A complete workflow for the analysis of full-size ChIP-seq (and similar) data sets using peak-motifs

Morgane Thomas-Chollier[1], Elodie Darbo[2], Carl Herrmann[2], Matthieu Defrance[3], Denis Thieffry[2,4] & Jacques van Helden[2,5]

[1]Department of Computational Molecular Biology, Max Planck Institute for Molecular Genetics, Berlin, Germany. [2]Technological Advances for Genomics and Clinics, Institut National de la Santé et de la Recherche Médicale (INSERM) U928 and Université de la Méditerranée, Marseille, France. [3]Centro de Ciencias Genomicas, Universidad Nacional Autónoma de México (UNAM), Cuernavaca, Mexico. [4]Institut de Biologie de l'Ecole Normale Supérieure—Centre National de la Recherche Scientifique Unité Mixte de Recherche (CNRS UMR) 8197 and INSERM U1024, Paris, France. [5]Laboratoire de Bioinformatique des Génomes et des Réseaux (BiGRe), Université Libre de Bruxelles, Bruxelles, Belgium. Correspondence should be addressed to M.T.-C. (thomas-c@molgen.mpg.de) or J.v.H. (jacques.van-helden@univmed.fr).

Published online 26 July 2012; doi:10.1038/nprot.2012.088

This protocol explains how to use the online integrated pipeline 'peak-motifs' (http://rsat.ulb.ac.be/rsat/) to predict motifs and binding sites in full-size peak sets obtained by chromatin immunoprecipitation–sequencing (ChIP-seq) or related technologies. The workflow combines four time- and memory-efficient motif discovery algorithms to extract significant motifs from the sequences. Discovered motifs are compared with databases of known motifs to identify potentially bound transcription factors. Sequences are scanned to predict transcription factor binding sites and analyze their enrichment and positional distribution relative to peak centers. Peaks and binding sites are exported as BED tracks that can be uploaded into the University of California Santa Cruz (UCSC) genome browser for visualization in the genomic context. This protocol is illustrated with the analysis of a set of 6,000 peaks (8 Mb in total) bound by the *Drosophila* transcription factor Krüppel. The complete workflow is achieved in about 25 min of computational time on the Regulatory Sequence Analysis Tools (RSAT) Web server. This protocol can be followed in about 1 h.

INTRODUCTION

The ChIP-seq technology[1,2] enables genome-wide detection of transcription factor binding sites and epigenetic marks. The method typically returns several millions of short sequence reads, which are mapped onto the reference genome and analyzed to extract peak regions (i.e., regions presenting a substantially high density of reads). The typical result is a list of several thousand peak regions of varying sizes (from a few tens of base pairs to several kilobases). Although various programs have been developed to perform read mapping and peak calling[3], 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 they typically restrict motif discovery to a few hundred peaks[4,5] or to the central-most part of the peaks[6].

To interpret genome-wide location data, there is a crucial need for time- and memory-efficient algorithms, interfaced as user-accessible tools to extract relevant information from high-throughput sequencing data[7,8]. For this purpose, we developed the software tool peak-motifs[9], which takes as input a set of peak sequences of interest ('test sequences'), discovers key motifs, compares them with transcription factor binding motifs from various databases, predicts the location of binding sites within the peaks and exports them in a format suitable for visualization in the UCSC Genome Browser (see **Box 1** for abbreviations). Notably, all these steps, including motif discovery, are performed on the full-size sets of peak sequences, without restrictions on peak number or width.

Workflow

The main analytical steps of the workflow are summarized hereafter and are shown in **Figure 1**.

Sequence purging (Steps 1–6). Input sequences are automatically purged to discard redundant fragments (peak overlaps, duplications), which would bias the estimation of the significance of overrepresented motifs.

Sequence composition. The distribution of sequence lengths provides a useful way to detect outlier peaks (i.e., exceptionally long peaks that may 'dilute' the motif signal) or irregular length distributions resulting from problems during the peak-calling procedure. Such indications may lead to the need for redoing the preprocessing in order to refine the peaks (e.g., by splitting large peak regions into individual peaks with PeakSplitter[10]) before using peak-motifs. Nucleotide and dinucleotide compositions are computed and displayed in the form of heat maps and positional profiles (**Box 2**).

Motif discovery (Steps 7–9). The workflow combines four word-based pattern-discovery algorithms that rely on two complementary criteria (overrepresentation and positional bias) to detect exceptional words (oligonucleotides) and spaced pairs of words (dyads; **Box 3**). Significant words are used as seeds to build probabilistic description of motifs (position-specific scoring matrices), indicating residue variability at each position of the motif.

Motif comparisons (Steps 10–12); motif databases. Discovered motifs are compared with one or several public databases of annotated motifs to predict associated transcription factors. Comparison results are displayed as multiple motif alignments to highlight matches with several annotated motifs (e.g., factors belonging to the same family, composite motifs bound by protein complexes). A personal collection of motifs may also be provided, such as the licensed TRANSFAC database (http://www.biobase-international.com/gene-regulation).

Motif comparisons (Steps 10–12); reference motif(s). ChIP-seq experiments may target transcription factors for which some

Box 1 | Abbreviations

BED: a standard format for files describing a list of genomic features (e.g., peaks, sites, gene coordinates and so on).
ChIP-seq: a combination of chromatin immunoprecipitation and massively parallel sequencing to characterize the DNA fragments bound to a protein of interest.
FASTA: a standard format for sequence files.
GEO: the Gene Expression Omnibus database.
PSSM: position-specific scoring matrix (sometimes referred to as position-weight matrices).
RSAT: Regulatory Sequence Analysis Tools.
UCSC: University of California Santa Cruz, the institution hosting the genome browser used in this protocol.

binding motifs have already been characterized and annotated in specialized databases[11–13]. Even in such cases, it is notable to discover motifs in peak sets because motifs discovered from large peak collections are generally more robust than those annotated from a handful of binding sites, leading to a substantial refinement of their predictive power[14]; and the discovery of additional motifs may reveal transcription factors interacting with the targeted factors[15]. Users can enter one or several reference motifs (i.e., motifs expected to be found in the result) to identify discovered motifs that match motifs known to be bound by the targeted factor.

Comparisons of discovered motifs with reference or database motifs are represented as matrix similarity graphs, where nodes represent motifs and edges their similarity. To grasp the groups of similar motifs returned by the different algorithms, the result can be browsed with standard network visualization programs (e.g., Cytoscape[16]).

Binding site predictions (Steps 13 and 14). Sequences are scanned with the discovered motifs to locate binding sites, and their positioning within peaks is analyzed (coverage, positional distribution along peaks).

Result visualization (Steps 15–18). Peak-motifs generates an HTML report summarizing the main results and giving access to each separate result file. The report page includes links, allowing users to upload input peaks and predicted sites to the UCSC Genome Browser[17] in order to visualize them in their genomic context.

Applications of the method

This protocol is applied but not limited to the analysis of peak sequences generated from ChIP-seq experiments. Data sets resulting from similar experiments (ChIP-PET[18], ChIP-on-chip[19], CLIP-seq[20]) can also be studied with this workflow. More generally, this approach is appropriate to motif discovery tasks applied to large collections of sequences, such as motif analysis in sets of promoter sequences centered on the transcription start site (e.g., ±250 bp around the transcription start site) or motif discovery around termination sites[21].

Main advantages of peak-motifs

Time efficiency. The processing time of the word-counting algorithms increases linearly with sequence size, whereas the complexity

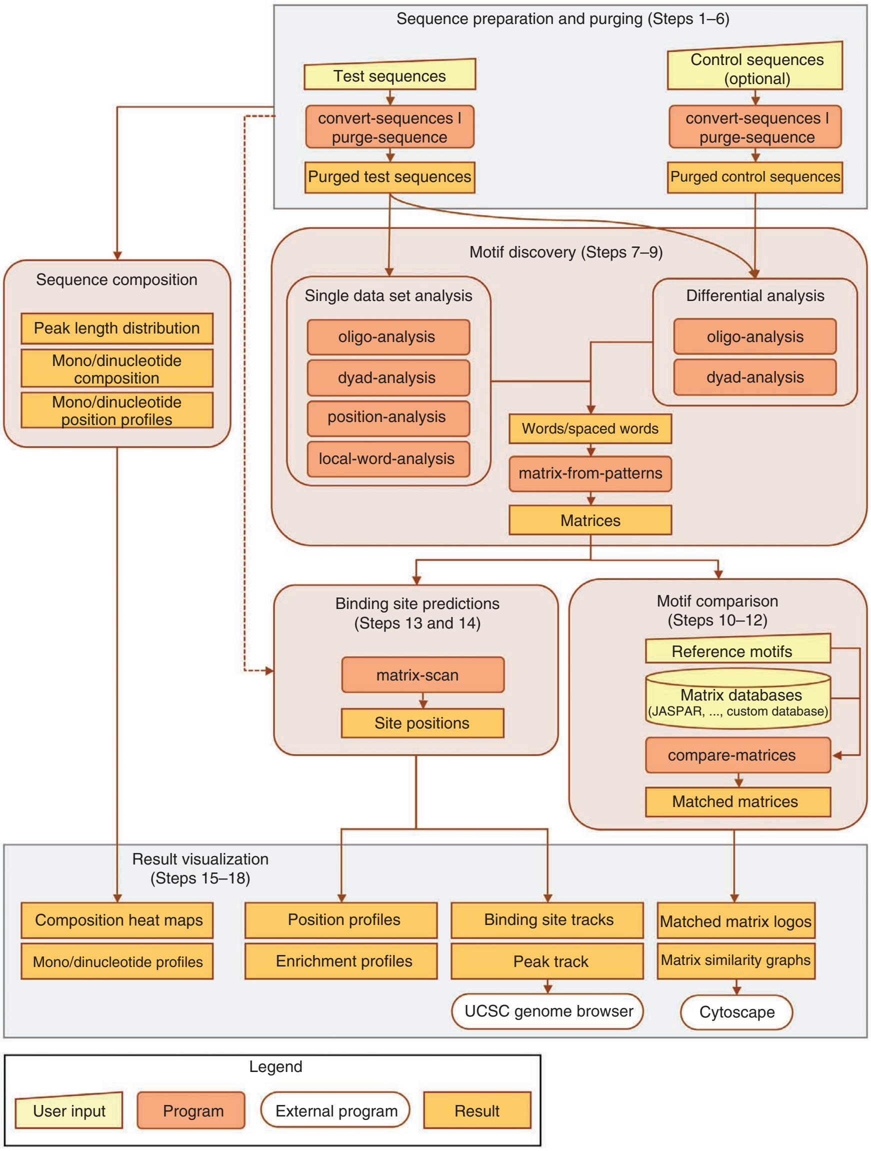


Figure 1 | Flowchart describing the peak-motifs workflow.

Box 2 | Sequence composition and background models

The choice of an appropriate background model is one of the most important criteria for predicting cis-regulatory elements. The analysis of sequence compositions in nucleotides and oligonucleotides provides useful hints for the choice of this model. **Figure 6** shows the sequence compositions of two collections of peak sequences obtained by ChIP-seq with two orthologous proteins (*Drosophila* CBP and mouse p300) that act as cofactors by interacting with multiple transcription factors. The heat maps indicate the probability to observe a given residue ('suffix', displayed in columns) following another residue ('prefix', displayed in rows). The *Drosophila* heat map shows typical aggregative tendency of As and Ts: after a 'T' prefix, there is a much higher probability to observe another T (33.4%) than an A (18.6%). A striking feature of the mouse heat map is the avoidance of CpG dinucleotides, typical of mammalian sequences: the probability of observing a G after a C is only 8%, whereas it is 30% after any other residue.

Such dependencies have an important impact on the computational analysis of cis-regulatory elements: the probability of a given site will strongly differ depending on the genomic context in which it is found. For instance, for the sequence ATCGCGAT, the probability estimated from dinucleotide composition is 1.4926×10^{-5} in *Drosophila* CBP peaks and 9.9424×10^{-7} in mouse p300 peaks. The same sequence is thus expected to occur by chance once every 66 kb in *Drosophila* p300 peaks (dist = 1/freq = $1/1.4926 \times 10^{-5}$ = 66,997), as compared with once per Mb in mouse CBP peaks ($1/9.9424 \times 10^{-7}$ = 1,005,793).

Nucleotide composition not only depends on the organism but also on the sequence type (promoters, introns, coding exons and so on) and on local particularities of the sequences. For example, the positional profiles of dinucleotide occurrences further show a specific depletion of AA, TT, TA and AT in the centers of the *Drosophila* CBP peaks.

The peak-motifs pipeline automatically computes sequence composition for words of various sizes in order to estimate the background probabilities. Background models based on simple nucleotide composition are not suited, as they fail to capture dependencies between adjacent nucleotides. Markov models of order 1 take such dependencies into account by estimating the probability of each residue depending on the preceding nucleotide (**Fig. 6**). By extension, a Markov model of order *m* can be built by computing the probabilities of each residue as a function of the *m* preceding residues. The oligo-analysis program uses such Markov models to estimate the expected frequency of longer oligonucleotides (e.g., hexanucleotides) on the basis of the frequencies of shorter words (e.g., tetranucleotides). Higher-order background models are more stringent, and return less false positives, but can result in a loss of sensitivity for small data sets.

of most other algorithms (usually based on multiple local alignments) is quadratic or worse. Our benchmarking showed that peak-motifs is able to treat peak sets of several tens of Mb in a few minutes on a personal computer[9]. Consequently, we impose no restriction on the size of the data sets analyzed on the web server. To the best of our knowledge, peak-motifs is currently the only web-supported tool applying motif discovery to full-sized data sets (**Supplementary Table 1**).

User-friendliness. Whereas each component of peak-motifs can be used as a separate tool of the RSAT suite, their organization within the pipeline makes them available for non-experts, with a user-friendly interface and preselected parameters suited for analyzing ChIP-seq data. Results are reported as a summary web page with expandable sections and links to the detailed results of each analysis step.

Multiple motif detection. The detection of multiple motifs provides clues about composite motifs and potential cofactors.

Reliability. The significance tests underlying pattern detection ensure a control of the rate of false positives, with suitable multitesting corrections.

Motif comparisons. Discovered motifs can be compared with user-specified reference motifs (i.e., the motifs expected to bind to the pulled-down factor) or with several public motif databases that can facilitate the identification of transcription factors with the potential to bind to each discovered motif.

