trace2dbest

version 3.0
User Guide

turning sequence chromatograph traces into expressed sequence tags

Alasdair Anthony, Ralf Schmid, John Parkinson, Mark Blaxter

Institute of Evolutionary Biology
School of Biological Sciences
University of Edinburgh
Ashworth Laboratories, King's Buildings, Edinburgh, EH9 3JT, UK

nematode.bioinf@ed.ac.uk
http://www.nematodes.org/
http://envgen.ceh.ox.ac.uk/

developed with funding from
the Natural Environment Research Council
Environmental Genomics Thematic Programme

Contents

1 What is trace2dbest and what does it do?

1.1 What is trace2dbest and what does it do?

trace2dbest is a computer program that takes sequencing chromatograph trace files from expressed sequence tag projects, as produced by fluorescent sequencing machines (such as ABI Prism or Amersham MegaBace) and processes them into quality-checked sequences, ready for submission to the public repository for expressed sequence tags, dbEST. trace2dbest will also help you create Publication, Library and Contact files, needed in addition to EST files for dbEST submissions. If you are not sure what an expressed sequence tag or EST is, please see the dbEST introductory pages at:

http://www.ncbi.nlm.nih.gov/dbEST/index.html

1.2 Why use trace2dbest?

trace2dbest simplifies the process of getting your sequences from the sequencer to the dbEST database. With only a few sequences, its possible to do this by hand, relying on manual editing, and individually-tailored responses to possible errors and other issues. When processing a lot of sequences it is easier to let a computer do most of the work. The high-throughput genome sequencing centres have developed software tools that can be adapted for use in a low- or medium-throughput setting. We have bundled some of these tools together into one program, called trace2dbest. We hope that using trace2dbest will be easy, and that it makes the process of generating and using ESTs exciting and rewarding.

1.3 What does trace2dbest do with my sequence traces?

trace2dbest uses the base-calling program phred [1] to get a raw sequence from your trace files. phred assigns a quality score to each of the bases it calls, based on the strength of the signal, the shape of the peak and the local environment of the peak. trace2dbest then takes this raw sequence through several stages of trimming, the end result being a good quality EST sequence.

trace2dbest uses the program cross_match to identify and trim vector sequence and, optionally, *E. coli* sequence. trace2dbest will also trim adapter, poly(A) tail and low quality bases from the sequence. All these trimming stages have parameters that can be adjusted by the user.

After sequence trimming the user has the option to preliminary annotate the sequences based on BLAST [2] searches. Finally, trace2dbest will create a dbEST EST file for it, based on information provided by the user. Once all the sequences have been processed, you have the option of mailing the completed submission file directly to dbEST. The files will also be saved locally.

1.4 Why do I have to use a controlled sequence naming scheme?

While trace2dbest is useful in isolation, it is also designed to be used in an integrated set of programs (called PartiGene [3]) which take EST sequence traces through a series of informatic analyses to produce a partial genome database of annotated sequences. A consistent naming scheme for all the sequences is required so that the programs can perform the analyses properly. This also allows the software to extract information from the file name rather than having to be told by the user what to do. For example, trace2dbest will extract the plate number and plate coordinates from each file name and insert this information into the dbEST EST file.

1.5 What naming schemes can I use?

trace2dbest accepts two naming schemes, the NERC environmental genomics (EG) scheme and the full STRESSGENES scheme. A naming scheme is essentially a series of tags separated by the underscore ' _ ' character. A trace file using the NERC EG naming scheme looks like:

```
Lr_adE_02A05
```

and one using the full STRESSGENES naming scheme would look like:

```
CcLL03b01a02f2_AbaRb
```

The basic principles of the both schemes are very similar. The NERC scheme has three parts separated by underscore characters. We describe the NERC EG scheme here (for the STRESSGENES scheme see http://legr.liv.ac.uk/). The first tag must be two characters and is used to indicate the species (or other major project identifier). The second tag, which may be from 3 to 5 letters long, indicates the library from which the clone sequenced was derived. The third tag indicates the address of the clone in terms of micro-titre plate number and row/column. Thus, in the example above, 'Lr' indicates the species (*Lumbricus rubellus*), 'adE' the library (adult Edinburgh) and '02A05' the 96-well or 384-well plate coordinates (plate 02, row A, column 05). Three digits identifiers for plates (such as plate 002) can also be used.

