BIOINFORMATICS APPLICATIONS NOTE

Vol. 26 no. 22 2010, pages 2908–2909
doi:10.1093/bioinformatics/btq557

Sequence analysis

Advance Access publication October 5, 2010

TURNIP: tracking unresolved nucleotide polymorphisms in large hard-to-assemble regions of repetitive DNA sequence

Robert P. Davey[1,2,*,†], Stephen A. James[1], Jo Dicks[2] and Ian N. Roberts[1]

[1]National Collection of Yeast Cultures, Institute of Food Research and [2]Department of Computational and Systems Biology, John Innes Centre, Norwich Research Park, Colney, Norwich, NR4 7UA, UK

Associate Editor: Joaquin Dopazo

ABSTRACT

Summary: TURNIP comprises a suite of Perl scripts and modules that facilitates the resolution of microheterogeneity within hard-to-assemble repetitive DNA sequences. TURNIP was originally developed for the *Saccharomyces* Genome Resequencing Project (SGRP) within which the ribosomal DNA (rDNA) of 36 strains of *S.cerevisiae* were analysed to investigate the occurrence of potential polymorphisms. Here, ‘partially resolved SNPs’, or pSNPs, as well as indels, were found to be far more prevalent than previously suspected. More generally, the TURNIP software ascertains degrees of variation between large tandem repeats within a single locus, offering insights into mechanisms of genome stability and gene conversion in any organism for which genome sequence data are available.

Availability: The TURNIP source code, results files and online help are available at http://www.ncyc.co.uk/software/turnip.html .

Contact: robert.davey@bbsrc.ac.uk

Received on June 29, 2010; revised on August 27, 2010; accepted on September 28, 2010

1 INTRODUCTION

The drive to continually decrease the cost of DNA sequencing to achieve better sequence value-for-money has led to a whole host of new sequencing platforms and analytical techniques. High-throughput ‘next-generation’ platforms such as ABI SOLiD, Roche 454 and Solexa (Illumina, Inc.) can generate so much data in a single run that where previously major concerns centered around read quality and coverage, now the challenges are shifting towards data storage, data management and timely analysis.

Assembly strategies and tools have had to keep up with the deluge of nucleotide bases, and as a result many are now highly sophisticated. Alongside (and often coupled with) these tools are a large number of SNP calling algorithms that aim to categorise point variation using reference-based or *de novo* assembly. However, these algorithms are not appropriate in all cases, such as dealing with repetitive regions.

An extreme case of a large repetitive genomic region is the ribosomal DNA (rDNA) tandem repeat array. For example, in the brewer’s yeast *Saccharomyces cerevisiae* this array, which typically comprises between 50 and 200 copies of a 9 kb repeat unit depending on the strain, can occupy up to 60% of chromosome XII on which it is located at a single locus (Kobayashi *et al.*, 1998; Petes, 1979). In fact the rDNA consensus of *S.cerevisiae* strain S288c, the first yeast and indeed first eukaryote to have its almost entire genome sequenced, is only derived from the terminal left- and right-hand rDNA repeat copies (Goffeau *et al.*, 1996). Thus, the complete sequence of this essential housekeeping fragment of the genome is neither straightforward to extract nor analyse as the sequence reads cannot be reliably assigned to a given repeat.

Analytical tools have been developed to cope with relatively small arrays comprising short repetitive sequences, e.g. the TRAP algorithm (Tammi *et al.*, 2003), but elucidating the order and variability of large repeats is still a complicated and unsolved problem. Single molecule sequencing may provide a solution, but in the interim we present TURNIP, a new software suite to address the problem of resolving variability in large repetitive genomic regions.

2 SYSTEMS AND METHODS

TURNIP uses a previously derived consensus sequence as a reference and compares aligned ‘seed’ read bases at each position, recording those that comply with set threshold parameters, and hence functions in a similar fashion to other SNP calling algorithms such as the Illumina SNP Caller (Bentley *et al.*, 2008) and an analagous process within the MAQ software (Li *et al.*, 2008).

