Frederick National Laboratory
for Cancer Research

sponsored by the National Cancer Institute

Cancer Type/Site Classification using Deep-Learning (Preliminary presentation slides)

S. Ravichandran
BIDS, FNLCR

(in preparation)

DEPARTMENT OF HEALTH AND HUMAN SERVICES • National Institutes of Health • National Cancer Institute

Frederick National Laboratory is a Federally Funded Research and Development Center operated by Leidos Biomedical Research, Inc., for the National Cancer Institute

Acknowledgements

- **NCI-DOE Pilot-1 Team**
  - Maulik Shukla

- **BIDS**
  - Drs. George Zaki, Andrew Weissman, Mark Jensen and Eric Stahlberg
  - Amar Khalsa, Dr. Deb Hope
  - Colleagues who reviewed the material

Feel free to follow-along

CBIIT

- https://cbiit.github.io/sdsi/workshops (landing site; creation in progress)

Github

- https://github.com/ravichas/ML-TC1 (in progress)

Introduction

- This is part of the NCI-DOE knowledge/capability transfer efforts

- Share tools/techniques/solutions for cancer related problems. We often take a test-case and show how it works

- You will be able to take the test-case (code/scripts) and tune it to your needs

Motivation: Cancer Prediction vs Cancer Detection

- **Cancer Prediction has been the major focus**
  - Prognosis, Recurrence, Susceptibility

- **Cancer Detection (classification of tumors/cancers) is lagging behind Prediction and we would like to share an application that might be useful**
  - Detect/Identify cancer type at an early stage

Goal(s)/Questions

- **Take unstructured genomic expression data from tumor/cancer samples and apply Deep-Learning to create Cancer types/site(s) classifier models**

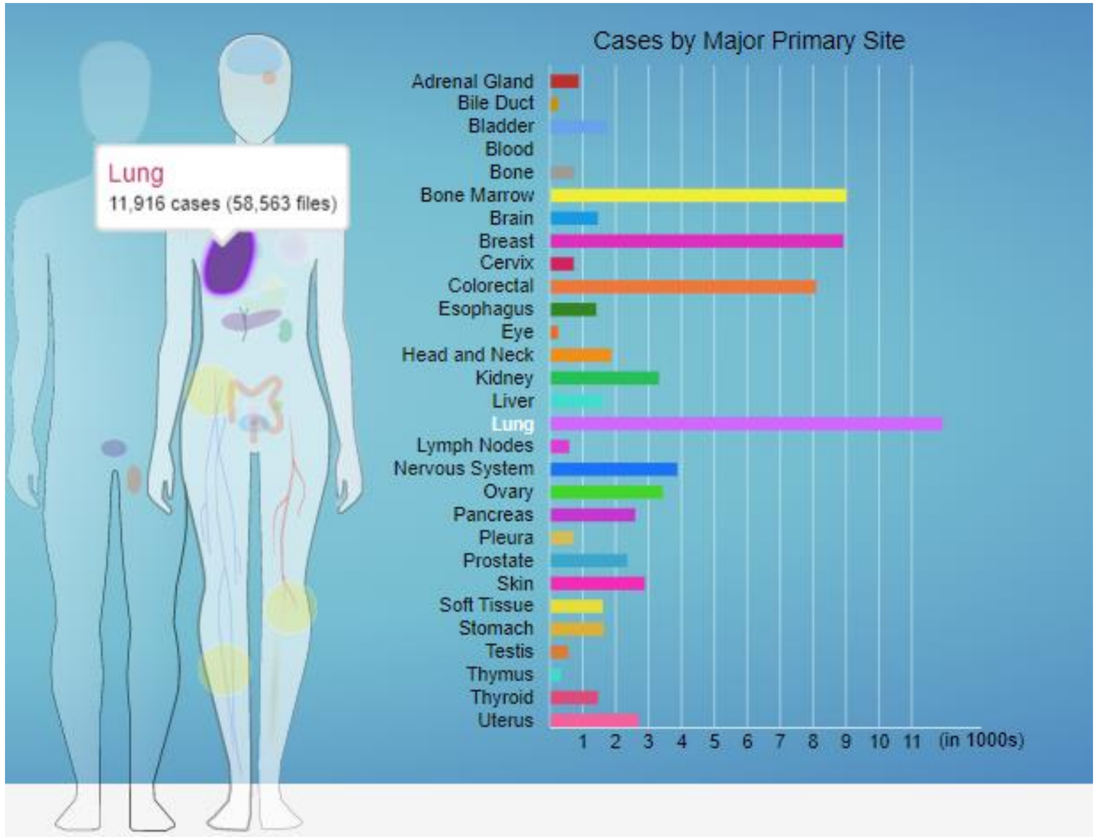
- **Are the expression profiles unique?**

- **Can we use the model as early cancer type detection**
  - Improving chance of early detection cure/survival?

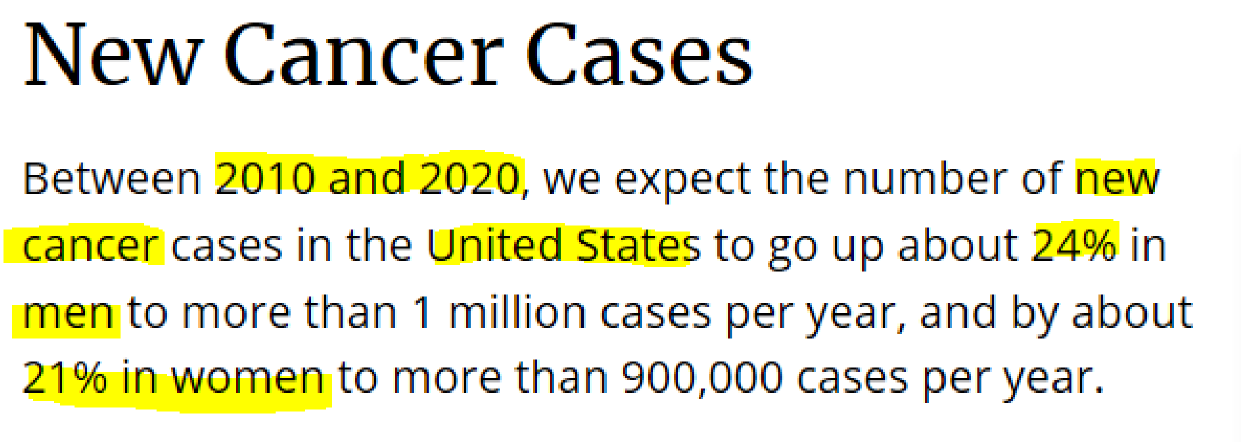
Cancer Burden

- **Cancer is a group of diseases and world-wide risk**
- **Acquired or somatic changes causes 90-95% of caner (all types)**
  - *Source TCGA*
- **~ 200 forms of cancer**
  - *DOI: 10.5114/wo.2014.47136*
- **For 2020**
  - ~1.8M new cancer cases are expected
  - ~600K deaths will occur

Figure from Genomic Data Commons



Expected New Cases/Deaths in 2020



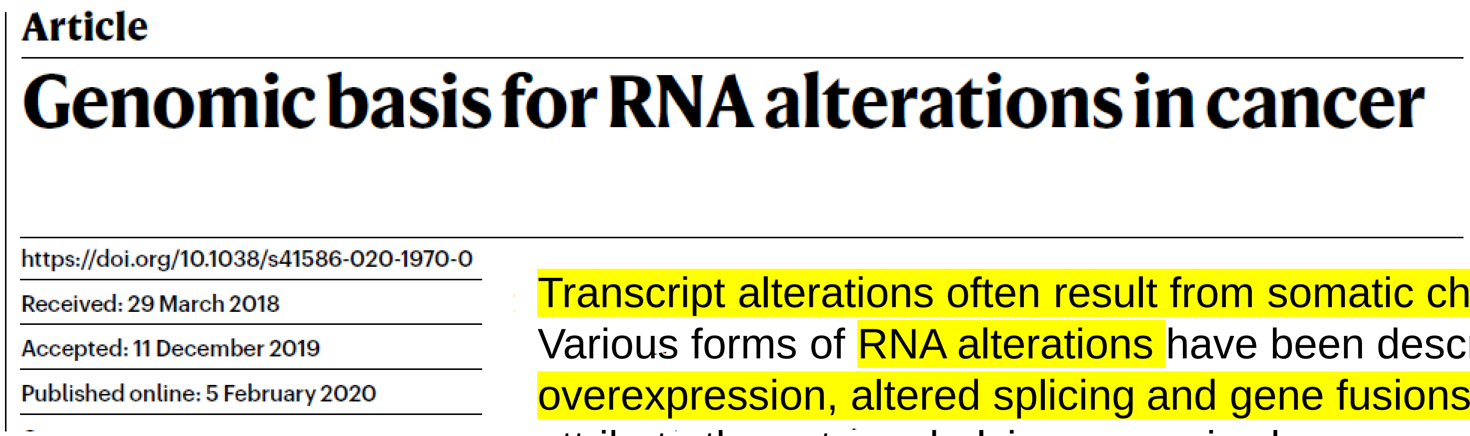
New Cancer Cases

Between 2010 and 2020, we expect the number of new cancer cases in the United States to go up about 24% in men to more than 1 million cases per year, and by about 21% in women to more than 900,000 cases per year.

US population gender	Cancers that are expected to increase
Men	Prostrate, Kidney, Liver and Bladder
Women	Lung, Breast, Uterine and Thyroid

Dynamic genomic changes result in Cancer

Cancer Genome (changes) → Transcript alterations



Article

Genomic basis for RNA alterations in cancer

https://doi.org/10.1038/s41586-020-1970-0

Received: 29 March 2018

Accepted: 11 December 2019

Published online: 5 February 2020

Transcript alterations often result from somatic changes in cancer genomes. Various forms of RNA alterations have been described in cancer, including overexpression, altered splicing and gene fusions; however, it is difficult to attribute these to underlying genomic changes owing to heterogeneity among patients and tumor types, and the relatively small cohorts of patients for whom samples have been analyzed by both transcriptome and whole-genome sequencing.

Expression changes in oncogenes; What type of changes?

