Analysis of Human/Mouse Tissue/Cell line expression patterns

Mike Beer, Dongwon Lee JHU 7/12/2012

ackn: Alessandra Breschi, Roderic Guigo, Tom Gingeras, others contributing data/prev analysis

- To what degree are regulatory programs conserved between human and mouse?
- To what degree can we compare regulatory programs in cell lines and mouse tissues?
- Should we be looking for cell line specific expressed genes, or tissue specific expressed genes, housekeeping genes, or something in between?
- For this presentation: regulatory program = coexpression

Unsupervised approach:

- cluster 3 datsets independently: human cell line, human tissues (HBM), and mouse tissues
- use human-mouse orthology compare sets of coexpressed genes

Data: 17 mouse tissues, 16 human tissues (HBM)

	mouse	human(HBM)
1	Adrenal	adipose
2	Colon	adrenal
3	Duodenum	brain
4	GenitalFatPad	breast
5	Heart	colon
6	Kidney	heart
7	LgIntestine	kidney
8	Liver	liver
9	Lung	lung
10	MammaryGland	lymphnode
11	Ovary	ovary
12	SubcFatPad	prostate
13	SmIntestine	skeletalmuscle
14	Spleen	testes
15	Stomach	thyroid
16	Testis	whiteblood
17	Thymus	

average replicates where available, use log(RPKM) for expression level

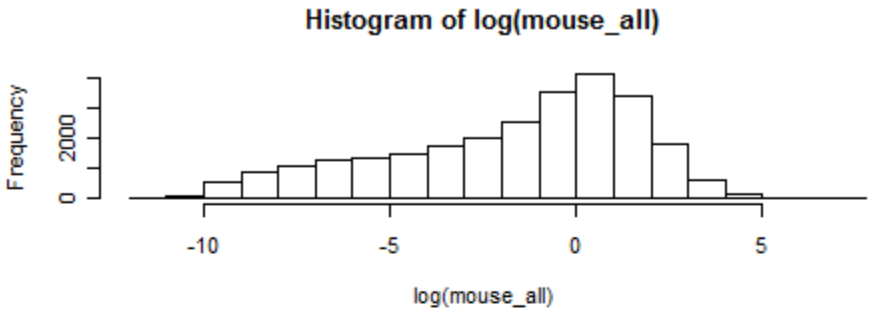
94 human cell line conditions:

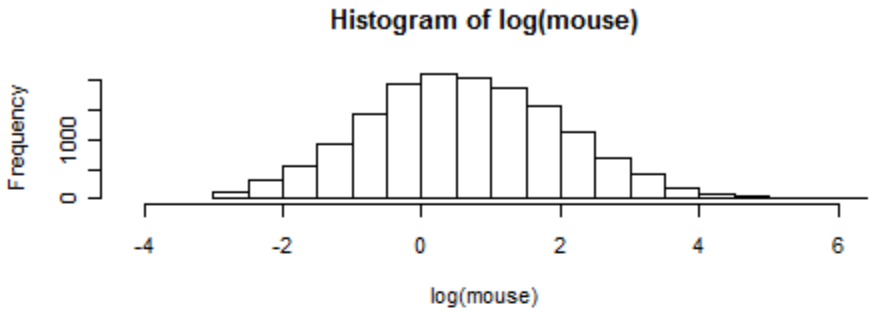
1	PolyA	NHLF	cell
2	PolyA	HSMM	cell
3	PolyA	H1-hESC	cell
4	PolyA	H1-hESC	cell
5	PolyA	GM12878	cell
6	PolyA	NHEK	cell
7	PolyA	HeLa-S3	cell
8	PolyA	HepG2	cell
9	PolyA	K562	cell
10	PolyA	HUVEC	cell
11	PolyA	MCF-7	cell
12	NonPolyA	K562	nucleus
13	PolyA	NHLF	cell
14	PolyA	HeLa-S3	cytosol
15	total	K562	nuclplasm
16	total	K562	nucleolus
17	NonPolyA	GM12878	nucleus
18	PolyA	A549	cell
19	PolyA	HeLa-S3	nucleus
20	total	K562	chromatin
21	PolyA	AG04450	cell
22	PolyA	GM12878	cytosol
23	PolyA	K562	nucleus
24	NonPolyA	HSMM	cell
25	NonPolyA	AG04450	cell
26	NonPolyA	GM12878	cytosol
27	NonPolyA	SK-N-SH_RA	cell
28	PolyA	SK-N-SH_RA	cell
29	NonPolyA	HUVEC	nucleus
30	NonPolyA	HepG2	nucleus
31	NonPolyA	HUVEC	cell
32	PolyA	HUVEC	cell

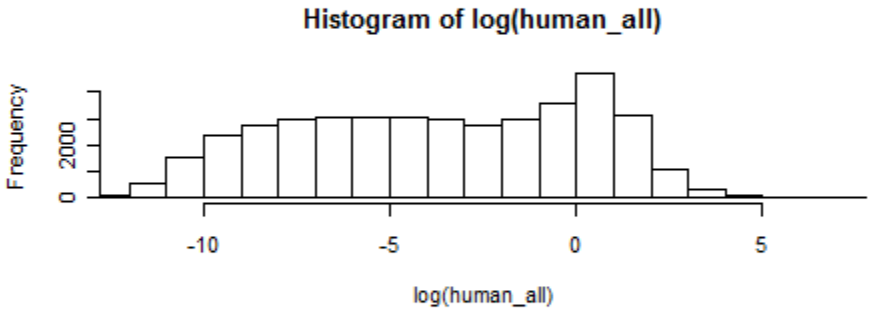
33	NonPolyA	NHEK	nucleus
34	NonPolyA	NHEK	cell
35	NonPolyA	HeLa-S3	nucleus
36	PolyA	NHEK	nucleus
37	PolyA	GM12878	cell
38	PolyA	K562	cytosol
39	NonPolyA	NHLF	cell
40	PolyA	BJ	cell
41	NonPolyA	HepG2	cytosol
42	NonPolyA	BJ	cell
43	NonPolyA	MCF-7	cell
44	NonPolyA	HepG2	cell
45	NonPolyA	K562	cytosol
46	PolyA	HeLa-S3	cell
47	PolyA	HepG2	nucleus
48	PolyA	HSMM	cell
49	PolyA	K562	cell
50	NonPolyA	GM12878	cell
51	NonPolyA	K562	cell
52	NonPolyA	A549	cell
53	PolyA	H1-hESC	cell
54	PolyA	NHEK	cell
55	NonPolyA	H1-hESC	cell
56	PolyA	HUVEC	cytosol
57	PolyA	HepG2	cell
58	PolyA	NHEK	cytosol
59	NonPolyA	HeLa-S3	cell
60	PolyA	GM12878	nucleus
61	PolyA	MCF-7	cell
62	PolyA	HUVEC	nucleus
63	PolyA	HepG2	cytosol
64	total	HCH	cell

