

Applications of high-throughput identification of tissue expression profiles and specificity

Introduction

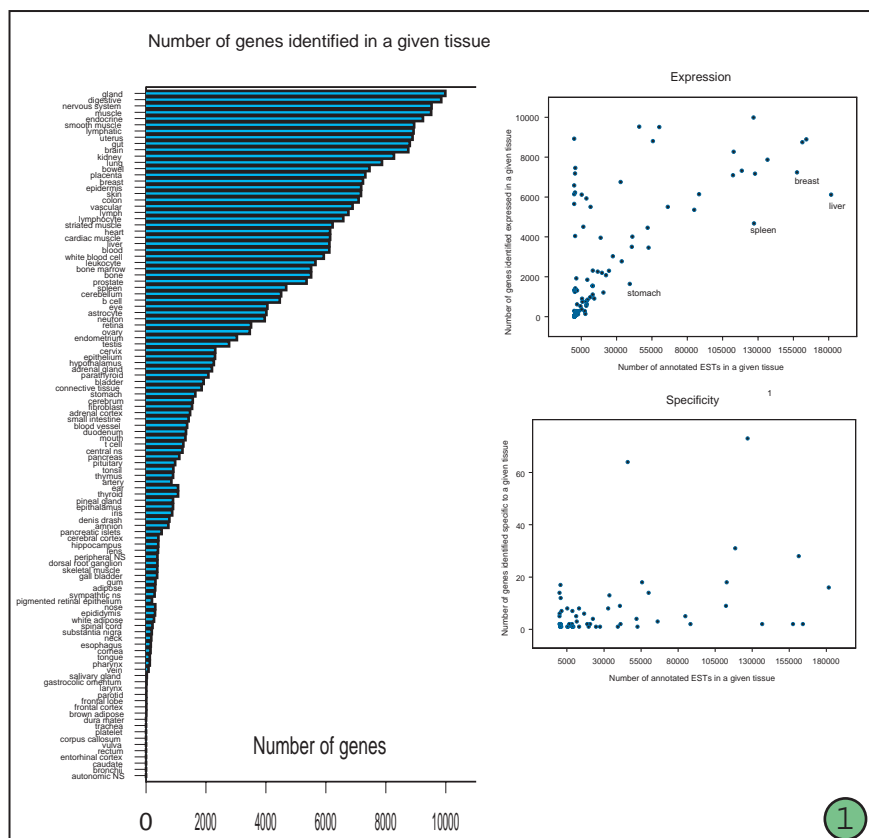
Organisms such as mammals do not express every single gene encoded by their genome in each of their cells. Rather, the various cell types of the organism express particular subsets of the genes in the genome. Cell types are further organized into tissues, and tissues constitute the organs that carry out various physiological functions. The detailed mechanisms of gene products underlying the functioning of this complex organization are today largely unknown. Several methods, including SAGE [Vel95] and microarray technology [Sho01] can be applied to the study of differential gene expression in the various cell types, in different tissues.

We recently developed **TissueInfo**, a high-throughput method to identify the tissue expression profile of the genes in an organism's genome, as well as the tissue specificity of a query sequence [Skr01]. The method carefully organizes the data publicly available in dbEST [Bog93] and is purely computational. With 80% coverage of the benchmark considered, **TissueInfo** achieves an accuracy of 76% when the tissue specificity of a gene is predicted and 89% when its expression in a given tissue is predicted. These results make possible the application of **TissueInfo** to the complete sequences available in the public draft of the human genome.

Here, we present applications of **TissueInfo** to genome-wide analysis of tissue expression, gene discovery, construction of tissue targeted microarray clone sets. Other applications include the assembly of training sets for the ab-initio prediction of tissue expression and specificity (promoter analysis).

Materials and Methods

Transcript sequences from the human genome project were obtained from the NCBI (ftp://ncbi.nlm.nih.gov/genomes/H_sapiens/mRNA/). The set of transcript sequences contained 25,612 transcripts in two sets: reference sequences (accession number starting with 'NM_', 12,685 sequences) which have been manually curated by NCBI staff, and transcript sequences computationally derived from the public assembly of the human genome (accession numbers starting with XM_, 12,927 sequences). As many transcripts overlap among these two datasets, we did the analysis with reference sequences only. Transcript sequences were masked to remove human repeats [Jur00] and used to search the human EST sequences with Tera-BLAST (Timelogic). The resulting data were filtered with timgblast as described in the **TissueInfo** publication [Skr01] with minimum-length=100 and max-error=0.05. We extracted a list of 104 human tissues from the **TissueInfo** annotation of dbEST. This list covers most tissues and organs represented in dbEST and adds anatomical groupings of tissues and organs, such as head (contains brain, eye, ear, etc.) or brain (contains hypothalamus, cortex, hippocampus, etc.). We used *tiquery* to compute the tissue information associated with each transcript.



