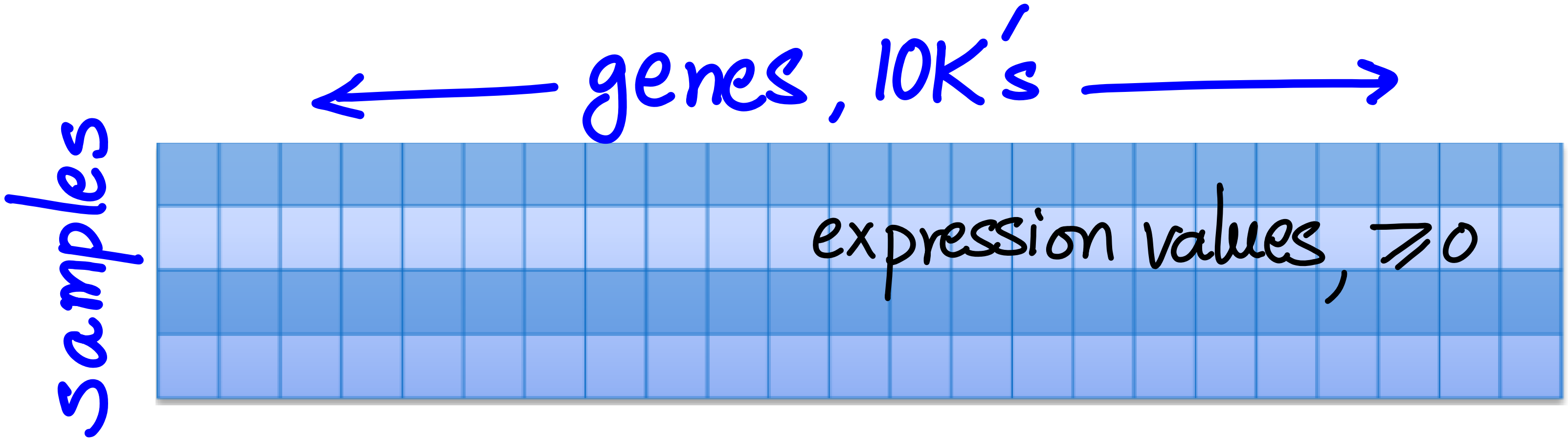
TOPOLOGICAL FEATURES IN CANCER GENE EXPRESSION DATA

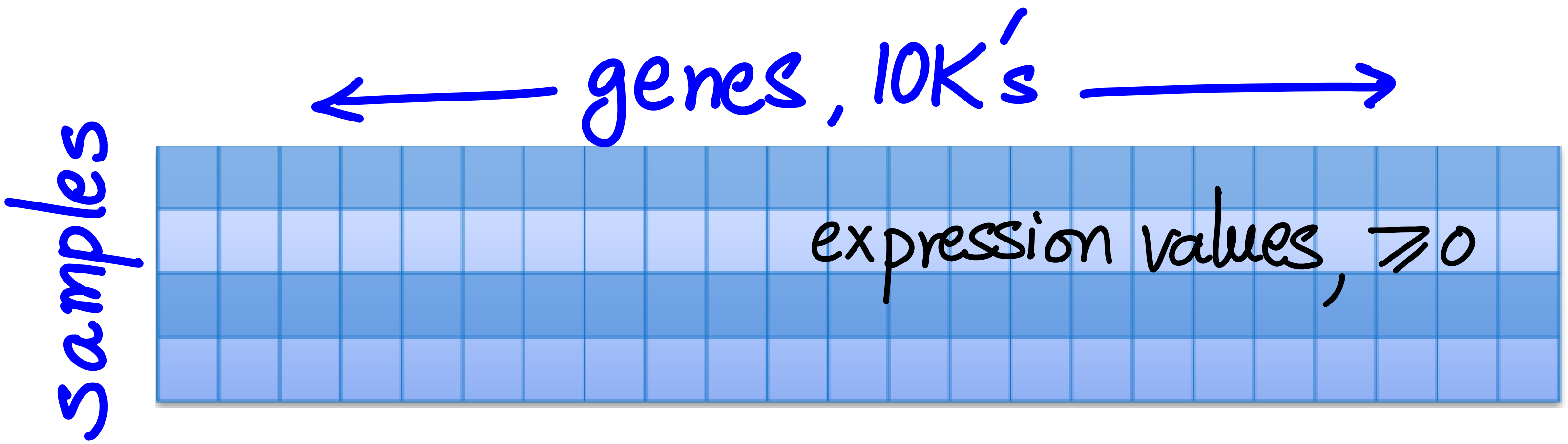
Svetlana Lockwood
Bala Krishnamoorthy
Washington State University

PSB, 2015;   arXiv: 1410.3198

CANCER GENE EXPR. DATA

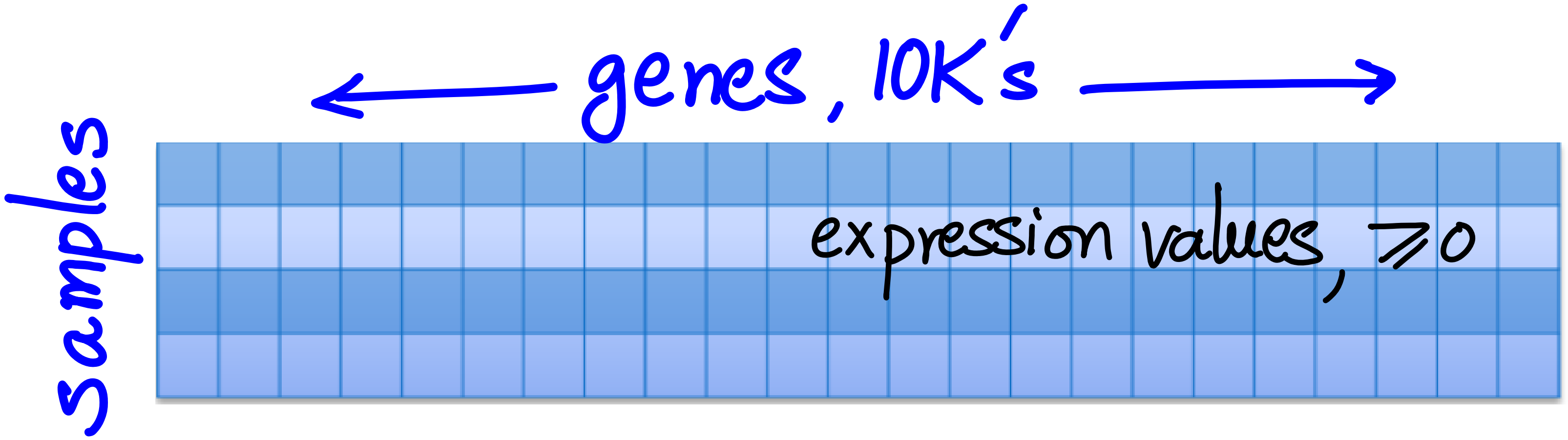


CANCER GENE EXPR. DATA



- ~50,000 genes
- 10s to a few 100 samples

CANCER GENE EXPR. DATA



- ~50,000 genes
- 10s to a few 100 samples
- select a few "biologically relevant" genes ?

