SeqQual version 1.0 – User guide

Contacts
For any question, please contact Tiange Lang at langtiange@xtbg.org.cn or Pauline Garnier-Géré at pauline.garnier-gere@pierroton.inra.fr

Using SeqQual scripts

The SeqQual website (https://github.com/derF/SeqQual/) provides an overview and documents that detail the scripts usage (with examples of shell command files for many of them). This might be all you need to use some scripts. This document gives a few more details on how to organize and install the scripts on a linux computer system, especially if you are not familiar with this type of environment.

Introduction

The first scripts of SeqQual were developed a few years ago, before the advent of next-generation sequencing techniques, to automatically process large amounts of chromatogram sequence data (*.ab1/abd/scf files), with the main objective of decreasing the time-consuming steps of checking 1) the existence of polymorphisms (SNPs, SSRs, INDELs) and heterozygote genotypes for diploid sequences (which were commonly performed "by eye" in various chromatogram editors), and 2) the data quality before starting population genetic analyses. There is no limit in the amount of data (i.e. fragments, genes) that can be processed, apart from the memory and space available in your computing environment. The scripts developed for these tasks are called **SeqQual-part1-scripts**.

SeqQual was later extended with scripts to deal with large *.ace assembly files generated from next-generation DNA sequence data (El Mujtar et al. 2014, Brousseau et al. 2014), with the same objective of integrating quality information on a per nucleotide basis (using either *.phd.1 files on single reads or large *.phd.ball.1 or *.qual files) while this was not commonly done. These scripts constitute **SeqQual-part2-scripts** of the pipeline.

Since most SeqQual output files are **fasta alignment files where accepted nucleotides are those above a particular quality threshold (usually the Phred score),** with the ones below that score being replaced by question marks (?)**,** a large number of tools have also been developed to further filter, trim, post-treat those fasta files, and compute summary statistics on various polymorphisms, or change their formats (for example into Arlequin input files, http://cmpg.unibe.ch/software/arlequin35/). Output formats including (?) can also be uploaded in DNAsp (http://www.ub.edu/dnasp/). However, please note that DNAsp will exclude any position with at least one missing base for most analyses. These scripts are the **SeqQual-part3-fastools-scripts.**

SeqQual scripts can be used without being familiar to Unix/Linux computer environments, by following the instructions given in the *.pdf describing the scripts usage (on the website), in the *.sh example files (on the website), and in the rest of the help document. Many websites can however give a good initiation to unix/linux systems:
http://www.ee.surrey.ac.uk/Teaching/Unix/ (and links therein)

http://www.ryanstutorials.net/linuxtutorial/

Installation and how to get started?

1) For simplicity, scripts for any part (1), 2) or 3) or all of them together can be unzipped under your home directory:
   unzip **SeqQual-part2**.zip

The scripts will automatically be located under the home/SeqQual (or ~/SeqQual) folder.

