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

Francisco Prosdocimi[1], Denize Altiva de Oliveira Lopes[2], Fabiano Cruz Peixoto[3], José Miguel Ortega[2]*

1- Laboratório de Biodiversidade e Evolução molecular. Depto. Biologia Geral, ICB-UFMG
2- Laboratório de Biodados. Depto. Bioquímica e Imunologia, ICB-UFMG
3- Laboratório de Computação Científica, UFMG

José Miguel Ortega *
miguel@icb.ufmg.br
Laboratório de Biodados. Sala N4-202.
Departamento de Bioquímica e Imunologia, ICB, UFMG
Av. Antônio Carlos, 6627 C.P. 486
31.270-010 Belo Horizonte, MG, Brazil
Tel: +55 31 3499-2654
Fax: +55 31 3499-2570

Running Head
Number of reads and consensus quality

* To whom correspondence should be addressed

SUMMARY
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 are used to sequence the same sample many times. Therefore, considering base-calling errors as rare events, the re-sequencing same molecule and the further assembling of reads is 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 the costs and achieve this high-fidelity sequence? In the present work, both the effect of the number of reads and PHRED trimming parameters were observed in order to verify the accuracy and size of the final consensus sequence. Hundreds of single-pool reaction pUC18 reads were generated and the assembled into consensus with CAP3 software. Using local alignment against the pUC18 cloning vector published sequence, the position and number of errors in the consensus were identified and stored in MySQL databases. We verified that stringent PHRED trimming parameters efficiently reduces the number of errors, although it also reduces the size of the final consensus. Moreover we observed the poor effect of re-sequencing on reducing the number of errors, although this procedure was capable to enlarge slightly the consensus size.

Keywords: Sequencing reads, trimming, assembling, consensus, CDS, PHRED, CAP3.

Introduction

Most of the recent developments in the field of genomics and bioinformatics dealt with data generated from genome sequencing projects and it is well-known that all genomes are build *in silico* by the superposition of thousands of overlapping reads joined together by assembly software such as PHRAP (Green, 1999) or CAP3 (Huang and Madan, 1999). Many assembly software, including the two cited here, take advantage of base quality values determined by base-caller algorithms such PHRED (Green and Ewing, 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 software are also used to cluster EST (Expressed Sequence Tag) data, in this case focusing 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 this consists in the fact that assembled molecules from genome projects are allowed to enter in genome databases whilst assembled ESTs are restricted to specific project websites and they are not allowed to be integrated in any of the very best known public molecular databases. Nevertheless, not so rare is the evolution of an EST project to a full-length cDNA sequencing project, such as the Mammalian Gene Collection (Strausberg *et al*., 1999; MGC Program Team, 2002), where selected clones are introduced into a pipeline of dedicated sequencing to eliminate any ambiguities from the reads generating a consensus sequence. The edited sequences can then be deposited into databases distinct from dbEST (Boguski *et al*., 1993), such as GenBank and GenPept, becoming therefore 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 possible alternative is the manual editing of the consensus with software such as Consed (Gordon *et al*., 1998), a procedure that shall be encouraged instead of any automated alternatives, although the 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 and (ii) the amount of errors admitted during trimming procedures. Usually, for genome projects, trimming seems to not be recommended because high quality regions are often overlapping low quality ones. However, this is clearly not the situation in partial sequencing of cDNA molecules, since all reads are expected to start at the same position and, most important, the low quality regions are concentrated in the edges of the sequences.

In this work we report the analysis of sample re-sequencing (from 2 up to 10 times) 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). Data obtained indicated that trimming efficiently reduces the number of errors but affects the size of the consensus, while the impact of having a large number of reads is not as remarkable as it could intuitively be expected.

Methodology

Sequencing reactions

The sequencing reaction premix was made in a single pool and divided on tubes for the PCR sequencing reaction. After that, products were joined together on the same tube, mixed, and sequenced in 96-well plates with MegaBACE equipment. Three laboratories from the Federal University of Minas Gerais (UFMG) that integrate the Minas Gerais Genome Network provided the 846 processed ESD files used in this work.

Base calling and trimming

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

Sequence assembly

From the 846 ESD files, 1,000 groups of two sequences were randomly taken and assembled with CAP3 software. The same procedure was done for groups of 3, 4, 5, 6, 7, 8, 9 and 10 sequences. Therefore, we have done 9,000 sequence group draws and assemblage.

