Knowledge Enrichment Analysis for Human Tissue-Specific Genes Uncover New Biological Insights

Xiu-Jun Gong[1*], Hua Yu[1], Chun-Bai Yang[1], Yuan-Fang Li[2]

[1]School of Computer Science and Technology, Tianjin University, Weijin Rd No. 92, Nankan, Tianjin, 300072, China

[2]Clayton School of IT, Faculty of Information Technology,Monash University, Wellington Rd, Clayton, Vic 3800, Australia

Summary

The expression and regulation of genes in different tissues are fundamental questions to be answered in biology. Knowledge enrichment analysis for tissue specific (TS) and housekeeping (HK) genes may help identify their roles in biological process or diseases and gain new biological insights.In this paper, we performed the knowledge enrichment analysis for 17,343 genes in 84 human tissues using Gene Set Enrichment Analysis (GSEA) and Hypergeometric Analysis (HA) against three biological ontologies: Gene Ontology (GO), KEGG pathways and Disease Ontology (DO) respectively.The analyses results demonstrated that the functions of most gene groups are consistent with their tissue origins. Meanwhile three interesting new associations for HK genes and the skeletal muscle tissuegenes are found. Firstly, Hypergeometric analysis against KEGG database for HK genes disclosed that three disease terms (Parkinson's disease, Huntington's disease, Alzheimer's disease) are intensively enriched.Secondly, Hypergeometric analysis against the KEGG database for Skeletal Muscle tissue genes shows that two cardiac diseases of "Hypertrophic cardiomyopathy (HCM)" and "Arrhythmogenic right ventricular cardiomyopathy (ARVC)" are heavily enriched, which are also considered as no relationship with skeletal functions.Thirdly, “Prostate cancer” is intensively enriched in Hypergeometric analysis against the disease ontology (DO) for the Skeletal Muscle tissue genes, which is a much unexpected phenomenon.

1 Introduction

Exploring the tissue-specificity of gene expressions and their regulations is always a main theme in biology [1].Housekeeping (HK) genes are constitutively expressed in all tissues while Tissue Specific (TS) genes are expressed at a much higher level in a single tissue type than in others. HKgenes serve as valuable experimental controls in gene and protein expression experiments, while TS genes tend to represent distinct physiological processes and are frequent candidates for biomarkers or drug targets [1 – 3].

In past years, many investigations have been conducted the expression and regulation mechanisms of HK and TS genes [4 – 8]. Emerging technologies including Microarray technology [9], Expressed Sequence Tag(EST) [10], Serial Analysis of Gene Expression (SAGE) [10], and Chip-Seq technology [11] enable the analysis of expression patterns of HK and TS genes at genome scale. The analysis for genomic and epigenetic features of two groups of genes may shed light on the mechanisms by which cells maintain basic and tissue-specific functions [10], [12 – 17]. Also several tissue specific databases such as TiGER[18], TiProD [19], TissueInfo [20], TissueDistributionDBs [21] and TiSGeD [22] provide web

[*] To whom correspondence should be addressed. Email: gongxj@tju.edu.cn

doi:10.2390/biecoll-jib-2012-194

services to extract the TS and HK genes and their genomic features by predefined threshold values. However, there is no widely accepted criterion to judge which gene is expressed in a specific tissue only by the expression patterns. More, the integrative behaviors of extracted groups of genes, for instance, the biological functions and disease associations with these extracted groups of genes, have not been extensively studied. Such information is useful in the identification of HK and TS genes but also in the discovery of new biological insights.

The gene set enrichment analysis aims to seeking enrichment with a group of genes to assess whether subsets of this group show statistical significance of biological attributes against backgroundknowledge.Hypergeometric analysis and GSEAare two commonly used algorithms to acquire such enrichment terms and assess statistical significance. Hypergeometric analysis [23] calculates the possibility of the enrichment of one biological term in a target set against the background set with the assumption that genes are sampled from a hypergeometric distribution. GSEA [24] relies on phenotype labels to distinguish gene expression patterns under different conditions with the consideration of degree of change in expression values. Both of them have been widely used for the gene set enrichment analysis of gene groups. Dezso et al. [25] conducted enrichment analysis of HK and TS gene sets using the hypergeometric distributionand Gene Set Enrichment Analysis (GSEA) algorithms to determine enrichment across four functional ontologies: canonical pathway maps, gene ontology (GO) processes, GeneGo (GG) process networks, and diseases.Their analysis showed that thebiological functions of HK and TS are consistent with tissue origin.

