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

Evaluation of window cohabitation of DNA sequencing errors and lowest PHRED quality values

Francisco Prosdocimi[1], Fabiano Cruz Peixoto[2] and José Miguel Ortega[3]

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

Genet. Mol. Res. 3 (4): 483-492 (2004)
Received October 4, 2004
Accepted December 6, 2004
Published December 30, 2004

ABSTRACT. When analyzing sequencing reads, it is important to distinguish between putative correct and wrong bases. An open question is how a PHRED quality value is capable of identifying the miscalled bases and if there is a quality cutoff that allows mapping of most errors. Considering the fact that a low quality value does not necessarily indicate a miscalled position, we decided to investigate if window-based analyses of quality values might better predict errors. There are many reasons to look for a perfect window in DNA sequences, such as when using SAGE technique, looking for BLAST seeding and clustering sequences. Thus, we set out to find a quality cutoff value that would distinguish non-perfect windows from perfect ones. We produced and compared 846 reads of pUC18 with the published pUC consensus, by local alignment. We then generated a database containing all mismatches, insertions and gaps in order to map real perfect windows. An investigation was made to find the potential to predict perfect windows when all bases in the window show quality values over a given cutoff. We conclude that, in window-based applications, a PHRED quality value cutoff of 7 masks most of the errors without masking real correct windows. We suggest that the putative wrong bases be indicated in lower case, increasing the information on the sequence databases without increasing the size the files.

Key words: DNA sequence quality, PHRED, Quality window, SAGE, BLAST

INTRODUCTION

Base caller algorithms are as important as sequencing machines for the identification of the sequence of bases in DNA molecules. They are responsible for the analysis of the raw data generated by the sequencing equipment and for the production of the sequence of bases putatively related to the original molecule, as well as the quality values determined for each of them (Prosdocimi et al., 2002). The best-known and most widely used base caller algorithm is PHRED, written by Green and Ewing (Ewing et al., 1998; Ewing and Green, 1998). An approach frequently used by researchers looking for miscalled bases in DNA sequences is the choosing of a minimum quality value based on intuition, considering the significance of the PHRED quality value (PQV). PQV 20 is the most widely used, and operationally it means that a base has one chance in a hundred to be miscalled. However, a low quality value does not necessarily cohabit with a miscalled position (Ewing and Green, 1998; Prosdocimi et al., 2003).

Beyond the use of a quality cutoff for single bases, many applications can make use of the quality value for a number of bases in tandem, or a window of bases. There are many reasons for researchers to look for a perfect window (PW) in a DNA sequence, defined as a sequence of called bases that putatively do not contain any mismatch or gap (insertion/deletion). This PW is particularly important in the SAGE technique, which consists of single pass sequencing of concatenated fragments of the cDNA tail subsequent to a given restriction site (Velculescu et al., 1995). The bases juxtaposed to the restriction site constitute a tag that has been assigned to genes. One single error on a SAGE tag (containing 14 nucleotides) can generate incorrect associations and false positives (and negatives) in the gene expression inference. Thus, it is quite important to be able to establish an appropriate quality cutoff, under which a window lacks, probabilistically, the potential to be entirely correct, reducing the number of false inferences.

BLAST is another application that could take advantage of PW; it is possible to choose only the perfect windows to be used as a BLAST seeding window (Altschul et al., 1997). In BLAST execution, if one of the letters in the sequence is represented by lower case, it is possible to avoid seeding on them, using, in the stand-alone version, the flag - UT (see README in documentation for stand-alone BLAST). Thus, the alignments will only seed on uppercase PWs, since putatively incorrectly called bases are represented in lower case.

In order to evaluate if the lowest PQV could correctly mask non-perfect windows, we analyzed 846 single-pool reads of pUC18. Aligning the reads to the published sequence for this cloning vector, a database of all mismatches, insertions and gaps generated by the entire sequencing procedure was built. Different window sizes were tested in order to find the best fit between real perfect windows (RPWs) and predicted perfect windows (PPWs), the ones not containing a PQV equal to or below the chosen cutoff. We also evaluated which PQV cutoff showed the best potential to identify the position of sequencing errors without masking, or spoiling, correct windows, so that it could be used in various applications.

