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

ChromaPipe: a pipeline for analysis, quality control and management for a DNA sequencing facility

T.D. Otto[1,2], E.A. Vasconcellos[1,2], L.H.F. Gomes[1,3], A.S. Moreira[1], W.M. Degrave[1], L. Mendonça-Lima[1] and M. Alves-Ferreira[1]

[1]Laboratório de Genômica Funcional e Bioinformática,
Instituto Oswaldo Cruz, Fiocruz, Rio de Janeiro, RJ, Brasil
[2]Fundação Ataulpho de Paiva, Rio de Janeiro, RJ, Brasil
[3]Faculdade de Medicina, Universidade Federal do Rio de Janeiro,
Rio de Janeiro, RJ, Brasil

Corresponding author: T.D. Otto
E-mail: otto@ioc.fiocruz.br

Genet. Mol. Res. 7 (3): 861-871 (2008)
Received June 2, 2008
Accepted August 11, 2008
Published September 23, 2008

ABSTRACT. Optimizing and monitoring the data flow in high-throughput sequencing facilities is important for data input and output, for tracking the status of results for the users of the facility, and to guarantee a good, high-quality service. In a multi-user system environment with different throughputs, each user wants to access his/her data easily, track his/her sequencing history, analyze sequences and their quality, and apply some basic post-sequencing analysis, without the necessity of installing further software. Recently, Fiocruz established such a core facility as a "technological platform". Infrastructure includes a 48-capillary 3730 DNA Sequence Analyzer (Applied Biosystems) and supporting equipment. The service includes running samples for large-scale users, performing DNA sequencing reactions and runs for medium and small users, and participation in partial or full genome projects. We implemented a workflow that fulfills these requirements for small and high throughput users. Our implementation also includes the monitoring of data for continuous quality improvement (reports by plate, month

and user) by the sequencing staff. For the user, different analyses of the chromatograms, such as visualization of good quality regions, as well as processing, such as comparisons or assemblies, are available. So far, 180 users have made use of the service, generating 155,000 sequences, 35% of which were produced for the BCG Moreau-RJ genome project. The pipeline (named ChromaPipe for Chromatogram Pipeline) is available for download by the scientific community at the url http://bioinfo.pdtis.fiocruz.br/ChromaPipe/. The support for assembly is also configured as a web service: http://bioinfo.pdtis.fiocruz.br/Assembly/.

Key words: Sequencing pipeline; Chromatogram processing; DNA sequencing

INTRODUCTION

Since the mid-1970s, DNA sequencing has been a key technique for recombinant DNA technology and more recently also for genome projects. Initially, two different techniques were developed for DNA sequencing: the Maxam-Gilbert (Maxam and Gilbert, 1977) and the Sanger approach (Sanger et al., 1977). The first one is also known as "chemical sequencing" because this method requires radioactive labeling of the (usually) 5' end of one strand, chemical treatment for partial nucleotide-specific cleavage of the DNA, clean-up and polyacrylamide gel electrophoresis of the DNA fragments followed by autoradiography to acquire the corresponding banding pattern and resulting sequence. The Sanger sequencing technique is based on the incorporation of dideoxynucleotide triphosphates (ddNTPs) as DNA chain terminators in a primer-driven complementary strand synthesis using modified DNA polymerases. In the beginning, incorporation of radiolabeled dNTPs was used, more recently substituted by the use of fluorescent ddNTPs or primers. New techniques are being developed using new methods, such as pyrosequencing (Margulies et al., 2005), sequencing through hybridization, use of nanopores, synthesis with multiplexing, fluorescent *in situ* sequencing, and sequencing by oligonucleotide ligation and detection, already triggering a quantum leap in the generation of sequence data.

DNA sequencing remains indeed a key technology for biotechnological approaches, generating data concerning recombinant products and whole genome analysis for a wide variety of organisms from mammals to pathogens, such as bacteria, protozoa and fungi. Other applications of DNA sequencing are genotyping, single nucleotide polymorphism mapping and pharmacogenomics. Currently, sequencing is also being used for metagenomics, which claims to analyze the genomes of all organisms existing in a specific environmental sample.

Several institutions in Brazil established core facilities mainly through participation in government funded genome projects, widening their scope and offering both large-scale and high-throughput DNA sequencing services for institutional research projects. Recently, Fiocruz established such a core facility as a "technological platform" within the institutional PDTIS program (Program for Technological Development of Products for Health). Infrastructure includes a 48-capillary 3730 DNA Sequence Analyzer (Applied Biosystems, USA) and supporting equipment such as centrifuges, thermal cyclers, etc. The aim is to support not only

large-scale users such as in genome projects, but also small- and medium-scale users. Researchers within the latter two categories do not need to perform DNA sequencing reactions, which are done by trained technologists. The users receive the resulting sequencing chromatograms and derived sequences automatically by e-mail and are able to perform basic analyses.