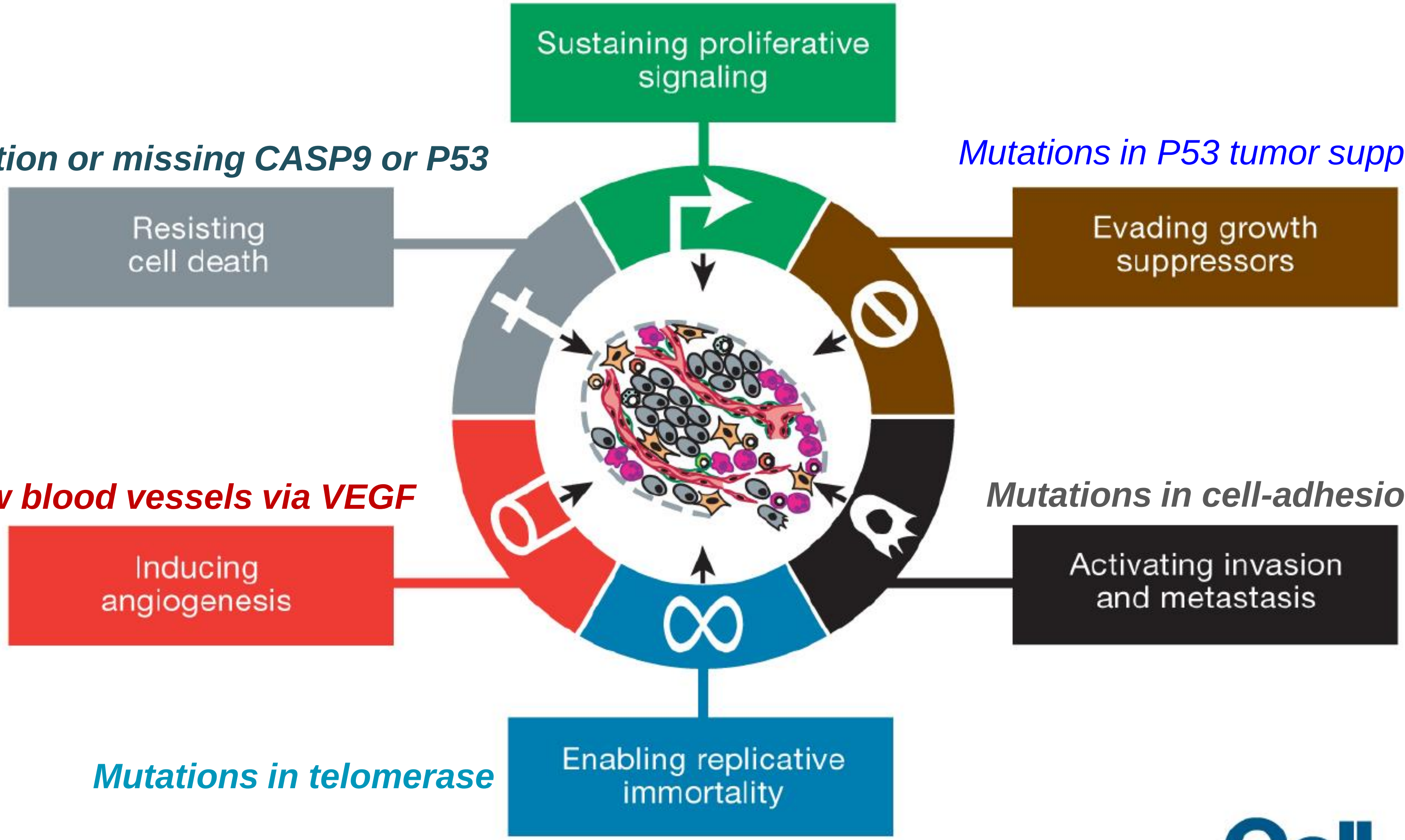
Hallmarks of cancer: Acquired capabilities (mutations) that drive cancer

Hallmarks of Cancer: The Next Generation

src Oncogene mutations

Mutation or missing CASP9 or P53

Mutations in P53 tumor suppressor


Sustaining proliferative signaling
Resisting cell death
Evading growth suppressors
Inducing angiogenesis
Activating invasion and metastasis
Enabling replicative immortality

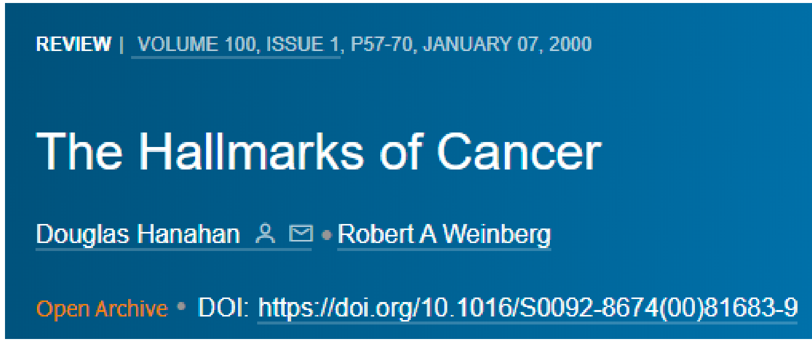
New blood vessels via VEGF

Mutations in cell-adhesion CDH1

Mutations in telomerase

Hanahan and Weinberg, 2011

Cell PRESS


REVIEW | VOLUME 100, ISSUE 1, P57-70, JANUARY 07, 2000

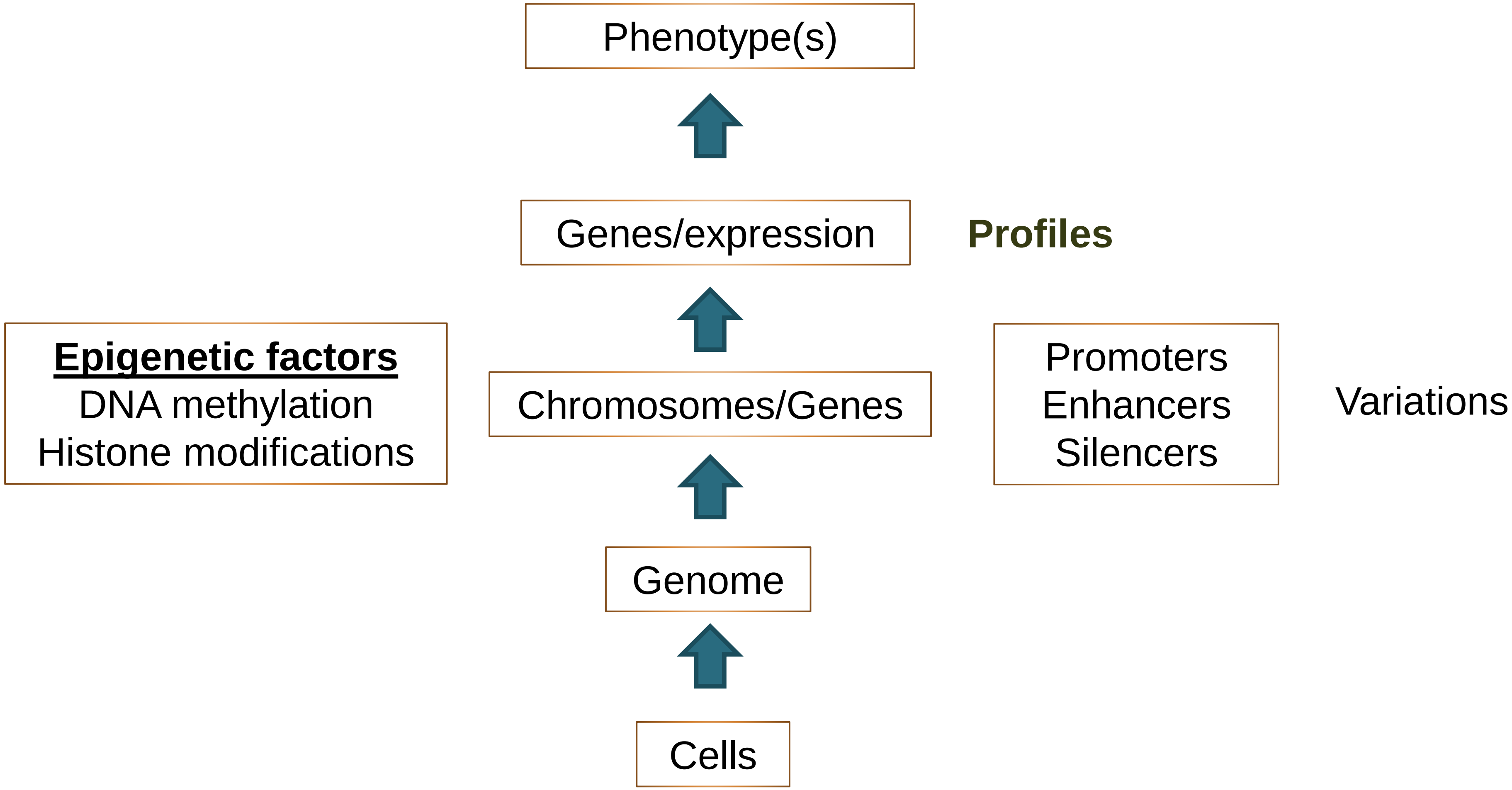
The Hallmarks of Cancer

Douglas Hanahan • Robert A Weinberg

Open Archive • DOI: https://doi.org/10.1016/S0092-8674(00)81683-9

Expression changes in oncogenes; Six capabilities; Overview of Genotype/phenotypes? ➡

Influence of genomic features on phenotypes: An overview



Phenotype(s)

