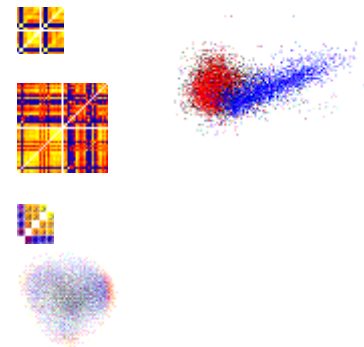# RNAnet a Map of Human Genes

alt.splice

Mathematical and Biological Sciences,

Essex University

W. B. Langdon

Introduction

- RNAnet is a firefox based web tool to explore human Affymetrix mRNA data in GEO using Ensembl gene and exon definitions

- Correlation heatmaps and scatter plots 2757 GeneChips for
  - All probes in a probeset
  - All probes uniquely mapped to an exon
  - Arbitrary (user defined) list of probe
  - All exons (approx 290 million)

- Web based detection of spatial flaws and single CEL file at a time quantile normalisation.

- Conclusions

RNAnet A Map of Human Gene Expression

- GEO contains 10s of thousands of published Affymetrix GeneChips.
- So far, RNAnet uses all the HG-U133 2+
- Affymetrix probes mapped to Ensembl exons (and so to Ensembl genes).
- Typically 2-11 PM probes per exon.
- Correlation between all probes in exon (should be free of alternative splicing etc)
- Best probe chosen (no G-spots!). 24132 exons, at least one exon from 14288 Ensembl genes
- All v all exon correlations calculated across 2757 arrays (290 million correlations)

RNAnet - Results

- How to present 290 million numbers?
  - Static, eigen analyse (PCA) group exons with similar interactions.
  - Interactive, user selects genes or exons of interest.
 - Combined. PCA map guides user as to what might be interesting.

Correlations between Human Exons
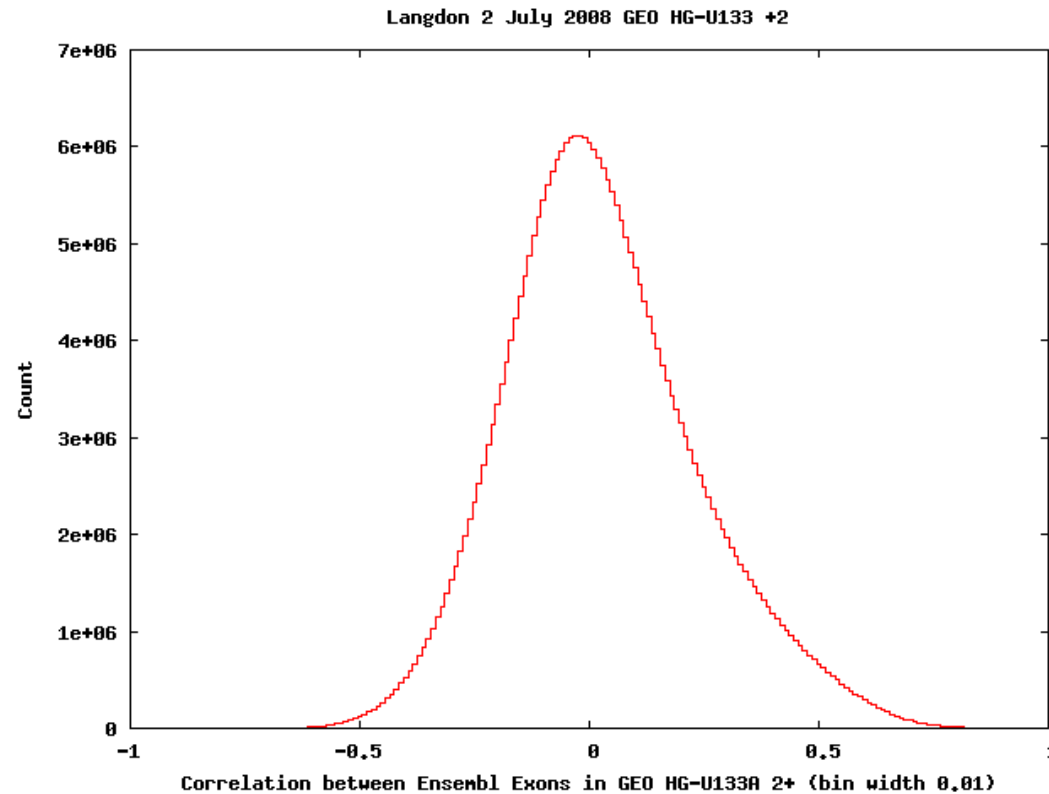
- Most correlations small.
- Slightly more +correlations than –
- Correlations are statistically significant

Eigen analysis of Human Exon Correlations

- Correlation of 24132×24132 exons as square symmetric matrix.

- Eigen values fall rapidly. Power law typical of "small world" networks. Only first 500 needed.

- Most eigen vectors short, <8000.

- Eigen analysis can approximate 300 million correlations with 400thousand numbers.

