

PartiGene v2.2 - A tool for generating partial genomes

(last updated 02/11/04)

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

**The Institute of Evolutionary Biology
University of Edinburgh**

**for the Natural Environment Research Council
Environmental Genomics Thematic Programme Data Centre
(EGTDC)**

**Ashworth Laboratories, King's Buildings, Edinburgh, EH9 3JT, UK
p +44 131 650 6760 +44 131 650 6761 f +44 131 650 7489**

nematode.bioinf@ed.ac.uk

<http://www.nematodes.org/>

<http://www.earthworms.org/>

<http://www.tardigrades.org/>

1. Introduction

PartiGene is part of the Edinburgh-EGTDC developed EST-software pipeline consisting of trace2dbest, PartiGene, prot4est and annot8r. PartiGene is a menu-driven, multi-step software which takes sequences (usually ESTs) and creates a non-redundant set of sequence objects (putative genes) which we term a partial genome. The process consists of the following segments :

1. Downloading of sequences from public databases on a species-specific basis
 2. Pre-processing sequences
 3. Clustering on the basis of sequence similarity using our clustering software (CLOBB) into groups of sequences which putatively derive from the same gene
 4. Assembly of the clusters into contigs using the public domain software phrap
 5. Simple annotation on the basis of BLAST to local in-house databases
 6. Creation of HTML summary tables
 7. Creation of a local SQL database allowing for the formulation of more complex queries to facilitate data mining
-

2. Requirements

PartiGene is designed to run on machines using the RedHat Linux operating system like, for example Bio-Linux, but should also easily port to other Linux / Unix distributions. PartiGene requires the presence of third party public domain software including :

Perl - typically installed on most unix workstations - http://www.perl.org/
BioPerl - http://www.bioperl.org/
PostgreSQL - included with most unix distributions - http://www.postgresql.org/
CLOBB2.pl - http://www.nematodes.org/CLOBB
phrap - http://www.phrap.org
cross_match - http://www.phrap.org
Standalone BLAST - http://www.ncbi.nlm.nih.gov/
wget utility - included with most unix distributions - http://www.wget.org/
A web browser - any of opera, netscape, mozilla or galeon

The users PATH environmental variable must also be updated to include these programs.

3. The Configuration file

A configuration file `.partigene.conf` is created in the user's home directory the first time PartiGene is started. In the `.partigene.conf` file several directories are specified according to Bio-Linux standard locations. Depending on your local system setup you may have to edit some of the entries using a text editor. The standard `.partigene.conf` file looks like this :

```
BLASTDB=/home/db/blastdb      ## Location of local BLAST databases
VECTOR=/usr/software/phrap/phrap/vector.seq
                                ## Vector file used for screening
DATABASE=newdb                ## Name of the postgresQL database
SEQREP=/home/user1/project     ## Base Location of directories containing
                                ## phred called sequence/quality files
QUALSCORE=15                  ## phred Quality score to be used when the
                                ## original files can't be found
SEQLINK=<a href=http://srs6.ebi.ac.uk/srs7bin/cgi-bin/wgetz?-e+[embl-
acc:PARTISEQ]>                ## Hyper text link for your sequences
                                ## The word PARTISEQ gets substituted
                                ## for your sequence names
```

To improve accuracy of cluster assembly, PartiGene attempts to find the original phred-called sequence and quality files. If it can't find these files, it will assume a quality score, defined in the configuration file, for each base during the assembly (phrap) phase of the process.

We recommend that you use `trace2dbest` for processing your raw sequence data. This will allow PartiGene to semi-automatically search for the relevant quality files. If you have not used `trace2dbest` we recommend that you set up a unique project directory for each species you are looking at in a common area, PartiGene will then try and find the files in the following directories:

```
SEQREP/project/library_name/
```

where :

`SEQREP` - is defined in the configuration file (`/home/user1/project` in the example above)

`project` - is obtained from user input

`library_name` - is obtained from the sequence name (see "4. Naming schemes" below)