Automation. All the operations can be readily integrated in automatic workflows, either as stand-alone applications or as web services invoked from a remote client via a SOAP/WSDL (simple object access protocol/web service definition language) programmatic interface (see ref. 22 for a description of RSAT web services).

Main limitations

Optimizing parameters and interpreting results. The workflow combines many analytic steps, each depending on several parameters that may strongly affect the outcome. In order to fully exploit the richness of the results, tuning these parameters and interpreting the results may require some experience to go beyond the superficial analysis of motif logos and predicted site maps. The goal of this protocol is precisely to guide users about the choice of parameters and the interpretation of the results.

Output redundancy. The output presents motifs in a redundant form, as the same motif can be discovered by multiple algorithms. We, however, chose to maintain this partly redundant presentation because detecting a motif by several independent programs indicates the robustness of the result. For example, a motif can be found both overrepresented ('oligo-analysis'[23], 'dyad-analysis'[24]) and concentrated in the center of the peaks ('position-analysis'[21], 'local-word-analysis')[9,25].

Input peak regions. The motif-discovery algorithms consider all input peak regions as equivalent and cannot take into consideration the actual peak shape. Such information provided as a coverage file can be taken into account by other programs such as ChIPMunk[26].

Comparison with other tools

A comparison of peak-motifs with other available tools for analyzing motifs in ChIP-seq peak sequences is available in the original publication of peak-motifs[9]. We provide here as **Supplementary Table 1** an updated version of the Table 1 from this publication listing the tasks, algorithms and usability properties of popular tools.

Box 3 | Motif discovery algorithms

Peak-motifs combines several previously described motif discovery algorithms that detect exceptional words on the basis of distinct criteria (**Fig. 12**): global overrepresentation of words (oligo-analysis[23]) or spaced word pairs (dyad-analysis[24]), local overrepresentation of words in positional windows (local-word-analysis) or heterogeneity of the word count distribution along the peak sequences (position-analysis[21]). A great advantage of these word-based algorithms is their low memory requirements and their linear time complexity regarding the data set size (i.e., computing time increases linearly with the sizes of the peak sequences). In the publication describing peak-motif performances[9], we showed that the oligo-analysis program is able to treat a 100-Mb sequence set in no more than 3 min on a MacBook laptop.

Several words returned by these algorithms can reveal fragments or variants of a same motif. Thus, the raw result (a list of scored words) has to be further processed in order to obtain a suitable description of the full motifs. For this, significant words are aligned to build PSSMs that can be used to scan sequences and predict binding sites.

Global overrepresentation of words (oligo-analysis)

The oligo-analysis program[23] (**Fig. 12**) counts the number of occurrences of each oligonucleotide ('word') of a given length (typically 6 or 7 nt, also called 6-mer or 7-mer) in the test set ('observed occurrences'), and compares it with the number of occurrences that would be expected by chance, according to a given background model. In 'single-set analysis' mode, background models for motif discovery are estimated from the oligonucleotide composition of the test sequences, with a Markov model of order m smaller than the word length minus one ($m < k-1$). The order of the Markov model should be adapted to the size of the sequence data set: we recommend low-order models ($m = 1$) to increase the sensitivity for small data sets (a few hundred kb), and higher-order models ($m = k-2$, where k is the oligonucleotide length) to increase the specificity for large sequence sets (≥ 1 Mb).

Optionally, a second sequence set ('control sequences') can be entered to estimate the random expectation of each word by the frequency of the same a word in the control set. The statistical significance of the overrepresentation is computed with the binomial distribution.

Global overrepresentation of spaced pairs of words (dyad-analysis)

Spaced motifs are characteristic of some classes of transcription factors that bind DNA in the form of homodimers or heterodimers. The dyad-analysis program[24] extends the principle of the oligo-analysis program, by counting the number of occurrences of pairs of trinucleotides separated by a spacing of fixed width but variable content. In 'single-set analysis' mode, the expected frequency of each dyad is estimated by the product of the frequencies of the two monads (trinucleotides) in the test set. In 'test versus control' mode, dyad frequencies are measured in the control set and used as estimates of prior probabilities of the same dyads in the test set. The program applies the binomial test to estimate the overrepresentation of each pair of trinucleotides with all possible spacing values from 0 to 20.

Positional biases (position-analysis)

The position-analysis program[21] (**Fig. 12b**) detects exceptional words on the basis of their positional biases, i.e., nonhomogeneous distribution relative to some reference point. For the analysis of peaks, positions are computed relative to peak centers. For other applications, reference positions can be chosen at the right extremity of the sequence (e.g., to detect upstream transcriptional), or yet at the left extremity (e.g., for the analysis of 3′ untranslated regions) (these variations are not relevant for ChIP-seq data analysis and are thus not considered in peak-motifs).

The program counts the observed number of occurrences of each oligonucleotide in nonoverlapping windows, and compares it with the count that would be expected from a homogeneous repartition. As the peaks can have variable lengths, the homogenous distribution is generally nonflat: expected occurrences typically decrease on both sides with increasing distances from peak centers (**Fig. 12b**, green curve). The significance of the difference between the observed and the homogeneous distributions is estimated with a χ^2-test.

Local overrepresentation (local-words)

The local-words program[25] detects overrepresented words in positional windows of variable or fixed size (**Fig. 12c**). In each positional window, occurrences of all oligonucleotides are counted and compared with those expected under an assumption of homogeneous distribution. The significance is estimated with the binomial distribution, where the prior probability is estimated from frequency per position in the whole sequence set.

Statistical significance of exceptional words

All the above programs return lists of words, each associated with a P value (binomial or χ^2-test depending on the program). The P value represents the nominal risk of false positive (i.e., the probability for one particular word (oligonucleotide or dyad) to show a given level of overrepresentation or positional bias by chance, according to the background model). As each analysis evaluates the significance of several thousands of words, a multitesting correction is applied by converting the P value into an E value (E value = P value × number tested words), which represents the expected number of false positives. This E value is in turn converted into a significance index $sig = -\log10(E \text{ value})$, providing an intuitive feeling of the reliability of the result (the higher the better).

Building matrices from lists of words

Each of the word-based motif discovery algorithms described above returns a set of exceptional words (oligonucleotides or dyads) sorted by significance. This list generally includes groups of mutually overlapping words, which reveal shifted fragments and variable residues of the same motif. These words are then aligned, and each group of assembled words (assembly) is used as seed to build a PSSM. The final result of the motif discovery is thereby a set of such PSSMs, which can be used to scan sequences and predict binding sites. See our previous protocols for the principle of matrix building from words[45] and sequence scanning with matrices[46].

The original publication describing the peak-motifs program[9] also provides a comparative analysis of time efficiency and a detailed analysis of motifs found on benchmark data sets. As mentioned in the original publication[9], the comparison focuses on web-interfaced software tools. Several alternative tools can be used under the Unix shell, in MATLAB[27] or as R functions[28]. Such tools, however, remain of poor usability for 'wet-lab' life-science researchers, to whom this protocol is primarily addressed.

The tool whose functionalities are most similar to those of peak-motifs is MEME-Chip[5], which combines various programs of the MEME suite in order to discover motifs and predict binding sites in a set of peaks obtained from ChIP-seq experiments. An important limitation of MEME-Chip is that the time cost of the motif discovery step, relying on the MEME program, increases as the square power of the sequence size. To circumvent this problem, MEME-Chip restricts the analysis to the 600 top peaks, clipped to 200 bp. MEME-Chip, however, also integrates DREME[6], a word-based motif discovery program based on the same principle as the oligo-analysis[23] component of peak-motifs. As for any bioinformatics analysis, it is advisable to run a few alternative programs in order to assess the robustness of the results. MEME-Chip currently constitutes the most elaborated tool to complement peak-motifs for the extraction of motifs from ChIP-seq peaks. Moreover, as several of the other ChIP-seq analysis tools cited in **Supplementary Table 1** delegate motif discovery to the MEME algorithm[4,29,30], analyses using the MEME-ChIP workflow will probably return the same motifs as these alternative workflows.

Experimental design

Input peak sequences. The peak files should contain sequences of reasonably well-defined peaks. We explain hereby three traps to avoid:

(1) Make sure that the sequence file contains peak sequences and not the raw reads. A peak file should have a size in the range of several megabytes, whereas a read file with millions of reads has a size of hundreds of megabytes to a few gigabytes. It is crucial to run peak-motifs on peak sequences, as the reads generally correspond to short fragments (typically 30 bp) on the left and on the right sides of the actual binding sites[3], and they are thus not expected to contain the actual binding sites. In addition, files containing several million reads are too large for online treatment. Files containing read sequences should first be treated with a read-mapping program (e.g., Bowtie[31]) that will align the reads on the reference genome. The resulting mapped reads should be processed with a peak-calling program (e.g., MACS[32]) to obtain the peak coordinates, and the corresponding sequences can finally be retrieved from specialized online resources (UCSC, Galaxy). Note that the processing of raw reads (read mapping) and the identification of peaks (peak-calling) are beyond the scope of this protocol. A detailed review of peak-calling software tools can be found in the study by Pepke *et al.*[3].

(2) The program expects peak sequences (in FASTA format) and not peak coordinates (BED files). If you dispose of peak coordinates in BED format, the RSAT tool 'fetch-sequences' can be used to retrieve the corresponding genomic sequences from the UCSC Genome Browser. See TROUBLESHOOTING for further information about other ways to obtain sequences from a coordinate file.

(3) Depending on the peak-calling program used, peaks may span several hundreds to thousands of base pairs. Long peak regions often result from the merging of a series of neighboring peaks. In this case, peak-motifs will perform better if these peaks are refined into subpeaks (the actual peak-shaped segments of the long peak regions), for example, with PeakSplitter[10]. This will increase the performance of position-analysis and local-word-analysis, as both algorithms search for motifs with positional biases, which are diluted when the peak regions are too broad.

Negative control sets. A negative control consists of checking the ability of a motif-discovery tool to return a negative answer when it is fed with sequences containing no specific signal. The RSAT suite offers a variety of tools to build data sets for such controls. For the analysis of microarray clusters, a typical negative control consists of analyzing the promoters of randomly selected genes (RSAT tool random-genes).

For ChIP-seq peaks, RSAT considers the following different approaches, which are suitable or not depending on the motif discovery method:

(1) Artificial sequences. The RSAT program 'random-sequences' allows users to generate artificial sequences with a composition of nucleotides or oligonucleotides mimicking that of a reference organism (higher-order Markov orders are supported, as explained in **Box 2**). Such sequences are by construction devoid of significant motifs, and a good motif-discovery program should return an empty answer once the background model is chosen in a consistent way for random sequence generation and motif discovery. Such artificial sequences are, however, moderately informative, because Markov models may not capture the complexity of biological sequences. This is particularly true for vertebrate genomes, whose composition shows strong local heterogeneities[33].

(2) Selection of random genome fragments. The 'random-genome-fragments' tool was recently added to the RSAT suite to address the specific problem of negative control for peak collections. The program takes as input a set of peak sequences and randomly selects genomic fragments of the same sizes. Random genome fragments are not ideal for programs designed to discover overrepresented motifs (oligo-analysis and dyad-analysis). Indeed, such programs can discover motifs in random collections of genome fragments, which may be biologically relevant because they correspond to recurrent motifs involved in global mechanisms (e.g., chromatin conformation). However, the significance of such global motifs should remain lower than that observed when analyzing collections of peaks pulled down with a specific transcription factor. Note that random-genome-fragments always provides a good negative control for programs that rely on position bias (position-analysis, local-word-analysis), as there is no reason for motifs to occupy particular positions along the sequences picked from random genomic locations.

Single-set analysis versus differential analysis. The pipeline can be used with two alternative modes: single-set analysis ('test set') or differential analysis ('test versus control'). In single-set analysis, the program discovers exceptional motifs (overrepresented and/or position biased) by comparison with background models built from the test set itself (**Box 2**).

Differential analysis aims to detect motifs that are overrepresented in one peak set relative to another. The concept of 'control set' presented here differs from the 'negative control': in differential analysis, the control set is typically a set of peaks resulting from a ChIP-seq experiment done with the same transcription factor but in a different tissue, at a different developmental stage or for a different experimental condition. Thus, the same collection of ChIP-seq peaks could be used as control for one analysis and as test set for another analysis. As an illustration, in the original description of the peak-motifs method[9], we applied differential analysis to discover tissue-specific motifs in peaks obtained with the mouse coactivator p300 in four different tissues (heart, midbrain, forebrain and limb, respectively). In this analysis, heart (test) versus midbrain (control) returned motifs corresponding to transcription factors expressed in the heart, whereas midbrain (test) versus heart (control) returned motifs bound by factors expressed in the brain. We foresee many potential applications of such differential analyses, which can be applied whenever ChIP-seq experiments have been led in several alternative experimental conditions.

Study case. To illustrate this protocol, we use a ChIP-seq data set obtained by pulling down the transcription factor Krüppel in 2- to 3-h-old embryos of *Drosophila melanogaster*[34]. Encoded by the gap gene *Krüppel* (*Kr*), this transcription factor plays a central role in anteroposterior patterning during early embryogenesis. Importantly, the products of gap genes and maternal factors, such as Bicoid (Bcd) and Hunchback (Hb), are known to bind to neighboring sites on the genome, within regulatory regions or enhancers driving precise spatiotemporal gene expression patterns.

This study case will illustrate how motif discovery can identify the motif corresponding to the targeted transcription factor, but also highlight potential cofactors. The starting point of our procedure is a set of ~6,000 peak coordinates returned by the peak-calling algorithm MACS, with a P value threshold of 1×10^{-5}. Because peak-calling can return very long regions that are unlikely to correspond to single binding sites, we will truncate longer peaks to a maximum of 2,000 bp.

Additional ChIP-seq study cases from vertebrates are accessible on the supporting website (http://rsat.bigre.ulb.ac.be/data/published_data/peak-motifs_Protocol_2012/), and discussed further at the end of the ANTICIPATED RESULTS. For each data set, we provide the peak sequences and the known motifs (for users wishing to rerun some analyses), as well as the result reports. These data sets were used in the original peak-motifs publication[9], which provides a complete discussion on the motifs found and their biological relevance.

