ReMap 2018: an updated atlas of regulatory regions from an integrative analysis of DNA-binding ChIP-seq experiments

Jeanne Chèneby, Marius Gheorghe, Marie Artufel, Anthony Mathelier, Benoit Ballester

▶ **To cite this version:**

Jeanne Chèneby, Marius Gheorghe, Marie Artufel, Anthony Mathelier, Benoit Ballester. ReMap 2018: an updated atlas of regulatory regions from an integrative analysis of DNA-binding ChIP-seq experiments. Nucleic Acids Research, 2018, 46 (D1), pp.D267-D275. 10.1093/nar/gkx1092 . hal-01646201

HAL Id: hal-01646201
https://amu.hal.science/hal-01646201v1

Submitted on 10 Jan 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

Published online 8 November 2017 *Nucleic Acids Research, 2018, Vol. 46, Database issue* ***D267–D275***
doi: 10.1093/nar/gkx1092

ReMap 2018: an updated atlas of regulatory regions from an integrative analysis of DNA-binding ChIP-seq experiments

Jeanne Chèneby[1,2], Marius Gheorghe[3], Marie Artufel[1,2], Anthony Mathelier[3,4] and Benoit Ballester[1,2,*]

[1]INSERM, UMR1090 TAGC, Marseille F-13288, France, [2]Aix-Marseille Université, UMR1090 TAGC, Marseille F-13288, France, [3]Centre for Molecular Medicine Norway (NCMM), Nordic EMBL Partnership, University of Oslo, 0318 Oslo, Norway and [4]Department of Cancer Genetics, Institute for Cancer Research, Oslo University Hospital Radiumhospitalet, 0310 Oslo, Norway

Received September 15, 2017; Revised October 19, 2017; Editorial Decision October 20, 2017; Accepted October 20, 2017

ABSTRACT

With this latest release of ReMap (http://remap.cisreg.eu), we present a unique collection of regulatory regions in human, as a result of a large-scale integrative analysis of ChIP-seq experiments for hundreds of transcriptional regulators (TRs) such as transcription factors, transcriptional co-activators and chromatin regulators. In 2015, we introduced the ReMap database to capture the genome regulatory space by integrating public ChIP-seq datasets, covering 237 TRs across 13 million (M) peaks. In this release, we have extended this catalog to constitute a unique collection of regulatory regions. Specifically, we have collected, analyzed and retained after quality control a total of 2829 ChIP-seq datasets available from public sources, covering a total of 485 TRs with a catalog of 80M peaks. Additionally, the updated database includes new search features for TR names as well as aliases, including cell line names and the ability to navigate the data directly within genome browsers via public track hubs. Finally, full access to this catalog is available online together with a TR binding enrichment analysis tool. ReMap 2018 provides a significant update of the ReMap database, providing an in depth view of the complexity of the regulatory landscape in human.

INTRODUCTION

Transcription factors (TFs), transcriptional coactivators (TCAs) and chromatin-remodeling factors (CRFs) drive gene transcription and the organization of chromatin through DNA binding. TFs specifically bind to DNA sequences (TF binding sites) to activate (activators) or repress (repressors) transcription, TCAs enhance gene transcription by binding to activator TF. While CRFs modify the chromatin architecture to allow DNA access for transcription machinery proteins. In recent years, the development of high-throughput techniques like chromatin immunoprecipitation followed by sequencing (ChIP-seq) (1) has allowed to experimentally obtain genome-wide maps of binding sites across many cell types for a variety of DNA-binding proteins. The popularity of ChIP-seq has led to a deluge of data in current data warehouses (2,3) for TFs, TCAs and CRFs, collectively named transcriptional regulators (TRs). The rapid accumulation of ChIP-seq data in public databases provides a unique and valuable resource for hundreds of TR occupancy maps. There is a strong need to integrate these large-scale datasets to explore the transcriptional regulatory repertoire. Unfortunately, the heterogeneity of the pipelines used to process these data, as well as the variety of underlying formats used, challenge the analysis processes and the underlying detection of TF binding sites (TFBSs). Integrative studies would offer significant insights into the dynamic mechanisms by which a TF selects its binding regions in each cellular environment.

ReMap has been the first large scale integrative initiative to study these data, offering significant insights into the complexity of the human regulatory landscape (4). The ReMap 2015 resource created a large catalog of regulatory regions by compiling the genomic localization of 132 different TRs across 83 different human cell lines and tissue types based on 395 non-ENCODE datasets selected from Gene Expression Omnibus (2) and ArrayExpress (3). This catalog was merged with the ENCODE multi-cell peaks (5), generating a global map of 13M regulatory elements for 237 TRs across multiple cell types. However, since the 2015 publication of ReMap, an even greater number of ChIP-seq assays has been submitted to genomic data repositories.

*To whom correspondence should be addressed. Tel: +33 4 91 82 87 39; Fax: +33 4 91 82 87 01; Email: benoit.ballester@inserm.fr

© The Author(s) 2017. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Downloaded from https://academic.oup.com/nar/article-abstract/46/D1/D267/4602873 by SCDU Mediterranee user on 10 January 2019

Here, we introduce the ReMap 2018 update, which includes the integration of 2829 quality controlled ChIP-seq datasets for TFs, TCAs and CRFs. The new ChIP-seq datasets ($n = 1763$, defined as 'Public' for non-ENCODE) as well as the latest ENCODE ChIP-seq data ($n = 1066$) have been mapped to the GRCh38/hg38 human assembly, quality filtered and analyzed with a uniform pipeline. In this update, we propose a unified integration of all public ChIP-seq datasets producing a unique atlas of regulatory regions for 485 TRs across 346 cell types, for a total of 80M DNA binding regions. Each experiment introduced in this release has been assessed and manually curated to ensure correct meta-data annotation. Our ReMap database provides DNA-binding locations for each TR, either for each experiment, at cell line or primary cell level, or at the TR level in a non-redundant fashion across all collected experiments. This update represents a 2-fold increase in the number of DNA-binding proteins, 7-fold in the number of processed datasets, 4-fold in the number of cell lines/tissue types and 6-fold in the number of identified ChIP-seq peaks. While the first version of the ReMap catalog covered 26% (793 Mb) of the human genome, the regulatory search space for ReMap 2018 covers 46% (1.4Gb).

