keep methods updated, unlike for a static publication. All methods have been submitted to the European Bioinformatics Institute's Experimental Factor Ontologies.

Any human naming convention will be limited. While naming conventions are useful, we do not want to lose the quirkiness and joy of method names such as BLESS[4] and Rapture[5]! Developers of new or modified methods could include a statement such as "This method is similar to method X," making the new method easier to find and group with related methods. Ultimately the community is responsible for the systematic organization of NGS methods to ensure the continued health and growth of genomics. We hope that our proposals here and the suggested naming conventions help in these efforts.

ACKNOWLEDGMENTS

We acknowledge the contribution of the many authors whose methods are referenced in this manuscript and apologize that restrictions on the number of references meant we did not list their work.

AUTHOR CONTRIBUTIONS

J.H. and J.R. contributed equally to this work. J.H. wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

James Hadfield[1,2] & Jacques Retief[3]

[1]Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge, UK. [2]Precision Medicine and Genomics, IMED Biotech Unit, AstraZeneca, Cambridge, UK. [3]Illumina Inc., San Diego, California, USA. Precision Medicine and Genomics, IMED Biotech Unit, AstraZeneca, Cambridge, UK.
e-mail: james.hadfield@cruk.cam.ac.uk, james.hadfield@astrazeneca.com

1. Retief, J. D. and Maxkwee K. *For all you seq,* https://www.illumina.com/content/dam/illumina-marketing/documents/applications/ngs-library-prep/ForAllYouSeqMethods.pdf (Illumina, 2014).
2. Anonymous. *Nat. Methods* **8**, 521 (2011).
3. Tang, F. *et al. Nat. Methods* **6**, 377–382 (2009)
4. Crosetto, N. *et al. Nat. Methods* **10**, 361–365 (2013).
5. Ali, O.A. *et al. Genetics* **202**, 389–400 (2016).

motifStack for the analysis of transcription factor binding site evolution

To the Editor: A sequence motif is a short recurring pattern with biological significance such as a DNA-recognition sequence for a transcription factor (TF), an mRNA splicing signal, or a functional region of a protein domain. Many high-throughput experimental approaches and computational tools have been developed to discover motifs from a population of functional sequences such as TF binding sites[1]. TF binding motifs are often represented as position weight matrices (PWMs) and visualized as sequence logos (**Supplementary Note**).

To facilitate classification and comparison of motifs, researchers have developed motif alignment and clustering tools such as STAMP[2], Tomtom[3], and MatAlign[4]. However, existing tools for the visualization of similarities or differences within groups of motifs are limited by their flexibility in displaying trees (STAMP), the number of motifs supported (DiffLogo[5]), or the ability to display motif logo alignments (Cytoscape[6]).

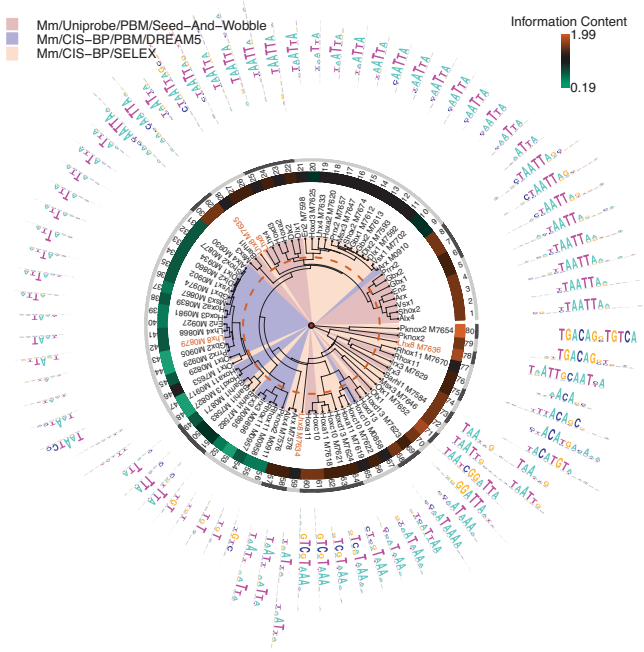

Figure 1 | Effects of experimental and computational methods on motif clustering. Motifs for a set of mouse HD TFs present in three different data sets are depicted as a radial phylogenic tree using motifStack with distance threshold of 2.5; each data set uses a different combination of experimental and computational motif generation methods (see **Supplementary Note** for details). Tree branches are colored to highlight the source of each motif. The inner ring is colored to indicate the information content (IC) of the motifs. The alternating light and dark gray colors in the second ring delineate different motif clusters.

