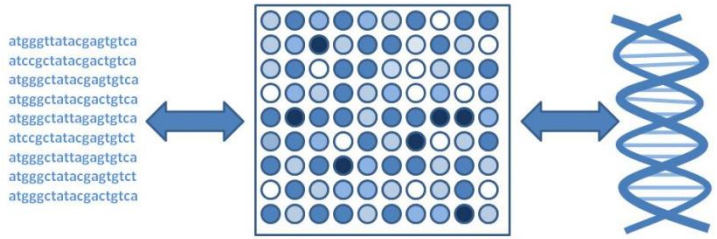
Correlate

A method for the integrative analysis of two genomic data sets

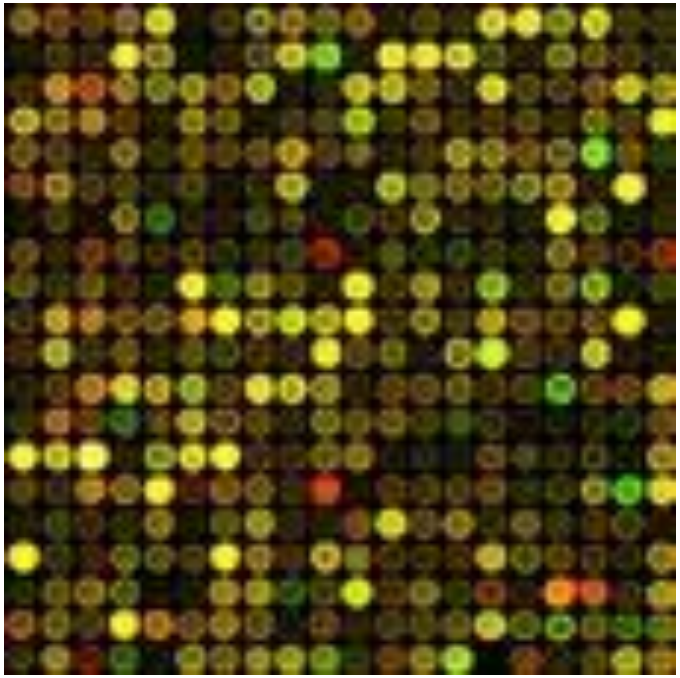


Sam Gross, Balasubramanian Narasimhan,
Robert Tibshirani, and Daniela Witten

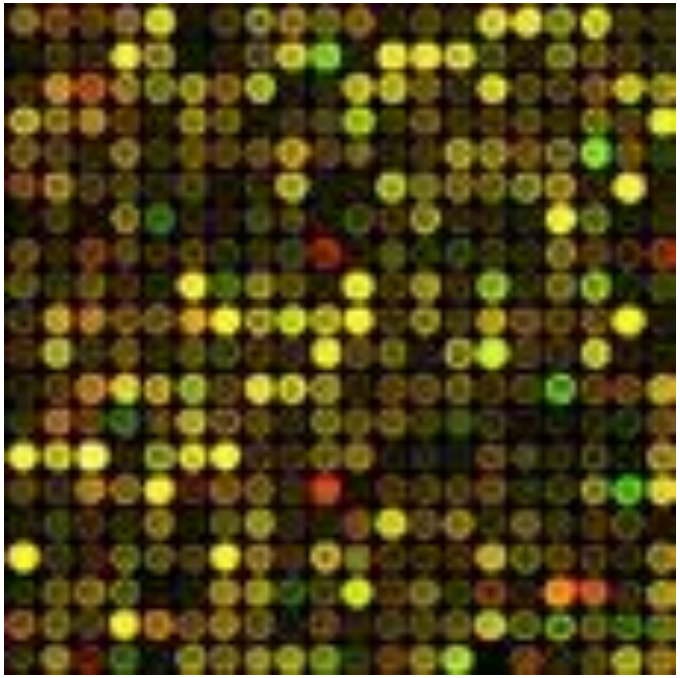
February 19, 2010

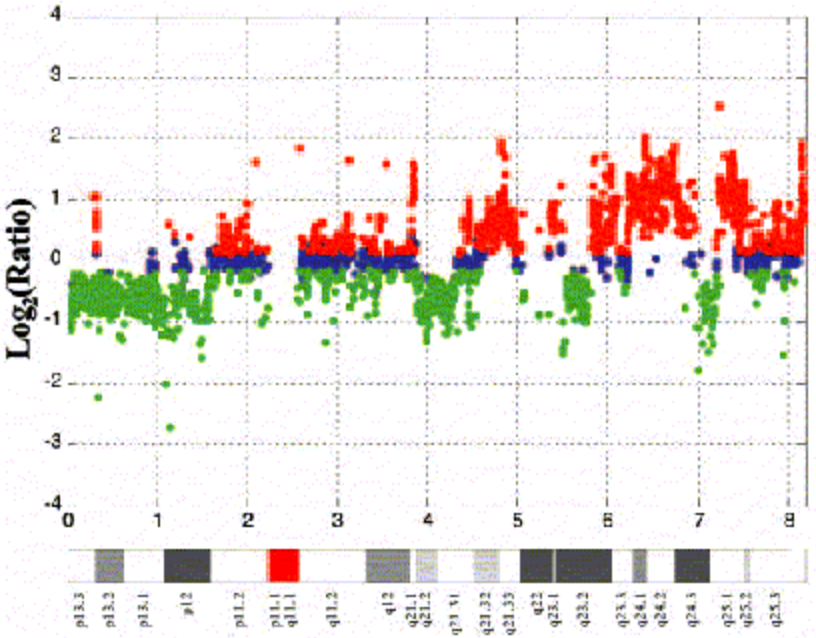
- Introduction
- Sparse Canonical Correlation Analysis
- `Correlate`: an Excel add-in that implements sparse CCA

A world of data

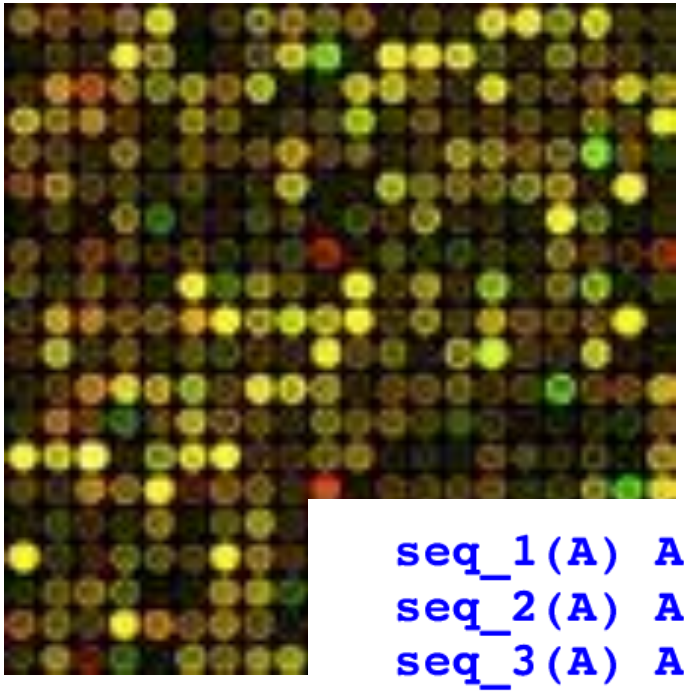


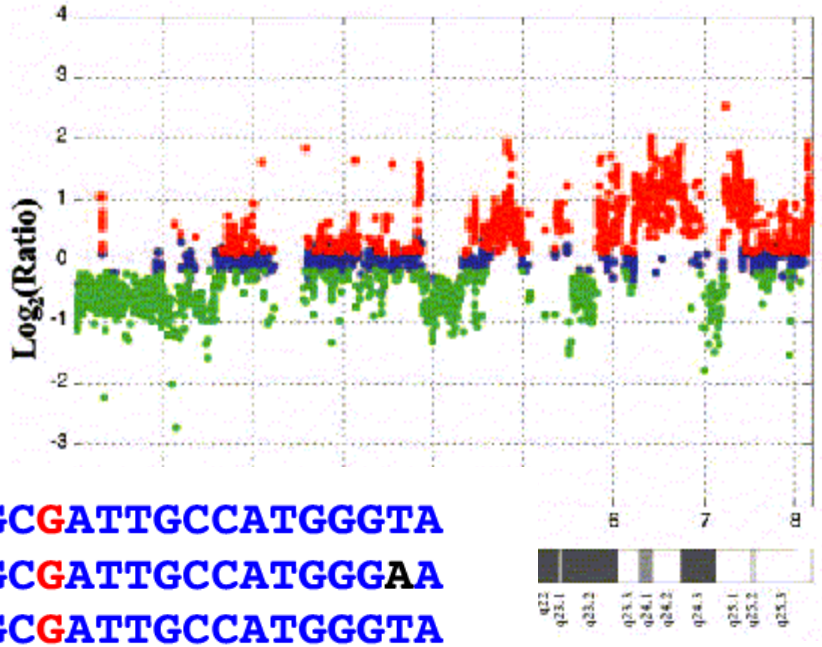
A world of data





A world of data



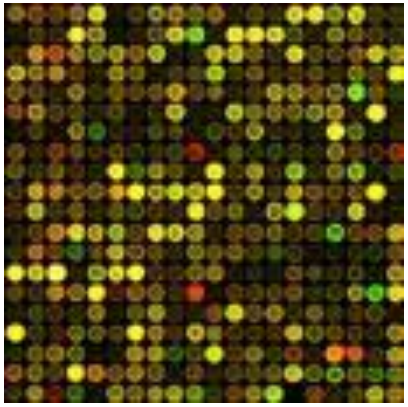



```
seq_1(A)  ATGCGGCGATTGCCATGGGTA
seq_2(A)  ATGCGGCGATTGCCATGGGAA
seq_3(A)  ATGCGGCGATTGCCATGGGTA
seq_1(B)  ATGCGGCAATTGCCATGGGTA
seq_2(B)  ATGCGGCAATTGCCATGGGTT
seq_3(B)  ATGCGGCAATTGCCATGGGTA
Contig    ATGCGGCGATTGCCATGGGTA
```

