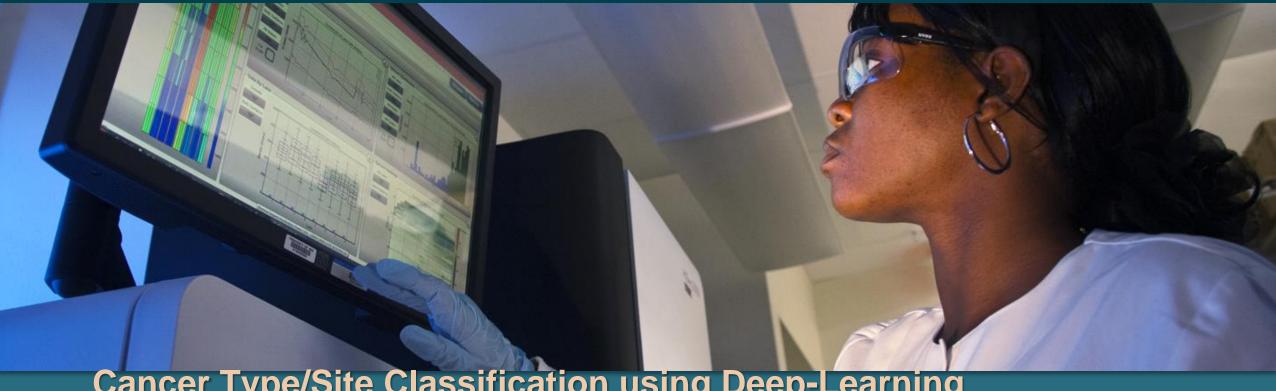
Frederick National Laboratory for Cancer Research

sponsored by the National Cancer Institute



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

S. Ravichandran BIDS, FNLCR

(in preparation)



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 testcase 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 <u>Prediction</u> has been the major focus
 - Prognosis, Recurrence, Susceptibility

- Cancer <u>Detection</u> (classification of tumors/cancers) is lagging behind <u>Prediction</u> 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?

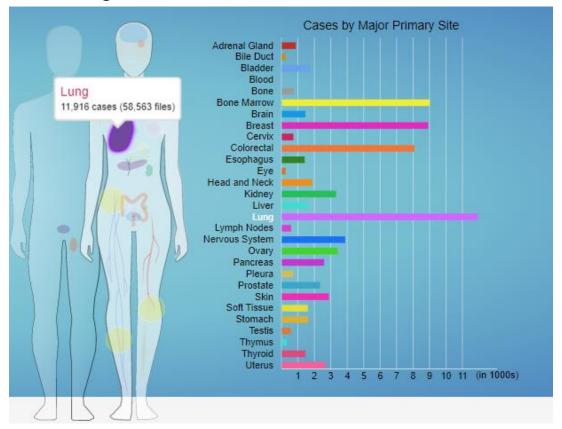
- Can we use the model as early cancer type detection
 - Improving chance of early detection cure/survival?



Cancer Burden

- Cancer is a group of diseases and worldwide 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.

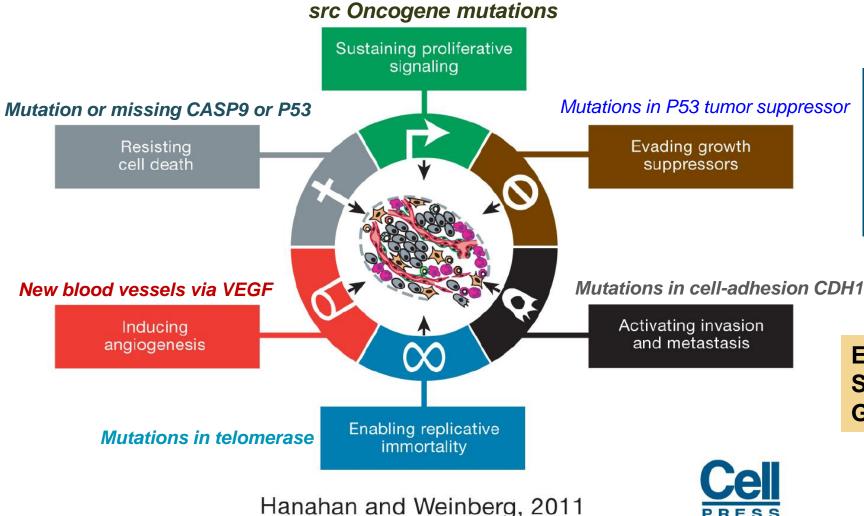
Expression changes in oncogenes; What type of changes?