1.6 How do I rename my sequence files to fit the scheme?

The program rename_file.pl comes as part of the trace2dbest package and helps you to rename trace files according to the naming schemes. rename_file.pl replaces one text string that you supply with another one and can also transform serially numbered files into files numbered as if from a 96 well plate (so that 001 becomes A01 and 096 becomes H12). If you run rename_file.pl you will get the following options list:

```
Usage : rename_file.pl <list of arguments>
-dir <txt> - set directory of traces <dir> to <txt>
-add <txt> - <txt> gets added to the beginning of each tracefile
         in directory <dir>
-end <txt> - <txt> gets added to the end of each tracefile
-txt <txt1> - <txt1> gets removed from each file
-sub <txt2> - (only with -txt set) txt1 is replaced by txt2
-format - Traces are reformatted to correct 96 well nomenclature.
         Single digits are replaced by double digits and row ID set
         to uppercase. In addition, if your files do not contain
         plate coordinates, but are numbered sequentially e.g.
         trace1, trace2, trace3 etc. this option will convert the
         numbers into 96 well format (it assumes 1-12 refer to row
         A columns 1-12 etc.)
-help - Get more detailed help
```

Thus, to change a set of filenames in a directory (such as 'my_traces') from an incorrect format (such as Apisadultw03F03) to a correct one (Am_AW1_03F03), you would type:

```
rename_file.pl -dir my_traces -txt ApisadultW -sub Am_AW1_
```

For more information type 'rename_file.pl -help'.

1.7 What are 'Library', 'Contact', 'Publication' and 'EST' files?

The public EST repository, dbEST, simplifies data deposition by splitting the information across a set of four types of linked files (for details of the EST submission process see http://www.ncbi.nlm.nih.gov/dbEST/how_to_submit.html). Each individual EST has an 'EST' file: this file holds the sequence and basic information such as the name of the library, the name of any publication existing or planned describing the dataset, and the name of a contact person who can be contacted for more details. Rather than repeating all the information on these three topics in each and every EST file, dbEST holds the data in linked files, called, 'Lib' for library information, 'Pub' for publication information and 'Cont' for contact information. To get a set of ESTs into dbEST one has to submit these three files along with the sequences. The simplifying feature of this is that once your 'Cont' file is in dbEST, any subsequent EST submissions you make (next week, month, year) need only to refer to this file to access the same contact information. The same is true of the 'Pub' and 'Lib' files.

2 SETTING UP trace2dbest

2.1 What do I need to run trace2dbest?

trace2dbest is a pipeline program processing each input file through a series of steps. Some of these steps are built-in to trace2dbest, while others rely on using external programs. In addition to the program itself (trace2dbest.pl) you will need :

(1) a set of sequence chromatographs. These need to follow a consistent naming scheme (see above). We provide a renaming script, rename_file.pl, that can help you to adjust sets of file names. For example, many sequencing services add a lane or capillary number to each trace name; this can simply be removed.

(2) a computer running a UNIX/Linux-based operating system (for example Bio-Linux as provided by the EGTDC Oxford). Almost all UNIX/Linux operating systems come with the programming language Perl pre-installed. trace2dbest is written in Perl, it requires Perl version 5.4 or later.

(3) the sequence chromatographic trace base-calling program phred [1], and the vector sequence matching software cross_match. phred, cross_match and a third program phrap come as a package available under a free academic licence from the programs' author, Phil Green, at the University of Washington. Please go to http://www.phrap.org/ to get a copy. The software is emailed to you after filling in an agreement form (http://www.phrap.org/consed/academic_agreement.txt). Simply follow the instructions provided by the authors to install these programs.

(4) (optionally) the sequence similarity search suite BLAST [2]. BLAST is the standard sequence similarity search program and is available from the NCBI via ftp://ftp.ncbi.nih.gov/blast/executables/ . Simply follow the instructions for download and installation given in this directory.

(5) (optionally) local sequence databases for BLAST. One section of the trace2dbest process allows you to perform a basic, BLAST-based annotation of the sequences before they are submitted. If you want to do this, you should make sure you have the required sequence databases available to you. For most purposes we recommend the Uniprot (Uniprot/Swiss-Prot or both Uniprot/Swiss-Prot and Uniprot/TrEMBL) database from http://www.ebi.ac.uk/uniprot/database/download.html. Once you have the database on your local computer you the need to run the formatdb command of the BLAST suite to get it ready for searching.

