

Genome analysis

SCARF: maximizing next-generation EST assemblies for evolutionary and population genomic analyses

Michael S. Barker^{1,2,*}, Katrina M. Dlugosch^{1,†}, A. Chaitanya C. Reddy¹, Sarah N. Amyotte¹ and Loren H. Rieseberg^{1,2}

¹Botany Department, University of British Columbia, Vancouver, B.C. V6T 1Z4, Canada and ²Department of Biology and Center for Genomics and Bioinformatics, Indiana University, Bloomington, IN 47405, USA

Received on November 4, 2008; revised on December 18, 2008; accepted on January 1, 2009

Advance Access publication January 6, 2009

Associate Editor: Joaquin Dopazo

ABSTRACT

Summary: Scaffolded and Corrected Assembly of Roche 454 (SCARF) is a next-generation sequence assembly tool for evolutionary genomics that is designed especially for assembling 454 EST sequences against high-quality reference sequences from related species. The program was created to knit together low-coverage 454 contigs that do not assemble during traditional *de novo* assembly, using a reference sequence library to orient the 454 sequences.

Availability: SCARF is freely available at <http://msbarker.com/software.htm>, and is released under the open source GPLv3 license (<http://www.opensource.org/licenses/gpl-3.0.html>).

Contact: msbarker@indiana.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Transcriptome and expressed sequence tag (EST) sequencing with next-generation systems such as Roche's 454 platform promises to make genomics a cost-effective endeavor in a wide range of model and non-model organisms. A hallmark of this new technology is the high megabase yield of sequenced DNA produced at a fraction of the cost of traditional Sanger sequencing, and an opportunity to generate *de novo* EST collections and large population genomic datasets at relatively low cost. However, 454 currently generates shorter reads than Sanger sequencing and these suffer from a homopolymer repeat problem (Margulies *et al.*, 2005), both of which confound the complete assembly of data from inherently non-contiguous and low depth sequencing strategies such as ESTs. Incomplete *de novo* assembly leads to short and multiple, non-overlapping unique gene assemblies (unigenes) representing the same gene. These are particularly acute problems for evolutionary and population genomics, where data analyses, such as tests for positive selection, are weakened by short lengths or inaccurate gene family phylogenies. Thus, to increase unigene length and more accurately assemble the correct number of genes we would like

to leverage all available sequence data to connect the disparate non-overlapping fragments and produce a scaffolded assembly for each unigene. Further, we would like to use data that we consider reliable to correct the homopolymer repeats present in 454 reads. Most currently available assembly tools for next-generation data are focused on genome resequencing and were not designed for the problems posed by sequencing efforts that generate many individual contigs that do not ultimately overlap. To resolve these assembly limitations for EST and similar datasets, we developed SCARF—Scaffolded and Corrected Assembly of Roche 454.

2 DESCRIPTION

SCARF is based on a relatively simple algorithm to match 454 contigs with reference sequences and generate a scaffolded contig. Although SCARF is capable of assembling raw 454 reads, it is preferred that reads are assembled with a *de novo* assembler prior to SCARFing. This reduces the total number of sequences that SCARF must handle and also provides assembly of reads that do not have a reference sequence. Similar to TGICL (Perlea *et al.*, 2003), SCARF first uses NCBI's megablast (Zhang *et al.*, 2000) to efficiently cluster 454 contigs with a reference sequence. These reference clusters are subsequently processed in parallel to increase overall speed. The clustered contigs are accurately aligned against the reference and each other through iterative Smith–Waterman alignments (Smith and Waterman, 1981). During this process, homopolymer repeats longer than a user-specified threshold that may optionally be checked and corrected against the reference. A consensus sequence for each reference cluster is generated based on the aligned 454 contigs using (i) the highest quality score for a base to resolve discrepancies or (ii) majority rule if quality scores are equal or no quality file is provided. Gaps between 454 contigs revealed through alignment to the reference are filled in with '-' or a character selected by the user, thus knitting together non-overlapping fragments of the same gene. These 'SCARFed' contigs are printed out to a FASTA formatted file along with non-SCARFed sequences that aligned to the reference and may have been altered by homopolymer repeat correction. Log files and summary statistics of the program run are also reported. Finally, SCARFed contigs may optionally be used to extend the reference sequences to produce a maximally complete EST dataset.

SCARF was designed to minimize assembly time and take advantage of multi-core computer architecture when available.