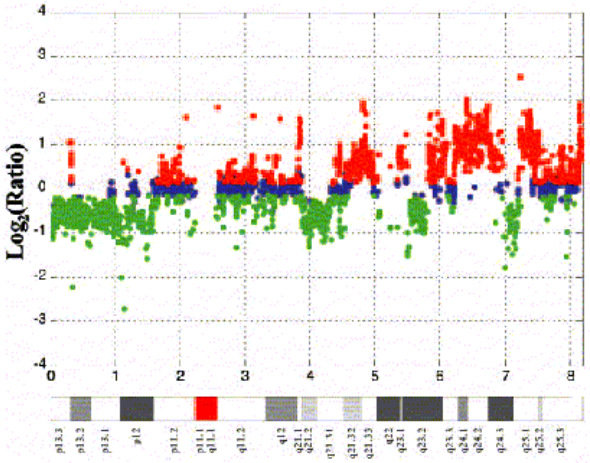
SNP

sequencing errors or paralogs

Statistical analyses



There are great statistical methods for the analysis of gene expression, DNA copy number, and SNP data sets.




```
seq_1(A)  ATGCGGCGATTGCCATGGGTA
seq_2(A)  ATGCGGCGATTGCCATGGGAA
seq_3(A)  ATGCGGCGATTGCCATGGGTA
seq_1(B)  ATGCGGCAATTGCCATGGGTA
seq_2(B)  ATGCGGCAATTGCCATGGGTT
seq_3(B)  ATGCGGCAATTGCCATGGGTA
Contig    ATGCGGCGATTGCCATGGGTA
              SNP ↑      ↑ ↑
                  sequencing errors or paralogs
```

An integrative approach

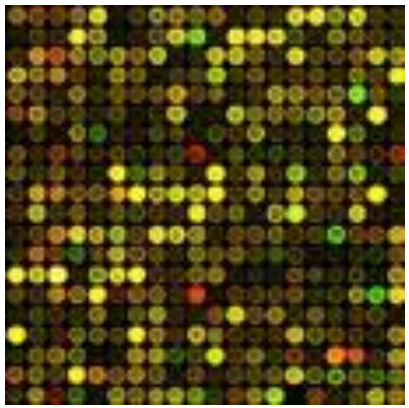
- But what if we have access to multiple types of data (for instance, gene expression and DNA copy number data) on a single set of samples?

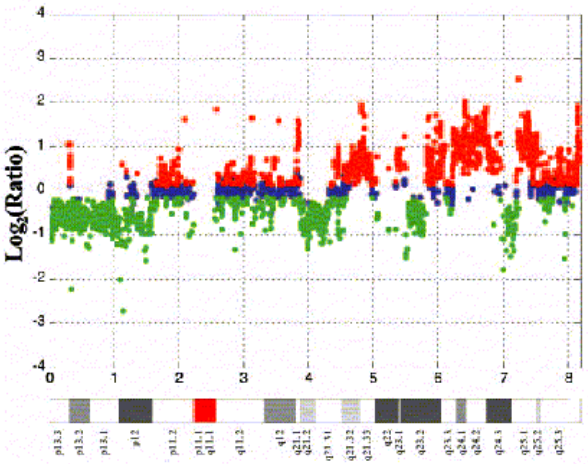
An integrative approach

- But what if we have access to multiple types of data (for instance, gene expression and DNA copy number data) on a single set of samples?
- The data types can be apples and oranges: for instance, imaging data and gene expression data

An integrative approach

- But what if we have access to multiple types of data (for instance, gene expression and DNA copy number data) on a single set of samples?
- The data types can be apples and oranges: for instance, imaging data and gene expression data





Introduction

- In this talk, we'll consider the case of DNA copy number and gene expression measurements on a single set of samples.

Introduction

- In this talk, we'll consider the case of DNA copy number and gene expression measurements on a single set of samples.
- Sparse CCA gives us a tool that can be used to answer the question:
  Can we identify a ***small set of gene expression measurements*** *that is correlated with a* ***region of DNA copy number gain/loss****?*

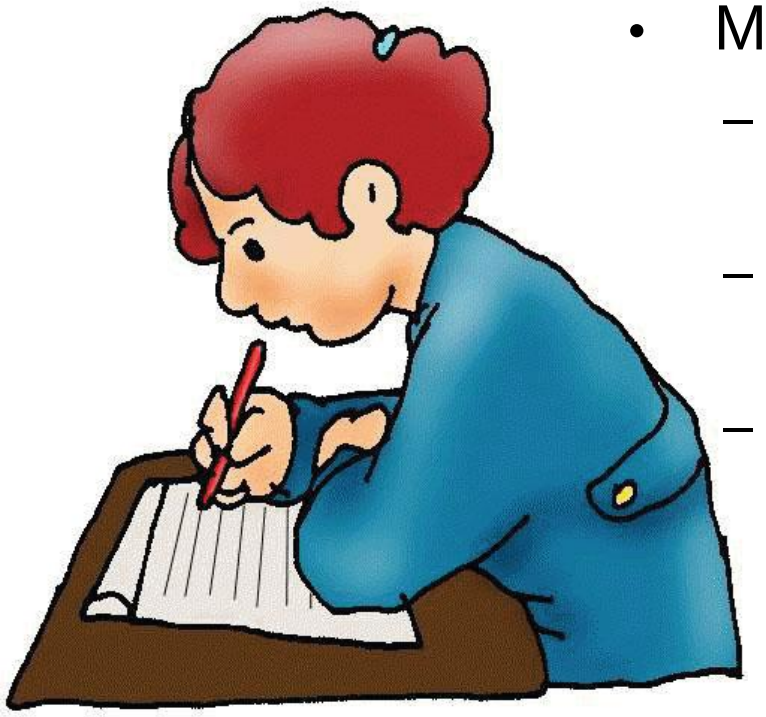
Introduction

- In this talk, we'll consider the case of DNA copy number and gene expression measurements on a single set of samples.
- Sparse CCA gives us a tool that can be used to answer the question:
  *Can we identify a **small set of gene expression measurements** that is correlated with a **region of DNA copy number gain/loss**?*
- `Correlate` provides an easy way to apply that method using Microsoft Excel

Canonical Correlation Analysis (CCA)

- CCA is a classical statistical method
- Suppose we have n samples and p+q features for each sample
  - Let the sample be a group of n kids
  - Let the first p features be their scores on a set of p tests: reading comprehension, Latin, math…
  - Let the next q features be the amount of time they spend on certain activities per week: team sports, watching TV, reading…

CCA



- **The question:** How are the q activities associated with scores on the p exams?
- Maybe
  - More Reading ⇔ Better Reading Comprehension Scores
  - More Reading And Less TV ⇔ Even Better Reading Comprehension Scores
  - More Reading, More team sports, More Homework, and Less TV ⇔ Good Scores on all tests

CCA

- Canonical correlation analysis allows us to discover relationships like this between the sets of variables.
- For instance, perhaps

  0.6*ReadingComp + 0.8*Math + .743*Latin

  is **highly correlated** with

  2*TeamSports − 11*TV + 8*Reading + 234*Homework

CCA

- CCA looks for linear combinations of variables in the two groups that are **highly correlated** with each other.
- Let **X** be a matrix with *n* columns - one for each student - and *p* = 3 rows, one for each test (Reading Comprehension, Math, Latin).
- And let **Y** be a matrix with *n* columns and *q* = 4 rows, one for each activity (Team Sports, TV, Reading, Homework).
- Statistically, we seek vectors **u** and **v** such that Cor(**X'u**, **Y'v**) is big. We can think of the components of **u** and **v** as weights for each variable.

CCA

- Thus, the output tell us that

  0.6*ReadingComp + 0.8*Math + .743*Latin

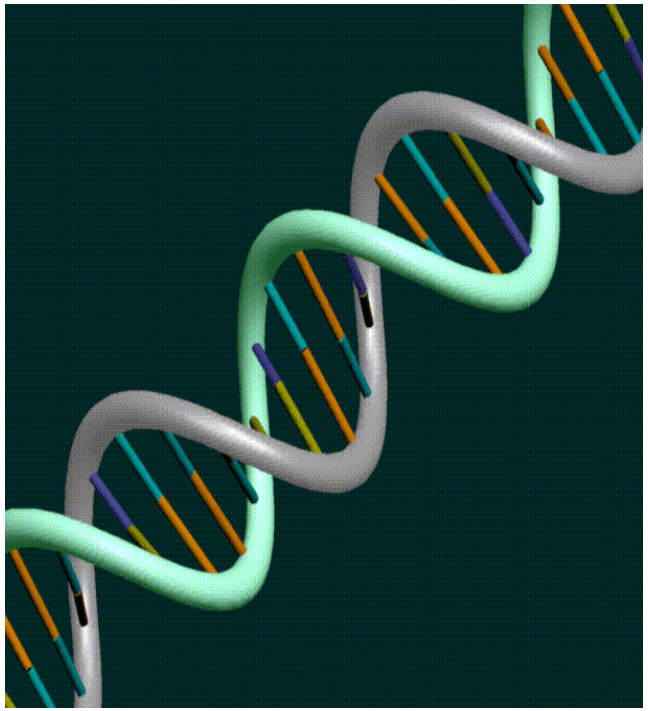
  is **highly correlated** with

  2*TeamSports − 11*TV + 8*Reading + 234*Homework

- Here,
  - u = (0.6, 0.8, 0.743)’
  - v = (2, −11, 8, 234)’

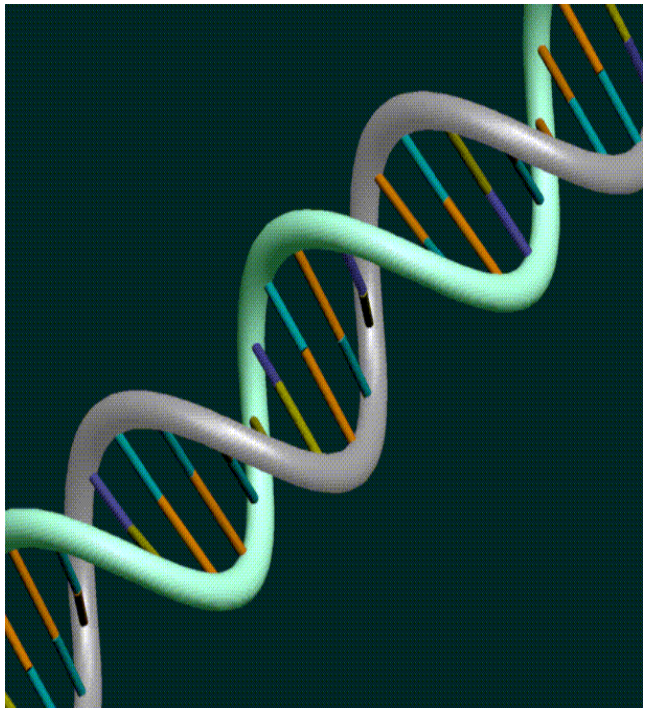
Why is it useful?

