Journal of Biomolecular Techniques
17:138–144 © 2006 ABRF

Evaluation of Methods for Sequence Analysis of Highly Repetitive DNA Templates

John W. Hawes,[1] Kevin L. Knudtson,[2] Helaman Escobar,[3] George S. Grills,[4] Timothy C. Hunter,[5] Emily Jackson-Machelski,[6] Heather Lin,[7] David S. Needleman,[8] Rashmi Pershad,[9] Glenis J. Wiebe[10]

[1]Miami University, Oxford, OH; [2]University of Iowa, Iowa City, IA; [3]University of Utah, Salt Lake City, UT; [4]Cornell University, Ithaca, NY; [5]Vermont Cancer Center, Burlington, VT; [6]Washington University School of Medicine, St. Louis, MO; [7]DigiSapien, Redmond, WA; [8]USDA ARS Eastern Regional Research Center, Wyndmoor, PA; [9]University of Texas MD Anderson Cancer Center, Houston, TX; [10]Max Planck Institute of Molecular Cell Biology and Genetics, Dresden, Germany

The DNA Sequencing Research Group (DSRG) of the ABRF conducted a study to assess the ability of DNA sequencing core facilities to successfully sequence a set of well-defined templates containing difficult repeats. The aim of this study was to determine whether repetitive templates could be sequenced accurately by using equipment and chemistries currently utilized in participating sequencing laboratories. The effects of primer and template concentrations, sequencing chemistries, additives, and instrument formats on the ability to successfully sequence repeat elements were examined. The first part of this study was an analysis of the results of 361 chromatograms from participants representing 40 different laboratories who attempted to sequence a panel of difficult-to-sequence templates using their best in-house protocols. The second part of this study was a smaller multi-laboratory evaluation of a single robust protocol with the same panel of templates. This study provides a measure of the potential success of different approaches to sequencing across homopolymer tracts and repetitive elements.

Key Words: DNA sequencing, repetitive elements, difficult templates, protocols.

The DNA sequences of eukaryotic and prokaryotic genomes are replete with regions containing homopolymer tracts and repetitive elements (reviewed in reference 1). The ability to successfully sequence these regions remains a challenge,[2–4] despite advancements in DNA sequencing chemistries and instruments.

Polyacrylamide slab gel–based DNA sequencing instruments have been the dominant technology used by DNA sequencing centers until recently.[5,6] The desire for improved automation, increased throughput, and sequencing quality has led to the development and use of capillary-based instruments.[6–8] While this has improved the capabilities of many DNA sequence analysis laboratories, it has also raised the level of expectations placed upon them. Advances in more sensitive dye systems that permit detection of smaller amounts of template and improved base calling methods have accompanied instrument advances.[9]

Dye-terminator chemistry was introduced over a decade ago[10,11] and has become the chemistry preferred by most automated sequencing centers. An early study that examined the accuracy of dye-primer vs. dye-terminator chemistries suggested that dye-primer chemistry gave longer read-length accuracy even when the dye-terminator results were manually edited.[12] Despite the poorer performance of the dye-terminator chemistry, users continued to embrace it because of its versatility and convenience. A subsequent study showed the use of dye-terminator chemistry was more successful than dye-primer chemistry in its ability to generate sequence data across homopolymer tracts and repetitive elements.[2] In addition, it was demonstrated that higher annealing temperatures and longer denaturation improved the ability to sequence through these difficult regions. Reagent additives have been shown to be important tools in successfully sequencing through homopolymer and repetitive types of difficult-to-sequence templates.[13]

Address correspondence and reprint requests to: Glenis Wiebe, Max Planck Institute of Molecular Cell Biology and Genetics, Pfotenhauerstr. 108, 01307 Dresden, Germany (email: wiebe@mpi-cbg.de).

There are currently no standard protocols or guidelines regarding the sequence analysis of templates with repetitive elements. Therefore, the DNA Sequencing Research Group (DSRG) of the ABRF undertook a study to assess how DNA sequencing core facilities could/would handle a set of well-defined difficult repeat templates. The goal of the ABRF DSRG study was to determine whether repetitive templates could be accurately sequenced using the equipment and chemistries currently utilized in participating laboratories, and whether it could be demonstrated that the uses of certain conditions or instruments provide improved sequencing quality when compared to others. This study examined the chemistries, additives, instrument formats, and reaction conditions used by DNA sequencing facilities to sequence through templates containing difficult repeat regions.

