MetamORF: A repository of unique short Open Reading Frames identified by both experimental and computational approaches for gene-level and meta-analysis

Sebastien A. Choteau[1,2], Audrey Wagner[1], Philippe Pierre[2,3,4], Lionel Spinelli[1,2] and Christine Brun[1,5,*]

[1] Aix-Marseille Univ., INSERM, TAGC, Turing Centre for Living Systems, Marseille, France

[2] Aix-Marseille Univ., CNRS, INSERM, CIML, Turing Centre for Living Systems, Marseille, France

[3] Institute for Research in Biomedicine (iBiMED) and Ilidio Pinho Foundation, Department of Medical Sciences, University of Aveiro, 3810-193 Aveiro, Portugal.

[4] Shanghai Institute of Immunology, School of Medicine, Shanghai Jiao Tong University, Shanghai, China.

[5] CNRS, Marseille, France

*To whom correspondence should be addressed. Tel: +33 4 91 82 87 12; Email: christine-g.brun@inserm.fr

ABSTRACT

The development of high-throughput technologies revealed the existence of non-canonical short open reading frames (sORFs) on most eukaryotic RNAs. They are ubiquitous genetic elements highly conserved across species and suspected to be involved in numerous cellular processes. MetamORF (http://metamorf.hb.univ-amu.fr/) aims to provide a repository of unique sORFs identified in the human and mouse genomes with both experimental and computational approaches. By gathering publicly available sORF data, normalizing it and summarizing redundant information, we were able to identify a total of 1,162,675 unique sORFs. Despite the usual characterization of ORFs as short, upstream or downstream, there is currently no clear consensus regarding the definition of these categories. Thus, the data has been reprocessed using a normalized nomenclature. MetamORF enables new analyses at loci, gene, transcript and ORF levels, that should offer the possibility to address new questions regarding sORF functions in the future. The repository is available through an user-friendly web interface, allowing easy browsing, visualization, filtering over multiple criteria and export possibilities. sORFs could be searched starting from a gene, a transcript, an ORF ID, or looking in a genome area. The database content has also been made available through track hubs at UCSC Genome Browser.

INTRODUCTION

Short open reading frames (sORFs) are usually defined as sequences delimited by a start and a stop codon and potentially translatable into proteins of less than 100 amino acids (1–8). They are present on all classes of transcripts (including presumptive long non-coding RNAs) and have been identified on most eukaryotic RNAs (2, 5, 8–15). In addition, their sequence often begin with a non-canonical start codon (8). Consequently, they have long been overlooked and interest in their possible regulatory functions has only raised recently with the advent of the ribosome profiling method that strongly suggests their translation (1, 3, 5, 6, 16–22).

Several sORFs biotypes have been defined according to their location on RNAs. For instance, upstream ORFs (uORFs) are located in the 5' untranslated regions of mRNAs and have been defined as sORFs whose start codon precedes the main coding sequence (CDS) (6, 8, 17, 18, 23). They are conserved across species (5, 6, 11, 21, 24), although they seem to be less conserved than canonical protein-coding ORFs (25). To date, uORFs have been essentially reported as gene expression *cis*-regulatory elements that regulate the efficiency of translation initiation of the main CDS, notably alleviating the repression of translation during cellular stress (13, 17, 18, 20, 23, 26). Moreover, the discovery of uORF-, and more generally sORF-encoded peptides led to the assumption that they may also play functional roles in *trans* (2–4, 7, 9, 10, 18, 24, 27–30), for instance as ligands of major histocompatibility complex (MHC) molecules (12, 22, 23). Very interestingly, uORF-encoded peptides have also been shown to form protein complexes with the protein encoded by the main CDS of the same mRNA (31) and it has been suggested that polycistronic sequences may exist in Eukaryotes (24, 31). Furthermore, given the increasing evidence on the regulatory functions of peptides encoded by sORFs located within mRNAs, introns of pre-mRNAs, lncRNAs, primary transcripts of miRNAs or rRNAs (2, 8–15, 26), there is an urgent need to study sORFs *(i)* individually, and *(ii)* at the whole proteome scale. Indeed, the latter should reveal important sORFs features, thus enabling the characterization and the identification of their functions. However, the fact that *(i)* the publicly available data are scattered across different databases and *(ii)* datasets are aligned on different genome builds, differently annotated and formatted, calls for an uniformized resource where each sORF is individually described. With this in mind, we have built a resource database of publicly available sORFs identified in the human and mouse genomes, by gathering information from computational predictions, Ribo-seq and proteomic experiments. The curation of data, their homogenization in order to merge the redundant information into unique entries, the completion and computation of missing information (*e.g.*, sequences, Kozak contexts) and the re-annotation of sORF classes represent the added value of this database. Notably, this enables the analysis at locus, gene, transcript, and ORF levels, as well as groups of them. In this work, we propose *(i)* a pipeline to regularly update the content of the database in a reproducible manner, *(ii)* a database content that can be fully downloaded for custom computational analyses and *(iii)* a user-friendly web interface to ease data access to biologists.

MATERIAL AND METHODS

MetamORF pipeline and database development

Inclusion criteria for publicly available sORF-related data

A total of 18 data sources, either *H. sapiens* and *M. musculus* original datasets or re-processed publicly available sORFs repositories, have been considered for inclusion in our database (Supplementary Table S1) (5, 7, 11, 12, 14, 15, 17–22, 32–37). These data sources provide results from computational predictions, Ribo-seq experiment analyses and mass spectrometry (proteomics / proteogenomics) analyses. The data sources not providing the absolute genomic coordinates of the ORF start and stop codons (5, 17, 20, 32–34) or fully included in another data source considered here (21), have been discarded. Databases that did not allow export of their content in a single file or to automate the download of all the files from their website, have also been discarded (19, 35). Despite their short size, it has been noticed that sORFs can be spliced. Theoretical lengths of the ORFs have been computed as the distance between the start and stop codons, eventually removing the intron length(s) when information about the ORF splicing was provided. Due to splicing, the theoretical length and the one reported by the data source may be different. Data sources harboring this difference for more than 95% of their entries were discarded as this indicates the splicing information was missing (10). Finally, data sources for which we were not able to perform this assessment as they were not providing information regarding the splicing of the ORF and did not provide any ORF length (15, 36) have not been included as well. Hence, the database has been made by collecting data from six distinct sources (Figure 1), including either original datasets (Table 1 and Supplementary Table S2) (11, 12, 14, 18, 22) or reprocessed data (37), and discarding 12 of them (Supplementary Table S1). Notably, we have included data from sORFs.org (37), considered as the main and most comprehensive repository of sORFs identified by genome-wide translation profiling (Ribo-seq), that currently integrates re-processed data from 73 original publications.

