A Versatile Algorithm for Finding Patterns in Large Cancer Cell Line Data Sets

James Jusuf, Phillips Academy Andover

May 21, 2017

MIT PRIMES

The Broad Institute of MIT and Harvard

Introduction

- A quest to understand the cancer genome
  - Discover pathways that can be targeted in cancer treatment
  - Predict valuable information about cancer patients based on known genetic indicators
- An explosion in the amount of available data
  - Human Genome Project (1990-2003)
  - Databases: COSMIC (2004), TCGA (2005), CCLE (2012)

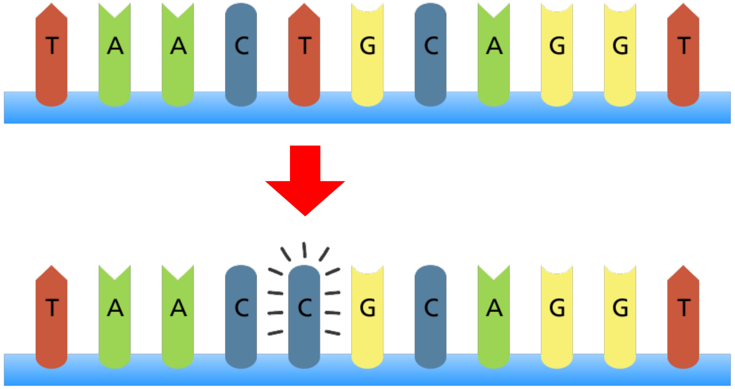
The more data the better: larger sample sizes allow us to detect patterns in with more reliability

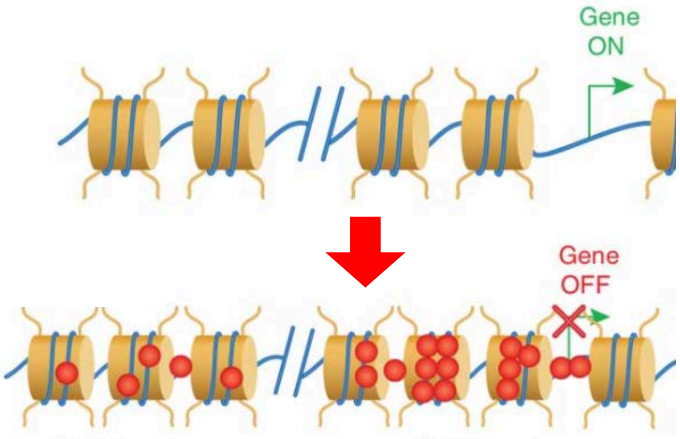
Introduction

- Our study focuses on two widely studied phenomena in cancer genomics/epigenomics:

Fig. 1A: Mutations: variations in DNA sequence

Fig. 1B: Methylation: amount and location of methyl ($–CH_3$) groups attached to DNA, which regulate gene expression





Goals

Find genes and cancer types in which a specific mutation affects a cell's methylation profile

Create an algorithm to quantify the correlation between methylation and the mutation of a given gene in a given cancer type

In the future:

- Apply the algorithm to other variables, e.g. exon splicing and mutations
- Further investigate any potentially significant patterns we discover in the laboratory

What is unsupervised clustering?

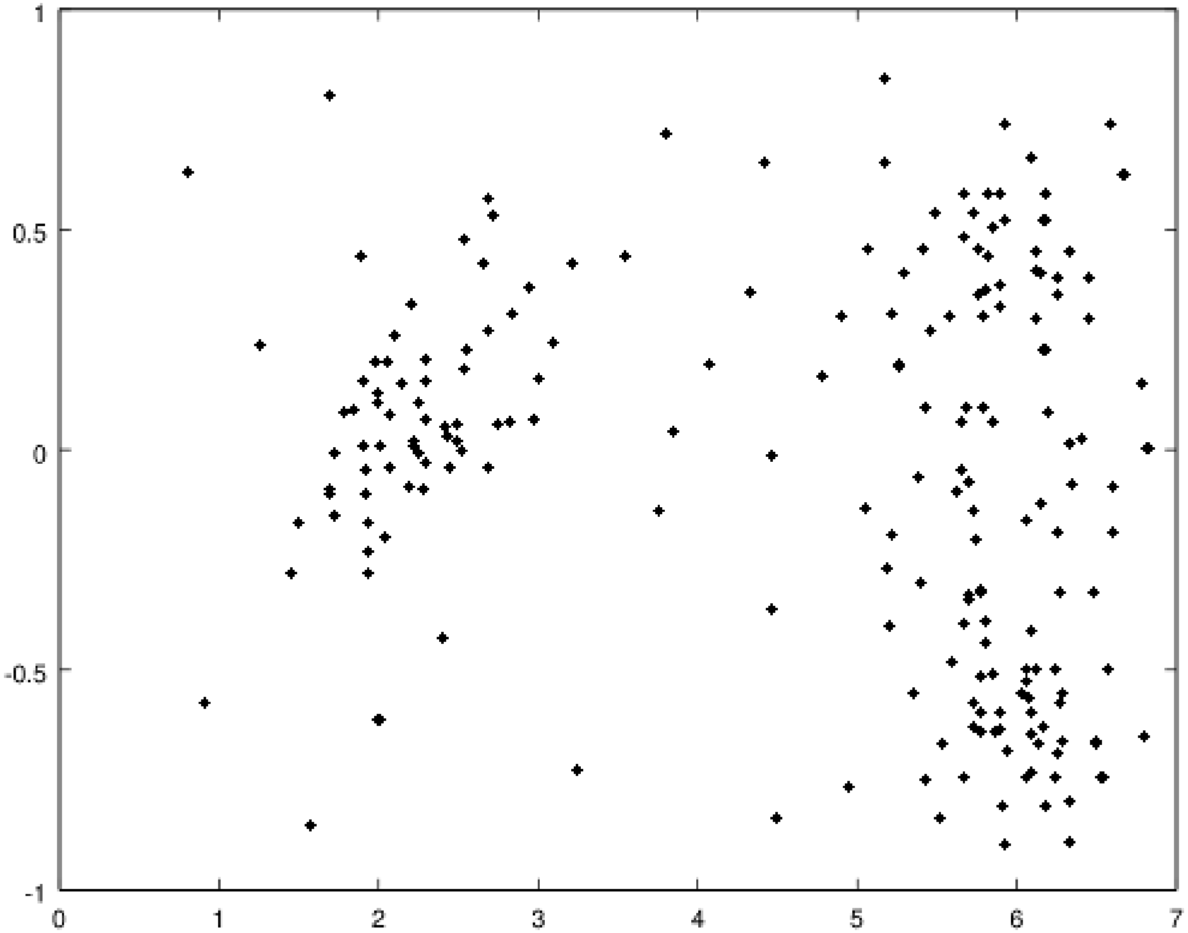


Fig. 2A: Some arbitrary data

What is unsupervised clustering?

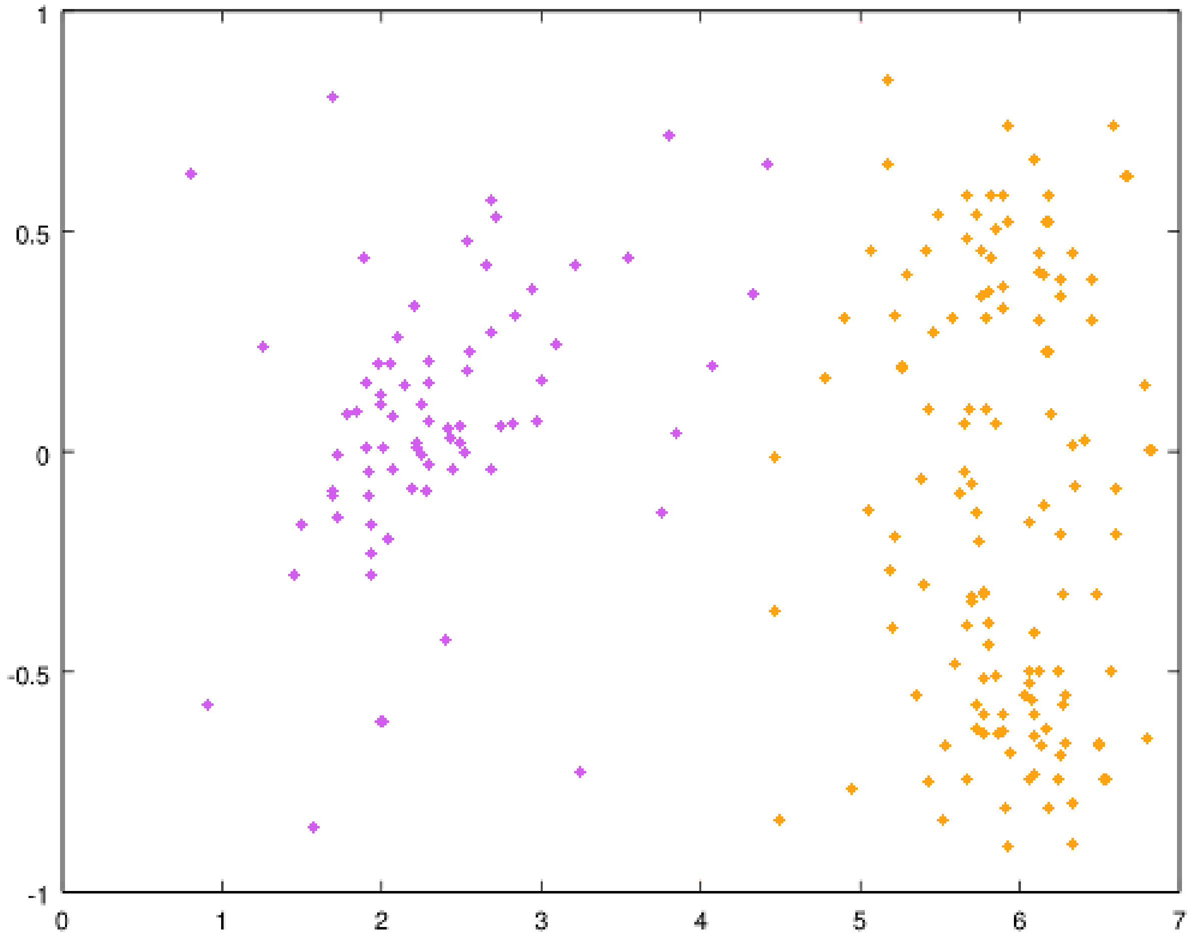


Fig. 2B: Data partitioned into 2 clusters

What is unsupervised clustering?

