AFRICAN CENTERS OF EXCELLENCE IN BIOINFORMATICS

KAMPALA, UGANDA

Gene Regulatory Networks 2

Today's Instructor

Yunhua Zhu,
Ph.D in stem cell biology

Computational Genomics Specialist – Transcriptomics

Bachelor in biochemistry @NUS
Ph.D in stem cell biology @NUS
Postdoc in neurodegeneration, single cell biology @JHU
Bioinformatician @ NIH

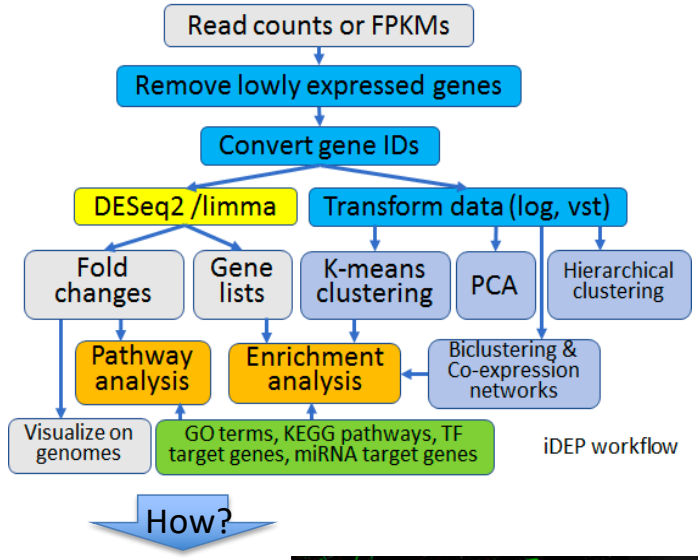
Recent projects
- Single cell analysis of cancer stem cells
- Single cell analysis of immue response in lung cancer
- Bulk RNA-seq on irradiation, HIV blood samples

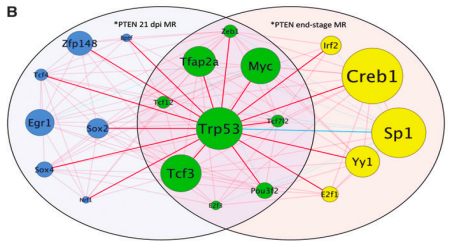
- PLEASE JOIN THE DISCUSSION

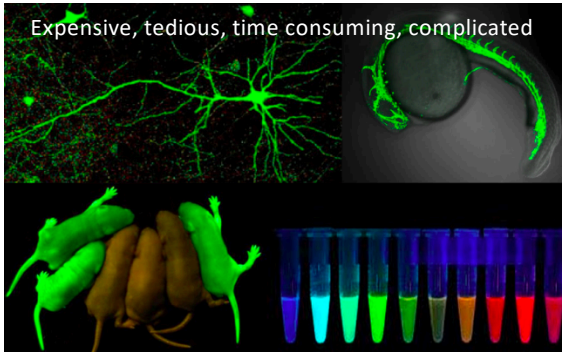
- Bioinformatics and Computational Biosciences Branch (BCBB), NIAID
- National Institutes of Health, Bethesda, MD USA.
- Contact our team via email:
  - Github: git clone **https://github.com/niaid/Gene_Regulatory_Networks**
  - Googledoc, **https://tinyurl.com/zhu-GRN**
  - Instructors: zhuy16@nih.gov
  - Server: ssh username@137.63.194.9:
    /home/bcbb_teaching_files/zhuy/cisTarget_databases
  - Email: bioinformatics@niaid.nih.gov
  - Linkedin: https://tinyurl.com/zhu-linkedin

Why study gene regulatory networks

- Why
  - High throughput methods are just screening procedures
  - Gene lists do not inform relationship between genes
  - Regulatory information is hidden in the statistical relationships
  - Down-stream validation is costly, tediously and risky
- Aims
  - To find regulatory network, to inform finding key candidate genes/master regulators for validation
- Challenges
  - Data are massive, and mostly static that do not reflect dynamic regulation
  - Relationships can take various forms, linear or non-linear







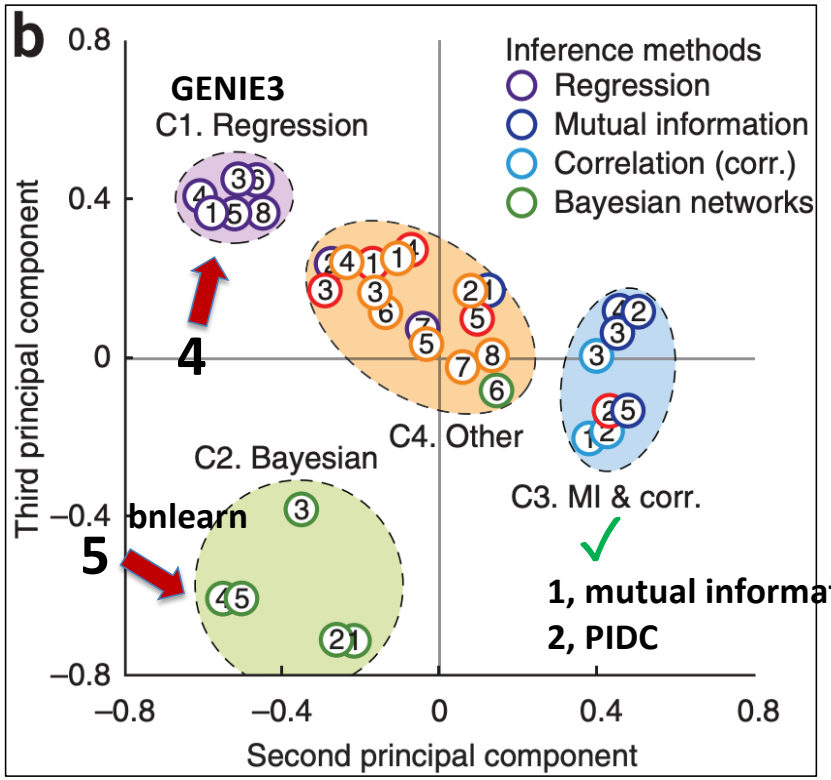
Overview of Tools in GRN inference

Software	ARACNE	NetworkInference/ PIDC	bnlearn	GENIE3	iRegulon	SCENIC
semantics	Mutual information	Partial Information Decomposition	Bayes theory	Random Forest, Regression tree	Promoter and TF binding sequence, database	Combination of regression and promoter sequence
years published	2006	2017	2009	2010	2014	2017
No. of cited	2179	82	894	658	337	265
FullName/ explanation	Algorithm for the Reconstruction of Accurate Cellular Networks	Using proportional unique contribution (PUC) to a target gene	Bayes net structure and parameter learning, causality	GEne Network Inference with Ensemble of trees	reverse-engineer the transcriptional regulatory network with regulatory sequence analysis	single-cell regulatory network inference and clustering
Implementation	GUI (geWorkbench)	Julia	R	R	GUI (Cytoscape)	R, Python
type of experiment	Microarray, bulk RNA-seq	Single cell data	General	single cell data	a list of gene names	single cell data
input format	.exp and csv(ref)	csv, matrix	csv, matrix	csv	a list	csv/loom file
output	xgmml	network file	directed network file	network file	network file/binding sequences	network file/heatmap

Today

Next week

Method 4, iRegulon – prediction of upstream transcriptional factors based on promoter sequences



4

5

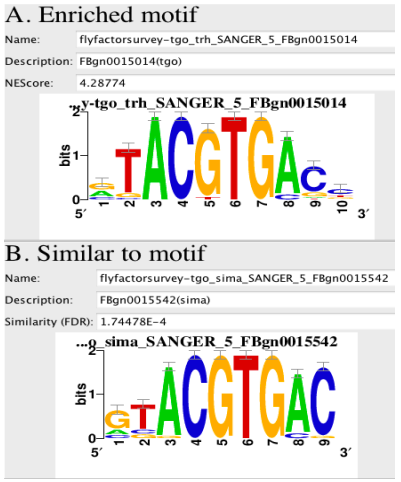
1, mutual information

2, PIDC

Based on statistical relationship between expression values of genes

3

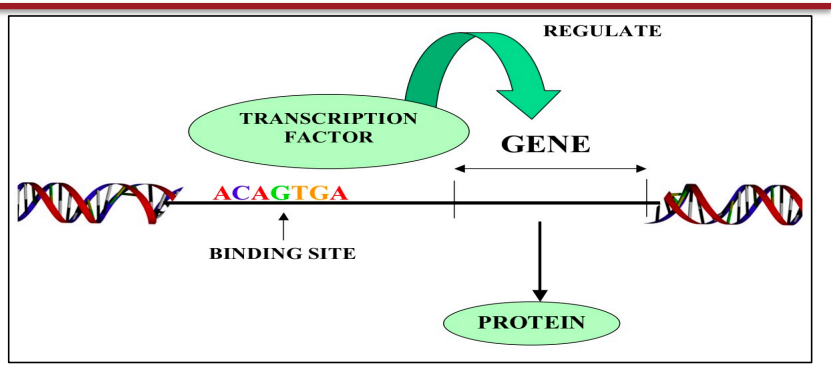
iRegulon/cis-Target

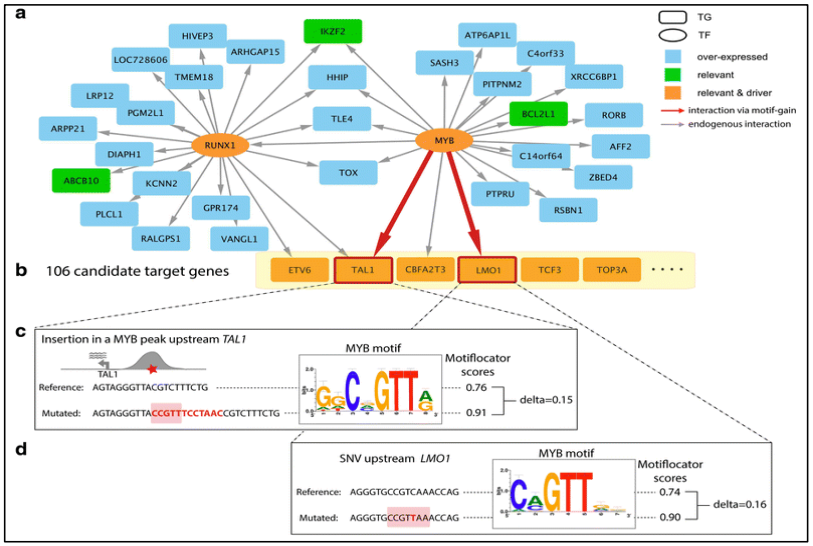


Using database to predict upstream transcriptional regulators

Introduction to iRegulon

- Transcriptional activity correlates with TF binding on promoter of target genes.
- One transcriptional factor activates a group of genes to achieve a biological function.
- TF expression increases it activity.
- Transcriptional factor has conserved binding sequence/motif.
- Many of the binding sequences have been validated.





Kalender Atak et al. Genome Medicine (2017) 9:80

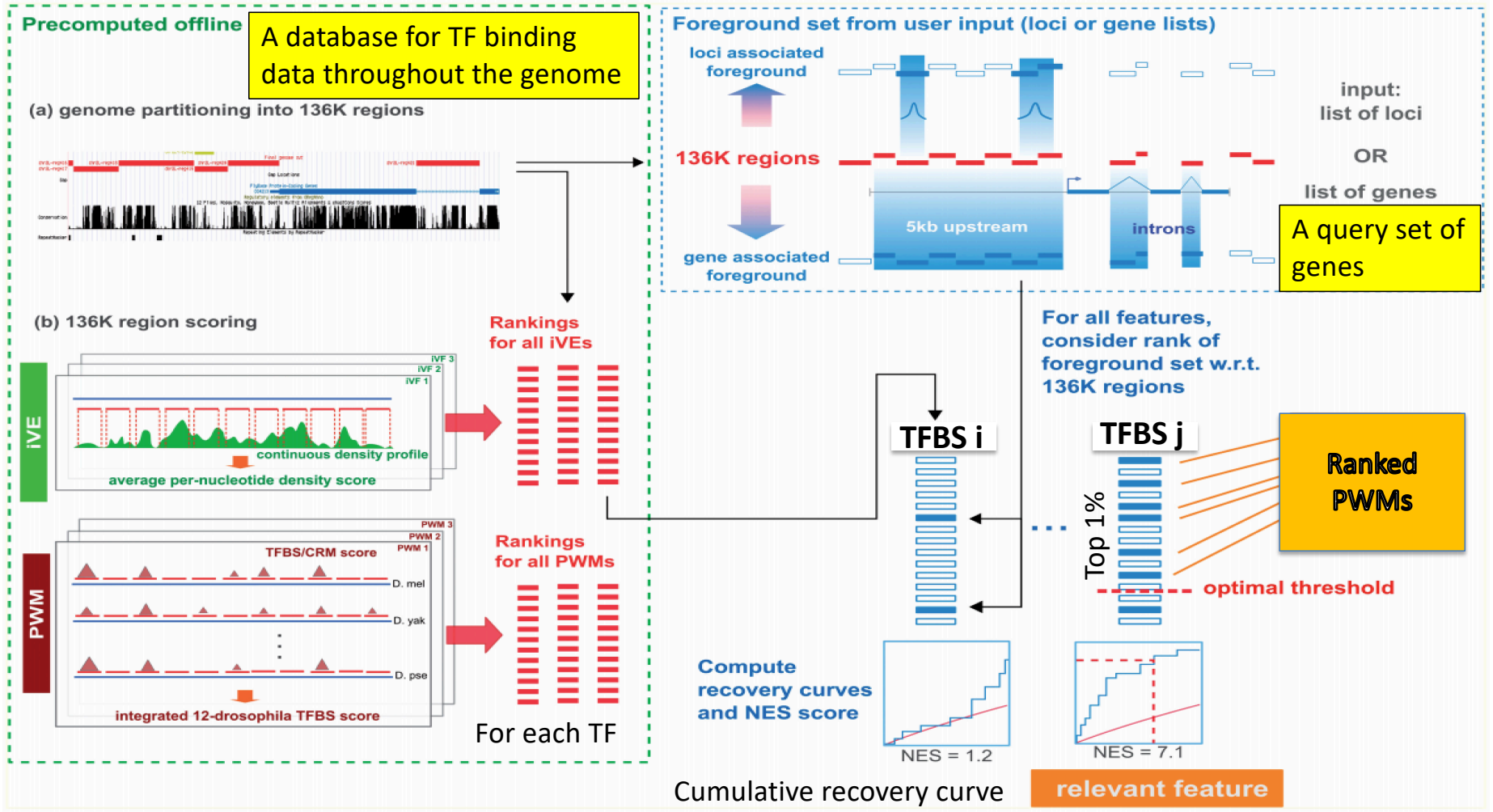
Source of the TFBS/Motif database

Table 1. Description of the motif and track collections used.

