Published online 24 October 2017 *Nucleic Acids Research, 2018, Vol. 46, Database issue* ***D175–D180***
doi: 10.1093/nar/gkx995

MGA repository: a curated data resource for ChIP-seq and other genome annotated data

René Dréos[1], Giovanna Ambrosini[1,2], Romain Groux[1], Rouayda Cavin Périer[1] and Philipp Bucher[1,2,*]

[1]Swiss Institute of Bioinformatics (SIB), CH-1015 Lausanne, Switzerland and [2]Swiss Institute for Experimental Cancer Research (ISREC), School of Life Sciences, Swiss Federal Institute of Technology (EPFL), CH-1015 Lausanne, Switzerland

Received August 15, 2017; Revised October 10, 2017; Editorial Decision October 11, 2017; Accepted October 20, 2017

ABSTRACT

The Mass Genome Annotation (MGA) repository is a resource designed to store published next generation sequencing data and other genome annotation data (such as gene start sites, SNPs, etc.) in a completely standardised format. Each sample has undergone local processing in order the meet the strict MGA format requirements. The original data source, the reformatting procedure and the biological characteristics of the samples are described in an accompanying documentation file manually edited by data curators. 10 model organisms are currently represented: *Homo sapiens*, *Mus musculus*, *Danio rerio*, *Drosophila melanogaster*, *Apis mellifera*, *Caenorhabditis elegans*, *Arabidopsis thaliana*, *Zea mays*, *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*. As of today, the resource contains over 24 000 samples. In conjunction with other tools developed by our group (the ChIP-Seq and SSA servers), it allows users to carry out a great variety of analysis task with MGA samples, such as making aggregation plots and heat maps for selected genomic regions, finding peak regions, generating custom tracks for visualizing genomic features in a UCSC genome browser window, or downloading chromatin data in a table format suitable for local processing with more advanced statistical analysis software such as R. Home page: http://ccg.vital-it.ch/mga/.

INTRODUCTION

Next-generation sequencing (NGS) technologies such as ChIP-seq have become increasingly common and invaluable tools for scientists in many fields. During the past ten years, as sequencing cost have decreased and NGS has increased in popularity, primary data archives, such as GEO (1) and ArrayExpress (2), have grown exponentially in size. Archiving raw sequencing data, however, is not sufficient to make it immediately usable by the scientific community. Over the years, we and others have developed user-friendly tools that facilitate the analysis of such complex data (3–6). Nonetheless, the majority of these tools do not provide access to preprocessed published data, leaving this time-consuming and resource-hungry task to individual scientists interested in analyzing them. This is in effect a barrier to many scientists that are willing to analyse public data but do not have time for or are not familiar with the pre-processing tools and methods.

Here, we present a database of NGS-data and a variety of other genome annotations in a completely standardized format along with machine-readable sample annotations (metadata). The MGA database does not store raw sequence files but instead lists of base positions in the genome corresponding, for instance, to the 5'-end of a mapped sequence read from a ChIP-seq experiment. These are stored in text files with a completely standardized format. Data is from published studies of various NGS technologies such as ChIP-seq, MNase-seq, GRO-seq, etc. Moreover, the MGA database contains genome annotations derived from manual curation efforts or computational integration of information from multiple primary sources. Those include transcription start sites, intron–exon boundaries, natural variants, sequence derived feature such as transcription factor putative binding sites, short repeats, etc. The database covers 10 species for a total amount of >24 000 samples. All samples are manually curated and manually annotated. The tight interconnection with the tools developed by our group, namely the ChIP-Seq (3) and Signal Search Analysis, SSA (7), servers, makes it a powerful resource for the analysis and interpretation of public data. To our knowledge this is the first resource that offers a large collection of NGS data along with annotation and sequence-derived features.