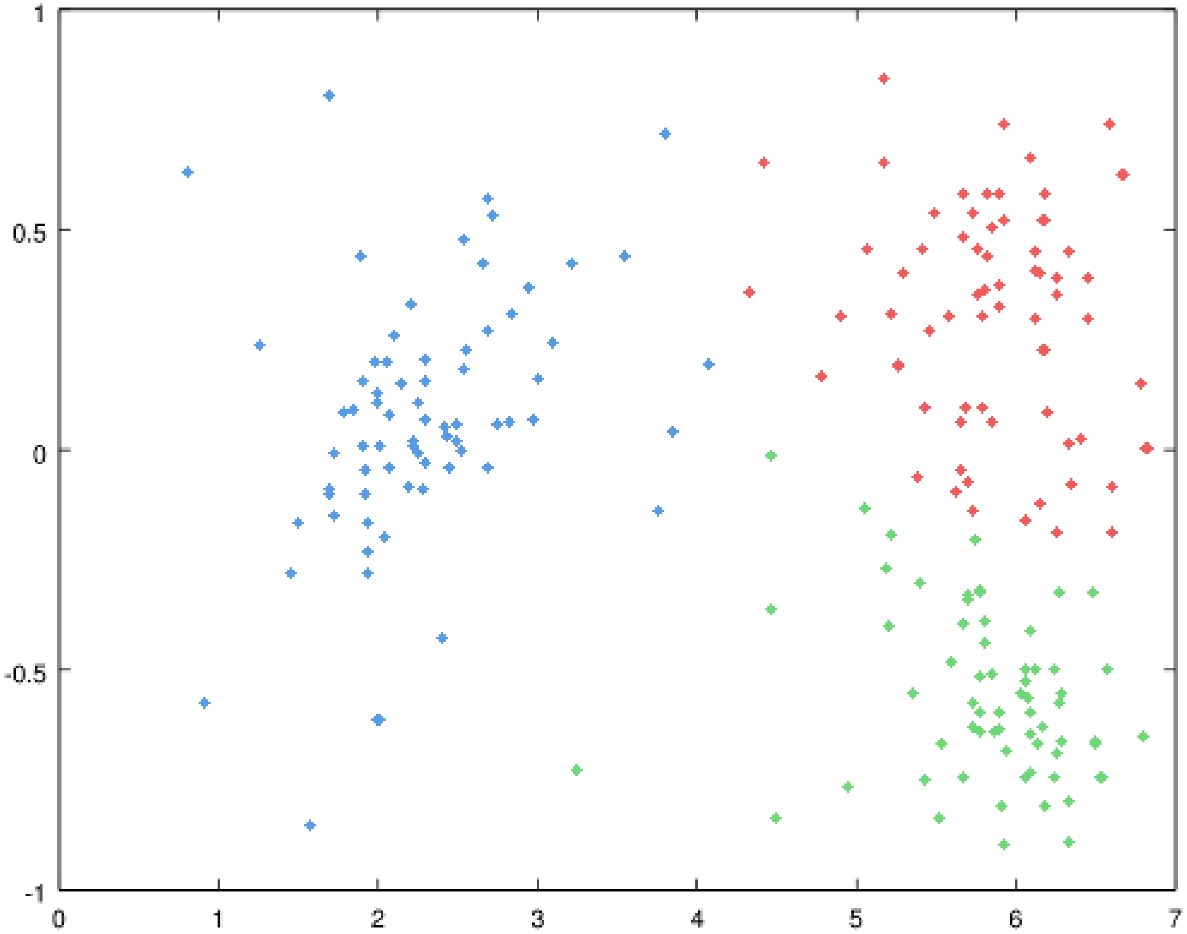


Fig. 2C: Data partitioned into 3 clusters

What is unsupervised clustering?

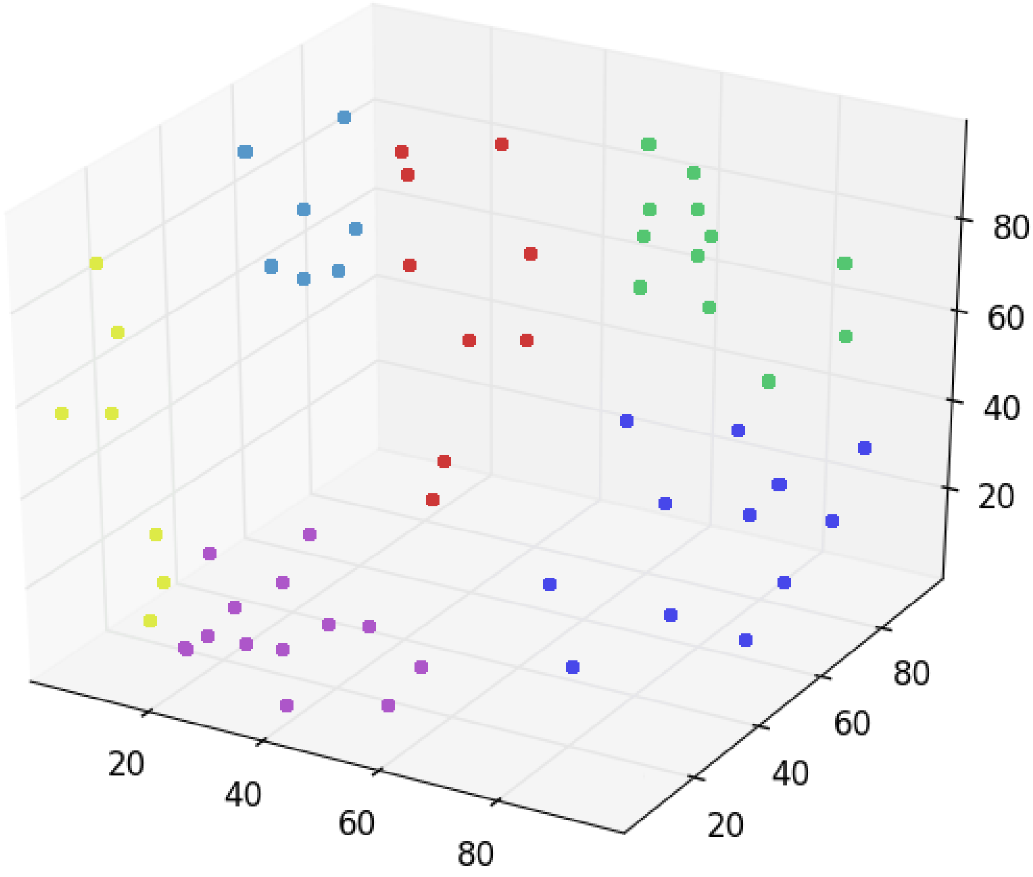
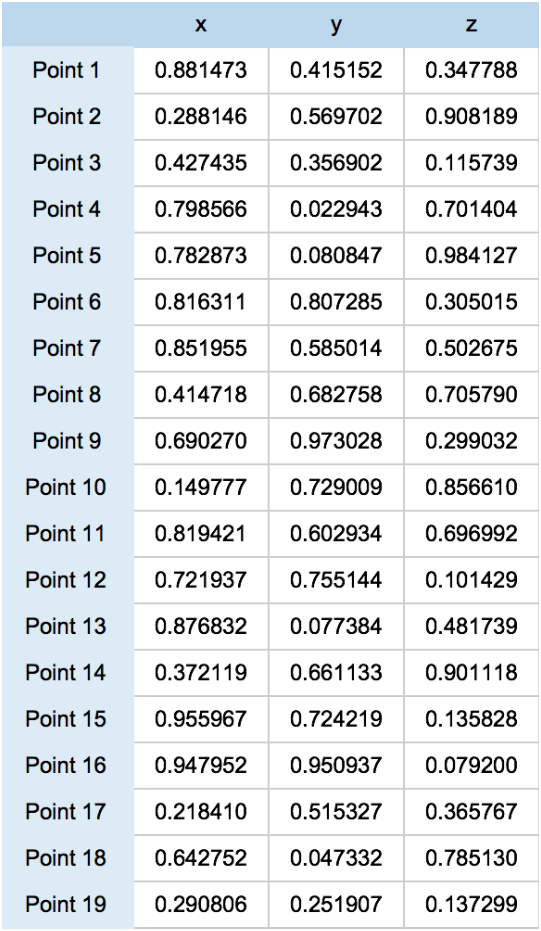


Fig. 3: Clustering can be generalized into any number of dimensions

How does clustering work mathematically?



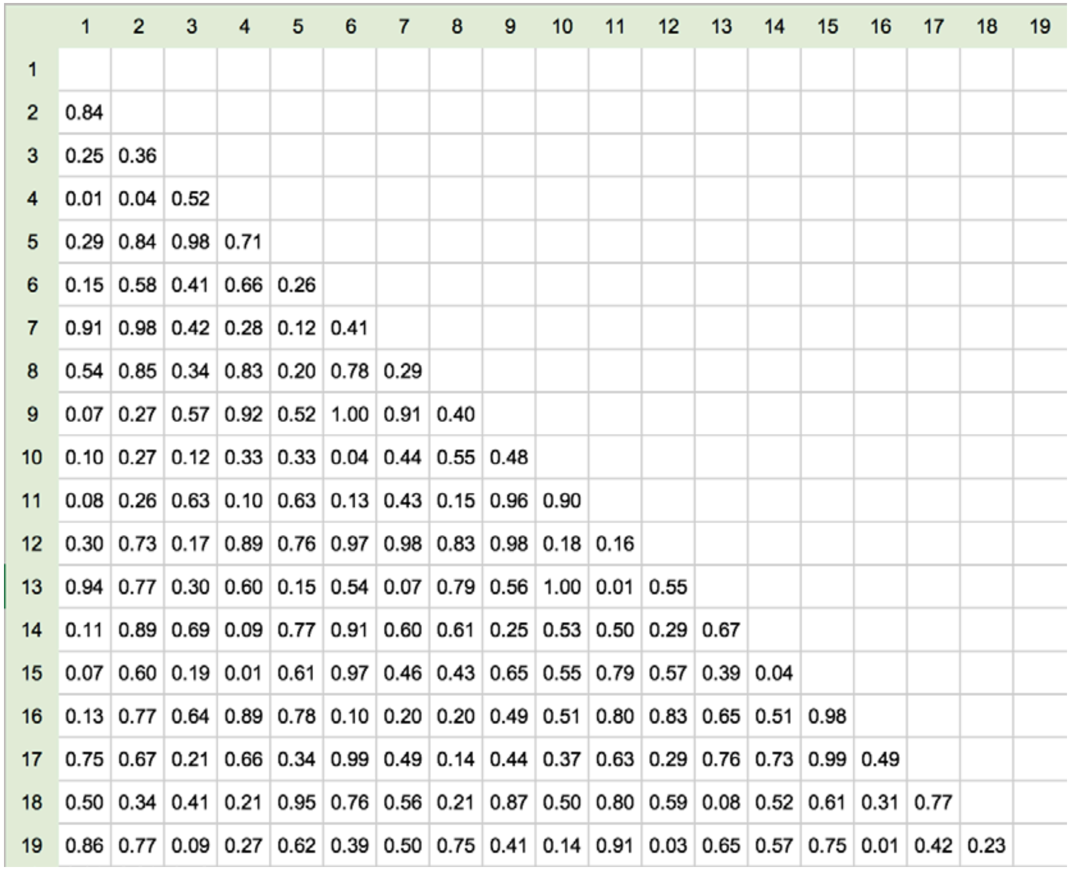
	x	y	z
Point 1	0.881473	0.415152	0.347788
Point 2	0.288146	0.569702	0.908189
Point 3	0.427435	0.356902	0.115739
Point 4	0.798566	0.022943	0.701404
Point 5	0.782873	0.080847	0.984127
Point 6	0.816311	0.807285	0.305015
Point 7	0.851955	0.585014	0.502675
Point 8	0.414718	0.682758	0.705790
Point 9	0.690270	0.973028	0.299032
Point 10	0.149777	0.729009	0.856610
Point 11	0.819421	0.602934	0.696992
Point 12	0.721937	0.755144	0.101429
Point 13	0.876832	0.077384	0.481739
Point 14	0.372119	0.661133	0.901118
Point 15	0.955967	0.724219	0.135828
Point 16	0.947952	0.950937	0.079200
Point 17	0.218410	0.515327	0.365767
Point 18	0.642752	0.047332	0.785130
Point 19	0.290806	0.251907	0.137299

Fig. 4A: Sample table of 3-dimensional data showing *x*, *y*, and *z*-coordinates of 19 points

Calculate the Euclidean distance between every pair of points:

$$d_{ij} = \sqrt{\Delta x_{ij}{}^2 + \Delta y_{ij}{}^2 + \Delta z_{ij}{}^2}$$

How does clustering work mathematically?