MATERIALS

EQUIPMENT

- A computer connected to the Internet and a web browser
- A collection of peak sequences of interest, hereafter called 'test sequences' (in FASTA format). See EQUIPMENT SETUP for an example of a FASTA file
- Optional: a collection of 'control sequences' (in FASTA format), for differential analysis
- Optional: reference motifs, i.e., one or more position-specific scoring matrices (PSSMs) representing already known motifs, against which discovered motifs should be compared (see EQUIPMENT SETUP)
- Optional: a custom motif database, against which the results should be compared (see EQUIPMENT SETUP)

EQUIPMENT SETUP

Supporting website The supporting website provides the data (sequences, matrices) required to run this protocol, as well as a copy of the result files: http://rsat.bigre.ulb.ac.be/rsat/data/published_data/peak-motifs_Protocol_2012/. All the file paths provided below for supporting material are relative to this base URL.

Peak collections Peak sequences can be produced by custom experiments, or obtained from publicly available data sets. Such data sets can be obtained, for example, from UCSC[17] (http://genome.ucsc.edu/), GEO[35] (http://www.ncbi.nlm.nih.gov/geo/) or Galaxy[36] (https://main.g2.bx.psu.edu/).

Download the peak sequences for testing this protocol The ChIP-seq data set used to illustrate this protocol was extracted from ref. 34. We retrieved the coordinates of the peak sequences from the GEO database (http://www.ncbi.nlm.nih.gov/geo/), entry GSM511084 (ref. 34), in BED format, and uploaded them in Galaxy to retrieve the corresponding sequences in FASTA format. The FASTA sequence file exported by Galaxy can be downloaded from the supporting website, above (file: data/sequences/peak_sequences/Kr_D.mel_E01-03h_Eisen_rep1.fasta). In the Galaxy export, the header of each sequence indicates its genomic coordinates, which are parsed by peak-motifs in order to generate custom tracks of the binding sites for the UCSC Genome Browser. For example, the header '>dm3_chr2L_26210_29479_+' indicates a region between positions 26,210 and 29,479 on the forward (+) strand of the left arm of the second chromosome (chr2L) of *D. melanogaster*.

! CAUTION Some peak-calling programs such as QuEST[37] export the positions of peak centers instead of their left and right limits. In such case, peak coordinates have to be extended by adding a fixed interval (e.g., ± 200 bp) around each peak center.

Reference matrices and custom motif databases (optional) Reference matrices correspond to the motifs previously known for the studied protein, expected to be found in the results. For the study case, we will use two reference motifs representing the binding specificity of Krüppel, obtained from JASPAR[11] and FlyReg[38], respectively . These reference matrices can be downloaded from the supporting website, above (file: data/matrices/Kr_JASPAR_FlyReg.tf).

In addition, discovered motifs can be matched against a user-loaded custom motif collection (e.g., a set of user-collected motifs, or a licensed database). The web interface requires for these user-loaded motifs to be provided as TRANSFAC-formatted files. This format was chosen because its syntax permits users to document matrices with detailed information (ID, description, bound factor, site sequences and others). Matrices coming from other sources can be converted from a wide variety of formats (JASPAR, MEME, MotifSampler, AlignACE, ClusterBuster and others) to the TRANSFAC format with the RSAT tool 'convert-matrix'[25].

PROCEDURE

Access the peak-motif web form

1| Open a connection to the RSAT web server (http://rsat.ulb.ac.be/rsat/). Depending on your geographic location, use one of the mirrors available from the main page. The left menu bar provides access to the various RSAT programs. In this menu, click on the title NGS-ChIP-seq and select the tool peak-motifs. This will open the peak-motifs form (**Fig. 2**).

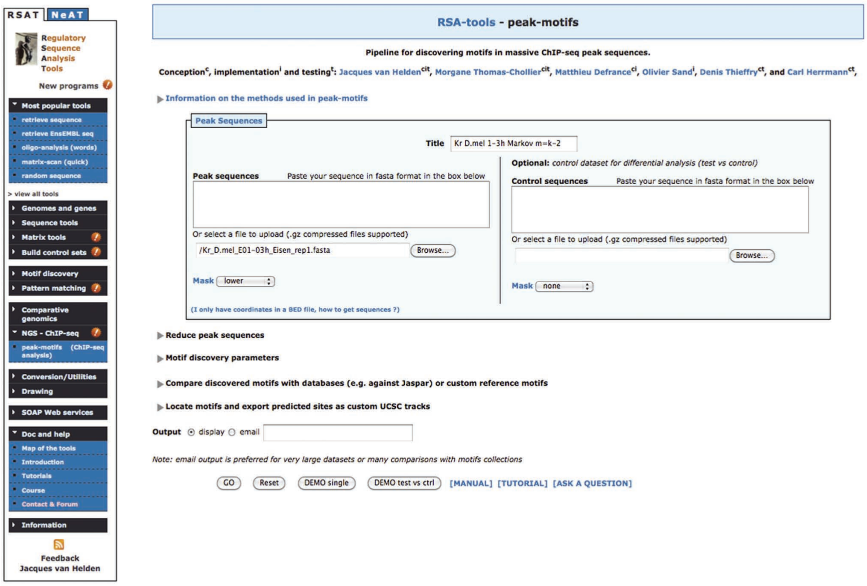


Figure 2 | Screenshot of the peak-motifs web form. By default, a simplified form is displayed. The four last sections indicated by gray arrowheads can be expanded to display the parameters for each analytic step.

RSA-tools - peak-motifs

Pipeline for discovering motifs in massive ChIP-seq peak sequences.

Conception[c], implementation[i] and testing[t]: Jacques van Helden[cit], Morgane Thomas-Chollier[cit], Matthieu Defrance[ci], Olivier Sand[i], Denis Thieffry[ct], and Carl Herrmann[ct],

Peak Sequences

Title: Kr D.mel 1-3h Markov m=k-2

Peak sequences — Paste your sequence in fasta format in the box below

Or select a file to upload (.gz compressed files supported): /Kr_D.mel_E01-03h_Eisen_rep1.fasta

Mask: lower

(I only have coordinates in a BED file, how to get sequences ?)

Optional: *control dataset for differential analysis (test vs control)*

Control sequences — Paste your sequence in fasta format in the box below

Or select a file to upload (.gz compressed files supported)

Mask: none

Output ● display ○ email

Note: email output is preferred for very large datasets or many comparisons with motifs collections

Input sequences

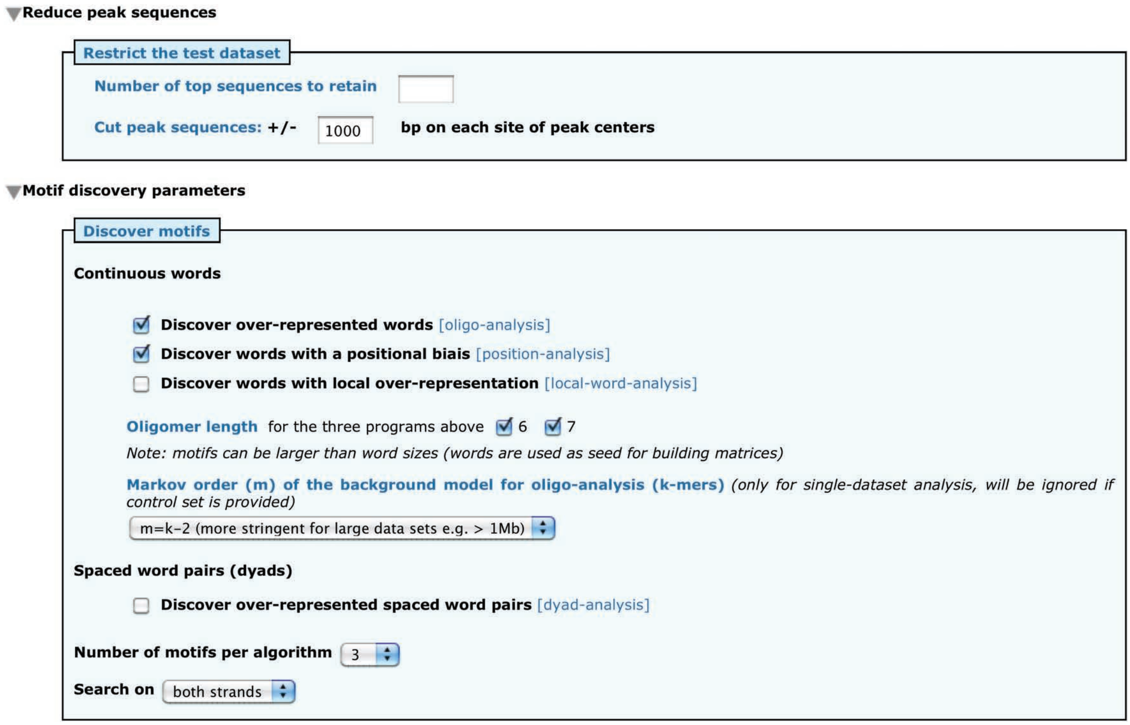
2| Specify a title for this analysis in the 'title' field of the top panel 'Peak sequences'. For the study case, type 'Kr D.mel 1–3 h Markov *m* = k – 2'. When performing a differential analysis using two data sets, the title can be formulated as 'treatment_vs_control', or 'factorX_vs_factorY', which will help you remember which data sets were given as input.

3| On the left side of the panel, under 'peak sequences', click on the 'browse' button and select the file containing the test sequences. Peak sequence(s) is the only mandatory option to run peak-motifs with default parameters. You can optionally perform a differential analysis by selecting a second sequence file with the browse button on the right side of the panel, under 'control sequences'.
▲ CRITICAL STEP We strongly advise using the 'browse' button to upload your file, rather than pasting the sequences in the box. The web browser will freeze or crash if thousands of peak sequences are pasted in the box.
▲ CRITICAL STEP If your sequence file is available on a web server (e.g., in your Galaxy account), you can directly type its URL in the text box 'URL of a sequence file available on a Web server', instead of entering your local copy with the 'browse' button. In this case, the sequences will be directly transferred from the external web server to RSAT, which avoids the double transfer (first from the server to your computer, then from your computer to RSAT).
? TROUBLESHOOTING

4| It is possible to mask repeated elements from the peak sequences via the 'mask' rolling menu (by default, the sequences are not masked). For the study case, sequences were downloaded from Galaxy and thus have repeated elements in lower case. Choose 'lower' in the 'mask' menu.

5| Peak-motifs offers the possibility to reduce the input data set by focusing on a given number of top sequences and/or by trimming the sequences to a desired length around peak centers. Click on 'Reduce peak sequences' to expand the hidden panel (**Fig. 3**). For the study case, the option 'Number of top sequences to retain' is left blank, as we will use all the peaks.
▲ CRITICAL STEP Several other ChIP-seq analysis tools automatically restrict the motif discovery step to a few hundred peaks, because the underlying algorithm would take several days or weeks to treat the whole data set. As all motif-discovery programs used in peak-motifs have a linear time response[9], the option 'Number of top sequences to retain' should generally be left blank. The option was actually included in the interface to ease comparison with third-party programs. In some cases, however, it can be interesting to analyze successively increasing numbers of top peaks, and to investigate the effect of peak number on the discovered motifs. Such analyses can help determine the best conditions for peak calling (assuming that the peaks have been sorted by decreasing calling score before fetching the sequences).

6| In the same panel, the option 'Cut peak sequences' restricts the analysis to the central region of each peak (e.g., peak center ± 200 bp), which is supposed to be dense in binding sites. This assumption is nevertheless highly dependent on the peak-calling program, and on whether the peak centers actually correspond to their summits (where the binding site is supposed to be found). We thus generally recommend leaving this option blank, or running both the restricted and complete analyses and comparing the results. However, in the Krüppel data set used as study case here, the peaks downloaded from GEO have highly variable sizes (up to 13,205 bp), which probably reflects a problem with the peak-calling procedure rather than the natural extension of Krüppel binding regions. For this particular study case, we will restrict this particular analysis to 1,000 bp on each side of peak centers.



▼Reduce peak sequences

Restrict the test dataset

Number of top sequences to retain

Cut peak sequences: +/- 1000 bp on each site of peak centers

▼Motif discovery parameters

Discover motifs

Continuous words

☑ Discover over-represented words [oligo-analysis]

☑ Discover words with a positional biais [position-analysis]

☐ Discover words with local over-representation [local-word-analysis]

Oligomer length for the three programs above ☑ 6 ☑ 7

Note: motifs can be larger than word sizes (words are used as seed for building matrices)

Markov order (m) of the background model for oligo-analysis (k-mers) *(only for single-dataset analysis, will be ignored if control set is provided)*

m=k-2 (more stringent for large data sets e.g. > 1Mb)

Spaced word pairs (dyads)

☐ Discover over-represented spaced word pairs [dyad-analysis]

Number of motifs per algorithm 3

Search on both strands

Figure 3 | Input sequence treatment (top) and motif discovery (bottom) options. An essential parameter is the choice of the background model, whose stringency should be adapted to the sequence size. The suggested way to fill in these options for the Krüppel case study is displayed.

Motif-discovery parameters

7| Click on 'Motif discovery parameters' to expand the motif-discovery option panel (**Fig. 3**). For the study case, keep the options oligo-analysis and position-analysis checked (the other algorithms may be checked for a full analysis, but this takes more time). Check the values 6 and 7 for 'Oligomer length', in order to detect significant hexanucleotides, as well as heptanucleotides.

▲ CRITICAL STEP The choice of motif discovery algorithms markedly affects the result. It is generally recommended to combine the analysis of overrepresentation (oligo-analysis) and positional bias (position-analysis). For this protocol, we do not activate the detection of spaced pairs (dyad-analysis) and locally overrepresented words (local-word) because they require more processing time, but they should be considered as suitable approaches for a full exploitation of ChIP-seq data sets.

8| The background model must be specified when analyzing a single set of peaks. In the case of differential analysis, the second set of peaks (control set) serves as background to estimate the random expectation of each oligonucleotide. In 'single-set analysis' mode, the background model is built from the test sequences based on frequencies of smaller words. For the study case, select the most stringent background model ($m = k-2$), as the complete sequence set is large (8 Mb). With this option, hexanucleotides ($k = 6$) will be analyzed with a fourth-order and heptanucleotides ($k = 7$) with a fifth-order Markov model.