For each of these sources, a set of features essential to properly characterize the sORFs, related to their location, length, sequences, environmental signatures and cell types (*i.e.* cell lines, tissues or organs) in which they are expressed, have been collected (see Table 2 for a full list of features considered for inclusion). When it was not provided by the source, the symbol of the gene related to the sORF was recovered using the transcript identifier (ID, if provided) or searching for the gene(s) or ncRNA(s) overlapping with the sORF coordinates in the original annotation version by querying Ensembl databases in their appropriate versions (v74, 75, 76, 80, 90) with pyensembl (v1.8.5, https://github.com/openvax/pyensembl). In addition to these features, information regarding the transcript(s) harboring the ORFs have been collected from the data sources when available. This is of particular interest as some ORF features, such as the ORF class, may depend on the transcript they are located on (*e.g.* an ORF may be located in the 5'UTR of a transcript and be overlapping with the CDS of another transcript). Finally, 3,379,219 and 2,066,627 entries from these six data sources have been collected and inserted in MetamORF for *H. sapiens* and *M. musculus*, respectively (Table 3).

Homogenization of genomic coordinates

As the data sources were providing genomic coordinates from different genome annotation versions (*e.g.* GRCh38 and GRCh37), all the genomic coordinates registered in our database have been lifted over the latest annotation version (GRCh38 for *H. sapiens*, GRCm38 for *M. musculus*) using pyliftover (v0.4, https://pypi.org/project/pyliftover). The lift-over has been considered as failed for an entry if *(i)* at least one of the coordinates (*i.e.* start, stop or one of the start or end exon coordinates) was located on a strand different from all the others, or *(ii)* the chromosome of the position changed during the lift-over, or *(iii)* the distance (in nucleotides) between the sORF start and stop codons has changed after the lift-over. All the entries for which the lift-over failed were removed from the database. Based on the previous assumptions, the lift-over failed for 709 ORFs (377 failed due to the last criteria) in *H. sapiens* and for none of the *M. musculus* entries (Table 3). The choice of such stringent criteria has been strengthened by the fact that these entries only represent less than 0.05% of the total number of entries for *H. sapiens* and are more susceptible to be unreliable entries.

Merge of redundant information

As our database aims to provide a repository of unique identified sORFs of the human and mouse genomes, all the redundant entries describing the same sORFs have been merged. In a first step, we identified all the sORF entries for which all the identification features were provided (chromosome, strand, start position, stop position, splicing status and splicing coordinates). sORFs sharing the same feature values were merged. In a second step, we identified all the remaining entries with only partial identification features provided: the chromosome as well as either *(i)* both the strand and the start position, or *(ii)* both the strand and the stop position, or *(iii)* both the start and the stop position. Those entries were merged to the best matching fully described entries identified in the first step. If no matching fully described entry was found, then the entries were removed. In order to keep track of the number of times a same sORF has been described in the original data sources, the initial number of entries merged together was registered for each sORF.

During this merging, information regarding the transcripts that harbor the sORFs have been registered too. Hence, when several sORFs were merged into one single entry in MetamORF, the resulting new sORF entry was registered as harbored by all the distinct transcripts related with the original entries. After this removal of redundant information, we were finally able to identify 664,771 and 497,904 unique sORFs for *H. sapiens* and *M. musculus,* respectively (Table 3).

It should be noticed that all unique sORF entries generated at this stage have been kept, including the ones describing ORFs longer than 100 amino acids. Entries describing such ORFs may be either coming from data sources that *(i)* did not remove the ORFs longer than 100 amino acids, or *(ii)* used a higher threshold or *(iii)* described the ORF as unspliced whilst it is actually susceptible to be spliced (and thus has a shorter sequence on the transcript than the one expected).

Completion of missing information and computation of relative coordinates

In the original data sources, the only information provided (when provided), on the transcripts was the transcript ID. Detailed information was retrieved from Ensembl databases (v90) through their REST API and inserted in our database: *(i)* the biotype, *(ii)* the transcript start and end genomic coordinates, *(iii)* the codon of the canonical coding sequence (CDS, for protein-coding transcripts only) start and stop genomic coordinates and *(iv)* the full nucleic sequence. In addition, the sequence flanking the start codon (20) has been recovered. As the sORF nucleic and amino acid sequences were not systematically provided by the data sources, these were downloaded from the Ensembl databases using their genomic coordinates.

Moreover, when the transcript ID was available, sORF start and stop relative coordinates have been computed on each of their transcript using AnnotationHub (v2.18.0, (39)) and ensembldb (v2.10.2, https://bioconductor.org/packages/release/bioc/html/ensembldb.html) R packages (R v3.6.0).

Standardization of the cell types and ORF classes

Cell types

Original data sources do not use a common thesaurus or ontology to name the cell types (*e.g.*, 'HFF' and 'Human Foreskin Fibroblast') or use non-biological meaning names (*e.g.*, sORFs.org (37) provides the name of the original publication as a cell type). In order to provide an uniform informative naming, we manually recovered the name of the cell line, tissue or organ used in these datasets and defined an unique name to be used in our database for each cell line, tissue or organ (Supplementary Table S3).

ORF classification

Despite the use of a common nomenclature by the wide majority of the scientific community to annotate the open reading frames, based on their relative position on their transcript (*e.g.*, short, upstream, downstream, overlapping), no clear consensus about the definitions of these categories nor their names has been defined so far (25). In order to homogenize this information in MetamORF, we created a new annotation of the ORFs using both the ORF length, transcript biotype, relative positions and reading frame information when available. In this annotation, a threshold of 100 amino acids has been used to define the "short ORFs", as this value is the most commonly used for historical reasons (2, 4, 6, 8, 24).

Computation of the Kozak contexts

The Kozak motif and context have been regarded as the optimal sequence context to initiate translation in all eukaryotes. We have thus assessed the Kozak context for each sORF, using the criteria defined in Hernandez et al. (40). Briefly, for each ORF to transcript association, the Kozak context was computed looking for regular expression characterizing an optimal, strong, moderate or weak Kozak context (Supplementary Tables S4 and S5). Kozak-alike contexts were also computed for non-ATG initiated sORFs looking for the same patterns with flexibility regarding nucleotides at +1 to +4 positions.

