PABS: An online platform to assist BAC-by-BAC sequencing projects

Sara Todesco[1], Davide Campagna[2], Fabrizio Levorin[2], Michela D'Angelo[2], Riccardo Schiavon[2], Giorgio Valle[1,2], and Alessandro Vezzi[1]

[1]*Department of Biology, University of Padova and* [2]*CRIBI Biotechnology Centre, University of Padova, Padova, Italy*

BioTechniques 44:60-64 (January 2008)
doi 10.2144/000112686

Genome sequencing projects are either based on whole genome shotgun (WGS) or on a BAC-by-BAC strategy. Although WGS is in most cases the preferred choice, sometimes the BAC-by-BAC approach may be better because it requires a much simpler assembly process. Furthermore, when the study is limited to specific regions of the genome, the WGS would require an unjustified effort, making the BAC-by-BAC the only feasible strategy. In this paper we describe an informatics pipeline called PABS (Platform Assisted BAC-by-BAC Sequencing) that we developed to provide a tool to optimize the BAC-by-BAC sequencing strategy. PABS has two main functions: (i) *PABS-Select, to choose suitable overlapping clones; and* (ii) *PABS-Validate, to verify whether a BAC under analysis is actually overlapping the neighboring BAC.*

The whole genome shotgun (WGS) strategy (1) is in most cases the preferred choice for genomic sequencing; however, in some cases the BAC-by-BAC approach (2) may be a better choice, especially when complex repeated regions must be resolved or when the study is limited to specific regions of the genome.

The BAC-by-BAC strategy consists of shotgun sequencing of individual adjacent BACs that cover the region of interest with a minimal but at the same time significant overlap between clones. To generate the minimal "tiling path," two approaches have been proposed (3): (*i*) the physical mapping approach, which requires the complex and laborious construction of a physical map (typically by BAC fingerprinting) to sort and select a series of clones (the "tiling path") before starting the sequencing process; and (*ii*) the walking approach, which requires direct sequencing without a priori knowledge of the clone position in the genome. In the latter case, the BAC library must be characterized by sequencing the ends of each insert, resulting in a database of BAC-end sequences (BES). After sequencing a BAC, it is possible to identify all the overlapping BES. Therefore the walking can start from "seed" BACs to extend bidirectionally on overlapping clones identified by their BES.

A key step in the BAC-by-BAC sequencing is the identification of reliable neighboring BACs. Often this process is difficult due to the presence of repeats, leading to misalignment of BACs and possible "jumps" along the genome. The analysis of repeats can be performed using RepeatMasker (www.repeatmasker.org) or similar tools able to identify known repeats. However, for those genomes not yet extensively studied, the repeated regions are not well characterized and their direct identification is impossible.

