

# DiSCoVER: Evaluating An Algorithm for Cancer-Related Medicine Recommendations

Gabriel Diaz

UCSD STARS

Mesirov Lab

Summer Research Conference 2019

UC San Diego

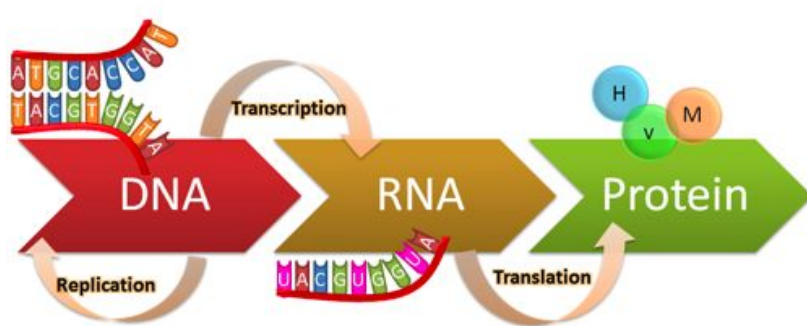
# Overview



- Background Information
- Proposed Solution
- Methods
- Results
- Future Steps

# Background Info

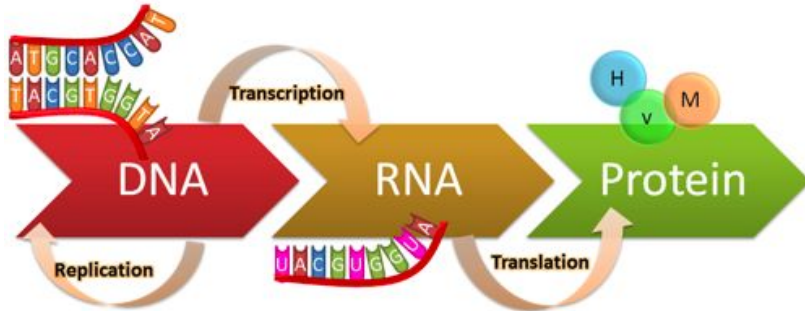
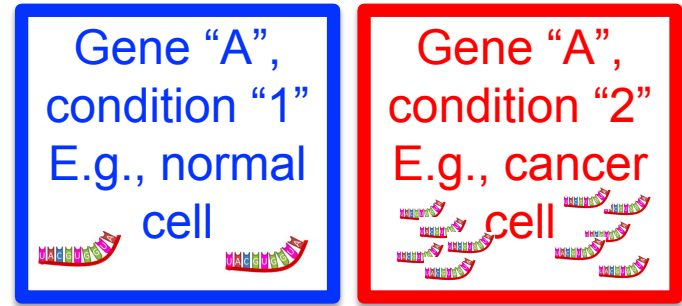
- Precision Medicine
- Protein Synthesis
- Gene Expression and Cancer



<https://genius.com/Biology-genius-the-central-dogma-annotated>

# Background Info

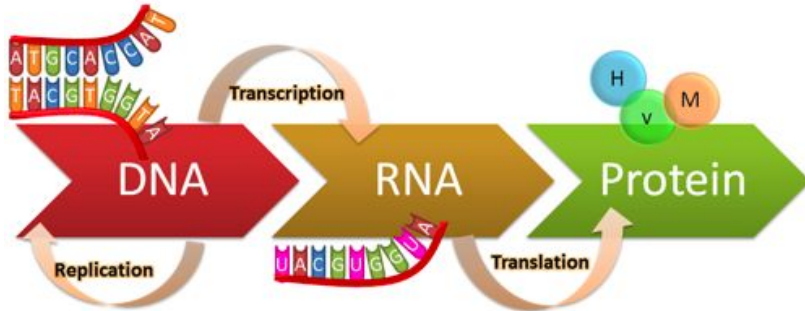
- Precision Medicine
- Protein Synthesis
- Gene Expression and Cancer
- GE Lists/Datasets



<https://genius.com/Biology-genius-the-central-dogma-annotated>

# Background Info

- Precision Medicine
- Protein Synthesis
- Gene Expression and Cancer  
GE lists/datasets