Genes/expression

Profiles

Epigenetic factors
DNA methylation
Histone modifications

Chromosomes/Genes

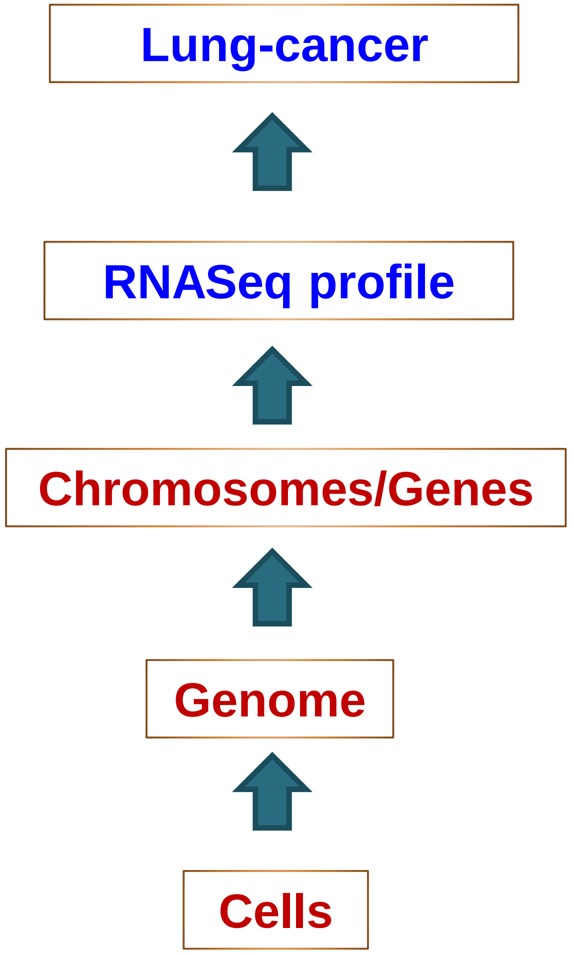
Promoters
Enhancers
Silencers

Variations

Genome

Cells

Influence of genomic features on phenotypes: An overview



Lung-cancer

RNASeq profile

Chromosomes/Genes

Genome

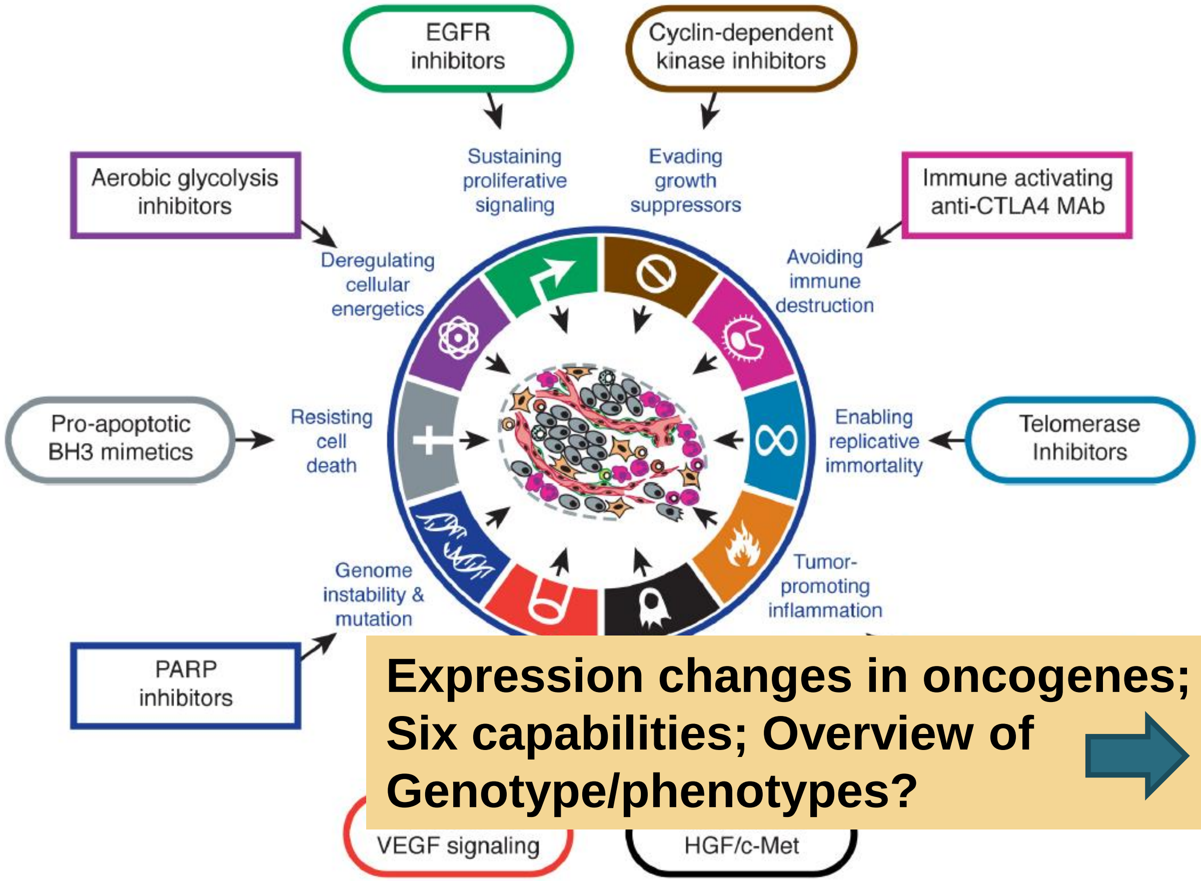
Cells

Diagnosis/treatment vs Prediction

Treatment vs Type-Prediction

- **Treatment**
  - Gene-centric (or a slice of pathway)
  - Imatinib targeting BCR/KIT

- **Detecting Type**
  - *“The architecture of occurring genetic aberrations such as somatic mutations, CNVs, changed gene expression profiles, and different epigenetic alterations, is unique for each type of cancer.”, DOI: 10.5114/wo.2014.47136*
  - Complex
  - Multi-gene centric


EGFR inhibitors
Cyclin-dependent kinase inhibitors
Aerobic glycolysis inhibitors
Sustaining proliferative signaling
Evading growth suppressors
Immune activating anti-CTLA4 MAb
Deregulating cellular energetics
Avoiding immune destruction
Pro-apoptotic BH3 mimetics
Resisting cell death
Enabling replicative immortality
Telomerase Inhibitors
Genome instability & mutation
Tumor-promoting inflammation
PARP inhibitors
VEGF signaling
HGF/c-Met

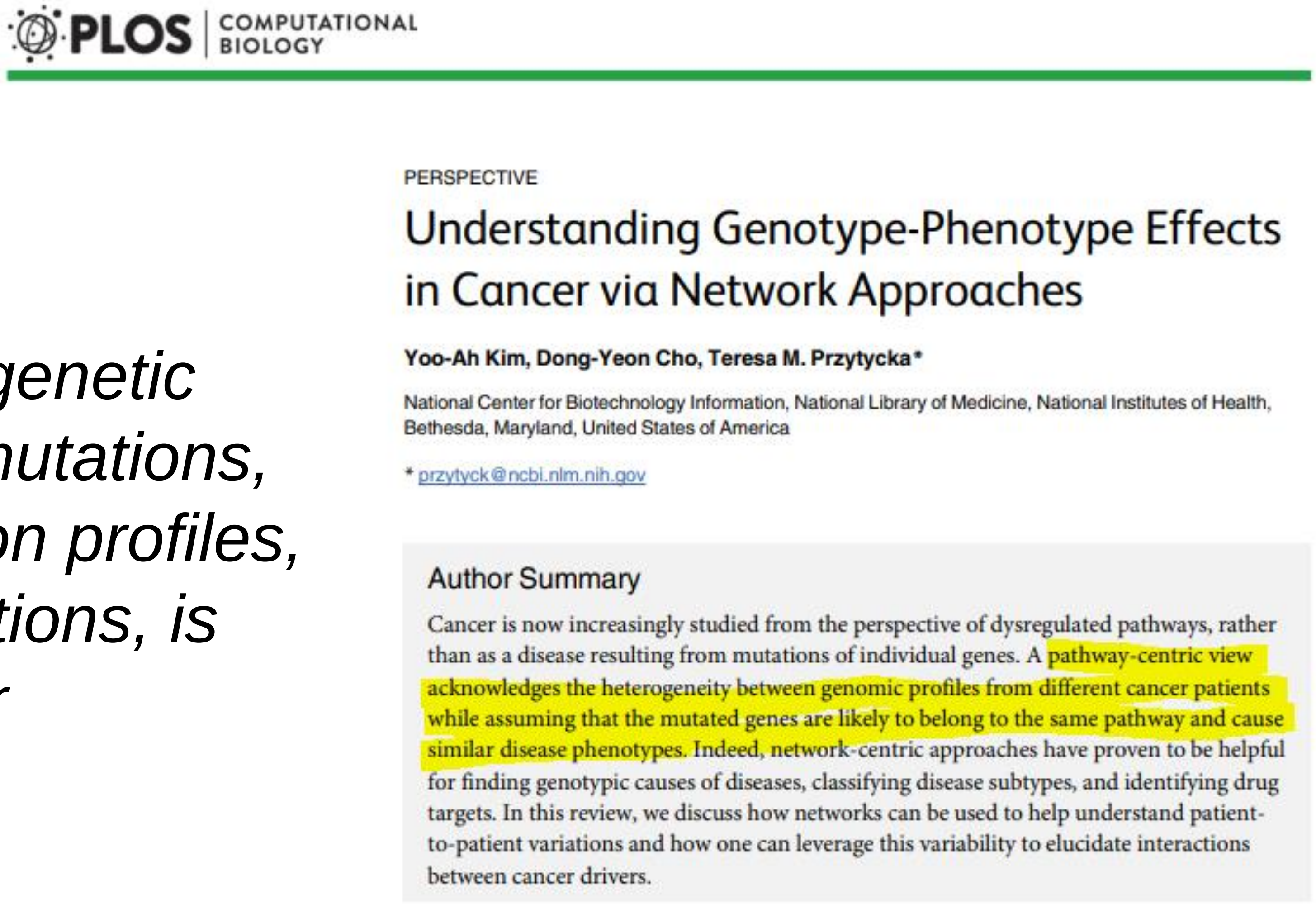
Expression changes in oncogenes; Six capabilities; Overview of Genotype/phenotypes?

Hanahan and Weinberg, 2011

Type-Prediction

The architecture of occurring genetic aberrations such as somatic mutations, CNV, changed gene expression profiles, and different epigenetic alterations, is unique for each type of cancer

