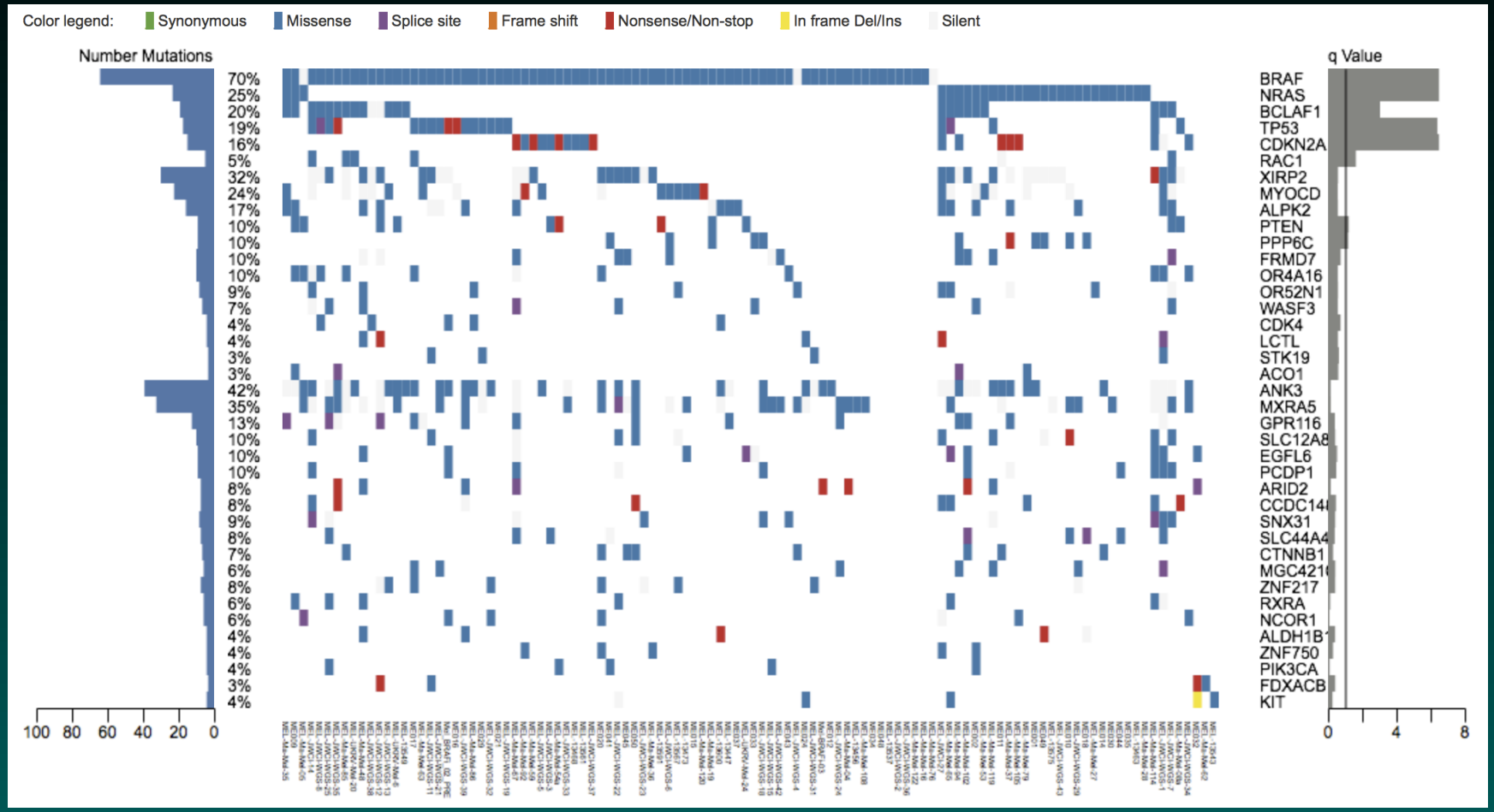
Since we have two copies of TP53, it often happens that one copy is mutated and the other is lost.