[*]To whom correspondence should be addressed. Tel: +41 21 6930956; Fax: +41 21 693 1850; Email: philipp.bucher@epfl.ch

DATABASE CONTENT

Data types, format and accessibility

The MGA repository currently contains NGS data mapped to a single, so-called primary assembly of the corresponding species, except for human and fruit fly where data is split over two or more assemblies (hg18, hg19 and hg38 for human, dm3 and dm6 for fruit fly). Each sample in the database belongs to one of the seven obligatory data types:

1. ChIP-seq: mapped sequence reads
2. ChIP-seq-peak: peak regions provided by the authors of the data
3. Transcription Profiling: RNA 5'-ends mapped with techniques such as CAGE, GRO-cap, etc.
4. DNase FAIRE etc.
5. DNA Methylation: various assays such as bisulfite sequencing, MBDCap, etc.
6. Sequence-derived: PWM matches, Natural Variants, Conservation scores, etc.
7. Genome Annotation: transcription start sites, transcription end sites, intron–exon boundaries.

An overview of the current contents of the MGA repository is shown in Figure 1. Currently, it contains more than 24'000 samples, with particular importance given to samples belonging to milestone projects such as ENCODE (8), RoadMap (9) and FANTOM5 (10). The large sample size of these projects is reflected in the database content with one quarter of samples belonging to ChIP-seq type (Figure 1A) and two thirds to human (Figure 1B).

The standard format used for all samples is called SGA (Simple Genome Annotation) (3). SGA is a tab delimited text file format providing information about the position of a feature in the genome, but unlike other formats such as BED or GFF, it represents features in the genome as single positions. We have chosen a single position format, because all the programs for DNA sequence and functional genomics data analysis use single positions as input. For example, in ChIP-seq experiments, the position given in the SGA file corresponds to the 5′-end of a mapped read. This convention does not compromise the accuracy of the annotation since the reads 3′-ends are defined by the sequencing machine and, *a priori*, have no biological meaning. Instead, the chipped DNA fragment ends are defined by the 5′-end positions of reads mapping on opposite strands. Columns are chromosome name (as RefSeq ID), feature name, position (as a single base), strand and counts. Reads that map to the same chromosome location are compacted into a single SGA line with the counts field corresponding to their number. Additional columns might be present for annotation (e.g. gene names for promoter lists) or peak scores.

Data from the same study are organized in series, as in GEO (1). A human-curated documentation file in HTML format annotates all data files belonging to the same series. Each series contains two additional machine-readable text files, one providing information about the series, the other one about individual samples. The first file contains the complete path to the data directory, a title, a literature or database reference in textual form plus, when available, GEO, ArrayExpress and/or PubMed IDs. The sample description file is a tab-delimited table with lines corresponding to samples. The first field of each line contains the name of the corresponding SGA file and is followed by fields containing a sample description and the feature name used in the SGA file. An additional field indicates whether the feature is 'oriented'. An SGA file is 'oriented' if the strand field is occupied by + and – signs; 'unoriented' SGA files have the strand field filled with zeros. GEO or ArrayExpress sample IDs are also included if applicable.

The MGA database is accessible via an anonymous FTP server (ftp://ccg.vital-it.ch/mga/) for download in SGA format.

Pre-processing pipeline and reproducibility

When processing published data, particular care is taken to reproduce mapped results as close as possible to the published methods. For this reason, whenever the authors provide mapped or processed data (e.g. peak regions of a ChIP-seq experiment), they are directly used and converted into SGA. When these are not available, a general pipeline is performed in order to map the reads to the corresponding genome assembly (often the same assembly used in the publication) using, when possible, the published mapping parameters. The pipeline used is for most part standardized and uses the aligners bowtie (11) or bowtie2 (for reads longer than 50 bp) (12) and the conversion utilities view and sort from SAMtools (13), bamToBed from BEDTools (14) and bed2sga from ChIP-Seq (3). Often samples need custom-made pipelines. Paired-end sequencing is treated as such allowing only matches that are within the expected fragment length. In such cases, the corresponding SGA file is often presented as unoriented and the midpoint of the fragment is taken as mapped position. When present, adapters or barcodes are trimmed before mapping. Sample technical replicas are often merged into a single file to facilitate the downstream analysis. All steps performed and file manipulations are then described in the corresponding documentation file.

