Genomic Exploration of the Hemiascomycetous Yeasts: 2. Data generation and processing

François Artiguenave[1,*], Patrick Wincker[1], Phillippe Brottier, Simone Duprat, Fabien Jovelin, Claude Scarpelli, Jean Verdier, Virginie Vico, Jean Weissenbach, William Saurin

GENOSCOPE, Centre National de Séquençage, 2 rue Gaston Crémieux, P.O. Box 91, F-Evry Cedex, France

Received 3 November 2000; accepted 9 November 2000

First published online 27 November 2000

Edited by Horst Feldmann

Abstract The generation of sequencing data for the hemiascomycetous yeast random sequence tag project was performed using the procedures established at GENOSCOPE. These procedures include a series of protocols for the sequencing reactions, using infra-red labelled primers, performed on both ends of the plasmid inserts in the same reaction tube, and their analysis on automated DNA sequencers. They also include a package of computer programs aimed at detecting potential assignation errors, selecting good quality sequences and estimating their useful length.

Key words: High-throughput sequencing; Sequence analysis; Genome; Hemiascomycete

1. Introduction

Genome shallow sequencing, based upon simple pass insert ends sequencing, rapidly provides data for comparisons to a reference genome. To obtain enough and useful data by single pass reads, the efficiency and quality of high-throughput sequencing is of special importance. The methods we have implemented for this project were determined by this constraint. First, in order to obtain the most information possible for every clone, the reads should be as long as possible. Secondly, the identification (i.e. name) of the sequences should be assigned with the smallest possibility of error, thereby allowing accurate interpretation on species identity and genome synteny.

2. Materials and methods

2.1. Biochemical methods

2.1.1. Plasmid DNA purification. 5 μl of the bacterial stock in glycerol/culture medium was used to inoculate 1.2 ml of 2×YT medium present in deep-well plates (96-well format, Polyfiltronics). These plates were incubated at 37°C with constant agitation during 16 h. The cells were then pelleted at 4000 rpm at 4°C in an Eppendorf 5810R centrifuge. After removing the medium by inversion, the plates were sealed with an adhesive pad and placed at −20°C for at least 1 h. Subsequently, each well received 100 μl of buffer P1 (Qiagen) supplemented with 50 μg/ml of RNase A stocked at 4°C, using a Hydra 96 dispenser (Robbins Scientific). The plates were placed on a rotary shaker (Bellco Instruments) for 15 min to allow complete resuspension. Then 200 μl of buffer P2 (Qiagen) was dispensed in each well using a second Hydra 96 dispenser, and the plates were placed on a rotary shaker for 7 min. Aliquots of 200 μl of P3 solution (Qiagen) at 4°C were then dispensed using a third Hydra 96 dispenser, the plates were placed 5 min under rotary agitation, and centrifuged for 46 min at 4100 rpm at 4°C in a Jouan KR422 centrifuge. The supernatants were then pipetted using a Hydra 96 robot, and dispensed into a new 96-deep-well plate. The DNA was precipitated using 300 μl of isopropanol and pelleted at 4100 rpm and 20°C in a Jouan KR422 centrifuge for 10 min. After a washing cycle with 70% ethanol, the pellets were dried and resuspended in 30 μl of TE buffer.

2.1.2. Sequencing reactions. All sequences were performed on LiCor 4200L DNA sequencers. The primary reason of this choice was the incomparable read lengths obtained with these machines, where 900–1000 bp of good quality (see Section 2.2.3) can routinely be produced for every read. A second reason came from the bi-directional protocol that was followed which enabled sequencing of both extremities of the DNA inserts in the same tube. This avoided the separate loading and analysis of the two opposite reactions, which can lead to many errors and thereby complicates the synteny analysis.

The reactions were performed with the Thermosequenase reagents kit (Amersham Pharmacia Biotech), 6% DMSO (Sigma), two primers labelled with either IRD700 or IRD800 (MWG Biotech), in a PE Biosystems 9700 Thermocycler using 96-well plates. After completion of the cycles, the reaction plates were dried under vacuum and resuspended in the loading buffer (Amersham Pharmacia Biotech). No plate transfers were performed at any step. The reactions were then loaded on 4.8% Rapid Gel XL gels (Amersham Pharmacia Biotech) and electrophoresed at 3200 V for 10–12 h. The collected data were analysed on a series of dedicated computers using the LiCor Base ImagIR v4.1 package.

