Module 4

Lecture 2

Introduction to RNA sequencing data

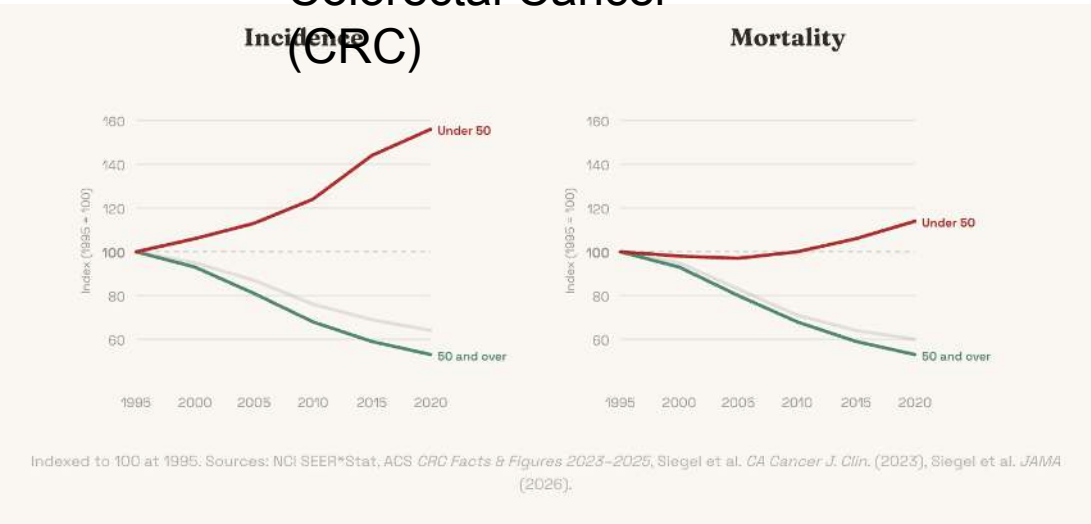
Sarah Groves, Ph.D.

Shayn Peirce-Cottler, Ph.D.

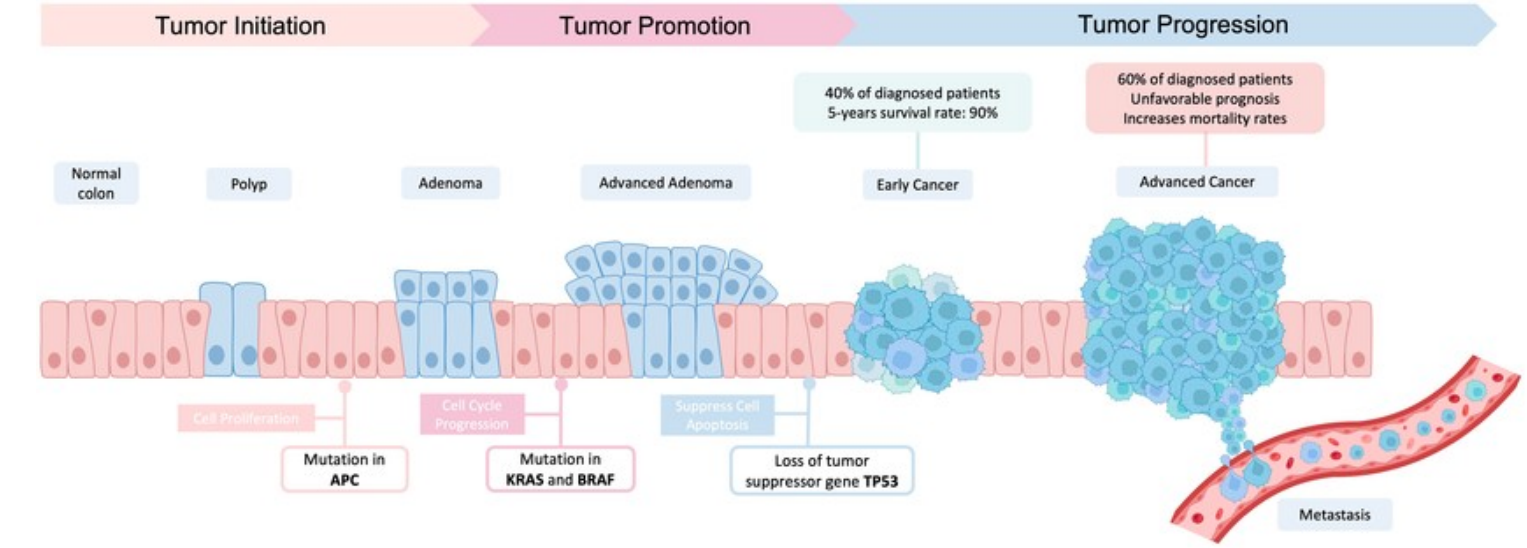


Exploring the hallmarks: everything is connected!

Colorectal Cancer (CRC)







- Try running with **only one of the hallmarks "active."**
  - What happens when you only turn on "sustained growth signals" – and what keeps the cell from dividing?
  - If the cell can only "resist apoptosis," does it eventually spread?
- Try **different combinations**
  - How many hallmarks do you need for the cancer to grow?
  - What about for it to metastasize (reach new sites)?
- Once you feel comfortable with the first version, try the advanced version (extra six hallmarks added)

https://pmc.ncbi.nlm.nih.gov/articles/PMC11714384/#bib14

https://www.hankgreen.com/crc

In class discussion

- What did you learn about the hallmarks and how they are connected?
  - Need sustained growth signals in combo with any other ones
  - Advanced sim: needed immune evasion
  - Need more than one: normal cells are redundant
- Which combination gave you the most interesting results and why?
  - Sometimes killer t cells can still kills the cancer cells
  - Immune evasion and genome instability

Tumor staging (clinical data)

TNM Staging

T = Primary Tumor

TX Primary tumor cannot be evaluated
T0 No evidence of primary tumor
Tis Carcinoma in situ (early cancer that has not spread to neighboring tissue)
T1, T2, T3, T4 Size and/or extent of the primary tumor

N = Regional Lymph Nodes

NX Regional lymph nodes cannot be evaluated
N0 No regional lymph node involvement
N1, N2, N3 Involvement of regional lymph nodes (number and/or extent of spread)

M = Distant Metastasis

MX Distant metastasis cannot be evaluated
M0 No distant metastasis
M1 Distant metastasis

STAGE

- Stage 0: Carcinoma in situ
- Stage I, Stage II, and Stage III: Higher numbers indicate more extensive disease: greater tumor size, and/or spread of the cancer to nearby lymph nodes and/or organs adjacent to the primary tumor.
- Stage IV The cancer has spread to another organ.