Finally, we give the community access to various options to visualize and browse our catalog, allowing users to navigate and dissect their genomic loci of interest co-occupied by multiple TRs in various cell types. Browsing the ReMap 2018 catalog using the Public Track hub, IGV data sever, Ensembl or UCSC sessions clearly exposes the abundance and intricacy of combinatorial regulation in cellular contexts.

This report presents the extensive data increase and regulatory catalog expansion of ReMap as a result of our large-scale data integration and genome-wide analysis efforts. The manual curation specific to the ReMap initiative offers a unique and unprecedented collection of TR binding regions. These improvements, together with several novel enhancements (search bars, data track displays, format and annotation), constitute a unique atlas of regulatory regions generated by the integration of public resources.

MATERIALS AND METHODS

Available datasets

ChIP-seq datasets were extracted from the Gene Expression Omnibus (GEO) (2), ArrayExpress (AE) (3) and ENCODE (5) databases. For GEO, the query '('chip seq' OR 'chipseq' OR 'chip sequencing') AND 'Genome binding/occupancy profiling by high-throughput sequencing' AND 'homo sapiens'[organism] AND NOT 'ENCODE'[project]' was used to return a list of all potential datasets, which were then manually assessed and curated for further analyses. For ArrayExpress, we used the query (Filtered by organism 'Homo sapiens', experiment type 'dna assay', experiment type 'sequencing assay', AE only 'on') to return datasets not present in GEO. Contrary to other similar databases (chip-atlas http://chip-atlas.org, (6,7)), ReMap meta-data for each experiment are manually curated, annotated with the official gene name from the HUGO Gene Nomenclature Committee (8) (www.genenames.org) and BRENDA Tissue Ontologies (9) for cell lines (www.ebi.ac.uk/ols/ontologies/bto). Datasets involving polymerases (Pol2 and Pol3), and some mutated or fused TFs (e.g. KAP1 N/C terminal mutation, GSE27929) were filtered out. A dataset is defined as a ChIP-seq experiment in a given GEO/AE/ENCODE series (e.g. GSE37345), for a given TF (e.g. FOXA1), and in a particular biological condition (e.g. LNCaP). Datasets were labeled with the concatenation of these three pieces of information (e.g. GSE37345.FOXA1.LNCAP).

A total of 3180 datasets were processed (Supplementary Table S1). Specifically, we analyzed 2020 datasets from GEO (1862) and ArrayExpress (158) repositories (July 2008 to May 2017). We define these non-ENCODE datasets as the 'Public' set, in opposition to ENCODE datasets (1160) (full list of experiments in Supplementary Tables S2 and 3).

ReMap 2015 contained the multi-cell peak calling processed from ENCODE release V3 (August 2013). For the ReMap 2018 update, we re-analyzed, starting from the raw data, all ENCODE ChIP-seq experiments for TFs, transcriptional and chromatin regulators, following the same processing pipeline as the Public set. We retrieved the list of ENCODE data as FASTQ files from the ENCODE portal (https://www.encodeproject.org/) using the following filters: Assay: 'ChIP-seq', Organism: 'Homo sapiens', Target of assay: 'TF', Available data: 'fastq' on 21 June 2016. Metadata information in JSON format and FASTQ files were retrieved using the Python *requests* module. We processed 1160 datasets associated to 161 TRs and 87 cell lines. We removed 2 TRs (POLR2A, POLR3G), and renamed TR aliases into official HGNC identifiers (e.g. p65 into RELA, see Supplementary Table) leading to a final list of 279 TRs from ENCODE.

ChIP-seq processing

Both ENCODE and Public datasets were uniformly processed and analyzed. Bowtie 2 (version 2.2.9) (10) with options –end-to-end –sensitive was used to align all reads on the human genome (GRCh38/hg38 assembly). For Public datasets, adapters were removed using TrimGalore (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/), trimming reads up to 30 bp. Polymerase chain reaction duplicates were removed from the alignments with samtools rmdup (11). For the ENCODE data, the adapter trimming step was not employed, as this data already passed certain quality assessment steps (https://www.encodeproject.org/data-standards/). TR binding regions were identified using the MACS2 peak-calling tool (version 2.1.1.2) (12) in order to follow ENCODE ChIP-seq guidelines (13), with stringent thresholds (MACS2 default thresholds, *P*-value: 1e-5). Input datasets were used when available. All peak-calling files are available to download. Among the 80M peaks identified, 99.5% of peaks (79 753 407) were below 1.5 kb in size (mean size: 286 bp, median size: 231 bp) and only 376 017 peaks were above 1.5 kb in size (mean size: 2209 bp, median size: 1859 bp).

Quality assessment

As raw data are obtained from various sources, under different experimental conditions and platforms, data quality differs across experiments. Since the ReMap 2015 release, our ChIP-seq pipeline assesses the quality of all

Downloaded from https://academic.oup.com/nar/article-abstract/46/D1/D267/4602873 by SCDU Mediterranee user on 10 January 2019

datasets, unlike similar databases (chip-atlas http://chip-atlas.org, (6,7)), (Supplementary Table S4). We compute a score based on the cross-correlation and the FRiP (fraction of reads in peaks) metrics developed by the ENCODE consortium (13) (Supplementary Figure S1). Descriptions of the ENCODE quality coefficients can be found on the UCSC Genome portal (http://genome.ucsc.edu/ENCODE/qualityMetrics.html). Our pipeline computes the normalized strand cross-correlation coefficient (NSC) as a ratio between the maximal fragment-length cross-correlation value and the background cross-correlation value, and the relative strand cross-correlation coefficient (RSC), as a ratio between the fragment-length cross-correlation and the read-length cross-correlation. The same methods and quality cutoffs were applied as in ReMap 2015 (4). Datasets not passing the QC were not included in the catalog of peaks available for download (http://remap.cisreg.eu).

DNA constraint scores

We provide the conservation profiles at the nucleotide level for each of the 485 TRs present in our catalog. We assessed the DNA constraint for each base pair by considering ±1 kb around the summit of each non-redundant peak (see below). Genomic Evolutionary Rate Profiling scores (GERP) were used to calculate the conservation of each nucleotide in a multi-species alignment (14). The computed GERP scores were obtained from the 24-way amniota vertebrates Pecan (15) multi-species alignment, and extracted from the Ensembl Compara database release v89 (16).