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



Hallmarks of Cancer: The Next Generation



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

The Hallmarks of Cancer

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

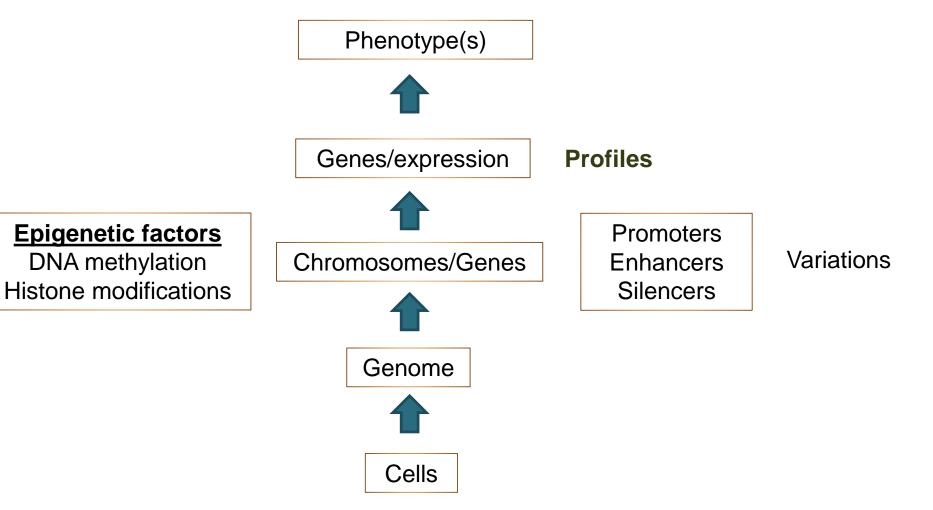
Expression changes in oncogenes; Six capabilities; Overview of **Genotype/phenotypes?**



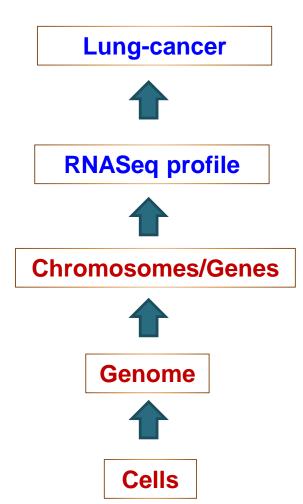


Influence of genomic features on phenotypes: An overview

DNA methylation



Influence of genomic features on phenotypes: An overview



Diagnosis/treatment vs Prediction





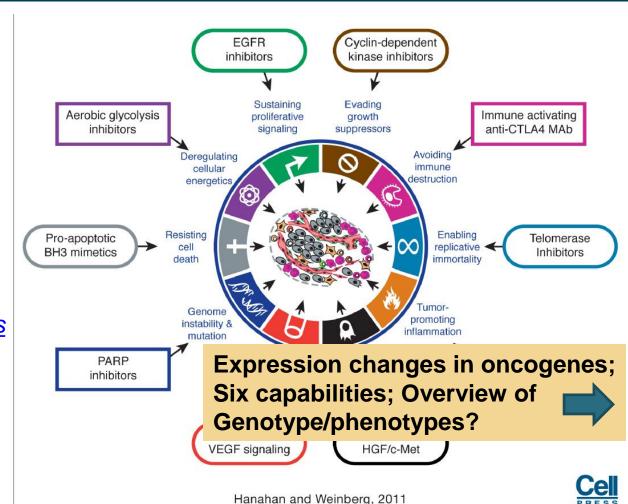
Treatment vs Type-Prediction

Treatment

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

Detecting Type

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



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

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.

Frederick National Laboratory for Cancer Research

What kind of data do we need?

