



## PartiGene—constructing partial genomes

John Parkinson\*, Alasdair Anthony, James Wasmuth, Ralf Schmid, Ann Hedley and Mark Blaxter

School of Biological Sciences, Ashworth Laboratories, King's Buildings, West Mains Rd, University of Edinburgh, Edinburgh EH9 3JT, UK

Received on July 31, 2003; revised on December 9, 2003; accepted on December 11, 2003  
Advance Access publication February 26, 2004

### ABSTRACT

Expressed sequence tags (ESTs) offer a low-cost approach to gene discovery and are being used by an increasing number of laboratories to obtain sequence information for a wide variety of organisms. The challenge lies in processing and organizing this data within a genomic context to facilitate large scale analyses. Here we present PartiGene, an integrated sequence analysis suite that uses freely available public domain software to (1) process raw trace chromatograms into sequence objects suitable for submission to dbEST; (2) place these sequences within a genomic context; (3) perform customizable first-pass annotation of the data; and (4) present the data as HTML tables and an SQL database resource. PartiGene has been used to create a number of non-model organism database resources including NEMBASE (<http://www.nematodes.org>) and LumbrBase (<http://www.earthworms.org/>). The packages are readily portable, freely available and can be run on simple Linux-based workstations.

**Availability:** PartiGene is available from <http://www.nematodes.org/PartiGene> and also forms part of the EST analysis software, associated with the Natural Environmental Research Council (UK) Bio-Linux project (<http://envgen.nox.ac.uk/biolinux.html>).

**Contact:** [jparkin@sickkids.ca](mailto:jparkin@sickkids.ca)

### INTRODUCTION

The advent of low-cost, high-throughput sequencing has permitted the generation of fully sequenced genomes of a number of model organisms including 122 prokaryotic and 17 eukaryotic species (<http://wit.integratedgenomics.com/GOLD/>). For these fully sequenced genomes, integrated databases are used to contextualize sequence data within a rich biological information environment. An increasing amount of sequence data is being generated from a range of other, non-model organisms. For eukaryotic species, these sequence data are typically in the form of expressed sequence tags (ESTs). Datasets range from just

a few hundred to as many as several hundred thousand sequences. There are over 180 species, with more than 1000 entries, in the database for ESTs (dbEST, <http://www.ncbi.nlm.nih.gov/dbEST/index.html>) (Boguski *et al.*, 1993). In general, these data are not well organized and are difficult to interpret in a genomic context.

Common problems include significant redundancy in the datasets (some genes may have been sequenced multiple times) and a lack of consistent annotation between projects. An effective way to overcome these problems is to group ESTs into clusters that represent genes and to provide annotations for the clusters. Since ESTs provide only a fraction of the available genes for a particular organism, we refer to these analysed datasets as partial genomes. Informatic solutions to produce partial genomes or 'gene indices' have been developed by several groups (Adams *et al.*, 1995; Boguski and Schuler, 1995; Sutton *et al.*, 1995; White and Kervalage, 1996; Christoffels *et al.*, 2001; Perlea *et al.*, 2003). The analysis of partial genomes has tended to involve complex integrated database solutions and/or a large amount of manual sequence annotation, both of which require a considerable investment in bioinformatic resources and make cross-species and between-lab integration difficult.

Our involvement in a wide range of different EST projects (Allen *et al.*, 2000; Daub *et al.*, 2000; Blaxter *et al.*, 2002; Kenyon *et al.*, 2003; Parkinson *et al.*, 2003) has led us to develop a generic, automated software pipeline, PartiGene, that handles an EST project from raw trace data through to a partial genome database ready for data mining. In this it goes beyond a simple EST-focussed LIMS system and other solutions to EST processing such as that of Paquola *et al.* (2003). PartiGene consists of three integrated scripts, based on the PERL scripting language, which rely on freely available public domain software. PartiGene is readily portable to most UNIX-based operating systems and is freely available from our Web server (<http://www.nematodes.org/PartiGene>). We have designed PartiGene to be freestanding, permitting installation and operation with a minimum of background expert knowledge. In addition to being portable and customizable, PartiGene offers further advantage over similar pipelines

\*To whom correspondence should be addressed at Programs in Genetics and Genomic Biology & Structural Biology and Biochemistry, Hospital for Sick Children, 555 University Avenue, Toronto, Ontario M5G 1X8, Canada.

in that it allows incremental updates to established partial genome datasets.

