Bioinformatics Advance Access published July 3, 2012

OmegaPlus: A Scalable Tool for Rapid Detection of Selective Sweeps in Whole-Genome Datasets

N. Alachiotis *, A. Stamatakis, P. Pavlidis *

The Exelixis Lab, Scientific Computing Group, Heidelberg Institute for Theoretical Studies

Associate Editor: Dr. Jeffrey Barrett

ABSTRACT

Summary: Recent advances in sequencing technologies have led to the rapid accumulation of molecular sequence data. Analyzing whole-genome data (as obtained from next-generation sequencers) from intra-species samples allows to detect signatures of positive selection along the genome and therefore identify potentially advantageous genes in the course of the evolution of a population.

We introduce OmegaPlus, an open-source tool for rapid detection of selective sweeps in whole-genome data based on linkage disequilibrium. The tool is up to two orders of magnitude faster than existing programs for this purpose and also exhibits up to two orders of magnitude smaller memory requirements.

Availability: OmegaPlus is available under GNU GPL at `http://www.exelixis-lab.org/software.html`.

Contact: pavlos.pavlidis@h-its.org

Supplementary information: Available at *Bioinformatics* online.

1 INTRODUCTION

Statistical tests that rely on selective sweep theory (Maynard Smith and Haigh, 1974) can localize targets of recent and strong positive selection by analyzing single nucleotide polymorphism (SNP) patterns in intra-species multiple sequence alignments. Next-generation sequencing technologies allow for the rapid and cheap extraction of data to assess whole-genome variation in single-species populations. However, applying such statistical tests to whole-genome data is challenging because of increased execution times and memory requirements.

Several studies in the last decade have focused on detecting positive selection in genomes from natural populations of individuals. Kim and Stephan (2002) developed a composite maximum likelihood (ML) framework to detect selective sweeps using the empirical mutation frequencies of polymorphic sites (site frequency spectrum). This approach was later adapted by Nielsen *et al.* (2005) for populations that have experienced past demographic changes. Kim and Nielsen (2004) proposed the ω statistic to accurately localize selective sweeps and developed a ML framework that uses linkage-disequilibrium (LD) information.

In this work, we introduce OmegaPlus, a high-performance implementation of the ω statistic. To the best of our knowledge, OmegaPlus represents the first scalable ω statistic implementation that *can* be applied to whole-genome data. Previous implementations were not suitable for whole-genome analyses: the implementation by Jensen *et al.* (2007) is only suitable for analyses of sub-genomic regions, while the implementation by Pavlidis *et al.* (2010) has excessive memory requirements even for moderate datasets.

2 THE LD PATTERN OF SELECTIVE SWEEPS

The LD is used to capture the non-random association of alleles or states at different alignment positions. Selective sweep theory predicts a pattern of excessive LD in each of the two alignment regions that flank a recently fixed beneficial mutation. This genomic pattern can be detected by using the ω statistic.

Assume a genomic region with S SNPs that is split into a left sub-region L and a right sub-region R with l and $S - l$ SNPs, respectively. The ω statistic is computed as follows:

$$\omega = \frac{(\binom{l}{2} + \binom{S-l}{2})^{-1}(\sum_{i,j \in L} r_{ij}^2 + \sum_{i,j \in R} r_{ij}^2)}{(l(S-l))^{-1} \sum_{i \in L, j \in R} r_{ij}^2}, \qquad (1)$$

where r_{ij}^2 is one common measure of LD that represents the correlation coefficient between sites i and j. Candidate regions for positive selection are characterized by high ω statistic values since the average LD is high within the L and R sub-regions but not across the beneficial mutation (see Section 1 in the supplement).

3 FEATURES

OmegaPlus is an open-source C code for Windows and Linux platforms. We provide a sequential version (OmegaPlus) for Windows and Linux, as well as two parallel versions for Linux that exploit fine-grain (OmegaPlus-F) and coarse-grain (OmegaPlus-C) parallelism (see Sections 2-6 in the supplement).

3.1 Computational workflow

OmegaPlus calculates the ω statistic at N distinct positions in the alignment. For each position p, $1 \leq p \leq N$, let k_L and k_R be the number of SNPs to the left and to the right of p, respectively, which are contained within a user-defined genomic region G. Thus, a number of k_L and k_R regions are located to the left and right side of p, each containing a different number of SNPs. The ω statistic is computed for all $k_L \times k_R$ pairs of sub-regions and the pair for which ω statistic has the highest value is reported.

*to whom correspondence should be addressed

To accelerate ω statistic calculations, a two-dimensional matrix M, $M_{l,m} = \sum_{l \le i < j \le m} r_{ij}^2$, is calculated for each region G. Thereafter, all sums in Equation 1 are retrieved from M. This avoids redundant calculations and keeps the memory requirements proportional to k^2, where k is the total number of SNPs in G.

3.2 Input/Output

OmegaPlus is a command line tool that only requires five input arguments for a typical run: an alignment file (`-input`), the number p of positions at which the ω statistic will be calculated (`-grid`), a minimum (`-minwin`) and maximum (`-maxwin`) size for the L and R sub-regions, and a name for the specific run (`-name`). Additional command line flags, such as the number of threads (`-threads`) for the parallel versions, are described in the manual. Binary data (derived/ancestral states) in ms (Hudson, 2002) or MaCS (Chen *et al.*, 2009) format and DNA data in FASTA format can be analyzed. For ms and MaCS files, the length of the alignment (`-length`) also needs to be provided by the user.

