BIOINFORMATICS APPLICATIONS NOTE Vol. 18 no. 10 2002
Pages 1396–1397

GENOTRACE: cDNA-based local GENOme assembly from TRACE archives

*Eugene Berezikov, Ronald H.A. Plasterk and Edwin Cuppen**

Hubrecht Laboratory, Netherlands Institute for Developmental Biology, Uppsalalaan 8, 3584 CT, Utrecht, The Netherlands

Received on March 14, 2002; revised on May 2, 2002; accepted on May 5, 2002

ABSTRACT

Summary: GENOTRACE identifies the genomic organization for a cDNA using raw data from genome sequencing projects in progress (trace archives). Local genomic contigs are generated, allowing for example the design of PCR primers in intronic sequences to amplify coding regions of a gene, needed for example for mutation or SNP detection.

Availability: The package and examples of output files can be downloaded from http://rat.niob.knaw.nl/GENOTRACE

Contact: ecuppen@niob.knaw.nl

Although raw genome sequencing data is continuously becoming available in public databases, it usually takes several years before this data is assembled in a user-friendly accessible format. Furthermore, an increasing number of genomes will only be sequenced up to draft genome coverage of 4×. At this moment, genome sequencing projects of about 15 organisms are in progress and for these organisms raw sequencing data is available from public servers at NCBI (http://www.ncbi.nlm.nih.gov/Traces/) and Ensembl (http://trace.ensembl.org/). Among these, only for human (http://www.ensembl.org/Homo_sapiens/) and mouse (http://www.ensembl.org/Mus_musculus/) draft genome assemblies are available at present time. Most of the data generated in genome sequencing projects is produced by whole genome shotgun sequencing, resulting in random short (600–800 nucleotides) fragments (traces) without any contextual information. Several online search tools, like SSAHA (http://www.ensembl.org/ssaha) and BLAST (http://www.ncbi.nlm.nih.gov/blast/mmtrace.html) allow the query of trace archives with for example cDNA sequences, but the output is not very well suited for retrieving intronic sequences flanking the coding regions. GENOTRACE is designed to meet this problem. The tool searches a local database of sequencing traces with a piece of DNA, i.e. a cDNA, as input using SSAHA (Ning *et al.*, 2001). The ends of matching traces are used in additional rounds of searches, 'walking' both 5′ and 3′ into the genome. Matching sequences from all searches are retrieved, assembled in genomic contigs using phrap (Gordon *et al.*, 2001), and the initial search sequence is annotated in the results (Figure 1b). This tool is not designed for making complete assemblies of large genomic regions but is optimized for identifying the intron–exon organization and for providing exon-flanking intronic sequences. GENOTRACE output can be used in genetic studies to analyze individuals for codon-changing single nucleotide polymorphism's (SNPs) or mutations. To this end, synthetic oligo's for PCR amplification flanking the exon sequence are needed for mutation detection. For this specific purpose, we implemented a primer-picking program (EMBOSS, eprimer3; Rozen and Skaletsky, 2000), to design sets of primers with the GENOTRACE output file as input (application is available upon request).

Although primarily designed for organisms for which genome sequencing is in progress, this tool may also be valuable for organisms with an assembled finished genome. Assembled genomes consist of a set of large contigs, ideally one for every chromosome. In practice, however, small and large gaps are present in these assemblies. Sequence traces that do not match any other trace or those that end up in contigs that are too small may be excluded from the assembled genome. This will occur more frequently when genomes are sequenced only up to draft coverage. As our tool uses the raw sequencing output files, no information is omitted from analysis, potentially resulting in more complete genomic coverage of a specific gene of interest.

The package we describe here consists of two components (Figure 1a). The first component is a tool to make and keep updated a local copy of a trace archive for a specific organism. This program downloads trace files from the NCBI trace archive (ftp://ftp.ncbi.nlm.nih.gov/pub/TraceDB) for any of the organisms available and creates a local file structure that is suited for use by the second component. Trace archives from other locations, like for example for fugu (ftp://ftp.jgi-psf.org/pub/JGI_data/Fugu/) should be installed manually (instructions

*To whom correspondence should be addressed.