Langdon 2 July 2008 GEO HG-U133 +2

Eigenvalue
Power law x**(-5/4)
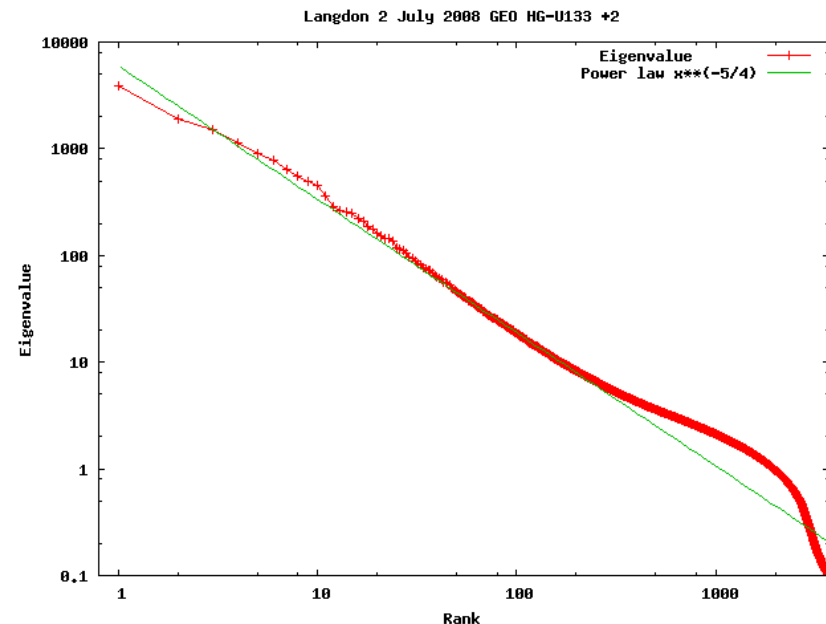
Systems Biology "Small World" Network

- The strong links (< -0.410 or >0.578) form a sparse network, which nonetheless connects nearly all exons.

- Average number of strong links is 32 (48)

- 2920 of exons have no strong links

- 350 exons are highly correlated with >2000 others.

Langdon 3 July 2008 GEO HG-U133 +2

#exons correlation < -0.410
#exons correlation > 0.578

Count (decibel bins)
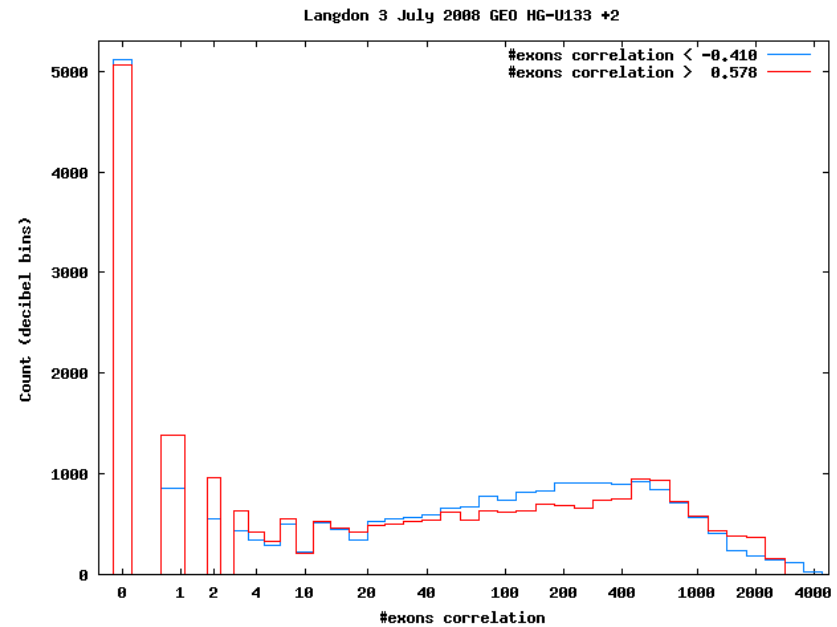
#exons correlation

Best Projection 24132 dimensions→2

- Linear projection onto first two eigenvectors (PCA) best preserves the relationships between 24132 exons.

- Exon placement given by PCA, colour given by average local correlation.

- Exons with weak correlations lie in centre, strong at edges, far from anti-correlated exons.

