

EFFECTS OF DIFFERENT DNA SEQUENCING METHODS EVALUATED USING A WEB BASED QUALITY CONTROL RESOURCE: THE ABRF DNA SEQUENCE RESEARCH GROUP 2001 STANDARD TEMPLATE STUDY



Grills, G.¹, Leviten, D.², Hall, L.³, Hawes, J.⁴, Hunter, T.⁵, Jackson-Machelski, E.⁶, Knudtson, K.⁷, Robertson, M.⁸, Thannhauser, T.⁹, Adams, P.S.¹⁰, Hardin, S.¹¹ and J. VanEe¹².

¹Albert Einstein College of Medicine, Bronx, NY; ²COS Corporation, Bothell, WA; ³Indiana University School of Medicine, Indianapolis, IN; ⁴University of Vermont, Burlington, VT; ⁵Washington University School of Medicine, Saint Louis, MO; ⁶University of Iowa, Iowa City, IA; ⁷University of Utah, Salt Lake City, UT; ⁸Cornell University, Ithaca, NY; ⁹Trudeau Institute, Saranac Lake, NY; ¹⁰University of Houston, Houston, TX.

ABSTRACT

The goal of this study was to analyze the effect of different DNA sequencing methods on the quality of resulting data. A wide variety of sequencing groups submitted data for pGEM, a standard quality control sequencing template. Sequence data was collected by FTP or HTTP and details of sequencing conditions were collected by web forms. The effect of factors such as different types of instrumentation and chemistries were examined. The current data were compared to data from our prior studies. Results of using common and new technologies were analyzed. In particular, results from capillary array sequencers such as the ABI 3700 were evaluated. A major aim of this study was to update and show the utility of our "Never Ending Story" (NES) database, a web based resource of sequencing data that we established in 1998 and made publicly available in a new easy to use format in 2000. The results of this study may be used for quality control, troubleshooting, and evaluation of new technologies.

INTRODUCTION

Goals: The overall goal of the Association of Biomolecular Resource Facilities (ABRF) DNA Sequence Research Group (DSRG) 2001 Standard Template Study was to analyze the effect of different DNA sequencing methods on the quality of sequencing results. We requested sequencing laboratories to submit the results of sequencing a standard pGEM template with any chemistry, run condition and machine type. The study examined both well established and relatively new sequencing methods. To evaluate the effects of new technologies, this study examined data collected from January 1998 to the end of April 2001.

NES Database: This analysis is a continuation of "The Standard Template Study: The Never Ending Story (NES)" that was established by the DSRG in 1998. The NES database web site was created last year. The NES is a web based resource of sequencing data that permits anonymous submission of sequencing data over the web. The database automatically does phred analysis of submitted data and allows on line queries of all data in the database. The database is located at <http://nes.biotech.cornell.edu>.

Applications of results: The results of this study may be used to: (1) anonymously evaluate the quality of sequencing results relative to that achieved in other laboratories; (2) systematically evaluate different instruments, chemistries and protocols when considering either equipment purchases or modifications to standard operating procedures; and (3) determine the causes and solutions to technical problems.

METHODS

Participation in this study was solicited through electronic bulletin boards. Participants submitted unedited chromatogram files of the results of sequencing pGEM-3Z(+/-) template with the M13(+/-) forward primer. LICOR participants used the M13(+/-) forward primer. Sequence data was submitted anonymously via the web. Chromatogram files and information about the sequencing conditions were collected on the NES web site at <http://nes.biotech.cornell.edu>. Data from instrument and reagent manufacturers was not included in this analysis.

The base composition of the pGEM-3Z(+/-) template from the M13(+/-) priming site was determined using SeqEd (Applied Biosystems, Foster City, CA). Potential secondary structures of the pGEM template were determined with mFold software (Merrill, G. and S.H. Hardin, *Nucleic Acids Res.* 28(7), E221, which identifies regions of self-complementarity and determines free energy values for such regions. Submitted sequences were compared to the known sequence using SeqEd. Alignments were trimmed at the 5' end to base +1 from the M13(+/-) priming site. A script (L.L. Albert Einstein Coll. of Med., Bronx, NY) was used to count the numbers of errors. Substitutions (both miscalls and ambiguities), insertions and deletions were considered errors.

