BIOINFORMATICS ORIGINAL PAPER Vol. 28 no. 6 2012, pages 777–783
doi:10.1093/bioinformatics/bts056

Sequence analysis Advance Access publication January 27, 2012

The Transcriptome Analysis and Comparison Explorer—T-ACE: a platform-independent, graphical tool to process large RNAseq datasets of non-model organisms

E. E. R. Philipp[1,*,†], L. Kraemer[1,*,†], D. Mountfort[2], M. Schilhabel[1], S. Schreiber[1] and P. Rosenstiel[1,*]

[1]Department of Cell biology, Institute of Clinical Molecular Biology, Christian-Albrechts-University Kiel, Schittenhelmstrasse 12, 24105 Kiel, Germany and [2]Cawthron Institute, 98 Halifax Street East Nelson 7010, Private Bag 2, Nelson 7042, New Zealand

Associate Editor: Ivo Hofacker

ABSTRACT

Motivation: Next generation sequencing (NGS) technologies allow a rapid and cost-effective compilation of large RNA sequence datasets in model and non-model organisms. However, the storage and analysis of transcriptome information from different NGS platforms is still a significant bottleneck, leading to a delay in data dissemination and subsequent biological understanding. Especially database interfaces with transcriptome analysis modules going beyond mere read counts are missing. Here, we present the Transcriptome Analysis and Comparison Explorer (T-ACE), a tool designed for the organization and analysis of large sequence datasets, and especially suited for transcriptome projects of non-model organisms with little or no *a priori* sequence information. T-ACE offers a TCL-based interface, which accesses a PostgreSQL database via a php-script. Within T-ACE, information belonging to single sequences or contigs, such as annotation or read coverage, is linked to the respective sequence and immediately accessible. Sequences and assigned information can be searched via keyword- or BLAST-search. Additionally, T-ACE provides within and between transcriptome analysis modules on the level of expression, GO terms, KEGG pathways and protein domains. Results are visualized and can be easily exported for external analysis. We developed T-ACE for laboratory environments, which have only a limited amount of bioinformatics support, and for collaborative projects in which different partners work on the same dataset from different locations or platforms (Windows/Linux/MacOS). For laboratories with some experience in bioinformatics and programming, the low complexity of the database structure and open-source code provides a framework that can be customized according to the different needs of the user and transcriptome project.

Contact: e.philipp@ikmb.uni-kiel.de; l.kraemer@ikmb.uni_kiel.de; p.rosenstiel@mucosa.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on June 27, 2011; revised on January 20, 2012; accepted on January 24, 2012

1 INTRODUCTION

Recent advances in sequencing technology have led to an increasing accumulation of transcriptomic (RNAseq) data. Although most of the data has been generated in model organisms for which a high number of annotated sequences and complete genomes are available, increasing transcriptomic sequence data for non-model organisms are generated, for which only little or no *a priori* sequence information exists. The analysis of the obtained sequences and/or contigs, therefore, relies on comparative analysis with annotated genes or protein domains of other organisms. A multimodal comparison with a high number of sequences from different organisms and databases, e.g. NCBI, UniProtKB, Gene Ontology (GO), KEGG (Ashburner *et al.*, 2000; Kanehisa and Goto, 2000), should be pursued to gain a wider picture of the putative biological role of a specific sequence. This approach, however, increases the quantity of information per sequence and adds to the already large amount of data. Not only bioinformatic processing of sequences (i.e. cleaning, assembly and annotation) is still a bottleneck in RNASeq, but so is the subsequent analysis of annotated sequences. The analyzing scientist is faced with the problem of organizing the manifold BLAST hits, protein domains, GO terms and KEGG pathway information assigned to ten thousands of individual sequences, together with the mere sequence information such as nucleotide and translated protein sequence, protein domain organization or, in case of assembled contigs, read coverage.

To date a number of software solutions exist for the assembly and annotation of transcriptomes. Major efforts have been put into the development of reference-guided and *de novo* assembly tools. Several assemblers such as MIRA (Chevreux *et al.*, 2004), Newbler (Roche/454 Life Sciences), CAP3 (Huang and Madan, 1999), Velvet (Zerbino and Birney, 2008) and others were developed by researchers together with software and sequencing companies, of which some are especially designed for different NGS applications e.g. 454, Illumina or SOLID. It must be noted, that *de novo* RNA sequence assembly is a particularly difficult bioinformatic problem, as e.g. the existence of transcript isoforms usually results in many different contigs that cannot be merged in a simple fashion (for review see Kumar and Blaxter, 2010; Martin and Wang, 2011). Beyond the mere sequence assembly, a suite of software solutions for clustering (Partigen), protein prediction (prot4EST) and GO,

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

KEGG and EC annotation (annot8r, Blast2Go, AutoFACT) of non-model organism EST data has been developed by different working groups (Conesa *et al.*, 2005; Koski *et al.*, 2005; Parkinson *et al.*, 2004; Schmid and Blaxter, 2008; Wasmuth and Blaxter, 2004). To work with the results tables of the various software tools, non-bioinformatically skilled biologists, however, rely on user-friendly front ends which preferably work on various operating systems (Windows/Linux/MacOS). Within the Generic Model Organism Database project (GMOD, gmod.org), the web front end TRIPAL and also Gbrowse was developed for the Chado database structure. Also annot8r recommends visualization by Gbrowse, whereas Blast2Go has an integrated user interface to view and manage the annotated sequence data. These currently available interfaces are, however, predominantly focused on the visualization of genomes and their associated annotation-, expression- or publication data. Concerning transcriptome studies, however, databases and interfaces especially designed for transcriptome studies with analysis modules going beyond a mere display of read count numbers are still missing. Ideally, a software tool for transcriptome analysis will link all available information for a single transcript, and also enable a transcriptome wide overview and analysis on transcript and read count (expression) level. Information and results files should be exportable in a common format and the tool should allow an implementation of additional analysis modules for customization to specific needs. In order to serve these needs, we developed the software tool T-ACE with a Windows/Linux/MacOS interface to organize and analyze large amounts of transcriptome data, especially of non-model organisms with limited sequence information.