DOI: 10.5114/wo.2014.47136



PLOS | COMPUTATIONAL BIOLOGY

PERSPECTIVE

Understanding Genotype-Phenotype Effects in Cancer via Network Approaches

Yoo-Ah Kim, Dong-Yeon Cho, Teresa M. Przytycka*

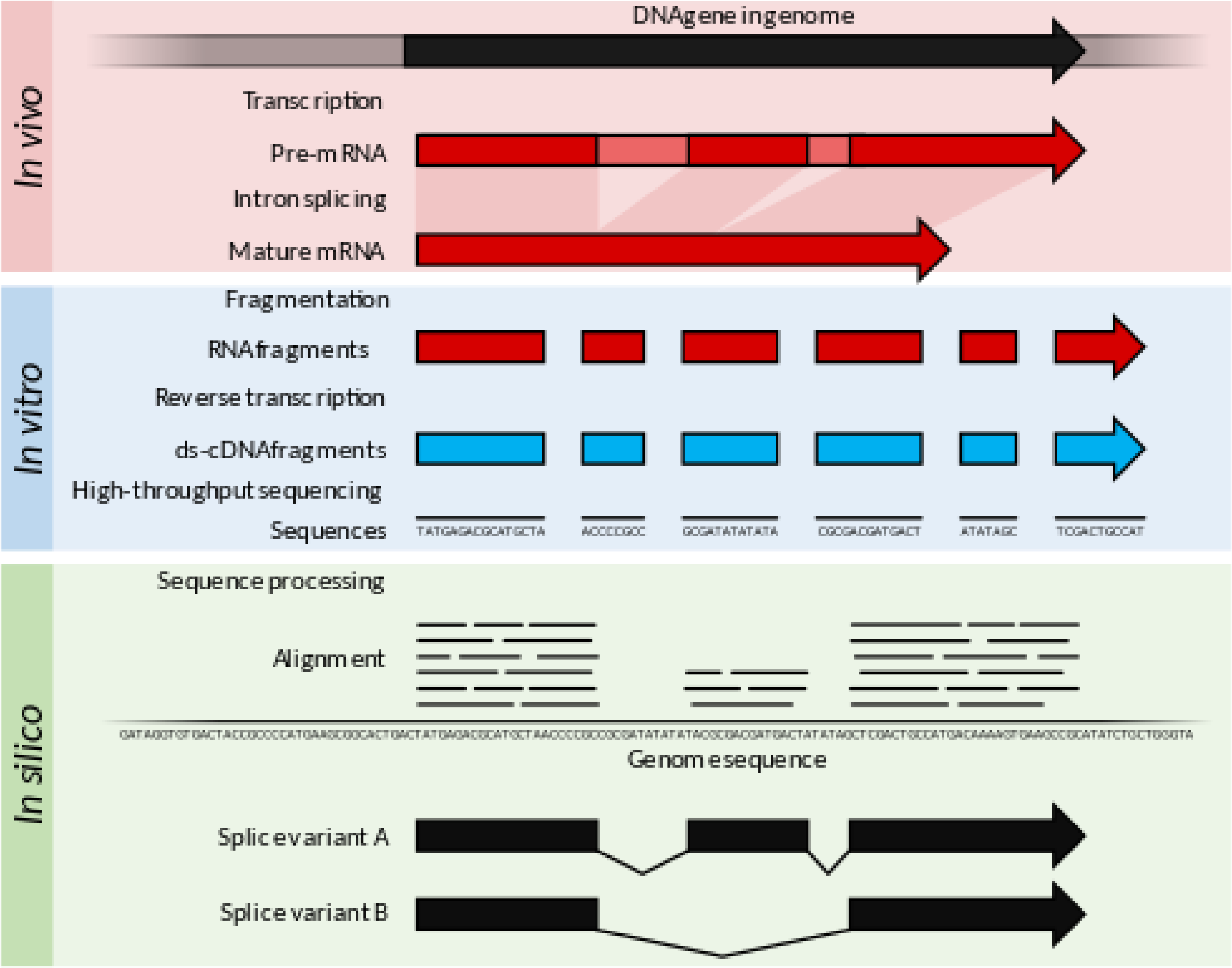
National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland, United States of America

* przytyck@ncbi.nlm.nih.gov

Author Summary

Cancer is now increasingly studied from the perspective of dysregulated pathways, rather than as a disease resulting from mutations of individual genes. A pathway-centric view acknowledges the heterogeneity between genomic profiles from different cancer patients while assuming that the mutated genes are likely to belong to the same pathway and cause similar disease phenotypes. Indeed, network-centric approaches have proven to be helpful for finding genotypic causes of diseases, classifying disease subtypes, and identifying drug targets. In this review, we discuss how networks can be used to help understand patient-to-patient variations and how one can leverage this variability to elucidate interactions between cancer drivers.

What kind of data do we need?



NGS

NGS

READS

Data source: The Cancer Genome Atlas (TCGA)

- NIH launched TCGA Pilot Project – a public funded project

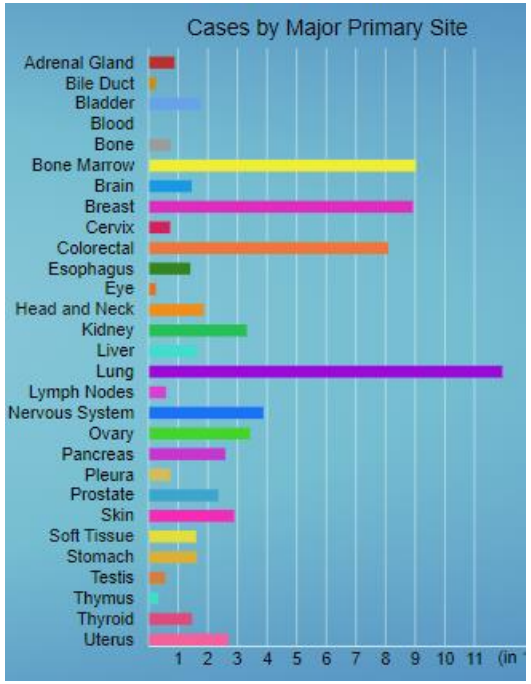
- Goal of creating a comprehensive “atlas” of cancer genomic profiles.

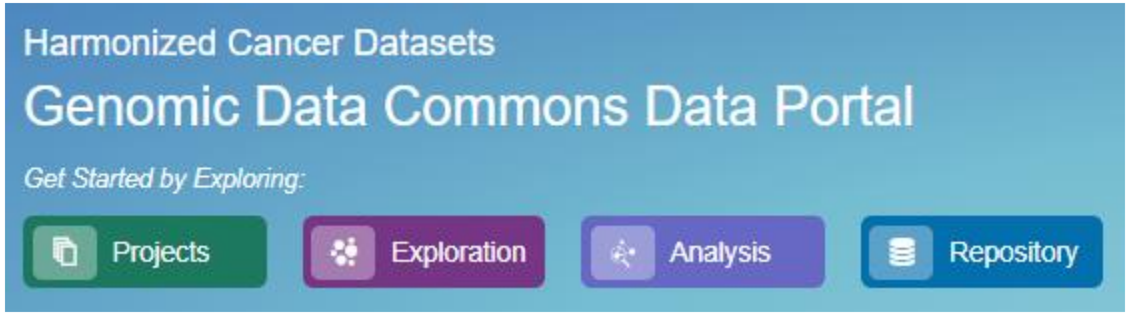
- Large cohorts of over 30 human tumors through large-scale genome sequencing and integrated multi-dimensional analyses.

- Contains Microarray and NGS data
  - RNASeq
  - miRNA seq
  - SNP based platforms
  - …..
- TCGA data is available via GDC

Data Harmonization: GDC

- **Data and metadata is submitted to the GDC in standard data types and file formats. Other data sources (Ex. TCGA) are also included**
- **Data are harmonized against a common reference genome (GRCh38)**
- **For this workshop, we will focus on TCGA Genomic expression data from GDC**




Harmonized Cancer Datasets

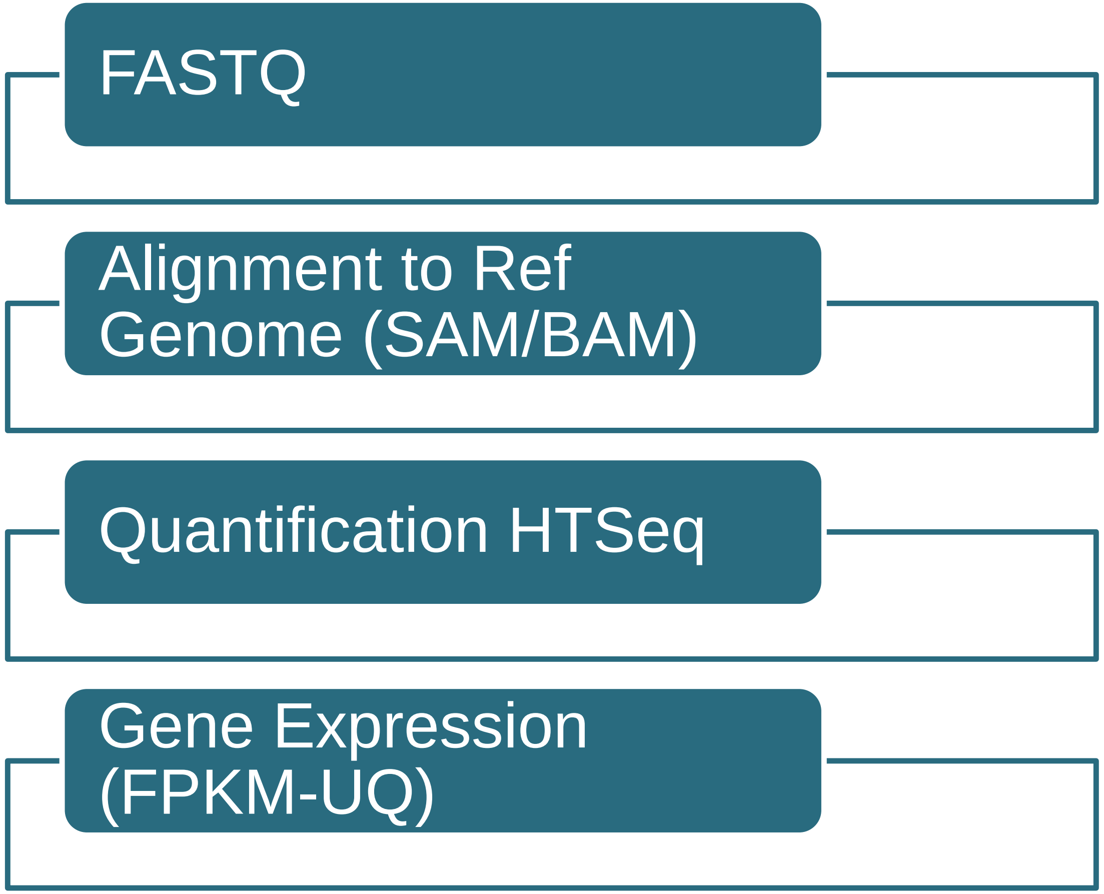
Genomic Data Commons Data Portal

Get Started by Exploring:

Expression Data Quantification

- **RC_g: Number of reads mapped to the gene**
- **RC_{g75}: The 75th percentile read count value for genes in the sample**
- **L: Length of the gene in base pairs; Calculated as the sum of all exons in a gene**

$$\mathrm{FPKM\text{-}UQ} = \frac{RC_g \times 10^9}{RC_{g75} \times L}$$



FASTQ

Alignment to Ref Genome (SAM/BAM)

