File: PHRED.DOC

PHRED Documentation

1. Introduction.

Phred reads DNA sequencer trace data, calls bases, assigns quality values to the bases, and writes the base calls and quality values to output files. Phred can read trace data from chromatogram files in the SCF, ABI, and ESD formats. It automatically determines the file format, and whether the chromatogram file was compressed using gzip, bzip2, or UNIX compress. After calling bases, phred writes the sequences to files in either FASTA format, the format suitable for XBAP, PHD format, or the SCF format. Quality values for the bases are written to FASTA format files or PHD files, which can be used by the phrap sequence assembly program in order to increase the accuracy of the assembled sequence.

I have tested phred base calling and quality value accuracies for

data from the following sequencing machines.

  ABI models 373, 377, and 3700
  Molecular Dynamics MegaBACE
  LI-COR 4000

I have tested the phred base calling accuracy only for data from
the following sequencing machines.

  ABI model 3100
  Beckman CEQ

Significant differences in this release

- quality value lookup table for ABI 3700 dye terminator chemistry data (phred still uses the quality value lookup table for ABI 373/377 dye primer data when it processes ABI 3700 dye primer chromatograms. I have insufficient dye primer data for this calibration)

- quality value lookup table for MegaBACE dye terminator data processed with the Cimarron version 3.0012 base caller (phred still uses the quality value lookup tables for Cimarron version 1.53 processed dye primer data when it processes Cimarron version 3.0012 dye primer data. I have insufficient dye primer data for this calibration)

- phred reads the trace processor software version from the ABI, ESD, and SCF format chromatograms in order to select the best base processing parameter and quality value lookup table for the Cimarron processed MegaBACE data. Note that only phred writes the trace processing software version string in the SCF file comments, identifying it with the new 'label' TPSW.

- sequencing machine and chemistry specific basecalling parameters improve base calling accuracy.

  This change makes the phred base calling depend on correct identification of the chromatogram 'source', which means that phred must match the primer ID string in the chromatogram with a string in the (included) 'phredpar.dat file' in order for it to process the chromatogram correctly.

  The '-exit_nomatch' option forces phred to exit immediately if it cannot match the chromatogram primer ID string with a 'phredpar.dat' entry.

  The '-process_nomatch' option allows phred to process a chromatogram with a non-matching primer ID string if the 'phredpar.dat' file contains an entry for "__no_matching_string__".

  phred's 'error' messages reflect these changes. This document includes a summary of the general program flow and the

consequent messages phred writes to 'stderr', which is your terminal normally, and the log file when you use the '-log' option.

The 'peak prediction' has increased 'resolution' for detecting peak spacing changes (which causes phred to run about 2 times slower).

- trace 'noise' value calculated and stored in the phd file header. Phred calculates the ratio of the total uncalled-base peak area to the total called-base peak area within the high quality base segment of the read. It stores this value in the phd file header with the label 'TRACE_PEAK_AREA_RATIO'. The high quality base segment of the read is determine using the modified Mott trimming algorithm described below.

- phred checks the SCF file private data block for the Beckman CEQ 'fingerprint'

- phred base calling is tuned for Beckman CEQ; however, the quality value lookup table is not designed for the Beckman CEQ because I have insufficient data for quality value calibration. (Phred uses the ABI 3700 quality value lookup tables for the Beckman CEQ data.)

- processes ABI and SCF format files with no bases stored in them.

- phred exits if the PHRED_PARAMETER_FILE environment variable is not set or it cannot read the 'phredpar.dat' file successfully.

- earlier phred versions interpreted a chromatogram primer ID string consisting entirely of non-printing characters as an empty string. Now phred does no interpretation.

- when phred runs with the '-id <chromat_dir>' option and the <chromat_dir> contains subdirectories, phred no longer tries to process the subdirectories, so it will not warn of an 'unknown file type' for the subdirectories.

- more sensitive compression motif detection for ABI 3700 dye primer data

- the '-v 1' option causes phred to write the command line and the time phred starts running to both stdout and stderr.

2. Acknowledgements.

Phred benefits from ideas developed by LaDeana Hillier, Mike Wendl, Dave Ficenec, Tim Gleeson, Alan Blanchard, and Richard Mott.

3. Algorithms.

Phred uses simple Fourier methods to examine the four base traces in

the region surrounding each point in the data set in order to predict a series of evenly spaced predicted locations. That is, it determines where the peaks would be centered if there were no compressions, dropouts, or other factors shifting the peaks from their "true" locations.

Next phred examines each trace to find the centers of the actual, or observed, peaks and the areas of these peaks relative to their neighbors. The peaks are detected independently along each of the four traces so many peaks overlap. A dynamic programming algorithm is used to match the observed peaks detected in the second step with the predicted peak locations found in the first step.

