

# Investigating open reading frames in known and novel transcripts using ORFanage

Received: 23 March 2023

Accepted: 5 July 2023

Published online: 31 July 2023



Ales Varabyou<sup>1,2</sup>✉, Beril Erdogan<sup>1,3</sup>, Steven L. Salzberg<sup>1,2,3,4</sup> & Mihaela Pertea<sup>1,2,3</sup>✉

ORFanage is a system designed to assign open reading frames (ORFs) to known and novel gene transcripts while maximizing similarity to annotated proteins. The primary intended use of ORFanage is the identification of ORFs in the assembled results of RNA-sequencing experiments, a capability that most transcriptome assembly methods do not have. Our experiments demonstrate how ORFanage can be used to find novel protein variants in RNA-seq datasets, and to improve the annotations of ORFs in tens of thousands of transcript models in the human annotation databases. Through its implementation of a highly accurate and efficient pseudo-alignment algorithm, ORFanage is substantially faster than other ORF annotation methods, enabling its application to very large datasets. When used to analyze transcriptome assemblies, ORFanage can aid in the separation of signal from transcriptional noise and the identification of likely functional transcript variants, ultimately advancing our understanding of biology and medicine.

Approximately 20,000 protein-coding genes have been annotated for the human genome<sup>1–5</sup>. Although a single isoform is often the source of the dominant protein<sup>6–8</sup>, many human gene loci express isoforms that encode different protein sequences, some of which may be tissue-specific<sup>9–12</sup>. The NCBI RefSeq database, for example, contains an average of 6.9 isoforms for each human protein-coding gene, which encode an average of 4.4 distinct protein sequences. The RefSeq annotation of the model organism *Arabidopsis thaliana* has on average 1.8 isoforms with 1.5 unique protein variants, respectively.

RNA-sequencing (RNA-seq) technology has allowed an unprecedented look at the transcriptome in a wide variety of species, with multiple studies reporting large numbers of previously unknown transcripts for protein-coding genes<sup>3,13–16</sup>. Consistent with previous reports about alternative splicing events<sup>17</sup>, most of the novel transcripts reported in RNA-seq studies are observed in protein-coding regions<sup>18,19</sup>. Alternative splicing events can alter the translated protein through exon skipping, frame-shifting and other changes<sup>20</sup>. These events and their effects on translated proteins are an essential component of genome biology<sup>9</sup>.

Changes in protein sequences may also be characteristic of disease states<sup>10,21–24</sup> or of specific tissues<sup>9,25,26</sup>. For example, splicing-induced changes in protein sequences have been associated with cancer development and progression, from activation of proto-oncogenes<sup>27</sup> to genome-wide splicing alteration in certain cancer types<sup>28,29</sup>. One example of why it is important to annotate all protein isoforms in the human genome is the widespread usage of exome sequencing in clinical settings. Exome capture methods have been extensively used to interrogate genetic variants and their associations with diseases, such as finding the genetic cause of a rare form of pediatric epilepsy<sup>30</sup>, or identifying driver mutations in cancer<sup>31</sup>. The technology is heavily dependent on the correct annotation of coding regions, and any exons that are unannotated will simply be missed by exome studies.

However, many observed novel transcripts are likely to represent transcriptional noise<sup>32</sup>; for example, the original CHES database assembled ~29 million transcript variants from 10,000 RNA-seq experiments, of which fewer than 2% were kept in the final annotation<sup>3,4</sup>. The ability to accurately identify non-functional isoforms can be a valuable tool in differentiating signal from noise in RNA-seq data, which

<sup>1</sup>Center for Computational Biology, Johns Hopkins University, Baltimore, MD, USA. <sup>2</sup>Department of Computer Science, Johns Hopkins University, Baltimore, MD, USA. <sup>3</sup>Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, USA. <sup>4</sup>Department of Biostatistics, Johns Hopkins University, Baltimore, MD, USA. ✉e-mail: [ales.varabyou@jhu.edu](mailto:ales.varabyou@jhu.edu); [mpertea@jhu.edu](mailto:mpertea@jhu.edu)

**Table 1 | Summary of differences between ORFs found by ORFanage and the originally annotated ORFs for all transcripts in RefSeq and GENCODE protein-coding genes**

| Reference annotation                                                                        | RefSeq  | GENCODE |
|---------------------------------------------------------------------------------------------|---------|---------|
| ORFanage finds the same ORF as reference                                                    | 117,212 | 63,966  |
| ORFanage finds a different ORF that matches MANE perfectly                                  | 2,212   | 786     |
| No ORF annotated on reference transcript, ORFanage finds an ORF that matches MANE           | 1,194   | 147     |
| No ORF annotated on reference transcript, ORFanage finds an ORF that is different from MANE | 9,240   | 35,393  |
| Other combinations                                                                          | 5,836   | 27,994  |
| Total number of protein-coding transcripts                                                  | 135,694 | 128,286 |

Comparisons to the MANE annotation refer to the ORFs from the MANE gene set, which is fully contained within both RefSeq and GENCODE.

is currently complicated by artifacts from computational methods, such as alignment and assembly errors, as well as the amount of noise inherently present in the data<sup>32</sup>.

Although many methods have been implemented for searching and assembling transcripts from RNA-seq data<sup>33,34</sup>, none of them identifies open reading frames (ORFs) based on similarity to the original protein at the locus. A number of methods, including TransDecoder<sup>34</sup> and GeneMarkS-T<sup>35,36</sup>, have been developed for ab initio ORF annotation (Methods and Table 2), but these methods were designed to find ORFs without the use of reference annotation as a guide. Other previous approaches only identified the longest ORF, sometimes requiring it to have the same start or stop codon positions as a reference<sup>35,37–39</sup>. None of these approaches consider the similarity of the resulting protein to previously known translations of the transcript.

In this Article we present ORFanage, a highly efficient and sensitive method to search for ORFs in protein-coding transcripts, guided by reference annotation to maximize protein similarity within genes.

Results

Accuracy of reference ORF reconstruction

ORFanage utilizes protein-coding gene annotation by identifying ORFs in query transcripts that have the maximal sequence identity with a user-provided set of reference ORFs. This approach presumes that proteins produced by different transcripts at the same locus should be as similar as possible<sup>8,40</sup>. In our first set of experiments, we tested the ability of ORFanage to reconstruct the GENCODE and RefSeq protein-coding annotation given an annotation that includes one canonical ORF at each protein-coding gene locus. For these experiments, we used the MANE database to define the canonical ORFs, because MANE was created by the developers of GENCODE and RefSeq to be a ‘universal standard’<sup>6</sup> of human protein-coding genes, and because both GENCODE and RefSeq contain every gene in MANE. These experiments illustrate how ORFanage can produce a set of ORFs at a locus that better agree with a chosen reference annotation, conserving the protein sequences and making annotation more internally consistent.

As shown in Fig. 1a, many gene transcripts in both RefSeq and GENCODE are annotated with ORFs that differ from the canonical variant; for example, 65% of ORFs in the RefSeq human annotation and 36% in GENCODE differ from the MANE ORF (Fig. 1a). In principle, the presence of an ORF that differs from MANE does not imply an error; however, if another ORF can be found in the same transcript that has closer identity to MANE, then an error seems possible. Furthermore, 8% of RefSeq and 43% of GENCODE transcripts in protein-coding loci have no ORFs annotated at all. By re-annotating each of the reference datasets using ORFanage, we identified numerous cases where a different ORF was more similar to the canonical protein. One example, from the *ZNF180* gene, is shown in Fig. 1f.

