Future works on AGCT

* Visualization
  * 3D graphs, PCA balls, validation graphs, etc.
* Methodology : Cluster inference
* Tests on performance & memory

Finding Gene Network regulated by the toxicity equivalent factor (TEF) of TCDD and TCDF chemicals .

Algorithms for clustering

* **Affinity propagation,** Bergman K-means, EM, Hierarchical clustering

Computer global optimum with small and sensitive clusters.

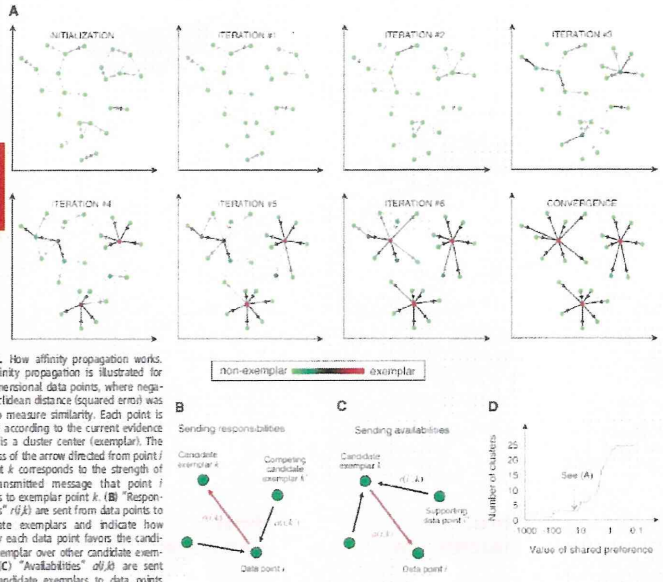
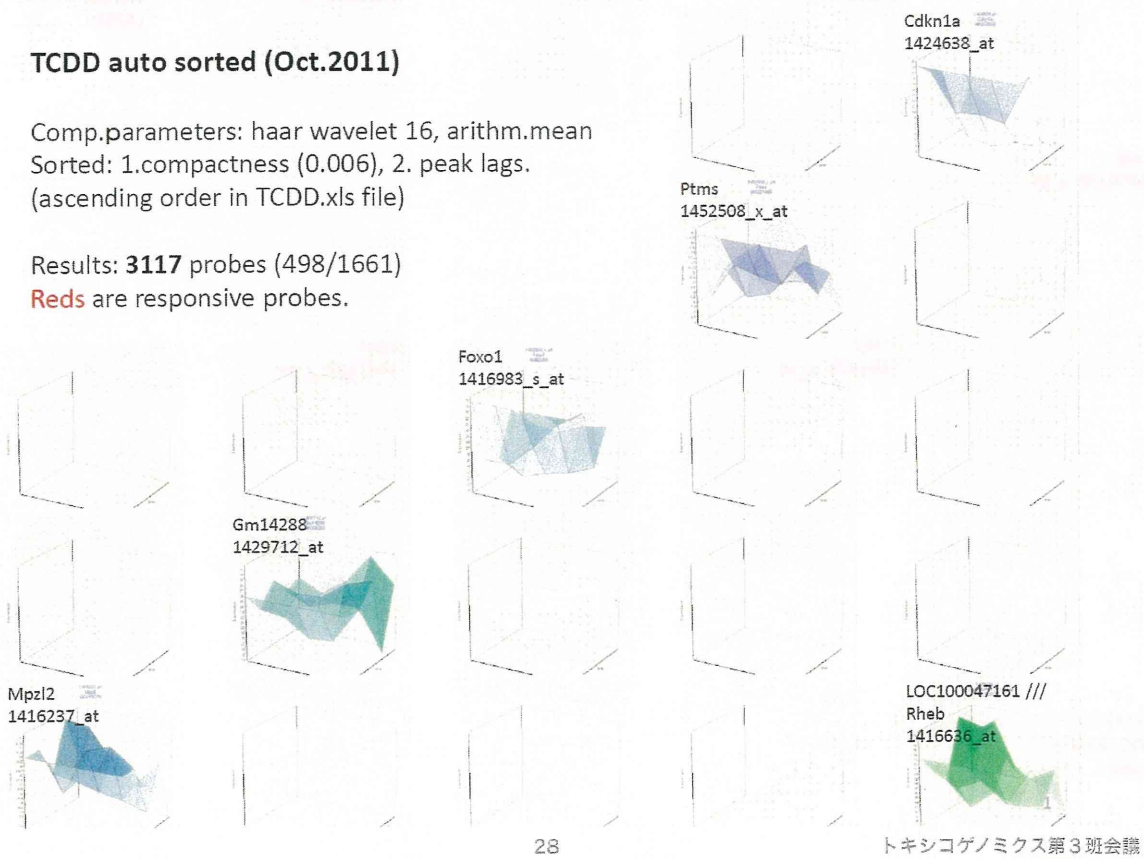


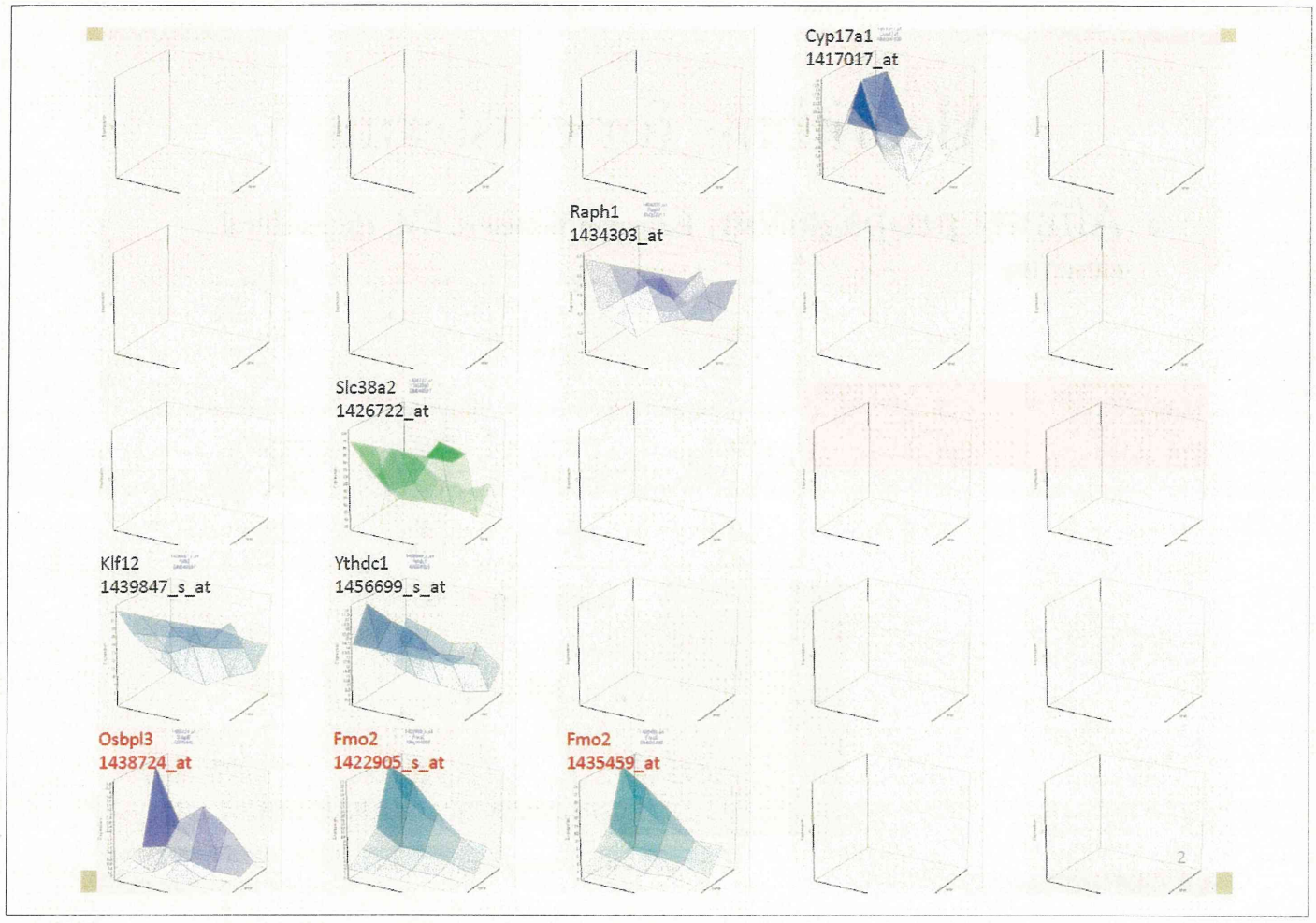
Fig. 1. How affinity propagation works. (**A**) Affinity propagation is illustrated for two-dimensional data points, where negative Euclidean distance (squared error) was used to measure similarity. Each point is colored according to the current evidence that it is a cluster center (exemplar). The darkness of the arrow directed from point *i* to point *k* corresponds to the strength of the transmitted message that point *i* belongs to exemplar point *k*. (**B**) "Responsibilities" $r(i,k)$ are sent from data points to candidate exemplars and indicate how strongly each data point favors the candidate exemplar over other candidate exemplars. (**C**) "Availabilities" $a(i,k)$ are sent from candidate exemplars to data points and indicate to what degree each candidate exemplar is available as a cluster center for the data point. (**D**) The effect of the value of the input preference (common for all data points) on the number of identified exemplars (number of clusters) is shown. The value that was used in (**A**) is also shown, which was computed from the median of the pairwise similarities.

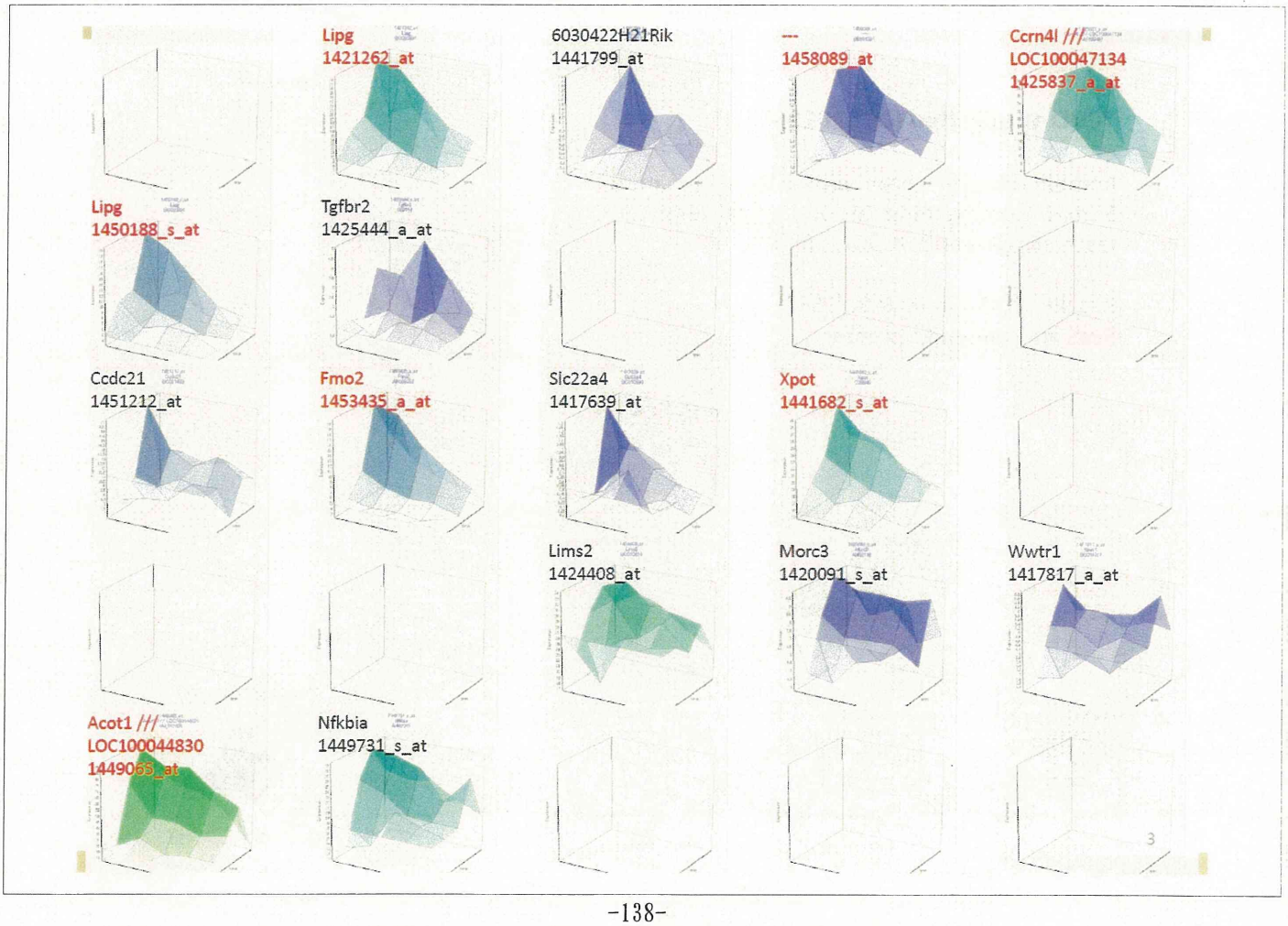
TCDD auto sorted (Oct.2011)

