

# ISH data interaction matrix analysis

Avinash Das

June 18, 2012

## 1 Initial Analysis

### 1.1 Sample matrix

I performed some initial analysis on the sample interaction data present at [http://www.eurexpress.org/ee/tools/EurExpressCluster\\_01.zip](http://www.eurexpress.org/ee/tools/EurExpressCluster_01.zip). At first, I tested whether there are certain classes of genes those are expressed in different anatomy ( those are significantly more than expected, like the housekeeping genes analysis similar to [?]). Figures 1 , 2 and 3 respectively shows the histogram of strongly expressed, mildly expressed ( genes those are highly or mildly expressed) and weakly expressed genes ( highly, mildly or weakly expressed). I have also overlay corresponding null distribution for each of the cases. I assumed binomial distribution with mean equal to observed mean as the null distribution. As seen from the figure the class genes are significantly compared to null distribution, particular in (1) classes where a gene is expressed in fewer number of anatomies (Probably this is not very interesting). (2) classes where a gene is expressed many tissues.

Next, I performed some annotation analysis whether these over represented genes are enriched in any particular annotation. I further divided the genes which are strongly expressed in many anatomies (and occur more than expected by null distribution P-value < E-8) into following groups:

- Strong40: The genes which are strongly expressed in more than 40 anatomy are enriched in membrane related annotation. The annotation have two prominent clusters: one with membrane and trans-membrane function; other with bi-sulfide related functions.
- Strong20: The genes which are strongly expressed in more than 20 anatomy are enriched in homestatis, adhesion and DNA repair functions, in addition to membrane and bi-sulfide functions.
- Strong8: The genes which are strongly expressed in more than 8 anatomy are enriched in growth factor activity, glycoprotien and muscle development in addition to functionalities of strong40 and strong20.

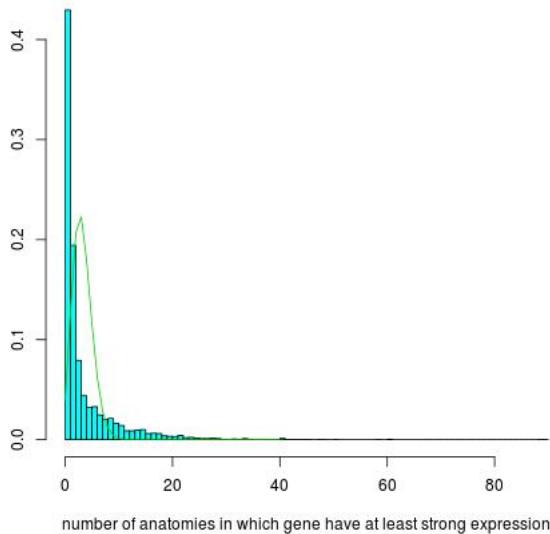


Figure 1: Strongly expressed gene histogram

It seems the genes, which are strongly expressed across many anatomies, are enriched in house-keeping functions (note I have reported only those annotation clusters which are significantly compared to all mouse genes as background).

## 1.2 Interaction matrix 5510x811

I performed analysis similar to performed on the sample matrix. At first, I compared the distributions of number of tissue types in which gene expressed with its null distribution. The null distribution assumed to be binomial distribution (with  $n = 811$  and  $p = \frac{\text{non-zero entries in interaction matrix}}{\text{total size of interaction matrix}}$ ). I performed 3 separate analysis 1. Strong expression only (figure 4), 2. Medium or strong expression (figure 5)and 3. Weak, medium or strong expression (figure 6).

## 1.3 Unsupervised clustering and visualization

The complete matrix is shown in figure /reffig:intM. Since the matrix is sparse, any structure is not discernible. In order to find any prominent structure in the interaction matrix. I performed unsupervised clustering by two method: 1. Principal component analysis (figure 8, 9) 2. Hierarchical clustering (figure /reffig:intMHierarchical).

