

Application Note Gene Expression Section

CoCo: A web application to display, store and curate ChIP-on-chip data integrated with diverse types of gene expression data.

Charles Girardot^{1,*}, Oleg Sklyar², Sophie Grosz¹, Wolfgang Huber² and Eileen E. M. Furlong¹

¹ European Molecular Biology Laboratory, D-69117 Heidelberg, Germany

² European Bioinformatics Institute, European Molecular Biology Laboratory, Cambridge CB10 1SD, United Kingdom

Associate Editor: John Quackenbush

ABSTRACT

Motivation: CoCo, ChIP-On-Chip Online, is an open-source web application that supports the annotation and curation of regulatory regions and associated target genes discovered in ChIP-on-chip experiments. CoCo integrates ChIP-on-chip results with diverse types of gene expression data (expression profiling, *in situ* hybridisation) and visualises them within a genomic context. Regulatory relationships between the transcription factor-bound regions and putative target genes can be stored and expanded throughout different sessions.

Availability: <http://furlonglab.embl.de/methods/tools/coco>

Contact: charles.girardot@embl.de

1 INTRODUCTION

Chromatin Immunoprecipitation followed by microarray analysis (ChIP-on-chip) is a very powerful *in vivo* method to systematically identify regulatory regions bound by a transcription factor (TF) at a genome-wide level (Ren et al., 2000; Sandmann et al., 2006).

Once significantly enriched regions have been extracted from ChIP-on-chip data, further evaluation is essential to determine the genomic landscape surrounding each putative regulatory region, assign the binding event to a putative target gene and assess the regulatory potential on the target gene's expression. In higher eukaryotes, enhancer regions can act at large distances from their target genes, including within introns of neighbouring loci or 3' to the regulated gene (Nobrega et al., 2003). Thus, assuming that a TF-bound region is regulating the closest gene will often select the wrong target gene, especially in gene-dense regions. This initial step of linking a TF-bound region to a correct target gene is fundamental for inferring regulatory relationships and subsequent network analysis, but yet has been largely ignored.

We present CoCo, ChIP-on-Chip online, a web-based tool dedicated to ChIP-on-chip data visualisation, analysis and knowledge storage. CoCo integrates diverse types of meta-data, including the genomic context around the TF-bound regions, *in situ* hybridisation data indicating the tissues where neighbouring genes are expressed, and expression profiling data indicating the response of surrounding genes to different perturbations.

Genome browsers such as GBrowse (Stein et al., 2002), UCSC (Kent et al., 2002) or Ensembl (Stalker et al., 2004) can be used to upload and visualise ChIP-on-chip datasets as custom annotation tracks, and in some simple use cases they provide an alternative for the task at hand. CoCo is a specialised tool, designed for more

complex analysis projects. It allows the dynamic evaluation of multiple cut-offs, offers more specialised and sophisticated representations of expression data (*in situ* patterns and microarray profiles) and is able to represent time courses of ChIP-on-chip data together with the expression data and major genomic features. These facilities are essential for the assignment of target genes to putative regulatory regions. The discovered regions and target gene assignments can be saved, edited and shared between collaborators, allowing for curation over time and experiments.

2 CORE FEATURES

2.1 Uploading files and creating a configuration

A *configuration* contains all datasets that will be visualised together. These include ChIP-on-chip datasets (including *mock* datasets i.e. ChIP-on-chip data using pre-immune serum), microarray expression profiling data, *in situ* hybridisation patterns and genome annotations. To accommodate temporal and spatial information, a configuration is given both a developmental stage term list and an anatomy term list. Finally, "sticky" feature lists (e.g. microarray features of genomic regions containing repetitive sequences) can be defined. As all uploaded data files are stored, they can be readily shared.

2.2 Data Presentation and Browsing

Once a configuration is created, data is visualised using interactive images that are browsed in the fashion of a genomic browser.

The *overview page* shows the distribution of TF-bound regions along each chromosome, allowing for a quick identification of positional bias. In addition, a table summarizes all TF-bound regions together with their enrichment values across all the ChIP-on-chip datasets loaded in the configuration.

The genomic region to be displayed can be specified by searching for a gene, a microarray feature, a genomic position or by following one of the various links offered on the overview page.