THE CHALLENGE

* too many degrees of freedom

THE CHALLENGE

- * too many degrees of freedom
- * concentration of measure in high dim.
  (Beyer et.al, 1999)

THE CHALLENGE

✱ too many degrees of freedom

✱ concentration of measure in high dim. (Beyer et.al, 1999)

✗ clustering, PCA, ... might not work

THE CHALLENGE

- too many degrees of freedom
- concentration of measure in high dim. (Beyer et.al, 1999)
  - clustering, PCA, ... might not work
- "higher order" method ?

OUR RESURT

✓ "dualize" the data

- genes in samples space

OUR RESURT

✓ "dualize" the data
— genes in samples space

✓ use persistent homology to find
loops (≡ "holes")

OUR RESURT

✓ "dualize" the data
  — genes in samples space

✓ use persistent homology to find
  loops (≡ "holes")

✓ genes forming loops implicated in cancer

OUR RESURT

✓ "dualize" the data
  — genes in samples space

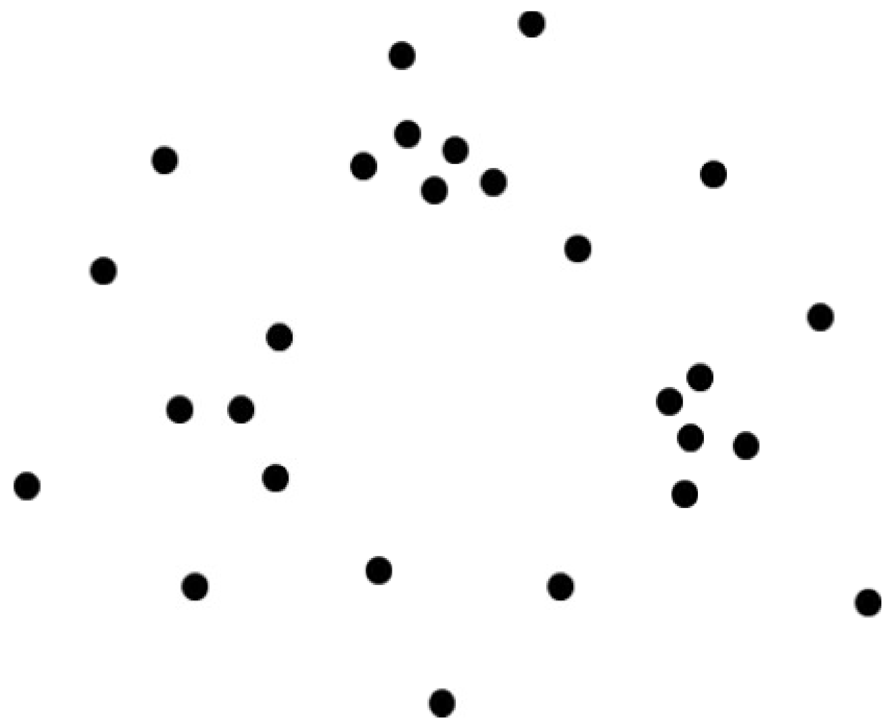
✓ use persistent homology to find
  loops (≡ "holes")

✓ genes forming loops implicated in cancer

✱ a method for data exploration...

