Online journal ISSN - 1676-5680
GMR Genetics and Molecular Research
EVOLUTION AND TECHNOLOGY

Effects of sample re-sequencing and trimming on the quality and size of assembled consensus sequences

F. Prosdocimi[1], D.A.O. Lopes[1], F.C. Peixoto[2], M.M. Mourão[3], L.G.G. Pacífico[4], R.A. Ribeiro[5] and J.M. Ortega[1]

[1]Laboratório de Biodados, Departamento de Bioquímica e Imunologia, ICB-UFMG, Belo Horizonte, MG, Brasil
[2]Laboratório de Computação Científica, UFMG, Belo Horizonte, MG, Brasil
[3]Laboratório de Genética-Bioquímica, Departamento de Bioquímica e Imunologia, UFMG, Belo Horizonte, MG, Brasil
[4]Laboratório de Imunologia de Doenças Infecciosas, Departamento de Bioquímica e Imunologia, UFMG, Belo Horizonte, MG, Brasil
[5]Laboratório de Biodiversidade e Evolução Molecular, Departamento de Biologia Geral, UFMG, Belo Horizonte, MG, Brasil
Corresponding author: J.M. Ortega
E-mail: miguel@icb.ufmg.br

Genet. Mol. Res. 6 (4): 756-765 (2007)
Received August 03, 2007
Accepted September 25, 2007
Published October 05, 2007

ABSTRACT. The production of nucleic acid sequences by automatic DNA sequencer machines is always associated with some base-calling errors. In order to produce a high-quality DNA sequence from a molecule of interest, researchers normally sequence the same sample many times. Considering base-calling errors as rare events, re-sequencing the same molecule and assembling the reads produced are frequently thought to be a good way to generate reliable sequences. However, a relevant question on this issue is: how many times the sample needs to be re-sequenced to minimize costs and achieve a high-fidelity sequence? We examined how both the number of re-sequenced reads and PHRED trimming parameters affect the accuracy and size of final consensus sequences. Hundreds of single-pool reaction pUC18 reads were generated and assembled into consensus sequences with CAP3 software. Using local alignment against the published pUC18 cloning vector sequence, the position and number of errors in the consensus were identified and stored in MySQL databases. Stringent PHRED trimming parameters proved to be efficient for the reduction of errors; however, this procedure also decreased consensus size. Moreover, re-sequencing did not have a clear effect on the removal of consensus errors, although it was able to slightly increase consensus.

Key words: Sequencing reads, Trimming, Assembling, Consensus, Codifying sequences, PHRED, CAP3

INTRODUCTION

Many of the recent developments in the genomics and bioinformatics field deal with data generated from genome-sequencing projects; it is well known that genomes are build *in silico* by the superposition of thousands of overlapping reads joined together by assembly software, such as PHRAP (Green, 1998) or CAP3 (Huang and Madan, 1999). Some assembly softwares, including these ones two, take advantage of base quality values determined by base-caller algorithms, such as PHRED (Ewing and Green, 1998; Ewing et al., 1998), in order to produce more reliable consensus sequences. Although their main application consists in the production of huge genomic sequences, assembly softwares are also used to cluster expressed sequence tag (EST) data. In this latter case, the project focus is shifted to gene discovery based on single-pass, partial sequencing of cDNA molecules, aiming to analyze the transcriptome (Adams et al., 1991; Franco et al., 1997). One interesting issue about clustering consists in the fact that assembled molecules from genome projects are allowed to enter in genome databases while assembled ESTs are restricted to specific project websites, and they are not allowed to be integrated into any of the best-known public molecular databases. Nevertheless, evolution of an EST project to a full-length cDNA sequencing project is not rare, such as the Mammalian Gene Collection (Strausberg et al., 1999; MGC Program Team, 2002), in which selected clones are introduced into a pipeline of dedicated sequencing to eliminate ambiguities from the reads and generate high-quality consensus sequences. Consequently, these manually edited consensus sequences are allowed to be deposited into GenBank and/or GenPept databases, becoming targets for ordinary BLAST similarity searches. Ideally, a combination of forward and reverse reads should be used in EST-sequencing projects, but many of the selected cDNA clones are larger than the distance that could be covered in both orientations with the simple alternative of using vector-anchored primers. Thus, the question that rises is whether or not a sufficiently large number of reads could be assembled into an error-free consensus and, if so, what would be the cost/benefit relationship between the number of samples sequenced and the efficiency in the production of this high-quality consensus, which could be promptly deposited as a partial cDNA sequence, either 5' or 3'.