Comp.parameters: haar wavelet 16, arithm.mean
Sorted: 1.compactness (0.006), 2. peak lags.
(ascending order in TCDD.xls file)

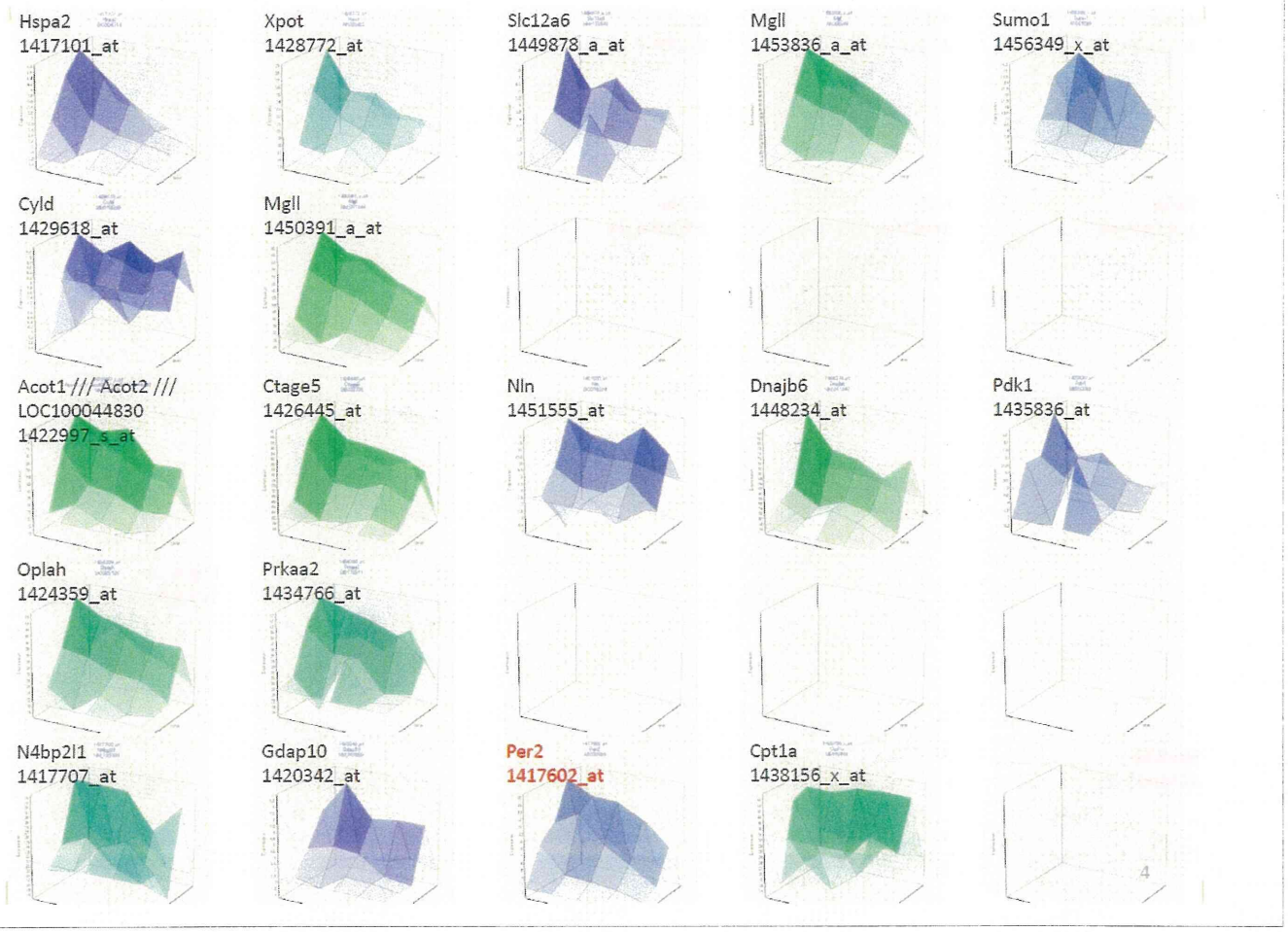
Results: **3117** probes (498/1661)
Reds are responsive probes.