Local alignment against pUC18 published sequence

All the generated consensus sequences were compared to the pUC18 published 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 of data

Since the data did not fit normal distribution, non-parametric ANOVA statistical tests were performed. So, we have run Kruskal-Wallis median tests to analyze the number of errors and size of the consensus generated when using trimming cutoff parameter at 1% or not trimmed sequences.

Results

Here, the efficiency of re-sequencing on the production of error-free consensus was evaluated by sampling thousands of groups containing two up to ten reads from a collection of 846 reads of the pUC18 cloning vector. Reads were base called with PHRED software and assembled with CAP3. Usually, during PHRED processing of data, no trimming of the low quality portion of the reads is performed (denoted by nT - no trimming - in figures). By aligning the 9,000 consensus produced with the published pUC18 sequence using BLASTn program, the errors in these *in silico* sequences (sometimes called contigs) were evaluated. It is noteworthy that BLASTn alignments do not elongate over the low quality portion of the reads, therefore errors per sequence tend to a maximum.

We decided to include additional data applying a trimming cutoff with PHRED internal algorithm - *trim_alt*, 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 amount of errors in consensus sequences.

Figure 1a presents the average number of errors per consensus sequences assembled by CAP3 and accessed with BLASTn, showing that trimming can reduce errors to less than one per sequence. Re-sequencing (increasing from two up to ten reads) was expected to reduce significantly the number of errors from the consensus assembled but, in fact, the reduction was not as significant as one might suppose. For a detailed analysis two regions of the curves where the differences between data were either maximized or minimized (trimming cutoff of 1% or nT). These data were also statistically analyzed and they are shown in Figures 1b and 1c. For the consensus generated from nT reads, the best results were observed with the assembling of ten reads, although the cost/benefit over the use of three reads is clearly higher. When reads were trimmed with cutoff 0.01 (1%), the effect of increasing the number of reads from 2 to 10 was a 4.3 fold reduction on errors per molecule (up to 24% of the initial amount, figure 1b). However, with no trimming (figure 1c) the reduction was of 1.5 fold (64% of the initial amount remaining) and not significant from 3 up to 8 reads. Thus, trimming most efficiently decreases the errors while maintaining the highest responsiveness to the increase on the number of reads.

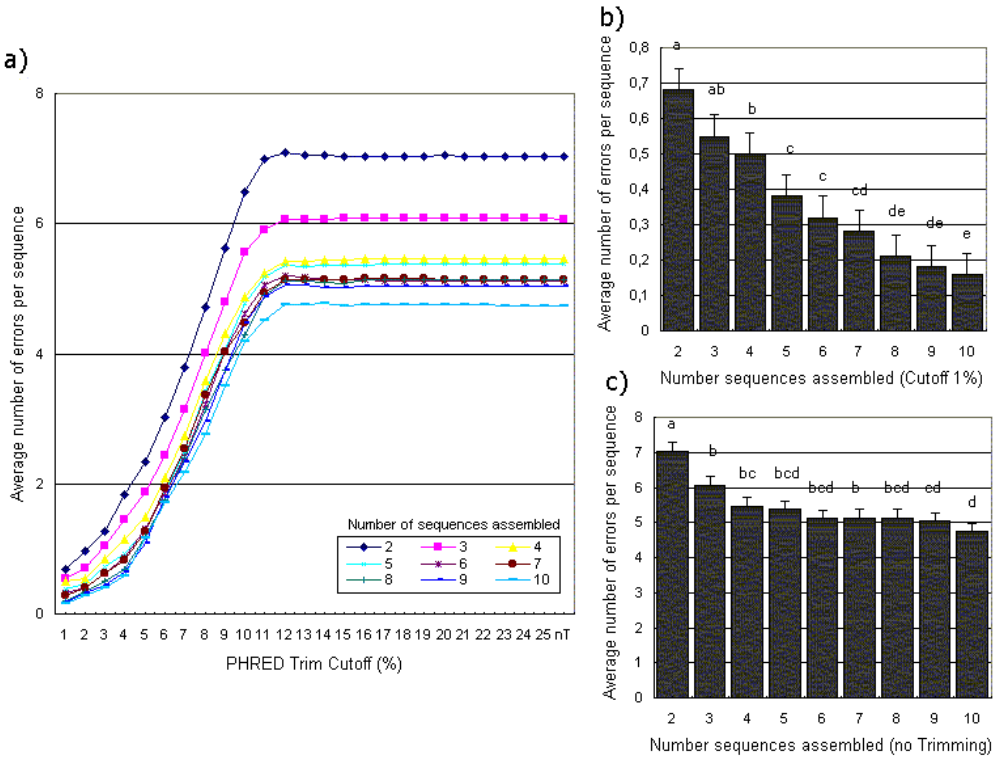

Figure 1 – a) Average number of errors per sequence when different number of reads (2 to 10) were assembled with CAP3 and aligned with pUC18 published sequence using BLAST software, sorted by PHRED trim cutoff percentage. b) and c) Zoom for most interesting data.