Gene "A",  
condition "1"  
E.g., normal  
cell



Gene "A",  
condition "2"  
E.g., cancer  
cell



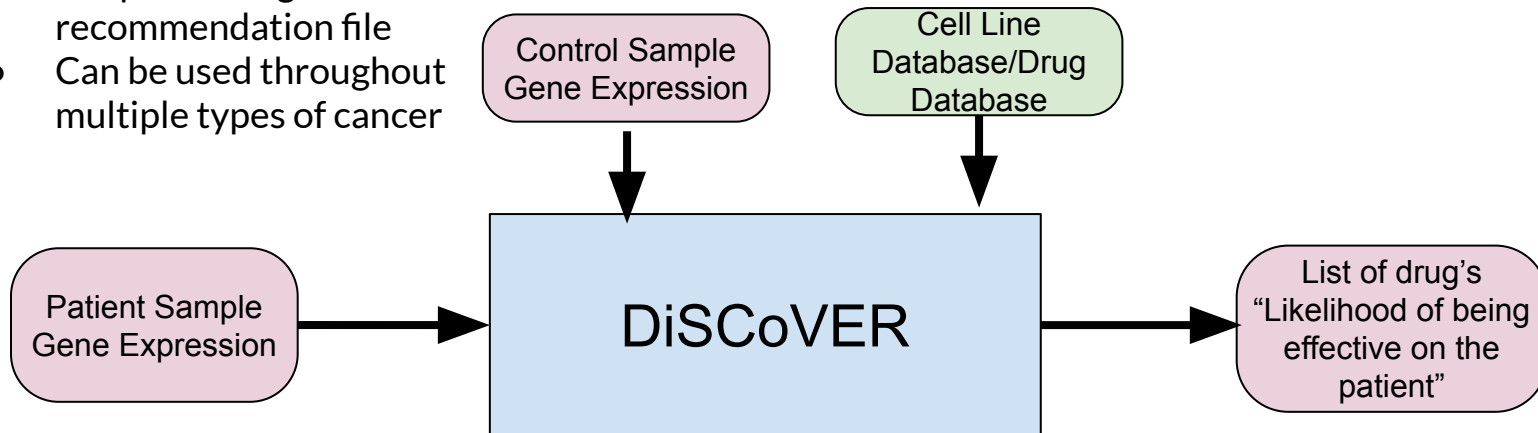
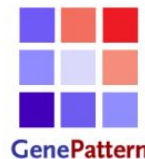
Rows:  
Genes

Name	TCGA-A7-A0CE-01
TSPAN6	7.023572
DPM1	5.210623
SCYL3	4.545135
C1orf112	4.982679
FGR	2.144122
CFH	4.67671
FUCA2	7.331436
GCLC	5.163128
NFYA	5.415756
STPG1	4.159719
NIPAL3	6.384105
LAS1L	6.187223
ENPP4	3.863163
SEMA3F	5.744268
ANKIB1	6.885191
KRIT1	5.720907
RAD52	3.838566
BAD	3.730756
LAP3	6.832838
CD99	6.452564
HS3ST1	0.959698
MAD1L1	5.675192
LASP1	7.668393
CNV11	4.388715

Columns:  
Cell Sample  
From Patients

# Background Info: DiSCoVER

- “Disease-model Signature vs. Compound-Variety Enriched Response”
- Input: 2 gene expression files
- Output: 1 drug recommendation file
- Can be used throughout multiple types of cancer





# Issue At Hand

Scientists sometimes need to preprocess data before running a computer analysis.  
For DiSCoVER, this may result in inaccurate recommendations.

## Solution

Find relationships between the outputs of DiSCoVER, where one output comes from a raw dataset and the other comes from a preprocessed dataset. If both outputs are not closely related, the accuracy of DiSCoVER's recommendations can be questionable.