- How does this apply to genomics and bioinformatics?



Why is it useful?

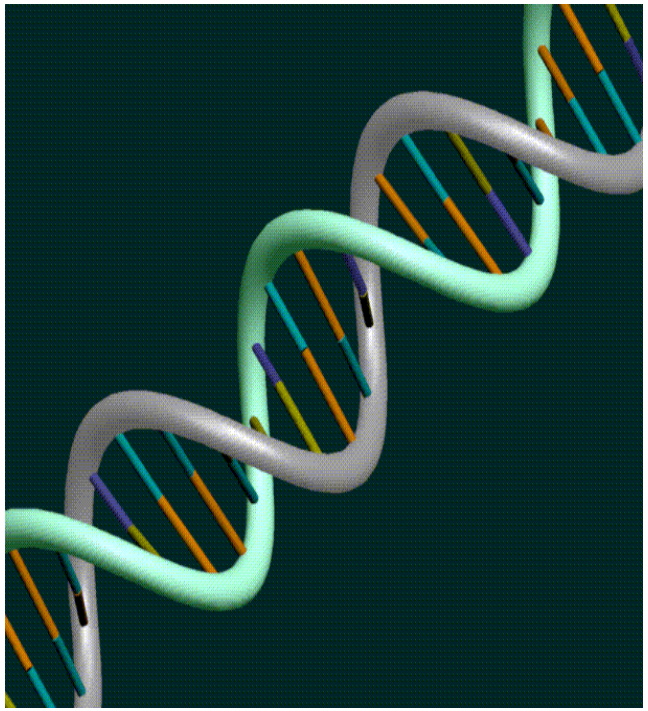
- How does this apply to genomics and bioinformatics?
- If we have copy number and gene expression measurements on the same set of samples, we can ask:



Why is it useful?

- How does this apply to genomics and bioinformatics?
- If we have copy number and gene expression measurements on the same set of samples, we can ask:

  Which genes have expression that is ***associated*** *with which regions of DNA gain or loss?*



Sparse CCA

- This is almost the question that CCA answers for us...
  - But, CCA will give us a linear combination of genes that is associated with a linear combination of DNA copy number measurements
  - These linear combinations will involve every gene expression measurement and every copy number measurement

Sparse CCA

- This is almost the question that CCA answers for us...
  - But, CCA will give us a linear combination of genes that is associated with a linear combination of DNA copy number measurements
  - These linear combinations will involve every gene expression measurement and every copy number measurement
- What we really want is this:
  - A short list of genes that are associated with a particular region of DNA gain/loss

Sparse CCA

- From now on:
  - **X** is a matrix of gene expression data, with samples on the columns and genes on the rows
  - **Y** is a matrix of copy number data, with samples on the columns and copy number measurements on the rows

X	Sample1	Sample2	...	SampleN
Gene1	.	.	...	.
Gene2	.	.	...	.
...	...	...	...	.
GeneP	.	.	.	.

Y	Sample1	Sample2	...	SampleN
CGH Spot1	.	.	...	.
CGH Spot2	.	.	...	.
...	...	...	...	.
CGH SpotQ	.	.	.	.

Sparse CCA

- CCA seeks weights **u**, **v** such that Cor(**X'u**, **Y'v**) is big

Sparse CCA

- CCA seeks weights **u**, **v** such that

  Cor(**X'u**, **Y'v**) is big
- Sparse CCA seeks weights **u**, **v** such that Cor(**X'u**, **Y'v**) is big, and most of the weights are zero

Sparse CCA

- CCA seeks weights **u**, **v** such that

  Cor(**X'u**, **Y'v**) is big
- Sparse CCA seeks weights **u**, **v** such that Cor(**X'u**, **Y'v**) is big, and most of the weights are zero
- **u** contains weights for the gene expression data, and **v** contains weights for the copy number data

Sparse CCA

- CCA seeks weights **u**, **v** such that

  Cor(**X'u**, **Y'v**) is big
- Sparse CCA seeks weights **u**, **v** such that Cor(**X'u**, **Y'v**) is big, and most of the weights are zero
- **u** contains weights for the gene expression data, and **v** contains weights for the copy number data
- Since the columns of **Y** are copy number measurements along the chromosome, then we want the weights in **v** to be smooth (not jumpy)

Sparse CCA

- By imposing the right penalty on **u** and **v**, we can ensure that
  - The elements of **u** are sparse
  - The elements of **v** are sparse and smooth
  - (Remember: **u** contains weights for the gene expression data, and **v** contains weights for the copy number data)

Sparse CCA

- By imposing the right penalty on **u** and **v**, we can ensure that
  - The elements of **u** are sparse
  - The elements of **v** are sparse and smooth
  - (Remember: **u** contains weights for the gene expression data, and **v** contains weights for the copy number data)
- We can also constrain **u** and **v** such that their weights are positive or negative

Sparse CCA, mathematically

We choose weights **u** and **v** to maximize

$$\text{Cor}(\mathbf{X}'\mathbf{u}, \mathbf{Y}'\mathbf{v}) \text{ subject to } \sum_i |u_i| \leq c_1,$$
$$\sum_j (|v_j| + |v_{j+1} - v_j|) \leq c_2$$

This is a **lasso** constraint on u and a **fused lasso** constraint on v.

For small values of c_1 and c_2, some elements of **u** and **v** are exactly zero, and **v** is smooth.

For the statisticians: the criterion

Assume that the features are standardized to have mean 0 and standard deviation 1.

$$\text{maximize}_{\mathbf{u},\mathbf{v}}\ \mathbf{u}'\mathbf{X}\mathbf{Y}'\mathbf{v}$$

$$\text{subject to } \mathbf{u}'\mathbf{u} \le 1,\ \mathbf{v}'\mathbf{v} \le 1,\ P_1(\mathbf{u}) \le c_1,\ P_2(\mathbf{v}) \le c_2$$

Here, P_1 and P_2 are convex penalties on the elements of **u** and **v**.

For the statisticians: biconvexity

$$\text{maximize}_{u,v}\ \mathbf{u}'\mathbf{X}\mathbf{Y}'\mathbf{v}$$

$$\text{subject to } \mathbf{u}'\mathbf{u} \le 1,\ \mathbf{v}'\mathbf{v} \le 1,\ P_1(\mathbf{u}) \le c_1,\ P_2(\mathbf{v}) \le c_2$$

- With **u** fixed, the criterion is convex in **v**, and with **v** fixed, it's convex in **u**.
- This suggests a simple iterative optimization strategy:
  1. Hold **u** fixed and optimize with respect to **v**.
  2. Hold **v** fixed and optimize with respect to **u**.

For the statisticians: the algorithm

$$\text{maximize}_{u,v}\ \mathbf{u}'\mathbf{X}\mathbf{Y}'\mathbf{v}$$

$$\text{subject to } \mathbf{u}'\mathbf{u} \le 1,\ \mathbf{v}'\mathbf{v} \le 1,\ P_1(\mathbf{u}) \le c_1,\ P_2(\mathbf{v}) \le c_2$$

- Initialize **v**.
- Iterate until convergence:

1. Hold **v** fixed, and optimize:
$\text{maximize}_u\ \mathbf{u}'\mathbf{X}\mathbf{Y}'\mathbf{v}$ subject to $\mathbf{u}'\mathbf{u} \le 1,\ P_1(\mathbf{u}) \le c_1$.
2. Hold **u** fixed, and optimize:
$\text{maximize}_v\ \mathbf{u}'\mathbf{X}\mathbf{Y}'\mathbf{v}$ subject to $\mathbf{v}'\mathbf{v} \le 1,\ P_2(\mathbf{v}) \le c_2$.

For the statisticians: the penalties

$$\text{maximize}_{u,v}\ \mathbf{u}'\mathbf{X}\mathbf{Y}'\mathbf{v}$$

$$\text{subject to } \mathbf{u}'\mathbf{u} \le 1,\ \mathbf{v}'\mathbf{v} \le 1,\ P_1(\mathbf{u}) \le c_1,\ P_2(\mathbf{v}) \le c_2$$

If P_1 is a lasso or L_1 penalty, $P_1(\mathbf{u})=||\mathbf{u}||_1$, then to update $\mathbf{u}$:

$$\mathbf{u}=S(\mathbf{X}\mathbf{Y}'\mathbf{v},\ d)/||S(\mathbf{X}\mathbf{Y}'\mathbf{v},d)||_2,$$

where $d \ge 0$ is chosen such that $||\mathbf{u}||_1=c_1$.

Here, S is the *soft-thresholding operator*: $S(a,c)=\text{sign}(a)(|a|-c)_+$.

For the statisticians: the penalties

$$\text{maximize}_{u,v}\ \mathbf{u}'\mathbf{X}\mathbf{Y}'\mathbf{v}$$

$$\text{subject to } \mathbf{u}'\mathbf{u} \le 1,\ \mathbf{v}'\mathbf{v} \le 1,\ P_1(\mathbf{u}) \le c_1,\ P_2(\mathbf{v}) \le c_2$$

If P_2 is a fused lasso penalty:

$$P_2(\mathbf{v}) = \sum_j (|v_j| + |v_{j+1} - v_j|) \le c_2,$$

then the update is a little harder and requires software for fused lasso regression.

