Nature Precedings

A Target Restricted Assembly Method (TRAM) for Phylogenomics

Kevin P. Johnson[1,*], Kimberly K. O. Walden[2], and Hugh M. Robertson[2]

[1]Illinois Natural History Survey, University of Illinois, 1816 S. Oak Street, Champaign, IL 68120 USA

[2]Department of Entomology, University of Illinois, 505 S. Goodwin Ave., Urbana, IL 61801 USA

ABSTRACT

While next generation sequencing technology can produce sequences covering the entire genome, assembly and annotation are still prohibitive steps for many phylogenomics applications. Here we describe a method of Target Restricted Assembly (TRAM) of a single lane of Illumina sequences for genes of relevance to phylogeny reconstruction, i.e. single copy protein-coding genes. This method has the potential to produce a data set of hundreds of genes using only one Illumina lane per taxon.

INTRODUCTION

Next generation sequencing technologies are revolutionizing the field of genomics. New DNA sequencing methods such as 454 and Illumina pyrosequencing can produce billions of base pairs of DNA sequence in a highly automated fashion for a fraction of the costs of traditional targeted Sanger sequencing (Hudson, 2008). For many phylogenetic questions, targeted sequencing of one or a few genes has proved insufficient to provide phylogenetic resolution with good support. Even "phylogenomic" studies that have included a large number of genes sequenced using PCR targeted sequencing have not resolved trees with good branch support at all nodes (Hackett *et al.*, 2008; Regier *et al.*, 2010). For difficult phylogenetic problems, DNA sequence datasets with many additional genes would be desirable in such cases, but the costs and labor of traditional targeted sequencing become prohibitive.

Completely assembled genome sequences are also not a panacea for solving difficult phylogenetic problems. Genome assembly and annotation is still a time and labor-intensive process. For most phylogenetic problems, complete genome sequences will not be necessary to provide well supported and resolved phylogenetic trees. One restricted genome approach used with success recently is to sequence expressed sequence tags (ESTs), which are cDNA copies of mRNA transcripts (Philippe and Telford, 2006). This restricts the proportion of the genome examined to the "transcriptome", which represents only the protein coding portions of the genome, and may potentially be more ideal for phylogenetic analysis. However, generation of very large EST datasets requires a large quantity of mRNA, which typically must be extracted from specially preserved or frozen very fresh starting material. For uncommon or small-bodied organisms, this might be prohibitive.

*To whom correspondence should be addressed.

e-mail: kjohnson@inhs.uiuc.edu

Here we extend and generalize a hybrid technique to identify protein-coding targets of interest in DNA sequences generated from next-generation sequencing technologies, a Target Restricted Assembly Method (TRAM). A version of this method, using human promoter regions as targets for mammalian low coverage sequence reads, was previously developed by Bainbridge *et al.* (2007). Using sequence reads from single lanes of Illumina sequencing runs, we first show that using targeted *tblastn* searches we can recover protein-coding sequence reads of interest from a single Illumina lane. These sequences can then be assembled locally, to provide a longer sequence (contig) of the gene of interest. By targeting genes that have been shown to be 1:1 orthologs across genomes for a relevant group (Kriventseva *et al.*, 2008), a dataset of a very large number (>500) of single copy nuclear protein coding genes can be produced for a fraction of the cost of producing such a dataset using targeted sequencing techniques. In addition, for most phylogenetic studies that might make use of this method, nuclear protein coding genes are likely to be the ideal genes to resolve phylogenetic relationships (i.e. species level and higher). Here we first outline the principles and procedures involved and then illustrate the method when 1) the target sequence is known and 2) the target sequence is unknown.

METHODS AND RESULTS

The Target Restricted Assembly Method

The Target Restricted Assembly Method (TRAM) consists of a series of steps to identify and assemble sequences of genes of interest from raw sequence reads generated by next generation sequencing technologies. These steps may require modification for particular applications, but outline the basic procedure.

Compile sequences of genes/proteins of interest from a species closely related to focal species. For most applications, the sequences of the target genes in the focal species will be unknown. Thus, it will be necessary to use sequences of such genes from closely related species. Sequence similarity between the two species will determine, in part, the success rate of this method (see Examples below). For protein coding genes, the amino acid sequences will generally be preferred to DNA sequences because they are more conserved. For most higher level phylogenetic problems, nuclear protein coding genes are the preferred data source because of their relatively high level of sequence conservation relative to non-coding regions and the relative ease with which they can be aligned. These sequences could either

be sequences from EST data sets or genome sequences in which protein coding genes have been annotated.

Use the target sequence(s) in a BLAST search of the raw read data to identify matches. This step can be done via local blast of Illumina data consisting of the raw unassembled reads. For protein coding sequences a *tblastn* search of the raw reads will be appropriate. These BLAST searches should identify potential matches to the target sequence among the raw reads.

Assemble the potential matches using relevant sequence assembly software. Once the list of matching reads from the BLAST search is obtained, these reads should be retrieved and imported into a relevant sequence assembly software. For the purposes of our examples (below), we retrieved the reads as a FASTA file and imported these sequences into Sequencher (GeneCodes), which we used to perform the assemblies. For protein coding sequences, typically this will result in subassemblies for each exon.

Examples

*Target Sequence Known***.** The DNA sequence for the odorant receptor 1 (Or1) gene (1440 bp) for the Argentine ant (*Linepithema humile*) has been previously determined (Robertson, unpublished data). We use the inferred protein sequences for this species, another ant (*Pogonomyrmex barbatus*), and the honey bee (*Apis melifera*) as *tblastn* queries against a single lane of Illumina reads from this Argentine ant. These reads averaged around 75 bp and approximately 20 million reads were available as blast targets. We assembled the all the blast matches into contigs using Sequencher with default settings. These local assemblies were then assembled against the actual DNA sequence for the gene, which included both exons and introns.