# Methods



Collect Open-Source Data  
From TCGA

Parse/Upload each  
Dataset to DiSCoVER

Rank/Compare  
Scores



# Methods

Collect Open-Source Data  
From TCGA

Parse/Upload each  
Dataset to DiSCoVER

Rank/Compare  
Scores

## Cancer Patients

Rows:  
Genes

	3	Name	Descriptic	TCGA-BH	TCGA-BH	TCGA-AC	TCGA-AC	TCGA-A7	TCGA-BH	TCGA-AC	TCGA-BH	TCGA-A7	TCGA-BH
4	TSPAN6	tetraspan	5.451506	6.672107	5.138829	3.082438	4.622452	5.399731	6.140938	3.983999	6.270467	5.883384	5.684891
5	DPH1	dolichyl-p	5.020696	5.020114	4.818484	5.780905	3.739142	4.743943	5.051848	4.526725	4.841457	4.790968	4.90422
6	SCYL3	SCYL like	4.355102	5.994743	4.399643	0.686101	6.577639	5.209043	3.695598	4.508038	4.926447	4.064534	4.77751
7	C1orf112	chromoso	4.35263	5.035165	3.09634	3.71777	4.285981	2.942706	1.875086	3.775341	2.210549	2.555197	3.315104
8	FGR	FGR proto	3.821978	2.454533	3.6258	1.72305	2.820446	2.699476	3.552944	4.004718	3.649853	5.206798	1.616612
9	CFH	complem	5.384736	5.473612	7.703547	5.644724	5.449173	6.943087	5.983759	5.688422	7.026805	7.717385	4.893814
10	FUCA2	fucosidasi	6.35433	5.309	5.787524	6.251058	5.085452	5.86883	5.543541	6.272818	5.903556	6.591877	5.79131
11	GLIC	glutamate	5.241141	4.494142	5.489878	5.019147	5.187508	5.536651	4.707675	4.973192	5.445259	6.107687	5.400271
12	NPYA	nuclear tr	5.911804	5.888342	5.15795	5.820049	6.03714	5.313134	4.58315	5.050929	5.03863	4.838376	4.22944
13	STPG1	sperm tail	3.335837	2.703195	2.926843	3.644922	4.285981	4.04197	3.718697	2.805567	3.395456	3.292163	3.10023
14	NIPAL3	NIPA like	5.082405	5.572081	5.571339	5.157626	7.121923	5.016579	5.684153	4.445131	5.779612	4.868737	5.797464
15	LAS1L	LAS1 like	5.727618	5.714696	5.359993	6.043019	4.589202	5.401952	5.596084	5.813025	5.645125	5.161409	5.867734
16	ENPP4	ectonucle	3.627162	7.47121	4.247486	4.524369	5.047636	4.963414	3.415976	5.460455	4.557558	5.383731	3.630466
17	SEMA3F	semaphor	6.400524	7.03908	6.592476	6.478607	5.913614	5.69193	6.128414	6.628243	5.428525	5.395457	6.93588
18	ANKB1	ankyrin re	5.902118	6.470039	6.511698	6.607785	6.073741	5.830891	4.620423	6.030689	5.701398	6.247803	5.824012
19	KRT11	KRT11, ani	4.855094	5.743971	5.246199	4.940129	5.200989	4.559864	4.278999	4.174845	5.11362	5.30065	4.77751
20	RAD52	RAD52 ho	1.246809	3.53694	3.123497	1.925077	4.707096	2.008339	2.690739	1.761781	2.730834	2.724126	3.067224
21	BAD	BCL2 asso	5.157952	3.968607	4.199374	4.63335	3.806676	5.031963	6.335296	5.322885	4.54328	4.741949	5.389202
22	LAP3	leucine ar	7.591645	6.106904	6.904103	4.466514	6.172213	6.393231	6.517296	7.580614	6.82839	7.210952	6.253975
23	CD99	CD99 mol	8.495822	6.786305	7.758889	7.938474	6.502367	8.794252	7.929157	7.543141	7.624574	9.082229	6.74329
24	H53ST1	heparan s	2.461576	1.959496	2.384288	0.913526	2.690151	2.061779	3.520443	3.290198	2.489062	2.512762	1.300998
25	MAD1L1	MAD1 mit	5.095767	4.401178	3.622631	3.467526	3.233614	4.780834	5.290231	4.698384	4.226276	4.516602	5.000468

