Large-scale GO analysis reveals gene expression level and variation are functional and location related

Simon M. Lin[1], Pan Du[1], and Warren A. Kibbe[1]

[1]Robert H. Lurie Comprehensive Cancer Center, Northwestern University, Chicago, IL, 60611, USA

Abstract

The CAMDA 2007 META-analysis data set encompass many different tissue types in normal and disease states, which prompted us to ask a fundamental biological question: What are the essential genes in a living cell? Why the expression levels of some genes fluctuate a lot while the others are under tight control? Are there any relations between gene functions or locations with the gene expression levels and variation? We present a systems biology overview of life processes based on this analysis. We also model low-variance housekeeping genes (LoV-HKGs), which have practical use as internal control genes for RT-PCR experiments.

Introduction

Ever since the emerging of the microarray technology, studies of a compendium of tissues have been attempted. A year 2000 study by Warrington and colleagues at Affymetrix identified cell maintenance genes required at all different developmental stages from fetal to adult (Warrington, et al., 2000). The HUGE index project at Harvard studied 19 normal human tissues using 59 arrays to further define housekeeping genes (Hsiao, et al., 2001).

The CAMDA 2007 META-analysis data set, which contains about 6,000 microarrays of a variety of tissues in both normal and pathological states, offers a unique opportunity to further answer a few fundamental questions of life:

- What are the essential genes in a living cell?
- Why are the expression levels of some genes tightly regulated while others appear to fluctuate in response to a variety of developmental, environmental or temporal signals?

To answer these questions, we used Gene Ontology (GO) (Ashburner and Lewis, 2002) to analyze the gene lists derived from some basic statistical properties of the CAMDA 2007 META-analysis data set.

Although the methods we utilized are very well established, the findings are intriguing. On the theoretical side, we present a systems biology overview of cellular processes that are exposed by these microarray experiments. One the pragmatic side, we suggest a list of internal control genes to be used in RT-PCR experiments, genes that among the CAMDA set of microarrays have a low variance.

Housekeeping Genes: revisited and redefined

The word of "housekeeping gene" has been extensively used in the literature in related contexts but with completely different meanings. Thus, we first revisit the usage of this word.

In some contexts, housekeeping genes refer to constitutively expressed genes, regardless of tissues types (Hsiao, et al., 2001) or developmental stages (Warrington, et al., 2000). They are essential transcripts to keep the cell alive.

In other contexts, housekeeping genes refer to the internal control genes used to normalize mRNA levels between different samples using RT-PCR, because usually RT-PCR can only quantify a relative expression level (Silver, et al., 2006). More recently, housekeeping genes are also attempted to facilitate the normalization of microarray results, especially for small diagnostic microarrays.

We attempt to clarify the issue above by redefining the housekeeping genes (HKGs) and introducing the concept of low-variance housekeeping genes (LoV-HKGs):

> **Housekeeping genes (HKGs)**: Genes constitutively found in all human cells. They are required for the maintenance of the basic cellular functions. HKG is a qualitative classification of the transcriptome into the housekeeping ones vs. the non-essential ones. Although a gene can be a HKG, its expression level can still vary significantly by tissue type, developmental stage, and environment.
>
> **Low-variance housekeeping genes (LoV-HKGs)**: A subset of HKGs with low variances of expression levels. LoV-HKGs can be used to calibrate the measurements of gene expressions made by either RT-PCR or microarrays. A LoV-HKG is based on the quantitative measurement of the variance between a large set of conditions.

A comparison of HKGs and LoV-HKGs is summarized in Table 1.

Table 1. Comparison of HKGs and LoV-HKGs.

	HKGs	LoV-HKGs
Biological meaning	Constitutively expressed	Constantly expressed
Indicated by	High prevalence	Low variance
Identified by	P% call	IQR of expression

Data Set and Statistical Methods

The CAMDA 2007 META-analysis data set was used in this study. It contains 5896 microarrays of a variety of tissues in both normal and pathological states:

1142 expression profiles from more than 80 cell lines,
1278 expression profiles from tens of normal tissues and cell types,
3476 expression profiles related to various diseases and syndromes.

The gcRMA normalized file from ArrayExpress was used to assess the level of expression of each gene. Arguably appropriate, we used the log2 intensity as a proxy for the true copy number of each transcript, because the 11-probe-per-transcript design of Affymetrix can presumably average out the differences between the binding affinities of the probes.