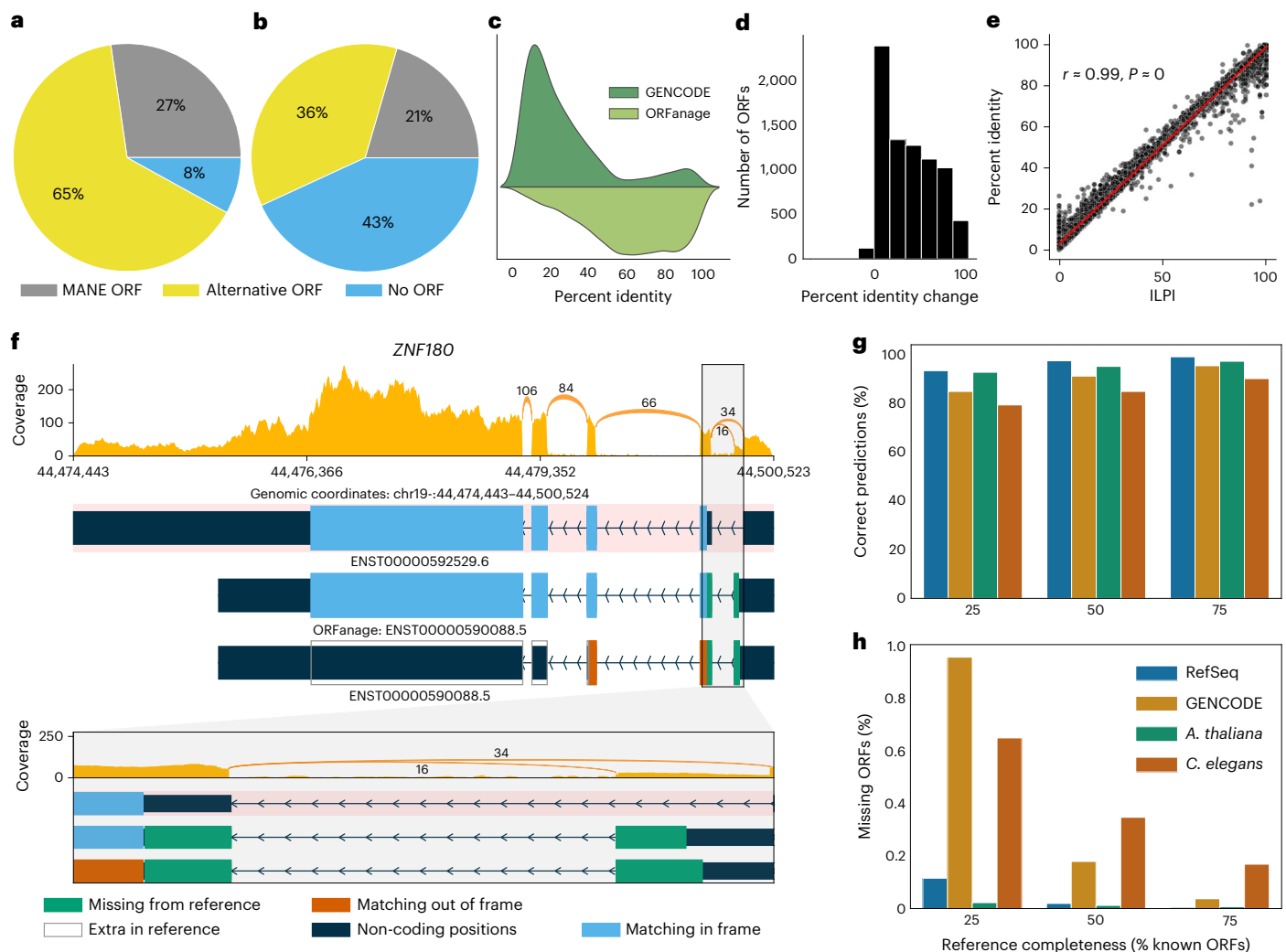
Although we found that ORFs in a large majority of transcripts in the RefSeq human annotation were in agreement with those predicted by ORFanage (117,212 out of 135,694), there were some striking differences, as illustrated in Table 1 and Supplementary Table 2. For example, we identified 2,122 transcripts in which an ORF annotated by RefSeq could be replaced by the canonical version from MANE without alterations. Similarly, 786 of the ORFs in the GENCODE human annotation could be replaced by their canonical variants from MANE. Even though alternative translations may be present at those transcripts, because GENCODE and RefSeq both recognize MANE as a standard<sup>6</sup>, it seems appropriate to choose the MANE ORFs over the alternative variants in accordance with established curation guidelines<sup>41</sup>.

In our analysis we purposefully refrained from filtering candidate ORFs, opting to report one best candidate ORF for every transcript where some sequence similarity was observed to the reference annotation. This allowed us to investigate all cases where analyzed annotations were inconsistent with the MANE reference at the cost of potential false discoveries. However, our software provides users with the ability to fine-tune the results through parameter settings such as the percent identity score, matching the translation initiation site, and other customizable criteria. These options enable users to refine the identification of valid ORFs and limit the number of false positives.

As a result, we also found thousands of transcripts for which no ORF was listed, even though they were annotated under protein-coding genes and even though a candidate ORF was identified by ORFanage (examples are provided in Supplementary Figs. 3 and 4). In GENCODE, we found an ORF that at least partially overlapped the MANE ORF in 35,540 out of 55,328 of these transcripts, including 147 transcripts that contained a perfect match to the MANE ORF. Although the RefSeq database had fewer protein-coding transcripts with no ORF listed, we still found 10,434 transcripts for which our method predicted an ORF, including 1,194 with a perfect match to MANE (Table 1 and Supplementary Table 2).

We also looked at transcripts where both ORFanage and the reference annotation differed from MANE (5,301 in RefSeq and 7,957 in GENCODE). For these transcripts we computed the percentage of in-frame positions shared between the annotated proteins and the MANE protein and observed that in 613 RefSeq and 7,005 GENCODE transcripts, ORFanage produced a protein that was closer to MANE (Fig. 1c,d). In many cases the differences were minor, affecting only start coordinates or conserving different segments of the reference protein. In some cases, though, such as *ZNF180*, as shown in Fig. 1f, ORFanage identified an ORF that conserved nearly all of the MANE protein sequence, while the protein encoded by the GENCODE ORF had no overlap with MANE. However, higher similarity of ORFs is not the only criterion for assessing ORF validity and other methods may be necessary to validate any novel sequences. Yet, in the absence of additional data, the similarity criterion can be successfully applied, as shown in our evaluation.

When ORFanage found an ORF that differed from the one chosen by RefSeq or GENCODE, the ORFanage sequence had an equal or higher proportion of codons that matched MANE (Fig. 1c,d), a property that is guaranteed by the algorithm. We confirmed these results by performing global alignments of the proteins to the MANE variants using EMBOSS Stretcher<sup>42</sup>. The higher percent identity is a consequence of the metric that ORFanage maximizes, which we term ‘in-frame length percent identity’ (ILPI). Following ORF identification via the algorithm described in Fig. 2, to compute ILPI, our method first computes the total number of positions in an ORF that are in the same frame as the reference, thus coding for the same codons, which determines the in-frame length (IL). The ILPI is then computed as the fraction of IL of the total length of the reference coding sequence (CDS). As illustrated in Fig. 1e, the correlation between ILPI and percent identity computed via the Smith–Waterman algorithm is very high.



**Fig. 1 | Overview of irregularities in reference database ORF annotation.**

**a,b**, Differences in ORFs at MANE loci as currently annotated for RefSeq (**a**) and GENCODE (**b**) annotations. Circular charts show, for each dataset, the proportions of transcripts annotated with the same ORF as MANE (gray), those with an alternative ORF not matching MANE (yellow), and transcripts in MANE loci that lack an annotated ORF (blue). **c**, Percent identity computed between the MANE protein and alternative ORFs as predicted by GENCODE (dark green) and ORFanage (light green). **d**, Histogram of the change in percent identity when replacing the GENCODE ORF with the ORFanage ORF. **e**, Pearson correlation coefficient ( $r$ ) and  $P$  value (two-sided,  $t$ -distribution) between percent identity computed via traditional alignment and ILPI computed by ORFanage, illustrating the close similarity between the two metrics (10,000 random samples).

**f**, A detailed look at alternative ORFs annotated by GENCODE and ORFanage for the *ZNF180* gene. At the top is the MANE isoform, shaded in pink, with its ORF shown in blue. Below it are two versions of an alternative isoform, with the ORFs annotated by ORFanage (middle) and GENCODE (bottom). Blue regions show where the protein sequence matches the MANE isoform, and green and orange show regions that are additional (green) or out of frame (orange) compared to MANE. At the bottom is a zoomed-in view of the first intron and flanking ORF regions. **g,h**, Overview of the impact that completeness of reference annotation has on the accuracy of ORFanage: percent of correctly inferred ORFs given different fractions of known reference ORFs for four organisms (**g**) and percentage of known ORFs that ORFanage failed to identify for different levels of reference completeness (**h**).

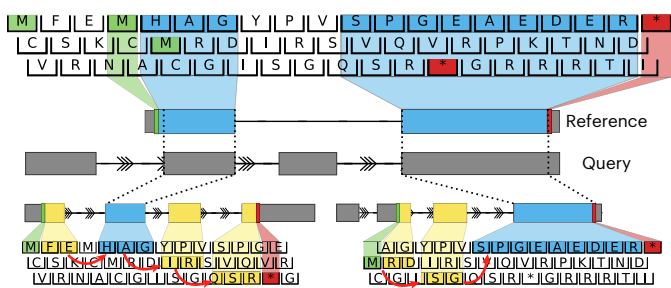
We then took a closer look at the 44,532 GENCODE transcripts where ORFanage found a different ORF. We found that ORFs identified by ORFanage often contained many novel positions (that is, not matching MANE). More specifically, nearly 22% of positions in these novel ORFs are marked as potentially coding only by our method, and although many of these positions could be artifacts of partial transcript models included in the GENCODE annotation, some are likely to represent new functional variants of known proteins<sup>14,23</sup>.

It is also worth noting here that when guided by protein-coding annotation such as MANE, ORFanage can reconstruct the ORFs present in GENCODE or RefSeq faster and more accurately than *ab initio* ORF finders like TransDecoder or GeneMarkS-T (Table 2 and Methods).