Phred evaluates the trace surrounding each called base using four or five quality value parameters to quantify the trace quality. It uses a quality value lookup table to assign the corresponding quality value. The quality value is related to the base call error probability by the formula

```
QV = - 10 * log_10( P_e )
```

where P_e is the probability that the base call is an error.

Phred uses data from a chemistry parameter file called 'phredpar.dat' in order to identify dye primer data. For dye primer data, phred identifies loop/stem sequence motifs that tend to result in CC and GG merged peak compressions. It reduces the quality values of potential merged peaks and splits those peaks that have certain trace characteristics indicative of merged CC and GG peaks. In addition, the chemistry and dye information are passed to phrap.

4. Building and installing.

The INSTALL file describes the steps for building and installing phred.

Copy the phred parameter file, called 'phredpar.dat', to a directory that is accessible by phred users and set the environment variable 'PHRED_PARAMETER_FILE' to the full path name of the file. For example, if you copy 'phredpar.dat' to '/usr/local/etc/PhredPar' and you are using the C shell then issue the command

```
% setenv PHRED_PARAMETER_FILE /usr/local/etc/PhredPar/phredpar.dat
```

It is most convenient to set the environment variable in the system-wide shell startup (cshrc or equivalent) file.

You can rename the phred parameter file but the PHRED_PARAMETER_FILE environment variable must reflect the new name.

With Windows NT you give the command

```
% set PHRED_PARAMETER_FILE=\usr\local\etc\PhredPar\phredpar.dat
```

in the DOS command window in which you will run phred.

Note: if you compile phred on a SUN Solaris OS using the BSD C compiler in the directory '/usr/ucb', you will find that the '-id' command line option fails (phred reports that it cannot read files, and it prints the name of each file it fails to read; however, the name it prints lacks the first few characters of the true name of the file). If this occurs, recompile phred using either the optional C compiler in the directory /opt/SUNWspro/bin or the GNU C compiler.

5. Running phred.

Phred uses command line options to control input, processing, and output. The command line options are delimited by a dash, "-".

The command line options are

Input Options

-id <directory name> Read and process files in <directory name>.

-if <file name> Read and process files listed in the file <file name>. Each line in <file name> must specify a valid path to a single input file.

-zd <directory name> Location of compression program. If -zd is omitted, phred uses the current path to search for the compression program.

-zt <directory name> Directory where chromat is uncompressed. If -zd is omitted, phred uses /usr/tmp. When phred processes a compressed file, it uncompresses the chromat into this temporary directory before it reads the file. It subsequently deletes the uncompressed file in the temporary directory.

Processing Options

-nocall Disable phred base calling and set the current sequence to the ABI base calls that are read from the input file. By default, the current sequence is set to the phred base calls. This affects the base trimming and output options.

-trim_alt <enzyme sequence> Perform sequence trimming on the current sequence. Bases are trimmed from the start and end of the sequence on the basis of

	trace quality. Specifically, for each base, the phred error probability is subtracted from the default value of 0.05 (or the value set using the '-trim_cutoff' option), and the resulting values are summed to find the maximum scoring subsequence. Furthermore, the subsequence must have a minimum number of bases. In addition, <enzyme sequence> specifies a short base sequence (typically the recognition sequence of the restriction enzyme sequence used for subcloning) that is used to trim bases off the start of the current sequence. You can specify a NULL enzyme sequence using empty double quotes, "". (We recommend that you use '-trim_alt' rather than '-trim' option described below because we believe that '-trim' trims off too many good bases).
-trim_cutoff <value>	Set trimming error probability for the '-trim_alt' option and the trimming points written in the phd files. The default value is 0.05.
-trim_fasta	Trim sequences written to sequence and quality value FASTA files. Set trimming information in the FASTA headers to reflect the high quality of the sequence, and append the string 'trimmed' to the header.
-trim_scf	Trim sequence, quality values, and base locations written to SCF file. Append the string 'trimmed' to the comments.
-trim_phd	Trim sequence, quality values, and base locations written to PHD files. Also set the first and last high quality base locations specified in the 'TRIM' comment field to the numbers of the first and last bases of the trimmed sequence (the first base in the sequence is base number zero). Finally set the error probability cutoff value in the 'TRIM' comment field to -1.00 to indicate that the sequence is trimmed, and that the trim points may be unrelated to the error probability cutoff value.
-trim_out	Trim information in the FASTA, SCF, and PHD output files. This is equivalent to specifying '-trim_fasta', '-trim_scf', and '-trim_phd' on the command line.
-trim <enzyme sequence>	Perform sequence trimming on the current sequence. Bases are trimmed from the start and end of the sequence on the basis of