Genome coverage, non-redundant peak sets and CRMs

Genome coverages were computed using the BedTools suite (17) (version 2.17.0) using the 'genomecov' function with the option -max 2 that combines all positions with a depth ≥2 binding locations. Full details of the ReMap 2015 and 2018 genome coverage are available in Supplementary Table S5. ReMap also provides a catalog of discrete, non-redundant binding regions for each TR, a specificity not found in other databases (chip-atlas http://chip-atlas.org, (6,18)). We used BedTools to merge overlapping peaks (with at least 1 bp overlap) identified in different datasets for the same TR. The summit of the resulting peaks was defined as the average position of the summits of the merged peaks. Those peaks made of at least two or more peaks for a given factor are defined as non-redundant peaks. We observed a mean variation of 77 bp between the summits of the non-redundant peaks and the individual peak summits (Supplementary Figure S2). Similarly, to obtain the *cis*-regulatory modules (CRMs) in the genome, overlapping peaks of all TRs were merged using BedTools. Regions bound by several TRs are called CRMs, whereas regions bound by only one TR are labeled as singletons.

Roadmap human epigenome annotations

Two sets of chromatin accessibility data were used to better characterize the ReMap atlas. We employed BedTools for overlap analyses allowing a minimum of 10% overlap. The NIH Roadmap Epigenomics Mapping Consortium (19) data were downloaded from the roadmap data portal (http://egg2.wustl.edu/roadmap). Delineation of DNaseI-accessible regulatory regions were accessed from http://egg2.wustl.edu/roadmap/web_portal/DNase_reg.html#delieation. BED files with coordinates of each region type for each epigenome separately are available for 81 232 promoter regions (1.44% of genome), 2 328 936 putative enhancer regions (12.63% of genome) and 129 960 dyadic promoter/enhancer regions (0.99% of genome). The core 15-state model of chromatin combinatorial interactions between different chromatin marks was downloaded from http://egg2.wustl.edu/roadmap/web_portal/chr_state_learning.html#core_15state. Chromatin state definitions and abbreviations are: 1 Active TSS (TssA), 2 Flanking active TSS (TssAFlnk), 3 Transcr. at gene 5′ and 3′(TxFlnk), 4 Strong transcription (Tx), 5 Weak transcription (TxWk), 6 Genic enhancers (EnhG), 7 Enhancers (Enh), 8 ZNF genes + repeats (ZNF/Rpts), 9 Heterochromatin (Het), 10 Bivalent/poised TSS (TssBiv), 11 Flanking bivalent TSS/Enh (BivFlnk), 12 Bivalent enhancer (EnhBiv), 13 Repressed Polycomb (ReprPC), 14 Weak repressed Polycomb (ReprPCWk) and 15 Quiescent/low (Quies).

DATA COLLECTION AND CONTENT

Integration of data sources

The 2018 release of the ReMap database reflects significant advances in the number of binding regions, the number of TFs, transcriptional co-activators, chromatin regulators and overall the total number of datasets integrated in our catalog. We initially selected, processed and analyzed 3180 ChIP-seq datasets against TRs from GEO, AE and ENCODE. To ensure consistency and comparability, all datasets were processed from raw data, through our uniform ChIP-seq workflow that included read filtering, read mapping, peak calling and quality assessment based on ENCODE quality criterions. As the quality of ChIP-seq experiments vary significantly (20,21), we incorporated a critical data quality filtering step in our pipeline—not implemented in other databases (chip-atlas http://chip-atlas.org (6,7,18)). Specifically, we considered four quality metrics, two metrics independent of peak calling for assessing signal-to-noise ratios in a ChIP-seq experiment and two metrics based on peak properties. Following ENCODE ChIP-seq guidelines and practices (13), we used the NSC and the RSC (see 'Materials and Methods' section). Further, we used the FRiP and the number of peaks in the dataset (see 'Materials and Methods' section). After applying our quality filters based on these four ChIP-seq metrics we retained 2829 datasets (89%): 1763 datasets from GEO and ArrayExpress and 1066 from ENCODE (Figure 1A and Supplementary Figure S1). The significant increase of data is spread across almost all TFs when compared to ReMap 2015 (Figure 1B). Nevertheless, we observe TFs (e.g. AR, ESR1, FOXA1) and CRFs (e.g. BRD4, EZH2) displaying a larger data growth than other DNA-binding proteins. The majority of TRs show additional datasets integrated in ReMap 2018 (Figure 1B, dark blue bars).

Downloaded from https://academic.oup.com/nar/article-abstract/46/D1/D267/4602873 by SCDU Mediterranee user on 10 January 2019