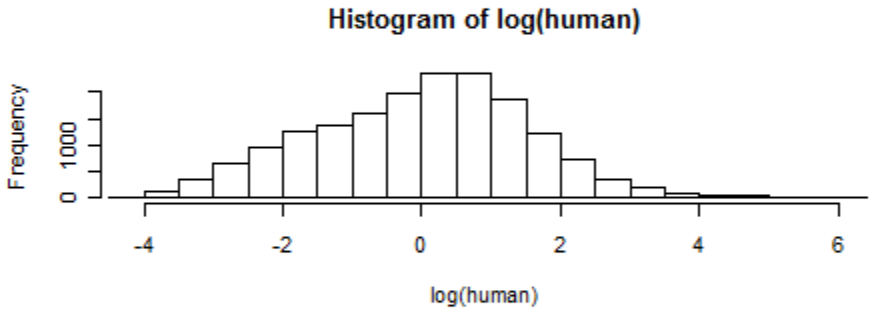
65	total	NHEM-f_M2	cell
66	total	HFDPC	cell
67	PolyA	SK-N-SH	cell
68	PolyA	MCF-7	nucleus
69	PolyA	MCF-7	cytosol
70	total	hMSC-BM	cell
71	total	IMR90	cell
72	total	HVMF	cell
73	total	SkMC	cell
74	total	hMSC-AT	cell
75	NonPolyA	CD20+	cell
76	total	HSaVEC	cell
77	total	HPIEpC	cell
78	PolyA	SK-N-SH	nucleus
79	total	HAoEC	cell
80	PolyA	SK-N-SH	cytosol
81	total	HWP	cell
82	total	NHEM_M2	cell
83	PolyA	A549	nucleus
84	PolyA	Monoc-CD1	cell
85	total	HOB	cell
86	PolyA	IMR90	nucleus
87	PolyA	CD20+	cell
88	PolyA	A549	cytosol
89	total	HPC-PL	cell
90	total	NHDF	cell
91	NonPolyA	Monoc-CD1	cell
92	PolyA	IMR90	cytosol
93	PolyA	IMR90	cell
94	total	hMSC-UC	cell

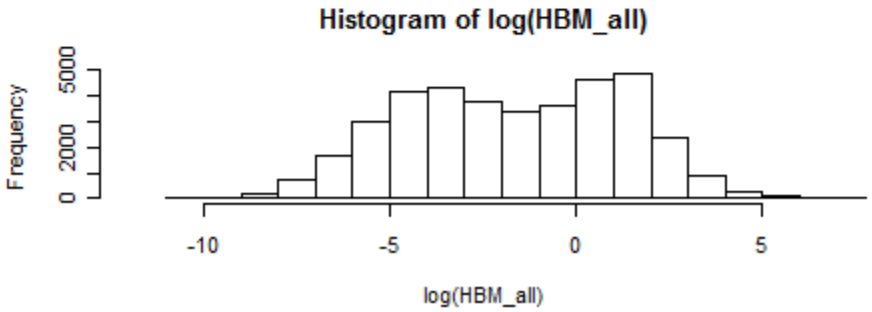
Datasets have variable amounts of low expressed genes, dominate clustering if not removed.
Mean expression across conds: Soln: Remove genes unless RPKM>1 in at least one condition.

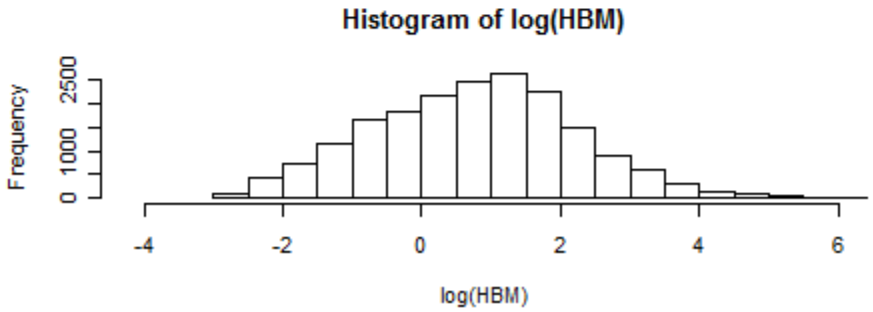












genes: 31916 mouse
43576 human
37913 HBM

genes: 15508 mouse
17649 human
19026 HBM

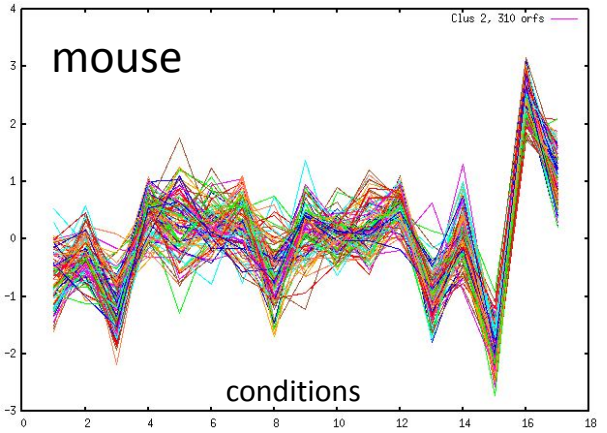
Cluster using kmeans, ask for 30 or 50 clusters with strict threshold for coexpression of a cluster C<0.8 (mouse, HBM) C<0.7 (human)