MATERIAL AND METHODS

Sequencing reactions

Three laboratories from the Universidade Federal de Minas Gerais (UFMG), which together make up the Rede Genoma de Minas Gerais network, provided the sequences. The reactions were made in a single pool and divided into tubes for the PCR sequencing reaction. After the PCR sequencing reaction, the sequences were joined again in the same tube, mixed, and then divided on three 96-well sequencing plates. Each plate was run three times on a MegaBACE sequencing equipment, yielding a total of 864 reads. Eight hundred and forty-six processed ESD files were obtained.

Base calling

All ESD files were processed by PHRED, without trimming, and a total of 840,134 bases were called.

Local alignment against the pUC18 published sequence

All the sequences generated were compared to the published pUC18 sequence (24.8% A, 25.2% C, 25.5% G, 24.5% T) using the local alignment algorithm SWAT (Smith and Waterman, 1981). Parser scripts written in PERL were built to populate MySQL tables with the position of errors in the reads, identified through the differences in the alignment results. The SWAT algorithm was run with the DNA matrix mat70, and 156,301 bases were removed from the analysis, since they did not show valid alignment to the pUC18 published sequence. The number of bases removed was similar to what was obtained with a PHRED trimming procedure using a trim cutoff parameter of 0.16 (data not shown).

Window-based analysis

RPW and PPW were defined for different window lengths, in order that they could be used in various applications. Table 1 lists the applications and their respective default window length. The PPW were compared to the RPW ones to identify which PQV cutoff (from 5 up to 15) should be used to mask the majority of the errors without masking (and then spoiling) the correct windows.

Error-main weighted analysis

Some researchers might choose to preferentially mask the real errors, even if this is coupled with undesirable masking of correct windows (spoiled windows). Taking this point into consideration, an index called weighted correctness (WC) was created. There are two types of incorrectly classified windows: the ones containing errors that were not masked (not masked windows, NMW) and the ones with no errors but which were masked because all their bases were under a certain PQV cutoff (spoiled windows, SW). WC will relate and weight NMW and SW according to the researcher's choice. Considering PSW as the percentage of SW divided

Table 1. Window sizes, which were analyzed, and related applications.

Window (number of bases)	Acronym	Application
1	Win1	Base calling
6	Win6	Restriction site at vector clipping
11	Win11	BLASTn seeding step
14	Win14	SAGE
28	Win28	MEGABLAST seeding step
40	Win40	Unigene clustering

by the total percentage of windows classified as wrong and PNW as the percentage of NMW divided by the total percentage of windows classified as correct, we can calculate WC as indicated below. The WC value is therefore a measure of the number of errors (NMW and SW) with weights associated with each type of error (Weight1 for NMW and Weight2 for SW).

$$\mathrm{WC} = 100 - \left(\frac{(\mathrm{Weight1} \times \mathrm{PNW}) + (\mathrm{Weight2} \times \mathrm{PSW})}{\mathrm{PNW} + \mathrm{PSW}} \right)$$

RESULTS

Data characterization

Analyzing Figure 1, it is possible to see that some PQV are preferred over others during base calling by PHRED, and the number of called bases for each PQV does not show a clear decay as the quality value increases.

The percentage of RPW, the ones with no sequencing errors when aligning the reads to the pUC18 published sequence, was counted for each window size (Table 2). The inverse correlation between the window size and the number of RPWs was expected, since a larger window is likely to shelter more errors than a smaller one.

Table 2. Proportion of real perfect windows (RPW) for each window size.