Different pipelines to help in the management and control of sequencing workflows have been proposed, such as Kaleidaseq (Dedhia and McCombie, 1998), VSQual (Binneck et al., 2004) or MAGIC-SPP (Liang et al., 2006). A huge amount of automated post-processing programs is available to minimize the bottleneck associated with the increase in data, resulting from new sequencing techniques. Just to name a few, preAssemble (Adzhubei et al., 2006) is helping in the assembly, POSA (Aerts et al., 2004) is a Perl pipeline to automate the assembly, and SABIA (Almeida et al., 2004), GARSA (Dávila et al., 2005), or Gene Projects (Carazzolle et al., 2007) are tools that assist from assembly to annotation. For work with ESTs, different tools were also proposed (Ayoubi et al., 2002; Berezikov et al., 2002; Hotz-Wagenblatt et al., 2003; Mao et al., 2003; Paquola et al., 2003; Matukumalli et al., 2004). Most programs have in common the quality trimming step, vector masking and the assembly of the EST data. The differences lie in implementation, data presentation and accessibility, and storage. Some programs have extra features such as annotation, automatic sequence submission or correlation of EST with microarray data. The programs use a relational database or flat files to organize the data. Most of them are based on tools such as Phred (Ewing and Green, 1998; Ewing et al., 1998) or the Staden Package (Staden, 1996) for basecalling, CAP3 (Huang and Madan, 1999), or Phrap (http://www.phrap.org) for assembly, and are programmed in PERL (http://www.perl.org).

Nevertheless, few of the existing tools have satisfied our needs to attend to also low-throughput users, and an easy workflow from sequencing data delivery up to post-processing is necessary. Therefore, we implemented a user-friendly tool for our necessities. All steps are automated - from user registration up to the possibility of sequence analysis without specific knowledge of software installation. For quality control, reports and surveys are generated automatically. This service has been available for over three years and has led to a considerable increase in sequencing quality. So far, around 155,000 sequences were obtained, from both plasmid and polymerase chain reaction products. Overall, about 72% of the templates yielded good-quality sequences with an average length of 680 nt ($Q \geq 20$ - Q20 represents a probability of 1% error), from 180 users.

The package is available on our web page (http://bioinfo.pdtis.fiocruz.br/ChromaPipe/) for local installation.

METHODOLOGY AND RESULTS

System description

We implemented our pipeline on an AMD Athlon server with 1 GB of memory and 200 GB of disc space. The automated sequencer is a 48-capillary 3730 DNA Sequence Analyzer (Applied Biosystems) with supporting equipment. Our tool is divided into three parts: sequence delivery, post-sequence analysis and monitoring (Figure 1). The tool consists of straightforward PERL programming, using the open source models such as the GD library (http://www.boutell.com/gd/) or the CGI package (http://stein.cshl.org/WWW/software/CGI). Our relational database scheme has two tables, one for the users and one for the samples. We

use an Oracle database (http://www.oracle.com/index.html), but an MySQL database (http://www.mysql.com) can also be used. It can be queried by the tools described below through a web server (e.g., Apache: http://www.apache.org) or by direct SQL statements.