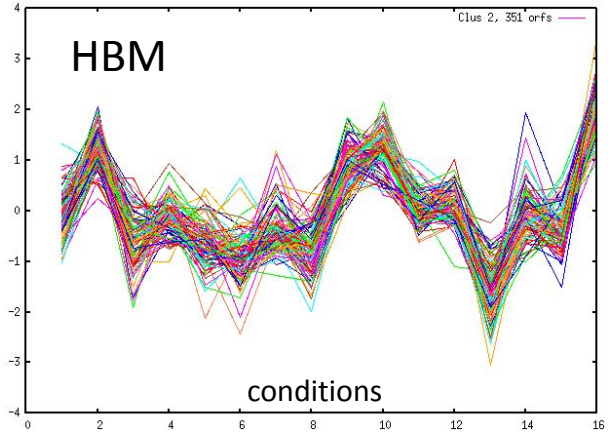
Restrict to genes which have 1-1 ortholog in mouse-human (11009 genes)

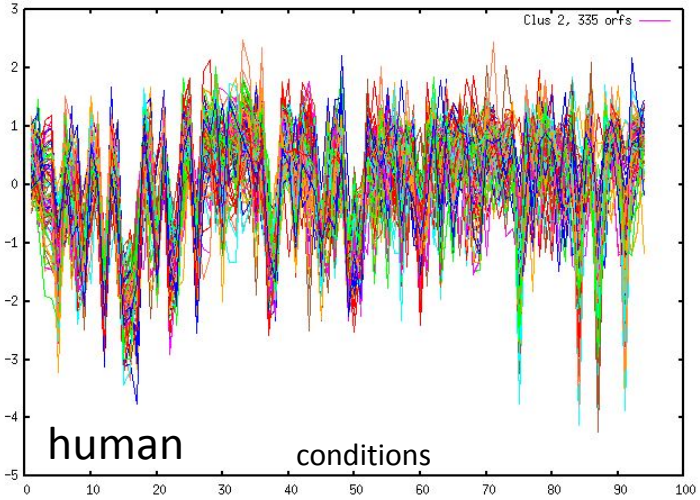
With 50 clusters, n ~ 10,000-11,000 get put in a cluster
With 30 clusters, n ~ 6,000-7,000 get put in a cluster

n ~ 100-300 genes/cluster

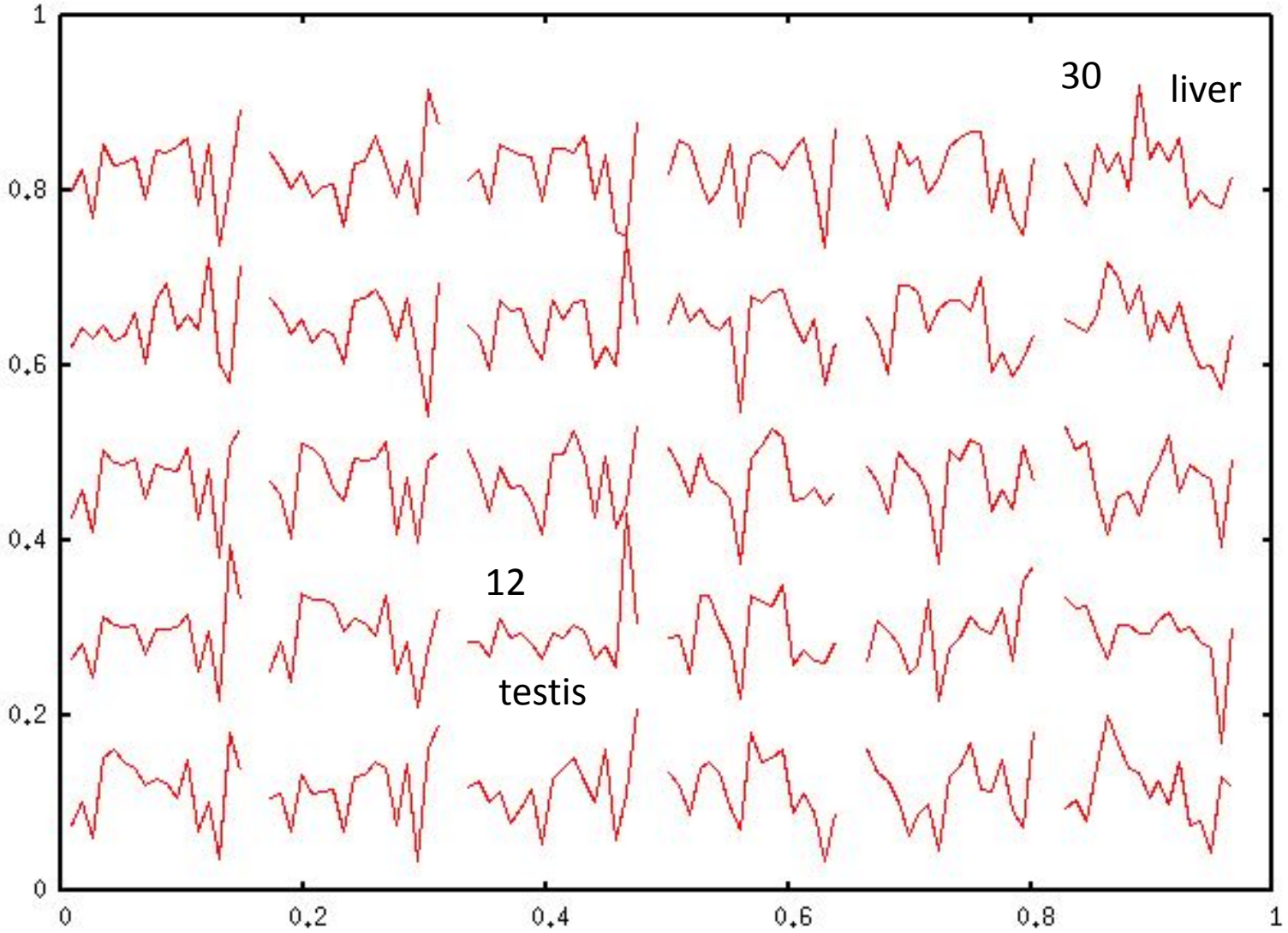
$$C_{x,y} = \frac{1}{n}\sum_{i=1}^{n}\frac{(x_i - \bar{x})(y_i - \bar{y})}{\sigma_x \sigma_y}$$