Raw Dataset

	3	Name	Descriptic	TCGA-BH	TCGA-BH	TCGA-AC	TCGA-AC	TCGA-A7	TCGA-BH	TCGA-AC	TCGA-BH	TCGA-A7	TCGA-BH
4	TSPAN6	tetraspan	5.451506	6.672107	5.138829	3.082438	4.622452	5.399731	6.140938	3.983999	6.270467	5.883384	5.684891
5	DPH1	dolichyl-p	5.020696	5.020114	4.818484	5.780905	3.739142	4.743943	5.051848	4.526725	4.841457	4.790968	4.90422
6	SCYL3	SCYL like	4.355102	5.994743	4.399643	0.686101	6.577639	5.209043	3.695598	4.508038	4.926447	4.064534	4.77751
7	C1orf112	chromoso	4.35263	5.035165	3.09634	3.71777	4.285981	2.942706	1.875086	3.775341	2.210549	2.555197	3.315104
8	FGR	FGR proto	3.821978	2.454533	3.6258	1.72305	2.820446	2.699476	3.552944	4.004718	3.649853	5.206798	1.616612
9	CFH	complem	5.384736	5.473612	7.703547	5.644724	5.449173	6.943087	5.983759	5.688422	7.026805	7.717385	4.893814
10	FUCA2	fucosidasi	6.35433	5.309	5.787524	6.251058	5.085452	5.86883	5.543541	6.272818	5.903556	6.591877	5.79131
11	GLIC	glutamate	5.241141	4.494142	5.489878	5.019147	5.187508	5.536651	4.707675	4.973192	5.445259	6.107687	5.400271
12	NPYA	nuclear tr	5.911804	5.888342	5.15795	5.820049	6.03714	5.313134	4.58315	5.050929	5.03863	4.838376	4.22944
13	STPG1	sperm tail	3.335837	2.703195	2.926843	3.644922	4.285981	4.04197	3.718697	2.805567	3.395456	3.292163	3.10023
14	NIPAL3	NIPA like	5.082405	5.572081	5.571339	5.157626	7.121923	5.016579	5.684153	4.445131	5.779612	4.868737	5.797464
15	LAS1L	LAS1 like	5.727618	5.714696	5.359993	6.043019	4.589202	5.401952	5.596084	5.813025	5.645125	5.161409	5.867734
16	ENPP4	ectonucle	3.627162	7.47121	4.247486	4.524369	5.047636	4.963414	3.415976	5.460455	4.557558	5.383731	3.630466
17	SEMA3F	semaphor	6.400524	7.03908	6.592476	6.478607	5.913614	5.69193	6.128414	6.628243	5.428525	5.395457	6.93588
18	ANKB1	ankyrin re	5.902118	6.470039	6.511698	6.607785	6.073741	5.830891	4.620423	6.030689	5.701398	6.247803	5.824012
19	KRT11	KRT11, ani	4.855094	5.743971	5.246199	4.940129	5.200989	4.559864	4.278999	4.174845	5.11362	5.30065	4.77751
20	RAD52	RAD52 ho	1.246809	3.53694	3.123497	1.925077	4.707096	2.008339	2.690739	1.761781	2.730834	2.724126	3.067224
21	BAD	BCL2 asso	5.157952	3.968607	4.199374	4.63335	3.806676	5.031963	6.335296	5.322885	4.54328	4.741949	5.389202
22	LAP3	leucine ar	7.591645	6.106904	6.904103	4.466514	6.172213	6.393231	6.517296	7.580614	6.82839	7.210952	6.253975
23	CD99	CD99 mol	8.495822	6.786305	7.758889	7.938474	6.502367	8.794252	7.929157	7.543141	7.624574	9.082229	6.74329
24	H53ST1	heparan s	2.461576	1.959496	2.384288	0.913526	2.690151	2.061779	3.520443	3.290198	2.489062	2.512762	1.300998
25	MAD1L1	MAD1 mit	5.095767	4.401178	3.622631	3.467526	3.233614	4.780834	5.290231	4.698384	4.226276	4.516602	5.000468

Preprocessed Dataset

- We are using RNA-Seq samples from 20 breast cancer (BRCA) patients.
- Rows represent genes, Columns represent cancerous or normal cells from each patient
- 755 columns by 40 rows

# Methods

Collect Open-Source Data  
From TCGA

Parse/Upload each  
Dataset to DiSCoVER

Rank/Compare  
Scores

TCGA Gene Expression Dataset (1 of 2)

