BIOINFORMATICS APPLICATIONS NOTE Vol. 16 no. 8 2000
Pages 743–744

gff2ps: visualizing genomic annotations

Josep F. Abril and Roderic Guigó*

Grup de Recerca en Informàtica Mèdica, Institut Municipal d'Investigació Mèdica (IMIM), Universitat Pompeu Fabra (UPF), C/ Dr. Aiguader, 80. 08003—Barcelona, Spain

Received on December 15, 1999; revised on February 18, 2000; accepted on February 24, 2000

Abstract

Summary: *gff2ps is a program for visualizing annotations of genomic sequences. The program takes the annotated features on a genomic sequence in GFF format as input, and produces a visual output in PostScript. While it can be used in a very simple way, it also allows for a great degree of customization through a number of options and/or customization files.*

Availability: *gff2ps is freely available at http://www1.imim.es/~jabril/GFFTOOLS/GFF2PS.html*

Contact: *jabril@imim.es*

Supplementary information: *http://www1.imim.es/~jabril/GFFTOOLS/GFF2PS.html*

As genomic sequences accumulate, visualization tools are becoming essential for the analysis and interpretation of sequence data. Recently, a format has been proposed for specifying genes and other features associated with genomic sequences, the General Feature Format (GFF, proposed by Durbin and Haussler, http://www.sanger.ac.uk/Software/GFF/). In GFF each feature on the genomic sequence is described in a single-line record that essentially specifies the type and position of the feature on the genomic sequence. A grouping field allows to define sets of features within the GFF file. A number of tools have already been developed to deal with GFF files (see also at GFF URL). We have developed a tool, `gff2ps`, which allows for visualization of GFF files. `gff2ps` is a program written in GNU awk (http://www.gnu.org/software/gawk/gawk.html) and PostScript, running on UNIX platforms, that generates a PostScript file given a GFF file.

The page description language PostScript is recognized as the current *de facto* industry standard for high-quality printing. PostScript provides both a printer-independent and a computer-system-independent means to describe integrated text and graphics, which can be put out on a variety of printers, plotters and workstation screens. The generation of PostScript output is very common in sequence analysis tools. Notably, we can cite the RSVP package by Searls (1993).

`gff2ps` plots the features from different sources specified on a GFF file in a number of parallel rows (the so-called tracks here) along the length of the output page(s) (see Figure 1 for examples). Actually these are 'virtual' pages (the so-called blocks here) allowing for several blocks to be included in a single physical page, or for splitting a single block in a number of physical pages. Features can be plotted in a variety of colors and shapes and those grouped together can be visually linked in a number of ways.

`gff2ps` allows for a substantial amount of customization through command line options, and configuration files. However, meaningful output in most cases, meaningful output can be obtained without the need of any customization, by simply calling `gff2ps` with the input GFF file. `gff2ps` assumes, by default, that the GFF file itself carries enough formatting information. The examples in the figure show the versatility of `gff2ps`. Additional examples can be found at the `gff2ps` web page, as well as a detailed User Manual.

One of the main advantages of `gff2ps` is its ability to manage many physical page formats, including user-defined ones. This allows, for instance, the generation of poster size genomic maps. As an example, we used `gff2ps` to display at the ISMB'99 meeting, the predictions submitted to the Genome Annotation Assessment Project (GASP1) (http://www.fruitfly.org/GASP1/). The GASP1 plot was generated on three B0 size posters from a GFF file of over 50 000 feature records. The program has also been used to obtain the poster figures of recent relevant papers in genomic research (Adams *et al.*, 2000; Reese, 2000).

Acknowledgments

We thank Moisès Burset and Genís Parra (IMIM) for their useful comments, Richard Bruskiewich (Sanger Center) for his helpful hints on the GFF format, also Elena Casacuberta and Amparo Monfort (CSIC) for motivating us to develop this tool. This work is supported by a grant from Plan Nacional de I+D, BIO98-0443-C02-01, and from a fellowship to J.A. from the Instituto de Salud Carlos III, 99/9345.

*To whom correspondence should be addressed.