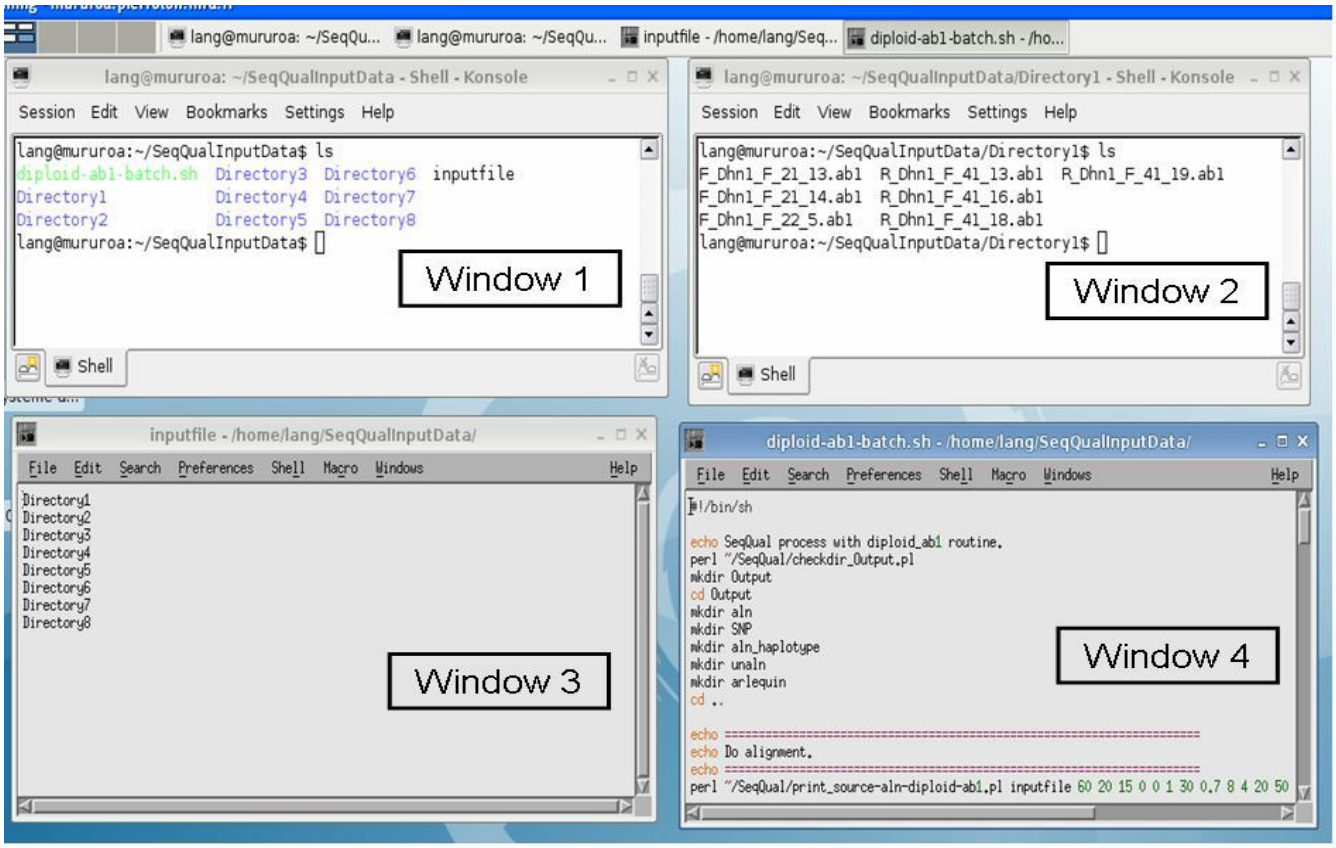
2) Put your data in one or more folders. For Sanger *.ab1 data, different or overlapping fragments from one region can be put in the same folder and SeqQual will assemble them using the Phrap software (but see **Warnings** below). If the input data are fasta files already (e.g. *.aln/*.fa) or *.ace files, put them in one single folder. In the folder above this(these) folder(s), create a small text file (e.g. "inputfile") containing the list of sub-folders (minimum of one), with one line per sub-folder name (see **windows 1 &3** below). This can be done by typing "ls > inputfile", and editing the file "inputfile" by keeping the chosen data folders.

3) Then choose one of the shell (*.sh) files shown as examples in the **SeqQual-part-*usage.pdf** documents that can be downloaded from the website (if they are unzipped under your home folder, they will be located under the home/shell folder and they already refer to the ~/SeqQual folder for calling scripts). This files can be edited, or you can extract or comment out (using the # tag) the command lines needed to run one or more scripts. "Echo" lines are here to follow what's being done on the screen, but they can be commented out. Some scripts should be used together, this is detailed the *usage.pdf documents. Files can be copied and executeed from above the sub-folders containing the data or somewhere else with the correct path:
   source ~/path/filename.sh , "path" being the location of the *.sh file.
   And if you are in the folder just above your data, you just have to type:
   source filename.sh