Higher-Order Structures

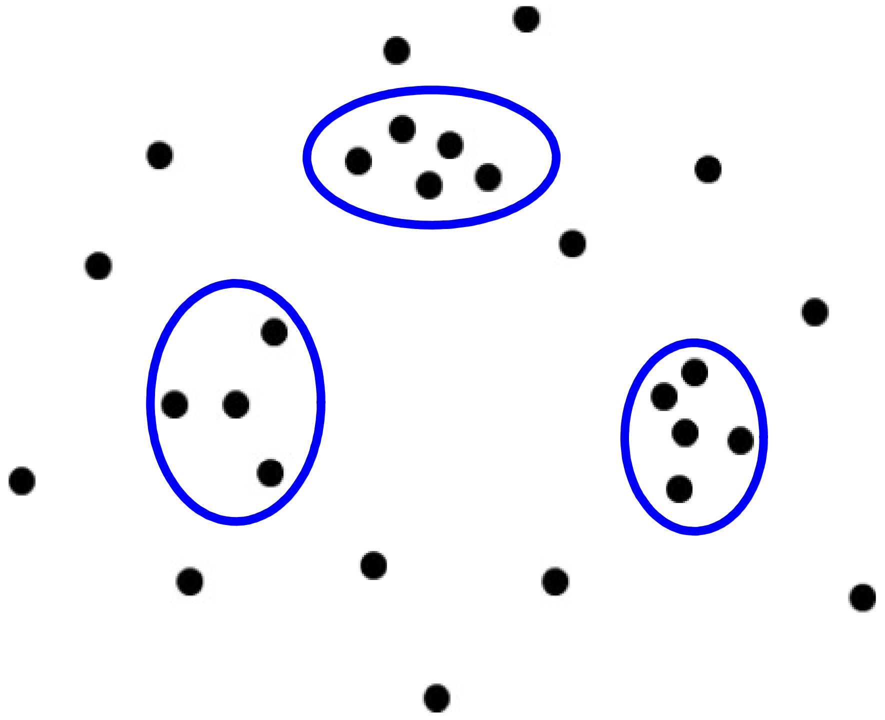
- 2D illustration



HIGHER-ORDER STRUCTURES

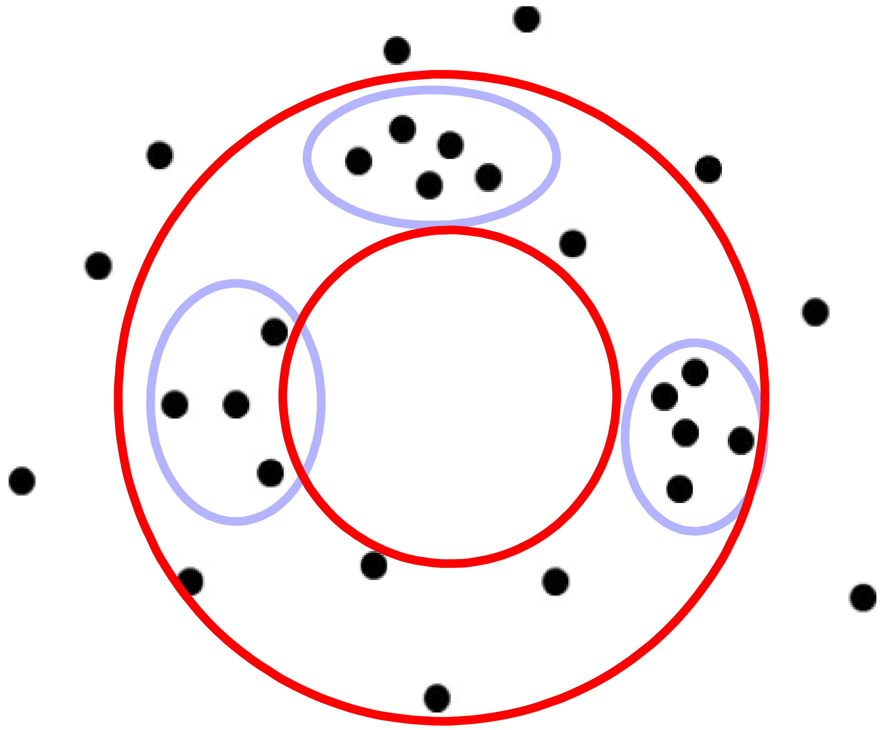
- 2D illustration

- traditional approach
  e.g., clustering
  - local structure



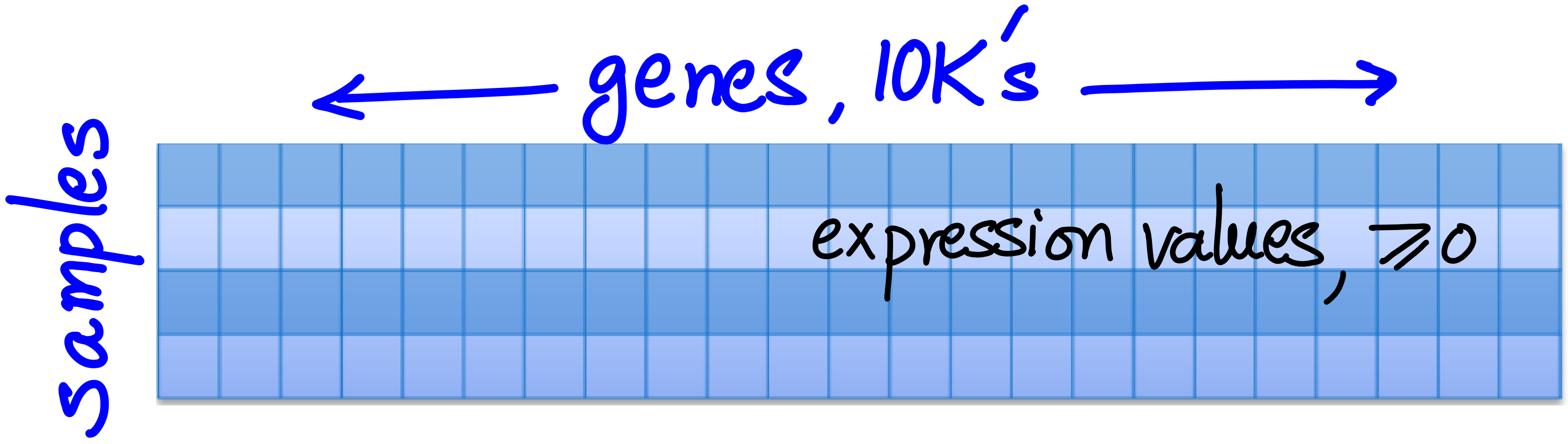
HIGHER-ORDER STRUCTURES

- 2D illustration
- traditional approach
  e.g., clustering
  – local structure
- miss higher order structure (loop)