Another potential alternative is the manual editing of consensus with software such as Consed (Gordon et al., 1998), a procedure that shall be encouraged in place of automated alternatives. However, the Consed operator would certainly benefit from additional information produced by automated tools, such as the expected number of errors per molecule as a function of i) the number of available reads clustered and ii) the quantity of errors admitted during trimming procedures. In genomic projects, trimming is not usually recommended because high-quality regions often overlap low-quality ones to close gaps. However, this is not the case when partial sequencing of cDNA molecules is done, since all reads are expected to start at the same position and, most importantly, the low-quality regions are concentrated at the edges of the sequences.

We analyzed the effects of sample re-sequencing (from 2 up to 10) and PHRED trimming parameters on assembled consensus' errors and size. All procedures were conducted using a set of 846 pUC18 one-direction reads, generated by a single-pool sequencing reaction (Prosdocimi et al., 2004). Assembling was conducted with CAP3 software and errors were analyzed with BLASTn (Altschul et al., 1997). Trimming efficiently reduced the number of errors, but it affected the size of the consensus, while the impact of re-sequencing was not as strong as it might be intuitively expected.

MATERIAL AND METHODS

Sequencing reactions

The sequencing reaction premix was made in a single pool and divided into several tubes for PCR-sequencing reactions. PCR products were joined together in the same tube, mixed, and sequenced in 96-well plates with MegaBACE equipment. Three laboratories of the Brazilian Federal University of Minas Gerais State (UFMG) participating in the Minas Gerais Genome Network provided 846 processed pUC18 ESD files.

Base calling and trimming

All pUC18 ESD files were processed by PHRED using variable trimming parameters. First, PHRED was run on each sequence with no trimming parameters (nT data). Then, PHRED was performed using "*-trim_alt*" parameter. When using "*-trim_alt*", the parameter "*-trim_cutoff* " was set from 0.01 (1%) to 0.25 (25%) for each read. This means that each read was trimmed 26 times with different PHRED trimming parameters. FASTA and QUAL files were stored.

Sequence assembly

One thousand groups of two sequences were randomly selected and assembled with CAP3 software from the 846 pUC18 ESD files. The same procedure was applied to groups of 3, 4, 5, 6, 7, 8, 9, and 10 sequences. In all, we selected and assembled 9000 sequence groups.

Local alignment against pUC18 published sequence

All the CAP3 consensus sequences were compared to the published pUC18 sequence (24.8% A, 25.2% C, 25.5% G, 24.5% T; accession number L08752) using the local alignment algorithm BLAST. Tabular output data (-m 8 option) was used to populate MySQL tables.

Statistical analyses

Since the data did not fit a normal distribution, non-parametric ANOVA statistical tests were performed. We have ran Kruskal-Wallis median tests to analyze the number of errors and size of the generated consensus.

RESULTS

The efficiency of re-sequencing in the production of error-free consensus was evaluated by sampling thousands of groups containing two up to ten reads from a collection of 846 pUC18 cloning vector reads. Reads were base called with PHRED software and assembled with CAP3. During PHRED processing, no trimming of reads' low-quality portion is normally performed (denoted by nT - no trimming - in figures). By aligning the 9000 consensus produced with the published pUC18 sequence using BLASTn program, the errors in these *in silico* sequences (sometimes called contigs) were evaluated. BLASTn alignments do not elongate over the low-quality portion of the reads; therefore, errors per sequence tend to a maximum.