Chromatograms were analyzed with phred software (Ewing, B. and P. Green, *Genome Res.* 8,196-199). Phred assigns base calls and quality values to each peak. The quality values correspond to the inverse probability of a correct base assignment. For example, a quality value of Q20 corresponds to approximately a 1 error in 10¹, or a 1% chance that the base call is not correct. The number of base calls with specific quality values was determined with grep (Brent Ewing, University of Washington, WA). Statistical analysis was done with SPSS (SPSS, Chicago, IL).

Submissions

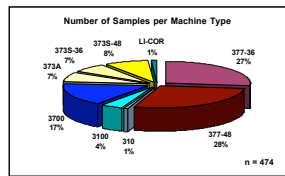


Figure 1. Summary of Submissions: Number of Samples Submitted for Different Machine Types. Machines are designated by model type and well-to-read length. The ABI machine configurations include the slab-gel based 373A, the 373-stretch with 36 cm plates (3735-36) or 48 cm plates (3735-48), the 377 with 36 cm plates (377-36) or 48 cm plates (377-48) and the capillary based 310, 3100, and 3700 instruments. The 377-36 4X and 2X run conditions are grouped together. 310 capillaries of different lengths are grouped together. Different well-to-read conditions for the LICOR are grouped together. A total of 474 unedited pGEM samples from 96 labs were submitted and analyzed for this study. Each lab submitted an average of 5.4 samples. 33% of labs submitted samples for more than one machine type. 210 samples were submitted in 1998, 116 samples in 1999, 5 samples in 2000 and 143 samples in the first four months of 2001.

Analysis of Standard Template

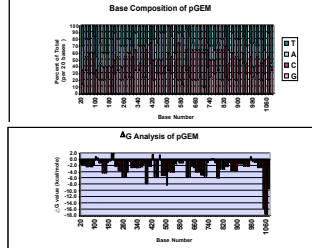


Figure 2. Analysis of pGEM as a Sequencing Template: Base Composition and Secondary Structure. (Top) pGEM-3Z(+/-) base content. Base content was calculated using a sliding 20-base window starting at the M13(+/-) priming site. pGEM has an average GC content of 54%. There is a 58% T-rich region from base +1070 to +1080. (Middle) Free energy AG values along the sequence of pGEM. The value for 10 base windows was calculated starting at the M13(+/-) priming site. pGEM has an average AG value of -2.7 kcal/mole from base +1 to +1040. From base +1040 to +1080, there is a marked decrease in average AG value to -1.1 kcal/mole. (Bottom) Inhibitory secondary structure of pGEM from base +1040 to +1080. There is a 32 base palindromic region from base +1040 to +1080.

CONCLUSIONS

Standard Template: pGEM-3Z(+/-) is an ideal sequencing substrate from base +1 of the M13(+/-) priming site up to base +1040. Inhibitory secondary structure may substantially decrease the success rate of obtaining read lengths longer than 1040 bases.

Machine Types: Longer well-to-read distance improves accuracy and quality on all machine types with standard template. Most machines give similar accuracy at less than 400 base read lengths. The ABI 377-48 and the LICOR instruments give the best read lengths, accuracy and quality. The ABI 3700 and 3100 can give overall sequence accuracy and quality as good or better than the ABI 377-36.

Dye Chemistry: ABI BigDyes v2 show an improvement in quality compared to results with previously available ABI dye chemistries.

Effects of Dilutions and Reaction Volumes: BigDye dye terminators maintain both accuracy and quality with the most common dilutions and reaction volumes submitted to this study. BigDyes with reduced reaction volumes or dilutions gave the best results overall with standard template.

