BIOINFORMATICS APPLICATIONS NOTE Vol. 24 no. 24 2008, pages 2918–2920
doi:10.1093/bioinformatics/btn542

Genome analysis

CARPET: a web-based package for the analysis of ChIP-chip and expression tiling data

Matteo Cesaroni[1,*], Davide Cittaro[2], Alessandro Brozzi[1], Pier Giuseppe Pelicci[1] and Lucilla Luzi[3,*]

[1]Department of Experimental Oncology, European Institute of Oncology, Via Ripamonti 435, 20141 Milano, [2]Cogentech, Consortium for Genomic Technologies, Via Adamello 16, 20139 Milano and [3]IFOM, FIRC Institute of Molecular Oncology Foundation, Via Adamello 16, 20139 Milano, Italy

Received on April 11, 2008; revised on August 29, 2008; accepted on October 17, 2008

Advance Access publication October 21, 2008

Associate Editor: Alfonso Valencia

ABSTRACT

Summary: CARPET (collection of automated routine programs for easy tiling) is a set of Perl, Python and R scripts, integrated on the Galaxy2 web-based platform, for the analysis of ChIP-chip and expression tiling data, both for standard and custom chip designs. CARPET allows rapid experimental data entry, simple quality control, normalization, easy identification and annotation of enriched ChIP-chip regions, detection of the absolute or relative transcriptional status of genes assessed by expression tiling experiments and, more importantly, it allows the integration of ChIP-chip and expression data. Results can be visualized instantly in a genomic context within the UCSC genome browser as graph-based custom tracks through Galaxy2. All generated and uploaded data can be stored within sessions and are easily shared with other users.

Availability: http://bio.ifom-ieo-campus.it/galaxy

Contacts: matteo.cesaroni@ifom-ieo-campus.it; lucilla.luzi@ifom-ieo-campus.it

1 INTRODUCTION

With the introduction of tiling array technology, it has become feasible for scientists to interrogate entire genomes. Such high-resolution DNA microarrays represent a universal framework exploitable through several diverse experimental approaches to extract the full significance of whole-genome data (Mockler *et al*., 2005). Indeed, tiling arrays have been widely employed to detect genome-wide protein–DNA interactions through chip hybridization of chromatin immunoprecipitation experiments (ChIP-chip) (ENCODE_Project_Consortium, 2007; Guenther *et al*., 2007; Wei *et al*., 2006) and, more recently, to perform expression studies, confirming the potential of these techniques for the characterization of the whole transcriptome (Kapranov *et al*., 2007). The integration of ChIP-chip and expression profiling data will be essential to decode the genetic and epigenetic networks that interlink DNA-binding protein regulators, transcriptional events and chromatin state, both in physiological and pathological conditions.

To support the analysis of such experimental approaches we have developed CARPET, a compilation of scripts integrated within the Galaxy2 platform (Blankenberg *et al*., 2007), which helps biologists to analyze ChIP-chip and expression tiling data independently and, if necessary, to merge the results for a more comprehensive understanding of the significance of the experimental data.

Different scripts handle the various steps along the proposed analysis procedure. Performed operations, maintained in sessions in Galaxy2, are stored on the server for 1 month and they can easily be shared with other users in much the same way as all the raw data and results.

Integration of the Galaxy2 platform with the UCSC Genome Browser permits users to visualize their ChIP or gene expression results, as custom tracks, in a genomic context.

2 GUIDELINE ANALYSIS STEPS

GFF or 'Pair files' from Nimblegen experiments are the natural input data for the package, but, upon proper reformatting, records coming from other platforms can be analyzed as well.

Although each CARPET script can be used as a stand-alone application, we propose a rational pipeline that can be applied to separately analyze ChIP-chip and expression tiling profile experiments and, finally, to merge them in order to extract additional biological and functional meaning from the results. Detailed instructions and a manual for using CARPET can be found on the web site.

2.1 Analysis of ChIP-chip experiments

2.1.1 Chip image visualization: ChipView Once a 'Pair file' is uploaded, it is possible to create and visualize an image simulating the chip surface; inspection of the distribution of signal over the chip is important to determine the presence of artefacts or hybridization problems (Fig. 1A).