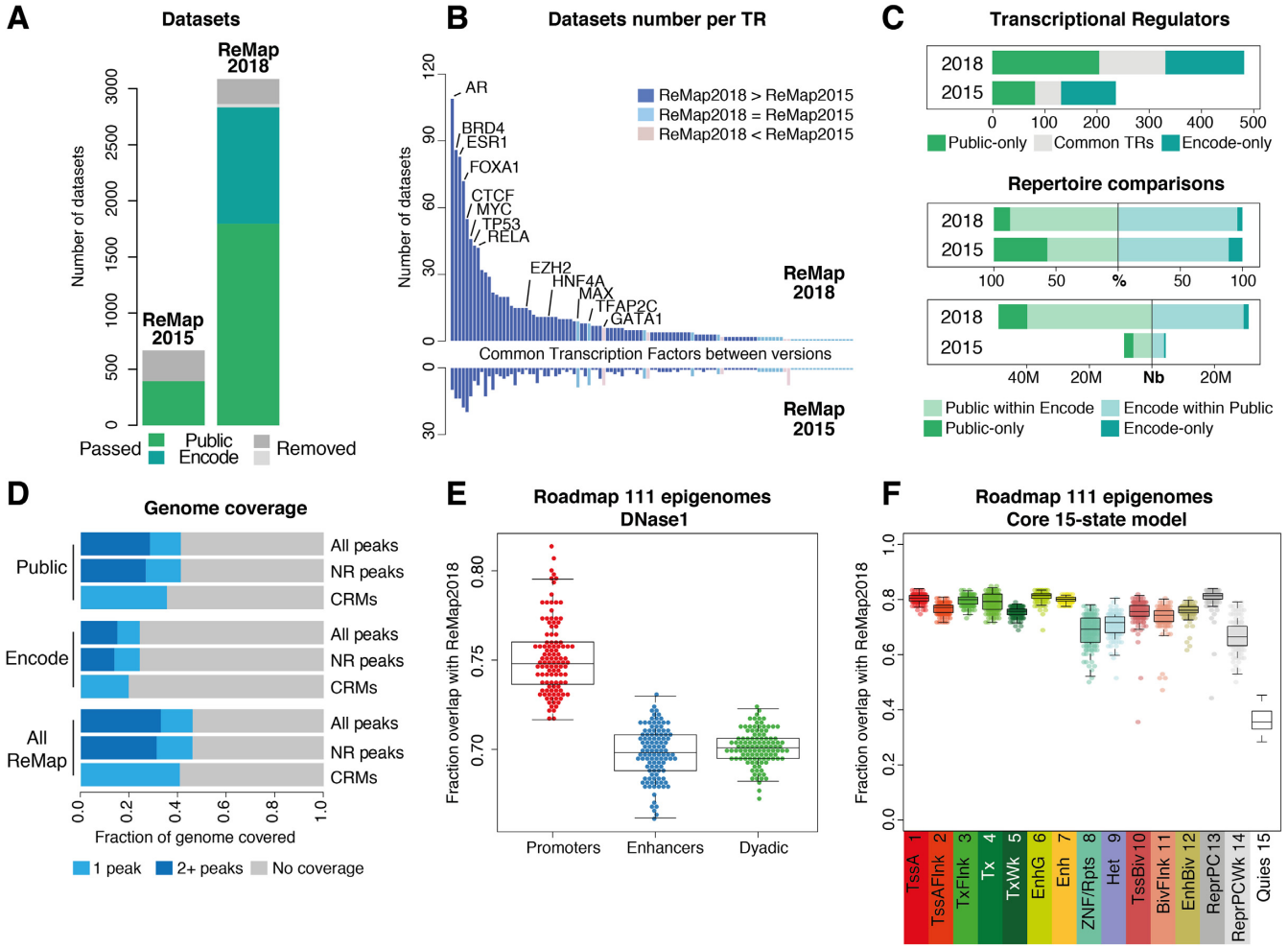

Figure 1. Overview of the ReMap database expansion. (**A**) Analyzed datasets growth in ReMap 2018 compared to ReMap 2015. (**B**) Evolution of the number of datasets per TRs, ranked across common between both ReMap versions. (**C**) Common TRs between Public and ENCODE sets of data (gray). Direct comparison of Public and ENCODE repertoire, defined as percentages (%), and as number (Nb) of peaks. (**D**) Genome coverage fraction of each ReMap dataset (NR non-redundant, CRM Cis Regulatory Modules). (**E**) Comparison of DNase I-accessible regulatory regions against the ReMap 2018, regions from the Roadmap Epigenomics Consortium defining promoter-only, enhancer-only or enhancer–promoter alternating states (Dyadic). Each dot represents the fraction overlap with ReMap 2018 for one of the 111 epigenomes. (**F**) Comparison of the Roadmap Epigenomics Consortium chromatin states annotations against the ReMap 2018 catalog, using the Core 15 chromatin states model, and a minimum overlap of 50% between regions. Each dot represents the overlap for one of the 111 epigenomes. Chromatin state definitions and abbreviations are as follows; 1 Active TSS (TssA), 2 Flanking active TSS (TssAFlnk), 3 Transcr. at gene 5′ and 3′(TxFlnk), 4 Strong transcription (Tx), 5 Weak transcription (TxWk), 6 Genic enhancers (EnhG), 7 Enhancers (Enh), 8 ZNF genes + repeats (ZNF/Rpts), 9 Heterochromatin (Het), 10 Bivalent/poised TSS (TssBiv), 11 Flanking bivalent TSS/Enh (BivFlnk), 12 Bivalent enhancer (EnhBiv), 13 Repressed Polycomb (ReprPC), 14 Weak repressed Polycomb (ReprPCWk), 15 Quiescent/low (Quies).

Regulatory catalog expansion

With all ChIP-seq data uniformly processed, the ReMap 2018 catalog displays ENCODE data down to the cell line and dataset level rather than the simpler multi-cell analysis provided by ENCODE DCC used in ReMap 2015. Our analyses produced 48 693 300 peaks for the Public-only (non-ENCODE) set across 331 TRs and 31 436 124 peaks for the ENCODE set across 279 TRs, leading to a final ReMap regulatory atlas of 80 129 705 peaks generated from 485 TRs (Figure 1C). We found 125 TRs common to the two sets, 154 proteins specific to ENCODE and 206 specific to the Public catalog (Figure 1C). We also found that 839 400 CRMs are shared between both catalogs. Taken separately, the ENCODE peaks overlaps by 96% the Public regions, and 87% of the Public peaks overlap ENCODE regions (Figure 1C). It suggests that merging both Public and ENCODE sets complements the annotation of DNA-bound regions, as it increases the number of regulatory regions in our atlas, hence improving the annotation of DNA-bound elements in the human genome (Figures 1C and 2).

Indeed, about 13% (405 Mb) of the human genome is covered by at least one feature only from the entire ReMap catalog and 33% (1.02 Gb) are covered by two or more features (Figure 1D and Supplementary Table S4). The Public-only and ENCODE-only sets cover the genome by two or more peaks by 28 and 15% respectively. The observed differences can be explained by the wide spectrum of cell lines and treatments included in the Public set (300 cell lines) compared to the ENCODE set (86 cell lines). As a comparison, the

Downloaded from https://academic.oup.com/nar/article-abstract/46/D1/D267/4602873 by SCDU Mediterranee user on 10 January 2019

