ReMap 2020: a database of regulatory regions from an integrative analysis of Human and Arabidopsis DNA-binding sequencing experiments

Jeanne Chèneby, Zacharie Ménétrier, Martin Mestdagh, Thomas Rosnet, Allyssa Douida, Wassim Rhalloussi, Aurélie Bergon, Fabrice Lopez, Benoit Ballester

Nucleic Acids Research, 2019 **1**
doi: 10.1093/nar/gkz945

ReMap 2020: a database of regulatory regions from an integrative analysis of Human and Arabidopsis DNA-binding sequencing experiments

Jeanne Chèneby, Zacharie Ménétrier, Martin Mestdagh, Thomas Rosnet, Allyssa Douida, Wassim Rhalloussi, Aurélie Bergon, Fabrice Lopez and Benoit Ballester*

Aix Marseille Univ, INSERM, TAGC, Marseille, France

Received September 14, 2019; Revised October 07, 2019; Editorial Decision October 08, 2019; Accepted October 09, 2019

ABSTRACT

ReMap (http://remap.univ-amu.fr) aims to provide the largest catalogs of high-quality regulatory regions resulting from a large-scale integrative analysis of hundreds of transcription factors and regulators from DNA-binding experiments in Human and Arabidopsis (*Arabidopsis thaliana*). In this 2020 update of ReMap we have collected, analyzed and retained after quality control 2764 new human ChIP-seq and 208 ChIP-exo datasets available from public sources. The updated human atlas totalize 5798 datasets covering a total of 1135 transcriptional regulators (TRs) with a catalog of 165 million (M) peaks. This ReMap update comes with two unique Arabidopsis regulatory catalogs. First, a catalog of 372 Arabidopsis TRs across 2.6M peaks as a result of the integration of 509 ChIP-seq and DAP-seq datasets. Second, a catalog of 33 histone modifications and variants across 4.5M peaks from the integration of 286 ChIP-seq datasets. All catalogs are made available through track hubs at Ensembl and UCSC Genome Browsers. Additionally, this update comes with a new web framework providing an interactive user-interface, including improved search features. Finally, full programmatically access to the underlying data is available using a RESTful API together with a new R Shiny interface for a TRs binding enrichment analysis tool.

INTRODUCTION

The rapid accumulation of experiments capturing protein–DNA interactions in public databases provides a unique and valuable resource for thousands of protein occupancy maps. The development of high-throughput methods like chromatin immunoprecipitation followed by sequencing (ChIP-seq) (1), or ChIP-seq with DNA digestion steps (ChIP-exo) (2) as well as DNA affinity purification sequencing (DAP-seq) (3) has allowed to experimentally obtain genome-wide maps of binding regions across many cell types for a variety of DNA-binding proteins. Integrating these thousands of large-scale experiments would allow to explore the depth of the transcriptional regulatory repertoire. Unfortunately, the heterogeneous experimental metadata annotations deposited in data-warehouse, the inconsistency in target name convention, the different cell type or tissue names, the variety of bioinformatics methods and underlying file formats challenge a global analysis process and the underlying mapping of TF binding regions.

ReMap has been the first large scale integrative initiative with dedicated curation and uniform data processing pipeline to reveal the complex architecture of the human regulatory landscape (4). The core foundation of the ReMap project rely on the manual curation and annotation of experiments metadata. Each experiment metadata introduced in ReMap has been assessed and manually curated to ensure correct target and biotype annotation. The ReMap 2015 database (4) introduced a catalog of 13 million (M) DNA binding regions by compiling the genomic localization of 237 different transcriptional regulators (TRs) across 83 different human cell lines and tissue types based on the integration of 395 datasets from Gene Expression Omnibus (5) and ENCODE (6,7). For ReMap 2018 we updated the catalogue by processing 2829 quality controlled ChIP-seq datasets leading to a unique atlas of regulatory regions for 485 TRs across 346 cell types, for a total of 80M DNA binding regions (8).

Here, we describe ReMap 2020 which introduces two unique regulatory catalogs for Arabidopsis (*Arabidopsis thaliana*), and includes a major expansion of the human regulatory catalog, along with new user-interface features. The Arabidopsis regulatory catalog is the result of curation, annotation and integration of 179 quality controlled ChIP-seq and 330 DAP-seq datasets for transcription factors and general components of the transcriptional machinery. Those datasets have been mapped to the TAIR10 Arabidopsis assembly and analyzed with a uniform pipeline. This unified

*To whom correspondence should be addressed. Tel: +33 4 91 82 87 28; Fax: +33 4 91 82 87 01; Email: benoit.ballester@inserm.fr

integration of Arabidopsis datasets lead to a unique atlas of 342 TRs across 20 biotypes, 12 ecotypes, for a total of 2.6M DNA binding regions. Additionally, we introduce a catalog of 33 Arabidopsis histone modifications and variants across 4.5M peaks from the integration of 286 ChIP-seq datasets.

The human atlas has been updated with the curation, annotation and integration of 2969 quality controlled ChIP-seq and 208 ChIP-exo datasets mapped to the GRCh38/hg38 assembly and analysed with the ReMap pipeline. In this update, we propose a unique atlas of regulatory regions for 1135 TRs across 602 cell types, in 5798 datasets, for a total of 165M DNA binding regions.

The ReMap 2020 human update represents a 1.7-fold increase in the number of cell lines/tissue types, and a 2-fold increase in the number of DNA-binding proteins, number of processed datasets and number of identified peaks. While the eMap 2018 human catalog covers 19% (0.6 Gb) of the human genome with more than five peaks, the regulatory search space for ReMap 2020 covers 34% (1 Gb, 5+ peaks).

Finally, the fully redesigned ReMap web-interface gives the community richer options to navigate and search our data, to visualize and browse all catalogs with public track hubs integrated in Ensembl and UCSC Genome Browsers. Additionally we updated our ReMapEnrich tool with a new R Shiny interface, while also allowing programmatic access to the underlying data with a RESTful API.