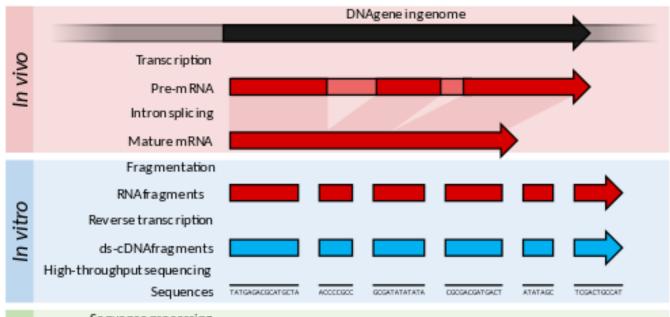
ta do we need?

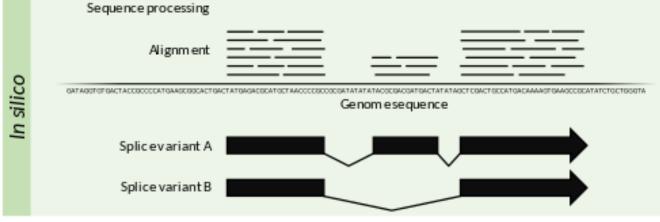
DNAgene ingenome

Transc ription

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 <u>30 human tumors</u> 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

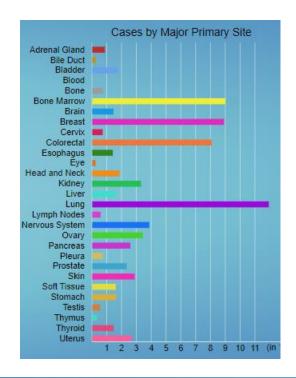
Frederick National Laboratory for Cancer Research

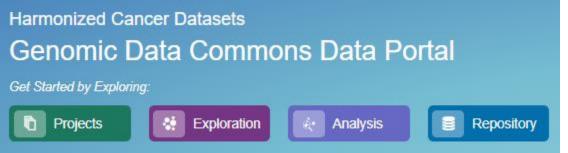
Data Harmonization: GDC

 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





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

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

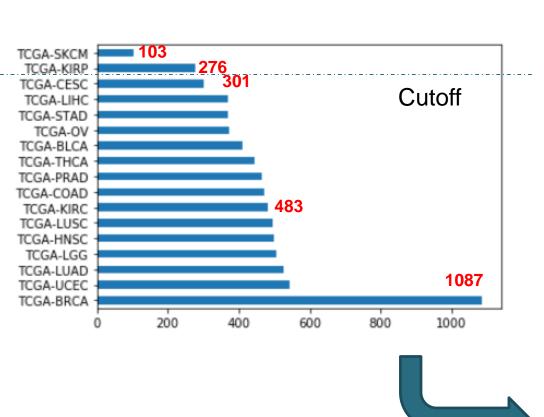
FASTQ Alignment to Ref Genome (SAM/BAM) **Quantification HTSeq** Gene Expression (FPKM-UQ)

Fragments Per Kilobase of transcript per Million mapped reads

Frederick National Laboratory for Cancer Research

National Cancer Institute

How much data for modeling?