The large number of missing annotations and overall observed improvements demonstrate the potential use of ORFanage at finding consistent ORFs in novel transcripts at protein-coding loci.

### Impact of reference transcripts on accuracy

In the next set of experiments, we set out to investigate how well our method can reconstruct a full set of protein sequences from subsets of reference data. We wanted to establish (1) how accuracy improves with an increase in the number of annotated ORFs at a locus and (2) the effects of choosing different subsets of known ORF variants on the accuracy of prediction. To answer the first question, we incrementally increased the number of ORFs provided to ORFanage as a reference. To address the second question, we repeated the experiment but randomly chose different sets of reference ORFs.



**Fig. 2 | Diagram illustrating the algorithm implemented in ORFanage.** ORFanage begins by computing overlaps between a reference ORF and query transcript. In the figure, dashed lines are used to connect matching intervals. For each overlap it extends coordinates towards the 3' and 5' ends based on suitable parameters. During extension, any changes to the exon structure may introduce shifting of the original frame (as indicated by red arrows). Once all intervals have been evaluated, ORFanage compares the results and reports the one with the highest score. In the figure, residues matching the reference are highlighted in blue, and mismatching residues are highlighted in yellow. In this example, ORFanage selects the longer ORF on the lower right, which has 10 out of 14 matching residues, compared to the ORF on the lower left, with only 3 out of 14 matching residues.

We repeated the iterative selection of reference transcripts ten times, providing 25%, 50% and 75% of the reference ORFs as a guide each time. We ran our analysis on the human genome annotation as well as *A. thaliana* and *Caenorhabditis elegans* using the same protocol. For the human genome, we evaluated both RefSeq and GENCODE, because the two databases differ substantially in their ORF annotations. For each test run, we ensured that at least one transcript remained unannotated at each locus and that any non-coding transcripts were removed before the evaluation.

The diversity of transcripts annotated for *A. thaliana* and *C. elegans* is much lower than for human reference annotations, with 1.8 and 1.4 transcripts per coding gene, respectively, compared to six and eight for RefSeq and GENCODE human annotations. Worth noting is that for *C. elegans*, only 4,440 suitable loci were identified based on the aforementioned criteria.

As expected, we observed an increase in accuracy as more reference annotation was provided. For the human genome, if we provided just a single reference ORF per locus (equivalent to 11% of all ORFs in RefSeq and 18% of all ORFs in GENCODE), ORFanage was able to correctly recreate 85% of the RefSeq ORFs and 81% of GENCODE ORFs. When we provided 75% of the reference ORFs, ORFanage correctly recreated close to 99% of RefSeq and 95% of GENCODE ORFs (Fig. 1g,h).

Even when ORFs were not identical to the original sources, the predictions produced by ORFanage were highly similar, averaging 81% for the non-identical predictions in the RefSeq dataset and 77% in GENCODE, respectively.

Because *A. thaliana* and *C. elegans* have fewer annotated reference ORFs per locus, our random permutations had smaller effects on the results. Nevertheless, in *A. thaliana*, ORFanage was able to correctly identify 91–97% of reference ORFs. For *C. elegans* the values were lower, ranging from 77% when a single random reference ORF was provided to 90% when guided by more complete annotations.

### Finding novel ORFs in assembled RNA-seq data

One of the main applications of ORFanage is to search for ORFs in datasets containing large numbers of transcripts that have not been assigned ORFs. ORFanage can annotate transcriptome assemblies from RNA-seq experiments, which often contain many novel splice variants, even for well-annotated genomes<sup>3,4,43</sup>. In these cases, ORFanage can identify candidate ORFs for protein-coding transcripts based on the

**Table 2 | Comparison of the true positive rate (TPR) of ORF annotation methods based on concordance with the GENCODE and RefSeq datasets**

|              | GENCODE              |                 | RefSeq               |                 |                |                 |
|--------------|----------------------|-----------------|----------------------|-----------------|----------------|-----------------|
|              | Execution time (min) |                 | Execution time (min) |                 | TPR            |                 |
|              | Multi-threaded       | Single-threaded | Multi-threaded       | Single-threaded | Multi-threaded | Single-threaded |
| ORFanage     | 0.28                 | 0.6             | 0.88                 | 0.33            | 1.1            | 0.94            |
| TransDecoder | 115                  | –               | 0.65                 | 175             | –              | 0.82            |
| GeneMarkS-T  | 100                  | 100             | 0.58                 | 85              | 85             | 0.71            |

TPR was computed as the percentage of all ORFs in each dataset that were reconstructed identically by the method. GeneMarkS-T times do not include conversion from reported format to genomic gene transfer format style.

conservation of known protein sequences at the locus using reference annotation as a guide.

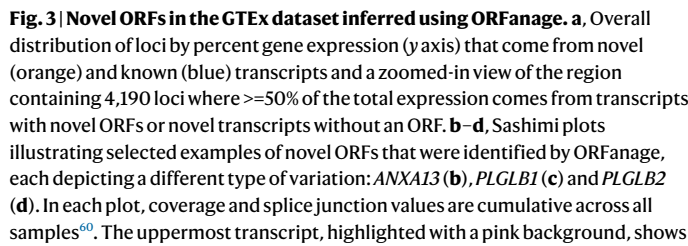
We next applied ORFanage to search for novel ORFs in experimental data, using data from the GTEx project<sup>44</sup>, a high-quality collection of poly-A selected RNA-seq samples across multiple human tissue types. We focused our experiments on 1,448 samples from brain tissue because these represented the most diverse collection of samples in the dataset. We ran ORFanage on the complete, unfiltered set of assemblies containing 6,674,316 isoforms that were assembled originally for the CHES human annotation database<sup>3,4</sup>.

We computed ORFs for all transcripts using the MANE annotation as the guide. For every MANE gene, we first identified all assembled transcripts overlapping that gene using gffcompare<sup>45</sup> (similarity codes ‘=’, ‘c’, ‘k’, ‘m’, ‘n’, ‘j’, ‘e’), yielding 4,256,346 transcripts. We then computed the total gene expression for each transcript using the sum of transcripts per million (TPM) values for that transcript across all samples.

In our search for novel ORFs, we took a conservative approach: if a transcript could accommodate an ORF from either RefSeq or GENCODE, we assigned that ORF to the transcript. Additionally, we removed ORFanage predictions for all transcripts marked as non-coding by either RefSeq or GENCODE. Because multiple distinct transcripts can contain the same novel ORF, we simplified our analysis by computing the total TPM aggregated across transcripts sharing the same ORF. In transcripts for which no ORF was assigned, we computed the total TPM as the sum of TPMs for that transcript across all samples. This selection left us with a total of 3,046,286 novel transcripts representing 1,006,547 ORF variants.

Next, to focus on highly expressed cases, we considered 4,190 loci where more than 50% of expression came from novel transcripts and ORFs (Fig. 3a). Many of the transcripts at these loci either had no valid ORF or else contained an ORF that was highly dissimilar from the canonical MANE protein. We therefore narrowed our focus to 462 loci where over 50% of expression was due to a single novel ORF. Of those, only 24 loci (Supplementary Table 1) were at least 70% identical to the MANE protein and had cumulative expression greater than 1,000 TPM across all samples (Fig. 3b–d and Supplementary Figs. 1 and 3). For example, in the *PLGLB* gene, an exon skipping event via a novel intron leads to the loss of the original start codon and a different, slightly longer N-terminal amino-acid sequence. Interestingly, we observed very similar exon skipping events in two different paralogs of this gene, *PLGLB1* and *PLGLB2*, as shown in Fig. 3c,d. In both cases, the alternative protein contains a different initial coding exon that replaces exon 1 of the MANE isoform, and in both cases, the majority of the expression comes from the alternative (non-MANE) isoform, suggesting that the MANE isoforms are not the dominant ones.

Another striking example of a novel ORF among these 24 loci occurs in the *ANXA13* gene (Fig. 3b,e), which is a member of the family of annexin genes responsible for the production of calcium-dependent



Nature Computational Science

comprehensive annotations of genomes of various ancestries that include not only transcripts but coding regions as well.

ORFanage can be used in conjunction with RNA-seq alignment and assembly to identify ORFs in novel transcripts, and to guarantee that those ORFs match the reference annotation as closely as possible. Whether using long-read alignments directly or assembled transcripts, this approach can uncover valuable insights into isoforms within protein-coding regions, leading to a better understanding of their effects on biological systems. And because RNA-seq datasets often produce large numbers of novel transcripts, the efficiency and scalability of ORFanage make it suitable for datasets of any size. We have recently applied our method to annotate ORFs in novel transcripts for the revised CHES3<sup>4</sup> catalog, and to help identify novel structurally stable isoforms that were then confirmed using AlphaFold2<sup>56</sup>.