2.2 Installing trace2dbest

If you are using Bio-Linux as your operating system, trace2dbest will already be installed and is ready to use. Otherwise install the executable trace2dbest.pl into a directory where you store your local executables. Typically these directories are named either `/usr/software/bin` or `/usr/local/bin`. If you want to run trace2dbest without typing the full path to the program, you need to make sure the installation directory is in your path. Finally, if there is more than one person using trace2dbest on the install machine, we recommend that you set up a directory named `/home/db/est_solutions` and ensure that all users have the appropriate permissions to write to this directory. The reasons for this are given in section 4 of this guide.

3 USING trace2dbest

3.0 Preparation

trace2dbest requires to enter information interactively. To make this process easier trace2dbest will make use of the perl Term::ReadLine::Gnu module, if it is found on your system. This module makes features such as filename completion and command history available. If you don't have this module we recommend that you download and install it from http://www.cpan.org/ (see Appendix 1).

Before starting trace2dbest you should first ensure that all the traces you wish to process are in a single directory that contains no other files. You should also check that all the trace files match one of the naming schemes described in section 1.5. As part of its quality checking system, trace2dbest will identify and remove any files in your trace directory that do not meet the specified naming scheme.

We recommend that you run trace2dbest from an empty directory, as this is where trace2dbest will initially write its output files. It is important that you do not try to run trace2dbest from your trace directory. If trace2dbest has been installed as described above you can start the program by typing `trace2dbest.pl`. This will take you to the main menu which offers you seven options. In the following few pages we will guide you through each of these options.

We suggest that you run these options in sequential order. You can exit the program between each of these options and later pick up where you have left off. There is no need to run option 1 'Setup configuration' and option 2 'Checks and tests' every time you are using trace2dbest. As soon as you have made sure that the program has been installed and configured properly (options 1 and 2) you can start processing your traces using option 3 'Process traces' . A screen shot from the trace2dbest welcome screen listing the seven options can be found on the next page.

```
#####################################################################
###                                                               ###
###                  trace2dbest -  Vs 3.0                        ###
###          a trace file processing and dbEST                    ###
###               sequence submission tool                        ###
###                                                               ###
###   A. Anthony, R. Schmid and colleagues for BANG 2005          ###
###                                                               ###
###   News, upgrades and help:   nematode.bioinf@ed.ac.uk         ###
###                                                               ###
#####################################################################

  Enter the number corresponding to the part of the trace2dbEST
  process you want to perform:
       1. Setup configuration.
       2. Checks and tests.
       3. Process traces.
       4. Blasts for preliminary annotation.
       5. Create or view submission information.
       6. Prepare dbEST submission files.
       7. Exit.
```

3.1 **Section 1 Setup configuration**

This option allows you to create or update the trace2dbest configuration file '.trace2dbest.conf'. The configuration file is saved in the user's home directory and stores information specific to the local setup of trace2dbest. '.trace2dbest.conf' is a simple text file and can also be edited with a standard text editor. The information stored includes:

- the location of the 'vector.seq' and 'ecoli.seq' files. These files are required for sequence trimming in section 3.
- the location of 'Libfile.db', 'Pubfile.db', 'Contfile.db' and 'ESTfile.db'. In these files library, contact information etc is stored to be re-used in later runs of the program. These files can be accessed and updated in section 5.
- phred parameters. This is an expert user option and, since we are not encouraging users to modify it, it is not accessible via the trace2dbest interface. However, phred parameters can be changed by modifying the configuration file using a text editor.

3.2 **Section 2 Tests and Checks**

This option allows you to check whether the relevant files as specified in '.trace2dbest.conf' can be found by trace2dbest and whether programs trace2dbest is relying on are available and configured correctly. If all tests and checks are successful you are ready to start processing traces.

First, trace2dbest checks for the availability of the programs phred, crossmatch, blastall and blastcl3. While the first two are essential the latter are only required in section 4 and therefore optional. Common errors are that these programs are either not installed or they are installed, but the installation directory is not in the user's path. Adding the installation directory to the $PATH variable should fix this (see Appendix 2).