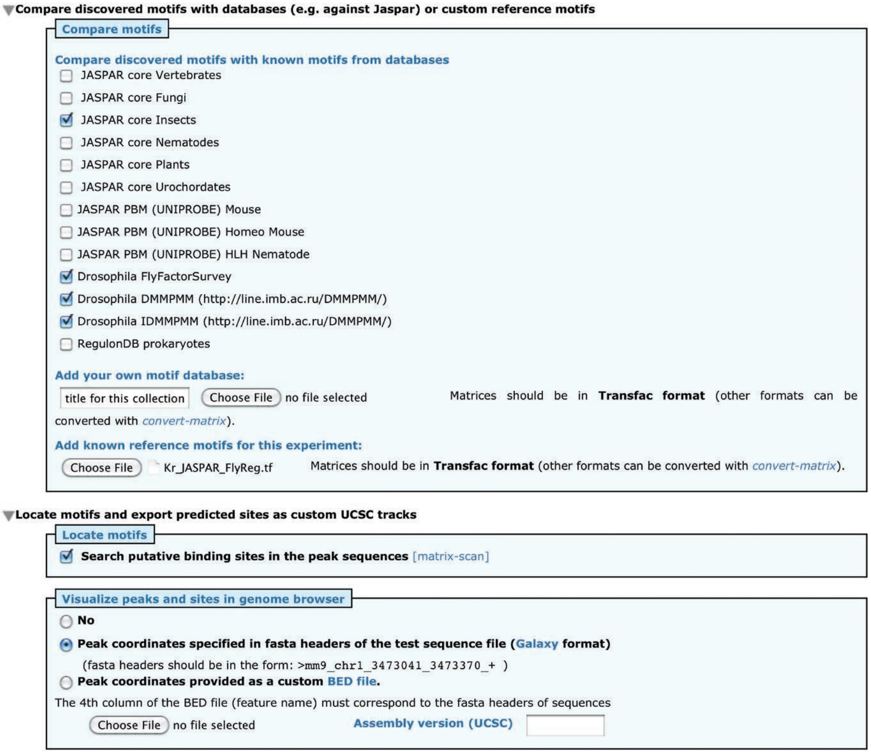
▲ CRITICAL STEP For single data sets, the background model must be chosen carefully, as this parameter strongly affects the results. **Box 2** explains how the background model is calculated from the input sequences, and **Box 3** explains how to select the Markov order depending on total test sequence sizes. The option 'Automatic' will select the most appropriate model according to the size of the test sequences. Note that the size limits chosen for selecting the Markov order are somewhat arbitrary and may thus be adapted to your particular data set.

9| Set the 'Number of motifs per algorithm' according to your needs. For the study case (Kr peaks in the *Drosophila* embryo), set this option to 3. Higher values (e.g., between 5 and 10) can be useful when analyzing data sets supposed to contain combinations of motifs bound by different transcription factors (e.g., peaks obtained by immunoprecipitation of a general coactivator such as p300).

▲ CRITICAL STEP Increasing the number of motifs has a cost in computing time (building matrices from significant words, comparisons with motif databases, peak scanning to detect site positions).

Comparisons of discovered motifs with motif databases and reference motifs

10| Click on 'Compare discovered motifs with databases' to reveal the 'Compare motifs' panel. This section displays a list of public motif databases, such as JASPAR[11], that are directly supported by peak-motifs. Each discovered motif will be compared with the selected collection(s) of motifs, in order to identify which transcription factors may correspond to these binding motifs, or to pinpoint the currently unknown motifs. Motif databases should be chosen according to the studied organism. For the study case, unselect the default database ('JASPAR core Vertebrates') and select all the databases related to *Drosophila* ('JASPAR core Insects', '*Drosophila* FlyFactorSurvey', '*Drosophila* DMMPMM' and '*Drosophila* IDMMPMM'), as illustrated in **Figure 4**. ▲ CRITICAL STEP Because of limitations in annotating resources, motif databases are very incomplete and should not be considered as comprehensive knowledge repositories. We strongly encourage users working on a specific factor to independently search the literature for documented binding motifs, and provide these to peak-motifs as reference motifs (see Step 12).



▼Compare discovered motifs with databases (e.g. against Jaspar) or custom reference motifs

Compare motifs

Compare discovered motifs with known motifs from databases

□ JASPAR core Vertebrates
□ JASPAR core Fungi
☑ JASPAR core Insects
□ JASPAR core Nematodes
□ JASPAR core Plants
□ JASPAR core Urochordates
□ JASPAR PBM (UNIPROBE) Mouse
□ JASPAR PBM (UNIPROBE) Homeo Mouse
□ JASPAR PBM (UNIPROBE) HLH Nematode
☑ Drosophila FlyFactorSurvey
☑ Drosophila DMMPMM (http://line.imb.ac.ru/DMMPMM/)
☑ Drosophila IDMMPMM (http://line.imb.ac.ru/DMMPMM/)
□ RegulonDB prokaryotes

Add your own motif database:

title for this collection | Choose File | no file selected — Matrices should be in **Transfac format** (other formats can be converted with *convert-matrix*).

Add known reference motifs for this experiment:

Choose File | Kr_JASPAR_FlyReg.tf — Matrices should be in **Transfac format** (other formats can be converted with *convert-matrix*).

▼Locate motifs and export predicted sites as custom UCSC tracks

Locate motifs

☑ **Search putative binding sites in the peak sequences** [matrix-scan]

Visualize peaks and sites in genome browser

○ **No**
● **Peak coordinates specified in fasta headers of the test sequence file (Galaxy format)**
(fasta headers should be in the form: >mm9_chr1_3473041_3473370_+)
○ **Peak coordinates provided as a custom BED file.**

The 4th column of the BED file (feature name) must correspond to the fasta headers of sequences

Choose File | no file selected — **Assembly version (UCSC)**

Figure 4 | Options for motif comparisons (top) and predicted sites visualization (bottom). The options are filled-in for the Krüppel case study.

11| If you make use of your own motif collections (e.g., licensed databases, custom matrices), make sure that they are formatted as TRANSFAC files (if not, use the tool 'convert-matrix' on the RSAT Web site) and upload the files by clicking on the 'browse' button of the section 'Add your own motif database'. A title should be specified for this custom database in the field on the left on the 'browse' button. For the study case, we will only use the public databases available on RSAT, and thus this option will be left blank.

12| One or several reference motifs can also be uploaded (in a single TRANSFAC-formatted file) by clicking on the 'browse' button in the section 'Add known reference motifs for this experiment'. For the study case, use the file data/matrices/Kr_JASPAR_FlyReg.tf downloaded from the supporting website (EQUIPMENT SETUP).

Search for binding sites and export as UCSC custom track

13| Click on 'Locate motifs and export predicted sites as custom UCSC tracks' to expand the panel with the options for searching putative binding sites in the peak sequences (**Fig. 4**, bottom). Check the box 'Search putative binding sites in the peak sequences'.

14| Optionally, the coordinates of test peaks and predicted sites can be exported as a custom track (BED file) that can be uploaded in the UCSC[17] or Ensembl[39] genome browsers, in order to visualize these putative binding sites in their annotated genomic environment. By default, this very helpful way to interpret the results is disabled, as it requires information in addition to peak sequences (genome assembly version and coordinates of the peaks). The required information can be provided in either of two ways: if your sequences have been fetched from Galaxy, check the radio button 'Peak coordinates specified in FASTA headers of the test sequence file (Galaxy format)'; otherwise, check the radio button 'Peak coordinates provided as a custom BED file', locate the .bed file indicating peak coordinates with the button 'Browse' and indicate the 'Assembly version (UCSC)'. For the study case, as sequences were previously downloaded from the Galaxy server, we simply need to check the second radio button.

Submit the form

15| Check the 'email output' option and provide your email address, in order to be notified when the results are ready. Alternatively, you can keep the 'display' output to obtain the results directly in the web browser. The email output is generally preferred for large data sets or when results are compared with many motif collections, because the whole processing can take around 20–30 minutes.

16| Click on the 'GO' button to run the analysis.

Viewing the results

17| A new page appears in place of the form, indicating that the task has been submitted to the server. A link to the results is displayed; click on this link to follow the analysis.

? TROUBLESHOOTING

18| Results are displayed on this page progressively, so that it is possible to start studying the results in the course of the analysis. The report page should be regularly refreshed to show the updated results. When the whole analysis is completed, the top of the page displays a summary of the results instead of the message 'Status: running …'.

? TROUBLESHOOTING

Troubleshooting advice can be found in **Table 1**.

TABLE 1 | Troubleshooting table.

Step	Problem	Cause	Solution
3	The peak files do not contain any sequence, but only genomic coordinates	Peak calling programs often return the genomic coordinates (generally in BED format), but not the sequences directly	There are several ways to retrieve the genomic sequences corresponding to the bed-specified coordinates: - using the RSAT 'fetch-sequences' tool; - from the Galaxy server (https://main.g2.bx.psu.edu/); - from the UCSC genome browser (http://genome.ucsc.edu/). We provide a step-by-step explanation at the bottom of the 'Peak sequences' panel, through the link 'I only have coordinates in a BED file, how to get sequences?'
17	The motif discovery programs return no or only weakly significant motifs	The order of the Markov model may be too high for the sequence size	Check the total sequence size in the 'Sequence Composition' box of the result page, and adapt the Markov order accordingly. If no motif is significant with the recommended background model, reducing the Markov order will increase the sensitivity, but this will be at the cost of specificity (you should expect more false positives, e.g., A+T-rich motifs)

● TIMING

The processing time depends on the server load (the number of jobs currently running on the server), on the selected tasks and on the total sequence size. For the Krüppel case study (6,003 peaks totaling 8 Mb), the complete analysis (sequence composition, motif discovery, motif comparisons, sequence scanning) took 17 min from the job submission to the reception of the email announcing the completion of the analysis. Note that results are progressively displayed on the website as they are produced, and the report page is refreshed every 120 s to indicate the progress of the analysis.

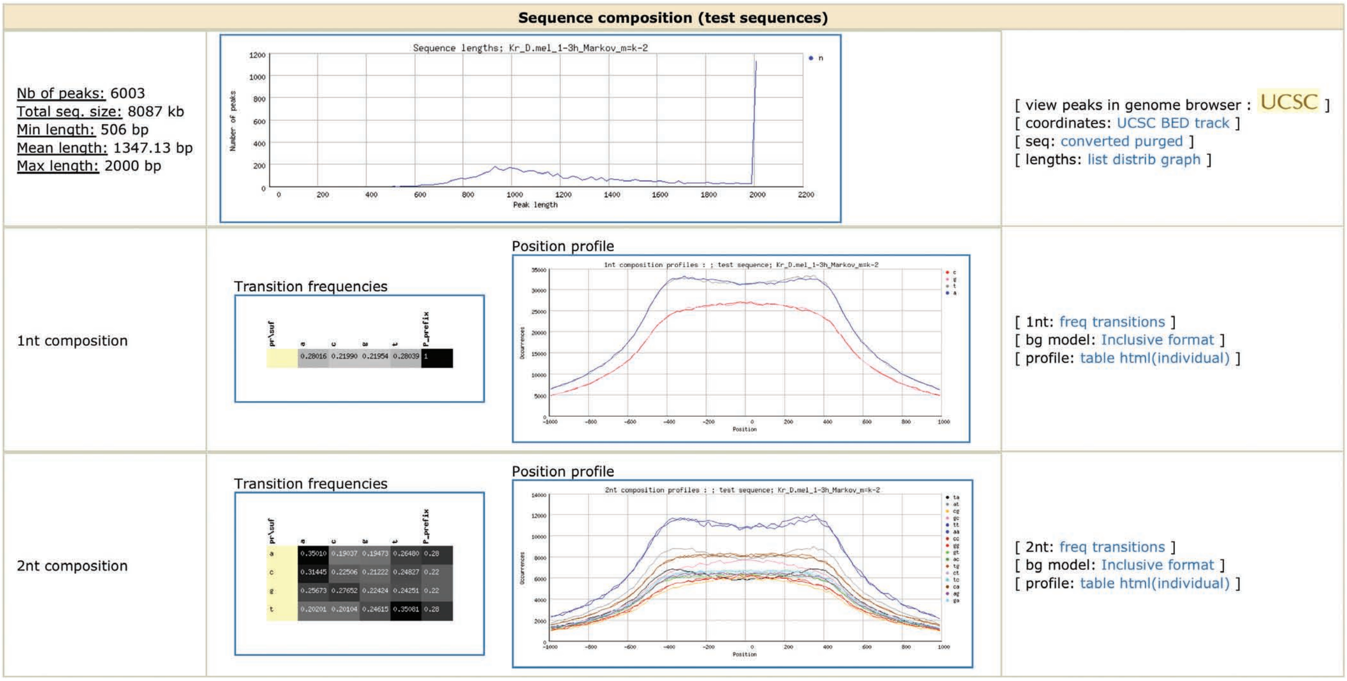
ANTICIPATED RESULTS

The result of peak-motifs is presented as a synthetic report with clickable links to the detailed result files. To ease the interpretation of the results, the report is organized in thematic sections as presented below.

For convenience, the synthetic report of the study case can be viewed on the supporting website (see EQUIPMENT SETUP; file: study_case_Kr/peak-motifs_synthesis.html).

Sequence length distribution and composition

The first section of the report (**Fig. 5**) shows the length distribution and composition of the peak sequences. The distribution of sequence lengths gives some hints about the pre-processing (peak-calling). The rightmost part of the report gives direct access to the sequence files. The link 'converted' gives access to the sequences obtained after clipping and selection of top sequences, if these options have been activated. 'Purged sequences' points to filtered sequences, where redundant fragments (peak overlaps, duplications) have been masked (replaced by 'N' characters). Motif discovery is performed on the purged sequences to avoid statistical biases because of redundant segments, whereas sequence scanning is done on the nonpurged sequences to locate all the putative binding sites.