In this paper we performed the knowledge enrichment analysis for 17,343 genes in 84 human tissues using Gene Set Enrichment Analysis (GSEA) and Hypergeometric Analysis (HA) against three biological ontologies: Gene Ontology (GO), KEGG pathways and Disease Ontology (DO).The analyses demonstrated that the functions of most gene groups are consistent with their tissue origins. Meanwhile three interesting associations for HK genes and the skeletal muscle tissue genes from the analysis might uncover new biological insights.

2 Methods

2.1 Data sources and pre-processing

We used the Human U133A/GNF1H Gene expression data to determine housekeeping and tissue-specific genes. The whole gene expression data and GNF1H annotation files linking Ensembl transcripts to their representative probes were downloaded from GNF (Genomics Institute of the Novartis Research Foundation) and U133A chip annotation files were downloaded from the Affymetrix website. The dataset contains 33,698 genes across 84 human tissues. To perform the enrichment analysis, all the genes in the dataset are unified as gene symbols. The probe sets which have no corresponding gene symbols are removed, and expression values with redundant probe IDs are averaged. Finally, 17,343 genes across 84 tissues are remained for further analysis.

Based on the purified gene expression dataset, we identified HK and TS genes. A gene whose expression value is no less than 3 standard deviations above the mean expression value in the give tissueis regarded as the TS gene. A gene whose expression value is more than 60 in no less than 75 tissues is regarded as the HK gene. Finally, we obtained1382 HK genes and 5,721 tissue-specific geneswith average 68.1 genes per tissue.

2.1.1 Knowledgebase preparations

To perform the enrichment analysis, we used three ontologies: Gene Ontology (GO) BP processes, KEGG Pathway, and Disease Ontology (DO).

The GO unifies the representation of gene and gene product attributes across all species using controlled vocabularies. It covers there domains: Cellular Component (CP), Molecular Function (MF) and Biological Process (BP). BP processconsists of operations or sets of molecular events with a defined beginning and end, pertinent to the functioning of integrated living units: cells, tissues, organs, andorganisms. This component is used to perform the biological process enrichment analysis in our study.

KEGG is a collection of five online databases dealing with genomes, enzymatic pathways, and biological chemicals. It describes networks of molecular interactions in the cells, and variants of them specific to particular organisms. We use the KEGG PATHAY database for the biological pathway enrichment analysis.

Disease ontology (DO)organises disease conceptsinto a directed acyclic graph (DAG) just like GO. Its mission is to integrate biomedical data for human diseases. DO has been used to annotate the human genome and show better performance than other current human disease annotations. We downloaded the DO from http://www.obofoundry.org/cgi-bin/detail.cgi?id=disease_ontology and compiled a map between genes and terms in DO. The Table 1 lists versions and sources of the three ontologies.

Table 1: knowledgebase and their versions

Knowledgebase	version
GO BP	c5.bp.v3.0.symbols.gmt
KEGG Pathway	c2.cp.kegg.v3.0.symbols.gmt
Disease Ontology (DO)	Access in May, 2011

2.1.2 Enrichment analysis

The enrichment analysis seeks to infer if a group of genes show statistical significance of biological attributes against the background knowledge.

We used the p-value as the enrichment levels of the HK genes and tissue-specific genes in Hypergeometric Analysis. P-value is calculated according to formula (1).

$$p(k) = \sum_{i=k}^{D} p(i, M, n, N) \qquad (1)$$

Where N and n are the numbers of genes in the background database and targeted gene set respectively, M and i are the numbers of genes labelled by a give biological term in the background database and targeted gene set respectively. The value $p(i, M, n, N)$ follows the Hypergeometric distribution:

$$p(i, M, n, N) = \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}} \qquad (2)$$

The p-value gives probability of having at least k marked elements in a sample of size n by random selection.

