T-STAG: resource and web-interface for tissue-specific transcripts and genes

Shobhit Gupta*, Martin Vingron and Stefan A. Haas

Computational Molecular Biology, Max Planck Institute for Molecular Genetics, Ihnestraße 63-73, D-14195 Berlin, Germany

ABSTRACT

T-STAG (tissue-specific transcripts and genes) is a resource and web-interface, designated to analyze tissue/tumor-specific expression patterns in human and mouse transcriptomes. It integrates our refined prediction of specific expression patterns both in genes as well as in individual isoforms with man–mouse orthology data. In combination with the features for combining/contrasting the genes expressed in different tissues, T-STAG implicates important biological applications, such as the detection of differentially expressed genes in tumors, the retrieval of orthologs with significant expression in the same tissue etc. Additionally, our refined categorization of expressed sequence tags (ESTs) according to the normalization of cDNA libraries allows searching for putative low-abundant transcripts. The results are tightly linked to our visualization tools, GeneNest (expression patterns of genes) and SpliceNest (gene structure and alternative splicing). The user-friendly interface of T-STAG offers a platform for comprehensive analysis of tissue and/or tumor-specific expression patterns revealed by the EST data. T-STAG is freely accessible at http://tstag.molgen.mpg.de.

INTRODUCTION

The complex differences in protein pools related to different cell types are a result of variation in the interpretation of the same genomic sequence. This variation is caused by various regulatory control mechanisms operating at transcriptional, post-transcriptional, translational and post-translational level. At the transcriptional level, control is achieved via transcription factors (1,2), which recognize certain *cis*-regulatory elements of the target genes and modulate their expression, occasionally in a tissue-specific manner (3,4). An additional regulatory mechanism is alternative splicing, which is controlled by exonic and intronic enhancers/silencers that allow differential expression of alternative mRNAs from the same primary transcript (5,6). Sometimes this mechanism also operates in a tissue-specific manner (7). Anomalous expression of the genes involved in these regulatory mechanisms is known to result in diseases (8,9).

Among the most popular methods to estimate and analyze expression patterns are serial analysis of gene expression [SAGE (10)] and expressed sequence tags [ESTs (11)] based methods. While both SAGE and ESTs are useful for analyzing expression patterns of genes [electronic northern (12)], a fraction of ESTs that cover isoform-specific parts [ASD (13), MAASE (14), ASAP (15) and SpliceNest (16)] also enable the detection of tissue/tumor specific alternative isoforms (17,18). However, the congruence between EST coverage and expression pattern is disturbed due to varying experimental protocols of EST generation (19), thereby implicating the need for a more refined methodology for estimating expression levels (20).

In the this paper, we describe T-STAG (tissue-specific transcripts and genes), a resource and web-interface, which integrates predictions of specific expression patterns of both genes as well as of individual isoforms, thus allowing to address additional or more specialized biological questions. Our detailed categorization of ESTs (normalized, disease related), the features to compare subsets of genes and the integration of tissue specific genes/isoforms implicate a wide range of applications. Among these are the detection of expression patterns of low-abundant transcripts and the identification of differential expression of genes in tumors. Above all, it allows for contrasting the tissue-specifically expressed isoforms with the background expression of all isoforms of the respective gene. The additional integration of man–mouse orthology data enables the comparison of expression profiles in orthologous genes. In combination with the user-friendly web-interface, T-STAG offers a platform for comprehensive analyses of expression patterns of genes as well as individual isoforms.

METHODS

T-STAG (http://tstag.molgen.mpg.de) is designed for detailed investigation of tissue/tumor specific expression in genes and

*To whom correspondence should be addressed. Tel: +49 30 8413 1163; Fax: +49 30 8413 1152; Email: gupta@molgen.mpg.de

Figure 1. The T-STAG query interface. The interface is arranged in three main sections: (i) basic information and chromosomal location: various gene identities, accessions, keywords and chromosomal location can be specified. (ii) Splicing information: this can be used to select types of splicing or define a quality cutoff for alternative splice prediction depending on the particular application. (iii) Tissue information: in this block, the user can specify the tissues of interest. Information related to second tissue can be used to specify additional tissues in which the candidate genes should (not) be expressed. The organism can be switched in order to look at human and mouse orthologs. To limit the selection to specifically expressed genes, the number of additional tissues with significant expression can be restricted.

transcripts predicted using EST data. The following resources are integrated via the web-database.