	trace quality. In addition, <enzyme sequence> specifies a short base sequence (typically the recognition sequence of the restriction enzyme sequence used for subcloning) that is used to trim bases off the start of the current sequence. You can specify a NULL enzyme sequence using empty double quotes, "". We recommend against using this option because we consider it to be too conservative. See the note below on the effect of using the trim option.
-nonorm	Disable phred trace normalization. This option is not recommended unless the base caller fails due to huge noise peaks extending over a large region at the start of the trace, as is characteristic of some dye terminator reactions.
-nosplit	Disable compressed peak splitting. By default, phred identifies and splits C and G peaks that may be a merged pair of peaks. Phred searches for compression prone loop/stem sequence motifs and attempts to confirm a compression using characteristics of the trace, primarily the size of the candidate peak.
-nocmpqv	Force phred to use the four parameter quality values. By default, phred uses five parameter quality values for dye primer data (only) in order to reduce the quality values of merged CC and GG peaks. (Phred uses the four parameter quality values for dye terminator chemistry data automatically. If phred cannot determine the chemistry, it uses the four parameter quality values.)
-ceilqv <ceil_qv>	Specifies a maximum quality value assigned to bases. Bases with quality value parameters that correspond to quality values greater than <ceil_qv> are assigned the value <ceil_qv>.
-beg_pred <trace_point>	Specifies the trace point at which to begin the peak prediction. This point should be in a region of 'good' trace where the peak spacing is even and representative of the peak spacing throughout the trace. In addition the peaks should be large and the noise low in the region, and the value of <trace_point> must not be within 100 points of the trace ends.
-exit_nomatch	When unable to match a chromatogram primer ID string with a 'phredpar.dat' file entry, exit

	immediately.
-process_nomatch	When unable to match a chromatogram primer ID string to a 'phredpar.dat' file entry, use the "__no_matching_string__" entry in the 'phredpar.dat' file to identify the chromatogram chemistry/dye/machine type. If you use this option and the 'phredpar.dat' file lacks the "__no_matching_string__" entry, phred exits immediately when it encounters a chromatogram with an unmatchable primer ID string. Use this option only when processing chromatograms from one type of sequencing machine running one type of sequencing chemistry. See the section describing the Phred parameter file below for more information before using this option.

Output Options

-st fasta	Set the output sequence file format to FASTA. (Default.) Trimming options affect the FASTA file; see the Notes below for more information.
-st xbap	Set the output sequence file format to XBAP.
-s	Write sequence output files with the names obtained by appending ".seq" to the names of the input files, and store them in the directory where phred is running.
-s <file name>	Write a sequence output file with the name <file name>. This option is valid for a single input file only.
-sd <directory name>	Write sequence output files with the names obtained by appending ".seq" to the names of the input files, and write them in the directory <directory name>.
-sa <file name>	Write a sequence output file in FASTA format with the name <file name>. The file contains the base calls of all the reads processed in this run of phred.
-qt fasta	Set the output quality file format to FASTA. (Default.) Trimming options affect the FASTA file; see the Notes below for more information.

-qt xbap	Set the output quality file format to XBAP. Trimmed off base quality values are omitted.
-qt mix	Set the output quality file format to FASTA. Base quality values for all bases are written (including those for trimmed off bases).
-q	Write quality output files with the names obtained by appending ".qual" to the names of the input files, and store them in the directory where phred is running. This option is valid for FASTA format output files only.
-q <file name>	Write a quality output file with the name <file name>. This option is valid for a single input file and a FASTA format output file only.
-qd <directory name>	Write quality output files with the names obtained by appending ".qual" to the names of the input files, and store them in the directory <directory name>.
-qa <file name>	Write a quality output file in FASTA format with the name <file name>. The file contains the quality values of all the reads processed in this run of phred.
-qr <file name>	Write a histogram of the number of high quality bases per read. This is meaningful when phred processes more than one read.
-c	Write SCF files with the trace data, the base calls of the current sequences, and the positions of the base calls. The SCF files have the names of the input files (phred will refuse to write the SCF file if you ask it to write the SCF file in the directory in which the input file resides).
-c <file name>	Write an SCF file with the trace data, the base calls of the current sequence, and the positions of the base calls. The SCF file has the name <file name>. This option is valid for a single input file only.
-cd <directory name>	Write SCF files with the trace data, the base calls of the current sequences,