The MGA repository contains any type of potentially useful genomic data that can be represented by the SGA format. Some special data sets have undergone substantial modifications relative to the original file, either to reduce the size or to make the representation compatible with an integer-based single-position format. The MGA offers compacted versions of the phastCons (15) and phyloP (16) tracks from the UCSC Genome Browser database to speed up the analysis of genomic conservation at the expense of some precision. When public peak lists are converted into SGA format the midpoint of the peak region or a designated 'summit' position is used as reference position in the SGA file, which is also unoriented. The purpose of offering published peak lists in addition to read alignment files, is to to help users to reproduce published results via our web servers and to generated similar plots with other data from the MGA repository using the same peak lists.

For the sake of reproducible computational research, the scripts used for generating the SGA files from the source data are provided for each sample on the FTP server and all third party software called by the scripts is identified by the download URL and version number. Any other custom-

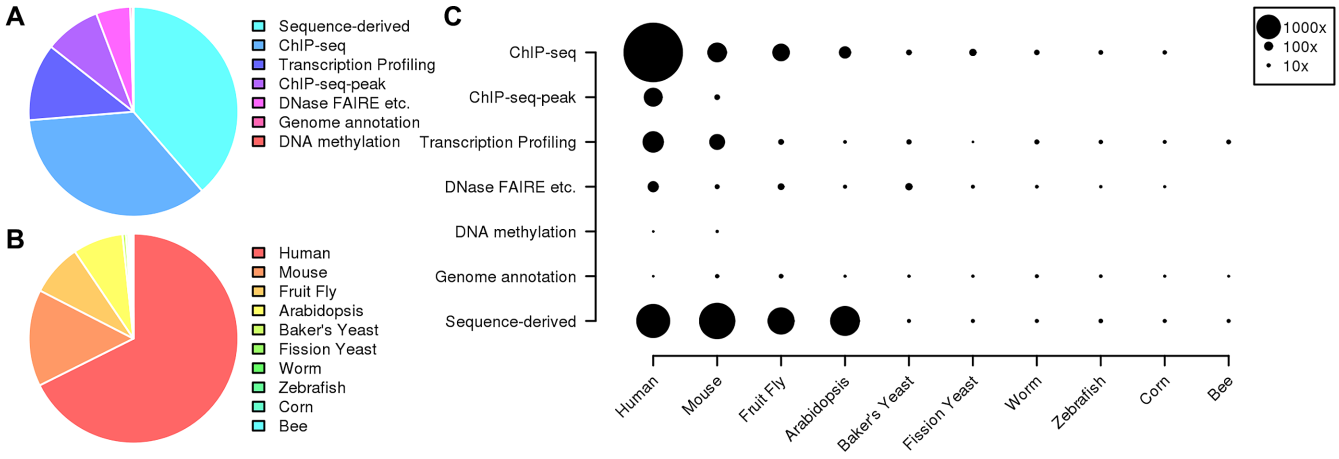

Figure 1. Content of the MGA repository. (**A**) Proportion of samples in the database grouped by type. (**B**) Proportion of samples grouped by organism. Assemblies belonging to the same organism are merged together. (**C**) Samples numbers stratified by type and organism. Dot areas are proportional to the total number of samples in that category. The corresponding numbers can be found in a weakly updated table posted on the MGA home page at http://ccg.vital-it.ch/mga.

made script that might be needed (often written in Perl) is provided in the same directory.

HOW TO USE THE MGA DATABASE

Query the database

The primary way to search for samples or experiments in the MGA database is through the HTML query page. It is a free text search that searches for assembly-specific samples. The text area dynamically shows samples names that match the input text as it is written. Upon hitting the search button, samples and series matching the search criteria are shown in a list-like fashion. Links to the series documentation page and source publication are provided next to the series name. For each sample, action buttons are provided for exporting the data or uploading the data to various analysis tools from the ChIP-seq and SSA servers. The precise function of these action buttons will be described further below.