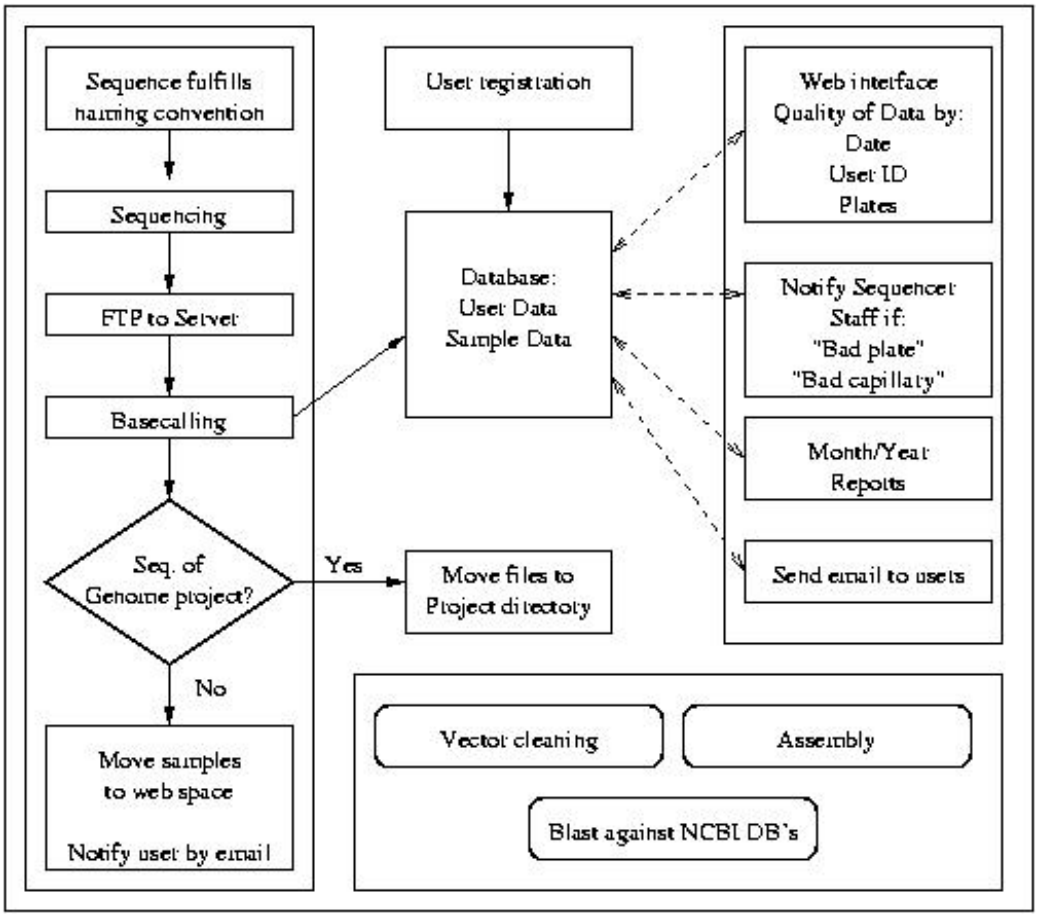

Figure 1. Main pipeline of the platform. In the center, the relational database is represented. The left column shows the user flow, from the naming convention up to the notification of the results (Figures 2 and 3). All sequencing data are saved in the database and can be retrieved by the sequencing staff - right column (Figures 5 and 6). The users can analyze their personal data - lower right box (Figure 4).

Automated analysis

The dataflow of our software is fully automated. Generated sequencing data are transferred from the sequencer computer to a server via an FTP protocol. Each chromatogram, e.g., FIO_F987_TDO_A1_F_B11.ab1, follows a naming convention: first the group name (FIO), the plate number (F987), the user ID (TDO) and finally the plate well (B11). ab1 is the ending attributed to the sequencer. Optionally, the DNA template name can be included (A1) as well

as the primer direction (e.g., F - forward, R - reverse, or W - walking). The main attributes of the chromatograms, such as the names of the sequences, ID, group, size of good-quality regions, and date are also stored in the database. To obtain the size of good-quality regions, each chromatogram is basecalled by Phred. For statistics, chromatograms without a sequence are considered "no reaction", and sequences smaller than 100 bp of Phred quality Q ≥ 20 are considered to be bad samples.

Before a user delivers DNA samples, he first registers online, giving his name, laboratory, telephone, and e-mail address. He/she then receives by e-mail a unique user ID. For each set of chromatograms the user receives an e-mail, with a link to an individual result page (Figure 2) for download and analysis.