In general, we recovered at least partial coverage for all eight exons independent of the divergence of the starting sequence: *Linepithema* (94% coverage), *Pogonomyrmex* (90%), *Apis* (90%). However, the number of reads recovered as tblastn matches were higher when more closely related species were used as the query target: 111 (*Linepithema*), 108 (*Pogonomyrmex*), and 89 (*Apis*) reads. Only 3 to 5 sites from the consensus assembly were a mismatch from the known sequence, and we cannot rule out native allelic variation in this case, given the relatively low coverage.

Target Sequence Unknown. Our next example involves Target Restricted Assembly in a more realistic application, where the target sequence is unknown in the species of interest (in this case the bumble bee, *Bombus impatiens*). Here we used a single lane of an Illumina sequencing run, with new technology where the average read length was around 125bp and which produced approximately 35 million reads. As targets we used genes that were strict 1:1 orthologs across all insects and crustacean genomes as identified in the Ortholog Database (http://cegg.unige.ch/orthodb). This database contains 898 such genes, so for illustrative purposes we used the first 10 genes recovered from the database.

Protein sequences of these genes from the honey bee (*Apis melifera*) were used as queries in *tblastn* searches against the *Bombus impatiens* Illumina raw reads. All the reads that were *tblastn* matches for each protein query were then assembled in Sequencher with default parameters. The length of these assemblies and number of contigs was highly variable, depending on the existence and size of introns. In general coverage ranged from 5-20X. In several cases, even though only the amino acid (i.e. exon) sequence was used as the query we were able to assemble these reads completely through intron sequences (see Table 1).

To assemble these sequences into complete protein coding DNA sequences, we downloaded the DNA sequences for *Apis melifera* for the 1:1 ortholog proteins from BeeBase (http://genomes.arc.georgetown.edu/drupal/beebase). We assembled the consensus sequence from each contig in the initial *Bombus* assemblies in Sequencher with the complete protein coding DNA sequence from *Apis*. Because the *Bombus* assemblies included at least partial intron sequences, we set the parameters to allow for large gaps and set the minimum match to 70%. We further aligned these sequences by removing unnecessary gaps, adjusting intron positions and removing overlap between contigs in non-homologous intron sequence (i.e. when the intron sequence was only partially represented on each end). In nearly all cases complete protein coding sequences were recovered from these assemblies (see Table 1). Reciprocal best *blastn* of these sequences against sequences in BeeBase recovered the DNA sequence from the original query protein in all 10 cases. Unlike the example with ants, some contigs were not included in the final assembly and it is likely that these represent other genes that are low level blast matches. Further optimization of BLAST cutoffs might be necessary, so that such sequences are not recovered. However, we found that these null matches do not interfere with the final assembly because they do not contain a high match for the DNA sequences themselves, and the entire protein coding DNA sequences were already represented in other contigs.

Table 1. Results of Target Restricted Assemblies for *Bombus* Genes

Gene	bp in *Apis*	# exons	# contigs	protein coverage
GB15017	1932	10	3	97%
GB11672	2043	11	3	100%
GB15983	1575	10	3	100%
GB13777	1062	6	2	100%
GB12202	1695	8	7	97%
GB19115	3930	19	12	98%
GB13857	891	5	4	100%
GB11458	1113	1	1	100%
GB17705	2316	1	1	100%
GB16928	1485	10	2	100%

DISCUSSION

We have shown complete or nearly complete protein coding DNA sequences can be recovered by a Target Restricted Assembly Method of raw sequences from a single Illumina lane. This extends the procedure outlined for human promoter regions (Bainbridge *et al.* 2007) to protein coding sequences in insects. Strictly orthologous genes, such as these, avoid the problems of paralogy associated with using genomic data for phylogenetics. In addition, because these are nuclear protein coding genes, alignment is relatively straightforward, and such data can be directly combined with inferred protein coding sequences from EST data sets generated from cDNA sequences, reversed transcribed from mRNA. For a modest scale phylogenetic project (20-30 species), the costs of data production would likely not be prohibitive, making phylogenomic data sets of hundreds of single-copy protein-coding genes with current technology very feasible.

Funding: KPJ was supported by NSF DEB-0612938. Gene Robinson provided the *Bombus impatiens* Illumina reads.

REFERENCES

Bainbridge,M.N. *et al.* (2007) THOR: targeted high-throughput ortholog reconstructor. *Bioinformatics* **23**, 2622-2624.

Hackett,S.J. *et al.* (2008) A phylogenomic study of birds reveals their evolutionary history. *Science* **320,** 1763-1768.

Hudson,M.E. (2008) Sequencing breakthroughs for genomic ecology and evolutionary biology. *Molecular Ecology Resources* **8**, 3-17.

Kriventseva,E.V. *et al.* (2008) OrthoDB: the hierarchical catalog of eukaryotic orthologs. *Nucleic Acids Research* **36**, D271-D275.

Philippe,H. and Telford,M.J. (2006) Large-scale sequencing and the new animal phylogeny. *Trends in Ecology and Evolution* **21**, 614-620.

Regier,J.C. *et al.* (2010) Arthropod relationships revealed by phylogenomic analysis of nuclear protein-coding sequences. *Nature* **463**, 1079-1083.