METHODS

Templates. Three mouse genomic clones containing repetitive elements flanked by an M13 primer site were used as templates for this study. These clones were isolated and characterized for finishing efforts for the Mouse Genome Project by the Genome Center of the Harvard Medical School-Partners Healthcare Center for Genetics and Genomics (Cambridge, MA) and donated for this study. Plasmid template samples were prepared using the maxi plasmid preparation method according to the manufacturer's protocol (Qiagen, Inc., Santa Clara, CA) and quantified via UV spectrophotometry.

Study design. The ABRF DSRG study was announced on the ABRF listserv and the ABRF website (http://www.abrf.org), and upon request, 10 µg of each of the three templates (denoted A, B, and C) along with the M13 (-20) forward primer (5'-TGTAAAACGACGGCCAGT-3') were provided to the study participants. No sequence information was provided other than that each template contained a repetitive element. Participating laboratories were requested to analyze each template using any chemistry, instrument, or condition of their choosing and were invited to try as many different conditions as the template amount provided would allow. Participants were asked to complete a survey that recorded all aspects of sample processing including reaction conditions, chemistries, additives, and instrument platforms along with their sequencing results.

Analysis. Sequencing results and surveys were collected via FTP and all data were analyzed for quality and read length using Phred, Phrap, and Consed.[14–16] Quality scores (q20 scores) were extracted from the Phred results and and plotted using custom Perl scripts (James VanEe, Cornell University, Ithaca, NY). After ranking all sequence submissions by q20, data were evaluated as follows: (1) quality sequence ending prior to the repeat regions, (2) quality sequence ending within the repeat regions, or (3) quality sequence beyond the repeat regions. The first 50 bases downstream of the M13 primer-binding site for each template were trimmed so all sequencing results were compared using a common origin.

RESULTS AND DISCUSSION

Templates selected for the study contained repeat sequences that varied in base composition and GC content (Figure 1). Study participants from 40 laboratories

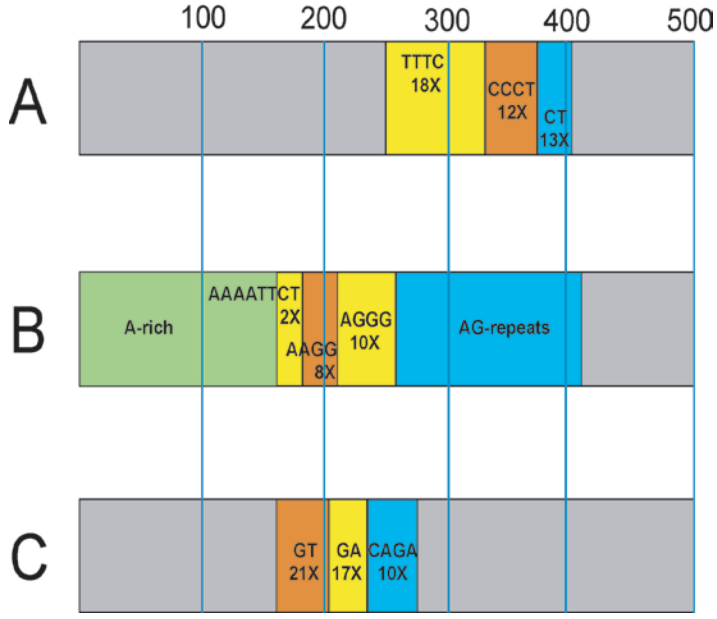

FIGURE 1

Illustration of the repetitive elements contained in the templates A, B, and C. The number of bases downstream of the M13 primer binding site is shown on top. Template A contained a 20-fold repeat of TTTC followed by a 12-fold repeat of CCCT followed immediately by a 26 base-pair repeat of CT. These repetitive elements began at approximately 250 base pairs from the M13 primer binding site. Template B contained an initial A-rich sequence containing two repeats of AAAATTCT at position 165 after the M13 primer binding site. This short repetitive element was followed by a 7-fold repeat of AAGG, a 30-fold repeat of AGGG, and an approximately 30-base-pair repeat of AG. Template C represented a greater level of diversity in the repetitive elements. This template contained a 44-base-pair repeat of GT starting at position 156 after the M13 primer binding site, followed immediately by a 32-base-pair repeat of AG and a 10-fold repeat of CCGA.