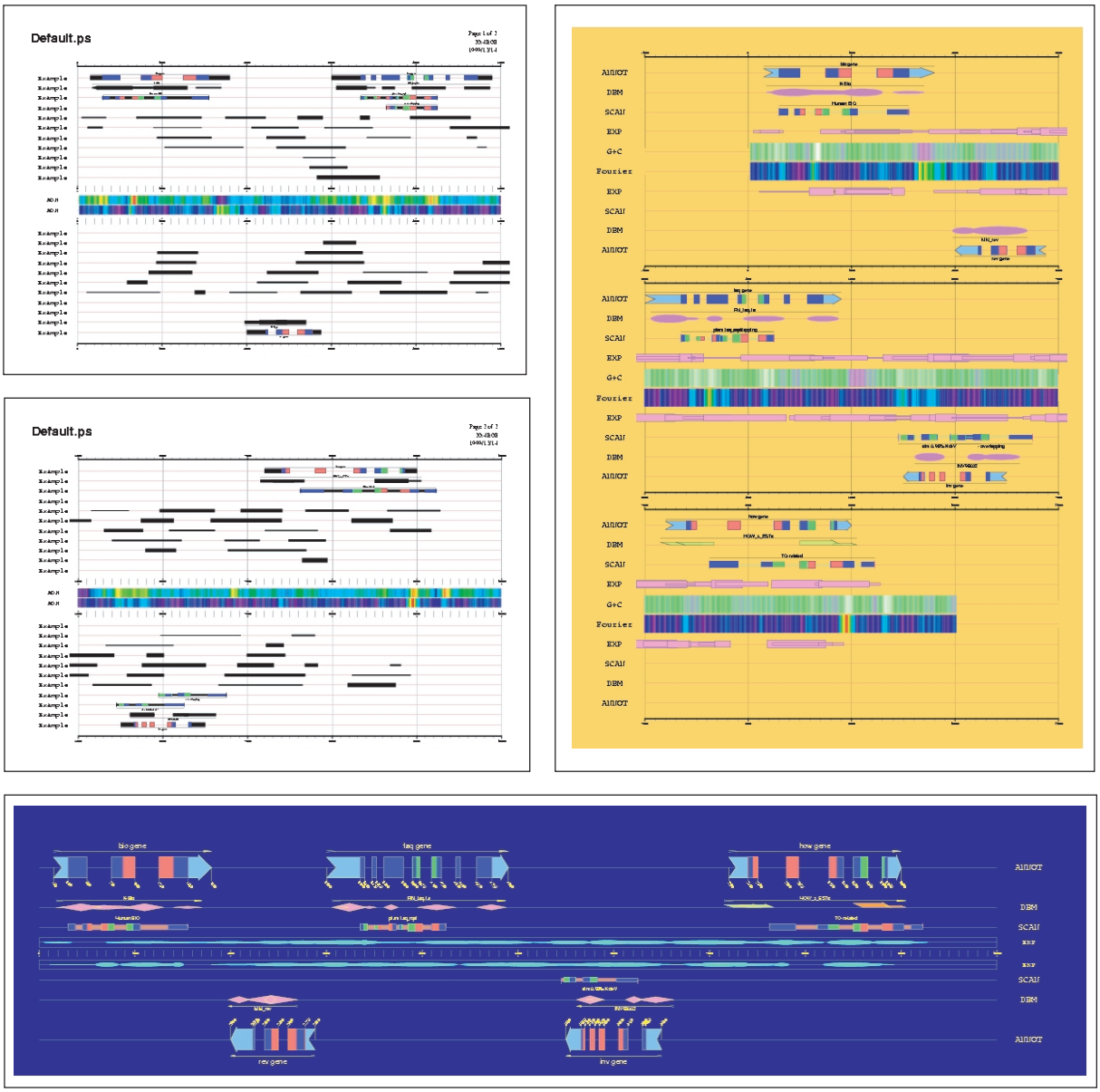

Fig. 1. Different views of the same input GFF file, using `gff2ps` with different configuration files and command-line options. The top two pages on the left were obtained using the default configuration (specifying only the number of pages, and output media size). By default, `gff2ps` makes a number of assumptions. Among others: (i) Features grouped from the GFF input file (ungrouped features are treated as a single element group) within the same source are plotted in the minimum number of tracks, guaranteeing that different groups do not overlap. (ii) The plot is fitted into a single block (assuming the length of the sequence to be the end of the most downstream feature), and the block is printed into a single physical page. (iii) Features for which the frame is specified are plotted using a two color code schema. The upstream half of the graphical element representing the frame of the feature and the downstream half the complement modulus 3 of its remainder. This is useful to check frame consistency between adjacent features (for instance, predicted exons). Two adjacent features are frame compatible when the color of the downstream half of the upstream feature matches the color of the upstream half of the downstream feature. (iv) If a score is provided for a feature, the feature is plotted with a height proportional to its score. (v) Obviously, all these default options can be overridden by the user. Notes: The real size color plots, the input GFF files, the configuration files and command line options used in each case, as well as additional examples can be found at: http://www1.imim.es/~jabril/GFFTOOLS/GFF2PS-Snapshots.html.

References

Adams,M.D. and Abril,J.F. *et al.* (2000) The genome sequence of *Drosophila melanogaster*. *Science*, **287**(5461), 2185–2195.

Reese,M.G., Hartzell,G., Harris,N.L., Ohler,U., Abril,J.F. and Lewis,S.E. (2000) Genome annotation assessment in *Drosophila melanogaster*. *Genome Res.*, **10**, 483–501.

Searls,D.B. (1993) Doing sequence analysis with your printer. *Comput. Appl. Biosci.*, **9**(4), 421–426.