This report presents the third ReMap release, which comes with two unique catalogs for Arabidopsis and an extensive data increase and regulatory catalog expansion of the human atlas as a result of our large-scale data integration and analysis efforts. The manual metadata curation engaged in the ReMap project offers a unique and unprecedented collection of DNA-binding regions for two species. This data expansion is supported by a range of new functionalities for better community access.

MATERIALS AND METHODS

Available Human and Arabidopsis datasets

New DNA-binding experiments such as ChIP-seq, ChIP-exo and DAP-seq were extracted from the NCBI Gene Expression Omnibus (GEO) (5) and ENCODE (6,7) databases. For GEO, the query '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. The same query with 'arabidopsis thaliana'[organism] was used to return all potential datasets. For each experiment ReMap metadata are manually curated and annotated with the official gene symbol. Materials and methods from published papers are often read when deposited metadata is insufficient. For human we used the HUGO Gene Nomenclature Committee (9) (www.genenames.org), BRENDA Tissue Ontologies (10) for cell lines (www.ebi.ac.uk/ols/ontologies/bto) as well as the Cellosaurus database (11) to homogenize cell names (e.g. MCF-7 not MCF7, Hep-G2, not HepG2, Hepg2 etc.). For Arabidopsis (*Arabidopsis thaliana*) we used the Ensembl Genome (12) gene symbols. Ecotypes and biotypes description were curated and homogenized when the information was available in the metadata or paper. Datasets involving polymerases (Pol2 and Pol3), and some mutated or fused TFs (e.g. KAP1 N/C terminal mutation, GSE27929) were filtered out. When multiple antibodies were pooled (eg: RUNX1 and RUNX3, GSE17954) we would name the target as RUNX1-3.

In ReMap, we define a dataset as a DNA-binding experiment in a given GEO/AE/ENCODE series (e.g. GSE37345), for a given TR (e.g. FOXA1), and in a particular biotype (e.g. LNCaP, K-562, Leaf, Seedling) in a given biological condition (e.g. 45 min DMSO, 21 days-wt-watered). Datasets were labeled with the concatenation of these pieces of information (e.g. GSE37345.FOXA1.LNCAP_45 min-DMSO). The core of ReMap data rely on ChIP-seq assays, but for this update we analysed a major human ChIP-exo experiment (GSE78099), and for Arabidopsis a major DAP-seq experiment (GSE60141).

A total of 7908 datasets were processed (Supplementary Table S1). Specifically, we analyzed 6498 human datasets deposited in public repositories from 1 July 2008 to 10 November 2018), and 1410 Arabidopsis datasets from 1 January 2009 to 2 February 2018 (full list of datasets in Supplementary Tables S5, S8, S11). For the ENCODE data in the 2020 update, we re-analyzed, starting from the raw data, all new ENCODE ChIP-seq experiments for TFs, transcriptional and chromatin regulators released since ReMap 2018 release (1 August 2016 to 5 February 2019), following the same processing pipeline. 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 5 February 2019. Metadata information in JSON format and FASTQ files were retrieved using the Python requests module. We processed 964 ENCODE datasets, 894 of whom passed our quality filters. We renamed TR ENCODE aliases into official HGNC identifiers (e.g. p65 into RELA, see Supplementary Table S6), and renamed cell lines to official BRENDA and Cellosaurus conventions (e.g. K562 into K-562, lost of modified names Supplementary Tables S6, S9, S12).

ChIP-seq processing

All human and arabidopsis ChIP-seq datasets were uniformly curated, processed and analyzed. Bowtie 2 (version 2.2.9) (13) with options -end-to-end -sensitive was used to align all reads on the human genome GRCh38/hg38 assembly and on the *A. thaliana* TAIR10 assembly. Adapters were removed using Trim Galore (https://github.com/FelixKrueger/TrimGalore), trimming reads up to 30 bp. Trim Galore is a wrapper around Cutadapt and FastQC to consistently apply adapter and quality trimming to FastQ files. Polymerase chain reaction duplicates were removed from the alignments with samtools rmdup (14). TR binding regions were identified using the MACS2 (15) peak-calling tool (version 2.1.1.2) to follow ENCODE ChIP-seq guidelines (13), with stringent thresholds (MACS2 default thresholds, *Q*-value: 1e–5, -g: with corresponding genome sizes). When available in the experiments, input datasets were used in the peak calling process. All peak-calling nar-

rowPeak or broadPeak files are available to download. For ChIP-seq, ChIP-exo and (amp)DAP-seq analyses, peak files containing <100 peaks were discarded.

Quality assessment

Data quality differs across experiments, as the data we process comes from various sources and are generated under different experimental conditions and platforms. Since the first release of ReMap 2015, our pipeline has assessed the quality of each dataset processed, unlike similar databases (Supplementary Tables S5, S8, S11). For ReMap 2020 the same quality pipeline and cutoffs were applied as in ReMap 2018 (8). Briefly, for both species, and all ChIP-seq and DAP-seq datasets processed for this update we computed a score based on the cross-correlation and the FRiP (fraction of reads in peaks) metrics developed by the ENCODE consortium (16) (Supplementary Figures S1–S3, ENCODE quality coefficients http://genome.ucsc.edu/ENCODE/qualityMetrics.html). Then 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. Datasets not passing the QC were not included in catalogs or BED files available for download. Rejected datasets are listed in (Supplementary Tables S2, S7, S10). For the human ChIP-exo data (GSE78099) we applied three post-processing steps as described in (16) material and methods. In brief, we filtered out peaks that would meet any of these criteria: MACS2 score <80 (equivalent to a $P = 1 \times 10^{-8}$); ratio of forward versus reverse strand reads >4; <20 reads over 500 bp per 15 million reads; normalized read count was less than twofold over the control. For the arabidopsis DAP-seq (17) data (GSE60141), we applied our standard ChIP-seq ReMap pipeline. Our quality assessment protocol could not be applied to ChIP-exo data, as the specificity of ChIP-exo peaks (extremely narrow) would not allow the computing of FRiP/RSC/NCS scores.