TABLE 1

DNA Sequencing Reaction Conditions Used to Generate the Top Three Quality Scores for Templates A, B, and C by Instrument

Template	Instrument[a]	Rank	Quality Score (q20)	WTR Length (cm)[b]	Primer (pmol)	Template (μg)	Reaction Volume (μL)	DNA Sequencing Chemistry[c]
A	377	1	837	48	3.2	300	20	BDTv3.1
		2	713	48	15.0	1140	40	BDTv3.1
		3	690	48	3.2	300	20	BDT v3.1
	3100	1	815	80	12.5	380	10	BDTv3.1
		2	765	80	5.0	630	20	BDTv3.1
		3	765	80	3.2	494	20	BDTv1.1
	3700	1	767	50	10.0	200	10	BDTv3.1
		2	715	50	10.0	200	10	BDTv3.1
		3	708	50	10.0	200	10	BDTv3.1
B	377	1	951	48	4.8	400	20	dGTPv3.0
		2	856	48	3.2	300	20	BDTv3.1
		3	793	48	5.0	800	20	dGTPv3.0
	3100	1	879	80	5.0	730	20	BDTv3.0:dGTPv3.0 (4:1)
		2	871	80	3.2	494	20	BDTv3.0:dGTPv2.0 (2:1)
		3	854	80	5.0	730	20	dGTPv3.0
	3700	1	784	50	4.0	200	15	BDTv3.1
		2	767	50	3.5	365	10	BDTv3.1
		3	758	50	3.5	365	10	BDTv3.0
C	377	1	908	48	4.8	400	20	dGTPv3.0
		2	820	48	3.2	300	20	BDTv3.1
		3	729	48	15.0	2550	40	dGTPv3.0
	3100	1	881	80	5.0	850	20	dGTPv3.0
		2	844	80	5.0	850	20	BDTv3.0:dGTPv3.0 (4:1)
		3	835	80	3.2	494	20	BDTv3.0:dGTPv2.0 (2:1)
	3700	1	805	50	10.0	400	10	BDv3.1
		2	788	50	10.0	400	10	dGTP v3.0
		3	777	50	4.0	200	15	BDT 3.1

[a]Instrument: Applied Biosystems Models 377, 3100, and 3700 sequencers.

[b]WTR Length: well-to-read length (cm).

[c]Chemistry: Applied Biosystems ABI Prism big dye terminator (BDT) and dGTP big dye terminator ready reaction mixes. Parentheses indicate the ratio of BDT to dGTP.

returned a total of 361 DNA sequencing results and surveys, consisting of 118 submissions each for templates A and B and 125 submissions for template C.

Ability to sequence repeats. Most participants were able to generate sequence through the repetitive sequence elements despite using a wide range of chemistries and instrument platforms. Ninety-one percent, 81%, and 79% of the sequence submissions gave successful results through the repetitive elements in templates A, B, and C, respectively. Only 33%, 43%, and 50% of the sequence submissions for templates A, B, and C, respectively, were able to generate successful sequence data beyond the repetitive element. The top three submissions for each template by instrument, based on q20, are shown in Table 1. Ranked phred quality scores for data from all participants are available to the public at: http://www.abrf.org/ResearchGroups/DNASequencing/EPosters/DSRG2003Study.pdf.

Effects of instrument. Data were submitted from participants that used slab gel-based or capillary-based instruments from five different sequencing instrument manufacturers. However, the majority (93%) of the results presented herein were acquired on Applied Biosystems