2 METHODS

2.1 Installation

All components of T-ACE are written as TCL scripts using the TCL/TK 8.5 software. Most of the scripts depend on additional TCL packages, such as: bwidget v1.9.2, tablelist, tclthread v2.6.5 (and libpgtcl v1.7, in case of the T-ACEpg version or the T-ACE_DB_Manager). For the full function of T-ACE, the additional software tools such as InterProScan v4.6 (Hunter *et al.*, 2009), NCBI-BLAST-2.2.25+, PHOBOS v3.3.2 (Mayer, 2007) and Primer3 v1.1.4 (Rozen and Skaletsky, 2000) are required. T-ACE is based on a PostgreSQL 8.4 database system. For non-local use, the PostgreSQL-server can be accessed via a PHP-enabled Apache 2.0 Web server. T-ACE and its necessary PostgreSQL-server runs on any standard computer, but the performance depends on the size of the examined dataset. The biggest dataset tested so far contains ∼120 000 contigs, 400 000 protein open reading frames (ORFs), over 3.1 million reads and according annotations. This database currently runs without difficulty on a dual-core unix system with 8 GB RAM. The T-ACE client can be executed on the same machine without memory issues or high processor load (4 GB RAM should be sufficient). Two different versions of T-ACE are currently available. The ‘T-ACE’ version accesses the PostgreSQL database through a php script, which has to run on the database server. With this version Pgtcl is not needed for running the T-ACE client. The ‘T-ACEpg’ version accesses the PostgreSQL database directly. For this version, the Pgtcl package is needed. In both versions, the T-ACE_DB_Manager accesses the database directly, therefore needs the Pgtcl package. A scheme of the T-ACE database is given in Supplementary Figure S1. After the required software is installed, T-ACE.tcl and T-ACE_DB_Manager.tcl should be executable. Detailed instructions, such as information about the additional software and its installation or how to set up the parent database, are described in the T-ACE manual and webpage (http://www.ikmb.uni-kiel.de/tace/).

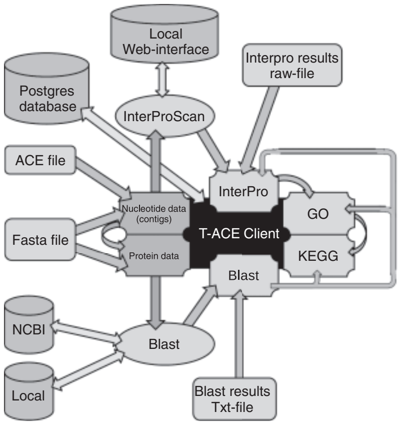


Fig. 1. Schematic overview of the generation and annotation of a T-ACE database. After creating a new database there are the following steps to perform: (i) adding sequence information: sequence information (protein or nucleotides) is added by uploading an ACE- or FASTA-file into the database. Uploading an ACE-file will not only add contig entries, but will also provide information about the positioning of the reads in each contig. Nucleotide/contig sequences can be directly translated into protein database entries. (ii) Blasting: nucleotide/contig or protein entries can be blasted at NCBI with the ‘NCBI-BLAST’-module or BLAST results imported into a database from a text file. The BLAST annotation can be used to deduce GO, InterProScan or KEGG annotations or Blast2GO results in .annot format can be imported into a database. (iii) InterProScan: contig or protein entries can be annotated by InterProScan, for this a custom installation of InterProScan with web interface is needed. Alternatively, InterProScan result files, in .raw format, can be imported into a database. InterProScan hits also lead to GO annotations, which can be used to deduce KEGG annotations.

2.2 Implementation

General comment: T-ACE was developed for the analysis and organization of transcriptome projects but is also helpful for the organization of small sequence datasets e.g. extracts of a large transcriptome databases. The current version of T-ACE does not provide an assembly function, thus data gained by NGS projects (e.g. 454, Illumina, SoliD) have to be assembled prior to the upload (e.g. using Newbler, Celera or TGICL). T-ACE currently accepts ACE or SAM files from different assembly and alignment programs. It must be emphasized that the choice of the assembler and the assembly strategy may result in slightly different models that represent a given transcriptome (Kumar and Blaxter, 2010; Martin and Wang, 2011). Depending on the type of study, the assembly strategy has to be carefully evaluated and highlighted assembly results have to be validated for each novel dataset. T-ACE does offer the option of an automatic BLAST against NCBI databases and InterProScan, but sequence annotation can be also undertaken outside T-ACE with sometimes more sophisticated annotation tools (e.g. Blast2Go) and the results files are loaded into T-ACE for further analysis (Fig. 1). For transcriptome comparisons, T-ACE is currently designed to work with a dataset composed of several transcriptomes of different treatments or tissues, which are assembled into a consensus transcriptome (Fig. 2). Subsequently, reads of the different transcriptomes are again mapped against the consensus transcriptome for information of transcriptome expression pattern i.e. number of reads per specific contig. Throughout the text, we will use the term ‘transcriptome’ for sequences (i.e. reads) gained from different samples and ‘consensus transcriptome’ for the contigs gained from the assembly of all ‘transcriptome’ sequences. Together with the annotations this builds the ‘database’.