However, both SNP Caller and MAQ do not record partially resolved polymorphisms, whereas TURNIP is specifically designed to accomplish this task by the following means. Region-specific sequences are found by performing a BLAST search (Altschul *et al.*, 1997) of reads against a supplied consensus region, and are then indexed by read ID for subsequent retrieval during the polymorphism calling process.

Using a sliding window approach (Fig. 1A), seed reads are anchored to a consensus using a less stringent BLAST than that carried out in the previous step for identifying region-specific sequences. These are then processed as 100mer (maximum) windows, in which the central 20mer is the target sequence for each window.

TURNIP may see reads that do not fully span the current window, and hence anchor regions that are smaller than 100 bp, in which case a minimal shortness threshold (default of 70 bp maximum resulting in a 20mer target and 25 bp flanks) can be set to ensure a large enough proportion of the read is available to anchor.

TURNIP uses MUSCLE (Edgar, 2004) to perform a gapped multi-alignment on each set of distinct subject (seed read) sequences and the window template (consensus) at each sliding 20mer position. The consensus template is retained in the alignment to identify gaps in the consensus that represent potential insertions in the seed sequences. If an alignment program other than MUSCLE is preferred, this may be changed easily within the TURNIP configuration.

[*]To whom correspondence should be addressed.

[†]Present address: The Genome Analysis Centre, Norwich Research Park, Colney, Norwich, NR4 7UA, UK.

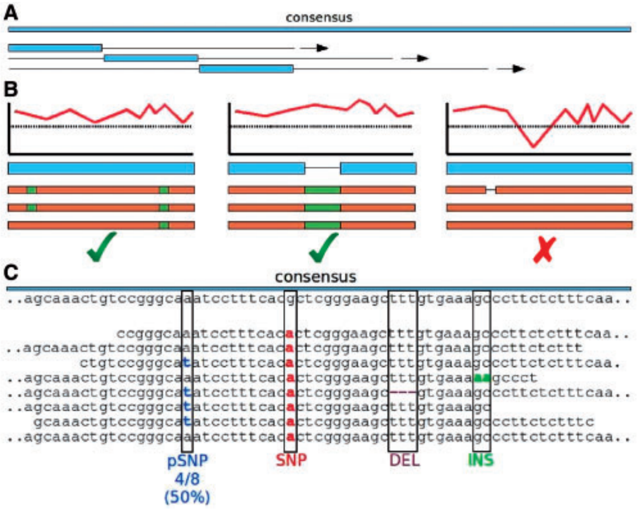
A
consensus
B
C
consensus
..agcaaactgtccgggcaaatcctttcacgctcgggaagctttgtgaaagccttctctttcaa..
ccgggcaaatcctttcacactcgggaagctttgtgaaagccttctctttcaa..
..agcaaactgtccgggcaaatcctttcacactcgggaagctttgtgaaagccttctcttt
ctgtccgggcatatcctttcacactcgggaagctttgtgaaagccttctctttcaa.
..agcaaactgtccgggcaaatcctttcacactcgggaagctttgtgaaaaagccct
..agcaaactgtccgggcatatcctttcacactcgggaagc---gtgaaagccttctctttcaa..
..agcaaactgtccgggcatatcctttcacactcgggaagctttgtgaaagc
gcaaactgtccgggcatatcctttcacactcgggaagctttgtgaaagccttctctttc
..agcaaactgtccgggcaaatcctttcacactcgggaagctttgtgaaagccttctctttcaa..
pSNP
4/8
(50%)
SNP
DEL
INS

Fig. 1. (**A**) Sliding window approach, depicting the central 20mer region anchored by longer flanking regions. (**B**) Seed read filtering procedures employed whereby quality scores are checked across each 20mer and rejected if any drop below a given threshold. (**C**) 'Stacking' of reads that align to a single copy consensus to ascertain SNP, indel and partial SNP (pSNP) variation. Variation is discarded if it is only resolved in a single read per 20mer window, e.g. the insertion and deletion would both be discarded here.