2.1.2 Preprocess for tiling: PPT PPT allows bi-weighted or quantile normalization of both ChIP-chip and expression tiling data. It can be applied potentially to any kind of platform experiments: it works with any number of replicates and any custom table-like file that contains at least Seq_IDs, start positions of the probes and a probe signal [log2(ratio) or paired Cy5-Cy3 raw values or single Cy5-Cy3 raw values]. PPT also calculates and compares correlations between replicates producing different plots taking help

*To whom correspondence should be addressed.

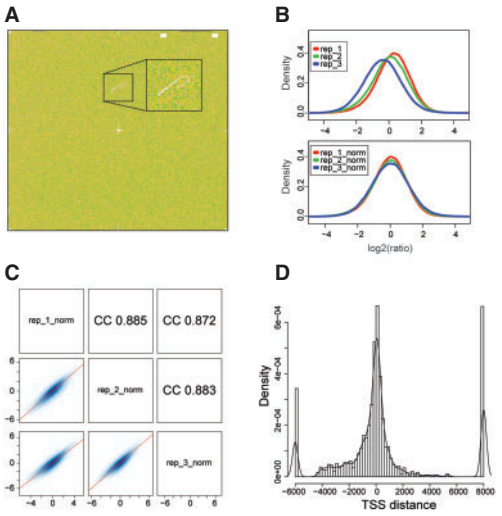

Fig. 1. Output examples of some CARPET applications. (**A**) ChipView simulated image of chip surface; (**B**) distribution of log2(ratio) of three ChIP-chip replicates before (upper panel) and after (lower panel) normalization; (**C**) graphs of the correlation between three replicates and (**D**) distribution of peaks around TSS generated by the GIN visualizer based on data coming from the GIN output.

from the Ringo Package (Toedling *et al*., 2007) (Fig. 1B and C). If the 'summarization' option is selected, the GFF file generated is suitable for the following peak identification step, directly through the PeakPicker tool.

2.1.3 Transformation from GFF to wiggle format: GFF2WIG To visualize the raw intensity data of each probe within the UCSC Genome Browser, GFF2WIG transforms GFF files into wiggle format (WIG). These can immediately be uploaded as a custom track in the Genome Browser through the corresponding link in the history frame.

2.1.4 Extraction of tiling-enriched regions: PeakPicker PeakPicker is a Perl script that is able to identify enriched regions (peaks) from a ChIP-chip experiment. A 'peak' is defined as the region where multiple probes, with a log2(ratio) greater than a user-defined threshold, are located genomically close to one another. PeakPicker allows the user to decide the minimal number of probes that must exceed the defined threshold, as well as the maximum distance allowed, in order to consider two probes as contiguous. Thereafter, a scoring analysis, or a scoring-plus statistical analysis, can be performed; in the latter case, both a peak score and a *P*-value are calculated. Statistical analyses are performed essentially as previously established (Scacheri *et al*., 2006). Thresholds on *P*-values can also be set, and neighbor-enriched regions can be joined together. The output of the analysis is a GFF file of the peak's coordinates and their corresponding score/*P*-value. Results can be immediately redirected to Genome Browser as a custom track for visualization.

2.1.5 Interval annotation: GIN The GIN annotator uses two files: a GFF file with genomic intervals and any user-preferred transcript annotation tables (e.g. RefSeq, UCSC genes) that can be easily downloaded from the UCSC Genome Browser database. GIN associates genomic interval queries with the matching interrogated transcripts. The output, for each interval, includes the name and absolute chromosome coordinates of the assigned transcriptional units, as well as a call describing its relative position with respect to the transcribed unit (e.g. first exon, fourth intron, promoter) and the relative distance from the putative TSS (transcription start site). Intervals that do not intersect any gene loci are annotated as 'intergenic'.

2.1.6 Distribution analysis of peaks around the TSS: GIN visualizer Knowledge of the location of binding elements with respect to the TSS of the associated transcriptional units often helps in the characterization of peculiar traits of regulatory DNA-binding proteins tested by ChIP. For this purpose, the GIN output file can be directly submitted to the visualizer to portray the distribution of peak-intervals around the TSS (Fig. 1D).

2.1.7 Comparison of two ChIP-chip datasets: Com&Uni Several binding factors or histone modifications are often ChIPed within the same experimental framework; therefore cross-comparison of experiments is a significant problem that needs to be addressed. Com&Uni extracts common or unique regions from two GFF files that were generated with PeakPicker. The program also permits users to add a choice of flanking regions to the original coordinates.