RNAnet, gene by cursor

Dragging cursor over dot reveals link to table of correlations for that exon.
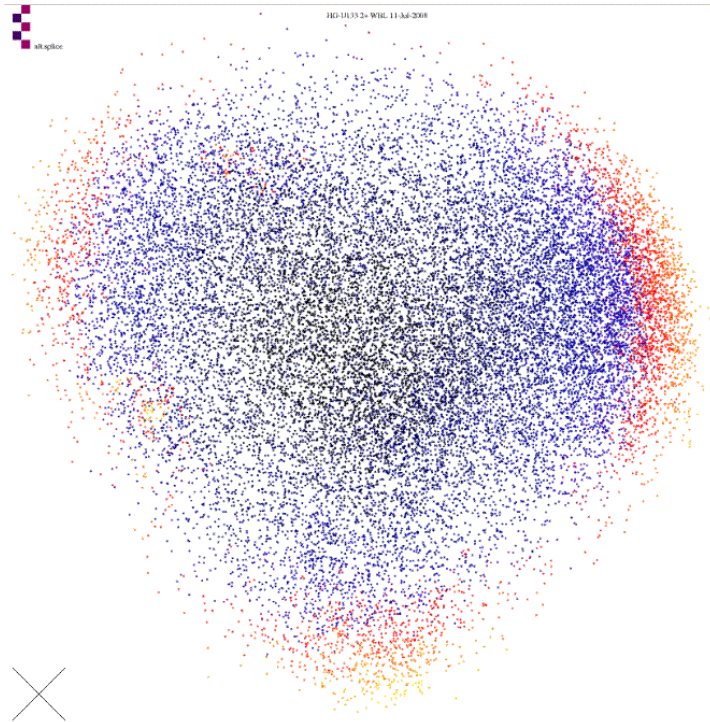
http://bslin11ha.essex.ac.uk/users/wlangdon/rnanet/exon.php?1442631#SMOC2_24_p
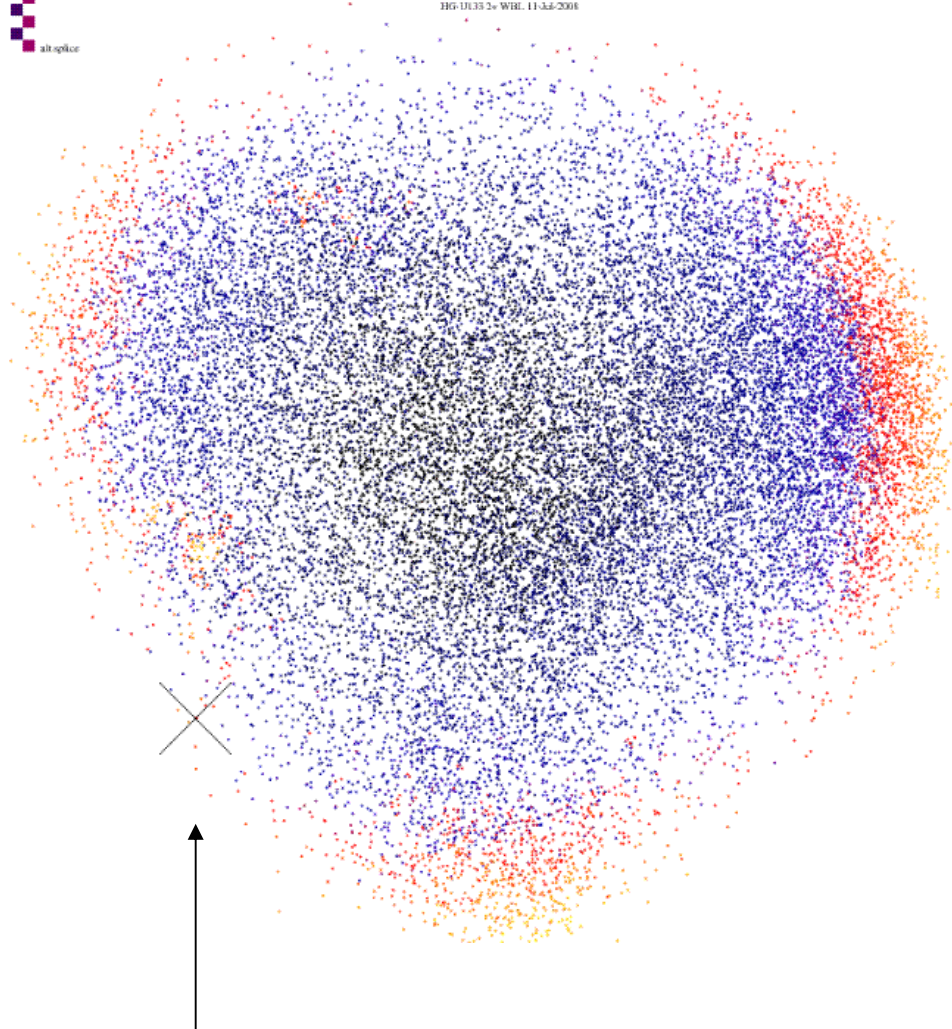
Ensembl id

Ensembl Gene name

Exon number in gene

Type of gene (p = protein coding)