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1																			
2	0.84																		
3	0.25	0.36																	
4	0.01	0.04	0.52																
5	0.29	0.84	0.98	0.71															
6	0.15	0.58	0.41	0.66	0.26														
7	0.91	0.98	0.42	0.28	0.12	0.41													
8	0.54	0.85	0.34	0.83	0.20	0.78	0.29												
9	0.07	0.27	0.57	0.92	0.52	1.00	0.91	0.40											
10	0.10	0.27	0.12	0.33	0.33	0.04	0.44	0.55	0.48										
11	0.08	0.26	0.63	0.10	0.63	0.13	0.43	0.15	0.96	0.90									
12	0.30	0.73	0.17	0.89	0.76	0.97	0.98	0.83	0.98	0.18	0.16								
13	0.94	0.77	0.30	0.60	0.15	0.54	0.07	0.79	0.56	1.00	0.01	0.55							
14	0.11	0.89	0.69	0.09	0.77	0.91	0.60	0.61	0.25	0.53	0.50	0.29	0.67						
15	0.07	0.60	0.19	0.01	0.61	0.97	0.46	0.43	0.65	0.55	0.79	0.57	0.39	0.04					
16	0.13	0.77	0.64	0.89	0.78	0.10	0.20	0.20	0.49	0.51	0.80	0.83	0.65	0.51	0.98				
17	0.75	0.67	0.21	0.66	0.34	0.99	0.49	0.14	0.44	0.37	0.63	0.29	0.76	0.73	0.99	0.49			
18	0.50	0.34	0.41	0.21	0.95	0.76	0.56	0.21	0.87	0.50	0.80	0.59	0.08	0.52	0.61	0.31	0.77		
19	0.86	0.77	0.09	0.27	0.62	0.39	0.50	0.75	0.41	0.14	0.91	0.03	0.65	0.57	0.75	0.01	0.42	0.23	

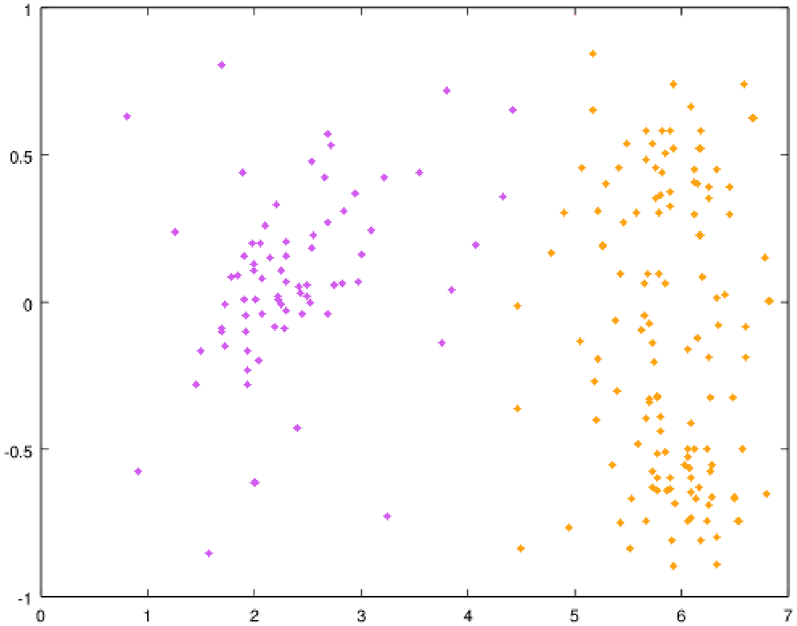
Fig. 4B: A distance matrix of the 19 data points

Back to our project

- Create an algorithm to quantify the correlation between methylation and the mutation of a given gene in a given cancer type

Methylation Cluster 1

All have mutation in Gene *X*



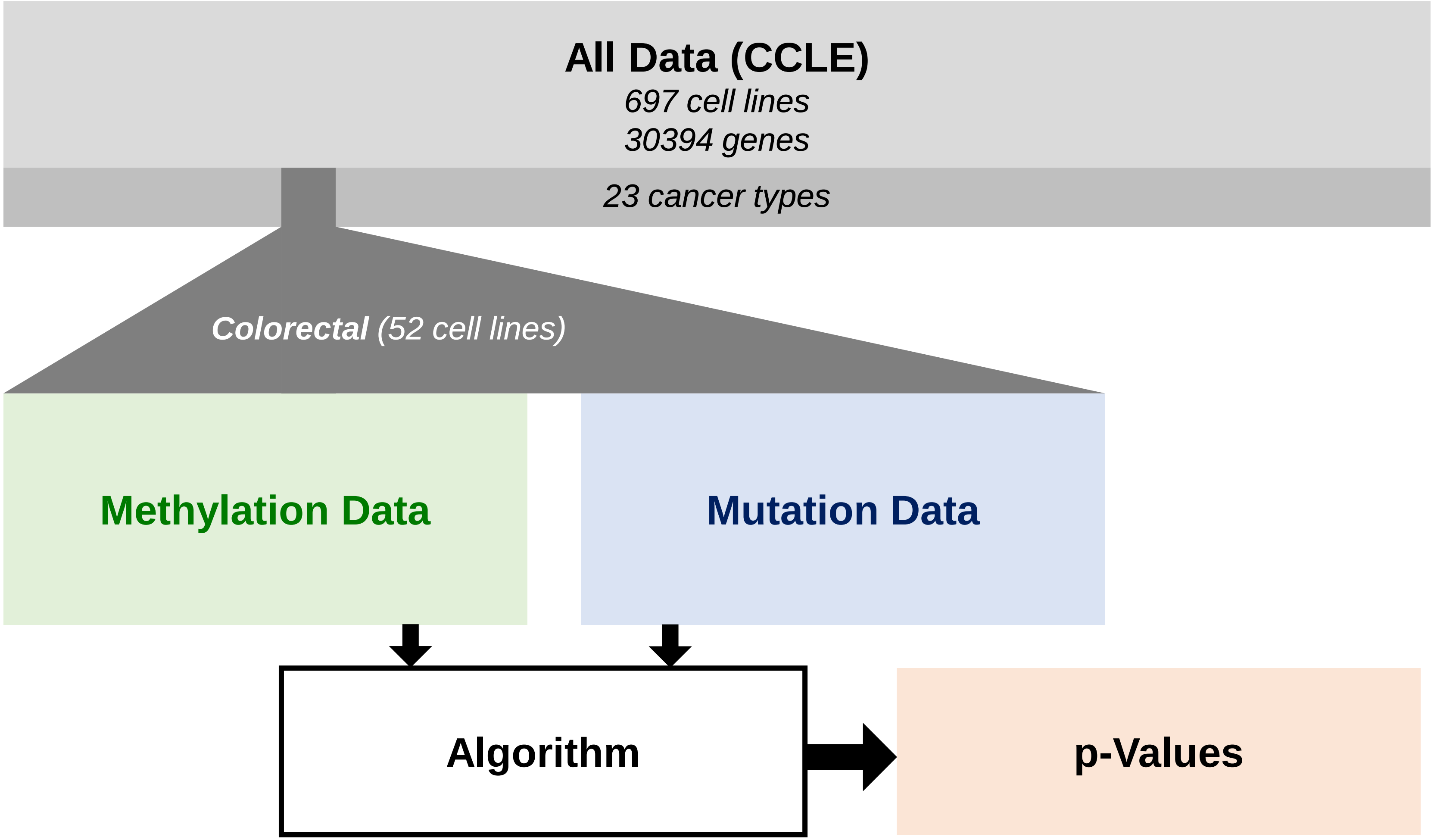
Methylation Cluster 2

No mutations in Gene *X*

Fig. 5: Sample cell lines clustered by methylation

- **Null hypothesis:** There is no relationship between methylation and mutation in [*gene*] among cells of [*cancer type*]

Cancer Cell Line Data



All Data (CCLE)