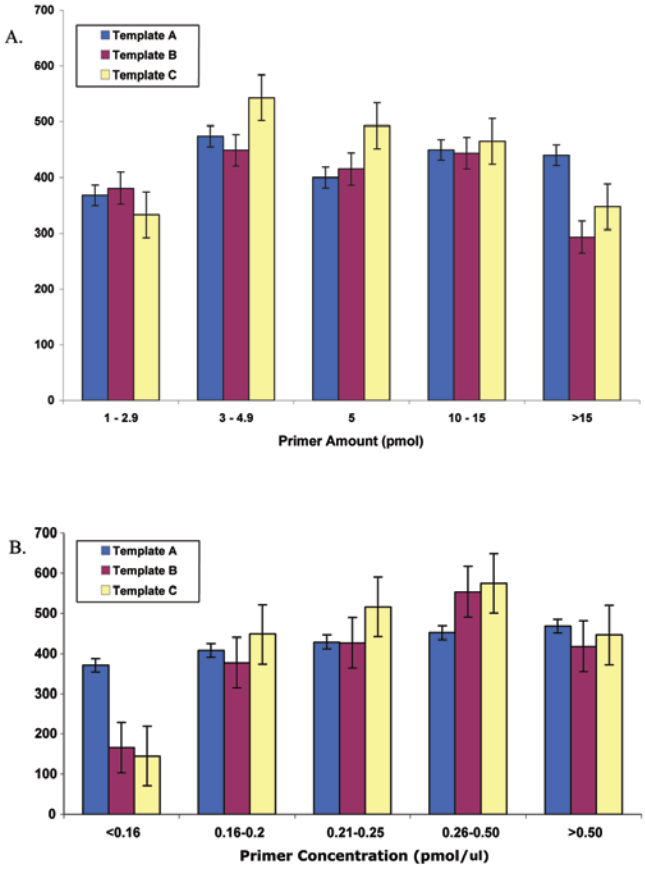

FIGURE 2

The effect of the amount of primer on sequence quality. **A**: Results displayed as the exact amount of primer added. **B**: Primer amounts adjusted for the reaction volume. The bars represent the average and standard errors of the q20 scores for each template (A, B, and C) for the amount of primer used in the sequencing reaction.

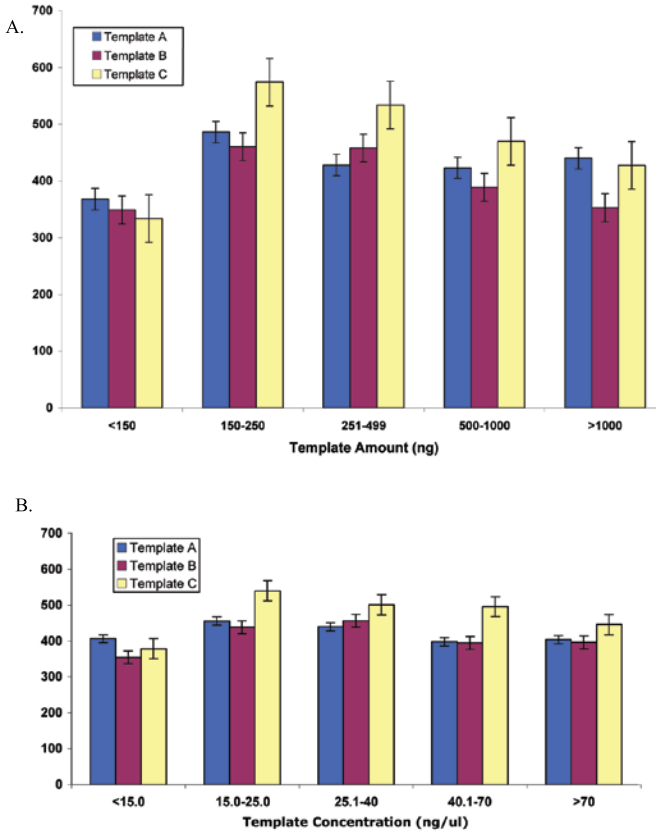

FIGURE 3

The effect of the amount of template on sequence quality. **A**: Results displayed as the exact amount of template added. **B**: Template amounts adjusted for the reaction volume. The bars represent the average and standard errors of the q20 scores for each template (A, B, and C) for the amount of template used in the sequencing reaction.

instruments, including the models 3100 (20%), 377 (38%), and 3700 (20%). Only 5% of the submissions were collected from the Applied Biosystems models 3730 or 3730xl autosequencers, because these instruments were released just prior to the launch of the study. Both gel- and capillary-based instruments were used to successfully generate sequence beyond the repeat regions in each of the three templates. Based on q20 scores, there was no significant difference between data quality generated from capillary and slab gel instruments (Table 1). As expected, the sequences generated using the longer capillaries or longer gels available for a given instrument, which allow for higher resolution, gave higher quality scores.

Effects of the amount of primer and template. The study participants used a wide range of primer (1 to 100 pmol) and template (100 ng to 2.6 μg) amounts, with reaction volumes ranging from 10 to 40 μL. The effect of the primer amount in the sequencing reaction was not significant. However, the use of less than 3 pmol or more than 15 pmol gave the poorest quality of sequence based on phred analysis (Figure 2A). The effect of template amount on quality of sequence was also not significant, but the use of less than 150 ng of template tended to yield poorer sequence quality (Figure 3A). Primer and template concentrations were also evaluated in order to account for differences in reaction volumes, but again, no significant effects were noted other than for those reactions using less than 0.16 pmol/μL of sequencing primer (Figures 2B and 3B).