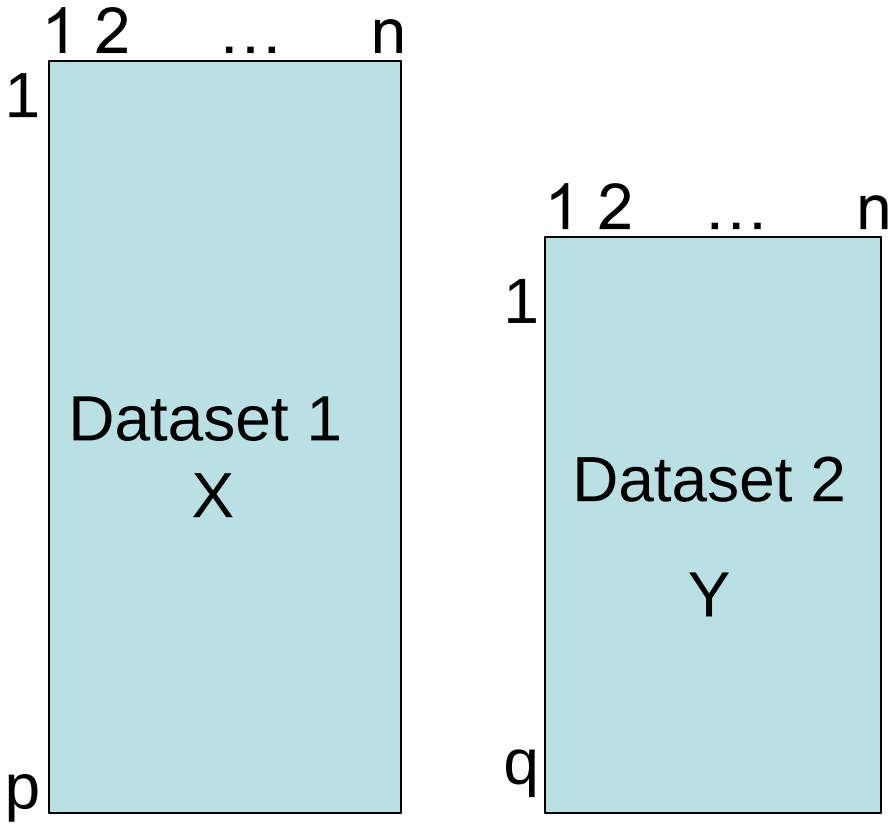
Sparse CCA results

- So what do we end up with?
  - A set of genes that is associated with a region (or regions) of DNA gain/loss
  - Weights for the gene expression measurements (can be constrained to all have the same sign)
  - Weights for the DNA copy number measurements, which will be smooth
  - We can get multiple (gene set, DNA gain/loss) pairs

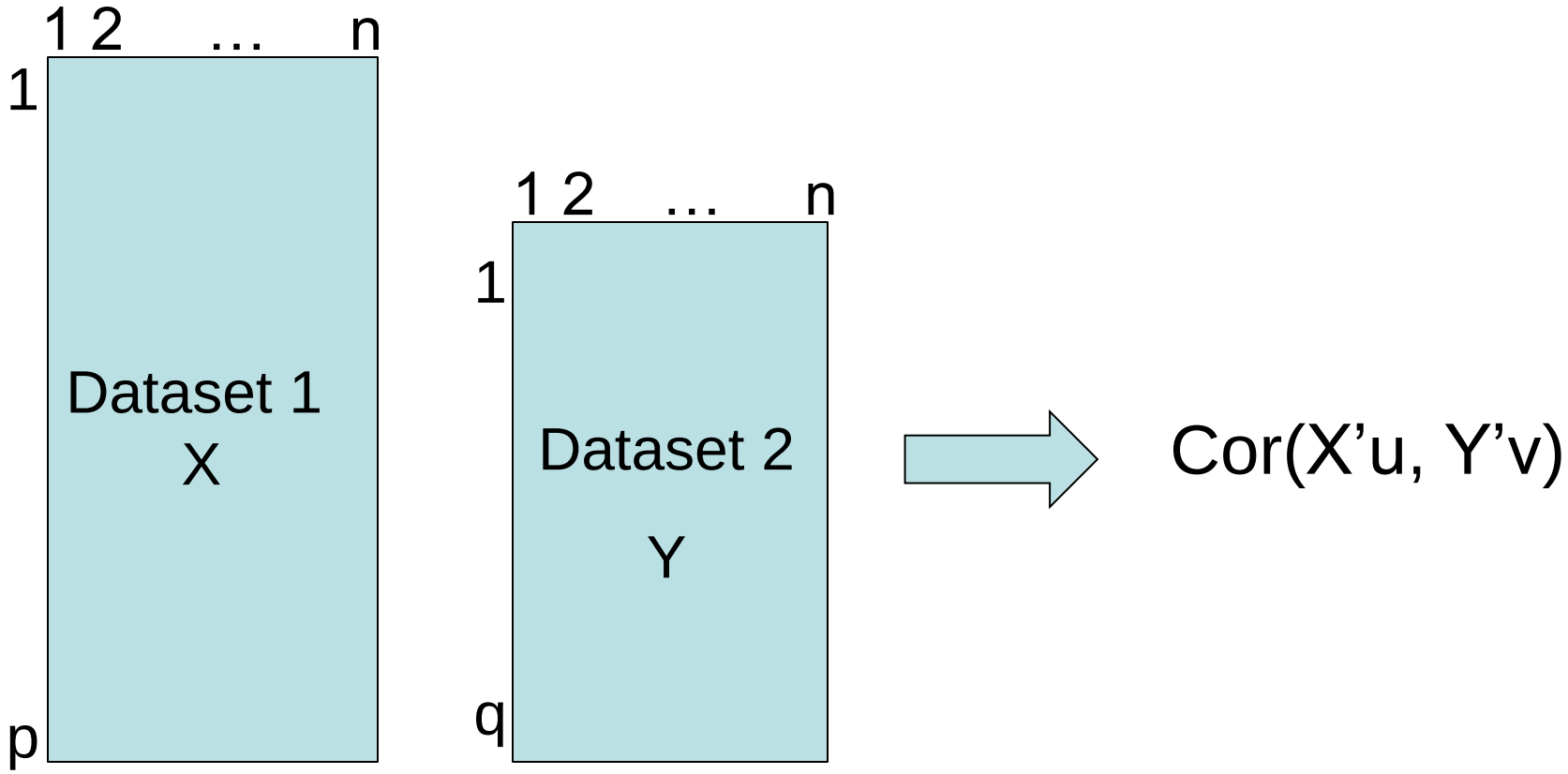
Sparse CCA results

- So what do we end up with?
  - A set of genes that is associated with a region (or regions) of DNA gain/loss
  - Weights for the gene expression measurements (can be constrained to all have the same sign)
  - Weights for the DNA copy number measurements, which will be smooth
  - We can get multiple (gene set, DNA gain/loss) pairs
- We use a permutation approach to get a p-value for the significance of the results