Inserted gaps in the template and subject sequences are spanned with surrounding known quality scores. Aligned seed read 20mers at a given window position are discarded if they have any single base below a set sequence quality score threshold (this does not mean that a particular read is discarded for the whole analysis, but just at the current 20mer window. See Fig. 1B). If they pass these constraints, the seed read 20mers are then compared with the consensus and called for variation (Fig. 1C).

In this way, the TURNIP software encapsulates, and contains improvements to, the method suggested by us and used in James *et al.* (2009). In particular, due to fine-scale alignment and filtering mechanisms employed within, TURNIP is able to resolve variable-length indels as well as point mutations, and other problematic features such as long homopolymeric (polyA and polyT) tracts. The TURNIP algorithmic process is described in greater detail in the online documentation.

Portions of the TURNIP software, such as BLAST, MUSCLE and the sliding window approach, can be configured to run concurrently on a multi-core/cluster environment and hence analyse multiple window positions simultaneously. As an example, strain YS4 from the Saccharomyces Genome Resequencing Project (SGRP) strain set, sequenced using Sanger whole-genome shotgun sequencing at approximately 1× coverage, has an estimated 100 repeat copies (James *et al.*, 2009) giving a complete dataset of just under 1 Mb. On a dual core 3 GHz desktop machine with 2 GB RAM, this dataset took 3 min to process using TURNIP.

At each stage of the analysis, TURNIP writes output files that allow it to use precomputed data in case of process interruption, or if certain variation-calling threshold parameters are changed and a simple reanalysis of the final alignments is needed. TURNIP produces simple text file output describing the positions, types and frequencies of the resolved polymorphisms, as well as SQL files for database importing and GFF3 for bulk loading into a GBrowse database (Stein *et al.*, 2002). A custom pie chart glyph and an example GBrowse configuration file are included in the TURNIP distribution.

The TURNIP suite also provides scripts to produce distance matrices and phylogenetic trees from the output variation frequency data. Trees generated from analysis of the SGRP *S.cerevisiae* dataset can be viewed and interactively modified by strain subset selection at the NCYC web site (see Availability). The trees generated from pSNP distances closely match those produced as part of the study of Liti *et al.* (2009), showing that analysis of the rDNA region of the yeast genome alone can be sufficient to produce informative phylogenies based on variation that is not fully resolved. This information could be used in future for the purposes of precision strain characterization and evolutionary studies including investigation into gene conversion in closely related lineages.

Funding: Biotechnology and Biological Sciences Research Council Tools and Resources Development Fund project no. BB/G000441/1 (to R.P.D.); BBSRC (to S.A.J., I.N.R., J.D.).

Conflict of Interest: none declared.

REFERENCES

Altschul,S. *et al*. (1997) Gapped blast and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Bentley,D. *et al*. (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, **456**, 53–59.

Edgar,R. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, **5**, 1–19.

Goffeau,A. *et al*. (1996) Life with 6000 genes. *Science*, **274**, 546–567.

James,S. *et al*. (2009) Repetitive sequence variation and dynamics in the ribosomal DNA array of Saccharomyces cerevisiae as revealed by whole-genome resequencing. *Genome Res.*, **19**, 626–635.

Kobayashi,T. *et al*. (1998) Expansion and contraction of ribosomal DNA repeats in saccharomyces cerevisiae: requirement of replication fork blocking (FOB1) protein and the role of RNA polymerase I. *Genes Dev.*, **12**, 3821–3830.

Li,H. *et al*. (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.*, **18**, 1851–1858.

Liti,G. *et al*. (2009) Population genomics of domestic and wild yeasts. *Nature*, **458**, 337–341.

Petes,T. (1979) Yeast ribosomal DNA genes are located on chromosome-xii. *Proc. Natl Acad.Sci. USA*, **76**, 410–414.

Stein,L. *et al*. (2002) The generic genome browser: a building block for a model organism system database. *Genome Res.*, **12**, 1599–1610.

Tammi,M. *et al*. (2003) TRAP: tandem repeat assembly program produces improved shotgun assemblies of repetitive sequences. *Comput. Methods Programs Biomed.*, **70**, 47–59.