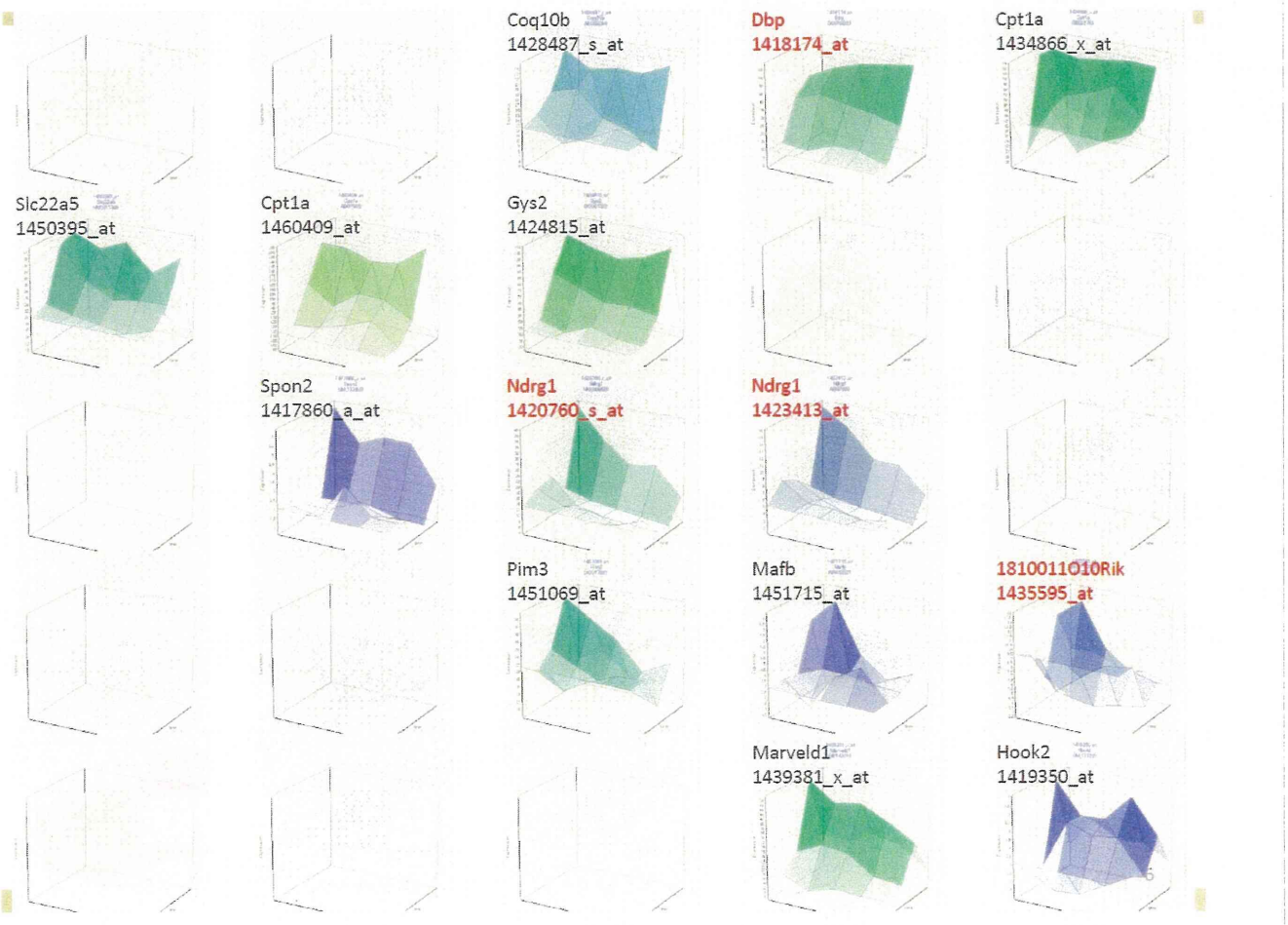


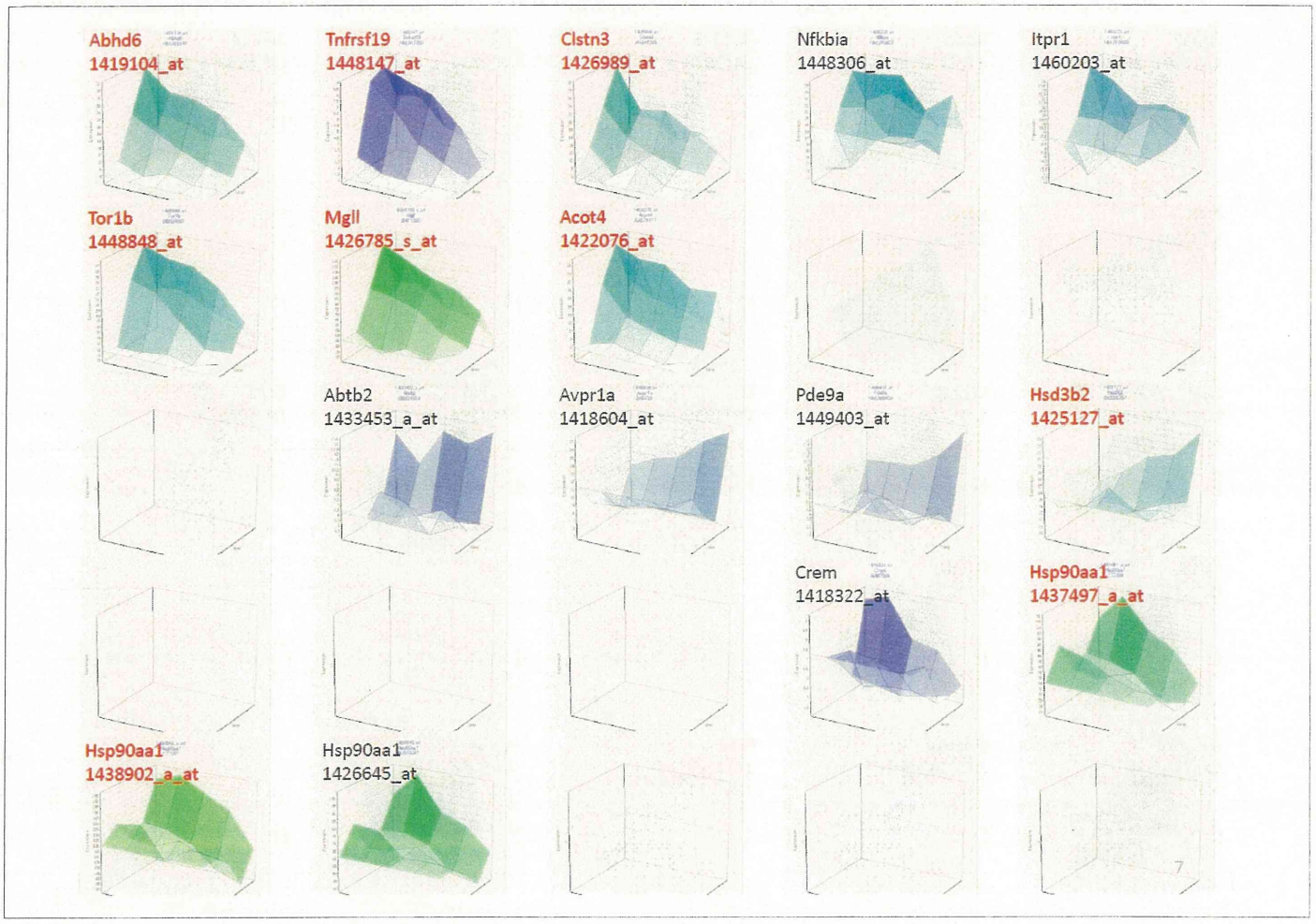

Cyp17a1
1417017_at
Raph1
1434303_at
Slc38a2
1426722_at
Klf12
1439847_s_at
Ythdc1
1456699_s_at
Osbpl3
1438724_at
Fmo2
1422905_s_at
Fmo2
1435459_at
2

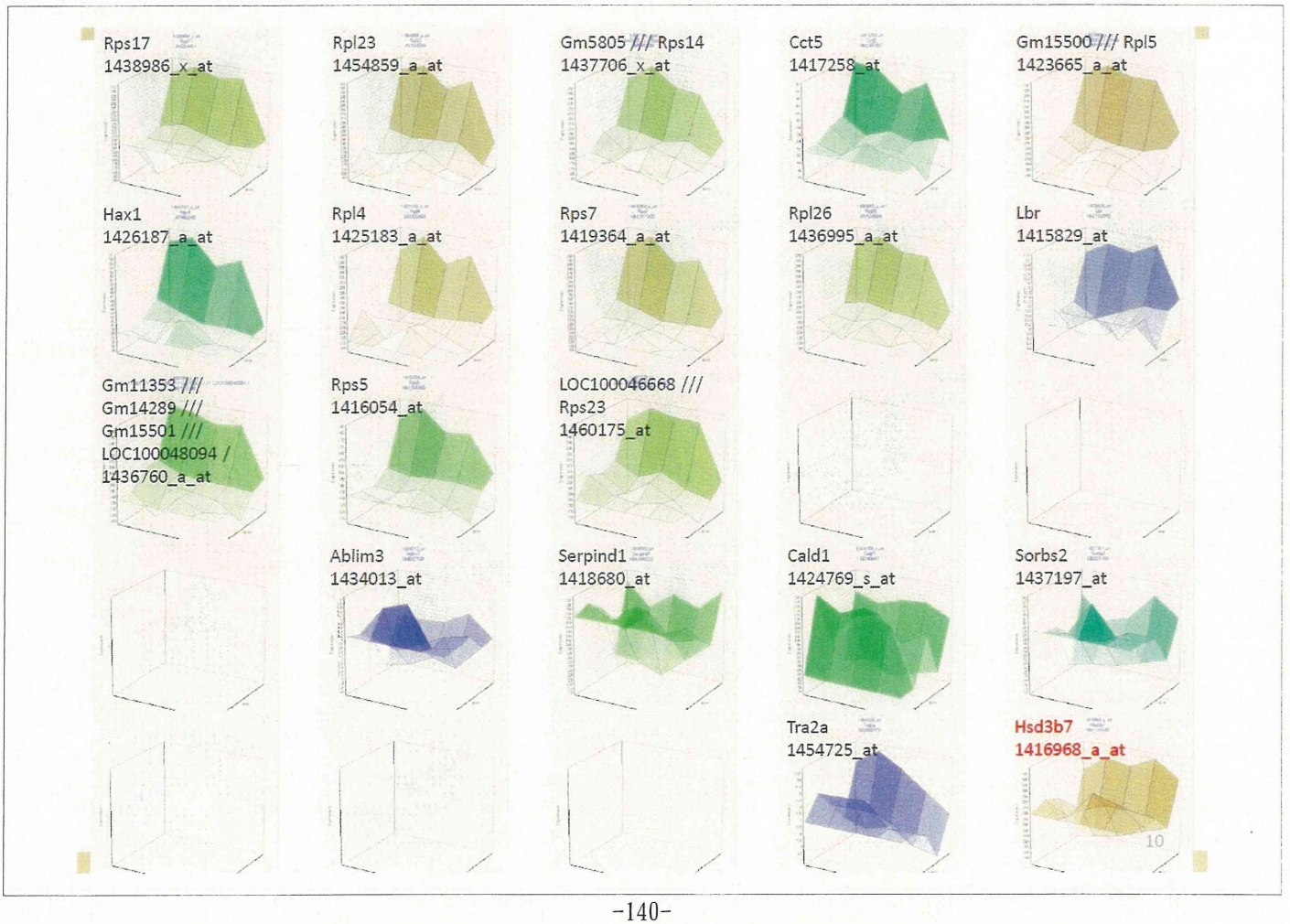

Lipg
1421262_at
6030422H21Rik
1441799_at

1458089_at
Ccrn4l ///
LOC100047134
1425837_a_at
Lipg
1450188_s_at
Tgfbr2
1425444_a_at
Ccdc21
1451212_at
Fmo2
1453435_a_at
Slc22a4
1417639_at
Xpot
1441682_s_at
Lims2
1424408_at
Morc3
1420091_s_at
Wwtr1
1417817_a_at
Acot1 ///
LOC100044830
1449063_at
Nfkbia
1449731_s_at
3


Hspa2
1417101_at
Xpot
1428772_at
Slc12a6
1449878_a_at
Mgll
1453836_a_at
Sumo1
1456349_x_at
Cyld
1429618_at
Mgll
1450391_a_at
Acot1 /// Acot2 /// LOC100044830
1422997_s_at
Ctage5
1426445_at
Nln
1451555_at
Dnajb6
1448234_at
Pdk1
1435836_at
Oplah
1424359_at
Prkaa2
1434766_at
N4bp2l1
1417707_at
Gdap10
1420342_at
Per2
1417602_at
Cpt1a
1438156_x_at
4


Coq10b
1428487_s_at
Dbp
1418174_at
Cpt1a
1434866_x_at
Slc22a5
1450395_at
Cpt1a
1460409_at
Gys2
1424815_at
Spon2
1417860_a_at
Ndrg1
1420760_s_at
Ndrg1
1423413_at
Pim3
1451069_at
Mafb
1451715_at
1810011O10Rik
1435595_at
Marveld1
1439381_x_at
Hook2
1419350_at

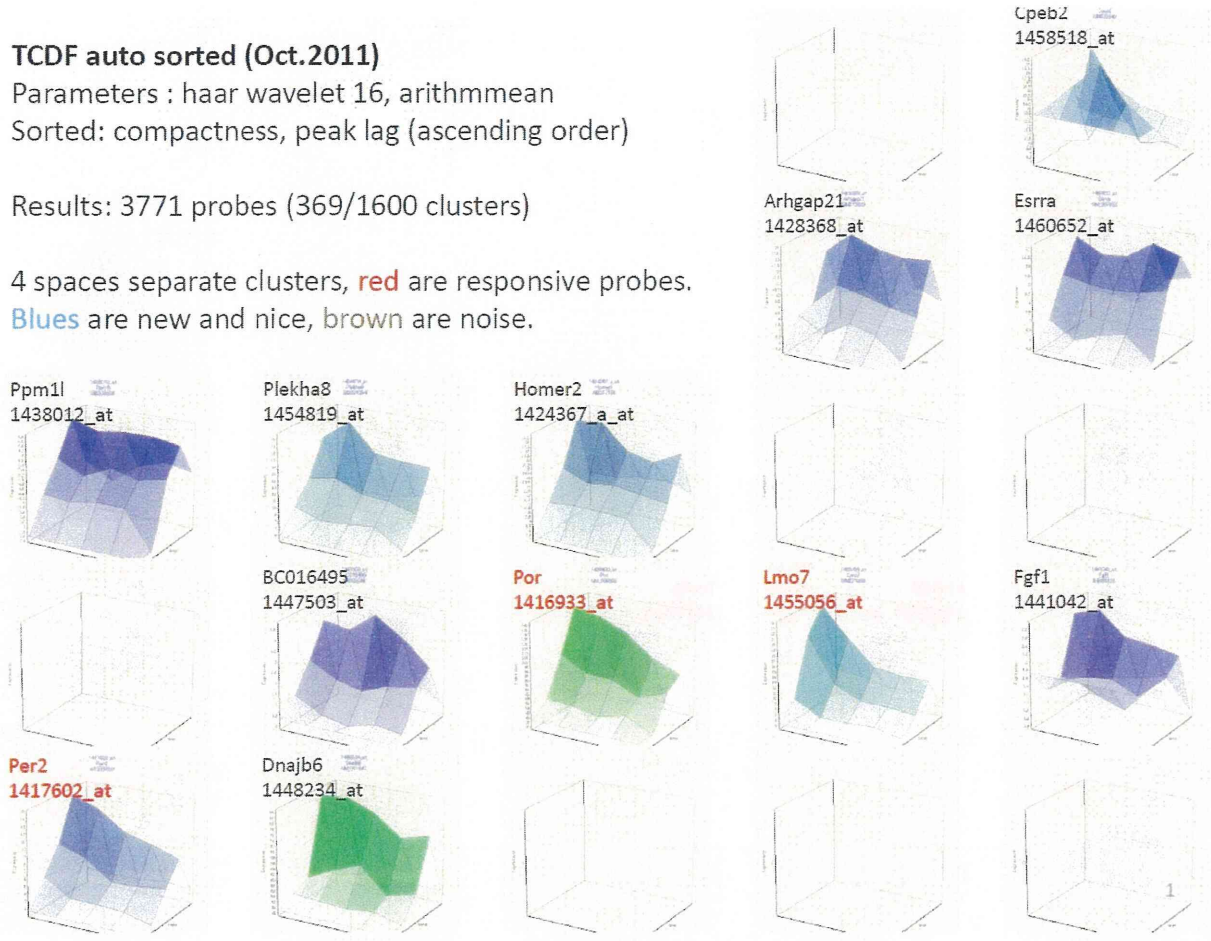

Abhd6
1419104_at
Tnfrsf19
1448147_at
Clstn3
1426989_at
Nfkbia
1448306_at
Itpr1
1460203_at
Tor1b
1448848_at
Mgll
1426785_s_at
Acot4
1422076_at
Abtb2
1433453_a_at
Avpr1a
1418604_at
Pde9a
1449403_at
Hsd3b2
1425127_at
Crem
1418322_at
Hsp90aa1
1437497_a_at
Hsp90aa1
1438902_a_at
Hsp90aa1
1426645_at
7

