SHORT ARTICLE

Transcription 3:5, 255-259; September/October 2012; © 2012 Landes Bioscience

D-peaks:
A visual tool to display ChIP-seq peaks along the genome

Sylvain Brohée* and Gianluca Bontempi

Machine Learning Group; Computer Science Department; Faculté des Sciences; Université Libre de Bruxelles; Brussels, Belgium

Keywords: ChIPseq, transcription factor, bioinformatics, high-throughput sequencing, visualization

ChIP-sequencing is a method of choice to localize the positions of protein binding sites on DNA on a whole genomic scale. The deciphering of the sequencing data produced by this novel technique is challenging and it is achieved by their rigorous interpretation using dedicated tools and adapted visualization programs. Here, we present a bioinformatics tool (D-peaks) that adds several possibilities (including user-friendliness, high-quality, relative position with respect to the genomic features) to the well-known visualization browsers or databases already existing. D-peaks is directly available through its web interface http://rsat.ulb.ac.be/dpeaks/ as well as a command line tool.

These very last years, researchers have been challenged by the development of novel techniques derived from high-throughput sequencing (e.g., genome, RNA or exome sequencing and epigenetic sequencing approaches). Among those, a very popular approach is ChIP-sequencing (ChIP-seq), which is currently widely used to analyze protein interactions (e.g., transcription factors and chromatin modifying enzymes) with DNA. ChIP-seq replaces now ChIP-chip as the method of choice allowing the exhaustive discovery of precise global DNA binding sites for a protein of interest.[1]

Briefly, ChIP-seq consists in chemically cross-linking DNA to proteins (among which is the protein of interest) to DNA with a chemical agent, then fragmenting the DNA into pieces of about 50 to 500 bp. The DNA pieces linked to the protein of interest are then immunoprecipitated using an antibody directed against this protein. Finally, the DNA pieces, enriched in the binding sites of the protein of interest, are sequenced (**Fig. 1**).

The following step is performed in silico. Indeed, it consists in identifying the portions of the genome that are enriched in short sequenced fragments (short reads), which are potential binding sites for the studied proteins. These regions are named peaks as they correspond to areas of the genome that are highly covered by the sequencing. This step, called peak-calling, is a big challenge in current bioinformatics as illustrated by the impressive number of tools dedicated to this task (for reviews see refs. 2 and 3). These tools not only specify where the peaks are located but also generally export the results by attributing a score to each position of the genome. This score generally corresponds to the number of sequenced reads (i.e., sequenced reads enrichment level) of a given genomic position. When high scores are plotted vs. the genomic position they can appear as peaks (**Fig. 3**).

Plotting the ChIP-seq peaks can be achieved using numerous bioinformatics libraries, tools like Seqminer[4] or browsers such as EnsEMBL, UCSC or Igv.[5-7] However, even if the possibilities offered by these well-established tools are numerous and impressive (mainly for their exhaustivity and flexibility), limitations do appear when the displayed figures are exported. Those figures are sometimes not very aesthetic and may not fulfill the classical criteria for a publication or a presentation. Moreover, the software mentioned above are generally so powerful and complete that it may seem difficult for the non-expert user to find the exact combination of options and manipulations that should be chosen to get a simple image of his/her ChIP-seq results that is comparable to the standards of the field. Finally, to our knowledge, none of these tools allow the user to display the chromosome coordinates from a given point (i.e., relative coordinates) allowing the user to view the distance between a peak and a feature of interest (e.g., transcription start or end site). Aware of these limitations, we developed D-peaks (draw-peaks), an user-friendly tool (with a few simple options) able to render several tracks of continuous values along the genes and the genome in high quality pictures and to specify coordinates relative to any position on the chromosome (**Fig. 2**). Its main advantages over the popular genome browsers mentioned above are its user-friendliness, the titles and labels easy customization, the possibility to display genomic relative coordinates and the high quality of the resulting figures.