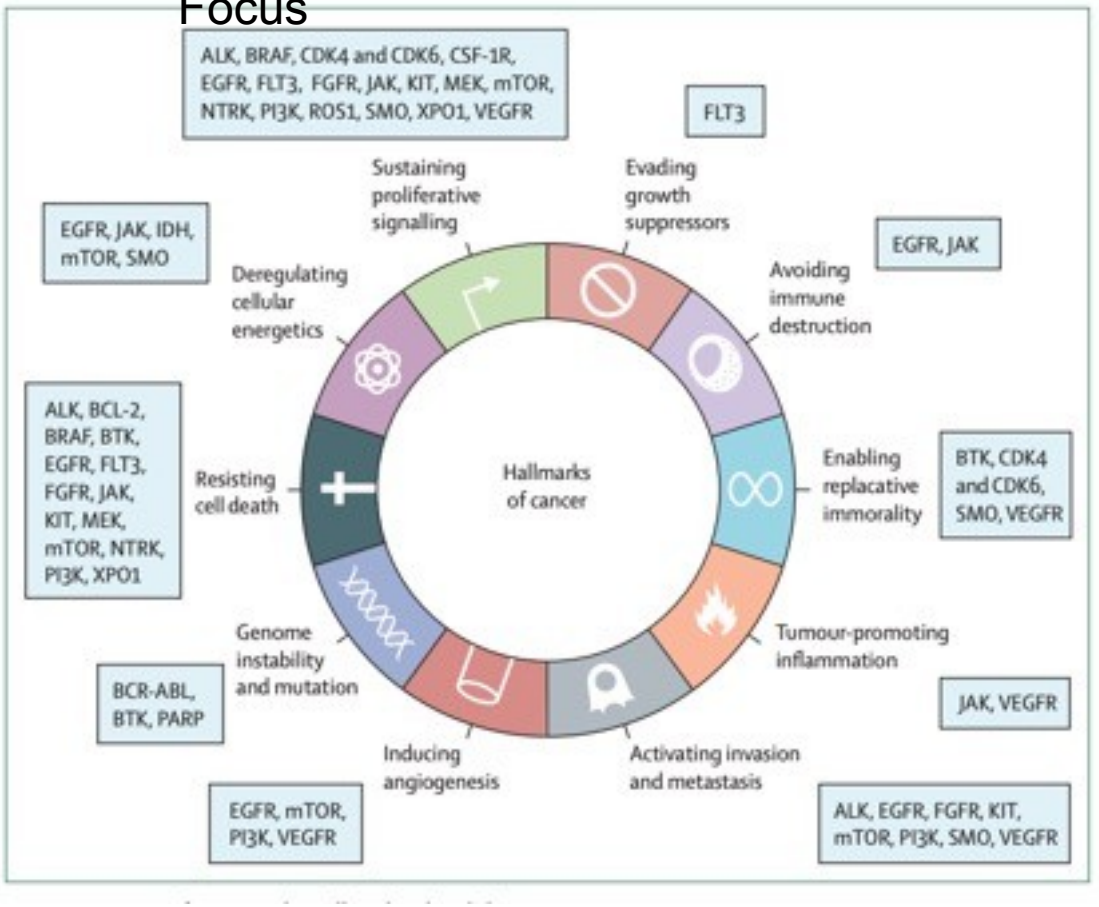
SUMMARY STAGE

- In situ: early cancer present only in the layer of cells in which it began
- Localized: limited to the organ in which it began
- Regional: spread to nearby lymph nodes or organs and tissues.
- Distant: spread to distant organs or distant lymph nodes.
- Unknown

Treatment options

- **Chemotherapy**: kills fast growing cells
- **Radiotherapy**: radiation applied to specific organ/region containing tumor
  - Causes DNA damage to kill proliferative cells
- **Targeted therapy**: Targeting a specific deregulated or mutated gene, like growth factor receptors that are overactive
  - **Hormone therapy**: targeting hormones that drive tumor growth
  - **Epigenetic therapy**: targeting epigenetic mechanisms like DNA methylation/gene silencing
- **Immunotherapy**: Enlisting immune cells to help kill the tumor

Targeted therapies by Hallmark Focus



https://doi.org/10.1016/S0140-6736(20)30164-1

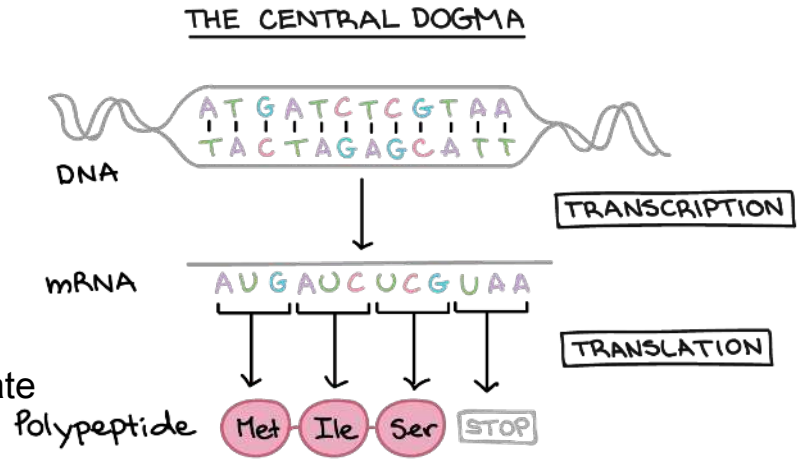
Preview of what's coming… how do we know which rules are broken in different cancers?