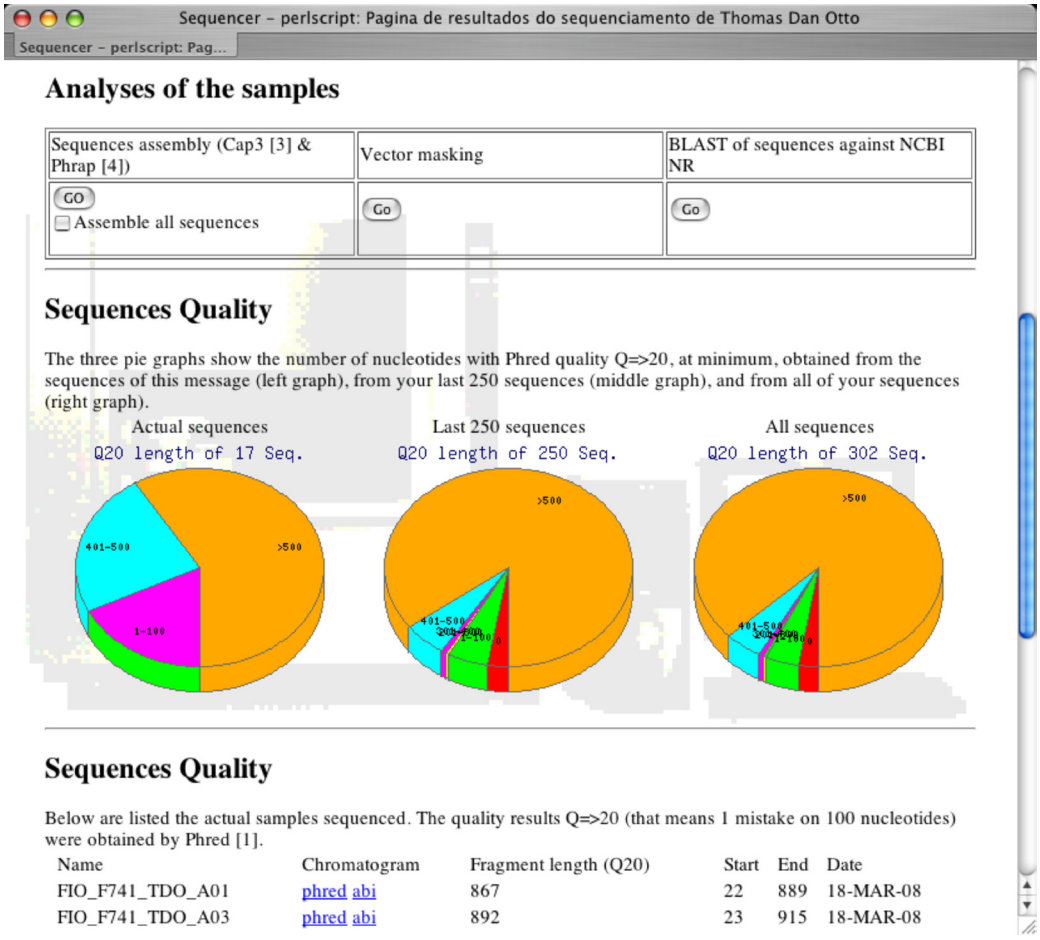

Figure 2. Notification of sequenced samples. The user can analyze generated sequences (Figure 4), monitor the history of his sequencing results or see each chromatogram individually (Figure 3).

Analysis on-demand

On the individual results page (Figure 2), the user can view his/her individual sequencing history or analyze each chromatogram online (Figure 3), for example, visualizing the chromatogram using third party software such as TraceViewer (Durbin, 1999). The nucleotide sequences obtained are given in FASTA format, with optional vector fragment masking (cross_match: http://www.phrap.com). Based on Phred quality, each base is colored accordingly, blue if the quality is equal to or above 20, red when it is below 10 and green when in between. The same chromatogram is also subject to basecalling with the ABI-basecaller (Figure 3B; http://www.appliedbiosystems.com/). The latter basecaller usually returns a better result, but not always. The user can decide on the results with which he/she wants to continue the analysis, using the appropriate links.

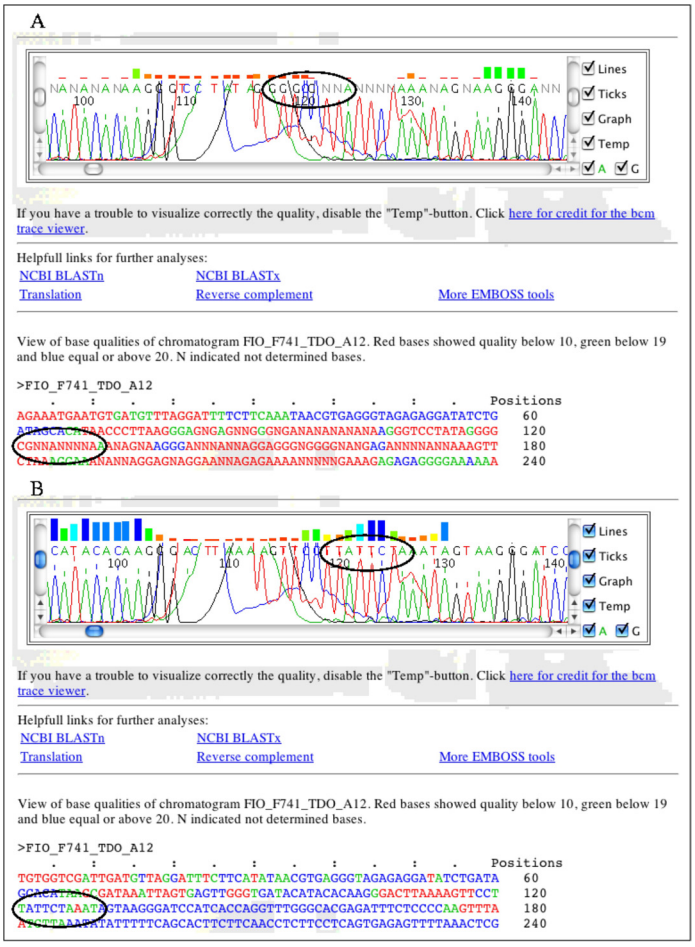

Figure 3. View of an individual chromatogram. Links to other online services are also given, as well as the sequence in FASTA format, with or without masking of vector fragments. The color-coding helps the user to visualize the quality of the bases (Red: Q < 10; Green: 10 < Q < 20; Blue: Q ≥ 20). **A.** Chromatogram basecalled with Phred. **B.** Basecalled with the ABI-basecaller. The circled regions show differences in basecalling.