Open ReMap pipeline

A common issue with bioinformatics workflow is that it normally evolves at a different speed than data is published in the literature. We are making the code of our ReMap pipeline available to GitHub in the ReMap Github organisation (https://github.com/remap-cisreg). As the ReMap project expands dramatically and regulatory catalogs for other species are requested, it becomes essential to enable joint efforts between the ReMap team and external teams. This is essential for future collaborative production efforts. Briefly, our pipelines uses SnakeMake either in a Conda or Singularity environment depending on the HPC resources, where Torque and Slurm managers are both supported. Details of the pipeline are published on the repository.

Genome coverage

Genome coverages were computed using the BedTools suite (17) (version 2.26.0) using the 'genomecov' function with the option -max 100 that combines all positions with a depth ≥100 binding locations. Full details of the ReMap 2018 and 2020 genome coverage are available in Supplementary Table S4 for both species. Genomic regions covered by at least five peaks were considered as regulatory (Figure 1E, K, blue/green and light blue/green), uncertain regulatory regions (Figure 1K, grey), or not covered (Figure 1K, light grey).

Non-redundant peak sets and CRMs

For each target, ReMap provides non-redundant binding regions, a specificity not found in other databases (Supplementary Table S3). As the number of redundant peaks increases, and to improve the genomic accuracy of these non-redundant peaks, we updated our method. For a given TR, all peak lengths were truncated to the median size of all peaks for this TR. Then, to find clusters of redundant peaks, we used BedTools to intersect overlapping truncated peaks across different datasets (with at least 25% overlap, both ways). Once the clusters of overlapping peaks identified, non-redundant peaks are computed by averaging start, end and summits coordinates of all peaks in a cluster using original ReMap peaks lengths. For a given factor across all experiments, the non-redundant peak set consists of the computed non-redundant peaks plus singletons, and are available for download from the ReMap website. *cis*-regulatory modules (CRMs) were obtained by merging regions of all non-redundant peaks using BedTools. Regions bound by several TRs are called CRMs, whereas regions bound by only one TR are labeled as singletons.

HUMAN REGULATORY CATALOG EXPANSION

This 2020 release of the ReMap human database exhibits significant growth in the number of datasets, the number of transcriptional regulators and overall in the number of binding regions integrated in our catalog. In this update, we curated processed and analyzed 3424 new human ChIP-seq and 222 ChIP-exo datasets against TRs from GEO and ENCODE. Since the first ReMap release, we ensure consistency and comparability across datasets by processing from the raw data, through our uniform ReMap pipeline that includes read filtering, read mapping, peak calling and quality assessment (see 'Materials and Methods' section). Unlike other databases, the core foundation of ReMap lie in the manual curation of metadata, involving reading materials and methods as submitted experiments annotation is heterogeneous. In addition, we run a critical data quality filtering step in our pipeline to address varying quality of DNA-binding experiments (18,19). After applying our quality filters we retained 2969 datasets (82%): 2767 ChIP-seq and 208 ChIP-exo datasets (Figure 1A and Supplementary Figure S1). This leads to an updated human regulatory atlas totalizing 5798 datasets. The uniform data processing contributes to a final ReMap 2020 human regulatory atlas of 164 732 372 peaks generated from 11 135 TRs (Figure 1B). Our analyses produced 163 741 896 peaks set across 927 TRs for ChIP-seq and 990 476 peaks set across 208 TRs for ChIP-exo.

This update shows a 2-fold increase in the number of TRs and number of peaks. The significant data growth is