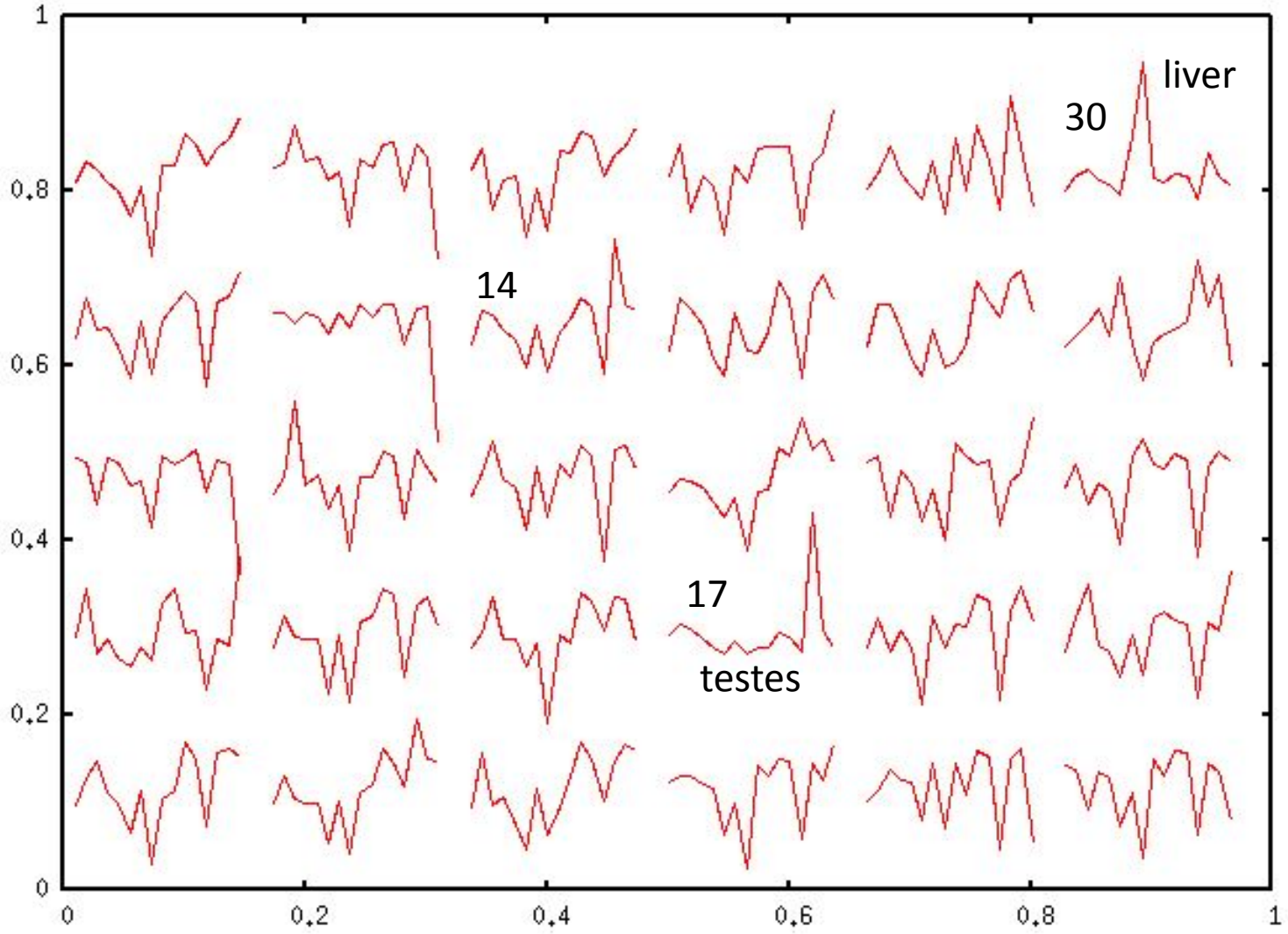




mean expr patterns of 30 mouse clusters. Tissue exclusive expression is the exception