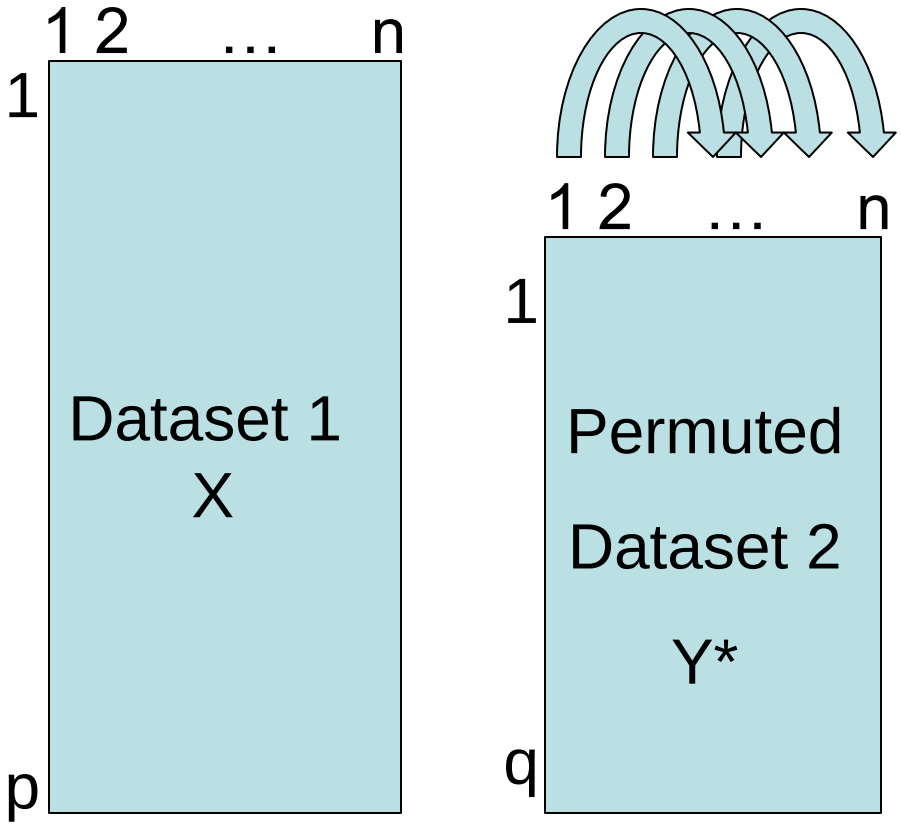
Permutation approach



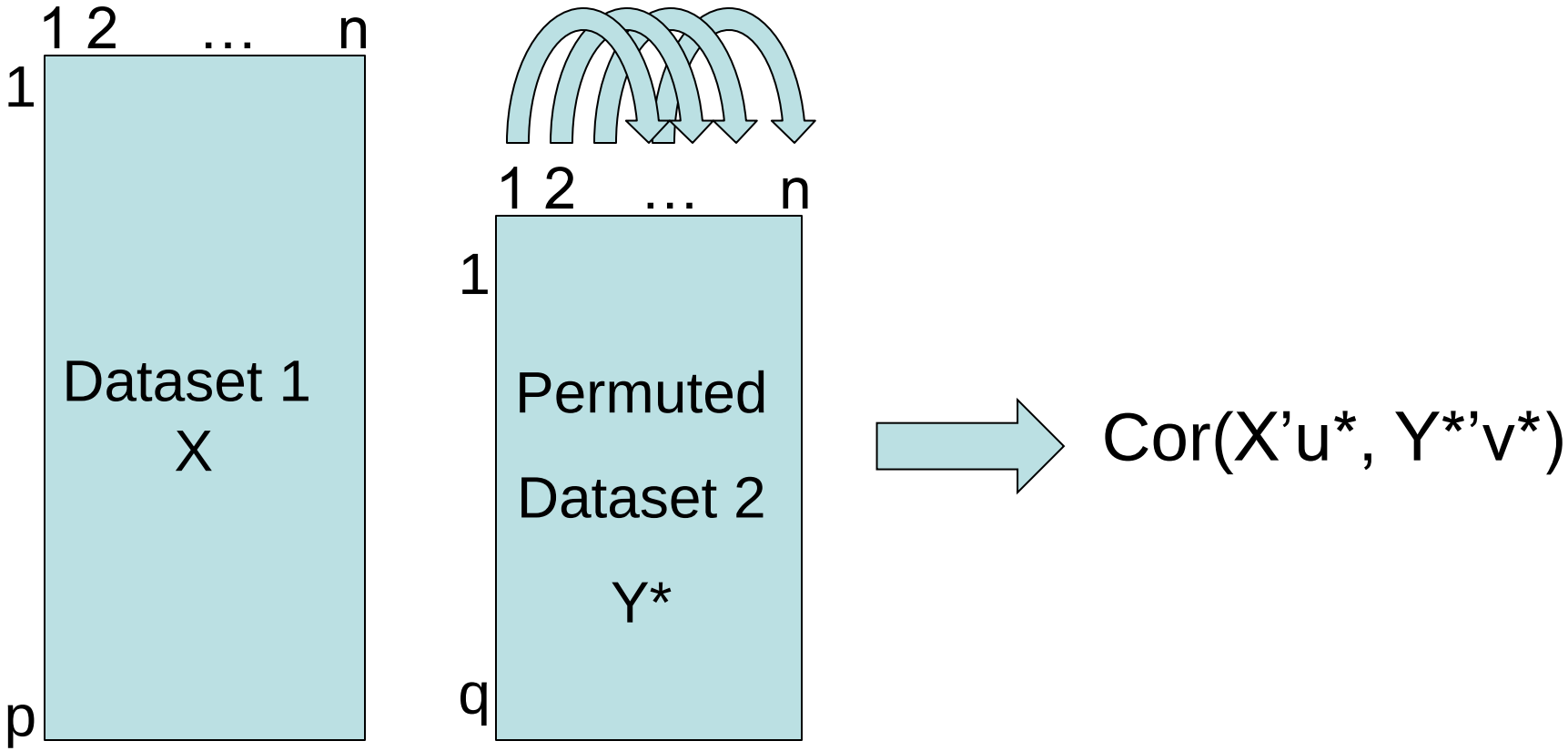
Permutation approach



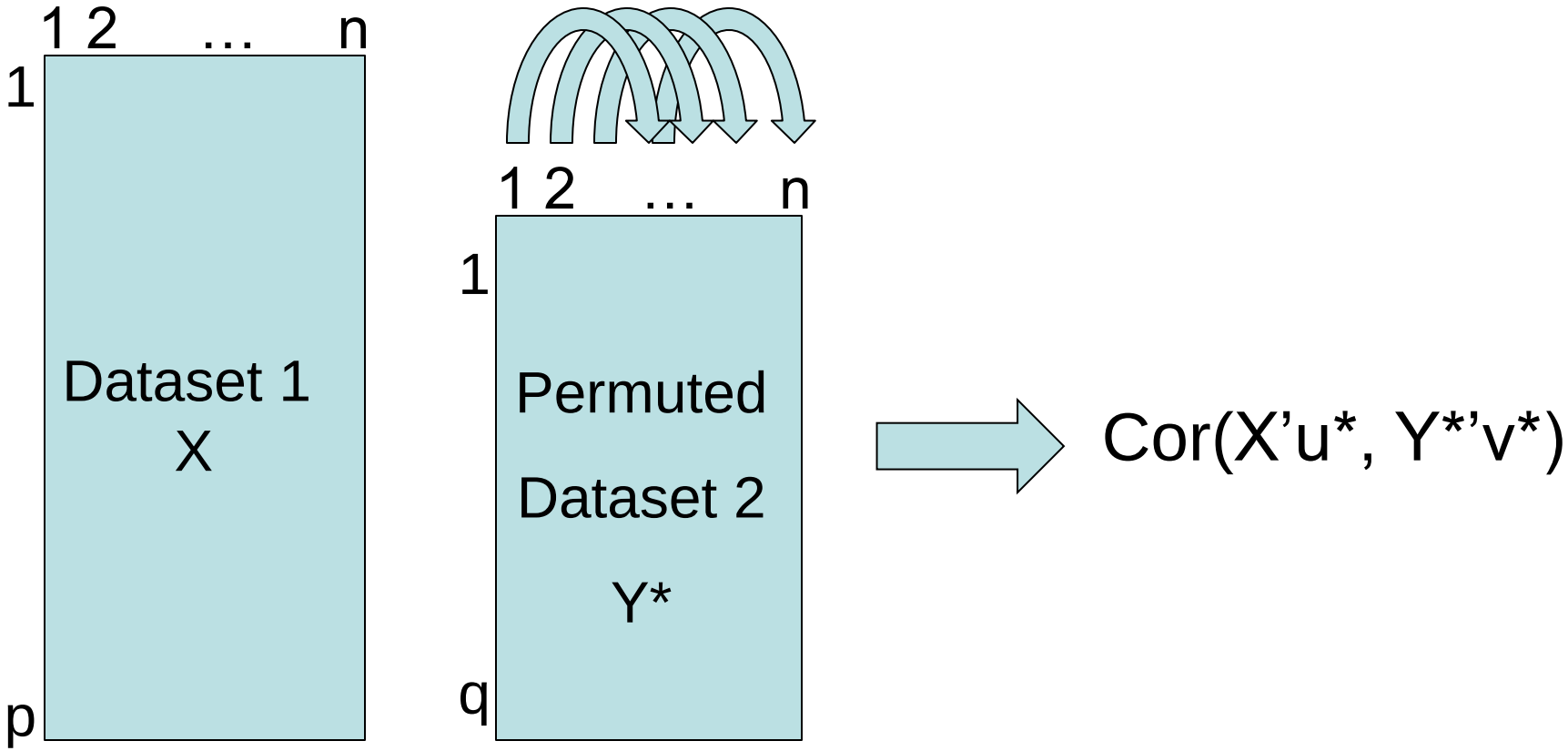
Permutation approach



Permutation approach



Permutation approach



1. Repeat 100 times.

2. Compare Cor(X'u, Y'v) to {Cor(X'u*, Y*'v*)}.

Extensions

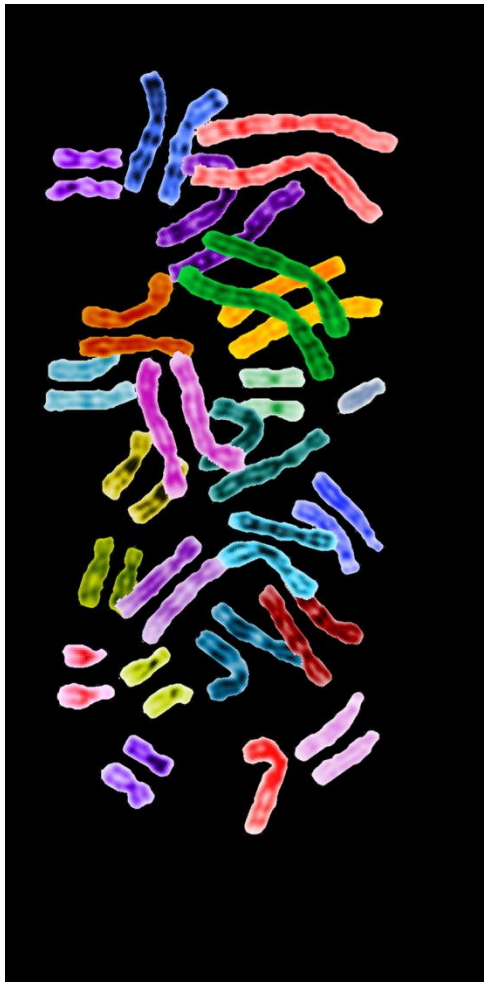
These ideas have been extended to the following cases:

- More than two data sets
- A supervising outcome (e.g. survival time or tumor subtype) for each sample

Data

- Applied to breast cancer data:
  - n = 89 tissue samples
  - p = 19672 gene expression measurements
  - q = 2149 DNA copy number measurements
  - Chin, DeVries, Fridlyand, et al. (2006) Cancer Cell 10, 529-541.
- Look for a region of copy number change on chromosome 20 that's correlated with the expression of some set of genes