Available post-processing tools can be seen in Figure 2 (top). For vector masking and quality clipping, cross_match (standard parameters) and Phred (-trim_alt 20, -trim_cutoff 20) are used. The results are available as FASTA files for cut-and-paste or direct download, with optional data regarding trimmed quality and vector masking. A similarity analysis, using BLAST (Altschul et al., 1997) against the actual NCBI NR databases (http://www.ncbi.nlm.nih.gov) can be done. The results are parsed and the first three hits are listed, including links to NCBI. Also online, an assembly of all or a subset of the chromatograms can be performed (Figure 4). The basecalling is done by Phred, and the assembly by CAP3 and Phrap. The user can analyze the assembly, download it and visualize the contig quality, the singlets or the complete assembly. Also, a graphic view of the contig quality is included, showing low-quality regions or possible polymorphisms (superimposed peaks in a high-quality region). The assemblers use the quality of the chromatograms, and if the user follows the name convention, the information about mate pairs is considered.

Other chromatograms or reference sequences in FASTA format can be included in the assembly, and subsets of chromatograms can be assembled separately by selecting chromatograms from a generated list. All post-processing data are stored as flat files on the web server for at least 6 months.

The assembly tool can also be used separately: http://bioinfo.pdtis.fiocruz.br/Assembly/.

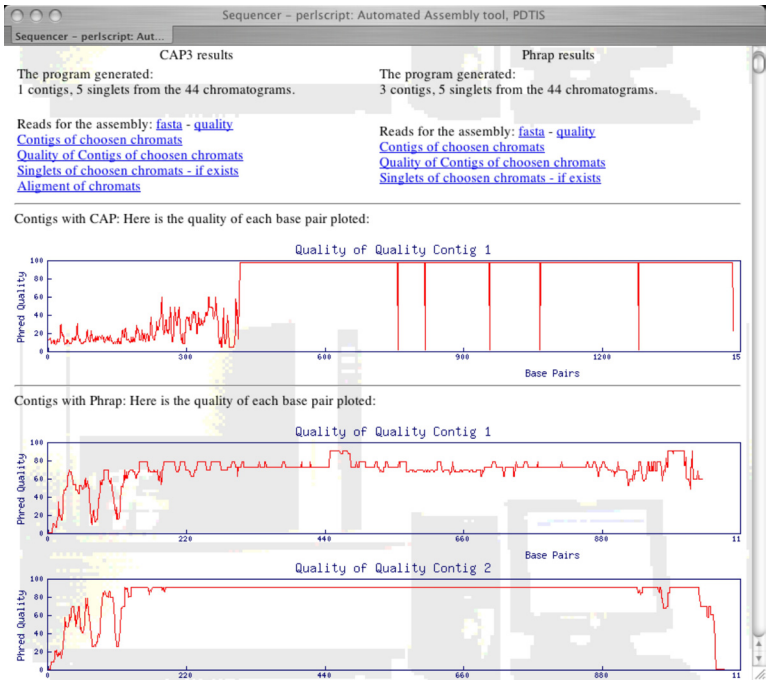

Figure 4. Assembly view. The user can select which sequences should be entered in the assembly process. New chromatograms can be included. The assembly is done with CAP3 and Phrap. The result files for the singlets and contigs are in FASTA format and the alignments are in text format. Furthermore, a quality graph is included, which shows the resulting Phred quality of each base pair in the contigs. With these chromatograms, Phrap and CAP3 generate different results.

Monitoring and feedback

All sequencing data are stored in the database. Different possibilities are available for querying. Figure 5 shows the possibility to query by date, user and/or plate. Each chromatogram can be analyzed as shown in Figure 3. For overall statistics, views are available for monthly, periodic, yearly, or all-time reports (Figure 6).

Apart from the overall statistics, different quality scans are performed daily. If a user had more than 50% bad sequencing results (sequences with less than 100 bp of Q20), or one of the sequencer's capillaries returns bad results 4 times, the staff is contacted via e-mail. Furthermore, a control sample is run on each plate to evaluate the performance of the sequencer.