The detection of each gene on an array was identified by the Affymetrix MAS5 present-call method. Briefly, the hybridization intensities of 11 perfect-match probes were tested against corresponding mismatch probes to suggest specific hybridization of each gene. $P < 0.04$ was used as a cutoff of a Present call, $P > 0.06$ was defined as Absent, and in between was defined as Marginal.

To analyze the biological significance of a list of genes, the hyper-geometric test of each GO term was used.

Analysis of the expression level of genes

It has been established that certain types of proteins, such as membrane proteins, are hardly measurable by proteomics due to their poor solubility. However, the detection bias of microarrays has

not been fully characterized. We can postulate that genes with low expression level (estimated by the mean of each gene across all samples shown as "Present") are intrinsically harder to detect because the true signal produced by these genes is closer to background noise and will be more susceptible to other effects like cross-hybridization from a gene with significant homology to the probes involved and a much higher expression level. As shown in Supplementary Table 3 and 4 and Figure 1, we found that in general, signaling genes are expressed at a low level, whereas genes that produce components of the protein synthesis machineries are abundantly expressed. The result agrees with the stage-wise amplification model in the signaling regulation. It also suggests that the receptor genes on the cell member are also challenging for microarray to detect due to their low expression levels. While we believe that proteomics and gene expression microarrays provide differing pictures of the dynamic molecular processes in a cell, both of these techniques appear to have sensitivity problems with signaling pathways, which are crucial in understanding the transitions from one state to another. However, we believe we can model the states themselves without this information.

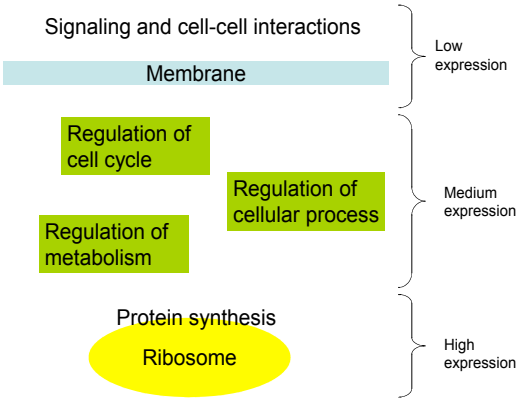

Figure 1. The expression level of different group of genes

Present% analysis suggests a spectrum of biological activities from organ-development genes to housekeeping genes

Next, we categorized genes into five groups according to their percentage detectability: constituently expressed (p% > 0.95, i.e., the gene is detectable in more than 95% of the 5896 arrays in the data set), rarely expressed (p% < 0.05), and three regulated groups (p% between 0.05-0.33, 0.33-0.67 and 0.67-0.95). A striking pattern was found. Generally, protein synthesis and metabolism are fundamental life processes (Supplementary Table 1), which happens in the cytoplasm and ribosome (Supplementary Table 2). In contrast, genes involved in organ development or organ-specific processes are only detectable in a very small portion of samples. Based on our analysis, a pyramid of life processes is illustrated in Figure 2. This is very much in keeping with our quantitative and intuitive understanding of cell biology.

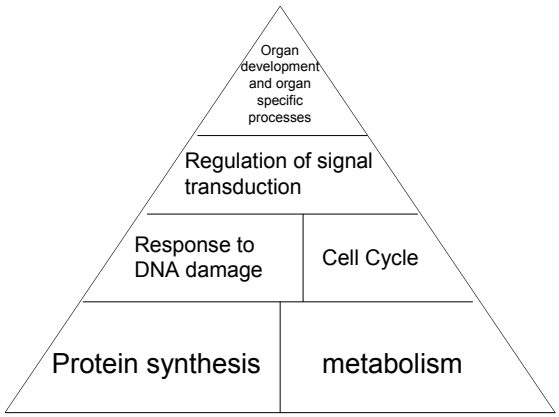

Figure 2. A pyramid of life processes. Protein synthesis and metabolism are constituent processes.

Analysis of the variations of genes

We further asked the quantitative question of gene regulation: why some genes are tightly regulated (showing little variation of expression levels) while other genes have big fold-changes across arrays? To answer this question, we calculated the variance of each gene.

