A visual system for DNA sequencing quality control

Eliseu Binneck[1*]
Adriane Wendland[6]
Agda Maria Rodrigues Morales[4]
Álvaro Manoel Rodrigues Almeida[1]
Carlos Alberto Arrabal Arias[1]
Cesar Augusto da Silveira[1]
João Flávio Veloso Silva[1]
José Renato Bouças Farias[1]
Juliana C. Molina[2]
Julio César Pedroso[5]
Michèle Claire Breton[2]
Noelle Giacomini Lemos[2]
Norman Neumaier[1]
Polyana Kelly Martins[3]
Renata Fuganti[2]
Renata Stolf[2]
Silvana Regina Rockenbach Marin[1]
Alexandre Lima Nepomuceno[1]
[1]Embrapa Soja – CNPSo. *Corresponding author: binneck@cnpso.embrapa.br
[2]Universidade Estadual de Londrina - UEL
[3]Universidade Federal de Viçosa - UFV
[4]Universidade Filadelfia - UNIFIL
[5]Universidade Estadual de Maringá – UEM
[6]Esalq/USP

ABSTRACT

There is a need for a versatile visual system for quality control of reads generated on a small to medium-scale DNA sequencing effort. Some automatic sequencers have a software system to basecalling and display the quality of reads, but often a more accurate program like Phred [1, 2], currently the most widely used basecalling software, is required to measure the error probability associated with each base through chromatogram analysis.

The principal use of Phred analysis is to produce the input files to programs that perform sequence trimming, clustering or assembly and finishing process, although it also can be useful for a control evaluation of the reads on the sequencing lab. The inconvenience is the fact that the raw text outputs of Phred are not easily readable and informative for laboratorists. We aimed to solve this task by writing a series of Perl [3, 5] multiplatform programs that make a system to read the Phred `.fasta` and `.fasta.qual` and Cross_match [4] `.fasta.screen` output files and produce a set of web based visual intuitive reports.

The system generates reports of the reads in FASTA colored format with visual quality information for each base: red stands for Phred score < 10; green stands for Phred score >= 10 and < 20; blue stands for Phred score >= 20 and < 30; and, black stands for Phred score >= 30.

Also, a general report for each 96 wells plate is produced on a plate shape figure where the sequence quality is reported as a colored button for each well. As default, green stands for an insert fragment of 200 or more bases with Phred score >= 20, yellow stands for a vector fragment of 200 or more bases with Phred score >= 20 if the first statement was not true, and red stands for a lower quality sequence. These parameters (Phred score and fragment size) are adjustable by the user on the command line. This report functions as a fully clickable map, giving access to each sequence on FASTA colored format as described above.

The reports are automatically produced and placed on a local Intranet to be accessed by the laboratory people. An example is at http://www.cnpso.embrapa.br/bioinformatica.

REFERENCES

[1]Ewing, B.; Hillier, L.; Wendl, M.C. and Green, P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res. 1998 Mar, 8(3):175-85.

[2]Ewing, B. and Green, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res. 1998 Mar, 8(3):186-94.

[3]Stein, L.D. Using Perl to facilitate biological analysis. Methods Biochem Anal. 2001, 43:413-49.

[4]The Phred/Phrap/Consed System Home Page. http://www.phrap.org

[5]Wall, L.; Christiansen T. and Schwartz, R.L. Programming Perl (2nd ed.).O 'Reilly, 1996.