2.1.3. Controlling the trace identities. Faithful sequence identification is of great importance for any random sequence tag project based on single pass sequencing. In the present case, it was necessary to control the identity of the sequenced plates to avoid misinterpretations due to sequencing clones belonging to the wrong species. It was also necessary to ensure that all reads were correctly tracked to avoid incorrect synteny assessments. We chose to re-sequence six clones for every 96-well plate, and compare these sequences to those obtained in the first experiment. The positions of the six clones are indicated in Fig. 1. This experiment enabled us to check all the potential exchange of plates across the whole process of sequencing, as well as the potential errors in the loading of the sequencers. When the two sequence versions were different (see Section 2.2.7.2), the whole plate was newly sequenced to resolve the ambiguity.

2.2. Sequence data processing

Sequence data processing involves a number of treatments which are designed in such a way that the resulting data set is obtained in a timely manner, and ultimately allows the extraction of biologically significant information. The volume of data produced in high-throughput sequencing centres, such as GENOSCOPE, raises difficulties with regard to data management, data integrity and treatment

*Corresponding author.
E-mail: artigue@genoscope.cns.fr

[1] These authors contributed equally to this work.

0014-5793/00/$20.00
PII: S0014-5793(00)02273-0

automation. In this section, we present the processing scheme, designed at GENOSCOPE, which strives to meet the above-stated goals.

Processing starts with the conversion of trace files to raw sequence files, and terminates with the final sequence being exported out of the sequencing project to its final destination like the web site or sequence databanks. Tasks such as base calling, quality checking, vector scanning, contamination checking, empty vector detection, plate contamination, and other error detection (Fig. 2 and below) are applied sequentially. These operations have previously been referred to as pre-processing or pre-assembly in shotgun sequencing projects [1,2]. The near-full automation of these tasks considerably reduced the time spent in managing the sequences, and is now being used for most of GENOSCOPE's projects.

2.2.1. Transfer. Two individual sequencers share an associated pilot PC computer running OS/2®. Gel image analysis and lane tracking are performed, after data collection, on separated and dedicated computers. SCF trace files [3], which are generated during this analysis, are automatically transferred from the analysis machines to the Compaq Tru64™ UNIX® servers by way of FTP (File Transfer Protocol) or CAP (Appleshare server for UNIX®). Upon arrival on the servers, the trace files are base-called using *Phred* [4] and a quality check is run by analysing the output [5]. Sequences not meeting our production quality criteria (at least a 100 bases window with more than 75 bases called with a quality over 20) are discarded. Sequence names are extracted from the validated SCF trace files and checked for conformance to the GENOSCOPE-standardised nomenclature. This nomenclature allows easy parsing of the template information, project and library identifications, which are then used to dispatch the trace files to the appropriate project directory.

2.2.2. Initial set-up. Subsequent analyses are carried out in a individual project environment allowing the various parameter settings to meet project-specific objectives. The software environment is set-up during project initialisation by editing configuration files describing each of the different steps to be performed. The configuration takes into account biological information concerning project materials (origin of the DNA, sequencing vectors, cloning vectors, insert size, etc.) as well as other project-related data such as project code, library code, number and format of plates or verifications to carry out.

2.2.3. Sequence analysis. Base calling and quality clipping: this module serves to determine the nucleotides from the trace files and to identify poor quality regions at the start and the end of each read. Information on the type of sequencer used for each sequence can be extracted from the trace files using a program called *scf2ps* (available upon request). This information is used to determine which quality clipping method will be applied, depending on the sequencer. For samples read on LI-COR® machines, the LI-COR® base calling is extracted from the trace file using *scf2ps*. The start of a useful region is fixed as the beginning of the first 60 bp window on the read, sliding from 5′ to 3′, with no more than one ambiguous base (non-ATGC). The end of the region is the last base of the first 60 bp window, sliding from 3′ to 5′, which satisfies the same criteria.

