Physician-Friendly Biological Topic Modeling using dLDA

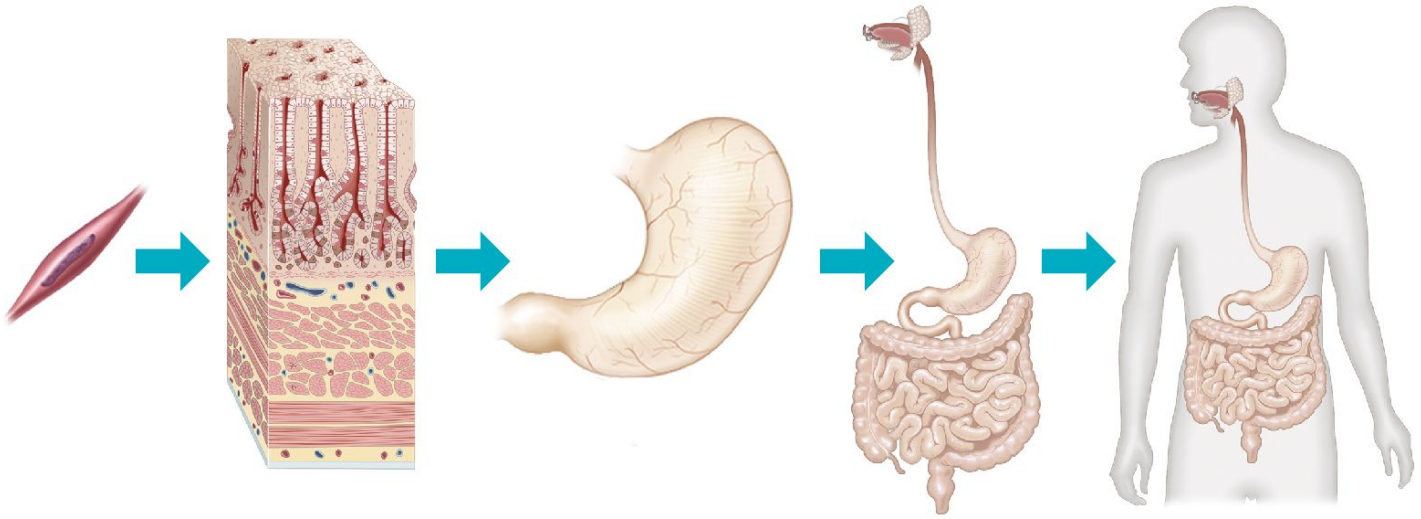
Sarah Davis, Amira Aissou

amii
ALBERTA MACHINE
INTELLIGENCE INSTITUTE

Intro to mRNA / Cancer Genetics

The Cell is the basic functional unit of life

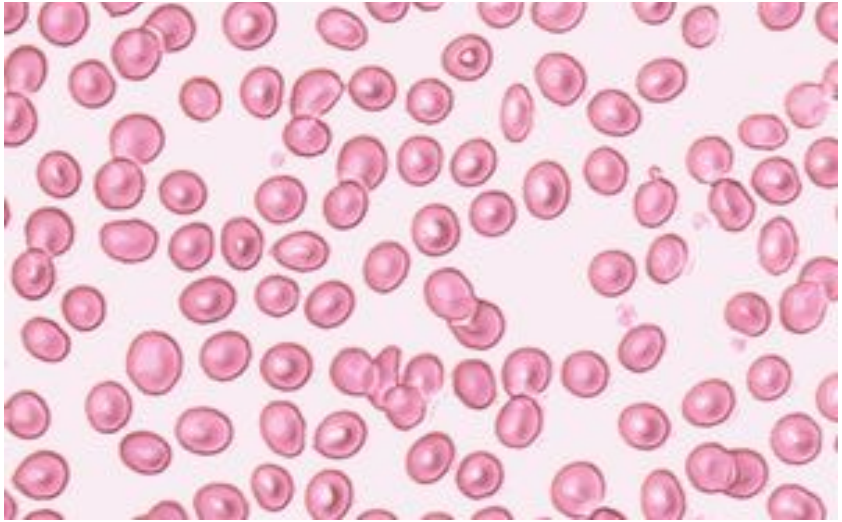
Levels of organization



© Encyclopædia Britannica, Inc.

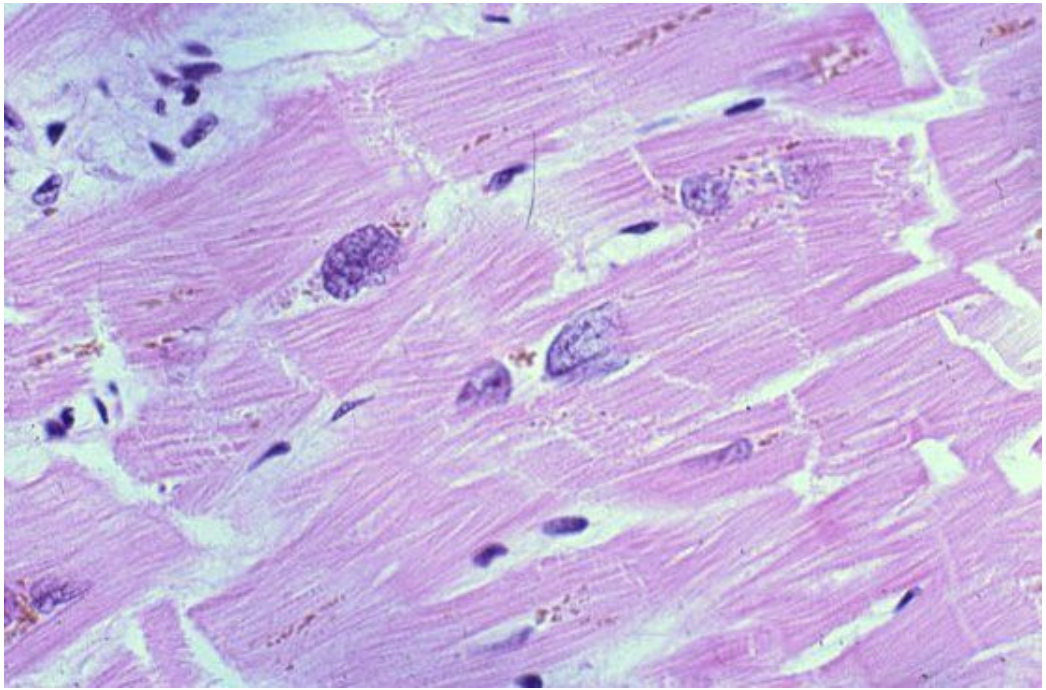
Cells can specialize in performing certain tasks...

What kind of cells are these?



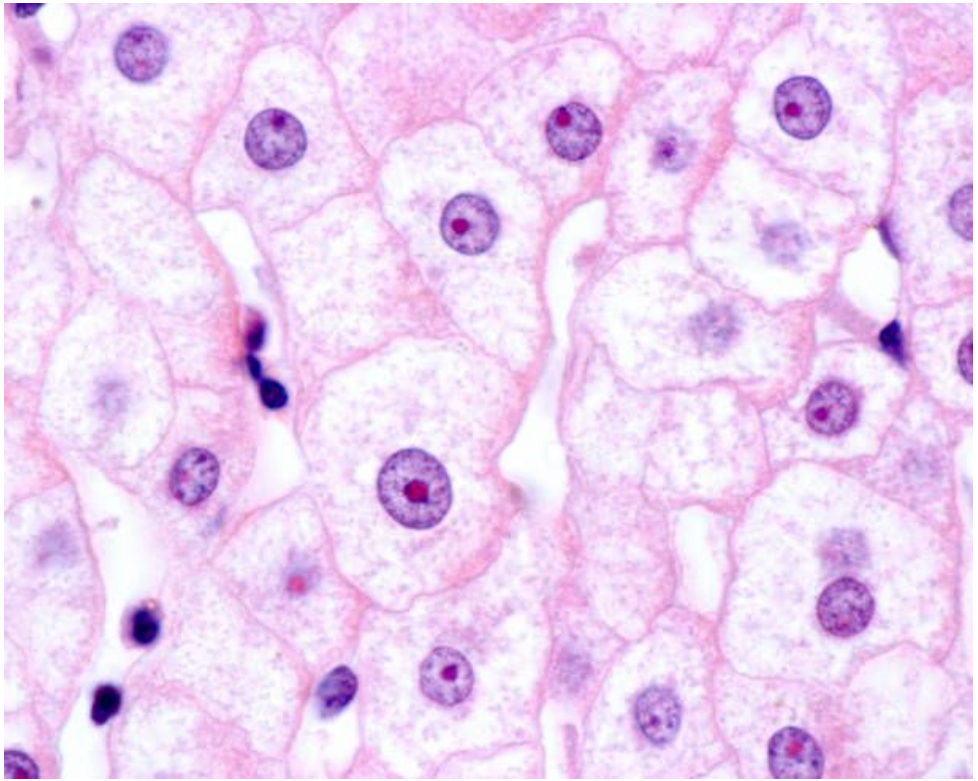
A. Neutrophils
B. Red Blood Cells
C. Platelets
D. Macrophages
E. Lymphocytes

What kind of cells are these?



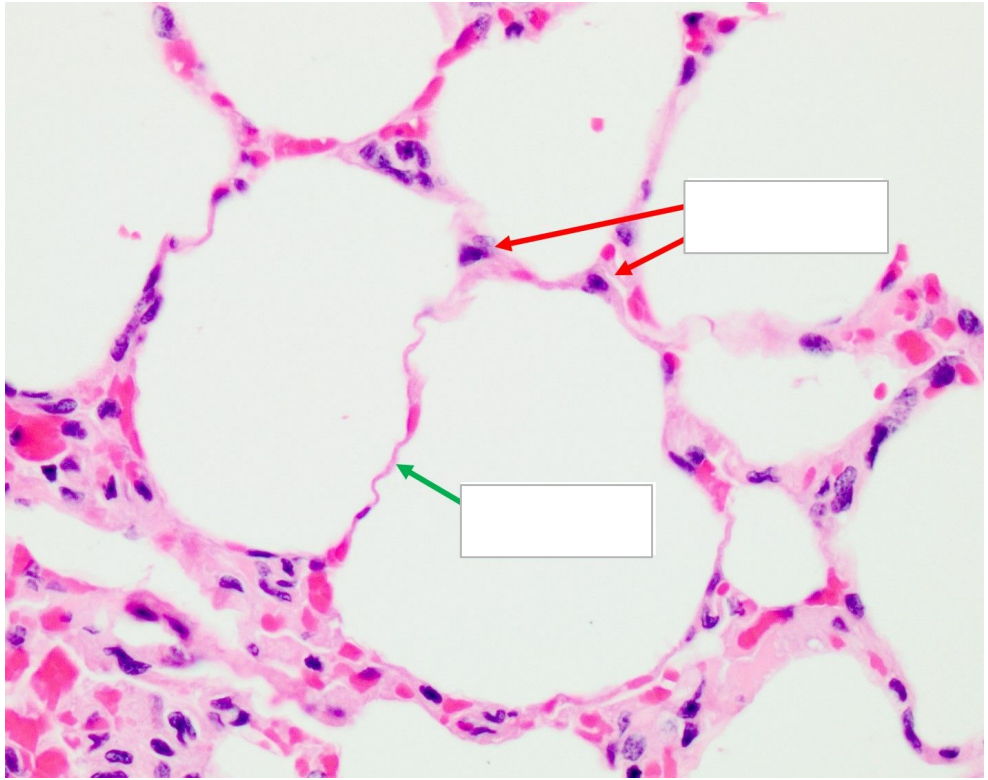
A. Skeletal Muscle Cells from the Rectus Abdominis
B. Longitudinal Muscularis Externa of the small bowel
C. Cardiac Muscle Cells
D. Muscle cells in the Tongue

What kind of cells are these?