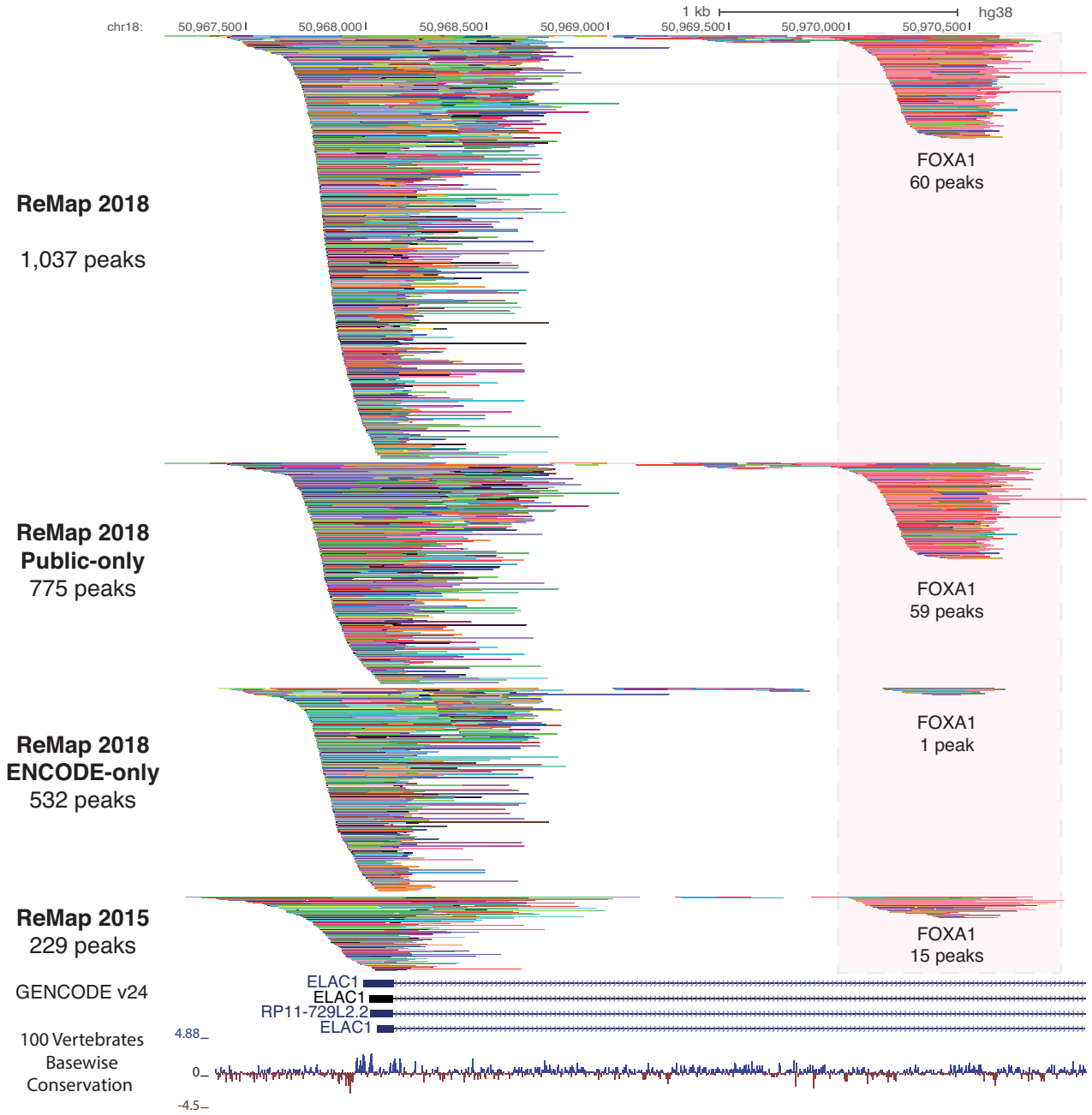

Figure 2. ReMap ChIP-seq binding pattern of 2829 datasets. A genome browser example of the ChIP-seq binding peak depth of the ReMap 2018 catalog compared to ReMap 2015 at the vicinity of the ELAC1 promoter (chr18:50,967,094-50,970,983). The tracks and peaks displayed are compacted to thin lines so the depth of ReMap 2018 bindings can be compared to ReMap 2015. A full and un-compacted screenshot is available as Supplementary Figures S2 and 3. On this location the ReMap 2018 catalog contains 1307 peaks, whereas the ReMap 2015 contains 229 peaks (ReMap 2015 lifted to GRCh38/hg38 assembly). The following genome tracks correspond to the GENCODE v24 Comprehensive Transcript Set and the 100 vertebrates base-wise conservation showing sites predicted to be conserved (positive scores in blue), and sites predicted to be fast-evolving (negative scores in red). A detailed view of the redundant peaks for a FOXA1 site is available in Figure 3.

ReMap 2015 catalog covered 10% (321 Mb) of the genome with one feature only, and 15% (471 Mb) with at least two or more features. Between the two ReMap versions, we observe that the fraction of the human genome covered by one feature remains extremely stable (+84 Mb from 2015 to 2018), whereas the fraction covered by two or more regulatory features increases by 545 Mb. With ReMap 2018, we increase the range of the regulatory space, and provide binding regions for similar TRs at a greater depth, revealing tight and dense co-localization sites (Figures 2 and 3).

Overlap with *cis*-regulatory genomic regions

Using the NIH Roadmap 111 epigenomes analyses, we asked whether the DNase I defined regions as well as the core 15 chromatin states model would better characterize

Downloaded from https://academic.oup.com/nar/article-abstract/46/D1/D267/4602873 by SCDU Mediterranee user on 10 January 2019