Gene expression estimates. The EST clusters (genes) and the annotation of EST libraries is derived from GeneNest database based on Unigene build 161 (August 2003) for human and Unigene build 118 (December 2002) for mouse (21). The tissue distribution of ESTs in a cluster relative to random background is translated into numerical estimates (*P*-values) of the likelihood of observing such a tissue distribution by chance (Haas, S. A. *et al*., manuscript in preparation). Therefore, a low *P*-value for a given gene–tissue pair reflects either significant and/or specific expression of the gene in the respective tissue.

Transcript expression data. The GeneNest (22) consensus sequences are mapped to the genome sequence (Human: April 2003 freeze of HUGO and Mouse: February 2002 freeze from the Mouse Genome Sequencing Consortium) and alternative isoforms are predicted with confidence values, using the EST coverage and splice signal indicators as a measure of reliability [SpliceNest (16,23)]. Parts of these putative transcripts that are specifically covered either by ESTs related to a single tissue or only by ESTs derived from tumor-related libraries are then labeled as tissue- or tumor-specific splice events, respectively (20).

Man–mouse orthologs. The human and mouse protein sequences are taken from RefSeq. Pairs of sequences with best bidirectional PBLAST alignment scores are defined as orthologs. The corresponding mRNA sequences are then inferred using TBLASTN of protein sequences with the respective reference sequences, thereby providing a link to the Unigene clusters.

Finally, the gene expression estimates, the transcript expression data and the man–mouse ortholog data is integrated via a relational database system (postgres). In order to enhance the

practical applicability of the resource, a user-friendly web interface (Figure 1) is designed with download option to facilitate integration of the data into external applications.

RESULTS AND DISCUSSION

Tissue-specific expression of genes and splice isoforms

Tissue-specific regulation of gene/isoform expression is known to play critical functional roles as in the case of many known genes, such as complement regulator *CD46* (24) and phosphodiesterase [*PDE7* (25)], apart from being associated with certain general mechanisms [tissue-specific RNA surveillance (26)]. The tissue-specific genes identified via the T-STAG database frequently include several known ones, as in the case of eye-specific genes, in which 19 out of the top 20 have already been described to be functionally related to eye (e.g. rhodopsin, crystallin, opticin etc.). A similar evaluation performed for alternative isoforms also revealed a number of already known tissue-specific splice events among the top ranking matches. For example, most of the known genes containing putative kidney-specific transcripts are experimentally described to contain kidney-related isoforms [*AFP* (27), *SLC22A8* (28), *WNK1* (29), *GLS* (30)]. The so far unannotated tissue-specific genes/isoforms with significant EST evidence need to be further analyzed, some of which after being screened for functional and/or possibly tissue-specific domains could be (partially) annotated.

Rare genes/alternative isoforms and disease-related genes/isoforms

The EST data provides an estimate of the low-abundant genes/ isoforms by the virtue of the differing protocols of EST

generation. Owing to the inherent over-representation of rare transcripts in normalized libraries (19), isoforms and genes that are represented only by such libraries are likely to be lowly expressed. This property of normalized libraries is utilized in the T-STAG to filter out those transcripts that are likely to be lowly expressed. A large fraction of the tissue-specific alternative isoforms is observed to be such lowly expressed ones (20), which even though in low abundance may still have crucial functions. For example, in one of the alternative isoforms of gene *WNK1*, an alternative promoter controls the expression of a kidney-specific and kinase-defective isoform (29).

