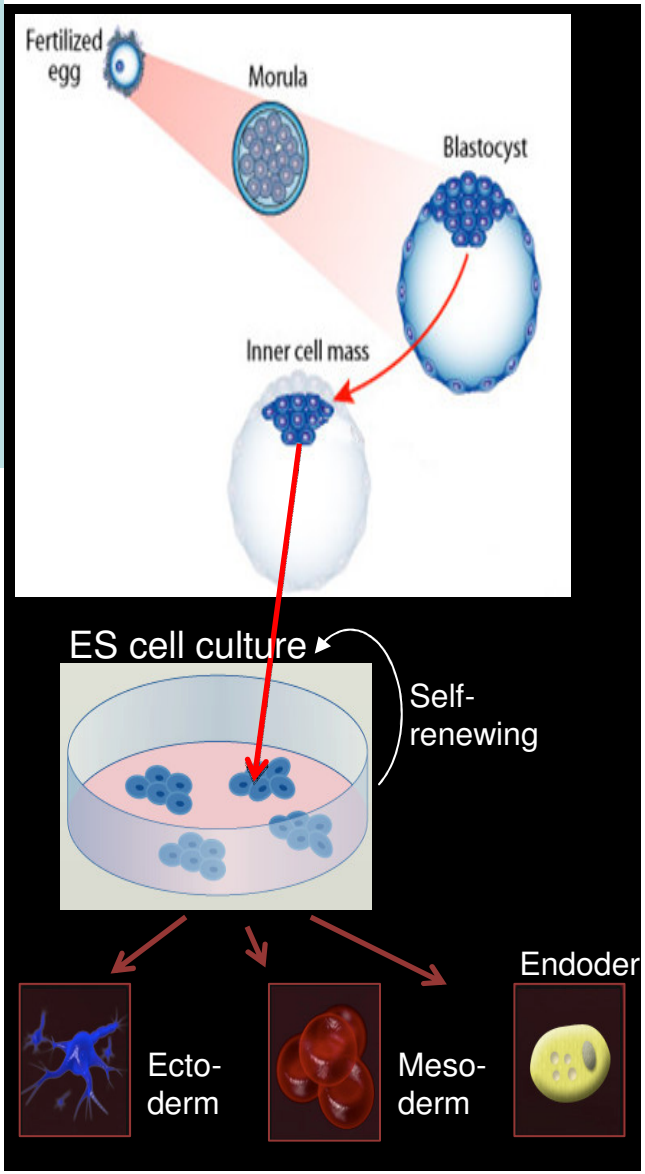
Signed weighted gene co-expression network analysis of transcriptional regulation in murine embryonic stem cells

Steve Horvath
University of California, Los Angeles

Acknowledgement:
Dissertation work of Mike J Mason
Guoping Fan, Kathrin Plath, Qing Zhou



Contents

- Weighted Gene Co-Expression Network Analysis
- Application to stem cell data

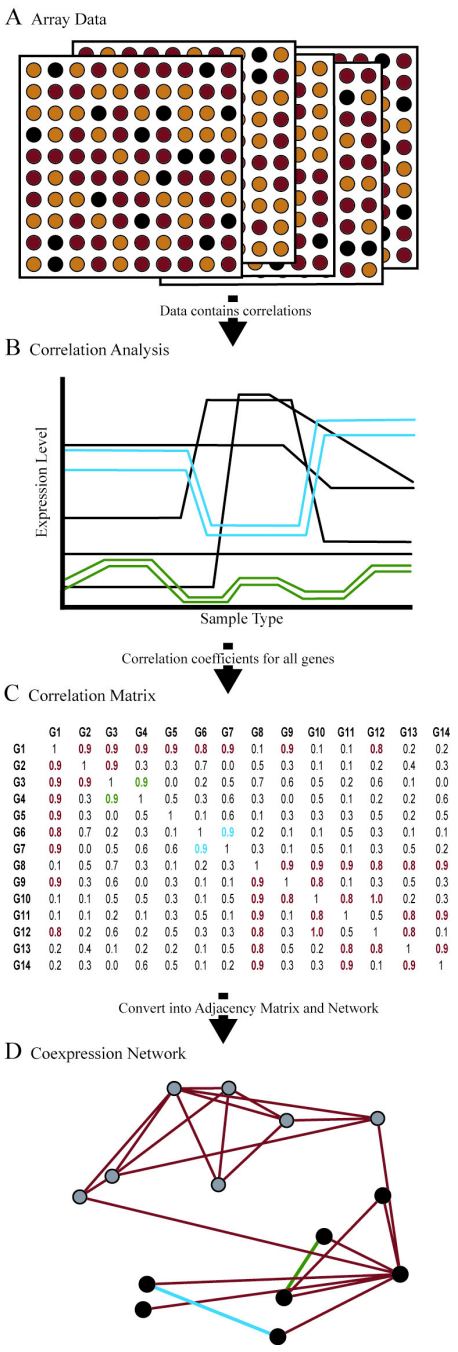
How to construct a weighted gene co-expression network?

Bin Zhang and Steve Horvath (2005) "A General Framework for Weighted Gene Co-Expression Network Analysis", Statistical Applications in Genetics and Molecular Biology: Vol. 4: No. 1

Undirected Network =Adjacency Matrix

- A network can be represented by an adjacency matrix, $A=[a_{ij}]$, that encodes whether/how a pair of nodes is connected.
  - A is a symmetric matrix with entries in [0,1]
  - For unweighted network, entries are 1 or 0 depending on whether or not 2 nodes are adjacent (connected)
  - For weighted networks, the adjacency matrix reports the connection strength between gene pairs

Figure 1



	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	G11	G12	G13	G14
G1	1	0.9	0.9	0.9	0.9	0.8	0.9	0.1	0.9	0.1	0.1	0.8	0.2	0.2
G2	0.9	1	0.9	0.3	0.3	0.7	0.0	0.5	0.3	0.1	0.1	0.2	0.4	0.3
G3	0.9	0.9	1	0.9	0.0	0.2	0.5	0.7	0.6	0.5	0.2	0.6	0.1	0.0
G4	0.9	0.3	0.9	1	0.5	0.3	0.6	0.3	0.0	0.5	0.1	0.2	0.2	0.6
G5	0.9	0.3	0.0	0.5	1	0.1	0.6	0.1	0.3	0.3	0.3	0.5	0.2	0.5
G6	0.8	0.7	0.2	0.3	0.1	1	0.9	0.2	0.1	0.1	0.5	0.3	0.1	0.1
G7	0.9	0.0	0.5	0.6	0.6	0.9	1	0.3	0.1	0.5	0.1	0.3	0.5	0.2
G8	0.1	0.5	0.7	0.3	0.1	0.2	0.3	1	0.9	0.9	0.9	0.8	0.8	0.9
G9	0.9	0.3	0.6	0.0	0.3	0.1	0.1	0.9	1	0.8	0.1	0.3	0.5	0.3
G10	0.1	0.1	0.5	0.5	0.3	0.1	0.5	0.9	0.8	1	0.8	1.0	0.2	0.3
G11	0.1	0.1	0.2	0.1	0.3	0.5	0.1	0.9	0.1	0.8	1	0.5	0.8	0.9
G12	0.8	0.2	0.6	0.2	0.5	0.3	0.3	0.8	0.3	1.0	0.5	1	0.8	0.1
G13	0.2	0.4	0.1	0.2	0.2	0.1	0.5	0.8	0.5	0.2	0.8	0.8	1	0.9
G14	0.2	0.3	0.0	0.6	0.5	0.1	0.2	0.9	0.3	0.3	0.9	0.1	0.9	1

Steps for constructing a co-expression network

A) Gene expression data

B) Measure concordance of gene expression with a Pearson correlation

C) The Pearson correlation matrix is either dichotomized to arrive at an unweighted adjacency matrix → unweighted network