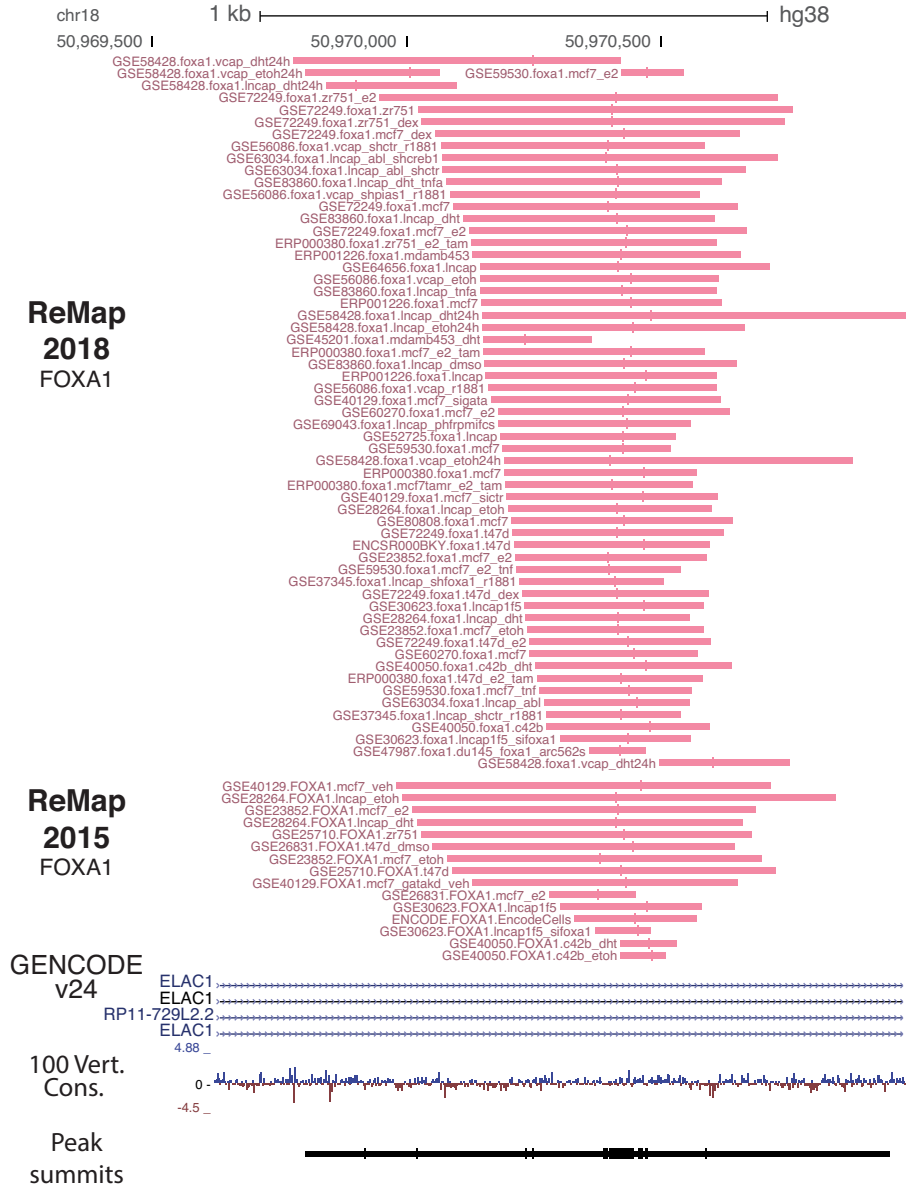

Figure 3. FOXA1 ChIP-seq peaks pattern evolution across ReMap versions. Detailed view of the FOXA1 peaks present in ReMap 2018 (60 peaks) compared to the FOXA1 peaks in ReMap 2015 (15 peaks) found at the genomic location chr18:50,969,638-50,970,931 in the first intron of the ELAC1 gene. Those 60 FOXA1 peaks are derived from GEO, ArrayExpress and ENCODE ChIP-seq across multiple cell lines. Interestingly, it can be noted that the peak summits (vertical bars) of each peak aggregate closely from each other, defining precisely the DNA binding location. Those aggregations of the FOXA1 summits are an illustration of what is globally observed for peaks of different TFs across the genome.

the ReMap atlas (Figure 1E and F). The Roadmap consortium defined a total of 3.5M DNase I-accessible regulatory regions by merging all DNase I hypersensitive regions across epigenomes, which were then annotated using the core 15-state model focusing on chromatin states for promoters, enhancers and dyadic (promoter + enhancer) ambiguous regions (see 'Materials and Methods' section). Among these three categories, the ReMap atlas could recapitulate on average 75.2% of the Roadmap promoter regions, 69.8% of enhancer regions and 70.1% of dyadic regions from the Roadmap annotation. Looking at the core 15-state model, we observe that the ReMap catalog recapitulates more than 70% of the regions covered by each state (Enhancer Genic (81%), Enhancer (80%) and TSS active (80%) states) with the exception of quiesent state (36%). Taken together, these results suggest that some promoter and enhancer activities from Roadmap may be cell type specific, as about 20–30% of those regions seem specific to Roadmap consortium cells. The ReMap initiative results from a large-scale integration of hundreds of diverse cell types, and leads to a regulatory landscape illustrating the large regulatory circuitry of those cells. The constant integration of novel data will allow for a greater definition of the regulatory space across the genome.

Downloaded from https://academic.oup.com/nar/article-abstract/46/D1/D267/4602873 by SCDU Mediterranee user on 10 January 2019

Large regulatory atlas

The ReMap database provides a large view of a unique regulatory landscape constituted by 80M binding regions forming 1.6M CRMs. The genomic organization of our occupancy map reveals dense co-localizations of sites forming tight clusters of heterogeneous binding sites with variable TRs complexity (Figure 2). For instance, the regulatory regions observed in the vicinity of the ELAC1 promoter illustrate the ReMap 2018 expansion ($n = 1037$ peaks). It highlights how the regulatory repertoire can be complemented by merging both Public and ENCODE sources. We observe a large cluster of peaks at the ELAC1 promoter followed by two clusters at +500 bp and 1 kb from the transcription start site. The third cluster exemplifies how integrating data from different sources improves genome annotations, as few peaks are available from ENCODE at this location. Additionally, this cluster was detailed in our previous ReMap publication (4) and consisted of 15 FOXA1 ChIP-seq peaks from different cells, antibodies, and laboratories (Figure 3). In this update, we consolidate this FOXA1 binding location with 60 peaks. The summit of each peak is represented by vertical bars aggregated closely from each other, providing an information about the putative location of the DNA binding site. The clustering of FOXA1 peaks and summits illustrates our genome-wide repertoire. However, this FOXA1 example shows overlapping sites derived from various experimental conditions, and therefore does not reflect the total number of discrete binding regions across the genome. To address redundancy between datasets, we merged binding regions for the same TR, resulting in a catalog of 35.5M peaks for all TRs combined. These merged peaks, defined as non-redundant peaks, are made of at least two or more peaks and singletons for a given factor across all experiments, and are available for download from the ReMap website. The TRs with the most merged binding regions across cell types are AR, FOXA1, CTCF and ESR1 (Supplementary Figure S6). These results indicate that most bindings are shared across different ChIP-seq experiments, either for similar or for different cell types. Overall, our ReMap update provides a unique opportunity to identify complex regulatory architectures containing multiple bound regions. We observe that by adding more cell lines, more experiments and more DNA-binding proteins, we increased the genome regulatory space and its depth (Figure 2), but also refined the current annotations of bound regions (Figure 3).