ORFanage can also be a valuable aid to isolating the true signal from noisy transcriptome data. Assuming that proteins produced from alternative transcripts need to remain similar for genes to function correctly<sup>56</sup>, the ORF structures in the observed isoforms should be similar as well. Our approach can identify transcripts that cannot accommodate a similar ORF to the reference, serving as a noise-filtering step in RNA-seq analysis.

## Methods

ORFanage is based on the direct comparison of intervals that make up the exonic structures of query and reference transcripts. This optimization technique does not require sequence alignment or pre-computed genome indices, greatly reducing the computational burden of running the tool and making the analysis far more efficient than an alignment-based approach. We have tested ORFanage on datasets comprising tens of millions of transcripts assembled from thousands of RNA-seq experiments<sup>3,4</sup> and found that it runs robustly on these data.

### Creating bundles of transcripts

ORFanage operates on ‘bundles’ of data, defined as the union of a set of overlapping reference ORFs with a set of query transcripts that overlap one or more of the reference ORFs. To reduce the impact of annotation errors such as readthrough transcription, ORFanage only loads CDS coordinates for each reference transcript, discarding non-coding exonic coordinates.

Once both the reference and query datasets are loaded into memory and sorted internally, bundling is done in linear time by iterating over transcripts and collecting groups of all overlapping transcripts. This technique is insensitive to any information on gene boundaries, and readthrough transcription, commonly present in RNA-seq assemblies, may lead to several genes being combined into a single locus. In some cases, genes may genuinely overlap, and in such instances ORFanage might compare the ORFs of unrelated genes, possibly leading to incorrect inferences. To combat this problem, ORFanage gives users the option to group transcripts by gene IDs.

### Interval comparison

For each query transcript in a bundle, ORFanage performs a comparison to each reference CDS. For each pair being compared, an intersection is computed to identify all intervals that belong to both the query and the reference. The process is performed for all reference transcripts, and duplicate intervals are removed.

After a set of candidate overlaps is identified, ORFanage continues to search for the optimal start and end coordinates for each interval, discarding any incomplete ORFs in the process. We define a valid ORF as an uninterrupted sequence of three-base codons that begins with a start codon (usually ATG in humans), ends with a stop codon (TAA, TAG and TGA in humans), and does not contain any other stop codons other than the final one. Although only one valid stop codon can be found by extending any given ORF, multiple start (ATG) codons may

be present in a single ORF. In ORFanage, an optimal start codon is the one that maximizes the number of bases that are in the same frame as the reference ORF while minimizing the number of coordinates that do not match or that match out-of-frame (Fig. 2).

After all intervals have been examined, if multiple distinct ORFs are plausible, ORFanage performs a heuristic selection of the optimal ORF based on a series of configurable steps. Internally, for every unique ORF, the software computes three scores, which are applied successively to each set of candidate ORFs to find the best result:

- The IL, defined as the number of positions that are shared with the reference in the same coding frame
- ILPI, defined as the fraction of IL with respect to the length of the reference ORF
- The length of the ORF

When maximizing ILPI, ORFanage will prioritize ORFs that have as little novel sequence present as possible, where ‘novel’ is defined as sequence that is not present in the reference ORF. When maximizing IL instead, ORFanage might select longer ORFs with more novel sequence if that choice increases the number of matches with the reference. Alternatively, users may specify via optional parameters that conserving the position of the start codon takes priority over conservation of the remaining protein sequence, forcing the algorithm to select ORFs whose start codon matches the reference protein whenever possible. Worth noting here is that 568 out of 19,058 ORFs in the MANE database use a start codon that is not the longest ORF.

### Additional parameters

As shown in our analysis, the ILPI metric is an effective function to assess which ORF to pick for a given isoform and corresponds closely to percent identity. It is not identical to the familiar percent identity measure, which would be more expensive to compute. For applications that might require it, ORFanage includes support for computing a Smith–Waterman alignment between the reference ORF and the ORFs identified by ORFanage, as part of the final validation of the ORFs. ORFanage also includes an option to measure evolutionary conservation of any ORF by computing PhyloCSF scores. This option is implemented via an integrated PhyloCSF++ module<sup>57,58</sup>. Finally, ORFanage contains a multi-threading option, under which it can process each bundle in parallel, further speeding its runtime. In our tests, ORFanage was able to process 4,256,346 collected from 1,448 brain samples of the GTEx dataset, using the MANE annotation as a reference, in 7 min using 24 cores of an Intel Xeon 6248R 3-GHz processor, with all other parameters set to defaults. A random individual sample from the same dataset (SRR598396) was processed in 8 s.

### Datasets

Studies of the human genome account for a large proportion of transcriptomic data being generated today, and several annotation databases are available for these studies. For our evaluation of ORFanage on the human genome, we used both the RefSeq (release 110) and GENCODE (release 41) annotations<sup>1,2</sup>.

To investigate the utility of ORFanage on other organisms, we focused on the well-studied *A. thaliana* and *C. elegans* genomes, both of which have highly curated annotations of the transcriptome and proteome. Because, for each of these two genomes, only a single reference annotation was available, we chose to investigate how well ORFanage could reconstruct the ORFs using a bootstrapping technique, which allowed us to evaluate the concordance of annotated ORFs with the ones inferred by ORFanage.

For our evaluations on GTEx data, we used 1,448 poly-A selected RNA-seq samples representing 13 brain regions (age ≥ 20 years) from GTEx release 7<sup>44</sup>. Samples were aligned with HISAT2<sup>59</sup>, assembled with StringTie 2<sup>33</sup>, and merged with gffcompare. Coverage and splice

junction summaries were extracted using the TieBrush suite<sup>60</sup>.

### Data preparation

Although ORFAnage can handle several types of exception to the normal rules governing ORFs, such as alternative (non-ATG) start codons, selenoproteins and otherwise overlapping genes, for our evaluations we removed these exceptions so as to measure the accuracy on genes that conform to standard rules.

We began by choosing a set of genes to be used as a reference for human annotation. The MANE database<sup>6</sup> was created by the developers of RefSeq and GENCODE as a resource of human genes where both databases agree precisely on the complete exon–intron structure as well as on the CDS of every gene in the database. MANE contains one canonical transcript for nearly every protein-coding gene, plus a small number (62 in release 1.0) of medically relevant transcripts that differ from the canonical ones. In our reference set, we included all genes in MANE except for (1) genes with non-ATG start codons, (2) selenoproteins and (3) polycistronic genes. For our evaluations of both RefSeq and GENCODE, we retained only transcripts corresponding to the remaining MANE genes.

In some cases, manual curation might have altered RefSeq or GENCODE to create unusual ORFs. For example, some partial transcripts have been manually curated to show usage of an alternative start codon, despite other ORFs at the locus containing a canonical start codon. Because we do not know whether such exceptions are intentional, we decided to avoid penalizing RefSeq or GENCODE and filtered out such cases, as follows. First, we used *gffread*<sup>45</sup> to identify and remove all transcripts that did not contain valid start and stop codons. Second, we searched for all pairs of overlapping ORFs that were labeled with different gene IDs and removed all such occurrences. In addition, for the RefSeq dataset we also removed 846 genes that had transcripts with known exceptions as annotated by NCBI. In the end, our filtering resulted in the removal of 1,423 genes out of 20,442 genes from RefSeq (release 110) and 1,869 genes out of 20,427 from GENCODE (version 41).

For the *A. thaliana* and *C. elegans* annotation datasets<sup>61,62</sup>, we used the primary model organism annotation as the reference, after filtering out genes with non-ATG start codons, selenoproteins and polycistronic genes.

### Comparison of ORF-finding methods

To evaluate the various ORF annotation methods against ORFAnage, we generated two transcript sets from the GENCODE and RefSeq datasets using the process outlined in the Data preparation section. For each set, we created two files: one with the original ORFs preserved, and another with only transcript models, devoid of any CDS records. The first file served as our control, and the second was used as input for all ORF annotation methods. The detailed results of this comparison are provided in Table 2.

Worth noting here is that both TransDecoder and GeneMarkS-T ORF-finding methods used in our comparison are designed for finding ORFs de novo, without a need for guide annotation, and as such serve a different niche of applications than ORFAnage (for example, annotation of species for which no previous annotation is available).

**TransDecoder.** TransDecoder, part of the Trinity package<sup>34</sup>, can also find ORFs in a set of transcripts. Although originally intended as an ab initio method for finding ORFs in de novo transcriptome assemblies, the results can be improved by using homology searching against a protein database of choice. Since its original release, the software has been adapted for use with transcript models that are assembled by programs such as StringTie<sup>33</sup> or Cufflinks<sup>63</sup>.

Because all transcripts in our analysis had confident strand assignment, we made sure to use the ‘-S’ option to ensure the software did not consider ORFs on the opposite strand to the one annotated. Second, we built a protein database using the MANE dataset for blastp search

against the candidate ORFs predicted by TransDecoder. These protein alignments were used to select the best candidate ORF during the second stage of the TransDecoder execution.