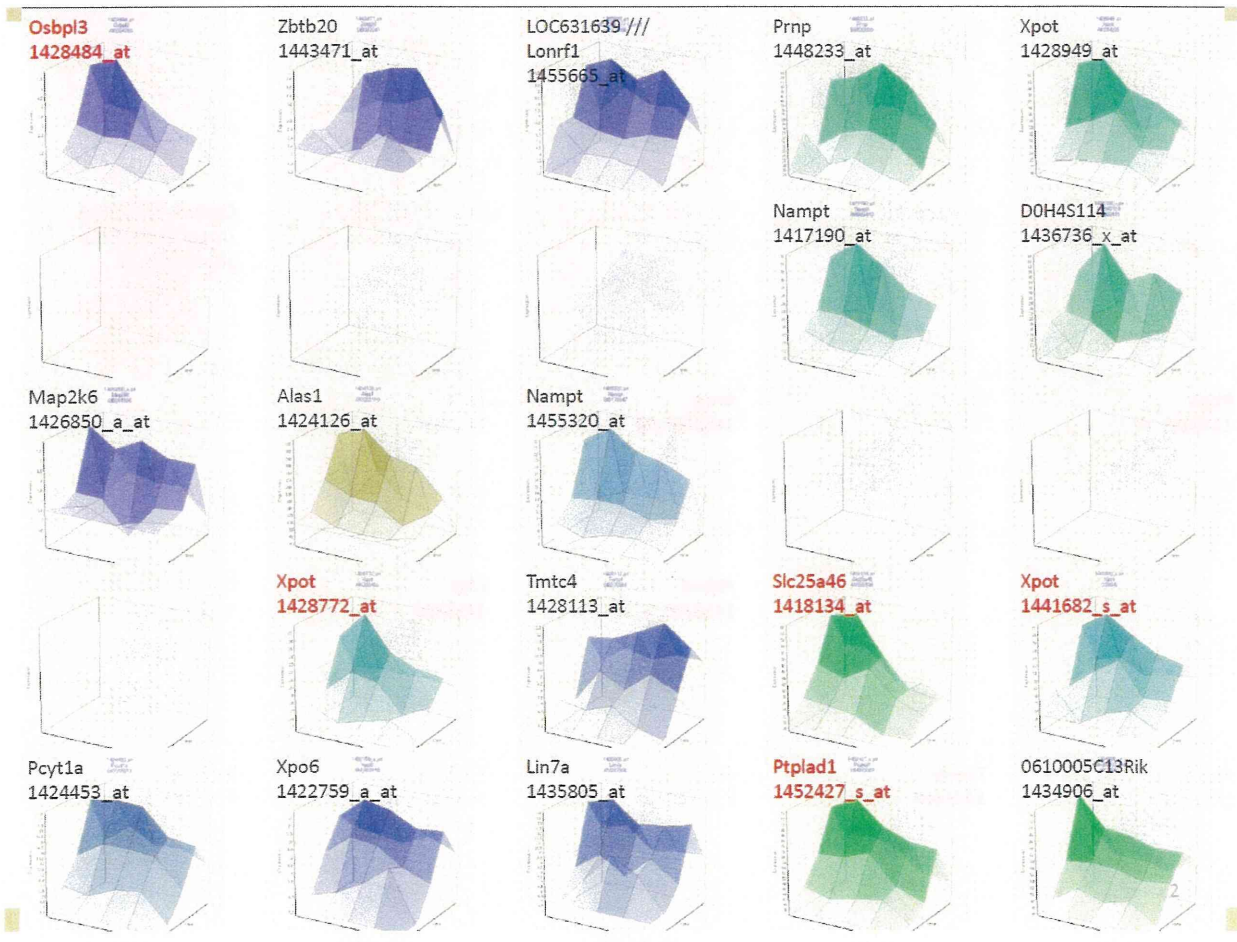

Rps17
1438986_x_at
Rpl23
1454859_a_at
Gm5805 /// Rps14
1437706_x_at
Cct5
1417258_at
Gm15500 /// Rpl5
1423665_a_at
Hax1
1426187_a_at
Rpl4
1425183_a_at
Rps7
1419364_a_at
Rpl26
1436995_a_at
Lbr
1415829_at
Gm11353 ///
Gm14289 ///
Gm15501 ///
LOC100048094 /
1436760_a_at
Rps5
1416054_at
LOC100046668 ///
Rps23
1460175_at
Ablim3
1434013_at
Serpind1
1418680_at
Cald1
1424769_s_at
Sorbs2
1437197_at
Tra2a
1454725_at
Hsd3b7
1416968_a_at
10

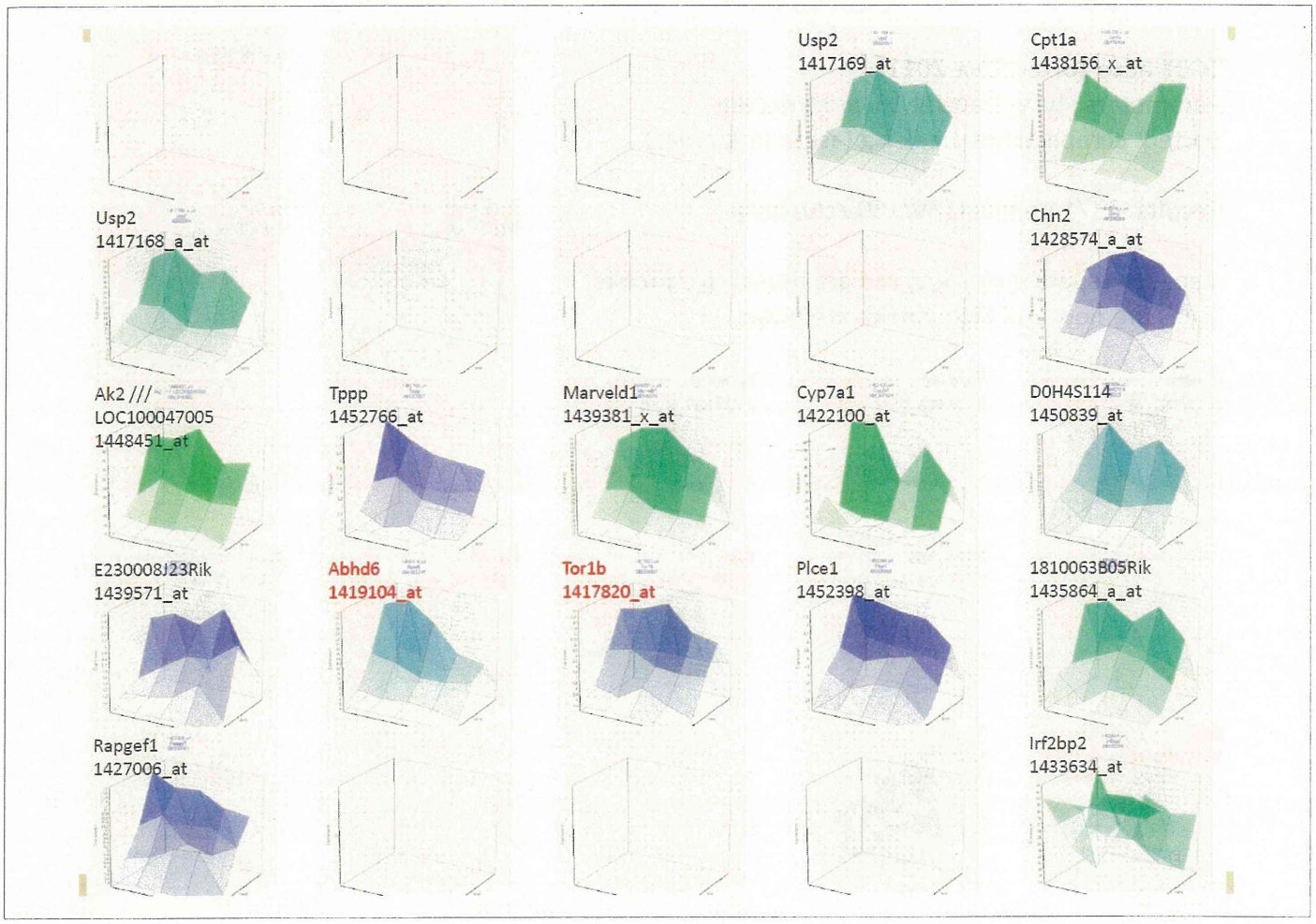
TCDF auto sorted (Oct.2011)
Parameters : haar wavelet 16, arithmmean
Sorted: compactness, peak lag (ascending order)

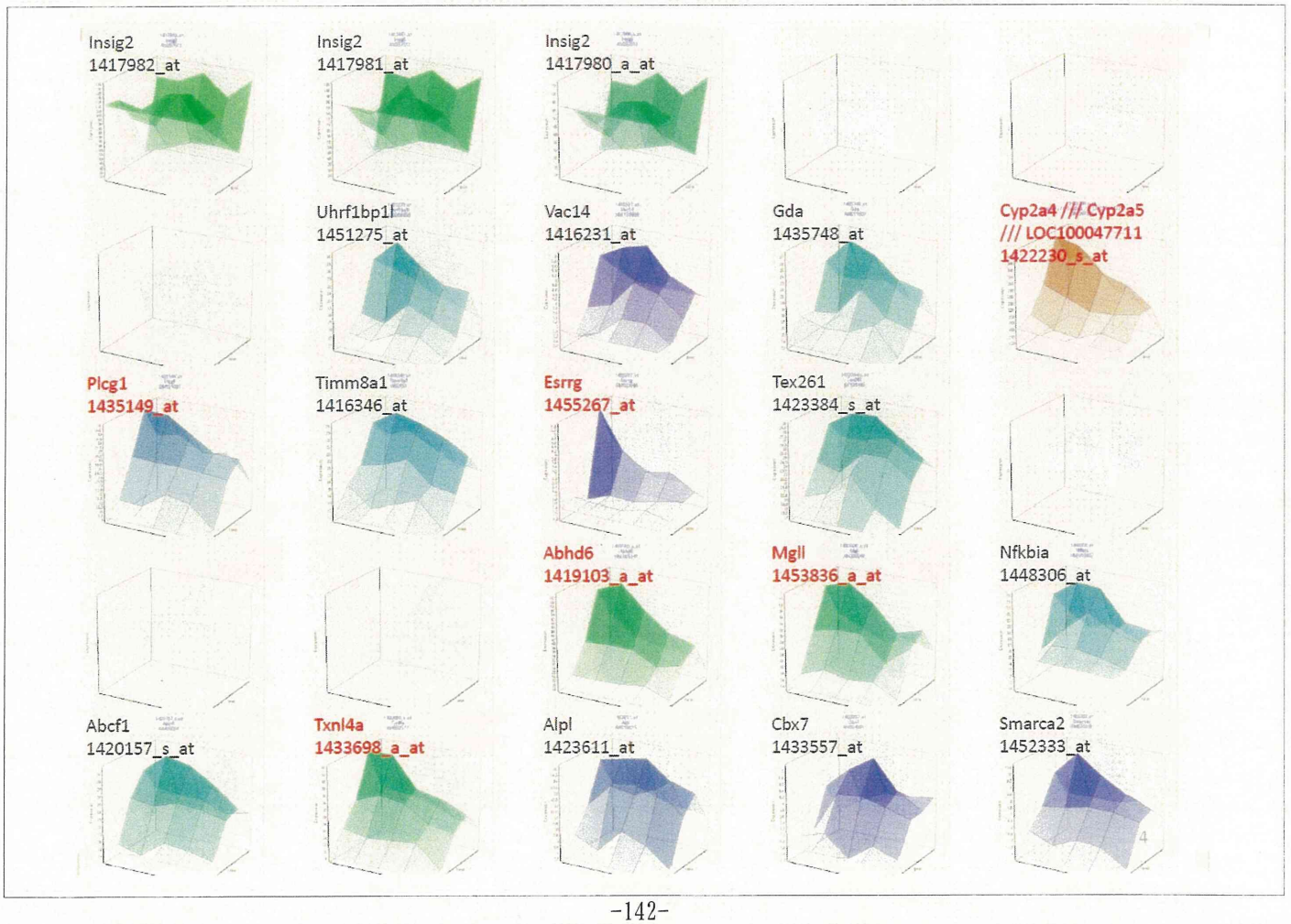
Results: 3771 probes (369/1600 clusters)