IMPLEMENTATION AND PUBLIC ACCESS

Web display

ReMap provides free public access to all data at http://remap.cisreg.eu. The results presented here provide an informative annotation for 80M ChIP-seq peaks coming from public data sources, which are derived from 485 TRs across 346 diverse cell lines. This catalog provides an unparalleled resource for dissecting site-specific TF bindings (e.g. FOXA1 in Figures 2 and 3) or genome-wide binding analyses. The ReMap web interface displays informations about the integrated TRs (description, classification, external references to Ensembl gene IDs, UniProt, RefSeq, WikiGene, JASPAR, FactorBook, TF Encyclopedia and other resources), peaks, and datasets (quality assessment, read mapping and peak calling statistics, conservation score under peaks). The interface provides a simple 'Dynamic Search' available from the TRs, Cell lines and Download pages and is the entry point for users to search for specific data. The search form allows users to narrow their searches based on gene aliases, dataset names or IDs, cell line names or ontology. For example, entering 'Oct' as search term in the 'Dynamic Search' returns three TFs POU2F2, POU2F1, POU5F1 having various 'OCT' aliases. Additionally, one could use the search box in the Cell or Download page to search for specific cell types containing the 'Colo' term for instance, or 'GSE66218' for a precise experiment from the Download page. Moreover, we provide a tool that allows the annotation of genomic regions provided by users. Those regions are compared against the ReMap catalog returning statistical enrichments of TR bindings present within user-provided input regions compared to random expectations. It allows for the study of over-represented TR binding regions.

Browsing and downloading data

Updates made in ReMap 2018 reflect significant improvement in the variety of genome navigation options. As the ReMap 2015 UCSC session was popular, we now provide more data navigation alternatives. The content of the ReMap database can be browsed through four options: (i) across two mirror sites of the UCSC Genome Browser (22) where a public session has been created (Figure 2 and Supplementary Figure S3), (ii) across three Ensembl Genome Browser mirrors (16) (Supplementary Figure S4), (iii) using the ReMap public track hub (23) or (iv) using the IGV data server (24) (Supplementary Figure S5). For each option, we provide four tracks, the full ReMap catalog containing all peaks, the Public-only peaks, the ENCODE-only peaks and a track containing only peaks above 1.5 kb. As the ReMap catalog expanded, it is crucial to allow visual exploration of regulatory regions across different platforms combined with public or user-specific genome-wide annotations. In addition, the entire ReMap 2018 catalog, as well as the Public-specific or ENCODE-specific peaks, have been compiled into BED files allowing further interpretations and computational analyses.

FUTURE DIRECTIONS

Next-generation sequencing technologies are playing a key role in improving our understanding of regulatory genomics. As ChIP-seq technology is applied to a broader set of cell lines, tissues and conditions, we will continuously maintain and update the database. In the near future, we propose on adding to the ReMap portfolio different peak-caller analyses to further consolidate the peak repertoire. Also, we aim to provide direct access to aligned reads through a FTP server, allowing users to upload and navigate aligned raw data of their choice. We plan on releasing a Bioconductor R-package for genomic region enrichment analyses for large genomic catalogs such as ReMap, which will be replacing our current web enrichment tool. In

Downloaded from https://academic.oup.com/nar/article-abstract/46/D1/D267/4602873 by SCDU Mediterranee user on 10 January 2019

the coming year, we would like to provide a Bioconductor R-package to search and download ReMap data for a specific study, to get genomic range objects, raw counts and/or metadata used for a specific study. Overall, determining the best approach to curate and annotate ChIP-seq data with a very broad level of submitted annotations and metadata into a simple-to-use, easy-to-analyze and up-to-date system will become a focus for the ReMap project.

CONCLUSION

The 2018 release of ReMap maintains the long-term focus of providing the research community with the largest catalog of high-quality regulatory regions by integrating all available ChIP-seq data from DNA-binding assays. The usefulness of ReMap is exemplified by the last release of the JASPAR database (25), for which ReMap ChIP-seq peaks were used to derive 45 new TF binding profiles that were incorporated in the 2018 release of the vertebrate CORE collection (Khan *et al*. 2018), providing a 9% increase from JASPAR 2016 (26) by solely relying on the ReMap 2018 catalog. Although new datasets are constantly added to repositories, we believe that our ReMap atlas will help in better understanding the regulation processes in human. In this update, we have (i) widely expanded the collection of datasets curated and analyzed from public sources with now 485 TFs, transcriptional co-activators and chromatin regulators; (ii) uniformly processed and integrated the ENCODE ChIP-seq data; (iii) enhanced the website usability by allowing dynamic search of TRs, aliases, cell lines and experiments, (iv) expanded the genome browsing experience by integrating ReMap in all UCSC and Ensembl Genome Browsers mirror sites and provided a Track Hub for data integration in other platforms; (v) improved the capacity to download all ReMap files in bulk or individually.

AVAILABILITY

ReMap 2018 can be accessed through a web interface at http://remap.cisreg.eu. Downloads are available in BED format for the entire ReMap catalog, the Public-only peaks, the ENCODE-only peaks, and in FASTA and BED formats for each TR. In addition, UCSC and Ensembl Genome Browsers users can navigate ReMap across their mirror sites, use ReMap in UCSC public sessions, or use the public track hub. Finally, Integrative Genome Browser (IGV) users have the option of loading an IGV optimized dataset directly in the application.

FEEDBACK

The ReMap team welcomes your feedback on the catalog, use of the website and use of the downloadable files. Please contact us at benoit.ballester@inserm.fr or remap@cisreg.eu for development requests. We thank our users for their feedback to make ReMap useful for the community.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

French Ministry of Higher Education and Research (MESR) PhD Fellowship (to J.C.); Norwegian Research Council (to A.M., M.G.); Helse Sør-Øst (to A.M., M.G.); University of Oslo (to A.M., M.G.). Funding for open access charge: Institut national de la santé et de la recherche médicale (INSERM).

Conflict of interest statement. None declared.