**GeneMarkS-T.** GeneMarkS-T<sup>36</sup> is another ab initio ORF finding method included in several prominent pipelines for annotating ORFs in transcriptome assemblies. Contrary to the TransDecoder, the method relies less on the longest ORFs to initiate search and more on other features, refining its choice the 5’ AUG as the translation initiation site.

We applied GeneMarkS-T to both RefSeq and GENCODE datasets, similarly ensuring that the strand information is kept true to the reference via the ‘-strand direct’ option.

### Execution time

Some methods include multiple separate steps and commands that need to be executed to annotate ORFs. When measuring runtime, we recorded the total time it took to run all commands specified by each method. However, because GeneMarkS-T reports CDS coordinates relative to the transcript in which they were found, we developed our own custom script to convert transcriptomic coordinates to genomic ones. Although we did not add the conversion time to the total runtime of the method, depending on the implementation, this step could considerably increase the runtime.

Both ORFAnage and the costly blastp alignment step in TransDecoder can make proper use of multi-threading, yet GeneMarkS-T cannot. Nonetheless, primarily because of how slow TransDecoder was without multi-threading enabled, we allowed both ORFAnage and TransDecoder to use 30 threads concurrently. For ORFAnage we provide both single and multi-threaded performance measurements (Table 2).

### Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this Article.

### Data availability

No new sequencing data were created for this study. The sequencing data used in this study are available through the GTEx project ([phs000424.v7.p2](https://phs000424.v7.p2)). GTEx data were first analyzed as part of the CHES project and the details can be found in the corresponding resources and publications (<http://ccb.jhu.edu/chess/>). The datasets analyzed in this study are (1) GENCODE annotation build version 41 ([https://www.gencodegenes.org/human/release\\_41.html](https://www.gencodegenes.org/human/release_41.html)); (2) RefSeq annotation build 110 ([https://www.ncbi.nlm.nih.gov/genome/annotation\\_euk/Homo\\_sapiens/110/](https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Homo_sapiens/110/)); (3) MANE joint annotation build version 1.0 ([https://ftp.ncbi.nlm.nih.gov/refseq/MANE/MANE\\_human/](https://ftp.ncbi.nlm.nih.gov/refseq/MANE/MANE_human/)); (4) *A. thaliana* annotation ([https://ftp.ncbi.nlm.nih.gov/genomes/refseq/plant/Arabidopsis\\_thaliana/all\\_assembly\\_versions/GCF\\_000001735.3\\_TAIR10/](https://ftp.ncbi.nlm.nih.gov/genomes/refseq/plant/Arabidopsis_thaliana/all_assembly_versions/GCF_000001735.3_TAIR10/)); and (5) *C. elegans* genome annotation ([https://ftp.ncbi.nlm.nih.gov/genomes/refseq/invertebrate/Caenorhabditis\\_elegans/all\\_assembly\\_versions/GCF\\_000002985.6\\_WBcel235/](https://ftp.ncbi.nlm.nih.gov/genomes/refseq/invertebrate/Caenorhabditis_elegans/all_assembly_versions/GCF_000002985.6_WBcel235/)). Source data are provided with this paper.

### Code availability

All code required to reproduce the data generated within the study from public sources is provided at [https://github.com/alevar/ORFAnage\\_tests](https://github.com/alevar/ORFAnage_tests). The core method is implemented in C++ and based on the GFFutils<sup>45</sup> and KSW2<sup>64,65</sup> libraries. The code and test data are available for download at <https://github.com/alevar/ORFAnage/releases/tag/1.0> (<https://doi.org/10.5281/zenodo.8102912>)<sup>66</sup>. Jupyter notebooks used to generate all results described in the manuscript are provided separately at [https://github.com/alevar/ORFAnage\\_tests](https://github.com/alevar/ORFAnage_tests) (<https://doi.org/10.5281/zenodo.8102918>)<sup>67</sup>. All additional software methods used in this study and their versions and appropriate references are listed in Methods.

## References

- O'Leary, N. A. et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion and functional annotation. *Nucleic Acids Res.* **44**, D733–D745 (2016).
- Frankish, A. et al. GENCODE: reference annotation for the human and mouse genomes in 2023. *Nucleic Acids Res.* **51**, D942–D949 (2023).
- Pertea, M. et al. CHESS: a new human gene catalog curated from thousands of large-scale RNA sequencing experiments reveals extensive transcriptional noise. *Genome Biol.* **19**, 208 (2018).
- Varabyou, A. et al. CHESS 3: an improved, comprehensive catalog of human genes and transcripts based on large-scale expression data, phylogenetic analysis and protein structure. Preprint at *bioRxiv* <https://doi.org/10.1101/2022.12.21.521274> (2022).
- Salzberg, S. L. Open questions: how many genes do we have? *BMC Biol.* **16**, 94 (2018).
- Morales, J. et al. A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. *Nature* **604**, 310–315 (2022).
- Rodriguez, J. M. et al. APPRIS: annotation of principal and alternative splice isoforms. *Nucleic Acids Res.* **41**, D110–D117 (2013).
- Tress, M. L., Abascal, F. & Valencia, A. Alternative splicing may not be the key to proteome complexity. *Trends Biochem. Sci.* **42**, 98–110 (2017).
- Wang, E. T. et al. Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470–476 (2008).
- Djebali, S. et al. Landscape of transcription in human cells. *Nature* **489**, 101–108 (2012).
- Reyes, A. & Huber, W. Alternative start and termination sites of transcription drive most transcript isoform differences across human tissues. *Nucleic Acids Res.* **46**, 582–592 (2018).
- Sinitcyn, P. et al. Global detection of human variants and isoforms by deep proteome sequencing. *Nat. Biotechnol.* <https://doi.org/10.1038/s41587-023-01714-x> (2023).
- Glinos, D. A. et al. Transcriptome variation in human tissues revealed by long-read sequencing. *Nature* **608**, 353–359 (2022).
- Park, E., Pan, Z., Zhang, Z., Lin, L. & Xing, Y. The expanding landscape of alternative splicing variation in human populations. *Am. J. Hum. Genet.* **102**, 11–26 (2018).
- Zhang, S. et al. New insights into *Arabidopsis* transcriptome complexity revealed by direct sequencing of native RNAs. *Nucleic Acids Res.* **48**, 7700–7711 (2020).
- Roach, N. P. et al. The full-length transcriptome of *C. elegans* using direct RNA sequencing. *Genome Res.* **30**, 299–312 (2020).
- Zhao, S. Alternative splicing, RNA-seq and drug discovery. *Drug Discov. Today* **24**, 1258–1267 (2019).
- Kiyose, H. et al. Comprehensive analysis of full-length transcripts reveals novel splicing abnormalities and oncogenic transcripts in liver cancer. *PLoS Genet.* **18**, e1010342 (2022).
- Leung, S. K. et al. Full-length transcript sequencing of human and mouse cerebral cortex identifies widespread isoform diversity and alternative splicing. *Cell Rep.* **37**, 110022 (2021).
- Matlin, A. J., Clark, F. & Smith, C. W. Understanding alternative splicing: towards a cellular code. *Nat. Rev. Mol. Cell Biol.* **6**, 386–398 (2005).
- Tazi, J., Bakkour, N. & Stamm, S. Alternative splicing and disease. *Biochim. Biophys. Acta* **1792**, 14–26 (2009).
- Garcia-Blanco, M. A., Baraniak, A. P. & Lasda, E. L. Alternative splicing in disease and therapy. *Nat. Biotechnol.* **22**, 535–546 (2004).
- Cummings, B. B. et al. Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. *Sci. Transl. Med.* **9**, eaal5209 (2017).
- Merkin, J., Russell, C., Chen, P. & Burge, C. B. Evolutionary dynamics of gene and isoform regulation in mammalian tissues. *Science* **338**, 1593–1599 (2012).
- Boulet, A. et al. The mammalian phosphate carrier SLC25A3 is a mitochondrial copper transporter required for cytochrome c oxidase biogenesis. *J. Biol. Chem.* **293**, 1887–1896 (2018).
- Kim, H. K., Pham, M. H. C., Ko, K. S., Rhee, B. D. & Han, J. Alternative splicing isoforms in health and disease. *Pflüg. Arch. Eur. J. Physiol.* **470**, 995–1016 (2018).
- Frampton, G. M. et al. Activation of MET via diverse exon 14 splicing alterations occurs in multiple tumor types and confers clinical sensitivity to MET inhibitors MET Exon 14 alterations confer response to targeted therapy. *Cancer Discov.* **5**, 850–859 (2015).
- Kahles, A. et al. Comprehensive analysis of alternative splicing across tumors from 8,705 patients. *Cancer Cell* **34**, 211–224 (2018).
- Brooks, A. N. et al. A pan-cancer analysis of transcriptome changes associated with somatic mutations in U2AF1 reveals commonly altered splicing events. *PLoS ONE* **9**, e87361 (2014).
- Allen, A. S. et al. De novo mutations in epileptic encephalopathies. *Nature* **501**, 217–221 (2013).
- Cancer Genome Atlas Research Network. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N. Engl. J. Med.* **368**, 2059–2074 (2013).
- Varabyou, A., Salzberg, S. L. & Pertea, M. Effects of transcriptional noise on estimates of gene and transcript expression in RNA sequencing experiments. *Genome Res.* **31**, 301–308 (2021).
- Kovaka, S. et al. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol.* **20**, 278 (2019).
- Haas, B. J. et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* **8**, 1494–1512 (2013).
- Vitting-Seerup, K. & Sandelin, A. The landscape of isoform switches in human cancers. *Mol. Cancer Res.* **15**, 1206–1220 (2017).
- Tang, S., Lomsadze, A. & Borodovsky, M. Identification of protein coding regions in RNA transcripts. *Nucleic Acids Res.* **43**, e78 (2015).
- Vitting-Seerup, K., Porse, B. T., Sandelin, A. & Waage, J. spliceR: an R package for classification of alternative splicing and prediction of coding potential from RNA-seq data. *BMC Bioinformatics* **15**, 81 (2014).
- Kang, Y. et al. CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. *Nucleic Acids Res.* **45**, W12–W16 (2017).
- Singh, U. & Wurtele, E. S. orfipy: a fast and flexible tool for extracting ORFs. *Bioinformatics* **37**, 3019–3020 (2021).
- Tress, M. L., Abascal, F. & Valencia, A. Most alternative isoforms are not functionally important. *Trends Biochem. Sci.* **42**, 408–410 (2017).
- Cunningham, F. et al. Ensembl 2022. *Nucleic Acids Res.* **50**, D988–D995 (2022).
- Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European molecular biology open software suite. *Trends Genet.* **16**, 276–277 (2000).
- Steijger, T. et al. Assessment of transcript reconstruction methods for RNA-seq. *Nat. Methods* **10**, 1177–1184 (2013).
- Lonsdale, J. et al. The genotype-tissue expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).
- Pertea, G. & Pertea, M. GFF utilities: GffRead and GffCompare. *F1000Res.* **9**, 304 (2020).
- Moss, S. E. & Morgan, R. O. The annexins. *Genome Biol.* **5**, 219 (2004).
- Gerke, V. & Moss, S. E. Annexins: from structure to function. *Physiol. Rev.* **82**, 331–371 (2002).
- McCulloch, K. M. et al. An alternative N-terminal fold of the intestine-specific annexin A13a induces dimerization and regulates membrane-binding. *J. Biol. Chem.* **294**, 3454–3463 (2019).