## METHODS

### Software and hardware tools

The creation and presentation of partial genomes described here was undertaken on an Intel workstation (Dual Processor Pentium III, 750 MHz) running Red Hat Linux 7.1. PartiGene has also been tested on more recent versions of Red Hat Linux (8.0 and 9.0) and is expected to be fully portable to most UNIX distributions and hardware architectures. PartiGene uses the PERL scripting language, installed as default on most systems: a PERL interpreter of version 5.005 or later is required. In addition to the scripts presented here, PartiGene requires the installation of a number of other publicly available tools (freely available unless otherwise noted): phred, phrap and cross\_match (<http://www.phrap.org>; a license is required for commercial users); DECODER (contact the authors, [rgscerg@gsc.riken.go.jp](mailto:rgscerg@gsc.riken.go.jp); a license is required for commercial users); ESTscan (<http://www.isrec.isb-sib.ch/ftp-server/ESTScan/>); PostgreSQL (<http://www.postgresql.org>); NCBI BLAST (<http://www.ncbi.nlm.nih.gov/BLAST/>); Bioperl (<http://www.bioperl.org>); and EMBOSS (<http://www.hgmp.mrc.ac.uk/Software/EMBOSS/>).

### Overview

We were concerned with producing a software solution that provided ease of use while maintaining best practice for EST analysis. Therefore we have written a pipeline that takes raw sequence trace (chromatogram) data, performs base calling and vector and low quality sequence removal, preparation of dbEST submission files, clustering into putative genes, consensus sequence prediction, peptide prediction and sequence similarity annotation. The analysed data can be viewed as flat files (in HTML format) or as a standard-format SQL database. Throughout, we have implemented 'best practice' based on our experience with generating and analysing EST sequences. PartiGene is divided into three segments that process the raw sequence traces (trace2dbest), generate the partial genomes (PartiGene) and derive peptide predictions (prot4est). The input may be in the form of raw sequence chromatographic trace data, processed sequence data or more simply the name of the target species for which EST data are available in dbEST. The output can include dbEST submission files, HTML tables describing each putative gene object and/or a set of SQL database tables that may be readily queried using the SQL interpreter.

### Process 1: from raw trace data to dbEST submission

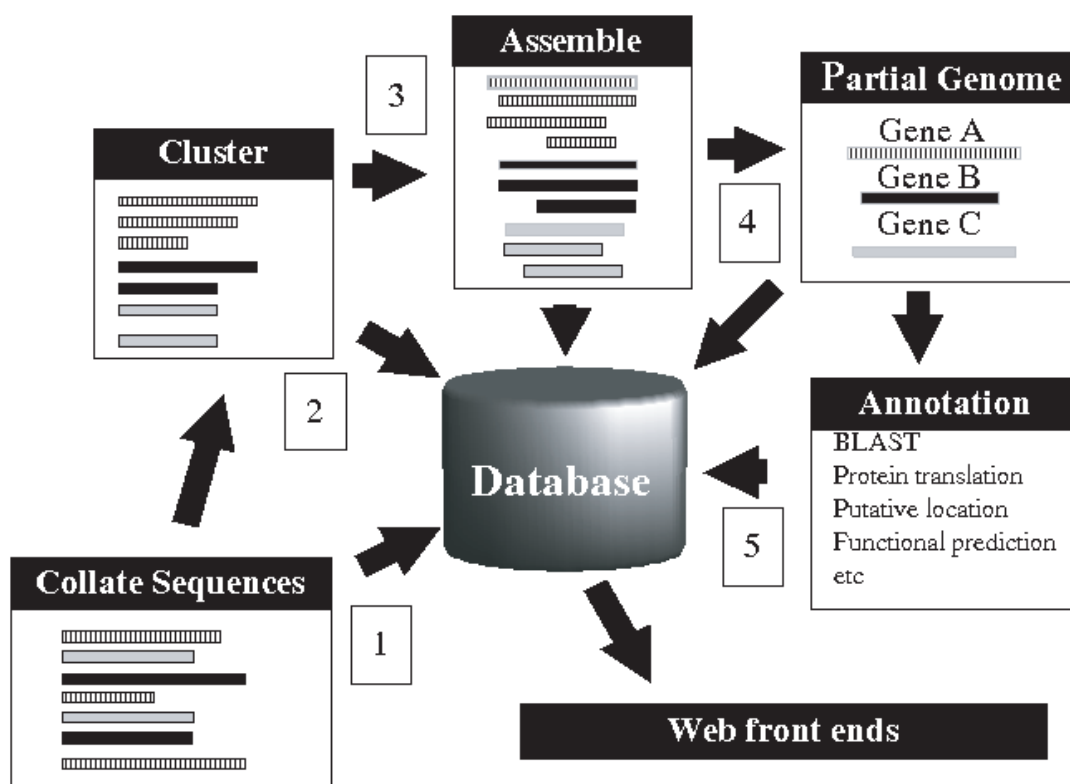
The first script, trace2dbest, is an interactive pipeline script that takes raw sequencer trace data and converts them into