*NB: if the *.sh files need executable rights, you can add them with the unix command "chmod+x filename.sh".*

In the case where some scripts are likely to take a few hours or more, you may want to disconnect the process from the window in which a *.sh file has been launched. For this, you can use:
nohup ./filename.sh (with filename.sh files having the executable rights, see just above)
or use nohup ./$(source) filename.sh

The figure below gives an example of one possible Unix/linux environment for beginners: Window 1 is the screenshot of a terminal window that shows the folder ~/SeqQualInputData that contains the sub-folders Directory1 to Directory 8, in which data are located (for example, see the *.ab1 files in Directory1 (Window 2).



Window 3 shows the content of the text file named “inputfile” (listed in Window1) which is simply the folder list on which you want to run a few scripts and Window 4 shows the content of the start of a *.sh file used here on diploid chromatograms, which you can also see listed in Window 1.

4) Output files are usually fasta alignment files with *.aln extension (or other common fasta extensions) which are usually stored **in a new Output folder.** For chromatogram data, each alignment name in this folder correspond to the subfolder name containing the original data.

5) When you launch a series of scripts, associated log scripts can also be ran to document the parameter values, options used or actions performed in dated log files

Other softwares required

For **SeqQual-part2** or **–part3 scripts**, only perl and bioperl are required.

a) perl (probably already installed, see http://www.misc-perl-info.com/install-perl.html)
b) bioperl: http://bioperl.org/ (or this suite of applications can be automatically downloaded from your application manager tool)
They can also be used under Windows OS via Cygwin (set of software tool that provide a Unix-like environment, see http://www.cygwin.com/), once you have also installed perl and bioperl for cygwin (for example Active perl at http://www.activestate.com/activeperl/downloads and : http://bioperl.org/INSTALL.WIN.html)

For the **SeqQual-part1 scripts** which deal with chromatogram sequence data, i.e. *.ab1/abd/scf files, the phred/phrap/Consed suite is also needed, see at: http://www.phrap.org/consed/consed.html#howToGet, and see the installation instructions of the authors. Concerning the use of PHRED, and the use of the phredpar.dat parameter file, please refer to the PHRED install procedure for more explanations:

Summarised extract of the Phred installation document:

Set the 'PHRED_PARAMETER_FILE' environment variable.
Phred attempts to read a parameter file that identifies sequencing reaction chemistries/dyes. The 'PHRED_PAMETER_FILE' environment variable specifies the file path.
For example, create the folder /usr/local/etc(or "bin")/Phred and copy the phred parameter file, called 'phredpar.dat', to this folder. Then set the environment variable 'PHRED_PARAMETER_FILE' to the full path name of the file.
If you are using the C shell then type the command

% setenv PHRED_PARAMETER_FILE /usr/local/etc/Phred/phredpar.dat

If you are using sh, or variants such as bash, set the 'PHRED_PARAMETER_FILE' variable with the commands:

$ PHRED_PARAMETER_FILE=/usr/local/etc/Phred/phredpar.dat
$ export PHRED_PARAMETER_FILE= /usr/local/etc/Phred/phredpar.dat

We recommend that you set the PHRED_PARAMETER_FILE variable in the system-wide shell startup files (cshrc or equivalent).

Additionnally, for diploid sequence data, the POLYPHRED software is needed, which can be obtained at http://droog.gs.washington.edu/poly_get.html

An academic license is available to users at academic and nonprofit research institutions at no charge. See:
http://depts.washington.edu/ventures/UW_Technology/Express_Licenses/PolyPhred.php

Various older and more recent versions of the above programs have been used on different systems without any problems of compatibility, for example phred versions 0.071220.b (type phred –V to check yours), phrap versions 1.080812 (type phrap –v to see yours) , polyphred versions 5.04 and 6.18 (type polyphred).