The surprising effect of increasing the number of reads on molecules trimmed under cutoff of 1%, which corresponds to PHRED 20, as opposed to the low effect on not trimmed reads, what is recommended for assembly, lead us to investigate the nature of these errors. Data presented on figure 2a show that at 1% cutoff, mismatches were minimum even when using only two reads. However, from 10% cutoff up to no trimming (nT), the number of mismatches decreased in similar proportion as for total errors as more reads were assembled (compare figures 2a and 1a). In contrast, when we analyzed the gaps, the opposite was observed: under PHRED 20, gaps were efficiently reduced as the number of reads increases, but this was not observed for non trimmed or poorly trimmed reads (figure 2b). This last observation is concordant with data showing that high-quality errors are mainly those generated by the insertion (Prosdocimi *et. al.*, 2003), thus producing gaps on the alignment. Therefore, the effect of efficient reduction in the number of reads under PHRED 20 is due to the decrease in the number of gaps and the proportion of decrease of the number of either mismatches or gaps under PHRED 10 up to no trimming is rather similar and low. Curiously, the number of gaps is minimum under 4% cutoff for 2 reads or 2% cutoff for 10 reads.

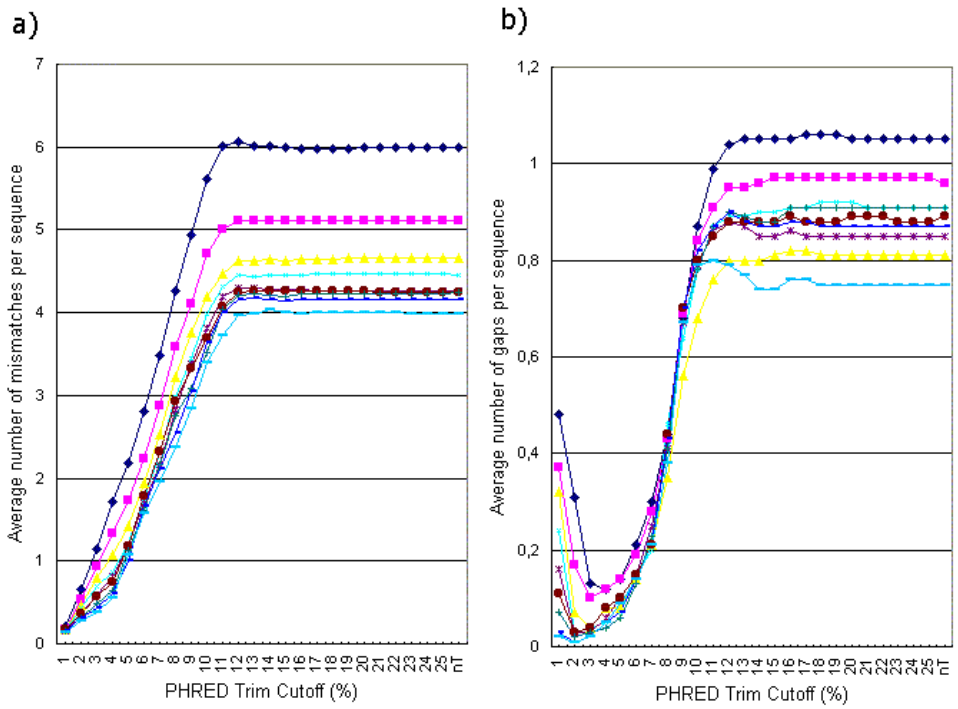

Figure 2 – a) Average number of mismatches per sequence when different number of reads (2 to 10) were assembled with CAP3 and aligned with pUC18 published sequence using BLAST software, sorted by PHRED trim cutoff percentage. b) Average number of gaps.

Although under PHRED 10 up to no trimming the use of more than two reads did not significantly improved the quality of the obtained sequences, trimming in this range as opposed to 1% cutoff increased the size of the consensus, thus producing a realistic benefit (Figure 3a). Resultant assembled consensus sequences were greater than 500 bp. Moreover, data presented in Figures 3b and 3c depict that consensus size is more responsive to the number of reads under PHRED 20 (1% cutoff, Figure 3b) than when using non trimmed reads (Figure 3c).