- We'll use RNAseq data (gene expression data)
- Mutations in oncogenes and tumor suppressors → abnormal gene expression patterns
- Hallmarks organize data into "gene expression modules" = related genes that tend to be correlated or anti-correlated
  - Often upregulated (turned UP) or turned OFF altogether

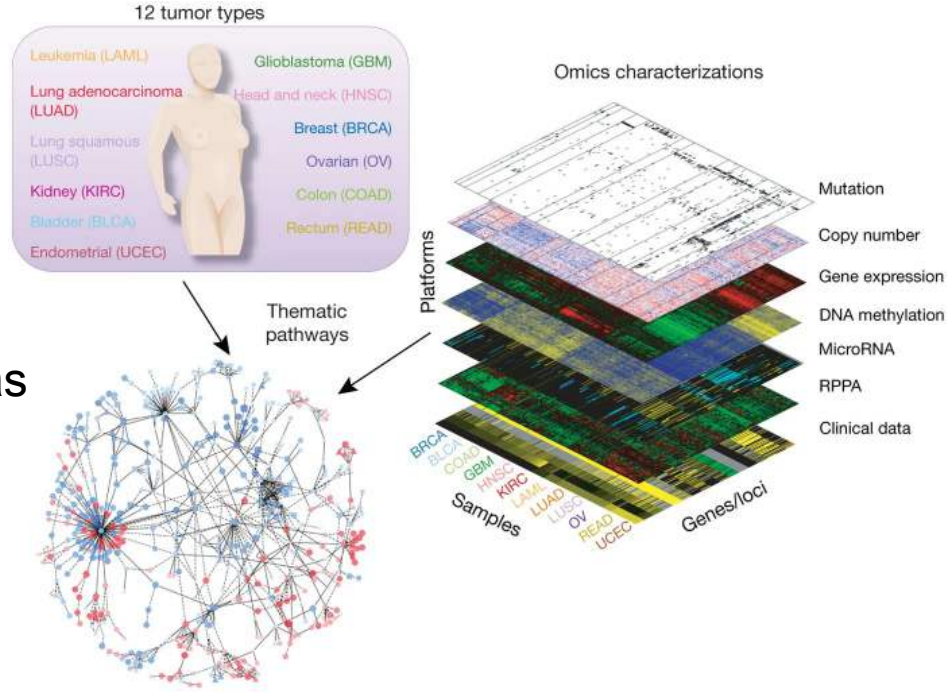
Mutations on this level (genome)

Affect "cell identity" on RNA level (transcriptome) "Live feed"

Proteins are made and accumulate to act as little machines "Highlights reel"



The Cancer Genome Atlas (TCGA)



Deciding on a project goal: choices to make!

- What **hallmark** of cancer will you focus on?
- Do you want to focus on a **single cancer type** or look for patterns **across cancer types**?
- Start to think about what kind of question you want to ask. Example question “scaffolds”:
  - Is hallmark X predictive of survival in Y cancer type?
  - Is hallmark X related to hallmark Y in Z cancer type?
  - Can different cancer types be distinguished by X hallmark alone?
  - Are there distinct groups of cancer types with hallmark X expression that respond to treatment in different ways?
  - Does patient history of X predict expression of hallmark Y in Z cancer?
  - Does expression of X hallmark predict Y tumor stage or Z treatment response?

Due Friday by midnight:
First GitHub Check-in with disease background, data background, and first pass at a question/project goal

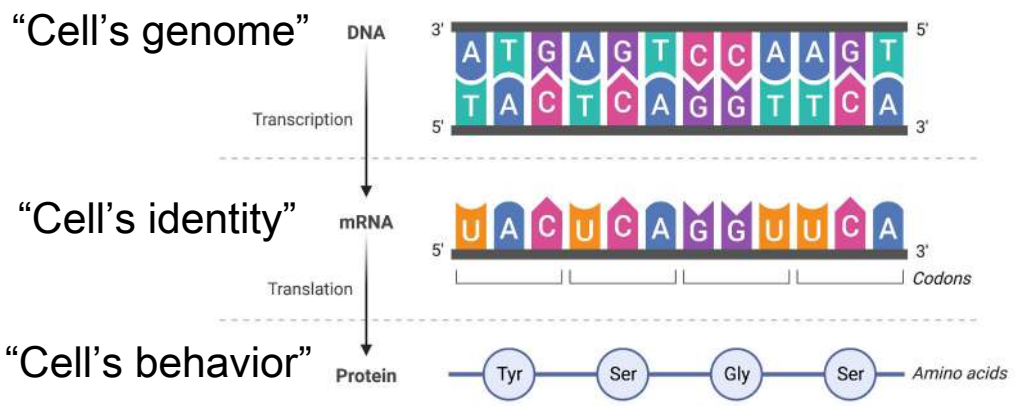
Learning Objectives for Module 3: Cancer

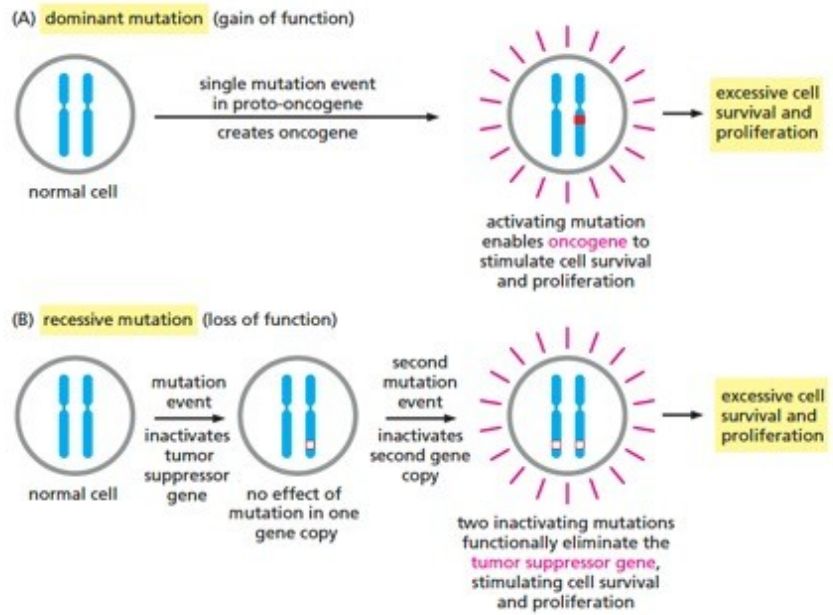
- *I understand how machine learning methods use **optimization** of a cost function to determine the best answer to my research question.*
- *I can use **pandas and scikit-learn** to work with **RNA-seq data**, apply a machine learning method to a dataset, and validate the model.*
- *I understand the difference between **supervised** and **unsupervised** learning methods and know when to use them.*
- *I know how to **avoid overfitting** and determine if my model is "good enough" using quantitative metrics.*