The program generates three text files: an information file, a warning file, and a report file. The information file contains details about program execution, such as the command line and the run time. The warning file reports conflicting SNP positions (SNPs which refer to the same alignment positions) when binary data are analyzed. Finally, the report file provides the highest ω statistic value for each position p. A detailed description of the input and output options is provided in the manual.

4 USAGE EXAMPLE

To demonstrate the ability to efficiently process genome-datasets, we used an off-the-shelf laptop with an Intel Core i5 CPU @ 2.53 GHz with 4 GBs main memory running Ubuntu 10.04.3. The X chromosome of 37 *Drosophila melanogaster* genomes sampled in Raleigh, North Carolina was analyzed. The sequences (available from the Drosophila Population Genomics Project at `www.dpgp.org`) were converted from fastq to fasta using a provided script (available at `www.dpgp.org`) and Solexa quality score threshold of 30. OmegaPlus was called as follows:

```
./OmegaPlus-F -maxwin 100000 -minwin 1000
  -name DPGP -input DPGP.fa -grid 10000 -threads 4.
```

We calculated the ω statistic at 10,000 equidistant positions, assuming that the maximum length of the L and R sub-regions to the left and right of a beneficial mutation is 100,000 base pairs. Thus, a selective sweep may affect at most 200,000 base pairs in total. The minimum length of L and R is set to 1,000 base pairs. A few entries of the report file are shown below. Each row denotes an alignment position p and the OmegaPlus score. In this example, there is a peak at position 1,079,493.

```
1077252 1.640796
1079493 48.282642
1081734 2.086421
```

The alignment comprised 37 sequences and 22,422,827 sites (339,710 SNPs). To analyze such an alignment, OmegaPlus used less than 2% of the available 4 GBs of main memory. Note that a peak memory requirement of 20.8% was recorded for some seconds to temporarily store the alignment in memory, discard non-polymorphic sites, and compress the remaining SNPs. Using both physical cores on the laptop and hyper-threading (two more virtual cores), the fine-grain version of OmegaPlus required 15 minutes. Assuming that a selective sweep only affects at most 60,000 base pairs (`-maxwin 30000`) instead of 100,000, the analysis only takes 5 minutes.

5 PERFORMANCE

An accuracy evaluation of the ω statistic and comparison to SweepFinder (Nielsen *et al.*, 2005) has been presented in Pavlidis *et al.* (2010). To provide a performance comparison, we generated a simulated dataset with 500 sequences and 70,000 SNPs using Hudson's ms (see Section 5 in the supplement), and employed a 12-core AMD Opteron CPU running at 2.2 GHz. OmegaPlus analyzed this dataset in 71.8 seconds occupying 75.8 MBs of main memory, while SweepFinder required 783.2 seconds and 2.7 MBs of main memory. Note however that SweepFinder employs the site frequency spectrum and not the ω statistic. The ω statistic implementation by Pavlidis *et al.* (2010), which attempts to process the entire alignment at once, required more than 2 days and allocated approximately 30 GBs of main memory.

To evaluate the parallel implementations, we generated a dataset with 1,000 sequences and 500,000 alignment sites. For this dataset size, fine-grained parallelism scales better on DNA data. Coarse-grained parallelism exhibits similar behavior on both data types and scales better on binary data (see Section 6 in the supplement).

6 CONCLUSION

We presented OmegaPlus to accurately identify selective sweeps in whole-genome data using the ω statistic. We intend to further improve the memory requirements as well as to support additional file formats and methods to calculate LD. We also plan to set up a web-server. User support is provided via the OmegaPlus Google group `http://groups.google.com/group/omegaplus`.

ACKNOWLEDGEMENT

Funding: Part of this work was funded by the DFG grant STA 860/2.

REFERENCES

Chen, G. K., Marjoram, P., and Wall, J. D. (2009). Fast and flexible simulation of dna sequence data. *Genome Res*, **19**(1), 136–142.

Hudson, R. R. (2002). Generating samples under a wright-fisher neutral model of genetic variation. *Bioinformatics*, **18**(2), 337–338.

Jensen, J. D., Thornton, K. R., Bustamante, C. D., and Aquadro, C. F. (2007). On the utility of linkage disequilibrium as a statistic for identifying targets of positive selection in nonequilibrium populations. *Genetics*, **176**(4), 2371–2379.

Kim, Y. and Nielsen, R. (2004). Linkage disequilibrium as a signature of selective sweeps. *Genetics*, **167**(3), 1513–1524.

Kim, Y. and Stephan, W. (2002). Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics*, **160**(2), 765–777.

Maynard Smith, J. and Haigh, J. (1974). The hitch-hiking effect of a favourable gene. *Genet Res*, **23**(1), 23–35.

Nielsen, R., Williamson, S., Kim, Y., Hubisz, M. J., Clark, A. G., and Bustamante, C. (2005). Genomic scans for selective sweeps using snp data. *Genome Res*, **15**(11), 1566–1575.

Pavlidis, P., Jensen, J. D., and Stephan, W. (2010). Searching for footprints of positive selection in whole-genome snp data from nonequilibrium populations. *Genetics*, **185**(3), 907–922.