Sequence lengths: Kr_D.mel_1-3h_Markov_m=k-2
n
Number of peaks
Peak length
1nt composition profiles : ; test sequence; Kr_D.mel_1-3h_Markov_m=k-2
2nt composition profiles : ; test sequence; Kr_D.mel_1-3h_Markov_m=k-2
Occurrences
Position

Sequence composition (test sequences)

Nb of peaks: 6003
Total seq. size: 8087 kb
Min length: 506 bp
Mean length: 1347.13 bp
Max length: 2000 bp

[view peaks in genome browser : UCSC]
[coordinates: UCSC BED track]
[seq: converted purged]
[lengths: list distrib graph]

1nt composition

Transition frequencies

pr\suf	a	c	g	t	P_prefix
	0.28016	0.21990	0.21954	0.28039	1

Position profile

[1nt: freq transitions]
[bg model: Inclusive format]
[profile: table html(individual)]

2nt composition

Transition frequencies

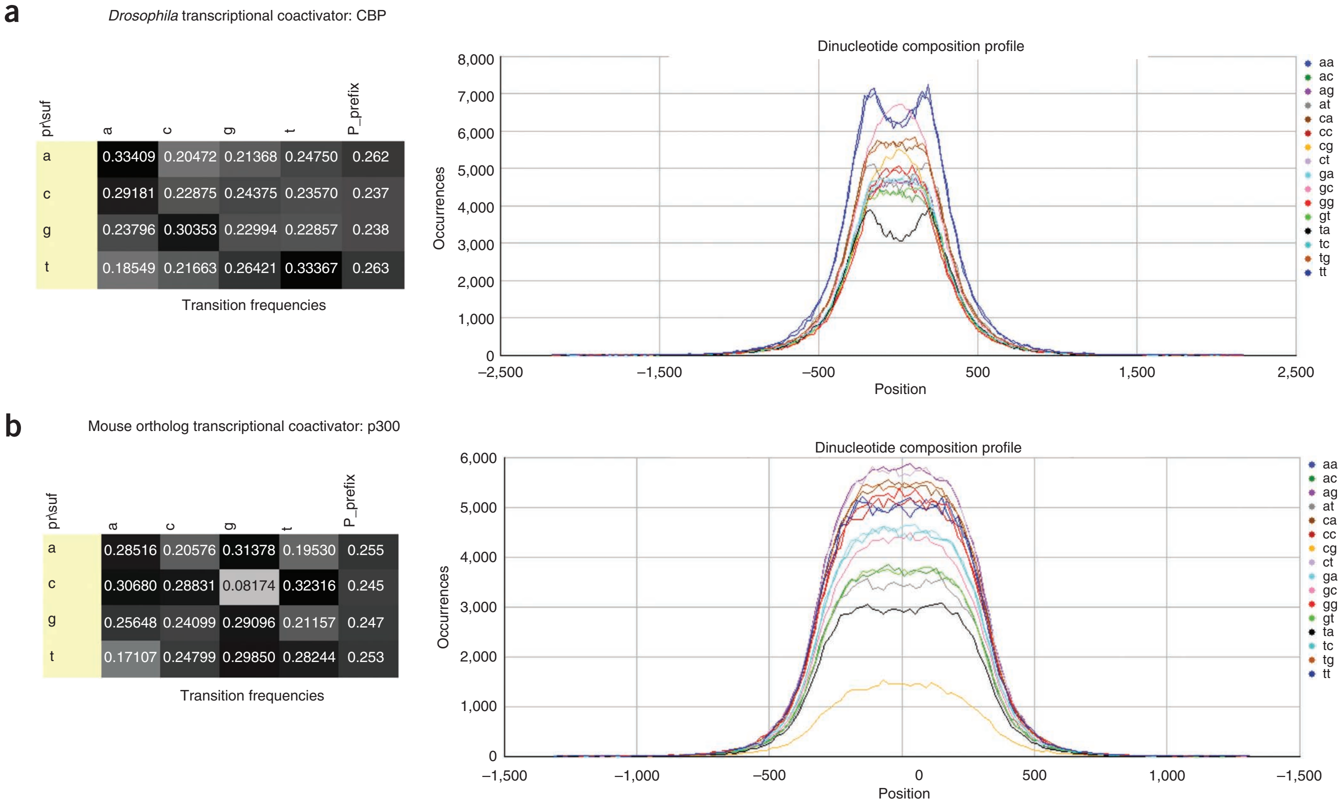
pr\suf	a	c	g	t	P_prefix
a	0.35010	0.19037	0.19473	0.26480	0.28
c	0.31445	0.22506	0.21222	0.24827	0.22
g	0.25673	0.27652	0.22424	0.24251	0.22
t	0.20201	0.20104	0.24615	0.35081	0.28

Position profile

[2nt: freq transitions]
[bg model: Inclusive format]
[profile: table html(individual)]

**Figure 5 | ** Sequence lengths and composition. From top to bottom, distribution of peak lengths, nucleotide and dinucleotide composition heat maps (left) and position profiles (right).

In the study case, the original peaks ranged from 506 bp to 13,205 bp, but they were clipped to a maximal size of 2 kb (1,000 bp on each side of each peak center). As evidenced by the abrupt rise at the right end of the distribution, ~1,100 peaks were clipped at 2,000 bp. Even after clipping, the mean peak length is still 1,347 bp, which exceeds by far the length



Drosophila transcriptional coactivator: CBP

pr\suf	a	c	g	t	P_prefix
a	0.33409	0.20472	0.21368	0.24750	0.262
c	0.29181	0.22875	0.24375	0.23570	0.237
g	0.23796	0.30353	0.22994	0.22857	0.238
t	0.18549	0.21663	0.26421	0.33367	0.263

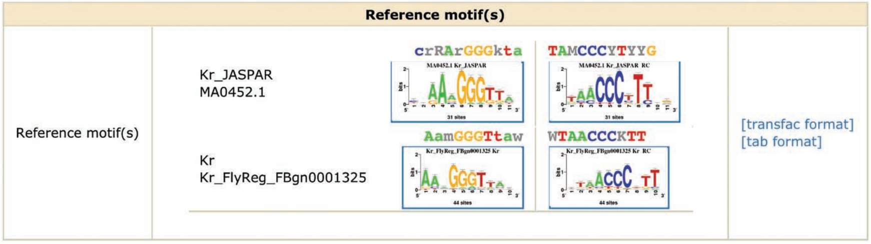
Transition frequencies

Mouse ortholog transcriptional coactivator: p300

pr\suf	a	c	g	t	P_prefix
a	0.28516	0.20576	0.31378	0.19530	0.255
c	0.30680	0.28831	0.08174	0.32316	0.245
g	0.25648	0.24099	0.29096	0.21157	0.247
t	0.17107	0.24799	0.29850	0.28244	0.253

Transition frequencies

Figure 6 | ** Dinucleotide composition and derived background models. (a**) Peaks bound by *Drosophila* transcriptional the co-regulator CBP (GEO sample GSM439463). (**b**) Peaks bound by the mouse cofactor p300 (GEO sample GSM559652), ortholog of *Drosophila* CBP. The heat maps (on the left) represent transition frequencies between prefix (rows) and suffix (columns) residues (the last column lists the frequencies of single nucleotides).



Reference motif(s)

Reference motif(s)

Kr_JASPAR
MA0452.1

crRArGGGkta | TAMCCCYTYYG

Kr
Kr_FlyReg_FBgn0001325

AamGGGTtaw | WTAACCCKTT

[transfac format]
[tab format]

Figure 7 | Reference motifs. Reference motifs can be entered to indicate which motifs are expected to be found (the 'correct' answer). Note that reference motifs are ignored during the motif discovery step; they are used *a posteriori* for validating the discovered motifs.

of a single binding site, or even the lengths of typical *Drosophila* enhancers. This unexpectedly large peak size is likely to result from suboptimal choices for the peak-calling procedure in the original publication. In principle, in such conditions, we would recommend redoing the peak-calling procedure with some alternative programs[3] and testing the effect of their parameters on the distribution of peak lengths. We will, however, pursue the analysis of this data set to highlight the interest of combining multiple criteria (occurrences and positions) for discovering exceptional motifs.

The nucleotide and dinucleotide compositions shown in **Figure 5** are typical of *Drosophila* noncoding sequences (**Box 2**; **Fig. 6**). The positional profiles show the depletion of some nucleotides (A and T) and dinucleotides (AA, TT, AT, TA) at the center of the peaks, suggesting a general avoidance of A/T-rich sequences in the Krüppel sites.

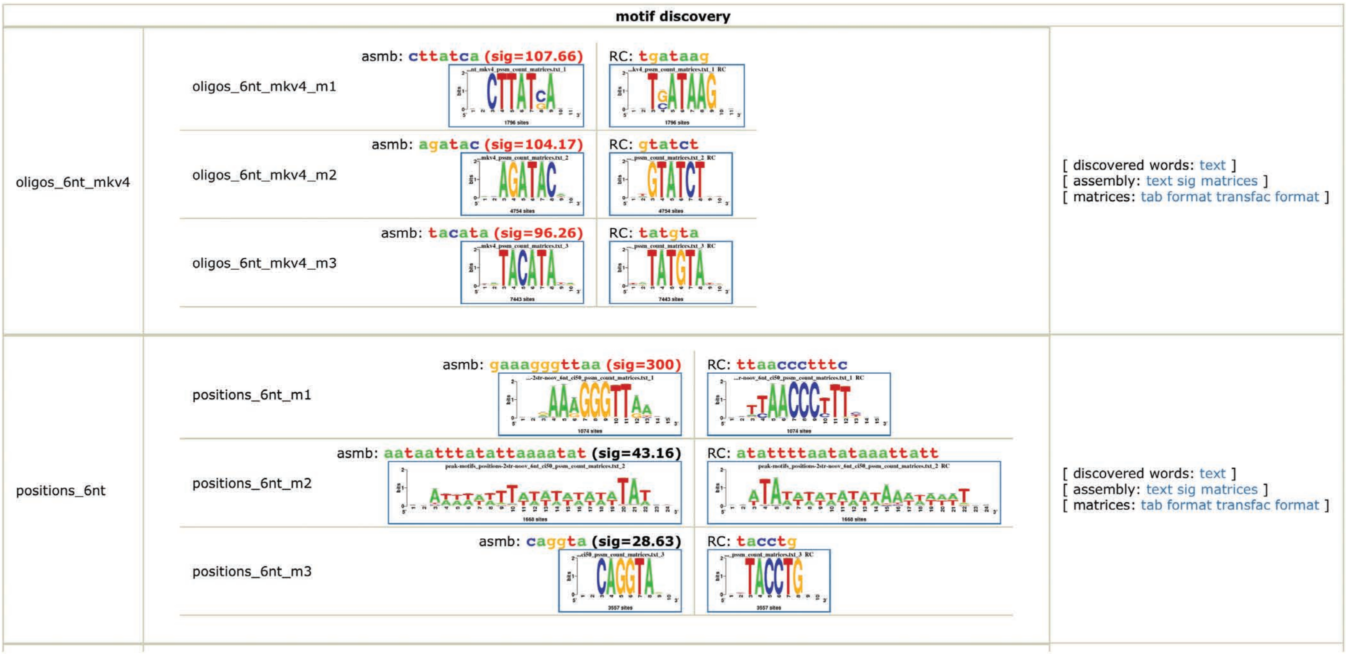
Reference motifs

For the study case, we use as reference two Krüppel binding motifs extracted from JASPAR[11] and FlyReg[38], respectively. Logos are displayed in both direct and reverse-complementary orientations (**Fig. 7**). The colored consensus sequences are shown above the logos, and can be searched for in the HTML output using the text search function of the browser.

Discovered motifs (by algorithm)

Figure 8 shows the discovered motifs, grouped by algorithm (to display it, click on the triangle on the right of the title 'Discovered motifs (by algorithm)'). Each motif is represented by its direct and reverse complementary logos, its colored consensus and its significance. The score associated with each motif is the binomial significance returned by oligo-analysis[23] for the most significant oligonucleotide used as seed to build the matrix.

In the present case, several highly significant motifs have been found, including a motif discovered by position-analysis, aaagggttaa, which is strongly similar to the canonical Krüppel motif. This view facilitates the comparison between the outputs of different algorithms.



motif discovery

oligos_6nt_mkv4

oligos_6nt_mkv4_m1 — asmb: cttatca (sig=107.66) | RC: tgataag

oligos_6nt_mkv4_m2 — asmb: agatac (sig=104.17) | RC: gtatct

oligos_6nt_mkv4_m3 — asmb: tacata (sig=96.26) | RC: tatgta

[discovered words: text]
[assembly: text sig matrices]
[matrices: tab format transfac format]

positions_6nt

positions_6nt_m1 — asmb: gaaagggttaa (sig=300) | RC: ttaaccctttc

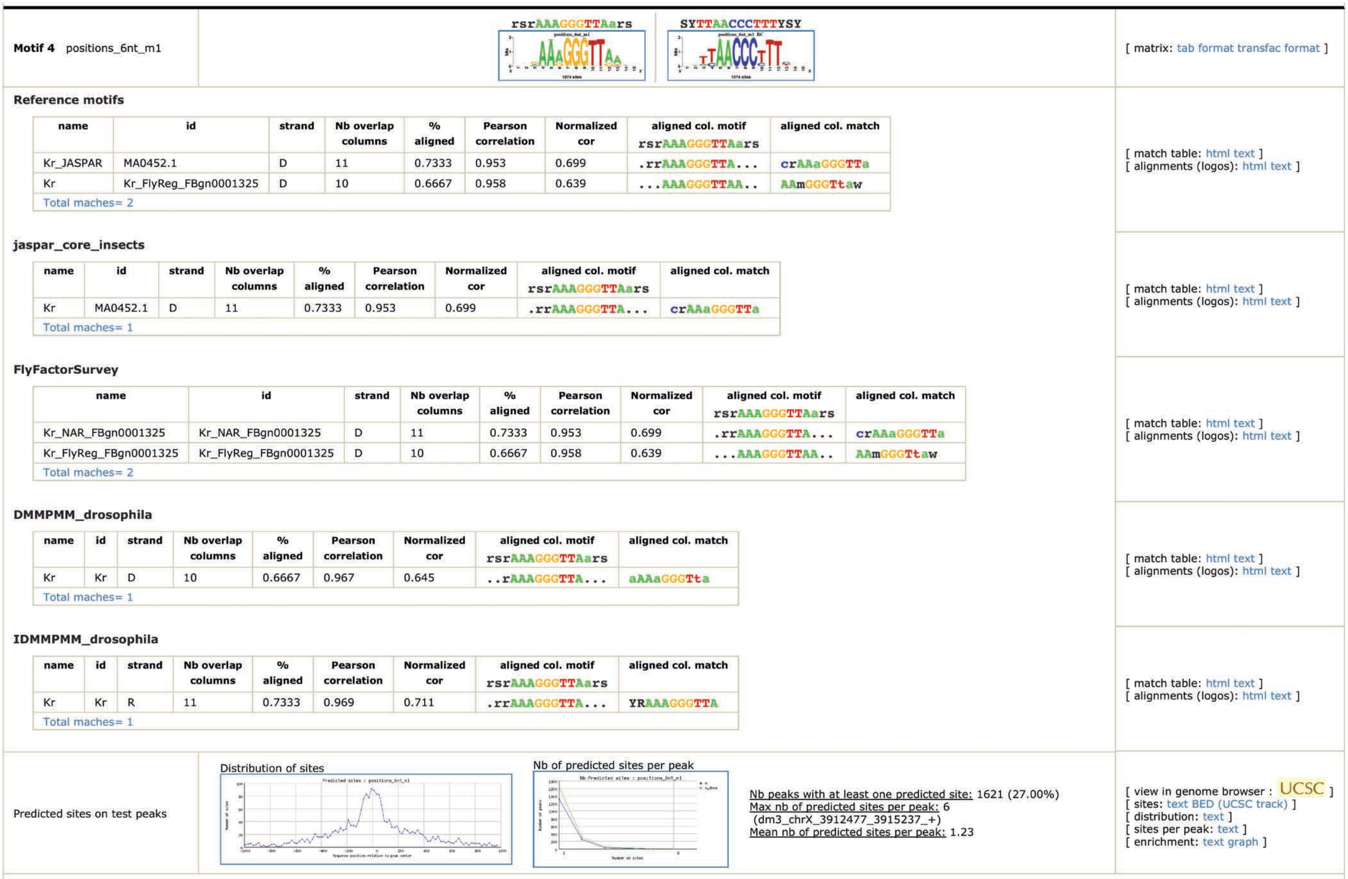
positions_6nt_m2 — asmb: aataatttatattaaaatat (sig=43.16) | RC: atattttaatataaattatt

positions_6nt_m3 — asmb: caggta (sig=28.63) | RC: tacctg

[discovered words: text]
[assembly: text sig matrices]
[matrices: tab format transfac format]