Second, trace2dbest looks whether the PHRED_PARAMETER_FILE environmental variable is set and whether the file does exist. If this test fails, see phred installation instructions or appendix for details.

Finally, trace2dbest checks whether the files defined in the configuration file '.trace2dbest.conf' are present. If any of these tests fails and the implemented auto correction mechanism also fails, please re-run section 1 and/or make sure you have installed 'vector.seq' and 'ecoli.seq' in the location specified in '.trace2dbest.conf'.

3.3 **Section 3 Process traces**

In this section, trace2dbest will run phred to base call the traces, then run cross_match to identify vector and, optionally, to identify *E. coli* contaminations. You can choose between two options in this section: 'Standard' and 'Advanced'. If an error occurs in this section, the error message will be reported to the screen. You should then check the logfiles to identify the cause of the problem.

The 'Standard' option assumes standard parameters for all settings. All you are asked for is the directory which contains your tracefiles. After entering this trace2dbest will process your traces using standard parameters for vector and *E. coli* screening. You will be updated about the progress made, and finally presented with a summary of the results.

The 'Advanced' option will allow you to modify standard parameters to increase or decrease the level of stringency for the trimming steps. First you need to indicate which naming scheme you have used, details are given in section 1.5. Then you are asked for the full path to your traces. trace2dbest will check that every file in the specified directory matches the selected naming scheme. You will be notified of any files that do not match the naming scheme, at which point you may exit trace2dbest and edit the trace file directory or you may continue and trace2dbest will remove (delete) these files.

Next you are asked to enter an adapter sequence. You may use a regular expression to represent the adapter sequence if you wish. trace2dbest will scan the raw sequence for the adapter sequence you have entered. If it is found the adapter sequence and everything before it (upstream) will be trimmed off. If you do not wish to trim adapter sequence, just press return.

trace2dbest will then offer to display the entries present in the vector.seq file. This file is used by cross_match to scan the raw sequence for vector. If your vector sequence is not included in the vector.seq file then simply add this sequence (in FASTA format) to the file. Next, you are asked if you would like to trim stray *E. coli* sequence from your ESTs. trace2dbest uses cross_match with stringent parameters to trim *E. coli* sequence, however EST sequence that is very similar to part of the *E. coli* genome may inadvertently be trimmed off. *E. coli* screening will add to the sequence processing time.

In the next section you can set the various parameters that control how the traces will be processed. trace2dbest has default values for all these parameters. The defaults are shown in brackets (). To select the default value for any of the parameters, just hit return. For more details on the cross_match parameters, see the cross_match documentation.

When defining the number of bases in a poly(A) tail, you should enter a number between 1 and 99 (inclusive). trace2dbest will scan all but the first 150 bases of the sequence for continuous stretches of As equal to or longer than the length you specify. trace2dbest will also scan the reverse sequence for stretches of Ts in a similar way. If found, the poly(A) tail and all sequence after it will be trimmed and this event recorded in the POLYA field of the EST file. If poly(T) tails are found then all the sequence before it is trimmed.

You are also given the option of trimming 'spliced leader' sequences from the EST sequences. The *C. elegans* spliced leader 1 sequence is preloaded and to use this enter yes. To use another spliced leader sequence just enter its sequence. As soon as trace2dbest has received all relevant information the traces will be processed. Details of the processing for all sequences can be found in the 'process' directory.

3.4 **Section 4 BLASTs for preliminary annotation**

In this section you have the option of adding BLAST annotation to your sequences, subject to a bit score cut-off. This annotation is generated by taking the description, score and e-value of the top BLASTx hit. If you wish to add such annotation, you have two options, remote BLAST via NCBI (limited to ~400 BLAST searches / day) or local BLAST.

If you choose to carry out a local BLAST then you will be asked to enter the location of your BLAST databases. trace2dbest will then present you with a list of all the protein BLAST databases in this directory. You should select one database by entering the appropriate number.

3.5 Section 5 – Create or view submission information

trace2dbest uses a set of files (Libfile.db, Contfile.db, Pubfile.db and ESTfile.db) to store information which is required in the submission process to dbEST. In this option you can view the entries present in each of these files, or add new entries to these files. If you choose to add a new entry you will be asked to enter some details regarding the sequences you are processing. Most of the information needed is self explanatory. For further details see the dbEST introductory pages at:

http://www.ncbi.nlm.nih.gov/dbEST/index.html

Whenever a new entry is created you will be given the opportunity to save it for future use.