697 cell lines

30394 genes

23 cancer types

Colorectal *(52 cell lines)*

Methylation Data

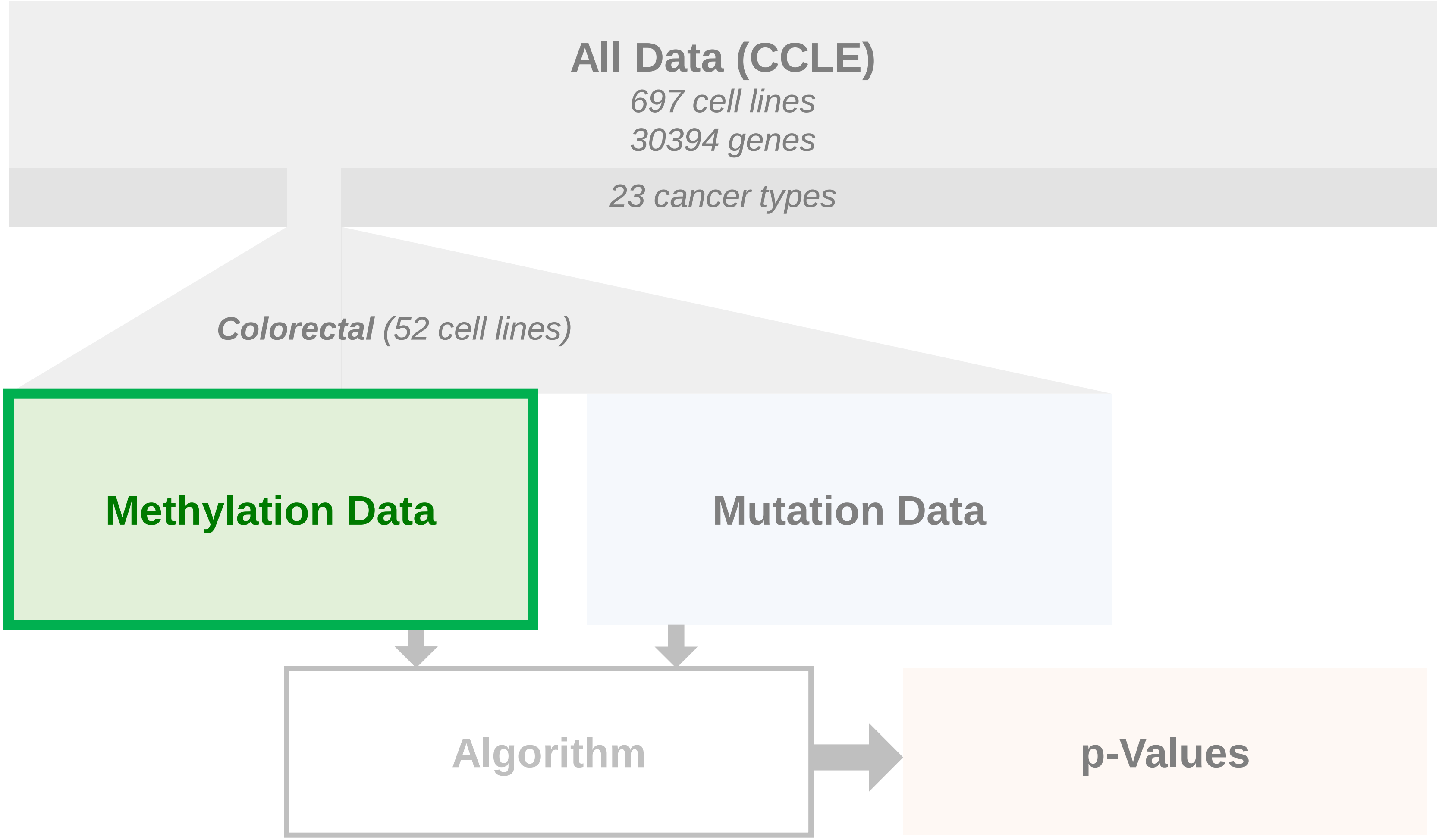
Mutation Data

Algorithm

p-Values

Fig. 6: Data pipeline for analyzing the correlation between methylation and mutation in a given cancer type

Cancer Cell Line Data



All Data (CCLE)

697 cell lines

30394 genes

23 cancer types

Colorectal *(52 cell lines)*

Methylation Data

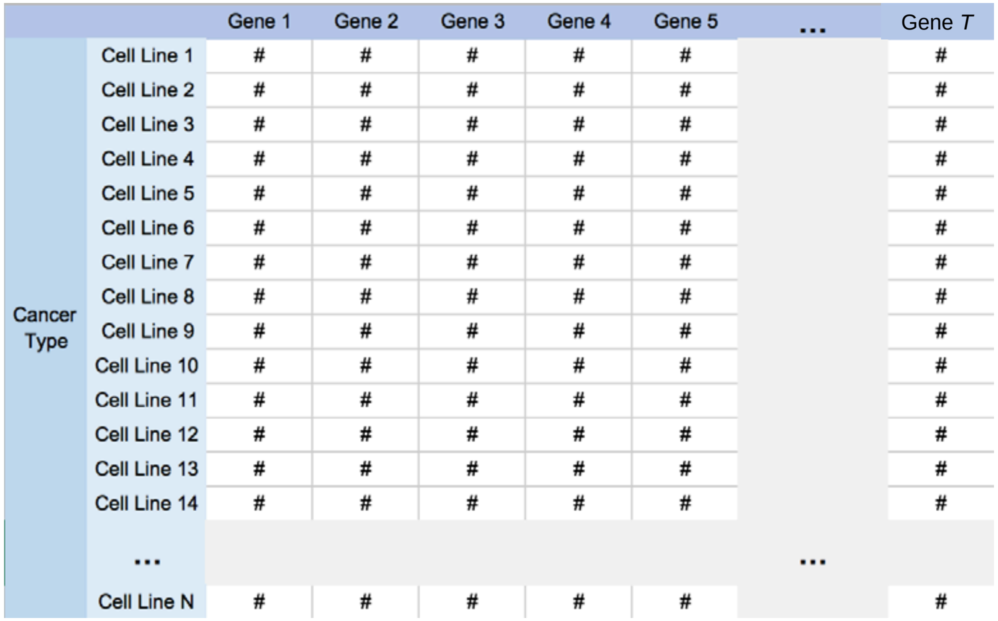
Mutation Data

Algorithm

p-Values

Fig. 6: Data pipeline for analyzing the correlation between methylation and mutation in a given cancer type

Methylation Data



		Gene 1	Gene 2	Gene 3	Gene 4	Gene 5	...	Gene *T*
Cancer Type	Cell Line 1	#	#	#	#	#		#
	Cell Line 2	#	#	#	#	#		#
	Cell Line 3	#	#	#	#	#		#
	Cell Line 4	#	#	#	#	#		#
	Cell Line 5	#	#	#	#	#		#
	Cell Line 6	#	#	#	#	#		#
	Cell Line 7	#	#	#	#	#		#
	Cell Line 8	#	#	#	#	#		#
	Cell Line 9	#	#	#	#	#		#
	Cell Line 10	#	#	#	#	#		#
	Cell Line 11	#	#	#	#	#		#
	Cell Line 12	#	#	#	#	#		#
	Cell Line 13	#	#	#	#	#		#
	Cell Line 14	#	#	#	#	#		#
	...						...	
	Cell Line N	#	#	#	#	#		#

Fig. 9A:
Methylation data

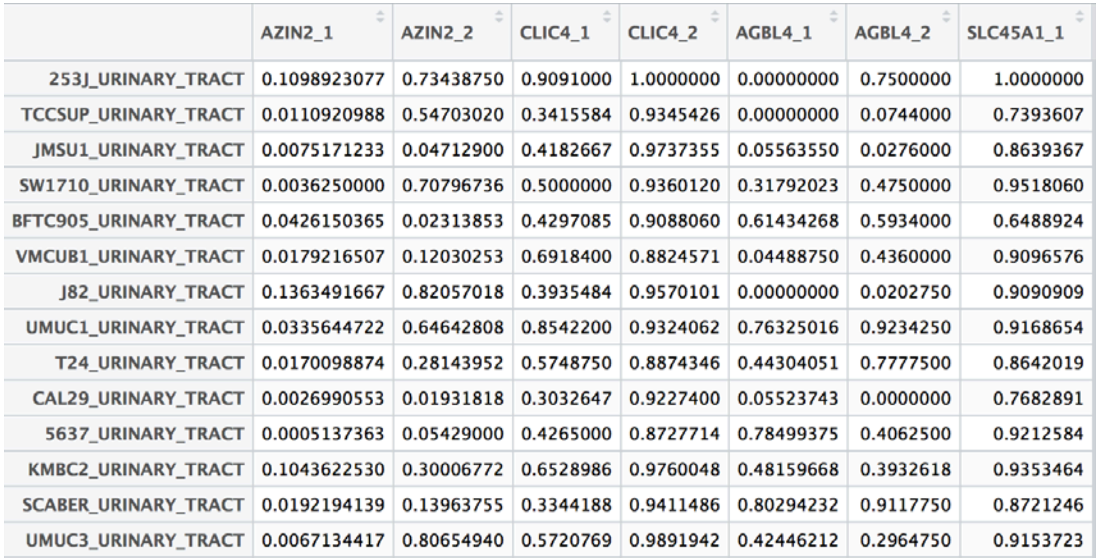
Calculate the Euclidean distance between every pair of cell lines, now in *T* dimensions:

$$d_{ij} = \sqrt{\Delta m_{1ij}^{\;2} + \Delta m_{2ij}^{\;2} + \cdots + \Delta m_{Tij}^{\;2}}$$

Methylation Data

Gene Name

Cell Line



	AZIN2_1	AZIN2_2	CLIC4_1	CLIC4_2	AGBL4_1	AGBL4_2	SLC45A1_1
253J_URINARY_TRACT	0.1098923077	0.73438750	0.9091000	1.0000000	0.00000000	0.7500000	1.0000000
TCCSUP_URINARY_TRACT	0.0110920988	0.54703020	0.3415584	0.9345426	0.00000000	0.0744000	0.7393607
JMSU1_URINARY_TRACT	0.0075171233	0.04712900	0.4182667	0.9737355	0.05563550	0.0276000	0.8639367
SW1710_URINARY_TRACT	0.0036250000	0.70796736	0.5000000	0.9360120	0.31792023	0.4750000	0.9518060
BFTC905_URINARY_TRACT	0.0426150365	0.02313853	0.4297085	0.9088060	0.61434268	0.5934000	0.6488924
VMCUB1_URINARY_TRACT	0.0179216507	0.12030253	0.6918400	0.8824571	0.04488750	0.4360000	0.9096576
J82_URINARY_TRACT	0.1363491667	0.82057018	0.3935484	0.9570101	0.00000000	0.0202750	0.9090909
UMUC1_URINARY_TRACT	0.0335644722	0.64642808	0.8542200	0.9324062	0.76325016	0.9234250	0.9168654
T24_URINARY_TRACT	0.0170098874	0.28143952	0.5748750	0.8874346	0.44304051	0.7777500	0.8642019
CAL29_URINARY_TRACT	0.0026990553	0.01931818	0.3032647	0.9227400	0.05523743	0.0000000	0.7682891
5637_URINARY_TRACT	0.0005137363	0.05429000	0.4265000	0.8727714	0.78499375	0.4062500	0.9212584
KMBC2_URINARY_TRACT	0.1043622530	0.30006772	0.6528986	0.9760048	0.48159668	0.3932618	0.9353464
SCABER_URINARY_TRACT	0.0192194139	0.13963755	0.3344188	0.9411486	0.80294232	0.9117750	0.8721246
UMUC3_URINARY_TRACT	0.0067134417	0.80654940	0.5720769	0.9891942	0.42446212	0.2964750	0.9153723

(30387 more columns)