3	Name	DescrptTCGA-BH-TCGA-BH-TCGA-AC-TCGA-A7-TCGA-BH-TCGA-AC-TCGA-BH-TCGA-A7-TCGA-BH-
4	TSPAN6	tetraspanin 5.451506 6.672107 5.138829 3.082438 4.622452 5.399731 6.140938 3.983969 6.270467 5.883384 5.684891
5	DPW1	dolichyl-p 5.020696 5.002114 4.813494 5.789095 3.739742 4.743943 5.051848 4.526725 4.841457 4.790968 4.90422
6	SCN3	SCN3 like 4.355102 5.94741 4.199443 6.080101 6.577639 5.209441 3.605596 4.580838 4.592447 4.064334 4.77751
7	C1orf112	chromoso 4.33263 5.055165 3.09634 3.71777 4.265981 2.942706 1.875086 3.773341 2.210549 2.555197 3.315104
8	FGR	FGR proto 3.821978 2.454531 3.6258 1.72305 2.820446 2.699476 3.552944 4.004718 3.649853 5.206798 1.616612
9	CFH	complexi 5.384736 5.473612 7.703547 5.644724 5.449173 6.943087 5.983729 5.688422 7.020805 7.717385 4.893814
10	FUSC2	fucosidase 6.33413 5.309 5.78724 6.251058 5.033452 5.86831 5.943464 6.272218 5.903516 6.91877 5.79131
11	GLUC	glutamate 5.241141 5.494142 5.469878 5.019147 5.187508 5.526651 4.707875 4.973932 3.445259 6.107687 5.400271
12	NFYA	nuclear tr 5.911804 5.888342 5.55795 5.820049 6.010714 5.331434 4.58315 5.050929 5.05863 4.838376 6.228944
13	STPGL1	stpm tall 3.333837 2.781129 2.508463 3.444022 4.265961 4.04157 3.718077 2.805347 3.395456 3.292163 3.10023
14	NIPAL3	NIPA like 5.082405 5.572083 5.757139 5.157626 7.121923 5.026579 5.684153 4.445131 5.779612 4.868737 5.797464
15	LAS1L	LAS1 like 5.727618 5.714696 5.310993 6.043019 4.589202 5.401592 5.956084 5.813025 5.645125 5.516149 5.867734
16	ENP94	ectonucle 5.627162 7.47121 4.247466 4.524369 5.047636 4.963414 3.413976 5.460455 4.355758 5.381731 6.830466
17	SCNA3F	semaphor 6.400524 7.01906 6.503476 6.478607 5.913614 5.49193 6.128414 6.632643 5.420525 5.395457 6.91588
18	ANKK1	ankyrin re 5.902118 6.470039 6.511698 6.607785 6.073741 5.830891 4.620423 6.030689 5.701398 6.247803 5.824012
19	KRT11	KRT11, ank 4.853094 5.743971 5.246139 4.948129 5.200969 4.559884 4.276999 4.174845 5.11382 5.30005 4.77751
20	RAOS2	RAOS2 tro 1.246809 3.33404 3.121487 3.105077 4.707096 2.008339 2.807731 1.761781 2.758454 2.724126 3.067214
21	BAD	BCL2 asso 5.157952 3.968607 4.199374 4.63335 3.866676 5.031963 6.355296 5.322885 4.54328 4.741949 5.389202
22	LAP3	leucine ar 7.591645 6.109904 6.904103 6.466514 6.17223 6.393231 6.517296 7.580614 6.82839 7.210952 6.253975
23	CD99	CD99 mol 4.495822 6.786205 7.708989 7.898426 6.500367 6.794252 7.921537 7.543141 7.634674 9.082226 6.74219
24	HIS3T1	hepatan s 2.465176 1.959496 2.384288 0.911536 2.696151 2.961779 3.520449 3.265198 2.489052 2.512762 1.350998
25	MAD1L1	MAD1 mit 5.095767 4.401178 3.622631 3.467526 3.233614 4.780834 5.290211 4.698384 4.226276 4.516662 5.000468

40 Columns

Control Sample  
Gene Expression



Patient Sample  
Gene Expression



DiSCoVER

# Collect Open-Source Data From TCGA

## Parse/Upload each Dataset to DiSCoVER

## Rank/Compare Scores

[illegible]

Cancerous  
Tissue  
Sample  
("01")

Name	TCGA-A7-A0E-01
TSPAN6	7.023572
DPM1	5.210623
SCYL3	4.545135
C1orf112	4.982679
FRG	2.144122
CFH	4.67671
FUCA2	7.331436
GLCL	5.163128
NFYA	5.415756
STPG1	4.159719
NIPAL3	6.384105
LAS1L	6.187223
ENPP4	3.863163
SEMA3F	5.744268
ANKIB1	6.885191
KRIT1	5.720907
RAD52	3.838566
BAD	3.730756
LAP3	6.832388
CD99	6.452564
H53ST1	0.959698
MDAD11	5.675192
USP1	7.668393
ENX1	4.388375

Normal/Control  
Tissue Sample  
("11")

Name	TCGA-A7-A0CE-11
TSPAN6	6.183838
DPM1	5.055648
SCYL3	4.318682
C1orf122	2.071605
FGF3	3.547935
CFH	6.275871
FUCA2	5.886361
GLYC4	5.222425
NPY2	4.792672
STPG1	3.507554
NIPAL3	6.067216
LAS1L	5.622791
ENPP4	3.974335
SEMA3F	6.10737
ANKIB1	5.419838
KRIT1	5.074776
RAD52	2.947542
BAD	5.120678
LAP3	6.931336
CD99	7.408174
H53SF1	2.224762
MD12L1	6.682871
LASP1	7.994374
SNX11	4.707592