Results and Evaluation

Three figures are grouped under point 1. The histogram on the left shows the number of genes identified expressed in a given tissue. The scatter plot titled "Expression" shows that there is a correlation between the number of genes predicted to be expressed in a given tissue and the number of ESTs present in dbEST, annotated as being prepared from this tissue. On the contrary, the same scatter plot obtained for genes specific to a given tissue shows no discernable correlation. Two tissues lie far below the median line: liver and spleen, suggesting that the estimate of the number of genes expressed in these two tissues will only slightly be improved as more ESTs are sequenced from these tissues.

Point 2 shows an evaluation of our results on a test set of 113 genes. Genes were selected to be included in the test set if identified as expressed in all the tissues of a cluster from the following list: [kidney & spleen & liver] (24 genes) [placenta & testis & spleen & liver] (1), [skin & heart & uterus] (11) [brain & kidney] (10), [colon & spleen & liver] (7), [lung & liver] (7), [brain&prostate] (6), [lung & brain] (6), [pituitary & placenta] (3), [placenta & breast](6) [placenta & spleen & liver] (6), [brain & spleen & liver] (5), [colon & bone marrow] (3) [lung & gut] (6) [nervous system & brain] (12). Each gene is tested for expression in each tissue belonging to a cluster and the literature is searched to check the identification. In 50% of the cases, the verification is not possible because the information is not available in the literature. In 77% of the cases where verification is possible, **TissueInfo** was accurate in identifying that a gene was expressed in a given tissue. These results are in good agreement with our previous evaluation of **TissueInfo** [Skr01].

Gene discovery

TissueInfo has applications in gene discovery as illustrated on point 3. Genes identified to be specific to retina (9) and ear (2) have been shown in greater detail. The 9 genes identified in the tables as specific of retina are verified in the literature. The two genes identified as specific to ear are dentin and a dentin homolog. Dentin is a protein in teeth, but no EST library have been made from teeth. What **TissueInfo** identifies is therefore a new organ in which the dentin gene could be expressed. Interestingly, missense mutations in the dentin gene are also associated with progressive hearing loss [Pat01], suggesting that the dentin gene, or a close homolog, could be involved in hearing. Furthermore, [Xia01] provides evidence that dentin is expressed in the mouse inner ear.

Our recent identification of the Sac sensory receptor gene candidate [Max01], is another example where the prediction of restricted tissue expression, or other specific expression profiles, can be pivotal in the identification of a gene candidate.

Novel statistical features

The histogram shown in point 4 plots the distribution of hits in dbEST for all the genes of the analysis set (genes with less than 3 hits are not shown). The distribution of hits shows two major peaks: one below 10 hits and a second centered on 80 hits. We verified that this feature is conserved in genes specific to a given tissue and noticed that most genes which are specific to a tissue and have more than 80 hits in dbEST are hormones or secreted proteins (6 over 7 of known function). This new type of results is now being analyzed for the possible physiological basis of the bi-modal distribution.

Construction of custom microarray clone sets

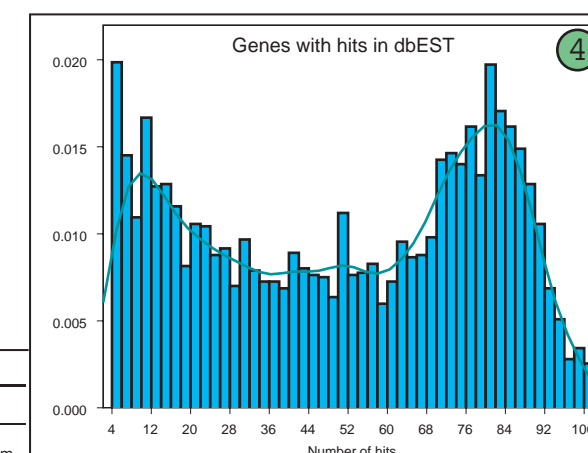
We have shown that the inferences produced by application of **TissueInfo** to genes for which data are not directly available in the literature should be of the same accuracy as those for genes described more completely. In addition, the method is scalable and therefore can be used to complement the data available in the literature and serve as a valuable resource for the selection of clones to build custom microarrays (e.g. liver or brain arrays).