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

sponsored by the National Cancer Institute

Expression data from a sample

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



Sample300

Data Preparation

Sample1

Breast Cancer • • • • • • • • •

Sample4

60,484 transcripts

Temporary Temp
00000079113.3 460.214441 000000000007914.0 1060.00000000000000000000000000000000
0000007918-1 5 34-96 10000 000007918-1 6 34-96 10000 000007918-1 6 34-96 10000 000007918-1 6 34-96 10000 000007918-1 6 34-96 10000007918-1 6 34-96 10000007918-1 6 34-96 100000007918-1 6 34-96 1000000000000000000000000000000000000
000007394.1 0 8885.6569 000014807.0 10 25999.2 151 000014807.0 10 25999.2 151 0000014807.1 10 49797.0 10 25999.2 151 0000015894.1 10 49797.0 10 25999.2 151 0000015984.1 10 49797.0 10 25999.2 151 0000015984.1 10 49797.0 10 25999.2 151 0000015984.1 10 49797.0 10 25999.2 151 0000015984.1 10 49797.0 10 25999.2 151 00000015984.1 10 49797.0 10 25999.2 151 00000015984.1 10 49797.0 10 25999.2 151 00000015984.1 10 49797.0 10 25999.2 151 00000015984.1 10 49797.0 10 25999.2 151 00000015984.1 10 49797.0 10 25999.2 151 00000015984.1 10 49797.0 10 25999.2 151 00000015984.1 10 49797.0 10 25999.2 151 00000015984.1 10 49797.0 10 25999.2 151 000000015984.1 10 49797.0 10 25999.2 151 00000015984.1 10 49797.0 10 25999.2 151 00000015984.1 10 49797.0 10 25999.2 151 00000015984.1 10 49797.0 10 25999.2 151 00000015984.1 10 49797.0 10 25999.2 151 00000015984.1 10 49797.0 10 25999.2 151 00000015984.1 10 49797.0 10 25999.2 151 000000015984.1 10 49797.0 10 25999.2 151 000000015984.1 10 49797.0 10 25999.2 151 0000000015984.1 10 49797.0 10 25999.2 151 00000000015984.1 10 49797.0 10 25999.2 151 000000000015984.1 10 49797.0 10 25999.2 151 00000000000000000000000000000000000
00000001915.5 6886.56568 0000001915.5 6886.56568 0000001915.4 6 6 191736 0000001915.4 6 6 191736 0000001916.1 6 191736 000001916.1 6 19 6 190001 000001916.1 6 19 6 190001 000001916.1 6 19 6 190001 000001916.1 6 19 6 190001 000001916.1 6 19 6 19 6 19 6 19 6 19 6 19 6 1
0000014618 0 20099 396
0000002754.4 0 0000015486.1 2 1 0.04917198 0000015486.1 2 1 0.04917198 0000015486.1 3 0.04917198 0000015486.3 0 0.04917198 0000015486.3 0 0.0502 00000001548.3 0 0.0502 0000001548.3 0 0.0502 0000001548.3 0 0.0502 0000001548.3 0 0.0502 00000001548.3 0 0.0502 000000001548.3 0 0.0502 000000000000000000000000000000000
0000015461, 12 004 97.7748 0000015461, 12 004 97.7748 00000015461, 13 00000015461, 13 00000015461, 13 00000015461, 13 0000000015461, 13 000000015461, 13 000000015461, 13 000000015461, 13 000000015461, 13 000000015461, 13 000000015461, 13 000000015461, 13 000000015461, 13 000000015461, 13 000000015461, 13 000000015461, 13 000000015461, 13 000000015461, 13 000000015461, 13 0000000015461, 13 0000000015461, 13 0000000015461, 13 0000000015461, 13 00000000015461, 13 00000000015461, 13 00000000015461, 13 00000000015461, 13 00000000015461, 13 000000000015461, 13 000000000015461, 13 0000000000015461, 13 0000000000015461, 13 00000000000015461, 13 0000000000000000000000000000000000