...nt_mkv4_pssm_count_matrices.txt_1
...kv4_pssm_count_matrices.txt_1 RC
1796 sites
...mkv4_pssm_count_matrices.txt_2
...pssm_count_matrices.txt_2 RC
4754 sites
...mkv4_pssm_count_matrices.txt_3
...pssm_count_matrices.txt_3 RC
7443 sites
...-2str-noov_6nt_ci50_pssm_count_matrices.txt_1
...r-noov_6nt_ci50_pssm_count_matrices.txt_1 RC
1074 sites
peak-motifs_positions-2str-noov_6nt_ci50_pssm_count_matrices.txt_2
peak-motifs_positions-2str-noov_6nt_ci50_pssm_count_matrices.txt_2 RC
1668 sites
...ci50_pssm_count_matrices.txt_3
...pssm_count_matrices.txt_3 RC
3557 sites
bits

Figure 8 | Discovered motifs grouped by algorithm. Motifs discovered by the oligo-analysis program (top) and the position-analysis program (bottom); only two selected motifs are shown. The motif identifier (first column) indicates the algorithm, oligonucleotide length and Markov model order (only for oligo-analysis). For each program, the three best-scored motif logos are displayed. The last column contains links to intermediate results (overrepresented words, assemblies of overlapping words and significance matrices), as well as to matrices in TRANSFAC and tab formats.



Motif 4 positions_6nt_m1

rsrAAAGGGTTAars

SYTTAACCCTTTYSY

[matrix: tab format transfac format]

Reference motifs

name	id	strand	Nb overlap columns	% aligned	Pearson correlation	Normalized cor	aligned col. motif rsrAAAGGGTTAars	aligned col. match
Kr_JASPAR	MA0452.1	D	11	0.7333	0.953	0.699	.rrAAAGGGTTA...	crAAaGGGTTa
Kr	Kr_FlyReg_FBgn0001325	D	10	0.6667	0.958	0.639	...AAAGGGTTAA..	AAmGGGTtaw
Total maches= 2								

[match table: html text]
[alignments (logos): html text]

jaspar_core_insects

name	id	strand	Nb overlap columns	% aligned	Pearson correlation	Normalized cor	aligned col. motif rsrAAAGGGTTAars	aligned col. match
Kr	MA0452.1	D	11	0.7333	0.953	0.699	.rrAAAGGGTTA...	crAAaGGGTTa
Total maches= 1								

[match table: html text]
[alignments (logos): html text]

FlyFactorSurvey

name	id	strand	Nb overlap columns	% aligned	Pearson correlation	Normalized cor	aligned col. motif rsrAAAGGGTTAars	aligned col. match
Kr_NAR_FBgn0001325	Kr_NAR_FBgn0001325	D	11	0.7333	0.953	0.699	.rrAAAGGGTTA...	crAAaGGGTTa
Kr_FlyReg_FBgn0001325	Kr_FlyReg_FBgn0001325	D	10	0.6667	0.958	0.639	...AAAGGGTTAA..	AAmGGGTtaw
Total maches= 2								

[match table: html text]
[alignments (logos): html text]

DMMPMM_drosophila

name	id	strand	Nb overlap columns	% aligned	Pearson correlation	Normalized cor	aligned col. motif rsrAAAGGGTTAars	aligned col. match
Kr	Kr	D	10	0.6667	0.967	0.645	..rAAAGGGTTA...	aAAaGGGTta
Total maches= 1								

[match table: html text]
[alignments (logos): html text]

IDMMPMM_drosophila

name	id	strand	Nb overlap columns	% aligned	Pearson correlation	Normalized cor	aligned col. motif rsrAAAGGGTTAars	aligned col. match
Kr	Kr	R	11	0.7333	0.969	0.711	.rrAAAGGGTTA...	YRAAAGGGTTA
Total maches= 1								

[match table: html text]
[alignments (logos): html text]

Predicted sites on test peaks

Distribution of sites

Nb of predicted sites per peak

Nb peaks with at least one predicted site: 1621 (27.00%)
Max nb of predicted sites per peak: 6
(dm3_chrX_3912477_3915237_+)
Mean nb of predicted sites per peak: 1.23

[view in genome browser : UCSC]
[sites: text BED (UCSC track)]
[distribution: text]
[sites per peak: text]
[enrichment: text graph]

positions_6nt_m1
positions_6nt_m1 RC
1074 sites
Predicted sites : positions_6nt_m1
Number of sites
Sequence position relative to peak center
Nb Predicted sites : positions_6nt_m1
Number of peaks
Number of sites

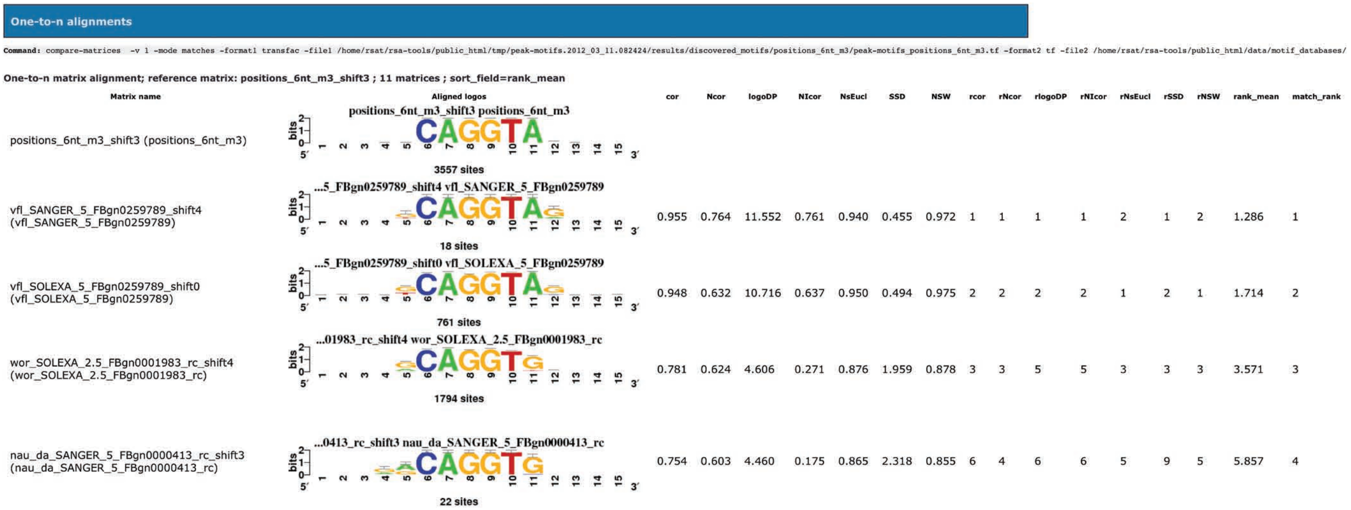
Figure 9 | Discovered motifs with motif comparisons. The snapshot displays the summary of the motif comparison step for a position-biased motif detected by position-analysis (motif identifier: positions_6nt_m1). The first row displays the logos (direct and reverse complementary) and colored consensus of the discovered motif, along with links to the matrix files. The next rows summarize the results of comparisons between the discovered motif and those provided as reference or found in the selected databases. The summary table displays up to three matches per comparison, indicating the identifier and names of the matched motif, the matching strand, the number of aligned columns and various similarity metrics. The last two columns display colored consensus restricted to the aligned positions: aligned positions in the discovered motif (the full consensus is indicated in the table header for comparison; nonaligned positions are replaced by dots), and aligned positions in the matched motif from the database. The complete list of matches with the detailed matching statistics can be accessed by clicking the link below the comparison summary ('Total matches = …'). Furthermore, tables showing all the correlation statistics, count matrices and logos alignments are available through links in the left panel. The last part of the summary per motif contains information about predicted sites. The plot 'Distribution of sites' shows the number of occurrences (*y* axis) per position (*x* axis) along the centered peak sequences. The next plot indicates the number of peaks (*y* axis) having 1, 2, 3, … sites (*x* axis), respectively. Some general statistics are also provided (peak coverage, mean number of sites per peak).

On the right hand side of this section, various links provide access to the primary results: the detailed list of significant words, their assembly into longer motifs and the resulting matrices in various formats (for more details, see the corresponding sections of the RSAT tutorials (http://tagc.univ-mrs.fr/rsa-tools/tutorials/).

Discovered motifs (with motif comparison)

Figure 9 shows one of the discovered motifs described in the previous section, compared with the motifs provided as references or found in the selected motif databases. For each database, the first three best matches are displayed (additional matches can be accessed by clicking the links 'match table' and 'alignment logos' on the right). The table summarizes information about the alignments: percentage of motifs aligned, Pearson correlation and normalized Pearson correlation. One should be aware that a high correlation coefficient can be misleading, because it might be obtained from a partial alignment of the matrices (e.g., the last column of the discovered matrix matching the first column of the reference motif). The goal of the normalized correlation is to avoid this effect by weighting the correlation according to the mutual coverage of the two compared motifs. The colored consensus indicates the aligned parts of each motif.

In this study case, the Krüppel motif is detected by its positional bias, but escapes detection by the oligo-analysis program. We interpret this as a consequence of the very large size of the peaks obtained from the GEO database. Consistently, when peaks are trimmed to the 200 central-most base pairs (100 bp on each side), the Krüppel motif appears as significantly overrepresented (it is detected by the oligo-analysis program) but its positional bias is not detected anymore, because sites are dispersed along the entire width of the trimmed peaks.



One-to-n alignments

```
Command: compare-matrices  -v 1 -mode matches -format1 transfac -file1 /home/rsat/rsa-tools/public_html/tmp/peak-motifs.2012_03_11.082424/results/discovered_motifs/positions_6nt_m3/peak-motifs_positions_6nt_m3.tf -format2 tf -file2 /home/rsat/rsa-tools/public_html/data/motif_databases/
```

One-to-n matrix alignment; reference matrix: positions_6nt_m3_shift3 ; 11 matrices ; sort_field=rank_mean

Matrix name	Aligned logos	cor	Ncor	logoDP	NIcor	NsEucl	SSD	NSW	rcor	rNcor	rlogoDP	rNIcor	rNsEucl	rSSD	rNSW	rank_mean	match_rank
positions_6nt_m3_shift3 (positions_6nt_m3)	positions_6nt_m3_shift3 positions_6nt_m3 3557 sites																
vfl_SANGER_5_FBgn0259789_shift4 (vfl_SANGER_5_FBgn0259789)	...5_FBgn0259789_shift4 vfl_SANGER_5_FBgn0259789 18 sites	0.955	0.764	11.552	0.761	0.940	0.455	0.972	1	1	1	1	2	1	2	1.286	1
vfl_SOLEXA_5_FBgn0259789_shift0 (vfl_SOLEXA_5_FBgn0259789)	...5_FBgn0259789_shift0 vfl_SOLEXA_5_FBgn0259789 761 sites	0.948	0.632	10.716	0.637	0.950	0.494	0.975	2	2	2	2	1	2	1	1.714	2
wor_SOLEXA_2.5_FBgn0001983_rc_shift4 (wor_SOLEXA_2.5_FBgn0001983_rc)	...01983_rc_shift4 wor_SOLEXA_2.5_FBgn0001983_rc 1794 sites	0.781	0.624	4.606	0.271	0.876	1.959	0.878	3	3	5	5	3	3	3	3.571	3
nau_da_SANGER_5_FBgn0000413_rc_shift3 (nau_da_SANGER_5_FBgn0000413_rc)	...0413_rc_shift3 nau_da_SANGER_5_FBgn0000413_rc 22 sites	0.754	0.603	4.460	0.175	0.865	2.318	0.855	6	4	6	6	5	9	5	5.857	4

Figure 10 | Motif comparisons. Logo alignments and scores of the matches between a motif discovered in Krüppel peaks and those annotated in the FlyFactorSurvey database[44]. To highlight partial correspondences, the logo of the discovered motif is aligned with the logos of the matched database motifs. The table contains multiple similarity metrics: cor, Pearson correlation; Ncor, normalized Pearson correlation; logoDP, dot product between the logo scores; NIcor, normalized correlation between information content values; NsEucl, normalized Euclidian distance; SSD, squared sum of deviations; NSW, normalized Sandelin-Wasserman. The next columns indicate the ranks of the previous columns (rcor, rank of the cor; rNcor, rank of Ncor and so on). The rank mean provides a robust measure of the overall similarity between two motifs.

In other cases (e.g., using mouse samples proposed as DEMO on the website), the correct motifs are consistently detected by several motif discovery approaches, indicating their robustness[9]. For this protocol, we deliberately chose a more difficult example to illustrate the variety of possible results, and to demonstrate the importance of combining different statistical criteria (global overrepresentation, position bias, local overrepresentation) in order to increase the sensitivity of motif discovery.