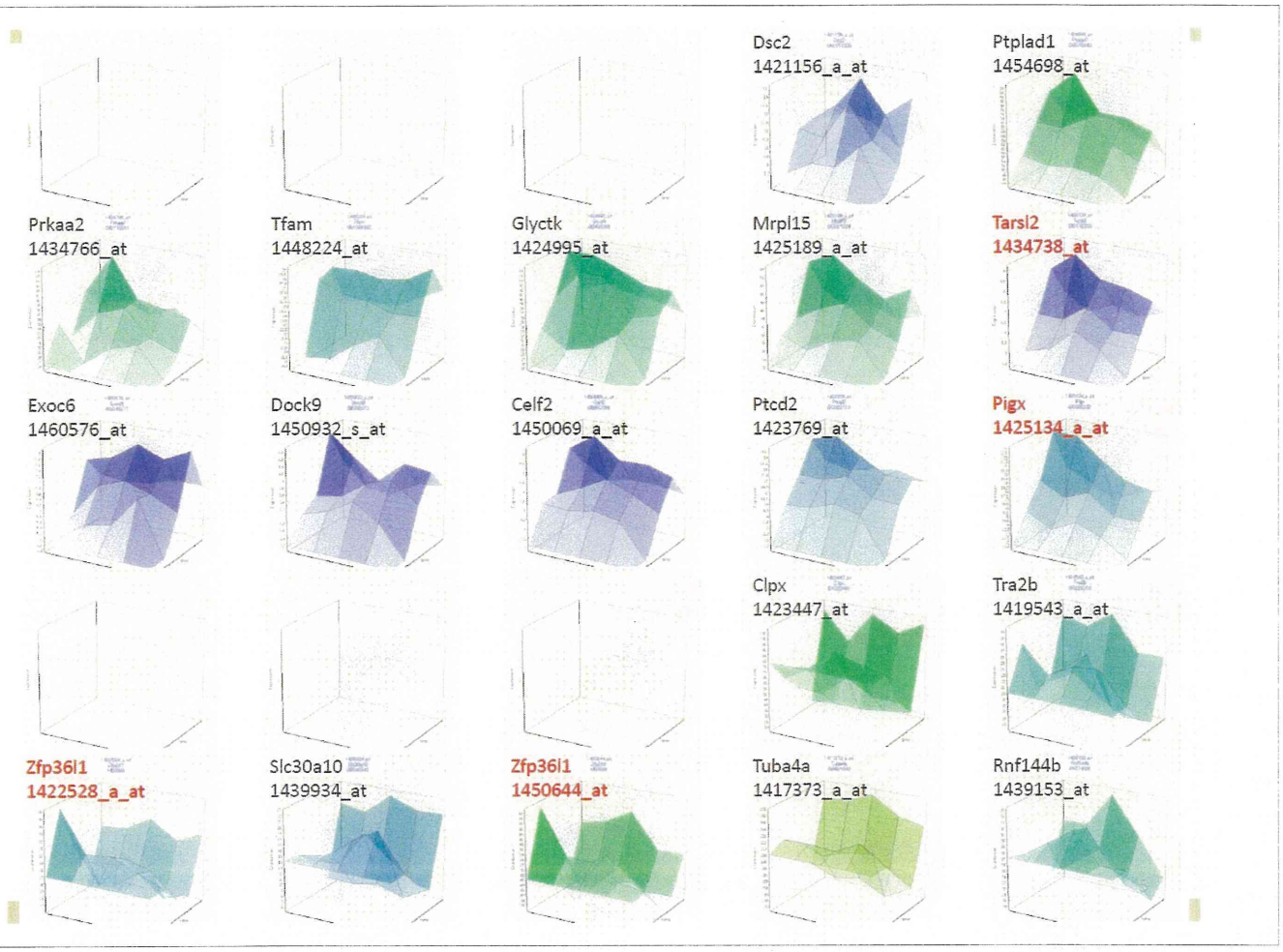
4 spaces separate clusters, red are responsive probes.
Blues are new and nice, brown are noise.

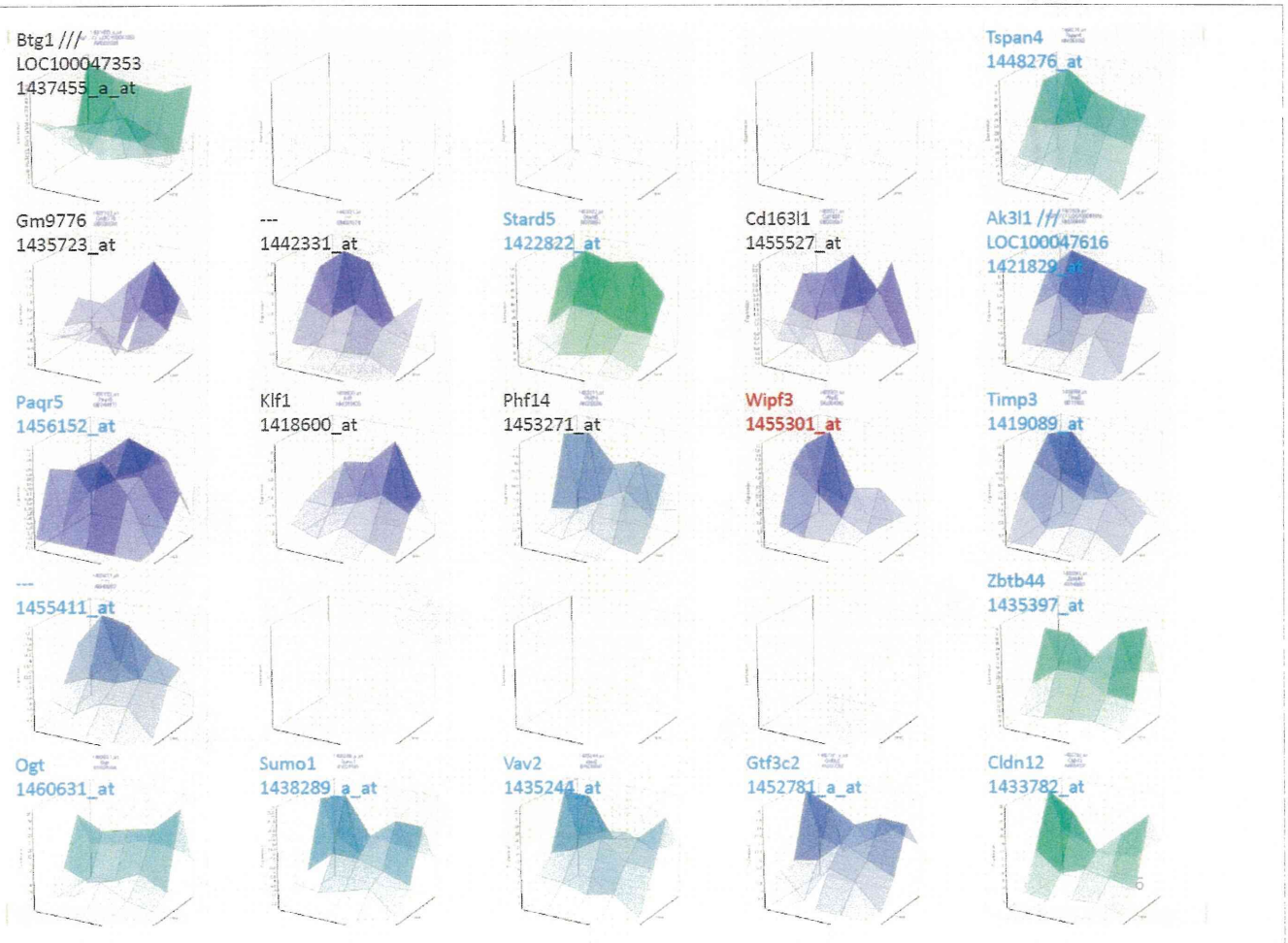





Usp2
1417169_at
Cpt1a
1438156_x_at
Usp2
1417168_a_at
Chn2
1428574_a_at
Ak2 ///
LOC100047005
1448451_at
Tppp
1452766_at
Marveld1
1439381_x_at
Cyp7a1
1422100_at
D0H4S114
1450839_at
E230008J23Rik
1439571_at
Abhd6
1419104_at
Tor1b
1417820_at
Plce1
1452398_at
1810063B05Rik
1435864_a_at
Rapgef1
1427006_at
Irf2bp2
1433634_at

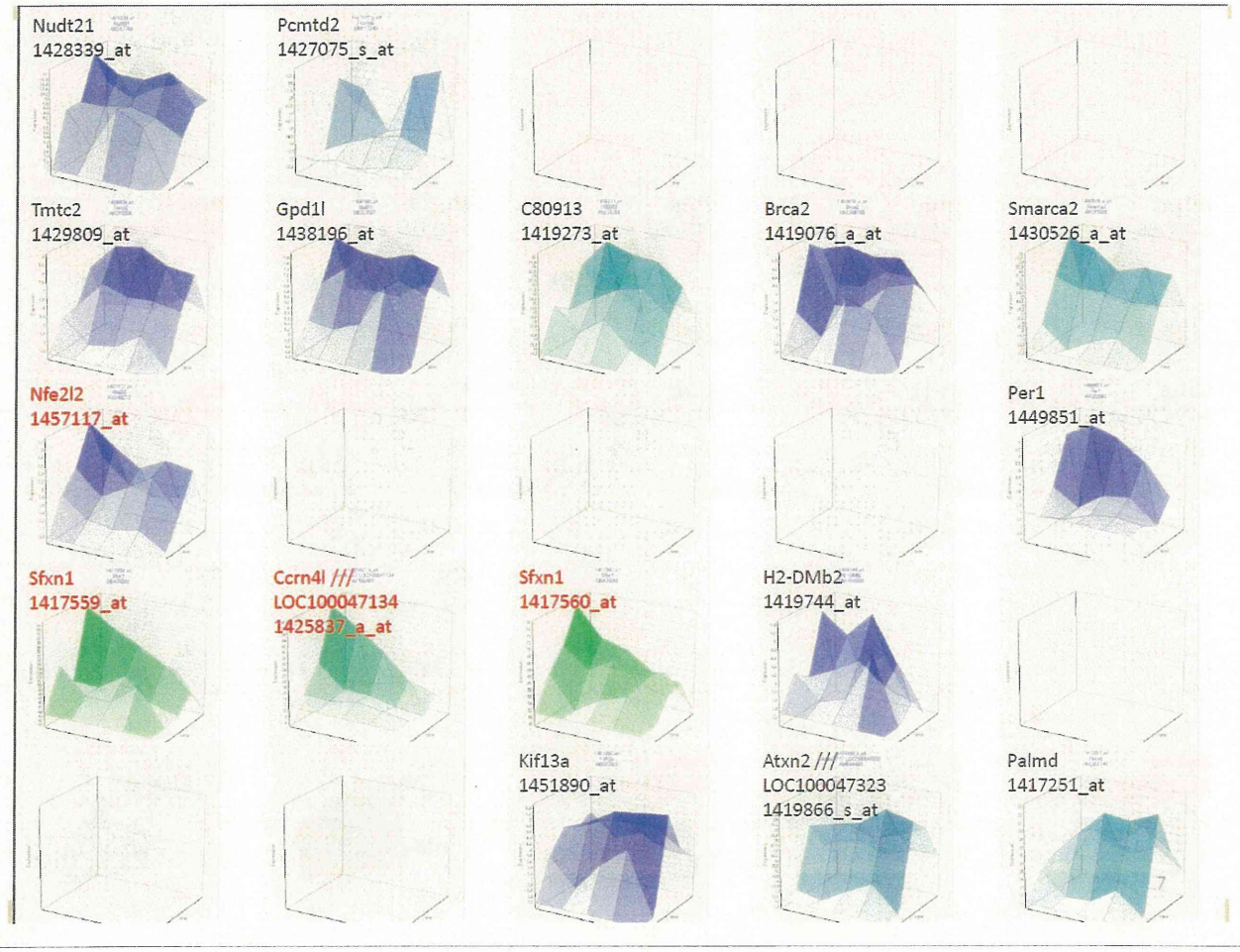

Insig2
1417982_at
Insig2
1417981_at
Insig2
1417980_a_at
Uhrf1bp1l
1451275_at
Vac14
1416231_at
Gda
1435748_at
Cyp2a4 /// Cyp2a5
/// LOC100047711
1422230_s_at
Plcg1
1435149_at
Timm8a1
1416346_at
Esrrg
1455267_at
Tex261
1423384_s_at
Abhd6
1419103_a_at
Mgll
1453836_a_at
Nfkbia
1448306_at
Abcf1
1420157_s_at
Txnl4a
1433698_a_at
Alpl
1423611_at
Cbx7
1433557_at
Smarca2
1452333_at