In this paper we describe the implementation of PABS for the International

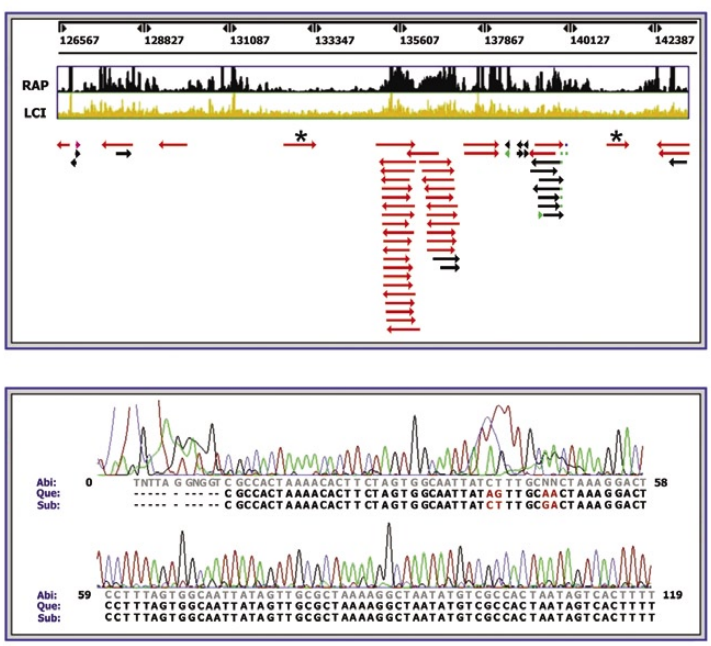

Figure 1. Screenshots from PABS-Select. (A) After uploading the initial sequence (typically the sequence of the BAC or the end to be extended) the application returns a graphical representation of the sequence, including the Repeat Analysis Program (RAP) Index (reflecting the repetitiveness of a region) and the Low Complexity Index (LCI, indicating the presence of low complexity regions such as homopolymers and microsatellites). To simplify the figure, only the terminal 16 kb of a 143 kb BAC insert are shown. The entire database of BAC-end sequences (BES) is preloaded on the system, thus allowing an automatic BLASTn search to align on the initial BAC all the matching BES, represented by arrows in the figure. This gives an immediate view of the possible overlapping BACs, the arrows pointing to the direction of the overlap. The extent of each arrow represents the region of overlap, while the color indicates the BLASTn score: red = >200; violet = 200–80; green = 80–50; blue = 50–40 and black = <40. The final aim is to find at each end of the input sequence a suitable overlapping BAC. Therefore, the best candidates will be those corresponding to BES with the following features: (*i*) direction toward the end of the initial BAC; (*ii*) position in a region with low RAP and LCI indexes; and (*iii*) appropriate extent of the overlap. The asterisks indicate two suitable candidates. By clicking on an arrow, the BES electropherogram aligned to the input sequence is displayed, as partially shown in (B). The query sequence (Que) corresponds to the initial BAC taken as input, while the subject (Sub) is the aligned BES as stored in the database. Moreover, the "Abi" sequence refers to the same BES, generated with the standard Applied Biosystems (Foster City, CA, USA) base caller. This allows an accurate inspection of any discrepancy between the two aligned sequences; for instance, the mismatching bases between query and subject (red colored) would indicate considerably different sequences, but the analysis of the electropherogram shows a likely perfect match of the two sequences.

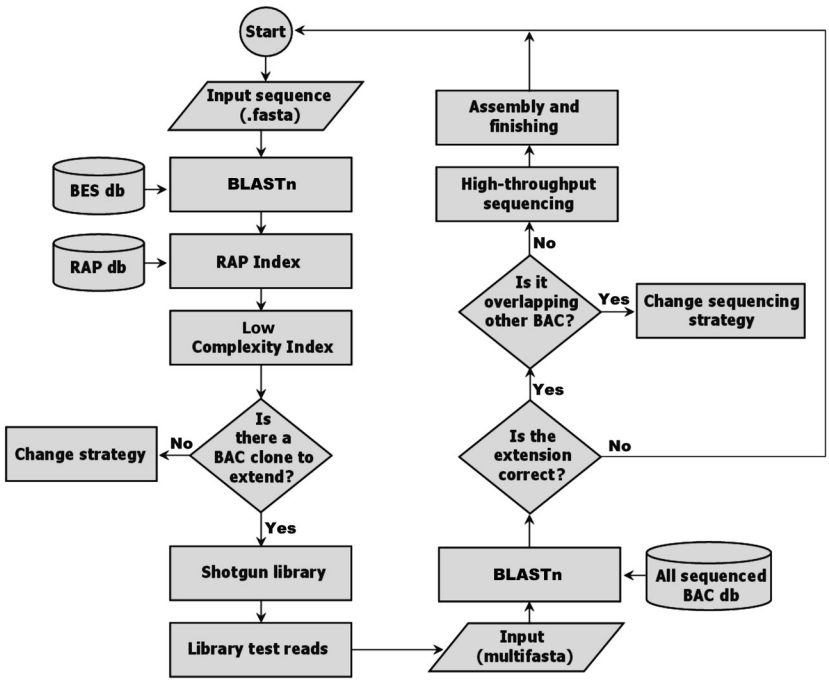

Figure 2. Schematic overview of the dataflow used in PABS. Databases are drawn as bins, rectangles represent applications; direction of dataflow is indicated by connectors. The identification of candidate extension clone is based on BLASTn analysis of the input sequence against the BAC-end sequences (BES) database, and on calculation of Repeat Analysis Program (RAP) Index and Low Complexity Index. The candidate BAC clones are then shotgun sequenced. A first set of 96 clones from the shotgun library is sequenced and a multifasta format of these is processed by the PABS-Validate, using BLASTn against different types of databases.

Tomato Genome Project. The project is based on a BAC-by-BAC sequencing strategy and relies on a BES database (more than 310,000 sequences), but lacks a robust physical map (4,5). Our group is involved in the sequencing of chromosome 12; at the time of writing, it has successfully used PABS for 33 rounds of walking, without any error in the extension process.

PABS uses BLASTn (6) to analyze a fully or partially sequenced clone (hereafter referred to as "initial BAC") against the BES database. PABS-Select takes as "input sequence" the initial BAC (either the complete sequence or the end under investigation) and returns a graphical representation of the position and orientation of the BES (represented as oriented arrows) overlapping the input sequence (Figure 1A).