MetamORF softwares and languages

The pipeline used to build MetamORF has been developed using Python (v2.7) with SQLAlchemy ORM (sqlalchemy.org, v1.3.5). The database has been handled using MySQL (mysql.com, v8.0.16). Docker

(docker.com, v18.09.3) and Singularity (singularity.lbl.gov, v2.5.1) environments have been used in order to ensure reproducibility and to facilitate deployment on high-performance clusters (HPCs).

The MetamORF web interface has been developed using the Laravel (laravel.com, v7.14.1) framework with PHP (v7.3.0), JavaScript 9, HTML 5 and CSS 3. The NGINX (v1.17.10) web server PHP server (v7.3.0) were deployed with Docker (docker.com, v18.09.3) and Docker-compose (v1.24.0) to ensure stability.

DATABASE CONTENT, ACCESSIBILITY AND WEB INTERFACE

A new repository of short ORF-related data

MetamORF describes 664,771 and 497,904 unique ORFs in the human and mouse genomes respectively, providing at least the information necessary to locate the ORF on the genome, its sequence and the gene it is located on (excepted for intergenic ORFs). Extensive information related to the transcripts is provided respectively for 614,997 (~93%) and 497,904 (100%) sORFs for the human and mouse genomes respectively. These features allowed us to classify 630,953 (~95%) human ORFs and 497,904 (100%) mouse ORFs in at least one class (Table 3, Figure 2, Supplementary Figure S1). Interestingly, it should be noticed that a large proportion (36% and 52% respectively for *H. sapiens* and *M. musculus*) of ORFs are using an alternative frame to the main CDS. In addition, nearly 23% of the ORFs are located on non-coding RNAs for both species.

User-friendly web interface and genome tracks

To provide users with a clear, fast and easy-to-use database, MetamORF can be queried through an user-friendly web interface at http://metamorf.hb.univ-amu.fr. A tutorial as well as a documentation page are available online. Briefly, the users may search for sORFs contained in the database starting with a gene symbol (symbol, alias, ID), transcript ID (ID, name), ORF ID, or screening a particular genomic area. The data is made accessible through four types of pages: *(i)* a "gene-centric" page (Figure 3), allowing to visualize information related to all transcripts and sORFs on a gene, *(ii)* a "transcript-centric" page, allowing to browse information related to a transcript gene and all its sORFs, *(iii)* an "ORF-centric" page allowing to fetch information related to all transcripts and gene that harbor the chosen ORF and finally *(iv)* a "locus" page allowing to get information related to all sORFs located in a particular locus. It is possible to navigate from one to another page easily to get extensive information about an sORF, a gene or a transcript (Supplementary Figure S2).

In each page, the results can be filtered on: *(i)* the identification method (computational prediction, ribosome profiling or mass spectrometry), *(ii)* the start codon, *(iii)* the Kozak context (as previously defined), *(iv)* the genomic length (defined as the sum of lengths of each exon constituting the ORF), *(v)* the transcript biotype (according to the Ensembl definitions), *(vi)* the ORF annotation (as previously defined) and *(vii)* the cell type (Supplementary Table S3).

All results can be exported in an easily-parsable format (comma-separated values file, CSV), as well as in FASTA or BED format.

On ORF, transcript and locus pages, a link allowing to easily visualize all the ORFs localized in a particular area on the UCSC genome browser (41) is proposed. We also implemented genome track hubs, allowing to use UCSC genome browser advance options, such as filtering on ORF categories, transcript biotypes, cell types and transcript IDs.

In addition to this user-friendly interface, it is possible to download from the website the content of the full MetamORF database at BED, and FASTA formats.

DISCUSSION AND CONCLUSION

MetamORF contains data about 1,162,675 unique sORFs for the human and mouse genomes identified by both experimental and computational approaches. Whilst the Ribo-seq is considered by most as the "gold standard" method to identify sORFs experimentally, the added value of predictive computational approaches, proteomics and peptidomics to characterize such biological sequences remains certain. Because these technologies are offering complementary information at genomic, transcriptomic and proteomic scales, we decided to include data from both experimental and computational experiments in our database. Nevertheless, data coming from distinct data sources may be difficult to compare, in particular because they are not necessarily using the same genome annotation and definitions of ORF classes and Kozak contexts, for instance. By homogenizing this information, MetamORF offers the possibility to compare datasets coming from different sources. We noticed that information regarding the Kozak context is missing most of the time and start flanking sequences are usually not provided. Hence, MetamORF provides there a new interesting set of information.

It should be noticed that a large amount (~80 %) of the sORFs contained in our database have been described in the sORFs.org repository (37). Despite being the most prominent sORF database and offering the community data processed in a normalized way using their own workflow, it has already been highlighted that sORFs.org does not provide metagene analyses (1). In addition, such analysis is made difficult by the absence of gene names as well as the high redundancy of information contained in the sORF.org database (37), an issue we addressed with MetamORF. Hence, in comparison with existing resources, MetamORF allows analyses at ORF, transcript, gene and loci levels. In addition, it opens the possibility of studying sORFs as a group, at a global scale.

The resource is accessible at http://metamorf.hb.univ-amu.fr and provides an intuitive querying interface to enable wet lab researcher to easily question this large set of information. Moreover, the implementation of MetamORF content in track hubs allows both quick and advanced visualization of data through the UCSC genome browser. Finally, the database content may be exported at various convenient formats widely used by the scientific community (e.g. FASTA, BED).

We believe that MetamORF is of interest not only to bioinformaticians working on short ORFs but also to a wider community, including any biologist who may benefit from knowledge regarding the sORFs located on their gene, transcript or region of interest. As ribosome profiling becomes more appreciated and proteomics starts to allow accurate identification of short peptides, new data describing sORFs in various conditions are expected to be published in the next years, and our database is expected to grow accordingly. As a conclusion, we believe that MetamORF should help to address new questions in the future, in particular regarding the regulatory functions of the sORFs as well as the functions of the short peptides they may encode.