REFERENCES

1. Mardis,E.R. (2007) ChIP-seq: welcome to the new frontier. *Nat. Methods*, **4**, 613–614.
2. Barrett,T., Wilhite,S.E., Ledoux,P., Evangelista,C., Kim,I.F., Tomashevsky,M., Marshall,K.A., Phillippy,K.H., Sherman,P.M., Holko,M. *et al.* (2013) NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res.*, **41**, D991–D995.
3. Kolesnikov,N., Hastings,E., Keays,M., Melnichuk,O., Tang,Y.A., Williams,E., Dylag,M., Kurbatova,N., Brandizi,M., Burdett,T. *et al.* (2015) ArrayExpress update–simplifying data submissions. *Nucleic Acids Res.*, **43**, D1113–D1116.
4. Griffon,A., Barbier,Q., Dalino,J., van Helden,J., Spicuglia,S. and Ballester,B. (2015) Integrative analysis of public ChIP-seq experiments reveals a complex multi-cell regulatory landscape. *Nucleic Acids Res.*, **43**, e27.
5. ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
6. Zhou,K.-R., Liu,S., Sun,W.-J., Zheng,L.-L., Zhou,H., Yang,J.-H. and Qu,L.-H. (2017) ChIPBase v2.0: decoding transcriptional regulatory networks of non-coding RNAs and protein-coding genes from ChIP-seq data. *Nucleic Acids Res.*, **45**, D43–D50.
7. Yevshin,I., Sharipov,R., Valeev,T., Kel,A. and Kolpakov,F. (2017) GTRD: a database of transcription factor binding sites identified by ChIP-seq experiments. *Nucleic Acids Res.*, **45**, D61–D67.
8. Yates,B., Braschi,B., Gray,K.A., Seal,R.L., Tweedie,S. and Bruford,E.A. (2017) Genenames.org: the HGNC and VGNC resources in 2017. *Nucleic Acids Res.*, **45**, D619–D625.
9. Gremse,M., Chang,A., Schomburg,I., Grote,A., Scheer,M., Ebeling,C. and Schomburg,D. (2011) The BRENDA Tissue Ontology (BTO): the first all-integrating ontology of all organisms for enzyme sources. *Nucleic Acids Res.*, **39**, D507–D513.
10. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.
11. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G., Durbin,R. and 1000 Genome Project Data Processing Subgroup (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
12. Zhang,Y., Liu,T., Meyer,C.A., Eeckhoute,J., Johnson,D.S., Bernstein,B.E., Nusbaum,C., Myers,R.M., Brown,M., Li,W. *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, **9**, R137.
13. Landt,S.G., Marinov,G.K., Kundaje,A., Kheradpour,P., Pauli,F., Batzoglou,S., Bernstein,B.E., Bickel,P., Brown,J.B., Cayting,P. *et al.* (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.*, **22**, 1813–1831.
14. Davydov,E.V., Goode,D.L., Sirota,M., Cooper,G.M., Sidow,A. and Batzoglou,S. (2010) Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput. Biol.*, **6**, e1001025.
15. Paten,B., Herrero,J., Beal,K., Fitzgerald,S. and Birney,E. (2008) Enredo and Pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. *Genome Res.*, **18**, 1814–1828.
16. Aken,B.L., Achuthan,P., Akanni,W., Amode,M.R., Bernsdorff,F., Bhai,J., Billis,K., Carvalho-Silva,D., Cummins,C., Clapham,P. *et al.* (2017) Ensembl 2017. *Nucleic Acids Res.*, **45**, D635–D642.
17. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.
18. Mei,S., Qin,Q., Wu,Q., Sun,H., Zheng,R., Zang,C., Zhu,M., Wu,J., Shi,X., Taing,L. *et al.* (2017) Cistrome Data Browser: a data portal for ChIP-Seq and chromatin accessibility data in human and mouse. *Nucleic Acids Res.*, **45**, D658–D662.

Downloaded from https://academic.oup.com/nar/article-abstract/46/D1/D267/4602873 by SCDU Mediterranee user on 10 January 2019

19. Roadmap Epigenomics Consortium, Kundaje,A., Meuleman,W., Ernst,J., Bilenky,M., Yen,A., Heravi-Moussavi,A., Kheradpour,P., Zhang,Z., Wang,J. *et al.* (2015) Integrative analysis of 111 reference human epigenomes. *Nature*, **518**, 317–330.
20. 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.*, **1418**, 243–265.
21. Marinov,G.K., Kundaje,A., Park,P.J. and Wold,B.J. (2014) Large-scale quality analysis of published ChIP-seq data. *G3 (Bethesda)*, **4**, 209–223.
22. Tyner,C., Barber,G.P., Casper,J., Clawson,H., Diekhans,M., Eisenhart,C., Fischer,C.M., Gibson,D., Gonzalez,J.N., Guruvadoo,L. *et al.* (2017) The UCSC Genome Browser database: 2017 update. *Nucleic Acids Res.*, **45**, D626–D634.
23. Raney,B.J., Dreszer,T.R., Barber,G.P., Clawson,H., Fujita,P.A., Wang,T., Nguyen,N., Paten,B., Zweig,A.S., Karolchik,D. *et al.* (2014) Track data hubs enable visualization of user-defined genome-wide annotations on the UCSC Genome Browser. *Bioinformatics*, **30**, 1003–1005.
24. Thorvaldsdóttir,H., Robinson,J.T. and Mesirov,J.P. (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.*, **14**, 178–192.
25. Khan,A., Fornes,O., Stigliani,A., Gheorghe,F.N., Castro-Mondragon,J.A., van der Lee,R., Bessy,A., Chèneby,J., Kulkarni,S.R., Tan,G. *et al.* (2018) JASPAR 2018: update of the open-access database of transcription factor binding profiles andits web framework. *Nucleic Acids Res.*, doi:10.1093/nar/gkx1126.
26. Mathelier,A., Fornes,O., Arenillas,D.J., Chen,C.Y., Denay,G., Lee,J., Shi,W., Shyr,C., Tan,G., Worsley-Hunt,R. *et al.* (2016) JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, **44**, D110–D115.

Downloaded from https://academic.oup.com/nar/article-abstract/46/D1/D267/4602873 by SCDU Mediterranee user on 10 January 2019