Example datasets: in the following, we will give examples for different features of T-ACE using two independent datasets generated by 454 pyrosequencing (Roche/454 Life Sciences). One dataset was generated

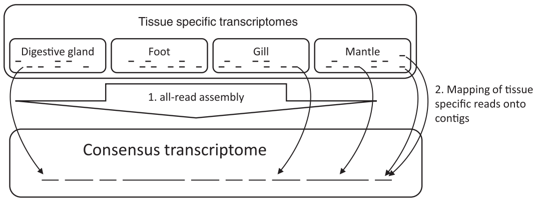

Tissue specific transcriptomes
Digestive gland
Foot
Gill
Mantle
1. all-read assembly
2. Mapping of tissue specific reads onto contigs
Consensus transcriptome

Fig. 2. Illustration of the assembly approach undertaken for the *M.galloprovincialis* dataset generated by Craft *et al.* (2010). Tissue-specific sequence reads were combined and assembled into a consensus transcriptome using the 'GS De novo Assembler 2.3' from Roche (Newbler, Roche/454 Life Sciences) and the TGICL (Cap3) assembler. Subsequently, tissue-specific reads were mapped against the consensus transcriptome using AMOScmp (http://sourceforge.net/apps/mediawiki/amos/index.php?title=AMOScmp) for the information of tissue-specific read number per contig i.e. tissue-specific transcript expression.

by Craft *et al.* (2010) for the marine mussel *Mytilus galloprovincialis* and was downloaded from the MG-RAST portal (download 12/10; http://metagenomics.nmpdr.org/; Meyer *et al.*, 2008). The second dataset contains sequences of the marine invader species *Sabella spallanzanii* and was sequenced at the Institute of Clinical Molecular Biology (ICMB Kiel, Germany) in cooperation with the Cawthron Institute (Dr D. Mountfort, New Zealand). Both datasets can be used to test T-ACE in the cloud. Detailed instructions how to load the T-ACE client and enter the databases are given on the T-ACE webpage (http://www.ikmb.uni-kiel.de/tace/) and in video tutorials on the webpage. The *M.galloprovincialis* database contains sequences of foot, gill, digestive gland and mantle tissue. Sequences were quality controlled and cleaned for primer and adapter sequences (Smart primer sequences and 454 adapters) as well as polyA tails by 'seqclean' and 'cln2qual' (TGI—The Gene Index Project) before the assembly. Reads <40 bp after the quality control and cleaning were excluded from further sequence assembly and annotation. The trimmed reads were assembled with the 'GS De novo Assembler 2.3' from Roche (Newbler, Roche/454 Life Sciences), as a first step. The standard parameters of the 'GS De novo Assembler 2.3' were used for this initial assembly ('minimum overlap length' = 40; 'minimum overlap identity' = 90). Afterwards the resulting contigs and singletons were further assembled in multiple rounds using the TGICL (Cap3) assembler. The 'minimum overlap length' varied between 40 and 300 bp and the 'minimum overlap identity' between 85 and 100%. In total, 104 123 MG-RAST sequences (average length 211 bp) of *M.galloprovincialis* were assembled into 12 827 contigs (average length 279 bp) and 40 972 singletons (average length 207 bp). Using AMOS (http://sourceforge.net), reads originating from the different tissues were subsequently mapped against the generated contigs and the *Mytilus edulis* mitochondrion genome (GI:55977238), which resulted in the final contigs and assigned reads from the different tissues (Fig. 2). The *S.spallanzanii* dataset contains 86 490 read sequences deduced from fan tissue which were assembled into 4714 contigs and 21 086 singletons. More detailed information of both datasets is given on the T-ACE webpage. For both datasets, putative gene names and protein domains were assigned to all contigs by using the BLASTx algorithm against the UniProtKB/Swiss-Prot and UniRef100 protein databases of UniProt Knowledgebase (UniProtKB, http://www.expasy.org/sprot/) with a cut off $e \leqslant 10^{-3}$,as well as tBLASTx ($e \leqslant 10^{-3}$) and BLASTn ($e \leqslant 10^{-10}$) against the NCBI nucleotide database (http://www.ncbi.nlm.nih.gov). The *S.spallanzanii* dataset was further analyzed for conserved domains by running the assembled contigs through InterProScan (1). GO terms were deduced from BLAST and InterProScan results and KEGG information from GO and BLAST results. Datasets as well as required databases (e.g. reference list, GO term list, etc.) for performing the below described analyses can be downloaded from the T-ACE webpage (http://www.ikmb.uni-kiel.de/tace/#Package/).

3 RESULTS

3.1 Working with T-ACE

Single sequence reads or assembled contigs, with their corresponding reads (e.g. ACE and SAM files), can be uploaded into the database together with information of independently performed BLAST- and protein domain-annotations. The current version of T-ACE, for example, supports the Blast2Go annot. and InterProScan files. Alternatively, BLAST annotation and protein domain identification (InterProScan) can be performed within T-ACE. When performed in T-ACE, GO and KEGG information is deduced from the BLAST and InterProScan entries. The nucleotide and translated protein sequence of individual sequences is then linked with the respective annotations. Different characters of a single sequence such as nucleotide and amino acid sequence, protein domain composition, ORF detection or contig coverage are visualized and can be curated manually within T-ACE. Such possibilities allow a detailed inspection and refinement of the results and goes beyond what is provided by other database interfaces, such as TRIPAL, Gbrowse or tbrowse (http://code.google.com/p/tbrowse/). At the whole database level, a complete overview of the sequence, GO term and KEGG pathway composition can be calculated and visualized within the interface. Data can be easily extracted for further external analysis and graphical processing. To identify genes of interest within the database, the tool offers the option for key word- or BLAST-searches. T-ACE is especially designed for the comparison of transcriptomes of non-model organisms without genome or large sequence information. Software solutions for such transcriptome comparisons are currently still missing, but are urgently needed due to the decreasing costs and increasing sequence numbers in NGS. T-ACE offers a first solution by using the information of the number of transcriptome specific single reads assigned to a defined contig, and changes in contig expression can be analyzed and visualized (further details below). In T-ACE, database organization and analysis is combined. To organize sequences and results, the tool enables individual databases to be set up, to sort sequences into projects or export data on the level of FASTA files or annotation tables for further analysis in external software solutions. T-ACE consists of a PostgreSQL database and a TCL Client interface, which allows multiple users to access the Postgres server from different computers and/or various operating systems (Windows/Linux/MacOS). Results saved within the database or in individual projects can thus be interchanged between different partners working on Windows, Linux or Mac platforms. On the one hand, T-ACE is attractive for laboratory environments, which have only a limited amount of bioinformatic support and for cooperating partners working on the same dataset from different locations. On the other, for laboratories with expertise in bioinformatics and programming, T-ACE provides a framework that can be extended by adding further modules, customized according to the different needs of the user and transcriptome project. The low complexity of the database allows an easy understanding of its structure and therefore facilitates the extension and integration of new tables and functions.