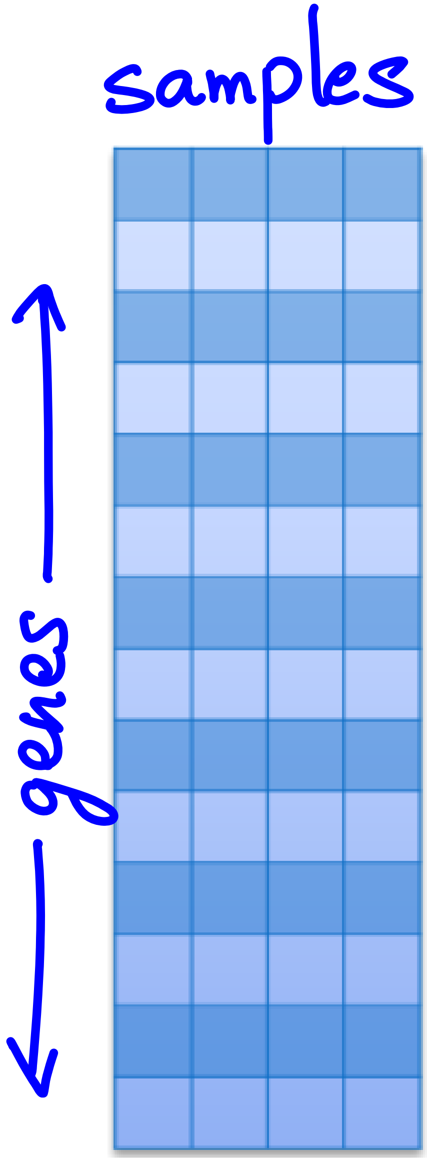
HI-DIM: DUAL SPACE

- instead of



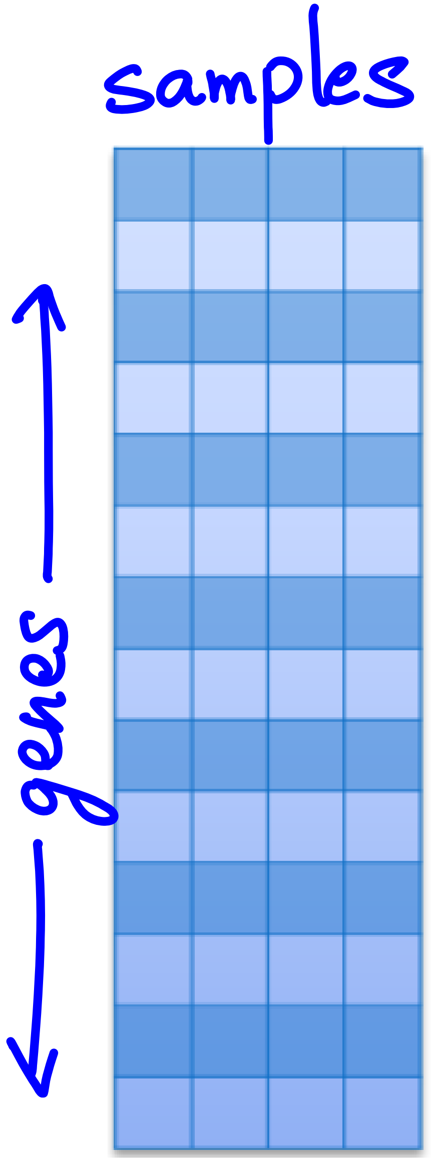
HI-DIM: DUAL SPACE

- use



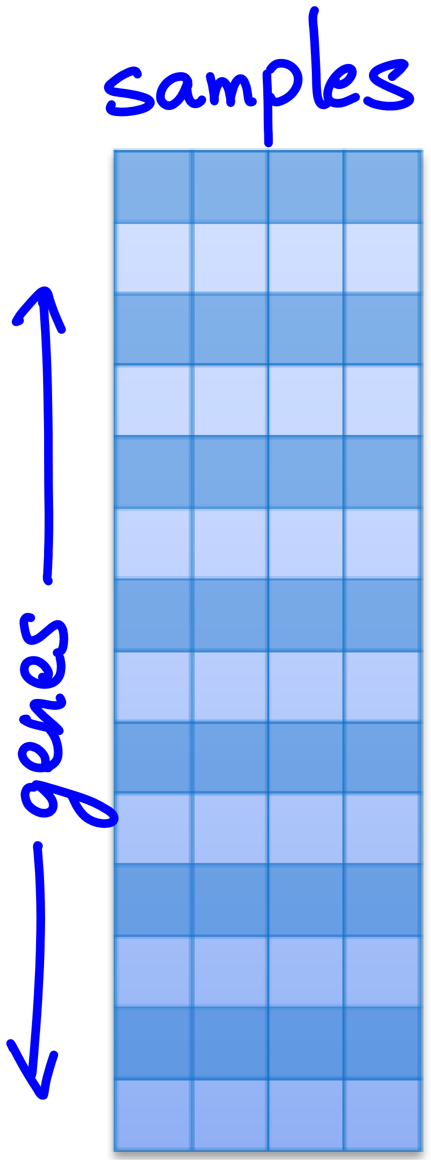
HI-DIM: DUAL SPACE

- use
- gene expressions considered across patients



HI-DIM: DUAL SPACE

- use
- gene expressions considered across patients
- pairwise distances are much more meaningful



PERSISTENT HOMOLOGY

* characterizes significant hi-dimensional "holes"

PERSISTENT HOMOLOGY

- characterizes significant hi-dimensional "holes"
- points → pairwise distances → simplicial complex → filtration → persistence diagrams/barcodes

PERSISTENT HOMOLOGY

* characterizes significant hi-dimensional "holes"

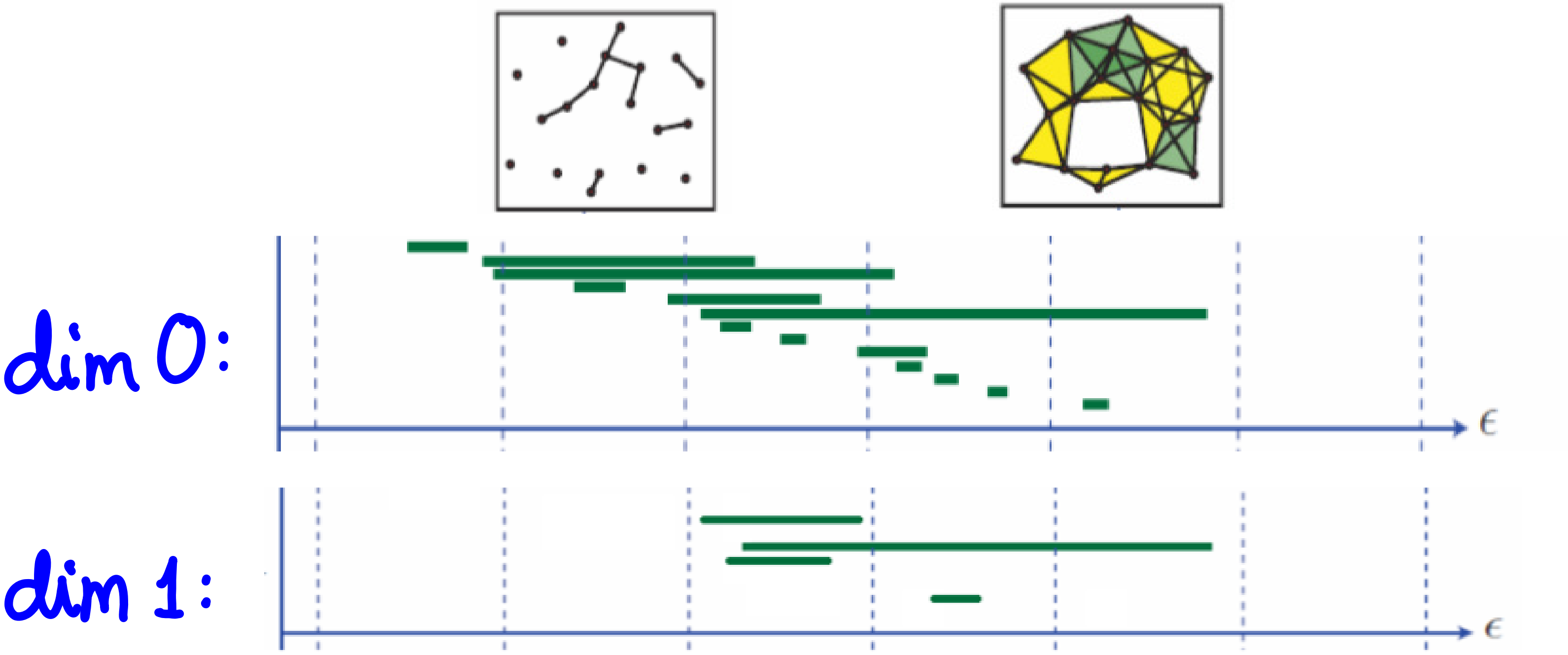
* points → pairwise distances → simplicial complex → filtration → persistence diagrams/barcodes

  - dim 0: connected components
  - dim 1: loops (around holes)
  - dim 2: enclosed voids

    ⋮

PERSISTENT HOMOLOGY

- characterizes significant hi-dimensional "holes"
- points → pairwise distances → simplicial complex → filtration → persistence diagrams/barcodes



(image adapted from Ghrist, 2008)

WITNESS COMPLEX

de Silva & Carlsson, 2004

- size of complex grows fast w/ # points in data D

WITNESS COMPLEX

de Silva & Carlsson, 2004

- ✱ size of complex grows fast w/ # points in data D
- ✱ define complex on $L \subset D$, subset of landmarks
  - edge $[\ell_0, \ell_1]$ is in complex if $\exists\, v \in D$ s.t. $\ell_0, \ell_1 \in L$ are 2 nearest neighbors of v

Witness Complex

de Silva & Carlsson, 2004

- size of complex grows fast w/ # points in data D
- define complex on $L \subset D$, subset of landmarks
  - edge $[\ell_0 \ell_1]$ is in complex if $\exists v \in D$ s.t. $\ell_0, \ell_1 \in L$ are 2 nearest neighbors of v
  - p-simplex $[\ell_0 \ell_1 \dots \ell_p]$ is in complex if $\exists v \in D$ s.t. $\ell_0, \dots, \ell_p \in L$ are $(p+1)$ nearest neighbors of v

WITNESS COMPLEX

de Silva & Carlsson, 2004

* size of complex grows fast w/ # points in data D

* define complex on $L \subset D$, subset of landmarks
  - edge $[l_0 l_1]$ is in complex if $\exists\, v \in D$ s.t. $l_0, l_1 \in L$ are 2 nearest neighbors of v
  - p-simplex $[l_0 l_1 \dots l_p]$ is in complex if $\exists\, v \in D$ s.t. $l_0, \dots, l_p \in L$ are $(p+1)$ nearest neighbors of v
  - v is the witness for the p-simplex