Quantification HTSeq

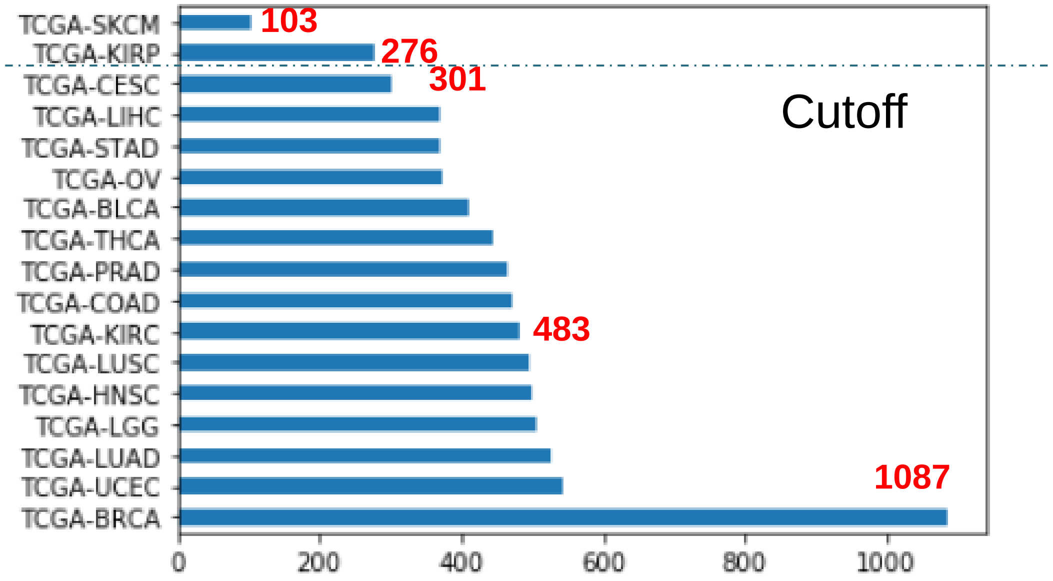
Gene Expression (FPKM-UQ)

Fragments Per Kilobase of transcript per Million mapped reads

How much data for modeling?

Frederick National Laboratory for Cancer Research

sponsored by the National Cancer Institute



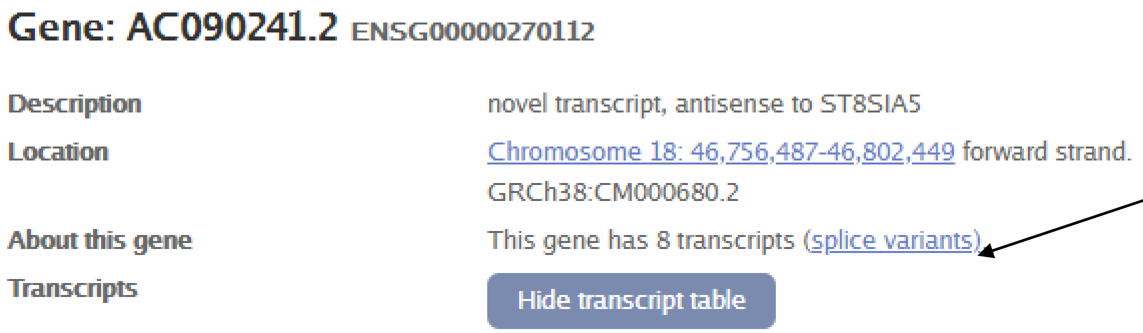
CODE	Cancer Site/Type
BRCA	Breast invasive carcinoma
UCEC	Uterine Corpus Endometrial Carcinoma
LUAD	Lung adenocarcinoma
LGG	Brain Lower Grade Glioma
HNSC	Head and Neck squamous cell carcinoma
LUHSC	Lung squamous cell carcinoma
KIRC	Kidney renal clear cell carcinoma
PRAD	Prostate adenocarcinoma
COAD	Colon adenocarcinoma
THCA	Thyroid carcinoma
BLCA	Bladder Urothelial Carcinoma
OV	Ovarian serous cystadenocarcinoma
STAD	Stomach adenocarcinoma
LIHC	Liver hepatocellular carcinoma
CEC	Cervical squamous cell carcinoma and endocervical adenocarcinoma

300 samples each

Expression data from a sample

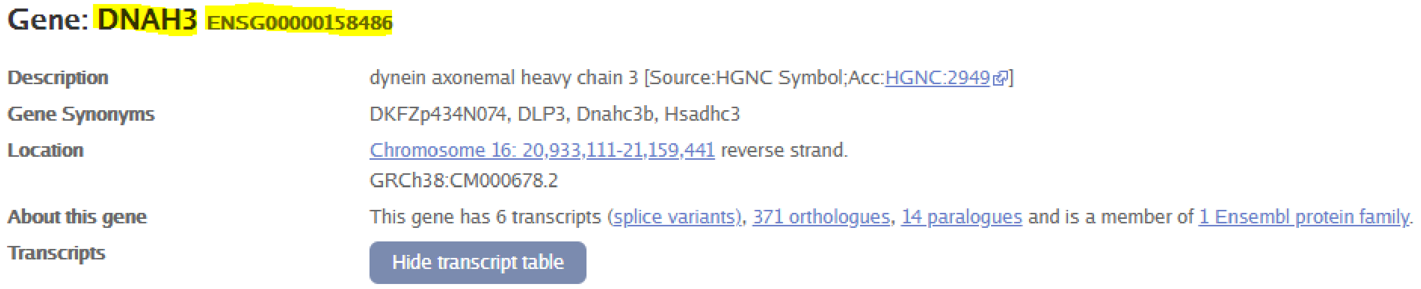
Frederick National Laboratory for Cancer Research

sponsored by the National Cancer Institute



Gene: AC090241.2 ENSG00000270112

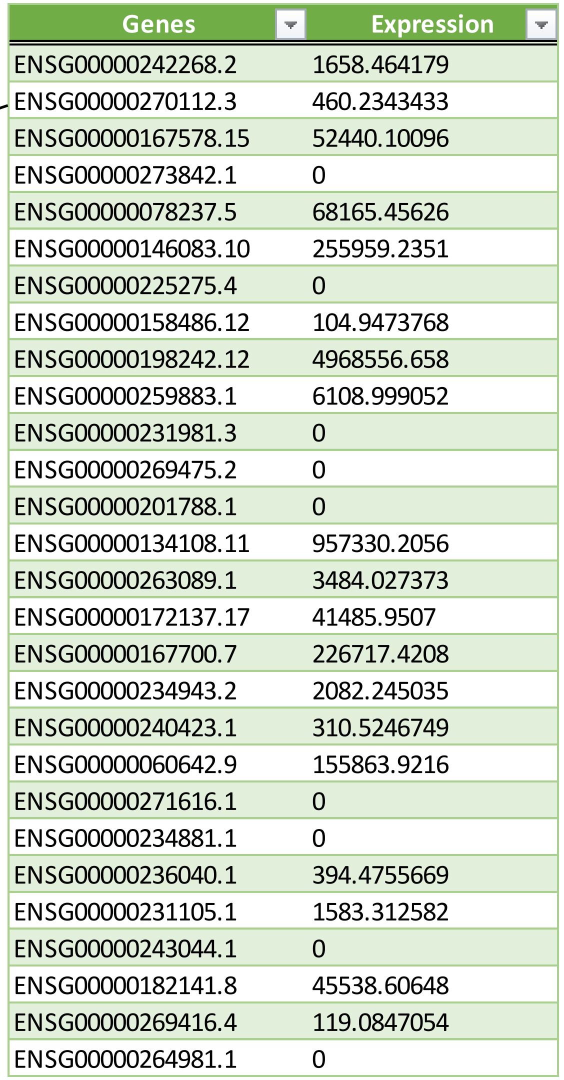
Description	novel transcript, antisense to ST8SIA5
Location	Chromosome 18: 46,756,487-46,802,449 forward strand. GRCh38:CM000680.2
About this gene	This gene has 8 transcripts (splice variants)
Transcripts	Hide transcript table



Gene: DNAH3 ENSG00000158486

Description	dynein axonemal heavy chain 3 [Source:HGNC Symbol;Acc:HGNC:2949]
Gene Synonyms	DKFZp434N074, DLP3, Dnahc3b, Hsadhc3
Location	Chromosome 16: 20,933,111-21,159,441 reverse strand. GRCh38:CM000678.2
About this gene	This gene has 6 transcripts (splice variants), 371 orthologues, 14 paralogues and is a member of 1 Ensembl protein family.
Transcripts	Hide transcript table

TCGA-BRCA



Genes	Expression
ENSG00000242268.2	1658.464179
ENSG00000270112.3	460.2343433
ENSG00000167578.15	52440.10096
ENSG00000273842.1	0
ENSG00000078237.5	68165.45626
ENSG00000146083.10	255959.2351
ENSG00000225275.4	0
ENSG00000158486.12	104.9473768
ENSG00000198242.12	4968556.658
ENSG00000259883.1	6108.999052
ENSG00000231981.3	0
ENSG00000269475.2	0
ENSG00000201788.1	0
ENSG00000134108.11	957330.2056
ENSG00000263089.1	3484.027373
ENSG00000172137.17	41485.9507
ENSG00000167700.7	226717.4208
ENSG00000234943.2	2082.245035
ENSG00000240423.1	310.5246749
ENSG00000060642.9	155863.9216
ENSG00000271616.1	0
ENSG00000234881.1	0
ENSG00000236040.1	394.4755669
ENSG00000231105.1	1583.312582
ENSG00000243044.1	0
ENSG00000182141.8	45538.60648
ENSG00000269416.4	119.0847054
ENSG00000264981.1	0

60,483 transcripts

Data Preparation

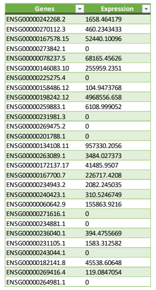
Frederick National Laboratory for Cancer Research

sponsored by the National Cancer Institute

Breast Cancer

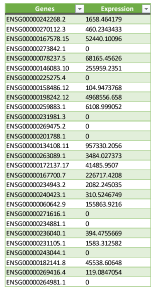
60,484 transcripts

Sample1



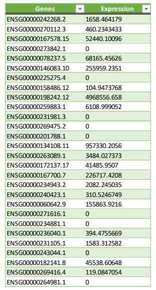
Genes	Expression
ENSG00000242268.2	1658.464179
ENSG00000270112.3	460.2343433
ENSG00000167578.15	52440.10096
ENSG00000273842.1	0
ENSG00000078237.5	68165.45626
ENSG00000146083.10	255959.2351
ENSG00000225275.4	0
ENSG00000158486.12	104.9473768
ENSG00000198242.12	4968556.658
ENSG00000259883.1	6108.999052
ENSG00000231981.3	0
ENSG00000269475.2	0
ENSG00000201788.1	0
ENSG00000134108.11	957330.2056
ENSG00000263089.1	3484.027373
ENSG00000172137.17	41485.9507
ENSG00000167700.7	226717.4208
ENSG00000234943.2	2082.245035
ENSG00000240423.1	310.5246749
ENSG00000060642.9	155863.9216
ENSG00000271616.1	0
ENSG00000234881.1	0
ENSG00000236040.1	394.4755669
ENSG00000231105.1	1583.312582
ENSG00000243044.1	0
ENSG00000182141.8	45538.60648
ENSG00000269416.4	119.0847054
ENSG00000264981.1	0