tissue	#no literature	#literature	%no literature	#positive identification	%accuracy
kidney	14	20	41.2	14	70.0
liver	18	32	36.0	27	84.4
brain	14	25	35.9	24	96.0
colon	7	3	70.0	2	66.7
spleen	21	11	65.6	2	18.2
pituitary	0	3	0.0	2	66.7
placenta	8	8	50.0	7	87.5
breast	6	0	100.0	0	unknown
lung	14	5	73.7	3	60.0
heart	8	3	72.7	3	100.0
testis	0	1	0.0	1	100.0
prostate	3	3	50.0	3	100.0
Total	113	114	49.8	88	77.2

accession number	hit#	tissue specific?	null count	tissue summary	gene name	number of genes	tissue
NM_000172	25	yes	1	eye,retina	retinal transducin	73	gland
NM_000440	22	yes	0	endometrium,retina	phosphodiesterase 6A	64	nervous system
NM_000541	39	yes	4	eye,retina	retinal S-antigen	46	digestive
NM_000539	51	yes	2	eye,retina	rhodopsin	31	placenta
NM_000554	15	yes	0	retina	cone-rod homeobox	28	brain
NM_004312	15	yes	1	retina	retinal arrestin	18	kidney
NM_006671	6	yes	0	retina	retinal glutamate transporter	17	gut
NM_021728	15	yes	1	retina	OTX2	18	eye
NM_020366	12	yes	3	retina	retinitis pigmentosa GTPase regulat	17	eye

- dentin matrix acidic phosphoprotein (DMP1)
"DSPP, a gene encoding dentin sialophosphoprotein is processed into two proteins: dentin sialoprotein (DSP) and dentin phosphoprotein (DPP). ... Notably, missense mutations in DSPP are also associated with progressive hearing loss." [Pat01,Xia01]
- similar to dentin matrix acidic phosphoprotein

number of genes	tissue
73	gland
64	nervous system
46	digestive
31	placenta
28	brain
18	kidney
17	gut
17	eye
16	liver
14	muscle
14	lymphatic
13	testis
12	bowel
9	vascular
9	retina
9	colon
8	pancreas
8	lymph
8	blood
7	white blood cell
7	lens
6	lymphocyte
6	hypothalamus
5	prostate
5	pituitary
5	leukocyte
4	parathyroid
4	b cell
3	bone marrow
3	bone
2	uterus
2	striated muscle
2	smooth muscle
2	small intestine
2	neuron
2	lung
2	heart
2	ear
2	duodenum
2	cerebellum
2	central ns
2	cardiac muscle
2	breast
2	astrocyte
1	thyroid
1	t cell
1	stomach
1	salivary gland
1	pineal gland
1	pancreatic islets
1	ovary
1	fibroblast
1	epithelium
1	epithalamus
1	epididymis
1	endometrium
1	denis drash
1	connective tissue
1	adrenal gland



Genes specific to a given tissue w/ hit# >= 80:				Immediate early genes:			
AC	name	hit#	specific to	hormone	AC	name	hit#
NM_000515	GH1	81	pituitary	yes	NM_006365	c-fos	96
NM_001317	CSH1	99	placenta	yes	NM_001964	zif268	91
NM_003235	TG	98	thyroid	yes	NM_004839	Homer-2B	66
NM_002581	PAPPA	94	placenta	yes	NM_004907	ETR101	73
NM_020991	CSH2	97	placenta	yes	NM_003897	IER3	93
NM_002581	FTPI2	94	placenta	no/secreted			
NM_024603	orphan	95	head,brain	?			
NM_000670	ADH4	99	liver	no/enzyme			

References

[Vel95] Velculescu, V.E., et al., Serial analysis of gene expression. Science, 1995. 270(5235): p. 484-7.
[Sho01] Shoemaker, D.D., et al., Experimental annotation of the human genome using microarray technology. Nature, 2001. 409(6822): p. 922-7.
[Skr01] Skrabanek, L. and F. Campagne, TissueInfo: high-throughput identification of tissue expression profiles and specificity. submitted, 2001.
[Bog93] Boguski, M.S., T.M. Lowe, and C.M. Tolstoshev, dbEST--database for "expressed sequence tags". Nat Genet, 1993. 4(4): p. 332-3.
[Max01] Max M, Shanker YG, Huang L, Rong M, Liu Z, Campagne F, Weinstein H, Damak S, Margolske RF., Tas1r3, encoding a new candidate taste receptor, is allelic to the sweet responsiveness locus Sac. Nat. Genet. 2001. 28: p. 58-63.
[Sab93] Sabatini LM, Ota T, Azen EA. Mol Biol Evol 1993 May;10(3):497-511
[Opp88] Oppenheim FG, Xu T, McMillian FM, Levitz SM, Diamond RD, Offner GD, Troxler RF. J Biol Chem 1988 Jun 5;263(16):7472-7
[Xia01] Xiao S et al. Dentinogenesis imperfecta 1 with or without progressive hearing loss is associated with distinct mutations in DSPP. Nat Genet. 2001 Feb;27(2):201-4.
[Pat01] Patel P. Soundbites. Nat Genet. 2001 Feb;27(2):129-30.
[Jur00] Jurka, J. Repbase Update: a database and an electronic journal of repetitive elements. Trends Genet. 9:418-420 (2000)