Bioinformatics Advance Access published June 2, 2015

Gene Expression

A web application for the unspecific detection of differentially expressed DNA regions in strand specific expression data

José M. Juanes[1$], Ana Miguel[2,3$], Lucas J. Morales[2], José E. Pérez-Ortín[2,3],Vicente Arnau[1*]

[1]Departamento de Informática, Escola Tècnica Superior d'Enginyeria, Universitat de València, Burjassot, Spain.

[2]Departamento de Bioquímica y Biología Molecular, Facultad de Biología, Universitat de València, Burjassot, Spain.

[3]E.R.I. Biotecmed, Universitat de València, Burjassot, Spain.

Associate Editor: Dr. Ziv Bar-Joseph

ABSTRACT

Genomic technologies allow laboratories to produce large-scale data sets, either through the use of next-generation sequencing or microarray platforms. To explore these data sets and obtain maximum value from the data, researchers view their results alongside all the known features of a given reference genome. To study transcriptional changes that occur under a given condition, researchers search for regions of the genome that are differentially expressed between different experimental conditions. In order to identify these regions several algorithms have been developed over the years, along with some bioinformatic platforms that enable their use. However, currently available applications for comparative microarray analysis exclusively focus on changes in gene expression within known transcribed regions of predicted protein-coding genes, the changes that occur in non-predictable genetic elements, such as non-coding RNAs. Here, we present a web application for the visualization of strand-specific tiling microarray or next-generation sequencing data that allows customized detection of differentially expressed regions all along the genome in an unspecific manner, that allows identification of all RNA sequences, predictable or not.

Availability: The web application is freely accessible at http://tilingscan.uv.es/

Implementation: TilingScan is implemented in PHP and JavaScript.

Contact: vicente.arnau@uv.es

Supplementary material available at: xxxxxxx

1. INTRODUCTION

Eukaryotic genomes are composed by coding and non-coding DNA sequences. Until recently it was thought that most coding regions had been identified, as they encode either protein-coding genes, highly conserved ribosomal RNA, transference RNA or other minor non-coding RNA (ncRNA) species which open reading frames (ORF) are relatively easy to find (Brent, 2007). Only 5' and 3' extended transcribed regions outside the ORF and some introns were not easily predictable (Machado-Lima, et al., 2008; Bahrami-Samani *et al*., 2014). During the last years, however, evidence from different sources has made clear that most of the assumed "non-coding" DNA was, in fact, encoding non-predictable RNA molecules. These non-coding sequences have proven not only to be expressed, but also to play many different regulatory roles within the cell (Mattick and Makunin, 2006).

Current genome-wide methods, such as tiling microarrays and next generation sequencing (NGS) provide platforms for the study of transcriptomes. In most cases, transcriptomic analyses are performed to compare relative RNA abundance in different samples, such as wild type versus mutant or before versus after stress induction (Molina-Navarro, *et al*., 2008; see Pérez-Ortín, *et al*., 2012 for a review). Most current algorithms and bioinformatic applications for comparative analysis in tiling arrays or NGS exclusively search for expression changes within known transcribed regions of predicted genetic elements (see Suarez*et al*., 2009 and Bahrami-Samani *et al*., 2014, for reviews). Changes in transcript length, existence of unpredicted introns, antisense transcripts, or other unannotated ncRNAs are not detectable using these algorithms. To meet this need, we have developed TilingScan, a new web application for the visualization of microarray data that has been implemented with a new search algorithm for the detection of differentially expressed regions in an unsupervised manner. For this detection, we have developed a version of the geometric moving average algorithm, which corresponds to the class of methods of stationary random processes and constitutes one of the several methods developed for the detection of abrupt changes in a signal (Basseville and Nikiforof, 1993). This algorithm scans the signal based in scalable and sliding windows and searches for over- or under-expressed regions of a minimum size of nucleotides flanked by neutral regions (see supplementary material for detailed description). Rather than focusing on changes in protein-coding genes exclusively, the application detects changes that occur all along the genome, including unannotated ncRNA, and then assigns them to the specific genetic elements annotated, if any. In addition, provided that microarray data signals often display noisy profiles, TilingScan enables the application of a smoothing algorithm that is based on a Gaussian filter, allowing signal noise removal prior to the search and clearer visualization of the data. The application can be used to analyze NGS data provided that BAM files are converted into compatible files using freely available software as described in the tutorial.

[*]To whom correspondence should be addressed. [$]Both authors contributed equally to the paper

2. IMPLEMENTATION OF THE WEBSITE

The website is implemented in PHP and JavaScript. PHP was used for the detection of significant differentially expressed regions and JavaScript was used for interactive visualization of the regions. Depending on the nature of the data set, strand specific or non-strand specific, visualization can be selected. By uploading data sets to the server each user can create a new project that will be saved under a unique identifier provided by the website (Project ID) as well as under a personal identifier selected by the user. This ID allows easy access to every created project, allowing different research groups to collaborate by sharing the specific project ID. The website contains a tutorial and a set of demonstration data for the trial of the application prior to any use. Along with the visualization and the detection of regions, the website has been implemented with a smoothing algorithm for signal noise removal (*Gauss filter*), as well as with a tool for manual annotation of regions of interest (*Annotate!*).