Sample2



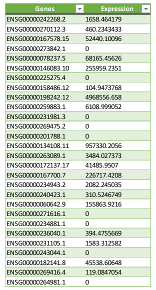
Genes	Expression
ENSG00000242268.2	1658.464179
ENSG00000270112.3	460.2343433
ENSG00000167578.15	52440.10096
ENSG00000273842.1	0
ENSG00000078237.5	68165.45626
ENSG00000146083.10	255959.2351
ENSG00000225275.4	0
ENSG00000158486.12	104.9473768
ENSG00000198242.12	4968556.658
ENSG00000259883.1	6108.999052
ENSG00000231981.3	0
ENSG00000269475.2	0
ENSG00000201788.1	0
ENSG00000134108.11	957330.2056
ENSG00000263089.1	3484.027373
ENSG00000172137.17	41485.9507
ENSG00000167700.7	226717.4208
ENSG00000234943.2	2082.245035
ENSG00000240423.1	310.5246749
ENSG00000060642.9	155863.9216
ENSG00000271616.1	0
ENSG00000234881.1	0
ENSG00000236040.1	394.4755669
ENSG00000231105.1	1583.312582
ENSG00000243044.1	0
ENSG00000182141.8	45538.60648
ENSG00000269416.4	119.0847054
ENSG00000264981.1	0

Sample3



Genes	Expression
ENSG00000242268.2	1658.464179
ENSG00000270112.3	460.2343433
ENSG00000167578.15	52440.10096
ENSG00000273842.1	0
ENSG00000078237.5	68165.45626
ENSG00000146083.10	255959.2351
ENSG00000225275.4	0
ENSG00000158486.12	104.9473768
ENSG00000198242.12	4968556.658
ENSG00000259883.1	6108.999052
ENSG00000231981.3	0
ENSG00000269475.2	0
ENSG00000201788.1	0
ENSG00000134108.11	957330.2056
ENSG00000263089.1	3484.027373
ENSG00000172137.17	41485.9507
ENSG00000167700.7	226717.4208
ENSG00000234943.2	2082.245035
ENSG00000240423.1	310.5246749
ENSG00000060642.9	155863.9216
ENSG00000271616.1	0
ENSG00000234881.1	0
ENSG00000236040.1	394.4755669
ENSG00000231105.1	1583.312582
ENSG00000243044.1	0
ENSG00000182141.8	45538.60648
ENSG00000269416.4	119.0847054
ENSG00000264981.1	0

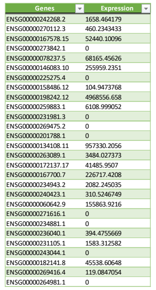
Sample4



Genes	Expression
ENSG00000242268.2	1658.464179
ENSG00000270112.3	460.2343433
ENSG00000167578.15	52440.10096
ENSG00000273842.1	0
ENSG00000078237.5	68165.45626
ENSG00000146083.10	255959.2351
ENSG00000225275.4	0
ENSG00000158486.12	104.9473768
ENSG00000198242.12	4968556.658
ENSG00000259883.1	6108.999052
ENSG00000231981.3	0
ENSG00000269475.2	0
ENSG00000201788.1	0
ENSG00000134108.11	957330.2056
ENSG00000263089.1	3484.027373
ENSG00000172137.17	41485.9507
ENSG00000167700.7	226717.4208
ENSG00000234943.2	2082.245035
ENSG00000240423.1	310.5246749
ENSG00000060642.9	155863.9216
ENSG00000271616.1	0
ENSG00000234881.1	0
ENSG00000236040.1	394.4755669
ENSG00000231105.1	1583.312582
ENSG00000243044.1	0
ENSG00000182141.8	45538.60648
ENSG00000269416.4	119.0847054
ENSG00000264981.1	0

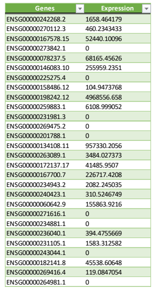
....

Sample297



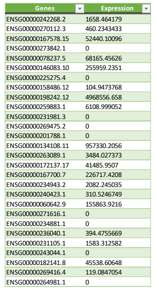
Genes	Expression
ENSG00000242268.2	1658.464179
ENSG00000270112.3	460.2343433
ENSG00000167578.15	52440.10096
ENSG00000273842.1	0
ENSG00000078237.5	68165.45626
ENSG00000146083.10	255959.2351
ENSG00000225275.4	0
ENSG00000158486.12	104.9473768
ENSG00000198242.12	4968556.658
ENSG00000259883.1	6108.999052
ENSG00000231981.3	0
ENSG00000269475.2	0
ENSG00000201788.1	0
ENSG00000134108.11	957330.2056
ENSG00000263089.1	3484.027373
ENSG00000172137.17	41485.9507
ENSG00000167700.7	226717.4208
ENSG00000234943.2	2082.245035
ENSG00000240423.1	310.5246749
ENSG00000060642.9	155863.9216
ENSG00000271616.1	0
ENSG00000234881.1	0
ENSG00000236040.1	394.4755669
ENSG00000231105.1	1583.312582
ENSG00000243044.1	0
ENSG00000182141.8	45538.60648
ENSG00000269416.4	119.0847054
ENSG00000264981.1	0

Sample298



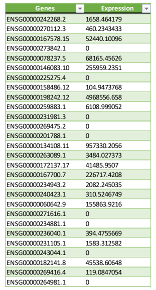
Genes	Expression
ENSG00000242268.2	1658.464179
ENSG00000270112.3	460.2343433
ENSG00000167578.15	52440.10096
ENSG00000273842.1	0
ENSG00000078237.5	68165.45626
ENSG00000146083.10	255959.2351
ENSG00000225275.4	0
ENSG00000158486.12	104.9473768
ENSG00000198242.12	4968556.658
ENSG00000259883.1	6108.999052
ENSG00000231981.3	0
ENSG00000269475.2	0
ENSG00000201788.1	0
ENSG00000134108.11	957330.2056
ENSG00000263089.1	3484.027373
ENSG00000172137.17	41485.9507
ENSG00000167700.7	226717.4208
ENSG00000234943.2	2082.245035
ENSG00000240423.1	310.5246749
ENSG00000060642.9	155863.9216
ENSG00000271616.1	0
ENSG00000234881.1	0
ENSG00000236040.1	394.4755669
ENSG00000231105.1	1583.312582
ENSG00000243044.1	0
ENSG00000182141.8	45538.60648
ENSG00000269416.4	119.0847054
ENSG00000264981.1	0

Sample299



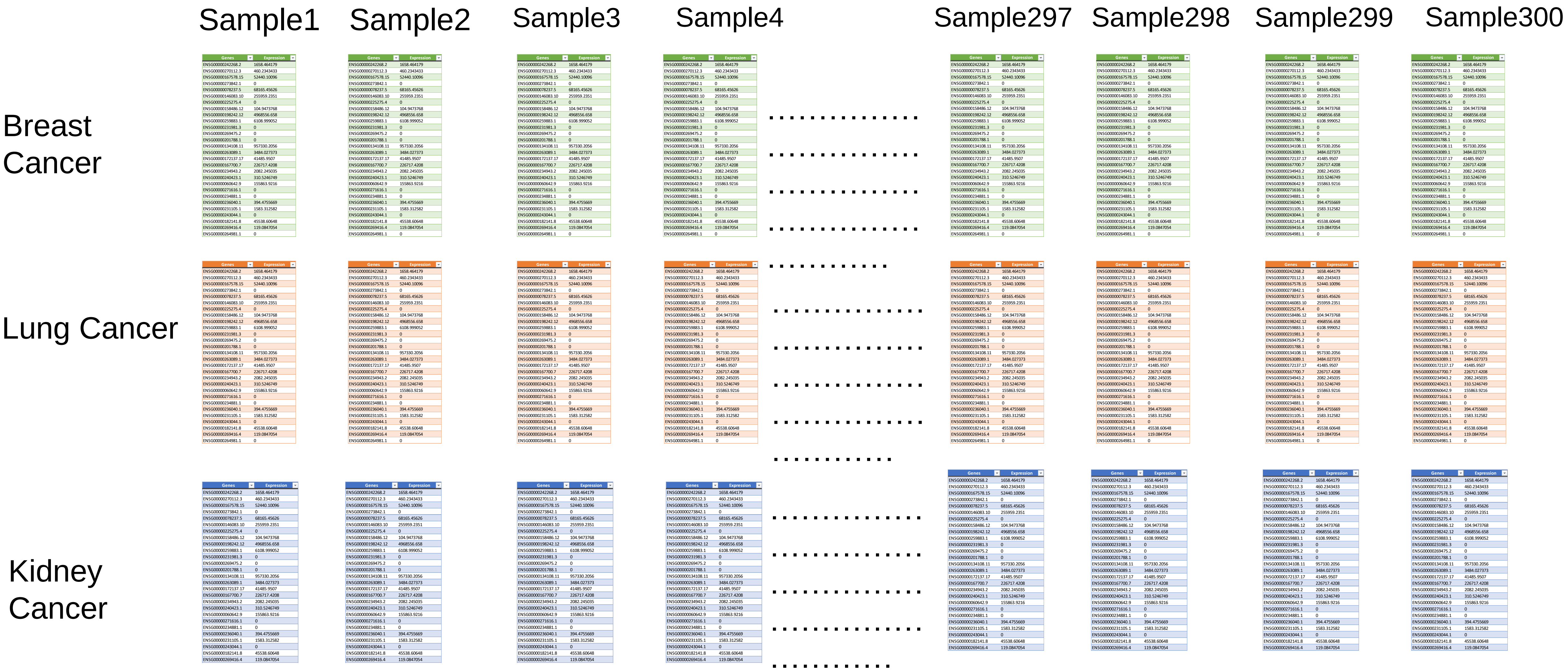
Genes	Expression
ENSG00000242268.2	1658.464179
ENSG00000270112.3	460.2343433
ENSG00000167578.15	52440.10096
ENSG00000273842.1	0
ENSG00000078237.5	68165.45626
ENSG00000146083.10	255959.2351
ENSG00000225275.4	0
ENSG00000158486.12	104.9473768
ENSG00000198242.12	4968556.658
ENSG00000259883.1	6108.999052
ENSG00000231981.3	0
ENSG00000269475.2	0
ENSG00000201788.1	0
ENSG00000134108.11	957330.2056
ENSG00000263089.1	3484.027373
ENSG00000172137.17	41485.9507
ENSG00000167700.7	226717.4208
ENSG00000234943.2	2082.245035
ENSG00000240423.1	310.5246749
ENSG00000060642.9	155863.9216
ENSG00000271616.1	0
ENSG00000234881.1	0
ENSG00000236040.1	394.4755669
ENSG00000231105.1	1583.312582
ENSG00000243044.1	0
ENSG00000182141.8	45538.60648
ENSG00000269416.4	119.0847054
ENSG00000264981.1	0