# 40 Columns

X 20




Parse/Upload each Dataset to DiSCoVER

# Rank/Compare Scores

---

Name	TCGA-A7-ADCE-01	Name	TCGA-A7-ADCE-11
TP53AN6	7.023572	TP53AN6	6.183838
DPM1	5.210623	DPM1	5.055648
CCYL1	4.545335	CCYL1	4.318682
CTF12	4.480879	CTF12	2.071000
CFH	2.144122	FCV3	3.547935
CFH	4.67071	CFH	6.275871
FUC2A	7.331436	FUC2A	5.886361
GOLC	5.163128	GOLC	5.22425
NPYFA	5.431795	NPYFA	4.796272
STG1	4.159719	STG1	3.507554
NP3A3	6.384105	NP3A3	6.027216
LAS1L	6.187223	LAS1L	5.867491
RPNH1	3.261361	RPNH1	3.794333
SEMA3F	7.442628	SEMA3F	6.10737
ANKB1	6.885191	ANKB1	5.419838
KRT11	5.720907	KRT11	5.20076
ADRS2	3.838566	ADRS2	2.947542
BAO	7.79794	BAO	5.042786
LAP3	6.832838	LAP3	6.931336
CD9	6.452564	CD9	7.408124
H3S1T1	5.976192	H3S1T1	2.224762
NADL1	6.755192	NADL1	4.682872
NPYFA	7.666393	NPYFA	7.594374
SNK11	4.386745	SNK11	4.707592



```

graph LR
    A[DiSCoVER] --> B[DiSCoVER]
  
```

Name	TCGA-A7-A0CE-01
TSNP6	7.023572
DPK1	5.210623
SCV1	5.454135
GLI2F12	4.996127
FCR	2.144122
GR	4.67671
FCU2	7.331436
GLC4	5.149128
GLC5	4.137564
STPG1	4.591719
NIPAL3	1.334705
LAS1	6.872233
ENP1A	3.186123
SEMA5F	7.042458
ANKB1	6.085191
KRT1	5.720907
RAD52	3.835566
BAD	3.797056
CDS9	4.652264
H031	3.959698
LAS1L	5.675192
MADP1	7.668393
TSNP6	7.023572
DPK1	5.210623
SCV1	5.454135
GLI2F12	4.996127
FCR	2.144122
GR	4.67671
FCU2	7.331436
GLC4	5.149128
GLC5	4.137564
STPG1	4.591719
NIPAL3	1.334705
LAS1	6.872233
ENP1A	3.186123
SEMA5F	7.042458
ANKB1	6.085191
KRT1	5.720907
RAD52	3.835566
BAD	3.797056
CDS9	4.652264
H031	3.959698
LAS1L	5.675192
MADP1	7.668393