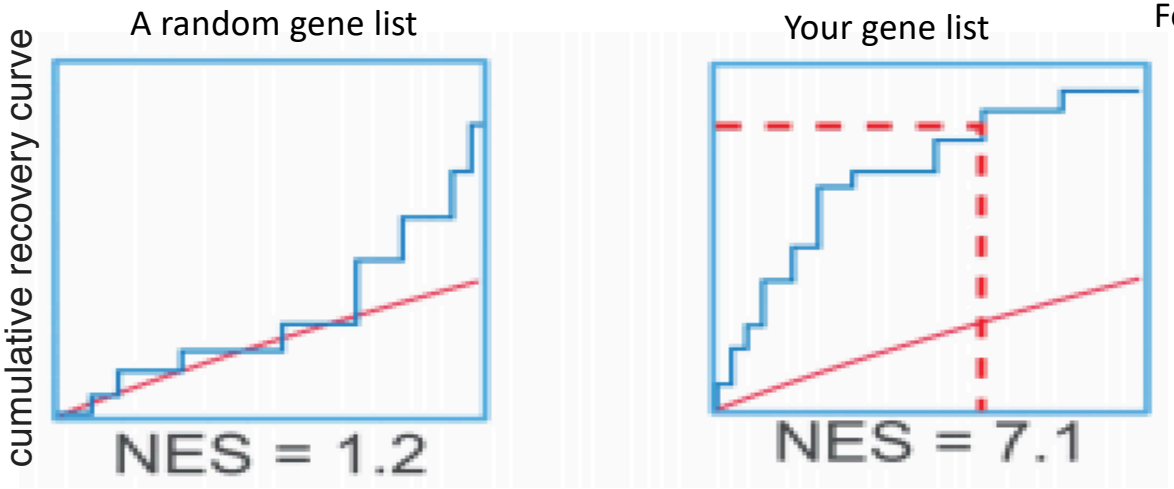
Source	Organism(s)	Type of motif	# motifs "6K"	# motifs "10K"	# tracks "1K ChIP"
Elemento [73]	Drosophila	Predicted (conserved)[a]	371	371	-
FlyFactorSurvey [75]	Drosophila	B-1H, others (e.g., FlyReg)	614	652	-
hPDI [77]	Human	Experimental	437	437	-
Jaspar [21]	Multiple species	Curated	1315	1315	-
SelexConsensus [76]	Drosophila	Curated (FlyReg)	38	38	-
Stark [74]	Drosophila	Predicted (conserved)[a]	228	228	-
Tiffin [76]	Drosophila	Predicted (gene sets)[a]	120	120	-
TRANSFAC PUBLIC [5]	Multiple species	Curated, ChIP-chip	398	398	-
TRANSFAC PRO [5]	Multiple species	Curated, ChIP-chip	1153	1850	-
YetFasco [78]	Yeast	Uniprobe, Curated, ChIP-chip	1709	1709	-
ENCODE [79]	Human	Predicted (from DHS)[a]	-	683	-
Factorbook [46]	Human	ENCODE ChIP-Seq motifs	-	79	-
Taipale [132]	Human, Mouse	HT-Selex	-	820	-
iDMMPMM [133]	Human	footprints, Selex, b1h, peaks	-	39	-
SwissRegulon [134]	Human	Curated	-	190	-
Wolfe [135]	Drosophila	ZFP motifs	-	36	-
HOMER [116]	Multiple species	ChIP-Seq Motifs, others (e.g. ENCODE)	-	1865	-
Dimers [136]	Human	Predicted dimers	-	603	-
ENCODE ChIP-Seq [23]	Human	-	-	-	999
Taipale ChIP-Seq [24]	Human	-	-	-	117
p53 and control ChIP-Seq (this study)	Human	-	-	-	2
Total			**6383**	11611 (**9713** nr)	**1118**

[a]Orphan motifs (unknown TFs).
nr = non-redundant.
doi:10.1371/journal.pcbi.1003731.t001

How iRegulon Work?

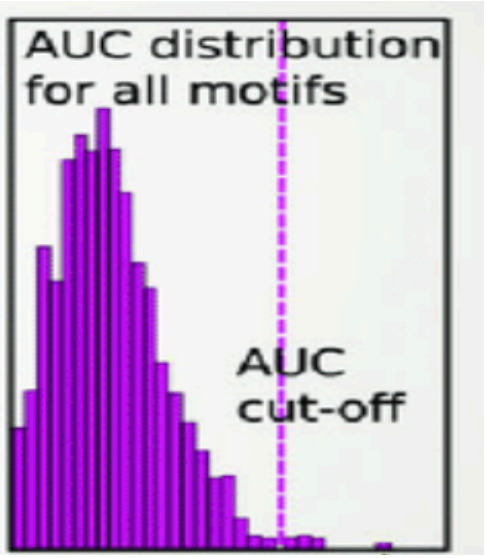


Normalized Enrichment Scores



For all TBBS/CRMs, **Area Under the Curve** (AUC) is calculated.
For top 1% of the rank. NES were calculated.

$$NES = \frac{AUC - AUCmean}{AUCstd}$$



Ranking of likelihood of all PWMs in all 137k loci, as a target for a TFBS.
(The ranking is based on a combination of motif clustering and comparative genomics.)

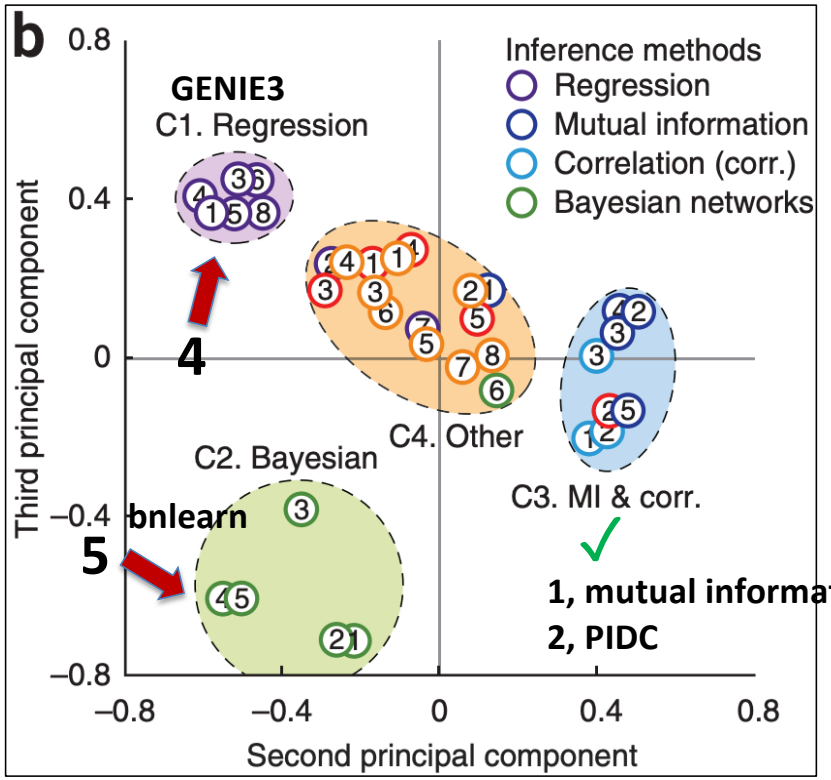
Demo on iRegulon

- How to import the gene list to Cytoscape
- Set parameters for iRegulon to run
- Interpret the output

Pro and cons

- Pros
  - Easy to generate hypothesis
  - Based on many coordinated genes in your data
  - Allow generation of a rich hypothesis
    - Binding sequence ---mutagenesis to test the TF binding.
    - Mechanism-- PCR to test the regulation of gene expression by the TF
    - Well informed on the biology as TF function is relatively well annotated
- Cons
  - The binding is mainly predicted by sequence enrichment, not guaranteed
  - Binding is dependent not only on sequence but also on DNA/histone modifications.
  - multiple TFs can bind to the same conserved sequence.

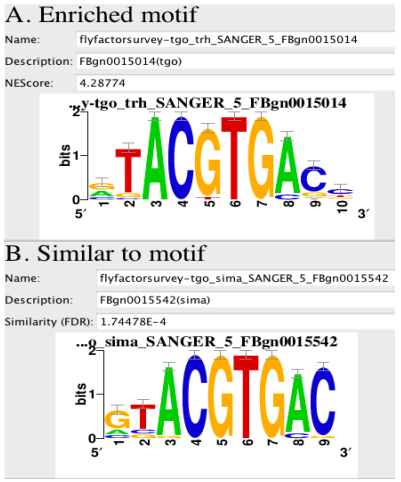
GENIE3--
GEne Network Inference with Ensemble of trees



Based on statistical relationship between genes

3

iRegulon

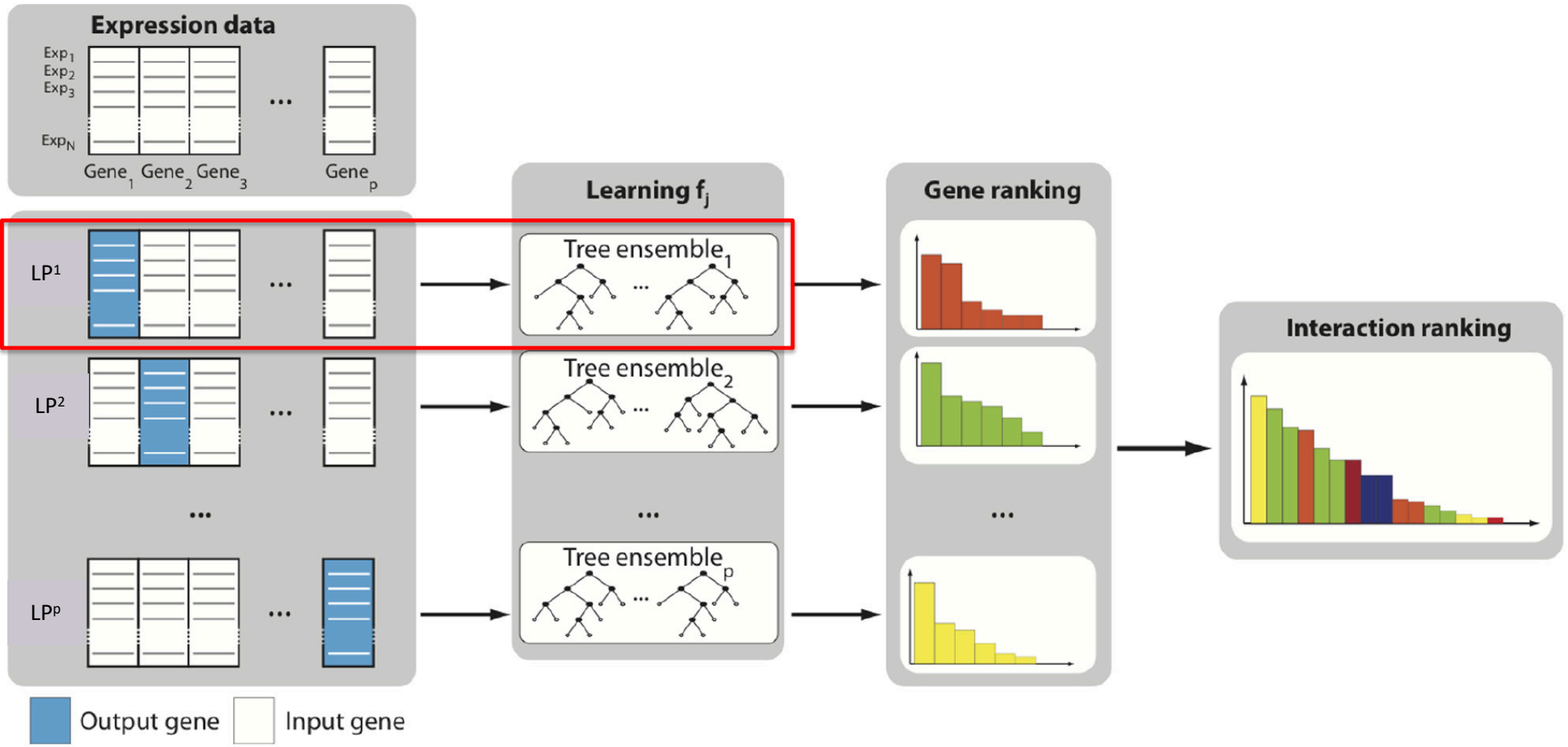


Using database to predict upstream transcriptional regulators

GENIE3-- GEne Network Inference with Ensemble of trees

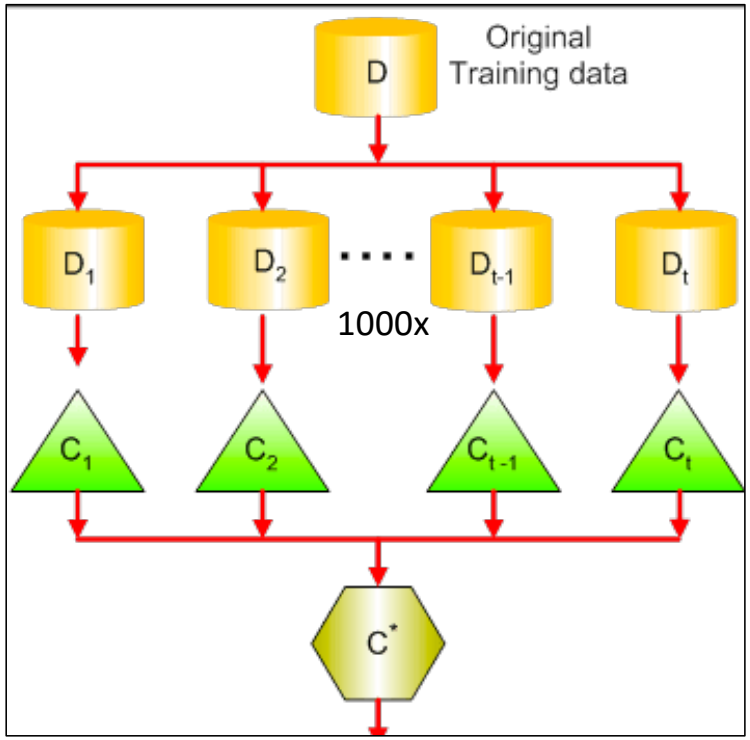
- Treat one gene a time as the target for regression.
- **Using bootstrapping/bagging to generate learning samples.**
- **Within each learning sample (LS), grow many regression trees, each provides an estimate of feature weight.**
- Average weights of each feature within the ensemble of trees to have a combined rank of feature importance of that gene.
- Combine all the feature-target interactions together. To have a rank of all gene—gene interaction throughout all genes in the list.