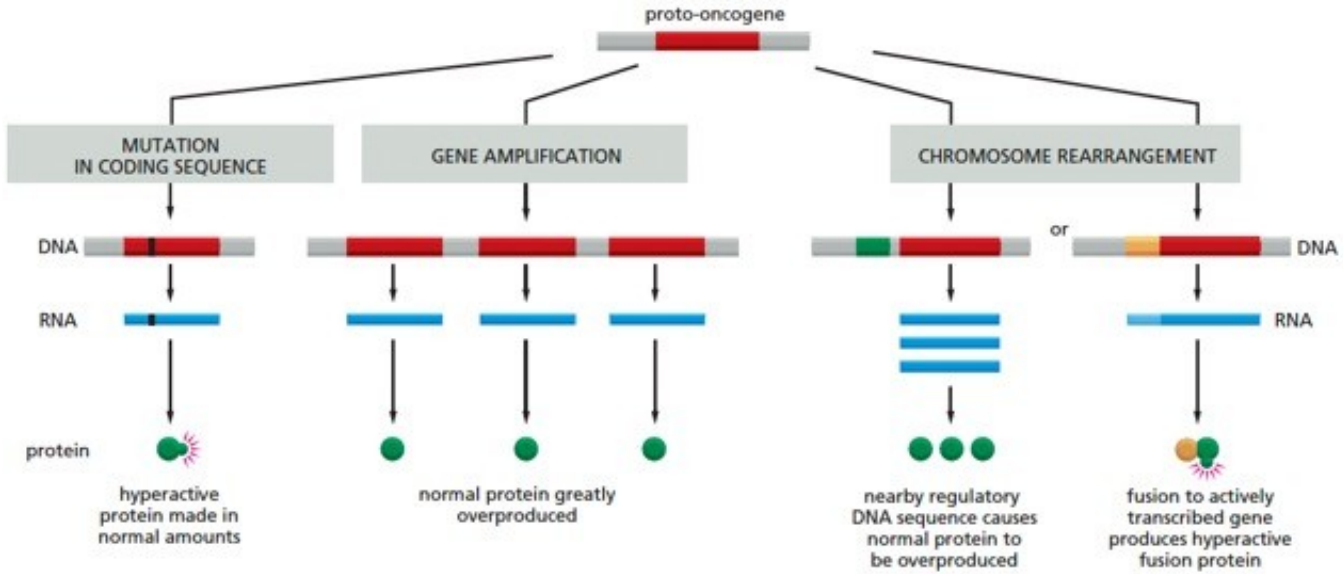
Getting the data

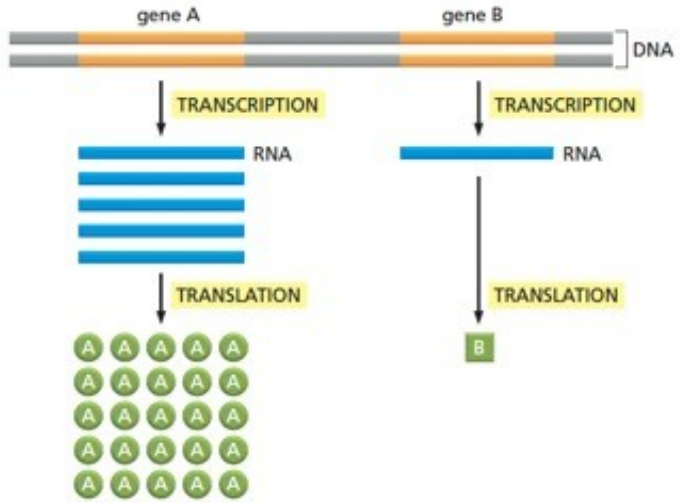
- Pull my upstream repo to get the metadata and sample EDA code
- IMPORTANT: >100MB is TOO BIG to push and pull from Github
- Instead, you should download the training and validation datasets from Canvas and add them to the **data** subfolder of your repository
  - .gitignore set up so they don't get tracked