As shown in Supplementary Table 5 and Figure 3, sensory perception, signal transduction and some regulation related genes have low variation. One possible reason is due to their low expression levels and mixture of backgrounds. It also matches the GO analysis results of expression means. Another finding is that groups of development related and response related genes have very high variations. And this matches our biological prior knowledge. By analyzing the GO of cellular components, we can see some genes located in the ribosome have a small variance (Supplementary Table 6). Comparing the results shown in Supplementary Table 2 and 4, we found ribosome related genes also have high Present % and expression levels. It suggests that ribosome related genes are good candidates for house keeping genes.

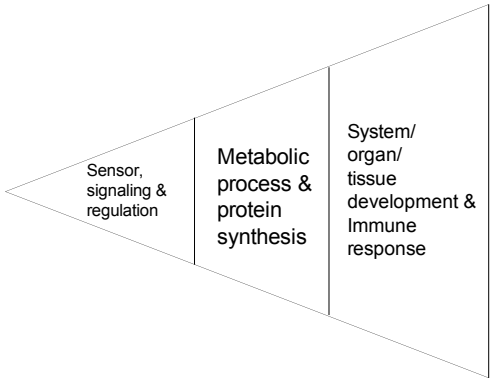

Figure 3. Magnitude of change for different biological processes.

Suggested LoV-HKGs as internal controls for RT-PCR experiments

A list of LoV-HKGs was identified from our study (see Table 2 for a subset of the list.). It has important applications for the selection of internal control genes for RT-PCR experiments.

Table 2. Suggested internal control genes for the normalization of RT-PCR experiments

Entrez ID	Gene Symbol	Gene Name	Present (%)[1)]	Mean Expression[2)]	Fold Change[3)]
6171	RPL41	ribosomal protein L41	99.02	14.81	1.32
6168	RPL37A	ribosomal protein L37a	99.95	14.67	1.22
6228	RPS23	ribosomal protein S23	99.88	14.58	1.37
23521	RPL13A	ribosomal protein L13a	99.88	14.58	1.47
6176	RPLP1	ribosomal protein, large, P1	99.98	14.54	1.54
6147	RPL23A	ribosomal protein L23a	99.68	14.53	1.35
9639	ARHGEF10	Rho guanine nucleotide exchange factor (GEF) 10	99.91	14.44	1.57
7178	TPT1	tumor protein, translationally-controlled 1	100	14.32	1.33
1915	EEF1A1	eukaryotic translation elongation factor 1 alpha 1	100	14.28	1.28
6232	RPS27	ribosomal protein S27 (metallopanstimulin 1)	100	14.23	1.47
6218	RPS17	ribosomal protein S17	99.91	14.16	1.65
6167	RPL37	ribosomal protein L37	100	14.13	1.57
60	ACTB	actin, beta	99.98	14.11	1.55
6222	RPS18	ribosomal protein S18	99.86	14.11	1.66
2597	GAPDH	glyceraldehyde-3-phosphate dehydrogenase	99.21	14.1	1.9

Present%[1)] : the percentage of arrays (total n= 5896) that the genes is detectable (estimated by the present calls by Affymetrix).
Mean Expression[2)] : mean of expression levels (log2).
Fold Change[3)] : the ratio between 75% percentile and 25% percentile of the expression levels across all the arrays for that gene, i.e., 2^IQR, which is a robust measurement of the variance.

In agreement de Jonge's recent publication (de Jonge, et al., 2007), we have found a group of ribosomal proteins that perform well as internal controls. In contrast to the de Jonge study, we have found traditional internal control genes, such as ACTB and GAPDH, also perform reasonably well for this task. In addition, we have found some novel candidate genes in the protein translation machinery, such as EEF1A1 and TPT1, which facilitate the function of ribosomal proteins. Surprisingly, we found a Rho GTPase, ARHGEF10, which was expressed with very low variance across arrays. This result is in keeping with the fundamental importance of Rho-dependent signaling in life processes.