Additional data were included to consider PHRED trimming. The internal algorithm "*-trim_alt*" was used, varying the trimming cutoff from 1 to 25% of accepted errors at the edge of reads in order to check the effect of this pre-processing on the quantity of errors observed in CAP3 consensus sequences.

Stringent PHRED trimming was able to reduce errors to less than one per consensus sequence (Figure 1A). Re-sequencing (increasing from 2 up to 10 reads) was expected to significantly reduce the number of errors seen in the consensus assembled; however, the reduction was not as great as one might suppose. In two regions where the differences among data were either maximized or minimized (trimming cutoff of 1% or nT), we present Kruskal-Wallis non-parametrical statistical analysis (Figure 1B and C). When reads were trimmed with a cutoff of 0.01 (1%), the effect of increasing the number of reads from 2 to 10 was a 4.3-fold reduction in errors per molecule (up to 24% of the initial amount, Figure 1B). However, without trimming (Figure 1C) there was a reduction of 1.5-fold (64% of the initial amount remaining) and there was no significant reduction from 3 up to 8 reads. Thus, trimming decreases the errors more efficiently, presenting higher responsiveness to error reduction than re-sequencing.

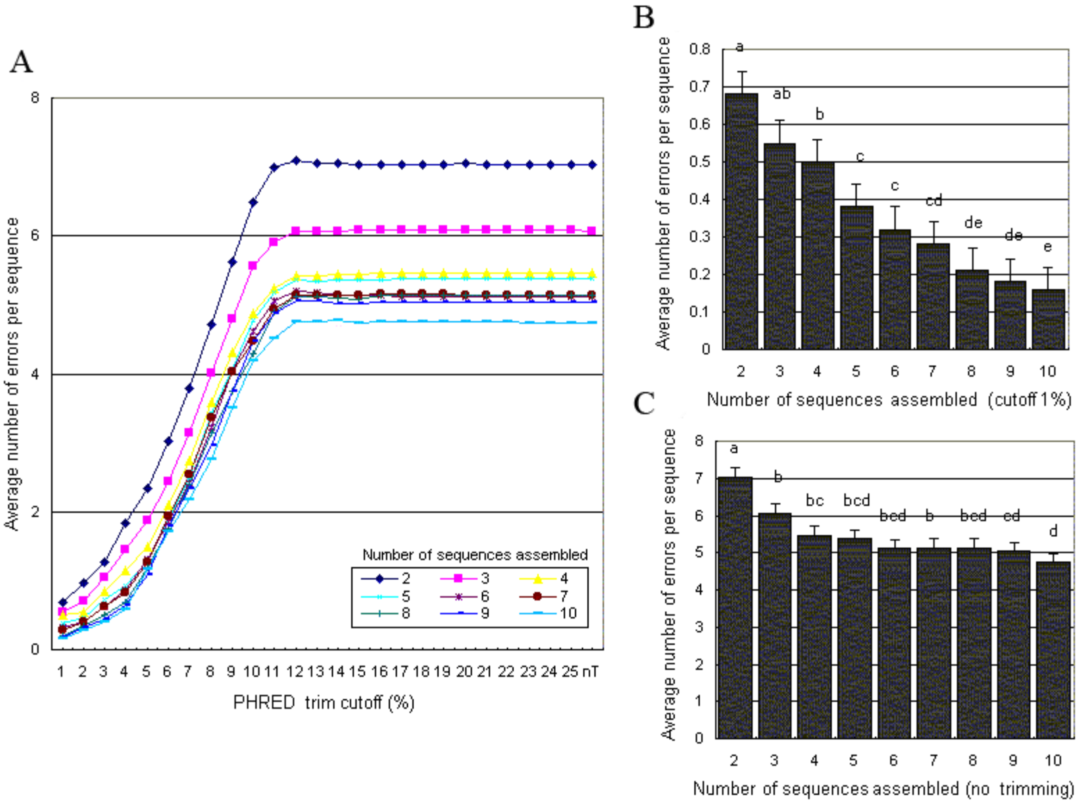

Figure 1. A. Average number of errors per sequence when different numbers of reads (2 to 10) were assembled with CAP3 and aligned against the published pUC18 sequence using BLAST software, sorted by PHRED trim cutoff percentage. **B.** Clusters of sequences trimmed with cutoff 1% (zoom). **C.** Clusters produced from non-trimming sequences (zoom).