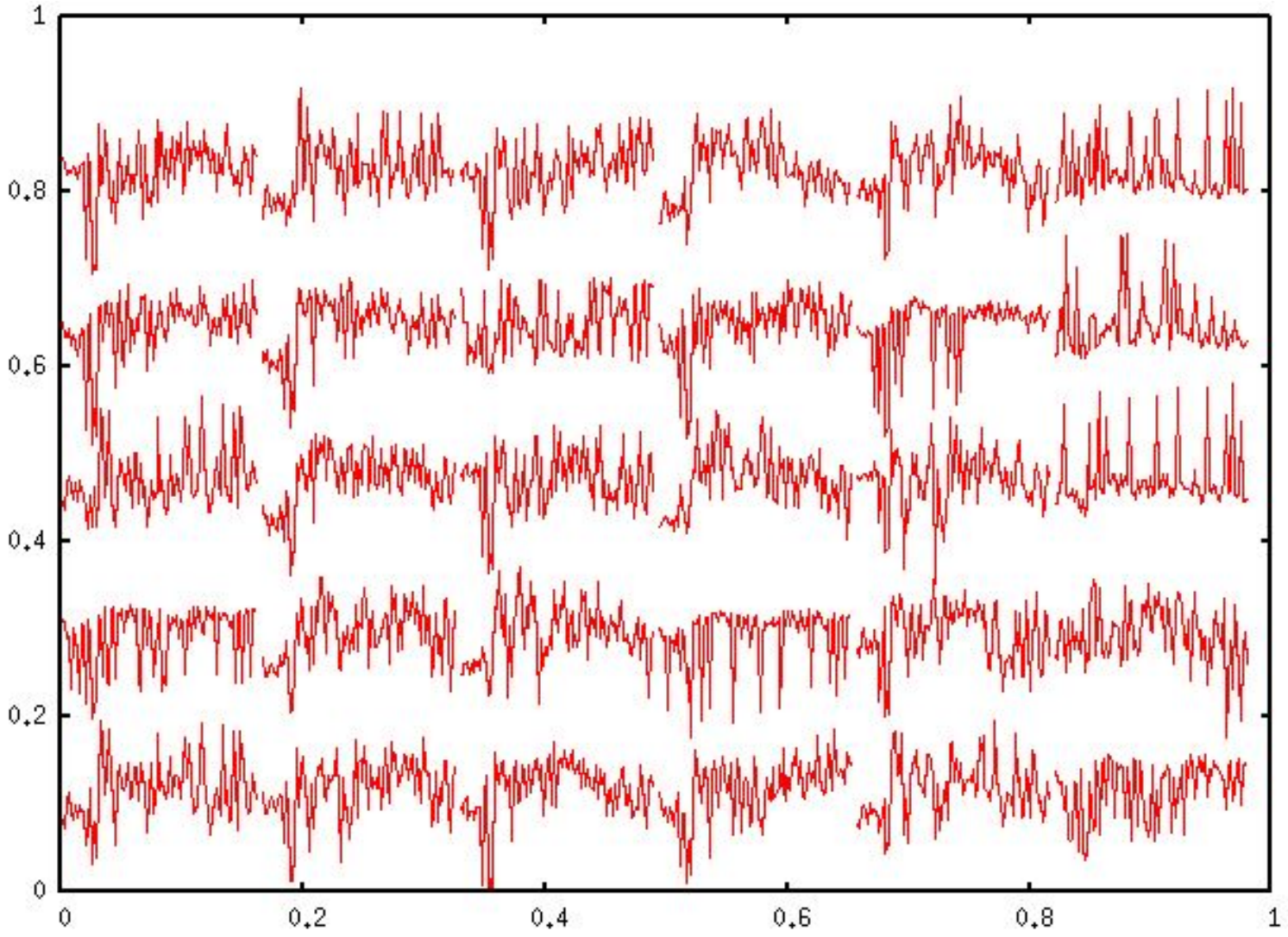


mean expr patterns of 30 HBM clusters. Tissue exclusive expression is the exception

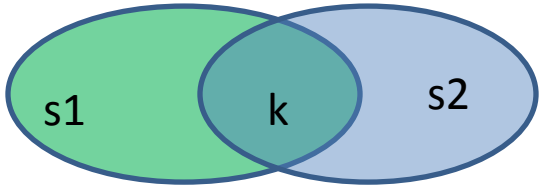


mean expr patterns of 30 human clusters

common modes: off in sets of conds
on in sets of conds



Clusters are significantly enriched for genes of common biological function GO terms (all datasets)



mouse clusters

Clus	Term	s1	k	pvalue
1	translation	256	28	2.30E-12
2	proteolysis involved in cellular prot	432	31	5.20E-10
3	RNA processing	343	37	2.70E-17
4	regulation of immune system proce	230	40	5.60E-28
5	gene expression	2455	92	6.60E-15
6	response to DNA damage stimulus	253	22	8.90E-11
7	vesicle-mediated transport	390	26	3.80E-09
8	protein import into nucleus, dockin	15	4	7.50E-05
9	regulation of gene expression	1852	54	1.30E-06
10	transcription	1789	46	9.00E-06
11	gene expression	2455	83	1.80E-14
12	spermatogenesis	202	15	5.90E-09
13	chromatin modification	203	15	1.20E-07
14	microtubule cytoskeleton	291	18	1.20E-09
15	cellular protein metabolic process	1842	46	6.80E-05
16	blood vessel development	236	20	4.00E-11
17	extracellular matrix	272	22	4.00E-11
18	extracellular matrix	272	29	1.90E-17
19	extracellular matrix	272	19	4.00E-09
20	immune system process	621	23	1.30E-05
21	nitrogen compound metabolic proc	2978	70	4.10E-12
22	chromosome, centromeric region	65	14	4.60E-15
23	nervous system development	733	20	7.30E-05
24	mitochondrion	831	22	2.00E-05
25	lysosome	142	10	1.50E-06
26	mitochondrial inner membrane	250	32	8.70E-27
27	response to virus	29	3	0.0021
28	gene expression	2455	31	0.001
29	mitochondrial inner membrane	250	31	3.60E-29
30	coagulation	65	8	6.70E-09