Black-blue low average correlation with 10 nearest neighbours. Red-Yellow strong correlation.
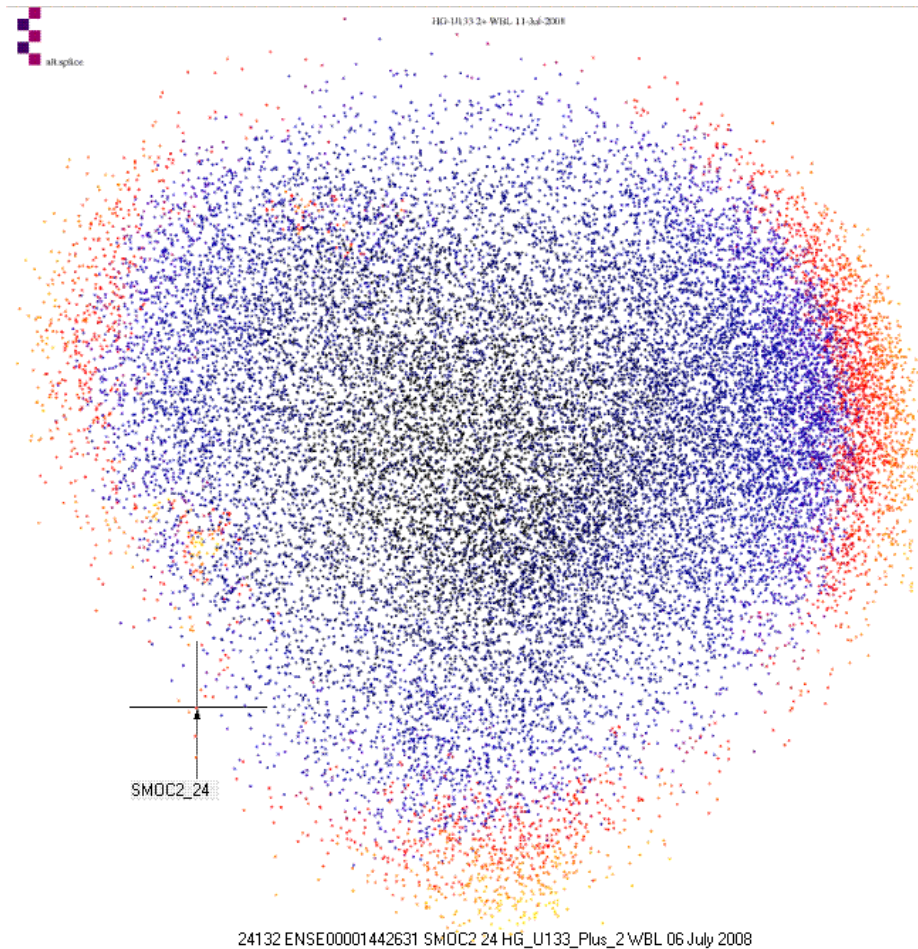
Using crosshairs

Expression of Human Genes

_	ENSE 1458959	UBR2 15		
_	ENSE 1080659	ADH4 1		
_	ENSE 1112712	NOVA2 1		
_	ENSE 1210566	USP28 1		
_	ENSE 889894	FANCM 23		
_	ENSE 1162040	PSG11 4		
_	ENSE 976713	Q96IP2 HUM		
_	ENSE 1434670	Q9P184 HUI		
_	ENSE 1212913	AQP11 2		
_	ENSE 1442631	SMOC2 24		

Find gene [   ]   [ ] Clear graphs -0.410

Up to ten exon/genes closest to × shown above.

W. B. Langdon, Essex

10

Find genes using text search

Expression of Human Genes

ENSE 1442631   SMOC2 24

Find gene   SMOC2   1   Clear graphs

Enter name of gene or Ensembl number.

Exon highlighted by crosshairs.

SMOC2_24

Locking gene saves it whilst search for others

Expression of Human Genes

1538 ENSE 1442631 SMOC2_24 ☑ 10
Find gene SMOC2 1 Clear graphs

SMOC2_24

24132 ENSE00001442631 SMOC2 24 HG_U133_Plus_2 WBL 06 July 2008

Click on small box next to exon/gene name.

Locking allows 9 other slots to be used to find other genes

Preloads correlations and displays thumbnail distribution of correlations. Number outside thresholds.

Highlight strong links
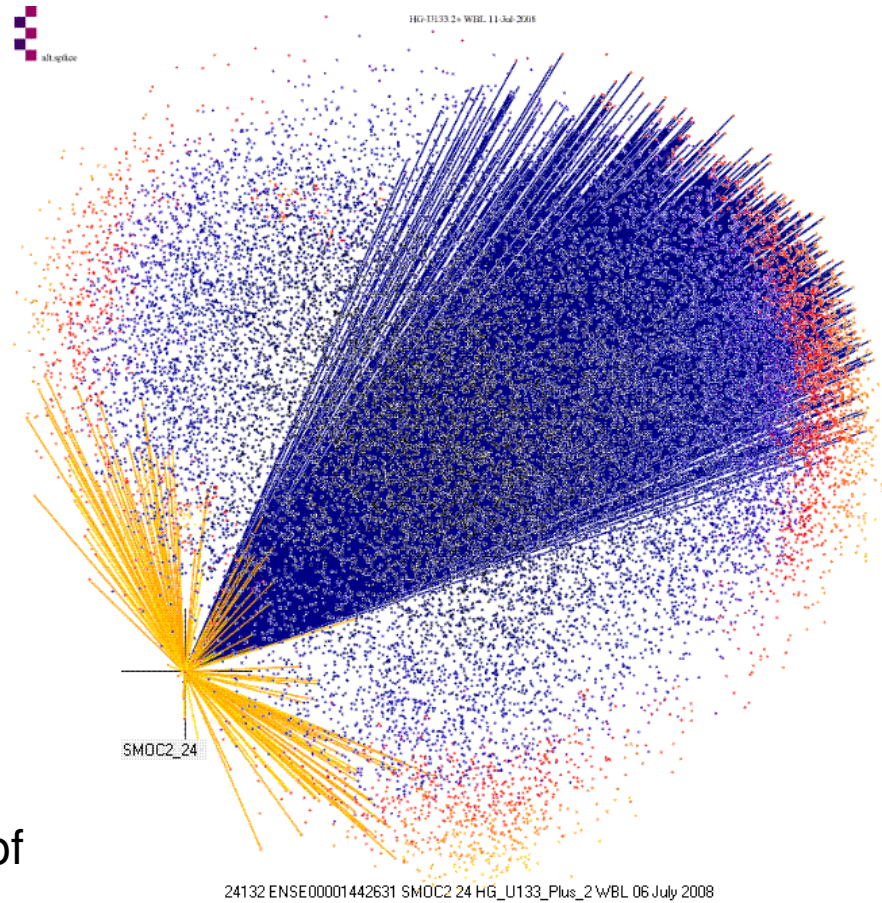
Press button to left of
exon/gene names.