Bootstrapping/ ensemble of trees



Building many small trees by bootstrapping/Bagging

- Question: how important is each other gene to the expression of target gene j
- From training data D, Derive multiple data sets/Learning sample (LS)
- From each data sets, build a a regression tree. To determine the importance of each other gene in determining gene j
- Combine all the feature importance to have a final one.



Bootstrapping/Bagging

Target gene

other genes

	Cst3	Ubb	Dbi	Malat1	Custom_(	Ppia	Fabp7	Aldoc	Rpl41	Actb	Tmsb4x	[illegible]	Ck[illegible]	Mt1	H[illegible]gn2	Eef1a1	Apoe	Slc1a3	Tspan7	Gstm1	Rn45s	Tuba1a	Cox7c	Fth1
C1	[illegible]	2837.97	12868.7	6472.14	33810.79	11533.01	15985.54	6012.69	15835.42	2564.16	812.17	15324.53	5591.03	2196.97	3271.17	2154.78	3454.3	1304.16	2458.26	6484.96	5812.14	124.6	9864.46	861
C2	[illegible]	6552.88	16688.47	3155.32	26170.54	9928.4	3058.52	4836.24	3215.21	2333.16	24.59	4523.57	18692.86	6497.76	3454.11	3064.79	3759.06	3422.24	5195.19	4184.75	2842.53	29.05	4654.4	2260
C3	[illegible]	5010.84	9920.7	5195.99	25021.09	7369.39	13253.52	118.62	79.13	3131.61	38.12	11239.78	8476.72	9211.29	3274.66	3586.28	4818.08	621.52	3606.3	10173.67	5722.82	1862.53	0	1132
C4	22518.94	10976.37	9269.34	10577.01	24678.19	8392.77	2289.83	6145.14	2117.77	4600.66	1258.42	5053.78	12701.38	6489.99	1794.51	3676.78	11929.93	4394.81	2673.44	5003.65	3614.04	3366.39	3227.89	1791
C5	[illegible]	18300.93	11012.57	7813.54	24664.25	9134.71	19037.38	15417.84	246.34	1785.77	1463.79	159.96	10998.27	2664.36	3211.03	3716.36	13208.09	5773.75	8841.08	6484.05	2821.56	992.69	633.44	2135
C6	45035.81	7503.7	15465.83	15742.18	24641.42	5876.89	2081.12	6127.73	3281.25	2311.31	11.09	7771.38	7110.98	5882.38	1202.19	1210.86	3441.37	8171.91	4565.17	3840.97	5687.78	128.69	0	155
C7	[illegible]	17524.64	5079.66	1553.14	23388.84	9246.24	2494.55	7229.04	83.76	2555.33	26.9	0	11273.35	10688.7	17.29	5043.56	7190.25	4884.42	5645.77	6752.9	2046.3	1443.9	1549.61	1457
C8	[illegible]	13294.21	15629.39	6255.72	23214.06	6377.92	9589.91	9727.12	7402.6	2194.46	419.92	3164.16	7883.68	6022.92	195.16	3218.22	5746.22	5754.6	7054.2	11326.48	2256.41	2280.83	1799.71	120
C9	[illegible]	11957.49	4054.95	9665.19	22324.25	6628.45	15734.99	18610.21	81.03	1403.86	182.19	0	9571.93	1396.93	1033.34	1679.16	6459.22	9157.3	8784.53	4288	3014.09	2068.73	0	16
C10	28773.24	7807.45	14077.06	18425.99	21404.47	4076.15	6706.01	20192.97	5111.03	1892.57	32.19	227.76	4856.78	4024.37	119.52	3164.64	6286.29	3667.08	13287.7	5719.8	4003.37	45.81	0	4140
C11	59296.6	11949.45	9850.14	[illegible]	[illegible]	4154.15	9846.89	10098.65	0	4282.25	1773.52	0	1944.25	1539.12	861.68	1181.9	2902.33	7176.5	2878.95	[illegible]	1347.52	1053.03	254.71	5767
C12	27066.48	16386.59	3933.27	[illegible]	[illegible]	13004.58	2766.45	3625.74	56.63	3310.38	[illegible]	0	4487.53	402.55	2205.37	5331.03	7636	8843.22	5563.64	[illegible]	[illegible]	4312.5	169.88	3362
C13	58581.6	17473.04	21105.9	760.5	19240.64	14206.8	28989.2	6207.65	121.27	3797.25	29.21	0	4652.62	22621.97	821.94	5796.2	4897.41	5067.96	7508.47	12310.76	5385.59	2877.31	3880.52	116
C14	5885.47	26852.53	14423.57	896.2	19126.04	9406.04	14444.33	4939.8	432.48	3738.73	1757.23	0	6909.67	3075.26	5462.22	107.31	85.62	4861.73	13.34	1463.49	1944.83	4283.63	129.74	1761
C15	7697.19	20252.39	5966.61	[illegible]	19081.58	10698.72	1528.02	7303.56	101.25	3895.1	2968.08	0	6234.44	928.49	8996.8	8059.14	3598.73	1623.47	3902.69	655.61	2230.83	3472.68	0	3904
C16	63423.02	4939.71	11587.48	4647.81	18064.62	5407.58	453.2	13627.09	163.91	2154.97	1079.26	0	5863.84	2039.13	1679.57	20	8147.01	6295.86	8238.75	2861.48	2235.23	1570.56	229.47	4506
C17	37623.68	8154.25	14200.53	17441.93	17870.31	10431.63	23545.8	4755.24	0	3630.82	3726.88	0	4578.13	2784.5	1974.59	280.02	4905.36	2960.4	3751.65	702.2	6523.99	6095.61	712.16	41
C18	14991.98	19156.08	20259.12	897.1	16158.44	6337.53	26127.42	10284.31	5606.18	2382.28	10.72	0	2462.27	13237.35	2768.26	6047.12	4150.64	3041.51	3031.86	5925.61	2324.71	1754.76	4872.04	2529
C19	36119.48	17359.78	9660.94	11935.03	16108.91	11435.47	1527.87	18275.44	10522.98	3056.47	23.47	0	8784.84	5137.45	4205.51	34.93	2313	2907.4	5622.8	3812.92	1457.8	2883.32	1644.22	2726
C20	22992.38	6511.3	12661.18	6849.2	16079.16	4714.21	2352.97	3133.54	2299.98	2098.69	3206.18	5645.4	2114.37	7913.66	346.05	7296.65	2681.19	4031.66	[illegible]	[illegible]	5771.25	859.92	4139.96	2667
C21	16006.61	26377.88	5224.08	7534.8	15632.02	11502.82	14505.76	14648.69	112.22	5178.27	1396.75	255.05	9384.81	179.5	1506.64	375.7	4166.95	6570.09	[illegible]	[illegible]	3528.25	10922.26	617.21	366
C22	48674.1	9903.64	11213.86	14894.99	15219	9156.72	7092.46	8311.5	0	6559.12	39.87	12573.04	6743.61	5458.02	36.41	1389.21	6483.98	8833.23	5098.73	5663.7	5539.01	6130.76	0	1973
C23	22133.84	7935.07	11468.22	12168.6	15078.11	6143.38	4138.59	587.87	0	3720.28	2720.53	9437.65	4004.74	9395.01	75.58	1610.1	1854.2	8120.99	4339.07	7106.98	6994.01	1989.87	4943.53	3378
C24	17477.16	7885.82	7263.56	14068.43	14758.02	5905.18	1290.67	5956.3	0	2355.03	2068.33	16292.5	6857.31	13291.3	585.08	361.65	3234.78	4153.96	926.63	2253.4	5652.71	6133.83	31.03	3750
C25	72284.97	12102.17	4635.61	18638.25	14671.26	[illegible]	[illegible]	3774.57	0	3426.27	1744.99	0	7204.42	3189.71	[illegible]	26.95	1221.59	16613.68	8117.34	8344.37	4950.34	1858.22	6159.55	35
C26	16586.81	27295.85	7285.31	5918.6	14386.26	[illegible]	[illegible]	8223.5	0	2593.93	18.05	0	6426.18	[illegible]	[illegible]	4700.35	3605.9	4285.93	4351.81	4562.32	1670.05	2238.96	299.65	1578
C27	87.24	49057.03	6585.44	861.27	14338.6	[illegible]	8638.24	31.59	3032.34	4202.53	15.22	0	462.32	143.15	3933.56	3669.47	62.79	158.37	57.12	26.89	5556.5	3222.44	14451.02	15
C28	62085.93	15165.13	8871.78	13409.7	14108.55	4336.75	8448.68	3841.31	37.34	2343.19	395.8	169.73	7624.54	942.37	25.69	3426.99	2669.76	7219.47	5708.88	6583.29	3490.5	611.16	0	2565
C29	35420.6	22968.24	8470.86	17743.87	13944.7	6124.75	1028.09	10793.26	3585.06	1798.61	24.5	2773.74	6581.18	5070.13	108.46	2480.67	2483.42	4946.95	6794.18	5077.71	1926.72	1042.93	1906.95	2054
C30	3783.66	37109.78	15888.6	1308.36	13714.35	13128.89	7386.6	18323.63	4365.59	6459.07	10.58	9659.69	12422.87	6322.46	6210.84	7515.46	249.28	824.68	1577.94	12644.65	2631.18	3275.34	2026.29	10479
C31	30558.53	12173.08	13754.65	17744.0	13620.63	7650.73	5063.56	10401.22	7842.53	1830.7	43.94	0	4245.47	7584.95	4675.48	1615.78	12734.93	6238.07	5563.88	4134.05	3377.11	7875.11	3921.27	818
C32	11461.73	6633.12	3607.31	11347.65	13533.05	10151.59	31.21	5997.45	225.76	2084.78	1985.04	718.33	9031.11	1412.35	4527.81	5020.12	2841.33	4787.42	4512.45	842.57	2149.41	2072.1	632.13	3579
C33	34223.25	12952.5	11797.97	25484.3	13155.97	1026.11	2576.93	6079.7	0	4113.37	7.84	999.17	7815.9	3351.07	103.07	41.98	3273.68	2849.23	11026.25	6417.36	1403.61	2413.31	3126.28	28
C34	38934.12	10048.69	9551.77	14269.56	12956.65	5336.78	80.78	4318.62	2654.36	2086.82	32.17	4894.4	7071.33	4002.29	43.68	38.75	4827.64	6994.37	4098.74	5486.5	4282.69	1770.33	3856.33	2206
C35	16331.95	9416.24	3448.03	18729.2	12904.53	15833.47	3251.26	9681.75	0	4382.14	0	0	3802.04	1224.51	1190.64	4840.31	6224.22	4175.47	3306.31	3139.48	2322	1546.68	224.67	1966
C36	5536.46	8368.96	2233.67	4478.9	12720.47	8313.25	24.86	27.71	5275.43	5036.1	1791.07	5286.31	2158.07	1406.36	19275	5554.84	258	968.79	14.79	25.17	2662.81	1371.04	2701.5	19