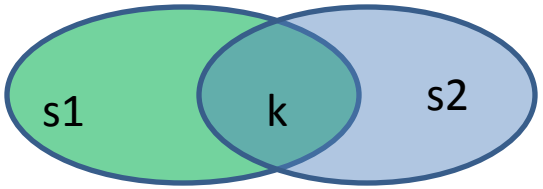
HBM clusters

Clus	Term	p-value
1	mRNA processing	5.30E-13
2	T cell activation	3.50E-25
3	cell adhesion	1.70E-21
4	AP-type membrane coat adaptor com	1.80E-05
5	chromatin modification	3.40E-19
6	gene expression	4.00E-11
7	RNA splicing	1.00E-17
8	synapse	1.30E-09
9	cell migration	2.00E-11
10	nervous system development	8.80E-11
11	ncRNA processing	1.10E-13
12	microtubule cytoskeleton	2.00E-05
13	Golgi apparatus	1.30E-06
14	M phase	1.10E-19
15	translational elongation	1.50E-30
16	small GTPase mediated signal transdu	2.00E-05
17	spermatogenesis	1.00E-29
18	cellular protein catabolic process	6.80E-15
19	RNA processing	7.60E-09
20	lysosome	3.00E-09
21	cell-cell junction	5.70E-06
22	endoplasmic reticulum	2.30E-13
23	immune system process	1.00E-07
24	anaphase-promoting complex-deper	3.90E-08
25	cilium	2.10E-09
26	cell adhesion	9.20E-06
27	small GTPase mediated signal transdu	8.50E-07
28	endoplasmic reticulum	4.70E-22
29	myofibril/sarcomere	2.80E-30
30	blood coagulation	2.60E-15

human cell line clusters

Clus	Term	p-value
1	RNA processing	3.60E-05
2	cell adhesion	7.20E-11
3	tRNA processing	3.00E-08
4	proteasome complex	6.90E-09
5	mitochondrion	1.70E-15
6	protein transport	1.10E-08
7	proteolysis involved in cellular protei	8.10E-08
8	protein transport	2.40E-08
9	endoplasmic reticulum	1.90E-07
10	chromatin modification	1.40E-07
11	mitochondrial inner membrane	5.00E-11
12	RNA processing	1.20E-07
13	ribosome	2.40E-24
14	extracellular matrix part	4.30E-16
15	ribosome/mitochondrion	1.40E-07
16	lysosome	8.00E-07
17	cell junction	3.90E-06
18	transcription	3.90E-07
19	protein localization	1.40E-06
20	gene expression	6.90E-16
21	RNA splicing	7.10E-16
22	mitochondrion	3.80E-09
23	mitochondrion	1.60E-10
24	regulation of signal transduction	1.30E-07
25	nuclear pore	4.90E-08
26	axonogenesis	2.60E-05
27	DNA replication	1.60E-29
28	leukocyte activation	1.60E-19
29	cell-cell junction	1.30E-07
30	immune system process	1.60E-10

Hypergeometric p-value for overlap of clusters are almost all significant: mouse vs. HBM, human vs. mouse, human vs HBM



mouse cluster / HBM cluster

		s1	s2	k	p-value
1	24	337	201	23	4.92E-08
2	6	310	314	43	2.85E-18
3	1	286	355	43	1.39E-17
4	2	279	351	159	3.16E-176
5	5	273	326	39	1.59E-16
6	5	263	326	29	9.07E-10
7	22	256	212	20	1.04E-07
8	1	253	355	39	2.10E-16
9	23	236	202	19	6.36E-08
10	6	235	314	34	3.50E-15
11	11	233	283	26	2.61E-10
12	14	227	262	29	1.02E-13
13	7	224	312	24	2.04E-08
14	25	217	199	26	1.40E-14
15	23	214	202	15	9.07E-06
16	3	207	337	45	3.94E-26
17	3	204	337	45	2.04E-26
18	3	200	337	38	5.42E-20
19	3	189	337	36	5.03E-19
20	23	187	202	20	1.96E-10
21	15	178	261	16	4.92E-06
22	14	157	262	33	2.40E-22
23	10	152	286	30	1.81E-18
24	9	150	289	16	1.94E-06
25	22	150	212	7	0.0259249
26	19	147	228	8	0.0113908
27	20	146	223	18	7.99E-10
28	15	115	261	9	0.00162546
29	29	110	138	2	0.402323
30	30	77	86	9	7.40E-09

HBM cluster / mouse cluster

		s1	s2	k	p-value
1	3	355	286	43	1.39E-17
2	4	351	279	159	3.16E-176
3	17	337	204	45	2.04E-26
4	6	333	263	24	1.39E-06
5	5	326	273	39	1.59E-16
6	2	314	310	43	2.85E-18
7	13	312	224	24	2.04E-08
8	23	306	152	15	2.10E-05
9	16	289	207	33	4.03E-17
10	23	286	152	30	1.81E-18
11	11	283	233	26	2.61E-10
12	8	278	253	20	6.25E-06
13	7	277	256	17	0.000256322
14	22	262	157	33	2.40E-22
15	20	261	187	18	4.54E-07
16	6	235	263	15	0.000517458
17	12	235	227	11	0.00949517
18	2	233	310	20	9.24E-06
19	3	228	286	23	2.50E-08
20	27	223	146	18	7.99E-10
21	14	215	217	15	2.25E-05
22	7	212	256	20	1.04E-07
23	20	202	187	20	1.96E-10
24	1	201	337	23	4.92E-08
25	14	199	217	26	1.40E-14
26	18	196	200	21	5.38E-11
27	4	159	279	12	0.000720185
28	30	154	77	4	0.022661
29	17	138	204	6	0.0432695
30	30	86	77	9	7.40E-09

