Bioinformatics, 34(19), 2018, 3377–3379
doi: 10.1093/bioinformatics/bty336
Advance Access Publication Date: 25 April 2018
Applications Note

Genome analysis

Cluster Locator, online analysis and visualization of gene clustering

Flavio Pazos Obregón[1,*], Pablo Soto[1], José Luis Lavín[2], Ana Rosa Cortázar[2], Rosa Barrio[3], Ana María Aransay[2,4] and Rafael Cantera[1,5]

[1]Developmental Neurobiology, IIBCE, 11600 Montevideo, Uruguay, [2]Genome Analysis Platform and [3]Functional Genomics Unit, CIC bioGUNE, 48160 Derio, Spain, [4]CIBERehd, ISCIII, 28029 Madrid, Spain and [5]Zoology Department, Stockholm University, 10691 Stockholm, Sweden

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on July 14, 2017; revised on April 16, 2018; editorial decision on April 22, 2018; accepted on April 24, 2018

Abstract

Summary: Genes sharing functions, expression patterns or quantitative traits are not randomly distributed along eukaryotic genomes. In order to study the distribution of genes that share a given feature, we present Cluster Locator, an online analysis and visualization tool. Cluster Locator determines the number, size and position of all the clusters formed by the protein-coding genes on a list according to a given maximum gap, the percentage of gene clustering of the list and its statistical significance. The output includes a visual representation of the distribution of genes and gene clusters along the reference genome.

Availability and implementation: Cluster Locator is freely available at http://clusterlocator.bnd.edu.uy/.

Contact: fpazos@iibce.edu.uy

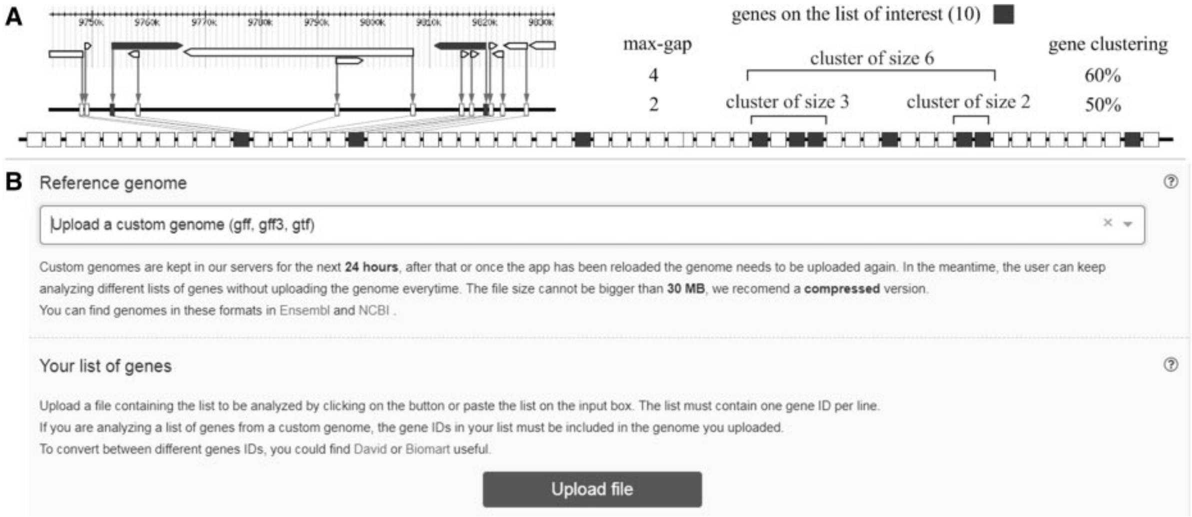
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

It has been well established that genes are not randomly located along eukaryotic genomes (Feuerborn and Cook, 2015; Hurst *et al.*, 2004). Diverse inter-correlations between gene location and gene expression, gene functions or quantitative traits have been found in all eukaryotes studied so far (De and Babu, 2010; Ghanbarian and Hurst, 2015). These correlations were first observed almost 20 years ago in the yeast Saccharomyces (Eisen *et al.*, 1998) and later on in nematodes, flies, mice, humans and other organisms (Michalak, 2008). Using diverse definitions for the term ‘cluster’, several studies have found clusters of co-expressed genes that share function, clusters of functionally related genes that share neighborhood in the genome or groups of nearby genes with similar expression patterns or related functions (Corrales *et al.*, 2017; Lee and Sonnhammer, 2003; Reimegård *et al.*, 2017; Thévenin *et al.*, 2014; Tiirikka *et al.*, 2014; Yi *et al.*, 2007). Thus, it is now accepted that the relative location of a gene in the genome is not independent of its biological function or its expression pattern.

In recent years, due to the refinement in genome annotation and a growing abundance of gene expression data, constructing lists of co-functional or co-expressed genes has become relatively easy. Nevertheless, there is a lack of tools allowing a straightforward statistical analysis of the way in which the genes on a list are clustered along the genome. Some tools that have been developed could provide insight on this (Aboukhalil *et al.*, 2013; Dottorini *et al.*, 2013; Yi *et al.*, 2007) but they were neither specifically designed to do so, nor currently available online or after request.