2.2.4. Sequencing-vector clipping. The sequencing vector is identified by using the *Lassap* (Gene-IT S.A., Le Chesnay, France) implementation of the Smith and Waterman algorithm [6] to compare the sample sequence to the expected short vector cloning site flanking sequence (20 bp). Alignments are considered valid only if they contain a minimum of eight bases which match exactly, and contain no more than six mismatches or gaps (threshold score = 40, match = 5, mismatch = −4, gap = −4). These parameters were adjusted to detect a short motif in a potentially poor quality region. Furthermore, to limit false positive results, the retained matches must involve the first 100 bases of the read and the correct extremity (typically the T7 or SP6 end) of the vector, depending on the primer used. Finally, adjusted clipping positions are calculated to correct possible incomplete alignments due to a poor quality sequence.

2.2.5. Sequencing-vector screening. Positive *blast* [7] matches between vector-clipped reads (based on the previous test) and the vector sequence are reported if the alignment percent identity is greater than 85% over at least 100 bases. *Blast* (*Lassap* implementation) is set-up to extend exact-matching seeds of at least 20 bp long (W = 20, S = 100).

2.2.6. Contamination screening. *Escherichia coli* transposon contaminants, such as Tn10, are identified by comparing the vector-clipped reads against databases using *blast*. In order to reduce the number of false positives, the *blast* parameters are tuned both on the organism sequenced, and the databank screened.

2.2.7. Plate processing. The following steps, neighbourhood test and plate control are carried out using the plate as the functional unit of comparisons. These aim to detect plate handling errors and cross-well contamination. Sequences passing these tests are written to a plate-specific FASTA-formatted file.

2.2.7.1. Neighbourhood test. Vector-clipped reads from neighbouring clones, primed from the same end of the vector, displaying significant homology are reported. Homology is checked using *blast* (*Lassap*) with the following parameters: match = 5, mismatch = −5, W = 20, S = 100 and X = 8. Hits are retained only if the starting positions of the matches on each of the sequences are within 100 bases of each other.

2.2.7.2. Misnamed plate and mistracked run detection. Six clones per 96-well plate are re-arrayed and independently sequenced (Fig. 1). Reads are processed as for the production sequences, and all checks, with the exception of the control step, are carried out. These sequences are then pooled to create a reference databank against which production sequences are compared. The test consists of a *blast* comparison (with the same parameters as for the neighbourhood test) between the plates of production sequences and the control bank. A plate is automatically validated if all of the duplicate reads, those

Table 1
Hemiascomycetes sequencing report

Species	PR[a]	CO	VE	COLI	RE	XC	VA	VA (%)
Saccharomyces uvarum	5 620	338	50	3	21	10	5 140	91.5
Saccharomyces exiguus	2 832	145	9	1	95	16	2 579	91.1
Saccharomyces servazzii	2 795	150	12	0	39	48	2 570	91.9
Zygosaccharomyces rouxii	5 402	287	235	10	55	154	4 934	91.4
Saccharomyces kluyveri	2 754	170	8	1	29	4	2 528	91.8
Kluyveromyces thermotolerans	3 007	200	124	17	16	20	2 653	88.5
K. lactis	6 144	386	46	18	64	4	6 080	99.0
Kluyveromyces marxianus	2 678	178	42	4	0	19	2 493	93.2
Pichia angusta	5 435	374	38	0	0	6	5 082	93.5
Debaryomyces hansenii	3 302	206	15	1	371	19	2 830	85.7
P. sorbitophila	6 533	313	818	19	11	898[b]	4 829	73.9
Candida tropicalis	2 838	176	52	2	21	8	2 541	89.5
Yarrowia lipolytica	5 221	348	78	2	12	71	4 940	94.6
Total	54 561	3 271	1 527	78	734	1 277	49 199	90.2

Results of data processing for individual yeast species are presented on each line. The last one shows results for the overall project. The first column contains the species names. Each of the following columns represents the number of reads which were: processed (PR); controls (CO); vector flagged (VE); *E. coli* tagged (COLI); redundant (RE, the insert ends were read more than once); failed due to probable cross-well contamination (XC, i.e. positive neighbourhood test, see text); validated for subsequent analyses (VA). The rightmost column is the percentage of valid reads (VA (%)).

[a]Processed reads do not include control sequences.

[b]High number of XC flagged sequence is a bias due to the high frequency of vector reads.