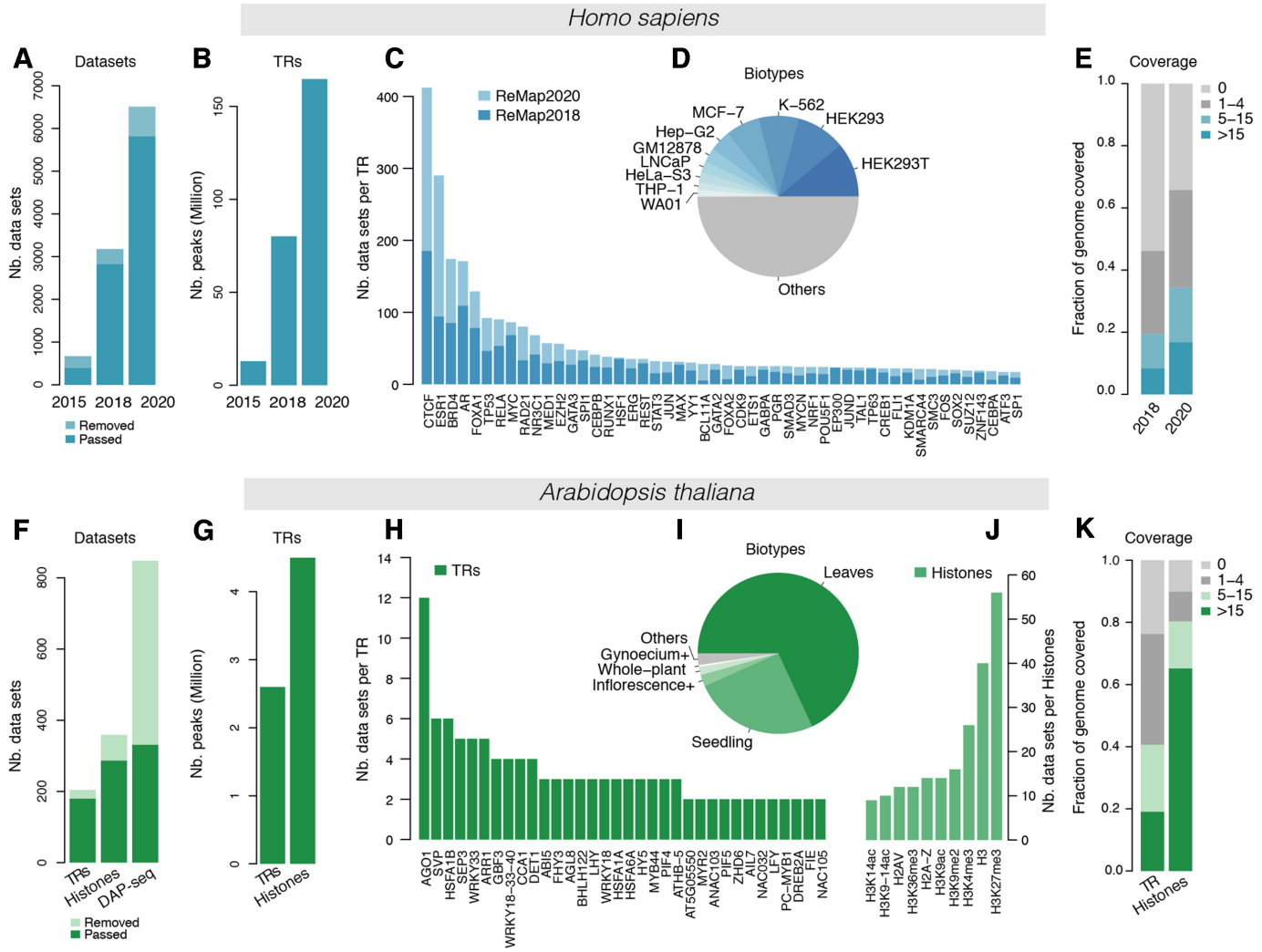

Figure 1. Overview of the ReMap database expansion. (**A**) Analyzed datasets growth in ReMap 2020 compared to 2018 and 2015 in human. (**B**) Transcriptional regulators (TRs) growth in ReMap 2020 compared to 2018 and 2015. (**C**, **D**) Evolution of the number of datasets across the top 50 TRs between ReMap 2020 and 2018. (**E**) Genome coverage fraction of each ReMap atlas by genomic regions covered by at least five peaks considered as regulatory (blue, light blue), potentially regulatory (grey), or not covered (light grey). (**F**) Analyzed TRs and Histones datasets in Arabidopsis. (**G**) Transcriptional regulators (TRs) in ReMap Arabidopsis. (**H, I**) Number of datasets for the top 30 TRs and top 5 biotypes. (**J**) Number of datasets for the top 10 Histone modifications and variants. (**K**) Genome coverage fraction of TRs and Histone ReMap catalogs, regions covered by at least five peaks considered regulatory (green, light green), potentially regulatory (grey), or not covered (light grey).

spread across almost all TRs when compared to ReMap 2018 (Figure 1C, light blue bars). More specifically, we observe that key TFs (e.g. ESR1, AR, FOXA1, TP53), transcriptional repressors (e.g. CTCF) and CRFs (e.g. BRD4) display larger data expansion than other DNA-binding proteins. Nevertheless, most of the top 50 TRs show additional datasets integrated in ReMap 2020 (Figure 1C, light blue bars). The top 10 most frequent biotypes correspond to the most common cell lines used in genomics (Figure 1D). The constant integration of a wide spectrum of cell lines and tissues will allow for a greater definition of the regulatory space across the genome. Indeed, ∼34.4% (106 Gb) of the human genome is covered by at least five features or more, and 17% (516 Mb) covered by 15 features or more (Figure 1E and Supplementary Table S4). As comparison, the ReMap 2018 catalog covered 19% (601 Mb) of the genome by at least five features, and 8% (257 Mb) with at least 15 or more features.

In this update, we expand the regulatory panorama revealing dense co-localized regulatory regions at unprecedented depth (Figure 2A). Indeed, the ReMap database shows an unprecedented landscape of the human regulatory abundance and complexity constituted by 165M binding regions forming 1.7M CRMs. The genomic organization of our atlas reveals dense co-localizations of peaks forming tight clusters of heterogeneous binding regions with variable TRs complexity (Figure 2A, Supplementary Figure S4). Since 2015 we highlight this regulatory complexity by observing the vicinity of the human ELAC1 promoter illustrating ReMap 2015, 2018 and 2020 catalog growth (n = 229; 1037; 2546 peaks respectively). We observe three clusters of peaks, one large at the promoter followed by two