The surprising effect of increasing the number of reads on molecules trimmed under a trim cutoff of 1% (corresponding to PHRED 20 trimming) led us to investigate the nature of these errors. At 1% trimming cutoff, mismatches were minimum, even when using only two reads (Figure 2A). However, from 10% cutoff up to no trimming (nT), the number of mismatches decreased in proportion to the number of errors as more reads were assembled (Figures 1A and 2A). In contrast, when gaps were analyzed, the opposite was observed: under PHRED 20, gaps were efficiently reduced as the number of reads increased, but this was not observed for non-trimmed or poorly trimmed reads (Figure 2B). This last observation is similar to findings that high-quality errors are mainly generated by insertions (Prosdocimi et al., 2003), thus producing gaps in the alignment. Therefore, the efficient reduction in the number of errors under PHRED 20 is due to a decrease in the number of gaps (insertions/deletions). Moreover, the decrease in mismatches and gaps under PHRED 10 up to no trimming is rather similar and low.

Curiously, the number of gaps was lowest with 4% cutoff for two reads and at 2% cutoff when 10 reads were clustered.

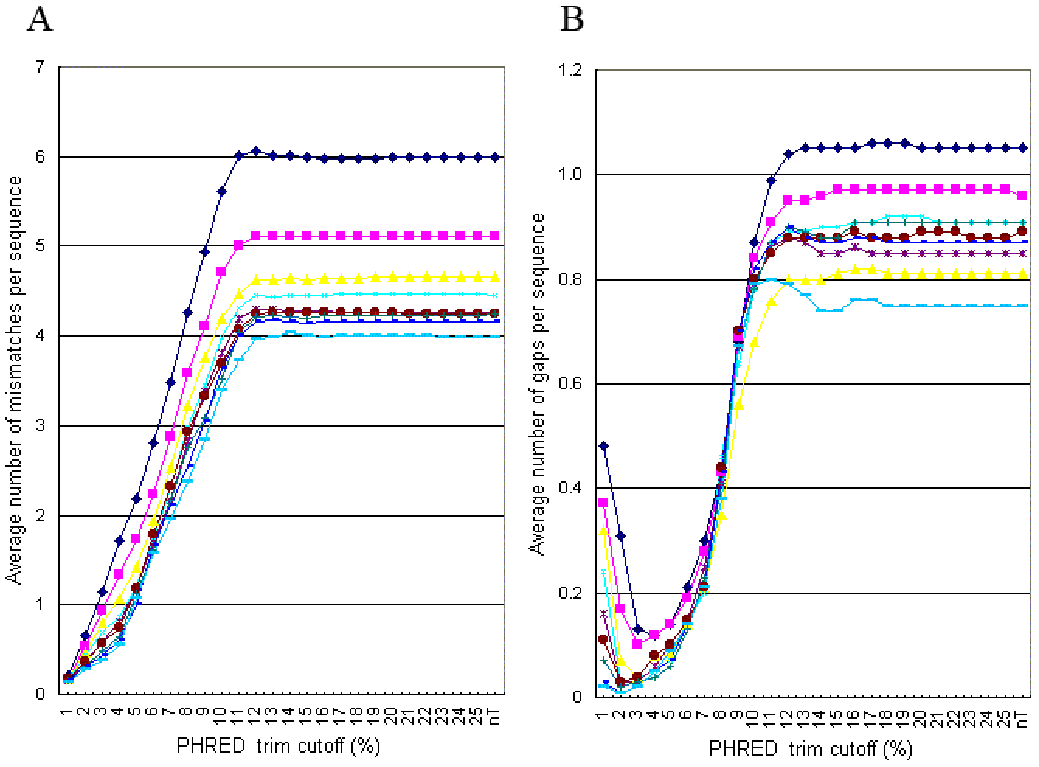

Figure 2. A. Average number of mismatches per sequence when different numbers of reads (2 to 10) were assembled with CAP3 and aligned with the published pUC18 sequence using BLAST software, sorted by PHRED trim cutoff percentage. **B.** Average number of gaps per sequence when different number of reads (2 to 10) were clustered. The colors indicating the number of reads clustered are the same as in Figure 1.