0000009841.2 4 98656.658 9 90002 9000009881.3 0 900002 900002 9000009881.3 0 9 900002 9000009881.3 0 9 900002 9000009881.3 0 9 9000009881.3 0 9 9000009881.3 0 9 9000009881.3 0 9 90000098881.3 0 9 90000098881.3 0 9 90000098881.3 0 9 90000098881.3 0 9 900000098881.3 0 9 90000098881.3 0 9 90000098881.3 0 9 90000098881.3 0 9 90000098881.3 0 9 90000098881.3 0 9 90000098881.3 0 9 900000098881.3 0 9 90000098881.3 0 9 90000098881.3 0 9 90000098881.3 0 9 90000098881.3 0 9 90000098881.3 0 9 90000098881.3 0 9 900000098881.3 0 9 90000098881.3 0 9 90000098881.3 0 9 90000098881.3 0 9 90000098881.3 0 9 90000098881.3 0 9 90000098881.3 0 9 90000000098881.3 0 9 90000098881.3 0 9 90000098881.3 0 9 90000098881.3 0 9 90000098881.3 0 9 90000098881.3 0 9 90000098881.3 0 9 9000000098881.3 0 9 90000098881.3 0 9 90000098881.3 0 9 90000098881.3 0 9 90000098881.3 0 9 90000098881.3 0 9 90000098881.3 0 9 9000000098881.3 0 9 90000098881.3 0 9 90000098881.3 0 9 90000098881.3 0 9 90000098881.3 0 9 90000098881.3 0 9 90000098881.3 0 9 90000000098881.3 0 9 90000098881.3 0 9 90000098881.3 0 9 900000098881.3 0 9 90000098881.3 0 9 900000098881.3 0 9 900000000000000000000000000000000
000002988.1 6089.99052 000002988.3 0 0 000002988.3 0 0 000002988.3 0 0 000002988.1 0 0 000001988.1 1 95730.266 000007198.1 1 95730.266 000007198.1 1 95730.266 000007198.1 1 95730.266 000007198.1 1 95730.266 000007198.1 1 95730.266 000007198.1 1 95730.266 000007198.1 1 95730.2666 000007198.1 1 95730.2666 000007198.1 1 95730.2666 000007198.1 1 95730.2666 000007198.1 1 95730.2666 000007198.1 1 95730.2666 000007198.1 1 95730.2666 000007198.1 1 95730.2666 000007198.1 1 95730.2666 000007198.1 1 95730.2666 0000007198.1 1 95730.2666 000007198.1 1 95730.2666 000007198.1 1 957300.2666 000007198.1 1 957300.2666 000007198.1 1 957300.2666 0000007198.1 1 957300.2666 0000007198.1 1 957300.2666 000007198.1 1 957300.2666 000007198.1 1 957300.2666 000007198.1 1 957300.26666 000007198.1 1 957300.26666 000007198.1 1 957300.26666 000007198.1 1 957300.26666 000007198.1 1 957300.26666 000007198.1 1 957300.26666 000007198.1 1 957300.26666 000007198.1 1 957300.26666 000007198.1 1 957300.26666 000007198.1 1 957300.26666 000007198.1 1 957300.26666 000007198.1 1 957300.266666 000007198.1 1 957300.266666 000007198.1 1 957300.266666 000007198.1 1 957300.266666 000007198.1 1 957300.2666666 000007198.1 1 957300.266666666666666666666666666666666666
0000021981 3 0 0000009475 2 0 000000019488 1 95730 2056 00000093991 1 9480 507373 00000572137 27 4885 5070 00000172137 27 22677 4888 000001948 1 306 24605 0000007496 1 306 24605 0000007496 1 306 24605 000007496 1 306 24769 000007496 1 384 475569 000007496 1 384 475569 000007496 1 384 475569 000007496 1 386 47566
00000059184 1 97310.2056 0000013881 1 97310.2056 0000013801 1 97310.2056 0000013801 1 97310.2056 0000033980 1 3486.072737 0000017217 17 4485.9970 0000017484 2 9700 0000001484 1 9300.24678 00000006644 9 1300.24678 0000007486 1 0 0000007486 1 0 0000007486 1 0 0000007486 1 0 0000007486 1 0 0000007486 1 0 0000007486 1 0 0000007486 1 0 0000007486 1 0 0000007486 1 0 0000007486 1 0 00000007486 1 0 0000007486 1 0 0000007486 1 0 0000007486 1 0 000000007486 1 0 00000007486 1 0 00000007486 1 0 00000007486 1 0 000000007486 1 0 00000007486 1 0 00000007486 1 0 00000007486 1 0 000000007486 1 0 00000007486 1 0 000000007486 1 0 000000007486 1 0 0000000000000000000000000000000000