49. Lillebostad, P. A. et al. Structure of the ALS mutation target annexin A11 reveals a stabilising N-terminal segment. *Biomolecules* **10**, 660 (2020).
50. Fernández-Lizarbe, S. et al. Structural and lipid-binding characterization of human annexin A13a reveals strong differences with its long A13b isoform. *Biol. Chem.* **398**, 359–371 (2017).
51. Mirdita, M. et al. ColabFold: making protein folding accessible to all. *Nat. Methods* **19**, 679–682 (2022).
52. Varadi, M. et al. AlphaFold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* **50**, D439–D444 (2022).
53. Finstermeier, K. et al. A mitogenomic phylogeny of living primates. *PLoS ONE* **8**, e69504 (2013).
54. Wall, J. D., Robinson, J. A. & Cox, L. A. High-resolution estimates of crossover and noncrossover recombination from a captive baboon colony. *Genome Biol. Evol.* **14**, evac040 (2022).
55. Shumate, A. & Salzberg, S. L. Liftoff: accurate mapping of gene annotations. *Bioinformatics* **37**, 1639–1643 (2020).
56. Sommer, M. J. et al. Structure-guided isoform identification for the human transcriptome. *eLife* **11**, e82556 (2022).
57. Pockrandt, C., Steinegger, M. & Salzberg, S. L. PhyloCSF++: a fast and user-friendly implementation of PhyloCSF with annotation tools. *Bioinformatics* **38**, 1440–1442 (2022).
58. Lin, M. F., Jungreis, I. & Kellis, M. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* **27**, 275–282 (2011).
59. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).
60. Varabyou, A., Pertea, G., Pockrandt, C. & Pertea, M. TieBrush: an efficient method for aggregating and summarizing mapped reads across large datasets. *Bioinformatics* **37**, 3650–3651 (2021).
61. Swarbreck, D. et al. The Arabidopsis Information Resource (TAIR): gene structure and function annotation. *Nucleic Acids Res.* **36**, D1009–D1014 (2007).
62. C. elegans Sequencing Consortium. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* **282**, 2012–2018 (1998).
63. Trapnell, C. et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7**, 562–578 (2012).
64. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
65. Suzuki, H. & Kasahara, M. Introducing difference recurrence relations for faster semi-global alignment of long sequences. *BMC Bioinformatics* **19**, 45 (2018).
66. Varabyou, A. ORFanage: reference guided ORF annotation 1.0.2. Zenodo <https://doi.org/10.5281/zenodo.8102912> (2023).
67. Varabyou, A. ORFanage evaluation notebooks. Zenodo <https://doi.org/10.5281/zenodo.8102918> (2023).
68. DeLano, W. L. PyMOL: an open-source molecular graphics tool. *CCP4 Newsl. Protein Crystallogr.* **40**, 82–92 (2002).

## Acknowledgements

This work was supported in part by the US National Institutes of Health under grants nos. R01 HG006677 (S.L.S.), R01 MH123567 (S.L.S.) and R35 GM130151 (S.L.S.) and by the US National Science Foundation under grant no. DBI-1759518 (M.P.). We would also like to thank C. Pockrandt for helpful discussions on phyloCSF++ implementation and usage.

## Author contributions

A.V. and B.E. conceived and developed the original idea. A.V. developed and implemented the final method and experiments. A.V., B.E., S.L.S. and M.P. conceptualized the study, methods and wrote the paper.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s43588-023-00496-1>.

**Correspondence and requests for materials** should be addressed to Ales Varabyou or Mihaela Pertea.

**Peer review information** *Nature Computational Science* thanks Liugo Wang and the other anonymous reviewer for their contribution to the peer review of this work. Primary Handling Editor: Fernando Chirigati, in collaboration with the *Nature Computational Science* team.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2023

## Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

### Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- |                                     |                                     |                                                                                                                                                                                                                                                            |
|-------------------------------------|-------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> | The exact sample size ( $n$ ) for each experimental group/condition, given as a discrete number and unit of measurement                                                                                                                                    |
| <input checked="" type="checkbox"/> | <input type="checkbox"/>            | A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly                                                                                                                                    |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> | The statistical test(s) used AND whether they are one- or two-sided<br><i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i>                                                               |
| <input checked="" type="checkbox"/> | <input type="checkbox"/>            | A description of all covariates tested                                                                                                                                                                                                                     |
| <input checked="" type="checkbox"/> | <input type="checkbox"/>            | A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons                                                                                                                                        |
| <input checked="" type="checkbox"/> | <input type="checkbox"/>            | A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> | For null hypothesis testing, the test statistic (e.g. $F$ , $t$ , $r$ ) with confidence intervals, effect sizes, degrees of freedom and $P$ value noted<br><i>Give <math>P</math> values as exact values whenever suitable.</i>                            |
| <input checked="" type="checkbox"/> | <input type="checkbox"/>            | For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings                                                                                                                                                           |
| <input checked="" type="checkbox"/> | <input type="checkbox"/>            | For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes                                                                                                                                     |
| <input checked="" type="checkbox"/> | <input type="checkbox"/>            | Estimates of effect sizes (e.g. Cohen's $d$ , Pearson's $r$ ), indicating how they were calculated                                                                                                                                                         |

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

### Software and code

Policy information about [availability of computer code](#)

Data collection No software was used for collecting data in this study.