So far, around 155,000 sequences were obtained, from both plasmid and polymerase chain reaction templates. Overall, about 72% of the templates yielded good-quality sequences with an average length of 680 nt ($Q \geq 20$). With plasmid DNA templates, about 75% yielded good-quality sequences with an average length of 750-1000 nt ($Q \geq 20$). Thus far, 180 users have used the facility. Around 50% of the users sequenced less than 100 samples, while 16% sequenced more than 1000 samples.

As can be seen in Figure 6, the user group 'FIO' shows more variable quality, as the low-throughput users are included in this group. When sequencing just a few samples a month, the first results are in general of poor quality. With time and assistance from the sequencing staff, the quality usually improves. In the group LBMI, composed of users who prepare and sequence whole plates, the success rate was around 95% (data not shown). As a general trend, sequencing plasmid templates generates longer sequences with higher quality. Protocols are available on the webpage: http://www.dbbm.fiocruz.br/PDTIS_Genomica/. The Leish group was a genome survey project of the protozoan *Leishmania braziliensis*, conducted shortly after installation of the new sequencer, and reflects somewhat the learning curve for use of the new facility. Of all reads, 35% were done for the *M. bovis* BCG Moreau-RJ genome project - the sequencing of the Brazilian vaccine against tuberculosis (Fiocruz - Fundação Ataulpho de Paiva collaboration). In the whole genome shotgun sequencing, more than 85% of sample

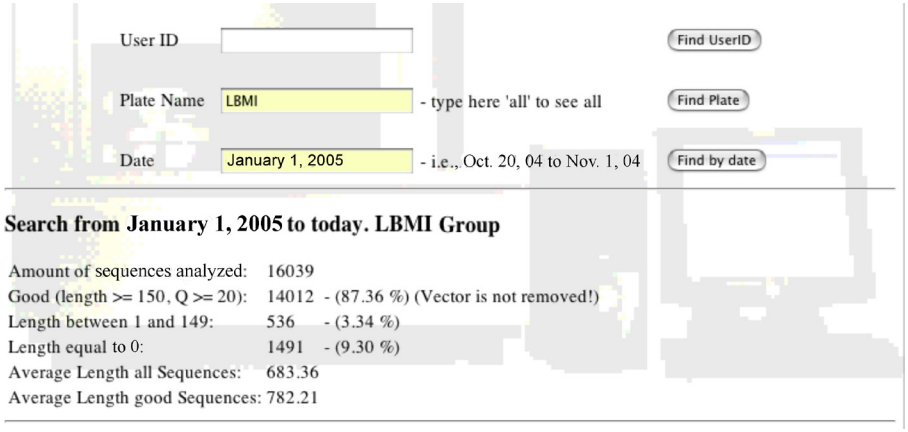

Figure 5. Page for querying chromatogram data. For each query, a short overview is given, and below a histogram and chromatogram are displayed, as shown in Figures 2 and 3.

were or good quality. This value dropped in the finishing phase, due to difficulties in the sequencing of some genomic regions with strong secondary structure or high GC bias.