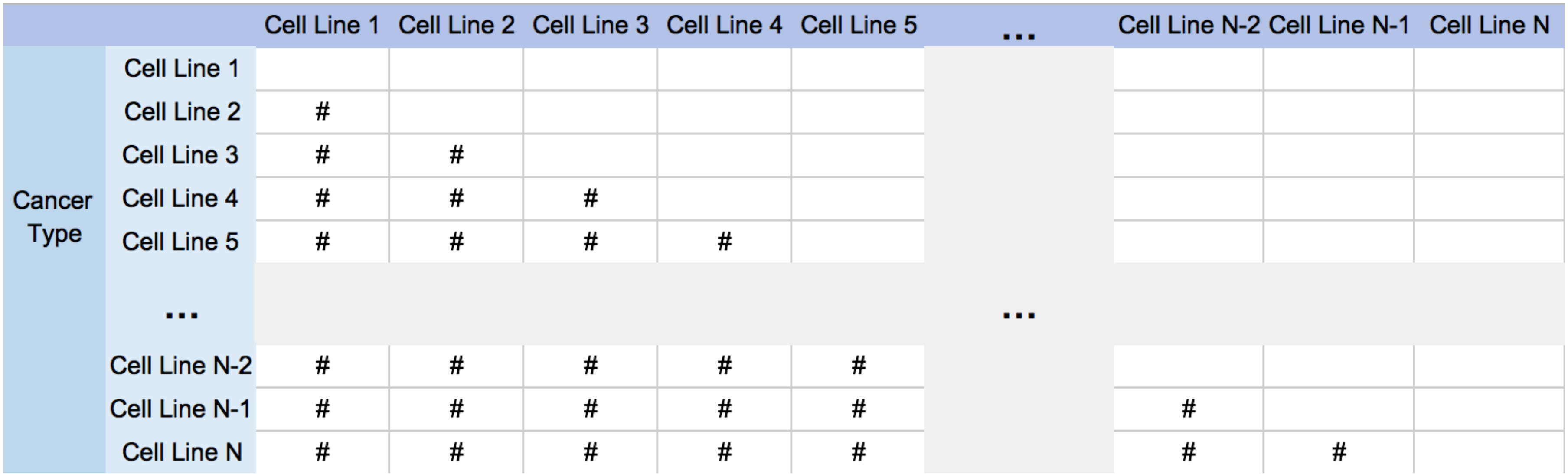
(9 more rows)

Fig. 9B: Section of methylation data for bladder cancer cell lines

Calculate the Euclidean distance between every pair of cell lines, now in *T* dimensions:

$$d_{ij} = \sqrt{\Delta m_{1ij}{}^2 + \Delta m_{2ij}{}^2 + \cdots + \Delta m_{Tij}{}^2}$$

Methylation Distance Matrix



		Cell Line 1	Cell Line 2	Cell Line 3	Cell Line 4	Cell Line 5	...	Cell Line N-2	Cell Line N-1	Cell Line N
Cancer Type	Cell Line 1									
	Cell Line 2	#								
	Cell Line 3	#	#							
	Cell Line 4	#	#	#						
	Cell Line 5	#	#	#	#					
	...						...			
	Cell Line N-2	#	#	#	#	#				
	Cell Line N-1	#	#	#	#	#		#		
	Cell Line N	#	#	#	#	#		#	#	

Fig. 9C: *N* x *N* Methylation distance matrix

Cancer Cell Line Data

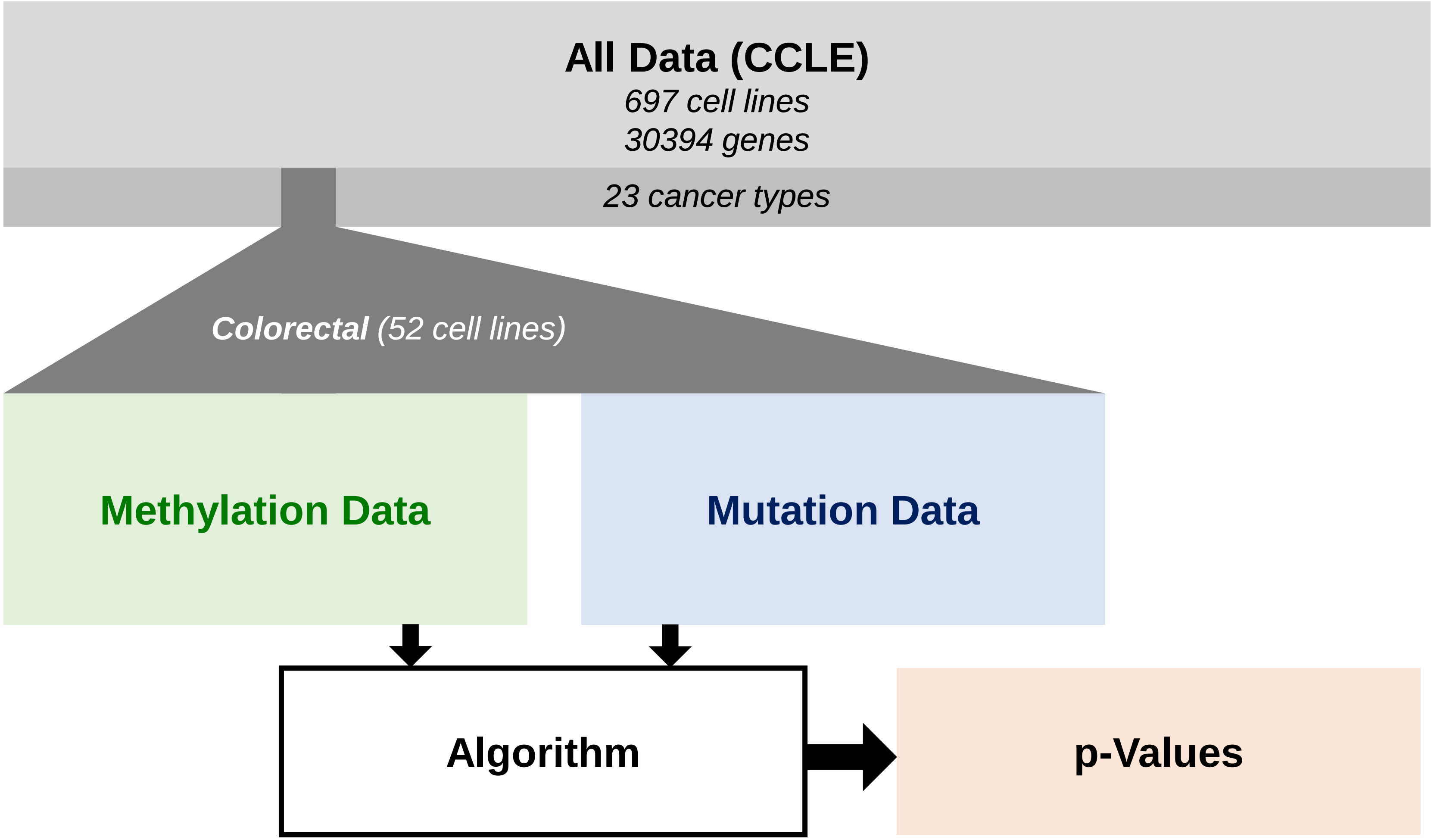
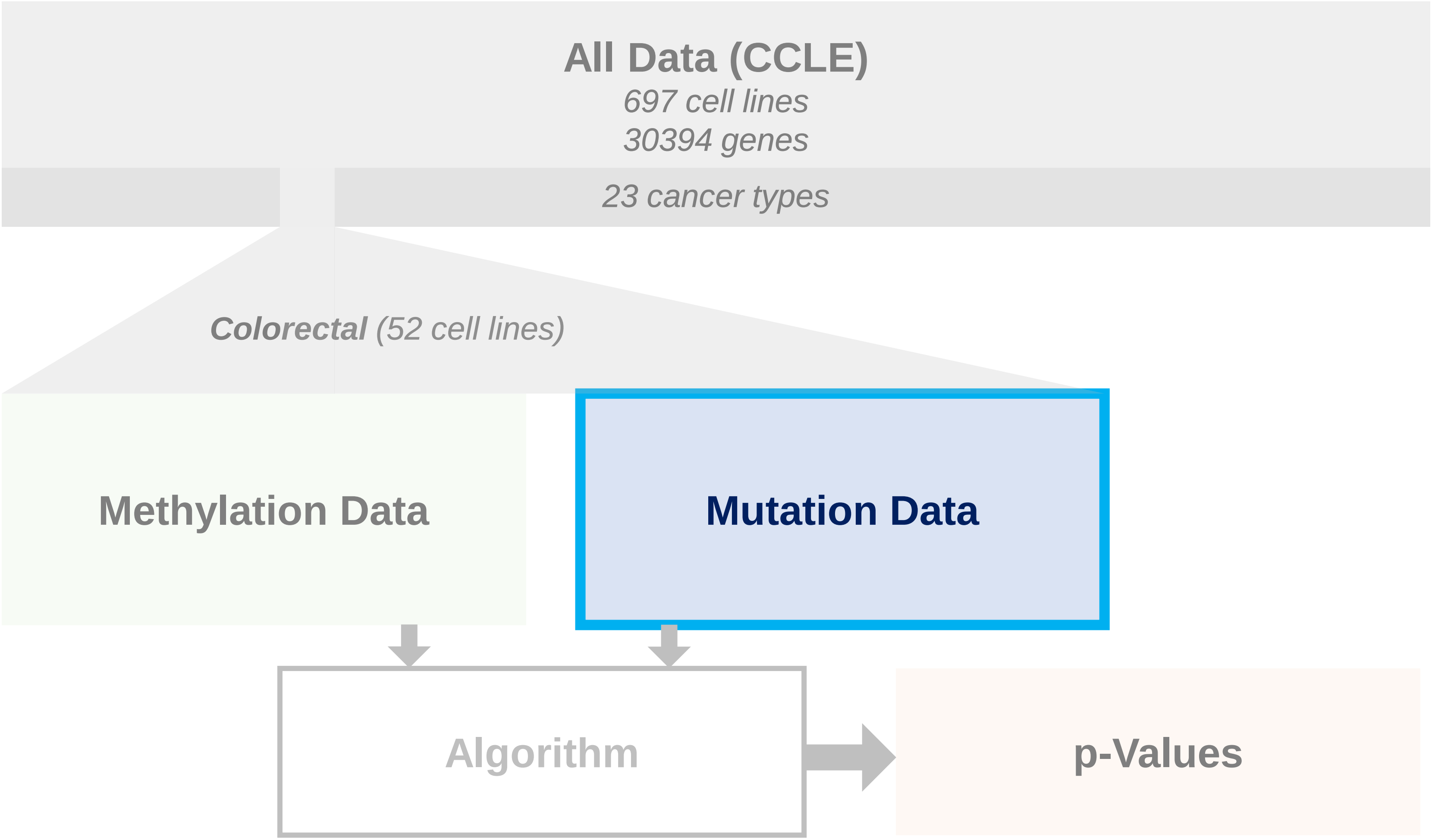


Fig. 6: Data pipeline for analyzing the correlation between methylation and mutation in a given cancer type