samples

LS 1

LS 2

LS 3

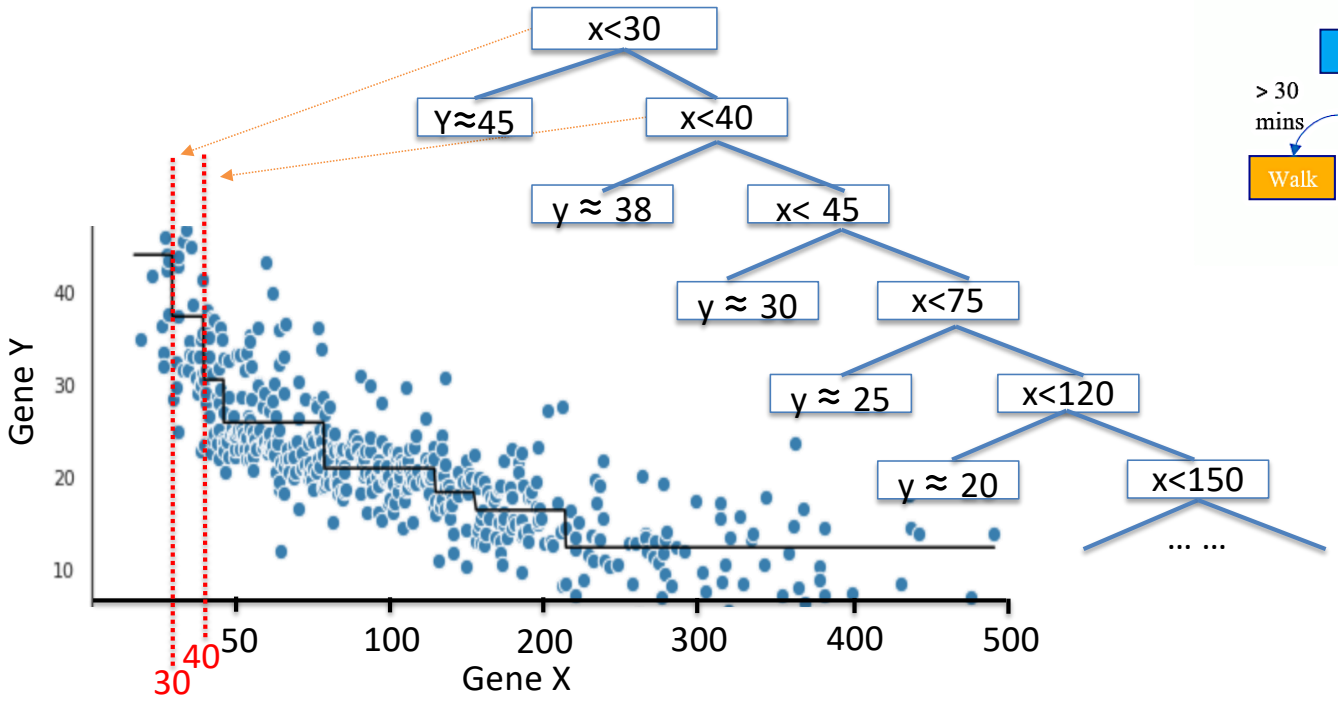
LS 4

LS 5

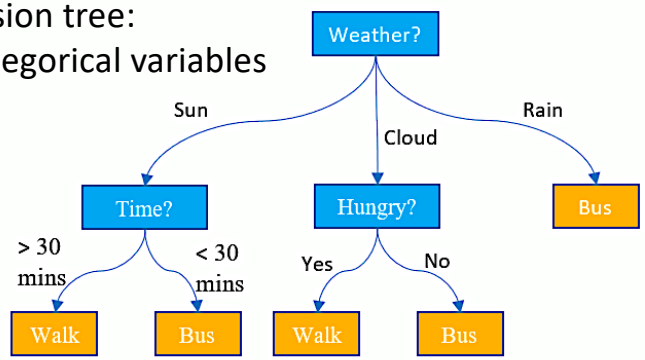
LS 6

Within each learning sample, grow a regression tree

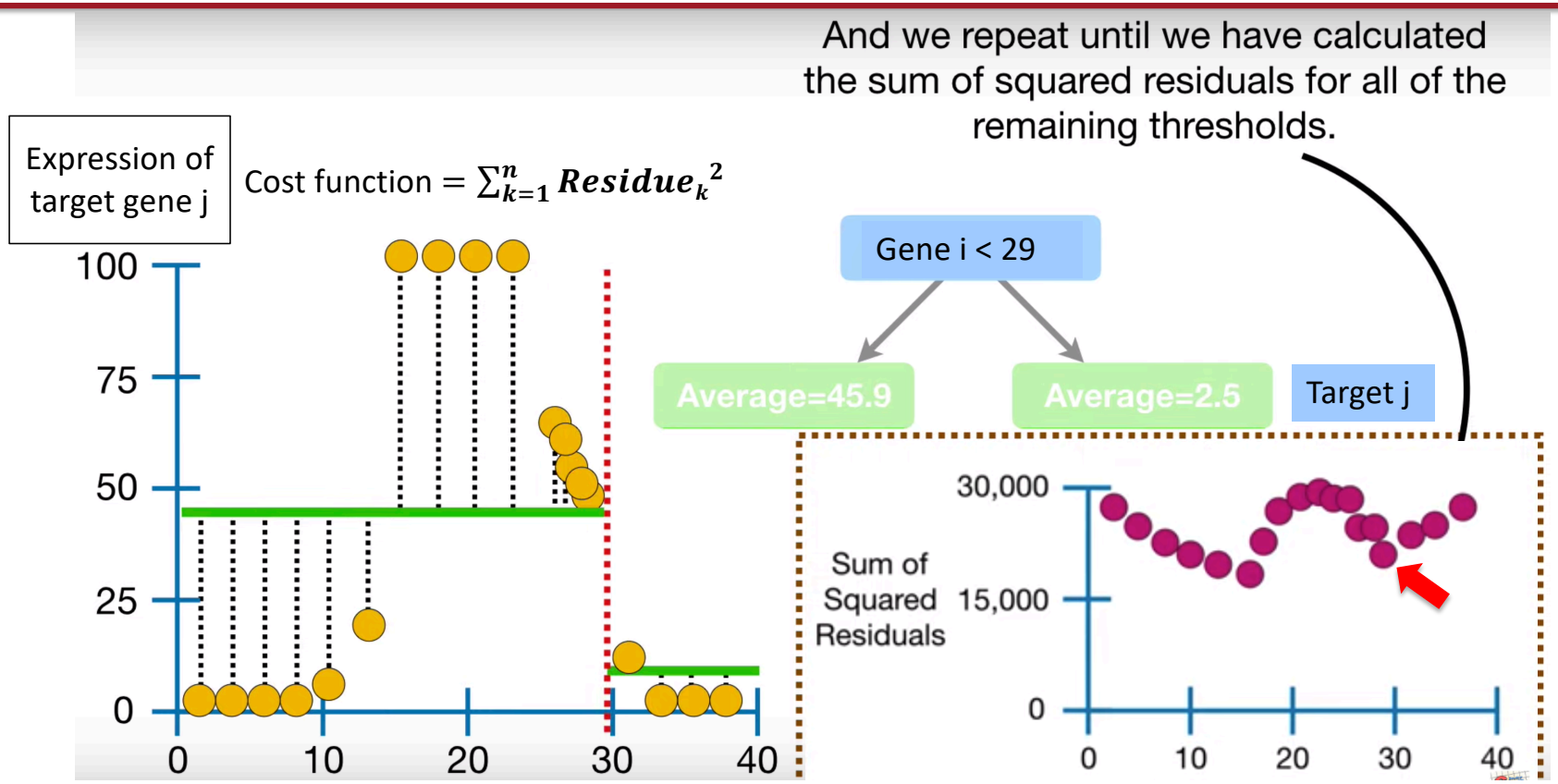
A regression tree:
For continuous variables



A decision tree:
For categorical variables

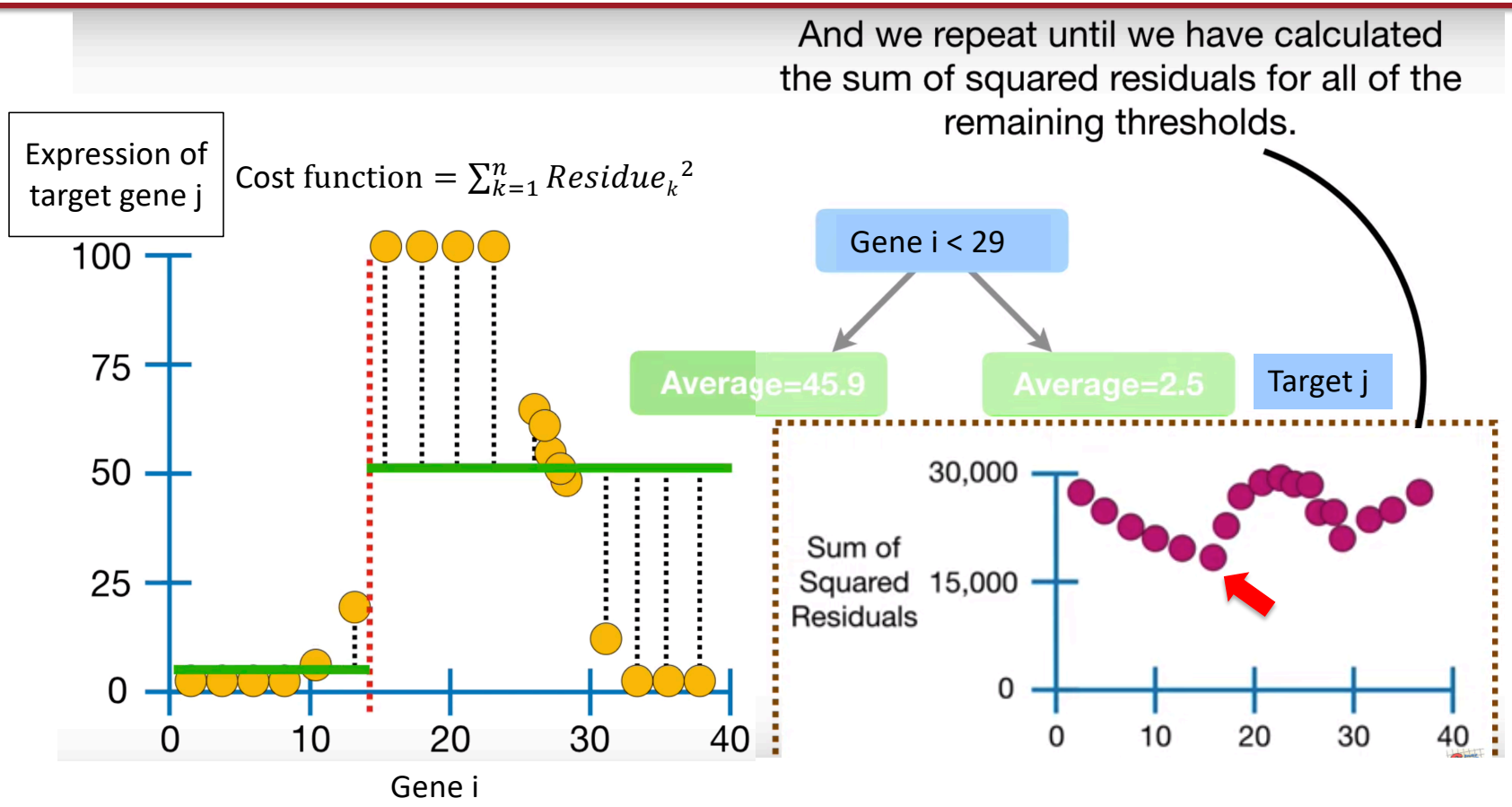


How a regression tree works



https://www.youtube.com/watch?v=g9c66TUylZ4

How a regression tree works



https://www.youtube.com/watch?v=g9c66TUylZ4

Build a tree from multiple variables

Grow a regression tree

1. For a given target gene j
2. For each of the feature gene i, determine the cost values of each possible split. Chose the lowest cost value as a potential split point.
3. ***Iterate through each of the rest features, get the lowest cost value.***
4. ***Choose the gene that provide the lowest cost to make a first split.***

5. For **each of the branch**, do the same 2→4 and iterate, until all the branch reaches the minimal number of samples.

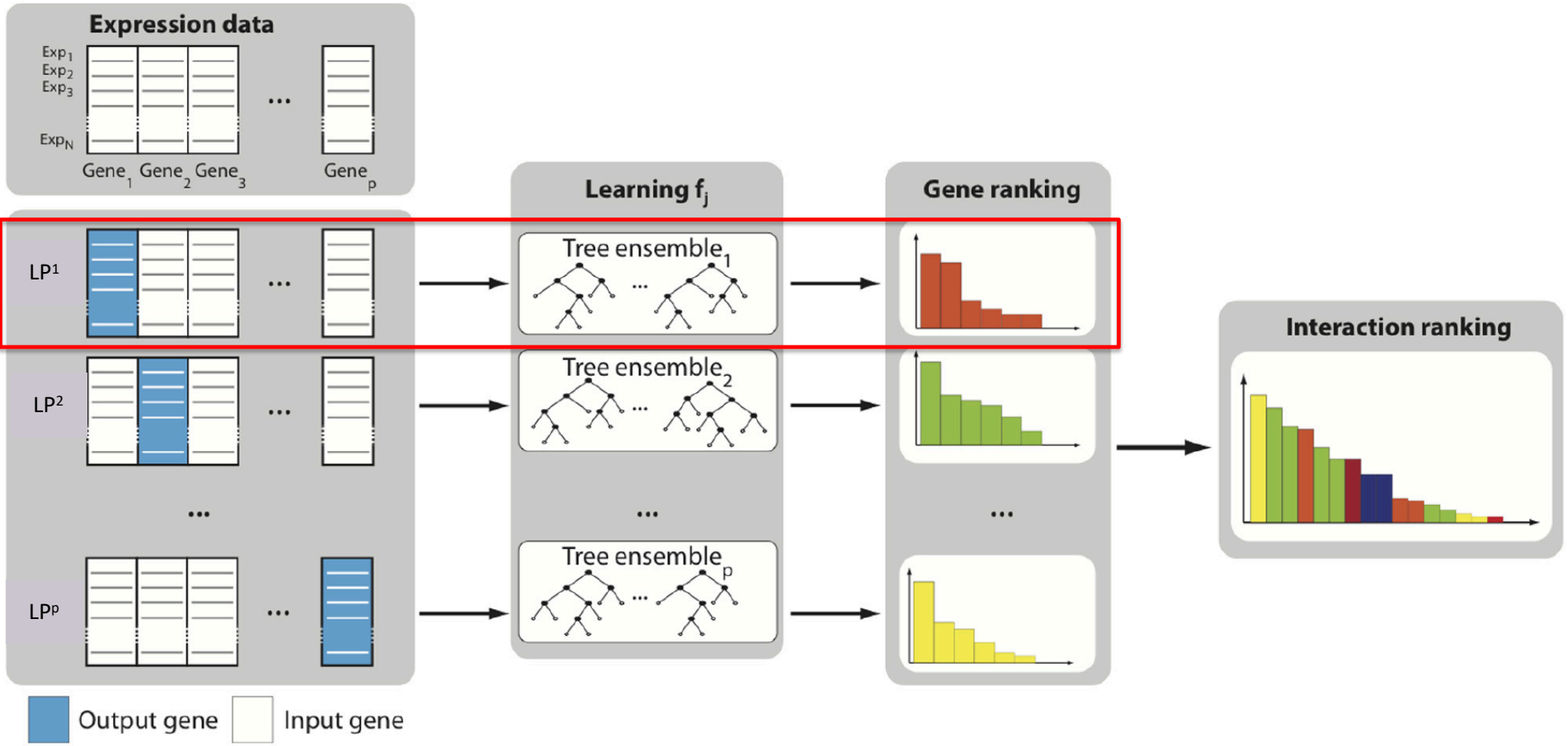
Determining feature importance

1. In a regression tree, **sum up the reduction of variances due to splitting** by each specific gene in all the branches. That is the score for the feature importance of that gene.
2. Average feature importance for each feature gene. Rank the feature genes by the scaled importance.