In the Hypergeometric Analysis, we used GeneCodis2 (http://genecodis.dacya.ucm.es/) for KEGG Pathway and GO BP, and fundo (http://django.nubic.northwestern.edu/fundo/) for the Disease Ontology (DO).

We used GSEA from http://www.broadinstitute.org/gsea/index.jsp for Gene Set Enrichment Analysis. The GO BP and KEGG pathway gene sets are both obtained from MSigDB. We compiled an annotation file for the Disease Ontology (DO) enrichment analysis.

3 Results

3.1 List of housekeeping genes and tissue-specific genes

We set up a fixed cutoff value to determine whether a gene is expressed in a specific tissue for the gene expression dataset; see the method part for details. Finally, 1382 housekeeping genes and 5,721 tissue-specific geneswith average 68.1 per tissue are identified. See Table 2 for more details.

3.2 Knowledge enrichment analysis for Housekeeping genes

We performed the HA against GO BP, KEGG Pathway and DO, for the enriched terms and their distributions (see Table 3 and Figure 1 respectively).

The results produced by HA for all the 84 tissues are accessible from http://cs.tju.edu.cn/faculty/gongxj/specifictome/HA.htm.

For the GO biological process, the top 10 terms can be classified approximatelyin three groups. The first group is related to central dogma with terms like "translational elongation", "translation", "RNA splicing" and "mRNA processing". These biological processes are vital in transmission of gene information and fulfilment of cells' functions. The second group is related to metabolism with terms like "electron transport chain", "mitochondrial ATP synthesis coupled proton transport", "mitochondrial electron transport, NADH to ubiquinone". Dezso et al. have performed a similar analysis of human HK and TS genes in 31 tissues [25]. In their analysis, metabolic terms concerned with electron transport, ATP synthesis are the most enriched ones. And for the third group, terms like "regulation of ubiquitin-protein ligase activity during mitotic cell cycle" and "anaphase-promoting complex-dependent proteasomal ubiquitin-dependent protein catabolic process" are involved. They are related to mitosis which is also fundamental to every tissue. All three groups are relevant to fundamental functions that maintain cells.

For the KEGG Pathway procedure, the enriched terms "Ribosome", "Oxidative phosphorylation", "Spliceosome" can be explained by its fundamental functions in metabolism or mitosis of HK genes. However a very interesting association is that three diseases terms: Parkinson's disease", "Huntington's disease” and "Alzheimer's disease” are intensively enriched. These three are typical neurodegenerative diseases caused by malformed proteins. Seemingly they are related with brain and nervous system and have nothing to do with the functions of housekeeping genes. However, [26] et al’showed that “A growing body of evidence now suggests that dysfunctionof autophagy causes accumulation of abnormal proteins and/or damaged organelles. Such accumulation has been linked to synaptic dysfunction, cellular stress and neuronal death.” Also they mentioned these three diseases are related to chronic nervous system disorders caused by dysfunction of autophagy. Therefore, we hypothesizethat the autophagy function of housekeeping genes might result in the three diseases.