Cancer Cell Line Data



All Data (CCLE)
697 cell lines
30394 genes

23 cancer types

***Colorectal** (52 cell lines)*

Methylation Data

Mutation Data

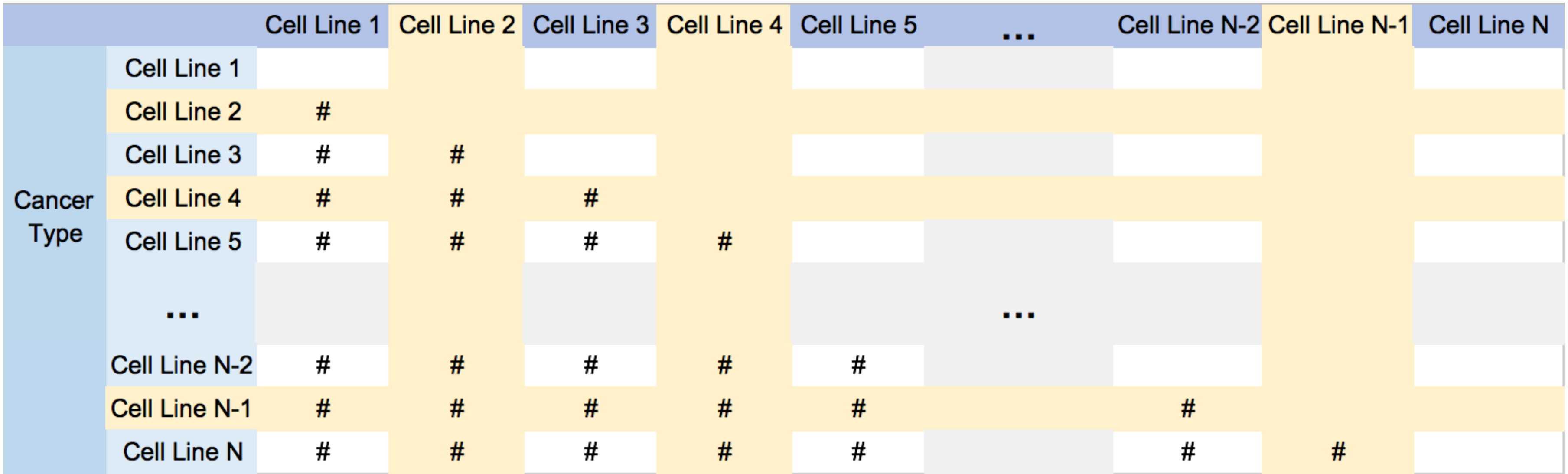
Algorithm

p-Values

Fig. 6: Data pipeline for analyzing the correlation between methylation and mutation in a given cancer type

Incorporating Mutation Data

Gene *A*

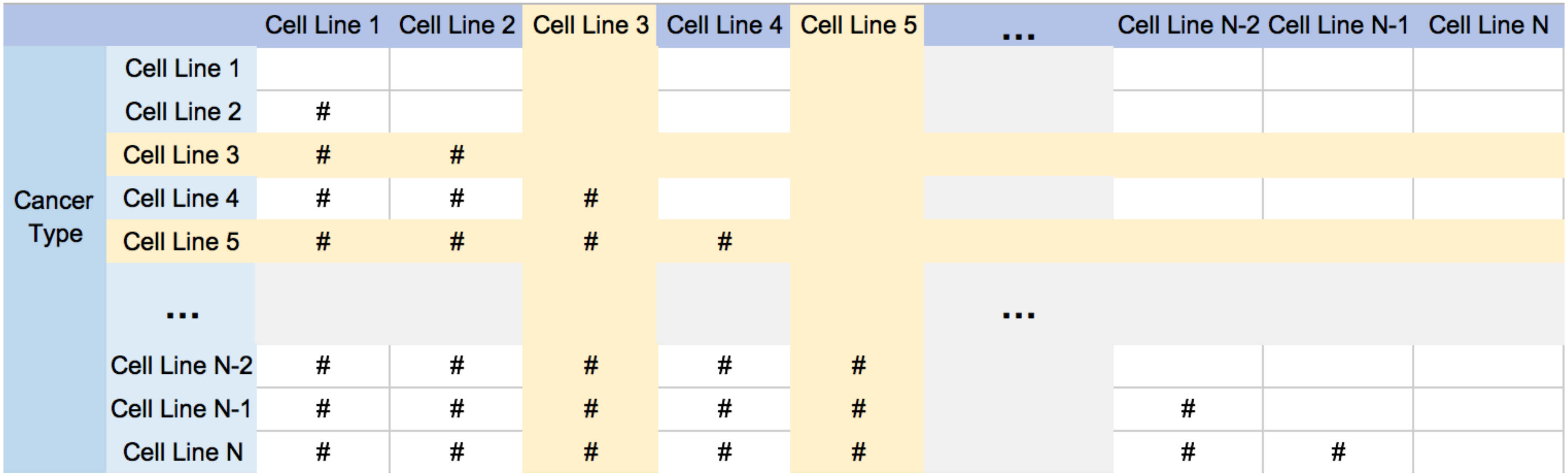


		Cell Line 1	Cell Line 2	Cell Line 3	Cell Line 4	Cell Line 5	…	Cell Line N-2	Cell Line N-1	Cell Line N
Cancer Type	Cell Line 1									
	Cell Line 2	#								
	Cell Line 3	#	#							
	Cell Line 4	#	#	#						
	Cell Line 5	#	#	#	#					
	…						…			
	Cell Line N-2	#	#	#	#	#				
	Cell Line N-1	#	#	#	#	#		#		
	Cell Line N	#	#	#	#	#		#	#	

Fig. 10A: Cell lines **2, 4, and *N*–1** have mutations in Gene *A*

Incorporating Mutation Data

Gene *B*

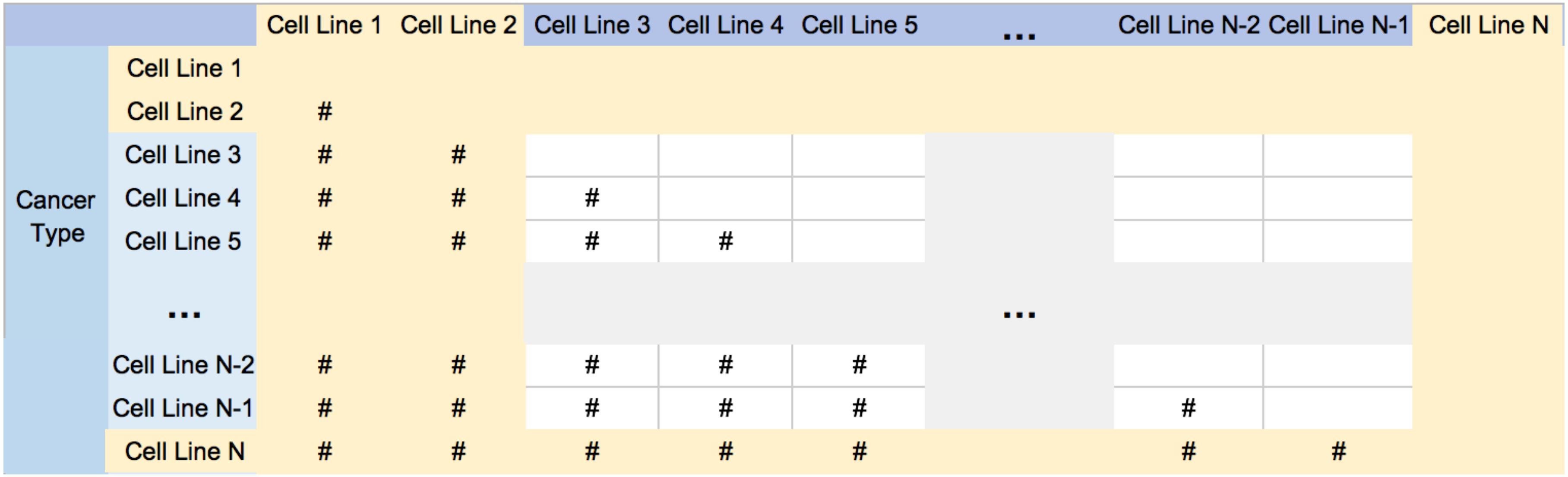


		Cell Line 1	Cell Line 2	Cell Line 3	Cell Line 4	Cell Line 5	...	Cell Line N-2	Cell Line N-1	Cell Line N
Cancer Type	Cell Line 1									
	Cell Line 2	#								
	Cell Line 3	#	#							
	Cell Line 4	#	#	#						
	Cell Line 5	#	#	#	#					
	...						...			
	Cell Line N-2	#	#	#	#	#				
	Cell Line N-1	#	#	#	#	#		#		
	Cell Line N	#	#	#	#	#		#	#	

Fig. 10B: Cell lines **3 and 5** have mutations in Gene *B*

Incorporating Mutation Data

Gene *C*

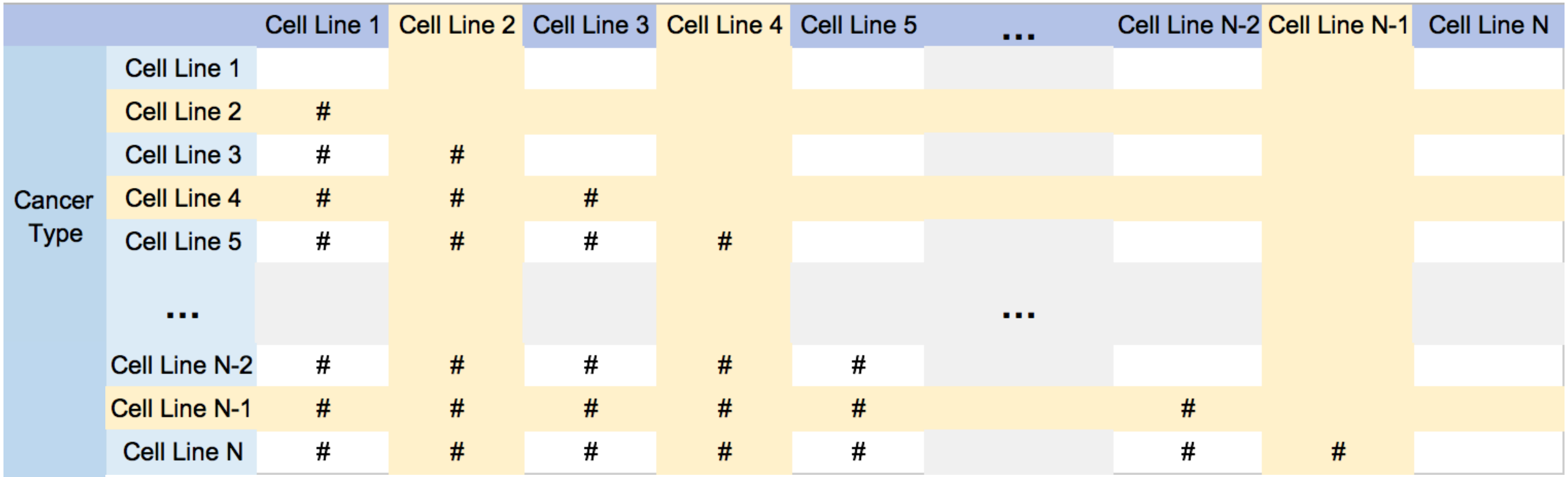


		Cell Line 1	Cell Line 2	Cell Line 3	Cell Line 4	Cell Line 5	…	Cell Line N-2	Cell Line N-1	Cell Line N
Cancer Type	Cell Line 1									
	Cell Line 2	#								
	Cell Line 3	#	#							
	Cell Line 4	#	#	#						
	Cell Line 5	#	#	#	#					
	…						…			
	Cell Line N-2	#	#	#	#	#				
	Cell Line N-1	#	#	#	#	#		#		
	Cell Line N	#	#	#	#	#		#	#	