Combine feature importance for all the target genes as a rank of all interactions.

1. For each of the target gene, repeat through all the steps above, until all the feature importance for all target genes are determined.
2. Combine all interactions and rank them by the scores.

Bootstrapping/ ensemble of trees



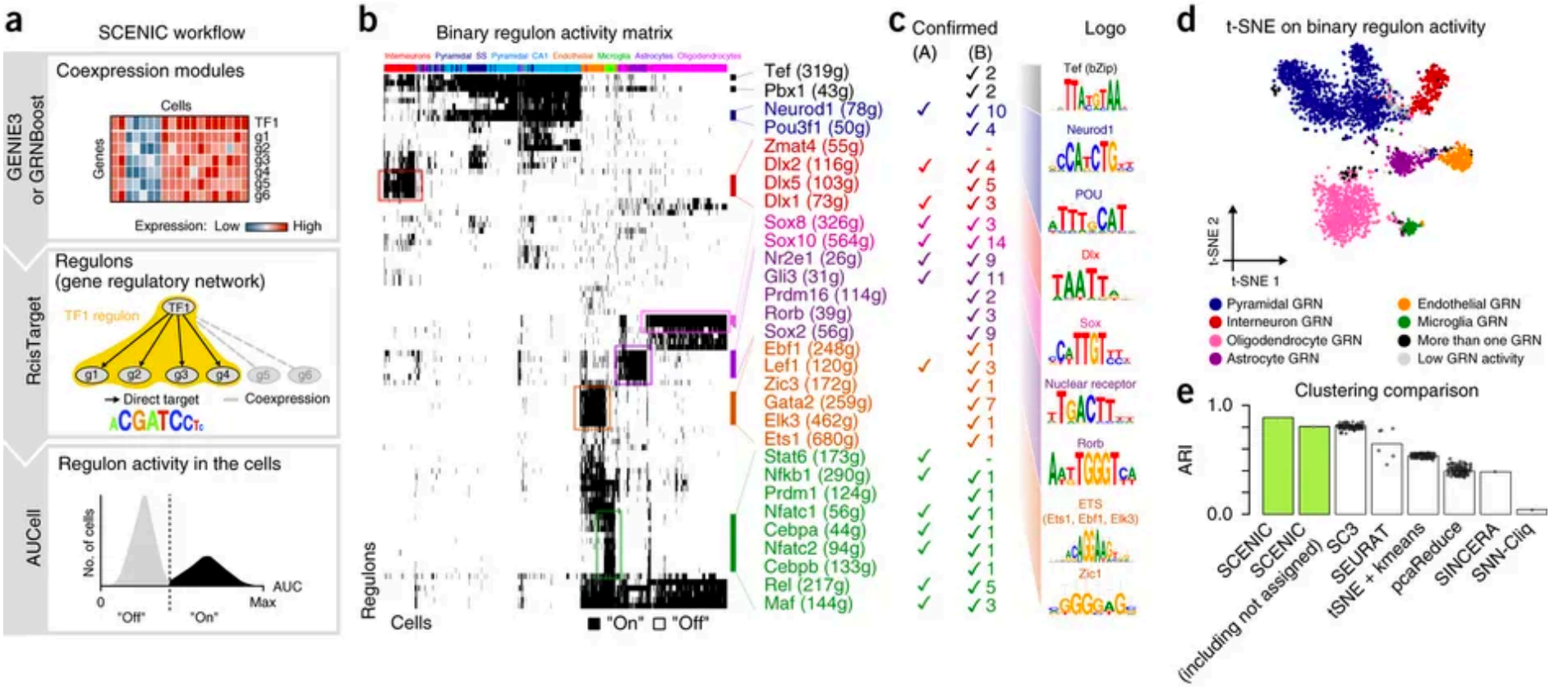
Discussion on Genie3

- Efficiently handle large dataset.
  - Parallel processing
  - Implemented in R and Python (even faster)
- Has been joined with cis-Target (**iRegulon**) to identify TF-target regulation in single cell data.

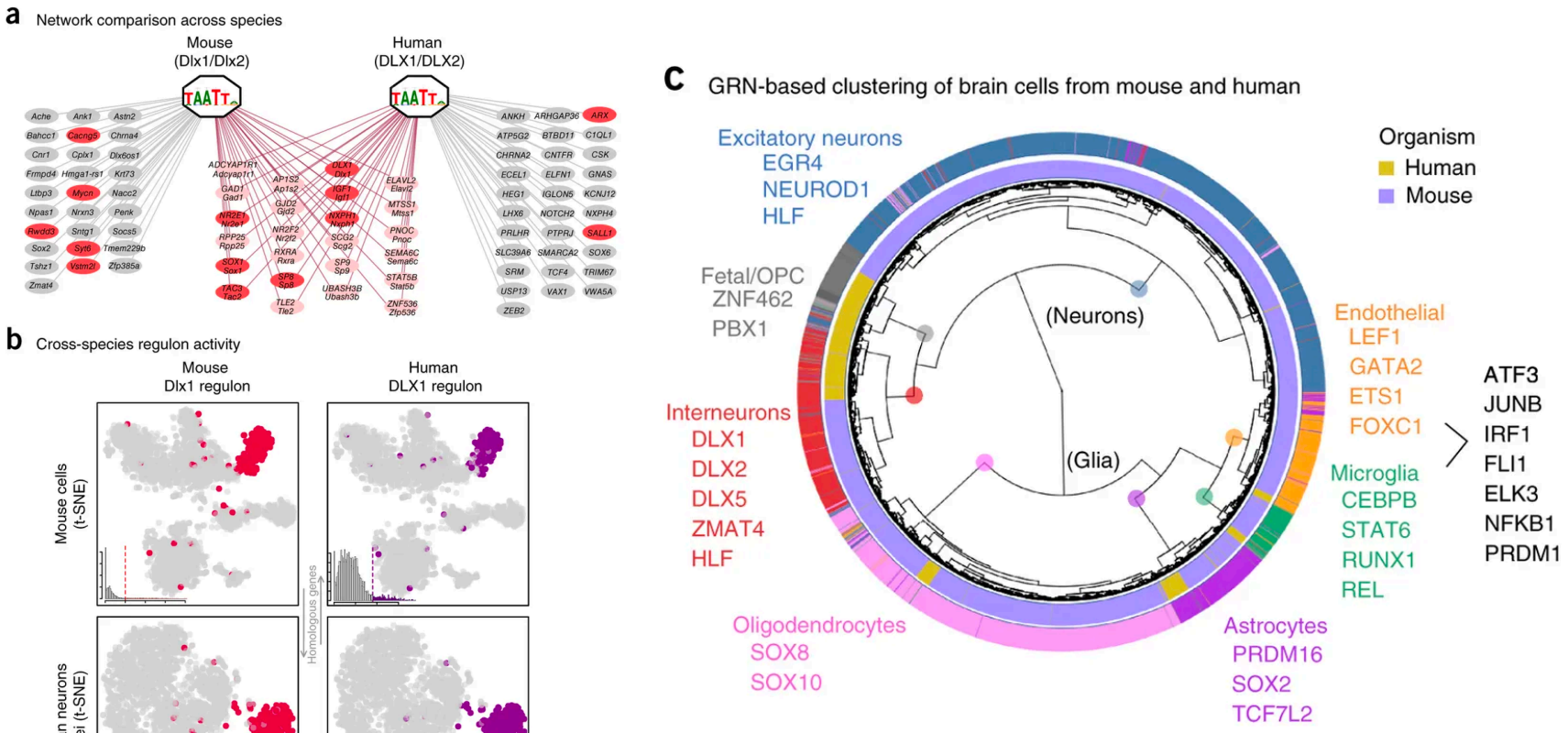
SCENIC

- Applied to single cell data.
- Integrate GENIE3 co-expression module detection with iRegulon-detection of TF regulation.

SCENIC workflow and results



What do you achieve?



Running SCENIC

- Run the installation may spent 2 hrs.
- Convert files to loom files.
- Download the cis-target file may take a long time. And how to set up these files before hand?
  - 1.1 G for mm9 mouse. Take 30 minutes to download.
- A dry run before your own data.

A side note: Object Oriented Programing in R

- R is a procedure oriented programing language.
  - A <- 4
  - B <- A+5
  - General/defined functions: B<- function(A)
- Many of the genomic R packages /Bioconductor packages uses OOP concepts in R.
  - S3, S4, R5, R6 objects in R are object oriented, to avoid to have too many environment variables.
  - Each has slots identified by ”@” operator, and each slot can have “$” fields.
  - Stored values are extracted by custom defined methods
  - Values can be mutated by applied methods and return back to its slot
  - If A is an OOP object, Function (A) will change some defined variable in certain slot in A.

Object-oriented programming (OOP)– internal structure of an S4 object: **scenicOptions**

@inputDatabaseInfo

getDatabaseInfo(scenicOptions, "org")

```
$org
[1] "mgi"
$datasetTitle
[1] ""
$cellInfo
[1] "int/cellInfo.Rds"
$colVars
[1] "int/colVars.Rds" NA
$int_01
[1] "int/cellColorNgenes.Rds" NA
```

OOP objects in R
- Commonly used in genomic analysis
- Classified to S3,S4, R5, R6 objects
- Customized names for class
- Structure of S4 objects
  - @ slots
  - $ separate fields within a slot
  - Manually defined Methods that operation on the object

@settings

getSettings(scenicOptions, "nCores")
getDatabases(scenicOptions)
getDbAnnotation(scenicOptions)

```
$dbs
10kb "mm9-tss-centered-10kb-7species.mc9nr.feather"
$dbDir
[1] "./cisTarget_databases"
$db_mcVersion
[1] "v9"
$verbose
[1] TRUE
$nCores
[1] 10
$seed
[1] 123
$devType
[1] "pdf"
$modules
$modules$weightThreshold
[1] 0.001
$regulons
list()
$aucell
$aucell$smallestPopPercent
[1] 0.25
$defaultTsne
$defaultTsne$dims
[1] 50
$defaultTsne$perpl
[1] 50
$defaultTsne$aucType
[1] "AUC"
$tSNE_filePrefix
[1] "int/tSNE"
```

@filenames

getIntName(scenicOptions, "genesKept")
loadInt(scenicOptions, "genie3ll", ifNotExits="null")

$int

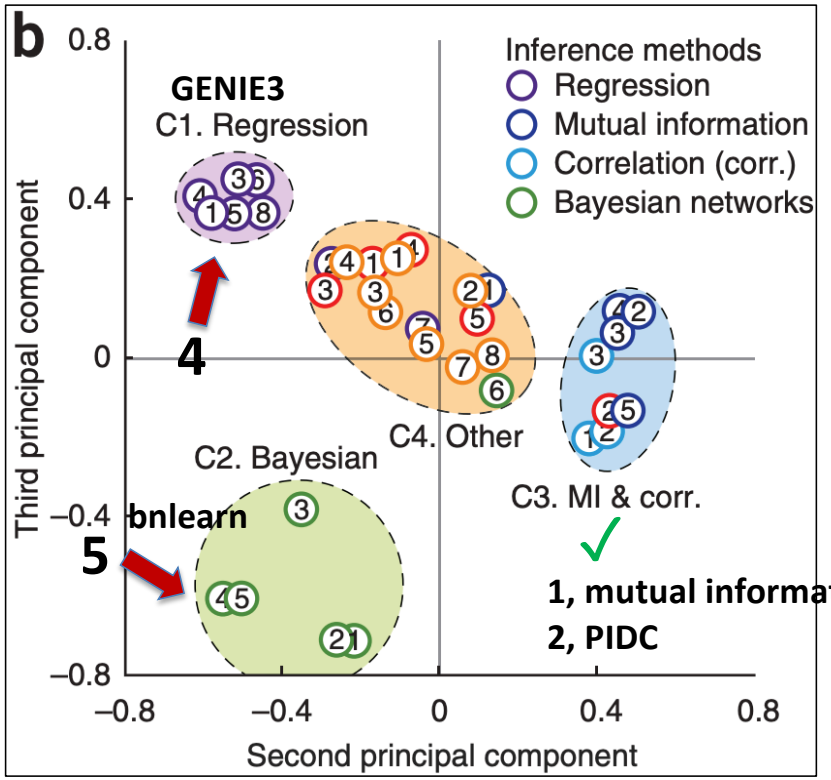
	fileName
genesKept	"int/1.1_genesKept.Rds"
corrMat	"int/1.2_corrMat.Rds"
genie3wm	"int/1.3_GENIE3_weightMatrix.Rds"
genie3ll	"int/1.4_GENIE3_linkList.Rds"
genie3weighPlot	"int/1.5_weightPlot"
tfModules_asDF	"int/1.6_tfModules_asDF.Rds"
tfModules_forEnrichment	"int/2.1_tfModules_forMotifEnrichmet.Rds"
motifs_AUC	"int/2.2_motifs_AUC.Rds"
motifEnrichment_full	"int/2.3_motifEnrichment.Rds"
motifEnrichment_selfMotifs_wGenes	"int/2.4_motifEnrichment_selfMotifs_wGenes.Rds"
regulonTargetsInfo	"int/2.5_regulonTargetsInfo.Rds"
regulons	"int/2.6_regulons_asGeneSet.Rds"
regulons_incidMat	"int/2.6_regulons_asIncidMat.Rds"
aucell_regulons	"int/3.1_regulons_forAUCell.Rds"
aucell_genesStatsPlot	"int/3.2_aucellGenesStats"
aucell_rankings	"int/3.3_aucellRankings.Rds"
aucell_regulonAUC	"int/3.4_regulonAUC.Rds"
aucell_thresholds	"int/3.5_AUCellThresholds.Rds"
aucell_thresholdsTxt	"int/3.5_AUCellThresholds_Info.tsv"
aucell_binary_full	"int/4.1_binaryRegulonActivity.Rds"
aucell_binary_nonDupl	"int/4.2_binaryRegulonActivity_nonDupl.Rds"
aucell_regulonSelection	"int/4.3_regulonSelections.Rds"
aucell_binaryRegulonOrder	"int/4.4_binaryRegulonOrder.Rds"

getOutName(scenicOptions, "s2_motifEnrichment")

