F1000 evaluations of Markova and Petrov, Genome Research, 2011.

Evaluated by Juan Escobar and Stephen Wright 16 Jun 2011 | Daniel Croll and Bruce McDonald

This study highlights important concerns associated with gene alignment methods for the inference of molecular evolution patterns, particularly the detection of positive selection on genes and sites within genes. It strengthens the idea that obtaining reliable alignments is one of the most important, yet often overlooked aspects of research in the genomics era.

Obtaining genomic data from large numbers of species has now become a feasible reality. In comparative genomics, these data serve to understand the way in which DNA sequences are affected by different evolutionary processes (e.g. selection, genetic drift and mutation bias). One important question that can be addressed is what fraction of genes and nucleotide sites have been subject to positive selection. However, a critical step to determine this is to obtain reliable alignments of orthologous sequences. While manual editing of sequence alignments after running programs was typically standard before genome-wide datasets, this is no longer feasible given the scale of genomic datasets, and we are increasingly reliant on the accuracy of automated aligners.

In this paper, the authors evaluate how the inference of positive selection (inferred through the ratio of non-synonymous to synonymous substitution rates) is affected by the alignment method. For this, they compare different aligner programs (AMAP, MUSCLE, ProbsCons, T-Coffee, ClustalW and Prank) using sequences from the 12 Drosophila Genomes Consortium. Surprisingly, they find that the number of genes (and sites within genes) inferred to be under positive selection varies substantially with the aligner program, with false positive rates generally extremely high, at 48%-82%. They identify sequence features that might be at the origin of false positive rates, including bad alignments due to annotation problems, misinference of intron positions, alternative splicing, amino acid repeats and the presence of indels in fast evolving pockets located in between well-conserved regions. The paper demonstrates that different alignment methodologies can have a crucial impact on the estimates of positive selection acting at the molecular level.
Although problems detected in this paper are serious (and to some extent currently inevitable), the study highlights the importance of building in measures of alignment quality and measures of uncertainty into analyses. In any case, obtaining alignments is a critical step in most molecular evolution studies and the problems introduced at this step should be fully considered in order to get precise estimates of the levels of selection acting at the genome level in different organisms.

Competing interests: None declared

- Cite this evaluation

Escobar J, Wright S: "This study highlights important concerns associated with gene alignment methods for the inference of molecular..." Evaluation of: [Markova-Raina P, Petrov D. High sensitivity to aligner and high rate of false positives in the estimates of positive selection in the 12 Drosophila genomes. Genome Res. 2011 Jun; 21(6):863-74; doi: 10.1101/gr.115949.110]. Faculty of 1000, 16 Jun 2011.
F1000.com/11045956

Short form
Escobar J, Wright S: 2011. F1000.com/11045956

Evaluated by:
Juan Escobar and Stephen Wright
University of Toronto, Canada
Plant Biology
16 Jun 2011
Rating 8

Must Read

This paper by Markova-Raina and Petrov is interesting because it reiterates a critical issue raised already by other authors (e.g. {1}) concerning evolutionary analyses at the genomic scale. The authors provided a careful and very insightful analysis of the fundamental step of properly aligning orthologous sequences found in different genomes. The authors showed that incorrectly aligned sets of orthologous genes of 12 Drosophila genomes greatly inflated the estimates of codons under positive selection.

The inference of selection acting on codons in a given gene critically depends on the proper alignment of the coding sequence to reflect, as accurately as possible, the evolutionary history of the gene. However, careful visual inspection and masking of unalignable segments is only feasible for a small number of genes. Analyses at the genomic scale require automation, including checks to ensure high alignment quality. The authors used the predicted number of positively selected sites in a well-defined set of orthologous genes to compare the performance of different aligners. Even though some aligners performed notably better than others (especially the phylogeny aware PRANK aligner), the rate of false positives was unacceptably high. High-level interpretations such as the overall strength of selection acting on genomes and the enrichment in certain Gene Ontology (GO) terms under accelerated evolution may be affected.

As Markova-Raina and Petrov discuss in their paper, there is no trivial workaround to avoid alignment errors. However, the authors showed that a combination of careful quality checks may go a long way in avoiding the majority of errors in large-scale genomic comparisons. As comparative genomics is increasingly applied to non-model organisms lacking extensive manual annotation and experimental data, the awareness of fundamental biases in evolutionary analyses is becoming even more important.

References:
{1} Wong et al. Science 2008, 319:473-6 [PMID:18218900].

Competing interests: None declared