Examples of shell files

Examples shell files (*.sh, set of command lines) are provided on the website for most scripts with detailed comments **(**https://github.com/derF/SeqQual/Shell_files_ex**),** making it easy to modify the parameters values in a linux text editor.

For checking input list of folders: → see *checkinput.sh*
All analyses are usually done in batch with a list of folder name(s) (listed for example in the file named "inputfile"). You can check for potential typing errors preventing script to run by typing **"source checkinput.sh"** in the folder where the file "inputfile" is located. An error message will be printed on the screen if there is any problem.

For cleaning old working folders: → see *clean.sh* and *clean-all.sh*
For several scripts organized in a pipeline, intermediate folders called 'mydata' are created, which include the folders' structure and all intermediate files needed for the

various scripts needed to run correctly. If the same (with different parameter values) or different scripts are used on the same data, the existing 'mydata' and 'Output' folders are renamed as 'mydata-suffix' and 'Output-suffix' with "suffix" being the current date and time. Therefore, after several runs of SeqQual scripts on the same data, the amount of files created and copied can be very voluminous, even though each file does not use much memory. So the clean*.sh scripts allow to remove all the old files stored in each subfolder as well as all the old "Output-…" folders, only keeping the latest ones for clean.sh and also removing the latest one for clean_all.sh. If you want to keep one of the old 'Output' folders before running clean.sh, you just need to rename them, with a name that does not begin with 'Output_', or just move them to another folder.

For chromatogram data (*.ab1/scf/abd files):
For haploid data chromatograms: → see *1.1-haploid-ab1.sh*
For diploid data chromatograms: → see *1.2-diploid-ab1.sh*
For ace files from haploid data (+phd files): → see *2.3-ace-qual.sh*
For ace files only without quality files: → see *2.1-ace-only.sh*
For ace files generated from large assembly of next-generation sequencing data (ex: 454) with fasta and fasta.qual: → see *2.3-ace-qual.sh*, *2.4-ace-qual2ball-combine.sh*.
For aligned fasta file which have been validated, in order to apply various post-processing steps (e.g. obtain SNP aligments and /or Arlequin input formats in batch): → see *3.1-fasta-diploid-data.sh* for diploid type data including IUPAC codes for heterozygotes, or *3.2-fasta-haploid-data.sh* for haploid type data.

In order to identify clusters or reads/sequences, a simple rule has been assumed in scripts, which consists in using an exclusive character chain at the start of each individual or sequence read in the fasta alignments (see comments in *3.2-fasta-haploid-data.sh* for more details, and pdf documents detailing the scripts usage on the website).

Examples of data and output

See **Data-examples** and **Results-examples** folders at https://github.com/derF/SeqQual/.
Most of the options have been activated in the example shell files used.

Speed and limits

From our own experience, thousands of genes and/or fragments, with chromatograms files organized in separate folders were processed in a few hours. Larger ace assembly files can take more time. Treating around 500000 reads for integrating quality in ElMujtar et al. 2014 needed a few hours. Steps of filtering and masking across a large number of fasta alignments files, especially if they are a few kb long can take a few hours to a few days longs, and regions can be split into 1 kb fragments to speed up processes in some cases (see split_aln_multi.pl for example).

Since many files/folders can be created or moved or copied from one folder to another in the process of running scripts, you need to be aware of the limits of your linux system. In a ext[234] filessystems computer, you have a fixed number of inodes. Each file or directory needs one inode. Type df –i to get your current count and limits. Also for the cp command for example, the max number of files that can be processed depends on the system and

version, on the number and size of the arguments and environment variable names. Use getconf ARG_MAX to get the limit.