Although trimming under PHRED 10 (10% cutoff) was found to reduce the number of errors, the trimming effect at this range also decreased the consensus size (Figure 3A). The resultant assembled consensus under PHRED 10 was smaller than 500 bp. Moreover, the consensus size was more responsive to the number of reads under PHRED 20 (1% cutoff, Figure 3B) than when using non-trimmed reads (Figure 3C).

We considered the fact that all reads start at some distance from the primer (Prosdocimi and Ortega, 2005) and progressively lose quality as they proceed away from the starting position. This might result in situations where a poor-quality edge of a single read represents the quality of the consensus, even if 10 sequences have been assembled. Thus, we conducted the experiment exemplified in Figure 4. First, three up to ten reads were assembled and the consensus was aligned to the individual reads used in the assembly. After that, any portions of the consensus generated by only one or two reads were eliminated to ensure that each position of the consensus would be covered by at least three reads.

The maximum number of errors per sequence diminished from ~6 (Figure 1A) to ~2.5 (Figure 5A). Again, re-sequencing from 3-10 reads produced only a small effect on the number of errors per consensus when non-trimmed individual reads (nT) were used (Figure 5C). Unex-

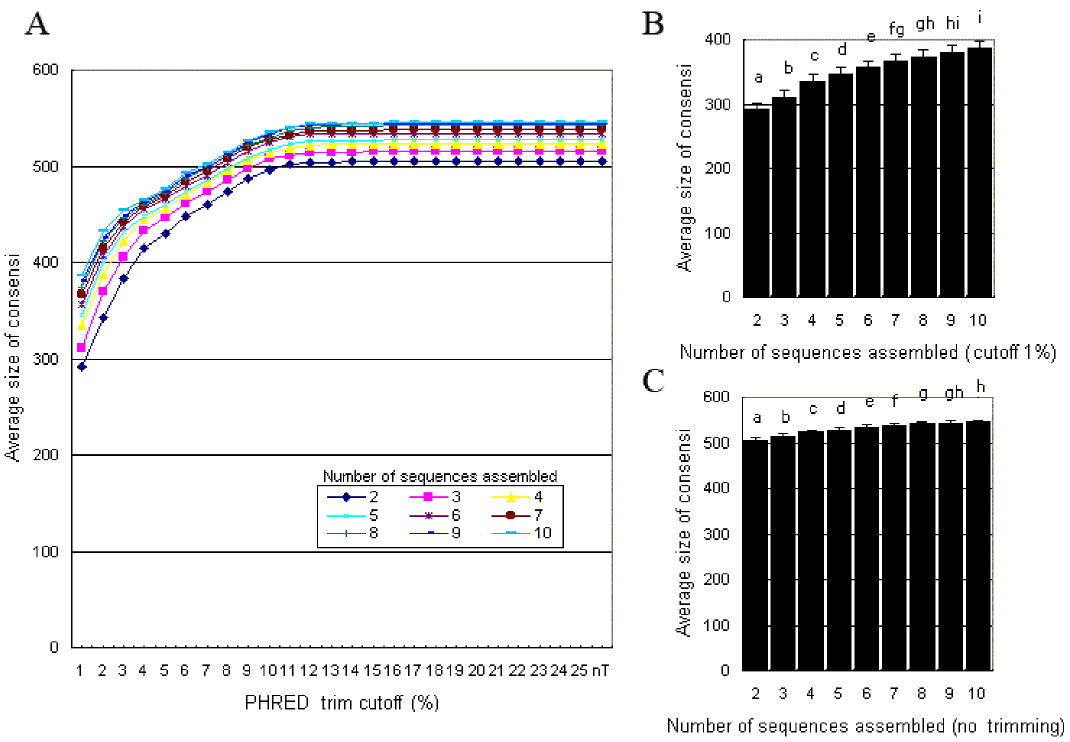

Figure 3. A. Average size of consensus sequences when different number of reads (2 to 10) were assembled with CAP3, sorted by PHRED trim cutoff percentage. **B.** Clusters of sequences trimmed with 1% cutoff (zoom). **C.** Clusters produced from non-trimming sequences (zoom).