How mutations can affect gene expression



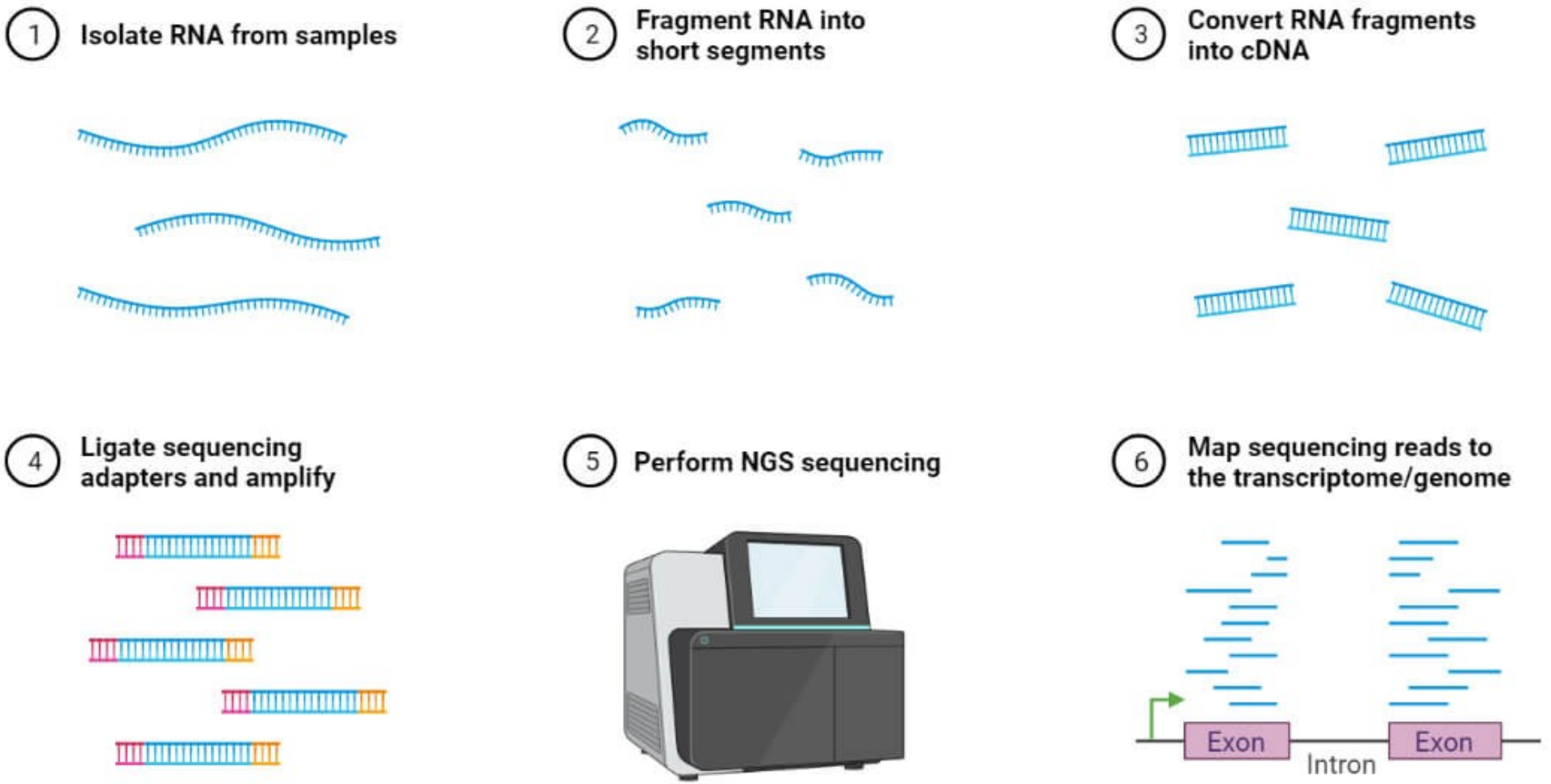






Cell identity can be measured with sequencing

A brief overview of RNA sequencing



https://microbenotes.com/rna-sequencing-principle-steps-types-uses/

https://www.rna-seqblog.com/rpkm-fpkm-and-tpm-clearly-explained/

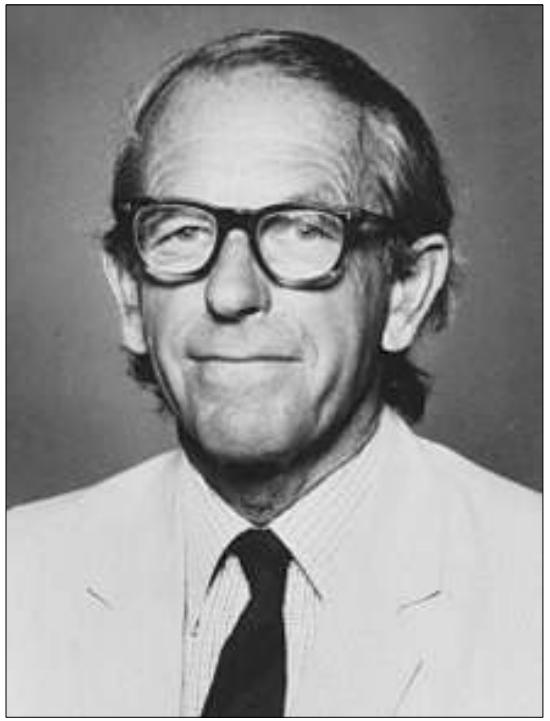
Sequencing DNA

5 Perform NGS sequencing



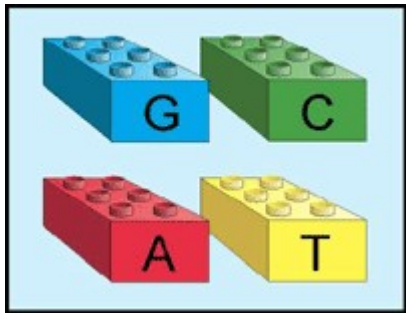
Sanger sequencing

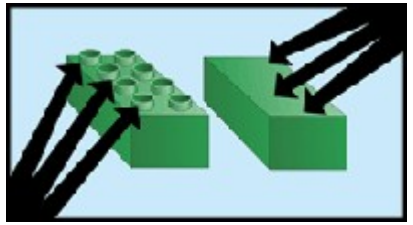
- First generation sequencing technology
- The technology used to sequence the human genome
- Still used widely today (though not for large-scale applications)
- Depends on **dideoxynucleotides** – analogous to a faulty lego that doesn't stack