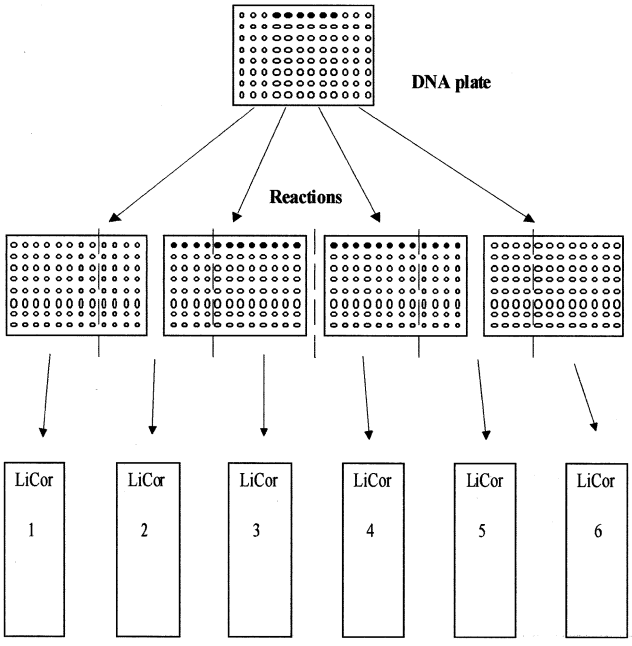

Fig. 1. Distribution of the sequenced clones from the original DNA plates to the LI-COR sequencers. One 96-well plate is split into four reaction plates each containing the four reactions for three columns of the first plate. One LI-COR sequencer is then loaded with 64 samples, corresponding to 2/3 of one reaction plate, or to 1/3 of two consecutive reaction plates. The positions of the clones re-sequenced for controlling the data are shown as dark spots.

present in both the reference bank and the plate to be checked, display a positive match. Ambiguous matches which involve different clones, or control sequences displaying no matches, reveal a potential problem during the sequencing process, such as plate inversion or tracking failure. Such situations require human intervention in order to resolve all ambiguities.

Finally, a FASTA-formatted file is written with all valid and trimmed sequences. A post-processing filter is applied to eliminate potential sequence redundancy (the same template read more than once). In this case, only the longest read is kept.

Table 2
Total number of clones for which at least one sequence was obtained as well as the number (and percentage) of clones with both insert ends successfully sequenced are reported for each individual yeast species

Species	Clones	Both ends sequenced clones
S. uvarum	2 705	2 435 (90.0%)
S. exiguus	1 389	1 190 (85.7%)
S. servazzii	1 392	1 178 (84.6%)
Z. rouxii	2 588	2 348 (90.7%)
S. kluyveri	1 315	1 213 (92.2%)
K. thermotolerans	1 439	1 221 (84.9%)
K. lactis	3 190	2 890 (90.6%)
K. marxianus	1 304	1 190 (91.3%)
P. angusta	2 674	2 410 (90.1%)
D. hansenii	1 510	1 320 (87.4%)
P. sorbitophila	2 666	2 159 (81.0%)
C. tropicalis	1 456	1 266 (87.0%)
Y. lipolytica	2 656	2 284 (86.0%)
Total	26 284	23 104 (87.9%)

The last line shows results for the overall project.

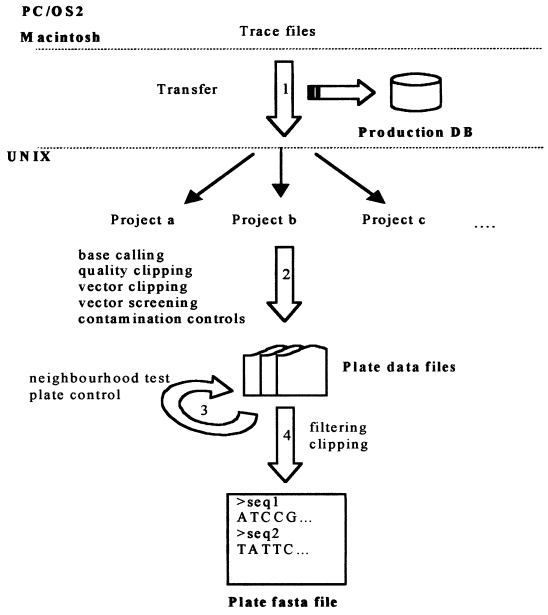

Fig. 2. Data processing involves numerous treatments applied sequentially: trace files are automatically transferred from the sequencers to the UNIX servers, and dispatched to project-specific directories (arrow 1). Within each project environment, sequences are processed independently and grouped by plate of origin (arrow 2) in a flat file database. This file contains the sequence itself and information acquired by the different operations. Upon completion of a plate, specific controls are carried out (arrow 3). Valid, vector-cleaned sequences are then written to a FASTA-formatted file (one file per plate, arrow 4).