Dsc2
1421156_a_at
Ptplad1
1454698_at
Prkaa2
1434766_at
Tfam
1448224_at
Glyctk
1424995_at
Mrpl15
1425189_a_at
Tarsl2
1434738_at
Exoc6
1460576_at
Dock9
1450932_s_at
Celf2
1450069_a_at
Ptcd2
1423769_at
Pigx
1425134_a_at
Clpx
1423447_at
Tra2b
1419543_a_at
Zfp36l1
1422528_a_at
Slc30a10
1439934_at
Zfp36l1
1450644_at
Tuba4a
1417373_a_at
Rnf144b
1439153_at

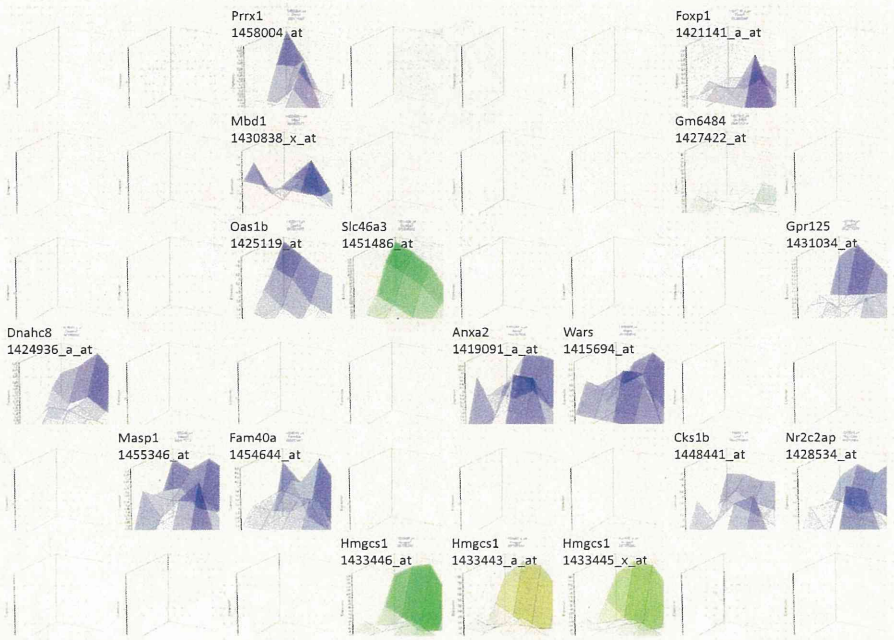

Btg1 ///
LOC100047353
1437455_a_at
Tspan4
1448276_at
Gm9776
1435723_at

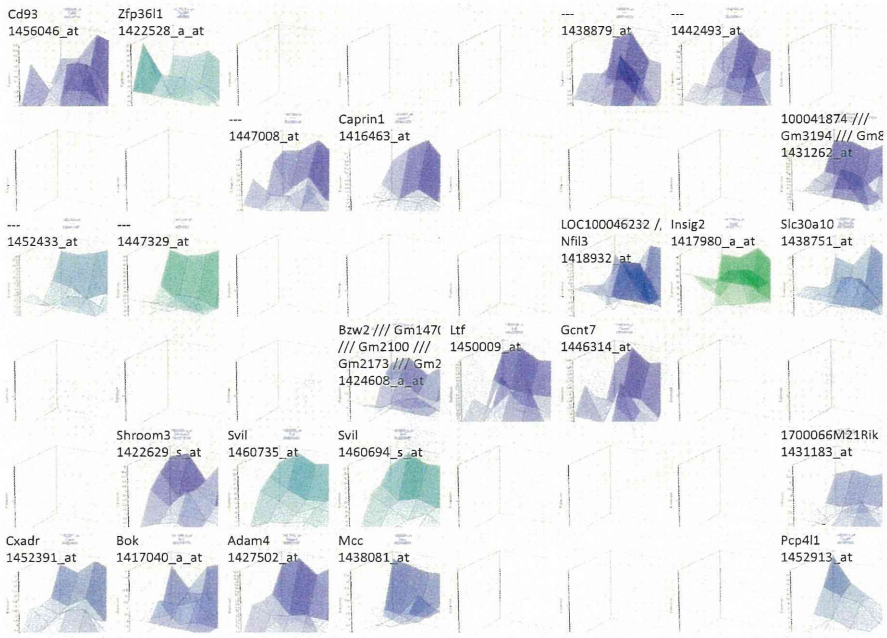
1442331_at
Stard5
1422822_at
Cd163l1
1455527_at
Ak3l1 ///
LOC100047616
1421829_at
Paqr5
1456152_at
Klf1
1418600_at
Phf14
1453271_at
Wipf3
1455301_at
Timp3
1419089_at

1455411_at
Zbtb44
1435397_at
Ogt
1460631_at
Sumo1
1438289_a_at
Vav2
1435244_at
Gtf3c2
1452781_a_at
Cldn12
1433782_at



TCDF 20,000 probes unsorted




Cd93
1456046_at
Zfp36l1
1422528_a_at

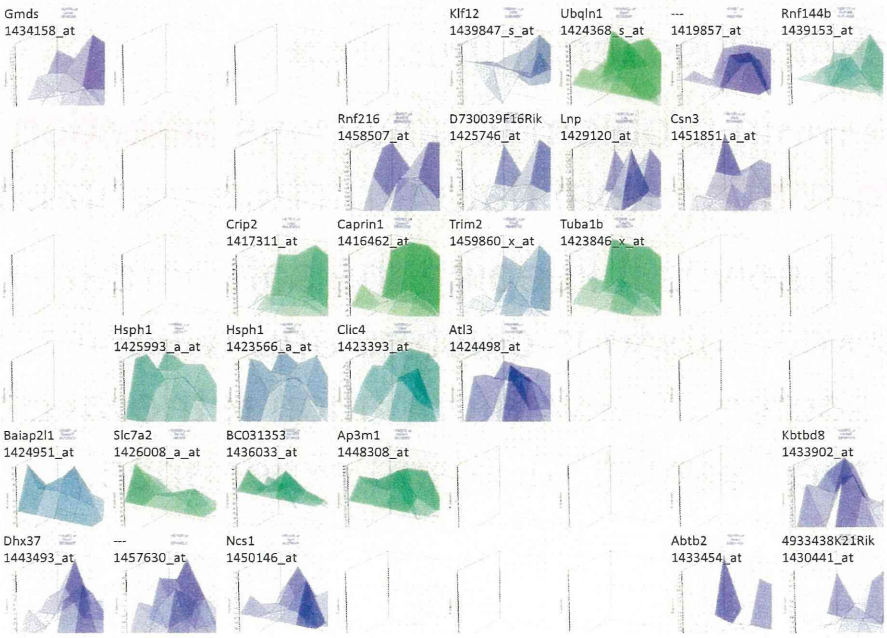
1447008_at
Caprin1
1416463_at

1438879_at

1442493_at
100041874 ///
Gm3194 /// Gm8
1431262_at

1452433_at

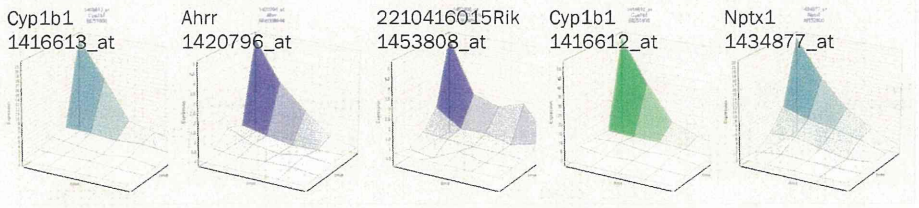
1447329_at
LOC100046232 /
Nfil3
1418932_at
Insig2
1417980_a_at
Slc30a10
1438751_at
Bzw2 /// Gm147(
/// Gm2100 ///
Gm2173 /// Gm2
1424608_a_at
Ltf
1450009_at
Gcnt7
1446314_at
Shroom3
1422629_s_at
Svil
1460735_at
Svil
1460694_s_at
1700066M21Rik
1431183_at
Cxadr
1452391_at
Bok
1417040_a_at
Adam4
1427502_at
Mcc
1438081_at
Pcp4l1
1452913_at


Gmds
1434158_at
Klf12
1439847_s_at
Ubqln1
1424368_s_at

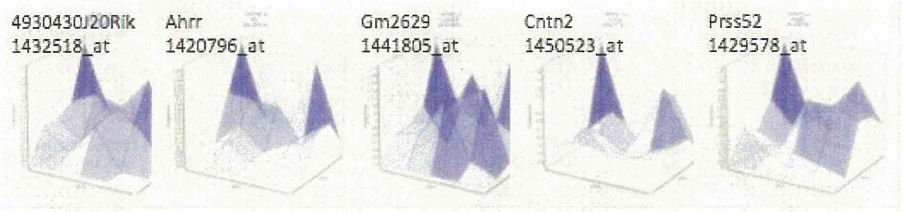
1419857_at
Rnf144b
1439153_at
Rnf216
1458507_at
D730039F16Rik
1425746_at
Lnp
1429120_at
Csn3
1451851_a_at
Crip2
1417311_at
Caprin1
1416462_at
Trim2
1459860_x_at
Tuba1b
1423846_x_at
Hsph1
1425993_a_at
Hsph1
1423566_a_at
Clic4
1423393_at
Atl3
1424498_at
Baiap2l1
1424951_at
Slc7a2
1426008_a_at
BC031353
1436033_at
Ap3m1
1448308_at
Kbtbd8
1433902_at
Dhx37
1443493_at