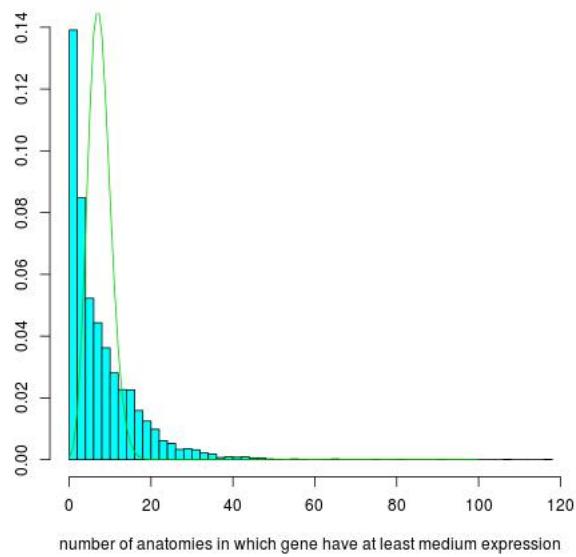


Figure 2: mildly expressed gene histogram

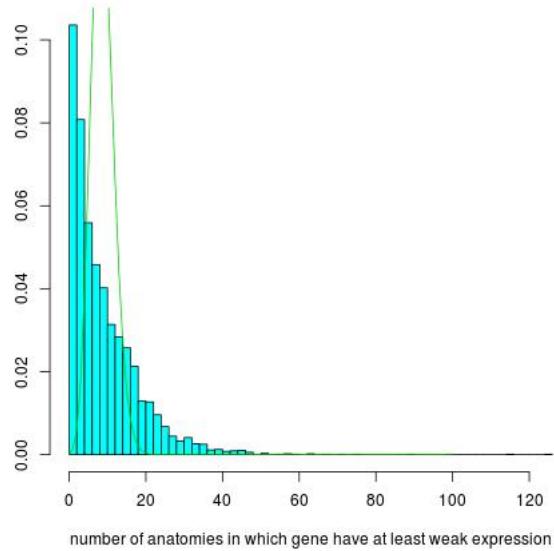


Figure 3: weakly expressed gene histogram

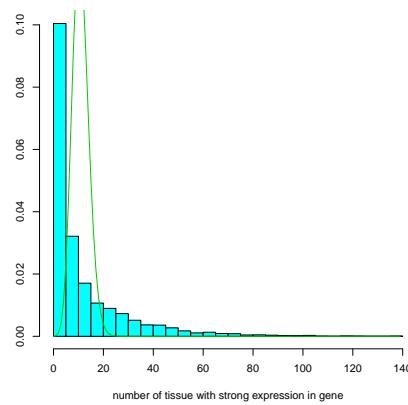


Figure 4: Distribution of number of tissues that have strong gene expression and corresponding null distribution.

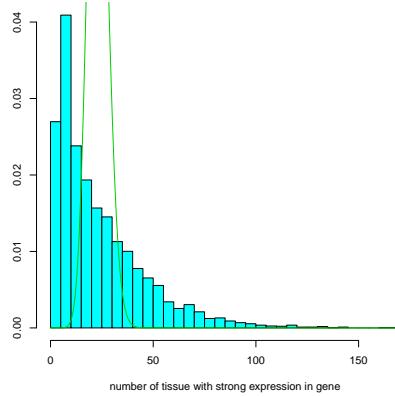


Figure 5: Distribution of number of tissue in which gene have strong or medium expression and null distribution.

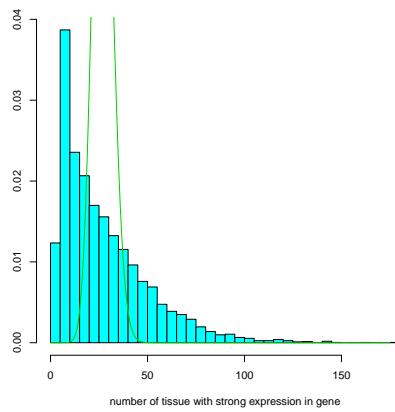


Figure 6: Distribution of number of tissue in which gene have expression and null distribution.