As mentioned above, together with the position of potential binding sites, continuous values scoring files (WIGGLE or WIG files) are generally generated by the peak-calling programs. Up to five (compressed or not) WIG files of total size smaller than 200 Mb (no restriction with the command line tool) can easily be submitted to D-peaks. As it is the case for some online browsers, the files remain stored on the server and thus must not be re-uploaded for each picture using the same data. Some options, such as the absolute (and the relative) genomic position, the DNA strand and some other aesthetic possibilities (labels and

*Correspondence to: Sylvain Brohée; Email: sbrohee@ulb.ac.be
Submitted: 08/28/12; Revised: 10/02/12; Accepted: 10/04/12
http://dx.doi.org/10.4161/trns.22457

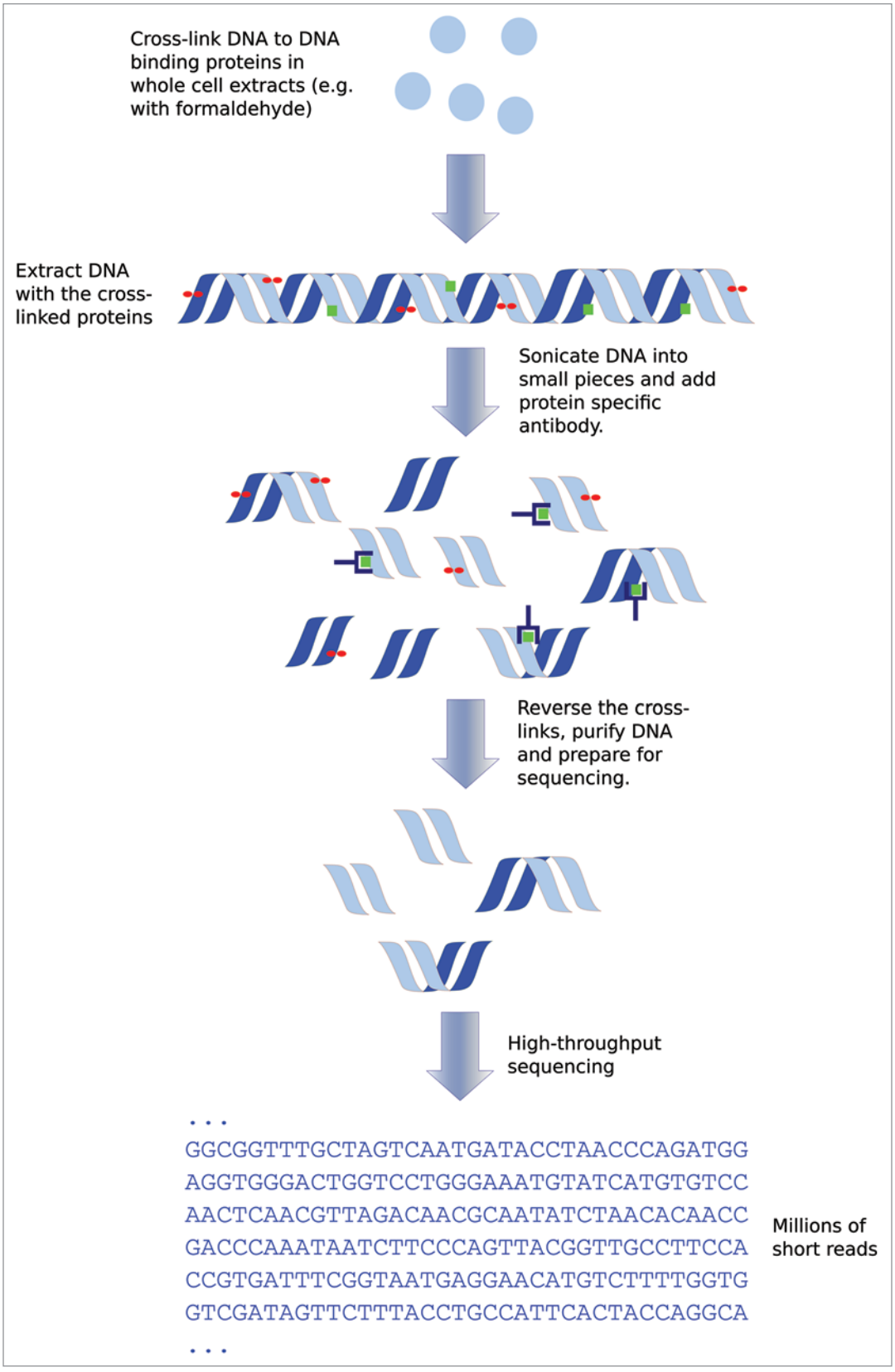

Figure 1. ChIP-seq principle. Classically, DNA and DNA binding proteins are cross-linked with formaldehyde in cell extracts. DNA is then extracted and sonicated into small pieces and these fragments, which are linked to the protein of interest are immunoprecipitated using a specific antibody. The cross-link reaction is then reversed and the precipitated DNA fragments are sequenced. The output of the sequencer consists in millions of short reads that should correspond to the DNA binding sites of the protein of interest.