DATA AVAILABILITY

Data sources are available on the editor's website or using the links provided in their original publications. The source code used to create the database, and the full technical documentation (source code documentation, manual, database structure, dockerfiles) are available on GitHub (https://github.com/TAGC-NetworkBiology/MetamORF). Full content of the database can be downloaded at BED and FASTA formats from MetamORF website and up-to-date version of track hubs may be download and/or used with your favorite genome browser providing the following link: http://metamorf.hb.univ-amu.fr/hubDirectory/hub.txt. The dump of the database is available on request.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

FUNDING

The project leading to this publication has received funding from the « Investissements d'Avenir » French Government program managed by the French National Research Agency (ANR-16-CONV-0001) and from Excellence Initiative of Aix-Marseille University.

CONFLICT OF INTEREST

The authors have no conflict of interest to declare.

ACKNOWLEDGEMENT

We thank Andreas Zanzoni for helpful discussions.

REFERENCES

1. Martinez,T.F., Chu,Q., Donaldson,C., Tan,D., Shokhirev,M.N. and Saghatelian,A. (2019) Accurate annotation of human protein-coding small open reading frames. *Nat. Chem. Biol.*, 10.1038/s41589-019-0425-0.

2. Andrews,S.J. and Rothnagel,J.A. (2014) Emerging evidence for functional peptides encoded by short open reading frames. *Nat. Rev. Genet.*, **15**, 193–204.

3. Aspden,J.L., Eyre-Walker,Y.C., Phillips,R.J., Amin,U., Mumtaz,M.A.S., Brocard,M. and Couso,J.-P. (2014) Extensive translation of small Open Reading Frames revealed by Poly-Ribo-Seq. *eLife*, **3**.

4. Saghatelian,A. and Couso,J.P. (2015) Discovery and characterization of smORF-encoded bioactive polypeptides. *Nat. Chem. Biol.*, **11**, 909–916.

5. Chew,G.-L., Pauli,A. and Schier,A.F. (2016) Conservation of uORF repressiveness and sequence features in mouse, human and zebrafish. *Nat. Commun.*, **7**, 11663.

6. Ingolia,N.T., Brar,G.A., Stern-Ginossar,N., Harris,M.S., Talhouarne,G.J.S., Jackson,S.E., Wills,M.R. and Weissman,J.S. (2014) Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. *Cell Rep.*, **8**, 1365–1379.

7. Hao,Y., Zhang,L., Niu,Y., Cai,T., Luo,J., He,S., Zhang,B., Zhang,D., Qin,Y., Yang,F., *et al.* (2018) SmProt: a database of small proteins encoded by annotated coding and non-coding RNA loci. *Brief. Bioinform.*, **19**, 636–643.

8. Pueyo,J.I., Magny,E.G. and Couso,J.P. (2016) New peptides under the s(ORF)ace of the genome. *Trends Biochem. Sci.*, **41**, 665–678.

9. Raj,A., Wang,S.H., Shim,H., Harpak,A., Li,Y.I., Engelmann,B., Stephens,M., Gilad,Y. and Pritchard,J.K. (2016) Thousands of novel translated open reading frames in humans inferred by ribosome footprint profiling. *eLife*, **5**, e13328.

10. Zanet,J., Chanut-Delalande,H., Plaza,S. and Payre,F. (2016) Small Peptides as Newcomers in the Control of Drosophila Development. *Curr. Top. Dev. Biol.*, **117**, 199–219.

11. Mackowiak,S.D., Zauber,H., Bielow,C., Thiel,D., Kutz,K., Calviello,L., Mastrobuoni,G., Rajewsky,N., Kempa,S., Selbach,M., *et al.* (2015) Extensive identification and analysis of conserved small ORFs in animals. *Genome Biol.*, **16**.

12. Laumont,C.M., Daouda,T., Laverdure,J.-P., Bonneil,É., Caron-Lizotte,O., Hardy,M.-P., Granados,D.P., Durette,C., Lemieux,S., Thibault,P., *et al.* (2016) Global proteogenomic analysis of human MHC class I-associated peptides derived from non-canonical reading frames. *Nat. Commun.*, **7**, 10238.

13. Couso,J.-P. and Patraquim,P. (2017) Classification and function of small open reading frames. *Nat. Rev. Mol.*, **18**, 575–589.

14. Samandi,S., Roy,A.V., Delcourt,V., Lucier,J.-F., Gagnon,J., Beaudoin,M.C., Vanderperre,B., Breton,M.-A., Motard,J., Jacques,J.-F., *et al.* (2017) Deep transcriptome annotation enables the discovery and functional characterization of cryptic small proteins. *eLife*, **6**.

15. McGillivray,P., Ault,R., Pawashe,M., Kitchen,R., Balasubramanian,S. and Gerstein,M. (2018) A comprehensive catalog of predicted functional upstream open reading frames in humans. *Nucleic Acids Res.*, **46**, 3326–3338.

16. Olexiouk,V., Crappé,J., Verbruggen,S., Verhegen,K., Martens,L. and Menschaert,G. (2016) sORFs.org: a repository of small ORFs identified by ribosome profiling. *Nucleic Acids Res.*, **44**, D324–D329.

17. Wethmar,K., Barbosa-Silva,A., Andrade-Navarro,M.A. and Leutz,A. (2014) uORFdb--a comprehensive literature database on eukaryotic uORF biology. *Nucleic Acids Res.*, **42**, D60-67.

18. Johnstone,T.G., Bazzini,A.A. and Giraldez,A.J. (2016) Upstream ORFs are prevalent translational repressors in vertebrates. *EMBO J.*, **35**, 706–723.

19. Wang,H., Yang,L., Wang,Y., Chen,L., Li,H. and Xie,Z. (2019) RPFdb v2.0: an updated database for genome-wide information of translated mRNA generated from ribosome profiling. *Nucleic Acids Res.*, **47**, D230–D234.

20. Andreev,D.E., Arnold,M., Kiniry,S.J., Loughran,G., Michel,A.M., Rachinskii,D. and Baranov,P.V. (2018) TASEP modelling provides a parsimonious explanation for the ability of a single uORF to derepress translation during the integrated stress response. *eLife*, **7**.

21. Fields,A.P., Rodriguez,E.H., Jovanovic,M., Stern-Ginossar,N., Haas,B.J., Mertins,P., Raychowdhury,R., Hacohen,N., Carr,S.A., Ingolia,N.T., *et al.* (2015) A regression-based analysis of ribosome-profiling data reveals a conserved complexity to mammalian translation. *Mol. Cell.*, **60**, 816–827.

22. Erhard,F., Halenius,A., Zimmermann,C., L'Hernault,A., Kowalewski,D.J., Weekes,M.P., Stevanovic,S., Zimmer,R. and Dölken,L. (2018) Improved Ribo-seq enables identification of cryptic translation events. *Nat. Methods*, **15**, 363–366.

