Cabanettes F, Klopp C. 2018. D-GENIES: dot plot large genomes in an interactive, efficient and simple way. PeerJ 6:e4958
https://doi.org/10.7717/peerj.4958

D-GENIES : Dot plot large GENomes in an Interactive, Efficient and Simple way

Floréal Cabanettes[1] **and Christophe Klopp**[1]

[1]**Plateforme Bio-informatique Toulouse Genopole, MIAT INRA, Castanet-Tolosan, France**

Corresponding author:
Christophe Klopp[1]

Email address: christophe.klopp@inra.fr

ABSTRACT

Dot plots are widely used to quickly compare sequence sets. They provide a synthetic similarity overview, highlighting repetitions, breaks and inversions. Different tools have been developed to easily generated genomic alignment dot plots, but they are often limited in the input sequence size. D-GENIES is a standalone and WEB application performing large genome alignments using minimap2 software package and generating interactive dot plots. It enables users to sort query sequences along the reference, zoom in the plot and download several image, alignment or sequence files. D-GENIES is an easy to install open source software package (GPL) developed in Python and JavaScript. The source code is available at `https://github.com/genotoul-bioinfo/dgenies` and it can be tested at `http://dgenies.toulouse.inra.fr/`.

INTRODUCTION

Dot plots are commonly used to visually compare two sets of sequences. They present insertions, deletions, inversions or repeats in an easily understandable manner. They can represent similarity differences using variable line thickness, line forms or colors. With the increasing numbers of genome assemblies produced there is a need for simple-to-use and efficient tools to produce dot plots of large genomes.

Existing Dot plot tools can be classified in two generations. The first, and oldest, comprises command line tools producing static graphics and includes among others tupple_plot Szafranski et al. (2006) and dot-matrix Sonnhammer and Durbin (1995). They usually chain two processing steps, the first of which produces a match files used in the second step which renders the graphical output. They are often limited to single sequence fasta files and do not enable any interaction with the produced graphic. Both mentioned tools are only running on Unix computers. The software packages of the second generation have been developed in java in order to be platform independent and user friendlier. They include tools such as JDotter Brodie et al. (2004), Gepard Krumsiek et al. (2007) and r2cat Husemann and Stoye (2009). The user interaction permits to add new dynamic features such as sequence orientation and ordering to maximize the diagonal alignment matches in order to ease the visual comparison. These tools are also limited in the size of processed sequences. For example, Guepard takes over an hour to align Human chromosomes 1 versus itself and plot the result.

A new generation of JavaScript based dot plot visualization tools emerges. One of its early members is `https://dnanexus.github.io/dot/`. To render the graphic, users have to generate the coordinate and index files. They can add annotations which will be displayed in the graph margins.

We present hereafter D-GENIES, an interactive, rapid and easy to use standalone and WEB application permitting to produce a complete human versus chimpanzee genome dotplot in one hour and ten minutes.

PROGRAM FEATURES

Fast dotplot computation

D-GENIES takes advantage of minimap2 Li (2017), one of the latest nucleic sequence alignment program which is able to map very large lowly similar multi-fasta files. D-GENIES can only produce dot plots for nucleic sequences. In order to limit memory consumption and lower processing time, the program

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26567v1 | CC BY 4.0 Open Access | rec: 23 Feb 2018, publ: 23 Feb 2018