Fig. 10C: Cell lines **1, 2, and N** have mutations in Gene *C*

The Algorithm

Gene *X*

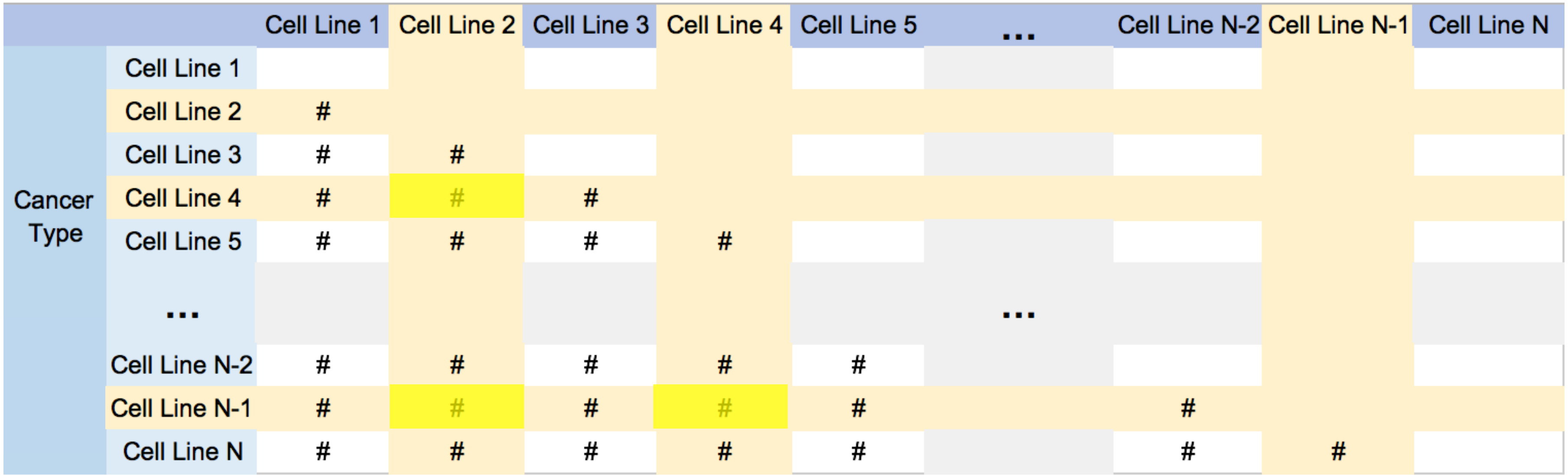


		Cell Line 1	Cell Line 2	Cell Line 3	Cell Line 4	Cell Line 5	…	Cell Line N-2	Cell Line N-1	Cell Line N
Cancer Type	Cell Line 1									
	Cell Line 2	#								
	Cell Line 3	#	#							
	Cell Line 4	#	#	#						
	Cell Line 5	#	#	#	#					
	…						…			
	Cell Line N-2	#	#	#	#	#				
	Cell Line N-1	#	#	#	#	#		#		
	Cell Line N	#	#	#	#	#		#	#	

- Suppose that *k* out of *N* cell lines have a mutation in gene *X*
- There are *N(N–1)/2* distances in the entire distance matrix
- There are *k(k–1)/2* distances between the cell lines with mutations

The Algorithm

Gene *X*



		Cell Line 1	Cell Line 2	Cell Line 3	Cell Line 4	Cell Line 5	…	Cell Line N-2	Cell Line N-1	Cell Line N
Cancer Type	Cell Line 1									
	Cell Line 2	#								
	Cell Line 3	#	#							
	Cell Line 4	#	#	#						
	Cell Line 5	#	#	#	#					
	…						…			
	Cell Line N-2	#	#	#	#	#				
	Cell Line N-1	#	#	#	#	#		#		
	Cell Line N	#	#	#	#	#		#	#	

- Call the distribution of all values in the matrix {*A*}
- Call the sample of *k(k–1)/2* distances between mutated cell lines {*K*}

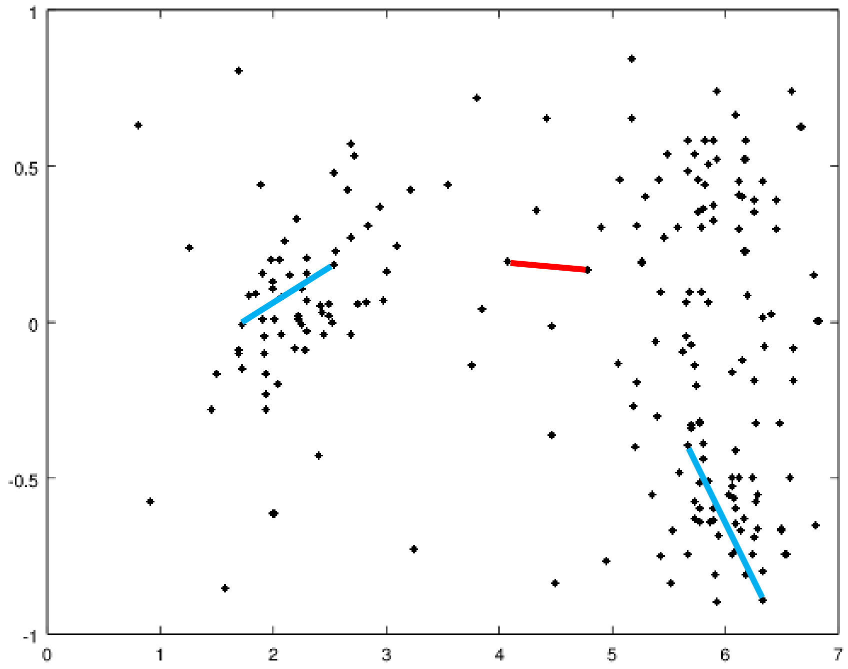
The Algorithm

- What is the probability that the sample {K} occurs entirely by random chance?
- Randomly select 10,000 samples of size $k(k–1)/2$ within the N x N half-matrix, and call these samples {R_1} thru {$R_{10,000}$}
- Use the **Kolmogorov-Smirnov (KS) test,** which returns the likelihood that a given sample is derived from some reference distribution (similarity score)
- The p-value will be the percentage of the time that KS({K}, {A}) exceeds KS({R_i}, {A}) as i ranges from 1 to 10,000
- Repeat process for each gene; p-values corrected for multiple hypothesis testing using Benjamini–Hochberg (BH)

Improving the Algorithm

- Instead of using the distance matrix to calculate p-values, we can use the **cophenetic distance matrix** based on hierarchical clustering instead
- Takes into account the proximity of other data points; amplifies biologically relevant signal and removes noise

Fig. 11: Example of cophenetic distance in 2D sample data



Visualizing the output

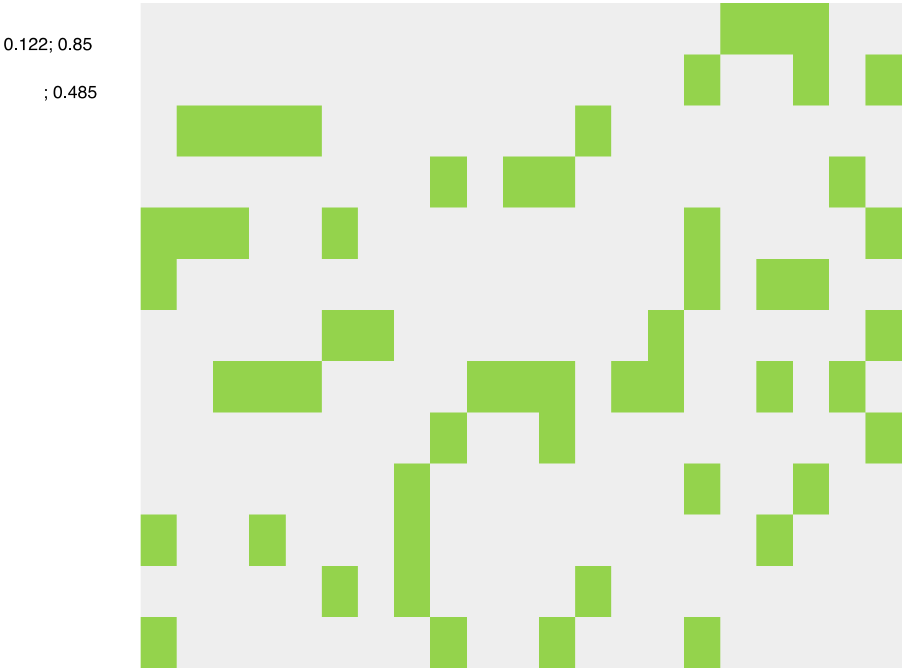


Fig. 12: Heatmap of deletion mutations in top-scoring genes in kidney cancer cell lines

Sample Results

Ly

and 2

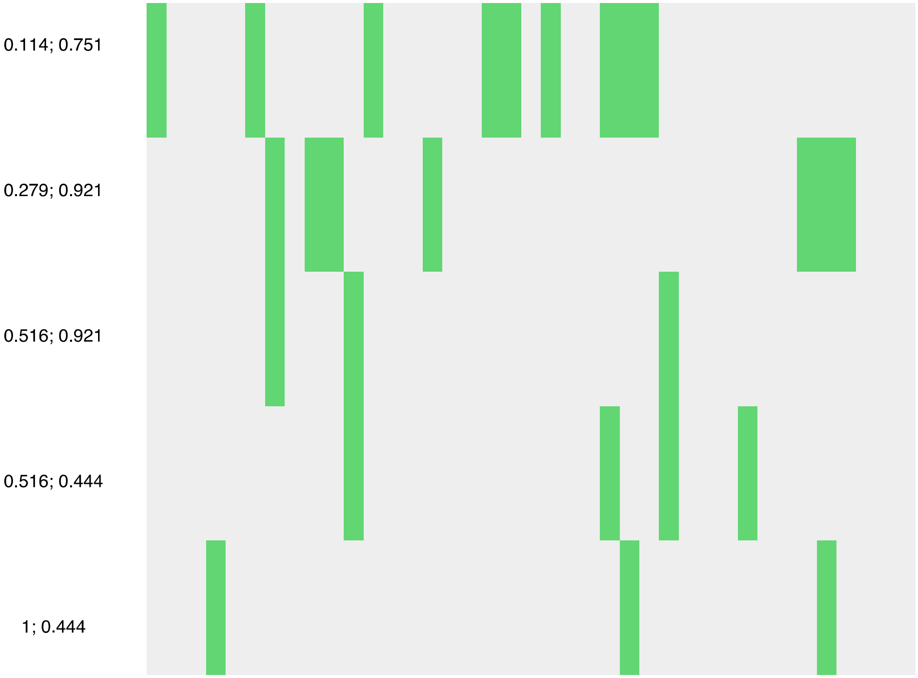


Fig. 13: *MYC* is a known oncogene in B-cell lymphomas and other cancers

Generalizing the algorithm

- Clustered cell lines based on exon splicing instead of methylation
- Find correlation between exon splicing patterns and mutations
- Could use algorithm for other applications in the future