23. Starck,S.R., Tsai,J.C., Chen,K., Shodiya,M., Wang,L., Yahiro,K., Martins-Green,M., Shastri,N. and Walter,P. (2016) Translation from the 5' untranslated region shapes the integrated stress response. *Science*, **351**, aad3867.

24. Crappé,J., Van Criekinge,W. and Menschaert,G. (2014) Little things make big things happen: A summary of micropeptide encoding genes. *EuPA Open Proteom.*, **3**, 128–137.

25. Orr,M.W., Mao,Y., Storz,G. and Qian,S.-B. (2019) Alternative ORFs and small ORFs: shedding light on the dark proteome. *Nucleic Acids Res.*, 10.1093/nar/gkz734.

26. Plaza,S., Menschaert,G. and Payre,F. (2017) In Search of Lost Small Peptides. *Annu. Rev. Cell Dev. Biol.*, **33**, 391–416.

27. Hazarika,R.R., Sostaric,N., Sun,Y. and van Noort,V. (2018) Large-scale docking predicts that sORF-encoded peptides may function through protein-peptide interactions in Arabidopsis thaliana. *PloS One*, **13**, e0205179.

28. Andreev,D.E., O'Connor,P.BF., Fahey,C., Kenny,E.M., Terenin,I.M., Dmitriev,S.E., Cormican,P., Morris,D.W., Shatsky,I.N. and Baranov,P.V. (2015) Translation of 5′ leaders is pervasive in genes resistant to eIF2 repression. *eLife*, **4**.

29. Zanet,J., Benrabah,E., Li,T., Pélissier-Monier,A., Chanut-Delalande,H., Ronsin,B., Bellen,H.J., Payre,F. and Plaza,S. (2015) Pri sORF peptides induce selective proteasome-mediated protein processing. *Science*, **349**, 1356–1358.

30. Cabrera-Quio,L.E., Herberg,S. and Pauli,A. (2016) Decoding sORF translation – from small proteins to gene regulation. *RNA Biol.*, **13**, 1051–1059.

31. Chen,J., Brunner,A.-D., Cogan,J.Z., Nuñez,J.K., Fields,A.P., Adamson,B., Itzhak,D.N., Li,J.Y., Mann,M., Leonetti,M.D., *et al.* (2020) Pervasive functional translation of noncanonical human open reading frames. *Science*, **367**, 1140–1146.

32. Rodriguez,C.M., Chun,S.Y., Mills,R.E. and Todd,P.K. (2019) Translation of upstream open reading frames in a model of neuronal differentiation. *BMC Genomics*, **20**, 391.

33. Sharipov,R.N., Yevshin,I.S., Kondrakhin,Y.V. and Volkova,O.A. (2014) RiboSeqDB – a repository of selected human and mouse ribosome footprint and RNA-seq data. *Virtual Biology*, **1**, 37-46–46.

34. Evans,V.C., Barker,G., Heesom,K.J., Fan,J., Bessant,C. and Matthews,D.A. (2012) De novo derivation of proteomes from transcriptomes for transcript and protein identification. *Nat. Methods*, **9**, 1207–1211.

35. Liu,W., Xiang,L., Zheng,T., Jin,J. and Zhang,G. (2018) TranslatomeDB: a comprehensive database and cloud-based analysis platform for translatome sequencing data. *Nucleic Acids Res.*, **46**, D206–D212.

36. Lee,S., Liu,B., Lee,S., Huang,S.-X., Shen,B. and Qian,S.-B. (2012) Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, E2424-2432.

37. Olexiouk,V., Van Criekinge,W. and Menschaert,G. (2018) An update on sORFs.org: a repository of small ORFs identified by ribosome profiling. *Nucleic Acids Res.*, **46**, D497–D502.

38. Yates,A.D., Achuthan,P., Akanni,W., Allen,J., Allen,J., Alvarez-Jarreta,J., Amode,M.R., Armean,I.M., Azov,A.G., Bennett,R., *et al.* (2020) Ensembl 2020. *Nucleic Acids Res.*, **48**, D682–D688.

39. Hansen,K.D., Sabunciyan,S., Langmead,B., Nagy,N., Curley,R., Klein,G., Klein,E., Salamon,D. and Feinberg,A.P. (2014) AnnotationHub: Large-scale hypomethylated blocks associated with Epstein-Barr virus-induced B-cell immortalization. *Genome Res.*, **24**, 177–184.

40. Hernández,G., Osnaya,V.G. and Pérez-Martínez,X. (2019) Conservation and Variability of the AUG Initiation Codon Context in Eukaryotes. *Trends Biochem. Sci.*, **44**, 1009–1021.

41. Kent,W.J., Sugnet,C.W., Furey,T.S., Roskin,K.M., Pringle,T.H., Zahler,A.M. and Haussler, a. D. (2002) The Human Genome Browser at UCSC. *Genome Res.*, **12**, 996–1006.

FIGURE LEGENDS

Figure 1 | MetamORF pipeline. This figure represents the workflow used to build MetamORF. First, the data from the sources selected has been inserted in the database and the absolute genomic coordinates have been homogenized from their original annotation version to the most recent (GRCh38 or GRCm38). Then the redundant information, *i.e.* the entries describing the same ORFs (same start, stop and splicing), have been merged, allowing to get one single and unique entry for each ORF detected on the human and mouse genomes. The missing information (sequences, transcript biotype) have been downloaded from Ensembl and the ORF relative coordinates have been computed. Finally the cell types and ORF classes have been normalized and the Kozak contexts have been computed using the sequences flanking the start codons.

Figure 2 | Count of ORFs in each class. The barplot represent the count of ORFs annotated for each class for (A) *H. sapiens* and (B) *M.* musculus. The percentages displayed over the bars indicates the proportion of ORFs annotated in the class over the total number of ORFs registered in the database for the species.

Figure 3 | MetamORF gene-centric view. The page displays the transcripts and the ORFs related to SGK3 gene. A filter has been applied to select exclusively the ORFs detected in HFF, Jurkat, RPE-1, HEK293 or HeLa cells. Other filters may be used and the results can be exported as CSV, FASTA or BED files.

TABLES

Table 1 | Information about the data sources used to build MetamORF. See supp. table S1 for more information about these data sources.