Export data in various formats

Data in the MGA repository can be exported in several formats such as BED, WIG and FASTA. These files are not permanently stored on the server but are dynamically generated when they are needed. Efforts have been made to reduce the generation time by implementing computationally expensive conversion tools in C rather than Perl. In the query output page, the sample descriptions are individually hyperlinked to a 'data hub', where users can upload the sample to a multitude of analysis tools or export the data in various formats. The series headers are linked to the series documentation page, which contains additional information on samples, and the original publication (if available).Via the data hub, it is also possible to lift over the data to other genome assemblies of the same species or across closely related species (e.g. human to mouse). For small data sets, e.g. peak lists, a menu appears for extracting sequences around them in FASTA format.

Two downstream tools, both part of the ChIP-Seq suite, are particularly useful for converting data in other formats. Both of them are reachable via the 'Analysis tools' menu of the data hub. The first one is named ChIP-Track and is used to generate UCSC Genome Browser tracks in WIG and BIGWIG formats for the whole genome or selected regions. Users can modify several parameters such as track colour, resolution and smoothing window. In the output page of this tool scientists can download the data in WIG format and, via a hyperlink, directly upload it to the UCSC Genome Browser for data visualisation. The tool is primarily intended to offer users the possibility to make publication-ready genome-browser snapshots displaying data from the MGA repository together with user-supplied tracks.

The second tool for data download is called ChIP-Extract. It enables computational biologists to extract relevant data from the MGA repository in table format for downstream processing with other tools (e.g. R software). The output is a matrix with each row representing a genomic region around a so-called reference feature, and each column a distance range relative to the reference position. Each cell then contains the number of instances of a so-called target feature that are found at a particular distance from the Reference. Examples of Reference features can be peak centers from a peak list, transcription start sites (TSS) etc. Target features are for example reads from a MNase-seq experiment indicating nucleosome positions. In addition to a tab-delimited text file, the ChIP-Extract server returns a graphical representation of the data as a heatmap (see Figure 2A for such an example). Note that ChIP-Extract is one of the programs from the ChIP-Seq server that requires two data samples (reference and target) as input. The reference and target features can be selected one after the other via the MGA query page by using the action buttons 'Set Reference' and 'Set Target'. Once both data sets are selected, the action button 'ChIP-Extract' sends the data to the corresponding program. The other ChIP-Seq server application that takes two input files is ChIP-cor. It generates a so-called aggregation plot (17), as shown in Figure 2B and C.