3.2 Functions of T-ACE

3.2.1 Whole database overview and analysis For a first overview of a database after the upload of sequences, basic information about the database content, such as the number and average length of nucleotide, protein and read sequences are displayed in the 'Database info' window (Supplementary Fig. S2). Number of annotation entries from externally performed analysis, or after BLAST and InterProScan annotation within T-ACE, are displayed according to the origin (e.g. GO, KEGG, Pfam). All databases can be exported directly from T-ACE, which facilitates the interchange or further analysis of data outside T-ACE. The 'Database info' window also gives information of the transcriptomes (e.g. different treatments, tissues, populations) included in the database. These are displayed in the 'run'-table and can be selected for whole database analysis and comparison between different transcriptomes (Supplementary Fig. S2). In the following, we will describe T-ACE functions by using a nucleotide database. It is also possible, however, to create a pure protein database or view a nucleotide database in a protein mode. By switching a nucleotide database into 'protein'-mode, the nucleotide sequence list of the 'Database browser' will be replaced by a list of all protein sequences contained in the database. In this way, annotations for distinct open reading frames of a contig can, for example, be reviewed.

Database sequence statistics: to get a first insight in how different RNAseq datasets/libraries compose the database, the nucleotide coverage and percentage of partially covered contigs of the consensus transcriptome can be calculated for transcriptomes selected within the 'run' table (Supplementary Fig. S3), and is executed via the 'Coverage' button. A more detailed statistical analysis about the database content is performed in the 'Database statistics' window. This menu allows an overview of sequence frequency on the level of reads or contigs, as well as contig and whole transcriptome coverage on the level of reads and nucleotides. Results are visualized as graphs within the tool or can be displayed as a list (Supplementary Fig. S4) and exported from T-ACE as txt/tables for analysis in external software solutions (e.g. Excel or GraphPad Prism).

Database GO statistics: in the GO statistics, the number of contigs of a specific GO term is listed and visualized for all levels of the GO tree (Supplementary Fig. S5). GO terms are deduced from the BLAST and InterProScan results of the consensus transcriptome and sorted into the different subcategories for molecular function, cellular component and biological process. Contigs detected within lower levels of the GO tree are also listed in the parent directory at higher levels of the GO tree but not counted twice. Alternatively, GO analysis can be conducted in Blast2GO and .annot files loaded into T-ACE. The GO table is graphically visualized or the list can be directly copied and imported in external software tools for graphical processing (e.g. for the generation of GO pie charts). Contigs belonging to a specific GO term can be directly exported by right mouse click into a new project file for detailed inspection. To investigate whether a group of specific contigs e.g. which show extremely high expression levels, represent distinct GO terms, entries of a project tab can be added as red bars to the GO tree diagram of the consensus transcriptome with the 'Compare'-button. A more detailed comparison of GO term patterns between and within RNAseq datasets can, however, be undertaken in the 'Run compare' tool, which is described below in the section of transcriptome comparisons.

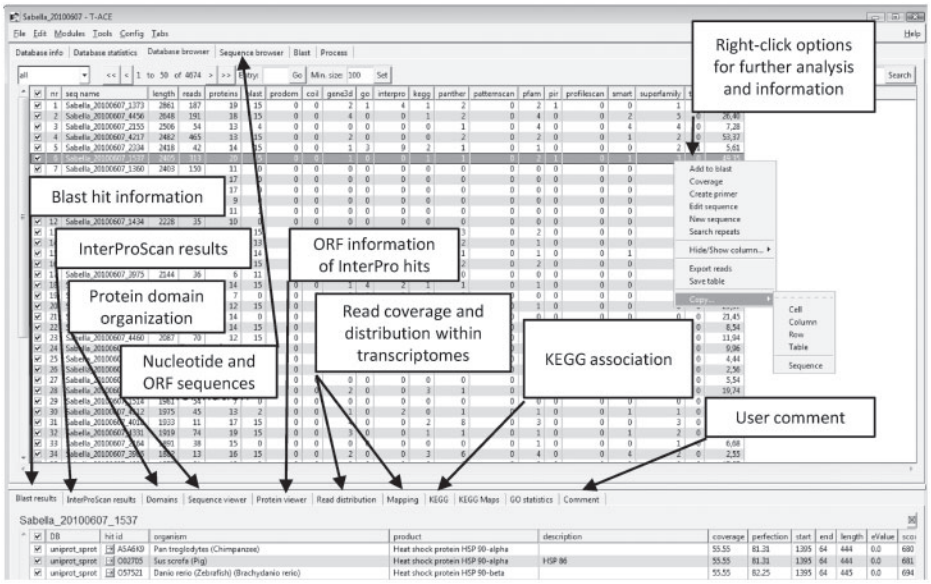
Database KEGG maps and composition: to analyze the gene composition and regulation of specific pathways within transcriptome data, the analysis of contigs on the basis of reference pathways from the Kyoto Encyclopedia of Genes and Genomes (KEGG) was implemented in T-ACE. KEGG pathway annotations for specific genes originate from the respective BLAST and/or GO annotations. Within the 'KEGG map' menu, an overview of all KEGG pathways identified in the consensus transcriptome is given together with the number of contigs covering this pathway. Further, the user gets an overview on which number and percentage (ko, ko%) of all pathway members within a specific reference pathway is covered by the contigs of the consensus transcriptome. Contigs are listed when clicking on the pathway to allow a more detailed investigation of the transcripts. Both tables have web links to either a pathway or a specific KEGG Orthology (KO)-ID (http://www.genome.jp/kegg/).