Adjust thresholds

HG-U133 2e WBL 11-Jul-2008

Expression of Human Genes

SMOC2_24

24132 ENSE00001442631 SMOC2 24 HG_U133_Plus_2 WBL 06 July 2008

29 | ENSE 1442631 | SMOC2_24 | 10

Find gene | SMOC2 | 1 | Clear graphs | -1 | 0.666 | Thresholds

Setting negative threshold to -1 prevents any negative correlations being plotted.

W. B. Langdon, Essex

14

SMOC2 (SPARC-related modular calcium-binding protein 2 precursor)
LRRC6 (Leucine-rich repeat-containing protein 6)

Expression of Human Genes

Search for gene lrrc6 by name. Two exons found.

No correlation with SMOC2.

Click on gene name for short text description (leads to Ensembl)

LRRC6 (Leucine-rich repeat-containing protein 6)
Correlation between exons 4 and 5 only

Expression of Human Genes

Click to data as table

Hyperlink on exon leads to table of strong correlations for that exon.

Can follow links to other exons, or to short description of gene.

Correlation between Human Exons and SMOC2[24] in GEO

The correlation cofficient across 2757 GeneChips taken from GEO was calculated for all human ENSE00001442631 (heatmap firefox*) is reported below.

Links to correlation heatmap. (firefox is dynamic)

Link to Ensembl information on this gene.

Correlation coefficient ENSEMBL id gene exon within gene

Corr	ENSEMBL id	gene
1.00	ENSE00001442631	SMOC2
0.76	ENSE00001464472	ABCA1[6]
0.74	ENSE00001212913	AQP11[1]
0.73	ENSE00000898471	na[1]
0.73	ENSE00001489755	CAMKK1[3]
0.72	ENSE00001446506	FCRL2[3]
0.71	ENSE00001162040	PSG11[4]
0.71	ENSE00001168361	TRIM4[3]
0.71	ENSE00000973532	KLHL3[10]
0.71	ENSE00001323787	USH1G[1]
0.71	ENSE00001233056	GPHB5[3]
0.70	ENSE00001169431	PRKCE[7]
0.70	ENSE00001244073	DSCR10[3]
0.70	ENSE00001539246	C17orf87[4]
0.70	ENSE00000976713	Q96IP2_HUMAN[5]
0.69	ENSE00000830509	WAPAL[4]
0.69	ENSE00001319337	PCSK6[2]
0.69	ENSE00001459224	JPH2[7]
0.69	ENSE00001366686	NHS[18]
0.69	ENSE00001268285	WFDC9[3]
0.69	ENSE00001285895	RCHY1[1]
0.68	ENSE00001161627	MYEOV2[2]
0.68	ENSE00001420899	KRTAP4-5[1]
0.68	ENSE00001316605	KCNH7[10]
0.68	ENSE00000889894	FANCM[5]
0.68	ENSE00001409328	RDH14[4]
0.67	ENSE00000882056	HYAL4[1]
0.67	ENSE00001524789	PGM5P1[11]
0.67	ENSE00000941114	P1768_HUMAN[5]
0.67	ENSE00000959102	FCRL5[3]
0.66	ENSE00001450679	C10orf80[21]
0.66	ENSE00001422916	DPH5[1]
0.66	ENSE00001415106	NP_115661.1[8]
0.66	ENSE00001126477	RLN3[3]
0.66	ENSE00001046480	NR_003081.1[1]
0.66	ENSE00001278386	CA5B[13]
0.66	ENSE00000856851	IL1F7[4]
0.66	ENSE00001299646	RGS6[15]

2 clicks into Ensembl

LRRC6 (Leucine-rich repeat-containing protein 6)
SMOC2 (SPARC-related modular calcium-binding protein 2 precursor)