	and the positions of the base calls. The SCF files are written in the directory <directory name> and have the same names as the input files.
-cp <number of bytes>	Store SCF trace data as 1 or 2 byte values. Defaults to 1 when the maximum trace value is less than 256, or to 2 when the maximum trace value is greater than or equal to 256. This is the trace precision.
-cv <version number>	Write SCF output file in SCF format version 2 or SCF format version 3. The default is version 2.
-cs	Always scale traces before writing them to an SCF output file. This ensures that the largest trace value has the largest value that can be stored in the SCF file. When the file trace precision is '1', the maximum value is 255, and when the precision is 2, the maximum value is 65535. Without this option, phred does not scale the trace unless (a) the trace was read from an ESD file or (b) the maximum trace value exceeds the value that can be stored in the SCF file at the precision used. Trace scaling ensures the maximum digital resolution for a given storage precision but it will make a uniformly low level trace appear to be a high level.
-p	Write a PHD file, which is used by the consed editor to display bases. A PHD file contains a set of comments used by consed for maintaining consistency between the chromat file, the .ace file and the PHD file, and it contains base data as triples consisting of the base call, quality, and position. Phred always writes the first version of the PHD file for a read, which has the name <filename>.phd.1. When a read is edited using consed, a new version of the phd is written by consed, for example, the second version has the name <filename>.phd.2. With the -p option, <filename> is the name of the input file.
-p <filename>	Write a PHD file with the name <filename>.phd.1. This option is valid for processing a single input file.
-pd <directory name>	Write PHD files in directory <directory name>. The PHD files have the names <filename>.phd.1

where <filename> is the name of the input file.

-d	Write a data file that is used for detecting polymorphic bases. The file has the name <filename>.poly where <filename> is the name of the input file. The first line of the file consists of the sequence name, the smallest amplitude normalization factor, and the amplitude normalization factors for the A, C, G, and T traces. One line for each called base follows the header line. The information on each line consists of the called base, the position of the called base, the area of the called peak, the relative area of the called peak, the uncalled base, the position of the uncalled base, the area of the uncalled base, the relative area of the uncalled base, and the amplitudes of the four traces at the position of the called base.
-dd <dirname>	Write polymorphism data files in directory <directory name>. The files have the names <filename>.poly where <filename> is the name of the input file.
-raw <sequence name>	Write <sequence name> in the header of the sequence output file and the quality output file. By default, the name of the input file is written in the headers of these files. This option is valid for a single input file only.
-log	Make phred append a log entry describing the processing run in the file "phred.log".

Miscellaneous

-v <n>	Verbose operation. You can control the level of verbosity with <n>, which ranges from 1 to 63. The value '1' cause phred to write the command line and the time it starts running to both the stdout and stderr.
-tags	Label common output with tags in order to facilitate output parsing.
-h, -help	Display a command line option summary.
-doc	Display phred documentation.
-V	Display phred version.

Examples

If you plan to use phred base calls and base quality information as input to the phrap assembly program and to the consed finishing program, we encourage you to use the phredPhrap Perl script that is part of the consed distribution. Please follow the documentation supplied with consed and then type:

phredPhrap

(with no arguments)

If you intend to use consed, you *MUST* use this perl script. Failure to use this script will result in many consed features not working correctly, including consed's autofinish function, user-defined consensus tags, tagging ALU and other repeats, and tagging vector sequence. Use the phredPhrap perl script.

An outline of the important processing steps performed by the script follows.

Let us say you want to call bases from the chromat files in subdirectory "chromat_dir", use phrap to assemble the contigs, and run consed to edit/examine the contigs. In this case you must ask phred to create "phd" output files, which are required by consed.

It runs phred with the options

```
% phred -id chromat_dir -pd phd_dir
```

which causes phred to read the chromat files in "chromat_dir" and write the "phd" files to "phd_dir". Next it makes FASTA files from the "phd" files by running the phd2fasta program.
For example,

```
% phd2fasta -id phd_dir -os seqs_fasta -oq seqs_fasta.screen.qual
```

Subsequently it screens out the vector in the sequences in "seqs_fasta" using cross_match:

```
% cross_match seqs_fasta vector.seq -minmatch 12 -minscore 20 -screen >
screen.out
```

which generates the screened sequence file "seqs_fasta.screen",

It runs phrap to perform the sequence assembly as follows:

```
% phrap seqs_fasta.screen -new_ace > phrap.out
```

Phrap writes the the assembled contigs to the file "seqs_fasta.screen.contigs", and creates a .ace file that can be used for importing the assembly to xbap, consed, or ace-mbly for

editing.

As another example, again you want to process the chromat files in subdirectory "chromat_dir", but now you want phred to write the base calls to a FASTA file named "seqs_fasta" and the base quality values to "seqs_fasta.qual". In this case you run phred with the options

```
% phred -id chromat_dir -sa seqs_fasta -qa seqs_fasta.qual
```

We recommend that you not use the trim option. Inaccurate bases called near the ends of the traces will not interfere with proper phrap assembly.

Refer to the file "phrap.doc", which is part of the phrap distribution, for information on cross_match and phrap.

Return values

Phred returns 0 for successful processing and for non-fatal errors. It returns -1 for 'fatal errors'. 'Fatal errors' include memory allocation failure and file write (usually due to no disk space) failure.