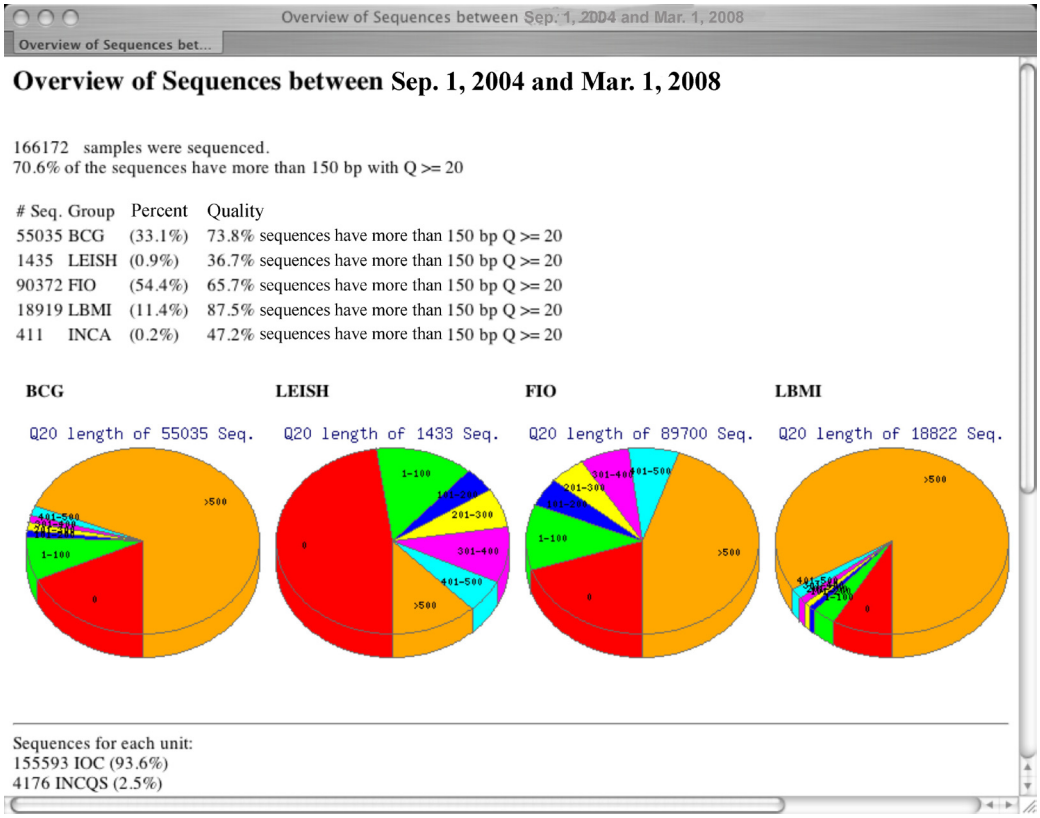

Figure 6. Report page of sequenced samples. Accounting is done according to user group and institution. Q indicates the Phred quality. This type of report can be generated for a monthly, periodic or annual overview.

Availability

The package can be downloaded on our website (http://bioinfo.pdtis.fiocruz.br/ChromaPipe/) and installed locally on a Linux server, for academic proposes. The installation is documented, but the following third party tools must be pre-installed: PhredPhrap, CAP3, BLAST, PERL, GD, MySQL database, and TraceViewer. All programs are either freely available or under an academic user agreement for non-commercial usage.

DISCUSSION

We present here an automated workflow for low- and high-throughput DNA sequencing at the Fiocruz/PDTIS core facility, starting from user and sample registration up to post-sequencing analysis. After online registration, the user delivers his/her samples to

the facility, and receives an e-mail when the samples are sequenced. On the result page, the user can perform all necessary analyses of chromatograms and further post-processing, such as BLAST comparisons, sequence trimming, vector masking, or assemblies. An important property of our interface is that the user does not depend on knowledge about computers or the installation of specific systems, or on a profound knowledge of sequence analysis. To avoid compatibility problems, the results are returned as text files for further processing. Most tools available in the literature are not designed for multiple user usage (Dedhia and McCombie, 1998; Binneck et al., 2004; Liang et al., 2006). They are designed for few users, mostly to check the flow of large genome projects. Also, the data structure of these tools is far more complex, enabling a lot of queries, which are normally not necessary for general core facilities.

Several tools, such as Magic-SPP (Liang et al., 2006), include post-processing applications. Many were implemented for EST analyses and include annotation (Hotz-Wagenblatt et al., 2003) or microarray analyses (Carazzolle et al., 2007), which is outside of our scope. Other tools, such as SABIA or GARSA, must be installed for each project. In some cases, the installation needs up to 20 other software packages. For our workflow, we need a normal Linux environment (Perl, Apache, MySQL) and four third party software (CAP3, PhredPhrap, TraceViewer, and BLAST). Also, few programs include a post-processing and a general sequence quality check, differentiated for different users and groups.

Many packages require users to learn how to handle programs and, in some cases, to learn Linux. Therefore, we implemented online post-processing. All data can be retrieved by cut-and-paste or downloaded from the browser. The advantage of our assembly approach is that the quality files and even the mate pair information are included in the analysis. Not all online assemblers use quality files, e.g., CAP3 (http://pbil.univ-lyon1.fr/cap3.php). For special, high-throughput analyses, a bioinformatics platform is also available (Otto et al., 2007), including genome assemblies, annotations, or more complex comparisons.

In conclusion, this paper describes a multi-user pipeline for DNA sequencing and chromatogram processing, analysis and quality control. Our tool is designed to attend to not only high-throughput users but also scientists who need to sequence and analyze a few samples without expert knowledge of this process.

ACKNOWLEDGMENTS

We thank the PDTIS, CAPES and CNPq for financial support. We appreciated the helpful comments of the anonymous referees.

REFERENCES

Adzhubei AA, Laerdahl JK and Vlasova AV (2006). preAssemble: a tool for automatic sequencer trace data processing. *BMC Bioinformatics* 7: 22.

Aerts JA, Jungerius BJ and Groenen MA (2004). POSA: perl objects for DNA sequencing data analysis. *BMC Genomics* 5: 60.