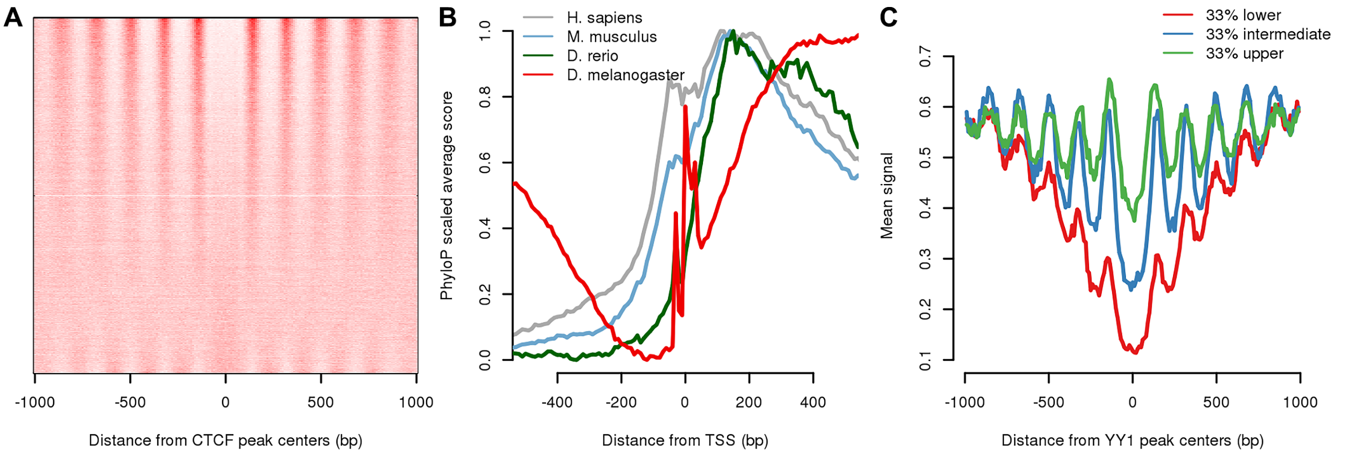

Figure 2. Examples of MGA data analysis. (**A**) Nucleosome organization for the lymphoblastoid cell line GM12878 around CTCF sites from the ENCODE 'Uniform TFBS' series. This plot was done using ChIP-Extract, sorting the CTCF sites according to similarity with the overall nucleosome pattern. (**B**) Distribution of conservation scores from PhyloP around TSS from the EPDnew database. *D. melanogaster* shows a distinctive distribution with sharp peaks corresponding to the TSS, TATA-box and DPE element. This plot was done using ChIP-Cor. (**C**) Example of reproducing a published figure (Figure 5A from (18)) showing the nucleosome organization around promoter-associated YY1 peaks stratified by YY1 binding strength evaluated as the number of YY1 sequencing reads around YY1 peaks. This plot was done using ChIP-Cor and ChIP-Extract. Detailed descriptions how to reproduce this figure can be found in Supplementary File 1.

Examples

Here we will outline a few examples of real-life non-trivial issues that might arise during an experiment analysis. The examples are not exhaustive and are intended to emphasise the tools' flexibility. A longer list of examples and analysis pipelines is resented in the tutorial section of the ChIP-Seq web site.

Three examples of the use of MGA samples are shown in Figure 2 whereas a detailed description of how to reproduce them can be found in Supplementary Data. Panel A shows the output of ChIP-Extract. Here, a peak list representing CTCF binding sites in the lymphoblastoid cell line GM12878 from the ENCODE series 'Uniform TFBS' is selected as Reference feature and used to study the nucleosome distribution around them in the same cell line. As expected and previously reported (18), the binding of CTCF to DNA is able to generate a region of tightly organized nucleosomes called nucleosome array. This figure has been copy-pasted from the ChIP-Extract output page.

Panel B shows the conservation scores around TSS coordinates from human, mouse, zebra fish and fly taken from the EPDnew database (19). The most striking difference revealed by this analysis is the very low level of sequence conservation in fly promoters immediately upstream of the TSS as compared to the vertebrate species. Note further the two sharp peaks in fly promoters, located ∼30 bp upstream and downstream from the TSS and corresponding to two well-known core promoter elements, the TATA-box and the DPE element. These results were generated with the ChIP-Cor tool from the ChIP-Seq server. The numerical results for the different species were individually downloaded to disk and combined in one plot with R.

The third example (Figure 2C) shows how to reproduce published results (figure 5A from (18)) using the ChIP-Seq tools. The Figure shows that promoter-associated YY1 sites that are strongly bound by the protein according to ChIP-seq are flanked by better-positioned nucleosomes as compared to weakly bound sites. The procedure to generate this figure required a few consecutive steps to first select YY1 peaks that are near promoters (this can be done using ChIP-Cor) and then to extract the nucleosome distribution and the YY1 reads around these peaks (both of them done using ChIP-Extract).

DISCUSSION

Here, we have presented a novel resource of published NGS and other genome annotation data, offered to the user in a common format and conforming to stringent data representation standards. The interconnection with other resources developed by our group for NGS and motif analysis allows scientists to investigate published data with few mouse clicks. The tools' flexibility makes it possible to perform innovative analyses and to gain new knowledge from published and unpublished data. These characteristics make this resource unique among the other existing databases.

Comparison with other resources

Table 1 shows a comparison between the MGA resource and other well-established NGS databases. Although all resources cover similar data types, the MGA stores unique genomic features such as transcription start sites, SNPs, conservation scores, etc. As shown in Figure 2B and C, given the flexibility of the ChIP-Seq tools, it is possible to perform similar types of analyses with different data types such as, for example, MNase data or conservation scores. Advanced users have also the possibility to download data in tabular format and perform further analyses using external software such as R. Other tools that have access to the MGA samples are used to download data in FASTA format and to lift over genomic coordinates to other assemblies. Moreover, the SSA server offers several tools for motif analysis expanding the possibilities for data interpretation.