3. Results and discussion

Sequence analysis results for the project, reported in Table 1, show overall low library biases and contamination with the exception of the high numbers of empty vectors detected in the *Pichia sorbitophila* reads. One can explain this observation by an instability of the DNA when cloned into *E. coli* [8]. Moreover, the plate-level tests revealed few handling errors. Specifically, there were four plate inversions and one plate replacement on a total of 327 plates sequenced. This amounts to 2.75% of errors due to plate misidentification, all of which were corrected. We also observed few ambiguities concerning only a part of a plate and attributed to some sequencing run problems. Twenty one such errors were identified and were either manually corrected or led to the erroneous data being

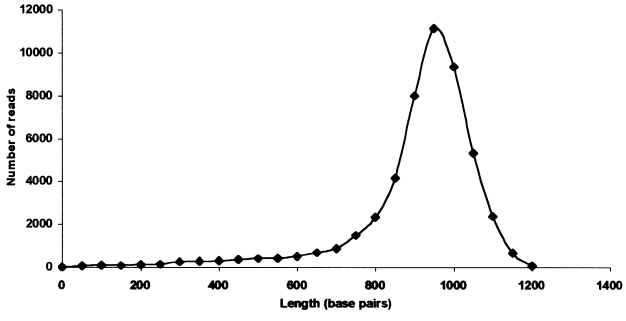

Fig. 3. Sequence length frequency distribution. Length was estimated on 49 400 sequences. Sequences are vector-clipped and trimmed according to our quality standard (see text). Average length is 910 bp.

discarded. Given the position of the control clones, we were only able to verify 2/3 of the runs (Fig. 1). However, we can estimate that the number of misidentified sequences remaining in the final data set does not exceed 0.35% (1/3 non-checked runs × 21 detected errors/1962 total runs). All the detected errors concern successive plates or runs, and never led to sequences from one species genome being attributed to another species. In addition, the control sequences were performed according to the same bi-directional protocol and failed to detect any error limited to one of the two twin reactions. This implies that robust synteny assessments can be concluded from this data set. Furthermore, the neighbourhood tests revealed an important bias during the sequencing of a first library for *Kluyveromyces lactis*. Therefore, data from this library were not included in the final pool of sequences and were substituted by a second library which was constructed de novo and sequenced.

In conclusion, the length and number of reads sequenced (Fig. 3) met the requirements of the sequencing project [9]. The bi-directional protocol used in the LI-COR® sequencers provided a high number of double-end sequenced clones (Table 2), and guarantees that end sequences can be unambiguously paired.

Acknowledgements: We are specially grateful to Michael Levy for critical reading of the manuscript and insightful comments. We also thank Marcel Salanoubat, Eric Pelletier, Olivier Jaillon and Thomas Bruls for helpful discussions. Part of this work was supported by a BRG Grant (ressources génétiques des microorganismes no. 11-0926-99).

References

[1] Wendl, C.M., Dear, S., Hodgson, D. and Hillier, L. (1998) Genome Res. 8, 975–984.
[2] Bonfield, J.K. and Staden, R. (1996) DNA Seq. 6, 109–117.
[3] Dear, S. and Staden, R. (1992) DNA Seq. 3, 107–110.
[4] Ewing, B.G., Hillier, L., Wendl, M.C. and Green, P. (1998) Genome Res. 8, 175–185.
[5] Ewing, B.G. and Green, P. (1998) Genome Res. 8, 186–194.
[6] Smith, T.F. and Waterman, M.S. (1981) J. Mol. Biol. 147, 195–197.
[7] Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) J. Mol. Biol. 215, 403–410.
[8] de Montigny, J., Spehner, C., Souciet, J.L., Tekaia, F., Dujon, B. et al. (2000) FEBS Lett. 487, 87–90 (this issue).
[9] Souciet, J.L., Aigle, M., Artiguenave, F., Blandin, G., Bolotin-Fukuhara, M. et al. (2000) FEBS Lett. 487, 3–12 (this issue).