An innovative feature of PABS is its ability to integrate the BES analysis with the presence of repetitive sequences. In particular, PABS identifies repeated regions with the Repeat Analysis Program (RAP) (7) and calculates the Low Complexity Index as one minus the Linguistic Complexity Index (8). The RAP Index gives an estimate of the "repetitiveness" of a DNA region. It is calculated for each position of the input sequence by means of a de novo analysis that does not require any previous knowledge about repeats. PABS displays the results of BLASTn and RAP, thus allowing a more reliable selection of adjacent clones. The choice will be addressed to BACs with a suitable overlap to the initial BAC and with the aligned BES positioned in a low-repeat region.

To make the selection easier and faster, PABS allows a direct visualization of the BES electropherogram aligned with the input sequence (Figure 1B). In this way the user can quickly evaluate sequences of poor quality that may be the cause of misleading BLASTn results. In addition, an automated procedure collects and summarizes all the available information on the candidate BACs (insert length, genetic markers, FISH data, sequencing status) to optimize the selection for the extension.

The selected BAC is then sequenced with a shotgun approach. To further validate the selection, we have designed PABS-Validate. Typically, the first set of 96 shotgun sequences produced from the selected BAC are submitted as a multifasta file to PABS-Validate and analyzed using BLASTn against three databases: the initial BAC, the finished BACs (i.e., all the finished BACs of the Tomato Genome Project), and the partially sequenced BACs (i.e., the BACs under sequencing). Three types of controls can be made: (*i*) some of the reads should fall into the overlapping region of the initial BAC, thus confirming a correct walking; (*ii*) no reads should significantly match other sequenced BACs belonging to different genomic regions, because this would indicate a possible jump to another region; and (*iii*) as an exception to the previous point, when several extensions are carried out simultaneously from different seeds, we expect that eventually the different walks could merge; therefore we must also consider this event and the consequent possibility to work out the extent of the overlap at the two ends of a bridging BAC.

A complete scheme of the PABS flowchart is represented in Figure 2.

In conclusion, PABS offers two main features:

- it makes the process of generating a reliable minimal tiling path of BACs more robust since it is specifically designed to deal with repetitive sequences;
- it allows a series of validations at the beginning of the shotgun sequencing of each BAC, minimizing the possibility of mistakes and optimizing the merging of overlapping BACs.

PABS is freely accessible at http://tomato.cribi.unipd.it/files/bioinformatics.html, where further detailed instructions are also available. At the moment, the pipeline has been implemented only for the Tomato Sequencing Project but its modular structure would allow easy adaptation to other projects

based on a clone-by-clone sequencing strategy.

COMPETING INTERESTS STATEMENT

The authors declare no competing interests.

ACKNOWLEDGEMENTS

This research is supported by the Fondo per gli Investimenti della Ricerca di Base (grant no. RBLA0345SF).

REFERENCES

1. **Venter, J.C., M.D. Adams, E.W. Myers, P.W. Li, R.J. Mural, G.G. Sutton, H.O. Smith, M. Yandell, et al.** 2001. The sequence of the human genome. Science *291*:1304-1351.
2. **Lander, E.S., L.M. Linton, B. Birren, C. Nusbaum, M.C. Zody, J. Baldwin, K. Devon, K. Dewar, et al.** 2001. Initial sequencing and analysis of the human genome. Nature *409*:860-921.
3. **Batzoglou, S., B. Berger, J. Mesirov, and E.S. Lander.** 1999. Sequencing a genome by walking with clone-end sequences: a mathematical analysis. Genome Res. *9*:1163-1174.
4. **Mueller, L.A., T.H. Solow, N. Taylor, B. Skwarecki, R. Buels, J. Binns, C. Lin, M.H. Wright, et al.** 2005. The SOL Genomics Network: a comparative resource for Solanaceae biology and beyond. Plant Physiol. *138*:1310-1317.
5. **Budiman, M.A., L. Mao, T.C. Wood, and R.A. Wing.** 2000. A deep-coverage tomato BAC library and prospects toward development of an STC framework for genome sequencing. Genome Res. *10*:129-136.
6. **Altschul, S.F., T.L. Madden, A.A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman.** 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. *25*:3389-3402.
7. **Campagna, D., C. Romualdi, N. Vitulo, M. Del Favero, M. Lexa, N. Cannata, and G. Valle.** 2005. RAP: a new computer program for de novo identification of repeated sequences in whole genomes. Bioinformatics *21*:582-588.
8. **Orlov, Y.L. and V.N. Potapov.** 2004. Complexity: an internet resource for analysis of DNA sequence complexity. Nucleic Acids Res. *32*:W628-W633.

Received 21 September 2007; accepted 26 October 2007.

Address correspondence to Alessandro Vezzi, Department of Biology, University of Padova, via Ugo Bassi 58/B, Padova, Italy. e-mail: sandrin@cribi.unipd.it