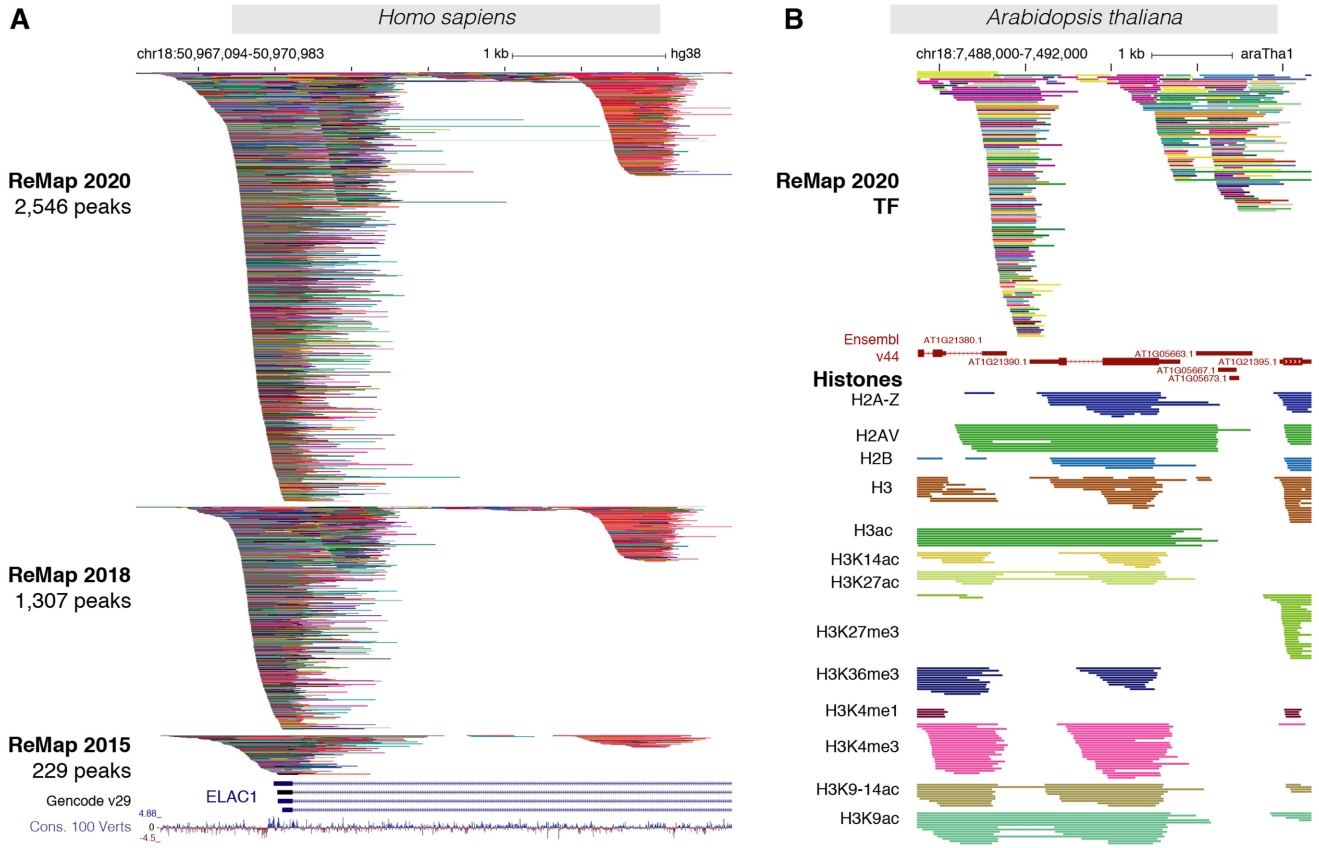

Figure 2. Genome browser views of both ReMap 2020 atlas. (**A**) ReMap 2020 human DNA–protein binding pattern of 5798 datasets. A genome browser example of the DNA-binding peak depth of the ReMap 2020 atlas compared to ReMap 2018 and 2015 at the vicinity of the ELAC1 promoter (hg38 chr18:50 967 094–50 970 983). The tracks displayed are compacted to thin lines so the depth of ReMap 2020 bindings can be compared to 2015. Uncompacted screenshot is available as Supplementary Figures S5. In this region ReMap 2020 displays 2546 peaks, 1307 peaks for ReMap2018, 229 peak for ReMap 2015 (lifted to GRCh38/hg38). The following genome tracks correspond to the GENCODE v29 annotation and the 100 vertebrates base-wise conservation showing sites predicted to be conserved (positive scores in blue). (**B**) A genome browser view of the first ReMap 2020 Arabidopsis TF and Histones modifications catalogs at the vicinity of the AT1G21390 gene (araTha1 chr1:7 488 000–7 492 000). The annotation genome track correspond to the Ensembl Genes v44 TAIR10 annotation. All peaks have been compacted for rendering and 13 (out of 33) histone modifications displayed.

clusters located at about +500 bp and +2 kb from the transcription start site. This third cluster has been detailed in our previous ReMap publications (4,8) to illustrate how integrating data from different sources improves genome annotations. Indeed this cluster contains two ENCODE peaks for FOXA1 in the 2020 update, and only one in previous versions. This update further consolidate the binding location with 93 FOXA1 peaks (60 peaks in 2018, 15 in 2015) across different cells, antibodies, and laboratories (Supplementary Figure S4). The summit of each peak is represented by a vertical bar, which when aggregated closely gives information about the putative location of the DNA binding site. This FOXA1 clustering shows overlapping peaks, and does not reveal the discrete repertoire of binding regions in the human genome. Therefore, to address redundancy between datasets, we merged peaks for the same TR, resulting in a catalog of 76M non-redundant peaks. The genomic accuracy of these non-redundant peaks have been improved with a new method to reduce peaks redundancy (see 'Materials and Methods' for details). Non-redundant peaks and CRMs annotations are computed across all ReMap datasets and biotypes, thus representing a multi-cellular multi-tissue regulatory map. Taken together, the 2020 update of the ReMap human catalog provides a unique opportunity to identify complex regulatory architectures in our genome, each containing hundreds or thousands of bound regulators. By adding more experiments and more DNA-binding proteins to the atlas, we increased the genome regulatory space (Figure 1), dramatically increased its depth (Figure 2), and refined current annotations of bound regions (Supplementary Figure S4).