6. Phred parameter file

Phred reads the 'primer ID' string in the chromatogram and tries to find the same name in the phred parameter file, which is mentioned in the 'Building and installing' section above. If it succeeds, the 'phredpar.dat' entry for the 'primer ID' identifies the sequencing reaction chemistry (primer or terminator), the dye type, and the sequencing machine type.

If phred cannot read the 'phredpar.dat' file, it exits immediately. The reasons that phred may not read the 'phredpar.dat' file include

- the PHRED_PARAMETER_FILE environment variable is unset
- the PHRED_PARAMETER_FILE environment variable is not set to a valid 'phredpar.dat' file

If phred cannot match the primer ID string to a 'phredpar.dat' entry, its operation depends on the command line options '-exit_nomatch' and '-process_nomatch'. The possible results are

- neither '-exit_nomatch' nor '-process_nomatch' is used

  phred skips to the next chromatogram without writing to an output file

- '-exit_nomatch' is used

phred exits immediately when it finds a chromatogram
with an unmatchable primer ID string
reports

- '-process_nomatch' is used

  phred looks for a "__no_matching_string__" entry in 'phredpar.dat'. If it finds this entry, it uses the entry to process the chromatogram. That is, the "__no_matching_string__" entry becomes the default machine/chemistry/dye type. The "__no_matching_string__" entry is commented out in the included 'phredpar.dat' file so, if you want to use the '-process_nomatch' option, you must remove the comment character (#) at the start of this line, and change the chemistry, dye, and machine types to the correct values. Use this option only if you use phred to process chromatograms from one type of sequencing machine running one type of sequencing chemistry. If phred cannot find the "__no_matching_string__" entry in the 'phedpar.dat' file, it exits immediately.

Additionally, when phred cannot find the 'primer ID' name in the 'phredpar.dat'file, it provides the information

```
unknown chemistry (xxxx) in chromat yyyy
add a line of the form
"xxxx"    <chemistry>      <dye type>      <machine type>
to the file zzzz
type 'phred -doc' for more information
```

where xxxx is the 'primer ID', yyyy is the chromatogram name, and zzzz is the 'phredpar.dat' file. In order to add the correct entry to 'phredpar.dat', you will need to know the sequencing chemistry type (primer or terminator), the dye name, and the type of sequencing machine. 'Cut' the entry template phred provides, 'paste' it into the 'phredpar'dat file, and add the correct chemistry, dye, and sequencing machine values in the indicate fields. You will find additional information about the acceptable form of entries in the header of the 'phredpar.dat' file.

The fields in the 'phredpar.dat' file are

field	value name
1	primer identification string
2	chemistry
3	dye
4	sequencing machine type

where the field values are separated by spaces or horizontal tabs.

The values phred recognizes are

value name	values

| ---------- | ------ |
primer ID string	primer name enclosed in double quotes
chemistry	primer, terminator, unknown
dye	rhodamine, d-rhodamine, big-dye, energy-transfer, bodipy, unknown
sequencing machine type	ABI_373_377, ABI_3100, ABI_3700, Beckman_CEQ_2000, LI-COR_4000, and MolDyn_MegaBACE

NOTES:

- phred treats the 'unknown' chemistry type the same as the 'terminator' chemistry type for base calling and quality value assignment; and it sets the chemistry type in the phd file header to 'unknown' (the chemistry type information in the phd file header is written in the FASTA sequence headers by the phd2fasta program in order to pass the information to phrap).

- phred does not use the dye type information for base calling or quality values but it writes the information in the phd file header (the dye type information in the phd file header is written in the FASTA sequence headers by the phd2fasta program in order to pass the information to phrap).

- SCF files created by the Beckman CEQ sequencer have no primer ID string but they have a special identifier in the private data block. Phred checks the SCF private data block for the Beckman identifier when the primer ID string is empty. If it finds the identifier, it sets the primer ID string to "BeckmanCEQ", and subsequently looks in 'phredpar.dat' for the corresponding entry.

- the 'MegaBACE Mobility File' entry in the phredpar.dat file specifies 'unknown' chemistry, rather than 'primer' or 'terminator' because some early MegaBACE software wrote 'MegaBACE Mobility File' for the primer ID string in both primer and terminator chemistry ABD files. You may want to change this value if you process exclusively primer or terminator chemistry MegaBACE data.

- phred considers a missing primer ID string to be an empty string so it will match it to the empty string entry in 'phredpar.dat', if the entry exists.

7. Sequence Trimming

First, a warning: in general, do not trim sequences that phrap will assemble. We introduced trimming capabilities in phred to allow identification of the high quality region of reads, and to permit trimming off low quality segments of reads that are not destined for a phrap assembly. We emphatically recommend against trimming reads for shotgun (or similar) sequencing projects. (Trimming may make

sense for single pass sequencing when the quality values will
be unavailable for subsequent analyses.)