Window	% RPW
Win1	96
Win6	82
Win11	74
Win14	71
Win28	61
Win40	56

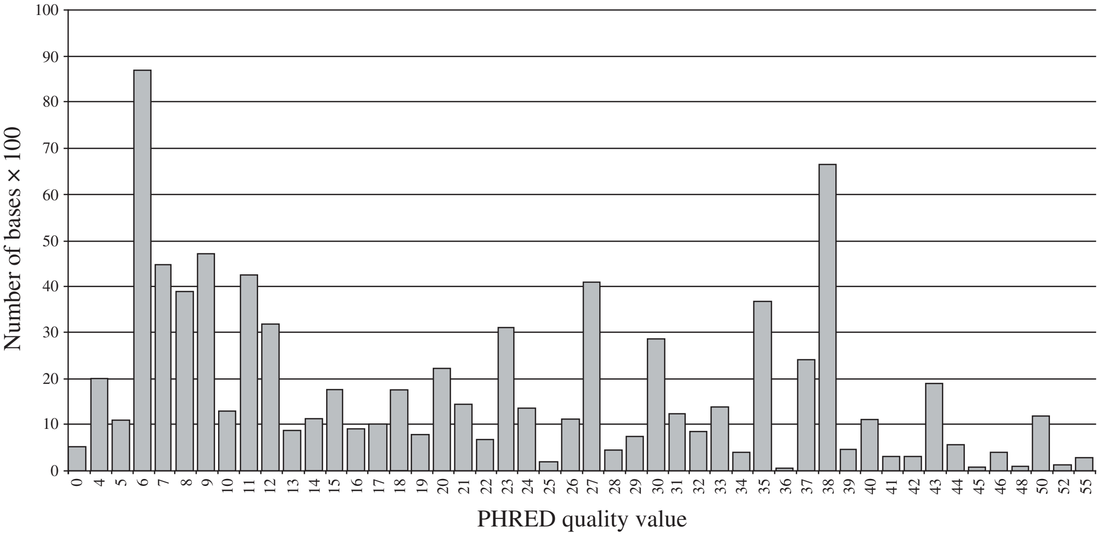


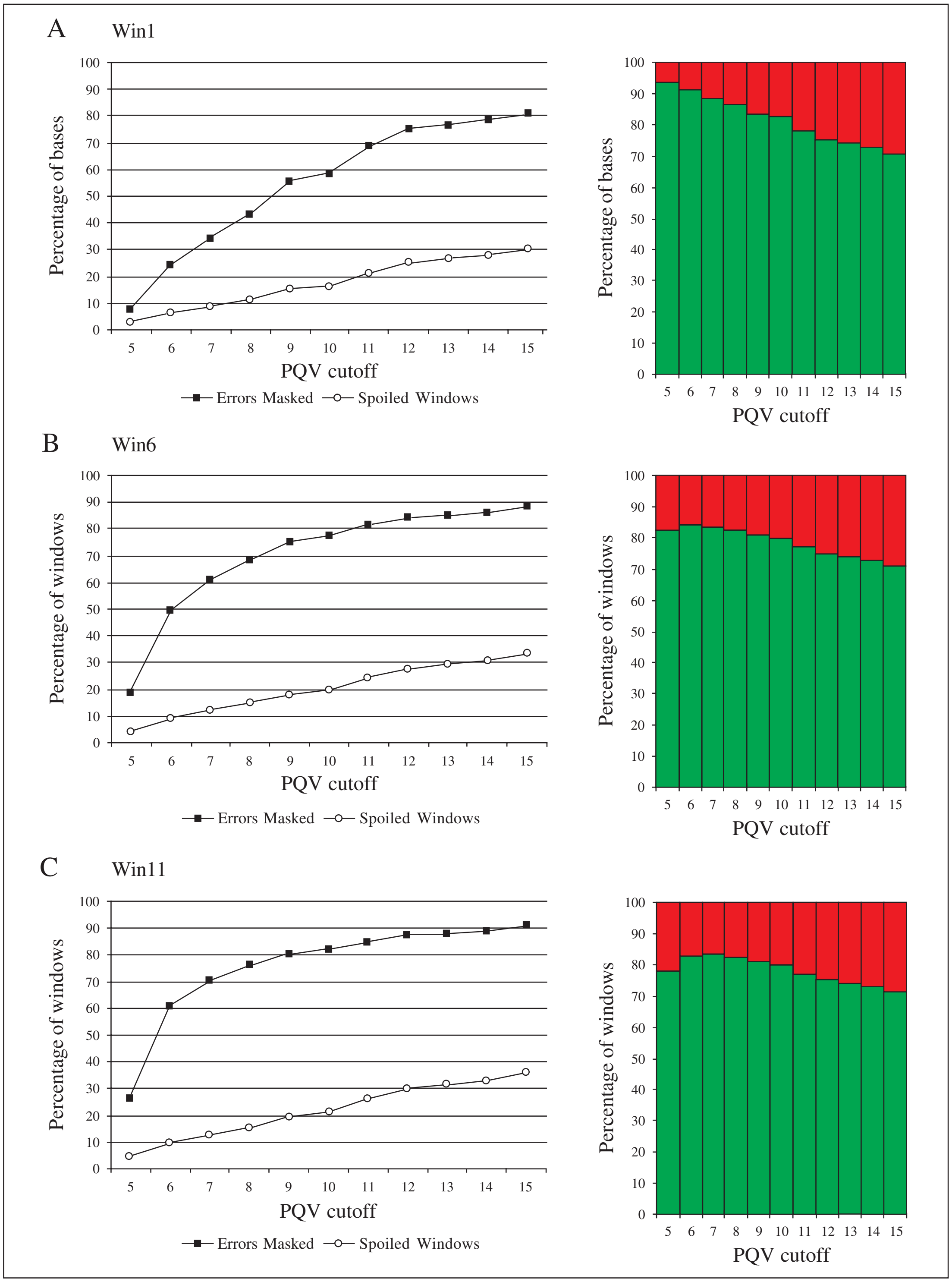
Figure 1. Number of bases called under each PHRED quality value.