206483_atpM2 v 243946_atpM1 Log Quantile Normalised HG_U133_Plus_2 2757 WBL 04 Aug correlation -0.018 (2706 cel files)
206483_atpM2 v 243946_atpM1 correlation 0.281 (9 cel files)

Expression of Human Genes

ENSE887822   ☐MM/PM
ENSE1442631   ☐MM/PM   probes
GSE5372

☐plot ☐clear
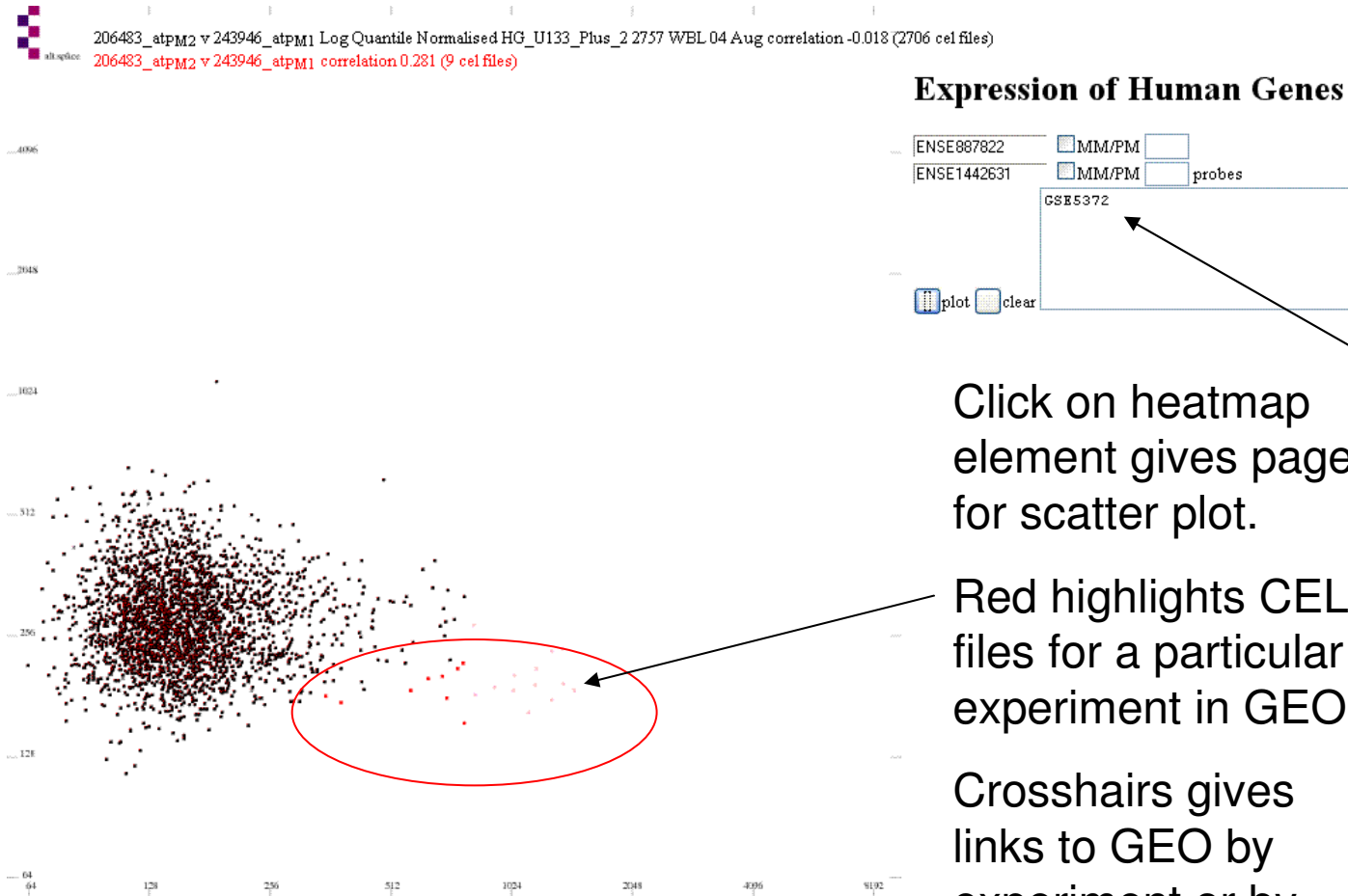
Click on heatmap element gives page for scatter plot.

Red highlights CEL files for a particular experiment in GEO.

Crosshairs gives links to GEO by experiment or by CEL file.

W. B. Langdon, Essex

19

GEO description of experiment

206483_atp$_{M2}$=771 243946_atp$_{M1}$=156 GSE5372 GSM122547

Crosshairs on scatter plot links to both GEO description of experiment and of individual CEL file.

NCBI · GEO · **Accession Display** ⓘ   Not logged in | Login ⓘ

GEO help: Mouse over screen elements for information.