Second, another warning: if you must trim sequences, we strongly
recommend that you use the '-trim_alt' option rather than the
'-trim' option because we believe that it generally preserves
more high quality bases, and it allows you to fine tune the
trimming using the '-trim_cutoff' option.

Phred uses two different algorithms to calculate trimming values.
The algorithm used and its effect depend on the trimming command line
options and the output file type.

The phd output file always contains trimming information in the
header. Phred calculates this trimming information using a modified
Mott algorithm (it does not trim off vector sequence so the trimming
information identifies the entire high quality segment of the read,
including high quality vector sequence). The trimming information
appears in the phd file header in the form

TRIM: <n1> <n2> <r1>

where <n1> is the first high quality base (where the first base in
the sequence is number zero) and <n2> is the last high quality base.
<r1> is the error probability cutoff value used to calculate the
trim points. The command line option '-trim_cutoff' affects the
phd file trimming information by setting the error probability cutoff
value used to calculate the base scores. If the sequence has fewer
than 20 high quality bases, the values <n1> and <n2> are set to -1.
If the '-trim_phd' or '-trim_out' option is used, <n1> and <n2>
are set to the numbers of the first and last bases in the trimmed
sequence (so <n1> is always zero), and <r1> is set to -1.00 to
indicate that the sequence is trimmed and that the error probability
cutoff value may be unrelated to the trim points.

The sequence, quality value, SCF, and PHD output files can be
affected by the trimming-related command line options. (Sequence
and quality value files are those created using the -s, -sa, -sd,
-q, -qa, and -qd options, SCF files are created using the -c and
-cd options, and PHD files are created using the -p and -pd
options). When phred runs without trimming-related options set,
it does not calculate trimming values for the sequence, quality
value, and SCF output files (and it does not 'trim' the values
stored in them).

The '-trim_alt' and '-trim' options select the trimming algorithm
used to calculate the trimming information used in the sequence,
quality value, and SCF output files. The algorithm used for the
'-trim_alt' option is based on the modified Mott algorithm: it uses
the base error probabilities calculated from the phred quality
values and the error probability cutoff (the cutoff can be adjusted
using the -trim_cutoff option). The algorithm used for the '-trim'
option is based directly on characteristics of the trace. It predates
phred and phred quality values. We believe that the '-trim' option
tends to be conservative, 'trimming off' more bases, in comparison to

the '-trim_alt' option. So we recommend using the '-trim_alt' algorithm. Both the '-trim_alt' and '-trim' options take an argument consisting of a restriction enzyme recognition sequence. If the argument is "" (null), phred finds the high quality segment of the read. If the argument is not null, and phred finds the beginning of the recognition sequence within the first 100 bases of the read, phred sets the left trim point to remove the sequence up to this point as well as low quality bases. Please note that the sequence must match the recognition sequence nearly exactly for phred to recognize it. Caution: this is not a substitute for vector masking. We recommend that you use cross_match to mask vector sequence in the reads. (The phredPhrap script automatically calls cross_match to mask vector in the reads.)

Selecting either '-trim_alt' or '-trim' causes phred to determine trimming information and to modify the sequence, quality value, and SCF files as follows.

The FASTA sequence header contains trimming information but the sequence is unaffected. The header has the form

```
>chromat_name   1323     15    548  ABI
```

where the sequence name immediately follows the header delimiter, which is ">", the first integer is the number of bases called by phred, the second integer is the number of bases 'trimmed off' the beginning of the sequence, the third integer is the number of bases 'remaining following trimming', and the string describes the type of input file.

The XBAP-type of sequence header contains trimming information, and the low quality bases are commented out.

For quality value file type option '-qt fasta' (default), the FASTA quality value header contains the same trimming information as in the FASTA sequence header and the quality values of the 'trimmed off' bases are set to zero.

For quality value file type option '-qt xbap', phred writes a XBAP-type of sequence header with trimming information followed by the quality values of the bases remaining after trimming on subsequent lines.

For quality value file type option '-qt mix', phred writes a FASTA quality value header with the same trimming information as in the FASTA sequence header followed by the quality values of all bases (without trimming).

The SCF file contains trimming information in the header, and the sequence, quality values, and trace locations of the called peaks are unaffected. The left clip is the number of bases to trim off the left end of the sequence and the right clip is the number of bases to trim off

the right end.

When the '-trim_fasta' or '-trim_out' option is used with the '-trim_alt' or '-trim' (and -s, -sa, -sd, -q, -qa, or -qd) option, phred writes the trimmed sequence to the sequence FASTA file and trimmed quality values to the quality value FASTA file; that is, it writes only the high quality bases and the corresponding quality values. In addition, it appends the string 'trimmed' to the FASTA headers and the trimming information in the header indicates that no (additional) bases are to be trimmed off. The option '-trim_fasta' is invalid with the '-qt xbap' and '-qt mix' options.