3.2.2 *Organization and overview of sequence information*

The organization and structured overview of sequence data within a database is an important component of T-ACE. In the 'Database Browser' window, all sequence entries of the database are listed together with the associated information such as sequence length and, in case of contigs, number of reads, as well as the number of different annotations (e.g. BLAST, GO, InterPro).

Working with single sequences and contigs: selection of a sequence entry will open detailed associated information like BLAST, GO and KEGG hits, InterProScan results, domain structure, read coverage and user-specific comments in different tabs in the lower windows of T-ACE (Fig. 3). Pop-up windows visualizing sequence read-coverage and ORF information can be opened by right click on single sequences (Fig. 4). Further options for processing single sequences can be selected e.g. by adding the sequence to the BLAST window or creating primers with Primer 3. This enables an immediate access and processing of different associated data of a sequence entry. Single and multiple sequences can be transferred and saved in a project file, which will help the user to organize groups of sequences. Project files can then be, for example, exchanged between research partners working on the same database but different platforms (Windows/Linux/MacOS).

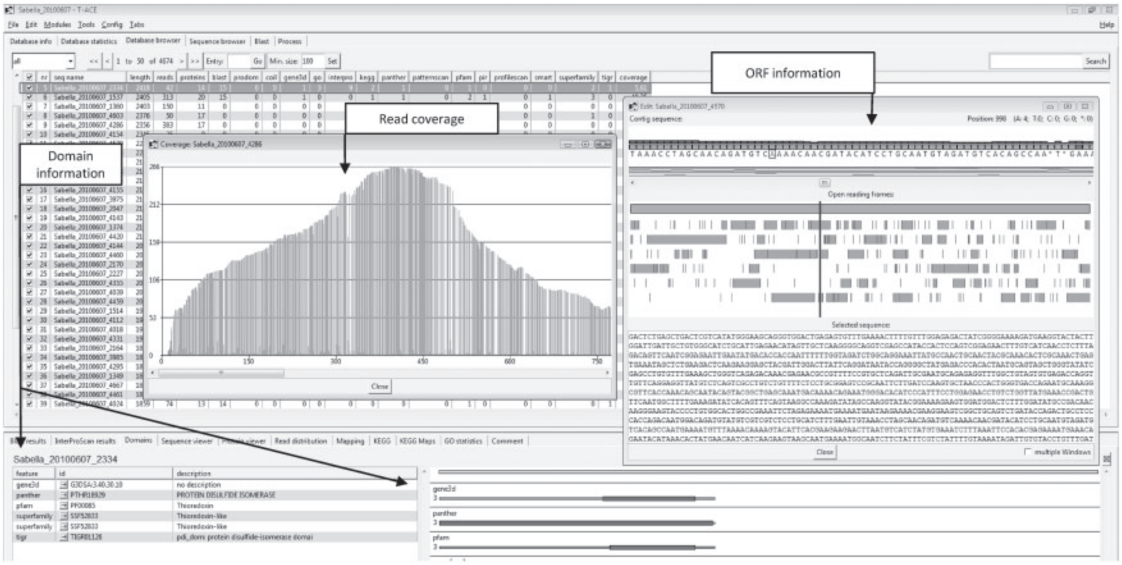
Search for target genes or protein domains: T-ACE offers database searches for target genes or protein domains either via keyword search or BLAST analysis. Keyword searches can be performed with user-specific filters for *e*-values or type of annotation in which should be searched (e.g. BLAST, KEGG). To perform BLAST searches with T-ACE, either a BLAST server or a local installation of NCBI- BLAST+ can be used and the required T-ACE databases have to be available as BLAST databases. For a local BLAST installation, this can easily be done by using the 'Create BLAST database'-module. The standard BLAST parameters are set through the 'BLAST parameters'-option in the 'Config'-menu. If a BLAST server is used for BLAST-analysis (selected in the 'Config'-menu 'BLAST configuration'), the 'Database'-combo box will contain a list of every database available to the user. When using the 'Local'-option, only databases situated in the BLAST_dbs folder in the T-ACE directory are listed in the 'Database'-combo box. BLAST results are displayed in a separate window in which the alignment of the BLAST matches can be given. By running the 'Mapping' option, the position of the different matches on the


Sabella_20100607 - T-ACE
File Edit Modules Tools Config Tabs
Database info | Database statistics | Database browser | Sequence browser | Blast | Process
Right-click options for further analysis and information
Blast hit information
InterProScan results
ORF information of InterPro hits
Protein domain organization
Read coverage and distribution within transcriptomes
Nucleotide and ORF sequences
KEGG association
User comment
Add to blast
Coverage
Create primer
Edit sequence
New sequence
Search repeats
Hide/Show column...
Export reads
Save table
Copy...
Cell
Column
Row
Table
Sequence
Blast results | InterProScan results | Domains | Sequence viewer | Protein viewer | Read distribution | Mapping | KEGG | KEGG Maps | GO statistics | Comment

Sabella_20100607_1537

☑	DB	hit id	organism	product	description	coverage	perfection	start	end	length	eValue	sco
☑	uniprot_sprot	A5A6K9	Pan troglodytes (Chimpanzee)	Heat shock protein HSP 90-alpha		55.55	81.31	1395	64	444	0.0	680
☑	uniprot_sprot	O02705	Sus scrofa (Pig)	Heat shock protein HSP 90-alpha	HSP 86	55.55	81.31	1395	64	444	0.0	681
☑	uniprot_sprot	O57521	Danio rerio (Zebrafish) (Brachydanio rerio)	Heat shock protein HSP 90-beta		55.55	82.25	1395	64	445	0.0	694

Fig. 3. Information assigned to single sequences or contigs are listed and visualized in different windows of T-ACE.