Since some other genes are activated in sequence, breaking any of them sets the system out of control.

Specific pathway breaks are “Hallmarks of Cancer”

Different genes drive different types of tumors.

Looking at SKCM (Melanoma)

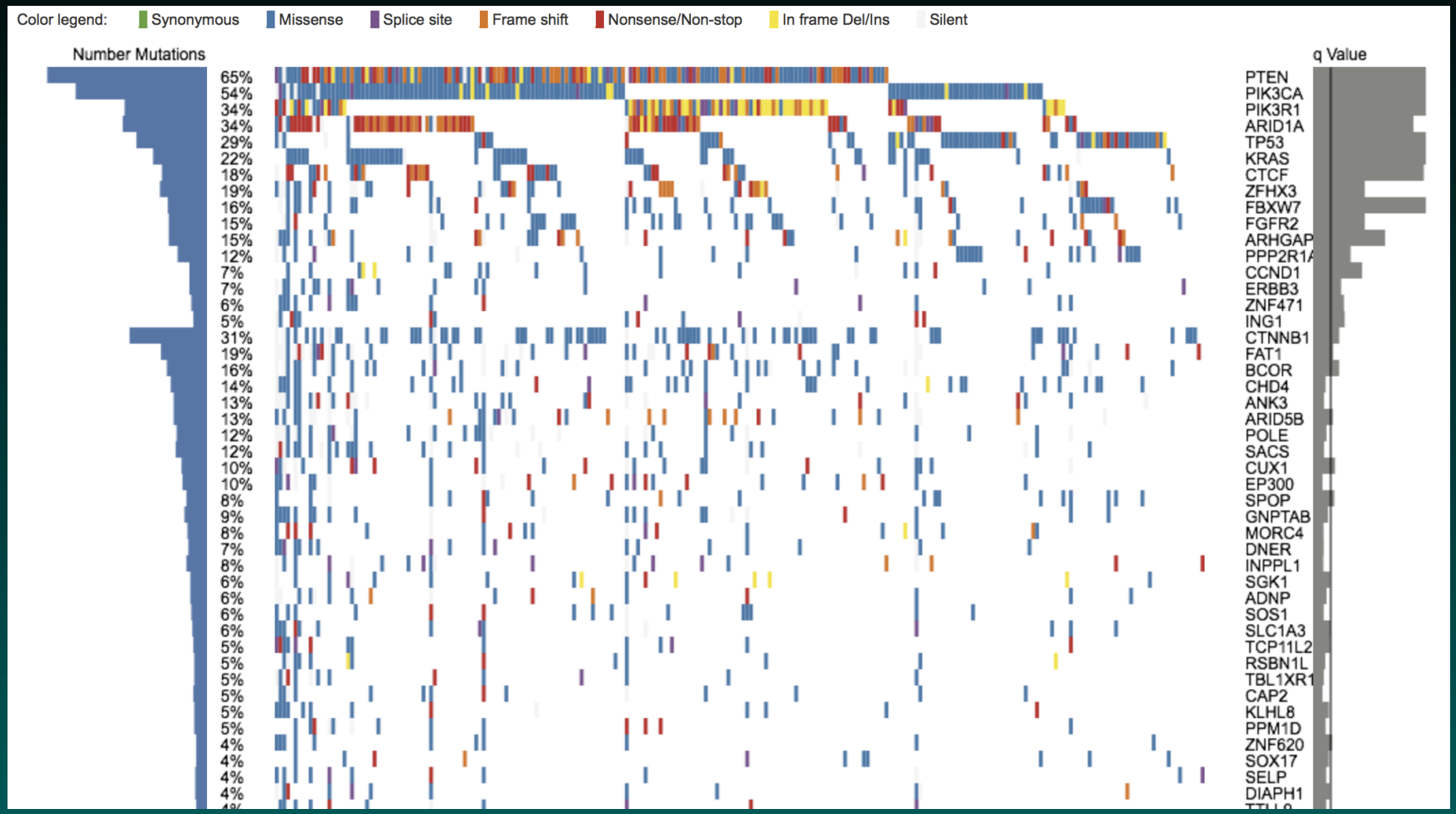


Does this suggest other genes that should be grouped in a pathway with BRAF?