RESULTS

Accuracy and Quality of Different Machine Types

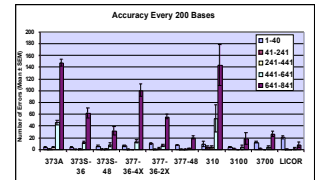


Figure 3. Accuracy and Quality of Different Machine Types. Machine configurations are differentiated by model type, well-to-read and speed of run conditions. The results for different chemistries and other run conditions are grouped together for each configuration. (Top) The average number of errors for each machine type for different length of reads, starting with base +1 to +40, and then the non-cumulative average number of errors for every 200 base interval up to +840 bases. Errors are defined as any type of error in base calling in the unedited sequence data, including miscalls, insertions, and ambiguities. (Middle) The total average number of errors for each machine type in the full range of +41 to +840 bases. (Bottom) Length of read: total number of bases detected by phred. Accurate basecalls: total number of unedited correct bases called by the ABI or LICOR analysis software from base +41 to +1000. Quality: total number of bases assigned a phred confidence value of Q20 (maximum number of bases possible to run with that machine configuration)/lanes used by the machine per sequence/run time.

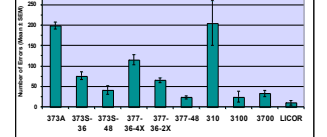


Figure 3. Accuracy and Quality of Different Machine Types. Machine configurations are differentiated by model type, well-to-read and speed of run conditions. The results for different chemistries and other run conditions are grouped together for each configuration. (Top) The average number of errors for each machine type for different length of reads, starting with base +1 to +40, and then the non-cumulative average number of errors for every 200 base interval up to +840 bases. Errors are defined as any type of error in base calling in the unedited sequence data, including miscalls, insertions, and ambiguities. (Middle) The total average number of errors for each machine type in the full range of +41 to +840 bases. (Bottom) Length of read: total number of bases detected by phred. Accurate basecalls: total number of unedited correct bases called by the ABI or LICOR analysis software from base +41 to +1000. Quality: total number of bases assigned a phred confidence value of Q20 (maximum number of bases possible to run with that machine configuration)/lanes used by the machine per sequence/run time.

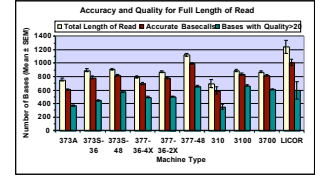


Figure 3. Accuracy and Quality of Different Machine Types. Machine configurations are differentiated by model type, well-to-read and speed of run conditions. The results for different chemistries and other run conditions are grouped together for each configuration. (Top) The average number of errors for each machine type for different length of reads, starting with base +1 to +40, and then the non-cumulative average number of errors for every 200 base interval up to +840 bases. Errors are defined as any type of error in base calling in the unedited sequence data, including miscalls, insertions, and ambiguities. (Middle) The total average number of errors for each machine type in the full range of +41 to +840 bases. (Bottom) Length of read: total number of bases detected by phred. Accurate basecalls: total number of unedited correct bases called by the ABI or LICOR analysis software from base +41 to +1000. Quality: total number of bases assigned a phred confidence value of Q20 (maximum number of bases possible to run with that machine configuration)/lanes used by the machine per sequence/run time.

Throughput of Different Machine Types

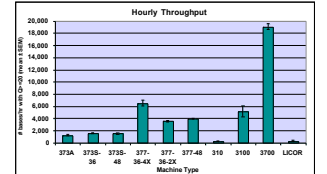


Figure 4. Throughput of Different Machine Types. The number of high quality bases that can be produced per hour by each machine type. Instrument throughput for each sequence is defined as: (the total number of bases with a phred quality of Q20)/(maximum number of lanes possible to run with that machine configuration)/lanes used by the machine per sequence/run time.