FIRST ARABIDOPSIS THALIANA REGULATORY CATALOGS

Arabidopsis transcriptional regulators atlas

This ReMap release comes with the first large scale Arabidopsis regulatory atlas for transcription factors and general components of the transcriptional machinery. Indeed, the extent of the regulatory space in the Arabidopsis genome has not yet been fully apprehended with systematic integration of public DNA-binding assays. Thus, to enable genome-wide identification of Arabidopsis regulatory ele-

ments we have collected, curated, uniformly processed and analysed 204 ChIP-seq and 848 DAP-seq datasets against TRs from GEO (Figure 1F and Supplementary Figure S2). We retained 509 datasets after quality control: 179 ChIP-seq and 330 DAP-seq datasets leading to a final Arabidopsis regulatory atlas of 2 645 004 peaks for 372 TRs (Figure 1F and G). While a study (20) was carried out by integrating TF peaks from few ChIP-seq experiments in flowers development, our ReMap atlas is the first to provide a global view of all detected TRs binding in a wide variety of biological contexts and variety of experiments (Figure 2B). The top three most represented TRs are Argonaute protein AGO1 (mRNA and chromatin binding), MADS-box protein SVP (Transcription repressor), Heat stress transcription factor HSFA1B (Figure 1H), while the two most represented tissues are leaves and seedlings (Figure 1I). About 40% (49 Mb) of the Arabidopsis genome is covered by at least five features or more, and 19% (22 Mb) covered by 15 features or more (Figure 1K and Supplementary Table S4). We present here a unique transcription factors occupancy map forming complex architecture in the plant genome revealed by the first large scale integration of public Arabidopsis DNA-binding experiments.

Arabidopsis histone modifications atlas

The ReMap Arabidopsis integration efforts have been complemented with the first release of histone modifications catalogue as a result of the uniform integration of ChIP-seq datasets. Histone modifications and variants regulate gene expression by remodelling the chromatin structure thus acting on chromatin accessibility (21–23). The integration of the histone modifications assays can be used to attribute functional properties to genomic regions as histone modification positioning often correlate with genomic features (6,24). We have collected, curated, processed and analysed 358 ChIP-seq datasets against histone modifications (Figure 1F and Supplementary Figure S3). We retained 286 datasets after quality control leading to a final atlas of 4 528 203 peaks covered by 33 histone modifications and variants (Figures 1F, G, 2B, Supplementary Figure S7). The top three most represented histones modifications and variants are H3K27me3, H3 and H3K4me3 (Figure 1J), with seedlings and leaves being most common represented tissues (Supplementary Figure S8). As histone modifications remodel large and broad chromatin regions, 80% (96 Mb) of the Arabidopsis genome is covered by at least 5 features or more (Figure 1K and Supplementary Table S4). This ReMap update comes with two unique Arabidopsis regulatory catalogues, one providing 372 transcription factors and general actors of the transcriptional machinery and a second catalogue of 33 of known histone modifications and variants.

A NEWLY DESIGNED WEB PORTAL

A new web interface

With this ReMap 2020 update, a new web interface has been designed to incorporate many of the features and characteristics present in contemporary websites. The new interactive website is designed with Laravel, a PHP web application framework for the development of web applications following the model–view–controller (MVC) architectural pattern. As template we used AdminLTE, a responsive HTML template based on the CSS framework Bootstrap 3. We used MySQL as a backend database to store ReMap datasets, targets, biotypes informations and external metadata. The site can be divided into six main sections: Search, Target and Biotype pages, Browsing data, the ReMapEnrich R Shiny tool, RESTful API and About pages. We have greatly improved the search functionalities of ReMap (Figure 3A). The search bar with advanced search options has been introduced into the main page for direct access to the data. The new search interface searches among all the main metadata, target synonyms, biotype synonyms to identify potential matches. Search results are presented in a paginated table which can be further filtered by a filter box. We have improved the Target and Biotype pages data visibility and usability by organising collections of blocks. Each Target and Biotype page contains an interactive paginated table for downloading separate dataset as narrow or broad peak files from macs2 (Figure 3B). The 'Browsing' page and browsing vignettes assist in the visual exploration of the regulatory catalogues, allowing researchers to navigate data in genome browsers. Finally, the About pages provide detailed information for each species on the experiments, pre-processing steps, quality results, as well as basic statistics of the catalogues. We believe that the new web interface and features incorporated in ReMap 2020 have improved our users experience. ReMap 2020 can be accessed at http://remap.univ-amu.fr or http://remap.cisreg.eu.

New genomic enrichment interface

Since ReMap 2015 an annotation tool has been deployed allowing users to query their regions of interests against our catalogue to identify enrichments of TRs. However a better web interface and statistics to compute genomic region enrichment analyses against ReMap catalogs was much needed. Associated with this release we introduced the ReMapEnrich R Shiny interface (Figure 3C) to identify significantly enriched regions from user defined regions against ReMap catalogs. This interface facilitates the interpretation of functional genomics, epigenomics and genomics data by providing common statistical functions and plot the enrichment analyses of ReMap TRs.

RESTful API

We have developed a RESTful web service API to provide programmatic access to ReMap metadata as well as additional resources such as target, biotype informations, and BED files (Figure 3D). The API has been implemented for querying the database with a batch list of TR, and/or biotypes. The RESTful API enables various bioinformatics tasks through a broad range of client software to access ReMap data programmatically. This web services is a convenient solution to integrate in an automated manner heterogeneous genomic information in complex work-