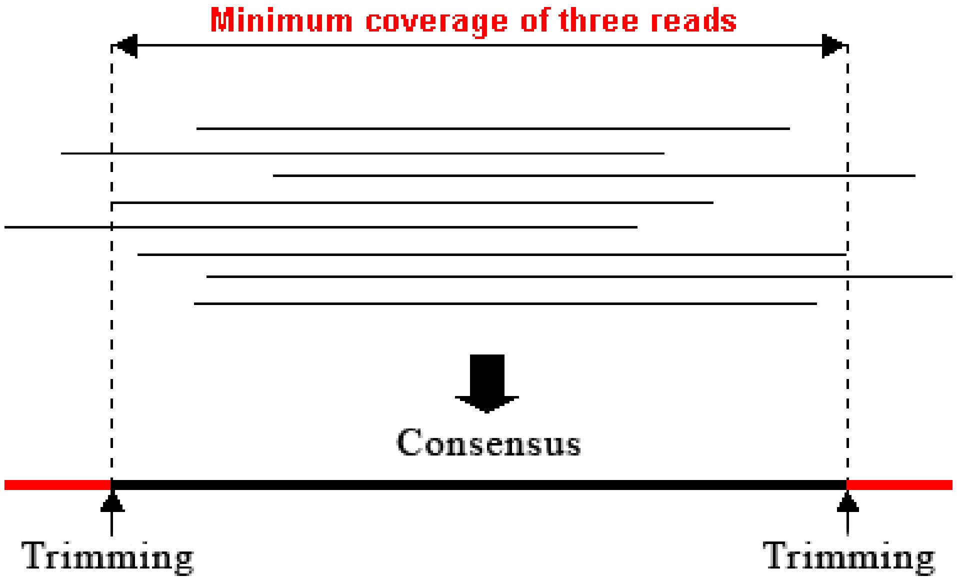

Figure 4. Methodology for consensus trimming.

pectedly, the assemblage of more than four reads increased the number of errors when PHRED 20 cutoff was used (Figure 3B) when analyzing these 3-read coverage consensus sequences. The number of errors in nT sequences, as compared to the simpler procedure (Figure 1A), was reduced more than 50% (from 5-7 down to 2.5 errors per molecule). However, this 50% reduction is still lower than the effect of trimming the reads with higher values such as PHRED 20 (Y-axis in Figures 1B, 5B and 1C, 5C).