Sabella_20100607 - T-ACE
File Edit Modules Tools Config Tabs
Database info | Database statistics | Database browser | Sequence browser | Blast | Process
ORF information
Read coverage
Domain information
Coverage: Sabella_20100607_4286
0
150
300
450
600
750
Close
Edit: Sabella_20100607_4570
Contig sequence:
Open reading frames
Selected sequence:
multiple Windows
Blast results | InterProScan results | Domains | Sequence viewer | Protein viewer | Read distribution | Mapping | KEGG | KEGG Maps | GO statistics | Comment
gene3d
panther
pfam

Sabella_20100607_2334

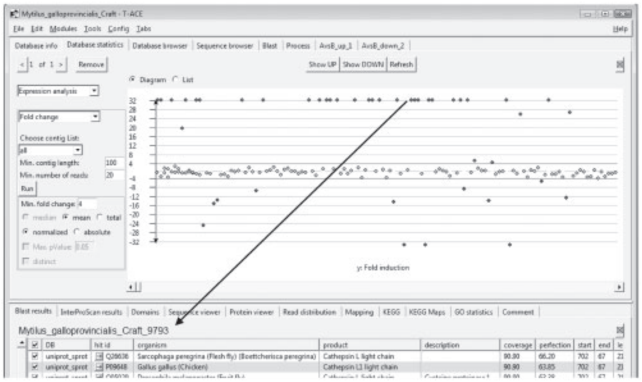
feature	id	description
gene3d	G3DSA:3.40.30.10	no description
panther	PTHR18929	PROTEIN DISULFIDE ISOMERASE
pfam	PF00085	Thioredoxin
superfamily	SSF52833	Thioredoxin-like
superfamily	SSF52833	Thioredoxin-like
tigr	TIGR01126	pdi_dom: protein disulfide-isomerase domai

Fig. 4. Visualization of the open reading frames, read coverage and domain organization of sequence entries of the *Sabella* database.

query can be visualized (Supplementary Fig. S6). Within BLAST, the selection of 'overlaps only' enables a result filter that, when activated, displays only BLAST hits which reach the start or end of the query sequence. This option is useful when searching for sequences elongating a given query.

3.2.3 Transcriptome comparison An important aspect when working with RNAseq datasets from different samples (e.g. different tissues or treatments) is the easy accessibility of information about transcriptomal changes in pathways described by GO/KEGG terms, changes in domain abundance and gene expression on the contig level. Besides T-ACE, only the Blast2Go tool offers a first statistical analysis between transcriptomes. This is, however, restricted to GO-terms. In T-ACE, the information of sequence reads from different transcriptomes assigned to specific contigs of the consensus transcriptome within the database is needed as a prerequisite for transcriptome analysis. In case of a *de novo* assembly, as done for *M.galloprovincialis* and *S.spallanzanii*, all reads from the different transcriptomes are assembled into contigs and the information of transcriptome-specific reads per database contig are extracted (Fig. 2). In T-ACE, different options for transcriptome comparisons are implemented that can be executed in the 'Database statistic' window or the 'Tool' drop-down menu.

Expression analysis: the 'Expression analysis' tool investigates the origin, i.e. transcriptomal dataset affiliation of reads contained in a contig. For this, transcriptomes of the run table in the 'Database



● Diagram ○ List

Choose contig List: all

Min. contig length: 100

Min. number of reads: 20

Min. fold change: 4

○ median ● mean ○ total

● normalized ○ absolute

☐ Max. pValue: 0.05

☐ distinct

Mytilus_galloprovincialis_Craft_9793

☑	DB	hit id	organism	product	description	coverage	perfection	start	end	le
☑	uniprot_sprot	Q26636	Sarcophaga peregrina (Flesh fly) (Boettcherisca peregrina)	Cathepsin L light chain		90.80	66.20	702	67	21
☑	uniprot_sprot	P09648	Gallus gallus (Chicken)	Cathepsin L1 light chain		90.80	63.85	702	67	21

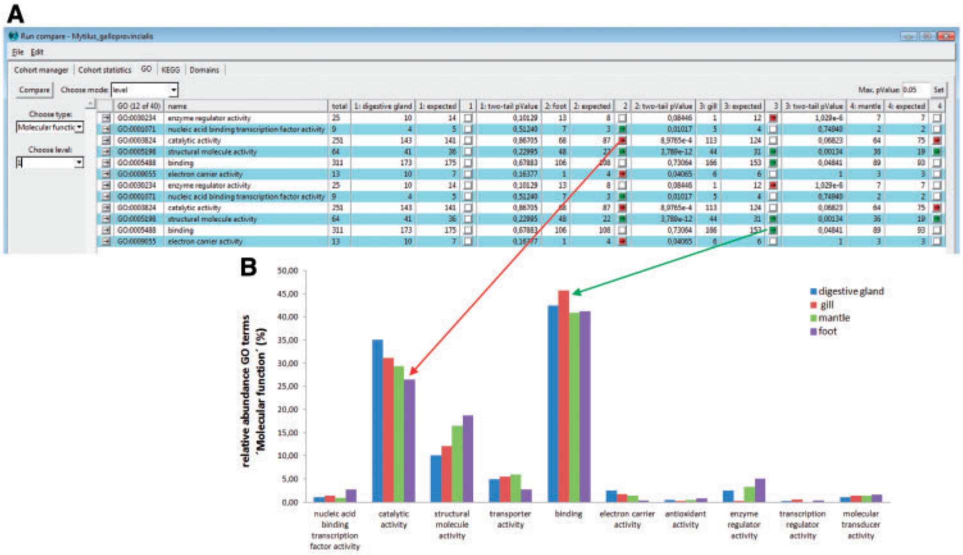
Fig. 5. Visualization of expression analysis when comparing different transcriptomes. As an example, the gill tissue transcriptome of the *M.galloprovincialis* database was set as group A and digestive gland tissue as group B and the fold change calculated for contigs >100 bp and containing minimum 20 reads. Each dot represents a sequence entry. Dots can be selected by mouse click and the information associated to the respective sequence is displayed in the different windows of T-ACE.