Looking at OV (Ovarian)

What breaks first?

Looking at UCEC (Endometrial)



PTEN in 65%! Actually, this is misleading, as would be similar plots for BRCA and SARC. There are subtypes.

What We're Giving You

We've pulled together all of the

mRNA Expression (log2 intensity, 10K samples)
Copy Number (log2 ratio, 20K samples)
Methylation (fraction meth at a site, 10K samples)

data across TCGA for the 25 most frequently mutated cancer genes (mutations are available from the tumor portal for about 5K samples).

We've also included overall survival information.

We're asking you to look for interactions.

More on TCGA

The Human Genome Project focused on assembling a draft sequence of a *working* human genome, all 3G bases of it. Looking at the number of ways it can *not* work is harder.

The goal of TCGA was to

- assemble several hundred samples of each of several dozen different tumor types,
- profile all of them with a variety of high-throughput assays focusing on different types of genomic measurements, and
- relate this information with the associated clinical data (e.g., type of therapy, survival, etc).

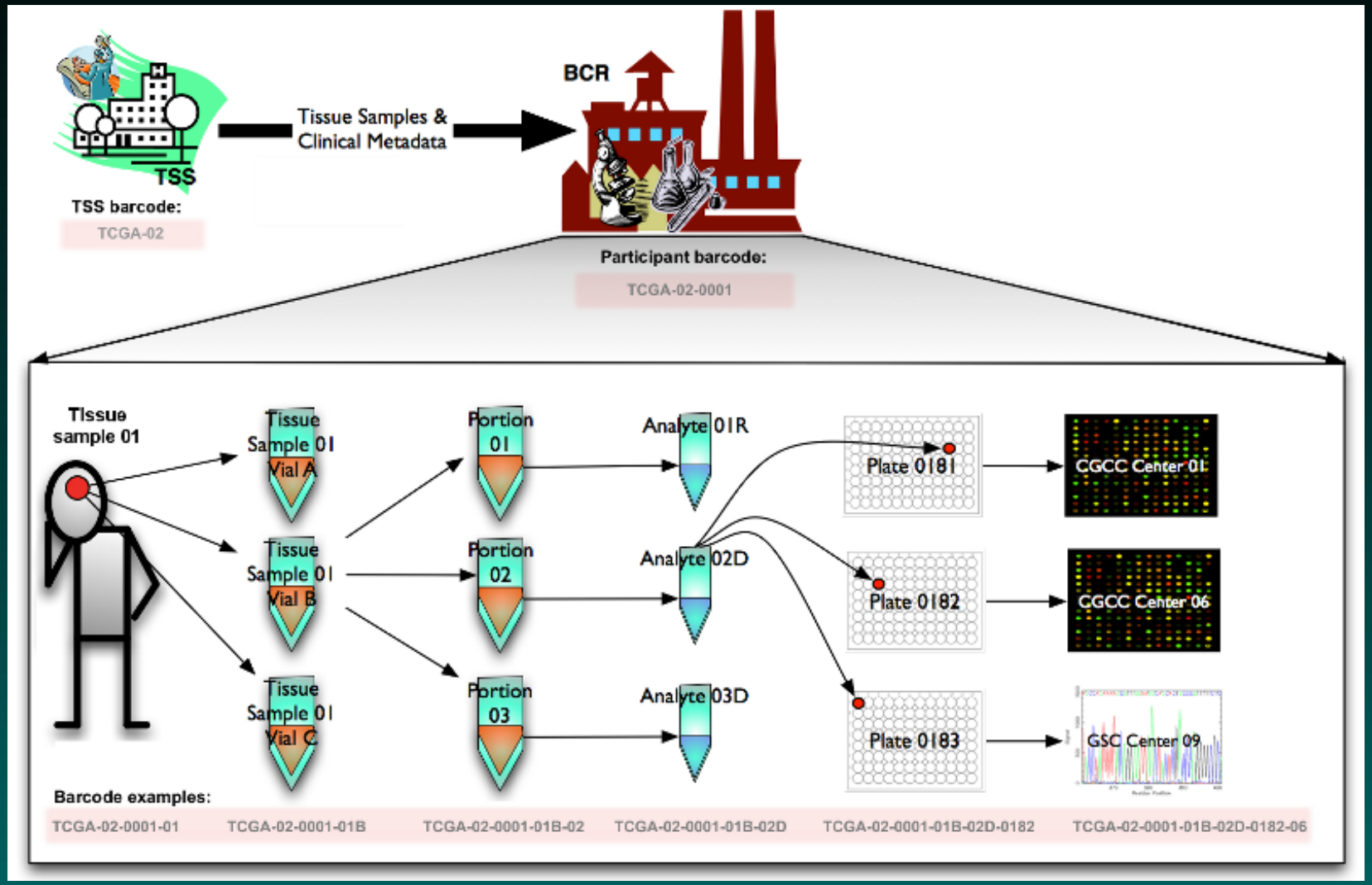
Challenges in Using TCGA Data

The data are sorted by disease type and assay, and then further broken down by the amount of processing employed:

