BIOINFORMATICS APPLICATIONS NOTE Vol. 23 no. 6 2007, pages 771–773
doi:10.1093/bioinformatics/btl641

*Gene expression*

# 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 Sklyar[2], Sophie Grosz[1], Wolfgang Huber[2] and Eileen E. M. Furlong[1]

[1]European Molecular Biology Laboratory, D-69117 Heidelberg, Germany and [2]European Bioinformatics Institute, European Molecular Biology Laboratory, Cambridge CB10 1SD, UK

Received on July 27, 2006; revised on November 17, 2006; accepted on December 15, 2006

Advance Access publication January 17, 2007

Associate Editor: John Quackenbush

## ABSTRACT

**Motivation:** CoCo, *C*hIP-*o*n-*C*hip *o*nline, 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* hybridization) and displays 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, as well as 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, *C*hIP-*o*n-*C*hip *o*nline, a web-based tool dedicated to ChIP-on-chip data visualization, analysis and knowledge storage. To support target gene assignment, CoCo integrates diverse types of meta-data, including the genomic context around the TF-bound regions, *in situ* hybridization data indicating the tissues where neighbouring genes are expressed, and expression profiling data indicating the response of surrounding genes to different perturbations.

As a specialized tool, CoCo implements several features adapted to ChIP-on-chip data visualization and provides key requirements for the management of complex analysis projects. To visualize large numbers of experiments simultaneously, the CoCo display is designed to pack essential features together, rather than stacking data in sequential tracks as in the generic data displays offered by genome browsers such as GBrowse (Stein *et al.*, 2002), UCSC (Kent *et al.*, 2002) or Ensembl (Stalker *et al.*, 2004). This compact visualization and interactive features, like the dynamic evaluation of cut-off selection, facilitate data interpretation and target gene assignment. Importantly, to assist continuity in data analysis, CoCo provides user file access management and a database for storing discovered regions and target gene assignments. This information can be later 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 visualized 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* patterns and genome annotations. To highlight relevant *in situ* data, the spatio-temporal focus of a configuration is defined by both a developmental stage controlled vocabulary (CV) list and an anatomy CV list. Finally, *sticky* feature lists (e.g. microarray features of genomic regions containing repetitive sequences) can be specified. As all uploaded data files are stored, they can be readily shared.

[*]To whom correspondence should be addressed.

## 2.2 Data presentation and browsing

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

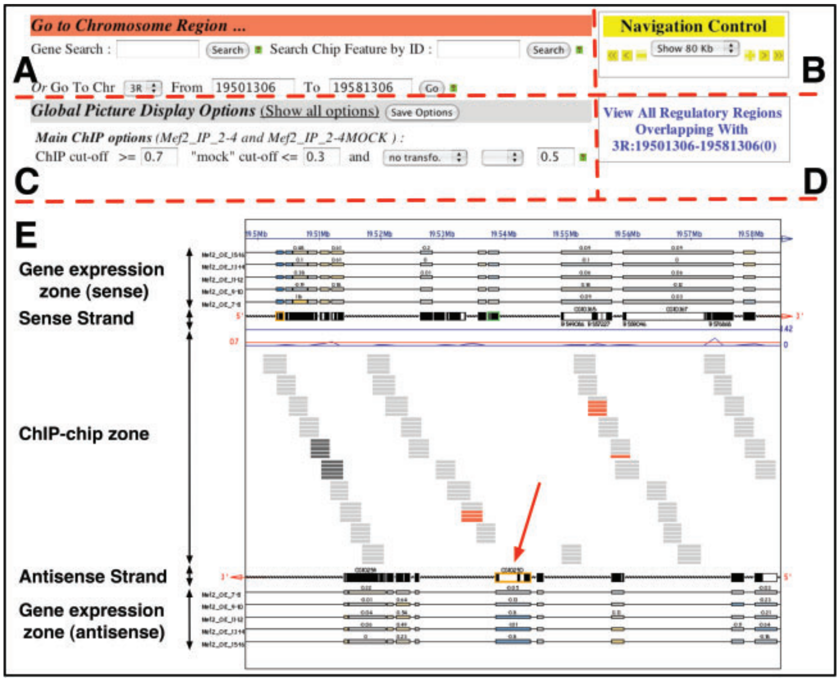
The *overview page* shows the distribution of TF-bound regions along each chromosome, allowing for a quick identification of positional bias. In addition, a table summarizes all TF-bound regions together with their enrichment values across all the ChIP-on-chip datasets loaded in the configuration.