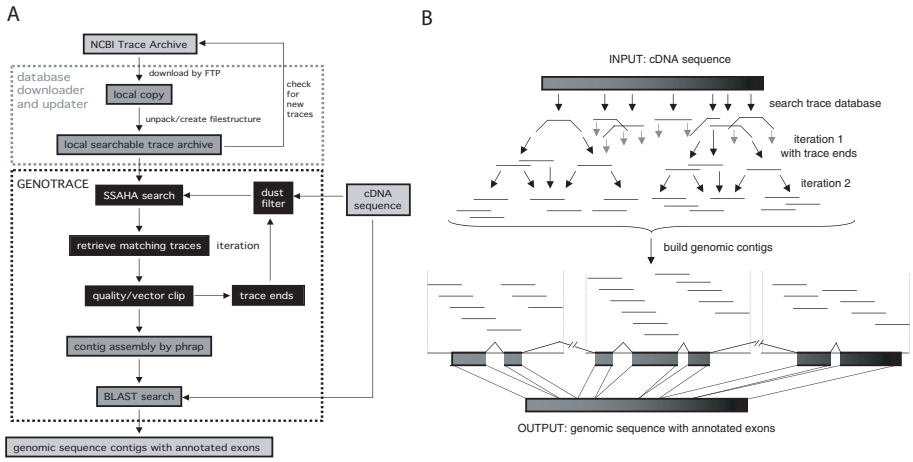

Fig. 1. (a) GENOTRACE pipeline. See the main text for details on the various components. (b) Schematic diagram of the GENOTRACE tool. A cDNA sequence is used as input for querying a sequence traces database. The ends of (partial) matching sequences are used to search the trace database again. This iteration step can be repeated a number of times, resulting in a 'walk' into the genomic DNA. As an output, matching traces from all steps are used to assemble one or more genomic contigs in which the exons for the initial cDNA are annotated.

are provided). The second component, GENOTRACE, uses a file containing one or more sequences in FASTA-format as input and searches, after filtering out simple repeats using DUST (Roma Tatusov and John Kuzio, unpublished), the complete trace archive for matching sequences using SSAHA (Ning *et al.*, 2001). Matching traces are retrieved, including the quality files and copied to a local directory. The retrieved sequences are clipped to remove low quality regions (phred score <20) and vector sequences using cross-match (Gordon *et al.*, 2001) and the UniVec database (ftp://ncbi.nlm.nih.gov/pub/UniVec/). The ends (150 nucleotides) of the resulting sequences are used, after DUST filtering, for a first iteration of SSAHA search against the trace archive. This iteration step can be repeated multiple times. Only the top six matches from a SSAHA search are considered in subsequent searches, to prevent the inclusion of enormous amounts of traces containing repetitive elements. These sequences are not suited for designing PCR primers anyway and in addition may complicate the last step in which all retrieved traces are used to build contigs using phrap. At present no special features are included to try to extend contigs beyond repetitive elements. Highly homologous genes and pseudogenes will probably be retrieved together with the gene of interest. However, due to the low degree of difference between such genes, they will end up in different genomic contigs during phrap assembly. Finally, an output-file is generated, showing the contigs that could be built including annotation of the input sequence. Identity scores for the matches between input sequence and genomic fragments are given in the header of each contig to be able to distinguish between highly homologous (pseudo)genes. Input (exon) sequences in the genomic contigs are shown in capitals and at the end of the file an overview of the input sequence is given illustrating the coverage of the input sequence in the genomic contigs.

GENOTRACE is written in Perl and can be run on any UNIX-based system. The package requires local installations of BioPerl (http://bio.perl.org), swat/cross_match/phrap (http://www.phrap.org), SSAHA (http://www.sanger.ac.uk/Software/analysis/SSAHA), NCBI Blast (ftp://ncbi.nlm.nih.gov/blast/), and DUST (ftp://ncbi.nlm.nih.gov/pub/tatusov/dust/). All these components are available freely or under an academic user agreement for non-commercial usage. In contrast to standard shotgun assembly methods, GENOTRACE does not require a very advanced computer setup. GENOTRACE was developed and is now used on a Pentium 1.9 GHz with 512 Mb internal memory and two hard disks of 80 Gb, running under Linux. With the configuration described 'GENOTRACEing' the rat traces database containing about 22×10^6 sequences (February, 2002) with a cDNA input sequence of about 2 kB and two iteration steps takes about 20–35 minutes.

ACKNOWLEDGEMENTS

This research was financially supported by the Dutch Ministry of Economic Affairs through the Innovation Oriented Research Program on Genomics

REFERENCES

Gordon,D., Desmarais,C. and Green,P. (2001) Automated finishing with autofinish. *Genome Res.*, **11**, 614–625.

Ning,Z., Cox,A.J. and Mullikin,J.C. (2001) SSAHA: a fast search method for large DNA databases. *Genome Res.*, **11**, 1725–1729.

Rozen,S. and Skaletsky,H. (2000) Primer3 on the WWW for general users and for biologist programmers. *Meth. Mol. Biol.*, **132**, 365–386.