Or transformed continuously with the power adjacency function → weighted network

Power adjacency function for constructing unsigned and signed weighted gene co-expr. networks

Unsigned network, absolute value

$$a_{ij} = | cor(x_i, x_j) |^{\beta}$$

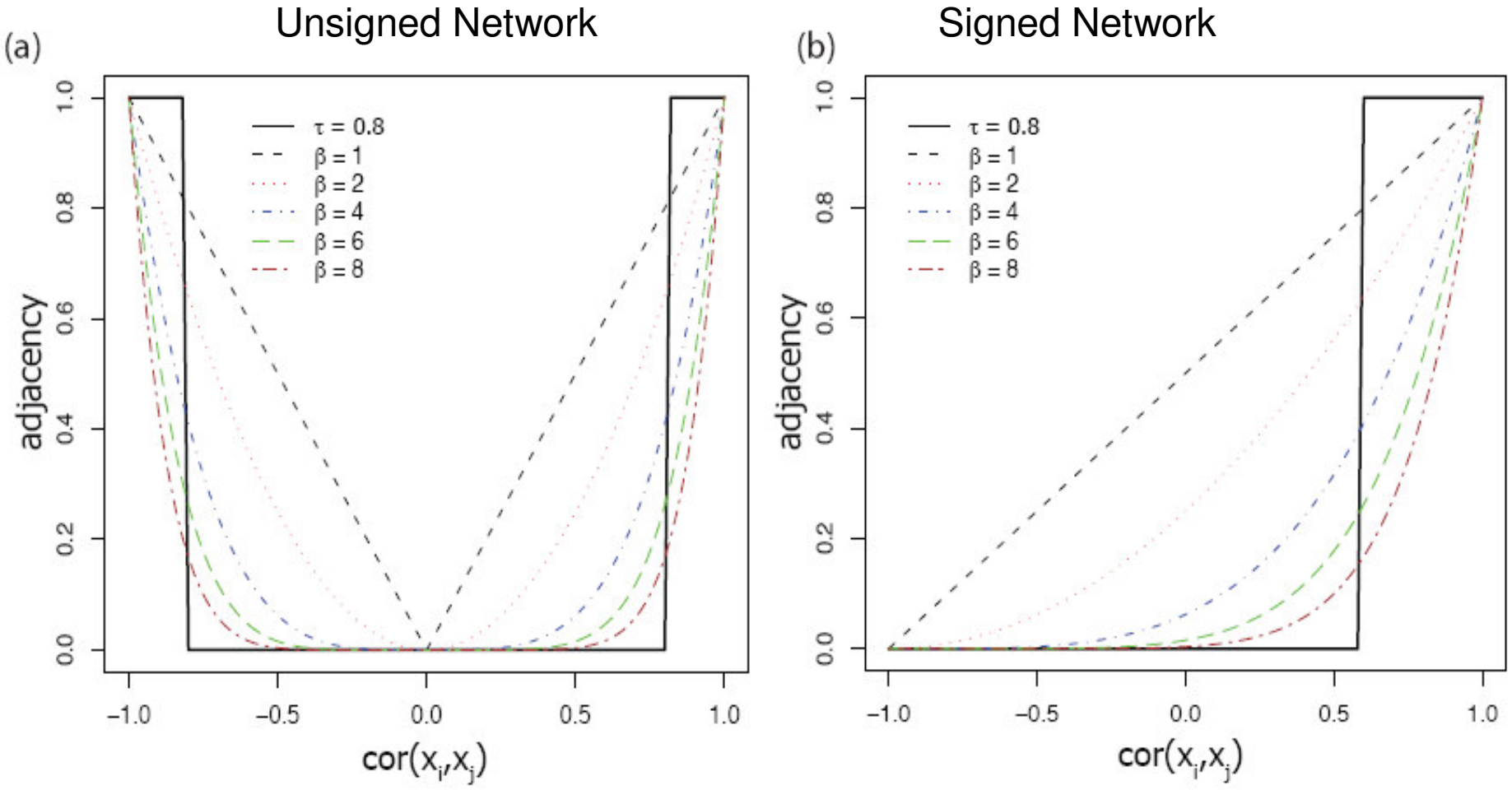
Signed network preserves sign info

$$a_{ij} = | 0.5 + 0.5 \times cor(x_i, x_j) |^{\beta}$$

Default values: beta=6 for unsigned and beta=12 for signed networks.

Alternatively, use the "scale free topology criterion" described in Zhang and Horvath 2005.

Comparing adjacency functions for transforming the correlation into a measure of connection strength



Why soft thresholding as opposed to hard thresholding?

1. Preserves the continuous information of the co-expression information
2. Results tend to be more robust with regard to different threshold choices

But hard thresholding has its own advantages:
In particular, graph theoretic algorithms from the computer science community can be applied to the resulting networks

Question: Are signed correlation networks superior to unsigned networks?

Answer: Overall, recent applications have convinced me that signed networks are preferable.

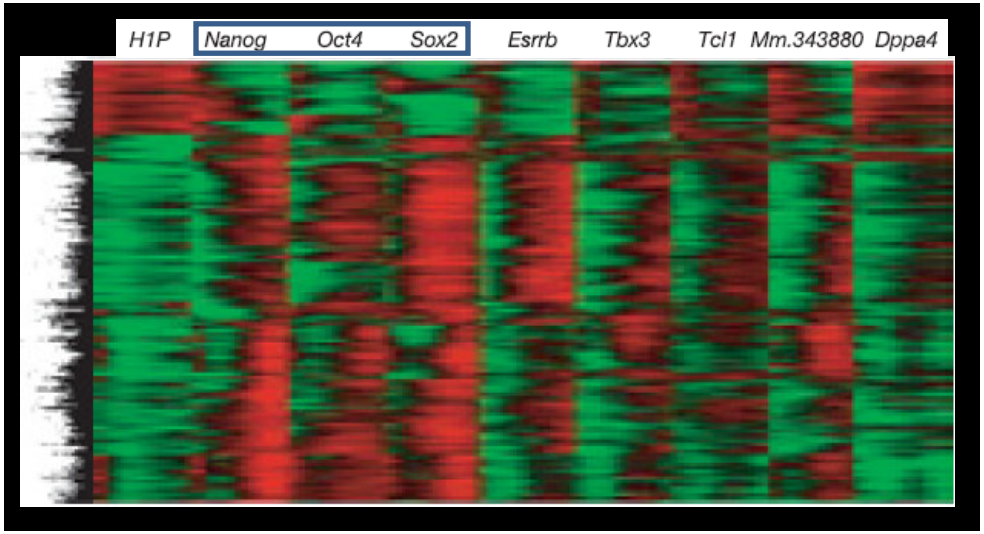
- For example, signed networks were critical in a recent stem cell application
- *Michael J Mason, Kathrin Plath, Qing Zhou, SH (2009) Signed Gene Co-expression Networks for Analyzing Transcriptional Regulation in Murine Embryonic Stem Cells. BMC Genomics 2009, 10:327*