Overall, the use of 3–5 pmol of primer and between 150 and 500 ng of template per sequencing reaction appeared to give consistently better results (Figures 2 and 3). These results agree with the Applied Biosystems protocol recommendation of using 3.2 pmol primer and between 200 and 500 ng template (for double-stranded

TABLE 2

Internal Study Results[a]

Template	Array Length (cm)	Quality Score (q20)	Quality Score Average
A	80	830	
A	80	765	
A	50	613	
A	50	526	
A	50	501	
A	50	440	
A	50	429	
A	50	376	
A	36	316	533
B	80	920	
B	80	703	
B	50	641	
B	50	637	
B	50	579	
B	36	546	
B	50	395	
B	50	388	
B	50	201	557
C	80	904	
C	80	849	
C	50	635	
C	50	629	
C	50	624	
C	50	617	
C	50	593	
C	50	354	
C	50	169	
C	36	32	541

[a]DNA sequencing reactions for Templates A, B, and C were prepared using the robust protocol and run on Applied Biosystems Model 3100 genetic analyzers.

products) when using their big dye terminator sequencing ready reaction kit.

Effects of reaction chemistries and additives. Rankings, based on q20, showed that Applied Biosystems BigDye Terminator (BDT) chemistries tended to produce higher quality sequences than the other chemistries used by the study participants (Table 1). Data with q20 reads beyond the repeat regions were noticeably more prevalent with the use of BDTv3.1 chemistry. Sequence reads beyond the repetitive region were also obtained using the Licor, Inc., and General Electric (formerly Amersham Biosciences) chemistries and instrumentation.

Perhaps more important than the choice of chemistries was the use of various additives in the sequencing reactions. The use of either DMSO or betaine had no noticeable effect on q20 scores with any of the three templates used in this study. However, there was a strong correlation with the use of dGTP in reaction mixtures and read length with all three templates. For example, only seven out of 125 submissions produced q20 scores over 800 for template C. All seven used either BDTv3.0 or BDTv3.1, and the top four submissions (longest read lengths) with template C used dGTP either alone or in combination with standard terminator mixtures. The choice of dGTP mixtures by over half of the participants indicates that many are aware of the potential benefit dGTP can have on the sequencing of difficult templates. The use of the dGTP chemistry on a wider

TABLE 3

Average Sequencing Quality Scores (Q20) for the Templates Used in the Internal and External Studies

Template/Study	Mean±SEM	Probability
Template A		
Internal	617±35	
External	431±15	P<0.0001
Template B		
Internal	662±48	
External	424±20	P<0.0001
Template C		
Internal	645±50	
External	476±19	P<0.002
Overall		
Internal	641±26	
External	444±11	P<0.0001

range of templates containing repeats warrants further examination.

Robust protocol for repeat sequence analysis. Based on the ABRF DSRG study results, an internal study was subsequently performed by DSRG members in which the same three repetitive templates were re-analyzed under a more controlled set of sequencing reaction conditions and chemistry choices. This was done in an effort to minimize lab-to-lab variation in methodology, and to test a defined protocol for robustness in sequencing of known difficult repeats. As with the external study, plasmid template samples were purified in one location, using standard maxi-prep methods. These templates were then distributed to members of the DSRG. A common set of reaction conditions and thermalcycling parameters was used to sequence these templates with the same primer used in the external study. The choice of reaction conditions was made based on examination of all of the data submitted by participants in the external study. These conditions were as follows:

Robust Protocol

Reaction:

300 ng template A, B, or C
5 pmol M13 forward primer
4 μL BDTv3.1 (Applied Biosystems)
2 μL 5X Sequencing Buffer (Applied Biosystems)
H_2O to 20 μL

Cycling:

Initial denaturation: 95°C, 5 min
30 cycles: 96°C 10 sec, 50°C 5 sec, 60°C 4 min
Rapid ramp and hold at 4°C

Post-reaction clean-up was performed by using either gel filtration or ethanol precipitation. Samples were resuspended in either water or formamide, and analyzed on an Applied Biosystems Model 3100 genetic analyzer, using instrument defaults for injection and run conditions. Phred q20 scores were used as a measure of sequencing quality. These data were evaluated based on the same criteria mentioned previously.