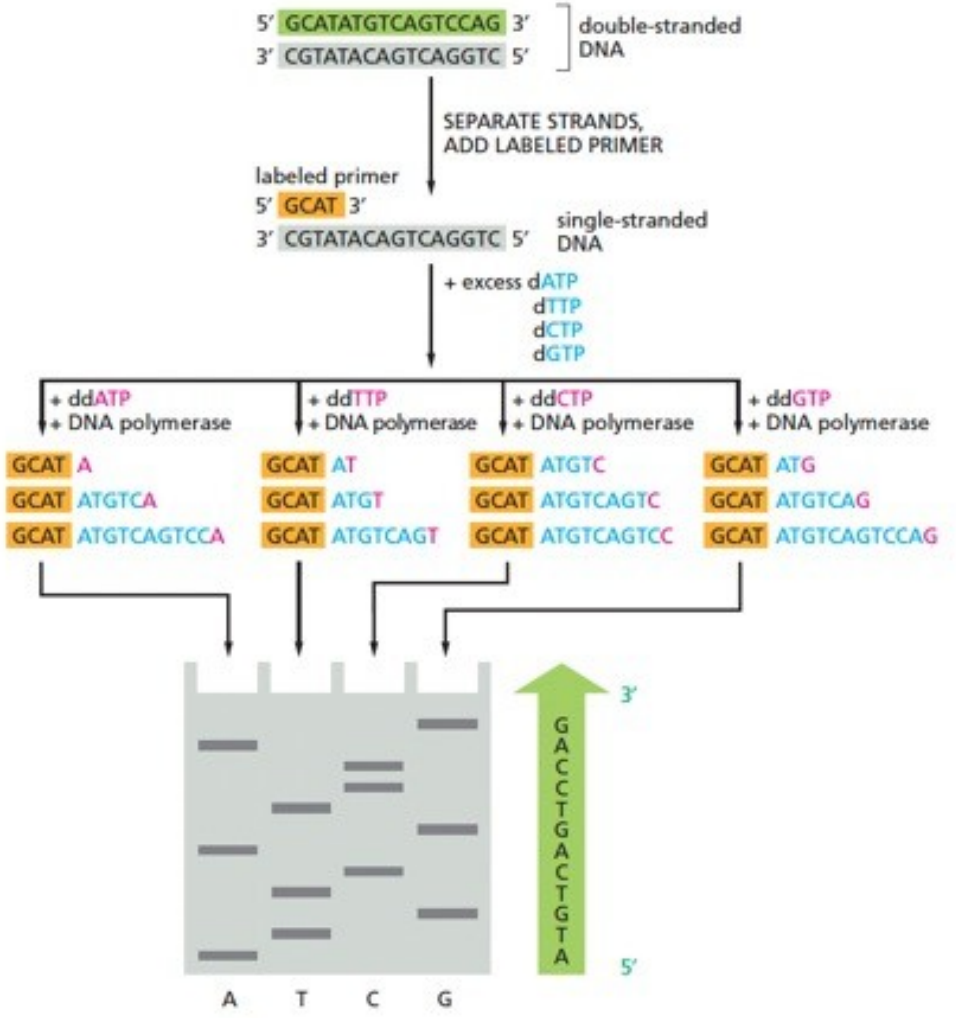
Frederick Sanger

1980 Nobel Prize for Sequencing Research





Sanger sequencing & automated Sanger sequencing



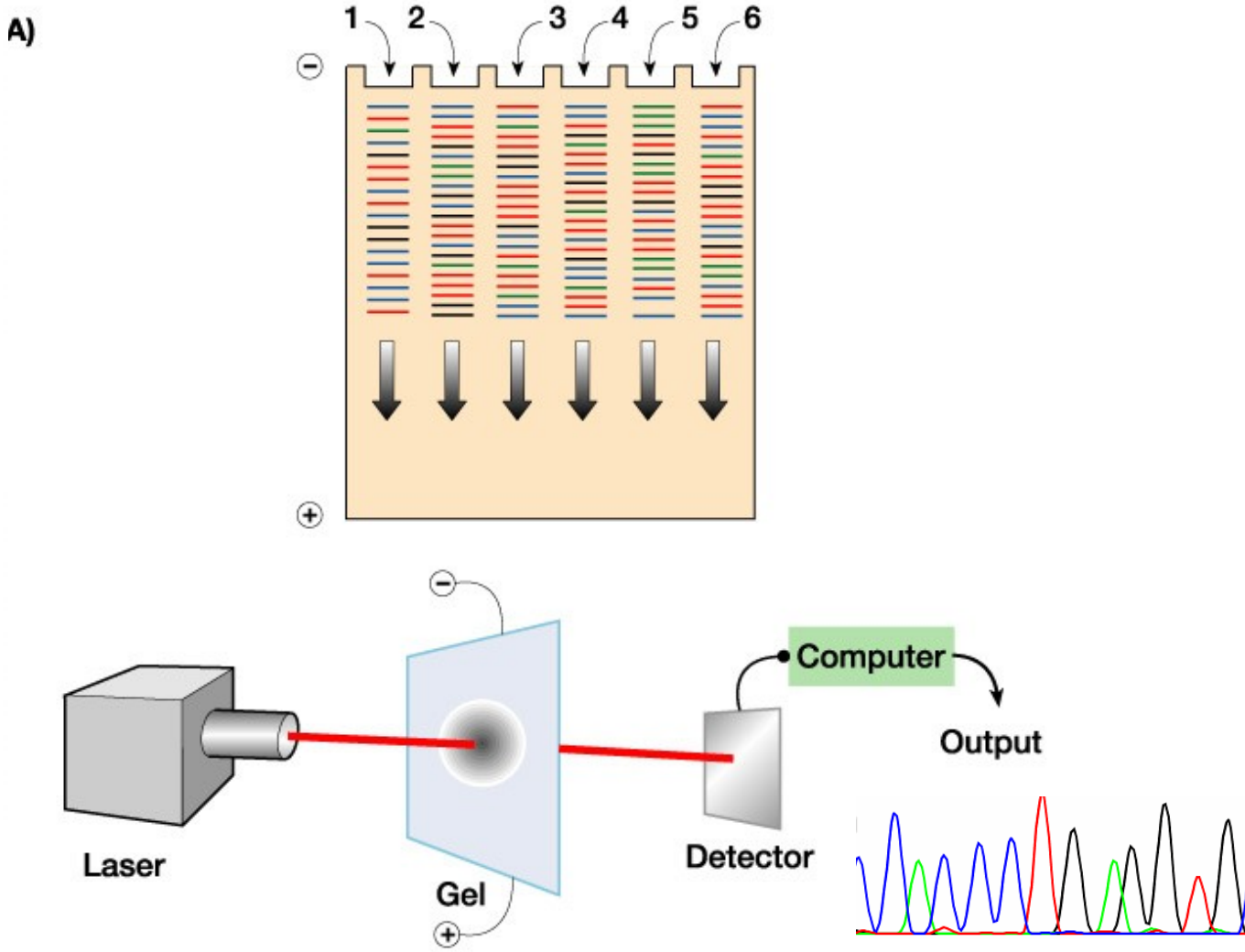
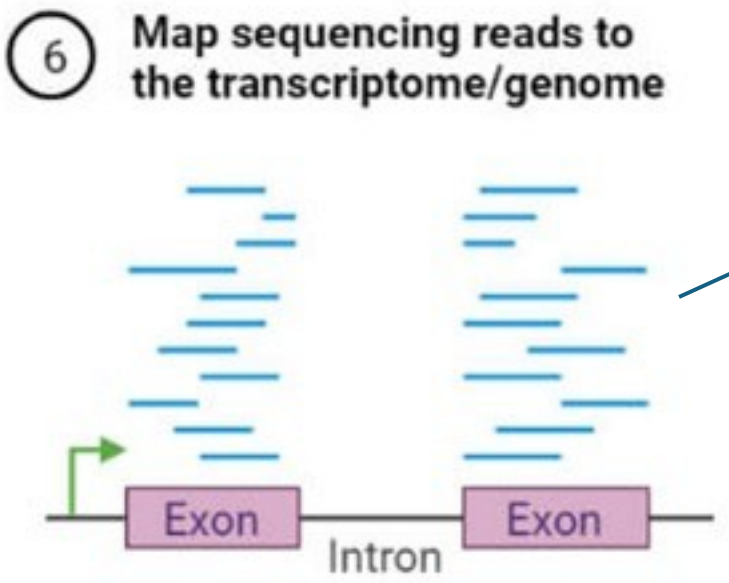