As an example, we will guide you through the creation of a new entry for the ESTfile.db. When asked for the sequencing primer, you should enter the **name** of the primer, followed, if desired, by the sequence in round brackets ( ). The primer name you give here will be entered in the SEQ_PRIMER field of the EST file. It will also be appended to the trace file name and entered in the EST# field of the EST file. The information for the forward and reverse PCR primers should be entered if you have it; otherwise hit enter and the field will be left blank. When requested, the date you would like your data to be made public should be entered in the form MM/DD/YYYY. For immediate release, just hit enter (the corresponding field in the EST file will be left blank). Please note that dbEST policy is to have a maximum hold on data of 6 months. When you have entered the information for the file, trace2dbest will format this according to dbEST standards and display the file on the screen. At this point you should check the file to ensure it is correct. To exit from the viewing program, type 'q'.

3.6 Section 6 – Prepare dbEST submission files

To create the final submission files trace2dbest uses information you have entered in section 5 or information from files already submitted to dbEST. If you choose the latter option then trace2dbest will ask you for a small amount of information relating to the file so that it can fill in the relevant parts of the EST file. (Please note therefore that the information you provide must match exactly that in the submitted file). trace2dbest gives you the option to view the merged submission files that it has created.

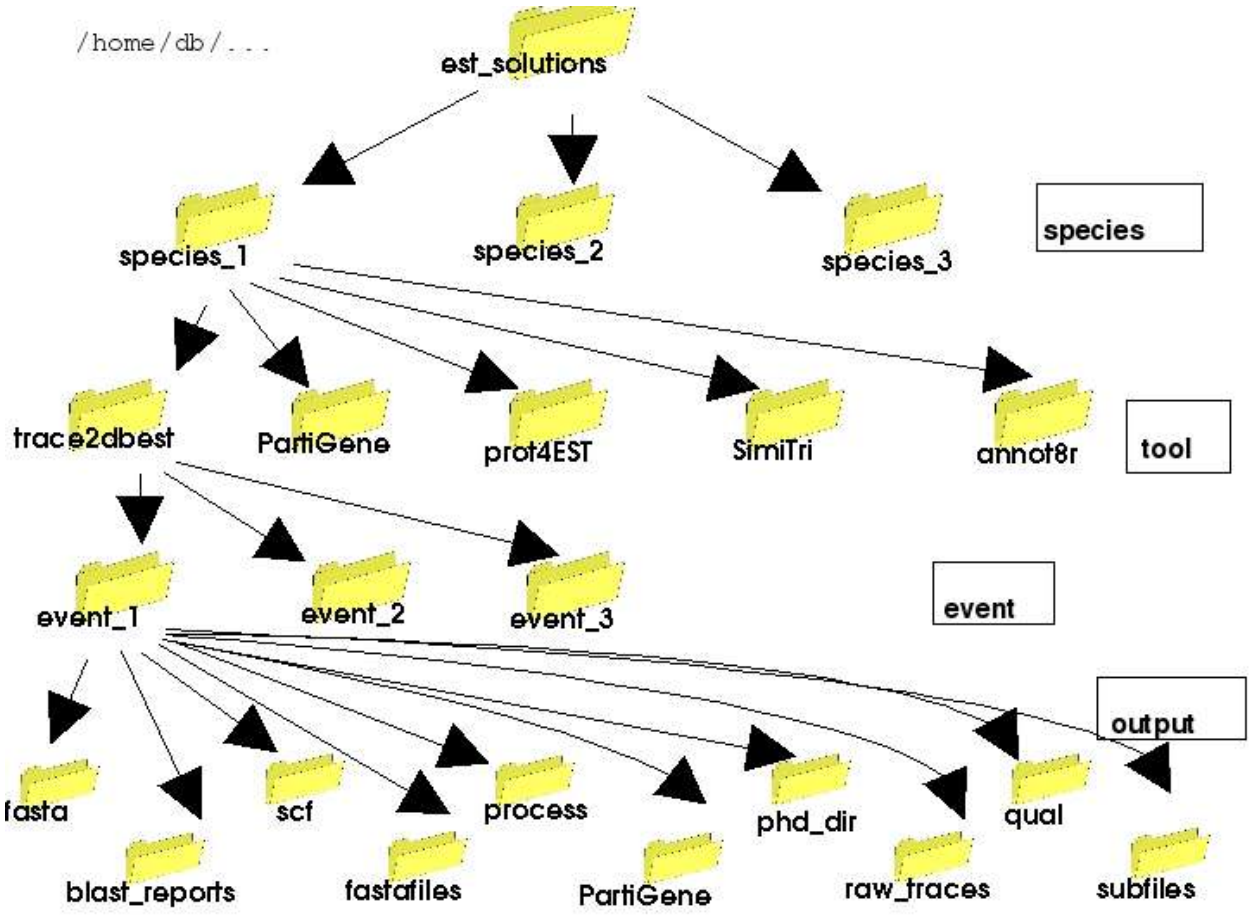
You will be given the option of emailing the submission file directly to dbEST from trace2dbest. To use this facility you will need to provide the name of your SMTP server for outgoing mail (your local computer support will be able to tell you this). Whether you choose to submit your files or not, they will automatically be saved. When saving your files trace2dbest will first try to save in the directory `/home/db/est_solutions`, if this does not exist, trace2dbest will save the files to your home directory, in a directory called est_solutions. trace2dbest will inform you of the exact location to which your files and all the other trace2dbest outputs have been saved (see next page for details).