PQV cutoff for window sizes

Considering the nature of the PQV, it becomes clear that, when raising the value of the PHRED quality cutoff by which bases are represented with lower case, an increasing number of sequencing windows will be masked, correctly or incorrectly. Here we defined windows containing at least one incorrect base as "errors masked" and windows containing only correct bases that also contain at least one PQV under the cutoff as "spoiled windows". Figure 2 shows the correlation between errors masked and spoiled windows when increasing the PQV cutoff. The graphs on the left side show the percentage of "errors masked" in filled squares and the percentage of "spoiled windows" in empty circles. On the right side of figure we also show the percentage of total window classification that was correct (green) or incorrect (red).

Our main purpose was to find the most adequate PQV cutoff that maximizes the number of windows that do contain errors and are masked by the lowest PQV (errors masked), while minimizing the number of correct windows incorrectly masked (spoiled windows). This PQV cutoff may be used to define unreliable lower case-containing windows. It is clear from Figure 2 that the percentage of errors that are masked tends to saturate, while the percentage of spoiled windows continues to raise. Moreover, it is possible to choose from the data plot a given PQV cutoff for the lower case representation that will mask over 80% of the windows containing errors without spoiling more than 20% of the correct windows. Furthermore, this relationship depends on the size of the working window (e.g., for the Win40, a PQV cutoff of 8 will mask 90% of error-containing windows while avoiding 20% of correct windows to be used by the application).