A. Bladder Epithelial cells
B. Esophageal Epithelial cells
C. Spleen cells
D. Liver cells
E. Lung cells

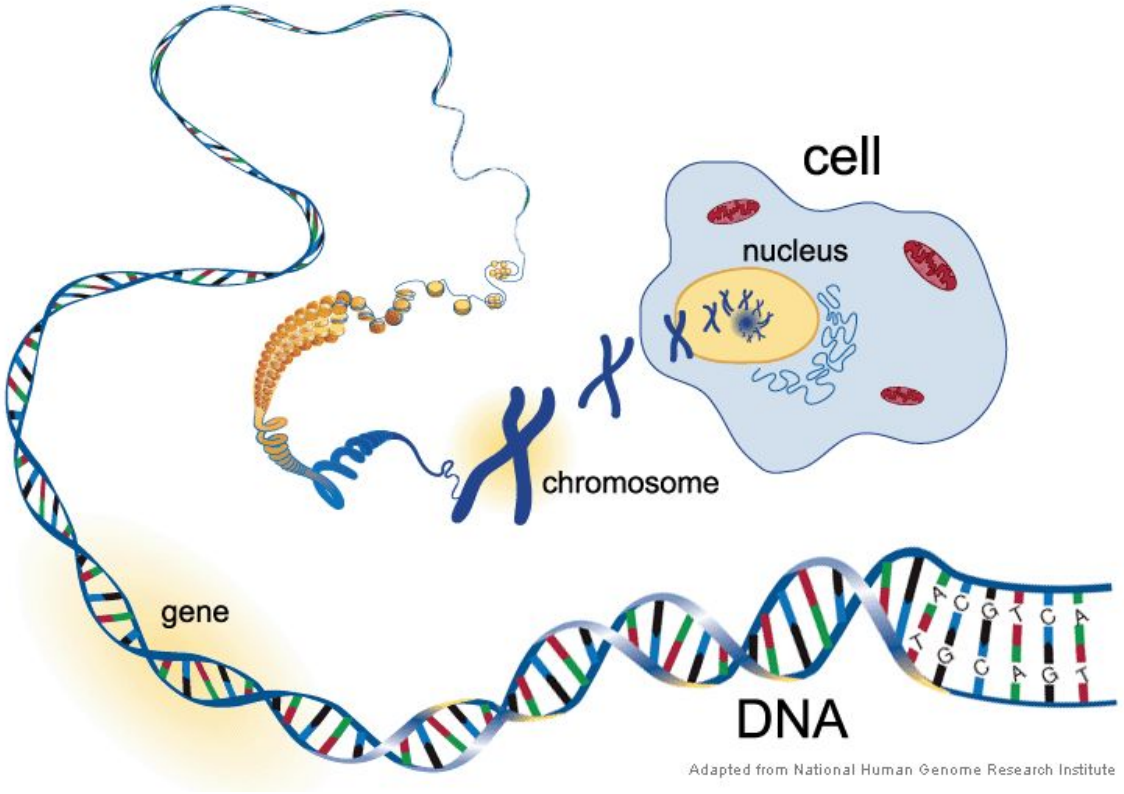
What kind of cell is the green arrow pointing at?



A. Type 1 pneumocyte (Lung cells)
B. Kupffer cell in the liver sinusoids
C. Lymphatic channel cell
D. Spleen wall cell

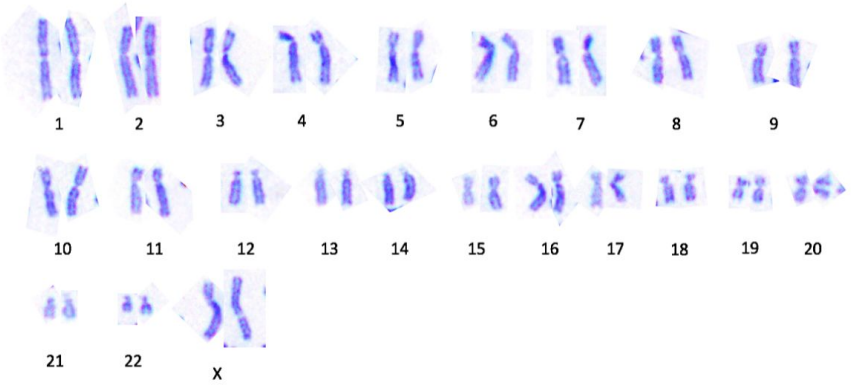
How can cells look and act different?

Genetics!

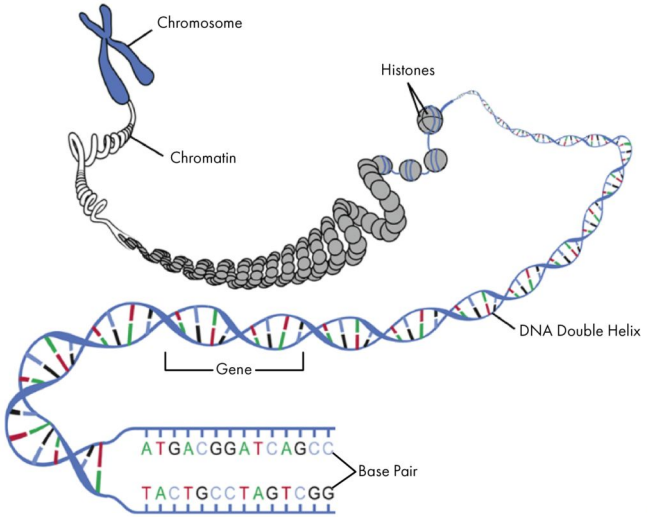

cell
nucleus
chromosome
gene
DNA
Adapted from National Human Genome Research Institute

The Human Genome is made up of...

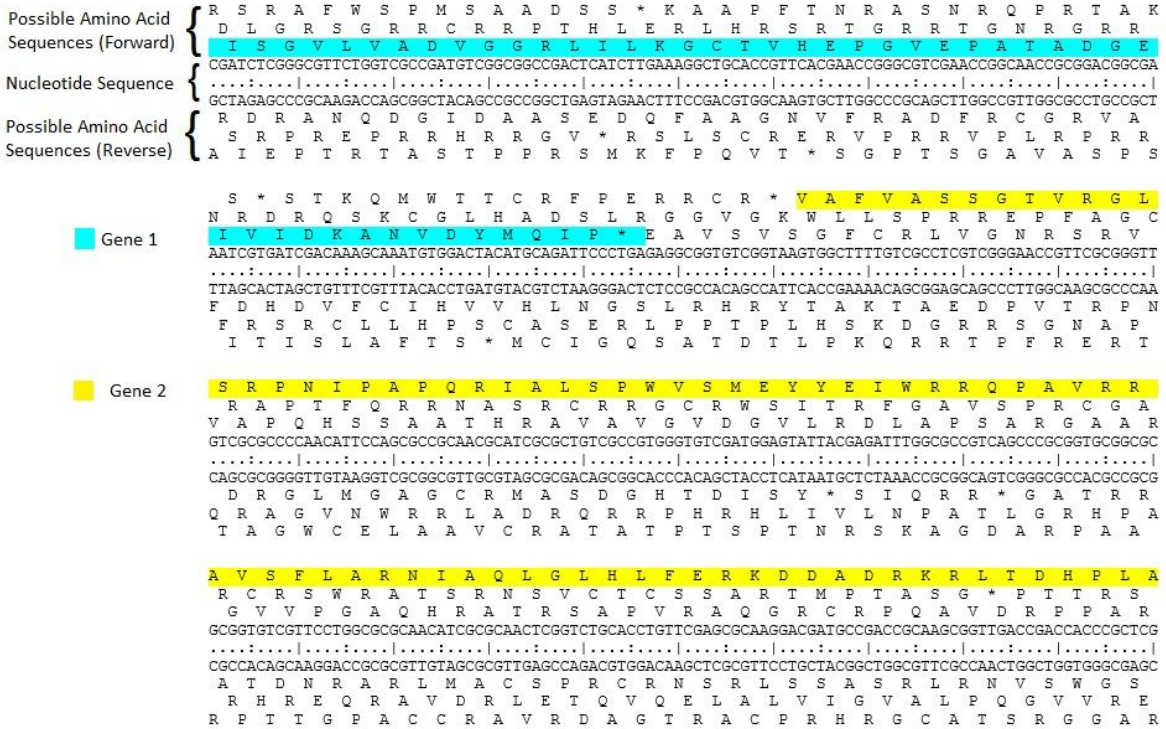
46 chromosomes (23 pairs)