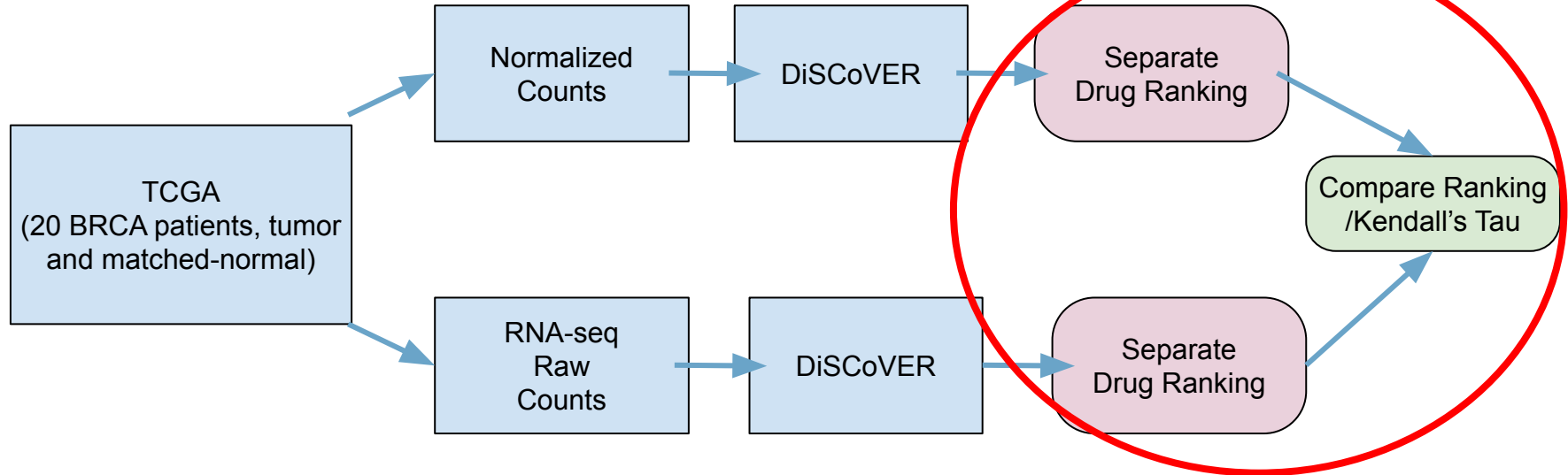
X 20

# Methods

Collect Open-Source Data  
From TCGA

Parse/Upload each  
Dataset to DiSCoVER

Rank/Compare  
Scores



# Methods

Collect Open-Source Data  
From TCGA

Parse/Upload each  
Dataset to DiSCoVER

Rank/Compare  
Scores

## Example Dataset

drug	TCGA-BH- A0B3-01_DiSCoVER_result	TCGA-BH- A0BC-01_DiSCoVER_result	TCGA-A7- A0CH-01_DiSCoVER_result	TCGA-BH- A0B7-01_DiSCoVER_result	TCGA-A7- A13G-01_DiSCoVER_result
gdsc_Motesanib	0.597	-0.596	-0.611	0.602	-0.686
gdsc_PD173074	0.590	-0.528	-0.548	-0.528	-0.591
gdsc_GW441756	0.573	-0.568	0.576	0.515	0.653
gdsc_Cetuximab	0.570	0.539	0.606	0.551	0.706
gdsc Afatinib	0.554	0.577	0.584	0.617	0.767

# Methods

Collect Open-Source Data  
From TCGA

Parse/Upload each  
Dataset to DiSCoVER

Rank/Compare  
Scores

## Example Dataset

5]:

	drug	TCGA-BH- A0B3-01_DiSCoVER_result	TCGA-BH- A0BC-01_DiSCoVER_result	TCGA-A7- A0CH-01_DiSCoVER_result	TCGA-BH- A0B7-01_DiSCoVER_result	TCGA-A7- A13G-01_DiSCoVER_result
0	ccle_17-AAG	515.0	186.0	186.0	350.0	183.0
1	ccle_AEW541	251.0	390.0	322.0	373.0	269.0
2	ccle_AZD0530	378.0	174.0	158.0	223.0	123.0
3	ccle_AZD6244	570.0	305.0	297.0	601.0	369.0
4	ccle_Erlotinib	565.0	92.0	64.0	150.0	42.0
5	ccle_Irinotecan	242.0	642.0	722.0	293.0	712.0

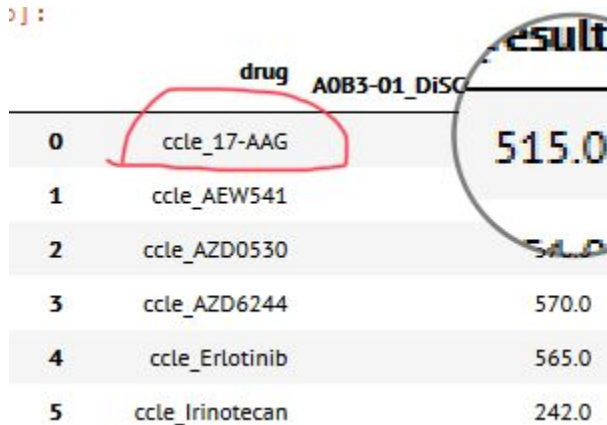
# Methods

Collect Open-Source Data  
From TCGA

Parse/Upload each  
Dataset to DiSCoVER

Rank/Compare  
Scores

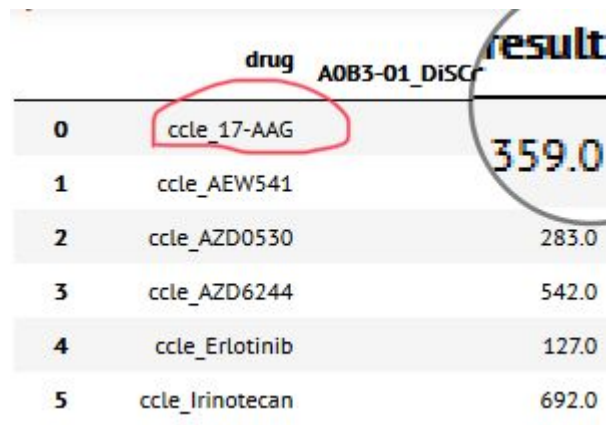
## Ranking of Raw Output



Ranking of Raw Output

	drug	A0B3-01_DiSCo	result
0	ccle_17-AAG		515.0
1	ccle_AEW541		
2	ccle_AZD0530		515.0
3	ccle_AZD6244		570.0
4	ccle_Erlotinib		565.0
5	ccle_Irinotecan		242.0