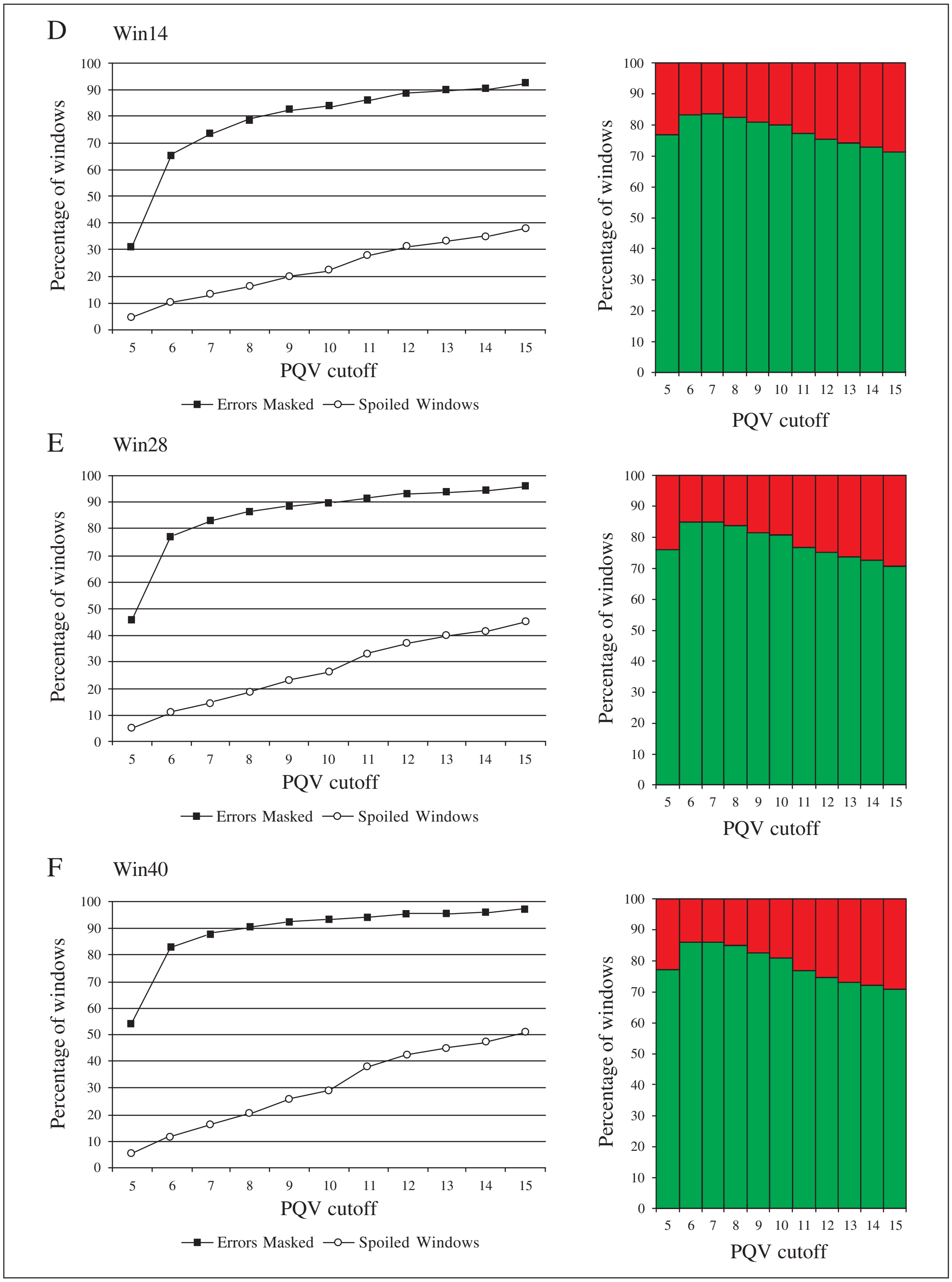
Another way of looking at the same data is the balance between correct and incorrect classification. Incorrect classification involves a RPW classified as wrong or an error-containing window classified as a PPW. These classifications are represented on the right side of Figure 2.



Continued on next page

Figure 2. Percentage of errors masked versus spoiled windows (left) and percentage of correct (green) and incorrect (red) window classification (right) for Win1, 6, 11, 14, 28, and 40. PQV = PHRED quality value.

Figure 2. Continued.



In general, a PQV cutoff of 6 or 7 maximizes the identification of errors, without spoiling a large proportion of correct windows. This best correlation is clear if one looks at the high inclination of the errors masked curve when the PQV cutoff increases from 5 to 6 and from 6 to 7. These inclinations are higher than all other increases of PQV cutoff, and they are much more accentuated than the ones observed for the spoiled window curves at the same cutoffs. However, in general, the PQV cutoff of 6 or 7 still gives a window classification error percentage close to 15% for windows larger than win1.

Error main weighted analysis

All the results for the last section were shown considering that an error that is not masked has the same importance (50-50%) as a correct window that has been incorrectly masked. However, one could argue that it is more relevant to mask the real errors, even if this has been coupled with a high number of spoiled correct windows. Thus, a weighted index WC was developed to obtain the data for many distinct relationships between the weights attributed to NMW and SW (see Methodology).

Data in Figure 3A is similar to the data shown in Figure 2 (green-red graphs). By analyzing the other plots (B, C and D), it is possible to see that the best value of PQV cutoff has been shifted to the right, as expected, when increasing the percentage of priority in error masking. In this way, the best quality cutoff tends to be higher than 7.