formatted dbEST submission objects. The script first asks the user for cDNA library-specific information, which may be entered interactively or recalled from a previous session. The program offers two levels of vector elimination stringency via cross\_match (P.Green, unpublished data) and also offers the opportunity to add some primary annotation to the dbEST submission in the form of the best similarity match found in a chosen protein database. After specifying a directory containing sequence traces, the process uses phred (Ewing and Green, 1998; Ewing *et al.*, 1998) to perform base calling. Any vector-derived sequence is removed, and user-specified leader/adaptor sequences may also be trimmed. Poly(A) tails are identified and deleted, and sequences that have more than 150 high-quality bases are used to create submission files. At this stage, a BLAST similarity search may be performed against a user-defined database to provide some preliminary annotation ('Similar to xyz...', with appropriate BLAST scores). Finally, the user is given the option to automatically submit the sequences to dbEST.

### Process 2: creating partial genome databases

An overview of the construction of a partial genome as implemented in the PartiGene script is given in Figure 1. The script operates as a series of menu-listed steps. Each step may be interrupted at any time and the process can simply be restarted from where it left off. The first step collates the sequences in fasta format. PartiGene is able to download complete species-specific datasets from dbEST. When databases are updated, downloaded sequences are compared against the existing database and only new sequences are extracted. As not all database sequences will necessarily have been processed through trace2dbest (e.g. ESTs submitted by other research groups), these ESTs are first screened for any possible contaminating vector sequence, poly(A) tails, quality (presence of *N* bases) and size. Non-insert sequences are trimmed, and only those sequences >100 bases in length are used in subsequent processing.

The next step involves clustering the sequences on the basis of sequence similarity into groups that putatively derive from the same gene using our freely available program, CLOBB (Parkinson *et al.*, 2002). CLOBB has an advantage over other clustering solutions in that it readily performs incremental updates of datasets maintaining previous cluster identities. Clusters that contain more than one sequence are then used to derive a consensus sequence (putative gene sequence). This assembly step, based on phrap (P.Green, unpublished data), offers the user the ability to incorporate sequence quality information (produced by the base calling package, phred, in the trace2dbest script). We have used phrap in preference to the alternative cap3 because phrap creates fewer contigs for large clusters and includes the 'single-stranded' regions at the ends of contigs (which are therefore longer). Our



**Fig. 1.** An outline of the whole PartiGene process. (1) Sequences are collated either from local sources or via automatic download from GenBank dbEST. (2) Sequences are clustered on the basis of sequence similarity using CLOBB into groups that putatively derive from the same gene. (3) Clusters containing more than one sequence are assembled into consensus sequences. (4) The partial genome consists of these consensus sequences along with those clusters that contain only one sequence (termed singletons). (5) Putative genes are annotated by performing custom BLAST searches, peptide predictions, etc. and collated in a central database.

analysis of the recent releases of phrap have not identified the issues of base insertion and non-majority base calls identified in earlier publications. PartiGene offers three options in building consensus sequences: (1) use no trace quality data; (2) use quality data (if available) only for clusters containing two sequences; and (3) use quality data (if available) for all clusters. If no traces are available locally, sequences are given a default, modifiable phred score of 15 for each base position. Our preliminary analyses suggest that option 2 yields the fewest, high-quality contigs per cluster in mixed-source datasets.