00000019488.1 0 97380.0566 0000034988.1 957380.0565 0000032898.1 3848.027273 00000172137.2 14485.5907 00000172137.2 14485.5907 00000172137.3 14485.5907 00000172137.3 15858.3
00000019488.1 0 97380.0566 0000034988.1 957380.0565 0000032898.1 3848.027273 00000172137.2 14485.5907 00000172137.2 14485.5907 00000172137.3 14485.5907 00000172137.3 15858.3
0000058981 1 3484.027273 0000077233.77 41488.9507 0000015733.77 41488.9507 0000015730.77 23617.4008 0000034043.1 2002.24605 0000004043.1 2002.24605 0000004043.1 2002.24605 0000004043.1 2002.24605 00000073646.1 392.4755669 0000023640.1 394.4755669 0000023640.1 1583.31582 00000036946.4 5588.66648 00000036944.8 45588.66648
0000017.197.17 44485.9507 0000016.770.0 7.285717.4088 00000024949.1 2 2087.27.4088 00000024949.1 3819.5246749 0000006404.9 15586.9216 00000027.8616.1 0 0000023890.1 105.1 1588.137582 00000023190.1 1558.1 1588.137582 00000023144.8 4 4558.66948 00000028144.8 4 4558.6648
0000017.197.17 44485.9507 0000016.770.0 7.285717.4088 00000024949.1 2 2087.27.4088 00000024949.1 3819.5246749 0000006404.9 15586.9216 00000027.8616.1 0 0000023890.1 105.1 1588.137582 00000023190.1 1558.1 1588.137582 00000023144.8 4 4558.66948 00000028144.8 4 4558.6648
0000016700.7 22677.4008 0000024949.3 2082.24605 00000040423.1 380.5246749 00000074561.5 0 0000074561.5 0 0000024881.1 0 0000023690.1 594.4755669 0000023105.1 1583.312582 00000043941.1 0 00000084394.1 5 45536.6048 0000088914.8 45536.6048
0000034943.2 2082.245035 00000040423.1 310.3246749 00000056542.9 155863.9216 00000273616.1 0 00000273616.1 1 0 0000023408.1 1 0 0000023400.1 1 394.4755699 00000031105.1 1583.312582 000000331141.8 45538.60548 00000083944.1 150.0847054
00000040423.1 310.5246749 0000060642.9 155863.9216 00000275651.5 0 00000234881.1 0 00000234881.1 1 0 00000234881.1 1583.312582 000000331405.1 1583.312582 000000331405.1 4 1583.912582 000000382141.8 45538.60648 000000859415.4 119.0847054
00000060642.9 155863.9216 00000234881.1 0 00000234881.1 0 00000234681.1 1583.312582 00000234106.1 1583.312582 00000243106.1 1583.312582 00000243044.1 0 00000122441.8 45538.60648 000000269416.4 119.0847054
00000271616.1 0 00000234881.1 0 00000234604.1 394.4755669 0000023105.1 1583.31282 00000018904.1 0 00000182141.8 45538.60648 000000569416.4 119.0847054
00000234881.1 0 00000236040.1 394.4755669 00000231105.1 1583.312582 00000243044.1 0 00000182141.8 45538.60648 000002699416.4 119.0847054
00000236040.1 394.4755669 00000231105.1 1583.312582 00000043044.1 0 00000182141.8 45538.60648 000000369416.4 119.0847054
00000231105.1 1583.312582 00000243044.1 0 00000182141.8 45538.60648 00000269416.4 119.0847054
00000243044.1 0 00000182141.8 45538.60648 00000269416.4 119.0847054
00000182141.8 45538.60648 00000269416.4 119.0847054
00000269416.4 119.0847054
00000204981.1 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