Like the ABRF DSRG external study, sequence quality appeared to be affected by capillary length such that the use of longer capillaries tended to give higher sequence quality (Table 2). There was also a higher overall success rate for the internal study as compared to the ABRF DSRG external study (Table 3). For each template, the use of the robust protocol gave significantly better sequencing quality than those protocols used in the external study. However, other factors such as user experience may have also contributed to the differences in the overall success rates between the studies. Therefore, the standardized protocol provided herein is a good starting point for sequencing through highly repetitive regions.

Additional information about the internal study may be found at http://www.abrf.org/ResearchGroups/DNA-Sequencing/Publications/DSRG2003_InternalStudy.pdf.

CONCLUSIONS

Three important trends were noted in the ABRF DSRG study: (1) the top two results from each template were achieved with long reads on a 377 and 3100, respectively, (2) the use of smaller amounts of primer (<3 pmol) and template (<150 ng) yielded poorer sequence quality, and (3) the BigDye v3.1 or BigDye dGTP kits (alone or in a mixture) were generally better at dealing with repetitive samples. Buffers, additives (DMSO or betaine), purification methods, loading media, and cycling conditions did not play statistically significant roles in the results submitted. However, each sample submitted used a slightly different protocol. The second, internal study was then designed to minimize the protocol differences as much as possible. Overall, a higher success rate of sequencing through repetitive regions was achieved by the protocol used in the internal study than from any of protocols used by participants in the ABRF DSRG study. Lab-to-lab variability affected the results even of the internal study, suggesting that variables such as instrument robustness, capillary array usage, reagent quality, and technical experience are important elements in successfully sequencing through difficult-to-sequence templates. However, the results of the internal study indicate that the standardized protocol presented here is a good starting point for sequencing through highly repetitive regions.

REFERENCES

1. Lewin B. *Genes VI*. New York: Oxford University Press, Inc. 1997.
2. Robbins CM, Hsu E, Gillevet PM. Sequencing homopolymer tracts and repetitive elements. *BioTechniques* 1996;20:862–868.
3. Langan JE, Rowbottom L, Liloglou T, Field JK, Risk JM. Sequencing of difficult templates containing poly(A/T) tracts: Closure of sequence gaps. *BioTechniques* 2002;33:276–280.
4. Stirling D. Technical notes for sequencing difficult templates. *Methods Mol Biol* 2003;226:401–402.
5. Wiebe GJ, Pershad R, Escobar H, Hawes JW, Hunter T, Jackson-Machelski E, et al. DNA Sequencing Research Group (DSRG) 2003—A general survey of core DNA sequencing facilities. *J Biomol Tech* 2003;14:230–236.
6. Marziali A, Akeson M. New DNA sequencing methods. *Annu Rev Biomed Eng* 2001;3:195–223.
7. Meldrum D. Automation for genomics, part two: Sequencers, microarrays, and future trends. *Genome Res* 2000;10:1288–1303.
8. Swanson D. The art of the state of nucleic acid sequencing. *The Scientist* 2000;14:23–26.
9. Pisano JM. The core of DNA sequencing. *The Scientist* 2002;16:41–45.
10. Rosenthal A, Charnock-Jones DS. New protocols for DNA sequencing with dye terminators. *DNA Seq* 1992;3:61–64.
11. Rosenthal A, Charnock-Jones DS. Linear amplification sequencing with dye terminators. *Methods Mol Biol* 1993;23:281–296.
12. Naeve CW, Buck GA, Niece RL, Pon RT, Robertson M, Smith AJ. Accuracy of automated DNA sequencing: A multi-laboratory comparison of sequencing results. *Biotechniques* 1995;19:448–453.
13. Adams PS, Dolejsi MK, Grills GS, McMinimy D, Morrison P, Rush S, et al. An analysis of techniques used to improve the accuracy of automated DNA sequencing of a GC-rich template. *J Biomol Tech* 1998;9:9–18.
14. Ewing B, Hillier L, Wendl MC, Green P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res* 1998;8(3):175–185.
15. Ewing B and Green P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res* 1998; (3):186–194.
16. Gordon D, Abajian C, Green P. Consed: A graphical tool for sequence finishing. *Genome Res* 1998;8(3):195–202.