info'-window are marked as defined groups (A or B). Results can be displayed as the number of 'A and 'B reads per contig on the linear or logarithmic scale as well as fold change in a list or graph (Fig. 5). If the compared transcriptomes are composed of a different number of assembled sequences, T-ACE gives the option to calculate the results on normalized values. The results can be filtered by e.g. setting minimum contig length, read number or the fold change threshold. Statistical analysis is undertaken when the A and B group each consists >2 transcriptomes. Fold change values are calculated for mean or median values and a P-value for significantly up- or downregulated contigs between groups is calculated by the Mann–Whitney U test. P-values are corrected for multiple testing using the Benjamini–Hochberg Step-Up false discovery rate-controlling procedure. Due to the still high sequencing costs, however, generally a low number of transcriptomes per treatment group (biological replicates) are generated, which is in most cases not suitable for statistical calculation and leading to non- significant P-values ($P > 0.05$) for up- or downregulated genes. To yet reduce the number of genes to be chosen for a subsequent more detailed investigation of expression changes e.g. by real time PCR, T-ACE holds the option to filter for 'distinct' differences. The number of reads per contig within groups A and B are ranked and the two groups are 100% distinctly different when the transcriptome with the lowest number of reads for a specific contig within one group (e.g. A) still has a higher number of reads compared with the transcriptome with the highest number of reads per contig in the other group (e.g. B). The percentage (%) of distinction can be set by the user. The obtained sequence entries for regulated genes can be transferred to separate project tabs by the 'Show UP'- and/or 'Show DOWN'-button for further analysis or export.

Analysis of GO term, KEGG pathway member or protein domain distribution within and between transcriptomes: in some cases, the composition of GO terms, KEGG pathway members or protein domains in different transcriptomal datasets or in the group of up- and downregulated genes of treatments can be more informative than investigating expression changes on the level of single genes. A high number of small expression changes of several genes within one pathway may not be detected as a major result when only looking at the single genes, but may be striking when investigated on the whole pathway level. The 'Run compare' tool compares multiple subsets of contigs (called cohorts) against the combined contigs of all subsets. This gives a statistical estimation on which GO terms, KEGG pathways or domains are enriched or depleted within the respective transcriptome or treatment group. Calculations can be conducted for the whole set of different GO terms and KEGG pathways, or for specific GO levels and KEGG pathways. P-values are calculated by performing the Fisher's exact test. In Figure 6A, a typical run compare table is shown with significant enriched- or depleted-GO terms marked in green (enriched) or red (depleted). Data obtained by the run compare tool can then be exported for further analysis and graphical display as, for example, undertaken for the GO subcategory 'Molecular function' for different tissues of *M.galloprovincialis* (Fig. 6B). A similar approach is undertaken in the Blast2Go software, which uses GOSSIP (Bluthgen *et al.*, 2005) to compare enriched GO terms between two datasets. BioMyn (http://www.biomyn.de/explore/) and skypainter (http://www.reactome.org/cgi-bin/skypainter2) are other online tools in which two lists of genes with identifiers can be uploaded for GO and pathway analysis. The analysis is, however, restricted to a limited number of species and excludes GO, KEGG and domain analysis of sequences without a proper gene BLAST hit i.e. with only a domain annotation. The advantage in T-ACE is that the two datasets to be compared do not need to be annotated and loaded separately but are within one T-ACE project. The calculation is not based on the gene identifier of a distinct species but on the previous performed GO, KEGG and domain annotation and further, not only overrepresentation but also underrepresentation of the respective terms or pathways is calculated.

3.3 Conclusion and outlook

The T-ACE was especially developed for NGS transcriptome projects of non-model organisms where significant *a priori* sequence information is missing on DNA and RNA level. We wanted to design a software tool, which goes beyond a webpage application for gene mining by keyword or BLAST searches. We also explicitly did not aim to compete with larger central sequence databases (e.g. short read archives of EBI and NCBI) that allow sharing of the unannotated data with the wider public. T-ACE was built in order to provide a graphical interface that can be used locally among different scientists in a single lab and also between collaborating laboratories in order to work with large amounts of transcriptome sequence data. T-ACE exhibits modules for manual curation and visualization of single contigs and underlying reads. Furthermore, statistical tools have been implemented for the analysis of differential expression or occurrence of GO, KEGG terms and protein domains comparison of transcriptomes. The relatively simple bioinformatic structure sets a framework for further integration of analysis modules and customization of the tool. Future development of T-ACE will particularly focus on such additional modules for transcriptome comparisons in the light of the increasing use of RNAseq data as 'virtual microarrays' i.e. a consensus transcriptome will be used as a 'virtual microarray' against which sequence data of short-read sequencing technologies (e.g. SoliD, Illumina) are mapped for gene expression analysis.