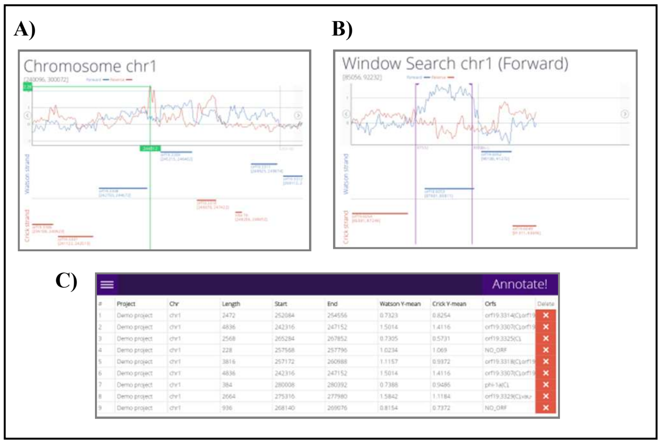
3. STRUCTURE AND USAGE OF THE WEBSITE

3.1. Visualization tool

The application enables the upload of data from tiling array (or data from NGS converted to BAM files) results as genome graphs, which display probe intensity alongside the genomic sequence, aligning it to a reference genome provided by the user. This allows customized visualization of either specific chromosomes or specific genes of interest using the *Visualize chromosome* and *Visualize gene* tools (Figure 1A). If signal smoothing is required, the application of the Gaussian filter can be manually selected prior to visualization, and permuted the desired number of times. If proper visualization of a specific region is required, a region of choice can be delimited and zoomed for X and Y scale adjustment. By delimiting a region manually and selecting the *Annotate* option, specific features of the selected region will be registered (such as chromosome, chromosomal coordinates, length, and mean intensity value on each strand) and listed in a table, which can be edited and downloaded for further use (Figure 1C).

3.2. Analysis tool for the detection of differentially expressed regions

For the detection of differential expression, the application has been implemented with a search algorithm that is based on a sliding window model (See supplementary material for description). Using the search tool (*Window search*) the application will accurately locate and identify differentially expressed regions, displaying both graphical outputs (Figure 1B) and the register of the above-mentioned region features (chromosomal coordinates, length, etc.). If a given region is detected at specific genomic coordinates that correspond to previously annotated genetic elements, the ORF names found within the region will also be recorded. These regions will not only be annotated automatically after the user runs the search analysis, but can also be selected and annotated manually if other regions of interest are found via the *Annotate!* tool.



#	Project	Chr	Length	Start	End	Watson Y-mean	Crick Y-mean	Orfs	Delete
1	Demo project	chr1	2472	252084	254556	0.7323	0.8254	orf19.3314(C),orf19	✕
2	Demo project	chr1	4836	242316	247152	1.5014	1.4116	orf19.3307(C),orf19	✕
3	Demo project	chr1	2568	265284	267852	0.7305	0.5731	orf19.3325(C),	✕
4	Demo project	chr1	228	257568	257796	1.0234	1.069	NO_ORF	✕
5	Demo project	chr1	3816	257172	260988	1.1157	0.9372	orf19.3318(C),orf19	✕
6	Demo project	chr1	4836	242316	247152	1.5014	1.4116	orf19.3307(C),orf19	✕
7	Demo project	chr1	384	280008	280392	0.7388	0.9486	phi-1a(C),	✕
8	Demo project	chr1	2664	275316	277980	1.5842	1.1184	orf19.3329(C),vau	✕
9	Demo project	chr1	936	268140	269076	0.8154	0.7372	NO_ORF	✕

Figure 1.TilingScan graphical outputs.(A) By selecting the chromosome or the gene of interest, signal intensity profiles will be shown, displaying the intensity profile on forward (Watson) strand in blue and the reverse (Crick) strand in red. Start and end chromosomal coordinates of the visualized region are shown in brackets (top, left). Intensity profiles are aligned to the selected reference genome (bottom). Sliding the mouse over the signal profile, X and Y coordinates for the selected data point will be shown (green rectangles on A panel), which correspond to the position in the genome (in bp) and the signal intensity value respectively. **(B)** Graphical output of a region detected using the demo data. Start and end chromosomal coordinates of the detected region are shown on the edges of the purple dashed vertical lines respectively. **(C)** Image section of the *Annotate!* tool displaying recorded features for some of the regions detected in a transcriptomic study performed in yeast under hypoxia.

ACKNOWLEDGEMENTS

Funding: J.E.P-O.is supported by grants from the Spanish MCINN (BFU2013-48643-C3-3-P), and from the Regional Valencian Government (PROMETEO 2011/088).The authors are thankful to Tianlu Li for critical revision of the manuscript.

Conflict of interest: none declared.

REFERENCES

Brent, M.R. (2007) How does eukaryotic gene prediction work?, *Nat. Biotech.* **25**, 883-885.

Bahrami-Samani E *et al.* (2014). Computational challenges, tools, and resources for analyzing co- and post-transcriptional events in high throughput.Wiley Interdiscip Rev RNA. 2014. doi: 10.1002/wrna.1274.

M. Basseville, I.V.Nikiforof. (1993) *Detection of abrupt changes: theory and application*. Prentice Hall, Englewood Clifss, NJ, U.S.A.

Machado-Lima, A., del Portillo, H.A. and Durham, A.M. (2008) Computational methods in noncoding RNA research, *J.Math.Biol.* **56**, 15-49.

Mattick, J.S. and Makunin, I.V. (2006) Non-coding RNA, *Hum. Mol. Genet.* **15**, R17-29.

Molina-Navarro, M.M., *et al.*(2008) Comprehensive transcriptional analysis of the oxidative response in yeast, *J. Biol.Chem.* **283**, 17908-17918.

Pérez-Ortín, J.E., de Miguel-Jiménez, L. and Chávez, S. (2012) Genome-wide studies of mRNA synthesis and degradation in eukaryotes, *Biochimi. Biophys. Acta* **1819**, 604-615.

Suárez, E., Burguete, A. and McLachlan, G.J. (2009) Microarray data analysis for differential expression: a tutorial, *PR Health Sci. J.* **28**, 89-104.