Table 2: The numbers of housekeeping genes and tissue-specific genes

Tissue name	# genes	Tissue name	# genes	Tissue name	# genes
Housekeeping	1382	721_B_lymphoblasts	475	Adipocyte	57
AdrenalCortex	32	Adrenalgland	20	Amygdala	57
Appendix	38	AtrioventricularNode	56	BDCA4+_DentriticCells	106
Bonemarrow	23	BronchialEpithelialCells	88	CD105+_Endothelial	46
CD14+_Monocytes	27	CD19+_BCells(neg._sel.)	95	CD33+_Myeloid	114
CD34+	133	CD4+_Tcells	29	CD56+_NKCells	186
CD71+_EarlyErythroid	226	CD8+_Tcells	60	CardiacMyocytes	105
Caudatenucleus	27	Cerebellum	19	CerebellumPeduncles	47
CiliaryGanglion	73	CingulateCortex	17	Colorectaladenocarcinoma	82
DorsalRootGanglion	22	FetalThyroid	10	Fetalbrain	93
Fetalliver	36	Fetallung	11	GlobusPallidus	19
Heart	178	Hypothalamus	20	Kidney	63
Leukemia_chronicMyelogenousK-562	32	Leukemia_promyelocytic-HL-60	4	Leukemialymphoblastic(MOLT-4)	16
Liver	313	Lung	83	Lymphnode	8
Lymphoma_burkitts(Daudi)	45	Lymphoma_burkitts(Raji)	117	MedullaOblongata	6
OccipitalLobe	8	OlfactoryBulb	33	Ovary	13
Pancreas	30	PancreaticIslet	43	ParietalLobe	14
Pituitary	50	Placenta	173	Pons	22
PrefrontalCortex	141	Prostate	77	Salivarygland	24
SkeletalMuscle	212	Skin	35	SmoothMuscle	80
Spinalcord	24	SubthalamicNucleus	30	SuperiorCervicalGanglion	233
TemporalLobe	5	Testis	82	TestisGermCell	43
TestisIntersitial	90	TestisLeydigCell	24	TestisSeminiferousTubule	13
Thalamus	35	Thymus	20	Thyroid	112
Tongue	51	Tonsil	7	Trachea	28
TrigeminalGanglion	51	Uterus	56	UterusCorpus	35
WholeBlood	138	Wholebrain	56	colon	50
pineal_day	49	pineal_night	62	retina	92
small_intestine	66				

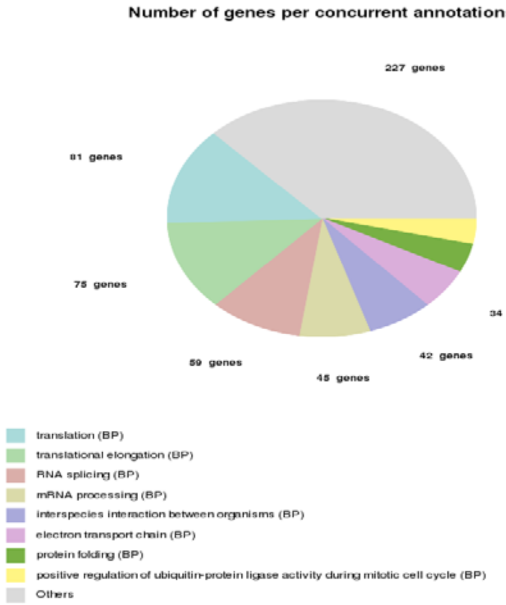

Figure 1: Number of genes per concurrent annotation in hypergeometric for GO biological process (generated by GeneCodis2)

Table 3: Enrichment terms of housekeeping genes using hypergeometric analysis

GO biological process	KEGG pathway	Disease ontology
translational elongation	Ribosome	Cancer
translation	Parkinson's disease	Colon Cancer
electron transport chain	Huntington's disease	Prostate Cancer
RNA splicing	Oxidative phosphorylation	Brain Tumor
mitochondrial ATP synthesis coupled proton transport	Alzheimer's disease	Down syndrome
mitochondrial ATP synthesis coupled proton transport	Spliceosome	Pancreas cancer
positive regulation of ubiquitin-protein ligase activity during mitotic cell cycle	Proteasome	Nasopharyngeal cancer
negative regulation of ubiquitin-protein ligase activity during mitotic cell cycle	Cardiac muscle contraction	Breast cancer
mRNA processing	Pathogenic Escherichia coli infection	Embryoma
	Antigen processing and presentation	Pancreas disease

For the Disease Ontology (DO) procedure, cancers and tumours take up most of highly enriched diseases. Cancer is caused by uncontrolled growth of cells which are related to the fundamental cellular functions like mitosis, autophagy and metabolism. See Figure 2 for more details.

3.3 Knowledge enrichment analysis for TS genes

We performedknowledge enrichment analysis for 84 human tissues using HA and GSEA against GO, KEGG pathway and DO. The functional annotations mostly are in consistence with their corresponding tissue's functions.

Because of the page limitation, we list ten top-scored terms for skeletal muscle and retina tissues in Table 4.