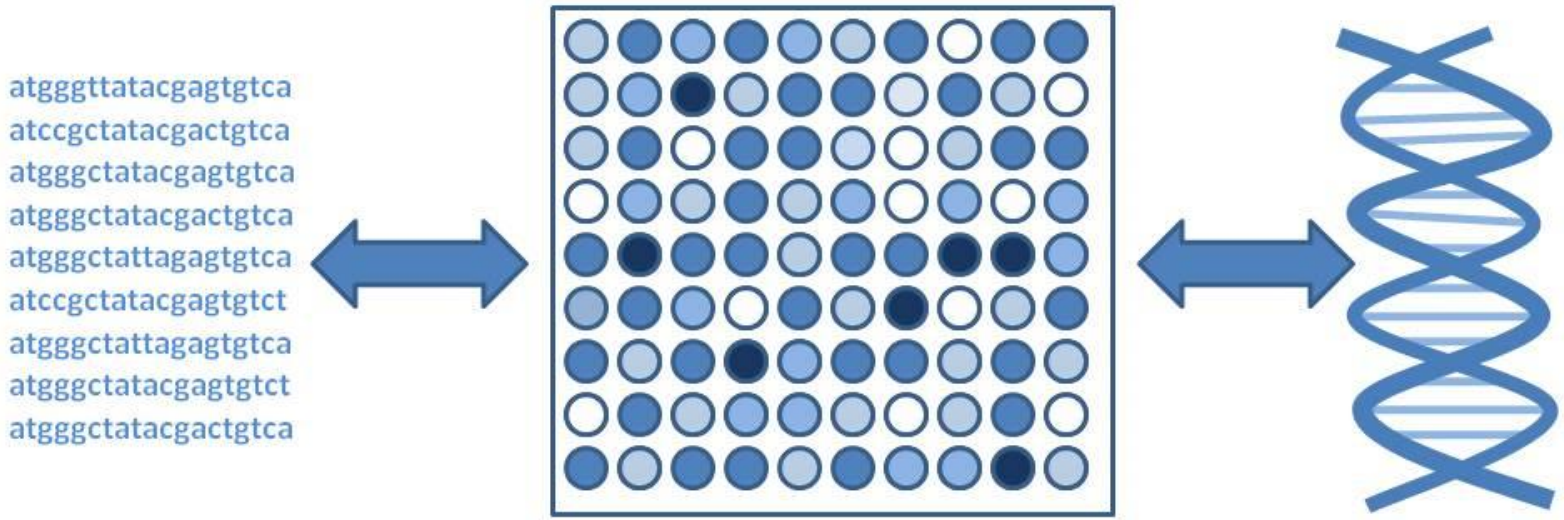
Example



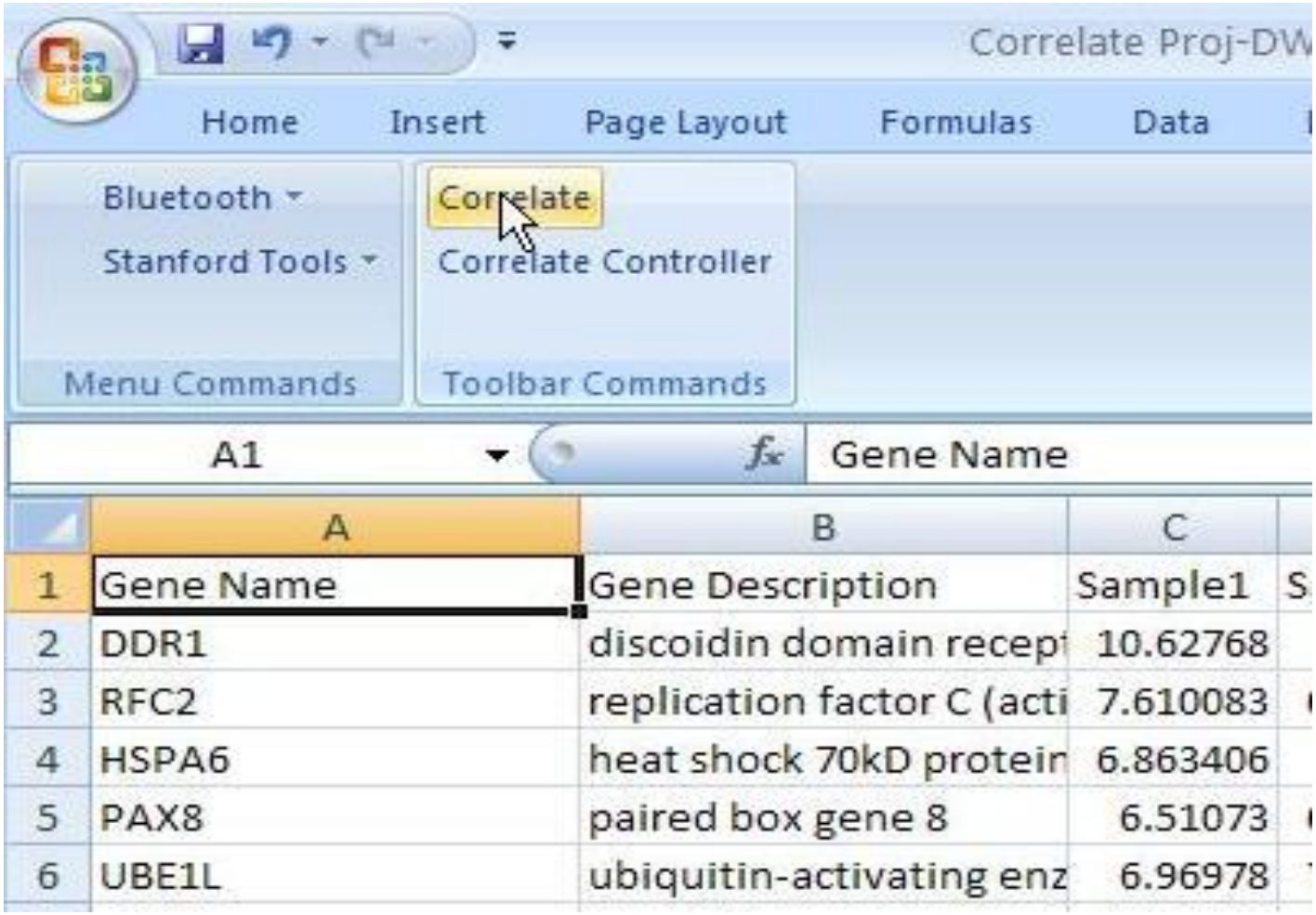
- Copy number data on chromosome 20
- Gene expression data from all chromosomes

- Can we find a region of copy number change on chromosome 20 that's correlated with the expression of a set of genes?

Correlate

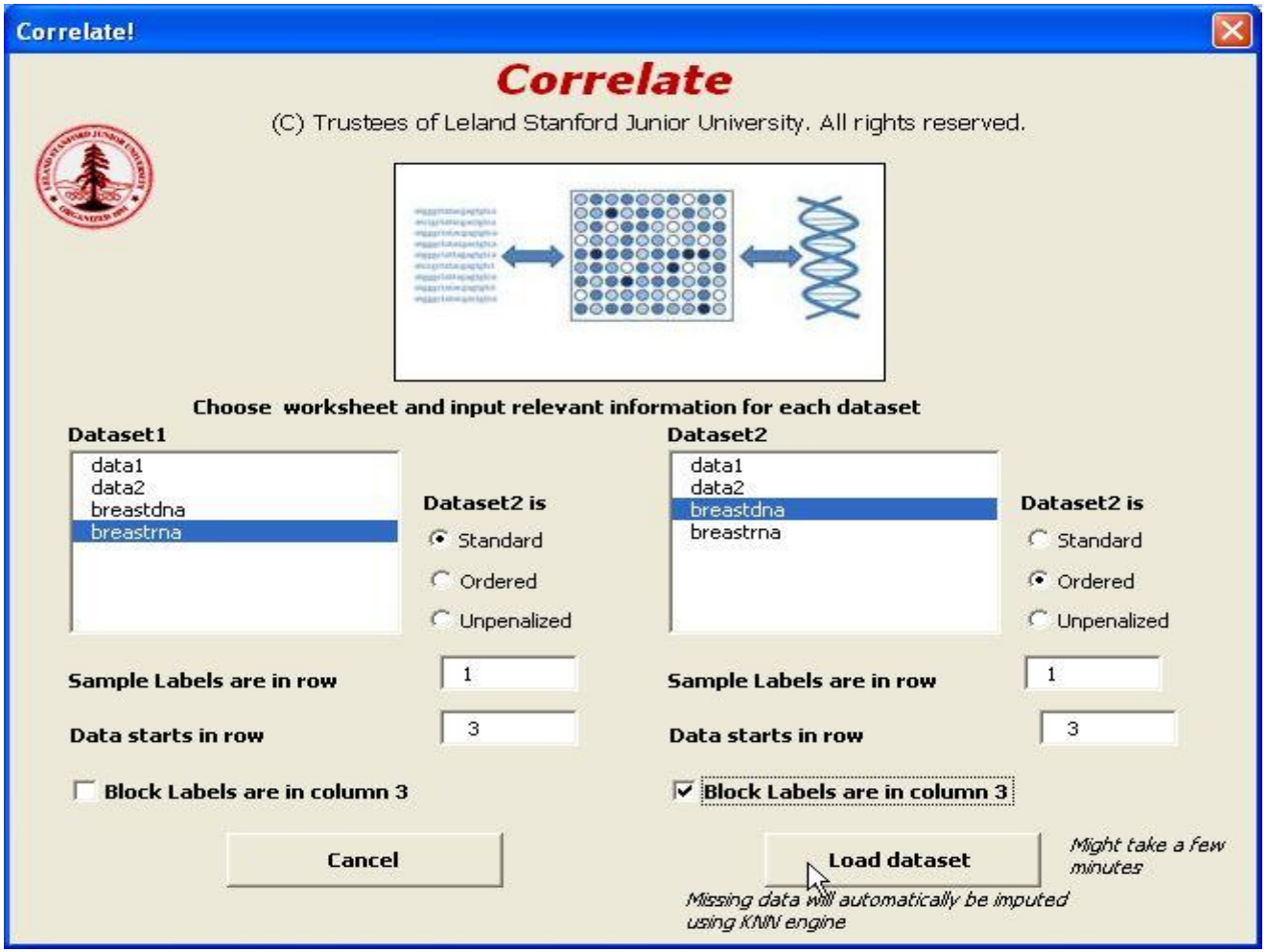

atgggttatacgagtgtca
atccgctatacgactgtca
atgggctatacgagtgtca
atgggctatacgactgtca
atgggctattagagtgtca
atccgctatacgagtgtct
atgggctattagagtgtca
atgggctatacgagtgtct
atgggctatacgactgtca

Correlate

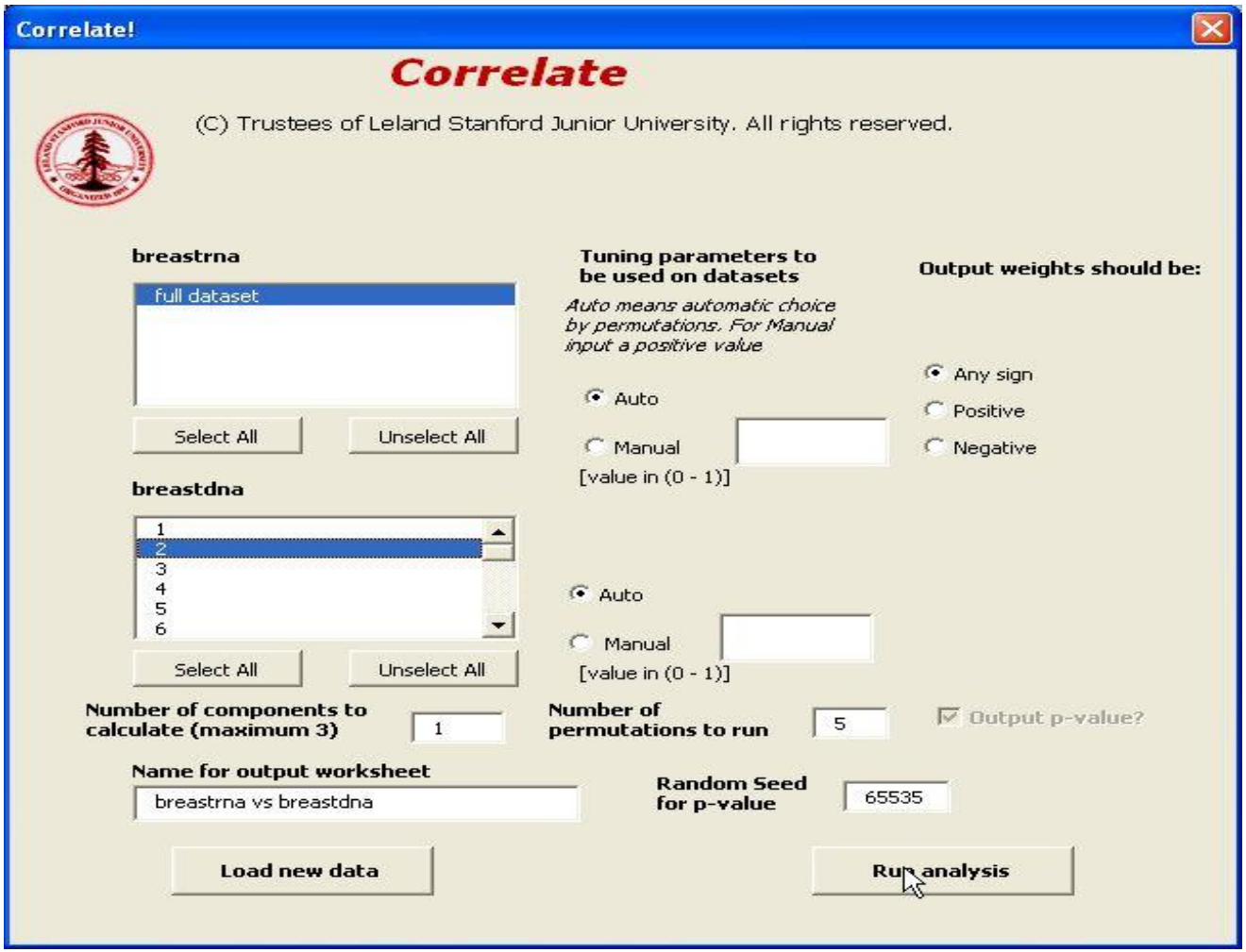


	A	B	C
1	Gene Name	Gene Description	Sample1
2	DDR1	discoidin domain recept	10.62768
3	RFC2	replication factor C (acti	7.610083
4	HSPA6	heat shock 70kD protein	6.863406
5	PAX8	paired box gene 8	6.51073
6	UBE1L	ubiquitin-activating enz	6.96978