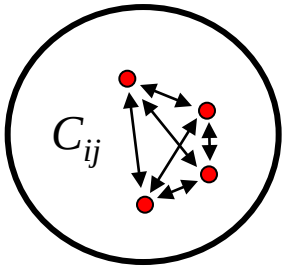
human / mouse

		s1	s2	k	p-value
1	11	373	233	34	4.35E-13
2	19	335	189	34	2.70E-17
3	21	307	178	25	2.22E-11
4	27	298	146	12	6.01E-04
5	1	275	337	48	2.17E-23
6	1	271	337	52	2.33E-27
7	2	260	310	26	1.91E-08
8	8	255	253	21	4.07E-07
9	25	255	150	11	7.24E-04
10	11	253	233	30	1.17E-14
11	7	239	256	17	4.23E-05
12	3	234	286	36	3.15E-18
13	26	234	147	23	5.01E-14
14	18	232	200	33	1.42E-20
15	28	226	115	17	1.64E-10
16	9	226	236	23	5.16E-10
17	16	225	207	21	1.35E-09
18	5	218	273	30	1.49E-14
19	7	218	256	25	3.77E-11
20	6	215	263	26	8.17E-12
21	3	198	286	18	3.90E-06
22	29	196	110	9	1.47E-04
23	1	195	337	33	6.45E-16
24	17	190	204	9	8.89E-03
25	2	187	310	24	4.66E-10
26	17	167	204	22	4.01E-13
27	22	151	157	79	6.52E-113
28	4	121	279	65	3.01E-73
29	19	107	189	1	0.844632
30	4	89	279	53	7.26E-63

human / HBM

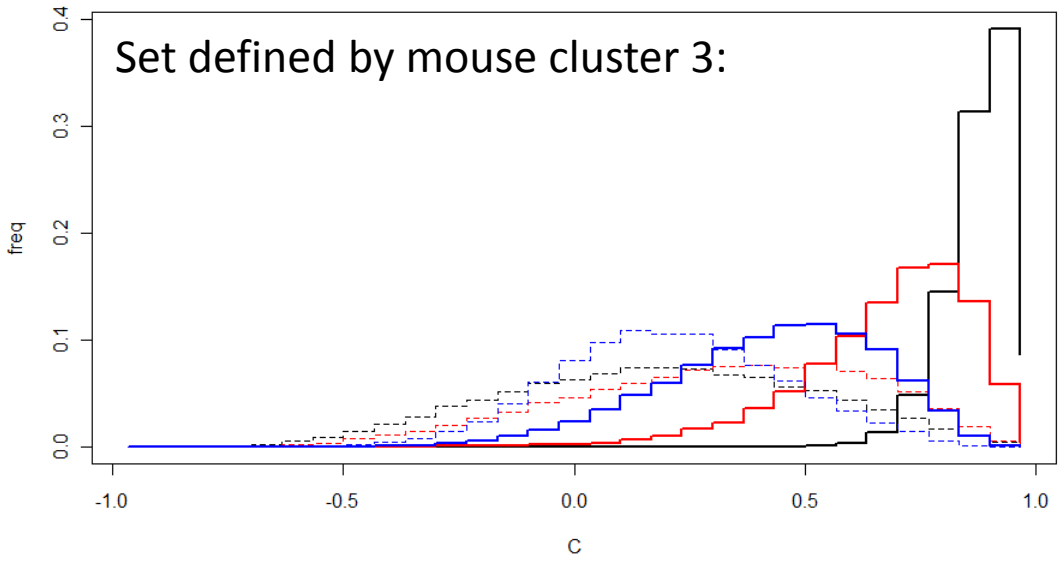
		s1	s2	k	p-value
1	11	373	283	36	5.62E-12
2	9	335	289	52	4.36E-26
3	7	307	312	34	7.61E-12
4	15	298	261	24	1.60E-07
5	13	275	277	25	2.59E-08
6	1	271	355	30	3.19E-09
7	1	260	355	34	2.58E-12
8	18	255	233	19	1.93E-06
9	18	255	233	15	3.41E-04
10	11	253	283	27	3.21E-10
11	6	239	314	17	4.93E-04
12	4	234	333	25	3.97E-08
13	15	234	261	21	1.62E-07
14	3	232	337	76	7.95E-59
15	15	226	261	16	9.45E-05
16	12	226	278	15	6.13E-04
17	9	225	289	37	1.45E-19
18	5	218	326	40	5.89E-21
19	16	218	235	17	4.14E-06
20	5	215	326	37	1.82E-18
21	7	198	312	25	3.26E-10
22	11	196	283	15	1.61E-04
23	1	195	355	24	1.62E-08
24	27	190	159	9	1.73E-03
25	6	187	314	21	7.97E-08
26	10	167	286	26	1.44E-13
27	14	151	262	35	3.42E-25
28	2	121	351	86	4.45E-104
29	21	107	215	17	2.29E-11
30	2	89	351	67	1.80E-83

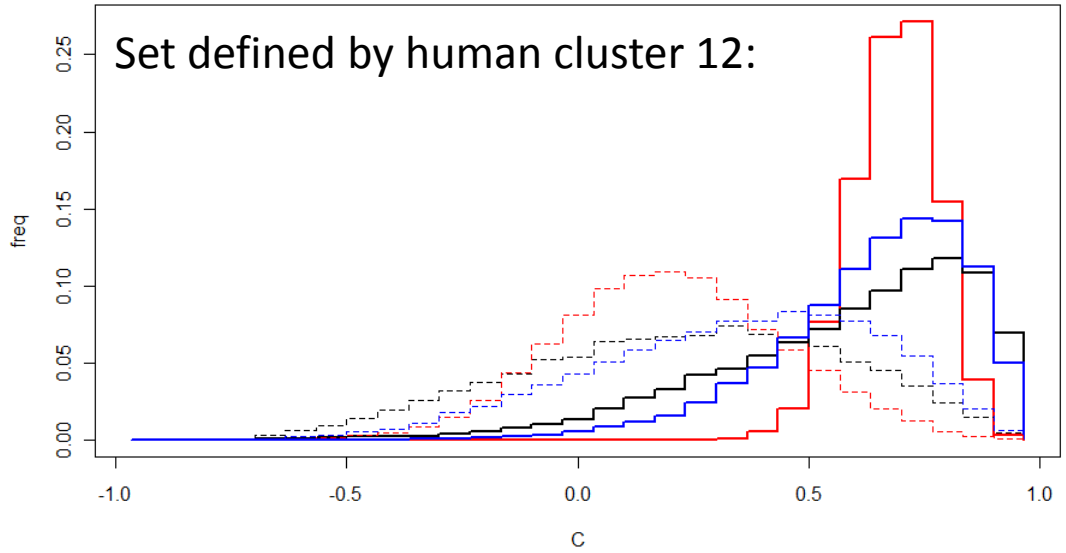
Independent of clustering:
How coexpressed is a set of coexpressed genes on another data set?