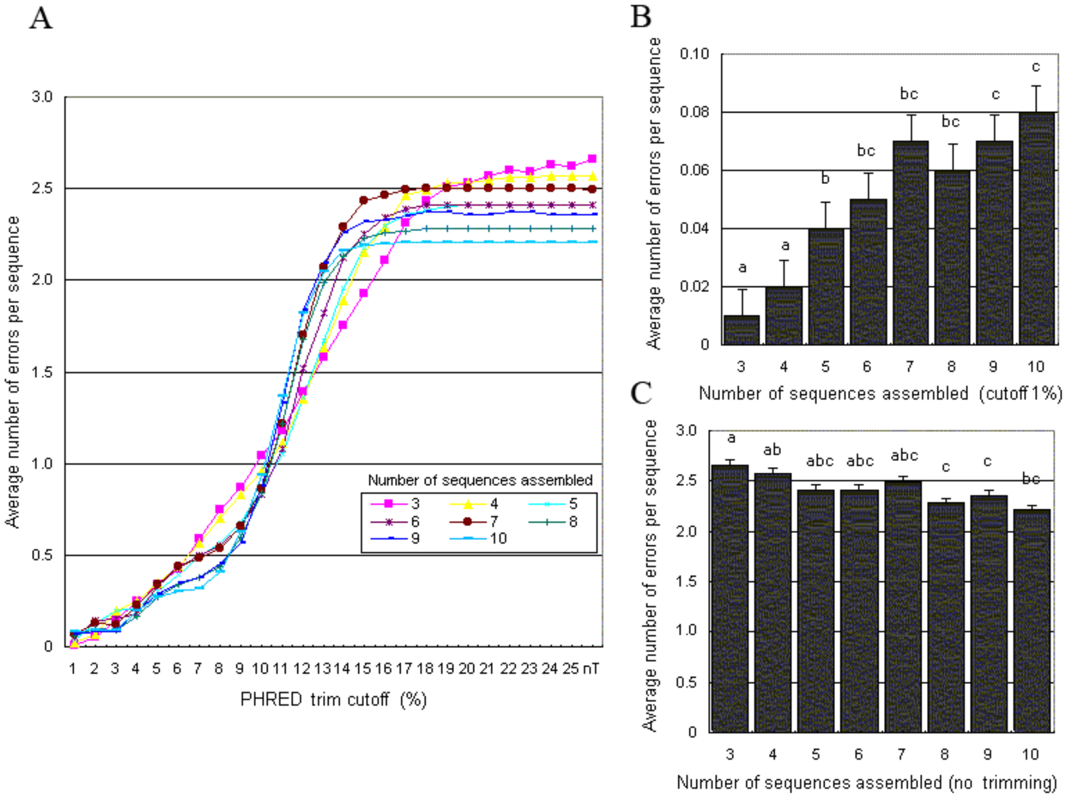

Figure 5. A. Average number of errors per molecule when different number of reads (2 to 10) were assembled with CAP3, trimmed for the regions containing at least three sequences and aligned to the published pUC18 sequence using BLAST software, sorted by PHRED trim cutoff percentage. **B.** Clusters of sequences trimmed with 1% cutoff (zoom). **C.** Clusters produced from non-trimming sequences (zoom).

DISCUSSION

The utilization of PHRED and CAP3 is common in both large- and small-scale genome-sequencing analysis. Some researchers have already addressed various aspects of the functioning of these algorithms (Ewing and Green, 1998; Ewing et al., 1998; Richterich, 1998; Huang and Madan, 1999; Chen and Skiena, 2000; Walther et al., 2001). However, as far as we know, this is the first detailed analysis of sample re-sequencing and trimming parameters on the quality and size of the assembled consensus. Although manual inspection is desirable, we have evaluated the automated procedure potential for providing to the operator qualified information about the expected occurrence of errors. This additional qualified information might be valuable, especially when inspecting 5' untranslated and N-terminal-coding regions of reads without significant similarity to deposited sequences.

We found that re-sequencing and assemblage of many reads (up to 10) do not reduce the average number of sequencing errors as much as might be expected (Figures 1 and 2). Stringent trimming procedures of reads have shown to be the best choice when the researcher aims to obtain a high-quality consensus sequence. However, the size of highly trimmed reads and their assemblage into consensus sequences are affected at the ranges shown in Figure 3. Consensus size reduction up to 40% was accompanied by a reduction of more than 10-fold in the number of errors per molecule due to trimming (PHRED 20), as compared to less than 10% gain in size and below 30% reduction in errors for non-trimmed (nT) reads by increasing the number of reads up to 10. This behavior might be restricted to the assemblage software used; a common alternative to CAP3 is Green's "phragment" assembly program (PHRAP; Green, 1998). We observed that consensus sequences assembled with PHRAP presented higher average number of errors than those produced by CAP3 (data not shown), as also found by other researchers (Huang and Madan, 1999), though overall results were similar.