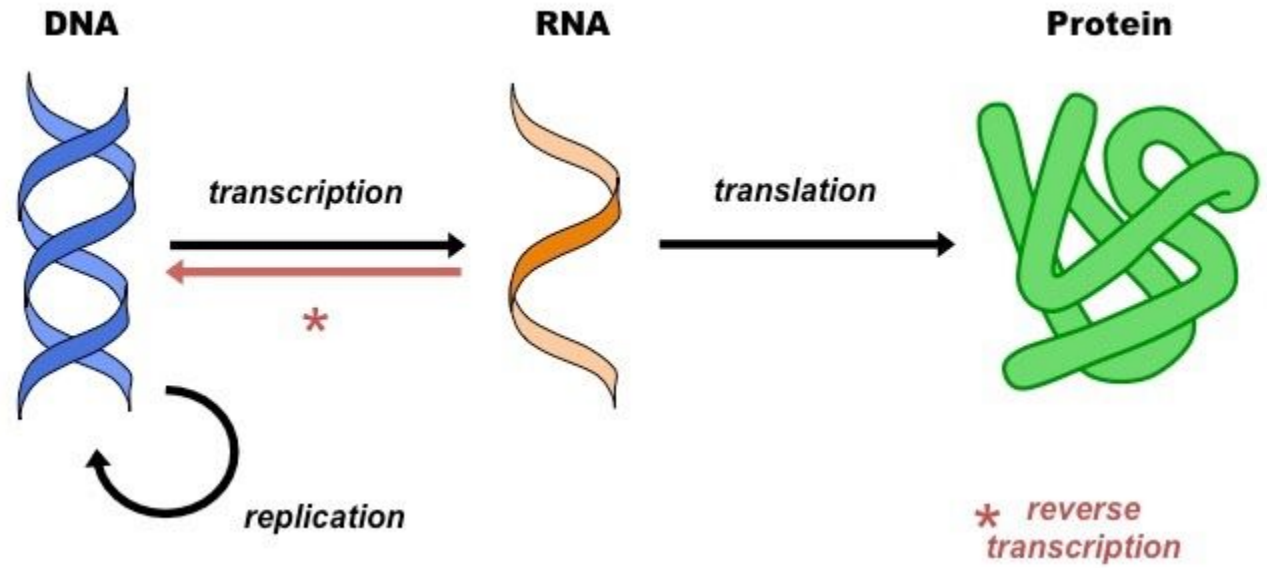
~ 3 200 000 000 nucleotides



Specific Nucleotide Sequences with functional properties are called genes



The Central Dogma of Biology



Genes can be protein-coding or non protein-coding……..

Approximately how many total genes (protein coding and non-protein coding) have we discovered so far?

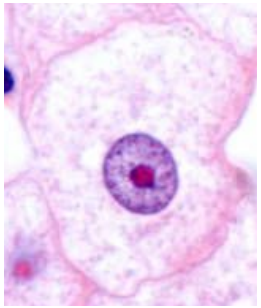
A. 10 000 - 20 000 genes
B. 30 000 - 40 000 genes
C. 60 000 - 70 000 genes
D. 90 000 - 100 000 genes

If all cells have the same genetic material, how can cells look and act differently?

Key Concept 1: Genes are NOT expressed equally

All cells have (roughly) the same genetic material and the same genes, BUT some **genes** can be **turned on** while others can be **turned off**.

For example:



Hepatocytes (Liver cells) will have **increased** expression of the **HPX gene** which codes for hemopexin



Type 1 Pneumocytes (Lung cells) will have **decreased** expression of the **HPX gene** which codes for hemopexin

What does this have to do with cancer?

What is Cancer?

Cancer is an **acquired** disease involving genetic mutations and **altered gene expression**.

All it takes is **one single cell** to have the right “cocktail” of mutations to become cancerous.

A cancer cell essentially becomes **immortal** and **divides out of control**; this may form a *solid mass tumor* or a *non-solid mass tumor* (like in leukemia). As the tumor grows it **invades healthy tissue** and becomes **less differentiated** compared to the original tissue.

The cancer can then **spread** to other places in the body and invade those tissues as well (called “metastases”)

Why is cancer a bad thing?

It can cause physical pain and suffering. Can prevent organs from effectively playing their role. This can give rise to complications which can cause suffering.

It can kill you. For you to remain alive, the cells in your body must receive oxygen and nutrients, and be able to get rid of waste and toxins. Anything that prevents this from happening will cause death.

How do we get rid of Cancer?

Surgery. A solid mass tumor can be surgically removed, but even if the cut is made well beyond the tumor periphery, there is almost always one single cancer cell left over.

Radiation. Usually done after surgery or in conjunction with other therapies. High energy beams (EMR or particles) designed to kill cells.

Chemotherapy. Using drugs to target a specific tumour's "weakness" to kill it. Many agents available and are dependant on the tumor type.

Immunotherapy. Improving or restoring body's natural cancer defence mechanisms.

The Big Problem: It will be a long time before we can fully understand all aspects of cancer genetics but people are still suffering and dying from this disease today.

There are many ways a cell can become cancerous.

Although there is a lot of research being done in cancer genetics, we are still a long ways away from identifying all the genetic factors and all the altered pathways.

How can we use our limited knowledge in the most optimal way possible?

The Big Problem: Cancer is so complex, how can we use our limited knowledge in the best way possible???

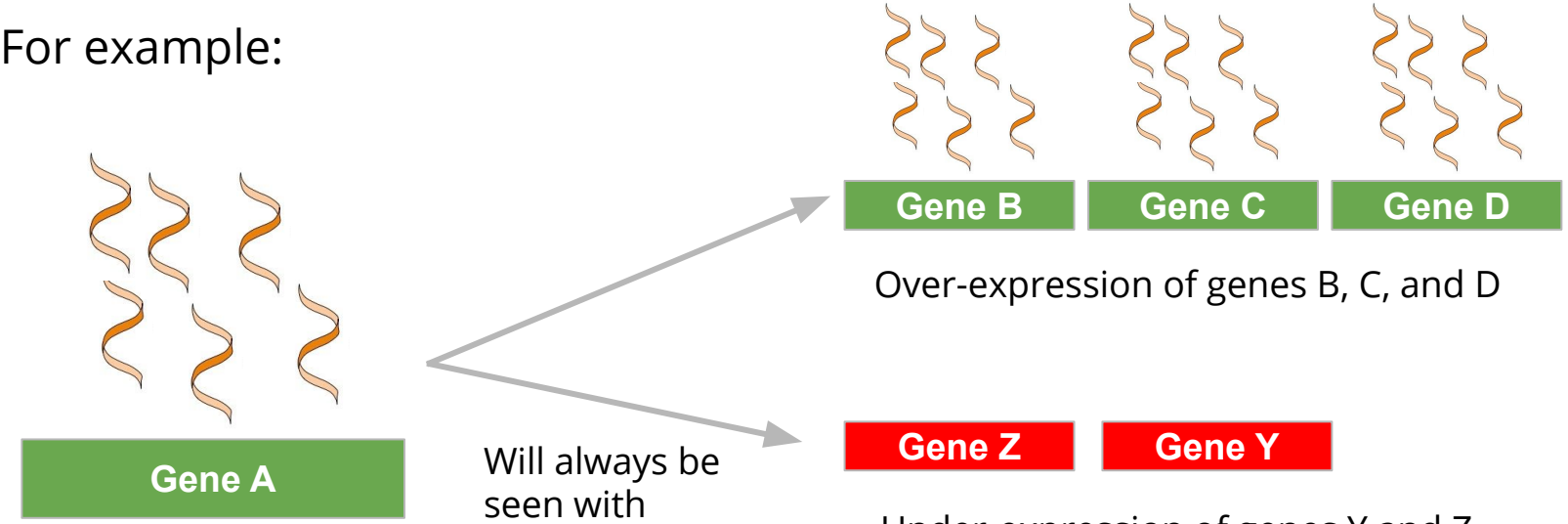
Since cancer is disease described by altered gene expression, maybe we can give this information to a machine? And it can help us?

Another big problem: The curse of dimensionality

There is TOO MUCH genetic information to describe a tumor. We need to somehow reduce the number of variables while preserving as much important information as possible.

Key Concept 2: Genes Move in Herds

For example:



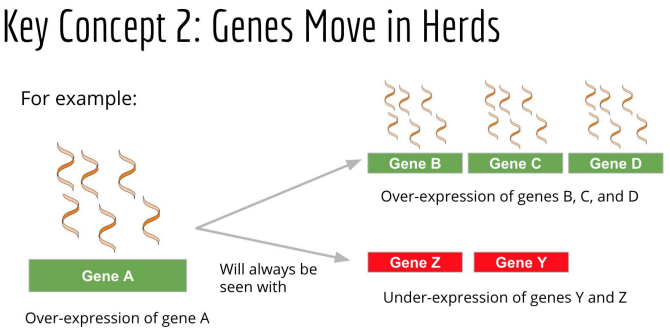
Can machines identify patterns in gene expression?

Another big problem: The curse of dimensionality

There is **TOO MUCH** genetic information to describe a tumor. We need to somehow reduce the number of variables while preserving as much important information as possible.

A way around this problem: Using the “genes move in herds” concept

Get the machine to identify patterns in gene expression and use each pattern as a single variable. These new variables (patterns) can be used to describe a patient’s tumor.



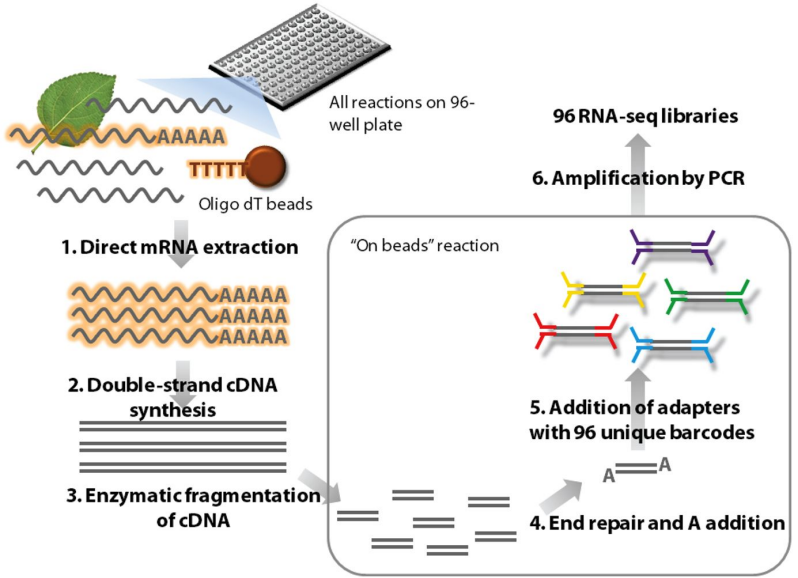
Data

How can we measure gene expression?