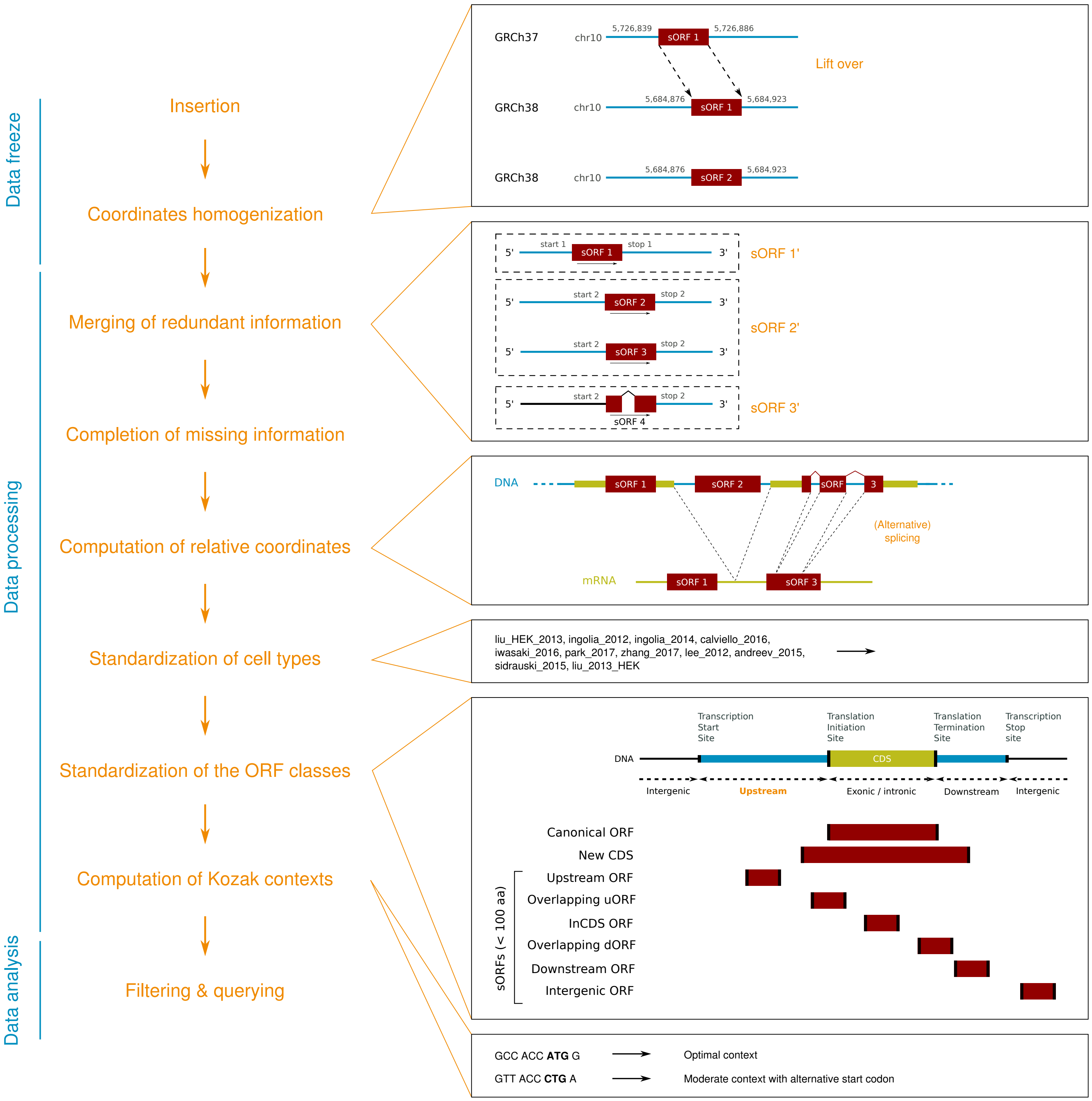
Publication	DOI
Mackowiak et al., 2015, Genome Biol. (11)	10.1186/s13059-015-0742-x
Erhard et al., 2018, Nat. Meth. (22)	10.1038/nmeth.4631
Johnstone et al., 2016, EMBO (18)	10.15252/embj.201592759
Laumont et al., 2016, Nat. Commun. (12)	10.1038/ncomms10238
Samandi et al., 2017, eLife (14)	10.7554/eLife.27860
Olexiouk et al., 2018, Nucl. Ac. Res. (37)	10.1093/nar/gkx1130

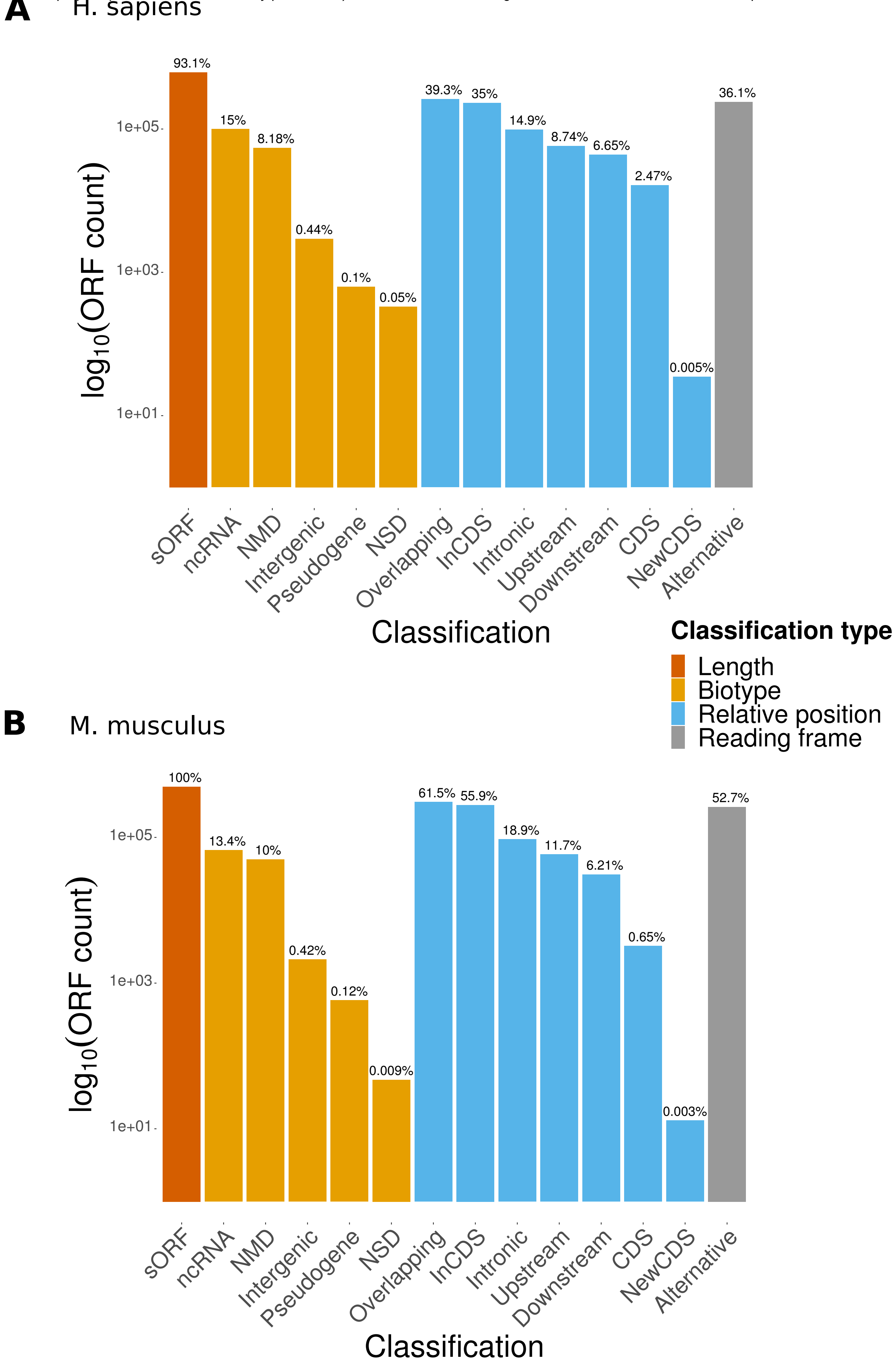
Table 2 | Features allowing to characterize the sORFs.