Here we present Cluster Locator, a free online and user-friendly tool that, given a list of protein-coding genes provided by the user, and after selecting a maximum gap allowed (see definition below), locates, quantifies and displays all the clusters formed by the genes on the list. The output of Cluster Locator, displayed in the browser and available to download, includes the number, size and position of the clusters identified, the identity and position of the genes in each cluster, as well as a statistical analysis of the results.



Reference genome

Upload a custom genome (gff, gff3, gtf)

Custom genomes are kept in our servers for the next **24 hours**, after that or once the app has been reloaded the genome needs to be uploaded again. In the meantime, the user can keep analyzing different lists of genes without uploading the genome everytime. The file size cannot be bigger than **30 MB**, we recomend a **compressed** version.
You can find genomes in these formats in Ensembl and NCBI .

Your list of genes

Upload a file containing the list to be analyzed by clicking on the button or paste the list on the input box. The list must contain one gene ID per line.
If you are analyzing a list of genes from a custom genome, the gene IDs in your list must be included in the genome you uploaded.
To convert between different genes IDs, you could find David or Biomart useful.

Fig. 1. (**A**) Definitions. Each genome is modeled as a collection of segments in which protein-coding genes are allocated side by side, without inter-genic regions or overlapping. The gap between two genes is the number of other genes starting points located between them. Given a max-gap, a cluster is the maximal set of genes for which the gap between adjacent genes is never larger than the max-gap. The size of a cluster is the number of genes that are included in the cluster. For a given max-gap, the percentage of gene clustering is the percentage of the genes on the list that belongs to any of the identified clusters. (**B**) Screenshot of Cluster Locator's User Interface showing the drop-down menu available to select or upload a custom genome and the button to upload the list of genes to be analyzed

2 Implementation

Cluster Locator is a web-based application implemented in Python 2.7 on the backend, and uses ReactJS and D3js libraries on the frontend. The backend is deployed on AWS Lambda, and the frontend static files are stored on AWS S3 Storage.

3 Definitions

We model each genome as a collection of segments corresponding to chromosomes (or chromosomal arms or scaffolds, depending on the reference genome used), in which protein-coding genes are allocated side by side, without intergenic regions or overlap (see Fig. 1A). The gap between two given genes is the number of other genes located between them, defined by the location of their starting points in the reference genome (Roy *et al.*, 2002). A **cluster** is the maximal set of genes for which the gap between adjacent genes is never larger than a given maximum gap ('**max-gap**', defined by the user). The **cluster size** is the number of genes that are included in the cluster. For a given max-gap, **the percentage of gene clustering** is the percentage of the genes on the list that belong to any of the identified clusters. Clusters can be formed only by genes in the same segment.

4 General features

Cluster Locator can analyze lists of protein-coding genes from any organism. The list can have a maximum of 1000 genes (one gene identifier per line) and can be provided either as a *txt or as a *csv file or directly pasted in an input box. The current version includes pre-loaded genomes of five organisms that are among the most studied, have well-annotated genomes, are widely spread through the eukaryotic phylogeny and diverse in terms of number and type of chromosomes (*Homo sapiens, Mus musculus, Drosophila melanogaster, Caenorhabditis elegans* and *Saccharomyces cerevisiae*). For each of these organisms there are at least 2 supported gene identifiers. The centromeres were taken into account to model the segments along which the clusters can be defined. The analysis of lists of genes from these five organisms is faster than of custom genomes because the statistical parameters are pre-computed. To analyze lists of genes belonging to any other organism, the user must upload the corresponding genome in gff, gff3 or gtf formats. In this case the analysis will not consider the centromeres when modeling the replication units and only the gene identifiers included in the uploaded genome will be supported.

The user's interface of Cluster Locator is very simple. First, a drop-down menu allows choosing the genome to which the genes in the list of interest belongs to (Fig. 1B). If the genome is not pre-loaded, the user can upload it in some of the supported formats, choosing '*Upload a custom genome*' in the same drop-down menu and browsing the genome file. Then the list of genes to be analyzed must be uploaded as a text file or directly pasted in an input text box. Finally, the user selects the max-gap that will be used to define clusters and hits the 'run' button. After a few seconds Cluster Locator displays the results that include the number, size and position of the clusters found in the analyzed list, the identity and position of the genes in each cluster, the percentage of gene clustering found, as well as the results of the statistical analysis. Below these results, which can be downloaded as a csv file, Cluster Locator displays a schematic visual representation of the way in which the genes on the analyzed list are distributed along the reference genome. For more details on the features and the user's interface see the User's Guide in Supplementary Material.

The results of an illustrative experiment are provided as Supplementary Material (SM 2). In the experiment, the gene clustering of the genes associated with the GO terms 'mitotic nuclear division' and 'proteolysis' were determined in the 5 pre-loaded organisms using Cluster Locator. These results indicate that the clustering of genes associated with specific biological processes could have an evolutionary meaning and provide a hint on the usefulness of the tool.