METHOD

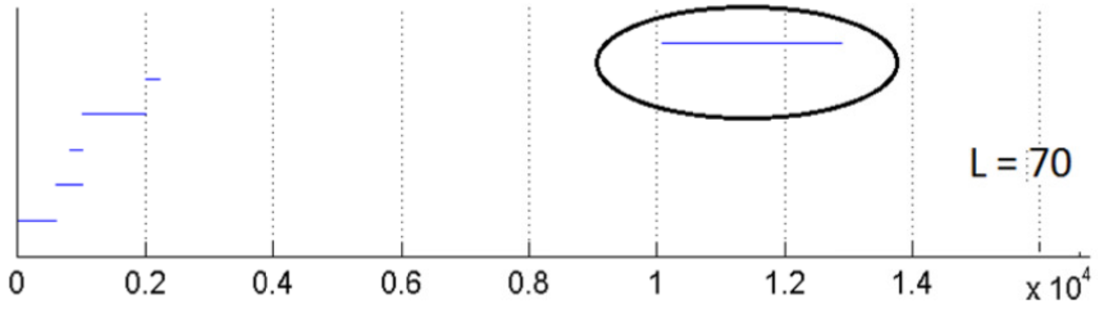
- for breast cancer (54,613 genes, 47 samples)

  dim-1 barcode

METHOD

* for breast cancer (54,613 genes, 47 samples)

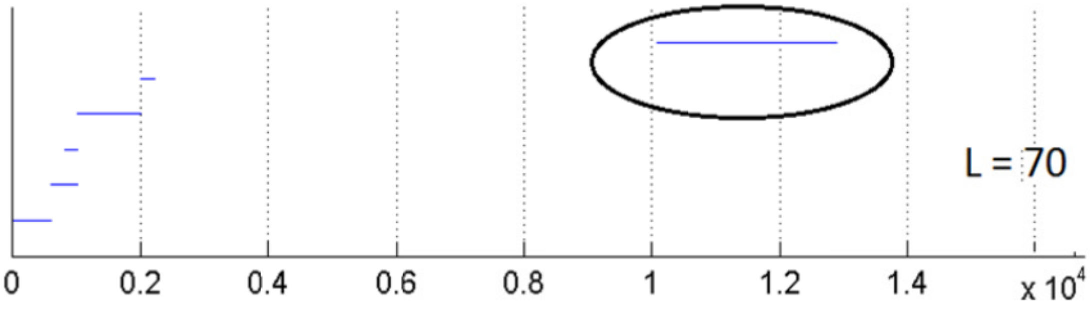
dim-1 barcode



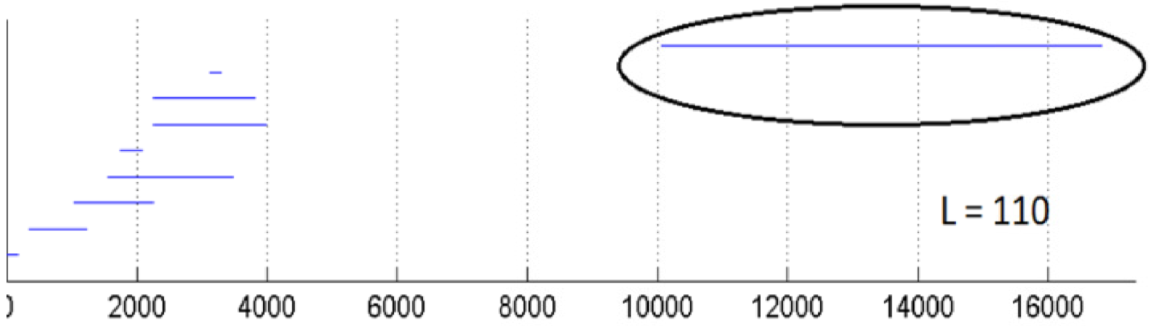
genes = 70

METHOD

* for breast cancer (54,613 genes, 47 samples)
  dim-1 barcode



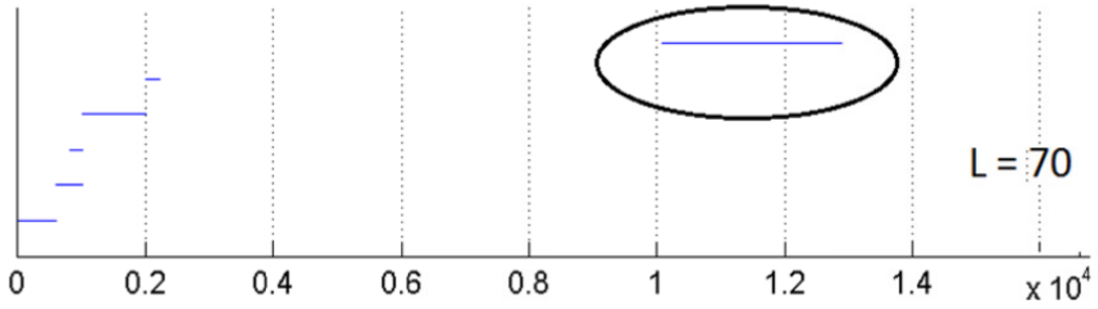
genes = 70



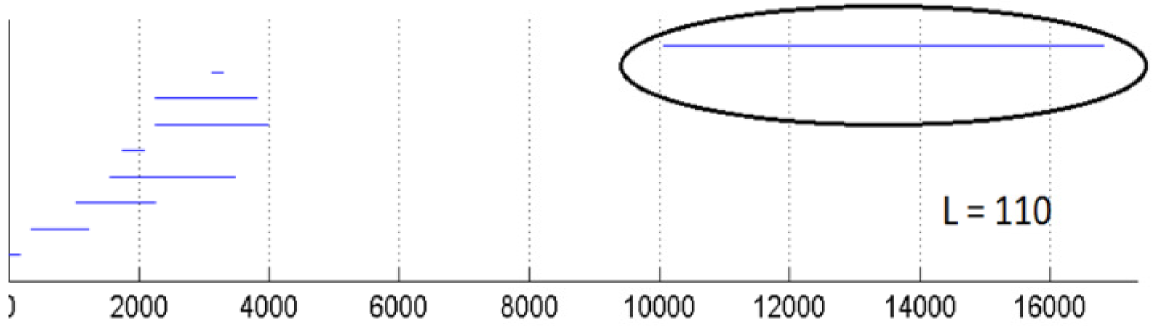
genes = 110

METHOD

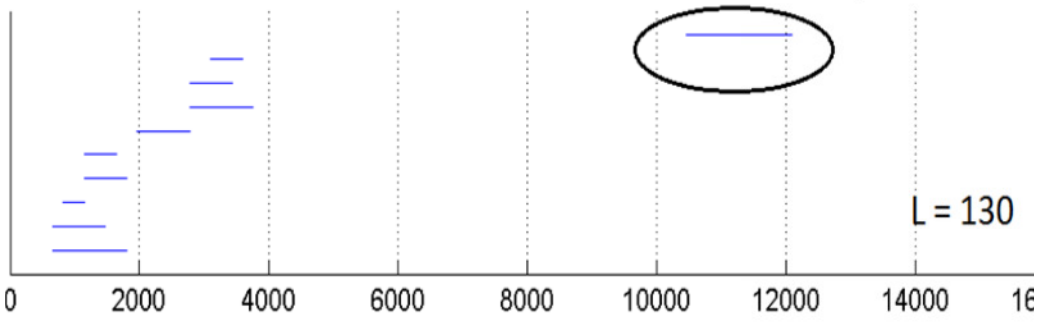
✱ for breast cancer (54,613 genes, 47 samples)
dim-1 barcode