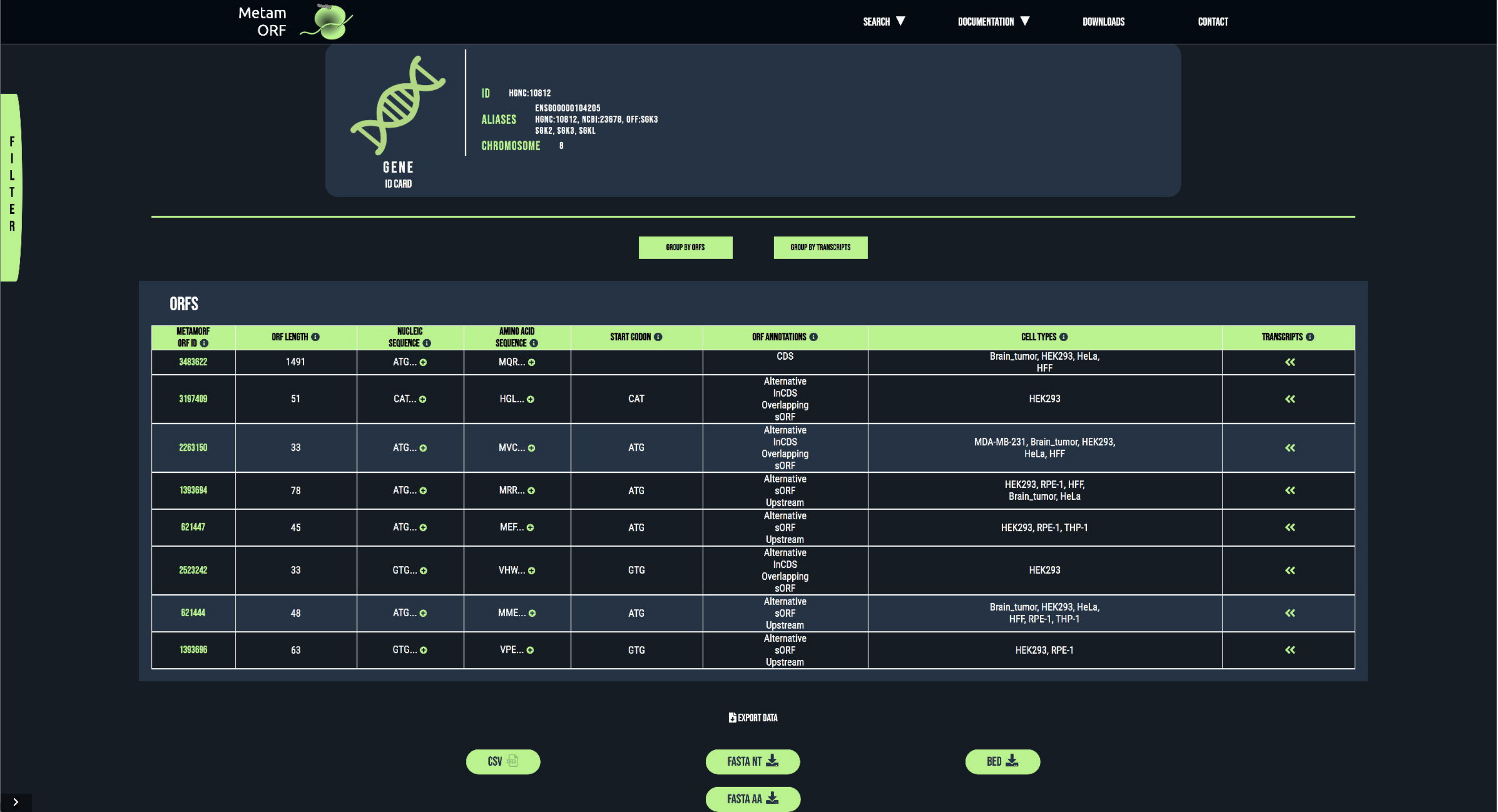
Family	Feature	Details
Location	Chromosome	The chromosome or scaffold on which the ORF is located
	Strand	The strand of the sORF
	Start	The absolute genomic coordinates of the start codon (position of the 1st nucleotide)
	ORF stop	The absolute genomic coordinates of the stop codon (position of the 3rd nucleotide)
	Splicing status	Is the sORF spliced?
	Splicing coordinates	The coordinates of the start and end of each exon constituting the sORF
	Transcript	The name or ID of the transcript(s) related to the sORF (eventually with transcript strand, start and end positions and biotype)
	Gene	The name, symbol, alias or ID of the gene(s) related to the sORF (when not intergenic)
Lengths	Length	The length of the sORF (in nucleotides)
	Putative sPEP length	The length of the (putative) sORF-encoded peptide in amino acids
Category	Category	The category to which the sORF belongs to (e.g. upstream or downstream)
Sequence signature	Start codon sequence	The nucleic sequence of the sORF start codon
	Nucleic sequence	The nucleic sequence of the sORF
	Amino acid sequence	The amino acid sequence of the (putative) sORF-encoded peptide
Environmental signature	Kozak context	Does a Kozak context has been identified for the sORF start codon?
Conservation	PhyloCSF score	The PhyloCSF score computed for the sORF
	PhastCons score	The PhastCons score computed for the sORF
Coding potential assessment	FLOSS class and score	The FLOSS class and score computed for the sORF
	ORF score	The ORF score computed for the sORF
Biological context	Cell Context	The cellular context in which the sORF has been identified or detected