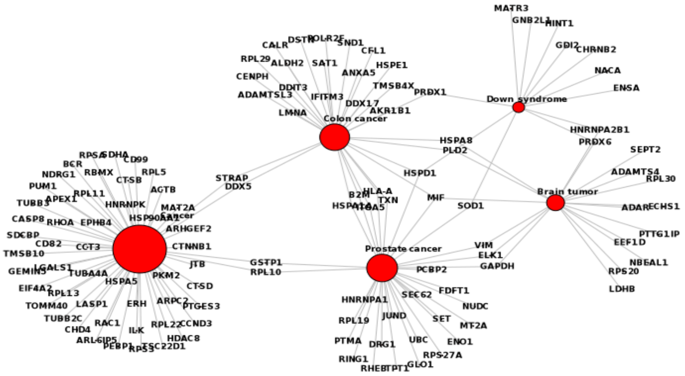

Figure 2: The mapping between genes and diseases in HA for DO enrichment analysis generated by FunDO.

Table 4: Enriched terms in HA and GSEA procedures for skeletal muscle and retina tissues

		Skeletal Muscle	Retina
GO biological process	HA	muscle contraction skeletal muscle contraction muscle organ development positive regulation of fast-twitch skeletal muscle fiber contraction regulation of striated muscle contraction striated muscle contraction	visual perception response to stimulus phototransduction, visible light phototransduction protein-chromophore linkage melanin biosynthetic process eye pigment biosynthetic process melanin biosynthetic process from tyrosine positive regulation of rhodopsin gene expression regulation of rhodopsin gene expression
	GESA	muscle development metal ion transport calcium ion transport di- tri- valent inorganic cation transport g protein signaling coupled to cyclic nucleotide second messenger cation transport regulation of muscle contraction cyclic nucleotide mediated signaling muscle cell differentiation second messenger mediated signaling	detection of external stimulus detection of abiotic stimulus sensory perception detection of stimulus involved in sensory perception detection of stimulus secondary metabolic process response to light stimulus neurological system process pigment biosynthetic process pigment metabolic process
KEGG pathway	HA	Hypertrophic cardiomyopathy (HCM) Calcium signaling pathway Dilated cardiomyopathy Cardiac muscle contraction Neuroactive ligand-receptor interaction Arrhythmogenic right ventricular cardiomyopathy (ARVC)	Tyrosine metabolism Retinol metabolism Melanogenesis

	GESA	neuroactive ligand receptor interaction	tyrosine metabolism
		cytokine cytokine receptor interaction	inositol phosphate metabolism
		tight junction	phenylalanine metabolism
		calcium signaling pathway	beta alanine metabolism
		pentose phosphate pathway	TGF beta signaling pathway
		heparansulfate biosynthesis	retinol metabolism
		maturity onset diabetes of the young	sphingolipid metabolism
		carbon fixation	phosphatidylinositol signaling system
		jak stat signaling pathway	melanogenesis
		alpha linolenic acid metabolism	histidine metabolism
disease ontology	HA	Prostate cancer	Retinitis pigmentosa
		"Myopathies, Nemaline"	Retinal disease
		Rhabdomyosarcoma	Uveomeningoencephalitic syndrome
		Vaccinia	Amaurosiscongenita of leber I
		Muscular atrophy	Macular degeneration
		Epilepsy	Eye disease
		Drug abuse	Glaucoma
		Amyotrophic lateral sclerosis	Melanoma
		Myopathy	Neoplasm metastasis
		Mucocutaneous lymph node syndrome	Uveitis
	GESA	myopathy	retinitis pigmentosa
		premature ovarian failure	age related macular degeneration
		attention deficit hyperactivity disease	retinal degeneration
		muscular dystrophy	glaucoma
		myelofibrosis	vitiligo
		disease of metabolism	retinal disease
		heart failure	autosomal dominant polycystic kidney
		rhabdomyosarcoma	primary open anle glaucoma
		thyroid neoplasm	hypercholesterolemia
		glucose intolerance	basal cell carcinoma

For the GO BP procedure, the top scored terms in skeletal muscle are all behaviours critical to skeletal muscle. In HA, muscle contraction is dominated. However, the results of GSEA procedure are more diversified,ranging from muscle development, differentiation, and ion transport tosignalling. Ion channels are critical in regulation of muscle behaviours. The enrichment intensity for ion related behaviours is in accordance with this statement.