$output

	fileName
s2_motifEnrichment	"output/Step2_MotifEnrichment.tsv"
s2_motifEnrichmentHtml	"output/Step2_MotifEnrichment_preview.html"
s2_regulonTargetsInfo	"output/Step2_regulonTargetsInfo.tsv"
s3_AUCheatmap	"output/Step3_RegulonActivity_heatmap"
s3_AUCtSNE_colAct	"output/Step3_RegulonActivity_tSNE_colByActivity"
s3_AUCtSNE_colProps	"output/Step3_RegulonActivity_tSNE_colByCellProps"
s4_boxplotBinaryActivity	"output/Step4_BoxplotActiveCellsRegulon"
s4_binaryActivityHeatmap	"output/Step4_BinaryRegulonActivity_Heatmap_"
s4_binarytSNE_colAct	"output/Step4_BinaryRegulonActivity_tSNE_colByActivity"
s4_binarytSNE_colProps	"output/Step4_BinaryRegulonActivity_tSNE_colByCellProps"
loomFile	"output/scenic.loom"

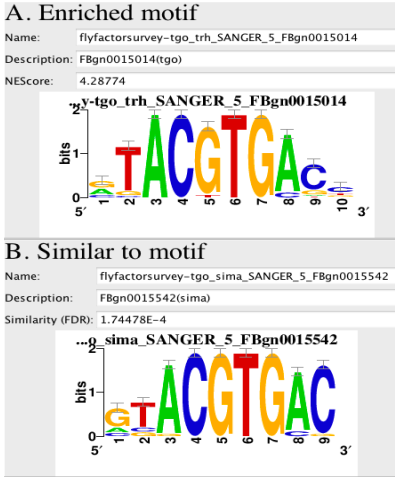
Method 5, Bayesian Network



Based on statistical relationship between genes

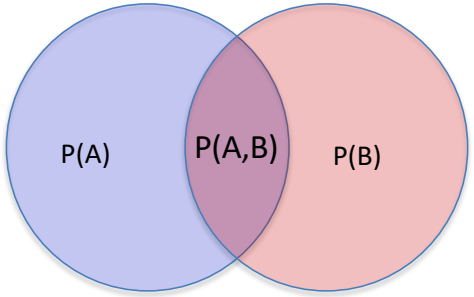
3

iRegulon



Using database to predict upstream transcriptional regulators

Method 5:
Bayesian Network – conditional dependence



P(A,B)=P(A|B)*P(B)=P(B|A)*P(A)

Bayes rule:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

Bayesian network – the conditional probability

If A and B are independent
A= winter (yes 25%, no 75%)
B= happy (yes 50%, no 50%)

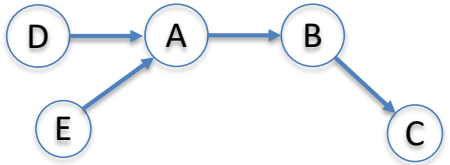
P(A) = 25%
P(B) = 50%
P(winter, happy) =
P(A,B)=25%*50%=12.5%
Correct

How about if B dependent on A:
A= winter (yes 25% or no 75%)
B= cold weather (overall yes 25%, no 75%)

P(A)=25%
P(B|A)=100% on winter and 0% on other seasons.
Overall 25%
The conditional probability of winter & cold is
P(A=winter, B=cold)=P(A)P(B)=25%*25%=6.25%
wrong!

P(winter, cold) =
P(A=winter;B=cold)=P(A)**P(B|A)**=25%*100%=25%

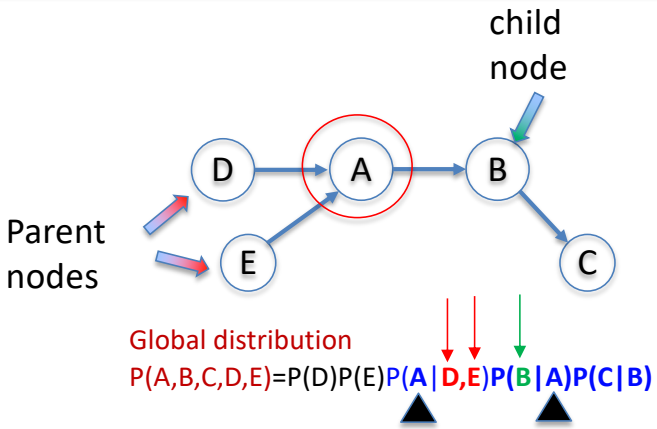
A= winter (yes or no 50% each)
B= cold weather (yes or no 100%)
C= wear sweater (yes or no, 50%)
D= month of a year
E= location on the earth



Global distribution
P(A,B,C,D,E)=P(D)P(E)P(A|D,E)**P(B|A)P(C|B)**

Parent nodes and child node

- To a given node
  - Conditional nodes (reasons or contributing factors) are called parents.
  - The consequent nodes are child nodes.
- 2 components in Bayes Net,
  - 1. Graph structure: edges (arrows, can be defined by a list of nodes and parents of each node).
  - 2. Parameters (θ): P(D), P(E), P(A|D,E), P(B|A), P(C|B)



Learn the network structure and parameters

Divide the Bayes net into two components:

- Θ: the conditional probabilities
- G: the structure
  - Can be decomposed to each node with parents (X_i and Π_{xi}).

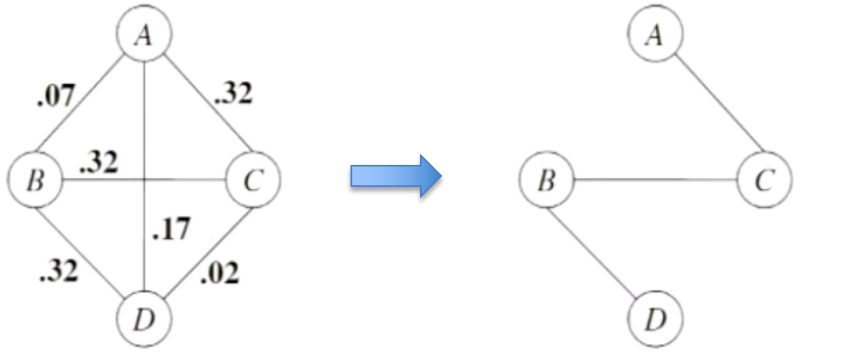
The problem | Step 1 | Step 2

$$\underbrace{P(\mathcal{M} \mid \mathcal{D}) = P(\mathcal{G}, \Theta \mid \mathcal{D})}_{\text{learning}} = \underbrace{P(\mathcal{G} \mid \mathcal{D})}_{\text{structure learning}} \cdot \underbrace{P(\Theta \mid \mathcal{G}, \mathcal{D})}_{\text{parameter learning}}$$

Probability of a model (M) given the data set (D).

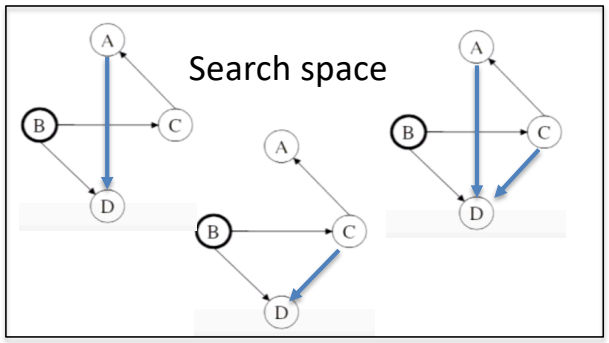
Learning the skeleton of a graph

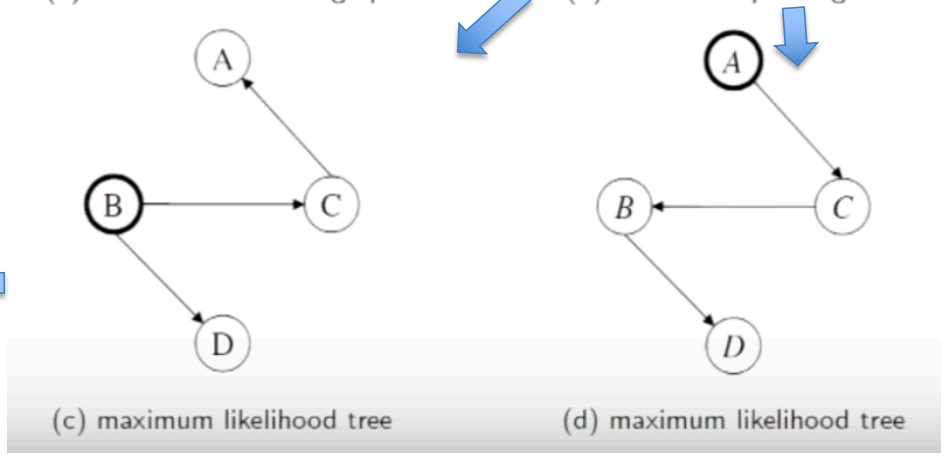
- Constructing Maximum Spanning Tree.
- Add direction through starting from a random node.



BIC score = Bayesian Information Criterion

$$\text{Score}(G|\mathcal{D}) \stackrel{def}{=} \text{LL}(G|\mathcal{D}) - \psi(N) \cdot ||G||$$





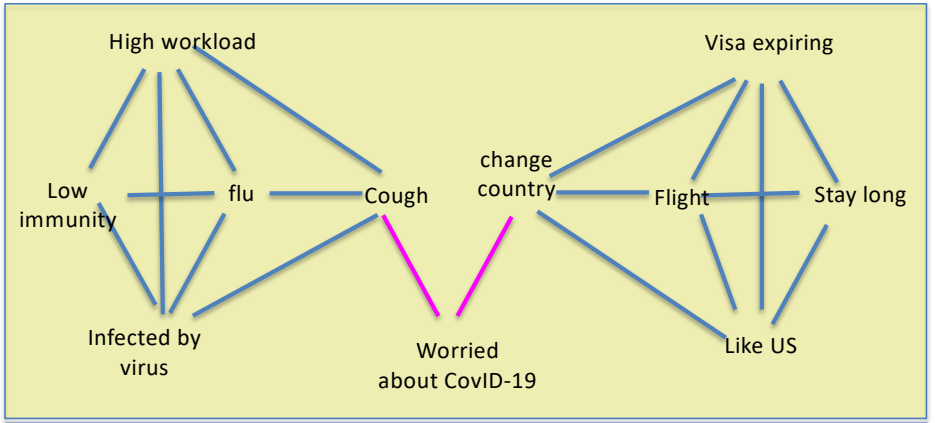
Assign directions: Induction of Causality algorithm

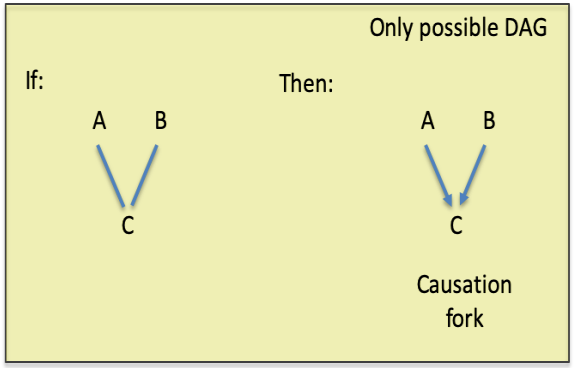
The Induction of Causality Algorithm:
If,

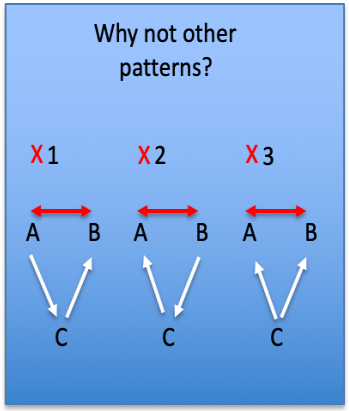
1. A, B are not connected through any other connection.
2. A and B have a common connection to C

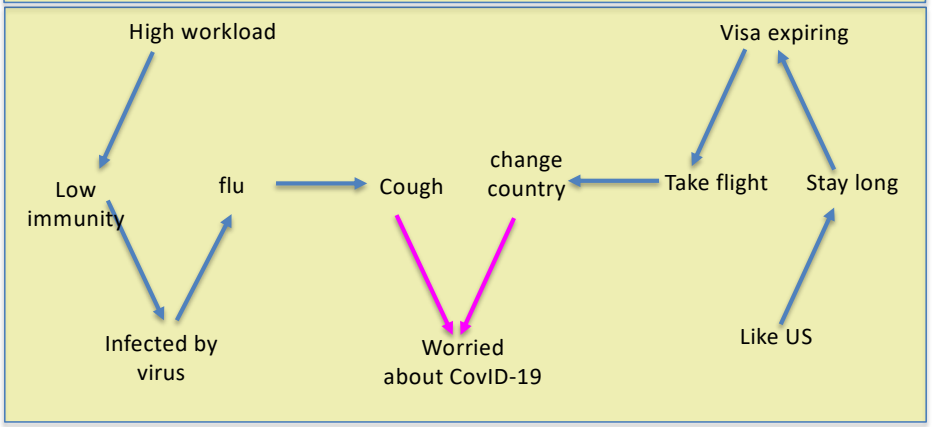
Then:

arrows point from A and B to C



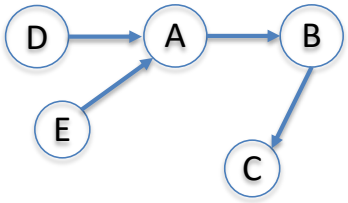






Curse of the dimensionality

- Complexity increases exponentially with respect to:
  - Number of bins each node can take.
  - Number of parent nodes that a node can have.
- Solutions
  - Breaking a huge network into conditionally independent units