Re-analysis of published microarray data sets

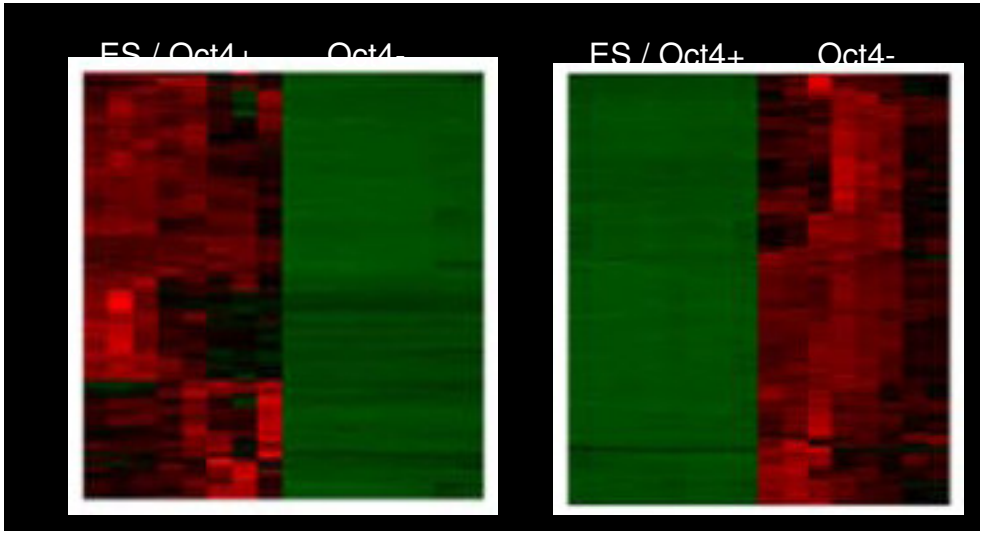
- Ivanova N, Dobrin R, Lu R, Kotenko L, Levorse J, DeCoste C, Schafer X, Lun Y, Lemischka I: Discecting self-renewal in stem cells with RNA interference.Nature 2006, 442:533-538
- Zhou Q, Chipperfield H, Melton DA, Wong WH: A gene regulatrory network in mouse embryonic stem cells. Proc Natl Acad Sci 2007, 104(42):16438-16443.

ES Cell Datasets Used

- Ivanova *et al.*: RNA knockdown of 8 TFs thought to play a role in pluripotency



- Zhou *et al.*: ES cell samples and differentiated cell samples sorted into Oct4 positive and negative groups

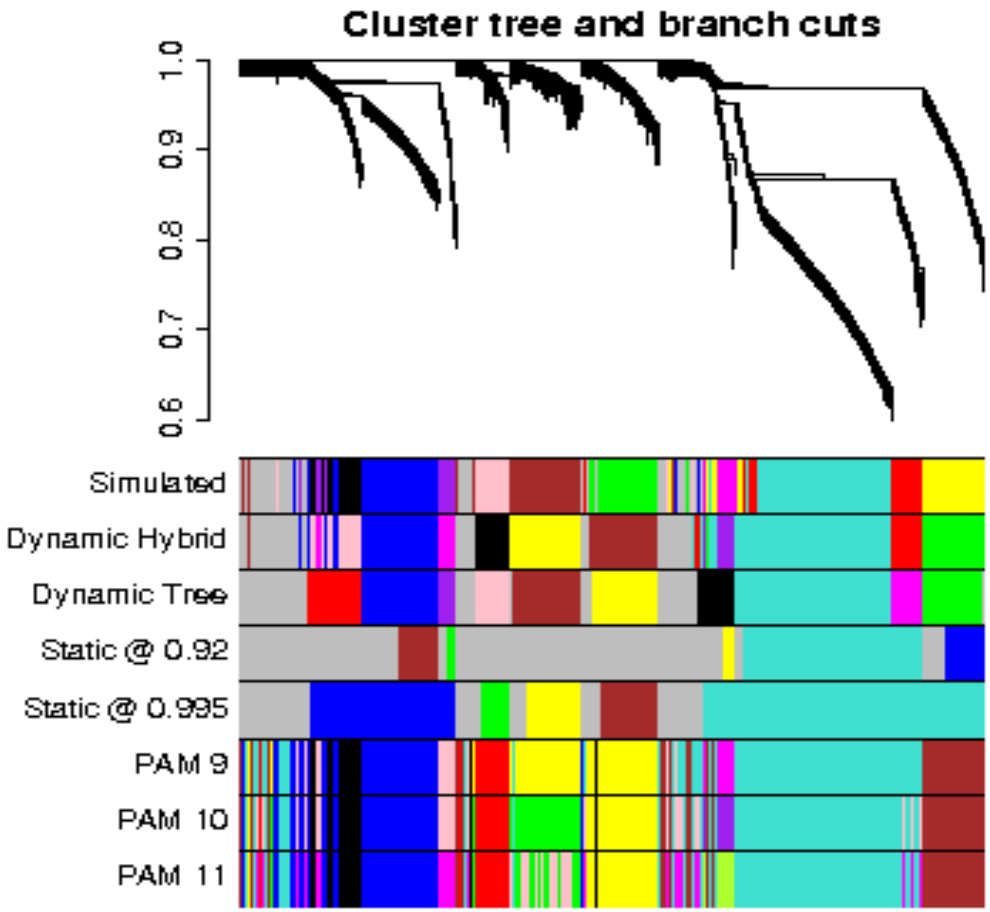


How to detect network modules?

As default, we define modules as branches of a cluster tree

- We use average linkage hierarchical clustering which inputs a measure of interconnectedness
  - often the topological overlap measure
- Once a dendrogram is obtained from a hierarchical clustering method, we define modules as branches using a branch cutting method
  - dynamicTreeCut R package (Peter Langfelder et al 2007)

How to cut branches off a tree?