4. Naming Conventions

Ideally PartiGene should be run using sequences obtained from EMBL (or GenBank) with the appropriate accession numbers. PartiGene will then try to find the original sequence and quality files on your system as above by referring to the specified clone name (typically found in the header of fasta format sequence files). To identify the original files from their clone names, we recommend that you adopt the NERC Environmental Genomics style format when initially naming your sequences

The preferred format for trace names is :

`\w\w_\w{2,5}_\d\d\w\d\d`

where `\w` represents any letter or number

`\d` represents any digit

`{2,5}` = minimum/maximum number of characters

The first two letters represent the species from which the sequences were derived, the middle letters represent the cDNA library used to derive the sequences and the last digits-letter-digits represent the plate number and well coordinates (usually in 96 well format).

E.g. for sequences derived from a cDNA library made from material derived from the adult nervous system of the earthworm *Lumbricus rubellus* library you might use :

`Lr_adN_01A01, Lr_adN_01A02 etc.`

5. Running the Program

It is recommended that prior to running PartiGene, you create a project specific directory and run PartiGene from within that directory. We suggest you use the directory scheme created by trace2dbest, for example :

`/home/user1/est_solutions/Lumbricus_rubellus/PartiGene`

PartiGene is started by issuing the command `PartiGene.pl`, after which you will be given a list of seven options. The process of creating a partial genome proceeds in a step-wise fashion through these options. You may quit the program after any of these stages and restart again at any point (this allows for e.g. addition of new BLAST analyses performed outwith the program - see 5.5 below). Note at several stages you will be asked for a species specific cluster identifier - this is a three letter string, typically the first two letters of the genus and species followed by C (for cluster) - e.g. for the earthworm *Lumbricus rubellus*, you might use the cluster identifier LRC. Sequence clusters are then identified by this three letter ID followed by a five digit number. If the identifier has previously been used, then new sequences will be used to update the existing database for that identifier.

5.0 PartiGene test and expert mode

Before using PartiGene we recommend to use the PartiGene test modus to find out whether the setup is OK. The test modus is started by issuing the command 'PartiGene.pl test'. The more experienced user is given the possibility to speed up data analysis in the PartiGene expert mode. The expert mode is entered by launching the command 'PartiGene.pl expert'.

5.1 Downloading sequences

We highly recommend that you use sequences downloaded from EMBL. If you do want to skip this step and use other sequences, simply place them in a directory called "sequences" in your project directory. If you are downloading from EMBL, you simply need to enter the species name, or alternatively the NCBI taxid (see <http://www.ncbi.nlm.nih.gov/Taxonomy>) for the respective species, and all sequences associated with that species will be downloaded and placed in a directory called sequences. The default setting is that just EST sequences will be downloaded. Alternatively there is an option to retrieve plus additional mRNA and DNA sequences. If you want to remove sequences for any reason (e.g. take out ribosomal sequences etc) then you can quit the program after this step and edit the contents of the sequences directory.

5.2 Preprocessing sequences

Some of the sequences submitted to dbEST are of dubious quality, therefore preprocessing of the downloaded sequences is recommended. In this step poly(A) and poly(T) tails are trimmed, vector contamination is eliminated and sequences which are simply too short for meaningful downstream analysis are removed from the data set. If all the sequences in question have been submitted via the trace2dbEST tool this step can be omitted.

5.2 Clustering sequences

You may have previously created a partial genome using sequences from the same species. In this case PartiGene will at first compare downloaded sequences with sequences already in your database, removing those which are already present. PartiGene uses a piece of software called CLOBB2.pl to cluster sequences into a non-redundant set of sequence objects. CLOBB2.pl uses megablast to cluster sequences into putative gene objects and allows for incremental updates. For more information on CLOBB see Parkinson J et al., (2002) BMC Bioinformatics. 3(1):31.

5.3 Sequence Assembly

Once clustered, the sequences need to be assembled into consensus (putative gene) sequences using the program phrap. During this process, you will be asked if you would like to process the sequences using quality information - this is best as it improves the quality of the predicted contigs. Ideally you might have some or all of the original sequence and quality files from phred processing (see section 3. above). Alternatively you can just assign a quality score (specified in the configuration file) to every base position. If you have used trace2dbest to process your trace files, PartiGene tries to find the relevant files automatically, if you have not used trace2dbest you need to point PartiGene to the respective directories as it is described in section 3.