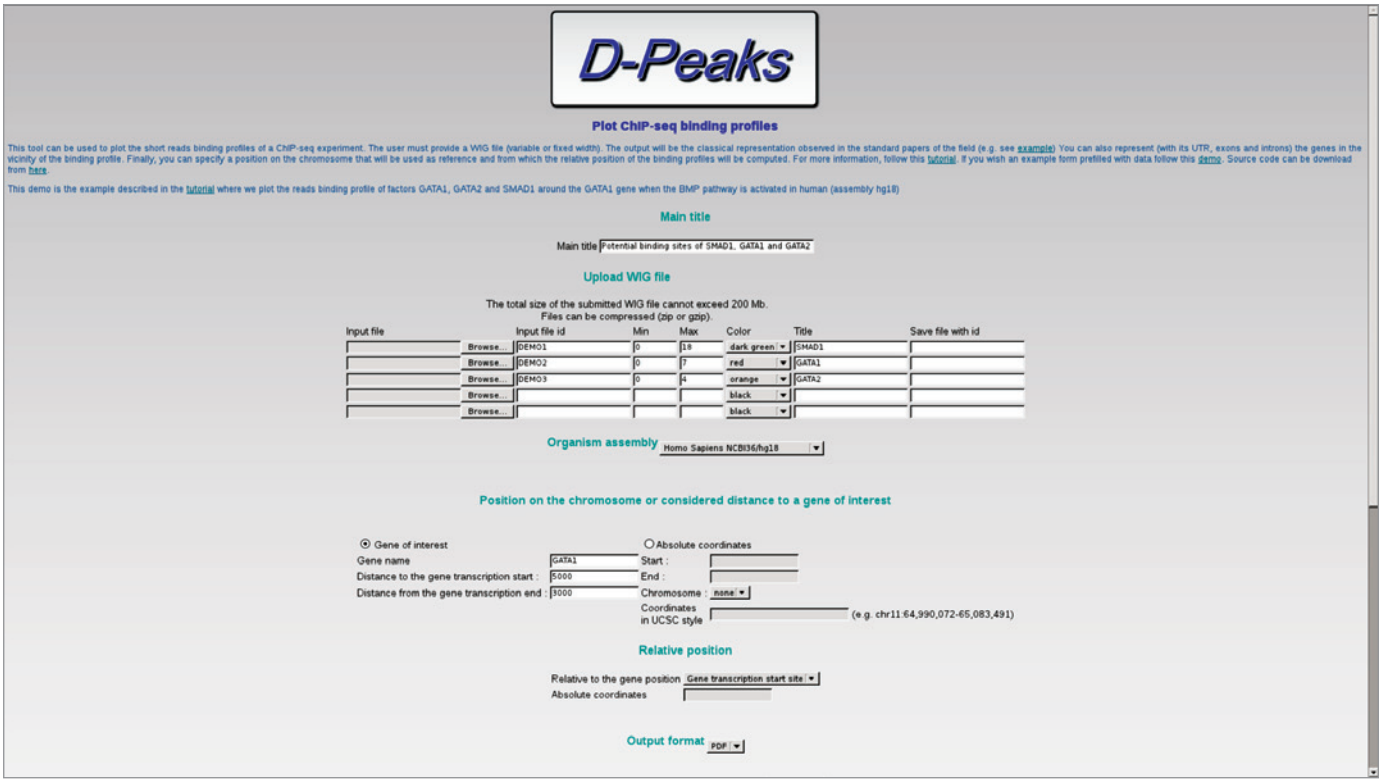

Figure 2. D-peaks main screen. Users are invited to submit the WIG files and to choose between various simple parameter possibilities to ensure an optimal rendering of their ChIP-seq data.

scale of the axes, colors of the peaks, etc.) must then be specified to obtain the resulting figure. For example, **Figure 3** displays a typical output of D-peaks where the binding sites of different transcription factors are visualized in the vicinity of the *Lefty1* gene.[8] Depending, on the number and size of uploaded scoring files, the figure is produced in seconds or minutes. Currently, the online version of D-peaks works with human, mouse, zebrafish and Drosophila assemblies, but other assemblies, genomes or features could easily be added in the future or on simple request. A guide, as well as a pre-filled demonstration form, is available from the main site helping the new user to use our tool online but also in command line.