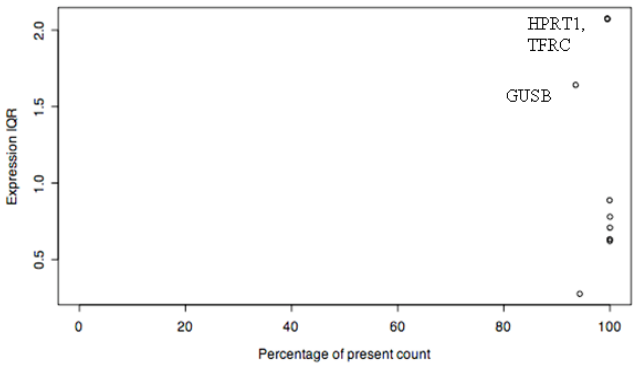

Figure 4. Traditional internal control genes used in RT-PCR studies.

We also investigated how the traditional internal control genes (ACTB, B2M, GAPDH,GUSB, HPRT1, PGK, PPIA, RPL13A, TBP, TFRC) perform in the CAMDA data set. As shown in Figure 4, HPRT1, TFRC, and GUSB, although expressed in most tissues (high p%), have considerable variations (judged by IQR).

Conclusion and discussion

Through large-scale gene ontology analysis, we found gene expression level and variation are functional and location related. Although the methods we utilized are very well established, the findings are intriguing. The results provide an overview of basic life processes. It also led to revisit the housekeeping genes. We found problems of previously defined house keeping genes. By defining low-variance housekeeping genes (LoV-HKGs), we identified a new list of LoV-HKGs, which provides both high percentage of present calls and low variation across samples. Considering all these results were based on the Affymetrix Hgu133a chips, the results could be probe design specific. In the next step, we will further evaluate these finding over the arrays in other Affymetrix versions or other platforms.

References

1. Ashburner, M. and Lewis, S. (2002) On ontologies for biologists: the Gene Ontology--untangling the web, *Novartis Found Symp*, **247**, 66-80; discussion 80-63, 84-90, 244-252.
2. de Jonge, H.J., Fehrmann, R.S., de Bont, E.S., Hofstra, R.M., Gerbens, F., Kamps, W.A., de Vries, E.G., van der Zee, A.G., Te Meerman, G.J. and Ter Elst, A. (2007) Evidence based selection of housekeeping genes, *PLoS ONE*, **2**, e898.
3. Hsiao, L.L., Dangond, F., Yoshida, T., Hong, R., Jensen, R.V., Misra, J., Dillon, W., Lee, K.F., Clark, K.E., Haverty, P., Weng, Z., Mutter, G.L., Frosch, M.P., Macdonald, M.E., Milford, E.L., Crum, C.P., Bueno, R., Pratt, R.E., Mahadevappa, M., Warrington, J.A., Stephanopoulos, G., Stephanopoulos, G. and Gullans, S.R. (2001) A compendium of gene expression in normal human tissues, *Physiol Genomics*, **7**, 97-104.
4. Silver, N., Best, S., Jiang, J. and Thein, S.L. (2006) Selection of housekeeping genes for gene expression studies in human reticulocytes using real-time PCR, *BMC Mol Biol*, **7**, 33.
5. Warrington, J.A., Nair, A., Mahadevappa, M. and Tsyganskaya, M. (2000) Comparison of human adult and fetal expression and identification of 535 housekeeping/maintenance genes, *Physiol Genomics*, **2**, 143-147.

Supplementary Table 1. Biological Processes of genes with varies degrees of Present % (top 10 categories at level 3 with p < 0.01)