4 trace2dbest output and where it is saved

4.1 Where are the files saved?

trace2dbest will try to save its output to the directory /home/db/est_solutions. If this is not possible, trace2dbest will save its outputs to the directory est_solutions in your home directory. We recommend that the directory /home/db/est_solutions is set up for two main

reasons. First, it means that all trace2dbest sessions run on a particular machine will be stored in one location, even if the program is run by different users. Secondly, it allows the other programs in the EGTDC EST pipeline to easily access the data produced by trace2dbest. The directory structure that will be set up within est_solutions is shown here:



To ensure that trace2dbest saves its output in the common area, you only need to create the est_solutions directory in /home/db and ensure that all users have read/write permissions to this directory. trace2dbest will create the species, tool, event and output subdirectories.

4.2 **trace2dbest output**

In the directory where trace2dbest has saved your files, you will find a comprehensive output consisting of ten directories and two files. These are described here:

dbEST_submission.txt contains the merged dbEST submission files. This file should be emailed to dbEST.

logfile contains progress information from various parts of the trace2dbest process. You generally shouldn't need to look in this file. If trace2dbest fails unexpectedly it may provide useful info.

blast_reports contains two files, blasts_full (containing the full BLAST reports for each sequence) and blasts_tophit (containing just the top hits), if you have chosen to do BLAST

searches.
fasta has two types of file for each sequence: .seq - raw (unprocessed) FASTA format sequence, produced by phred and .seqsc - the seq files with vector screened out and *E.coli* replaced by Zs and Xs respectively.
fastafiles contains .fsa files for each sequence processed. These are the processed sequences in FASTA format. These files may be used as the input sequence files in the 'sequences' directory of PartiGene, the next stage in the EST processing pipeline.
partigene This directory contains .seq and .qual files for use by PartiGene. These files are accessed automatically by PartiGene. You don't need to move or copy them.
phd_dir has the .phd files produced by phred. These files contain the base quality scores for the sequence.
process contains information on how trace2dbest has trimmed the sequences. The sub-directory traceinfo contains a file for each sequence which details the trimming performed on that sequence. Depending on what trimming has taken place, there may also be files such as polya_trim, vector_trim, quality_trim and adapter_trim, which detail the particular types of trimming. There is also a file giving summary statistics.
qual contains .qual files that contain a matrix of phred quality scores.
raw_traces has a copy of the trace files used for the this trace2dbest session.
scf contains the standard chromatogram format files produced by phred.
subfiles contains an individual EST submission file and processed sequence file for each trace processed.