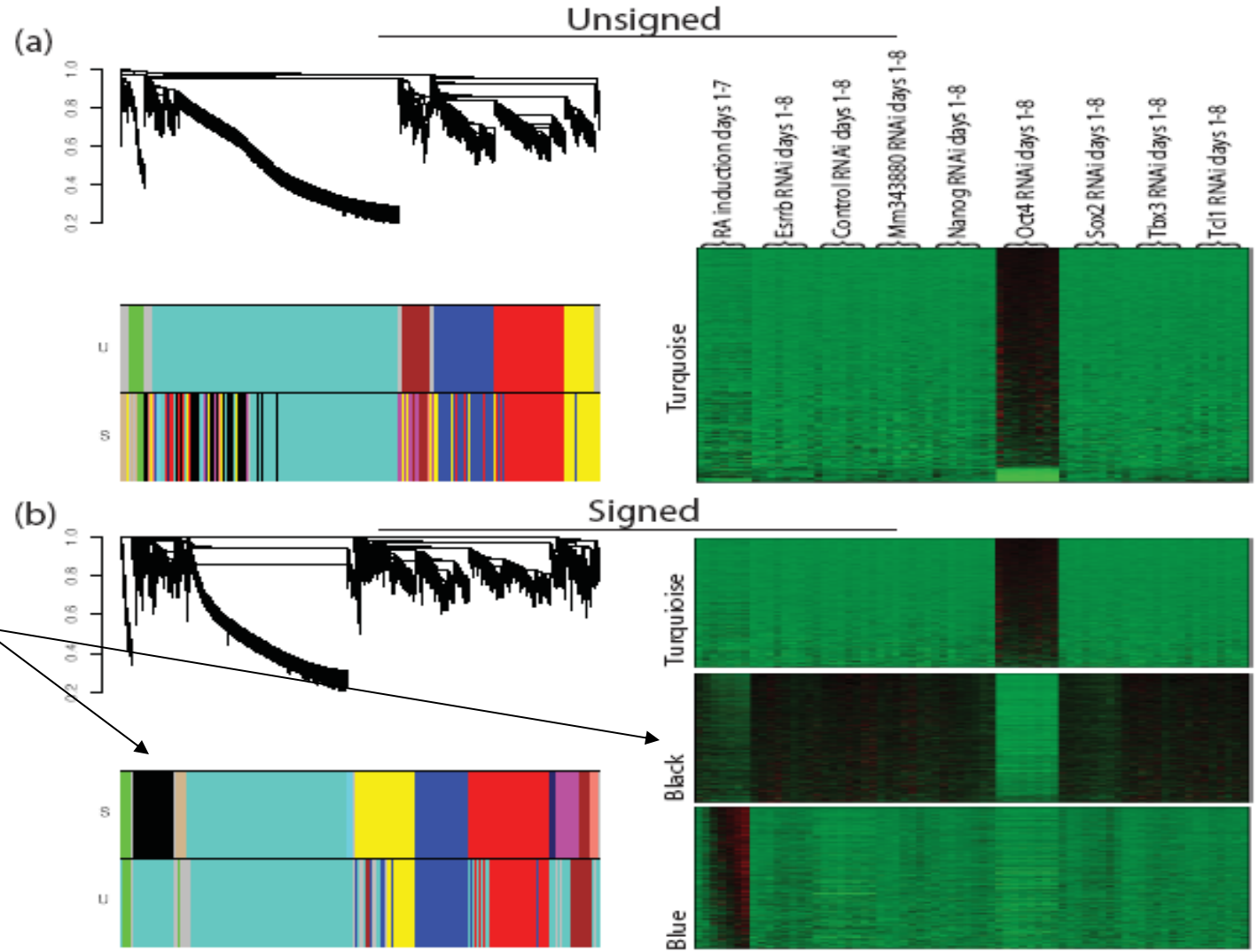
Module=branch of a cluster tree

Module genes are assigned the same color

Bioinformatics 2008 ***24****(5):719-720*

Signed WGCNA finds a pluripotency related module, which cannot be found in an unsigned network analysis

Pluripotency module



Question: How does one summarize the expression profiles in a module?

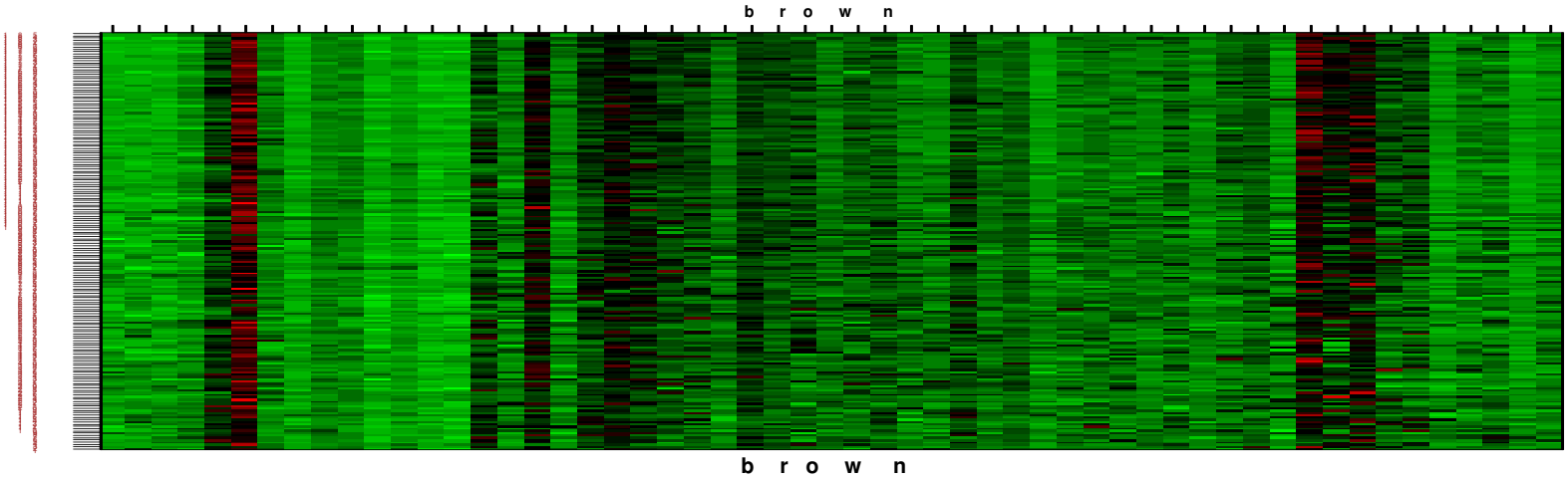
Math answer: module eigengene = first principal component

Network answer: the most highly connected intramodular hub gene

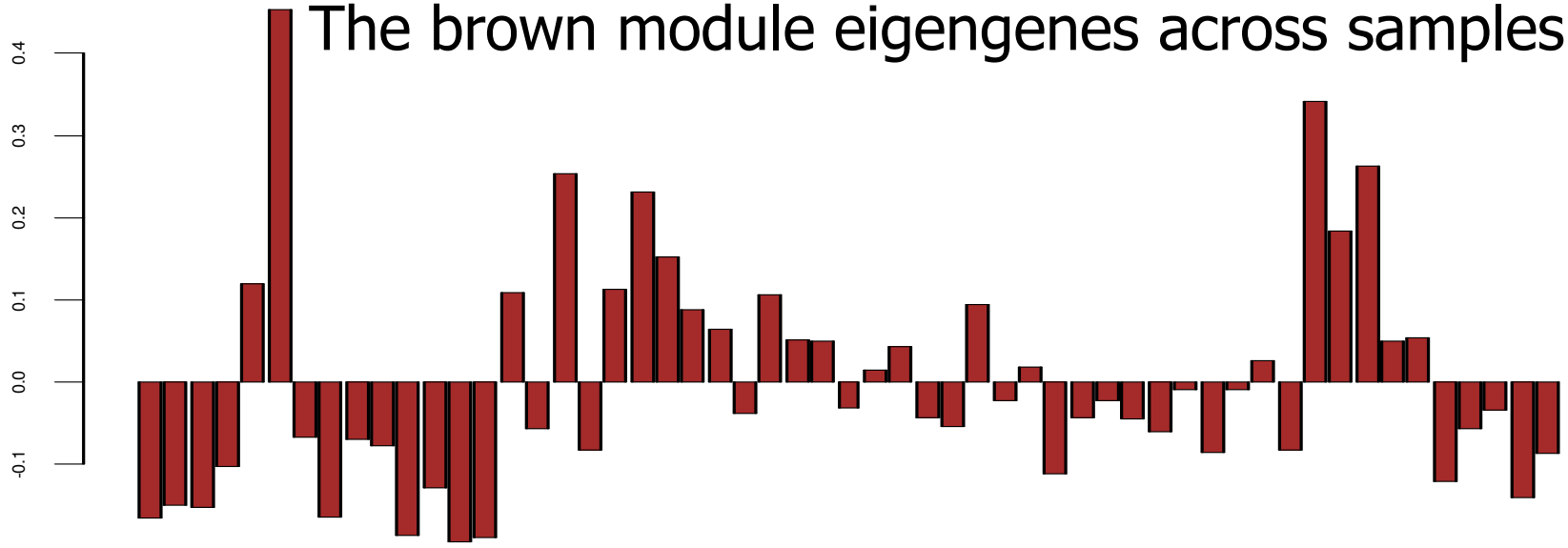
Both turn out to be equivalent

Module Eigengene= measure of over-expression=average redness

Rows,=genes, Columns=microarray



The brown module eigengenes across samples



Eigengene-based connectivity, also known as kME or module membership measure

$$k_{ME,i} = ModuleMembership(i) = cor(x_i, ME)$$

kME(i) is simply the correlation between the i-th gene expression profile and the module eigengene.

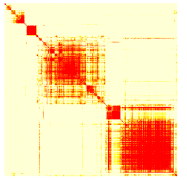
Very useful measure for annotating genes with regard to modules.

Module eigengene turns out to be the most highly connected gene

What is weighted gene co-expression network analysis?