1457630_at
Ncs1
1450146_at
Abtb2
1433454_at
4933438K21Rik
1430441_at

TCDD



TCDF



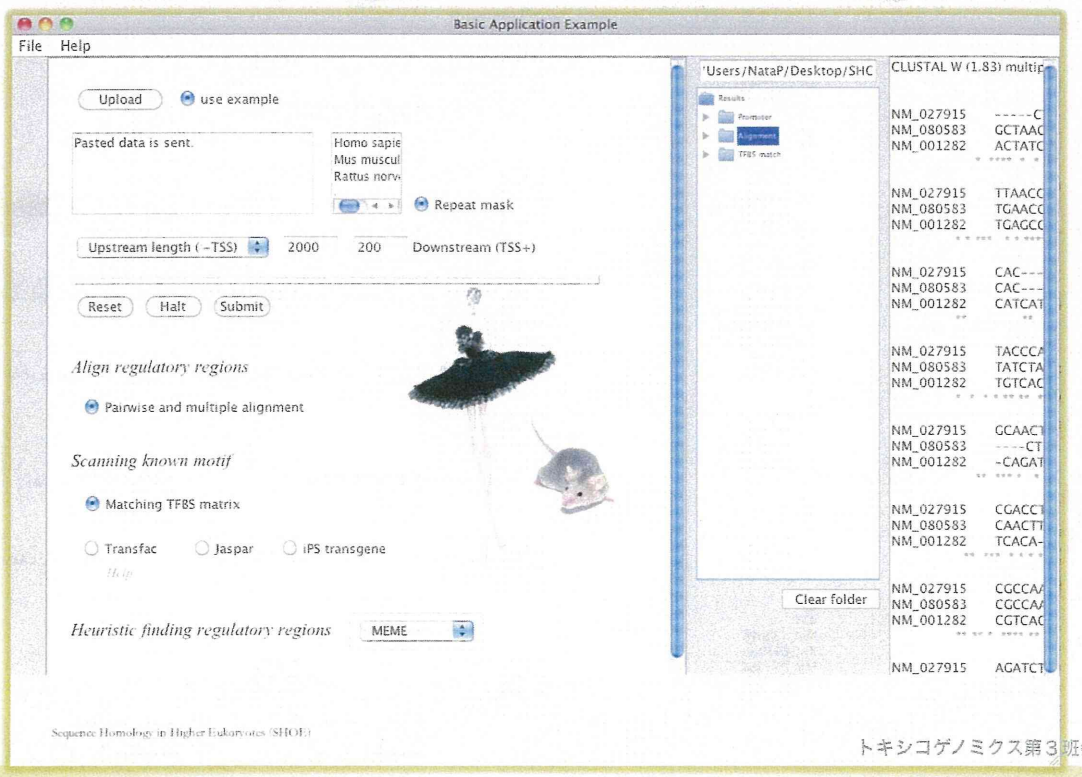
What was the problem?

* technical issues of interface
* the main problem to solve was **selection of information** from data structure computed.
* AGCT solves it keeping the soft border between obvious clusters and twilight zone, as source of new information retrieval basing on the User intuition and experience.

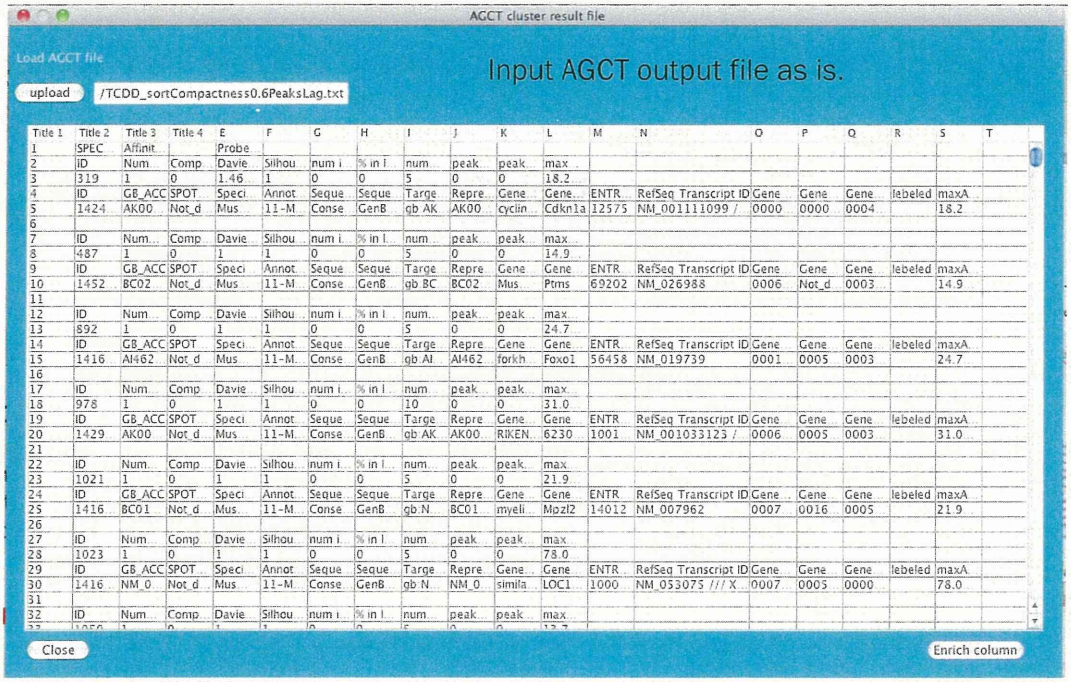
SHOE

Sequence Homology in Higher Eukaryotes

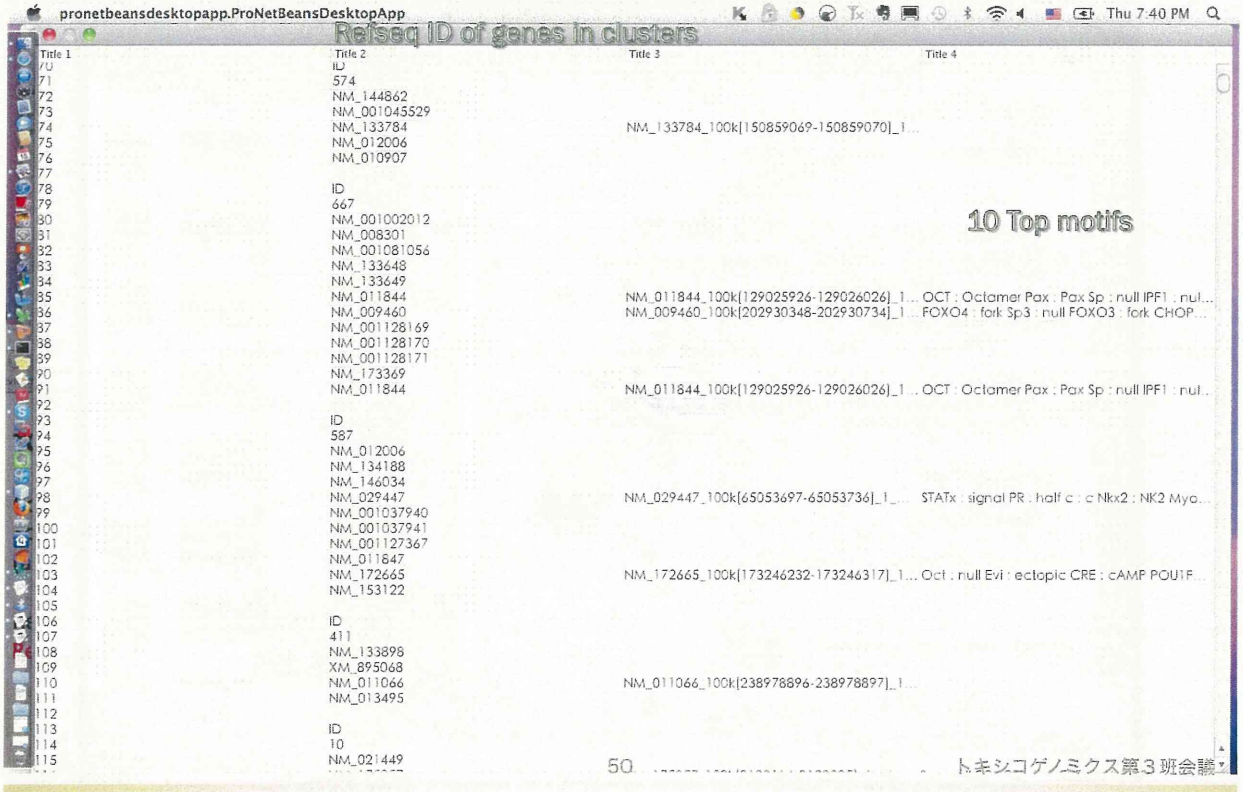
SHOE Phylogenetic footprinting tool



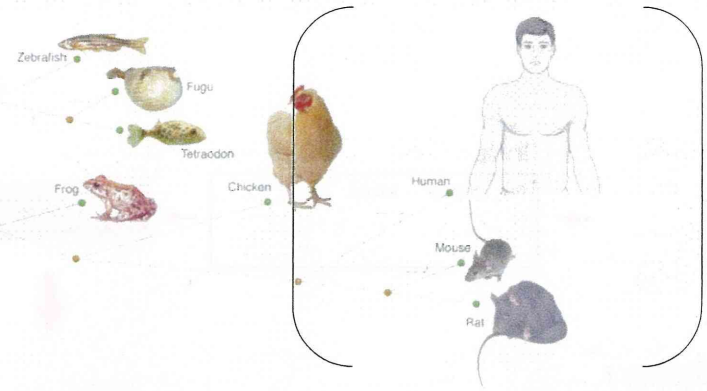
AGCT-Shoe Plug-in



SHOE RESULTS on HUMAN-MOUSE-RAT promoter analysis



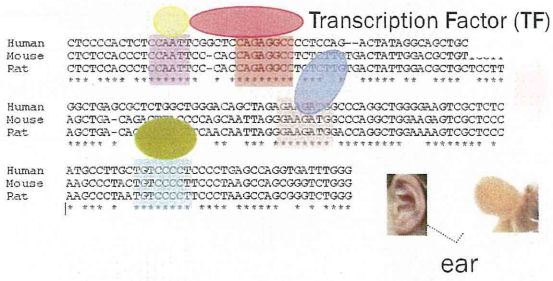
Phylogenetic footprinting finds evidence of functionality



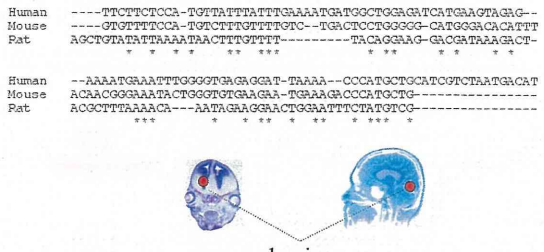
- TFs
- Brain-specific

(Suzuki et. al, 2004)

Regulation conserved !

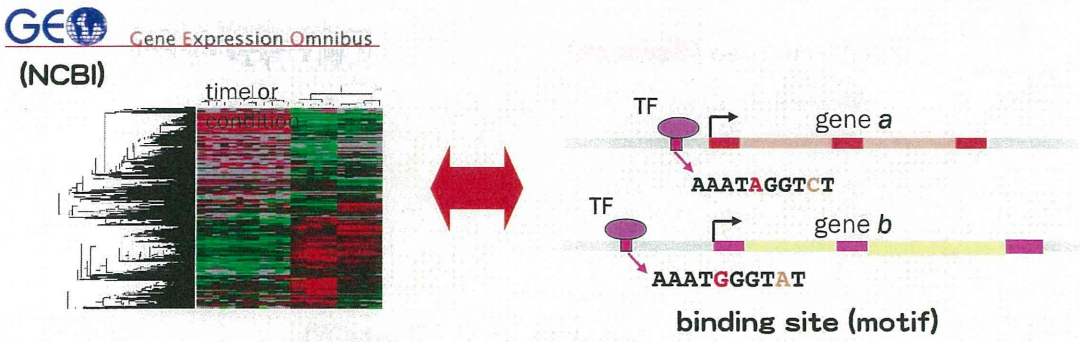


Regulation not conserved



Co-regulated genes

Similar way of expression suggests **regulation** by the same Transcription Factor (**TF**)



Motif discovery tools

MEME (Bailey *et.al.*)
Consensus (Stormo *et.al.*)
Gibbs sampler (Lawrence *et.al.*)
Yebis (Yada *et.al.*)....

&

Comparative genomics tools

MONKEY (Moses et.al)
FootPrinter (Tompa et.al)
PhyMe (Sinha et.al)
PhyloGibbs (Siddharthan et.al) ....