	GO (12 of 40)	name	total	1: digestive gland	1: expected	1	1: two-tail pValue	2: foot	2: expected	2	2: two-tail pValue	3: gill	3: expected	3	3: two-tail pValue	4: mantle	4: expected	4
⊞	GO:0030234	enzyme regulator activity	25	10	14	□	0,10129	13	8	□	0,08446	1	12	🟥	1,029e-6	7	7	□
⊞	GO:0001071	nucleic acid binding transcription factor activity	9	4	5	□	0,51240	7	3	🟩	0,01017	5	4	□	0,74940	2	2	□
⊞	GO:0003824	catalytic activity	251	143	141	□	0,86705	68	87	🟥	8,9765e-4	113	124	□	0,06823	64	75	🟥
⊞	GO:0005198	structural molecule activity	64	41	36	□	0,22995	48	22	🟩	3,789e-12	44	31	🟩	0,00134	36	19	🟩
⊞	GO:0005488	binding	311	173	175	□	0,67883	106	108	□	0,73064	166	153	🟩	0,04841	89	93	□
⊞	GO:0009055	electron carrier activity	13	10	7	□	0,16377	1	4	🟥	0,04065	6	6	□	1	3	3	□
⊞	GO:0030234	enzyme regulator activity	25	10	14	□	0,10129	13	8	□	0,08446	1	12	🟥	1,029e-6	7	7	□
⊞	GO:0001071	nucleic acid binding transcription factor activity	9	4	5	□	0,51240	7	3	🟩	0,01017	5	4	□	0,74940	2	2	□
⊞	GO:0003824	catalytic activity	251	143	141	□	0,86705	68	87	🟥	8,9765e-4	113	124	□	0,06823	64	75	🟥
⊞	GO:0005198	structural molecule activity	64	41	36	□	0,22995	48	22	🟩	3,789e-12	44	31	🟩	0,00134	36	19	🟩
⊞	GO:0005488	binding	311	173	175	□	0,67883	106	108	□	0,73064	166	153	🟩	0,04841	89	93	□
⊞	GO:0009055	electron carrier activity	13	10	7	□	0,16377	1	4	🟥	0,04065	6	6	□	1	3	3	□

B
relative abundance GO terms 'Molecular function' (%)
50,00
45,00
40,00
35,00
30,00
25,00
20,00
15,00
10,00
5,00
0,00
nucleic acid binding transcription factor activity
catalytic activity
structural molecule activity
transporter activity
binding
electron carrier activity
antioxidant activity
enzyme regulator activity
transcription regulator activity
molecular transducer activity
🟦 digestive gland
🟥 gill
🟩 mantle
🟪 foot

Fig. 6. (**A**) Example for a run compare table showing significant enriched (marked green) or depleted (marked red) GO terms, calculated for the subcategory 'Molecular function' in digestive gland, gill, mantle and foot tissue and compared with all GO terms found within the reference database (all tissues combined). GO terms were deduced form BLAST hits using the BLASTX algorithm against the UniProtKB/Swiss-Prot and UniRef100 protein databases of UniProt Knowledgebase (UniProtKB, http://www.expasy.org/sprot/) with a cut off $e \leqslant 10^{-3}$,as well as tBLASTx ($e \leqslant 10^{-3}$) and BLASTn ($e \leqslant 10^{-10}$) against the NCBI non-redundant protein (nr) database (http://www.ncbi.nlm.nih.gov). (B) Relative abundance of GO terms (%) for the subcategory Molecular function calculated from data obtained by the 'run compare' tool. Arrows indicate one significantly enriched and depleted GO term within the respective tissue.

Taken together, T-ACE has been designed for scientists in different laboratories working cooperatively on a central database and allows its users access from different terminals and platforms (Linux/Windows/MacOS).

ACKNOWLEDGEMENTS

The authors would like to thank the cell biology group and sequencing platform of the IKMB for technical support, especially Tanja Kaacksteen, Melanie Friskovec and Anita Dietsch. We thank Dr Georg Hemmrich-Stanisak for helpful comments while writing the manuscript as well as three anonymous reviewers.

Funding: DFG clusters of Excellence 'The Future Ocean' and 'Inflammation at Interfaces' and the DFG priority programme 1399 'Host-parasite covolution', Genomics Analysis platform.

Conflict of Interest: none declared.

REFERENCES

Ashburner,M. *et al.* (2000) Gene Ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.

Bluthgen,N. *et al.* (2005) Biological profiling of gene groups utilizing Gene Ontology. *Genome Inform.*, **16**, 106–115.

Chevreux,B. *et al.* (2004) Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. *Genome Res.*, **14**, 1147–1159.

Conesa,A. *et al.* (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, **21**, 3674–3676.

Craft,J.A. *et al.* (2010) Pyrosequencing of *Mytilus galloprovincialis* cDNAs: tissue-specific expression patterns. *PLoS One*, **5**, e8875.

Huang,X. and Madan,A. (1999) CAP3: A DNA sequence assembly program. *Genome Res.*, **9**, 868–877.

Hunter,S. *et al.* (2009) InterPro: the integrative protein signature database. *Nucleic Acids Res.*, **37**, D211–D215.

Kanehisa,M. and Goto,S. (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.*, **28**, 27–30.

Koski,L. *et al.* (2005) AutoFACT: An Automatic Functional Annotation and Classification Tool. *BMC Bioinformatics*, **6**, 151.

Kumar,S. and Blaxter,M. (2010) Comparing de novo assemblers for 454 transcriptome data. *BMC Genomics*, **11**, 571.

Martin,J.A. and Wang,Z. (2011) Next-generation transcriptome assembly. *Nat. Rev. Genet.*, **12**, 671–682.

Mayer,C. (2007) PHOBOS – a tandem repeat search tool for complete genomes. http://www.rub.de/spezzoo/cm.

Meyer,F. *et al.* (2008) The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*, **9**, 386.

Parkinson,J. *et al.* (2004) PartiGene—constructing partial genomes. *Bioinformatics*, **20**, 1398–1404.

Rozen,S. and Skaletsky,H. (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods Mol. Biol.*, **132**, 365–386.

Schmid,R. and Blaxter,M.L. (2008) annot8r: rapid assignment of GO, EC and KEGG annotations. *BMC Bioinformatics*, 2008, **9**, 180.

Wasmuth,J. and Blaxter,M. (2004) prot4EST: Translating Expressed Sequence Tags from neglected genomes. *BMC Bioinformatics*, **5**, 187.

Zerbino,D. and Birney,E. (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.*, **18**, 821–829.