Motif comparisons with multiple logo alignments

A more detailed view of the alignments is obtained by clicking on the html link on the right hand side ('Alignments (logos)'), which displays a HTML page with 'one-to-*n*' alignments between one discovered motif and one or several database motifs (**Fig. 10**). It is very advisable to visually check these results, as the human eye turns out to be more accurate in detecting similarities than any measure. Several similarity metrics are indicated. Besides the similarity score, we provide, for each measure, the rank of the motif with respect to this measure, and compute the mean rank, which provides a robust indicator of the motif correspondences measured by the multiple scores.

Comparison with databases of known motifs can provide valuable clues to interpret additional motifs discovered besides the binding motif of the immunoprecipitated factor. With the Krüppel study case, the motif rgCAGGTAra discovered by position-analysis matches the *vfl* motif reported in FlyFactorSurvey (**Fig. 10**), bound by the Zelda transcription factor (*vfl* stands for *vielfaltig*, synonym for the gene *Zelda*). Zelda is a master regulator of the maternal-to-zygotic transition that occurs from 1 to 3 h after fertilization[40]. Zelda is known to cooperate with other transcription factors (Dorsal[41], STAT92E (ref. 42)), and the gain and loss of Zelda binding sites in peaks obtained from six anteroposterior factors (Hb, Bcd, Kr, Gt, Kni, Cad) in *D. melanogaster* and *D. yakuba* strongly correlates with the variation in changes in binding of all these factors[34]. The analysis of discovered motifs thus suggests that Krüppel may be yet another factor interacting with Zelda on the enhancers of its target genes.

Predicted sites on peaks

The bottom of the per-motif summary (**Fig. 9**) indicates the positions of predicted sites (left) and statistics on the number of sites per peak (right). The spatial distribution of predicted sites can be very informative for some motifs. In our study case, the various motifs have very different profiles: the motif corresponding to Krüppel (positions_6nt_m1) presents a very sharp peak with the maximum coinciding with the peak centers, and a deviation of about ± 100 bp. This adequately confirms that the peak centers are indeed enriched in Krüppel binding sites, as Krüppel was targeted in the experiment.

Some other motifs show a 'volcano-like' profile, with a high enrichment on each side of the peaks center. These might correspond to transcription factors that are cofactors of Krüppel. For example, several AT-rich motifs (e.g., oligos_6nt_mkv4_m1) were detected by the oligo-analysis program as strongly overrepresented in the peak sequences, yet their positional profile shows a clear avoidance in the middle of the peaks, where the Krüppel-binding sites show the highest concentration. Altogether, these observations suggest that the peaks contain a high concentration of well-centered Krüppel sites flanked by AT-rich motifs, which might correspond to Krüppel partners such as Hunchback.

The graph on the right at the bottom of **Figure 9** shows the distribution of predicted sites per peak, enhanced with some indicative statistics: first, the number of peaks with at least one site indicates the coverage of the peaks by the discovered motif. Second, the mean and maximum numbers of sites per peak are also indicated (along with the name of the peak with the maximum number of sites). A very large number of sites could point to a motif found in repetitive elements, or yet to low-complexity motifs, which should be considered with caution because they are particularly enriched in a subset of the peaks, and not representative for the site collection as a whole.

In short, despite the large dispersion of peak sizes in the initial data set, peak-motifs identifies the correct motif, and it reveals some heterogeneity in AT composition at the center of the peaks, as well as a consistent avoidance of several AT-rich motifs, one of them potentially corresponding to Hunchback.

Visualization in the UCSC Genome Browser

By clicking on the UCSC button on the right side of the distribution profiles, predicted sites can be uploaded to the UCSC Genome Browser, in order to visualize them in their genomic context and compare them with other annotation tracks (**Fig. 11**). Note that the peak regions can also be uploaded as a custom track in UCSC, by clicking on the UCSC button at the right of the sequence composition profile (**Fig. 5**).

Downloading the results for further analysis

The result is maintained on the web server for 3 d only. It is, however, very easy to download the complete result folder, by clicking the link 'Download all results (peak-motifs_archive.zip)' in the result summary box on the top of the report form.

Note that the direct links to UCSC only work when called from the web page on the RSAT server, and not from your local copy. To enable future analysis from your computer, input peaks and results are also exported as separate BED files, which can be stored on your computed and uploaded later as custom tracks in the UCSC genome browser. Refer to the UCSC manual (http://genome.ucsc.edu/training.html) to learn how to manage custom tracks.

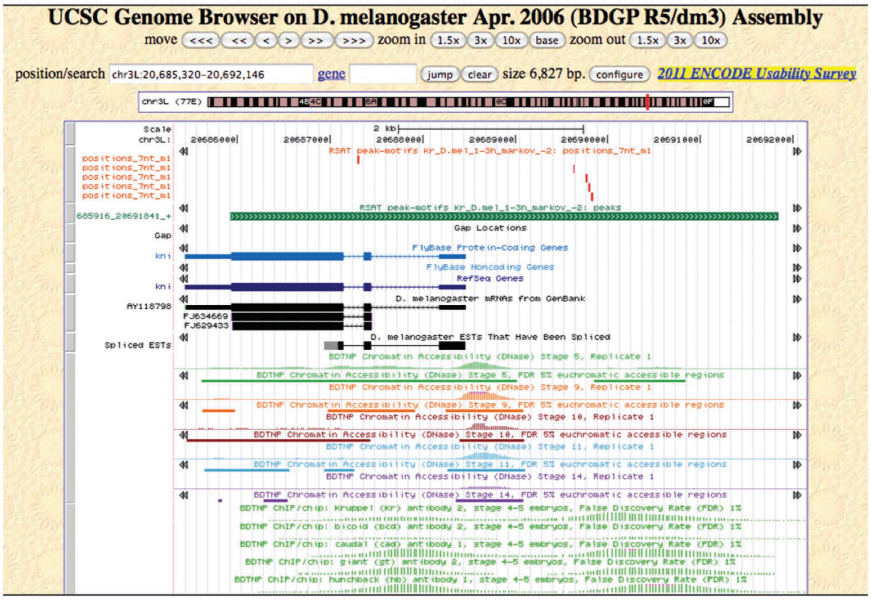
Interpretation of results for two additional study cases

Additional ChIP-seq study cases from vertebrates are accessible on the supporting website (http://rsat.bigre.ulb.ac.be/data/published_data/peak-motifs_Protocol_2012/). For each data set, we provide the peak sequences and the known motifs (for users wishing to rerun some analyses), as well as the result reports. These data sets were used in the original peak-motifs publication[9], which provides a complete discussion on the motifs found and their biological relevance.

Study case 2: motifs for interacting partners in Oct4 peaks. We will succinctly comment below the results obtained for the mouse Oct4 transcription factor, and guide their interpretation.

The peak lengths range from 39 bp to 839 bp, with a mean of 317 bp. These peak sizes are consistent with the usual selection of ChIP fragment sizes (~300 bp). In this case, the peak-calling was performed with an additional step to split longer peak regions into individual peaks (PeakSplitter[10]), which contributed to concentrate the binding signal in smaller regions, thereby producing better results for the motif discovery step. The sequence composition is characteristic of mouse, with an avoidance of the CpG dinucleotide (although less marked than in **Fig. 6**).

In this analysis, the four motif-discovery algorithms were used (oligo-analysis, local-words, position-analysis and dyad-analysis (see **Box 3** and **Fig. 12**)). There is some redundancy in the found motifs (e.g., local_words_6nt_m1, local_words_7nt_m1 and positions_7nt_m1 correspond to the same SOCT motif encoding the binding preference of the Sox2/Oct4 complex). This indicates the



UCSC Genome Browser on D. melanogaster Apr. 2006 (BDGP R5/dm3) Assembly

position/search chr3L:20,685,320-20,692,146 size 6,827 bp.

move <<< << < > >> >>>
zoom in 1.5x 3x 10x base
zoom out 1.5x 3x 10x
gene
jump
clear
configure
2011 ENCODE Usability Survey
chr3L (77E)
4B 4C
6A
8C
8F
Scale
chr3L:
2 kb
20686000
20687000
20688000
20689000
20690000
20691000
20692000
RSAT peak-motifs Kr_D.mel_1-3h_markov_-2: positions_7nt_m1
positions_7nt_m1
positions_7nt_m1
positions_7nt_m1
positions_7nt_m1
positions_7nt_m1
RSAT peak-motifs Kr_D.mel_1-3h_markov_-2: peaks
685916_20691841_+
Gap Locations
Gap
FlyBase Protein-Coding Genes
kni
FlyBase Noncoding Genes
RefSeq Genes
kni
D. melanogaster mRNAs from GenBank
AY118798
FJ634669
FJ629433
D. melanogaster ESTs That Have Been Spliced
Spliced ESTs
BDTNP Chromatin Accessibility (DNase) Stage 5, Replicate 1
BDTNP Chromatin Accessibility (DNase) Stage 5, FDR 5% euchromatic accessible regions
BDTNP Chromatin Accessibility (DNase) Stage 9, Replicate 1
BDTNP Chromatin Accessibility (DNase) Stage 9, FDR 5% euchromatic accessible regions
BDTNP Chromatin Accessibility (DNase) Stage 10, Replicate 1
BDTNP Chromatin Accessibility (DNase) Stage 10, FDR 5% euchromatic accessible regions
BDTNP Chromatin Accessibility (DNase) Stage 11, Replicate 1
BDTNP Chromatin Accessibility (DNase) Stage 11, FDR 5% euchromatic accessible regions
BDTNP Chromatin Accessibility (DNase) Stage 14, Replicate 1
BDTNP Chromatin Accessibility (DNase) Stage 14, FDR 5% euchromatic accessible regions
BDTNP ChIP/chip: Kruppel (Kr) antibody 2, stage 4-5 embryos, False Discovery Rate (FDR) 1%
BDTNP ChIP/chip: bicoid (bcd) antibody 2, stage 4-5 embryos, False Discovery Rate (FDR) 1%
BDTNP ChIP/chip: caudal (cad) antibody 1, stage 4-5 embryos, False Discovery Rate (FDR) 1%
BDTNP ChIP/chip: giant (gt) antibody 2, stage 4-5 embryos, False Discovery Rate (FDR) 1%
BDTNP ChIP/chip: hunchback (hb) antibody 1, stage 4-5 embryos, False Discovery Rate (FDR) 1%

Figure 11 | Predicted sites visualized in their genomic contexts on the UCSC genome browser. Positions of Krüppel-predicted sites are displayed in red, along with the corresponding peaks in green, in the light of relevant annotation tracks made available through the UCSC genome browser. Most of the information available for *Drosophila* has been generated by the ModENCODE consortium (http://www.modencode.org/).

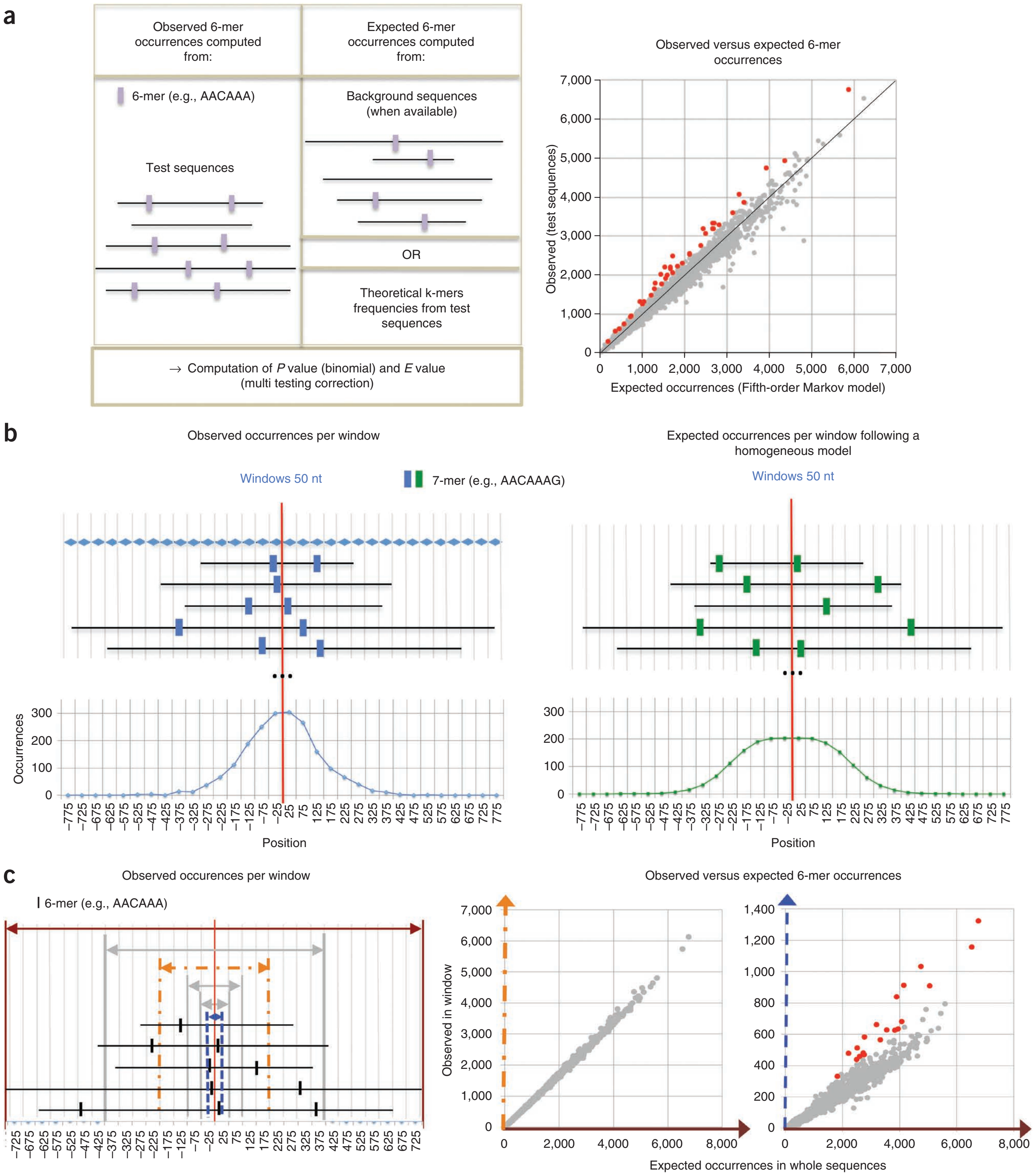


Figure 12 | Motif discovery approaches. Schematic representation of the criteria for detecting exceptional words. (**a**) Overrepresentation of words (oligo-analysis). Left, schematic view of the principle underlying the test of overrepresentation for a given word. Right, occurrences observed for each word in the test set (*y* axis) are compared with the occurrences expected according to the background model (*x* axis). Each dot represents a hexanucleotide (also called 6-mer). Hexanucleotides showing significant overrepresentation compared with the expectation are highlighted in red (binomial sig ≥ 10). (**b**) Positional bias (position-analysis). Left, sequences are aligned relative to the peak centers, and the occurrences of each word are counted in nonoverlapping windows of fixed width (indicated by gray vertical lines). The vertical red line represents the reference coordinate (i.e., peak centers). The blue diamonds indicate the positional windows in which word occurrences are counted. Right, positional distribution of word occurrences that would be expected under a homogeneous distribution. As peak sequences have varying widths, the number of sequences decreases with distance to peak centers, and the expected occurrences (green curve) decrease accordingly. (**c**) Local overrepresentation (local-word-analysis). Left, word occurrences are counted in windows of increasing widths centered on peak centers (indicated by the colored arrows). Right, these observed occurrences are compared with the occurrences that would be expected under a homogeneous model. In this example (Sox2 peaks[9]), the right-hand plot shows that the central 50-bp window (depicted with a dashed blue line) contains strongly overrepresented words (sig ≥ 5, highlighted in red), whereas a 400-bp central region (middle plot, yellow dotted and dashed line) does not show any significant local enrichment. The significant motifs of the central region (highlighted in red on the right-hand plot) correspond to different fragments of the Sox2 binding motif.

robustness of the result, as a given motif can be found by independent criteria (overrepresentation versus positional bias). The first motif (oligos_6nt_mkv4_m1) matches the two reference motifs of Oct4, but the logo alignment shows that the best-ranking match is the individual Oct4-binding motif (V$OCT_Q6, with consensus ATGyAAAt), whereas the second match (V$OCT4) actually represents the larger composite SOCT motif. In contrast, the motif local_words_6nt_m1 best matches this second V$OCT4 reference motif, suggesting that this motif is a SOCT motif rather than an individual Oct4 motif.