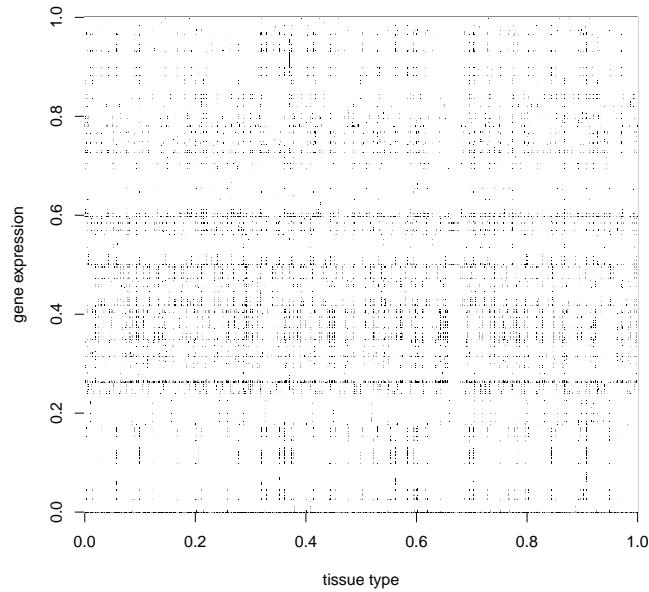


Figure 7: Interaction matrix

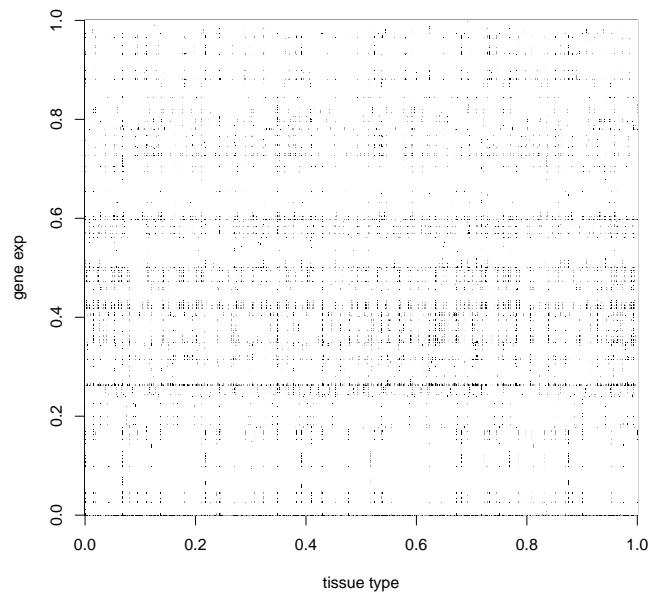


Figure 8: Principal component analysis with row arranged according to the first PC across gene.

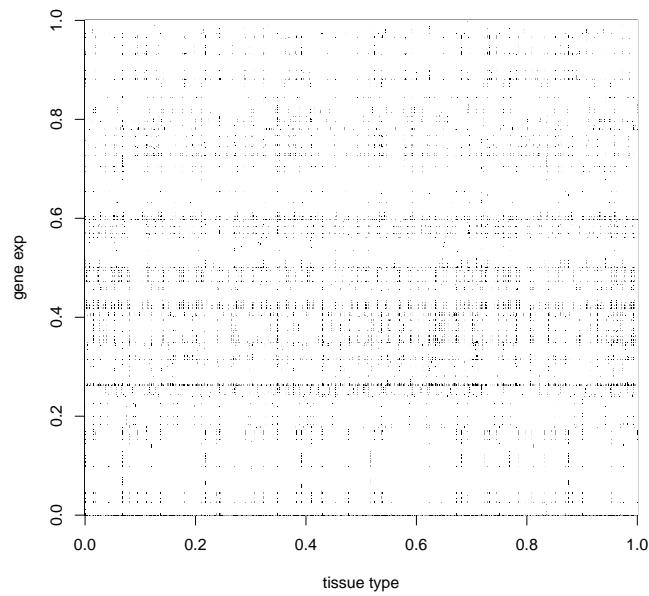


Figure 9: Principal component analysis across tissue and genes row and column arranged according to the first PC.

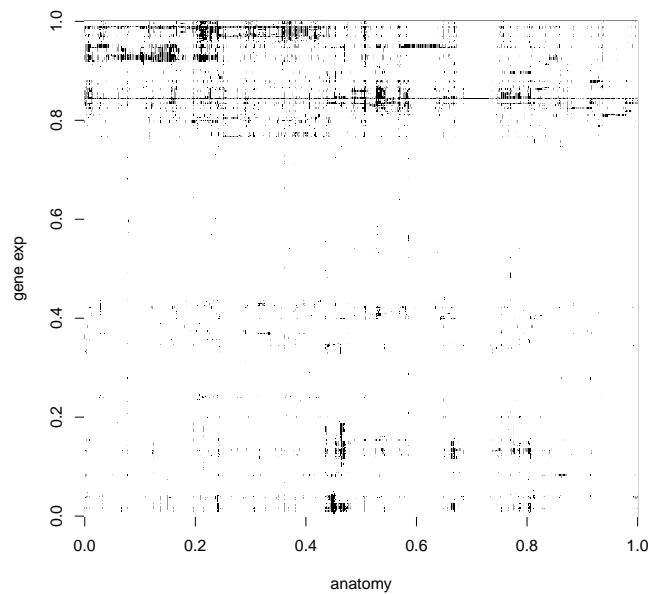


Figure 10: Hierarchical clustering with of interaction matrix.