< 0.05		0.05 – 0.33		0.33 – 0.67		0.67 – 0.95		> 0.95	
GO.Term	p.value	GO.Term	p.value	GO.Term	p.value	GO.Term	p.value	GO.Term	p.value
sensory perception	7.25E-20	signal transduction	6.10E-11	cell morphogenesis	2.31E-06	biopolymer metabolic process	7.48E-23	cellular macromolecule metabolic process	2.02E-30
cell-cell signaling	1.23E-17	system development	1.74E-06	regulation of cell cycle	3.23E-06	nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	6.77E-19	macromolecule biosynthetic process	1.53E-27
system development	1.75E-17	cell-cell signaling	1.25E-05	phosphorus metabolic process	4.97E-06	establishment of cellular localization	1.82E-11	protein metabolic process	1.85E-27
tissue development	2.10E-10	organ development	2.57E-05	organelle organization and biogenesis	9.86E-06	establishment of protein localization	1.83E-09	cellular biosynthetic process	2.04E-25
transmission of nerve impulse	6.90E-10	transmission of nerve impulse	6.57E-05	**cell cycle phase**	1.07E-05	regulation of cellular metabolic process	5.29E-09	**ribonucleoprotein complex biogenesis and assembly**	8.60E-25
organ development	3.19E-09	striated muscle contraction	1.02E-04	regulation of signal transduction	2.96E-05	protein metabolic process	2.55E-08	establishment of cellular localization	7.37E-18
signal transduction	9.67E-09	**cell migration**	3.26E-04	mitotic cell cycle	3.27E-05	macromolecular complex assembly	1.29E-07	establishment of protein localization	1.22E-14
detection of stimulus during sensory perception	3.87E-06	regulation of muscle contraction	4.10E-04	signal transduction	3.34E-05	cellular macromolecule metabolic process	1.45E-06	**biopolymer metabolic process**	8.17E-14
bone remodeling	8.25E-06	cell-cell adhesion	4.21E-04	biopolymer metabolic process	2.05E-04	**response to DNA damage stimulus**	2.71E-06	nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	1.07E-13
regulation of heart contraction	5.17E-05	**regulation of signal transduction**	7.13E-04	negative regulation of cellular process	3.13E-04	organelle organization and biogenesis	9.43E-06	cofactor metabolic process	6.18E-11

Supplementary Table 2. Cellular Components of genes with varies degrees of Present % (top 10 categories at level 3 with p < 0.01)

< 0.05		0.05 – 0.33		0.33 – 0.67		0.67 – 0.95		> 0.95	
GO.Term	p.value	GO.Term	p.value	GO.Term	p.value	GO.Term	p.value	GO.Term	p.value
plasma membrane part	2.83E-28	plasma membrane	3.89E-12	**intracellular membrane-bound organelle**	1.40E-05	intracellular	4.12E-77	**cytoplasm**	1.66E-66
plasma membrane	1.11E-25	**intrinsic to membrane**	2.42E-11	**nuclear envelope**	2.41E-03	**intracellular part**	6.03E-76	intracellular membrane-bound organelle	3.59E-55
intrinsic to membrane	8.00E-13	plasma membrane part	2.36E-10	endomembrane system	2.56E-03	membrane-bound organelle	2.38E-66	cytoplasmic part	2.13E-53
membrane fraction	1.81E-05	myosin II complex	8.59E-04	nuclear membrane part	3.13E-03	intracellular membrane-bound organelle	3.66E-66	**ribosome**	2.00E-40
nicotinic acetylcholine-gated receptor-channel complex	3.17E-04	T cell receptor complex	1.47E-04	fibrillar collagen	3.93E-03	intracellular organelle	6.93E-64	nuclear part	3.31E-31
		cytoplasmic vesicle	5.97E-04			cytoplasm	2.80E-24	cytosolic large ribosomal subunit (sensu Eukaryota)	6.69E-23
						cell part	1.26E-21	cytosolic small ribosomal subunit (sensu Eukaryota)	4.40E-20
						intracellular organelle part	1.40E-16	spliceosome	2.40E-18
						cytoplasmic part	1.20E-15	organelle membrane	3.79E-18
						nuclear part	2.83E-14	mitochondrial part	4.67E-17

Supplementary Table 3. Biological Processes of genes with different mean expression levels (log2) (top 10 categories at level 3 with $p < 0.01$)

< 5		5 – 10		> 10	
GO.Term	p.value	GO.Term	p.value	GO.Term	p.value
signal transduction	2.43E-10	mitotic cell cycle	2.62E-08	macromolecule biosynthetic process	1.36E-35
regulation of cellular metabolic process	2.19E-07	cell cycle phase	6.72E-07	cellular biosynthetic process	3.56E-27
sensory perception	6.03E-06	**regulation of cell cycle**	9.63E-07	ribonucleoprotein complex biogenesis and assembly	5.77E-16
cell-cell signaling	1.63E-05	**negative regulation of metabolic process**	3.48E-06	**protein metabolic process**	1.09E-14
meiotic recombination	1.16E-03	biopolymer metabolic process	1.40E-05	cellular macromolecule metabolic process	1.64E-13
cell-cell adhesion	1.96E-03	organic acid metabolic process	2.70E-05	generation of precursor metabolites and energy	3.81E-10
transmission of nerve impulse	2.32E-03	heterocycle metabolic process	5.44E-05	macromolecular complex assembly	5.95E-10
detection of chemical stimulus	4.79E-03	establishment of protein localization	6.95E-05	establishment of cellular localization	9.78E-08
embryonic pattern specification	5.91E-03	establishment of RNA localization	8.47E-05	regulation of biosynthetic process	1.97E-07
bone resorption	6.23E-03	**negative regulation of cellular process**	1.13E-04	regulation of protein metabolic process	3.19E-07