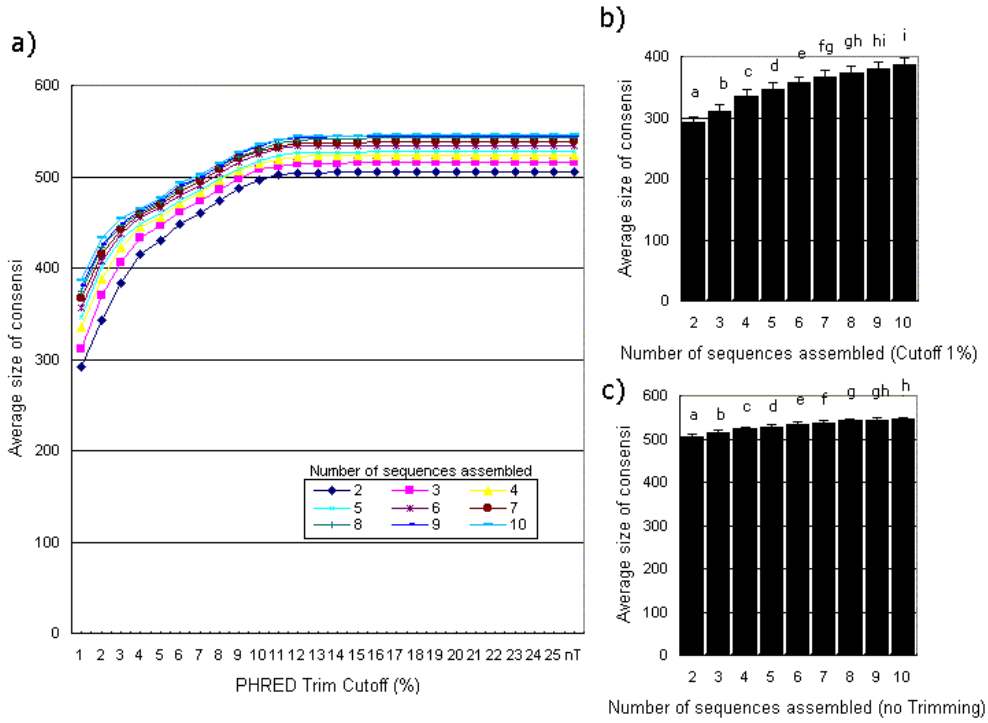

Figure 3 – a) Average size of consensi when different number of reads (2 to 10) were assembled with CAP3, sorted by PHRED trim cutoff percentage. b) and c) Zoom for most interesting data.

We considered the fact that all reads start at the primer and progressively loose quality as they proceed away from the starting position. This might result in situations where a poor-quality edge of a single read stands for the quality of the consensus, even if ten 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 individuals 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.

Analyses of the number of errors in these trimmed consensus sequences are presented in Figures 5a to 5c. The maximum number of errors per sequence diminishes from around 6 (Figure 1a) to up to 2.5 (Figure 5a). Again, increasing the number of used reads from three up to ten produced a small effect on the number of errors per consensus when non-trimmed individual reads (nT) were used (figure 5c). Intriguingly, the use of more than four reads raised the number of errors when using PHRED 20 cutoff

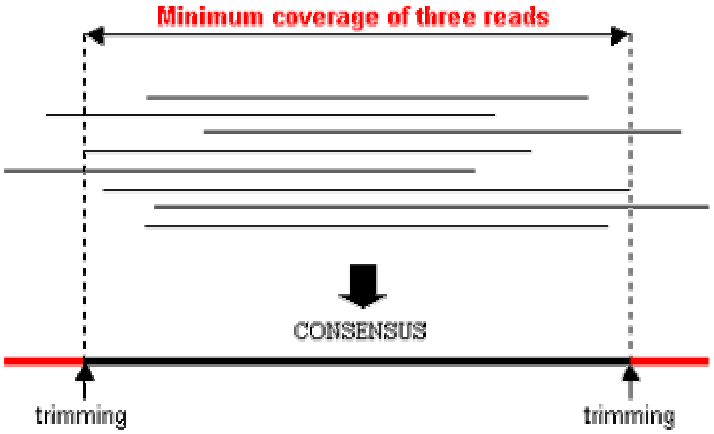

Figure 4 – Methodology for consensi trimming.