5.4 BLASTing

To provide some simple annotation to the putative gene objects, PartiGene offers the possibility of performing a series of sequence similarity searches using BLAST. Initially PartiGene will prepare the set of gene objects which can be used for BLASTing and place them in a directory called BLAST. PartiGene then queries the directory defined in your configuration file for available databases, you may select up to a maximum of five BLASTs to undertake. These can either be against single databases or you may combine a single BLAST search against several databases. Results from these BLASTs are stored in appropriately named sub-directories of the blast directory. All cluster sequences which are successfully BLASTed against the selected databases are moved to a sub-directory called passed in the directory blast - otherwise they will remain in the blast directory. Note that you can actually skip this step and perform a series of BLASTs outwith of the program - all you need to do to include them in the downstream processing is to place the results in an appropriately named sub-directory of the directory blast. For example you may wish to perform a BLASTN against the non-redundant nucleotide database "nr" and a BLASTX against the "swiss-prot" protein database. As this can take considerable computing time you may wish to perform these on external resources - simply perform the BLASTs making sure that you specify the output format to plain text, not html, and move the results to sub-directories named "nr" and "swprot" (or any other name you associate with the respective BLAST-results).

5.5 Creation of HTML tables

For viewing smaller datasets (typically < 1000) sequences, you can create a series of HTML tables which will list each cluster, the constituent sequences of the cluster and some simple annotation obtained from any BLASTs you may have undertaken. You also have the option to view the results if you have an appropriate browser.

5.6 Databasing

For larger datasets it is advisable to build a local database as this greatly facilitates access to the data (using database queries to select clusters of interest rather than having to scroll through many pages of hypertext). So long as you have a local instance of postgresQL - a freely available public domain databasing solution - running and configured correctly, PartiGene allows you to build a simple database from your results. PartiGene will construct the following tables :

cluster	contains information on each cluster (number of sequences, consensus (putative gene) sequences)
est	contains information of each sequence used to derive the clusters (which cluster/contig it is associated with, alignment and quality information with respect to the consensus sequence)
est_seq	contains the actual sequence for each sequence used in the clustering process
blast	contains information from BLASTs performed on the consensus sequences (type of BLAST, database, description of hit etc)
clone_name	look-up table for clone names
p4e_ind	the p4e tables contain peptide predictions derived from consensus sequences using the prot4EST tool
p4e_loc	Please note you need to run prot4EST before using these tables
p4e_hsp	For more on prot4EST see http:// www.nematodes.org/PartiGene

BLAST information for each cluster is obtained for each set of BLASTs found in the blast directory - so in theory you could perform step 5.4 several times and have a larger number of different BLAST results stored in your database.

Please note that for databasing the postgresql back end must be running. It is typically started on Red Hat Linux machines by logging on as root and issuing the command :

```
/etc/init.d/postgresql start
```

The user must also have permissions to create databases, this can be done by logging on as the postgres user (su postgres) and issuing the command :

```
createuser <username>
```

5.7 Beyond PartiGene

PartiGene is part of a pipeline to process and annotate EST datasets. For further annotation translation of the consensus sequences into peptides is required. We recommend that you use prot4EST for high quality translation. The predicted peptides then can be used as input for the annot8r suite of annotation tools. Furthermore, we are currently developing wwwPartiGene, a web interface which will allow remote access to the database you have created. If you are interested in a test version please contact us via the email address below. There is no support for users to query the database so far- we recommend that you try one of the online web tutorials on SQL programming (which are a lot easier than they sound). For constructing web pages similar to e.g. NEMBASE (see <http://www.nematodes.org/>) we recommend that you start with wwwPartiGene and use the web scripting language php for additional features.

If you would like more information on PartiGene or any issues raised here, please contact nematode.bioinf@ed.ac.uk and have a look at our web page <http://www.nematodes.org/PartiGene> for the latest developments. If you are working in an EG-Awardee lab and have questions or problems please contact the helpdesk helpdesk@envgen.nox.ac.uk.