RSAT peak-motifs: fast extraction of transcription binding motifs from full-size ChIP-seq datasets

Morgane Thomas-Chollier[1], Matthieu Defrance[2], Olivier Sand[3], Carl Herrmann[4], Denis Thieffry[4] and Jacques van Helden[4,5]

1. Department of Computational Molecular Biology, Max Planck Institute for Molecular Genetics, Ihnestrasse 73, 14195 Berlin, Germany. Email: morgane@bigre.ulb.ac.be
2. Centro de Ciencias Genomicas, Universidad Nacional Autónoma de México. Av. Universidad, Cuernavaca, Morelos 62210, Mexico. Email: defrance@ccg.unam.mx
3. CNRS-UMR8199 Institut de Biologie de Lille. Génomique et maladies métaboliques. 1, rue du Pr Calmette, 59000 Lille, France. Email: sand@good.ibl.fr
4. Technological Advances for Genomics and Clinics (TAGC), INSERM U928 & Université de la Méditerranée. Campus de Luminy, F - 13288 Marseille, France. Email: {thieffry,herrmann}@tagc.univ-mrs.fr
5. Laboratoire de Bioinformatique des Génomes et des Réseaux (BiGRe). Université Libre de Bruxelles, Campus Plaine, CP 263. Bld du Triomphe. B-1050 Bruxelles, Belgium. Email: Jacques.van.Helden@ulb.ac.be

Keywords: ChIP-seq, meta-analysis, motif discovery, transcription factor binding motifs.

ChIP-seq has become a method of choice to study binding preferences of transcription factors, and localization of epigenetic regulatory marks at a genomic scale. There is a crucial need for specialized software tools to make sense of these data. While various programs have been developed to perform read mapping and peak calling, the subsequent steps have not yet reached proper maturation: identifying relevant transcription factor binding motifs and the precise location of their binding sites remains a bottleneck. Most existing tools present limitations on sequence size, and typically restrict motif discovery to a few hundreds peaks.

We present a pipeline called peak-motifs, integrated in the Regulatory Sequence Analysis Tools (http://rsat.ulb.ac.be/rsat/), which takes as input a set of peak sequences, discovers exceptional motifs, compares them with motif databases, predicts binding site positions, and offers different visualization interfaces. The pipeline relies on tried-and-tested algorithms whose computing time increases linearly with sequence size, ensuring scalability to massive datasets of several tens of Mb. In addition to the website, peak-motifs can be used as stand-alone application, as well as SOAP/WSDL web services.

We assessed *peak-motifs* performances on several published datasets. In all cases, relevant motifs are disclosed. For example, we discovered individual Oct and Sox motifs in Sox2 and Oct4 peak collections, whereas the original study only found the composite Sox/Oct motif. For the generic transcriptional co-activator p300 examined in heart and midbrain, *peak-motifs* identified motifs bound by tissue-specific transcription factors consistent with these two tissues.

In summary, *peak-motifs* supports time-efficient and statistically reliable analysis of *complete* ChIP-seq datasets, while offering an online user-friendly and well-documented interface.