Owing to our annotation of tumor- and disease-associated EST libraries, the T-STAG database allows the retrieval of genes/isoforms that are significantly expressed in tumor-or disease-related tissues. However, in tumor cells an overall loss of control is observed in different parts of regulation machinery (31,32), leading to a large number of genes with abnormal expression levels. Therefore, it is more informative to focus only on those genes that show significant differential expression in tumors as compared with the normal cell types.

Comparing expression patterns

In order to detect such genes that are differentially expressed in tumors, in accordance with some microarray-based methods (33), the predicted tumor-specific genes can be contrasted with another set of genes that are significantly expressed in the respective healthy tissue (defined in the form of *P*-values). This can be achieved by using the subtraction feature of the T-STAG database. Several of the top ranking genes revealed in this fashion are already known cancer-related genes. In brain tumor, for example, 6 of the top 10 genes have already been described to be tumor-associated. These include some genes which are suggested as tumor markers [*OLIG1* (34) and *CRF* (35)].

Alternatively, by using the addition feature of T-STAG, anatomically or functionally related tissues can be grouped together. For example, heart and muscle, for which six genes with significant expression in both tissues are revealed. This set includes *titin*, which is already known to play a critical role for both heart (36) and skeletal muscle (37). In addition, seemingly non-related pairs of tissues might also have biologically meaningful set of genes in common. For example, in the case of eye and pineal gland, we identified a group of genes (*CRX*, *OTX2* and *PDE6*), which are already annotated to be functional in both tissues [*PDE6* (38)]. Furthermore, *OTX2* is a known transcription factor that regulates the expression of the gene *CRX* both in eye and in pineal gland (39), thereby hinting toward the existence of a common functional/regulatory pathway in these tissues. Some of the remaining genes in the dataset, most of which are currently annotated to be functional only in eye (such as *RCV1*, *RTDBN*, potassium voltage-gated channel etc.) are therefore potential candidates that may be regulated by the same molecular mechanism.

With respect to the analysis of individual isoforms, the addition and subtraction features of the T-STAG database can be applied to further categorize the tissue-specific isoforms. First of the two categories consists of tissue-specific isoforms related to those genes for which other transcripts show different/ubiquitous expression pattern. Tissue-specific expression observed in such transcripts is likely to be regulated at the level of splicing (40). In contrast, the second category comprises tissue-specific splice events that are observed in genes, for which all related transcripts are also highly expressed in the same tissue. These transcripts may reflect tissue-specific transcription (41,42), rather than tissue-specific splicing. Notably, such observations may be biased due to other post-transcriptional events, such as nonsense-mediated decay [(NMD) (43)], which might occur with different stringencies in different tissues. In our data, we observe a large number of tissue-specific transcripts for both these categories, e.g. 187 human brain-specific transcripts potentially undergo specific alternative splicing, while 91 specific transcripts are likely to be the consequence of specific regulation of entire genes.

Evolutionarily conserved expression patterns

The integration of orthology data with expression data enables the retrieval of evolutionarily conserved expression patterns in mouse and human. This provides an additional schema for defining orthologs in a more stringent fashion. However, the emergence of expression in additional tissues, like in the case of gene *ACRBP* which is expressed only in mouse testis but is additionally expressed in human brain, may reflect evolution of novel functions.

The web-interface

The interface (Figure 1) is user-friendly and flexible with possibilities to define cutoffs (*P*-values for gene expression, quality values related to alternative splicing) based on individual applications. Additional restricted datasets based on individual applications can be generated by providing keywords and/or chromosomal location, thereby enabling queries like 'Give me *All kinases* expressed in human and mouse brain'. The HTML output provides tight links to the visualization tools, GeneNest (EST resource and visualization) as well as SpliceNest (gene structure and alternative splice visualization), which allows a detailed inspection of candidate genes and transcripts.