Many methods -- our project uses illumina RNA seq

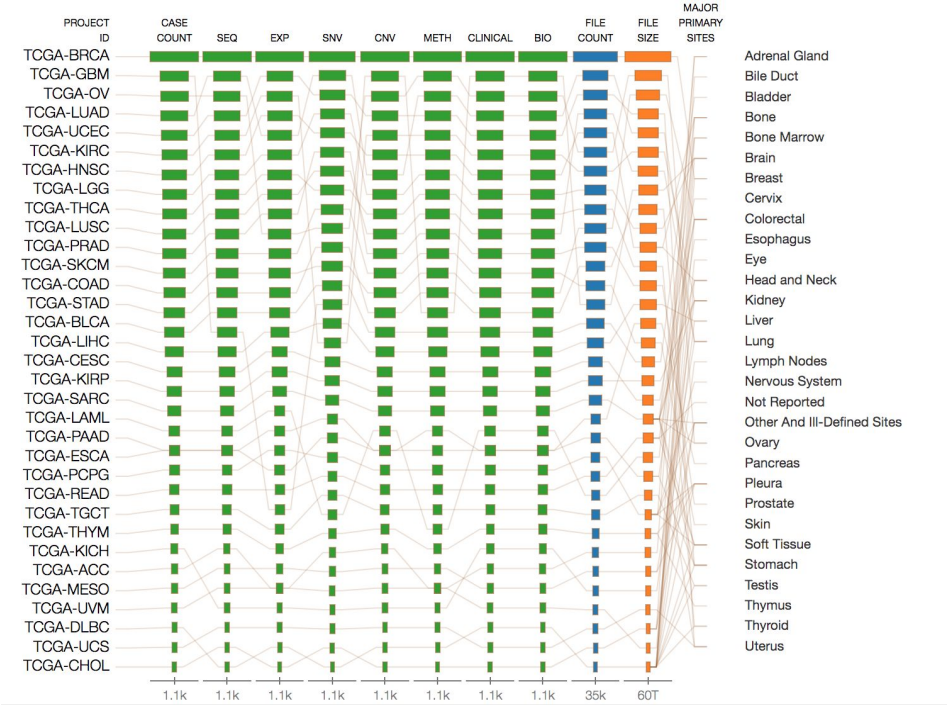
Measurement of the concentration of messenger RNA molecules from transcription → This relates to the degree of gene expression.

~20 000 mRNA measurements



Source of Cancer **Transcriptomic** and **Clinical** Information: **The Cancer Genome Atlas**





Transcriptomic Data available:

- 33 cancer types (solid mass tumors and non solid mass tumors)
- Metastasis
- Pseudo-normal

Transcriptomic Data used:

- 32 cancer types (Solid mass tumors only)
- Pseudo-normal

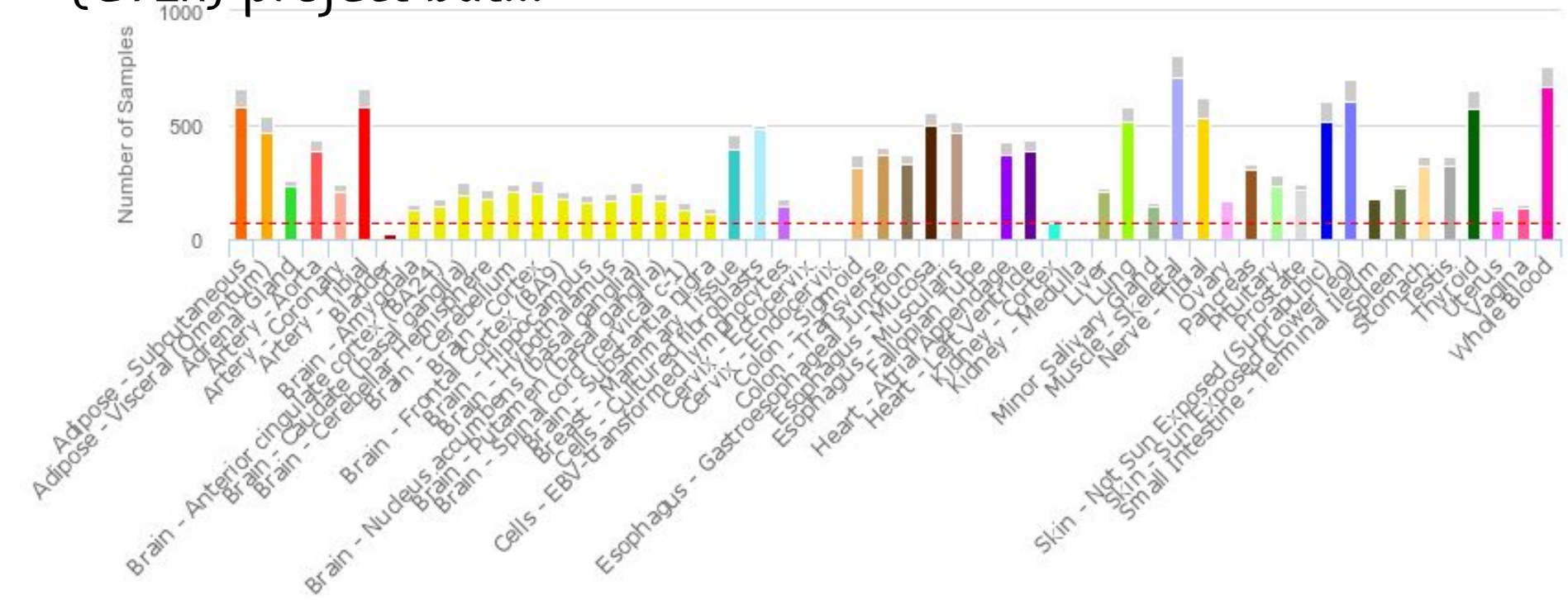
Clinical Data available:

- Age, vital status, days to death, days to last follow up, tumor site, path stage, T stage, N stage, M stage, sex, date of dx, rad therapy, histo type, pack yrs smoke, residual tumor, # of lymph nodes, race, ethnicity, etc.

Clinical Data used:

- Age, vital status, days to death, tumor site, sex, race, ethnicity.

We also wanted to include healthy data from the Genotype-Tissue Expression (GTEx) project but...



HREBA | Health Research Ethics Board of Alberta Cancer Committee



Home / Research Support / Research Ethics Office / Human Research Ethics

Research Ethics Office

- Research Ethics
- Animal Research Ethics +
- Human Research Ethics +
- Research Ethics Boards +
- Risk and Review Process
- Informed Consent
- The Use of Incentives In

Human Research Ethics

All research involving living human participants (or their data) or human biological materials conducted by staff or students affiliated with the University of Alberta, or involving University resources, must be reviewed and approved by one of the University's Research Ethics Boards (REB), or an REB desig... University before the research starts (ie. the University ha... cancer research to the HREBA-CC). Hum... may be funded or unfunded. ... and focus are... les,

REBs

- REB 1
- REB 2
- REB 3: Health Research Ethics Board - Health Panel
- REB 4: Health Research Ethics Board - Biomedical Panel

Veronica Coronado July 25, 2019 at 3:04 PM

To: Amira Aissiou Cc: Russ Greiner, Rene Dery, Pat Jones

Re: dbGaP data access request

VC

The database you are seeking access to is a "controlled-access" database. At UA you require ethics to use data within controlled-access databases. Please contact REO to get ethics in place.

HREBA | Health Research Ethics Board of Alberta Cancer Committee

Health Research Ethics Board of Alberta
Cancer Committee
1500, 10104 - 103 Avenue NW
Edmonton, Alberta, T5J 0H8
Telephone: (780) 423-5727
Fax: (780) 429-3509
Email: cancer@hreba.ca

Certification of Ethics Approval

This is to acknowledge that the following research has been reviewed and on behalf of the Health Research Ethics Board of Alberta (HREBA) – Cancer Committee (CC) I am granting approval for your site's participation in the research.

Ethics ID:	HREBA.CC-19-0290
Principal Investigator:	Russell Greiner
Co-Investigator(s):	Lynne-Marie Postovit
Student Co-Investigator(s):	There are no items to display
Study Title:	Using discretized Latent Dirichlet Allocation to Identify Patterns in mRNA Described Cancerous and Healthy Tissue
Sponsor:	Natural Sciences and Engineering Research Council

Effective: 12-Aug-2019 **Expires:** 11-Aug-2020

Sep 4 2019

Amira Aissiou
To: Russ Greiner
Re: APPROVAL of your request [#83020-1] for access phs000424/GRU

September 4, 2019 at 3:28 PM
All Mail - Google



!!!!!!🎊🎊🎊🎊🎊🎊🎊

Request List

Approved (1) | Closed (1)

#	Study, Consent	Status	Expiration	Actions
Project #22504: Using discretized Latent Dirichlet Allocation to Identify Patterns in mRNA Described Cancerous and Healthy Tissue				
83020-1	Genotype-Tissue Expression (GTEx) (phs000424.v8.p2) General Research Use (phs000424.v8.p2.c1), NHGRI	✓ Approved GRANTED	2020-09-04	Request Files Processing History Revise project

This is an automated message from NCBI dbGaP (the Database of Genotypes and Phenotypes) Authorized Access system. Do not reply back to this message or send email to dbgap-reply@ncbi.nlm.nih.gov

Dear Russ Greiner,

NIH has **APPROVED** your request [**#83020-1**] for the dataset **General Research Use** in **Genotype-Tissue Expression (GTEx)** access as part of your project titled #22504: "Using discretized Latent Dirichlet Allocation to Identify Patterns in mRNA Described Cancerous and Healthy Tissue " .

The following comments were provided:
Thank you for applying to dbGaP. Please note all external collaborators must independently apply for dbGaP approved access.

- Before accessing the data, please **REVIEW** the terms of access of the Data Use Agreement that you and have signed.
- All external collaborators must submit an independent Project Request from their institution and be approved to access the datasets before any data may be shared or exchanged

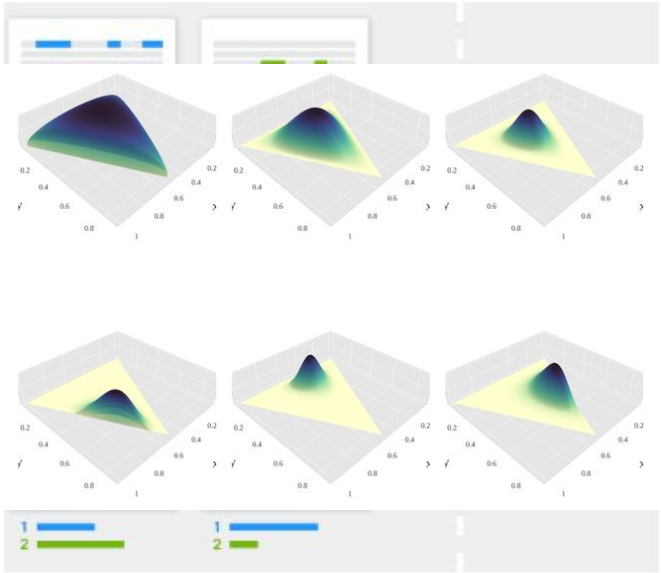
Intro to Topic Modeling / LDA

What is Topic Modeling?