The collection of clusters that contain only one sequence (termed singletons) and the sequence consensus created in the assembly step above form the partial genome of the selected organism. A first-pass annotation of these putative genes is afforded by customizable BLAST similarity analyses. The user may select up to five different searches against locally available databases. As performing a large number of BLAST searches can be time-consuming, the user may halt the PartiGene process and perform such analyses independently: the search results can be imported.

To view and review these analyses, PartiGene offers two levels of access. For smaller datasets, a series of HTML summary tables can be written that provide information on each cluster including constituent members and summary BLAST output (Fig. 2). It is recommended that this be only undertaken for smaller datasets (less than 1000 clusters). The final step of the PartiGene script involves importing the data into a local database. PostgreSQL is implemented for its availability, functionality, development and support. If a database has not yet been created, PartiGene will automatically generate appropriate tables. Sequence, cluster and annotation data are then automatically imported.

### Process 3: predicting protein translations

In current EST datasets, a small majority of the putative genes will have a BLAST similarity match to a database protein. However, a significant minority (up to 45%) will remain 'novel'. The majority of ESTs derive from protein-encoding genes, but translation of the putative genes identified by the PartiGene script is not trivial. Sequencing errors are common within ESTs and can lead to frameshift errors,

**Results Page**

Page 1   Page 2   Page 3   Page 4   Page 5

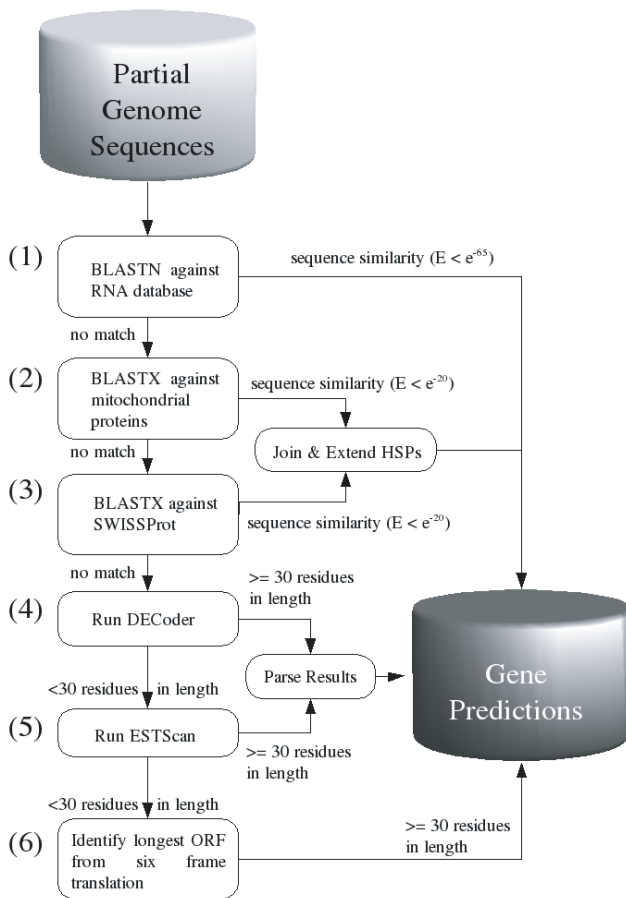
Cluster ID	No. seqs	List of sequences	BLASTX vrx <i>C. elegans</i>	BLASTX vrx SwissProt - nematode proteins
ZPC00001	2	AW773324 AW783743	C42C1.14 CE26911 status:Confirmed TR:Q95X53 protein_id:AAK72292.1 122 2e-29	O42846 60S ribosomal protein L34. Schizosaccharomyces pombe (Fission yeast). 122 7e-28
ZPC00003	11	AW773326 AW773333 AW773341 AW773348 AW773415 AW773500 AW773506 AW783696 AW783744 AW783767 AW783803	T25C8.2 CE16463 locus:act-5 Actins status:Confirmed TR:O45815 protein_id:CAB05817.1 338 9e-94	P53470 Actin 1. Schistosoma mansoni (Blood fluke). 336 8e-92
ZPC00004	2	AW773327 AW783688	ZK20.5 CE06608 locus:rpn-12 vegetative protein X like status:Partially_confirmed SW:Q23449 protein_id:CAA93778.1 204 3e-53	P48556 26S proteasome regulatory subunit S14 (P Homo sapiens (Human)). 131 5e-30
ZPC00007	2	AW773330 AW773461	Y48B6A.2 CE22117 locus:rpl-43 status:Confirmed TR:Q9U2A8 protein_id:CAB54440.1 125 4e-30	Q9VMU4 CG5827 protein (RH41593p) (RE23595p). Drosophila melanogaster (Fruit fly). 129 7e-30
		AW773331	ZK721.2 CE05106 locus:unc-27	Q9VWY3 CG7178 protein. Drosophila