Table 1. Comparison with other resources

	Cistrome	ChIP-Atlas	GeneProf	MGA
Organism	Hs, Mm [a]	Hs, Mm, Dm, Ce, Sc [a]	Hs, Mm, Dr, Ce, Dm, Sc, At, Gg, Ss, Os [a]	Hs, Mm, Dr, Dm, Ce, Ma, At, Zm, Sc, Sp [a]
Exp. Assays, genomic features	ChIP, DNase, ATAC	ChIP, DNase, MNase	ChIP, RNA, MNase	ChIP, DNase, ATAC, TSS, DNA-met, annotation, conservation, variation.
# of Samples	23’319	53’867	13’423	24’344
Downlaods, Export (format)	Signal (bigWig) Peaks (bed)	Signal (bigWig) Peaks (bed)	Signal (wig) Peaks (bed) Report (pdf)	Read-alignments (sga, bed) Peaks (sga, bed) Signal (bigwig) DNA sequence (fasta) Heatmaps (text table)
Query interface	Menu driven, free text-based	By sample ID (SRX)	Menu driven, free text-based	Menu-driven, free text-based
Lift-over	No	No	No	Yes
QC report	Yes	No	Yes	No
Metadata annotation	Yes	Yes	Yes	Yes
Visualization	WashU browser UCSC browser	IGV	Internal	UCSC browser
Integrated Analysis tools	Limited, only for peak files	Target genes, colocalisation, in-silico ChIP	Examine experiment analysis history	APs, Heatmaps, peak finder, DNA motif analysis

a) Note: Hs, *H. sapiens*; Mm: *M. musculus*; Dm: *D. melanogaster*; Dr: *D. rerio*; Ce: *C. elegans*; At: *A. thaliana*; Zm: *Z. mays*; Or: *O. sativa*; Sc: *S. cerevisiae*; Sp: *S. pombe*; Gg: *G. gallus*; Ss: *S. scrofa*.

Uniquely among the databases, MGA provides access to read alignment files. Although MGA does not always provide peak files, a fast and efficient peak finding tool is present in the ChIP-Seq server. ChIP-Peak accepts read alignments in SGA format and returns peak region in SGA, BED and FASTA formats. Users interested in using MGA data with other peak callers, such as MACS or MACS2 (20), can download samples data in BED format via the data hub page or, if interested in batch analysis of many samples, by downloading SGA files from the FTP server and converting them to BED using the sga2bed tool (a C script part of the ChIP-Seq toolbox (3)). (Instructions how to export data from MGA in BED format are posted on the MGA home page). Moreover, storing read alignment files allows scientists to generate UCSC tracks on the fly with the desired parameters without relying on pre-computed files with defined resolution.

Limitations and future extensions

Over the next few years we will continue to add samples to the database as soon as seminal studies are published and following requests from users. Moreover, we plan to cover more model organisms. High priority will be given to chicken (*Gallus gallus*), macaque (*Macaca mulatta*) and Xenopus (*Xenopus tropicalis*). In a second phase we plan to add rat (*Rattus norvegicus*) and rice (*Oryza sativa*). Further improving the query mechanisms is also on our agenda. By imposing a controlled vocabulary for all cell type, tissues, ChIP-seq targets and experimental assays through manual annotation efforts, we hope to be able to offer users the possibility to search for samples of interest in a more systematic manner.

DATABASE AVAILABILITY

The MGA repository is freely accessible without need for preregistration. Web-based access is provided via the MGA web site at http://ccg.vital-it.ch/mga/. Data files can be downloaded via FTP from ftp://ccg.vital-it.ch/mga/.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

ACKNOWLEDGEMENTS

Computationally intensive tasks were performed at the Vital-IT high-performance computing center of the Swiss Institute of Bioinformatics.

FUNDING

Swiss Government via a grant to the Swiss Institute of Bioinformatics (to R.D. and R.G.). Funding for open access charge: Swiss Government.

Conflict of interest statement. None declared.