CONCLUSIONS

T-STAG is a resource and web-interface that allows comprehensive evaluation of tissue/tumor-specific expression both on the level of genes as well as on the level of individual transcripts. The resource is currently available for human and mouse with integrated man–mouse orthology data. In combination with the respective gene expression estimates, it provides an opportunity to compare expression patterns between orthologous genes. The comparison capability of the resource resolves the differential expression of genes both with respect to different tissues and with respect to normal versus tumor cell types. T-STAG also provides opportunity to categorize the tissue-specific transcripts that are potentially regulated at the transcriptional level and those that are likely to be tissue-specifically spliced. In essence, coupled with a comprehensive user-friendly web-interface, the T-STAG aims at serving as a resource for detailed computational analysis of expression patterns derived from EST data.

Future developments

Future development will include the prediction of developmental stage specific genes and isoforms. We plan to extend the database to include other organisms. Additionally, we plan to compare the EST-based gene expression estimates with gene expression profiles derived from microarray data. The consensus between these two independent datasets would provide a platform for the detection of common regulatory motifs among coexpressed genes.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

REFERENCES

1. Remenyi,A., Scholer,H.R. and Wilmanns,M. (2004) Combinatorial control of gene expression. *Nature Struct. Mol. Biol.*, **11**, 812–815.
2. Spiegelman,B.M. and Heinrich,R. (2004) Biological control through regulated transcriptional coactivators. *Cell*, **119**, 157–167.
3. Perrone-Bizzozero,N. and Bolognani,F. (2002) Role of HuD and other RNA-binding proteins in neural development and plasticity. *J. Neurosci. Res.*, **68**, 121–126.
4. Teunissen,B.E. and Bierhuizen,M.F. (2004) Transcriptional control of myocardial connexins. *Cardiovasc. Res.*, **62**, 246–255.
5. Ladd,A.N. and Cooper,T.A. (2002) Finding signals that regulate alternative splicing in the post-genomic era. *Genome Biol.*, **3**, Review 0008.
6. Caceres,J.F. and Kornblihtt,A.R. (2002) Alternative splicing: multiple control mechanisms and involvement in human disease, *Trends Genet.*, **18**, 186–193.
7. Ladd,A.N., Nguyen,N.H., Malhotra,K. and Cooper,T.A. (2004) CELF6, a member of the CELF family of RNA-binding proteins, regulates muscle-specific splicing enhancer-dependent alternative splicing. *J. Biol. Chem.*, **279**, 17756–17764.
8. Jin,X., Turcott,E., Englehardt,S., Mize,G.J. and Morris,D.R. (2003) The two upstream open reading frames of oncogene *mdm2* have different translational regulatory properties. *J. Biol. Chem.*, **278**, 25716–25721.
9. Csoka,A.B., English,S.B., Simkevich,C.P., Ginzinger,D.G., Butte,A.J., Schatten,G.P., Rothman,F.G. and Sedivy,J.M. (2003) Genome-scale expression profiling of Hutchinson–Gilford progeria syndrome reveals widespread transcriptional misregulation leading to mesodermal/mesenchymal defects and accelerated atherosclerosis. *Aging Cell*, **3**, 235–243.
10. Tuteja,R. and Tuteja,N. (2004) Serial analysis of gene expression (SAGE): unraveling the bioinformatics tools. *Bioessays*, **26**, 916–922.
11. Adams,M.D., Kelley,J.M., Gocayne,J.D., Dubnick,M., Polymeropoulos,M.H., Xiao,H., Merril,C.R., Wu,A., Olde,B., Moreno,R.F. *et al.* (1991) Complementary DNA sequencing: expressed sequence tags and human genome project. *Science*, **252**, 1651–1656.
12. Stanton,J.A., Macgregor,A.B. and Green,D.P. (2003) Identifying tissue-enriched gene expression in mouse tissues using the NIH UniGene database. *Appl. Bioinformatics*, **2**, S65–S73.
13. Thanaraj,T.A., Stamm,S., Clark,F., Riethoven,J.J., Le Texier,V. and Muilu,J. (2004) ASD: the Alternative Splicing Database. *Nucleic Acids Res.*, **32**, D64–D69.
14. Zheng,C.L., Nair,T.M., Gribskov,M., Kwon,Y.S., Li,H.R. and Fu,X.D. (2004) A database designed to computationally aid an experimental approach to alternative splicing. *Pac. Symp. Biocomput.*, 78–88.