genes = 70



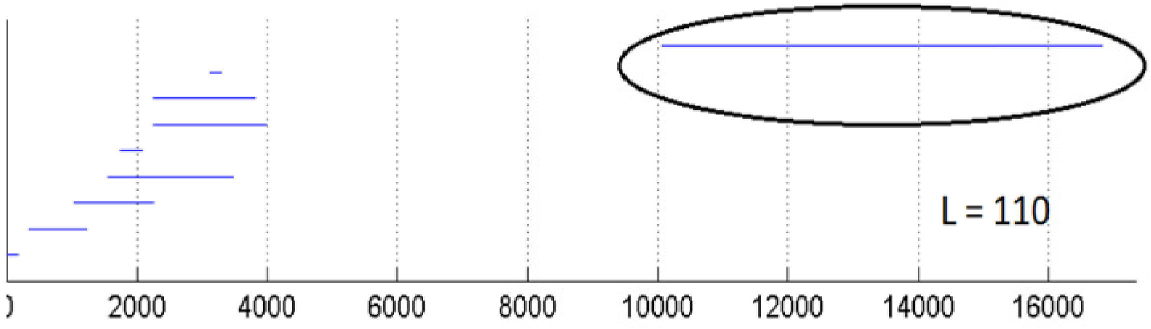
genes = 110



genes = 130

METHOD

- for breast cancer (54,613 genes, 47 samples)
  dim-1 barcode
- pick the longest loop(s)



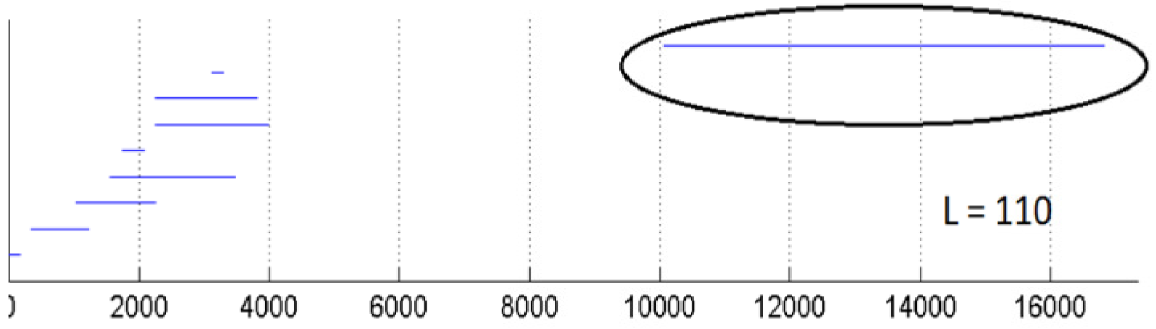
genes = 110

METHOD

- for breast cancer (54,613 genes, 47 samples)

dim-1 barcode

- pick the longest loop(s)



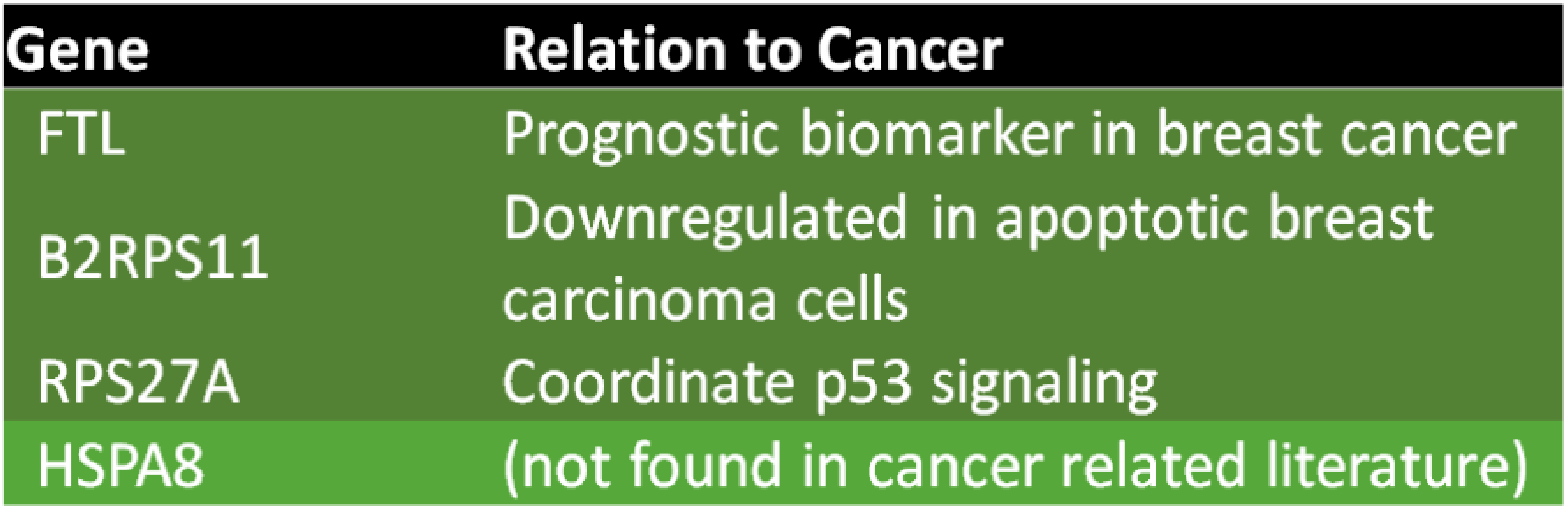
genes = 110

- Are genes in loop(s) relevant for cancer?