INSCIDENCEST79.1.5.5 SAME 10096	ENSG00000270112.3	460.2343433
MAGGINGOUTER-15 GERES ARGOS	ENSG00000167578.15	52440.10096
INCOGNOMINATION 259996-22953 100000000000000000000000000000000000	ENSG00000273842.1	0
INSTITUTE	ENSG00000078237.5	68165.45626
Decomposition Decompositio	ENSG00000146083.10	255959.2351
DESCRIPTION CONTROL	ENSG00000225275.4	0
Decorptions Graph 99902 Decorption Decorption Graph 99902 Decorption Decorption Graph 99902 Decorption		
NGG000001981 0		
Topic Topi	ENSG00000259883.1	6108.999052
INSTITUTE 1	ENSG00000231981.3	0
INSCIDENCIAL STATE AND ASSOCIATION AND ASSO		
DISCOMMONISMENT 1 MARCO 2773 DISCOMMONISMENT 2 MARCO 2773 DISCOMMONISMENT 2 MARCO 2774 AURIS 5107 DISCOMMONISMENT 2 MARCO 2774 AURIS 5105 DISCOMMONISMENT 2 MARCO 2774 AURIS 5105 DISCOMMONISMENT 2 MARCO 2774 DISCOMMONISMENT 2 MARCO 2774	ENSG00000201788.1	0
DRSG0000017311717 4488 5197 DRSG00000173717 238717408 DRSG00000173707 238717408 DRSG00000184012 238717408 DRSG000000184013 319.354819 DRSG00000018401 1 319.354819 DRSG00000018401 1 0 DRSG0000011801 1 1 0 DRSG0000011801 1 1 180 31.1582 DRSG0000011801 1 180 31.1582	ENSG00000134108.11	957330.2056
BASCOMOMISTAND. 7 23:577.408 BASCOMOMISTAND. 2 20:571.408 BASCOMOMISTAND. 2 20:52.24015 BASCOMOMISTAND. 1 3015.546749 BASCOMOMISTAND. 1 1506.3016 BASCOMOMISTAND. 1 1506.6016 BASCOMOMISTAND. 1 1506.6016 BASCOMOMISTAND. 1 1506.6016 BASCOMOMISTAND. 1 150.6015	ENSG00000263089.1	3484.027373
PRSCORDOUGHAND 2 2002 245035 PRSCORDOUGHAND 3 318 2454749 PRSCORDOUGHAND 3 318 2454749 PRSCORDOUGHAND 3 318 2457474 PRSCORDOUGHAND 3 318 245569 PRSCORDOUGHAND 1 394 275569 PRSCORDOUGHAND 1 318 31 2582 PRSCORDOUGHAND 4 1 30 258 265648 PRSCORDOUGHAND 4 1 50 258 265648 PRSCORDOUGHAND 4 1 50 258 265648 PRSCORDOUGHAND 4 1 10 267754		
BSG00000044421.1 310.5346749 bSG000000464.2 9 155863.9216 BSG000000273616.1 0 DSG00000024481.1 0 DSG00000024481.1 18.83.17582 DSG00000023105.1 1883.317582 DSG00000032104.1 0 DSG0000003214.8 45538.60648 DSG00000082141.8 45538.60754	ENSG00000167700.7	226717.4208
ENS G0000006642.9 155863.9216 BNS G00000272616.1 0 ENS G00000272868.1 0 ENS G00000236040.1 394.4755669 ENS G00000236040.1 0 ENS G00000243044.1 0 ENS G00000243044.1 0 ENS G00000243044.1 10 ENS G0000012414.8 45538.60648	ENSG00000234943.2	2082.245035
ENSG00000275656.1 0 ENSG00000234681.1 0 ENSG0000234690.1 394.4755669 ENSG000023105.1 1583.317582 ENSG00000243044.1 0 ENSG00000043044.1 0 ENSG00000182141.8 45538.6048 ENSG0000026946.4 119.0647054		
ENSG00000234881.1 0 ENSG0000023609.1 394.4755669 ENSG0000023105.1 1583.312582 ENSG0000043044.1 0 ENSG00000182141.8 45538.60648 ENSG00000189416.4 119.0847054	ENSG00000060642.9	155863.9216
ENSG00000236040.1 394.4755669 ENSG00000231105.1 1583.312582 ENSG00000243044.1 0 ENSG0000024304.8 45538.60648 ENSG00000269416.4 119.0847054	ENSG00000271616.1	0
ENGG00000231105.1 1583.312582 ENGG00000243044.1 0 ENGG0000182141.8 45538.60648 ENGG00000269416.4 119.0847054	ENSG00000234881.1	0
ENSG00000243044.1 0 ENSG00000182141.8 45538.60648 ENSG00000269416.4 119.0847054	ENSG00000236040.1	394.4755669
ENSG00000182141.8 45538.60648 ENSG00000269416.4 119.0847054	ENSG00000231105.1	1583.312582
ENSG00000269416.4 119.0847054	ENSG00000243044.1	0
	ENSG00000182141.8	45538.60648
ENSG00000264981.1 0		119.0847054
	ENSG00000264981.1	0