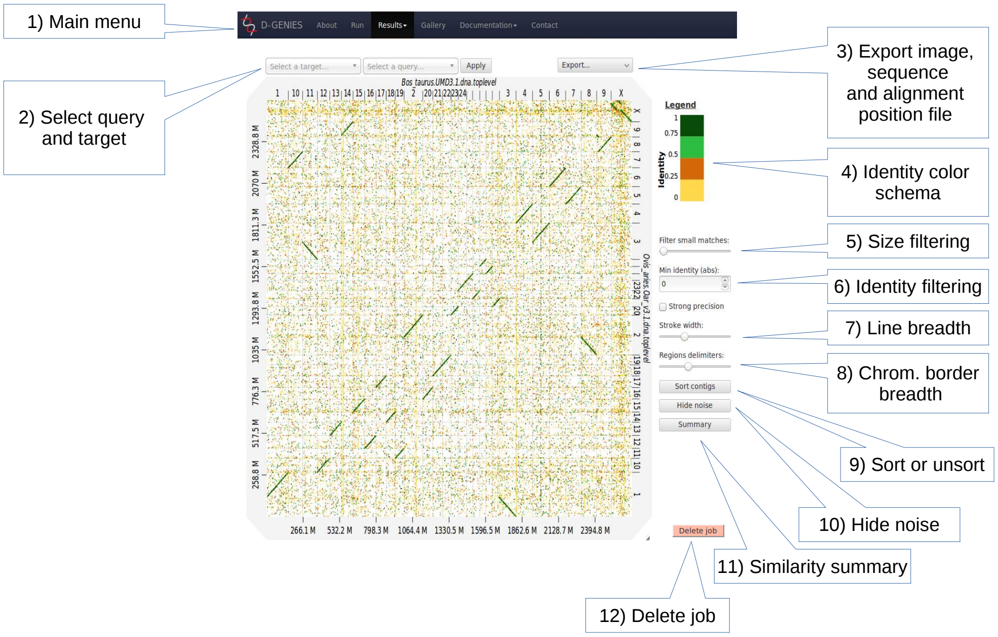


Figure 1. Result page view. 1) main menu to navigate D-GENIES pages. 2) reference and query sequence drop down selection boxes and button to zoom in the alignment. 3) Export menu to download image files (PNG and SVG), alignment, ordered query and unaligned query or reference fasta files. 4) Identity color panel. 5) Match size filtering slider. 6) Identity filtering entry and check boxes. 7) Line width slider. 8) Reference and query border horizontal and vertical border line slider. 9) Query sort and unsort button. 10) Noise filtering button. 11) Similarity summary button. 12) Delete job button.

splits large sequence queries, such as chromosomes, in ten mega-base chunks and merges consecutive matches to produce the final alignment file. Processing time and memory consumption are presented in the corresponding paragraph hereafter. minimap2 can easily be replaced by any other aligner generating PAF (Pairwise mApping Format) files.

Simple interactive user friendly interface

The PAF file is rendered in the dotplot by a javascript client developed with d3.js `https://d3js.org/`. To limit drawing time, only the hundred thousand largest alignments are shown in the dotplot.

Both, standalone and WEB application are accessed through a WEB-browser. The page top menu **(Fig 1.)** permits to launch a new alignment, visualize results, browse the example gallery or documentation and send an email for support. To produce a dot plot, a user clicks on the **Run** menu item and fills three input boxes. A modifiable job name is automatically attributed to the dot plot. The user email address is mandatory. The application will send a message once the dot plot is rendered. Both, query and target fasta files can be uploaded from the local machine or given URLs. Reference and query files can be compressed in gzip format. If no query file is provided, the reference will be aligned on itself and all trivial matches corresponding to same sequence and same positions will be removed. After hitting the submit button, the user can follow the upload and processing progression presented with different texts and progress bars. Once the job is ended, an email containing the result page link is sent to the user. The same link appears in the monitoring page. If a user has several stored results, they can be accessed using the drop down menu of the **Results** menu item.

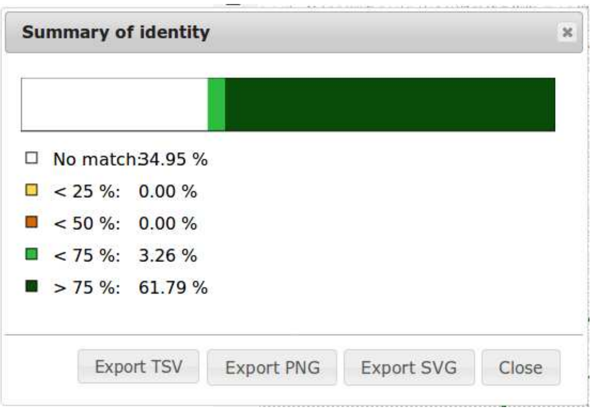
The result page (fig. 1), when first accessed, presents the dot plot following the fasta files sequence order. The alignment matches are presented as colored lines on the graphical panel. The colors correspond to similarity values which have been binned in four groups (less than 25%, between 25 and 50%, between

50 and 75% and over 75% similarity). For colorblind users, clicking on the color scale modifies the schema. Three color schema are already available and others can be easily added. The graphical panel top and right margins display sequence names. Depending on the sequence and name lengths, the names will be fully or partially presented. In order to ease visualization, all sequences smaller than 0.2 percent of the total length are merged in a unique super-sequence for which the margin is grayed. The left and bottom margins show the sequence size scales.

At the to top of the graphical panel, the user will find, on the left, two drop down text areas and a button enabling to select query and target sequences to zoom to, and on the right the **Export** menu. The other way of zooming in the graphical panel is to click on a given square or to push the CRTL key while turning the mouse wheel forward to zoom in and backward to zoom out. To come back to the initial view the user will click on the icon in the top right angle of the graphical panel, or press the escape key. The **Export** menu enables to retrieve the graphic as a PNG or SVG file, suited for publication, the PAF match file and the association table which links each query with the corresponding reference sequence, as well as the ordered query fasta file. The unaligned query and reference sequences as well as the query sequences reorganized following the reference organization can also be retrieved using this menu.

On the graphic right **(Fig 1.)**, users will have access to several buttons, sliders and input boxes enabling to change color schema, filter matches on their similarity and size or because they are seen as noise, modify match or border size as well as sort query sequences relatively to the reference. A match is considered noise if its size is small and its size frequency is quite high. Therefore we group matches by size bins, the number of bins corresponds to one tenth of the number of alignments, the bins are scanned in increasing size order to find the most represented one and from this one the one corresponding to one percent of its count is searched. All the alignments in bins smaller in size than this one are considered noise. The delete job button located at the bottom right of the diagram can be used to discard obsolete results.

If after sorting, the query sequence orientation does not not correspond to the users expectation, it can be changed by right clicking in the graphic and selecting **Reverse query**. Right clicking enables also to export the complete graphic in PNG or SVG format.