Data analysis ORFAnage (v1.0, <https://doi.org/10.5281/zenodo.8102912>), HISAT2 (v2.2), TieBrush (v0.0.6), TieCov (v0.0.6) and Sashimi package (v0.0.6), StringTie (v2.2.1), PhyloCSF++ (v1.2.0), gffcompare (0.12.6), gffread (v0.12.7), STRETCHER (v6.6.0.0), transdecoder (v5.7.0), GeneMarkS-T (3.20) and ColabFold (v1.3.0) were used in this study. All additional code used to generate and analyze the results of the study is available at: [https://github.com/alevar/ORFAnage\\_tests](https://github.com/alevar/ORFAnage_tests) (<https://doi.org/10.5281/zenodo.8102918>).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

### Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

Sequencing data used in this study is available through the GTEx project (phs000424.v7.p2). GTEx data was first analyzed as part of the CHES project and the

details can be found in the corresponding resources and publications (<http://ccb.jhu.edu/chess/>). The datasets analyzed in this study are 1. GENCODE annotation build version 41 ([https://www.gencodegenes.org/human/release\\_41.html](https://www.gencodegenes.org/human/release_41.html)); 2. RefSeq annotation build 110 ([https://www.ncbi.nlm.nih.gov/genome/annotation\\_euk/Homo\\_sapiens/110/](https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Homo_sapiens/110/)); 3. MANE joint annotation build version 1.0 ([https://ftp.ncbi.nlm.nih.gov/refseq/MANE/MANE\\_human/](https://ftp.ncbi.nlm.nih.gov/refseq/MANE/MANE_human/)); 4. A. thaliana annotation [https://ftp.ncbi.nlm.nih.gov/genomes/refseq/plant/Arabidopsis\\_thaliana/all\\_assembly\\_versions/GCF\\_000001735.3\\_TAIR10/](https://ftp.ncbi.nlm.nih.gov/genomes/refseq/plant/Arabidopsis_thaliana/all_assembly_versions/GCF_000001735.3_TAIR10/)) and 5. C. elegans genome annotation ([https://ftp.ncbi.nlm.nih.gov/genomes/refseq/invertebrate/Caenorhabditis\\_elegans/all\\_assembly\\_versions/GCF\\_000002985.6\\_WBcel235/](https://ftp.ncbi.nlm.nih.gov/genomes/refseq/invertebrate/Caenorhabditis_elegans/all_assembly_versions/GCF_000002985.6_WBcel235/)).

## Human research participants

Policy information about [studies involving human research participants and Sex and Gender in Research](#).

|                             |     |
|-----------------------------|-----|
| Reporting on sex and gender | N/A |
| Population characteristics  | N/A |
| Recruitment                 | N/A |
| Ethics oversight            | N/A |

Note that full information on the approval of the study protocol must also be provided in the manuscript.

## Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

## Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

|                 |                                                                              |
|-----------------|------------------------------------------------------------------------------|
| Sample size     | This is a computational methods paper and this point is not applicable here. |
| Data exclusions | N/A                                                                          |
| Replication     | N/A                                                                          |
| Randomization   | N/A                                                                          |
| Blinding        | N/A                                                                          |

## Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

|                   |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 |
|-------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Study description | Briefly describe the study type including whether data are quantitative, qualitative, or mixed-methods (e.g. qualitative cross-sectional, quantitative experimental, mixed-methods case study).                                                                                                                                                                                                                                                                                 |
| Research sample   | State the research sample (e.g. Harvard university undergraduates, villagers in rural India) and provide relevant demographic information (e.g. age, sex) and indicate whether the sample is representative. Provide a rationale for the study sample chosen. For studies involving existing datasets, please describe the dataset and source.                                                                                                                                  |
| Sampling strategy | Describe the sampling procedure (e.g. random, snowball, stratified, convenience). Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient. For qualitative data, please indicate whether data saturation was considered, and what criteria were used to decide that no further sampling was needed. |
| Data collection   | Provide details about the data collection procedure, including the instruments or devices used to record the data (e.g. pen and paper, computer, eye tracker, video or audio equipment) whether anyone was present besides the participant(s) and the researcher, and whether the researcher was blind to experimental condition and/or the study hypothesis during data collection.                                                                                            |
| Timing            | Indicate the start and stop dates of data collection. If there is a gap between collection periods, state the dates for each sample cohort.                                                                                                                                                                                                                                                                                                                                     |

## Data exclusions

If no data were excluded from the analyses, state so OR if data were excluded, provide the exact number of exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.

## Non-participation

State how many participants dropped out/declined participation and the reason(s) given OR provide response rate OR state that no participants dropped out/declined participation.

## Randomization

If participants were not allocated into experimental groups, state so OR describe how participants were allocated to groups, and if allocation was not random, describe how covariates were controlled.

## Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

## Study description

Briefly describe the study. For quantitative data include treatment factors and interactions, design structure (e.g. factorial, nested, hierarchical), nature and number of experimental units and replicates.

## Research sample

Describe the research sample (e.g. a group of tagged *Passer domesticus*, all *Stenocereus thurberi* within Organ Pipe Cactus National Monument), and provide a rationale for the sample choice. When relevant, describe the organism taxa, source, sex, age range and any manipulations. State what population the sample is meant to represent when applicable. For studies involving existing datasets, describe the data and its source.

## Sampling strategy

Note the sampling procedure. Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient.

## Data collection

Describe the data collection procedure, including who recorded the data and how.

## Timing and spatial scale

Indicate the start and stop dates of data collection, noting the frequency and periodicity of sampling and providing a rationale for these choices. If there is a gap between collection periods, state the dates for each sample cohort. Specify the spatial scale from which the data are taken

## Data exclusions

If no data were excluded from the analyses, state so OR if data were excluded, describe the exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.

## Reproducibility

Describe the measures taken to verify the reproducibility of experimental findings. For each experiment, note whether any attempts to repeat the experiment failed OR state that all attempts to repeat the experiment were successful.

## Randomization

Describe how samples/organisms/participants were allocated into groups. If allocation was not random, describe how covariates were controlled. If this is not relevant to your study, explain why.

## Blinding

Describe the extent of blinding used during data acquisition and analysis. If blinding was not possible, describe why OR explain why blinding was not relevant to your study.

Did the study involve field work?

☐ Yes

☐ No

## Field work, collection and transport

## Field conditions

Describe the study conditions for field work, providing relevant parameters (e.g. temperature, rainfall).

## Location

State the location of the sampling or experiment, providing relevant parameters (e.g. latitude and longitude, elevation, water depth).

## Access &amp; import/export

Describe the efforts you have made to access habitats and to collect and import/export your samples in a responsible manner and in compliance with local, national and international laws, noting any permits that were obtained (give the name of the issuing authority, the date of issue, and any identifying information).

## Disturbance

Describe any disturbance caused by the study and how it was minimized.

## Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

## Materials &amp; experimental systems

## Methods

- n/a Involved in the study
- ☒ ☐ Antibodies
- ☒ ☐ Eukaryotic cell lines
- ☒ ☐ Palaeontology and archaeology
- ☒ ☐ Animals and other organisms
- ☒ ☐ Clinical data
- ☒ ☐ Dual use research of concern

- n/a Involved in the study
- ☒ ☐ ChIP-seq
- ☒ ☐ Flow cytometry
- ☒ ☐ MRI-based neuroimaging

## Antibodies

Antibodies used

Describe all antibodies used in the study; as applicable, provide supplier name, catalog number, clone name, and lot number.

Validation

Describe the validation of each primary antibody for the species and application, noting any validation statements on the manufacturer's website, relevant citations, antibody profiles in online databases, or data provided in the manuscript.

## Eukaryotic cell lines

Policy information about [cell lines](#) and [Sex and Gender in Research](#)

Cell line source(s)

State the source of each cell line used and the sex of all primary cell lines and cells derived from human participants or vertebrate models.

Authentication

Describe the authentication procedures for each cell line used OR declare that none of the cell lines used were authenticated.

Mycoplasma contamination

Confirm that all cell lines tested negative for mycoplasma contamination OR describe the results of the testing for mycoplasma contamination OR declare that the cell lines were not tested for mycoplasma contamination.

Commonly misidentified lines  
(See [ICLAC](#) register)

Name any commonly misidentified cell lines used in the study and provide a rationale for their use.

## Palaeontology and Archaeology

Specimen provenance

Provide provenance information for specimens and describe permits that were obtained for the work (including the name of the issuing authority, the date of issue, and any identifying information). Permits should encompass collection and, where applicable, export.

Specimen deposition

Indicate where the specimens have been deposited to permit free access by other researchers.

Dating methods

If new dates are provided, describe how they were obtained (e.g. collection, storage, sample pretreatment and measurement), where they were obtained (i.e. lab name), the calibration program and the protocol for quality assurance OR state that no new dates are provided.

☐ Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.

Ethics oversight

Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

## Animals and other research organisms

Policy information about [studies involving animals](#); [ARRIVE guidelines](#) recommended for reporting animal research, and [Sex and Gender in Research](#)

Laboratory animals

For laboratory animals, report species, strain and age OR state that the study did not involve laboratory animals.

Wild animals

Provide details on animals observed in or captured in the field; report species and age where possible. Describe how animals were caught and transported and what happened to captive animals after the study (if killed, explain why and describe method; if released, say where and when) OR state that the study did not involve wild animals.

Reporting on sex

Indicate if findings apply to only one sex; describe whether sex was considered in study design, methods used for assigning sex. Provide data disaggregated for sex where this information has been collected in the source data as appropriate; provide overall numbers in this Reporting Summary. Please state if this information has not been collected. Report sex-based analyses where performed, justify reasons for lack of sex-based analysis.