Hypergeometric analysis against the KEGG databases for the Skeletal Muscle tissue shows that “Hypertrophic cardiomyopathy (HCM)" and "Arrhythmogenic right ventricular cardiomyopathy (ARVC)" areheavily enriched, which are considered as having no relationship with skeletal functions. These may result in the shared functions between the skeletal and cardiac muscles.

For the DO procedure, the most enriched terms of for skeletal muscle tissue genes in both methods are quite reasonable disorders related to skeletal muscle. "Myopathy", "Rhabdomyosarcoma", "muscular atrophy" etc. are all typical skeletal muscle diseases. In HA, the term “prostate cancer” has the highest score which is quite unexpected. However, a common function between skeletal muscle and prostate is that they're both androgen-regulated human tissues.Several studies focused on the effect of androgen on tissues such as skeletal muscle and prostate. Kumar R et al. pointed out in their study [27] that androgens are involved in the development of these tissues. And one regulator of androgens, namely the androgen receptor(AR) has association with prostate cancer. This is a potential reason for the enriched disease"prostate cancer" in skeletal muscle analysis.

The analysis results for the retina tissue are consistent with the one done by Dezaso et al. For GO BP, enriched terms can be classified in two groups. One includes the visual perception terms including phototransduction, detection and response to stimulus. The other is about retina pigment biosynthesis. The KEGG pathway result intensively concentrates on the biosynthesis of pigments and amino acids which are related to retina. Disease ontology analysis result agrees with the intuition which shows the connection with eye diseases like glaucoma and uveitis.

4 Discussion

4.1 Identification of housekeeping and tissue-specific genes

As for the identification of HK and TS genes based on expression datasets, there are several methods available. Dezso et al. used a 10 fold signal noise ratio criterion [25]. Hsiao et al. used present or absent calls to determine its presence [28]. Tu et al. and Su et al. used cutoff values to filter t HK and TS genes [9], [12]. We use a similar method with Su et al. and try several cutoff values to classify the two groups of genes. The distributions of the number of genes along with the number of tissues that the given genes are classified using cutoff values 40, 60, and 90 are drawn in Figure 3. Compared with the classification results of other researchers, the ideal distribution should be higher on both ends and lower in the middle region.

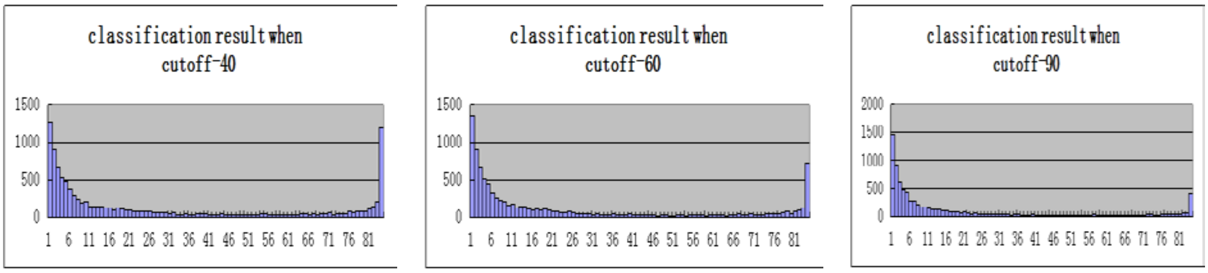

Figure 3: Relationship between the number of genes and the number of tissues that a given gene is classified using cutoff values 40, 60 and 90.

We took 60 as the cutoff value for HK gene classification. The genes with expression value more than 60 in no less than 75 tissues are considered as housekeeping gene. 1,382 out of the 17,343 genes are identified which correspond to the 10% estimation and also show their reasonableness in the functional analysis [12]. This criterion is not applicable to the identification of TS genes,because TS genes should be highly and uniquely expressed in a given tissue. There are genes that express on average much higher than others in all tissues. This results in the high variance of expression values of different genes in a tissue, while the expression values of a gene in different tissues tend to be less fluctuant. Wetake 3 standard deviations above the mean expression value of genes to be the cutoff for TS gene classifications.

