Bioinformatics Advance Access published October 11, 2006

THOR: Targeted High-throughput Ortholog Reconstructor

Matthew N. Bainbridge, René L. Warren, An He, Mikhail Bilenky, A. Gordon Robertson and Steven J. M. Jones*

British Columbia Cancer Agency (BCCA) Genome Sciences Centre, 675 West 10th Avenue
Vancouver, BC, Canada

Associate Editor: Martin Bishop

ABSTRACT

Summary: Low-coverage genomes (LCGs) are becoming an increasingly important source of data for phylogenetic studies. However, assembly of these genomes is time consuming, difficult and lags behind sequence generation. THOR is a fast, stringent application for targeted reconstruction of sequence orthologs in unassembled LCGs. Using a 4x coverage set of mouse whole-genome sequence reads, THOR could partially or completely reconstruct 416/1000 human promoter ortholog regions in ~7.3 min/promoter. THOR's reconstruction rate improves markedly with both higher-coverage, and less divergent target species.

Availability: THOR is implemented in java and is currently available as source code and as a web service (www.bcgsc.ca/services/thor) for reconstructing human sequences.

Contact: thor@bcgsc.ca

1 INTRODUCTION

At present there is a massive genome sequencing initiative with 28 diverse, mammalian genomes currently being or due to be sequenced at low (~2x) coverage (www.genome.gov/10002154). This represents an unprecedented amount of genetic information which is useful for comparative genomics and evolutionary studies. Many of these genomes are currently unassembled because of the time, expense and expertise required for such a task (Margulies, et al., 2005). Fortunately, many studies that would benefit from low-coverage genomic data require only one or a few short (1-10 kb) orthologous regions to be assembled.

Here we present THOR, a Targeted High-throughput Ortholog Reconstructor, an application that can rapidly reconstruct sequence ortholog regions in an LCG. Although comparative genome assemblies have been presented before (Pop, et al., 2004), THOR is an easy to use, computationally inexpensive application that ties a number of bioinformatics applications together into an efficient reconstruction pipeline.

2 IMPLEMENTATION

Given a set of low-coverage sequence reads (LCRs) from some organism (the source) and a set of human sequences (the targets), with known position on the human genome, THOR aligns all reads against all targets using wuBLAST (Gish, 1996 - 2005). The LCRs that best align against any particular target (up to a user defined minimum BLAST-score and maximum number of reads) are selected, trimmed of low quality bases, cleaned of any vector sequence and assembled using phrap (Green, 1994) with low stringency parameters (-minmatch 30 -minscore 60 -forcelevel 0 -vector_bound 0). Both assembled and unassembled reads (we shall refer to both types of reads as "contigs") are then aligned against the human genome using wuBLAST, with default search parameters. This "reciprocal BLAST" stage is used to ensure that the contigs match better to the target sequence than some other genomic region and represents a stringent filter of orthology. The contigs that pass the reciprocal BLAST test are then "assembled" using the target sequence as a template. In this approach we start with a sequence, S, which consists entirely of 'N's and is the same length as the target sequence. As contigs are found which align to the target (starting with the best BLAST scored contigs), the bases in S are filled in with 'real' bases from the source organism. When there are no more contigs to consider, S contains the reconstructed target in the source organism. Regions in S where there are no alignments remain as 'N's, and insertions are shown as lower-case letters. This technique allows contigs that overlap poorly, or not at all, to be placed into the target sequence in a meaningful manner. This template-driven assembly allows far more sequence to be reconstructed than relying solely on *de novo* assembly.

The total execution time of THOR for any particular species scales linearly with genome coverage, although execution time also depends on the divergence of the target species from human. For most genomes at low-coverage (< 3x), the reciprocal BLAST stage represents approximately half of the total execution time (data not shown). Although this is a computationally costly step, it is critical for ensuring the stringency of this method (see Results) and for the template-driven reconstruction.

With the exception of the p21 upstream reconstruction, which was executed on a heterogeneous computer grid, all executions of THOR presented here took place on a single 2 CPU (4 cores) AMD Opteron 275 @2.2GHz with 4GB of RAM. The maximum number of reads was set to 30, and the minimum BLAST score was set to 150.

3 RESULTS

We evaluated the ability of our application to reconstruct orthologs of non-coding human genomic regions. To do this, we selected 1000 promoter regions (2.6 Mb total length) from different human genes with known homologs in mouse as determined by multiple genome alignment (Schwartz, et al., 2003). We randomly selected a subset of the total available mouse read data (trace.ensembl.org) to simulate various (0.5x, 1x, 2x, 4x) levels of low-coverage.

Because of biases which occur when constructing genome libraries, and the presence of repeat regions and homologs in the human genome, it is difficult to determine how many reads we would reasonably expect to pass the reciprocal BLAST test (Wendl and Yang, 2004). To address this, we established base-lines for the expected reconstruction rates by replicating the mouse experiments using human whole-genome sequence read data at the same levels of coverage. Reconstructing "human on human" serves as a positive control for THOR, establishing the best possible outcome at a given level of coverage. Using 2x coverage human reads, THOR successfully reconstructed ~1.4 Mb of sequence data in 938 promoter regions (2.4 Mb total size) in ~87 hours (~5.6 min/promoter). Figure 1 shows the reconstruction rates of promoters and total bases using human reads at 0.5x, 1x, 2x, and 4x cov-

*To whom correspondence should be addressed: sjones@bcgsc.ca

erage and mouse reads at 0.5x, 2x and 4x. At the highest coverage, THOR reconstructs 98.8% (1.8 Mb) of the 1000 target sequences. As expected, the number of successfully reconstructed sequences using 2x coverage mouse reads was much lower than for human, yielding only 344 promoters (0.198 Mb). However, the total execution time also decreased significantly to 36.4 hours (~6.3 min/promoter). At 4x coverage, the number of reconstructed targets improves to 416 (0.274 Mb) with an associated increase in execution time of 14 hours. To determine our false positive rate we used wuBLAST to align the 416 reconstructed regions to the mouse genome and found 383 sequences (92%) successfully aligned to the correct region in mouse, for 2x coverage this number decreased slightly to 91%.