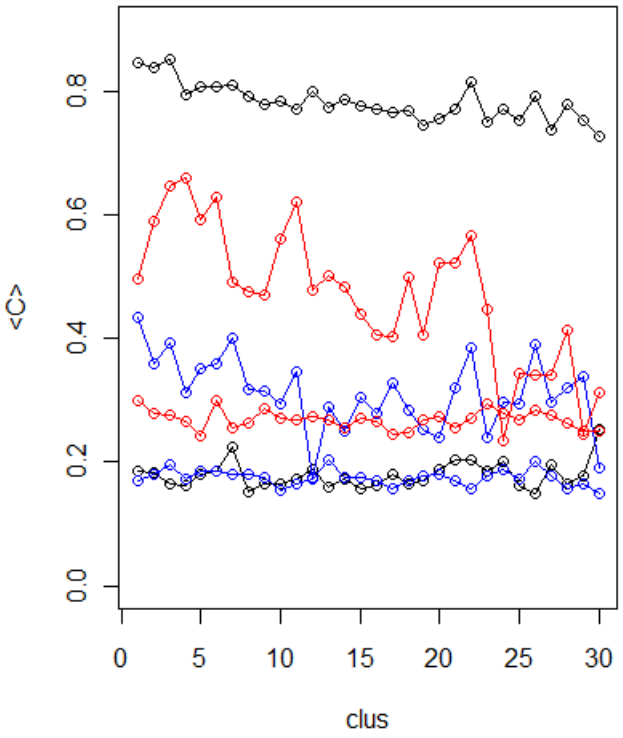
on mouse tissue data
on human cell data
on HBM tissue data

dashed curves are
random set of same size

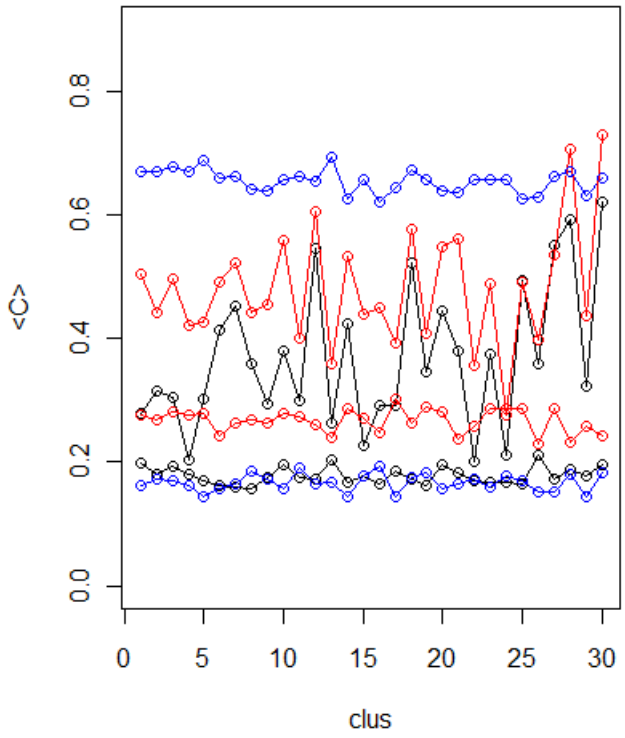




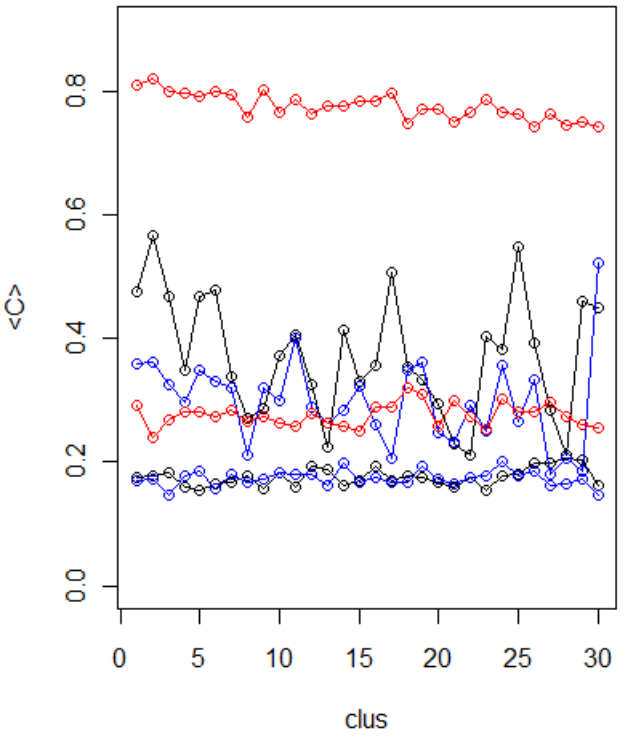
mouse tissue clusters



human cell clusters



HBM clusters



top 3: actual cluster
bottom 3: matched random set

on mouse tissue data
on human cell data
on HBM tissue data

1. mouse tissue clusters tightly coexpressed in HBM data, and still significantly coexpressed in human cell data
2. human cell clusters coexpressed in both HBM data and mouse tissue data, some variability
3. HBM clusters are only slightly more coexpressed in mouse tissue data compared to human cell line data
4. most genes are coexpressed across many conditions in all datasets