Sample300



Genes	Expression
ENSG00000242268.2	1658.464179
ENSG00000270112.3	460.2343433
ENSG00000167578.15	52440.10096
ENSG00000273842.1	0
ENSG00000078237.5	68165.45626
ENSG00000146083.10	255959.2351
ENSG00000225275.4	0
ENSG00000158486.12	104.9473768
ENSG00000198242.12	4968556.658
ENSG00000259883.1	6108.999052
ENSG00000231981.3	0
ENSG00000269475.2	0
ENSG00000201788.1	0
ENSG00000134108.11	957330.2056
ENSG00000263089.1	3484.027373
ENSG00000172137.17	41485.9507
ENSG00000167700.7	226717.4208
ENSG00000234943.2	2082.245035
ENSG00000240423.1	310.5246749
ENSG00000060642.9	155863.9216
ENSG00000271616.1	0
ENSG00000234881.1	0
ENSG00000236040.1	394.4755669
ENSG00000231105.1	1583.312582
ENSG00000243044.1	0
ENSG00000182141.8	45538.60648
ENSG00000269416.4	119.0847054
ENSG00000264981.1	0

Data Preparation

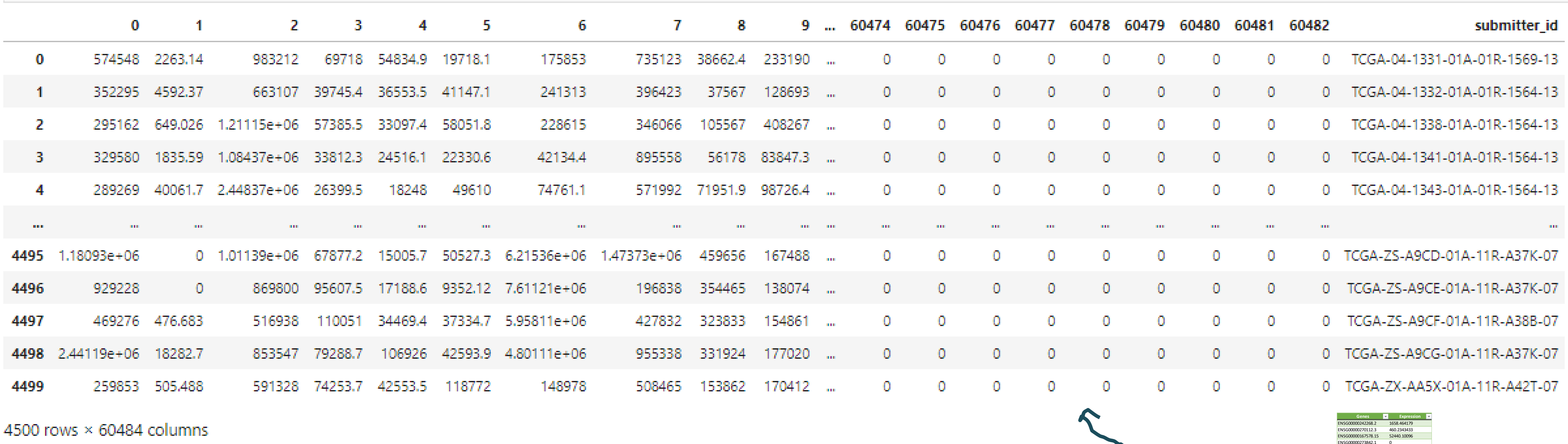

Sample1
Sample2
Sample3
Sample4
Sample297
Sample298
Sample299
Sample300
Breast Cancer
Lung Cancer
Kidney Cancer

Genes	Expression
ENSG00000242268.2	1658.464179
ENSG00000270112.3	460.2343433
ENSG00000167578.15	52440.10096
ENSG00000273842.1	0
ENSG00000078237.5	68165.45626
ENSG00000146083.10	255959.2351
ENSG00000225275.4	0
ENSG00000158486.12	104.9473768
ENSG00000198242.12	4968556.658
ENSG00000259883.1	6108.999052
ENSG00000231981.3	0
ENSG00000269475.2	0
ENSG00000201788.1	0
ENSG00000134108.11	957330.2056
ENSG00000263089.1	3484.027373
ENSG00000172137.17	41485.9507
ENSG00000167700.7	226717.4208
ENSG00000234943.2	2082.245035
ENSG00000240423.1	310.5246749
ENSG00000060642.9	155863.9216
ENSG00000271616.1	0
ENSG00000234881.1	0
ENSG00000236040.1	394.4755669
ENSG00000231105.1	1583.312582
ENSG00000243044.1	0
ENSG00000182141.8	45538.60648
ENSG00000269416.4	119.0847054
ENSG00000264981.1	0

Merged Sample Expression Data

Genes

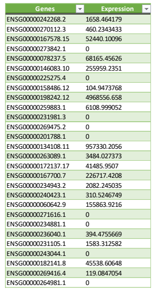
SAMPLES



	0	1	2	3	4	5	6	7	8	9	...	60474	60475	60476	60477	60478	60479	60480	60481	60482	submitter_id
0	574548	2263.14	983212	69718	54834.9	19718.1	175853	735123	38662.4	233190	...	0	0	0	0	0	0	0	0	0	TCGA-04-1331-01A-01R-1569-13
1	352295	4592.37	663107	39745.4	36553.5	41147.1	241313	396423	37567	128693	...	0	0	0	0	0	0	0	0	0	TCGA-04-1332-01A-01R-1564-13
2	295162	649.026	1.21115e+06	57385.5	33097.4	58051.8	228615	346066	105567	408267	...	0	0	0	0	0	0	0	0	0	TCGA-04-1338-01A-01R-1564-13
3	329580	1835.59	1.08437e+06	33812.3	24516.1	22330.6	42134.4	895558	56178	83847.3	...	0	0	0	0	0	0	0	0	0	TCGA-04-1341-01A-01R-1564-13
4	289269	40061.7	2.44837e+06	26399.5	18248	49610	74761.1	571992	71951.9	98726.4	...	0	0	0	0	0	0	0	0	0	TCGA-04-1343-01A-01R-1564-13
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
4495	1.18093e+06	0	1.01139e+06	67877.2	15005.7	50527.3	6.21536e+06	1.47373e+06	459656	167488	...	0	0	0	0	0	0	0	0	0	TCGA-ZS-A9CD-01A-11R-A37K-07
4496	929228	0	869800	95607.5	17188.6	9352.12	7.61121e+06	196838	354465	138074	...	0	0	0	0	0	0	0	0	0	TCGA-ZS-A9CE-01A-11R-A37K-07
4497	469276	476.683	516938	110051	34469.4	37334.7	5.95811e+06	427832	323833	154861	...	0	0	0	0	0	0	0	0	0	TCGA-ZS-A9CF-01A-11R-A38B-07
4498	2.44119e+06	18282.7	853547	79288.7	106926	42593.9	4.80111e+06	955338	331924	177020	...	0	0	0	0	0	0	0	0	0	TCGA-ZS-A9CG-01A-11R-A37K-07
4499	259853	505.488	591328	74253.7	42553.5	118772	148978	508465	153862	170412	...	0	0	0	0	0	0	0	0	0	TCGA-ZX-AA5X-01A-11R-A42T-07

4500 rows × 60484 columns

Transpose and add as a row



Genes	Expression
ENSG00000242268.2	1658.464179
ENSG00000270112.3	460.2343433
ENSG00000167578.15	52440.10096
ENSG00000273842.1	0
ENSG00000078237.5	68165.45626
ENSG00000146083.10	255959.2351
ENSG00000225275.4	0
ENSG00000158486.12	104.9473768
ENSG00000198242.12	4968556.658
ENSG00000259883.1	6108.999052
ENSG00000231981.3	0
ENSG00000269475.2	0
ENSG00000201788.1	0
ENSG00000134108.11	957330.2056
ENSG00000263089.1	3484.027373
ENSG00000172137.17	41485.9507
ENSG00000167700.7	226717.4208
ENSG00000234943.2	2082.245035
ENSG00000240423.1	310.5246749
ENSG00000060642.9	155863.9216
ENSG00000271616.1	0
ENSG00000234881.1	0
ENSG00000236040.1	394.4755669
ENSG00000231105.1	1583.312582
ENSG00000243044.1	0
ENSG00000182141.8	45538.60648
ENSG00000269416.4	119.0847054
ENSG00000264981.1	0

Quantifying mRNA abundance and Scaling

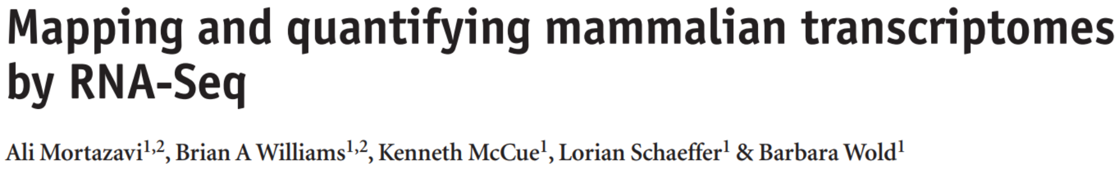
- GDC harmonization data is provided in FPKM-UQ

- In out code, FPKM-UQ is rescaled to TPM using the following formula.

$$\mathrm{TPM}_i = \left(\frac{\mathrm{FPKM}_i}{\Sigma_j \mathrm{FPKM}_j}\right) \cdot 10^6$$