4.2 Complementary results from Hypergeometric analysis and GSEA

Hypergeometric analysis and GSEA method both have their own advantages. Hypergeometric analysis is the classical way for enrichment analysis. It requires users' specification of target genes which are usually gained by classifications on expression value.It is straightforward and has the strength in integrating researchers' prior knowledge outside the expression data. However, the enrichment results are unreliable when no enough evidencesareavailable. GSEAonly requires specification of labels to distinguish genes in different conditions. It also takes into account all genes' expression values and focuses on a prior defined gene sets like cell cycle.This difference makes Hypergeometric analysis more subjective while GSEA more objective. The incorporation of both algorithms gives us more comprehensive functional annotations.

We can see this complementation in the analysis results. The 3 standard deviation above mean is a quite stringent cutoff. In Hypergeometric analysis, tissues having few genes, usually less than 40, have very few enriched terms or even none. However, GSEA results are stable and independent of the choice of cutoff. For the stringent cutoff, Hypergeometric analysis results tend to be clustered on some highly significant functions while GSEA tend tobe ranged more widely.

4.3 Disease ontology

Disease ontology was initially releasedin 2003 as part of a project of Northwestern University. It has been demonstrated its usefulness in several experiments and has also given very reasonable enrichment terms for many tissues. Osborne et al. produced an annotation of human genome with DO terms using GeneRif database [29]. Their validation data set suggested a much higher recall rate and precision rate against the widely used Online Mendelian Inheritance in Man(OMIM) annotations.However, relatively little attention has been given to DO. We could only find 16 articles from Pubmed using the key word “disease ontology” searching. There're also very few tools availableto make use of DO.

5 Conclusions

Knowledge enrichment analyses for HK and TS genes show that most of functions for these two groups of genes are consistent with their tissue origins. However some of them show unexpected functions, which may be vital clues for further exploring their functions or the mechanisms of some diseases. Our experiments also demonstrated that GSEA method is preferred in the verification of consistent knowledge, while HA in the discovery of ones for groups of genes. Therefore GSEA and HA can be used complementarily for the analysis of groups of genes.

Acknowledgements

This research is partly supported by the Natural Science Funding of China under grant number 61170177and innovation funding of Tianjin University.

References

[1] J. Zhu, F. He, S. Hu, and J. Yu. On the nature of human housekeeping genes. *Trends Genet*, 24 (10):481-4, 2008.

[2] N. Rosenfeld et al.. MicroRNAs accurately identify cancer tissue origin. *Nat. Biotechnology,* 26 (4):462-9, 2008.

[3] J. Gilbert, M. Haber, S. B. Bordow, G. M. Marshall, and M. D. Norris. Use of Tumor-Specific Gene Expression for the Differential Diagnosis of Neuroblastoma from Other Pediatric Small Round-Cell Malignancies. *Am. J. Pathol*., 155 (1):17-21, 1999.

[4] X. Yu, J. Lin, D. J. Zack, and J. Qian. Computational analysis of tissue-specific combinatorial gene regulation: predicting interaction between transcr iption factors in human tissues. *Nucleic Acids Research*, 34 (17):4925-36, 2006.

[5] K. E. Kouadjo, Y. Nishida, J. F. Cadrin-girard, M. Yoshioka, and J. St-amand. Housekeeping and tissue-specific genes in mouse tissues. *BMC Genomics,* 8:127, 2007.

[6] C. Zhang, Z. Zhang, J. Castle, S. Sun, and J. Johnson. Defining the regulatory network of the tissue-specific splicing factors Fox-1 and Fox-2. *Genes Dev.*, 22 (18):2550-63, 2008.

[7] D. Kuzmin et al.. Novel strong tissue specific promoter for gene expression in human germ cells. *BMC Biotechnology*, 10:58, 2010.

[8] T. Shlomi, M. N. Cabili, M. J. Herrgård, B. Ø. Palsson, and E. Ruppin. Network-based prediction of human tissue-specific metabolism. *Nature Biotechnology*, 26 (9):1003-10, 2008.