Figure 7-3 part 1 of 2 Human Molecular Genetics, 3/e. (© Garland Science 2004)

From raw sequencing reads to “gene counts”



Alignment (FASTQ to BAM)

Counts per exon (transcribed gene chunk)

Longer genes will generate more fragments even if the gene isn’t transcribed more
→ normalize by gene length

Tumor samples wont all have the same total number of reads
→ normalize by “library size”

Reads / length of gene = RPK

↓

Divide by $\frac{total\ RPK}{1{,}000{,}000}$

↓

TPM

Our data is log-transformed TPM values

Dataset and metadata

- The Cancer Genome Atlas RNA-seq dataset, re-processed by Rahman et al. (see paper on Canvas)
  - 24 cancer types (abbreviations are spelled out in paper)
- I subsetted the data to include:
  - ~50-100 tumors per cancer type (1802 out of 9264 total tumor samples)
- Metadata includes ~20 columns from original 526
  - Note that not all metadata has values across cancer types!

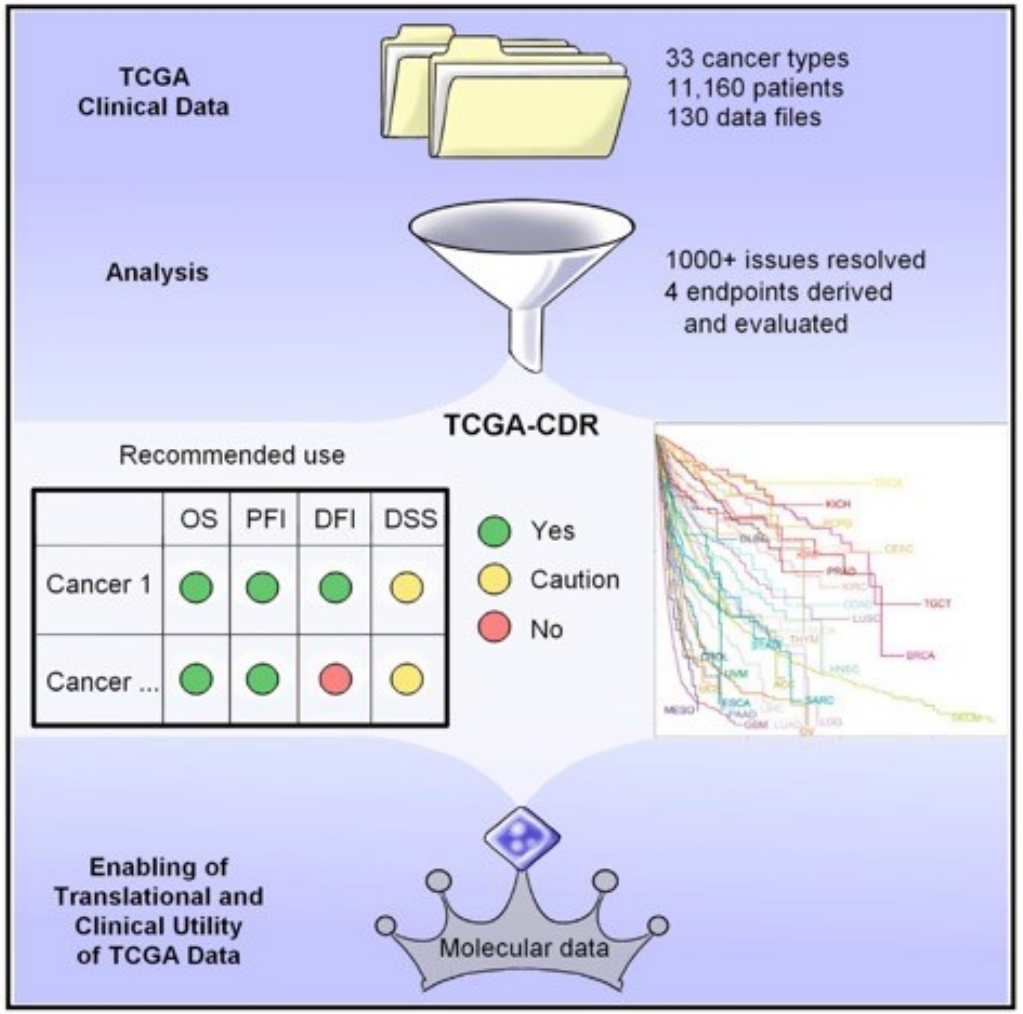
Data: genes X samples, $\log_2$(TPM+1)

	TCGA-E9-A1NI-01A-11R-A14D-07	TCGA-E2-A1LK-01A-21R-A14D-07	TCGA-BH-A0B2-01A-11R-A10J-07	TCGA-E2-A107-01A-11R-A10J-07	TCGA-LL-A5YN-01A-11R-A28M-07	TCGA-BH-A0DQ-01A-11R-A084-07	TCGA-D8-A73X-01A-11R-A32P-07	TCGA-AR-A0TP-01A-11R-A084-07
KRT5	0.594188	9.570116	6.591551	6.504049	6.328152	5.202264	6.187277	3.867826
KRT19	12.399565	4.379253	10.551527	13.090924	11.520876	11.789722	12.210669	8.995227
AGR2	11.478450	0.265591	8.512289	10.101441	10.436393	10.554898	10.715356	2.790289
KRT17	0.977721	10.326619	6.148319	6.757271	8.051663	5.613325	7.118941	5.277433
KRT6A	1.763508	9.266280	1.160773	2.529692	2.464750	1.461973	2.823227	1.908509