Table 3 | MetamORF most important statistics.

Feature		H. sapiens	M. musculus
Original data sources	**ORFs**	**1,344,978**	**1,249,176**
	Transcripts	101,597	85,653
	Predicted ORFs for which the transcript is missing	181,122	213,301
	ORFs detected by Ribo-seq for which the transcript is unknown	79,422	8,546
	ORFs detected by MS for which the transcript is unknown	54	0
	ORF to transcript associations	**3,379,219**	**2,066,627**
	ORFs predicted	202,309	222,705
	ORFs identified by ribosome profiling	1,142,669	1,026,471
	ORFs identified by mass spectrometry	166	0
ORFs for which the homogeneization of genomic coordinates failed		709	0
MetamORF database	**ORFs**	**664,771**	**497,904**
	Transcripts	90,406	63,147
	Predicted ORFs for which the transcript is missing	13,440	14,327
	ORFs detected by Ribo-seq for which the transcript is unknown	71,158	2
	ORFs detected by MS for which the transcript is unknown	48	0
	ORF for which the transcripts are missing	83,403	14,329
	ORF to transcript associations	**729,793**	**696,785**
	ORFs predicted	17,027	14,500
	ORFs identified by ribosome profiling	664,771	497,904
	ORFs identified by mass spectrometry	147	0
Genes harboring at least 1 sORF		23,767	15,869
ORFs having at least one class annotation (short, upstream...)		630,953	497,904


Data freeze
Data processing
Data analysis
Insertion
Coordinates homogenization
Merging of redundant information
Completion of missing information
Computation of relative coordinates
Standardization of cell types
Standardization of the ORF classes
Computation of Kozak contexts
Filtering & querying
GRCh37
chr10
5,726,839
sORF 1
5,726,886
Lift over
GRCh38
chr10
5,684,876
sORF 1
5,684,923
GRCh38
chr10
5,684,876
sORF 2
5,684,923
5'
start 1
sORF 1
stop 1
3'
sORF 1'
5'
start 2
sORF 2
stop 2
3'
5'
start 2
sORF 3
stop 2
3'
sORF 2'
5'
start 2
stop 2
3'
sORF 4
sORF 3'
DNA
sORF 1
sORF 2
sORF
3
(Alternative) splicing
mRNA
sORF 1
sORF 3
liu_HEK_2013, ingolia_2012, ingolia_2014, calviello_2016, iwasaki_2016, park_2017, zhang_2017, lee_2012, andreev_2015, sidrauski_2015, liu_2013_HEK
Transcription Start Site
Translation Initiation Site
Translation Termination Site
Transcription Stop site
DNA
CDS
Intergenic
Upstream
Exonic / intronic
Downstream
Intergenic
Canonical ORF
New CDS
Upstream ORF
Overlapping uORF
InCDS ORF
Overlapping dORF
Downstream ORF
Intergenic ORF
sORFs (< 100 aa)
GCC ACC ATG G
Optimal context
GTT ACC CTG A
Moderate context with alternative start codon


A
H. sapiens
log10(ORF count)
1e+05
1e+03
1e+01
93.1%
15%
8.18%
0.44%
0.1%
0.05%
39.3%
35%
14.9%
8.74%
6.65%
2.47%
0.005%
36.1%
sORF
ncRNA
NMD
Intergenic
Pseudogene
NSD
Overlapping
InCDS
Intronic
Upstream
Downstream
CDS
NewCDS
Alternative
Classification
Classification type
Length
Biotype
Relative position
Reading frame
B
M. musculus
log10(ORF count)
1e+05
1e+03
1e+01
100%
13.4%
10%
0.42%
0.12%
0.009%
61.5%
55.9%
18.9%
11.7%
6.21%
0.65%
0.003%
52.7%
sORF
ncRNA
NMD
Intergenic
Pseudogene
NSD
Overlapping
InCDS
Intronic
Upstream
Downstream
CDS
NewCDS
Alternative
Classification



GENE
ID CARD

ID HGNC:10812

ALIASES ENSG00000104205
HGNC:10812, NCBI:23678, OFF:SGK3
SGK2, SGK3, SGKL

CHROMOSOME 8

ORFS

METAMORF ORF ID	ORF LENGTH	NUCLEIC SEQUENCE	AMINO ACID SEQUENCE	START CODON	ORF ANNOTATIONS	CELL TYPES	TRANSCRIPTS
3483622	1491	ATG...	MQR...		CDS	Brain_tumor, HEK293, HeLa, HFF	«
3197409	51	CAT...	HGL...	CAT	Alternative InCDS Overlapping sORF	HEK293	«
2263150	33	ATG...	MVC...	ATG	Alternative InCDS Overlapping sORF	MDA-MB-231, Brain_tumor, HEK293, HeLa, HFF	«
1393694	78	ATG...	MRR...	ATG	Alternative sORF Upstream	HEK293, RPE-1, HFF, Brain_tumor, HeLa	«
621447	45	ATG...	MEF...	ATG	Alternative sORF Upstream	HEK293, RPE-1, THP-1	«
2523242	33	GTG...	VHW...	GTG	Alternative InCDS Overlapping sORF	HEK293	«
621444	48	ATG...	MME...	ATG	Alternative sORF Upstream	Brain_tumor, HEK293, HeLa, HFF, RPE-1, THP-1	«
1393696	63	GTG...	VPE...	GTG	Alternative sORF Upstream	HEK293, RPE-1	«

EXPORT DATA