DISCUSSION

PHRED software is the most widely used base caller in the genomics field, and its use has been extensively evaluated (Ewing and Green, 1998; Ewing et al., 1998; Richterich, 1998; Walther et al., 2001; Scheetz et al., 2003). Besides being well known that the lowest PQV does not necessarily stand for miscalled bases, the question of whether or not the lowest PQV cohabit in the same sequencing window with the sequencing errors has not been addressed. Our main purpose was to map the miscalled positions and the RPWs on sequencing reads based on a PQV cutoff. With a specific cutoff, e.g., PHRED 15, we decided to represent putatively miscalled bases with lower case letters, so that those with quality values less than or equal to 15, would be considered as unreliable. The advantage of translating bases to lower case letters is that, with the same number of bytes used to store standard sequences, we could add relevant information in an easy-to-see fashion. It would be particularly useful for the publication of single-passed sequences, such as ESTs, GSSs and SAGE tags. Currently, some researchers apply this procedure, but no study has been conducted to evaluate the appropriate cutoff.

Contrary to what was expected by intuition, the PQV cutoff was found to be more effective in masking errors without spoiling PWs at a quality value of 6 to 7. Based on the data that were collected, we would recommend the lower case masking of windows containing at least one base with a quality value of 7 or less, for most of the applications exemplified. However, if the researcher prefers to correctly mask all (or almost all) errors, even with a great chance of increasing the number of spoiled windows, he should use a higher cutoff value, varying from 9 up to 15 (see Figure 3 for a better understanding of this). Although these recommendations are valid for all the data, it is useful to inspect Figure 2, trying to fit the best PQV cutoff for a particular application, or for the most probable destination of the output.