- TPM has nice mathematical properties and a stable entity

https://docs.gdc.cancer.gov/Encyclopedia/pages/HTSeq-FPKM-UQ/

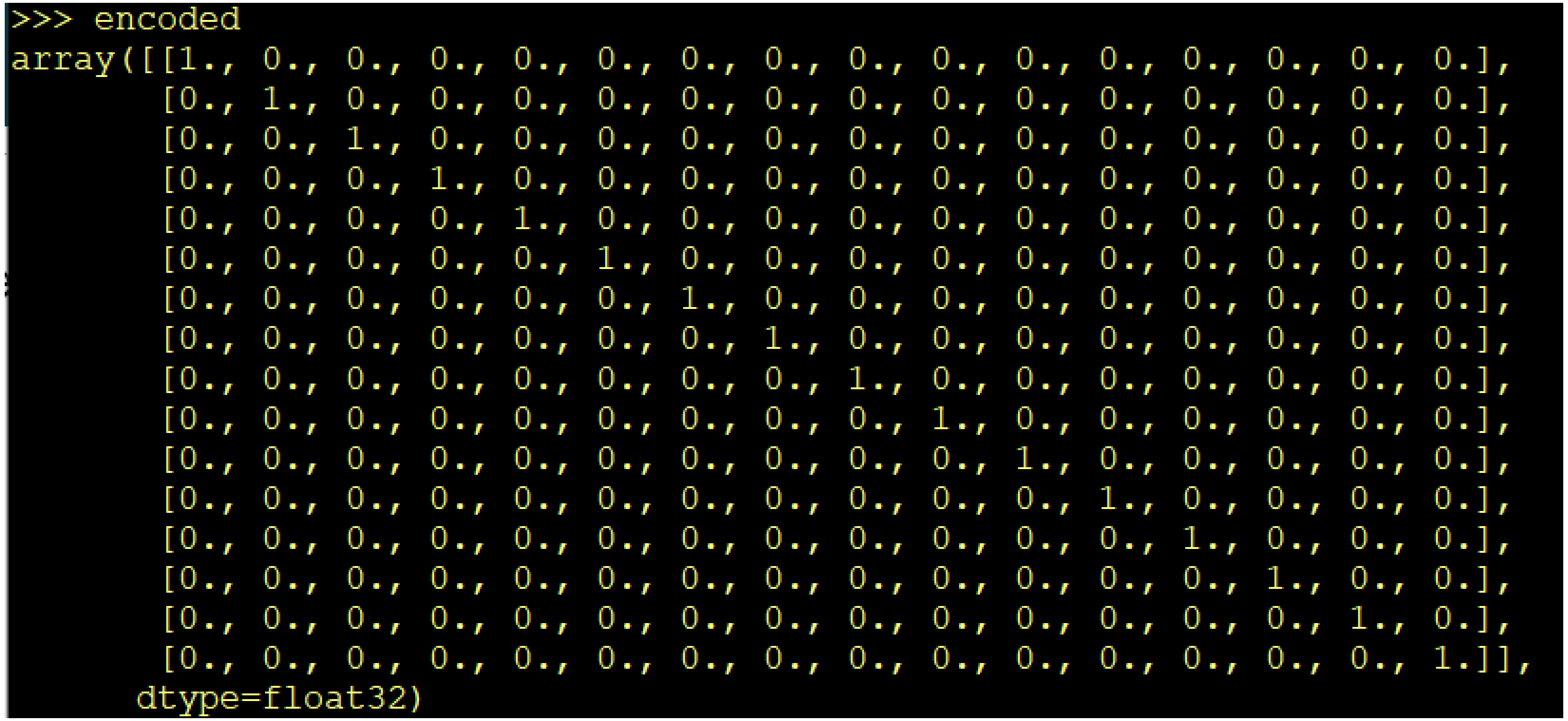


Mapping and quantifying mammalian transcriptomes by RNA-Seq

Ali Mortazavi[1,2], Brian A Williams[1,2], Kenneth McCue[1], Lorian Schaeffer[1] & Barbara Wold[1]

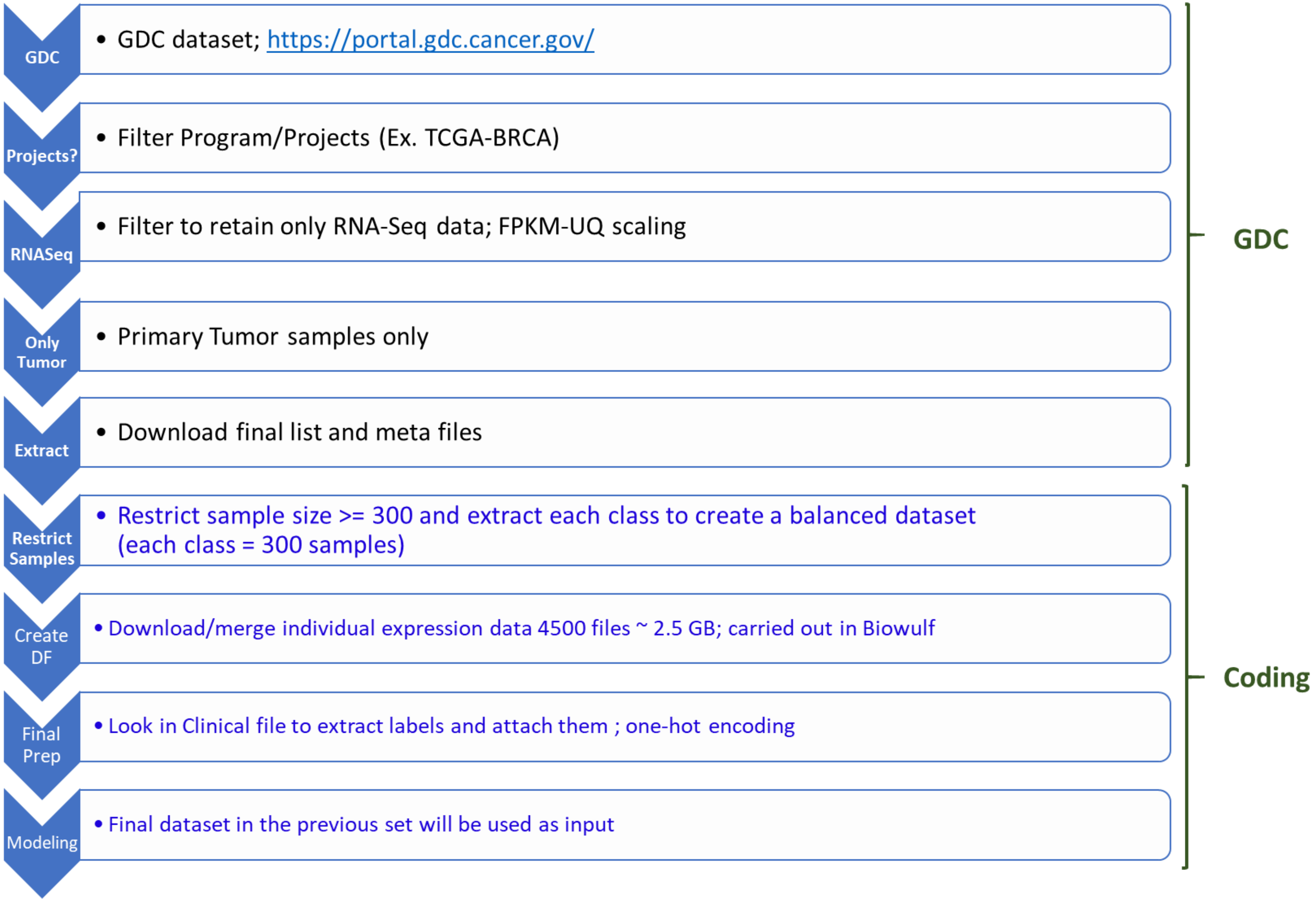
One-hot encoding to convert Cancer types to numbers

- **Convert each class to a numerical quantity**
  - BRCA to 0 ; LUAD to 1 etc.
  - 0, 1, 2, 3, …, 13, 14, 15



```
>>> encoded
array([[1., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0.],
       [0., 1., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0.],
       [0., 0., 1., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0.],
       [0., 0., 0., 1., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0.],
       [0., 0., 0., 0., 1., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0.],
       [0., 0., 0., 0., 0., 1., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0.],
       [0., 0., 0., 0., 0., 0., 1., 0., 0., 0., 0., 0., 0., 0., 0., 0.],
       [0., 0., 0., 0., 0., 0., 0., 1., 0., 0., 0., 0., 0., 0., 0., 0.],
       [0., 0., 0., 0., 0., 0., 0., 0., 1., 0., 0., 0., 0., 0., 0., 0.],
       [0., 0., 0., 0., 0., 0., 0., 0., 0., 1., 0., 0., 0., 0., 0., 0.],
       [0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 1., 0., 0., 0., 0., 0.],
       [0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 1., 0., 0., 0., 0.],
       [0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 1., 0., 0., 0.],
       [0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 1., 0., 0.],
       [0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 1., 0.],
       [0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 1.]],
      dtype=float32)
```

Data preparation steps summary



GDC

- GDC dataset; https://portal.gdc.cancer.gov/

Projects?

- Filter Program/Projects (Ex. TCGA-BRCA)

RNASeq

- Filter to retain only RNA-Seq data; FPKM-UQ scaling

Only Tumor

- Primary Tumor samples only

Extract

- Download final list and meta files

Restrict Samples

- Restrict sample size >= 300 and extract each class to create a balanced dataset (each class = 300 samples)

Create DF

- Download/merge individual expression data 4500 files ~ 2.5 GB; carried out in Biowulf

Final Prep

- Look in Clinical file to extract labels and attach them ; one-hot encoding

Modeling

- Final dataset in the previous set will be used as input

GDC

Coding

Before we break for hands-on

- **Python as the programming language for this workshop, but similar libraries are available in R or other languages**
- **Will use Jupyter Notebook for sharing the code**
  - With little effort one can convert the Python code into R and still use Jupyter Notebook

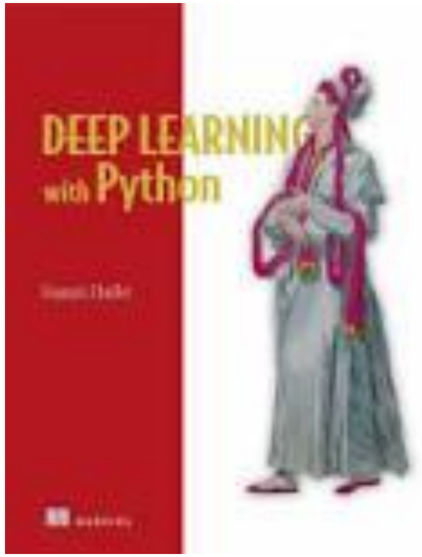
To be continued after Code-Review

https://github.com/ravichas/ML-TC1

Before we break for hands-on

- **Due to lack of time, I wont be covering the basics of Neural Network**

O'REILLY
Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow
Aurélien Géron



- **Following two are good books for beginners and up**

Convolutional Neural Networks

- **Preparation in progress**

Thanks

S. Ravichandran