[9] A. I. Su et al.. Large-scale analysis of the human and mouse transcriptomes. *Proc. Natl. Acad. Sci. U.S.A.*, 99 (7):4465-70, 2002.

[10] V. E. Velculescu, L. Zhang, B. Vogelstein, and K. W. Kinzler. Serial analysis of gene expression. *Science*, 270 (5235):484-7, 1995.

[11] A. Visel et al.. ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature*, 457 (7231):854-8, 2009.

[12] Z. Tu, L. Wang, M. Xu, X. Zhou, T. Chen, and F. Sun. Further understanding human disease genes by comparing with housekeeping genes and other genes. *BMC Genomics*, 7:31, 2006.

[13] E. Schilling. Analysis of tissue-specific & allele-specific DNA methylation. *PhD thesis*, University Hospital Regensburg, 2010.

[14] A. Grosso, A. Gomes, and N. Barbosa. Tissue-specific splicing factor gene expression signatures. *Nucleic Acids Reserach*, 36 (15):4823-32, 2008.

[15] E. T. Wang et al.. Alternative isoform regulation in human tissue transcriptomes, *Nature*, 456 (7221):470-6, 2008.

[16] J. Schug, W. Schuller, and C. Kappen. Promoter features related to tissue specificity as measured by Shannon entropy. *Genome Biol.*, 6 (4):R33, 2005.

[17] F. Song et al. Tissue specific differentially methylated regions (TDMR): Changes in DNA methylation during development. *Genomics*, 93 (2):130-9, 2009.

[18] X. Liu, X. Yu, D. J. Zack, H. Zhu, and J. Qian. TiGER : A database for tissue-specific gene expression and regulation. *BMC Bioinformatics*, 9:271, 2008.

[19] X. Chen, J.-min Wu, K. Hornischer, A. Kel, and E. Wingender. TiProD: the tissue-specific promoter database. *Nucleic Acids Research*, 34 (Database issue):D104-7, 2006.

[20] L. Skrabanek and F. Campagne. TissueInfo: high-throughput identification of tissue expression profiles and specificity. *Nucleic Acids Research*, 29 (21):E102-2, 2001.

[21] S. Kogenaru, C. D. Val, A. Hotz-Wagenblatt, and K.H. Glatting. TissueDistributionDBs: a repository of organism-specific tissue-distribution profiles. *Theoretical Chemistry Accounts*, 125:651-658, 2010.

[22] S.-J. Xiao, C. Zhang, and Z.-L. Ji. TiSGeD: a database for tissue-specific genes. *Bioinformatics*, 26 (9):1273-5, 2010.

[23] L. A. Pennacchio, G. G. Loots, M. A. Nobrega, and I. Ovcharenko. Predicting tissue-specific enhancers in the human genome. *Genome Research*, 17 (2):201-11, 2007.

[24] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, and B. L. Ebert. Gene set enrichment analysis : A knowledge-based approach for interpreting genome-wide. *Proc. Natl. Acad. Sci. U.S.A.*, 102 (43):15545-50, 2005.

[25] Z. Dezso et al.. A comprehensive functional analysis of tissue specificity of human gene expression. *BMC Biology*, 6:49, 2008.

[26] M. Xilouri and L. Stefanis. Autophagy in the central nervous system: implications for neurodegenerative disorders. *CNS Neurol Disord Drug Targets*, 9 (6):701-19, 2010.

[27] R. Kumar, H. Atamna, M. N. Zakharov, S. Bhasin, S. H. Khan, and R. Jasuja. Role of the androgen receptor CAG repeat polymorphism in prostate cancer, and spinal and bulbar muscular atrophy. *Life Sci.*, 88 (13-14):565-71, 2011.

[28] L. L. Hsiao et al.. A compendium of gene expression in normal human tissues. *Physiol. Genomics*, 7 (2):97-104, 2001.

[29] J. D. Osborne et al.. Annotating the human genome with Disease Ontology. *BMC Genomics*, 10 Suppl (1):S6, 2009.