We also evaluated THOR's ability to reconstruct our promoter set in organisms which are less divergent than mouse, which has a ~70% neutral substitution rate (NSR) compared to human (Cooper, et al., 2004). Using reads from elephant (2.3x coverage, 24% NSR) and rabbit, (2.6x coverage, 52% NSR) (Cooper, et al., 2005), THOR was able to reconstruct 737 promoters (0.59Mb) and 735 promoters (0.55Mb) in 35 hours and 45 hours for elephant and rabbit, respectively.

Because promoter regions may not be highly conserved across species, we also examined THOR's ability to reconstruct orthologs for protein coding genomic regions. Using 2x-coverage mouse reads, sequence reconstruction rates rose modestly to 40%, a 20% increase over using promoter sequences. However, the amount of sequence reconstructed rose to 41%, a 60% increase.

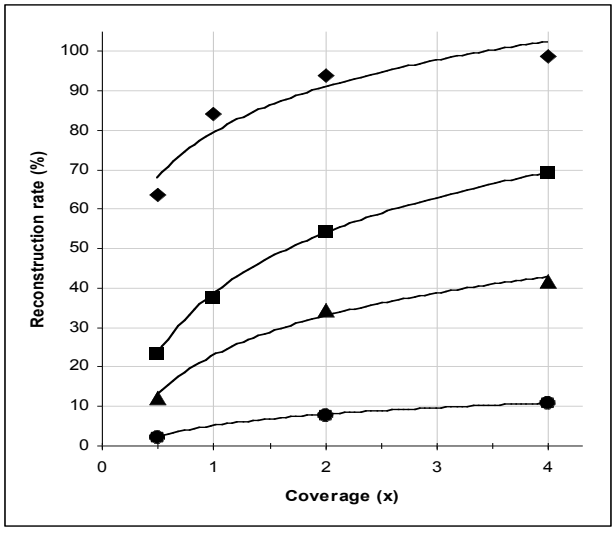

Fig. 1. Reconstruction rates of 1000 human promoters by THOR, at various levels of read coverage. Diamonds and triangles represent the number of reconstructed promoters, while squares and circles indicate the number of reconstructed bases for human and mouse respectively.

Lastly, in order to simulate the typical usage of the THOR web service, we reconstructed a 3Kb region of human genome immediately upstream of the tumour suppressor gene *p21* in 20 LC or unassembled genomes. These genomes range in coverage from 0.1x to 10x. Using these reads, THOR successfully managed to reconstruct all or some portion of the target in 11 LCGs. The median execution time for reconstruction was 540 seconds, with 3 executions taking over 1 hour and 2 taking less than 1 minute.

4 DISCUSSION

We have shown that THOR can reconstruct human orthologs using low-coverage reads in a fast, high-throughput, and stringent manner on modest commodity hardware. THOR's reconstruction rate and execution time is dependant on source genome coverage and the amount of conservation between the source and target sequences. As can be seen in Figure 1, increasing coverage provides decreasing relative gain in reconstruction rates. This figure also indicates that routinely sequencing genomes to 2x may be insufficient to discover the majority of conserved genomic elements, especially for organisms that are highly divergent from human.

Although we have only presented reconstructing human targets in LCGs from sequences derived from placental mammals (i.e. up to ~70% NSR), THOR can be used to reconstruct targets in any organism with at least a partially assembled genome. Internally, it has been used to reconstruct mammalian promoter orthologs for mouse and also nematode orthologs for *C. elegans* sequences (manuscripts in preparation) for use in the cisRED project (Robertson, et al., 2006).

In an ideal case, at 4x coverage, THOR can partially or completely reconstruct 98.8% of 1000 promoter target sequences. More realistically, 70% reconstruction rates should be achievable for most species. Further, strongly conserved genomic elements, such as protein coding regions, produce 20% higher reconstruction rates. Finally, we have also shown that even with highly divergent organisms, THOR reconstructs orthologs with a less than 10% false positive rate, meaning that even when reconstruction rates are low, the results are reliable.

ACKNOWLEDGEMENTS

This work was funded in part by Genome Canada. SJMJ is a scholar of the Michael Smith Foundation for Health Research.

REFERENCES

Cooper, G.M., et al. (2004) Characterization of evolutionary rates and constraints in three Mammalian genomes, *Genome Res*, **14**, 539-548.

Cooper, G.M., et al. (2005) Distribution and intensity of constraint in mammalian genomic sequence, *Genome Res*, **15**, 901-913.

Gish, W. (1996 - 2005) wublast.wustle.edu.

Green, P. (1994) www.phrap.org.

Margulies, E.H., et al. (2005) An initial strategy for the systematic identification of functional elements in the human genome by low-redundancy comparative sequencing, *Proc Natl Acad Sci U S A*, **102**, 4795-4800.

Pop, M., et al. (2004) Comparative genome assembly, *Brief Bioinform*, **5**, 237-248.

Robertson, G., et al. (2006) cisRED: a database system for genome-scale computational discovery of regulatory elements, *Nucleic Acids Res*, **34**, D68-73.

Schwartz, S., et al. (2003) Human-mouse alignments with BLASTZ, *Genome Res*, **13**, 103-107.

Wendl, M.C., et al. (2004) Gap statistics for whole genome shotgun DNA sequencing projects, *Bioinformatics*, **20**, 1527-1534.