Error messages

a) If your problems are linked to the use of a linux command, you should get an indicative error message
b) If the *.sh file you are using stops just after the launch or seem to work correctly then stops without producing any files in the Output folders:

First check that the "inputfile" file does not contain any error (e.g. a folder that does not exist). You can do a check of your folder list by using checkinput.sh (see above)

For SeqQual-**part1,** the names of the chromatogram sequence reads should not contain blank spaces. If that's the case, replace them by another character or remove them using the following unix command: rename 's/\ /\-/g' *.ab1 (will replace all white spaces with "-" in files with the ab1 extension in the directory in which command is launched)

Warnings and recommandations

a) About the alignments obtained

SeqQual is not an assembly pipeline, although the proposed default parameters in the **scripts of part 1** have been defined to work with a small number of sequences from unassembled Sanger data for one or several resequenced regions, which correspond to the large range of diversity expected **within a species or closely related species across genomic regions**. However, the quality of alignment obtained has to be decided by the user, and in particular with large datasets from different species, the alignments might not be optimal. In this case however, 2 options can be used: either change the values of the Phrap software to improve the alignments, or export the un-aligned fasta files that integrate quality and analyse them with other alignment softwares.

b) About SeqQual usage on chromatogram data

If you are processing chromatogram data, it might be a good idea to do a run with Phred score =30 then Phred score=20 (or 40 and 30 respectively), the comparison between both types of outputs is likely to be very informative in terms of possible false negative SNPs (and false positives)

c) Default parameters

Inappropriate parameter values can lead to misleading results and outputs. We are proposing a few defaults parameter sets that were adapted the cases that we treated, either on chromatogram data or ace assembly files (***see below List of parameters used***). For Sanger-based sequence data, the advantage of processing data automatically, even though it might not be optimal in the first run, is that alignments including quality at each nucleotide can be visualised, and thus any problem can be much more rapidly assessed than by checking chromatograms "by eye from the start" in a Chromatogram editor. Change of parameter values for Phrap can often allow solving an alignment problem. Most alignment problems usually arise when the overall quality is low, whether for haploid or diploid data.

d) Transfering files (*.sh or *.pl) from a Windows environment to your Unix server account via filezilla or ssh tranfer:

To avoid text format compatibility problems in linux in case of transfer from/via windows environments, please use the dos2unix utility (from the packages tofrodos or sysutils for Debian) by typing "dos2unix *.sh" and dos2unix *.pl" in their respective folders before using them.
Type: man dos2unix or dos2unix -h for more options
For files created under Mac, you can find the mac2unix command on the Web.

Summary of definitions for parameters used.

Phred parameters

Phred_quality_score: 20 or 30 (defaults values)
This is the quality score for each base corresponding to a peak in the chromatogram processed image by the sequence analyser software. A score of 20 corresponds to an error rate of approximately 1 in 100 bases, a score of 30 to 1 in 1000 bases etc…. Thus it is recommended to compare outputs with both 20 and 30 for medium quality sequence data. The same values for quality scores for next-generation sequencing data may have different probability meanings and recommended values can vary depending on the biotechnique, your experience, the software pre-processing the data and its parameters, but recommended values are usually above 20.

Polyphred parameters (*only in case of diploid data, comments below are based on practical experience and should be read with critical eyes***)**

Polymorphism_score (default 60): this is a score that is computed for each aligned base, and that affects the amounts of SNPs that will be detected. A default value is given initially, which serves as reference to compare computed values by the program for potential SNPs, since only those above the initial value will be considered as SNPs. From practical experience, high scores (above 60) correspond to reliable SNPs. However, this value does not affect the sequence data alignment outputs since those are based on the phred and heterozygote scores, and no output file indicating explicitly SNPs identified from this score are provided. Based on our data, the polymorphism scores and the occurrence of SNPs in the fasta alignment outputs were consistent.

Trim_quality_score: This is the quality score used in polyphred (with the same meaning than the phred score above) that is used to consider or not the reads' ends for detecting SNPs below a certain score. This score value should be below the Phred score asked for accepting nucleotides as valid in output alignments.

Phrap parameters used and their default values
(+in italics, extracts of the online help from the http://www.phrap.org/phredphrap/phrap.html website)

default_qual 20

Quality value to be used for each base **for initial alignment with Phrap and building of the consensus sequence**. A quality value of 15 corresponds to an error rate of approximately 1 in 30 bases, i.e. relatively accurate sequence.