When the '-trim_scf' or '-trim_out' option is used with the '-trim_alt' or '-trim' (and -c or -cd) option, phred writes the trimmed sequence, trimmed quality value, and trimmed called peak locations to the SCF output file. In addition, it appends the string 'trimmed' to the comment field and the left and right clip values are set to zero.

When the '-trim_phd' or '-trim_out' option is used with the '-trim_alt' or '-trim' (and -p or -pd) option, phred writes the trimmed sequence, trimmed quality value, and trimmed called peak locations to the PHD output file. In addition, when it writes the 'TRIM' field in the comment block (at the beginning of the file), it sets the values for the first and last high quality bases to the numbers of the first and last bases of the trimmed sequence (where the first base is number zero), and it sets the error probability cutoff value to -1.00. Setting the cutoff value to -1.00 indicates that the sequence is trimmed, and that the trim points may be unrelated to the error probability cutoff value.

The modified Mott trimming algorithm, which is used to calculate the trimming information for the '-trim_alt' option and the phd files, uses base error probabilities calculated from the phred quality values. For each base it subtracts the base error probability from an error probability cutoff value (0.05 by default, and changed using the '-trim_cutoff' option) to form the base score. Then it finds the highest scoring segment of the sequence where the segment score is the sum of the segment base scores (the score can have non-negative values only). The algorithm requires a minimum segment length, which is set to 20 bases.

8. Trace noise calculation

phred calculates a value related to the amount of 'noise' in the trace, and stores this value in the phd file header. The value is the ratio of the total uncalled-base peak area to the total called-base peak area within the high quality segment of the read. If the high quality region consists of fewer than 20 bases or the area of the called peaks is 0, the value is set to 100. The value appears in the phd file header with the label 'TRACE_PEAK_AREA_RATIO'. This value may be useful for identifying low quality traces due to low signal levels and due to template mixtures.

9. Phred program flow and messages

The following is a overview of the phred program flow and the most important associated messages. phred produces the shown messages when the '-tags' option is used, which I recommend for parsing the phred output with another program/script. phred produces similar messages when run without the '-tags' option.

a. read PHRED_PARAMETER_FILE environment variable

succeeds)

MESSAGE: none
RESULT: phred continues

fails)

MESSAGE: FATAL_ERROR: PHRED_PARAMETER_FILE environment variable not set
RESULT: phred exits immediately

b. read 'phredpar.dat' file

succeeds)

MESSAGE: none
RESULT: phred continues

fails)

MESSAGE: FATAL_ERROR: unable to read parameter file
RESULT: phred exits immediately

c. memory allocation for chromatogram reading

succeeds)

MESSAGE: none
RESULT: phred continues

fails)

MESSAGE: FATAL_ERROR: <chromat_name>: error while reading
RESULT: phred exits immediately

d. chromat file type identification

succeeds)

MESSAGE: none
RESULT: phred continues

fails)

MESSAGES: FILE_ERROR: <file_name>: file read error: unknown file type
RESULT: phred skips to next chromatogram (does not write phd file)

e. chromat reading

succeeds)

MESSAGE: PROCESS: <chromat_name>
RESULT: phred continues

fails)

MESSAGE: FILE_ERROR: <chromat_name>: file read error: <error_description>
RESULT: phred continues processing chromatogram but does not call bases **

f. data checking (does chromat contain a nonzero trace?)

succeeds)

MESSAGE: none
RESULT: phred continues

fails)

MESSAGES: FILE_ERROR: <chromat_name>: trace data missing

OR

FILE_ERROR: <chromat_name>: flat trace data
RESULT: phred continues but does not call bases **

g. chromatogram identification (matching chromatogram primer ID string with an entry in phredpar.dat)
Note: phred ignores match failures when '-nocall' option is used

succeeds)

MESSAGE: none
RESULT: phred continues and calls bases

fails)

o default: use neither '-process_nomatch' nor '-exit_nomatch' command line option

MESSAGE: FILE_SKIP_NOMATCH: <chromat_name>: unable to match primer ID string
RESULT: phred skips to next chromatogram without writing a phd file (or any other output file entry)

OR

- use '-exit_nomatch' command line option

MESSAGE: FATAL_ERROR: <chromat_name>: unable to match primer ID string
RESULT: phred exits immediately

OR

- use '-process_nomatch' option and '__no_matching_string__' entry exists in 'phredpar.dat' file

MESSAGE: FILE_PROCESS_NOMATCH: <chromat_name>: using '__no_matching_string__'
RESULT: phred continues, using the chemistry, dye type, and machine type given in the '__no_matching_string__' entry

OR

- use '-process_nomatch' option and no '__no_matching_string__' entry exists in 'phredpar.dat' file

MESSAGE: FATAL_ERROR: <chromat_name>: unable to match primer ID string
RESULT: phred exits immediately

** the resulting phd file has no bases

10. Phd files

Phred writes 'phd' files to store base calling information, including the sequence, quality values, and peak locations, when it is run with either the '-pd' or the '-p' options. Phred creates phd files with the name '<chromat_name>.phd.1' where the '1' at the end of the name is the version number of the phd file for that chromatogram. It always writes version '1' phd files, whereas 'consed' writes phd files with higher version numbers (it increments the version number each time it saves an edited read).