Topic modeling is a type of statistical modeling used to extract abstract “topics” from a collection of documents.

LDA is an example of topic modeling tool; It builds a topic-per-document model and a words-per-topic model using Dirichlet distributions.

Also a form of feature extraction / dimensionality reduction.



dLDA

- discretized Latent Dirichlet Allocation
- Proposed in Luke Kumar's Master's thesis
- Prepocesses mRNA data so it's more LDA friendly, including:
  - Normalization, discretization, and corrections
  - Splitting gene expression into upregulated and downregulated
  - Imputing missing values

	TCGA-2W-A8YY-01	TCGA-4J-AA1J-01	TCGA-BI-A0VR-01
A1BG|1	198.83	63.82	70.02
A1CF|29974	0	1	0
A2BP1|54715	2	10	0
A2LD1|87769	282.72	257.99	178.56
A2ML1|144568	164	7819	1587
A2M|2	7879.81	9592.92	7923.89
A4GALT|53947	320	2261	2319



	TCGA-2W-A8YY-01	TCGA-4J-AA1J-01	TCGA-BI-A0VR-01
A1BG|1_LOW	0	0	0
A1BG|1_HIGH	0	0	0
A2BP1|54715_LOW	8	4	7
A2BP1|54715_HIGH	0	0	0
A2LD1|87769_LOW	0	0	0
A2LD1|87769_HIGH	0	0	0
A2ML1|144568_LOW	0	0	0
A2ML1|144568_HIGH	0	3	0
A2M|2_LOW	0	0	0
A2M|2_HIGH	2	3	2
A4GALT|53947_LOW	0	0	0
A4GALT|53947_HIGH	0	0	0

Example



Documents

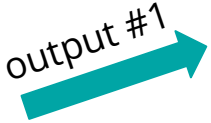
	"Film"	"Car"	...	"The"
Document 1	0	6	...	18
Document 2	8	1	...	12
Document 3	1	0	...	25

Word Count Representation

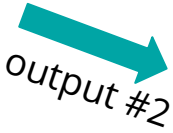
Example

	"Film"	"Car"	...	"The"
Document 1	0	6	...	18
Document 2	8	1	...	12
Document 3	1	0	...	25

Word Count Representation



LDA



	Topic 1	Topic 2	Topic 3
Document 1	10^-7	0.7	0.3
Document 2	0.8	10^-4	0.2
Document 3	0.1	10^-7	0.9

Topic Per Document Model

	"Film"	"Car"	...	"The"
Topic 1	10^-7	0.7		0.1
Topic 2	0.5	0.2		0.2
Topic 3	0.1	10^-7		0.8

Words Per Topic Model

Example

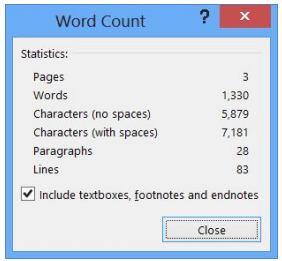
Can we apply this to biological systems…?



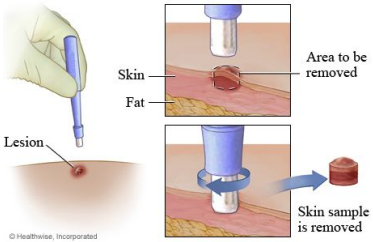
Document = Biopsy



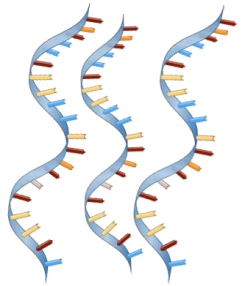
Words = Genes



Word Counts = mRNA Counts



CAPZB	HSPB1	MDH1
PRDX6	DNAJA1	CAPN2
CRYAB	HSPA1B	CAPN1
DES	HSPA8	MYPBH
PARK7	HSPA9	MYH7
ENO1	LDHB	MYH2



Example

	ABCA12	ANTXR1	...	ZIC2
Person 1	0	6	...	18
Person 2	8	1	...	12
Person 3	1	0	...	25

mRNA Count Representation

LDA

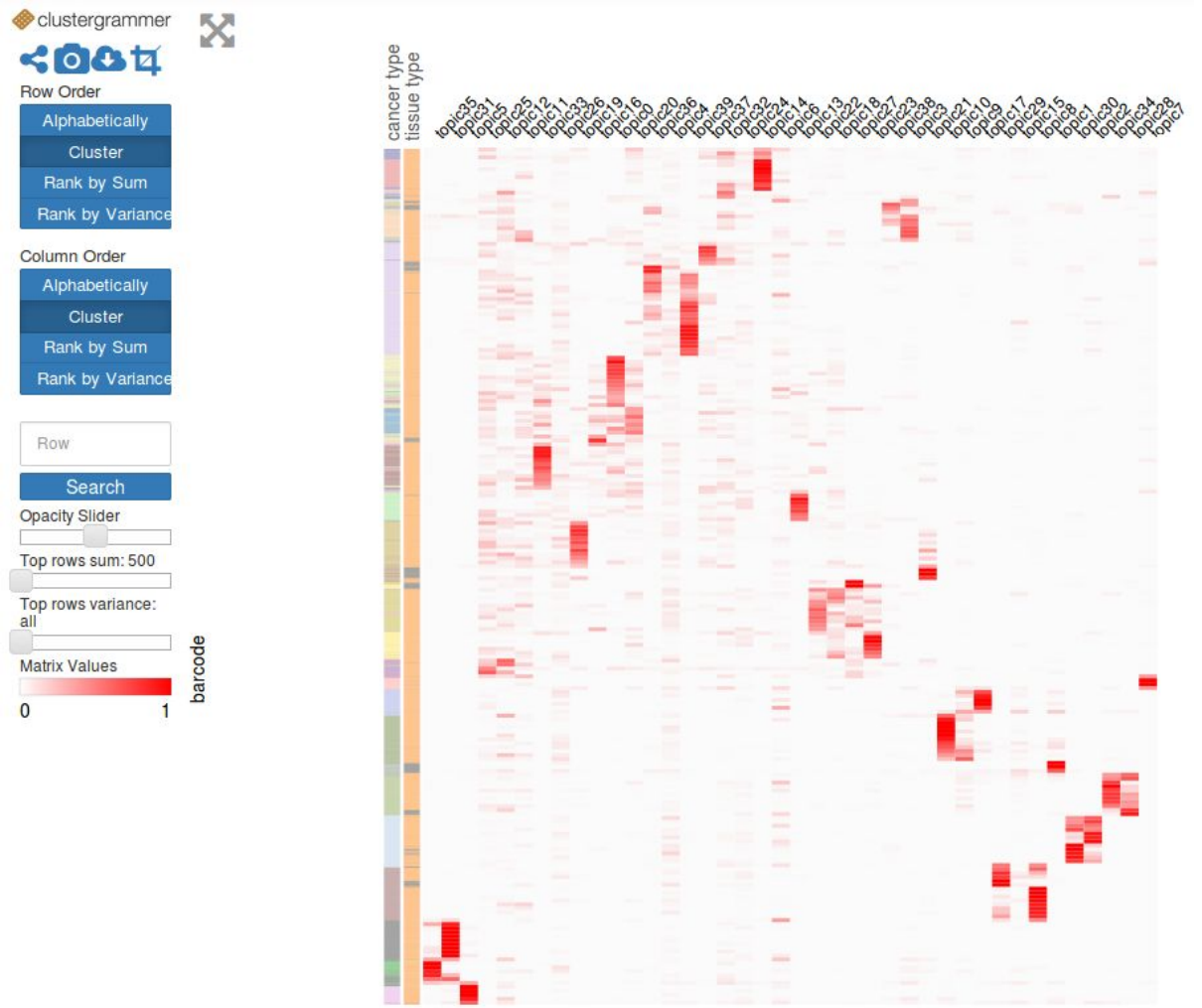
	Topic 1	Topic 2	Topic 3
Person 1	10^-7	0.7	0.3
Person 2	0.8	10^-4	0.2
Person 3	0.1	10^-7	0.9

Topic Per Biopsy Model

	ABCA12	ANTXR1	...	ZIC2
Topic 1	10^-7	0.7		0.1
Topic 2	0.5	0.2		0.2
Topic 3	0.1	10^-7		0.8

Gene Per Topic Model

Topic Per Biopsy Heatmap



ML Validation

Why?

- Since LDA can be used as a form of feature extraction, we wanted to make sure classification tasks on those features did reasonably well
  - We want as little data loss as possible
- Used simple classifiers (Logistic Regression, Random Survival Forests) to test accuracy

What Data did we Test it On?

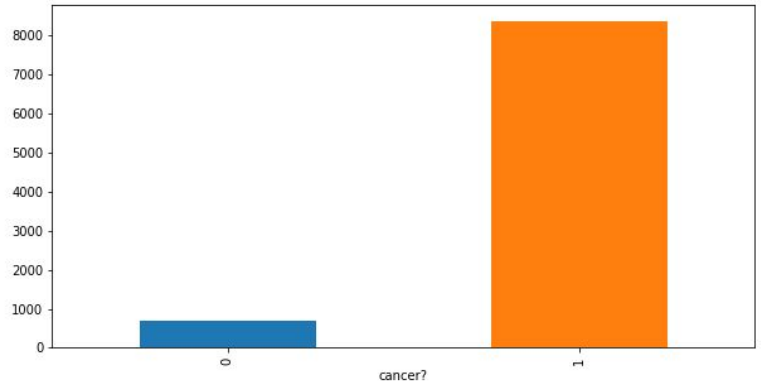
- Normalized mRNA counts, taken directly from TGCA (~20k attributes)
- Discretized, transformed data (as per Luke's thesis) (~26k attributes)
- Topics memberships / dLDA output (40 attributes)