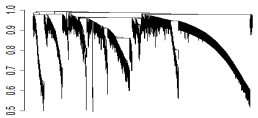
Construct a network

Rationale: make use of interaction patterns between genes



Identify modules

Rationale: module (pathway) based analysis

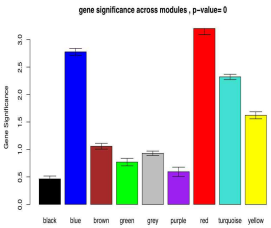


Relate modules to external information

Array Information: RNAi knock-out

Gene Information: gene ontology, DNA binding data, epigenetic

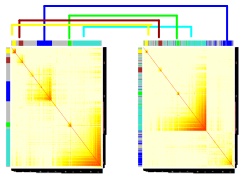
Rationale: find biologically interesting modules



Study Module Preservation across different data

Rationale:

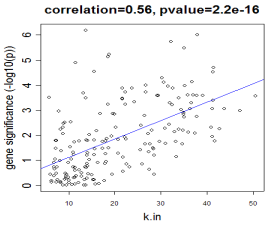
- Same data: to check robustness of module definition
- Example Ivanova versus Zhou data



Find the key drivers in *interesting* modules

Tools: intramodular connectivity kME

Rationale: experimental validation, novel genes



What is different from other analyses?

- **Emphasis on modules (pathways) instead of individual genes**
  - Greatly alleviates the problem of multiple comparisons
    - Less than 20 comparisons versus 20000 comparisons
- Use of intramodular connectivity kME to find key drivers
  - Quantifies module membership (centrality)
  - If the module is preserved, intramodular hub genes are preserved as well
- Module definition is based on gene expression data only
  - No prior pathway information is used for module definition
  - Two module (eigengenes) can be highly correlated
  - Typically defined by cutting branches of a cluster tree
- Emphasis on a unified approach for relating variables
  - Default: power of a correlation
- Technical Details: soft thresholding with the power adjacency function, topological overlap matrix to measure interconnectedness

How to relate modules to external data?

Oct4 RNAi knock out status gives rise to a gene significance measure

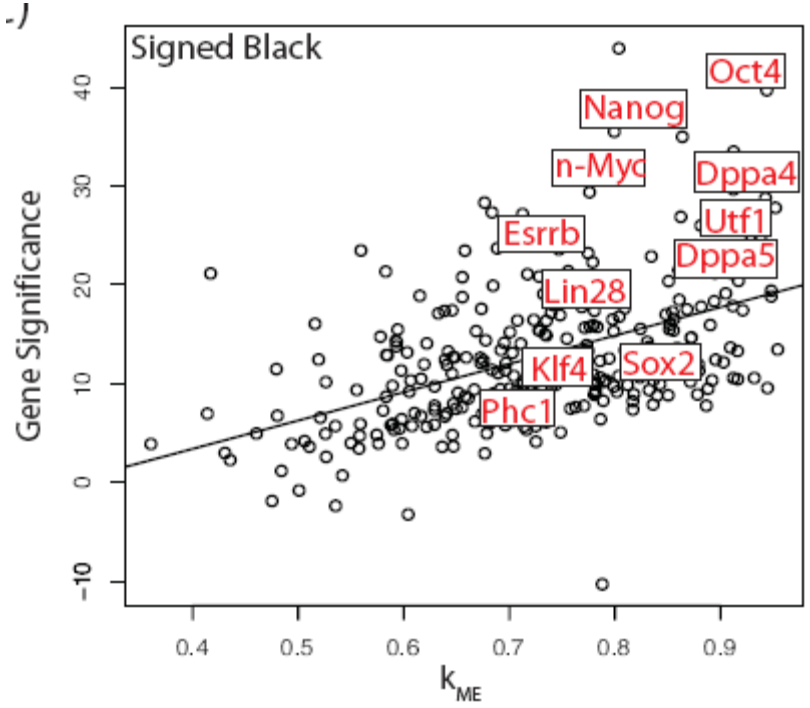
Possible definitions

- We defined a measure of gene significance (GS) as the t-statistic from the paired Student's t-test of expression in control RNAi samples and ES cell samples with RNAi knock down of Oct4 (paired by day of treatment)
- GS could also be a fold change
- GS(i)=|T-test(i)| of differential expression
- GS(i)=-log(p-value)

A gene significance naturally gives rise to a module significance measure

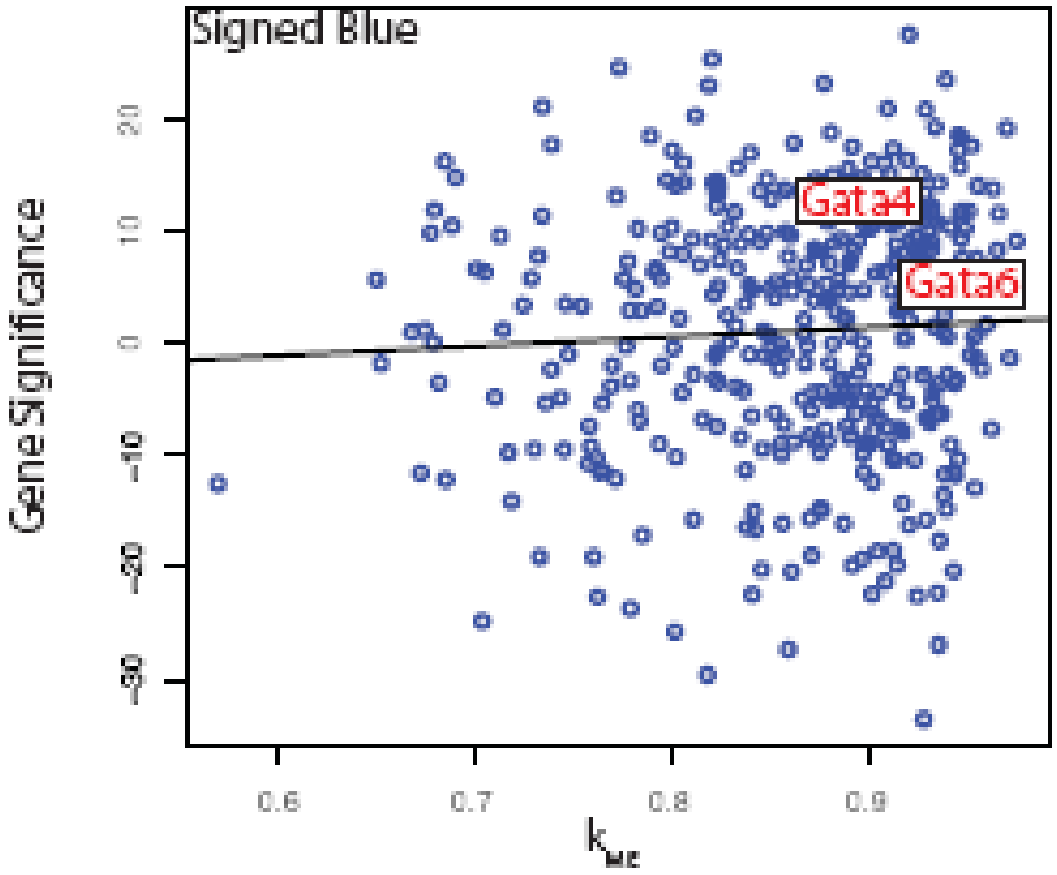
- Define module significance as mean gene significance
- Often highly related to the correlation between module eigengene and trait