ACKNOWLEDGMENTS

Available assistance provided by: L.L. Ewing Biotech, Napa, CA; George Arino (Albert Einstein College of Medicine, Gene, Switzerland); Tom Sebald, Tanya Fagan (University of Washington, Seattle, WA); John Collins (University of Washington, Seattle, WA); Thomas Schmitt, John B. Collins (University of Washington, Seattle, WA); and the ABRF DNA Sequence Research Group.

Dye Chemistry Comparison

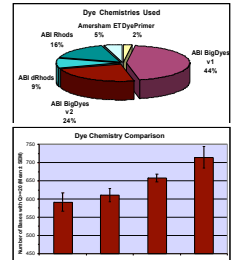


Figure 5. Comparison of Dye Chemistries. (Top) Types of dye chemistries used by submitted samples. 98% used dye terminator chemistry. 88% used ABI BigDyes terminator chemistry. dRhods refers to ABI Dichlororhodamine terminator chemistry. Rhods refers to the older ABI Rhodamine terminator chemistry. All Rhod samples were created prior to the introduction of dRhods and BigDyes. (Bottom) Phred quality results of sequencing pGEM on one machine type, the ABI 377-48, with BigDyes v1 (n=83), BigDyes v2 (n=18), dRhods (n=19), and rhods (n=9) terminator chemistry. These samples came from a total of 34 different labs.

Effects of Dilution & Rxn Vol.

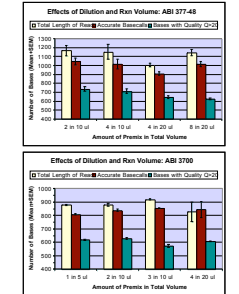


Figure 6. Effects of Dilution and Reaction Volume. The most common dilutions and reaction volumes submitted to this study were analyzed for ABI BigDyes terminator chemistry run on the 377-48 (n=81) and 3700 (n=68). The most common dilutions of this enzyme premix were: full volume (8 µl of enzyme premix in 20 µl total rxn), 1/2 volume (4 µl of premix in 20 µl or 10 µl total rxn), 1/4 volume (2 µl of premix in 10 µl total rxn), and 1/8 volume (1 µl of premix in 10 µl total rxn).