Sample Results

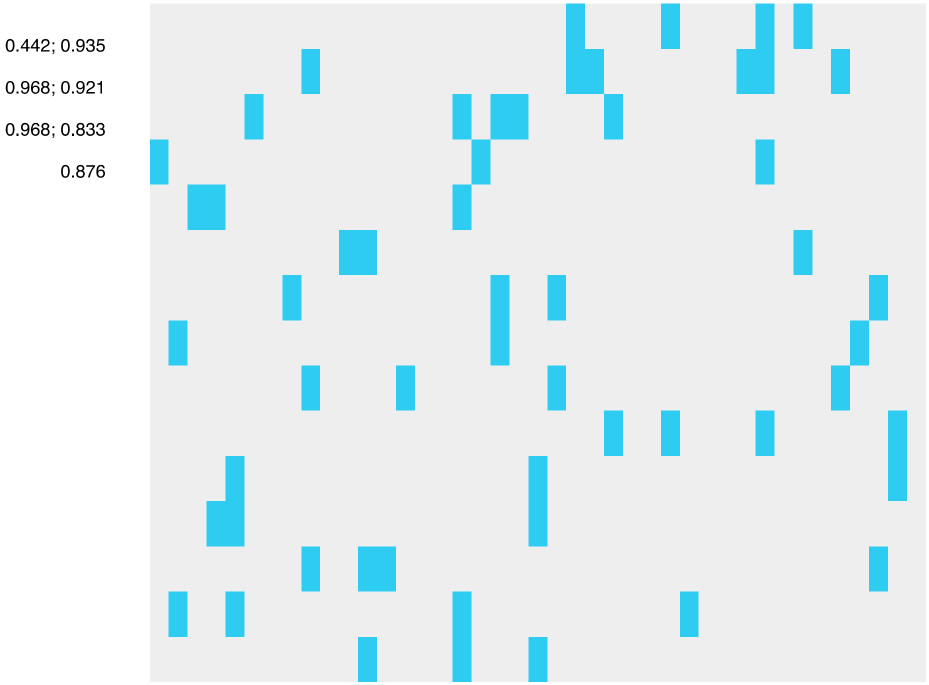
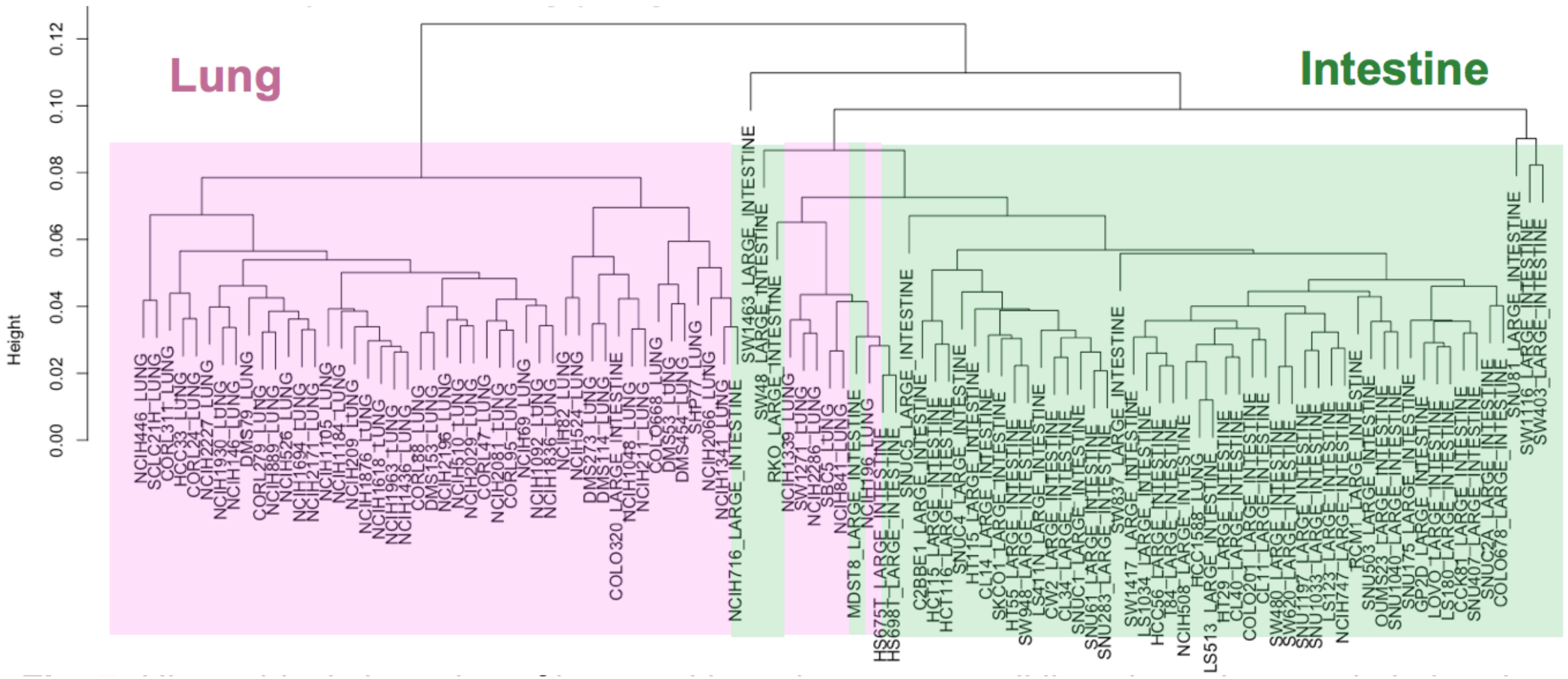


Fig. 14: *SF3B1* is known to affect splicing patterns in pancreatic cancer

Acknowledgements

I would like to thank:

- Dr. Mahmoud Ghandi
- The Broad-Novartis Cancer Cell Line Encyclopedia (CCLE) for providing the data used in this study
- The PRIMES program coordinators
- My parents


Lung
Intestine
Height
0.00
0.02
0.04
0.06
0.08
0.10
0.12
NCIH446_LUNG
SCLC21H_LUNG
CORL311_LUNG
HCC33_LUNG
CORL24_LUNG
NCIH2227_LUNG
NCIH1930_LUNG
NCIH146_LUNG
DMS79_LUNG
CORL279_LUNG
NCIH889_LUNG
NCIH526_LUNG
NCIH1694_LUNG
NCIH2171_LUNG
NCIH1105_LUNG
NCIH1184_LUNG
NCIH209_LUNG
NCIH1876_LUNG
NCIH1618_LUNG
NCIH1963_LUNG
NCIH1436_LUNG
CORL88_LUNG
DMS153_LUNG
NCIH2196_LUNG
NCIH510_LUNG
NCIH2029_LUNG
CORL47_LUNG
NCIH2081_LUNG
CORL95_LUNG
NCIH69_LUNG
NCIH1092_LUNG
NCIH1836_LUNG
NCIH82_LUNG
NCIH524_LUNG
DMS273_LUNG
DMS114_LUNG
COLO320_LARGE_INTESTINE
NCIH1048_LUNG
NCIH211_LUNG
COLO668_LUNG
DMS53_LUNG
DMS454_LUNG
SHP77_LUNG
NCIH2066_LUNG
NCIH1341_LUNG
NCIH716_LARGE_INTESTINE
SW1463_LARGE_INTESTINE
SW48_LARGE_INTESTINE
RKO_LARGE_INTESTINE
NCIH1339_LUNG
SW1271_LUNG
NCIH2286_LUNG
SBC5_LUNG
NCIH841_LUNG
MDST8_LARGE_INTESTINE
NCIH196_LUNG
HS675T_LARGE_INTESTINE
HS698T_LARGE_INTESTINE
SNUC5_LARGE_INTESTINE
C2BBE1_LARGE_INTESTINE
HCT15_LARGE_INTESTINE
HCT116_LARGE_INTESTINE
SNUC4_LARGE_INTESTINE
HT115_LARGE_INTESTINE
CL14_LARGE_INTESTINE
SKCO1_LARGE_INTESTINE
HT55_LARGE_INTESTINE
SW948_LARGE_INTESTINE
LS411N_LARGE_INTESTINE
CW2_LARGE_INTESTINE
CL34_LARGE_INTESTINE
SNUC1_LARGE_INTESTINE
SNU61_LARGE_INTESTINE
SNU283_LARGE_INTESTINE
SW837_LARGE_INTESTINE
SW1417_LARGE_INTESTINE
LS1034_LARGE_INTESTINE
HCC56_LARGE_INTESTINE
T84_LARGE_INTESTINE
NCIH508_LARGE_INTESTINE
HCC1588_LUNG
LS513_LARGE_INTESTINE
HT29_LARGE_INTESTINE
CL40_LARGE_INTESTINE
COLO201_LARGE_INTESTINE
CL11_LARGE_INTESTINE
SW480_LARGE_INTESTINE
SW620_LARGE_INTESTINE
SNU1197_LARGE_INTESTINE
SNU1033_LARGE_INTESTINE
LS123_LARGE_INTESTINE
NCIH747_LARGE_INTESTINE
RCM1_LARGE_INTESTINE
SNU503_LARGE_INTESTINE
OUMS23_LARGE_INTESTINE
SNU1040_LARGE_INTESTINE
SNU175_LARGE_INTESTINE
GP2D_LARGE_INTESTINE
LOVO_LARGE_INTESTINE
LS180_LARGE_INTESTINE
CCK81_LARGE_INTESTINE
SNU407_LARGE_INTESTINE
SNUC2A_LARGE_INTESTINE
COLO678_LARGE_INTESTINE
SNU81_LARGE_INTESTINE
SW1116_LARGE_INTESTINE
SW403_LARGE_INTESTINE