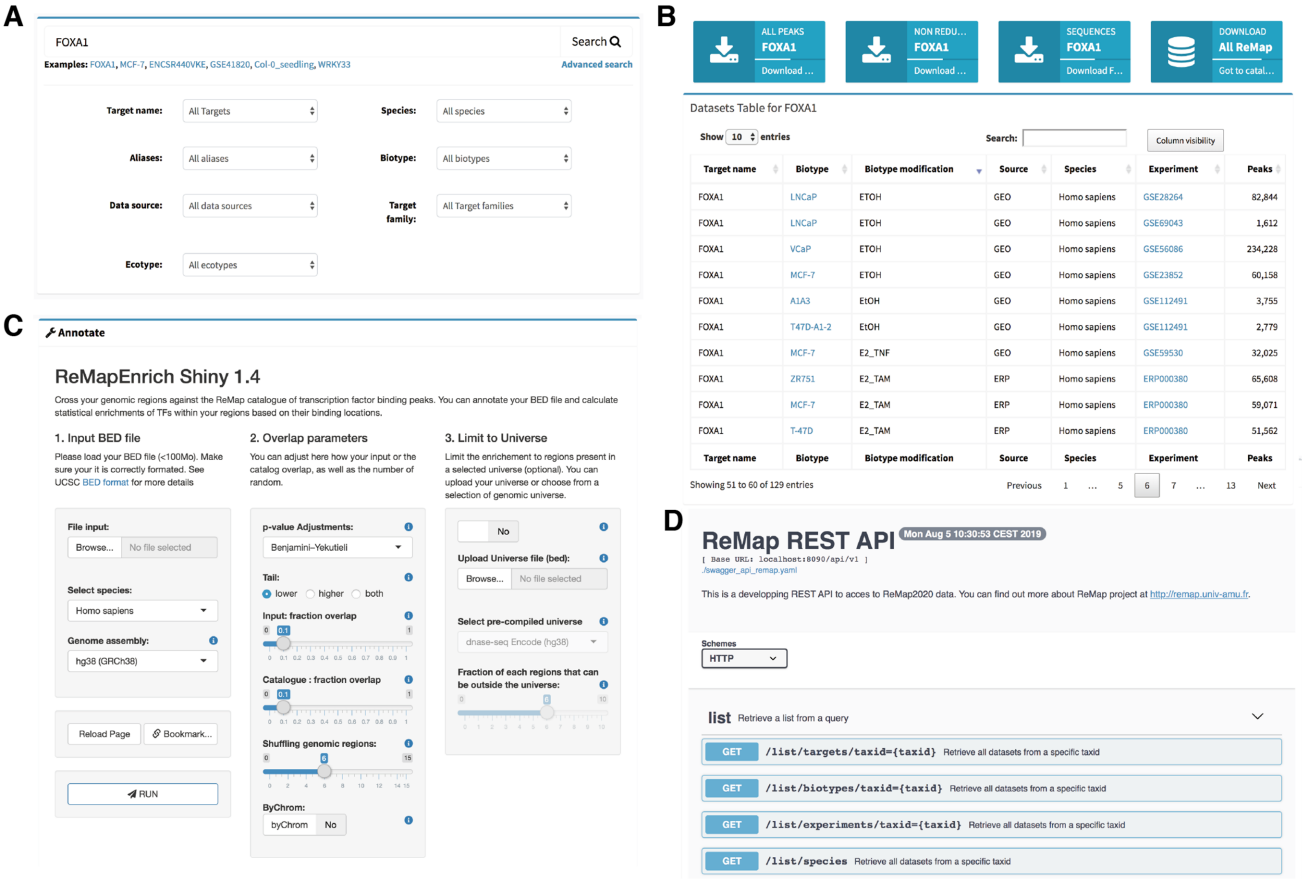

Figure 3. Overview of the ReMap 2020 new web interface with interactive searching activity. (**A**) A full search bar on the homepage and search page with advanced features. (**B**) For each targets or biotype page a responsive table allowing further filtering. (**C**) The ReMapEnrich R Shiny interface for plotting enrichment analyses. (**D**) List of queries available for the ReMap RESTful API.

flows without the need of local databases installations. The RESTful API is natively implemented in the Laravel Web Framework of ReMap, it returns the data in the convenient JSON format, and is accessible at http://remap.univ-amu.fr/rest_page.

Data download and genome browsers

Most users may prefer to download our atlas directly and query it locally, for instance to integrate with internal data or use in workflows. Each of the three ReMap database releases are available from the download page, where the entire catalogs are accessible in BED format in a variety of configurations: 'all' peaks, non-redundant peaks, CRM peaks. In addition, the download page is the entry point for searching and directly accessing BED files for specific targets, biotypes or datasets centric tables.

Since the 2018 update we provided data navigation options for visual exploration of the ReMap regulatory catalogs combined with public or user-specific genome-wide annotations. As the ReMap catalogs expand in size and complexity, and researchers routinely generate large scale genome-wide genomic data, it became crucial to provide navigation flexibility to centralise these data. The content of the different ReMap atlas can be browsed across major Genome Browsers such as Ensembl (25), Ensembl Genomes (12) and the UCSC Genome Browser (26) within public sessions or public hubs (Figure 2, Supplementary Figures S5–7). Additionally, for each ReMap species, track hubs have been deposited to the Track Hub Registry (https://trackhubregistry.org/) for open data integration in various platforms or browsers (27). Our goal is to facilitate researchers to discover the abundance and complexity of ReMap catalogs in combination with other biological tracks.

CONCLUSION AND FUTURE DIRECTIONS

The 2020 release of ReMap maintains the long-term focus of providing the research community with the largest catalogs of high-quality regulatory regions by integrating all available DNA-binding assays. The ReMap 2020 update comes with (i) a significant expansion of the human regulatory atlas; (ii) the first Arabidopsis regulatory atlas for transcriptional regulators; (iii) the first Arabidopsis catalog for histone modifications; (iv) a new website with improved user experience; (v) a new annotation tool with an R Shiny interface; (vi) a programmatic access to ReMap with a RESTful API; (vii) an updated genome browsing experience with Track Hubs data integration for different platforms. The

quality of ReMap data is illustrated in the latest version of the JASPAR database (28), for which peaks were used to calculate TF link profiles. We believe that our ReMap catalogs will help in better understanding the regulation processes in Human and Arabidopsis.

In the future, as new datasets are constantly added to repositories, we would like to engage the scientific community in the curation process to increase our capacity to introduce new regulatory profiles for different species and different DNA-binding assays. As a long term goal, we would like to call the regulatory community with the curation and production processes to provide the best annotated regulatory repertoire in different species.