(Figure 3b) in these 3-reads coverage consensus sequences. The total amount of errors in nT sequences, as compared to the simpler procedure of assembling the reads without taking on account the number of overlapping sequences (Figure 1a), is reduced by around 50% (from 5-7 to up to 2.5 errors per molecule), what it is still lower than the effect of trimming the reads with higher values of such as PHRED 20.

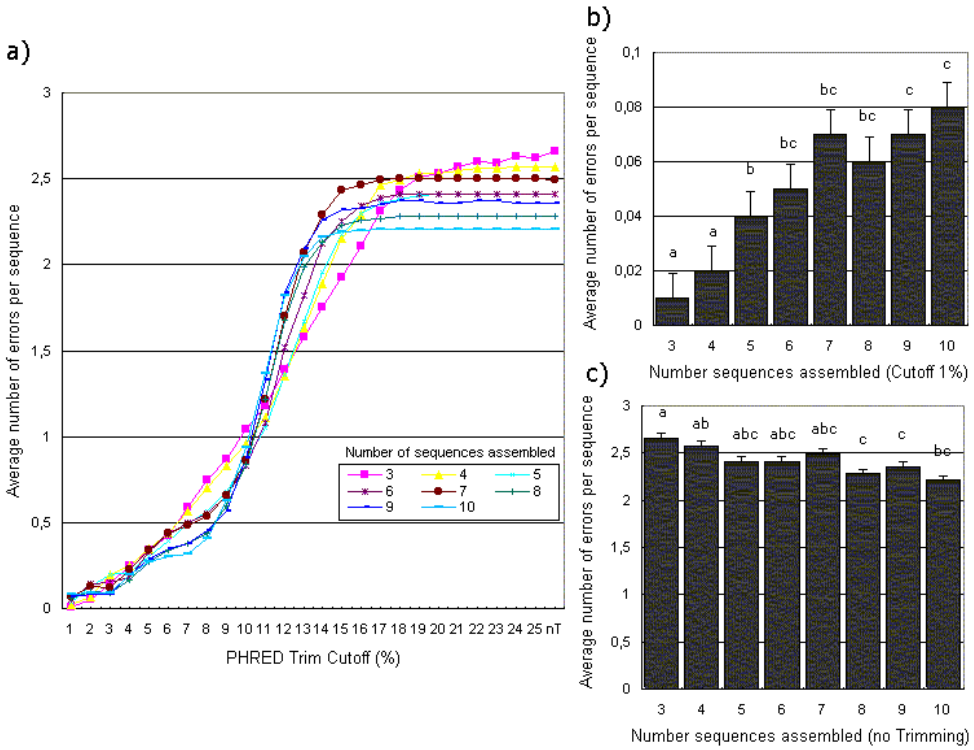

Figure 5 – a) Average number of errors per molecule when different number of reads (2 to 10) were assembled with CAP3, trimmed for the regions containing at least 3 sequences and aligned to pUC18 published sequence using BLAST software, sorted by PHRED trim cutoff percentage. b) and c) Zoom for most interesting data.

Discussion

Application of PHRED base caller and the assembler software CAP3 is common on both large-scale genome analysis and low scale sequencing as exemplified here. Some works have already addressed issues about their functioning (Ewing and Green, 1998; Ewing *et al*., 1998; Richterich, 1998; Huang and Madan, 1999; Schen and Skiena, 2000; Walther *et al*., 2001), but as long 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, here we evaluated the potential of automated procedures on giving the operator qualified information about the expected occurrence of errors. That might be valuable while inspecting 5'UTRs and N-terminal coding region without significant similarity to deposited sequences.

Our results show that the assemblage of large amount of reads (up to ten) does not reduce the average number of errors at the intensity that it might be possibly expected (Figures 1 and 2). It was evidenced that trimming procedures previous to the assemblage are the best choice when the aim is to obtain a high-quality sequence. However, the resultant sizes of these sequences are affected by stringent PHRED trimming cutoff parameters at the ranges shown in Figure 3. In our experiments, size reduction up to 40% was accompanied by a reduction of over 10 folds in errors per molecule due to trimming (PHRED 20), as compared to less than 10% gain in size and below 30% reduction of errors for non trimmed (nT) reads by increasing the number of reads up to ten.

This behavior could be restricted to the assembling software used. An alternative to CAP3 is the software PHRAP. We observed that consensus sequences assembled with PHRAP presented higher average number of errors than those produced by CAP3 (data not shown), in concordance with other published data (Huang and Madan, 1999), though the results obtained were very similar.