Almeida LG, Paixao R, Souza RC, Costa GC, et al. (2004). A system for automated bacterial (genome) integrated annotation - SABIA. *Bioinformatics* 20: 2832-2833.

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.

Ayoubi P, Jin X, Leite S, Liu X, et al. (2002). PipeOnline 2.0: automated EST processing and functional data sorting.

Nucleic Acids Res. 30: 4761-4769.
Berezikov E, Plasterk RH and Cuppen E (2002). GENOTRACE: cDNA-based local GENOme assembly from TRACE archives. *Bioinformatics* 18: 1396-1397.
Binneck E, Silva JF, Neumaier N, Farias JR, et al. (2004). VSQual: a visual system to assist DNA sequencing quality control. *Genet. Mol. Res.* 3: 474-482.
CAP3 Sequence Assembly Program. http://pbil.univ-lyon1.fr/cap3.php. Accessed November 2007.
Carazzolle MF, Formigheri EF, Digiampietri LA, Araujo MRR, et al. (2007). Gene Projects: a genome web tool for ongoing mining and annotation applied to CitEST. *Genet. Mol. Biol.* 30: 1030-1036.
CGI.pm - a Perl5 CGI Library. http://stein.cshl.org/WWW/software/CGI. Accessed November 2007.
ChromaPipe. The Chromatogram Pipeline. http://bioinfo.pdtis.fiocruz.br/ChromaPipe/. Accessed March 2008.
Chromatogram Assembly Online. http://bioinfo.pdtis.fiocruz.br/Assembly/. Accessed March 2008.
Dávila AM, Lorenzini DM, Mendes PN, Satake TS, et al. (2005). GARSA: genomic analysis resources for sequence annotation. *Bioinformatics* 21: 4302-4303.
Dedhia NN and McCombie WR (1998). Kaleidaseq: a Web-based tool to monitor data flow in a high throughput sequencing facility. *Genome Res.* 8: 313-318.
Durbin KJ (1999). BCM Trace Viewer. http://www.hgsc.bcm.tmc.edu/downloads/software/trace_viewer/. Accessed November 2007.
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.
GD Graphics Library. http://www.boutell.com/gd. Accessed November 2007.
Genome Sciences Department, University Washington. http://www.phrap.org. Accessed November 2007.
Hotz-Wagenblatt A, Hankeln T, Ernst P, Glatting KH, et al. (2003). ESTAnnotator: a tool for high throughput EST annotation. *Nucleic Acids Res.* 31: 3716-3719.
Huang X and Madan A (1999). CAP3: A DNA sequence assembly program. *Genome Res.* 9: 868-877.
Liang C, Sun F, Wang H, Qu J, et al. (2006). MAGIC-SPP: a database-driven DNA sequence processing package with associated management tools. *BMC Bioinformatics* 7: 115.
Mao C, Cushman JC, May GD and Weller JW (2003). ESTAP - an automated system for the analysis of EST data. *Bioinformatics* 19: 1720-1722.
Margulies M, Egholm M, Altman WE, Attiya S, et al. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437: 376-380.
Matukumalli LK, Grefenstette JJ, Sonstegard TS and Van Tassell CP (2004). EST-PAGE - managing and analyzing EST data. *Bioinformatics* 20: 286-288.
Maxam AM and Gilbert W (1977). A new method for sequencing DNA. *Proc. Natl. Acad. Sci. U.S.A.* 74: 560-564.
MySQL. http://www.mysql.com/. Accessed November 2007.
NCBI. National Center for Biotechnology Information. http://www.ncbi.nlm.nih.gov. Accessed November 2007.
Oracle database. http://www.oracle.com/index.html. Accessed November 2007.
Otto TD, Catanho M, Miranda AB and Degrave WM (2007). The PDTIS bioinformatics platform: from sequence to function. *RECIIS* 1: 286-294.
Paquola AC, Nishyiama MY Jr, Reis EM, da Silva AM, et al. (2003). ESTWeb: bioinformatics services for EST sequencing projects. *Bioinformatics* 19: 1587-1588.
Plataforma de Seqüenciamento de DNA PDTIS/FIOCRUZ. http://www.dbbm.fiocruz.br/PDTIS_Genomica. Accessed November 2007.
Sanger F, Nicklen S and Coulson AR (1977). DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U.S.A.* 74: 5463-5467.
Staden R (1996). The Staden sequence analysis package. *Mol. Biotechnol.* 5: 233-241.
The Apache Software Foundation. http://www.apache.org. Accessed November 2007.
The Perl Directory at Perl.org. http://www.perl.org. Accessed November 2007.