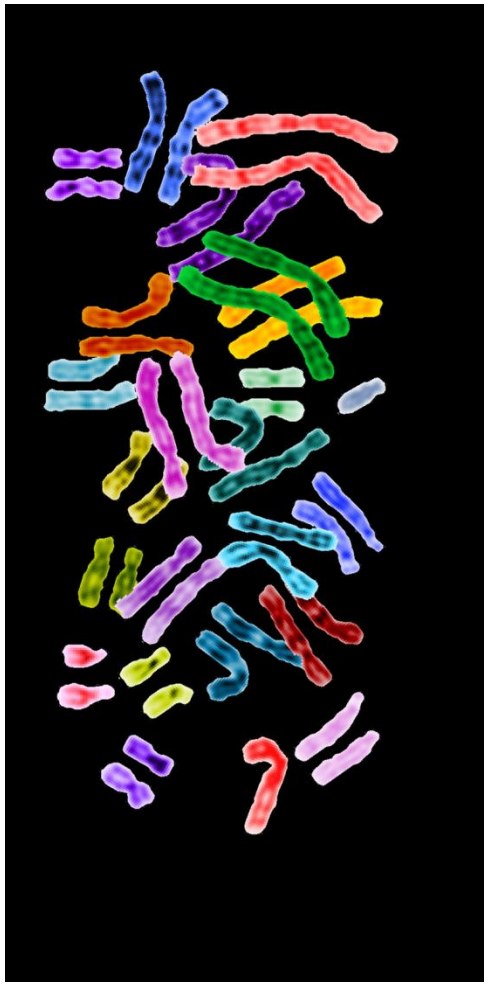
Correlate


Correlate!
Correlate
(C) Trustees of Leland Stanford Junior University. All rights reserved.
Choose worksheet and input relevant information for each dataset
Dataset1
data1
data2
breastdna
breastrna
Dataset2 is
Standard
Ordered
Unpenalized
Sample Labels are in row
1
Data starts in row
3
Block Labels are in column 3
Dataset2
data1
data2
breastdna
breastrna
Dataset2 is
Standard
Ordered
Unpenalized
Sample Labels are in row
1
Data starts in row
3
Block Labels are in column 3
Cancel
Load dataset
Might take a few minutes
Missing data will automatically be imputed using KNN engine

Correlate


Correlate!
Correlate
(C) Trustees of Leland Stanford Junior University. All rights reserved.
breastrna
full dataset
Select All
Unselect All
breastdna
1
2
3
4
5
6
Select All
Unselect All
Tuning parameters to be used on datasets
Auto means automatic choice by permutations. For Manual input a positive value
Auto
Manual
[value in (0 - 1)]
Auto
Manual
[value in (0 - 1)]
Output weights should be:
Any sign
Positive
Negative
Number of components to calculate (maximum 3)
1
Number of permutations to run
5
Output p-value?
Name for output worksheet
breastrna vs breastdna
Random Seed for p-value
65535
Load new data
Run analysis

Example



- Copy number data on chromosome 20
- Gene expression data from all chromosomes

- Can we find a region of copy number change on chromosome 20 that's correlated with the expression of a set of genes?

Correlate - chromosome 20

	A	B	C	D	E	F	G	H	I	J
1	Correlate	Output								
2										
3		dataset1	dataset2					summary statistics		
4	name	breastcgh	breastexp		# of sampl	111			component 1	
5	type	ordered	standard		K	1		dataset1 r	54	
6	penalty	0.068514	11.97554		random se	65535		dataset2 r	316	
7	output co	any sign	any sign		p-value	0		cors	0.900448	
8	# of rows	111	19672							
9										
10										
11										
12	first dataset output					second dataset output				
13	labels			weights		labels			weights	
14	1922	127	20	0		discoidin	DDR1	6	0	
15	1923	307	20	0		ribosomal	RPL10A	7	0	
16	1924	474	20	0		HLA-B ass	BAT1	1	0	
17	1925	614.133	20	0		ATP-bindi	ABCF1	2	0	
18	1926	1263.145	20	0		heat shocl	HSPCB	3	0	
19	1927	1972	20	0		ribosomal	RPS10	17	0	
20	1928	2535	20	0		desmopla	DSP	14	0	
21	1929	2898.852	20	0		glyoxalase	GLO1	17	0	
22	1930	3123	20	0		protein ty	PTP4A1	10	0	

Correlate - chromosome 20

	A	B	C	D	E	F	G	H	I	J
1	Correlate	Output								
2										
3		dataset1	dataset2					summary statistics		
4	name	breastcgh	breastexp		# of sampl	111			component 1	
5	type	ordered	standard		K	1		dataset1 r	54	
6	penalty	0.068514	11.97554		random se	65535		dataset2 r	316	
7	output co	any sign	any sign		p-value	0		cors	0.900448	
8	# of rows	111	19672							
9										
10										
11										
12	first dataset output					second dataset output				
13	labels			weights		labels			weights	
14	1922	127	20	0		discoidin	DDR1	6	0	
15	1923	307	20	0		ribosomal	RPL10A	7	0	
16	1924	474	20	0		HLA-B ass	BAT1	1	0	
17	1925	614.133	20	0		ATP-bindi	ABCF1	2	0	
18	1926	1263.145	20	0		heat shocl	HSPCB	3	0	
19	1927	1972	20	0		ribosomal	RPS10	17	0	
20	1928	2535	20	0		desmopla	DSP	14	0	
21	1929	2898.852	20	0		glyoxalase	GLO1	17	0	
22	1930	3123	20	0		protein ty	PTP4A1	10	0	

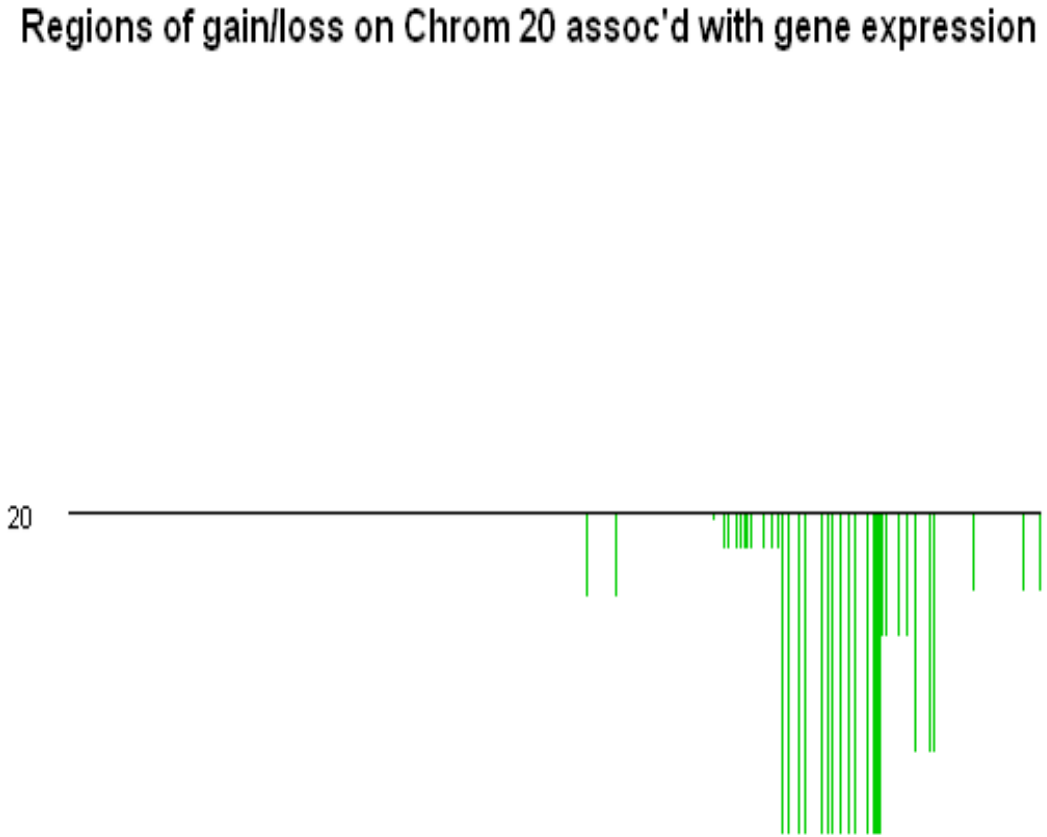
Correlate - chromosome 20

	A	B	C	D	E	F	G	H	I	J
1	Correlate	Output								
2										
3		dataset1	dataset2					summary statistics		
4	name	breastcgh	breastexp		# of sampl	111			component 1	
5	type	ordered	standard		K	1		dataset1 r	54	
6	penalty	0.068514	11.97554		random se	65535		dataset2 r	316	
7	output co	any sign	any sign		p-value	0		cors	0.900448	
8	# of rows	111	19672							
9										
10										
11										
12	first dataset output					second dataset output				
13	labels			weights		labels			weights	
14	1922	127	20	0		discoidin	DDR1	6	0	
15	1923	307	20	0		ribosomal	RPL10A	7	0	
16	1924	474	20	0		HLA-B ass	BAT1	1	0	
17	1925	614.133	20	0		ATP-bindi	ABCF1	2	0	
18	1926	1263.145	20	0		heat shoc	HSPCB	3	0	
19	1927	1972	20	0		ribosomal	RPS10	17	0	
20	1928	2535	20	0		desmopla	DSP	14	0	
21	1929	2898.852	20	0		glyoxalase	GLO1	17	0	
22	1930	3123	20	0		protein ty	PTP4A1	10	0	