Supplementary Table 4. Cellular Components of genes with different mean expression levels (log2) (top 10 categories at level 3 with $p < 0.01$)

< 5		5 – 10		> 10	
GO.Term	p.value	GO.Term	p.value	GO.Term	p.value
plasma membrane part	6.89E-12	intracellular membrane-bound organelle	4.82E-14	cytoplasm	2.64E-59
plasma membrane	7.27E-08	nuclear part	5.59E-09	**ribosome**	1.20E-53
intrinsic to membrane	2.64E-07	**cytoplasm**	4.67E-08	cytoplasmic part	1.56E-52
		nuclear lumen	1.74E-05	cytosolic large ribosomal subunit (sensu Eukaryota)	8.32E-33
		outer membrane	1.26E-04	cytosolic small ribosomal subunit (sensu Eukaryota)	7.58E-31
		cytoplasmic part	2.60E-04	intracellular non-membrane-bound organelle	2.50E-29
		endomembrane system	5.78E-03	mitochondrial membrane part	1.70E-12
		mitochondrial lumen	5.84E-03	heterogeneous nuclear ribonucleoprotein complex	2.23E-09
		nuclear membrane part	6.29E-03	mitochondrial envelope	4.82E-09
		organelle membrane	7.26E-03	mitochondrial part	1.12E-08

Supplementary Table 5. Biological Processes of genes with different fold change (defined based on expression IQR) (top 10 categories at level 3 with p < 0.01)

< 1.5		1.5 – 8		> 8	
GO.Term	p.value	GO.Term	p.value	GO.Term	p.value
sensory perception	1.79E-05	biopolymer metabolic process	1.03E-26	**system development**	4.56E-22
regulation of cellular metabolic process	2.30E-05	nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	1.35E-20	**response to wounding**	2.88E-20
meiotic recombination	9.16E-05	establishment of cellular localization	1.71E-16	**inflammatory response**	6.14E-16
transmission of nerve impulse	3.38E-04	establishment of protein localization	4.51E-16	**organ development**	1.93E-15
regulation of neurological process	1.79E-03	protein metabolic process	4.98E-14	Chemotaxis	1.36E-14
signal transduction	2.48E-03	cellular macromolecule metabolic process	6.75E-13	Locomotory behavior	2.28E-14
		ribonucleoprotein complex biogenesis and assembly	2.98E-11	**humoral immune response**	2.58E-12
		organelle organization and biogenesis	4.77E-11	**response to other organism**	9.10E-11
		macromolecule biosynthetic process	1.55E-08	cell-cell signaling	1.90E-10
		cellular biosynthetic process	5.47E-08	**tissue development**	1.90E-10

Supplementary Table 6. Cellular Components of genes with different fold change (defined based on expression IQR) (top 10 categories at level 3 with p < 0.01)

< 1.5		1.5 – 8		> 8	
GO.Term	p.value	GO.Term	p.value	GO.Term	p.value
ribosome	1.84E-03	intracellular membrane-bound organelle	4.09E-67	plasma membrane	6.73E-19
cytosolic large ribosomal subunit (sensu Eukaryota)	5.99E-03	cytoplasm	9.03E-34	plasma membrane part	5.30E-16
		nuclear part	1.12E-30	MHC class II protein complex	1.44E-06
		cytoplasmic part	1.90E-26	external side of plasma membrane	2.02E-06
		nuclear lumen	1.40E-19	intrinsic to membrane	2.34E-05
		organelle membrane	5.02E-18	fibrillar collagen	9.93E-05
		endomembrane system	1.10E-10	soluble fraction	3.01E-04
		spliceosome	1.74E-09	membrane fraction	2.45E-03
		mitochondrial part	1.11E-08	cytoskeletal part	3.33E-03
		organelle inner membrane	2.08E-07	growth cone	4.73E-03