**Fig. 2.** HTML summary table for displaying cluster and associated BLAST annotation for ESTs derived from the nematode *Zeldia punctata*. For each cluster, the number and list of ESTs are provided, along with a brief description of the top hit from a BLAST search to a list of user defined databases (in this instance *Caenorhabditis elegans* proteins and SwissProt, with nematode proteins extracted, were selected). The page features links to individual and cluster consensus sequences and the detailed BLAST output for each cluster.

which may not be corrected by consensus sequence prediction. We have therefore developed prot4est, which combines state of the art programs to produce accurate protein predictions from PartiGene-processed ESTs (Fig. 3). prot4est is a six tier system of prediction. The first three tiers involve the use of BLAST annotation. First, potential RNA genes are identified, tagged and removed from the dataset. Second, all remaining sequences are searched using BLASTX against a protein database of choice [we recommend SwissProt; Boeckmann *et al.* (2003)]. If a sequence is found to share significant sequence similarity (expectation ( $E$ )-value  $< e^{-20}$ ) to a database protein, the frame of translation used to resolve the match is assumed to be the correct frame of translation. The frame of translation for local regions within each of these sequences is determined, and using transeq (part of the EMBOSS package) a robust tiling path determined. A series of rules (see supplementary data on Web site) are then invoked to determine which, if any, potential start codons should be used. Potentially incorrect

stop codons caused through errors in the sequencing process are identified through comparison of the high-scoring pairs (HSPs) and may be ignored. Third, potential mitochondrial proteins are identified, and for these, further processing implements translations using the relevant mitochondrial genetic codes.

Many sequences will not share significant similarity to a database entry. In such cases, *de novo* prediction software must be employed. prot4est combines two of the more successful programs, DECODER (Fukunishi and Hayashizaki, 2001), and ESTScan (Iseli *et al.*, 1999), to obtain accurate peptide predictions. Both require training sets, in the form of annotated complete coding sequences and codon usage tables, to identify coding regions. prot4est will automatically download this information from the relevant Web-based resource. DECODER uses codon bias tables (<http://www.kazusa.or.jp/codon/>) and coding sequences, while ESTScan relies on the availability of protein coding sequences (typically available from EMBL/Genbank, e.g. <http://srs.ebi.ac.uk/>).



**Fig. 3.** How prot4est derives peptide sequence from low-quality EST data. Partial genome sequences derived from the PartiGene process are fed through a six tier system. (1) RNA genes are identified by BLASTN matches to a RNA gene database. (2) and (3) Nuclear and mitochondrially encoded protein-coding genes are identified on the basis of BLASTX similarity to known proteins. The BLAST output is analysed to allow extensions beyond the high scoring pairs. Sequences with significant sequence similarity to a known protein use the frame of translation designated by their common alignment to obtain an accurate peptide prediction. Peptides from sequences with no significant sequence similarity to a known protein are determined *de novo* using either DECODER (4) or ESTScan (5). If neither program predicts a peptide above a tunable length cutoff, six-frame translations of the sequence are identified (6) and the longest open reading frame extracted. Results each stage are collated in a central database.

We have determined that, for both these programs, reduced prior information significantly impacts translation quality (data not shown). If the acquired information is likely to be insufficient for accurate translations (less than 50 coding sequences for DECODER and less than 125 sequences for ESTScan), then the user is warned and the alternative of using data derived from a related species is offered.

In the fourth step, then, sequences are passed through DECODER, which requires the availability of quality files. These are generated as part of the PartiGene process above. For sequence consensus, the phrap derived quality file is used. For singletons the original trace quality file is used, or if this is not available, then as above, sequences are given a default, modifiable score of 15 for each base position. As DECODER only makes a prediction in the forward strand, a reverse complement of each singleton and consensus sequence is created to ensure that all frames are considered. DECODER was originally written for use on complete cDNAs, and it expects a start methionine, which may not always be present in incomplete EST sequences. prot4est therefore appends any peptide sequence upstream of the prediction made by DECODER, provided that no stop codons are encountered. If the peptide is less than 30 amino acids in length, the sequence is passed to the fifth step, ESTScan.