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 Fig. 1) all data are visualized 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 visualization 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 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 configuration. Finally, the upper and the lower regions of the image (Fig. 1E, gene expression zones) display gene expression values as colour-coded rectangles aligned with the corresponding genes.

## 2.3 Defining regulatory regions and target genes

Putative regulatory regions and associated target genes can be defined and stored in CoCo database by a simple click on enriched ChIP-on-chip features and genes in the genomic region view (Fig. 1E). They can also be imported from tab-delimited files. CoCo supports consistent knowledge accumulation in several ways. When creating a regulatory region, CoCo detects overlap with existing regulatory regions and suggests complementing them, rather than creating a new one. 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.



Go to Chromosome Region ...

Gene Search : [ ] Search Search Chip Feature by ID : [ ] Search

Or Go To Chr 3R From 19501306 To 19581306 Go

Global Picture Display Options (Show all options) Save Options

Main ChIP options (Mef2_IP_2-4 and Mef2_IP_2-4MOCK ) :

ChIP cut-off >= 0.7 "mock" cut-off <= 0.3 and no transfo. 0.5

Navigation Control

Show 80 Kb

View All Regulatory Regions Overlapping With 3R:19501306-19581306(0)

**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 link in panel **D**. The genomic region view (**E**) here shows a dMef2 ChIP-chip time series (Sandmann *et al.*, 2006) and illustrates the difficulty of correct target gene assignment. Five ChIP-on-chip time points are visualized together with five *Mef2* loss-of-function expression profiling experiments, BGDP and unpublished *in situ* data. Regions are described enriched if their enrichment (in log) is ≥0.7 and if associated mock enrichment is ≤0.3 (**C**). This view shows two distinct bound regions and suggests *nautilus* (red arrow) as the target gene for both. This conclusion is strengthened by the fact that (1) *nautilus* is expressed in the same cells as dMef2 at the developmental stages under study (indicated by the orange border of *nautilus*, deduced from *in situ* data) and (2) has a reduced expression in *Mef2* loss-of-function experiments (indicated by blue bars in the gene expression zone).

Finally, regulatory regions and gene assignments are given a confidence value that can be modified as knowledge accumulates. Regulatory regions are accessible through a flexible search interface or directly from the genomic region view (Fig. 1D). More details are available in the user manual.

## 3 CONCLUSION

CoCo provides dedicated functionality for the analysis of 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 discover and store regulatory relationships between TF-binding data and the potential target genes.

CoCo offers substantial advantages compared to the use of annotation tracks in generic genome browsers. First, CoCo allows the dynamic evaluation of cut-off selection and readily integrates *mock* values to evaluate feature enrichments (see filtering conditions in Fig. 1C). It also integrates *sticky* feature lists. In contrast, genome browsers typically use scores to colour features in shades and no filtering thresholds can be dynamically specified. Second, CoCo greatly facilitates *in situ* pattern integration using CVs for both stage and tissue. Third, CoCo automatically maps data extracted from user files to the genome version used by the tiling array in the configuration. This eliminates tedious work for users to update custom annotations when new genome versions are released. Finally, all data files as well as CV and *sticky* feature lists are stored on the server and can be shared between users, thereby making this information effortless to include in subsequent configurations. The discovered regions and target gene assignments can be saved, edited and shared. As visualization tools, genome browsers do not offer such possibilities.

Although CoCo has been primarily used with *Drosophila melanogaster*, it was developed to accommodate all organisms with annotated genomes. CoCo is optimized 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). Visualization 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 an open-source JAVA web 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. This work was supported by EMBL, and in part by the EU FP6 network MyoRes.

*Conflict of Interest*: none declared.

## REFERENCES

Gentleman,R.C. *et al*. (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*., **5**, R80.

Kent,W.J. *et al*. (2002) The human genome browser at UCSC. *Genome Res*., **12**, 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**, 2306–2309.

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.