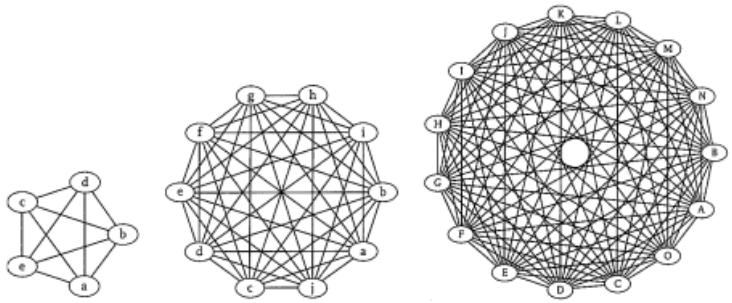
Global distribution

$$P(A,B,C,D,E) = P(D)P(E)P(A|D,E)P(B|A)P(C|B)$$

If each sample has 5 bins

Values to fully enumerate the distribution table: 5^5=1875

Enumerate the Bayesian network: 5 +5 +125 +25 +25=185



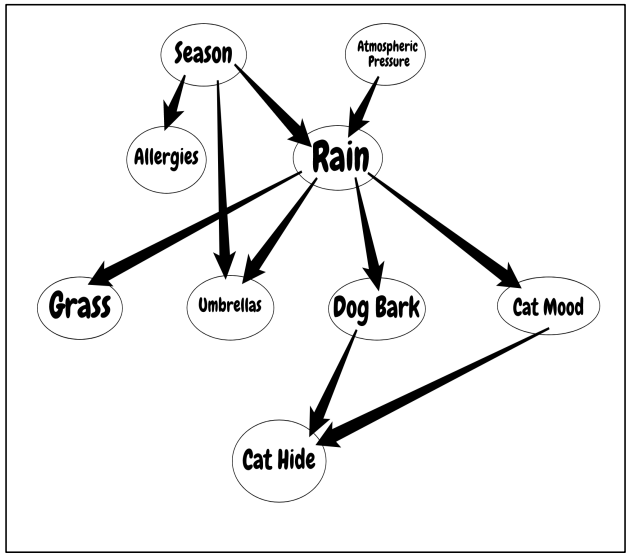
Discussion on Bayesian Network

Pros

- Easy interpretation

Cons

- Requires large computation
- Decomposition of Bayesian seems not efficient



$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

Demonstration on bnlearn in R

- bnlearn in R
  - Demonstrate the import of data
  - Data structure
  - Output structure
  - Demonstrate how the graph are calculated
  - Import in to Cytoscape for visualization

Thank you for your attention!

- Questions?
  - GitHub: https://github.com/niaid/Gene_Regulatory_Networks
  - Googledoc, https://tinyurl.com/zhu-GRN
  - My email: zhuy16@nih.gov
  - BCBB email: bioinformatics@niaid.nih.gov

The end

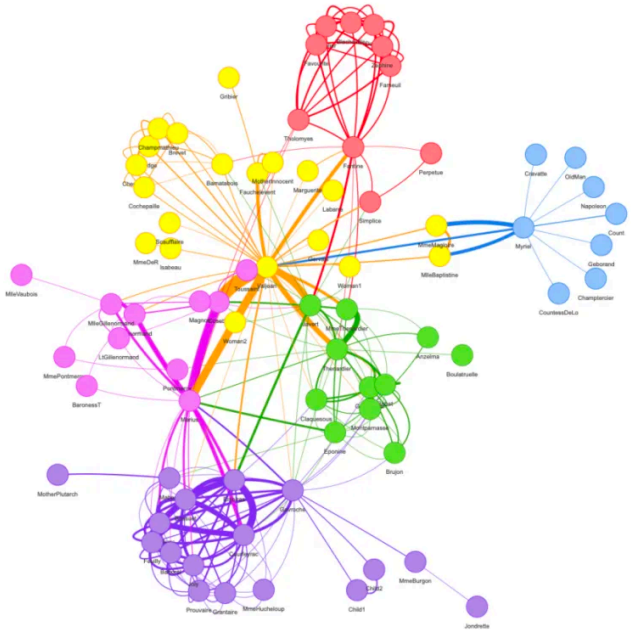
Other resources

- GraphPlot, http://juliagraphs.github.io/GraphPlot.jl/
- Gamma distribution, https://towardsdatascience.com/gamma-distribution-intuition-derivation-and-examples-55f407423840
- install.packages("igraph")
- install.packages("bnlearn")
- BiocManager::install("Rgraphviz")
- iGraph for network analysis https://www.bioss.ac.uk/people/helen/igraphIntro.html
- https://www.r-bloggers.com/interactive-network-visualization-with-r/

How to set up a tunel to server

- install.packages('IRkernel') # Don’t forget step 2/2!
- IRkernel::installspec()
- using Pkg
- Pkg.add("IJulia")
- user@local_machine$ ssh -N -L localhost:8888:localhost:8888 user@remote_mahcine
- Paste this to the url to access the Julia –”localhost:1234”

- > tpm=read.table("S_GSE71485_Single_TPM.txt",row.names=1,header=T)
- > Tpm=tpm[rowMeans(tpm)>100,]
- > write.table(Tpm,"S_GSE71485_Single_TPM.txt")
-



Parameters in Bayesian network

likelihood prior

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

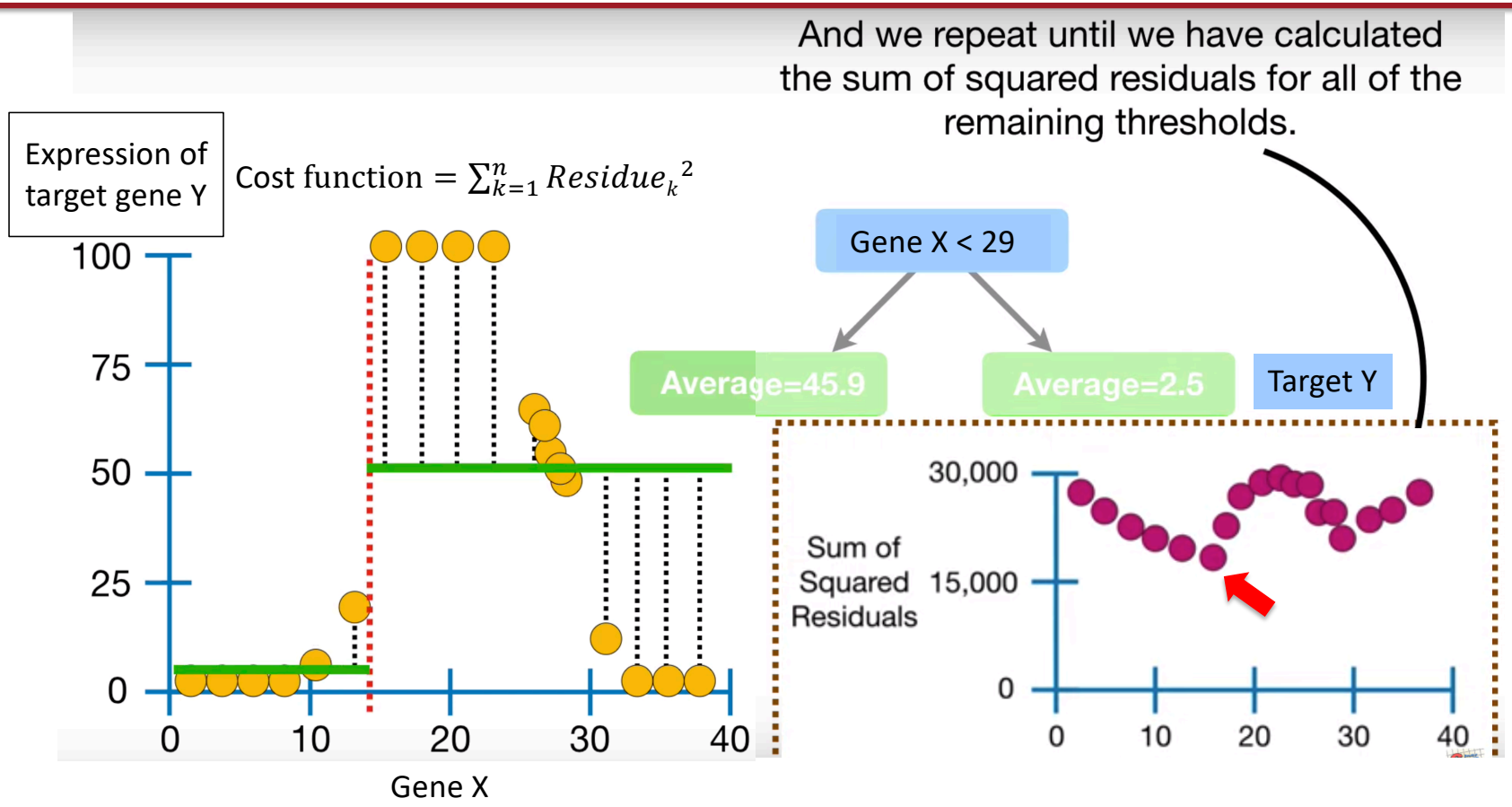
posterior Evidence

posterior likelihood prior

$$\mathbf{P(cold|winter)} = \frac{P(\boldsymbol{winter}|\boldsymbol{cold})P(cold)}{P(winter)} = \frac{P(\boldsymbol{winter}|\boldsymbol{cold})P(cold)}{P(winter)} = \frac{100\% * 25\%}{25\%} = 100\%$$

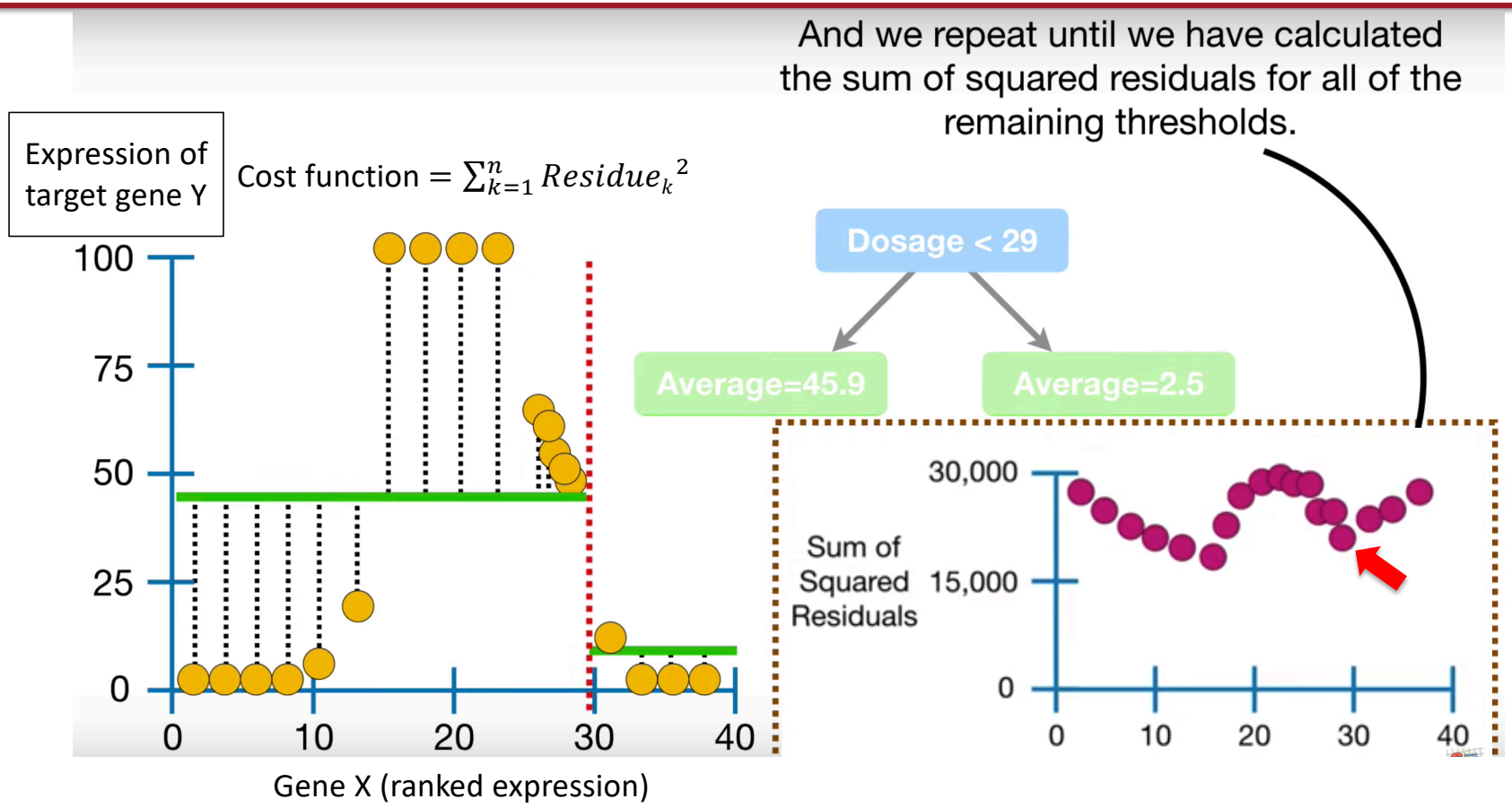
Evidence

How regression tree works



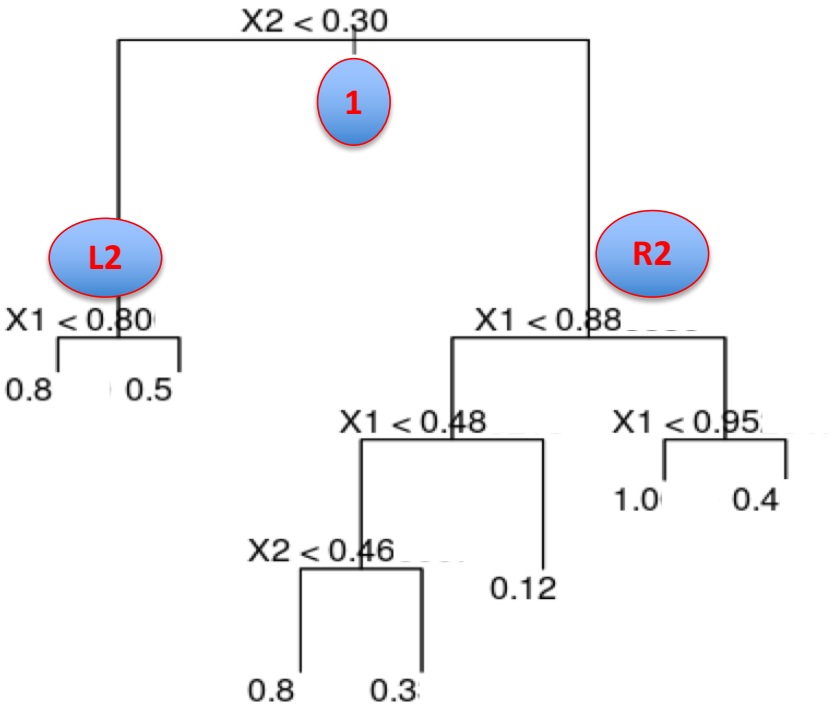
https://www.youtube.com/watch?v=g9c66TUylZ4