15. Lee,C., Atanelov,L., Modrek,B. and Xing,Y. (2003) ASAP: The Alternative Splicing Annotation Project. *Nucleic Acids Res.*, **31**, 101–105.
16. Coward,E., Haas,S.A. and Vingron,M. (2002) SpliceNest: visualization of gene structure and alternative splicing based on EST clusters. *Trends Genet.*, **18**, 53–55.
17. Xu,Q. and Lee,C. (2003) Discovery of novel splice forms and functional analysis of cancer-specific alternative splicing in human expressed sequences. *Nucleic Acids Res.*, **31**, 5635–5643.
18. Xu,Q., Modrek,B. and Lee,C. (2002) Genome-wide detection of tissue-specific alternative splicing in the human transcriptome. *Nucleic Acids Res.*, **30**, 3754–3766.
19. Bonaldo,M.F., Lennon,G. and Soares,M.B. (1996) Normalization and subtraction: two approaches to facilitate gene discovery. *Genome Res.*, **6**, 791–806.
20. Gupta,S., Zink,D., Korn,B., Vingron,M. and Haas,S.A. (2004) Prediction and experimental evaluation of tissue-specific alternative transcripts derived from EST data. *BMC Genomics.*, **5**, 72.
21. Wheeler,D.L., Church,D.M., Federhen,S., Lash,A.E., Madden,T.L., Pontius,J.U., Schuler,G.D., Schriml,L.M., Sequeira,E., Tatusova,T.A. and Wagner,L. (2003) Database resources of the National Center for Biotechnology. *Nucleic Acids Res.*, **31**, 28–33.
22. Haas,S.A., Beissbarth,T., Rivals,E., Krause,A. and Vingron,M. (2000) GeneNest: automated generation and visualization of gene indices. *Trends Genet.*, **16**, 299–300.
23. Gupta,S., Zink,D., Korn,B., Vingron,M. and Haas,S.A. (2004) Genome-wide identification and classification of alternative splicing based on EST data, *Bioinformatics*, **20**, 2579–2585.
24. Russell,S.M., Sparrow,R.L., McKenzie,I.F. and Purcell,D.F. (1992) Tissue-specific and allelic expression of the complement regulator CD46 is controlled by alternative splicing. *Eur. J. Immunol.*, **22**, 1513–1518.
25. Bloom,T.J. and Beavo,J.A. (1996) Identification and tissue-specific expression of PDE7 phosphodiesterase splice variants, *Proc. Natl Acad. Sci. USA*, **93**, 14188–14192.
26. Bateman,J.F., Freddi,S., Nattrass,G. and Savarirayan,R. (2003) Tissue-specific RNA surveillance? Nonsense-mediated mRNA decay causes collagen X haploinsufficiency in Schmid metaphyseal chondrodysplasia cartilage. *Hum. Mol. Genet.*, **12**, 217–225.
27. Poliard,A., Feldmann,G. and Bernuau,D. (1998) Alpha fetoprotein and albumin gene transcripts are detected in distinct cell populations of the brain and kidney of the developing rat. *Differentiation*, **39**, 59–65.
28. Sweet,D.H., Miller,D.S., Pritchard,J.B., Fujiwara,Y., Beier,D.R. and Nigam,S.K. (2002) Impaired organic anion transport in kidney and choroid plexus of organic anion transporter 3 (Oat3 (Slc22a8)) knockout mice. *J. Biol. Chem.*, **277**, 26934–26943.
29. Delaloy,C., Lu,J., Houot,A., Disse-Nicodeme,S., Gasc,J., Corvol,P. and Jeunemaitre,X. (2003) Multiple promoters in the WNK1 gene: one controls expression of a kidney-specific kinase-defective isoform. *Mol. Cell. Biol.*, **24**, 9208–9221.
30. Modi,W.S., Pollock,D.D., Mock,B.A., Banner,C., Renauld,J.C. and Van Snick,J. (1991) Regional localization of the human glutaminase (GLS) and interleukin-9 (IL9) genes by in situ hybridization. *Cytogenet. Cell Genet.*, **57**, 114–116.
31. Corn,P.G. and El-Deiry,W.S. (2002) Derangement of growth and differentiation control in oncogenesis. *Bioassays*, **24**, 83–90.
32. Malumbres,M. and Carnero,A. (2003) Cell cycle deregulation: a common motif in cancer. *Prog. Cell Cycle Res.*, **5**, 5–18.
33. Anglesio,M.S., Evdokimova,V., Melnyk,N., Zhang,L., Fernandez,C.V., Grundy,P.E., Leach,S., Marra,M.A., Brooks-Wilson,A.R., Penninger,J. and Sorensen,P.H. (2004) Differential expression of a novel ankyrin containing E3 ubiquitin-protein ligase, Hace1, in sporadic Wilms' tumor versus normal kidney. *Hum. Mol. Genet.*, **13**, 2061–2074.
34. Lu,Q.R., Park,J.K., Noll,E., Chan,J.A., Alberta,J., Yuk,D., Alzamora,M.G., Louis,D.N., Stiles,C.D., Rowitch,D.H. and Black,P.M. (2001) Oligodendrocyte lineage genes (OLIG) as molecular markers for human glial brain tumors. *Proc. Natl Acad. Sci. USA*, **98**, 10851–10856.