	TCGA-2W-A8YY-01	TCGA-4J-AA1J-01	TCGA-BI-A0VR-01
A1BG|1_LOW	0	0	0
A1BG|1_HIGH	0	0	0
A2M|2_HIGH	2	3	2
A4GALT|53947_LOW	0	0	0
A4GALT|53947_HIGH	0	0	0

	topic0	topic1	topic2
TCGA-3X-AAV9-01A-72R-A41I-07	0.1121745685	1.27557528445897E-06	1.27557528445897E-06
TCGA-3X-AAVA-01A-11R-A41I-07	0.1349116088	6.83708059821245E-05	9.29748224183867E-07
TCGA-3X-AAVB-01A-31R-A41I-07	0.3774889377	1.10884414087062E-06	1.10884414087062E-06
TCGA-3X-AAVC-01A-21R-A41I-07	0.1367651416	9.48154890590014E-07	9.48154890590014E-07
TCGA-3X-AAVE-01A-11R-A41I-07	0.1065216935	1.03263114415951E-06	1.03263114415951E-06

Cancer vs. No Cancer Task

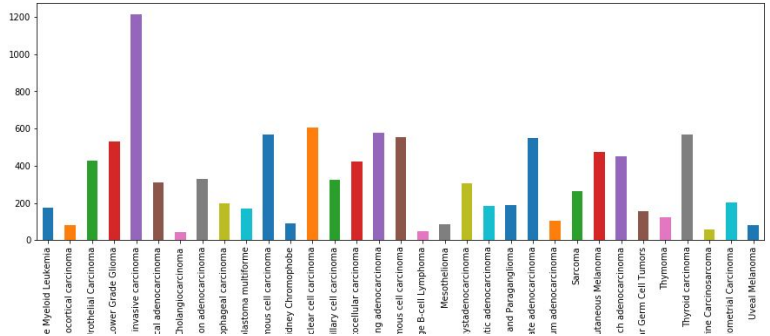
- Undersampled to account for class imbalance
- 5 fold cross validation
- No parameter tweaking



Accuracies		
	Logistic Regression	RSF
Counts	0.948	0.958
Discretized	0.982	0.962
Topics	0.945	0.930

Type of Cancer Task

- Used an multinomial approach for LogReg
- 5 fold cross validation



Accuracies		
	Logistic Regression	RSF
Counts	0.927	0.936
Discretized	0.968	0.916
Topics	0.894	0.913

Survival Prediction

- We tested to see how well we could predict survival from the topics

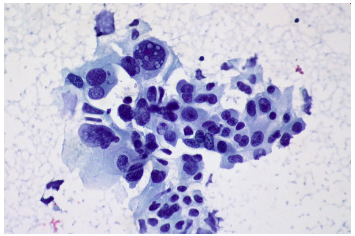
Predictor Name	Finished Time	Data Set	Status	Valid?	Public?	CI	χ^2	χ^2 p-value	
TCGA dLDA Output Data	about 1 year ago	9062 Instances 40 Features	Finished	yes	Private	0.75	7.88	0.97	Delete

- Model had a concordance index (pairwise comparison) of 0.75

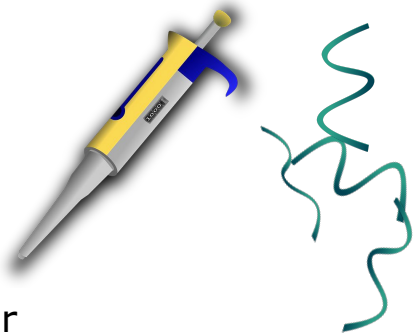
Gene Sets

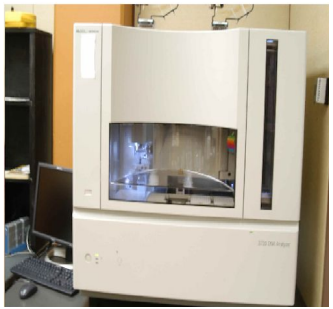
Wet lab Experiments are being conducted to measure gene expression in different conditions

For example:



epithelial lung cancer cell lines over-expressing the KRAS gene.





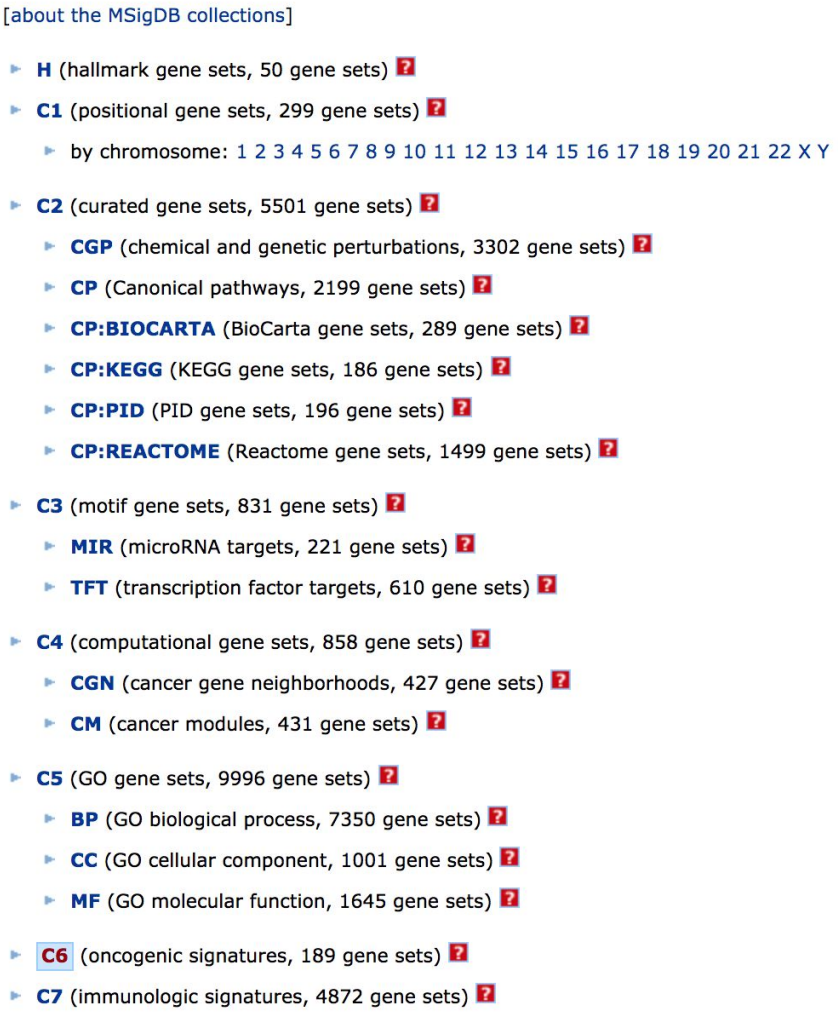
ABCC6
ABCG5
AFP
AKT2
ALLC
ASIP
ATG10
ATXN3L .....

Genes **up-regulated** in epithelial lung cancer cell lines over-expressing KRAS gene.

There are many genesets in the database for various conditions

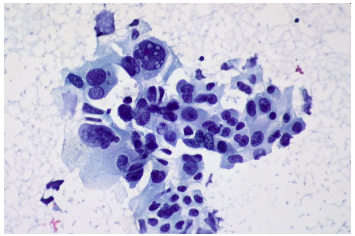
These genesets (~2300) can provide information on:

- Essential processes
- How single gene mutation alters expression of other genes
- Genetic expression of therapeutically resistant mice/cell lines
- Etc.



Format of genesets:

Lists of **over**expressed (up-regulated, **UP**) genes and **under**expressed (down- regulated, **DN**) genes



epithelial lung cancer cell lines over-expressing the KRAS gene.

KRAS.AMP.LUNG_UP.V1_UP	KRAS.AMP.LUNG_UP.V1_DN
ABCC6 ABCG5 AFP AKT2 ALLC ASIP ATG10 ...	… DTNB EMILIN2 ENPP1 EPB41 EPYC FCGR2B ...

Making the topics Physician-friendly by associating them to gene-sets

This allows the topics to be associated to specific pathways, conditions, and therapeutic response.

Thus, this has **descriptive** and **predictive** powers.

We have **gene sets**



We have **genes-per-topic** model

	ABCA12	ANTXR1	...	ZIC2
Topic 1	10^-7	0.7		0.1
Topic 2	0.5	0.2		0.2
Topic 3	0.1	10^-7		0.8

Can we associate the two to give a name to our nameless topic?

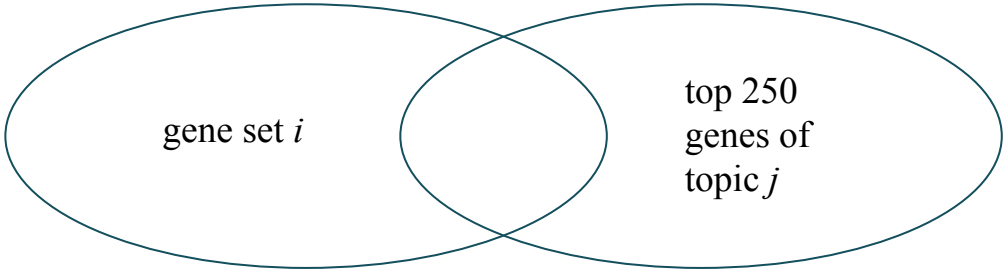
Associating Topics and Gene Sets

Attempt #1

Attempt #1

- Using a simple intersection between a gene set and the top **n** genes contributing to a topic
- The “score” between a topic and a gene set is the number of genes in the intersection

$$\text{score} = |\ \text{GS}_i\ \cap\ \text{Topic}_j\ |$$



- The top gene sets were then sorted by an oncology collaborator in an attempt to sort out what each topic meant

Topic	1	2	3	4
UP	Nasopharyngeal carcinoma (NPC), ovarian cancer (low metastatic potential), breast cancer, bile duct carcinoma, liver cancer	Brain specific genes	Migration, metastatis	Hypoxia response
	Endocrine therapy resistance in breast cancer	Mitochondrial, OXPHOS, metabolism genes	CDH1 (E-cad) targets	
	BMP, TGF-b, SMAD2/3 targets	Stem cell genes		
	Cell migration, EMT	UV response		
		Response to dox, progesterone, tosedostat, tretinoin		
DOWN	Epigenetically regulated genes (targets of histone methylating enzymes or chromatin remodellers)	Brain specific genes	MAPK targets	Immune, inflammatory response
	PRC2, SMARCA2, HDAC targets	Epigenetically regulated genes (PRC2, SMARCA2)	Epigenetically regulated genes	Colorectal, bile duct, nasopharyngeal, breast, prostate, advanced gastric, esophagus cancer
	Brain, head and neck cancer	Liver cancer	Endocrine therapy resistance	EMT genes
	Breast cancer, prostate, brain/head and neck cancer, liver cancer		Neuroactive ligand-receptor interaction, neuronal system	Gefitinib resistance
			Metabolism	KRAS/MAPK signaling