Scope: [Self ▾]   Format: [HTML ▾]   Amount: [Quick ▾]   GEO accession: [GSE5372]   [GO]
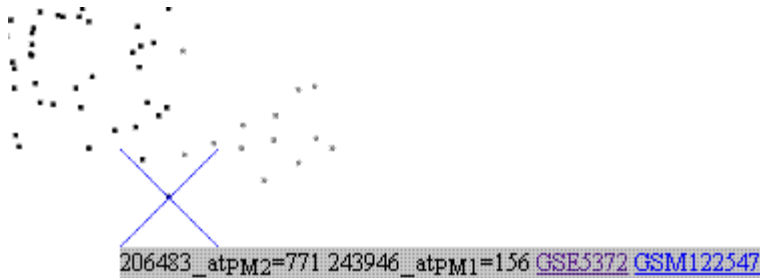
Series GSE5372   Query DataSets for GSE5372
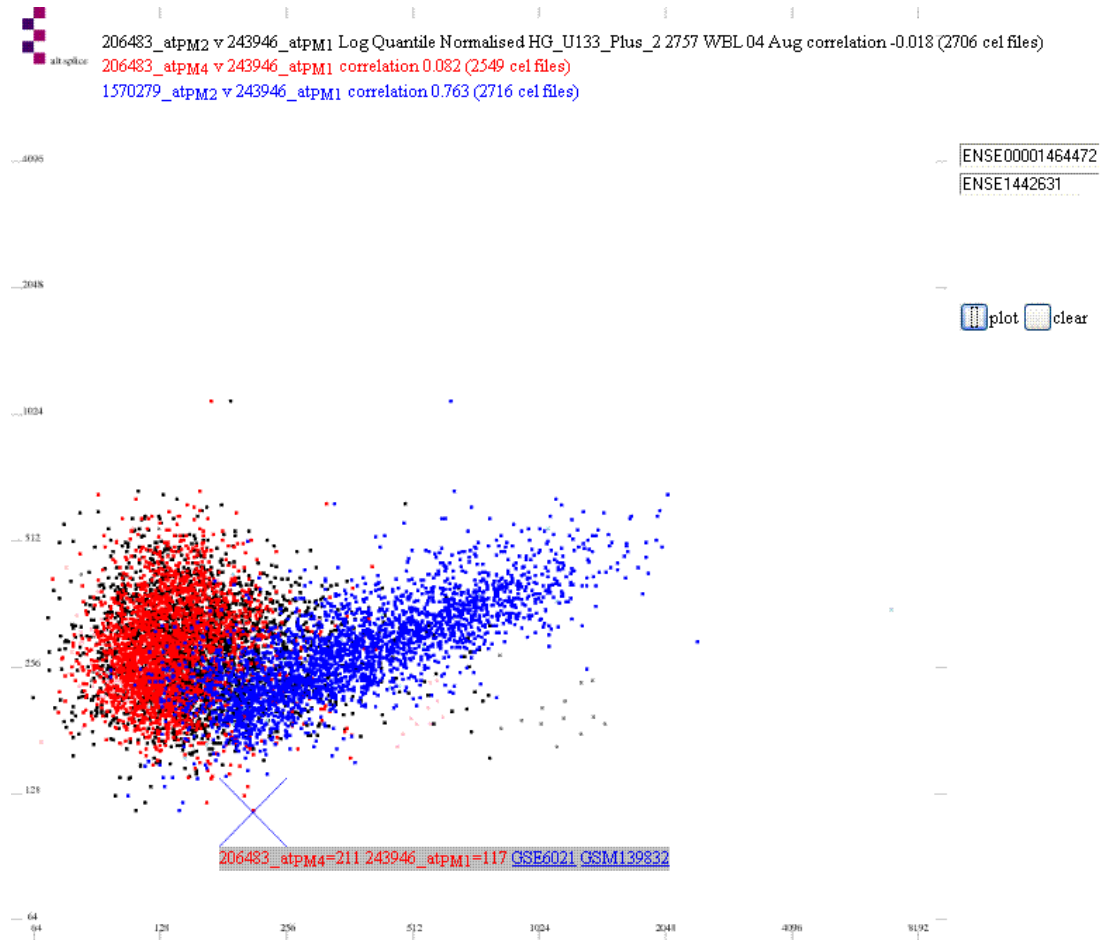
Status   Public on Dec 12, 2006
Title   airway epithelium, large airways, pre and post-mechanical injury
Organism(s)   Homo sapiens
Experiment type Expression profiling by array
Summary   Responses of the Human Airway Epithelium Transcriptome to In Vivo Injury
                To identify genes participating in repair of the human airway epithelium
                following injury, we used bronchoscopy and brushing to denude the airway
                epithelium of healthy individuals, sequentially sampled the same region 7
                and 14 days later, and assessed the recovered epithelium for relative
                levels of gene expression using Affymetrix high-density oligonucleotide
                microarrays with TaqMan PCR confirmation. Histologic assessment showed
                that the epithelium was denuded immediately following injury, at 7 days
                the epithelium was completely covered but partially de-differentiated, and
                by 14 days there was close to normal proportions of differentiated cells.
                Gene expression analysis was carried out with both the Affymetrix
                Microarray Suite 5.0 and Robust Multi-array Average algorithms, applying a
                multiple test correction to identify bona fide changes in gene expression. At
                day 7, there were substantial differences in the gene expression pattern
                compared to the resting epithelium, with a distinctive airway epithelial
                "repair transcriptome" of actively proliferating cells in the process of
                re-differentiation. The repair transcriptome at 7 days was dominated by
                genes encoding proteins involved in cell cycle regulation, transcription,
                signal transduction, metabolism and transport. Interestingly, the majority
                of cell cycle genes differentially expressed at day 7 belonged to the G2 and
                M late phases of the cell cycle, suggesting that the proliferating cells are
                relatively synchronized 1 wk following injury. At 14 days post-injury, the
                majority of the gene expression changes observed at day 7 were no
                longer observed, with the expression profile similar to that of resting
                airway epithelium. Using a class prediction algorithm, a group of 50 genes
                dominated by cell cycle genes, that represent a human airway epithelial
                "repair signature" was identified. These observations provide a baseline of
                the functional gene categories participating in the process of normal
                human airway epithelial repair that can be used in future studies of injury
                and repair in human airway epithelial diseases.
                Keywords: response to airway injury