Within a *genomic region view* (see Figure 1) all data are visualised together. In the central part of the image (Fig. 1E, ChIP-on-chip zone) each array feature is plotted as a red or grey rectangular bar depending on whether its enrichment value is above or below the user-defined threshold, respectively. When multiple ChIP-on-chip experiments are integrated into a configuration, the bars representing the tiling array features appear as stacked bars. This allows for visualisation of combinatorial binding when experiments from different TFs are used in one configuration, and of temporal enhancer occupancy when a time-series for one TF is used. Sticky

*To whom correspondence should be addressed.

features appear in dark grey. The sense and antisense genomic strands together with genes are shown above and below the central ChIP-on-chip zone, respectively (Fig. 1E). Genes are colour-coded according to available *in situ* patterns reflecting whether genes are expressed at any of the stages and/or anatomy specified in the con-

figuration. Finally, the upper and the lower regions of the image (Fig. 1E, gene expression zones) display gene expression values integrated into the configuration. Each expression dataset has its own track, where values are displayed as colour-coded rectangles aligned with the corresponding genes.



Fig 1. Target Gene Assignment Procedure in CoCo. The Query Toolbox (A) and the Navigation Control Panel (B) allow users to easily zoom in/out and locate genes. Thresholds are positioned in the Display Option Panel (C) and regulatory regions overlapping with the genomic region view on display can be accessed using the hyperlink in panel D. As an example, the genomic region view (E) shows a ChIP-chip time series using the *Drosophila melanogaster* Mef2 transcription factor in the developing embryo (Sandmann et al., 2006) and illustrates the difficulty of correct target gene assignment. Five ChIP-on-chip time points are visualised together with five *Mef2* loss-of-function expression profiling experiments, BGDP and unpublished *in situ* data. This view shows two distinct Mef2-bound regions and suggests *nautilus* (red arrow) as the target gene for both. This conclusion is strengthened by the fact that *nautilus* is expressed in the same cells as Mef2 at the developmental stages under study (indicated by the orange border of *nautilus*) and a reduced expression in *Mef2* loss-of-function expression profiling experiments (indicated by blue bars in the gene expression zone).

2.3 Defining Regulatory Regions and Target Genes

Putative regulatory regions and associated target genes can be defined and stored in CoCo by a simple click on enriched ChIP-on-chip features and genes in the genomic region view (Fig. 1E). Alternatively, they can also be imported from a simple tab-delimited file. CoCo supports consistent knowledge accumulation in several ways. When creating a regulatory region, CoCo detects overlap with existing regulatory regions and suggests complementing them, instead of creating new ones. The origin of regulatory regions ("experimental" for regions created in CoCo or a user-defined name for imported regions) is tracked and references to ChIP-on-chip results showing the binding of TFs on regulatory regions are maintained. Finally, regulatory regions and target gene assignments are given a confidence value that can be manually modified as knowledge accumulates. Regulatory regions are editable and accessible through a flexible search interface or directly from the genomic region view (Fig. 1D). Details can be found in the user manual.

3 CONCLUSION

CoCo provides dedicated functionality for ChIP-on-chip experiments, allowing for the integration of TF-bound regions with different types of gene expression data. It constitutes a user-friendly platform to build and store regulatory relationships between TF-binding data and the potential target genes.

Although CoCo has been primarily used with *Drosophila melanogaster*, it was developed to accommodate all organisms with annotated genomes. CoCo is optimised for small to medium size tiling arrays. Analysis of data generated by high-density oligonucleotide arrays (Affymetrix, NimbleGen) typically report sets

of statistically enriched genomic regions (i.e. consecutively enriched features). Visualisation of these TF-bound regions in the ChIP-on-chip zone (Fig. 1E), instead of all individual microarray features, is recommended and will soon be available in CoCo.

CoCo is a JAVA application and uses the gff3Plotter R-package of Bioconductor (Gentleman et al., 2004) to generate pictures.

ACKNOWLEDGEMENTS

We are grateful to all Furlong lab members for useful suggestions. We thank Julien Gagneur for his help and advice and Christian Boulin for constant support.

REFERENCES

- Gentleman R.C. et al. (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **5**(10):R80.
- Kent, W.J. et al. (2002) The Human Genome Browser at UCSC. *Genome Res.* **12**(6), 996-1006.
- Nobrega, M. A. et al. (2003) Scanning human gene deserts for long-range enhancers. *Science* **302**, 413.
- Ren, B. et al. (2000) Genome-wide location and function of DNA binding proteins. *Science* **290**(5500), 2306-9.
- Sandmann, T. et al. (2006) A temporal map of transcription factor activity: mef2 directly regulates target genes at all stages of muscle development. *Dev Cell* **10**, 797-807.
- Stalker, J. et al. (2004) The Ensembl Web site: mechanics of a genome browser. *Genome Res* **14**, 951-955.
- Stein, L. D. et al. (2002) The generic genome browser: a building block for a model organism system database. *Genome Res* **12**, 1599-1610.