FEEDBACK

The ReMap team welcomes your feedback on the catalogs, use of the website and use of the downloadable files. We thank our users for past and future feedback to make ReMap useful for the community. Please contact benoit.ballester@inserm.fr for development requests.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We would like to thank Lionel Spinelli for in depth scientific discussions in pipeline implementation and his expertise on Docker and Singularity usage with the HPC resource. We would like to thank Marius Gheorghe, Aziz Khan and Anthony Mathelier from NCMM Norway for constant scientific feedback, the UCSC Genome informatics groups for help with tack hubs, the Ensembl and Ensembl Plant group for help with the Homo sapiens and Arabidopsis thaliana track hub. This work was granted access to the HPC resources of Aix-Marseille Université financed by the project Equip@Meso [ANR-10-EQPX-29-01] of the program 'Investissements d'Avenir' supervised by the Agence Nationale de la Recherche.

FUNDING

French Ministry of Higher Education and Research (MESR) PhD Fellowship (to J.C.); Funding for open access charge: Institut National de la Santé et de la Recherche Médicale (INSERM).
Conflict of interest statement. None declared.

REFERENCES

1. Johnson,D.S., Mortazavi,A., Myers,R.M. and Wold,B. (2007) Genome-wide mapping of in vivo protein-DNA interactions. *Science*, **316**, 1497–1502.
2. Rhee,H.S. and Pugh,B.F. (2011) Comprehensive Genome-wide Protein-DNA interactions detected at single nucleotide resolution. *Cell*, **147**, 1408–1419.
3. Bartlett,A., O'Malley,R.C., Huang,S.-S.C., Galli,M., Nery,J.R., Gallavotti,A. and Ecker,J.R. (2017) Mapping genome-wide transcription-factor binding sites using DAP-seq. *Nat. Protoc.*, **12**, 1659–1672.
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. 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.* (2012) NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res.*, **41**, D991–D995.
6. ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
7. Davis,C.A., Hitz,B.C., Sloan,C.A., Chan,E.T., Davidson,J.M., Gabdank,I., Hilton,J.A., Jain,K., Baymuradov,U.K., Narayanan,A.K. *et al.* (2018) The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res.*, **46**, D794–D801.
8. Chèneby,J., Gheorghe,M., Artufel,M., Mathelier,A. and Ballester,B. (2018) ReMap 2018: an updated atlas of regulatory regions from an integrative analysis of DNA-binding ChIP-seq experiments. *Nucleic Acids Res.*, **46**, D267–D275.
9. Braschi,B., Denny,P., Gray,K., Jones,T., Seal,R., Tweedie,S., Yates,B. and Bruford,E. (2019) Genenames.org: the HGNC and VGNC resources in 2019. *Nucleic Acids Res.*, **47**, D786–D792.
10. Chang,A., Schomburg,I., Placzek,S., Jeske,L., Ulbrich,M., Xiao,M., Sensen,C.W. and Schomburg,D. (2015) BRENDA in 2015: exciting developments in its 25th year of existence. *Nucleic Acids Res.*, **43**, D439–D446.
11. Bairoch,A. (2018) The Cellosaurus, a cell-line knowledge resource. *J. Biomol. Tech. JBT*, **29**, 25–38.
12. Kersey,P.J., Allen,J.E., Allot,A., Barba,M., Boddu,S., Bolt,B.J., Carvalho-Silva,D., Christensen,M., Davis,P., Grabmueller,C. *et al.* (2018) Ensembl Genomes 2018: an integrated omics infrastructure for non-vertebrate species. *Nucleic Acids Res.*, **46**, D802–D808.
13. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.
14. 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.
15. Zhang,Y., Liu,T., Meyer,C.A., Eeckhoute,J., Johnson,D.S., Bernstein,B.E., Nussbaum,C., Myers,R.M., Brown,M. and Li,W. (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, **9**, R137.
16. 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.
17. Quinlan,A.R. (2014) BEDTools: The Swiss-army tool for genome feature analysis. *Curr. Protoc. Bioinforma.*, **47**, doi:10.1002/0471250953.bi1112s47.
18. 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.
19. Marinov,G.K., Kundaje,A., Park,P.J. and Wold,B.J. (2014) Large-scale quality analysis of published ChIP-seq data. *G3*, **4**, 209–223.
20. Chen,D., Yan,W., Fu,L.-Y. and Kaufmann,K. (2018) Architecture of gene regulatory networks controlling flower development in Arabidopsis thaliana. *Nat. Commun.*, **9**, 4534.
21. Bannister,A.J. and Kouzarides,T. (2011) Regulation of chromatin by histone modifications. *Cell Res.*, **21**, 381–395.
22. Venkatesh,S. and Workman,J.L. (2015) Histone exchange, chromatin structure and the regulation of transcription. *Nat. Rev. Mol. Cell Biol.*, **16**, 178–189.
23. Kouzarides,T. (2007) Chromatin modifications and their function. *Cell*, **128**, 693–705.
24. 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.
25. Cunningham,F., Achuthan,P., Akanni,W., Allen,J., Amode,M.R., Armean,I.M., Bennett,R., Bhai,J., Billis,K., Boddu,S. *et al.* (2019) Ensembl 2019. *Nucleic Acids Res.*, **47**, D745–D751.
26. Haeussler,M., Zweig,A.S., Tyner,C., Speir,M.L., Rosenbloom,K.R., Raney,B.J., Lee,C.M., Lee,B.T., Hinrichs,A.S., Gonzalez,J.N. *et al.*

(2019) The UCSC Genome Browser database: 2019 update. *Nucleic Acids Res.*, **47**, D853–D858.
27. 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.
28. Fornes,O., Castro-Mondragon,J.A., Khan,A., Lee,R. van der, Zhang,X., Richmond,P.A., Modi,B.P., Correard,S., Gheorghe,M., Baranašić,D. *et al.* (2019) JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, doi:10.1093/nar/gkz1001.