	Expression *
	1658.464179
	460.2343433
5	52440.10096
	0
	68165.45626
٥	255959.2351
	0
2	104.9473768
2	4968556.658
	6108.999052
	0
	0
	0
1	957330.2056
	3484.027373
7	41485.9507
	226717.4208
	2082.245035
	310.5246749
	155863.9216
	0
	0
	394.4755669
	1583.312582
	0
	45538.60648
	119.0847054

Sample297 Sample298

Genes	 Expression 	×
SG00000242268.2	1658.464179	
SG00000270112.3	460.2343433	
SG00000167578.15	52440.10096	
SG00000273842.1	0	
SG00000078237.5	68165.45626	
SG00000146083.10	255959.2351	
SG00000225275.4	0	
SG00000158486.12	104.9473768	
SG00000198242.12	4968556.658	
SG00000259883.1	6108.999052	
SG00000231981.3	0	
SG00000269475.2	0	
SG00000201788.1	0	
SG00000134108.11	957330.2056	
SG00000263089.1	3484.027373	
SG00000172137.17	41485.9507	
SG00000167700.7	226717.4208	
SG00000234943.2	2082.245035	
SG00000240423.1	310.5246749	
SG00000060642.9	155863.9216	
SG00000271616.1	0	
SG00000234881.1	0	
SG00000236040.1	394.4755669	
SG00000231105.1	1583.312582	
SG00000243044.1	0	
SG00000182141.8	45538.60648	
SG00000269416.4	119.0847054	
SG00000264981.1	0	

Sample299



Data Preparation

	Sample1	Sample2	Sample3	Sample4	Sample297	Sample298	Sample299	Sample300
Breast Cancer	Ceres Cer	Service Ser	Seminary Seminary	Common		Common Common		
Lung Cancer	Series				Section	Common Common		Series
					Genes ▼ Expression ▼ ENSCOOD002788.2 1658.464179	Genes	Genes Expression ENSGROUND42268.2 1558.464179	Genes
Kidney Cancer		General Company Compan		General	09G000007011.3 40 20 34443 09G000007194.5 52 24440 1006 09G000007194.5 52 2440 1006 09G000007194.5 52 2440 1006 09G00000000000000000000000000000000000	INCOGNOTOPICE 3	NewGOODCOMPATION 1 - 30-340-410	MONOCONDONIAL MAN MAN

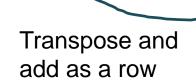


Merged Sample Expression Data

Genes

		0	1	2	3	4	5	6	7	8	9	6	0474	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
Щ	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
7	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
SAMP	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
3																						***
Z	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



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.

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

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



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
[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., 0., 1., 0., 0., 0., 0., 0., 0.]
 dtype=float32)
```

Frederick National Laboratory for Cancer Research

Data preparation steps summary

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

GDC • Filter Program/Projects (Ex. TCGA-BRCA) • Filter to retain only RNA-Seq data; FPKM-UQ scaling **GDC RNASeq** Primary Tumor samples only Tumor • Download final list and meta files **Extract** Restrict sample size >= 300 and extract each class to create a balanced dataset (each class = 300 samples) Samples • Download/merge individual expression data 4500 files ~ 2.5 GB; carried out in Biowulf Create Coding • Look in Clinical file to extract labels and attach them; one-hot encoding Prep

▼ Modeling

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



Before we break for hands-on

 Python as the programming language for this workshop, but similar libraries are available in R or other languages



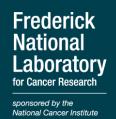


- Will use Jupyter Notebook for sharing the code
 - With little effort one can convert the Python code into R and still use Jupyter Notebook



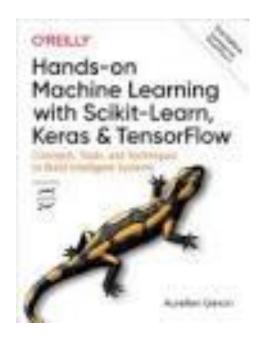
To be continued after Code-Review

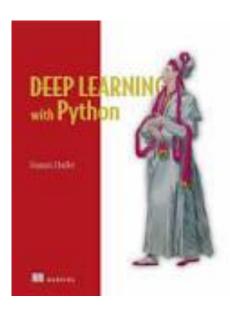
https://github.com/ravichas/ML-TC1



Before we break for hands-on

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





Following two are good books for beginners and up



Convolutional Neural Networks

Preparation in progress



Thanks

S. Ravichandran