The poor effect of re-sequencing on the average number of errors per sequence for non-trimmed reads was very equivalent when the type of error was investigated (Figures 2a and 2b), although the contribution of gaps seemed to count for the most of the reduction for PHRED 20 trimmed reads (Figures 2b and 1b). This observation is in agreement with our previous work that evidenced 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 cutoff (e.g. PHRED 20), due to the introduction of gaps by the assembler software, the expected improvement by using more reads shall concentrate on diminishing the occurrence of gaps.

The clipping of consensus regions formed by the assemblage of less than three overlapping reads produced sequences with less number of errors (Figure 5), suggesting a procedure that is possible to be implemented. Even under this treatment of the set of reads, the effect of sample re-sequencing from 3 to 10 times was even less significant.

Thus, we conclude that the production of a large number of reads from the same molecule in a single direction, rather than eliminate consensus errors, is more efficient to enlarge the size of the produced sequence (around 33% and 10%, for PHRED 20 trimmed and non-trimmed reads, respectively, Figures 3b and 3c). The set of evaluation presented here provide the data necessary for research groups to balance between the size of the automated certified sequences and the quality of the generated consensus sequences. With a brief inspection of Figures 1 and 3, it is possible to choose the best PHRED trimming cutoff parameter and the number of reads to be assembled and furthermore to predict the expected average number of errors and size of the resultant consensus sequences.

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

Acknowledgements

The authors wish to thank the "Rede Genoma de Minas Gerais" (supported by FAPEMIG and MCT/Brazil), especially Marina Mourao, Lucila Pacifico and Renata Ribeiro, for providing the sequences used in the analysis.

References

1. **Adams, M.D., Kelley, J.M., Gocayne, J.D., Dubnick, M., Polymeropoulos, M.H., Xiao, H., Merril, C.R., Wu, A., Olde, B., Moreno, B.F., Kerlavage, A.R., McCombie, W.R.** and **Venter, J.C.** (1991) Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* 252: 1651-1656.

2. **Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W.** and **Lipman, D.J.** (1997). Gapped BLAST, PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25, 3389-3402.

3. **Boguski, M.S., Lowe, T.M.** and **Tolstoshey, C.M.** (1993). dbEST- database for "expressed sequence tags". *Nat Genet* 4:332-3

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

5. **Ewing, B., Hillier, L., Wendl, M.C.** and **Green P**. (1998) Base-Calling of automated sequencer traces using phred. I. Accuracy Assessment. *Genome Res* 8: 175-185.

6. **Franco, G.R., Rabelo, E.M., Azevedo, V., Pena, H.B., Ortega, J.M., Santos, T.M., Meira, W.S., Rodrigues, N.A., Dias, C.M., Harrop, R., Wilson, A., Saber, M., Abdel-Hamid, H., Faria, M.S., Margutti, M.E., Parra, J.C.** and **Pena, S.D.** (1997). Evaluation of cDNA libraries from different developmental stages of *Schistosoma mansoni* for production of expressed sequence tags (ESTs). *DNA Res* 4: 231-240.

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

8. **Green, P**. (1998) Documentation for PHRAP and cross-match. http://www.phrap.org/phrap.docs/phrap.html

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

10. **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

11. **Prosdocimi, F., Peixoto, F.C.** and **Ortega, J.M.** (2003) DNA Sequences Base Calling by PHRED: Error Pattern Analysis. *R Tecnol Inf* 3: 107-110.

12. **Prosdocimi, F., Peixoto, F.C.** and **Ortega, J.M.** (2004) Evaluation of window cohabitation of DNA sequencing errors and lowest PHRED quality values. *Gen Mol Res* 3:483-492.
13. **README for stand-alone BLAST**. ftp://ftp.ncbi.nlm.nih.gov/blast/documents/blast.txt
14. **Richterich, P**. (1998) Estimation of errors in "raw" DNA sequences: a validation study. *Genome Res*. 8:251-259.
15. **Chen, T.** and **Skiena, S.S.** (2000) A case study in genome-level fragment assembly. Bioinformatics. 16:494-500.
16. **Strausberg, R.L., Feingold, E.A., Klausner, R.D.** and **Collins, F.S.** (1999) The Mammalian Gene Collection. *Science* 286: 455-457.
17. **Walther, D., Bartha, G.** and **Morris, M.** (2001) Basecalling with LifeTrace. *Genome Res*. 11:875-888.