ESTScan builds hidden Markov models based on coding sequence nucleotide patterns to derive peptide sequence. prot4est takes these predictions and again adds upstream and downstream in-frame ORF translations. A 30-residue cutoff is again applied. The sixth step takes the remaining sequences, generates a six-frame translation and identifies the longest open reading frame (ORF). If the length of this ORF is less than 30 residues, the sequence is deemed to be non-coding.

prot4est peptide predictions may be imported into the SQL database created by PartiGene. Further annotation of these protein data, including pI, molecular weight and putative location, may then be generated and imported. Comments on the accuracy of translation for certain regions within the sequence are also passed by prot4est to the database.

### Presentation of the partial genome

Although PartiGene offers the ability to view results in the form of simple HTML tables (Fig. 2), the creation of a local database provides a powerful resource for querying and presenting the data. PostgreSQL is based on the popular SQL syntax and provides an easily accessible interface to perform complex queries on the data. Alternatively, the user may wish to consider the use of Web-based forms to allow remote users, and those less experienced in computing, access to the data (Fig. 4). We have created a number of Web-accessible sites for presentation of our partial genome data including NEMBASE (<http://www.nematodes.org/nematodeESTs/nembase.html>), LophDB (<http://www.nematodes.org/Lopho/LophDB.php>) and LumbriBase (<http://www.earthworms.org>). Each site utilizes the Apache Web server (<http://www.apache.org>) to serve pages created using the PHP Web scripting language, which features a database interpreter (<http://www.php.net>). Examples of these scripts may be obtained from the authors.

(A)

Retrieve a specific cluster   (1)

Enter either its ID (e.g. ASC00088) or the accession number of one of its constituent sequences (e.g. AW165759) or a clone name (e.g. As\_tgz\_51H10).

or use the form below (see documentation) to identify nematode sequence clusters which match your search criteria.

(2)

Submit to perform search

Reset to reset form

Organism

Search against blast output

Ascaris lumbricoides  
Ascaris suum  
Brugia malayi  
Haemonchus contortus  
Necator americanus

Text

Minimum e score

To view all clusters for an organism just leave the text box blank

View results as a list ordered by :  abundance  blast score

[View Detailed Library](#)

(B)

6 Total Sequences

Sequence Types (1) EST : 6

Number of contigs : 1

Library ID	Stage	Number
6252	AD	6

ALC00229

Blast information for cluster ALC00229

blastx vrs nr	blastn vrs nr	blastn vrs dbest	Trace?	
Contig_01_ (NM_072693)	myosin	[Caenorhabditis elegans]	5e-09	
Contig_01_A59287	myosin heavy chain - fluke	(Schistosoma mansoni)	8e-08	
Contig_01_ (AJ306290)	myosin heavy chain	[Toxocara canis]	1e-08	
Contig_01_ (AB015484)	myosin heavy chain	[Dugesia japonica]	1e-08	
Contig_01_ (U40036)	myosin heavy chain	[Mytilus edulis]	5e-08	
Contig_01_ (AJ249991)	myosin heavy chain	[Mytilus galloprovincia]	5e-08	

If you would like one of these clones try emailing Claire

ALC00229 - contig : 1 Length of Contig : 594

6252 - AD : BU585672 / BU585852 / BU586118 / BU586385 / BU587028 / BU587052 /

0 100 200 300 400 500

Contig 1

BU585852 BU585672 BU586118 BU586385 BU587028

Trace? Yes No

**Fig. 4.** Screenshots from NEMBASE showing Web pages created using the php scripting language to submit user queries to the underlying postgresSQL database. (A) Annotation search page. This form may be used to retrieve individual clusters by their unique ID (1) or groups of clusters by keywords associated with their BLAST annotation (2). (B) Detailed cluster page. This page provides information on a single cluster including the number and source of constituent sequences (1), summaries of BLAST annotation (2) and graphical views of the alignment of individual sequences to the cluster consensus (3). For details on the interpretation of this information, please see the NEMBASE help pages.