2.2 Analysis of expression tiling experiments

2.2.1 Expression chip annotation: ENo (Expression notator) The first crucial step in expression tiling analysis is the allocation of probes to their corresponding gene locus: the ENo script annotates each probe of the entire chip using user-defined transcript annotation tables (e.g. RefSeq, UCSC genes) downloaded from the UCSC Genome Browser database.

2.2.2 Tiling expression analyzer: TEA Tiling expression analyzer (TEA) performs two different tasks, depending on the number of different experiments uploaded. For simple expression estimation, starting from the annotation file created with ENo, TEA calculates a value of expression based on the mean and/or the median of all probe signals associated with the exons of a particular transcriptional unit. In comparison experiments, the signal distribution of a gene in two different conditions (e.g. untreated versus treated) is analyzed, and fold-change and *P*-values are calculated. The user may also choose to operate a FDR correction (Benjamini *et al*., 1995). Many filters can be also applied, e.g. on the raw signal, fold-change and *P*-value.

2.2.3 Intersecting expression and ChIP-chip results: binding-expression correlation (BEC) Merging expression and ChIP-chip results permits researchers greater confidence when they formulate hypotheses regarding, for example, the mechanistic implications of a transcription factor (TF) that binds near, or within, putative target gene loci that are also transcriptionally regulated. Outputs of ChIP-chip results from PeakPicker and expression data from TEA can be rapidly compared using binding-expression correlation. For each gene, the program returns the number of peaks that match the strict transcriptional unit or the user-defined putative promoter region around the TSS.

Correlations between gene regulation or expression and TF binding can therefore immediately be evaluated.

3 OUTLOOK

CARPET provides a very powerful, user-friendly and comprehensive set of tools for ChIP-chip and expression tiling analysis, beginning from raw data and leading to data mining and visualization. Its main assets are that (i) it provides a collection of coordinated programs directly accessible through the web on our site; (ii) no knowledge is needed of programming languages, such as R or Perl, (iii) the integration of CARPET with the Galaxy2 environment makes data storage and sharing very easy and allows the direct graphical visualization of results as custom tracks in the UCSC Genome Browser and (iv) it facilitates the comparison of results obtained through different experimental approaches.

ACKNOWLEDGEMENTS

The authors are very grateful to the Galaxy2 team for conceiving, and maintaining the platform. We thank Lara Lusa for statistical suggestions, Alessandro Gardini, Simone Minardi and Lelio Lassandro for useful discussions and beta-testing of CARPET, Pascale Romano and Rosalind Gunby for proofreading the article, Myriam Alcalay, Gaetano Dellino and James Reid for critical reading of article.

Funding: Italian Association for Oncology Research (OGCG1139 to P.G.P.).

Conflict of Interest: none declared.

REFERENCES

Benjamini,Y. *et al*. (1995) Controlling the false discovery rate - a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B Meth.*, **57**, 289–300.

Blankenberg,D. *et al*. (2007) A framework for collaborative analysis of ENCODE data: making large-scale analyses biologist-friendly. *Genome Res.*, **17**, 960–964.

ENCODE_Project_Consortium. (2007) The ENCODE pilot project. *Genome Res.*, **17**, 667–964.

Guenther,M.G. *et al*. (2007) A chromatin landmark and transcription initiation at most promoters in human cells. *Cell*, **130**, 77–88.

Kapranov,P. *et al*. (2007) Genome-wide transcription and the implications for genomic organization. *Nat. Rev. Genet.*, **8**, 413–423.

Mockler,T.C. *et al*. (2005) Applications of DNA tiling arrays for whole-genome analysis. *Genomics*, **85**, 1–15.

Scacheri,P.C. *et al*. (2006) Statistics for ChIP-chip and DNase hypersensitivity experiments on NimbleGen arrays. *Methods Enzymol.*, **411**, 270–282.

Wei,C.L. *et al*. (2006) A global map of p53 transcription-factor binding sites in the human genome. *Cell*, **124**, 207–219.

Toedling,J. *et al*. (2007) Ringo–an R/Bioconductor package for analyzing ChIP-chip readouts. *BMC Bioinformatics*, **8**, 221.