Phrap help extract:

"Quality value to be used for each base, when no input .qual file is provided. Note that a quality value of 15 corresponds to an error rate of approximately 1 in 30 bases, i.e. relatively accurate sequence. If you are using sequence that is substantially less accurate than this and do not have phred-generated quality values you should be sure to decrease the value of this parameter."

Might affect the quality of the consensus sequence produced in the output alignment.

trim_start 0

Number of bases to be removed at beginning of each read.

force_level 0

Relaxes stringency to varying degree during final contig merge pass. Allowed values are integers from 0 (most stringent) to 10 (least stringent), inclusive.

bypass_level 0

Controls treatment of inconsistent reads in merge. Currently allowed values are 0 (no bypassed allowed; most stringent) and 1 (a single conflicting read may be bypassed).

For data from the same fragment a priori, 0 should be the default, this has to be tested.

maxgap 30

Maximum permitted size of an unmatched region in merging contigs, during first (most stringent) merging pass.

repeat_stringency 0.80

Controls stringency of match required for joins. Must be less than 1 (highest stringency), and greater than 0 (lowest stringency).

This parameter is very important and should be adjusted depending on the similarity/dissimilarity of the sequences that you expect. For within species data, > 0.8 should be low enough, except if you have diploid data with a very small part of the fragment that can be aligned before a heterozygote indel for example. In this case, 0.7 has been chosen as the default value but this can be modified further. The value 0.7 has been put in each *.sh example given, this can be increased if needed (mixture of paralogs sequences for example). For within species data in which fragments have been resequenced, we had to use a value as low as 0.5, with a maxgap parameter of 100 (see above) in particular cases where large indels were detected.

nodeseg 8

Minimum segment size (for purposes of traversing weighted directed graph).

Refer to Phrap documentation online.

nodespace 4

Spacing between nodes (in weighted directed graph).

qual_show 20

Cutoff for flagging "low_quality" regions in contig sequence and "high quality" discrepancies between read and contig. Bases in the .contigs and .ace file are lowercase if and only if their LLR-converted quality values are below this value.

max_subclone_size 5000
Maximum subclone size -- for forward-reverse read pair consistency checks.
A priori needed only in case of EST sequence data where lots of contiguing work is needed.

trim_score 20
Minimum quality score for identifying degenerate sequence at beginning & end of read.

trim_penalty -2
Penalty used for identifying degenerate sequence at beginning & end of read.

trim_qual 13
Quality value used in to define the "high-quality" part of a read, (the part that should overlap; this is used to adjust qualities at ends of reads).

confirm_length 8
Minimum size of confirming segment (segment starts at 3d distinct nuc following discrepancy).
NB: confirmed reads (i.e. reads matching some other read)
NB→ in case of a small region with a high number of polymorphism, this may affect the contiguing into more than one contig while only one fragment is expected.

confirm_trim 1
Amount by which confirming segments are trimmed at edges.

confirm_penalty -5
Penalty used in aligning against "confirming" reads.

confirm_score 30
Minimum alignment score for a read to be allowed to "confirm" part of another read.

Indexwordssize 10
Size of indexing (hashing) words, used in finding word matches between sequences. The value of this parameter has a generally minor effect on run time and memory usage.

SeqQual Core parameters/options

Three options are available and were used for modifying the output fasta alignment files, but they should be used with caution, even more for lower quality data. Those options can be used independently to one another.

a) **Truncate** start and end of alignment files.

In the aligned fasta files, you often see many missing data at start and end of alignment. Here we allow the user to truncate the start and the end of the alignment files according

to the number of the **NON-MISSING** data that he wants to keep in each column. If in one position, starting at both ends of a contig, the number of non-missing data is below the parameter given by the user, the position is removed (default 1 before initial checkings of outputs).

b) **Replace** isolated nucleotides by "???" in the alignment files.