The Black Module Contains Genes Involved in Pluripotency

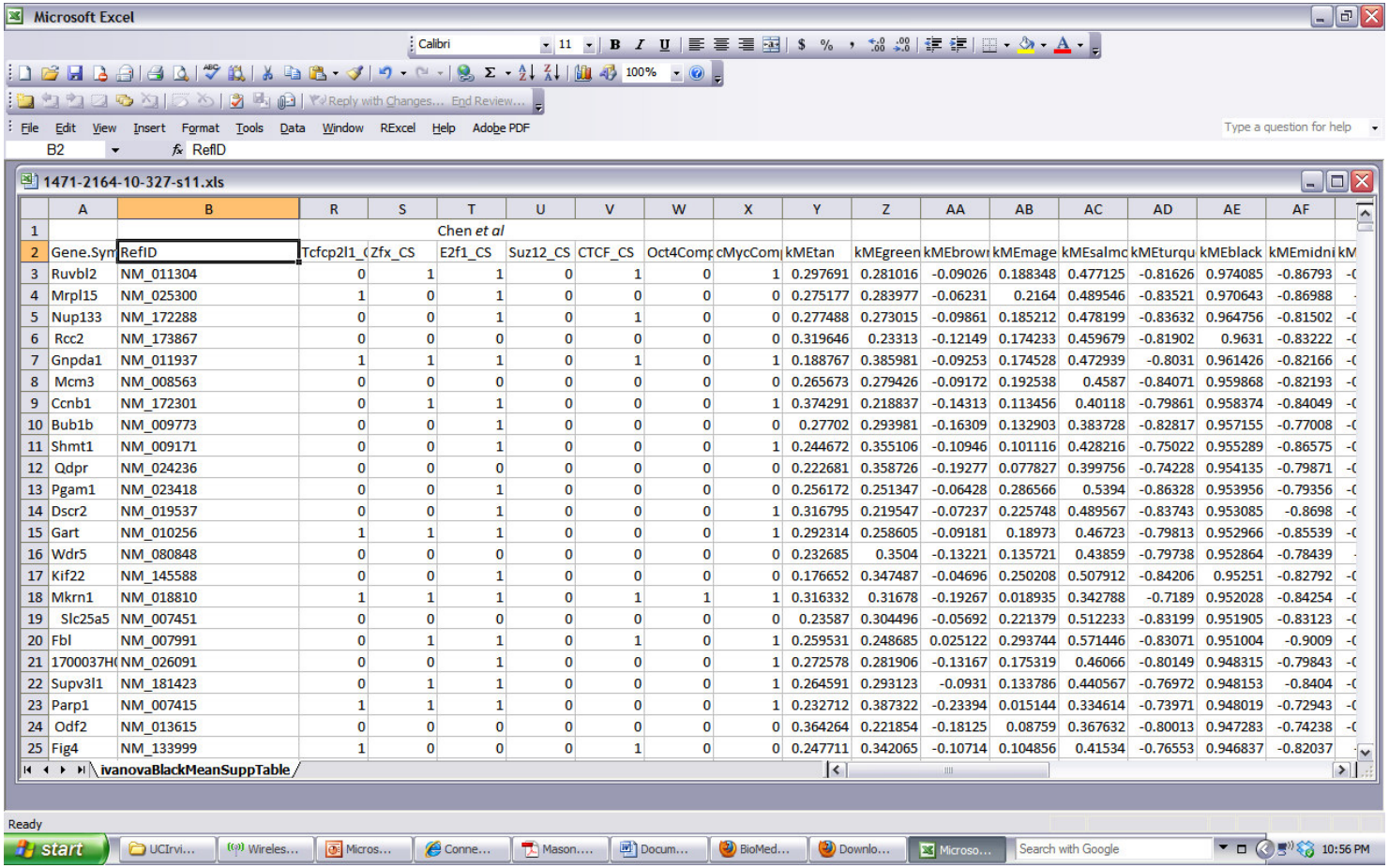


- The genes of this module are significantly more likely to be bound by key regulators of pluripotency and self-renewal

The blue module contains transcription factors involved in differentiation



Module Membership and Binding Information in the Signed Ivanova et al (2006) Network. This file contains module membership, kME, and binding data from Loh et al (2006), Boyer et al (2007), and Chen et al (2008) for each gene on the microarray.



	A	B	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF
1					Chen *et al*												
2	Gene.Sym	RefID	Tcfcp2l1_(	Zfx_CS	E2f1_CS	Suz12_CS	CTCF_CS	Oct4Comp	cMycCom	kMEtan	kMEgreen	kMEbrown	kMEmage	kMEsalmc	kMEturqu	kMEblack	kMEmidni
3	Ruvbl2	NM_011304	0	1	1	0	1	0	1	0.297691	0.281016	-0.09026	0.188348	0.477125	-0.81626	0.974085	-0.86793
4	Mrpl15	NM_025300	1	0	1	0	0	0	0	0.275177	0.283977	-0.06231	0.2164	0.489546	-0.83521	0.970643	-0.86988
5	Nup133	NM_172288	0	0	1	0	1	0	0	0.277488	0.273015	-0.09861	0.185212	0.478199	-0.83632	0.964756	-0.81502
6	Rcc2	NM_173867	0	0	0	0	0	0	0	0.319646	0.23313	-0.12149	0.174233	0.459679	-0.81902	0.9631	-0.83222
7	Gnpda1	NM_011937	1	1	1	0	1	0	1	0.188767	0.385981	-0.09253	0.174528	0.472939	-0.8031	0.961426	-0.82166
8	Mcm3	NM_008563	0	0	0	0	0	0	0	0.265673	0.279426	-0.09172	0.192538	0.4587	-0.84071	0.959868	-0.82193
9	Ccnb1	NM_172301	0	1	1	0	0	0	1	0.374291	0.218837	-0.14313	0.113456	0.40118	-0.79861	0.958374	-0.84049
10	Bub1b	NM_009773	0	0	1	0	0	0	0	0.27702	0.293981	-0.16309	0.132903	0.383728	-0.82817	0.957155	-0.77008
11	Shmt1	NM_009171	0	0	1	0	0	0	1	0.244672	0.355106	-0.10946	0.101116	0.428216	-0.75022	0.955289	-0.86575
12	Qdpr	NM_024236	0	0	0	0	0	0	0	0.222681	0.358726	-0.19277	0.077827	0.399756	-0.74228	0.954135	-0.79871
13	Pgam1	NM_023418	0	0	1	0	0	0	0	0.256172	0.251347	-0.06428	0.286566	0.5394	-0.86328	0.953956	-0.79356
14	Dscr2	NM_019537	0	0	1	0	0	0	1	0.316795	0.219547	-0.07237	0.225748	0.489567	-0.83743	0.953085	-0.8698
15	Gart	NM_010256	1	1	1	0	0	0	1	0.292314	0.258605	-0.09181	0.18973	0.46723	-0.79813	0.952966	-0.85539
16	Wdr5	NM_080848	0	0	0	0	0	0	0	0.232685	0.3504	-0.13221	0.135721	0.43859	-0.79738	0.952864	-0.78439
17	Kif22	NM_145588	0	0	1	0	0	0	0	0.176652	0.347487	-0.04696	0.250208	0.507912	-0.84206	0.95251	-0.82792
18	Mkrn1	NM_018810	1	1	1	0	1	1	1	0.316332	0.31678	-0.19267	0.018935	0.342788	-0.7189	0.952028	-0.84254
19	Slc25a5	NM_007451	0	0	0	0	0	0	0	0.23587	0.304496	-0.05692	0.221379	0.512233	-0.83199	0.951905	-0.83123
20	Fbl	NM_007991	0	1	1	0	1	0	1	0.259531	0.248685	0.025122	0.293744	0.571446	-0.83071	0.951004	-0.9009
21	1700037H(	NM_026091	0	0	1	0	0	0	1	0.272578	0.281906	-0.13167	0.175319	0.46066	-0.80149	0.948315	-0.79843
22	Supv3l1	NM_181423	0	1	1	0	0	0	1	0.264591	0.293123	-0.0931	0.133786	0.440567	-0.76972	0.948153	-0.8404
23	Parp1	NM_007415	1	1	1	0	0	0	1	0.232712	0.387322	-0.23394	0.015144	0.334614	-0.73971	0.948019	-0.72943
24	Odf2	NM_013615	0	0	0	0	0	0	0	0.364264	0.221854	-0.18125	0.08759	0.367632	-0.80013	0.947283	-0.74238
25	Fig4	NM_133999	1	0	0	0	1	0	0	0.247711	0.342065	-0.10714	0.104856	0.41534	-0.76553	0.946837	-0.82037