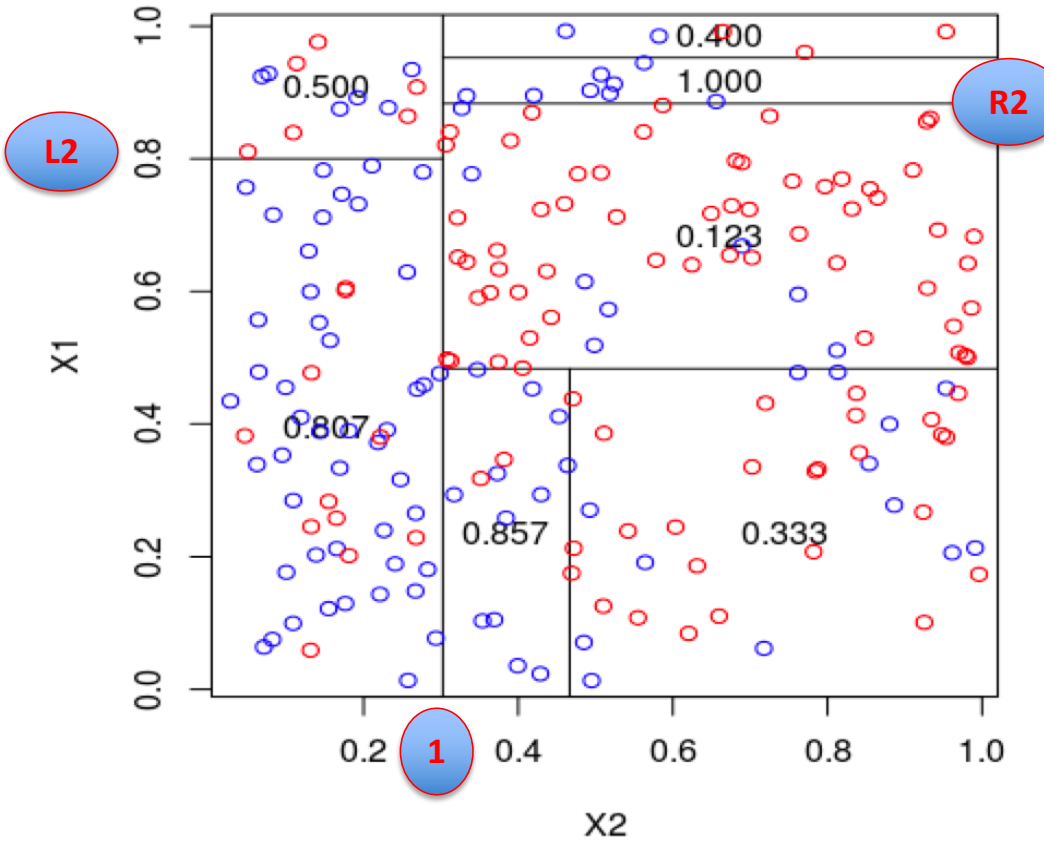
How regression tree works



https://www.youtube.com/watch?v=g9c66TUylZ4

A regression tree by two variables





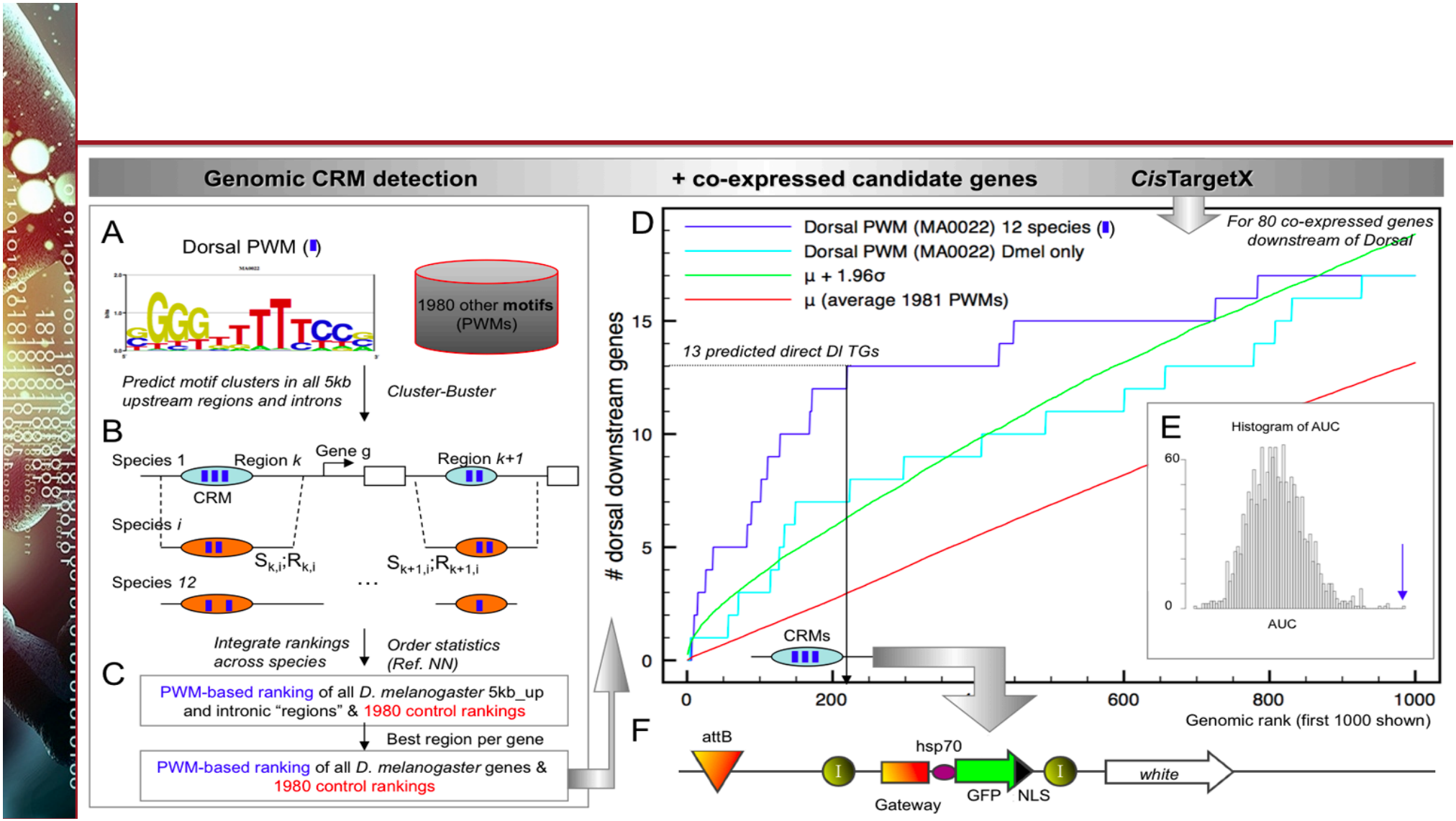
Target: The **% of o blue circles**
Variables: X1 and X2

Build a tree from multiple variables

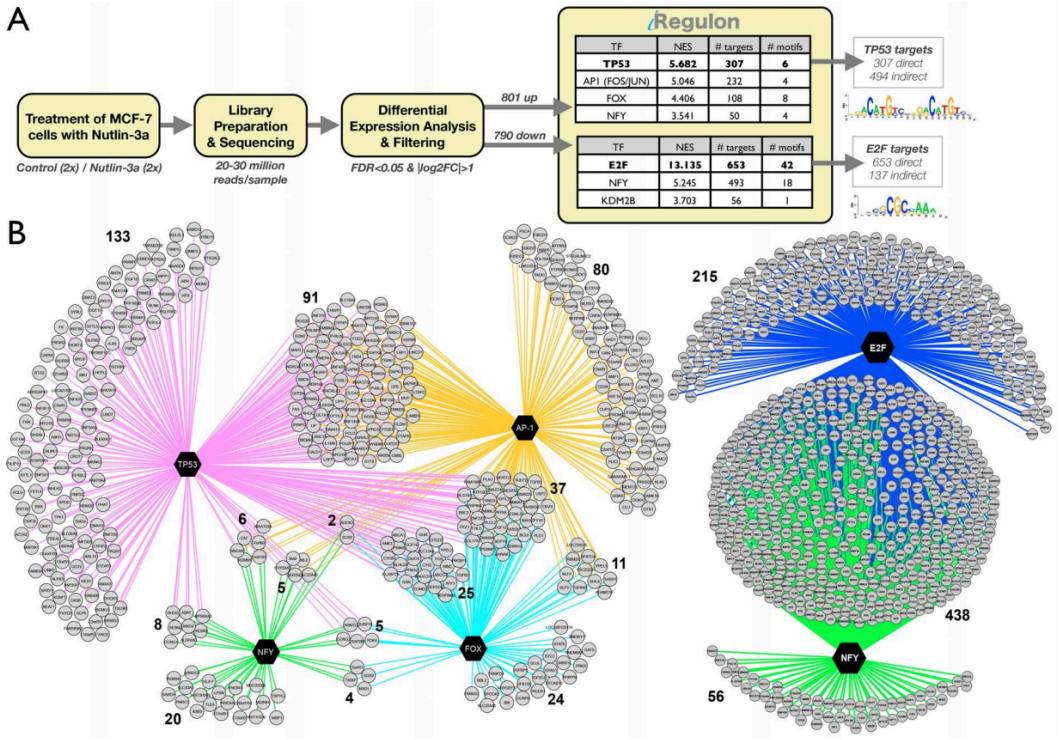
Procedures of using regression tree

1. Create 1000 learning samples using bagging/bootstraping
2. Within each learning sample
3. With a given target gene Y
4. For each of the feature gene, determine the cost values for each possible split. Chose the lowest cost value as a potential split point.
5. Iterate through each of the gene as predictor, get the lowest cost value.
6. Choose the gene that provide the lowest cost to make that first decision.
7. For each of the branch, do the same 2→4 and iterate, until all the branch reaches the minimal number of samples.
8. Sum up all the variances of Y explained by splitting by a specific gene in all the branches . That's the score for the feature importance.
9. Rank the genes by feature importance.

10. For each of the target gene, repeat through 1→7, until all the feature importance for all target genes are determined.
11. Combine all the ranks, and reranked all. That is the rank of all relationships within the transcriptome.


Genomic CRM detection
+ co-expressed candidate genes
CisTargetX
A
Dorsal PWM (MA0022)
1980 other motifs (PWMs)
Predict motif clusters in all 5kb upstream regions and introns
Cluster-Buster
B
Species 1
Region k
Gene g
Region k+1
CRM
Species i
Species 12
Integrate rankings across species
Order statistics (Ref. NN)
C
PWM-based ranking of all D. melanogaster 5kb_up and intronic "regions" & 1980 control rankings
Best region per gene
PWM-based ranking of all D. melanogaster genes & 1980 control rankings
D
Dorsal PWM (MA0022) 12 species
Dorsal PWM (MA0022) Dmel only
μ + 1.96σ
μ (average 1981 PWMs)
For 80 co-expressed genes downstream of Dorsal
13 predicted direct Dl TGs
dorsal downstream genes
CRMs
Genomic rank (first 1000 shown)
E
Histogram of AUC
AUC
F
attB
hsp70
Gateway
GFP
NLS
white

iRegulon


A
Treatment of MCF-7 cells with Nutlin-3a
Control (2x) / Nutlin-3a (2x)
Library Preparation & Sequencing
20-30 million reads/sample
Differential Expression Analysis & Filtering
FDR<0.05 & |log2FC|>1
801 up
790 down
iRegulon
TF NES # targets # motifs
TP53 5.682 307 6
AP1 (FOS/JUN) 5.046 232 4
FOX 4.406 108 8
NFY 3.541 50 4
TF NES # targets # motifs
E2F 13.135 653 42
NFY 5.245 493 18
KDM2B 3.703 56 1
TP53 targets
307 direct
494 indirect
E2F targets
653 direct
137 indirect
B
133
91
80
215
TP53
AP-1
E2F
37
6
2
11
5
25
8
5
NFY
FOX
438
4
24
20
56
NFY