## Field-collected samples

For laboratory work with field-collected samples, describe all relevant parameters such as housing, maintenance, temperature, photoperiod and end-of-experiment protocol OR state that the study did not involve samples collected from the field.

## Ethics oversight

Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

## Clinical data

Policy information about [clinical studies](#)

All manuscripts should comply with the ICMJE [guidelines for publication of clinical research](#) and a completed [CONSORT checklist](#) must be included with all submissions.

## Clinical trial registration

Provide the trial registration number from ClinicalTrials.gov or an equivalent agency.

## Study protocol

Note where the full trial protocol can be accessed OR if not available, explain why.

## Data collection

Describe the settings and locales of data collection, noting the time periods of recruitment and data collection.

## Outcomes

Describe how you pre-defined primary and secondary outcome measures and how you assessed these measures.

## Dual use research of concern

Policy information about [dual use research of concern](#)

### Hazards

Could the accidental, deliberate or reckless misuse of agents or technologies generated in the work, or the application of information presented in the manuscript, pose a threat to:

No | Yes

- ☒ ☐ Public health  
☒ ☐ National security  
☒ ☐ Crops and/or livestock  
☒ ☐ Ecosystems  
☒ ☐ Any other significant area

### Experiments of concern

Does the work involve any of these experiments of concern:

No | Yes

- ☒ ☐ Demonstrate how to render a vaccine ineffective  
☒ ☐ Confer resistance to therapeutically useful antibiotics or antiviral agents  
☒ ☐ Enhance the virulence of a pathogen or render a nonpathogen virulent  
☒ ☐ Increase transmissibility of a pathogen  
☒ ☐ Alter the host range of a pathogen  
☒ ☐ Enable evasion of diagnostic/detection modalities  
☒ ☐ Enable the weaponization of a biological agent or toxin  
☒ ☐ Any other potentially harmful combination of experiments and agents

## ChIP-seq

### Data deposition

- ☐ Confirm that both raw and final processed data have been deposited in a public database such as [GEO](#).  
☐ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

## Data access links

May remain private before publication.

For "Initial submission" or "Revised version" documents, provide reviewer access links. For your "Final submission" document, provide a link to the deposited data.

## Files in database submission

Provide a list of all files available in the database submission.

## Genome browser session

(e.g. [UCSC](#))

Provide a link to an anonymized genome browser session for "Initial submission" and "Revised version" documents only, to enable peer review. Write "no longer applicable" for "Final submission" documents.

## Methodology

|                         |                                                                                                                                                                             |
|-------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Replicates              | Describe the experimental replicates, specifying number, type and replicate agreement.                                                                                      |
| Sequencing depth        | Describe the sequencing depth for each experiment, providing the total number of reads, uniquely mapped reads, length of reads and whether they were paired- or single-end. |
| Antibodies              | Describe the antibodies used for the ChIP-seq experiments; as applicable, provide supplier name, catalog number, clone name, and lot number.                                |
| Peak calling parameters | Specify the command line program and parameters used for read mapping and peak calling, including the ChIP, control and index files used.                                   |
| Data quality            | Describe the methods used to ensure data quality in full detail, including how many peaks are at FDR 5% and above 5-fold enrichment.                                        |
| Software                | Describe the software used to collect and analyze the ChIP-seq data. For custom code that has been deposited into a community repository, provide accession details.        |

## Flow Cytometry

### Plots

Confirm that:

- ☐ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).
- ☐ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).
- ☐ All plots are contour plots with outliers or pseudocolor plots.
- ☐ A numerical value for number of cells or percentage (with statistics) is provided.

### Methodology

|                                                                                                                                                |                                                                                                                                                                                                                                                |
|------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Sample preparation                                                                                                                             | Describe the sample preparation, detailing the biological source of the cells and any tissue processing steps used.                                                                                                                            |
| Instrument                                                                                                                                     | Identify the instrument used for data collection, specifying make and model number.                                                                                                                                                            |
| Software                                                                                                                                       | Describe the software used to collect and analyze the flow cytometry data. For custom code that has been deposited into a community repository, provide accession details.                                                                     |
| Cell population abundance                                                                                                                      | Describe the abundance of the relevant cell populations within post-sort fractions, providing details on the purity of the samples and how it was determined.                                                                                  |
| Gating strategy                                                                                                                                | Describe the gating strategy used for all relevant experiments, specifying the preliminary FSC/SSC gates of the starting cell population, indicating where boundaries between "positive" and "negative" staining cell populations are defined. |
| <input type="checkbox"/> Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information. |                                                                                                                                                                                                                                                |

## Magnetic resonance imaging

### Experimental design

|                                 |                                                                                                                                                                                                                                                            |
|---------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Design type                     | Indicate task or resting state; event-related or block design.                                                                                                                                                                                             |
| Design specifications           | Specify the number of blocks, trials or experimental units per session and/or subject, and specify the length of each trial or block (if trials are blocked) and interval between trials.                                                                  |
| Behavioral performance measures | State number and/or type of variables recorded (e.g. correct button press, response time) and what statistics were used to establish that the subjects were performing the task as expected (e.g. mean, range, and/or standard deviation across subjects). |

## Acquisition

|                               |                                                                                                                                                                                           |
|-------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Imaging type(s)               | <i>Specify: functional, structural, diffusion, perfusion.</i>                                                                                                                             |
| Field strength                | <i>Specify in Tesla</i>                                                                                                                                                                   |
| Sequence & imaging parameters | <i>Specify the pulse sequence type (gradient echo, spin echo, etc.), imaging type (EPI, spiral, etc.), field of view, matrix size, slice thickness, orientation and TE/TR/flip angle.</i> |
| Area of acquisition           | <i>State whether a whole brain scan was used OR define the area of acquisition, describing how the region was determined.</i>                                                             |
| Diffusion MRI                 | <input type="checkbox"/> Used <input type="checkbox"/> Not used                                                                                                                           |

## Preprocessing

|                            |                                                                                                                                                                                                                                                |
|----------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Preprocessing software     | <i>Provide detail on software version and revision number and on specific parameters (model/functions, brain extraction, segmentation, smoothing kernel size, etc.).</i>                                                                       |
| Normalization              | <i>If data were normalized/standardized, describe the approach(es): specify linear or non-linear and define image types used for transformation OR indicate that data were not normalized and explain rationale for lack of normalization.</i> |
| Normalization template     | <i>Describe the template used for normalization/transformation, specifying subject space or group standardized space (e.g. original Talairach, MNI305, ICBM152) OR indicate that the data were not normalized.</i>                             |
| Noise and artifact removal | <i>Describe your procedure(s) for artifact and structured noise removal, specifying motion parameters, tissue signals and physiological signals (heart rate, respiration).</i>                                                                 |
| Volume censoring           | <i>Define your software and/or method and criteria for volume censoring, and state the extent of such censoring.</i>                                                                                                                           |

## Statistical modeling & inference

|                                                                           |                                                                                                                                                                                                                         |
|---------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Model type and settings                                                   | <i>Specify type (mass univariate, multivariate, RSA, predictive, etc.) and describe essential details of the model at the first and second levels (e.g. fixed, random or mixed effects; drift or auto-correlation).</i> |
| Effect(s) tested                                                          | <i>Define precise effect in terms of the task or stimulus conditions instead of psychological concepts and indicate whether ANOVA or factorial designs were used.</i>                                                   |
| Specify type of analysis:                                                 | <input type="checkbox"/> Whole brain <input type="checkbox"/> ROI-based <input type="checkbox"/> Both                                                                                                                   |
| Statistic type for inference<br>(See <a href="#">Eklund et al. 2016</a> ) | <i>Specify voxel-wise or cluster-wise and report all relevant parameters for cluster-wise methods.</i>                                                                                                                  |
| Correction                                                                | <i>Describe the type of correction and how it is obtained for multiple comparisons (e.g. FWE, FDR, permutation or Monte Carlo).</i>                                                                                     |

## Models & analysis

|                                               |                                                                                                                                                                                                                                  |
|-----------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| n/a                                           | Involvement in the study                                                                                                                                                                                                         |
| <input type="checkbox"/>                      | <input type="checkbox"/> Functional and/or effective connectivity                                                                                                                                                                |
| <input type="checkbox"/>                      | <input type="checkbox"/> Graph analysis                                                                                                                                                                                          |
| <input type="checkbox"/>                      | <input type="checkbox"/> Multivariate modeling or predictive analysis                                                                                                                                                            |
| Functional and/or effective connectivity      | <i>Report the measures of dependence used and the model details (e.g. Pearson correlation, partial correlation, mutual information).</i>                                                                                         |
| Graph analysis                                | <i>Report the dependent variable and connectivity measure, specifying weighted graph or binarized graph, subject- or group-level, and the global and/or node summaries used (e.g. clustering coefficient, efficiency, etc.).</i> |
| Multivariate modeling and predictive analysis | <i>Specify independent variables, features extraction and dimension reduction, model, training and evaluation metrics.</i>                                                                                                       |