The phd files phred creates begin with the line

BEGIN_SEQUENCE <sequence_name>

and end with the line

```
  END_SEQUENCE
```

Enclosed between these lines phred writes a header data block, which is enclosed between lines with the labels 'BEGIN_COMMENT' and 'END_COMMENT', and a read data block, which is enclosed between lines with the labels 'BEGIN_DNA' and 'END_DNA'. Thus the overall file structure is (the lines are indented here)

```
  BEGIN_SEQUENCE <sequence_name>

  BEGIN_COMMENT

    [comment block]

  END_COMMENT

  BEGIN_DNA

    [read data block]

  END_DNA

  END_SEQUENCE
```

The header data consists of a number of lines where each line begins with a label followed by a colon and one or more values. Currently, the phd header has the following information

header entry	description
CHROMAT_FILE: <string>	chromatogram file name
ABI_THUMBPRINT: <n>	an integer assigned by the ABI software
PHRED_VERSION: <string>	phred version used to create the file
CALL_METHOD: <string>	<string>="phred" unless run with '-nocall'
QUALITY_LEVELS: <n>	maximum quality value permitted
TIME: <string>	the time and date the file was created
TRACE_ARRAY_MIN_INDEX: <n>	the index for the first trace point (always 0)
TRACE_ARRAY_MAX_INDEX: <n>	the index for the last trace point (npoints-1)
TRIM: <n1> <n2> <r>	read trim points. See (a) below.
TRACE_PEAK_AREA_RATIO: <r>	trace noise level. See (b) below.
CHEM: <string>	chromatogram sequencing chemistry type
DYE: <string>	chromatogram sequencing dye type

(a) the 'TRIM' values consist of the first and last bases in the high quality read segment (where the first base of the read is zero) and the error probability used to calculate the trim points. The modified Mott algorithm is used to calculate the the trim points.

(b) the 'TRACE_PEAK_AREA_RATIO' is the ratio of the total uncalled-base peak area to the total called-base peak area within the high quality segment of the read. Thus this value indicates the level

of the 'background' signal as a fraction of the called-base peak area. This value will tend to be relatively high for traces with

- little or no signal
- a mixture of inserts

The read data block consists of one line for each read base. Each line has the three values

- the called base (a, c, g, t, or n)
- the quality value assigned to the base
- the location of the called-base peak in the trace

The values are separated from each other by a single space.

11. Notes

ESD Files

Phred reads processed MegaBACE ESD files. It cannot read the raw ESD files. It is important that you identify the dye chemistry correctly when you run the MegaBACE base caller so that phred can assign the right base to each trace. (This is important with ABI data as well.)

In order to obtain the best phred quality value accuracy with MegaBACE data, phred must use the quality value lookup tables designed for this data. Phred identifies the sequencing machine by reading the primer ID string in the chromatogram and matching it with an entry in the phredpar.dat file. The matching entry lists the chemistry, dye, and sequencing machine types. For example, the primer ID string of the form 'ET Primer' identifies a chromatogram as ET dye primer data generated on a MegaBACE sequencing machine. You can check that phred interprets the primer ID string correctly by using the '-v 63' option to have phred write diagnostic information to the screen.

LI-COR Data

Band Spread Ratio (BSR)

Phred reads SCF files created by the LI-COR gel processing software and has quality value lookup tables calibrated for traces processed with Band Spread Ratio (BSR) of 2.2. The LI-COR software writes a primer ID string in the SCF file that indicates the BSR value used in the trace processing, which for BSR=2.2 is

'DyePrimer{LI-COR_IR_2.2}'. Accordingly, the phredpar.dat file in this distribution has an entry with this string, which enables phred to recognize LI-COR traces processed with BSR=2.2, and to use the quality value lookup table designed for this LI-COR data. Phred has a quality value lookup table for data processed with BSR=2.2 only so the quality values for LI-COR traces processed with other BSR values will have reduced accuracy.

12. References

Brent Ewing, LaDeana Hillier, Michael C. Wendl, and Phil Green. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. 1998. Genome Research 8:175-185.

Brent Ewing and Phil Green
Base-calling of automated sequencer traces using phred. II. Error probabilities. 1998. Genome Research 8:186-194.

End: PHRED.DOC