METHOD

- for breast cancer (54,613 genes, 47 samples)
  dim-1 barcode

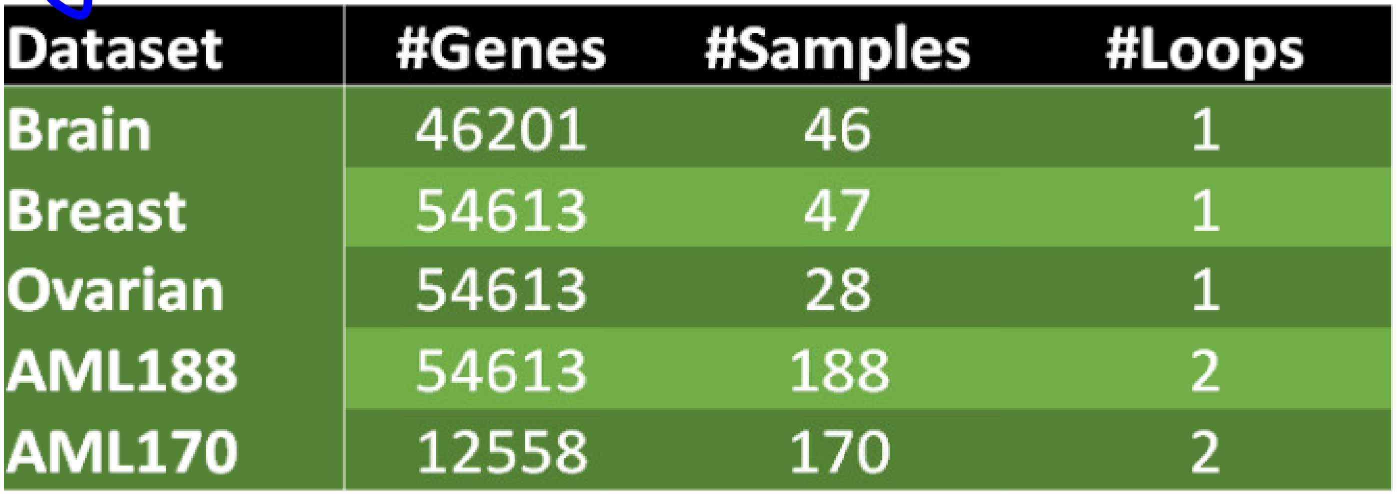
- genes in the breast cancer loop:



Gene	Relation to Cancer
FTL	Prognostic biomarker in breast cancer
B2RPS11	Downregulated in apoptotic breast carcinoma cells
RPS27A	Coordinate p53 signaling
HSPA8	(not found in cancer related literature)

RESULTS

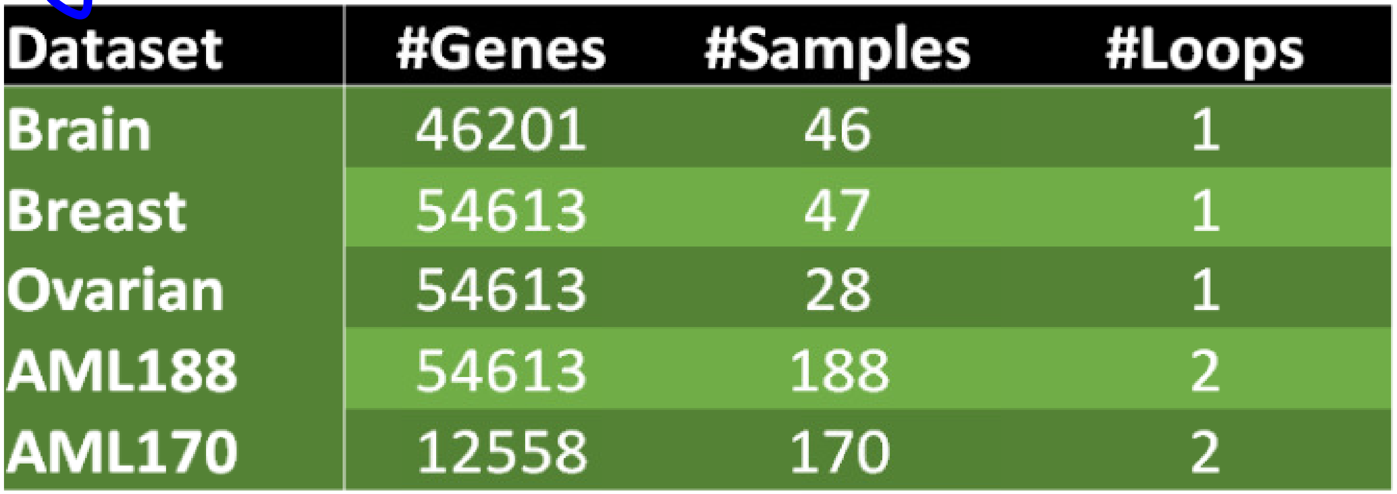
* analyzed five different cancer datasets



Dataset	#Genes	#Samples	#Loops
Brain	46201	46	1
Breast	54613	47	1
Ovarian	54613	28	1
AML188	54613	188	2
AML170	12558	170	2

RESULTS

- analyzed five different cancer datasets

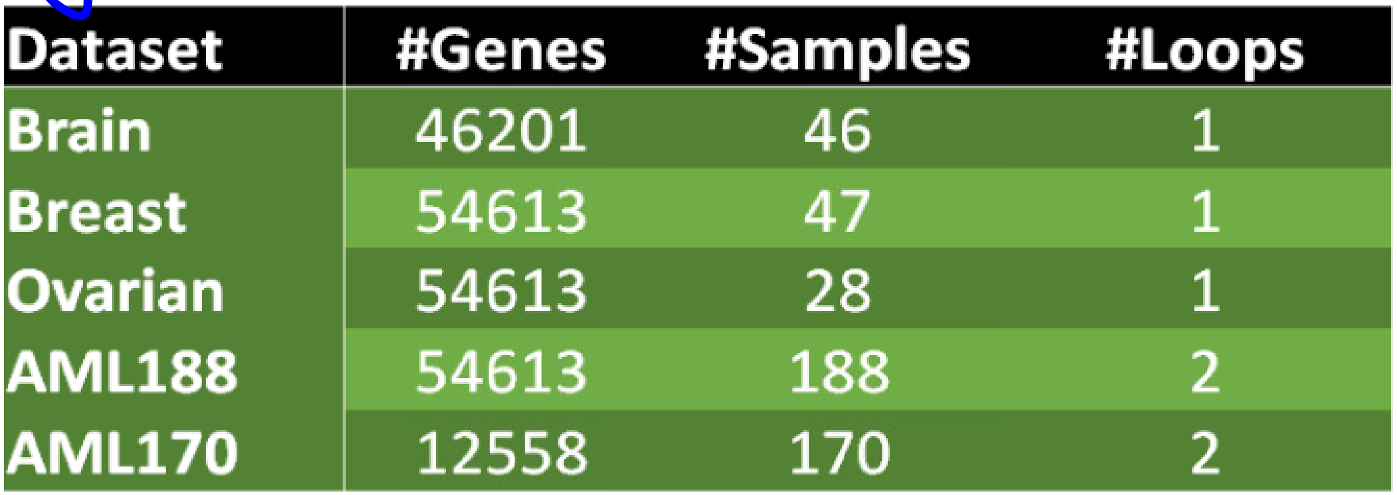


Dataset	#Genes	#Samples	#Loops
Brain	46201	46	1
Breast	54613	47	1
Ovarian	54613	28	1
AML188	54613	188	2
AML170	12558	170	2

- majority of loop genes implicated in cancer in all cases

RESULTS

* analyzed five different cancer datasets



Dataset	#Genes	#Samples	#Loops
Brain	46201	46	1
Breast	54613	47	1
Ovarian	54613	28	1
AML188	54613	188	2
AML170	12558	170	2

* majority of loop genes implicated in cancer in all cases

* selected landmarks (L), as well as loop genes do <u>not</u> have extreme expression values

OPEN QUESTIONS

- small groups (6-13) of genes forming loops cannot be found by other methods

OPEN QUESTIONS

- ✱ small groups (6-13) of genes forming loops cannot be found by other methods
- ✱ Does loop connectedness of genes imply functional connectedness?