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

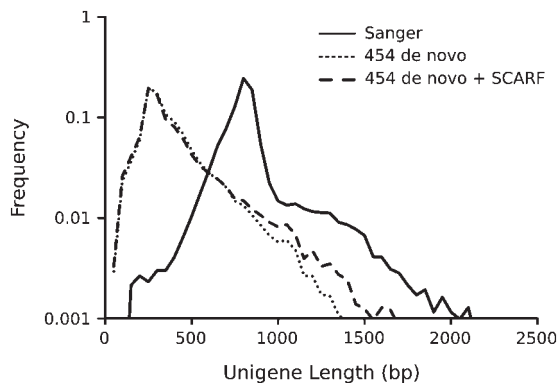


Fig. 1. Frequency distribution of unigene lengths of assembled Sanger ESTs (solid line), MIRA *de novo* assembled 454 ESTs (dotted line) and post-MIRA SCARF assembled 454 ESTs (dashed line) for *C. solstitialis*.

Currently, SCARF is provided as a multi-threaded and multi-processor command line program written in C with user options to define the number of threads and processors. We provide pre-compiled binaries of SCARF for Linux and OS X, and knowledgeable users should be able to compile the available source code on alternative operating systems. In our tests, SCARF assembled half plate 454 runs against approximately 30 000 Sanger ESTs in <5 min on dual core 2.5 GHz Intel Xeon or Core 2 Duo machines running Ubuntu 8.04. Because SCARF uses parallel processing and multi-threading its memory usage can be significant, and at least 4 GB of RAM is recommended. We also provide a perl script for generating the homopolymer repeat distribution to determine the repeat correction threshold. Further installation details, system requirements and user options are described in the documentation provided with SCARF.

3 EXAMPLE USAGE

Relative to *de novo* assemblies, 454 ESTs assembled with SCARF demonstrate significant improvements. To test the performance of SCARF, we used 23 267 Sanger ESTs of *Centaurea solstitialis* generated by the Compositae Genome Project (Barker *et al.*, 2008) and approximately 100 000 454 EST reads for four different individuals of this species. Following *de novo* assembly of each individual 454 dataset with MIRA (Chevreux *et al.*, 2004), we post-assembled these sequences with SCARF against the Sanger EST references. On average, SCARF yielded a 6.04% increase in unigene length (excluding gaps) and a 9.4% decrease in unigene number. Significantly, SCARF increased the number of unigenes >1 kb by an average of 96%, or 443.3 unigenes, across the four EST datasets (Fig. 1). Similar results were obtained using Newbler *de novo* assemblies (data not shown). Although SCARF performs well when the reference sequences are of the same species as the 454 reads, its ability to further assemble contigs decreases abruptly with

references from other species (Supplementary Fig. S1). We selected four Compositae species with approximately the same number of Sanger EST sequences as *C. solstitialis* (Barker *et al.*, 2008), and observed a roughly exponential decline in the number of SCARF contigs with increasing genetic divergence of the reference. Based on this analysis, SCARF should be able to reasonably improve EST assemblies with reference sequences from species with <0.1 K_s (synonymous site) divergence.

4 CONCLUSIONS

Next-generation sequencing technologies are now making evolutionary and population genomic studies tractable for many organisms, and ESTs sequenced by 454 and similar technologies will likely form a large proportion of these datasets. SCARF alleviates some of the problems posed by low-coverage next-generation sequencing by leveraging all available data to improve the accuracy, particularly unigene copy number and homopolymer repeat lengths, and assembly of these sequences. The individual datasets generated by SCARF can form the basis for population genomic analyses because unlike many other assembly tools, SCARF does not include the reference sequence in the consensus for each unigene permitting subsequent analyses with each individual's FASTA formatted assembly. SCARF should also be useful in metagenomic analyses to simultaneously identify, sort and assemble 454 reads from disparate species with available reference sequences. In general, low-depth or non-contiguous sequencing strategies will benefit from SCARF's flexible assembly requirements.

ACKNOWLEDGEMENTS

We thank Matt King for providing OS X compilations of SCARF.

Funding: National Science Foundation Plant Genome Award (No. 0421630 to L.H.R.); Natural Sciences and Engineering Research Council of Canada (No. 353026 to L.H.R.).

Conflict of Interest: none declared.

REFERENCES

- Barker, M.S. *et al.* (2008) Multiple paleopolyploidizations during the evolution of the Compositae reveal parallel patterns of duplicate gene retention after millions of years. *Mol. Biol. Evol.*, **25**, 2445–2455.
- Chevreux, B. *et al.* (2004) Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. *Genome Res.*, **14**, 1147–1159.
- Margulies, M. *et al.* (2005) Genome sequencing in open microfabricated high density picoliter reactors. *Nature*, **437**, 376–380.
- Pertea, G. *et al.* (2003) TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. *Bioinformatics*, **19**, 651–652.
- Smith, T.F. and Waterman, M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.
- Zhang, Z. *et al.* (2000) A greedy algorithm for aligning DNA sequences. *J. Comput. Biol.*, **7**, 203–214.