Correlate - chromosome 20

	A	B	C	D	E	F	G	H	I	J
1	Correlate	Output								
2										
3		dataset1	dataset2					summary statistics		
4	name	breastcgh	breastexp		# of sampl	111			component 1	
5	type	ordered	standard		K	1		dataset1 r	54	
6	penalty	0.068514	11.97554		random se	65535		dataset2 r	316	
7	output co	any sign	any sign		p-value	0		cors	0.900448	
8	# of rows	111	19672							
9										
10										
11										
12	first dataset output					second dataset output				
13	labels			weights		labels			weights	
14	1922	127	20	0		discoidin	DDR1	6	0	
15	1923	307	20	0		ribosomal	RPL10A	7	0	
16	1924	474	20	0		HLA-B ass	BAT1	1	0	
17	1925	614.133	20	0		ATP-bindi	ABCF1	2	0	
18	1926	1263.145	20	0		heat shocl	HSPCB	3	0	
19	1927	1972	20	0		ribosomal	RPS10	17	0	
20	1928	2535	20	0		desmopla	DSP	14	0	
21	1929	2898.852	20	0		glyoxalase	GLO1	17	0	
22	1930	3123	20	0		protein ty	PTP4A1	10	0	

Correlate - chromosome 20



Correlate - chromosome 20

Non-zero gene expression weights by chromosome

1	2	3	4	5	6	7	8	9	10	11
9	10	5	7	5	16	8	2	6	6	23

12	13	14	15	16	17	19	20	21	22	23
22	3	4	1	13	45	7	116	1	3	2

Correlate - chromosome 1

	A	B	C	D	E	F	G	H	I	J
1	Correlate! Output									
2										
3		dataset1	dataset2					summary statistics		
4	name	breastcgh	breastexp		# of sampl	136			component 1	
5	type	ordered	standard		K	1		dataset1 r	72	
6	penalty	0.03178	5		random se	65535		dataset2 r	44	
7	output co	any sign	any sign		p-value	0		cors	0.820605	
8	# of rows	136	19672							
9										
10										
11										
12	first dataset output					second dataset output				
13	labels			weights		labels			weights	
14	1	5918.606	1	0		translocat	TPR	1	-0.166	
15	2	6069	1	0		protoporp	PPOX	1	-0.11293	
16	3	6817	1	0		tuftelin 1	TUFT1	1	-0.10075	
17	4	7827.384	1	0		hypotheti	FLJ10359	1	-0.01374	
18	5	9421.348	1	-0.05214		hypotheti	FLJ12528	1	-0.00067	
19	6	10284	1	0		HLA-B ass	BAT1	1	0	
20	7	12042	1	0		major hist	HLA-DPB1	1	0	
21	8	13349	1	0		S-adenosy	AMD1	1	0	
22	9	14391	1	0		MCM3 mi	MCM3	1	0	

Correlate - chromosome 1

- All 44 non-zero gene expression weights are on chromosome 1
- Top 10:
  - splicing factor 3b, subunit 4, 49kD
  - HSPC003 protein
  - rab3 GTPase-activating protein, non-catalytic subunit (150kD)
  - hypothetical protein My014
  - UDP-Gal:betaGlcNAc beta 1,4- galactosyltransferase, polypeptide 3
  - glyceronephosphate O-acyltransferase
  - NADH dehydrogenase (ubiquinone) Fe-S protein 2 (49kD) (NADH-coenzyme Q reductase)
  - hypothetical protein FLJ12671
  - mitochondrial ribosomal protein L24
  - CGI-78 protein

Correlate – Conclusions

- Can be applied to any pair of data sets: SNP, methylation, microRNA expression data, and more….

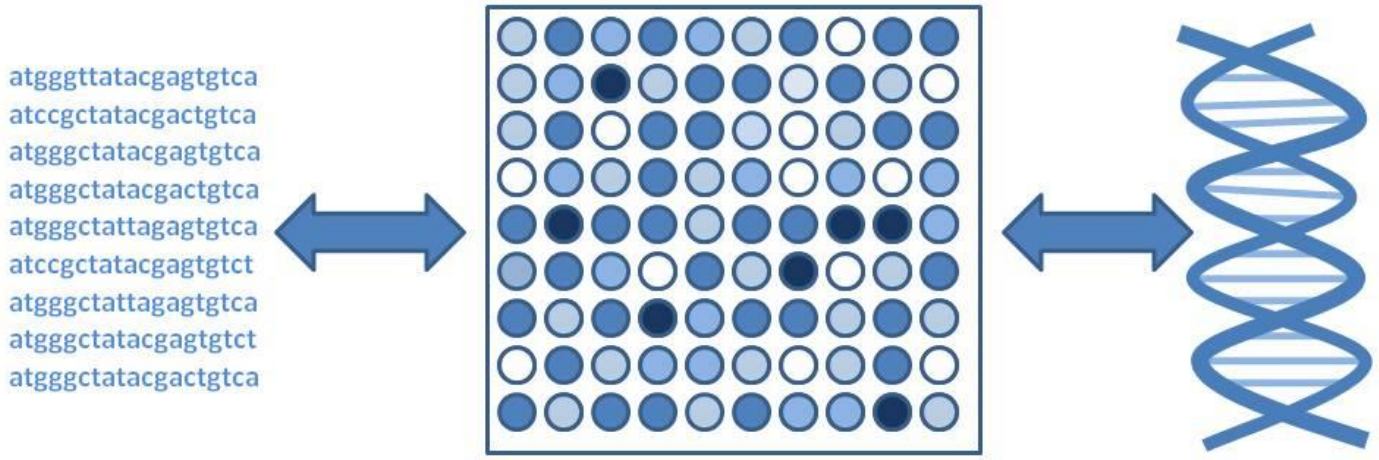
Correlate – Conclusions

- Can be applied to any pair of data sets: SNP, methylation, microRNA expression data, and more….
- Think broadly… a collaborator is using it to correlate image data and gene expression data in cancer. Linear combination of image features is highly predictive of survival!

Correlate – Conclusions

- Can be applied to any pair of data sets: SNP, methylation, microRNA expression data, and more….
- Think broadly… a collaborator is using it to correlate image data and gene expression data in cancer. Linear combination of image features is highly predictive of survival!
- A principled way to discover associations and perform an integrative analysis of two data sets.

Try it out!

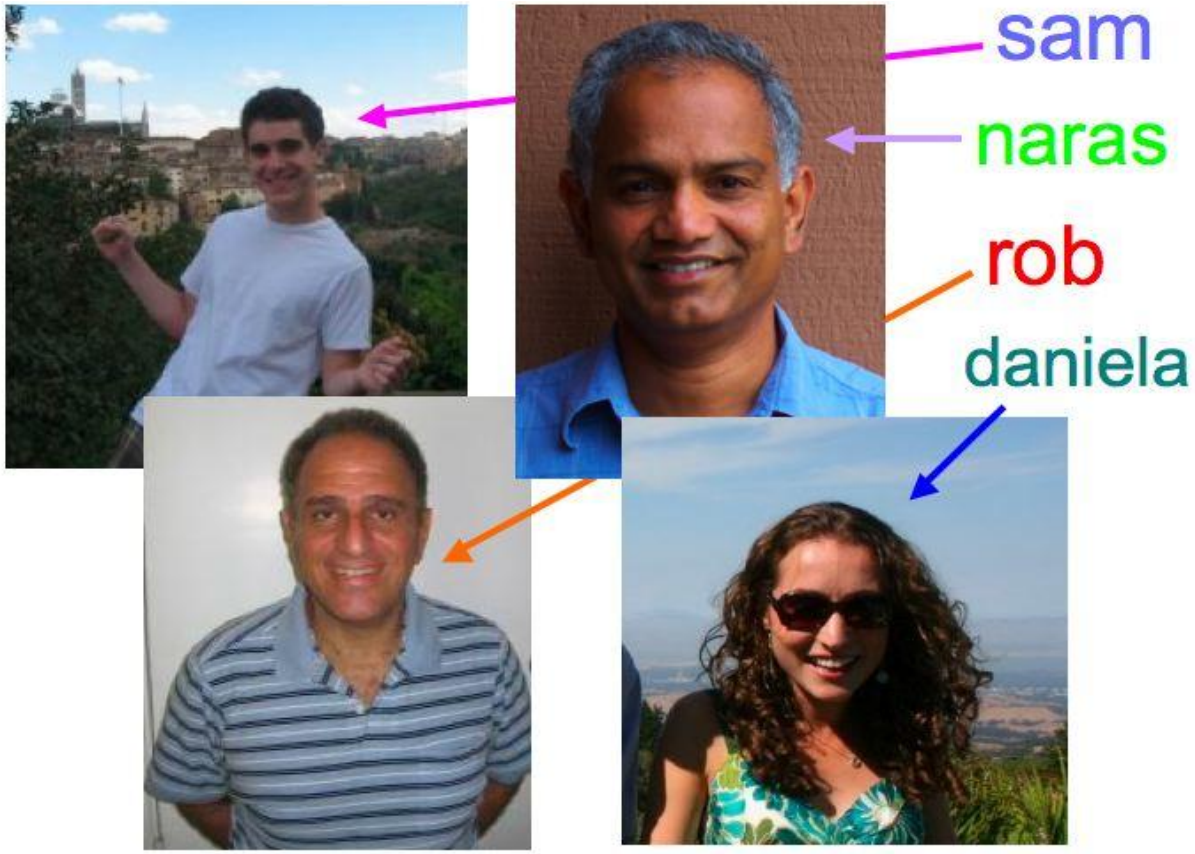


http://www-stat.stanford.edu/~tibs/Correlate/

Or google “Tibshirani”

Or, for R users: package PMA on CRAN

Acknowledgments



Sam Gross (Harvard),
Balasubramanian Narasimhan (Stanford),
and Robert Tibshirani (Stanford)

References

- Witten DM, Tibshirani R, and T Hastie (2009) A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* **10(3):** 515-534.

- Witten DM and R Tibshirani (2009) Extensions of sparse canonical correlation analysis, with applications to genomic data. *Statistical Applications in Genetics and Molecular Biology* **8(1):** Article 28.