OPEN QUESTIONS

- small groups (6-13) of genes forming loops cannot be found by other methods
- Does loop connectedness of genes imply functional connectedness?
  - hard to study coexpression of multiple genes

OPEN QUESTIONS

- ✱ small groups (6-13) of genes forming loops cannot be found by other methods

- ✱ Does loop connectedness of genes imply functional connectedness?
  - — hard to study coexpression of multiple genes

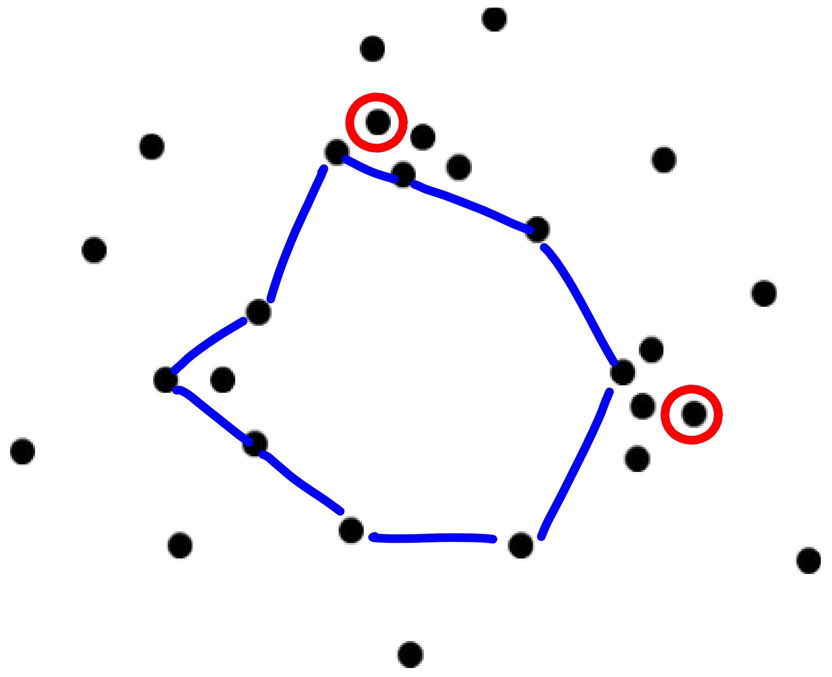
- ✱ Does dualization affect ability to prove results on structure/stability of data?

OPEN QUESTIONS

* a few relevant genes not included in loops

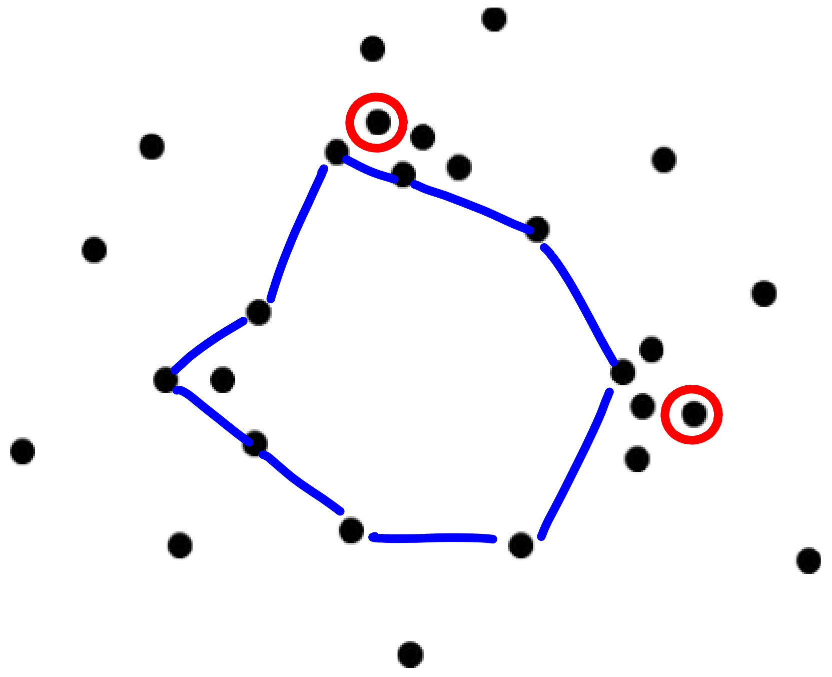
OPEN QUESTIONS

- a few relevant genes not included in loops



OPEN QUESTIONS

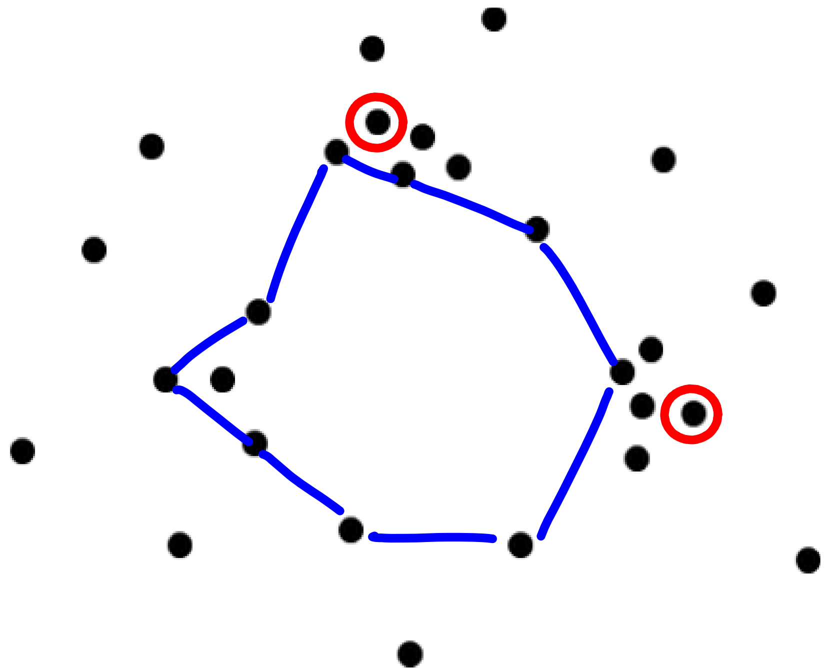
- a few relevant genes not included in loops



- Can we identify loop(s) with "all critical genes"?

OPEN QUESTIONS

- a few relevant genes not included in loops



- Can we identify loop(s) with "all critical genes"?
- Apply to other classes of data sets?