## ACKNOWLEDGEMENTS

The authors would like to thank the Natural Environmental Research Council Environmental Genomics Thematic Data Centre, especially Dr Dan Swan and Dr Bela Tiwari, for their invaluable comments during the creation of the PartiGene system. This work was supported by the Wellcome Trust and the Natural Environmental Research Council Environmental Genomics Thematic Research Programme. J.W. is supported by the Biotechnology and Biological Sciences Research Council and Astra Zeneca, UK.

## REFERENCES

- Adams,M.D., Kerlavage,A.R., Fleischmann,R.D., Fuldner,R.A., Bult,C.J., Lee,N.H., Kirkness,E.F., Weinstock,K.G., Gocayne,J.D., White,O. *et al.* (1995) Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. *Nature*, **377** (suppl.), 3–174.
- Allen,J.E., Daub,J., Guiliano,D., McDonnell,A., Lizotte-Waniewski,M., Taylor,D.W. and Blaxter,M. (2000) Analysis of genes expressed at the infective larval stage validates utility of *Litomosoides sigmodontis* as a murine model for filarial vaccine development. *Infect. Immun.*, **68**, 5454–5458.
- Blaxter,M., Daub,J., Guiliano,D., Parkinson,J., Whitton,C. and The Filarial Genome Project (2002) The *Brugia malayi* genome project: expressed sequence tags and gene discovery. *Trans. R. Soc. Trop. Med. Hyg.*, **96**, 7–17.
- Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.-C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I., Pilbout,S. and Schneider,M. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.
- Boguski,M.S. and Schuler,G.D. (1995) ESTablishing a human transcript map. *Nat. Genet.*, **10**, 369–371.
- Boguski,M.S., Lowe,T.M. and Tolstoshev,C.M. (1993) dbEST—database for 'expressed sequence tags'. *Nat. Genet.*, **4**, 332–333.
- Christoffels,A., van Gelder,A., Greyling,G., Miller,R., Hide,T. and Hide,W. (2001) STACK: Sequence Tag Alignment and Consensus Knowledgebase. *Nucleic Acids Res.*, **29**, 234–238.
- Daub,J., Loukas,A., Pritchard,D.I. and Blaxter,M. (2000) A survey of genes expressed in adults of the human hookworm, *Necator americanus*. *Parasitology*, **120**, 171–184.
- Ewing,B. and Green,P. (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.*, **8**, 186–194.
- Ewing,B., Hillier,L., Wendl,M.C. and Green,P. (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.*, **8**, 175–185.
- Fukunishi,Y. and Hayashizaki,Y. (2001) Amino-acid translation for cDNA with frame-shift error. *Physiol. Genomics.*, **5**, 81–87.
- Iseli,C., Jongeneel,C.V. and Bucher,P. (1999) ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **7**, 138–158.
- Kenyon,F., Welsh,M., Parkinson,J., Whitton,C., Blaxter,M.L. and Knox,D.P. (2003) Expressed sequence tag survey of gene expression in the scab mite *Psoroptes ovis* allergens, proteinases and free radical scavengers. *Parasitology*, **126**, 451–460.
- Paquola,A.C., Nishiyama,M.Y.Jr, Reis,E.M., da Silva,A.M. and Verjovski-Almeida,S. (2003) ESTWeb: bioinformatics services for EST sequencing projects. *Bioinformatics*, **19**, 1587–1588.
- Parkinson,J., Guiliano,D.B. and Blaxter,M. (2002) Making sense of EST sequences by CLOBBing them. *BMC Bioinformatics*, **3**, 31.
- Parkinson,J., Mitreva,M., Hall,N., Blaxter,M. and McCarter,J. (2003) 400,000 nematode ESTs on the net. *Trends Parasitol.*, **19**, 283–286.
- Pertea,G., Huang,X., Liang,F., Antonescu,V., Sultana,R., Karamycheva,S., Lee,Y., White,J., Cheung,F., Parvizi,B., Tsai,J. and Quackenbush,J. (2003) TIGR gene indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. *Bioinformatics*, **19**, 651–652.
- Sutton,G.G., White,O., Adams,M.D. and Kerlavage,A.R. (1995) TIGR assembler: a new tool for assembling large shotgun sequencing projects. *Gen. Sci. Technol.*, **1**, 9–19.
- White,O. and Kerlavage,A.R. (1996) TDB: new databases for biological discovery. *Methods Enzymol.*, **266**, 27–40.