Signed WGCNA finds Novel Pathways Involved in Pluripotency in Zhou dataset

p-value	Functional Groups	Black Module highly connected genes (kME)
1.98E-08	response to DNA damage stimulus; DNA damage; DNA repair	Msh6 (0.993), Rif1 (0.983), Mre11a (0.982), Setx (0.974), Xrcc5 (0.971), Chek1 (0.968), Xab2 (0.967), Xrn2 (0.967), Trp53 (0.959), Npm1 (0.958), Tdp1 (0.955), Bccip (0.954)
3.75E-08	Mitochondrion; transit peptide; Mitochondrion	Mrpl15 (0.992), Ppif (0.991), Mrps5 (0.987), Hspa9 (0.984), Coq3 (0.984), Tst (0.981), Mrpl45 (0.98), Akap1 (0.979), L2hgdh (0.978), Mrps31 (0.978), Chchd4 (0.976), Abce1 (0.975), Dci (0.975), Fpgs (0.974), Mrpl39 (0.973), Bdh1 (0.971)
5.83E-08	nucleus; biopolymer metabolic process; DNA binding; cellular metabolic process; Transcription regulation;	Msh6 (0.993), Pes1 (0.991), Zic3 (0.991), Uchl1 (0.99), Rnf138 (0.99), Rnf138 (0.99), Wdr36 (0.989), Pou5f1 (0.989), Rbpj (0.987), Glo1 (0.987), Tdgf1 (0.987), OTTMUSG00000010173 (0.986), Aarsd1 (0.986), Nup133 (0.985), Xpo1 (0.985), Xpo1 (0.985), Dnajc6 (0.985), Klhl13 (0.984), Dppa4 (0.984),
5.26E-04	cell cycle phase; cell cycle process; cell cycle; mitotic cell cycle; mitosis; cell division	Pes1 (0.991), Rif1 (0.983), Mre11a (0.982), Gtpbp4 (0.972), Chek1 (0.968), Mnat1 (0.966), Rcc2 (0.964), Gadd45gip1 (0.963), Rpa1 (0.961), Hells (0.96), Trp53 (0.959), Terf1 (0.959)

- Nup133 is ranked 29th by connectivity and 777th by fold change

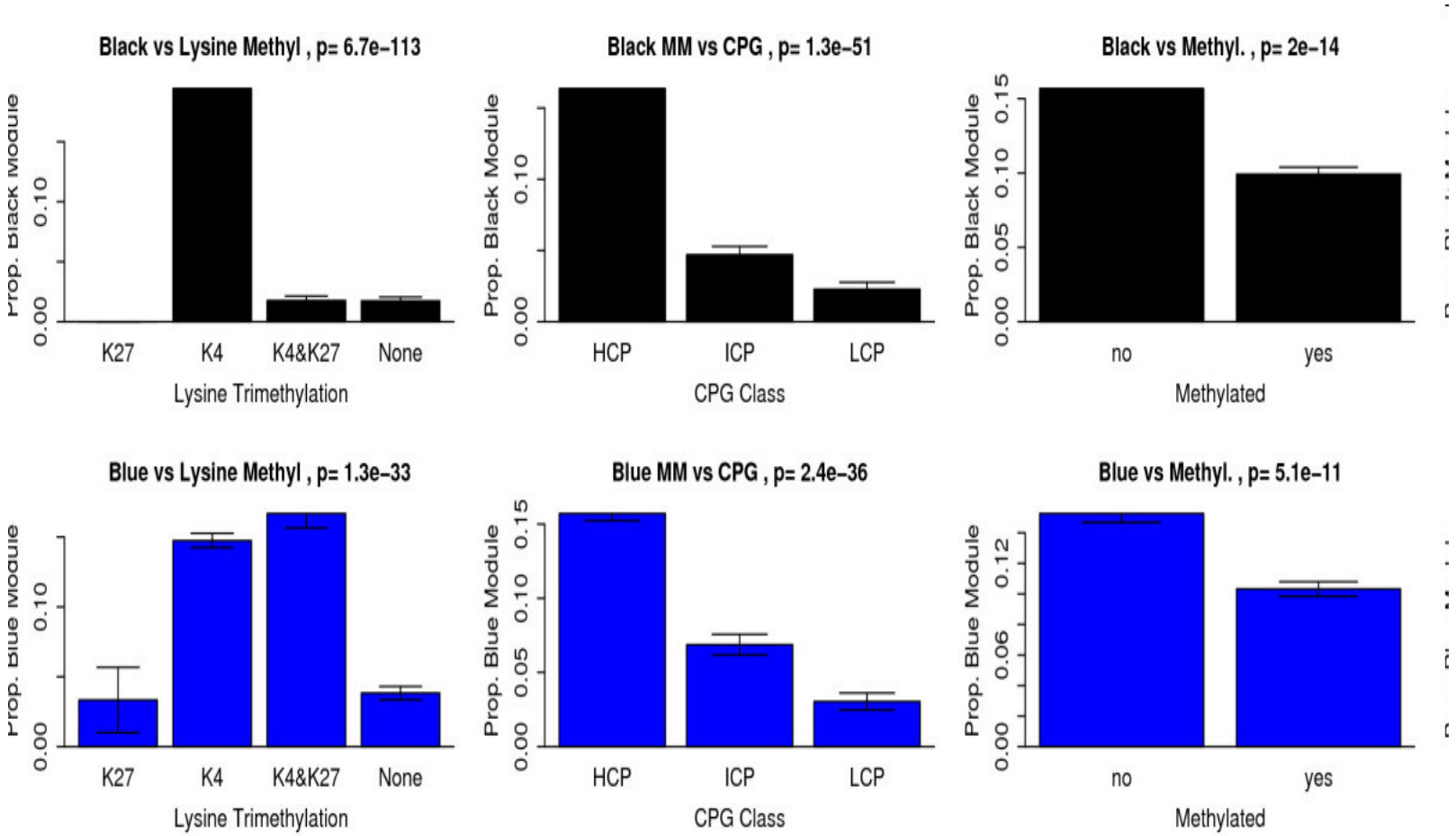
Nuclear Pore Composition Regulates Neural Stem/Progenitor Cell Differentiation in the Mouse Embryo

Floria Lupu[1], Annabelle Alves[2, 3], Kathryn Anderson[1], Valérie Doye[2, 3] and Elizabeth Lacy[1]

Epigenetic Regulation and Module Membership

- Recent studies suggest that chromatin structure and epigenetic modifications, like histone modification and DNA methylation, play a role in controlling gene expression during ES cell self-renewal and differentiation.
  - For example, gene repression by the PcG protein complex via histone H3 lysine 27 trimethylation (H3K27me3) is required for ES cell self-renewal and pluripotency.
- To understand how epigenetic variables contribute to the regulation of ES cells we studied the relationship of the pluripotency and differentiation modules with ES cell H3K4 and H3K27 trimethylation, DNA methylation, and CpG promoter content from previously published data sets.
- Data from Guenther et al. Cell 2007

- **Relating Module Membership to Epigenetic Regulation**.
- The y-axis reports the proportion of top 1000 genes that are known to belong to the group of genes defined on the x-axis.
- Histone H3K4me3 trimethylation status is abbreviated K4, H3K27me3 trimethylation status is abbreviated by K27.
- Note that genes with promoter CpG methylation are significantly ($p = 2.0 \times 10^{-14}$) under-enriched with respect to the top 1000 black module genes.