REFERENCES

1. Barrett,T., Troup,D.B., Wilhite,S.E., Ledoux,P., Evangelista,C., Kim,I.F., Tomashevsky,M., Marshall,K.A., Phillippy,K.H., Sherman,P.M. *et al.* (2011) NCBI GEO: archive for functional genomics data sets—10 years on. *Nucleic Acids Res.*, **39**, D1005–D1010.
2. Rustici,G., Kolesnikov,N., Brandizi,M., Burdett,T., Dylag,M., Emam,I., Farne,A., Hastings,E., Ison,J., Keays,M. *et al.* (2013) ArrayExpress update—trends in database growth and links to data analysis tools. *Nucleic Acids Res.*, **41**, D987–D990.
3. Ambrosini,G., Dreos,R., Kumar,S. and Bucher,P. (2016) The ChIP-Seq tools and web server: a resource for analyzing ChIP-seq and other types of genomic data. *BMC Genomics*, **17**, 938.
4. Liu,T., Ortiz,J.A., Taing,L., Meyer,C.A., Lee,B., Zhang,Y., Shin,H., Wong,S.S., Ma,J., Lei,Y. *et al.* (2011) Cistrome: an integrative platform for transcriptional regulation studies. *Genome Biol.*, **12**, R83.
5. Mendoza-Parra,M.A., Saleem,M.-A.M., Blum,M., Cholley,P.-E. and Gronemeyer,H. (2016) NGS-QC Generator: A Quality Control System for ChIP-Seq and Related Deep Sequencing-Generated Datasets. *Methods Mol. Biol. Clifton NJ*, **1418**, 243–265.

6. Halbritter,F., Kousa,A.I. and Tomlinson,S.R. (2014) GeneProf data: a resource of curated, integrated and reusable high-throughput genomics experiments. *Nucleic Acids Res.*, **42**, D851–D858.
7. Ambrosini,G., Praz,V., Jagannathan,V. and Bucher,P. (2003) Signal search analysis server. *Nucleic Acids Res.*, **31**, 3618–3620.
8. Consortium,T.E.P. (2004) The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science*, **306**, 636–640.
9. Kundaje,A., Meuleman,W., Ernst,J., Bilenky,M., Yen,A., Kheradpour,P., Zhang,Z., Heravi-Moussavi,A., Liu,Y., Amin,V. *et al.* (2015) Integrative analysis of 111 reference human epigenomes. *Nature*, **518**, 317–330.
10. Lizio,M., Harshbarger,J., Shimoji,H., Severin,J., Kasukawa,T., Sahin,S., Abugessaisa,I., Fukuda,S., Hori,F., Ishikawa-Kato,S. *et al.* (2015) Gateways to the FANTOM5 promoter level mammalian expression atlas. *Genome Biol.*, **16**, 22.
11. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.
12. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.
13. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G. and Durbin,R. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
14. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.
15. Pollard,K.S., Hubisz,M.J., Rosenbloom,K.R. and Siepel,A. (2010) Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.*, **20**, 110–121.
16. Siepel,A., Bejerano,G., Pedersen,J.S., Hinrichs,A.S., Hou,M., Rosenbloom,K., Clawson,H., Spieth,J., Hillier,L.W., Richards,S. *et al.* (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, **15**, 1034–1050.
17. Jee,J., Rozowsky,J., Yip,K.Y., Lochovsky,L., Bjornson,R., Zhong,G., Zhang,Z., Fu,Y., Wang,J., Weng,Z. *et al.* (2011) ACT: aggregation and correlation toolbox for analyses of genome tracks. *Bioinformatics*, **27**, 1152–1154.
18. J,W., J,Z., S,I., X,L., Tw,W., Mc,G., Bg,P., X,D., A,K., Y,C. *et al.* (2012) Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors., Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res Genome Res.*, **22**, 1798–1812.
19. Dreos,R., Ambrosini,G., Groux,R., Cavin Périer,R. and Bucher,P. (2017) The eukaryotic promoter database in its 30th year: focus on non-vertebrate organisms. *Nucleic Acids Res.*, **45**, D51–D55.
20. Feng,J., Liu,T., Qin,B., Zhang,Y. and Liu,X.S. (2012) Identifying ChIP-seq enrichment using MACS. *Nat. Protoc.*, **7**, 1728–1740.