Metadata: tumor samples X features

1	sample	cancer_type	bcr_patient_barcode	bcr_patient_uuid	patient_id	gender	race
2	TCGA-E9-A1NI-01A-11R-A14D-07	BRCA	TCGA-E9-A1NI	c65b835a-2d38-4250-a173-0780d2c2cf58	A1NI	FEMALE	WHITE
3	TCGA-E2-A1LK-01A-21R-A14D-07	BRCA	TCGA-E2-A1LK	1d27253f-b036-44e7-a04d-8da5bbf57419	A1LK	FEMALE	BLACK OR AFRICAN AMERICAN
4	TCGA-BH-A0B2-01A-11R-A10J-07	BRCA	TCGA-BH-A0B2	57a1604c-60b7-4b30-a75e-f70939532c5c	A0B2	FEMALE	WHITE
5	TCGA-E2-A107-01A-11R-A10J-07	BRCA	TCGA-E2-A107	eff2360e-399a-4167-ab2b-798e27bef739	A107	FEMALE	WHITE
6	TCGA-LL-A5YN-01A-11R-A28M-07	BRCA	TCGA-LL-A5YN	EF1B3332-CD7F-41BB-A2D3-2538E7BECC5C	A5YN	FEMALE	BLACK OR AFRICAN AMERICAN
7	TCGA-BH-A0DQ-01A-11R-A084-07	BRCA	TCGA-BH-A0DQ	a851ca14-7bbf-455b-a054-ec03979a4247	A0DQ	FEMALE	WHITE
8	TCGA-D8-A73X-01A-11R-A32P-07	BRCA	TCGA-D8-A73X	1BE375FE-5AE5-4E75-A729-69F86A661D9A	A73X	FEMALE	WHITE
9	TCGA-AR-A0TP-01A-11R-A084-07	BRCA	TCGA-AR-A0TP	ab89bb2e-da84-4116-a8cd-49aad7bbfd4d	A0TP	FEMALE	WHITE
10	TCGA-E2-A1IF-01A-11R-A144-07	BRCA	TCGA-E2-A1IF	718db047-9f32-48c4-98c9-515218441bca	A1IF	FEMALE	WHITE
11	TCGA-EW-A6SD-01A-12R-A33J-07	BRCA					
12	TCGA-AO-A03T-01A-21R-A034-07	BRCA	TCGA-AO-A03T	dd6b64a8-d546-4fd1-bbbd-ac9703791cae	A03T	FEMALE	WHITE
13	TCGA-OL-A5D8-01A-11R-A27Q-07	BRCA	TCGA-OL-A5D8	6D4860B2-21C6-4B7B-9093-2F32700D0612	A5D8	FEMALE	BLACK OR AFRICAN AMERICAN
14	TCGA-AR-A24X-01A-11R-A169-07	BRCA	TCGA-AR-A24X	491d5ebb-e9ab-4b17-acb8-630ff9fef0f2	A24X	FEMALE	WHITE
15	TCGA-B6-A0WV-01A-11R-A109-07	BRCA	TCGA-B6-A0WV	7588a286-6935-424a-bc5f-92e24c83eb37	A0WV	FEMALE	WHITE
16	TCGA-A8-A09W-01A-11R-A00Z-07	BRCA	TCGA-A8-A09W	452672e8-b6d9-45b5-bb02-e88e03e2dd66	A09W	FEMALE	[Not Available]
17	TCGA-C8-A8HQ-01A-11R-A36F-07	BRCA	TCGA-C8-A8HQ	6951FE6A-D7D1-4F41-85AF-76655B5D1F62	A8HQ	FEMALE	ASIAN

Survival data for each sample from Pan-Cancer Clinical Data Resource (Liu et al. 2018)



- **Overall survival** (OS): the period from the date of diagnosis until the date of death from any cause
  - Requires long follow ups not available for all TCGA samples
- **Disease-specific survival** (DSS): death from the diagnosed cancer type
  - Avoids counting deaths that are not due to the cancer
- **Progression-free interval**: the period from the date of diagnosis until the date of the first occurrence of a new tumor event (NTE)
- **Disease-free interval**: the period from the date of diagnosis until the date of the first new tumor progression event subsequent to the determination of a patient's disease-free status after their initial diagnosis and treatment

"Definitions and derivation of clinical survival outcome endpoints"

If you want to use this survival data, look at the paper to determine which metric can be used for your cancer type!

Using **pandas** to explore a dataset

- See example_EDA.py file for example **exploratory data analysis**
- What data is missing in the dataset or metadata?
- What is the minimum log(TPM+1) value? The maximum in the dataset?
- What is the average expression across all genes?
- Can you find the genes with the highest average expression?
- Can you make a scatterplot comparing the expression of two genes of interest?

NOTE: FOCUS YOUR ANALYSIS ON THE TRAINING DATA!!
WE WILL NOT NEED THE VALIDATION DATA UNTIL WE BUILD AN ML MODEL
WE WILL NOT NEED THE TEST DATA UNTIL EVALUATION OF THE MODEL

Homework

- **Due tomorrow 11:59pm:** Jupyter notebook check-in
  - **Disease background** (including **hallmark** chosen and **cancer type chosen**)
  - **Dataset** background
    - You should have information about what **genes** you will be looking at
    - For machine learning to “work well,” we want a good number of “features” (genes)
    - You should ensure that the gene/gene set you chose is actually in the dataset! Sometimes protein and gene names are different, so you may need to look for a few different names
    - Answer some of the exploratory data analysis questions as you write about the data.
  - Preliminary **question you’d like to answer**
- **Due 4/14 (by next class):** Andrew Ng videos (intro level) and comprehension quiz
  - These intro videos give overviews of supervised and unsupervised learning methods so we can hit the ground running next class!