The parsing of the genomic and the scoring files are computed using the Perl programming language, which in turns uses the R statistical environment[9] to draw the requested figures. The website consists in a simple PHP layer on top of the Perl script. R and Perl must thus be installed for D-peaks to work programmatically (i.e., in command line mode). This programming strategy allows D-peaks to be easily portable, as it does not depend on the web interface. However, when changing any input parameter, the resulting figure must be each time globally recomputed.

With the apparition of new techniques based on high-throughput sequencing, ChIP-seq has recently become one of the most successful genomics technique able to detect the localization of the DNA binding sites of a protein. However, even if several powerful, exhaustive but sometimes complicated browsers allow the visualization of these results, a nice and high quality rendering of these data is, to our knowledge, not easily achieved. We thus developed D-peaks, a ChIP-seq result analysis tool which draws a precise representation of several ChIP-seq experiments along the genomes in a few very simple steps. We are convinced that this tool may be of high interest to any scientist working and publishing in the ChIP-seq field, as indicated by some threads going in that direction on online specialized forums and as this type of representation has become a standard of the field.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Acknowledgments

We wish to thank Dr Cédric Blanpain (IRIBHM-Université Libre de Bruxelles) for giving us the initial impulse for developing D-peaks and Dr Antoine Bondue (IRIBHM-Université Libre de Bruxelles) for his help during the implementation and the redaction of this note.

Funding

S.B. is Chargé de Recherches at the Fonds National de la Recherche Scientifique de la Fédération Wallonie-Bruxelles de Belgique. G.B. is supported by grants from the la Fédération Wallonie-Bruxelles de Belgique: Actions de Recherche Concertées (ARC).