Mathematical Viability vs. Biological Viability

Viability

- We want to give a score or a probability for the strength of association between a gene set and a topic
- Once we have the scores for every gene set and topic combination (a 40x2300 matrix), we have two options for calculating Z-scores
  - Z scoring across topics
  - Z scoring across gene sets

Viability

- Z scoring across topics maintains the order of the scoring, and does not mathematically change what we are measuring
  - However, gene sets that are highly associated with every topic (e.g., housekeeping genes) can push out specific or more subtle gene sets important to the topic
- Z scoring across gene sets will pull out strong biological signals, but z-scores calculated this way are not necessarily mathematically comparable to each other anymore
  - However, much more useful for a physician

Formula #2

Formula 2.1

$$P(GS \mid topic) = \prod_{g \in GS^+} P(+g \mid topic) \prod_{g \in GS^-} P(-g \mid topic) \prod_{g \notin GS} P(0g \mid topic)$$

$$P(g+ \mid topic) = \frac{e^{\epsilon}}{e^{\epsilon} + e^{\delta} + e^{\gamma}}$$

$$P(g- \mid topic) = \frac{e^{\delta}}{e^{\epsilon} + e^{\delta} + e^{\gamma}}$$

$$P(0g \mid topic) = \frac{e^{\gamma}}{e^{\epsilon} + e^{\delta} + e^{\gamma}}$$

ϵ = the amount gene_iUP contributes to the topic

δ = the amount gene_iDN contributes to the topic

γ = some constant

Formula 2.1

- We were told by our oncology collaborator that our formula should not reward or penalize based on genes being not included in the gene set
  - Based on hypothesis testing (how including a gene in a gene set is determined) we can only draw statistical conclusions about genes that are **in** the gene set
  - We cannot say “this gene is not in the gene set, therefore it is not important”

- On to attempt #2...

Formula #2.2

$$P(GS \mid topic) = \frac{\prod_{g \in GS^+} P(+g \mid topic) \prod_{g \in GS^-} P(-g \mid topic)}{\mid GS \mid}$$

$$P(g+ \mid topic) = \frac{e^{\epsilon}}{e^{\epsilon} + e^{\delta}}$$

$$P(g- \mid topic) = \frac{e^{\delta}}{e^{\epsilon} + e^{\delta}}$$

ϵ = the amount gene_iUP contributes to the topic

δ = the amount gene_iDN contributes to the topic

Formula #2.2 Results

- Correcting for gene set size seemed natural -- each score was approximately 0.5*|GS|

- Despite this, formula suddenly favoured many of the same very small gene sets (each with 5-6 genes)

Z-scored Across Topics

Topic	Manual Description Based on Heatmap	Top 10 Positively Associated Gene Sets	Highest Z Scores
topic0	A bit of Cervical squamous cell carcinoma and endocervical adenocarcinoma, a lot of Head and Neck squamous cell carcinoma. Lots of randoms littered in too.	MYLLYKANGAS_AMPLIFICATION_HOT_SPOT_14	14.10717832
		MYLLYKANGAS_AMPLIFICATION_HOT_SPOT_30	12.24293131
		KONDO_COLON_CANCER_HCP_WITH_H3K27ME3	11.72141114
		MYLLYKANGAS_AMPLIFICATION_HOT_SPOT_11	10.80401355
		MYLLYKANGAS_AMPLIFICATION_HOT_SPOT_13	10.80295954
		MYLLYKANGAS_AMPLIFICATION_HOT_SPOT_25	10.80238371
		MYLLYKANGAS_AMPLIFICATION_HOT_SPOT_18	10.01239839
		MYLLYKANGAS_AMPLIFICATION_HOT_SPOT_16	9.45654739
		MYLLYKANGAS_AMPLIFICATION_HOT_SPOT_21	8.724966293
		MYLLYKANGAS_AMPLIFICATION_HOT_SPOT_8	8.72491547
topic1	Pseudonormal Kidney renal clear cell carcinoma and Kidney renal papillary cell carcinoma	MYLLYKANGAS_AMPLIFICATION_HOT_SPOT_14	14.10534277
		MYLLYKANGAS_AMPLIFICATION_HOT_SPOT_30	12.24340827
		KONDO_COLON_CANCER_HCP_WITH_H3K27ME3	11.72286876
		MYLLYKANGAS_AMPLIFICATION_HOT_SPOT_11	10.80425884
		MYLLYKANGAS_AMPLIFICATION_HOT_SPOT_13	10.80318687
		MYLLYKANGAS_AMPLIFICATION_HOT_SPOT_25	10.80175556
		MYLLYKANGAS_AMPLIFICATION_HOT_SPOT_18	10.01248595
		MYLLYKANGAS_AMPLIFICATION_HOT_SPOT_16	9.455951965
		MYLLYKANGAS_AMPLIFICATION_HOT_SPOT_27	8.725362279
		MYLLYKANGAS_AMPLIFICATION_HOT_SPOT_8	8.725159946
topic2	Half of the prostate adenocarcinoma cases	MYLLYKANGAS_AMPLIFICATION_HOT_SPOT_14	14.1056694
		MYLLYKANGAS_AMPLIFICATION_HOT_SPOT_30	12.24314159
		KONDO_COLON_CANCER_HCP_WITH_H3K27ME3	11.72279868
		MYLLYKANGAS_AMPLIFICATION_HOT_SPOT_11	10.80460802
		MYLLYKANGAS_AMPLIFICATION_HOT_SPOT_13	10.80319871
		MYLLYKANGAS_AMPLIFICATION_HOT_SPOT_25	10.80264697
		MYLLYKANGAS_AMPLIFICATION_HOT_SPOT_18	10.01253599
		MYLLYKANGAS_AMPLIFICATION_HOT_SPOT_16	9.45591008
		MYLLYKANGAS_AMPLIFICATION_HOT_SPOT_27	8.725045962
		MYLLYKANGAS_AMPLIFICATION_HOT_SPOT_8	8.724870505

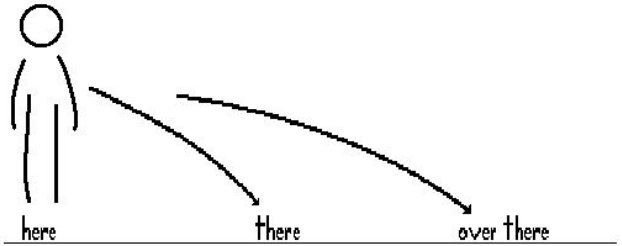
Z-scored Across Gene Sets

Topic	Manual Description Based on Heatmap	Top 10 Differentially Associated Gene Sets	Highest Z Scores
topic1	Pseudonormal Kidney renal clear cell carcinoma and Kidney renal papillary cell carcinoma	SU_KIDNEY	4.6023
		STEIN_ESRRA_TARGETS_RESPONSIVE_TO_ESTRO(	3.4517
		OHM_EMBRYONIC_CARCINOMA	3.218
		TANG_SENESCENCE_TP53_TARGETS	3.0469
		CROONQUIST_IL6_DEPRIVATION	2.9576
		MOOTHA_PYR	2.9514
		KOBAYASHI_EGFR_SIGNALING_24HR	2.8831
		FLORIO_NEOCORTEX_BASAL_RADIAL_GLIA	2.8428
		LE_NEURONAL_DIFFERENTIATION	2.6697
		MOLENAAR_TARGETS_OF_CCND1_AND_CDK4	2.5337
topic2	Half of the prostate adenocarcinoma cases	WANG_RESPONSE_TO_ANDROGEN	3.6497
		WANG_RESPONSE_TO_FORSKOLIN	3.4522
		YEGNASUBRAMANIAN_PROSTATE_CANCER	3.2029
		NELSON_RESPONSE_TO_ANDROGEN	3.1941
		WALLACE_PROSTATE_CANCER	2.9624
		TOMLINS_PROSTATE_CANCER	2.9545
		MIKKELSEN_NPC_ICP_WITH_H3K27ME3	2.8259
		DOANE_RESPONSE_TO_ANDROGEN	2.5982
		FARMER_BREAST_CANCER_CLUSTER_7	1.9544
		DOANE_BREAST_CANCER_CLASSES	1.8949
topic3	Strong in sarcoma, also strong in a jumble of random samples	HOLLERN_EMT_BREAST_TUMOR	3.4751
		COLDREN_GEFITINIB_RESISTANCE	3.2241
		LIEN_BREAST_CARCINOMA_METAPLASTIC_VS_DUC	2.9635
		IGARASHI_ATF4_TARGETS	2.816
		ENGELMANN_CANCER_PROGENITORS	2.7106
		NAKAMURA_TUMOR_ZONE_PERIPHERAL_VS_CENTI	2.633
		RODRIGUES_NTN1_TARGETS	2.5569
		ZWANG_EGF_INTERVAL	2.5342
		MYLLYKANGAS_AMPLIFICATION_HOT_SPOT_1	2.4765
		SABATES_COLORECTAL_ADENOMA_SIZE	2.4177

Formula #3

Using Distances





Subtracting the expression values for a gene in a topic and a gene in a geneset -- adding up the absolute values for each of these genes to find the total “distance” of the topic from the geneset.