The third motif reported by the local-word-analysis program (local_words_7nt_m3, with consensus crTATGCGCATAyg) is notable because it does not show any substantial similarity with known motifs from the selected collections. Yet, the predicted sites for this motif accumulate within ± 100-bp central regions of the peaks, thereby suggesting that this motif may be biologically relevant. Further investigation actually indicated that this motif is an alternative Oct4 motif uncovered in recent studies[9].

Study case 3: differential analysis—tissue-specific motifs in p300 peaks. This third study case is proposed as a 'DEMO test vs ctrl' on the website, and it intends to illustrate the use of peak-motifs to perform differential analyses, in which two sequence data sets are provided, one as a foreground data set and the second as a control data set. The demo uses p300 ChIP-seq data sets obtained in various mouse embryonic tissues[43]. As the foreground data set, we use 1,000 peaks obtained in heart embryonic tissue, with sizes ranging from 283 bp to 2,636 bp (mean: 814 bp), whereas 1,000 peaks obtained in limb tissue serve as control set, with sizes ranging from 276 to 2,301 bp (mean 679 bp). Hence, we are looking for motifs that are specifically enriched in heart tissue, by comparison with other tissues. When the tool is used in the 'test versus control' mode, oligo-analysis options are adapted to return *k*-mers that are markedly overrepresented in the first data set with respect to the second one. As the differential analysis only applies to word-counting algorithms, we run peak-motifs using the oligo-analysis program only, with word length of 6 and 7.

Among the ten significant oligonucleotides returned, we can distinguish two distinct motifs: the first is a GATA-like motif, with consensus sequence GATAA (detected with various oligonucleotide lengths, e.g., oligos_6nt_vs_ctrl_m1, oligos_7nt_vs_ctrl_m2 and others), whereas the second motif has a consensus sequence GACAG (e.g., oligos_6nt_vs_ctrl_m3, oligos_7nt_vs_ctrl_m3 and others). Comparison with the JASPAR and PBM databases identifies the first motif as belonging to the GATA family of transcription factors. These transcription factors have very similar DNA-binding motifs (GATAA). Interestingly, GATA6 transcription factor is expressed in the myocardium, and has the Gene Ontology annotation 'cardiac muscle cell differentiation' (oligos_7nt_vs_ctrl_m1 matches Gata6_1 of the PBM database with normalized correlation coefficient of 0.74 (not included when running the demo, but results are available in **Supplementary Table 1**)). The second motif (GACAG) has no match in the JASPAR or PBM database; however, compared with the TRANSFAC database, this motif is highly similar to the DNA-binding motif of MEIS1 (V$MEIS1_01, Ncor = 0.75), a homeobox transcription factor involved in angiogenesis, specifically expressed in lung, heart and brain, but not in limb.

Hence, we have identified two motifs that correspond to transcription factors believed to be involved in heart formation, but not limb development. In contrast, this differential analysis discards the motifs corresponding to general transcription factors including that of the ubiquitous p300 cofactor Sp1, known to be active in both[9]. It must be noted that this differential analysis is asymmetric: by switching the test and control sets above, one would obtain motifs specifically enriched in p300 peaks found in the limb, relative to p300 peaks in the heart.

Note: Supplementary information is available in the online version of the paper.

ACKNOWLEDGMENTS This work was supported by the Alexander von Humboldt foundation to M.T.-C.; the Agence Nationale de Recherche (ANR) partner of the ERASysBio+ initiative supported under the EU ERA-NET Plus scheme in FP7 to C.H.; ANR Young Researchers Grant 'CardiHox' to C.H.; the Belgian Program on Interuniversity Attraction Poles, initiated by the Belgian Federal Science Policy Office (project P6/25 (BioMaGNet)); EU-funded Cooperation in Science and Technology (COST) action (BM1006 'SEQAHEAD—Next-Generation Sequencing Data Analysis Network'); FP7 MICROME Collaborative Project (Microbial genomics and bio-informatics', contract number 222886-2). We acknowledge the colleagues who helped to install and maintain the RSAT Web servers: R. Leplae (ULB, Belgium), R. Zayas-Lagunas (UNAM, Mexico), E. Bongcam-Rudloff (Uppsala, Sweden), F.-X. Théodule (Aix Marseille Université, France), P. Vincens (Ecole Normale Supérieure, France) and F. Joubert (Pretoria, South Africa).

AUTHOR CONTRIBUTIONS J.v.H., M.T.-C. and M.D. initiated and developed the peak-motifs software tool. E.D., C.H. and D.T. contributed to improve the tool and analyzed the study case for this protocol. All authors edited the manuscript.

COMPETING FINANCIAL INTERESTS The authors declare no competing financial interests.

Published online at http://www.nature.com/doifinder/10.1038/nprot.2012.088. Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Robertson, G. *et al.* Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat. Methods* **4**, 651–657 (2007).
2. Johnson, D.S., Mortazavi, A., Myers, R.M. & Wold, B. Genome-wide mapping of *in vivo* protein-DNA interactions. *Science* **316**, 1497–1502 (2007).
3. Pepke, S., Wold, B. & Mortazavi, A. Computation for ChIP-seq and RNA-seq studies. *Nat. Methods* **6**, S22–S32 (2009).
4. Boeva, V. *et al. De novo* motif identification improves the accuracy of predicting transcription factor binding sites in ChIP-Seq data analysis. *Nucleic Acids Res.* **38**, e126 (2010).
5. Machanick, P. & Bailey, T.L. MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics* **27**, 1696 (2011).
6. Bailey, T.L. DREME: Motif discovery in transcription factor ChIP-seq data. *Bioinformatics* **27**, 1653–1659 (2011).
7. Rusk, N. Focus on next-generation sequencing data analysis. *Nat. Methods* **6**, S1 (2009).
8. McPherson, J.D. Next-generation gap. *Nat. Methods* **6**, S2–S5 (2009).
9. Thomas-Chollier, M. *et al.* RSAT peak-motifs: motif analysis in full-size ChIP-seq datasets. *Nucleic Acids Res.* **40**, e31 (2012).
10. Salmon-Divon, M., Dvinge, H., Tammoja, K. & Bertone, P. PeakAnalyzer: genome-wide annotation of chromatin binding and modification loci. *BMC Bioinformatics* **11**, 415 (2010).
11. Portales-Casamar, E. *et al.* JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **38**, D105–D110 (2010).

12. Wingender, E. The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation. *Brief Bioinform.* **9**, 326–332 (2008).
13. Gama-Castro, S. *et al.* RegulonDB version 7.0: transcriptional regulation of *Escherichia coli* K-12 integrated within genetic sensory response units (Gensor Units). *Nucleic Acids Res.* **39**, D98–D105 (2011).
14. Medina-Rivera, A. *et al.* Theoretical and empirical quality assessment of transcription factor-binding motifs. *Nucleic Acids Res.* **39**, 808–824 (2011).
15. Chen, X. *et al.* Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* **133**, 1106–1117 (2008).
16. Cline, M.S. *et al.* Integration of biological networks and gene expression data using Cytoscape. *Nat. Protoc.* **2**, 2366–2382 (2007).
17. Fujita, P.A. *et al.* The UCSC Genome Browser database: update 2011. *Nucleic Acids Res.* **39**, D876–D882 (2011).
18. Fullwood, M.J., Wei, C.L., Liu, E.T. & Ruan, Y. Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses. *Genome Res.* **19**, 521–532 (2009).
19. Lee, T.I. *et al.* Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* **298**, 799–804 (2002).
20. Sanford, J.R. *et al.* Splicing factor SFRS1 recognizes a functionally diverse landscape of RNA transcripts. *Genome Res.* **19**, 381–394 (2009).
21. van Helden, J., del Olmo, M. & Perez-Ortin, J.E. Statistical analysis of yeast genomic downstream sequences reveals putative polyadenylation signals. *Nucleic Acids Res.* **28**, 1000–1010 (2000).
22. Sand, O., Thomas-Chollier, M., Vervisch, E. & van Helden, J. Analyzing multiple data sets by interconnecting RSAT programs via SOAP Web services: an example with ChIP-chip data. *Nat. Protoc.* **3**, 1604–1615 (2008).
23. van Helden, J., Andre, B. & Collado-Vides, J. Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J. Mol. Biol.* **281**, 827–842 (1998).
24. van Helden, J., Rios, A.F. & Collado-Vides, J. Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. *Nucleic Acids Res.* **28**, 1808–1818 (2000).
25. Thomas-Chollier, M. *et al.* RSAT 2011: regulatory sequence analysis tools. *Nucleic Acids Res.* **39**, W86–W91 (2011).
26. Kulakovskiy, I.V., Boeva, V.A., Favorov, A.V. & Makeev, V.J. Deep and wide digging for binding motifs in ChIP-Seq data. *Bioinformatics* **26**, 2622–2623 (2010).
27. Agius, P., Arvey, A., Chang, W., Noble, W.S. & Leslie, C. High resolution models of transcription factor-DNA affinities improve *in vitro* and *in vivo* binding predictions. *PLoS Comput. Biol.* **6**, e1000916 (2010).
28. Mercier, E. *et al.* An integrated pipeline for the genome-wide analysis of transcription factor binding sites from ChIP-Seq. *Plos ONE* **6**, e16432 (2011).
29. Kuttippurathu, L. *et al.* CompleteMOTIFs: DNA motif discovery platform for transcription factor binding experiments. *Bioinformatics* **27**, 715–717 (2010).
30. van Heeringen, S.J. & Veenstra, G.J. GimmeMotifs: a *de novo* motif prediction pipeline for ChIP-sequencing experiments. *Bioinformatics* **27**, 270–271 (2011).
31. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
32. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
33. Sand, O., Turatsinze, J.V. & vanHelden, J. Evaluating the prediction of *cis*-acting regulatory elements in genome sequences. in *Modern Genome Annotation: The BioSapiens Network* (eds. Frishman, D. & Valencia, A.) (Springer, 2008).
34. Bradley, R.K. *et al.* Binding site turnover produces pervasive quantitative changes in transcription factor binding between closely related *Drosophila* species. *PLoS Biol.* **8**, e1000343 (2010).
35. Barrett, T. *et al.* NCBI GEO: archive for functional genomics data sets—10 years on. *Nucleic Acids Res.* **39**, D1005–D1010 (2011).
36. Goecks, J., Nekrutenko, A. & Taylor, J. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.* **11**, R86 (2010).
37. Valouev, A. *et al.* Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nat. Methods* **5**, 829–834 (2008).
38. Bergman, C.M., Carlson, J.W. & Celniker, S.E. *Drosophila* DNase I footprint database: a systematic genome annotation of transcription factor binding sites in the fruitfly, *Drosophila melanogaster*. *Bioinformatics* **21**, 1747–1749 (2005).
39. Flicek, P. *et al.* Ensembl 2011. *Nucleic Acids Res.* **39**, D800–D806 (2011).
40. Harrison, M.M., Li, X.Y., Kaplan, T., Botchan, M.R. & Eisen, M.B. Zelda binding in the early *Drosophila melanogaster* embryo marks regions subsequently activated at the maternal-to-zygotic transition. *PLoS Genet.* **7**, e1002266 (2011).
41. Kanodia, J.S. *et al.* Pattern formation by graded and uniform signals in the early *Drosophila* embryo. *Biophys. J.* **102**, 427–433 (2012).
42. Tsurumi, A. *et al.* STAT is an essential activator of the zygotic genome in the early *Drosophila* embryo. *PLoS Genet.* **7**, e1002086 (2011).
43. Blow, M.J. *et al.* ChIP-Seq identification of weakly conserved heart enhancers. *Nat. Genet.* **42**, 806–810 (2010).
44. Zhu, L.J. *et al.* FlyFactorSurvey: a database of *Drosophila* transcription factor binding specificities determined using the bacterial one-hybrid system. *Nucleic Acids Res.* **39**, D111–D117 (2011).
45. Defrance, M., Janky, R., Sand, O. & van Helden, J. Using RSAT oligo-analysis and dyad-analysis tools to discover regulatory signals in nucleic sequences. *Nat. Protoc.* **3**, 1589–1603 (2008).
46. Turatsinze, J.V., Thomas-Chollier, M., Defrance, M. & van Helden, J. Using RSAT to scan genome sequences for transcription factor binding sites and cis-regulatory modules. *Nat. Protoc.* **3**, 1578–1588 (2008).