The relatively small effect of re-sequencing on the average number of errors per sequence for non-trimmed reads continued, even when the type of error (mismatch or gap) was investigated (Figure 2A and B). However, the contribution of gaps seemed to count more than mismatches to error reduction when PHRED 20 trimmed reads were analyzed (Figures 1B and 2B). In a previous study, we had also found that mismatches are frequently associated with the lowest quality values while inserted bases often show higher quality values than mismatches (Prosdocimi et al., 2003). Thus, under stringent trimming cutoffs (e.g., PHRED 20), new strategies for consensus quality improvement should concentrate on diminishing the number of gaps.

The clipping of consensus regions formed by the assemblage of less than three overlapping reads produced sequences with smaller numbers of errors (Figure 5); this would be an easy and simple procedure to be implemented in future projects. Under these conditions, the effect of sample re-sequencing from 3 to 10 was found to be even less significant.

Thus, producing a large number of reads from the same molecule in a single direction, rather than eliminating consensus errors, is more efficient for enlarging the size of the resultant assembled sequence (around 33 and 10%, for PHRED 20 trimmed and non-trimmed reads, respectively; Figure 3B and C). The set of evaluation presented here provides data necessary for research groups to weigh the relative importance of automated consensus production as it affects size and quality. Inspection of Figures 1 and 3, can help choose the best PHRED trimming cutoff parameter and the number of reads to be produced and assembled; one can furthermore predict the expected average number of errors and size of the consensus sequences.

In general, high-quality sequences can be obtained with two reads (trimmed with PHRED 20) when size is not a constraint and the goal is to provide the operator with secure information about a specific portion of the read (e.g., when the correct translation start site is being investigated).

ACKNOWLEDGMENTS

The authors thank "Rede Genoma de Minas Gerais" (supported by FAPEMIG and MCT/Brazil) for providing a grant for much of this study.

REFERENCES

Adams MD, Kelley JM, Gocayne JD, Dubnick M, et al. (1991). Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* 252: 1651-1656.

Altschul SF, Madden TL, Schaffer AA, Zhang J, et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25: 3389-3402.

Chen T and Skiena SS (2000). A case study in genome-level fragment assembly. *Bioinformatics* 16: 494-500.

Ewing B and Green P (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* 8: 186-194.

Ewing B, Hillier L, Wendl MC and Green P (1998). Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* 8: 175-185.

Franco GR, Rabelo EM, Azevedo V, Pena HB, et al. (1997). Evaluation of cDNA libraries from different developmental stages of *Schistosoma mansoni* for production of expressed sequence tags (ESTs). *DNA Res.* 4: 231-240.

Gordon D, Abajian C and Green P (1998). Consed: a graphical tool for sequence finishing. *Genome Res.* 8: 195-202.

Green P (1998). Documentation for PHRAP and cross-match. http://www.phrap.org/phrap.docs/phrap.html. Accessed November 3, 2007.

Huang X and Madan A (1999). CAP3: A DNA sequence assembly program. *Genome Res.* 9: 868-877.

MGC (Mammalian Gene Collection) Program Team (2002). Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences. *PNAS* 99: 16899-16903.

Prosdocimi F and Ortega JM (2005). Accessing optimal primer distance from insert. *In Silico Biol.* 5: 469-477.

Prosdocimi F, Peixoto FC and Ortega JM (2003). DNA sequences base calling by PHRED: error pattern analysis. *R. Tecnol. Inf.* 3: 107-110.

Prosdocimi F, Peixoto FC and Ortega JM (2004). Evaluation of window cohabitation of DNA sequencing errors and lowest PHRED quality values. *Genet. Mol. Res.* 3: 483-492.

README for stand-alone BLAST. ftp://ftp.ncbi.nlm.nih.gov/blast/documents/blastall.html. Accessed January 1, 2006.

Richterich P (1998). Estimation of errors in "raw" DNA sequences: a validation study. *Genome Res.* 8: 251-259.

Strausberg RL, Feingold EA, Klausner RD and Collins FS (1999). The mammalian gene collection. *Science* 286: 455-457.

Walther D, Bartha G and Morris M (2001). Basecalling with life trace. *Genome Res.* 11: 875-888.