Details:

- **We only consider the genes present in the geneset** (no penalty if a topic has high expression of a gene and the geneset does not have that gene listed)
- We normalize the genes in the genesets and those included in the topic
- Subtract and add the absolute values



For example:

Genes in geneset 1:

GS1_UPREGULATED	GS1_DOWNREGULATED
G1 **G4**	**G2**

	G1_UP	G1_DN	G2_UP	G2_DN	G3_UP	G3_DN	G4_UP	G4_DN	...
GS1	**1**	0	0	**1**	0	0	**1**	0	0
GS1 (norm)	**0.33**	0	0	**0.33**	0	0	**0.33**	0	0

	G1_UP	G1_DN	G2_UP	G2_DN	G3_UP	G3_DN	G4_UP	G4_DN	...
GS1	1	0	0	1	0	0	1	0	0
GS1 (norm)	0.33	0	0	0.33	0	0	0.33	0	0

	G1_UP	G1_DN	G2_UP	G2_DN	G3_UP	G3_DN	G4_UP	G4_DN	...
Topic 1	0.2	0	0.3	0	0.1	0	0.2	0	0.2
Topic 1 (reduced to GS1)	0.2	0	0.3	0	0	0	0.2	0	0
Topic 1 (reduced to GS1 & norm)	0.2/0.7 = **0.29**	0	0.3/0.7 = **0.43**	0	0	0	0.2/0.7 = **0.29**	0	0

	G1_UP	G1_DN	G2_UP	G2_DN	G3_UP	G3_DN	G4_UP	G4_DN	...
GS1 (norm)	0.33	0	0	0.33	0	0	0.33	0	0

[redacted]

Topic 1 (reduced to GS1 & norm)	0.2/0.7 = **0.29**	0	0.3/0.7 = **0.43**	0	0	0	0.2/0.7 = **0.29**	0	0

[redacted]

GS1-T1	0.04	0	-0.43	0.33	0	0	0.04	0	0

Total Distance and interpretation

GS1-T1	0.04	0	-0.43	0.33	0	0	0.04	0	0

Total distance =
|0.04|+|0|+|-0.43|+|0.33|+|0|+|0|+|0.04|+|0|+|0|
= **0.84**

0 = Perfect match
1 = As good as random
2 = Perfect mismatch

Formula Results using Real Data

Z-scored Across Topics

Topic	Manual Description Based on Heatmap	Top 10 Positively Associated Gene Sets	Highest Z Scores	Gene Set Size Pos
topic0	A bit of Cervical squamous cell carcinoma and endocervical adenocarcinoma, a lot of Head and Neck squamous cell carcinoma. Lots of randoms littered in too.	NGO_MALIGNANT_GLIOMA_1P_LOH	4.171393336	17
		AMIT_EGF_RESPONSE_20_MCF10A	3.782747733	14
		TESAR_ALK_TARGETS_HUMAN_ES_4D	3.64005991	6
		AZARE_STAT3_TARGETS	3.583310691	24
		FU_INTERACT_WITH_ALKBH8	3.551263232	13
		STANHILL_HRAS_TRANSFROMATION	3.468205033	8
		BOWIE_RESPONSE_TO_TAMOXIFEN	3.316259776	18
		MYLLYKANGAS_AMPLIFICATION_HOT_SPOT_11	3.274314752	5
		VERRECCHIA_RESPONSE_TO_TGFB1_C2	3.22324829	25
		BILANGES_SERUM_AND_RAPAMYCIN_SENSITIVE_GENES	3.203866549	68
topic1	Pseudonormal Kidney renal clear cell carcinoma and Kidney renal papillary cell carcinoma	NGO_MALIGNANT_GLIOMA_1P_LOH	4.920680801	17
		NIKOLSKY_BREAST_CANCER_19P13_AMPLICON	4.609942666	5
		STANHILL_HRAS_TRANSFROMATION	4.602453026	8
		MARIADASON_RESPONSE_TO_BUTYRATE_CURCUMIN_SULINDAC_TSA_2	4.548056576	8
		DING_LUNG_CANCER_MUTATED_RECURRENTLY	4.20671183	6
		BILANGES_SERUM_AND_RAPAMYCIN_SENSITIVE_GENES	4.195700467	68
		SU_KIDNEY	4.099212079	15
		FU_INTERACT_WITH_ALKBH8	4.011429506	13
		TERAMOTO_OPN_TARGETS_CLUSTER_3	3.918689805	5
		HEIDENBLAD_AMPLIFIED_IN_BONE_CANCER	3.918689805	7
topic2	Half of the prostate adenocarcinoma cases	MYLLYKANGAS_AMPLIFICATION_HOT_SPOT_2	6.749595558	5
		NGO_MALIGNANT_GLIOMA_1P_LOH	4.889897141	17
		BILANGES_SERUM_AND_RAPAMYCIN_SENSITIVE_GENES	4.851918399	68
		FU_INTERACT_WITH_ALKBH8	4.779872552	13
		LUI_TARGETS_OF_PAX8_PPARG_FUSION	4.72221498	34
		LUI_THYROID_CANCER_CLUSTER_3	4.714528026	28
		GAUTSCHI_SRC_SIGNALING	4.410817839	8
		STANHILL_HRAS_TRANSFROMATION	4.399565675	8
		MOOTHA_ROS	4.155275012	7
		TESAR_ALK_TARGETS_HUMAN_ES_4D	4.104249046	6

Z-scored Across Gene Sets

Topic	Manual Description Based on Heatmap	Top 10 Positively Associated Gene Sets	Highest Z Scores
topic0	A bit of Cervical squamous cell carcinoma and endocervical adenocarcinoma, a lot of Head and Neck squamous cell carcinoma. Lots of randoms littered in too.	RICKMAN_HEAD_AND_NECK_CANCER_C	5.3920643452
		HUPER_BREAST_BASAL_VS_LUMINAL	3.7708374185
		MURAKAMI_UV_RESPONSE_24HR	3.6149125477
		AUJLA_IL22_AND_IL17A_SIGNALING	3.5810793645
		SEIKE_LUNG_CANCER_POOR_SURVIVAL	3.4718121538
		WHITE_NEUROBLASTOMA_WITH_1P36.3_DELETION	3.4347008381
		OKAWA_NEUROBLASTOMA_1P36_31_DELETION	3.3422453355
		HOLLERN_SQUAMOUS_BREAST_TUMOR	3.3243127532
		GHANDHI_BYSTANDER_IRRADIATION	3.2821639761
		RAY_ALZHEIMERS_DISEASE	3.2356020603
topic1	Pseudonormal Kidney renal clear cell carcinoma and Kidney renal papillary cell carcinoma	SU_KIDNEY	4.86183044
		DIERICK_SEROTONIN_FUNCTION_GENES	3.6159066589
		PIONTEK_PKD1_TARGETS	3.3160797147
		TRACEY_RESISTANCE_TO_IFNA2	3.2170266897
		HEIDENBLAD_AMPLIFIED_IN_BONE_CANCER	3.0487402201
		LI_WILMS_TUMOR_VS_FETAL_KIDNEY_1	3.0010993091
		KINNEY_DNMT1_METHYLATION_TARGETS	2.8708096801
		ZEMBUTSU_SENSITIVITY_TO_VINCRISTINE	2.7854068637
		MOOTHA_GLUCONEOGENESIS	2.7691528834
		HOLLERN_PAPILLARY_BREAST_TUMOR	2.726830187
topic2	Half of the prostate adenocarcinoma cases	KONDO_COLON_CANCER_HCP_WITH_H3K27ME3	4.0691480186
		GALI_TP53_TARGETS_APOPTOTIC	3.9821186486
		TOMLINS_PROSTATE_CANCER	3.3110617282
		YEGNASUBRAMANIAN_PROSTATE_CANCER	3.2866181273
		SETLUR_PROSTATE_CANCER_TMPRSS2_ERG_FUSION	3.0035614442
		MIKKELSEN_NPC_ICP_WITH_H3K27ME3	2.9552544513
		NELSON_RESPONSE_TO_ANDROGEN	2.8428952219
		WANG_RESPONSE_TO_ANDROGEN	2.8425803905
		MEISSNER_BRAIN_HCP_WITH_H3_UNMETHYLATED	2.7603066142
		WANG_RESPONSE_TO_FORSKOLIN	2.7014818033

Next Steps

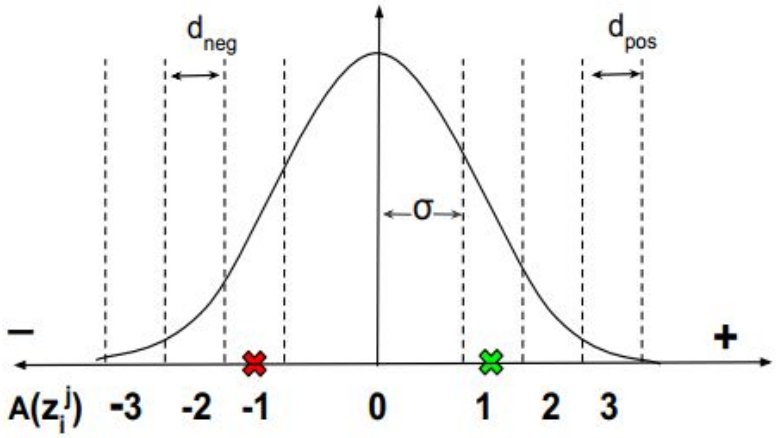
Next Steps

- Alternative methods for finding the probability of a geneset given a topic
- Testing the algorithm and geneset associations using only healthy tissue data
- Extracting features from a dataset comprised of cancer, pseudo-normal, and healthy (non-cancer) tissue

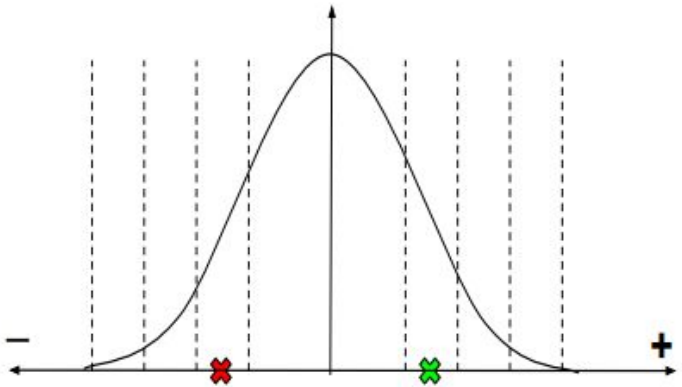
Submit a paper

Thank You!

Questions?



$A(\times) = A(\times) = [1]$



$B(\times) = [1, 0]$

$B(\times) = [0, 1]$

Formula Validation

Validation (Option #1)

- There are some types of cancer with many associated topics (clearly visualized in the heatmap)
- Select one of these cancers
- Pull out patients with that cancer who have high membership to only one of these topics
- See if patients from topic 1 tend to live longer than patients from topic 2, list potential gene sets that can explain the difference
- Check to see if enriched gene set + cancer have been reported previously in the literature

Validation (Option #2)

- Some TCGA patients have chemotherapy information associated with their record
- We have chemotherapy resistance gene sets
- If we can show, en masse, that patients who subscribe highly to a drug resistance topic and were treated with that drug have poorer prognosis than those who were treated with drugs they don’t have a resistance to, that speaks well for our formula

Validation

1) Have all the formula done with the distance/probability scores (or whatever we call it) for each gene set, given a topic

2) Go back and identify cancers that have more than one strong topic membership (ie. there are 3 topics corresponding to breast cancer, no?). We select a cancer to use for validation

3) Pull out the samples (patients) that correspond to different topics within that cancer and therefore are associated with different gene sets

4) See if the different gene sets more strongly associated with the different topics correlate with prognosis (ie. do patients belonging to topic 1 have better prognosis than topic 2, and if so, what are the potential gene sets that explain this difference?)

5) Then what we can do is see if the enrichment of that gene set in that particular cancer has been reported in the literature previously