W. B. Langdon, Essex   20

Colour allows multiple plots

LRRC6 v SMOC2

LRRC6 v SMOC2

ABCA1 V SMOC2

206483_atp$_{M2}$ v 243946_atp$_{M1}$ Log Quantile Normalised HG_U133_Plus_2 2757 WBL 04 Aug correlation -0.018 (2706 cel files)
206483_atp$_{M4}$ v 243946_atp$_{M1}$ correlation 0.082 (2549 cel files)
1570279_atp$_{M2}$ v 243946_atp$_{M1}$ correlation 0.763 (2716 cel files)

ENSE00001464472
ENSE1442631

plot clear

206483_atp$_{M4}$=211 243946_atp$_{M1}$=117 GSE6021 GSM139832

Dynamic Heatmap

Correlation of Expression of Human Genes in GEO

Exon ids for SMOC2 and LRRC6 (both) entered into ENSE lookup. Affymetrix probe ids generated.

Heatmap has links to scatter graph.

```
326 ENSE00000887822     ENSE lookup

243946_at.pm1* 206483_at.pm4* 206483_at.pm5 206483_at.pm1 206483_at.pm2*
```

☑ PM only ⫶ Heatmap ⬚ text done

```
                                                                        1 2 3 4 5
5 206483_at.pm2 528,195 1953 ACAACGGGTCATTTGGTCATCTGCA 154 1.43  0  6 6 9
4 206483_at.pm1 222,769 1917 GTGAAACCCGATAGTAGTTCTGCTA 116 1.44  0  6 6   9
3 206483_at.pm5 595,585 2030 GAAAACTACCTCGGACAGGAGCAGA 135 1.35     5   6 6
2 206483_at.pm4 1141,35 1998 ATCACAGGTGGTCAGCGAGCATTCA 136 1.28       5 6 6
1 243946_at.pm1 278,673    36 GATGCAATGGTGGTGTCCTCCAGAC 281 1.30     1   0 0
```

Dynamic Correlation Calculation

As previous slide, but as for table of numbers rather than heatmap.

Correlation of Expression of Human Genes in GEO

326 ENSE00000887822   ENSE lookup

243946_at.pm1* 206483_at.pm4* 206483_at.pm5 206483_at.pm1 206483_at.pm2*

☑PM only   Heatmap ☐ text Finished

```
-1 5 2757 HG_U133_Plus_2 WBL 04 August 2008
1170454 5.63789065768 0.264438211164 2734
350625 4.90948516321 0.245293624745 2570
350626 4.90572062077 0.301140110653 2724
350622 4.74945912595 0.365829802292 2719
350623 5.0387776394 0.356291585869 2729
0.0821820053463
0.0226527290256
-0.0445339024614
-0.0182262596875
0.505728501714
0.571402060642
0.624526466434
0.585644597001
0.604268377336
0.855221219804
```

Conclusions

- Correlations between human exons are highly non-random. The network formed by pairs of strongly correlated exons is a "small world":
    - No exon is strongly correlated with all others
    - 2% exons are strongly correlated with >1000 others
    - Most exons are strongly correlated with <48 others
    - By using multiple steps most exons can be reached
    - 12% of exons not strongly correlated with any others
    - Some "small world" power laws found.

- RNAnet is an interactive tool to explore RNA expression (protein gene and non-protien coding) of thousands of published 3' GeneChips covering virtually all medically interesting human tissues.

- Links to Ensembl and GEO.

- One CEL quantile normalisation and error detection.

- Future: extend to other arrays, other organisms, GO link?

END

Questions

- R code to average, detect flaws etc. under review. Release expected soon.

- Technical report CES-486.

- RNAnet, 150000 heatmaps, error detection, normalisation etc:

  http://bioinformatics.essex.ac.uk/users/wlangdon