5 Statistics

Cluster Locator provides two *P-values* after each analysis. One of these *P-values* is obtained performing the Kolmogorv-Smirnov test with the null hypothesis that the genes on the analyzed list are uniformly distributed along each replication unit of the reference genome. This *P-value* is always the same for a given list of genes and does not depend on the max-gap selected to define the clusters. The other *P-value* is associated with the percentage of gene clustering, which in turn depends on the max-gap selected to perform the analysis. This *P-value* is the probability of finding an equal or a more extreme percentage of gene clustering if the genes on the list are uniformly distributed along the genome. To estimate this probability we follow the empiric approach reported by Roy *et al.* (2002). First, Cluster Locator generates 1000 lists of the same size than the analyzed list with genes randomly picked out from the reference genome. Then, using the max-gap selected by the user, Cluster Locator determines the percentage of gene clustering in each of those lists. Next, Cluster Locator checks with the Kolmogorv-Smirnov test if the distribution of the obtained values is normal and if so, determines their empirical mean and standard deviation. With this, Cluster Locator calculates the *P-value* associated with the percentage of gene clustering found in the analyzed list.

Note that even if the null hypothesis that the genes on the analyzed list are uniformly distributed along the reference genome cannot be rejected, it is possible to find a significantly bigger percentage of gene clustering than the one expected by chance. Also note that the different max-gaps provide insights on different mechanisms of gene clustering, i.e. gene tandem duplications, chromosomal rearrangements, etc. We strongly discourage the search of the lowest p-value as a criterion to select the max-gap, as this might imply p-hacking.

6 Conclusions

Cluster Locator allows the non-experts to study how the genes on a list of interest are distributed in clusters and whether the percentage of gene clustering found in the list is statistically significant. Cluster Locator provides an easy, fast and reproducible way to test hypothesis about genome organization in relation to a particular function, gene expression level or quantitative trait.

Funding

This work was supported by the “National Agency of Research and Innovation” (ANII); the “Program for the Development of Basic Sciences” (PEDECIBA); the “Basque Department of Industry, Tourism and Trade” (Etortek and Elkartek Programs), the “Innovation Technology Department of Bizkaia County”, “CIBERehd Network” and “Spanish MINECO” for BFU2014-52282-P, BFU2017-84653-P and the “Severo Ochoa Excellence Accreditation” (SEV-2016-0644).

Conflict of Interest: none declared.

References

Aboukhalil,R. *et al*. (2013) Kerfuffle: a web tool for multi-species gene colocalization analysis. *BMC Bioinformatics*, **14**, 22.

Corrales,M. *et al*. (2017) Clustering of Drosophila housekeeping promoters facili-tates their expression. *Genome Res*., **27**, 1153–1161.

De,S. and Babu,M.M. (2010) Genomic neighbourhood and the regulation of gene expression. *Curr. Opin. Cell Biol*., **22**, 326–333.

Dottorini,T. *et al*. (2013) CluGene: a bioinformatics framework for the identification of co-localized, co-expressed and co-regulated genes aimed at the investigation of transcriptional regulatory networks from high-throughput expression data. *PloS One*, **8**, e66196.

Eisen,M.B. *et al*. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA*, **95**, 14863–14868.

Feuerborn,A. and Cook,P.R. (2015) Why the activity of a gene depends on its neighbors. *Trends Genet. TIG*, **31**, 483–490.

Ghanbarian,A.T. and Hurst,L.D. (2015) Neighboring genes show correlated evolution in gene expression. *Mol. Biol. Evol*., **32**, 1748–1766.

Hurst,L.D. *et al*. (2004) The evolutionary dynamics of eukaryotic gene order. *Nat. Rev. Genet*., **5**, 299–310.

Lee,J.M. and Sonnhammer,E.L.L. (2003) Genomic gene clustering analysis of pathways in eukaryotes. *Genome Res*., **13**, 875–882.

Michalak,P. (2008) Coexpression, coregulation, and cofunctionality of neighboring genes in eukaryotic genomes. *Genomics*, **91**, 243–248.

Reimegård,J. *et al*. (2017) Genome-wide identification of physically clustered genes suggests chromatin-level co-regulation in male reproductive development in Arabidopsis thaliana. *Nucleic Acids Res*., **45**, 3253–3265.

Roy,P.J. *et al*. (2002) Chromosomal clustering of muscle-expressed genes in *Caenorhabditis elegans*. *Nature*, **418**, 975–979.

Thévenin,A. *et al*. (2014) Functional gene groups are concentrated within chromosomes, among chromosomes and in the nuclear space of the human genome. *Nucleic Acids Res*., **42**, 9854–9861.

Tiirikka,T. *et al*. (2014) Clustering of gene ontology terms in genomes. *Gene*, **550**, 155–164.

Yi,G. *et al*. (2007) Identifying clusters of functionally related genes in genomes. *Bioinformatics*, **23**, 1053–1060.