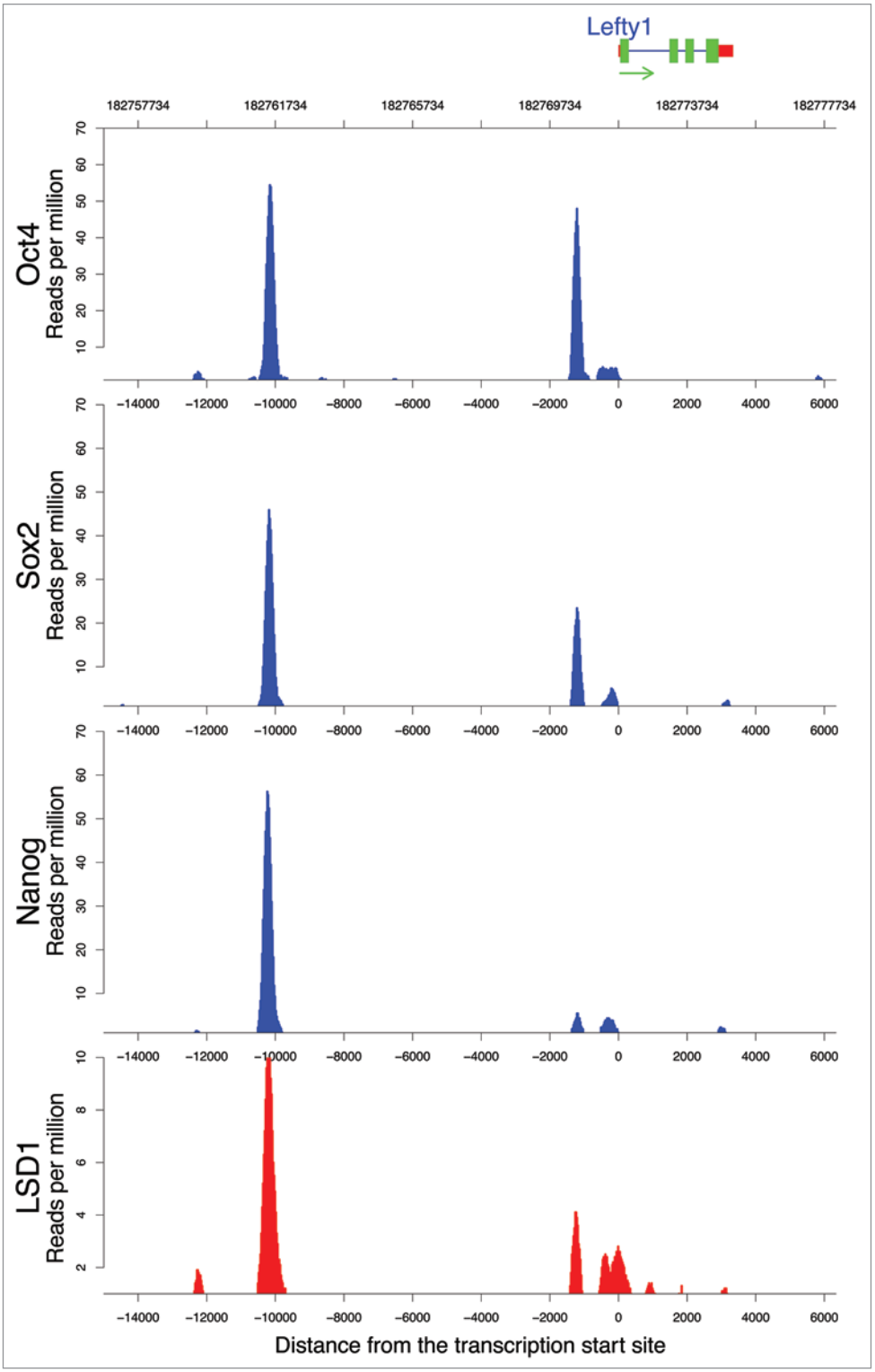

Figure 3. D-peaks results. This figure is a partial reproduction of a figure produced in Whyte et al (2012) where the authors showed that Sox2, Nanog and Oct4 (and other not shown factors) bind in the same regions that the histone demethylase LSD1 in the promoter of the *Lefty1* gene (here).[8] Note that the relative position of the peaks compared with the *Lefty1* transcription start site is clearly visible with our tool. Color code of the gene: Red, UTR; blue, introns; green, exons.

References

1. Liu ET, Pott S, Huss M. Q&A: ChIP-seq technologies and the study of gene regulation. BMC Biol 2010; 8:56; PMID:20529237; http://dx.doi.org/10.1186/1741-7007-8-56.
2. Laajala TD, Raghav S, Tuomela S, Lahesmaa R, Aittokallio T, Elo LL. A practical comparison of methods for detecting transcription factor binding sites in ChIP-seq experiments. BMC Genomics 2009; 10:618; PMID:20017957; http://dx.doi.org/10.1186/1471-2164-10-618.
3. Wilbanks EG, Facciotti MT. Evaluation of algorithm performance in ChIP-seq peak detection. PLoS One 2010; 5:e11471; PMID:20628599; http://dx.doi.org/10.1371/journal.pone.0011471.
4. Ye T, Krebs AR, Choukrallah MA, Keime C, Plewniak F, Davidson I, et al. seqMINER: an integrated ChIP-seq data interpretation platform. Nucleic Acids Res 2011; 39:e35; PMID:21177645; http://dx.doi.org/10.1093/nar/gkq1287.
5. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform 2012; PMID:22517427; http://dx.doi.org/10.1093/bib/bbs017.
6. Spudich GM, Fernández-Suárez XM. Touring Ensembl: a practical guide to genome browsing. BMC Genomics 2010; 11:295; PMID:20459808; http://dx.doi.org/10.1186/1471-2164-11-295.
7. Dreszer TR, Karolchik D, Zweig AS, Hinrichs AS, Raney BJ, Kuhn RM, et al. The UCSC Genome Browser database: extensions and updates 2011. Nucleic Acids Res 2012; 40(Database issue):D918-23; PMID:22086951; http://dx.doi.org/10.1093/nar/gkr1055.
8. Whyte WA, Bilodeau S, Orlando DA, Hoke HA, Frampton GM, Foster CT, et al. Enhancer decommissioning by LSD1 during embryonic stem cell differentiation. Nature 2012; 482:221-5; PMID:22297846.
9. R Development Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2011, ISBN 3-900051-07-0.