Appendix 0: *ChangeLog from version 2*

- A major change in architecture: rearrangement of the script into separate menus, allows user to run parts of the script without the need to give all information required for submission of EST-sequences
- trace2dbest has now a configuration file: this replaces some hard-coded parameters and gives more flexibility for using private db files
- new test and setup option
- using phred *-altrim* flag instead of *-trim. -altrim* is not quite as stringent as *-trim*, but *-trim* often looses bits of high quality sequence
- a few additions to the rename_file.pl script

Appendix 1: *Installing Term::ReadLine::Gnu from CPAN*

- Go to the CPAN search webpage (http://search.cpan.org)
- Search for Term::ReadLine::Gnu
- Follow the link "Term-ReadLine-Gnu-1.15" (please note: the version might be a more recent one by the time you are installing it)
- Download the latest version to your computer
- Unzip and untar the downloaded file using the command `gunzip Term-ReadLine-Gnu-1.15.tar.gz` followed by `tar xvf Term-ReadLine-Gnu-1.15.tar`
- Change your working directory to the `Term-ReadLine-Gnu-1.15` directory and follow the instructions given in the `README` file of this directory

Appendix 2: *Adding directories to your PATH*

This short paragraph is not intended to replace any introduction into Linux or UNIX where you will find much more detailed and complete information, but it will try to assist the unexperienced Linux/UNIX user in overcoming issues related to the setting of the user's PATH. Linux/UNIX needs to be told in which directories the operating system should look for executables. The collection of these directories is defined in the variable

PATH and there are plenty of ways to modify your PATH. To make matters even more confusing there are two main families of shells, c-type shells and bash shells, which offer somewhat different ways to manage your PATH. To find out what shell you are using type `echo $SHELL` .

In the following we will describe one way to modify your PATH for each of the two major shell families. If you are using a bash shell and have installed the executable `phred` in the directory `/usr/software/bin` you would add this directory to the line which is defining your PATH in your `~/.bashrc` file (see example below). You can edit your `~/.bashrc` file using any standard text editor.

```
export PATH=$PATH:/usr/software/exec:/usr/software/bin;
```

If you are using a c-type shell (csh and tcsh being the most common ones) you have to edit the ~/.cshrc file instead. Add the directory `/usr/software/bin` to the line

```
setenv PATH "/usr/software/exec:/usr/software/bin"
```

After modifying and saving the respective file you need to make the changes effective by running `source ~/.bashrc` or `source ~./cshrc` , respectively.

Appendix 3a: ***Common phred problems - phredpar.dat not found***

phred requires the parameter file phredpar.dat. The location of this file needs to be defined as an environmental variable. For a bash shell (see appendix 2) you would add the line

```
export PHRED_PARAMETER_FILE=/usr/software/phred/phredpar.dat;
```

to your ~./bashrc file (obviously the entry needs to point to the real location of the file).
For a c-like shell (see appendix 2) you need to add the line

```
setenv PHRED_PARAMETER_FILE "/usr/software/phred/phredpar.dat"
```

to your ~./cshrc file. Again you have to run `source ~/.bashrc` or `source ~./cshrc` , respectively, to make the changes effective.

Appendix 3b: ***Common phred problems - The 'unknown chemistry' problem***

Every time ABI releases a new “chemistry” you will encounter the "unknown chemistry" problem. When phred is reading a trace file it extracts information about chemistry (primer or terminator), about dye type (eg. rhodamine or big-dye) and about sequencing machines (eg. ABI377 or ABI3700). phred then compares this information with entries in the 'phredpar.dat' file. If phred doesn't find a corresponding entry in 'phredpar.dat' it will exit with the 'unknown chemistry' error message. To fix this, simply add a new entry to your 'phredpar.dat' file following the instructions given at the end of this file. An example for entries we have added recently to our phredpar.dat file can be found below.

```
# additional lines of machine definitionss
"DT {BD Set Any-Primer}.2 copy"  terminator big-dye  ABI_373_377
"DT3730POP7{BDv3}.mob"           terminator big-dye  ABI_3700
```

References:

[1a] Ewing B, Green P: Genome Research 8: 186-194 (1998).
[1b] Ewing B, Hillier L, Wendl M, Green P: Genome Research 8:175-185 (1998).
[2] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: J. Mol. Biol. 215: 403-10 (1990).
[3] Parkinson J, Anthony A , Wasmuth J, Schmid R, Hedley BA, Blaxter M: Bioinformatics 20: 1398–1404 (2004).