Ranking by Accuracy

TYPE	Machine	Chemistry	Dilution	Rxn Vol	Seqs	Errs	Expns
2736	373A	BigDye	1:1	10 µl	1000	10	\$100
2737	373A	BigDye	1:1	20 µl	1000	10	\$100
2738	373A	BigDye	1:1	40 µl	1000	10	\$100
2739	373A	BigDye	1:1	80 µl	1000	10	\$100
2740	373A	BigDye	1:1	160 µl	1000	10	\$100
2741	373A	BigDye	1:1	320 µl	1000	10	\$100
2742	373A	BigDye	1:1	640 µl	1000	10	\$100
2743	373A	BigDye	1:1	1280 µl	1000	10	\$100
2744	373A	BigDye	1:1	2560 µl	1000	10	\$100
2745	373A	BigDye	1:1	5120 µl	1000	10	\$100
2746	373A	BigDye	1:1	10240 µl	1000	10	\$100
2747	373A	BigDye	1:1	20480 µl	1000	10	\$100
2748	373A	BigDye	1:1	40960 µl	1000	10	\$100
2749	373A	BigDye	1:1	81920 µl	1000	10	\$100
2750	373A	BigDye	1:1	163840 µl	1000	10	\$100
2751	373A	BigDye	1:1	327680 µl	1000	10	\$100
2752	373A	BigDye	1:1	655360 µl	1000	10	\$100
2753	373A	BigDye	1:1	1310720 µl	1000	10	\$100
2754	373A	BigDye	1:1	2621440 µl	1000	10	\$100
2755	373A	BigDye	1:1	5242880 µl	1000	10	\$100
2756	373A	BigDye	1:1	10485760 µl	1000	10	\$100
2757	373A	BigDye	1:1	20971520 µl	1000	10	\$100
2758	373A	BigDye	1:1	41943040 µl	1000	10	\$100
2759	373A	BigDye	1:1	83886080 µl	1000	10	\$100
2760	373A	BigDye	1:1	167772160 µl	1000	10	\$100
2761	373A	BigDye	1:1	335544320 µl	1000	10	\$100
2762	373A	BigDye	1:1	671088640 µl	1000	10	\$100
2763	373A	BigDye	1:1	1342177280 µl	1000	10	\$100
2764	373A	BigDye	1:1	2684354560 µl	1000	10	\$100
2765	373A	BigDye	1:1	5368709120 µl	1000	10	\$100
2766	373A	BigDye	1:1	10737418240 µl	1000	10	\$100
2767	373A	BigDye	1:1	21474836480 µl	1000	10	\$100
2768	373A	BigDye	1:1	42949672960 µl	1000	10	\$100
2769	373A	BigDye	1:1	85899345920 µl	1000	10	\$100
2770	373A	BigDye	1:1	171798691840 µl	1000	10	\$100
2771	373A	BigDye	1:1	343597383680 µl	1000	10	\$100
2772	373A	BigDye	1:1	687194767360 µl	1000	10	\$100
2773	373A	BigDye	1:1	1374389534720 µl	1000	10	\$100
2774	373A	BigDye	1:1	2748779069440 µl	1000	10	\$100
2775	373A	BigDye	1:1	5497558138880 µl	1000	10	\$100
2776	373A	BigDye	1:1	10995116277760 µl	1000	10	\$100
2777	373A	BigDye	1:1	21990232555520 µl	1000	10	\$100
2778	373A	BigDye	1:1	43980465111040 µl	1000	10	\$100
2779	373A	BigDye	1:1	87960930222080 µl	1000	10	\$100
2780	373A	BigDye	1:1	175921860444160 µl	1000	10	\$100
2781	373A	BigDye	1:1	351843720888320 µl	1000	10	\$100
2782	373A	BigDye	1:1	703687441776640 µl	1000	10	\$100
2783	373A	BigDye	1:1	1407374883553280 µl	1000	10	\$100
2784	373A	BigDye	1:1	2814749767106560 µl	1000	10	\$100
2785	373A	BigDye	1:1	5629499534213120 µl	1000	10	\$100
2786	373A	BigDye	1:1	11258999068426240 µl	1000	10	\$100
2787	373A	BigDye	1:1	22517998136852480 µl	1000	10	\$100
2788	373A	BigDye	1:1	45035996273704960 µl	1000	10	\$100
2789	373A	BigDye	1:1	90071992547409920 µl	1000	10	\$100
2790	373A	BigDye	1:1	180143985094819840 µl	1000	10	\$100
2791	373A	BigDye	1:1	360287970189639680 µl	1000	10	\$100
2792	373A	BigDye	1:1	720575940379279360 µl	1000	10	\$100
2793	373A	BigDye	1:1	1441151880758558720 µl	1000	10	\$100
2794	373A	BigDye	1:1	2882303761517117440 µl	1000	10	\$100
2795	373A	BigDye	1:1	5764607523034234880 µl	1000	10	\$100
2796	373A	BigDye	1:1	11529215046068469760 µl	1000	10	\$100
2797	373A	BigDye	1:1	23058430092136939520 µl	1000	10	\$100
2798	373A	BigDye	1:1	46116860184273879040 µl	1000	10	\$100
2799	373A	BigDye	1:1	92233720368547758080 µl	1000	10	\$100
2800	373A	BigDye	1:1	184467440737095516160 µl	1000	10	\$100

Figure 7. Top Three Lab Submissions per Machine Type. Sequences were ranked first by the number of errors from base +1-840 and then by errors from base +1-1040. The most accurate sequence per lab for each machine type was ranked. More information on the run conditions for all files are available on the NES database web site. File names are anonymous identification numbers. Phred Q20: total number of base calls with this confidence value. LCR: longest continuous correct length of read. DT: Dye terminator. DP: Dye Primer.