This is just a small *ad hoc* option for a better visualisation, in case bad reads still have some clean peaks scattered throughout: it allows to replace isolated nucleotides (either one, two or three neighbour nucleotides) surrounded by at least 5 "?"s of missing data . It can be repeated twice in the *.sh file in needed, use with caution.

c) **Remove** spurious positions in the alignment files due to false insertions.

Positions are removed which only have insertions in the initial phrap steps, which turned out to be bad quality data later replaced by missing data (indicated by "?") in the alignment output files.
Increasing the *trim_score* (see above) here should also decrease the problem. This option is to be used with caution, in particular for medium quality data, as a base could artificially be removed in a gene and change its ORF etc. This option is more for helping preliminary visualisation of fasta outputs. Also in some rare cases, we noticed rare bugs of very low Phred scores in all bases of a few contiguous positions in starts and end of sequences, despite the high quality of the data, which created small regions with only "???" despite correct data and consensus.

Additional Comments on identified problems with the first alignment outputs and possible actions for changing parameters

Case of long stretches of bad quality sequences
→ Output fasta alignments will then contain small isolated stretches of DNA of decent quality, often badly aligned but not informative, so manual editing might be needed (replacing bad regions with missing data).

Haploid data
Case where you see columns of around 3 question marks "???" in the middle of alignments or a mixture of "???" and an excess of heterozygotes corresponding to one SNP → this means a possible amplification of paralogs, either on already diploid sequences, or run the diploid pipeline on haploid data to see if heterozygotes are identified. If too many are observed while data should be haploid, this is also the sign of paralog amplification or contamination.
Also you can drop the phred score a bit (from 30 to 25 or 20) if a high number of "???" are observed in otherwise correct alignments

Diploid data
If you have the same case than above → check by eye on a few alignments if this would not be a case where heterozygotes are not producing clear double peaks (due to differential amplification of both strands, or due to paralog amplification and then the second peaks are slightly shifted instead of completely overlapping)→ see if lowering the heterozygote score can help, but use with caution.

Case of heterozygote indels with diploid chromatogram data

Heterozygote indels are fairly easy to detect on fasta alignments output files where low quality is being masked. Such cases include for example the presence of homozygotes with either an insertion or deletion, and heterozygote indels on other individuals can be identified by a much higher frequency of "???" from around the area where the polymorphic indel is located. The availability of good quality forward and reverse sequences in case of only one heterozygote indel in one fragment greatly helped in their detection as well. More than one heterozygote polymorphic indels will result in series of reads with good quality parts and then parts of sequences with very poor quality and thus only missing data.

False positives

Cases of false positives are in general easily solved by adjusting the different scripts' parameters, either increasing the phred score, or using more or less stringent alignments parameters (e.g. repeat stringency parameter from the phrap parameters). In all the cases that we checked "by eye" (100s of them), very few false positives have been detected, apart from those due to possible alignment problems which are identified easily.

False negatives:

More false negatives can occur due not only to the performance of the programs used in the pipeline but also often from actual amplification problems. This is a difficult problem which has been encountered essentially with diploid data, and could be due to many reasons:

a) Unequal amplification of both DNA strands, which then produce 2 peaks but one with a much lower height and then a base call is the same as for one base only. SNPs corresponding to homozygotes for a rare allele are unlikely and should be carefully checked as some heterozygotes in that case are likely to have been missed.
b) For the same reason, some heterozygotes might be missed, and this is a problem that has been encountered on particular fragments.
c) In all case, increasing the phred score can help identify some heterozygotes which would then be replaced by "?" in the middle of very good quality reads.

Also, even if the lower quality can be visualised by a higher proportion of "???", sequences with low quality regions all along are prone to more frequent false negatives or positives.

Disclaimer

We have tested SeqQual on our own servers, and used it routinely on our own data, additionally to having used it on several datasets of colleagues that allowed us to validate the scripts. However, we can take no responsibility for any damage caused by running the programs, not the quality of data it produces.

Copyright notice