鼠 S H O E 人

STEP 1

DATASET of ORTHOLOGOUS PROMOTERS

refseq ID

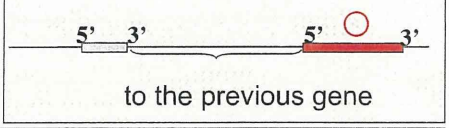
Human gene A
Human gene B
Mouse gene C
Rat gene D

Index file: 7,000genes

Gene	Human	Mouse	Rat
Acly	NM_198830	NM_134037	NM_016987
Aldoa	NM_184043	NM_007438	NM_012495
Fdft1	NM_004462	NM_010191	NM_019238
Gpd2	NM_000408	NM_010274	NM_012336

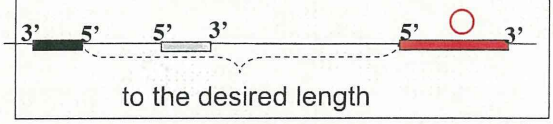
User' s orthologs

Option 1



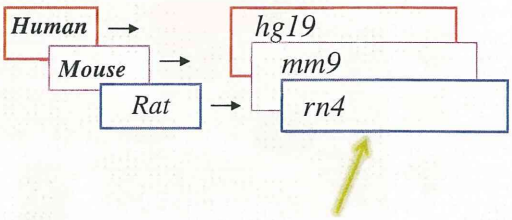
to the previous gene

Option 2



to the desired length

Genes locations INFO + Genomic Data



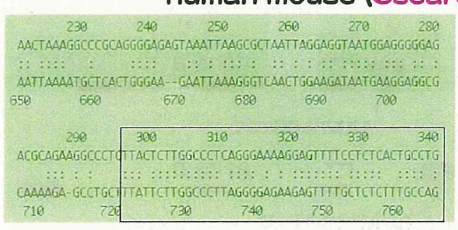
Updated!

鼠 S H O E 人

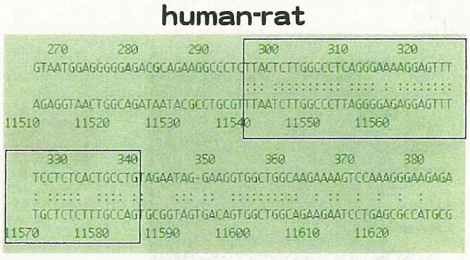
STEP 2

Construction of multiple alignment

human-mouse (Ssearch)



human-rat



human-mouse-rat (ClustalW)

STEP3

How good is the alignment?

Pattern frequencies tables for "good" and "random" alignments

74 patterns

A, T, G, C, - × AA, AT, AG, AC, A-, TT, TG, TC, T-, GG, GC, G-, CC, C-, --

835 orthologous alignments (238,800bp)

1260 random alignments (239,600bp)

$$MAscore = \log_{10}\frac{\Pr g1 * \Pr g2 * \Pr g3 \ldots \Pr gn}{\Pr r1 * \Pr r2 * \Pr r3 * \Pr rn} = \log_{10}\frac{\prod_{m} Pr(c|good\_alignment)}{\prod_{m} Pr(c|random\_alignment)},$$

c - probability of pattern in each column in *good alignment* and *random alignment* tables,
m - motif length.

BIOBASE BIOLOGICAL DATABASES TRANSFAC

STEP4

Calculation of PSSM score (PMscore)

498 human-rodent matrices

	①	②	③	④	⑤	⑥	⑦	⑧	⑨	⑩	⑪
A	4	5	3	0	4	3	3	2	1	1	1
C	1	2	0	0	0	0	0	1	3	4	6
G	2	2	7	2	3	7	0	4	3	1	1
T	3	1	0	8	3	0	7	3	3	4	2

$$PMscore = \sum_{i=1}^{m} \log_{10}\frac{count_{x_i} + pseudocount_{x_i}}{\sum_{x=A,T,G,C} count_{x_i} + \sum_{x=A,T,G,C} pseudocount_{x_i}}$$

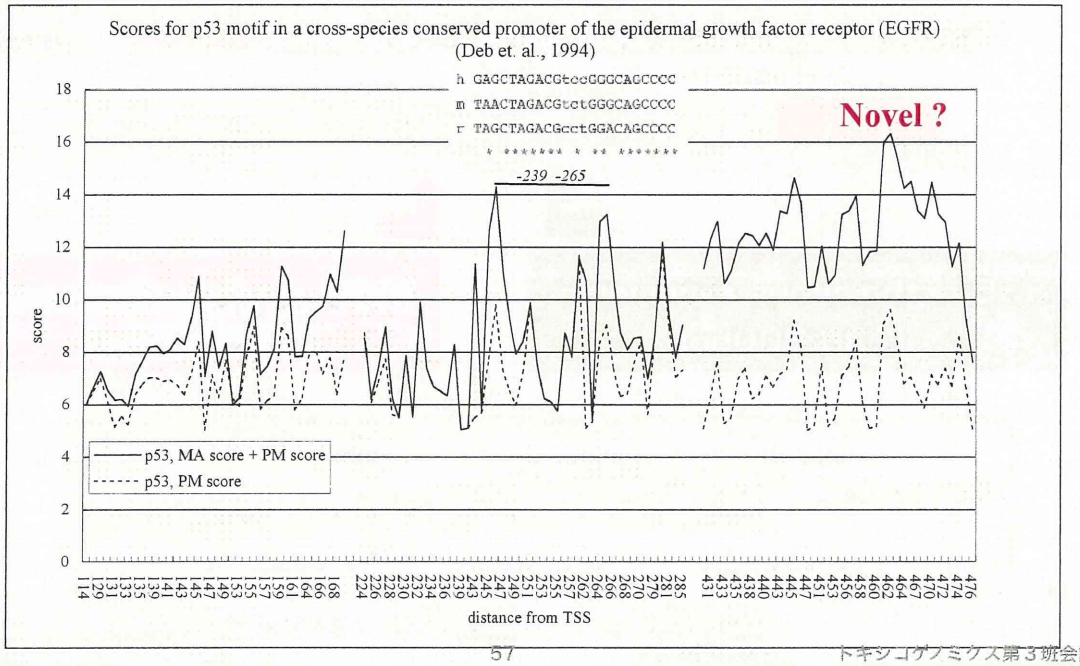
where ***pseudocount*** = 1
m is a motif length

Probability of motif at each position in human sequence

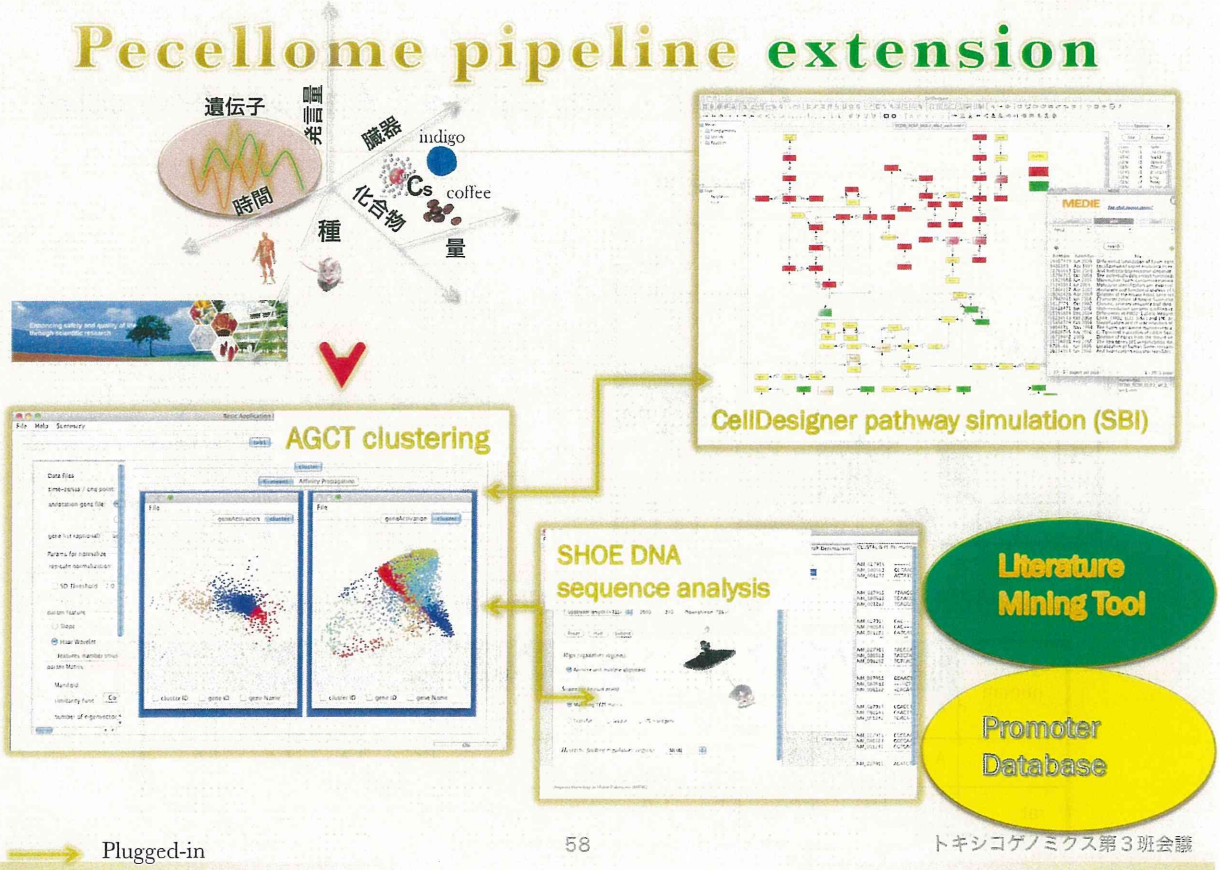
human	-0.48	-0.39	-0.27	-0.22	-0.57	-0.27	-0.27	-0.57	-0.57	-0.47	-0.33	-4.46
	A	A	G	T	G	G	T	T	C	C	C	
mouse	A	T	G	T	G	G	T	C	A	C	C	
rat	A	T	C	T	G	G	T	A	A	C	C	

Finding motif candidates

Multiple alignment score (MA) + Transfac PSSM score (PM)



Pecellome pipeline extension



Acknowledgements

* Hiroaki Kitano (Sony CSL)
* Keigo Oka (Tokyo University)
* Frank Nielsen (Sony CSL)
* Richard Nock (University of Martinique)

TCDD / TCDF

2012 Feb

ym

TCDD

- 2,3,7,8-Tetrachlorodibenzo-p-dioxin (TCDD)

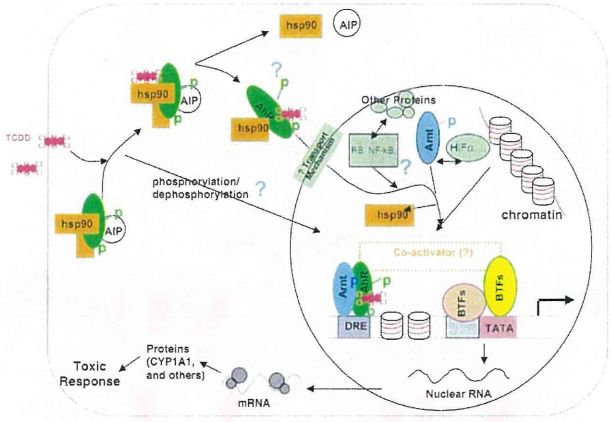


Figure 2-2. Cellular Mechanisms for Ah Receptor Action

http://www.epa.gov/ncea/pdfs/dioxin/nas-review/pdfs/part2/dioxin_pt2_ch02_dec2003.pdf