□ No match34.95 %
🟨 < 25 %: 0.00 %
🟧 < 50 %: 0.00 %
🟩 < 75 %: 3.26 %
■ > 75 %: 61.79 %

Figure 2. Example of identity summary

To ease dot plot comparison, clicking the **Summary** button generates a bar graph presenting the reference similarity profile **(Fig 2.)**, meaning the sums of the projections of the matches on the reference per similarity category divided by the total reference length. This graph is produced after sorting the query along the reference, removing included matches and noise filtering; result not shown on the graphical panel. It gives a realistic view of the overall reference and query similarity which is often not very precisely measured through visual inspection.

All these features are documented in the **Documentation** menu item of the main menu. The **Gallery** menu item give access to several examples also presented in the "processing time" section of this article.

Easy stand-alone or WEB server installation

D-GENIES can be installed and run as a stand-alone application on Unix or MS-Windows or as a WEB server on Unix only. It uses Flask framework `http://flask.pocoo.org/` back-end to serve WEB pages and submit jobs. In stand-alone mode only one process should be run in a given instance. In WEB server mode several processes will be run simultaneously. Three steps are time and disk space consuming when working with large genomes : file upload, alignment and data preparation. D-GENIES uses three mechanisms to ensure robustness.

When installed as a WEB server, it can use a computer cluster to run the memory and disk intensive processes through the DRMAA layer. It also uses a local scheduler storing jobs in a MySQL database which defines process order and manages concurrency on the available cores. Because files can be large and may saturate the server their size is tested before upload. DRMAA, MySQL parameters and maximum file size are set in the configuration file.

The file folder storing the input and output files can be cleaned using the delete job button in stand-alone or WEB server mode. A cron job deleting files having more than a given number of days can also be launched periodically in WEB server instances.

The software package can be installed using the - pip install dgenies - command and and run with the - dgenies run - command. By default, under Unix, all data is stored in the user .dgenies folder and if needed the application.properties configuration file located in the /etc/dgenies folder can be updated.

The source code can be downloaded from https://github.com/genotoul-bioinfo/dgenies.

Memory consumption and processing times

D-GENIES has been tested on various reference and query fasta files coming from Ensembl 91 `https://www.ensembl.org/`. Test results presented in **Table 1.** have been performed on a 32 cores Intel(R) Xeon(R) CPU E5-2670 v2 @ 2.50GHz with 256GB RAM server, using 4 cores for the minimap2 alignments. The results of the tests are presented in table 1.

Reference genome	Query genome	CPU time	Maximum RAM usage
Human (3.5 Gb)	Chimpanzee (3.4 Gb)	67 minutes 14 seconds	36 GB
Mouse (3.4 Gb)	Rat (3.0 Gb)	39 minutes 54 seconds	24 GB
Cow (2.6 Gb)	Sheep (2.5 Gb)	43 minutes 3 seconds	27 GB
A. thaliana (135 Mb)	A. lyrata (206 Mb)	1 minute 4 seconds	2GB
Poplar (417 Mb)	Vine (486 Mb)	3 minutes 21 seconds	8GB
Brassica rapa (284 Mb)	Brassica rapa (284 Mb)	2 min 52 s	8.3 Gb

Table 1. Processing time and memory consumption table for Ensembl 91 datasets.

WEB portal

D-GENIES can be tested using the `http://dgenies.toulouse.inra.fr/` portal which permits to process up to 3 Gb reference and query sequence fasta files.

CONCLUSION

New alignment algorithms and JavaScript visualization libraries enable to develop a third generation of dot plot applications. This generation is able to process large genomes in reasonable time and provides user-friendly graphical interfaces. Even if D-GENIES has been developed to process large genomes it is also suited for small or medium size genomes.

ACKNOWLEDGMENTS

We thank Céline Noirot, Léo Lamy and the Bioinfo Genotoul platform system engineers and users for there feedback and help during installation and tests as well as Damien Leroux for his precious advises concerning JavaScript and Flask.

REFERENCES

Brodie, R., Roper, R. L., and Upton, C. (2004). Jdotter: a java interface to multiple dotplots generated by dotter. *Bioinformatics*, 20(2):279–281.

Husemann, P. and Stoye, J. (2009). r2cat: synteny plots and comparative assembly. *Bioinformatics*, 26(4):570–571.

Krumsiek, J., Arnold, R., and Rattei, T. (2007). Gepard: a rapid and sensitive tool for creating dotplots on genome scale. *Bioinformatics*, 23(8):1026–1028.

Li, H. (2017). Minimap2: fast pairwise alignment for long dna sequences. *arXiv preprint arXiv:1708.01492*.

Sonnhammer, E. L. and Durbin, R. (1995). A dot-matrix program with dynamic threshold control suited for genomic dna and protein sequence analysis. *Gene*, 167(1):GC1–GC10.

Szafranski, K., Jahn, N., and Platzer, M. (2006). tuple_plot: Fast pairwise nucleotide sequence comparison with noise suppression. *Bioinformatics*, 22(15):1917–1918.