## Ranking of Preprocessed Output



Ranking of Preprocessed Output

	drug	A0B3-01_DiSCo	result
0	ccle_17-AAG		359.0
1	ccle_AEW541		
2	ccle_AZD0530		283.0
3	ccle_AZD6244		542.0
4	ccle_Erlotinib		127.0
5	ccle_Irinotecan		692.0



# Methods

Collect Open-Source Data  
From TCGA

Parse/Upload each  
Dataset to DiSCoVER

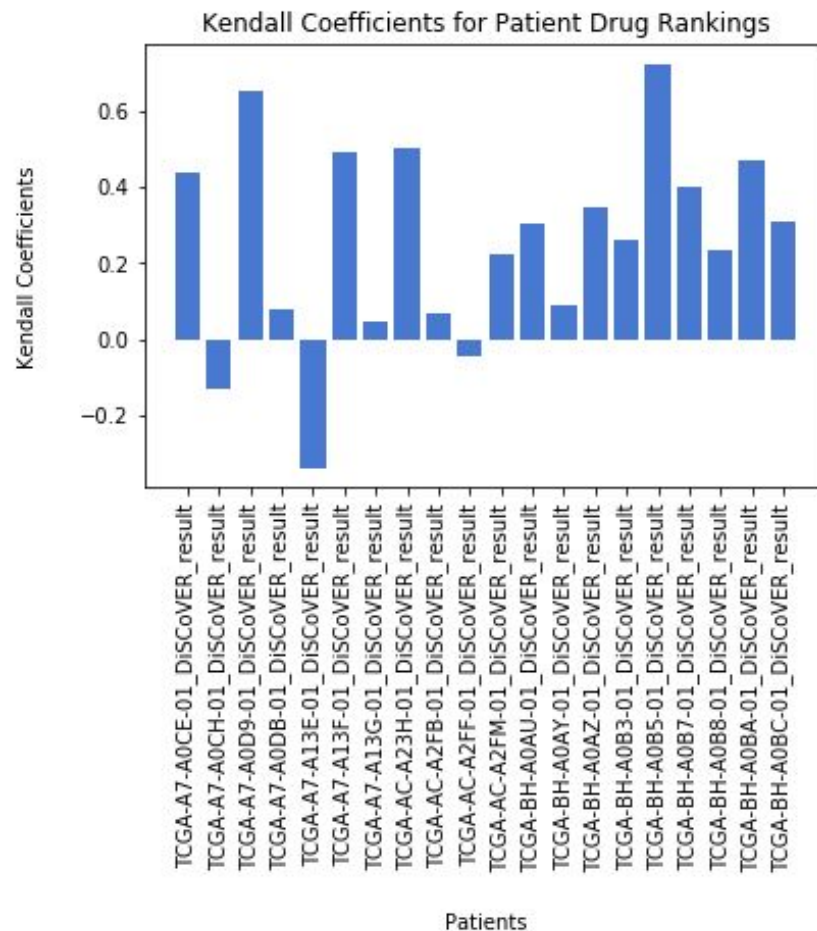
Rank/Compare  
Scores

## Kendall's Tau

- Represents degree of correlation between two columns of ranked data
- Can range from -1.0 to 1.0
- Based on concordant and discordant pairs

Drug	Patient1	Patient2	C	D
A	1	1	5	0
B	2	2	4	0
C	3	4	2	1
D	4	3	2	0
E	5	6	0	1
F	6	5	0	0
		Total:	13	2
		KT	0.733	

# Preliminary Results:





## Possible Areas of Further Research

- Analyze possible data preprocessing methods
- Determine potential sources of “noisy data” for recommendation scores



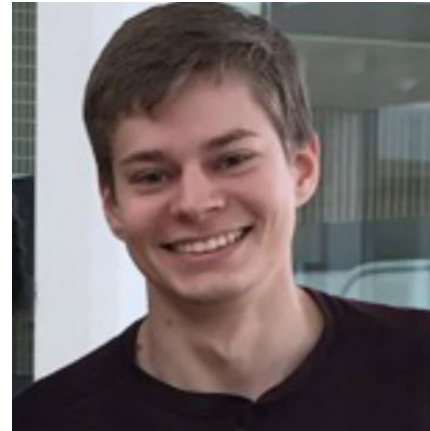
## Acknowledgments - Mesirov Lab



Dr. Jill Mesirov  
PI/Lab Director



Dr. Edwin Juarez  
Bioinformatics Programmer  
/Research Guide



Alex Wenzel  
PhD Candidate



Owen Chapman  
PhD Candidate



**Thank You!**