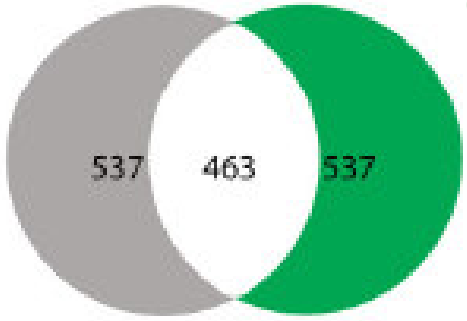
Analysis of variance
module membership (kME) versus epigenetic variables

Source of Variation in kME	kMEblack, Total Prop Var Explained = 8.3%		kMEblue, Total Prop Var Explained = 4.2%	
Source	Prop. Of Total Var	p-value	Prop. Of Total Var	p-value
Histone Trimethylation (K4, K27,K4&K27)	0.067	< 2.2E-16	0.034	< 2.2E-16
cMyc Complex	0.015	< 2.2E-16	0.002	2.6E-04
Oct4 Complex	0.003	8.0E-08	0.001	7.5E-03
CPG class (HCP, ICP, LCP)	0.002	4.9E-04	0.005	6.0E-10
PcG Bound	0.000	8.7E-02	0.000	1.9E-01
CpG Methylated	0.000	7.1E-01	0.001	2.2E-02

Comparison of gene screening based on kME versus screening based on differential expression

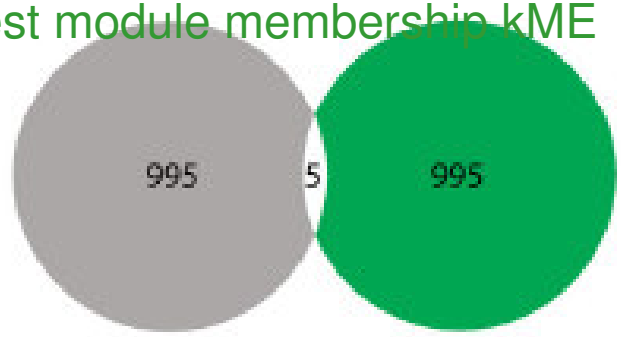
- Venn diagrams show the amount of gene overlap between the top 1000 black (pluripotency) module genes and the top 1000 genes most significantly down-regulated upon Oct4 RNAi (left)
- gene overlap between the top 1000 blue (differentiation) module genes and the 1000 genes most significantly up-regulated with Oct4 RNAi (right).
- Ivanova et al data set.

Grey: Standard differential expression analysis
Green: genes with highest module membership kME



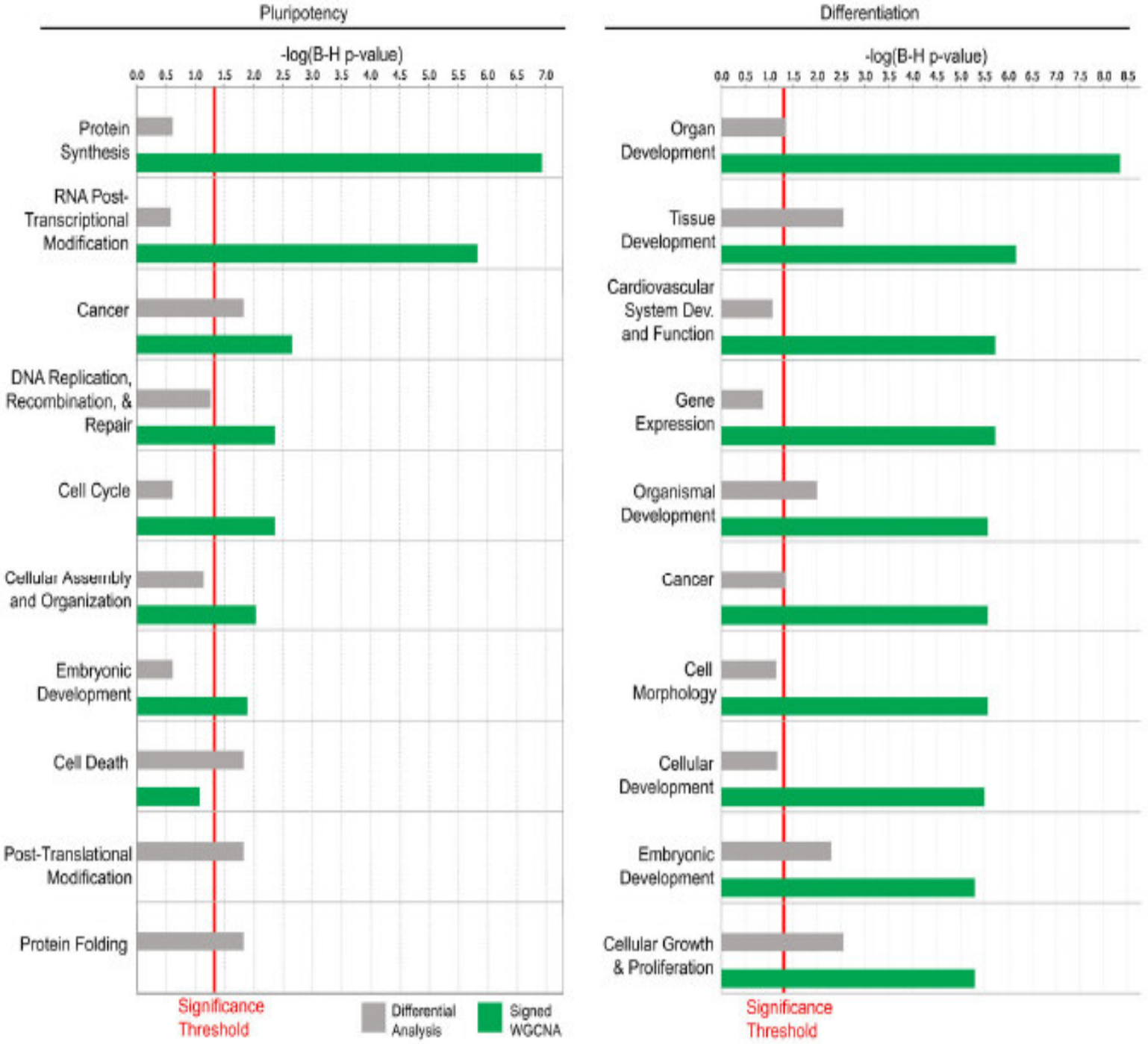
Pluripotency

Green=
Black module genes



Differentiation

Module genes (green) have more significant enrichment than those found by a standard differential expression analysis



Conclusion

- Signed WGCNA
  - has more consistent gene rankings between data sets,
  - is better able to identify functionally enriched groups of genes
- Focus on module eigengenes circumvents the multiple testing problems that plague standard gene-based expression analysis.
- kME =module membership is very useful
  - kME based gene screening identifies several novel stem cell related genes that would not have been found using a standard differential expression analysis
  - kME is valuable for annotating genes with regard to module membership and for identifying genes related to pluripotency and differentiation
  - Can be used as input of analysis of variance to dissect which factors contribute to module membership

Software and Data Availability

- R software tutorials etc can be found online
- Google search
  - weighted co-expression network
  - "WGCNA"
  - "co-expression network"
- http://www.genetics.ucla.edu/labs/horvath/CoexpressionNetwork

Acknowledgement

- Dissertation work of **Mike J Mason**
- Collaborators:

  Guoping Fan, Kathrin Plath, Qing Zhou

- WGCNA R package: Peter Langfelder