35. Reubi,J.C., Waser,B., Vale,W. and Rivier,J. (2003) Expression of *CRF1* and *CRF2* receptors in human cancers. *J. Clin. Endocrinol. Metab.*, **88**, 3312–3320.

36. Granzier,H., Labeit,D., Wu,Y., Witt,C., Watanabe,K., Lahmers,S., Gotthardt,M. and Labeit,S. (2003) Adaptations in titin's spring elements in normal and cardiomyopathic hearts. *Adv. Exp. Med. Biol.*, **538**, 517–530.

37. Siebrands,C.C., Sanger,J.M. and Sanger,J.W. (2004) Myofibrillogenesis in skeletal muscle cells in the presence of taxol. *Cell Motil. Cytoskeleton*, **58**, 39–52.

38. Holthues,H. and Vollrath,L. (2004) The phototransduction cascade in the isolated chick pineal gland revisited. *Brain Res.*, **999**, 175–180.

39. Nishida,A., Furukawa,A., Koike,C., Tano,Y., Aizawa,S., Matsuo,I. and Furukawa,T. (2003) Otx2 homeobox gene controls retinal photoreceptor cell fate and pineal gland development. *Nature Neurosci.*, **6**, 1255–1263.

40. Hanamura,A., Caceres,J.F., Mayeda,A., Franza,B.R.,Jr and Krainer,A.R. (1998) Regulated tissue-specific expression of antagonistic pre-mRNA splicing factors. *RNA*, **4**, 430–444.

41. Odom,D.T., Zizlsperger,N., Gordon,D.B., Bell,G.W., Rinaldi,N.J., Murray,H.L., Volkert,T.L., Schreiber,J., Rolfe,P.A., Gifford,D.K., Fraenkel,E., Bell,G.I. and Young,R.A. (2004) Control of pancreas and liver gene expression by HNF transcription factors. *Science*, **303**, 1378–1381.

42. Pikkarainen,S., Tokola,H., Kerkela,R. and Ruskoaho,H. (2004) GATA transcription factors in the developing and adult heart. *Cardiovasc. Res.*, **63**, 196–207.

43. Lewis,B.P., Green,R.E. and Brenner,S.E. (2004) Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. *Proc. Natl Acad. Sci. USA*, **100**, 189–192.