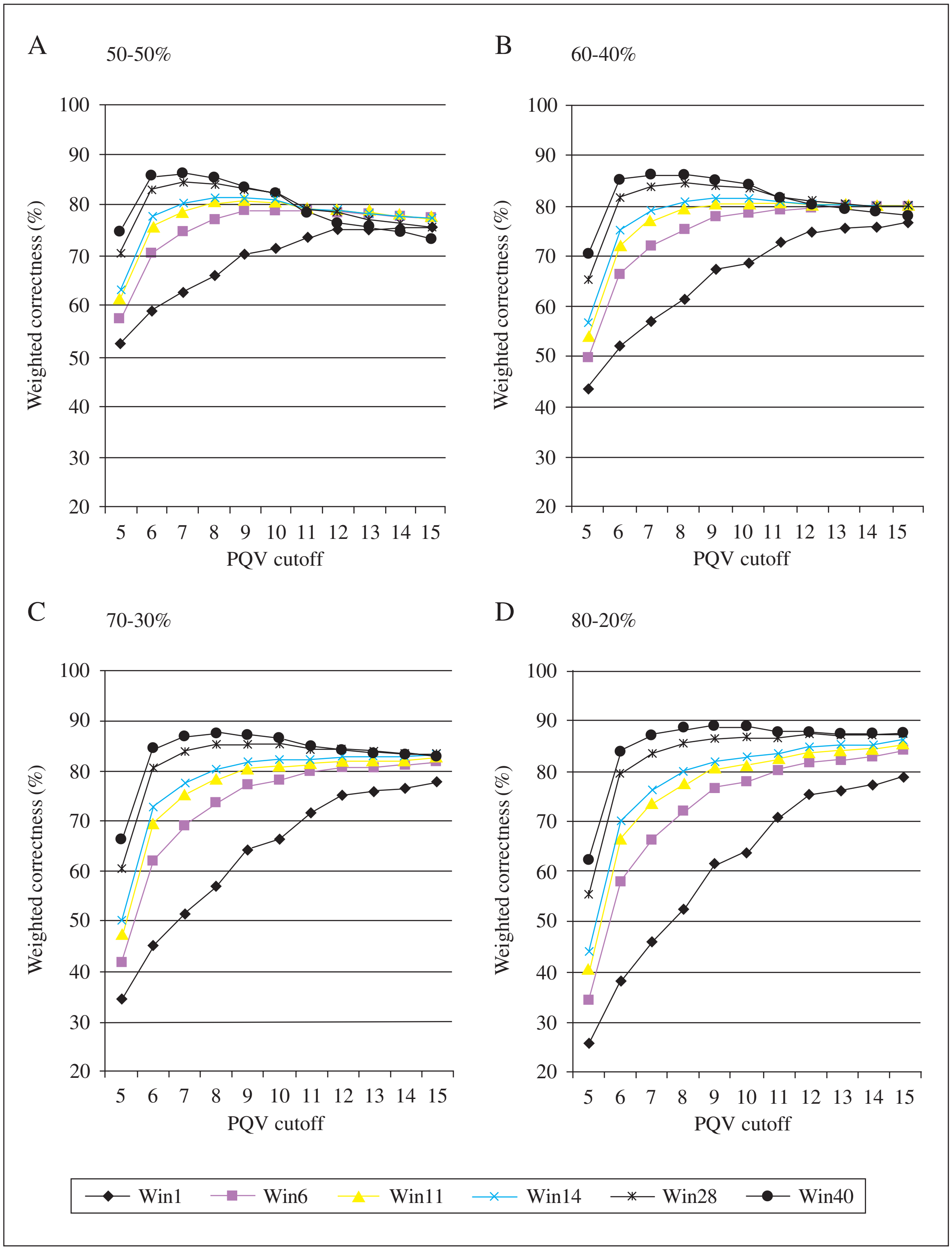


Figure 3. Distinct weights given to not masked windows and spoiled windows. The percentages in the graphs (Weight1-Weight2) indicate the relative importance of error masking as compared to the spoiling of correct windows (see Methodology). PQV = PHRED quality value.

Therefore, we conclude that the use of PQV cutoff masking frequently allows more than 85% of correct identification of windows. We have provided the data necessary for a sequencing research group to balance the number of the PQV cutoff if they preferentially desire either to mask errors or to allow for more correct windows. By examining Figure 3, it is possible to choose the best PQV cutoff for a specific application. We exclusively used sequences from the plasmid pUC18; similar studies using other templates and different sequencing machines, as well as other sequencing substrates (such as PCR or RT-PCR products), are necessary to determine if analogous results will be found.

ACKNOWLEDGMENTS

The authors thank the Rede Genoma de Minas Gerais (supported by FAPEMIG and MCT/Brazil), especially Marina Mourao, Lucila Pacifico and Renata Ribeiro, for the sequences and CENAPAD-MG/CO for the equipment used in the present study.

REFERENCES

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.

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, M.C. and **Green, P.** (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res. 8*: 175-185.

Green, P. (2004). PHRED Documentation. http://www.phrap.org/phrap.docs/phred.html. Acessed August 29, 2004.

Prosdocimi, F., Cerqueira, G.C., Binneck, E., Silva, A.F., Reis, A.N., Junqueira, A.C.M., Santos, A.C.F., Nhani-Júnior, A., Wust, C.I., Camargo-Filho, F., Kessedjian, J.L., Petretski, J.H., Camargo, L.P., Ferreira, R.G.M., Lima, R.P., Pereira, R.M., Jardim, S., Sampaio, V.S. and **Folgueras-Flatschart, A.V.** (2002). Bioinformática: Manual do usuário (in Portuguese). *Biotec. Cienc. Des. 29*: 18-31.

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.

README for stand-alone BLAST (2004). ftp://ftp.ncbi.nlm.nih.gov/blast/documents/blast.txt. Acessed August 29, 2004.

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

Scheetz, T.E., Trivedi, N., Roberts, C.A., Kucaba, T., Berger, B., Robinson, N.L., Birkett, C.L., Gavin, A.J., O'Leary, B., Braun, T.A., Bonaldo, M.F., Robinson, J.P., Sheffield, V.C., Soares, M.B. and **Casavant, T.L.** (2003). ESTprep: preprocessing cDNA sequence reads. *Bioinformatics 19*: 1318-1324.

Smith, T.F. and **Waterman, M.S.** (1981). Identification of common molecular subsequences. *J. Mol. Biol. 147*: 195-197.

Velculescu, V.E., Zhang, L., Vogelstein, B. and **Kinzler, K.W.** (1995). Serial analysis of gene expression. *Science 270*: 484-487.

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