Level 1: “Raw”
Level 2: Some processing (how?)
Level 3: “Final” data for analysis

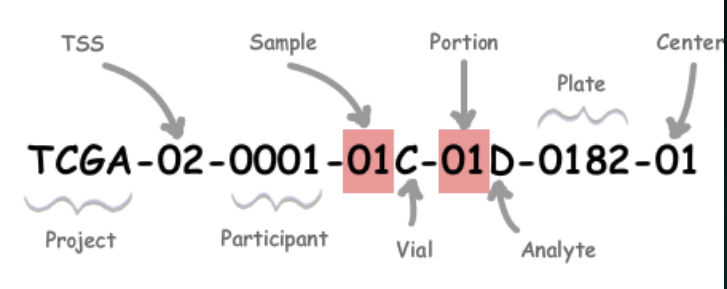
Essentially all Level 3 data is publicly available, but much of the Level 1 data is controlled access due to HIPAA concerns.

Challenges in Using TCGA Data



Sample Naming and Data Assembly

Barcodes



Code Tables Report

Sample Type

Select a Code Table: Sample type | Export Data | Help

Code	Definition	Short Letter Code
01	Primary solid Tumor	TP
02	Recurrent Solid Tumor	TR
03	Primary Blood Derived Cancer - Peripheral Blood	TB
04	Recurrent Blood Derived Cancer - Bone Marrow	TRBM
05	Additional - New Primary	TAP
06	Metastatic	TM
07	Additional Metastatic	TAM
08	Human Tumor Original Cells	THOC
09	Primary Blood Derived Cancer - Bone Marrow	TBM
10	Blood Derived Normal	NB
11	Solid Tissue Normal	NT
12	Buccal Cell Normal	NBC
13	EBV Immortalized Normal	NEBV
14	Bone Marrow Normal	NBM
20	Control Analyte	CELLC
40	Recurrent Blood Derived Cancer - Peripheral Blood	TRB
50	Cell Lines	CELL
60	Primary Xenograft Tissue	XP

What Questions Can TCGA Answer?

What breaks? How?

What tumor subtypes are there?

What genes are linked?

What gene changes are prognostic?

When should some therapies be tested?

...

What can you think of?

Where is RR important here?