We describe motifStack, a Bioconductor package to visualize the alignment of motifs as a phylogenetic tree. This tool facilitates the analysis of binding site diversity and conservation within families of TFs and the evolution of TFs among different species. motifStack can align DNA motifs; generate motif signatures for closely related motifs; and plot aligned motifs as a stack, a linear or a radial tree, or a word cloud of sequence logos (**Supplementary Fig. 1**). Different parameter settings can be used to generate diverse types of plots with color schema highlighting important data features (**Supplementary Fig. 2**).

To illustrate the utility of motifStack for providing insights into families of related motifs, we analyzed DNA-binding motifs determined for fruit fly, mouse, and human homeodomain (HD) TFs (**Supplementary Note** and **Supplementary Figs. 3**–**7**). The diversity and relative frequency of motifs from the fly collection are depicted as a linear dendrogram showing individual TF motifs and motif signatures for each cluster (**Supplementary Fig. 3a**); the common HD motif TAATTA is correctly identified as the largest cluster. Consistent with previous studies, comparisons of HD motifs from multiple species show that the binding specificities of HD TFs are frequently conserved between mammals and insects, but that species-specific outliers exist, which may indicate gain or loss of family members during evolution (**Supplementary Figs. 3** and **7**). We also find that different experimental and computational methods can yield different motifs for the same TF (**Fig. 1** and **Supplementary Fig. 3d**). Furthermore, changing the computational method used to build motifs from the same binding data can lead to artificial segregation of motifs for identical TFs within a motif alignment (**Supplementary Fig. 2o**).

In conclusion, motifStack facilitates the discovery of divergence and conservation within TF families during evolution through motif comparisons and can illustrate biases introduced into these motifs by different computational and experimental platforms.

Life Sciences Reporting Summary. Further information regarding the experimental design may be found in the **Life Sciences Reporting Summary**.

Data availability statement. The motifStack package is freely available at http://bioconductor.org/packages/release/bioc/html/motifStack.html. To make it easy to install and run, a Docker container has been created with motifStack and all its R and system dependencies already installed (https://github.com/jianhong/motifStack.documentation). In addition, all data and scripts are also included in the Docker container for reproducing the figures (**Supplementary Note**). Docker is an open-source software platform that allows applications to be readily installed and run on any system. The availability of motifStack with all its dependencies as a Docker container also facilitates the integration of the motifStack package into workflow pipelines that support Docker images.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS
We appreciate M. Enuameh and L. Ni for helpful discussion. S.A.W., M.H.B., and L.J.Z. were partly supported by National Human Genome Research Institute of the National Institutes of Health (R01HG004744-01).

AUTHOR CONTRIBUTIONS
J.O. developed the software. J.O. and M.H.B. performed the data analysis. J.O., S.A.W., M.H.B., and L.J.Z. contributed to software testing, results interpretation, and manuscript preparation. All authors read and approved the final manuscript.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

Jianhong Ou[1], Scot A Wolfe[1,2], Michael H Brodsky[1,3] & Lihua Julie Zhu[1,3,4]

[1]Department of Molecular, Cell and Cancer Biology, University of Massachusetts Medical School, Worcester, Massachusetts, USA. [2]Department of Biochemistry and Molecular Pharmacology, University of Massachusetts Medical School, Worcester, Massachusetts, USA. [3]Program in Molecular Medicine, University of Massachusetts Medical School, Worcester, Massachusetts, USA. [4]Program in Bioinformatics and Integrative Biology, University of Massachusetts Medical School, Worcester, Massachusetts, USA.
e-mail: julie.zhu@umassmed.edu or michael.brodsky@umassmed.edu

1. Das, M.K. & Dai, H.-K. *BMC Bioinformatics* **8**, S21 (2007).
2. Mahony, S. & Benos, P.V. *Nucleic Acids Res.* **35**, W253–W258 (2007).
3. Gupta, S., Stamatoyannopoulos, J.A., Bailey, T.L. & Noble, W.S. *Genome Biol.* **8**, R24 (2007).
4. Zhao, G. *et al. G3* **2**, 469–481 (2012).
5. Nettling, M. *et al. BMC Bioinformatics* **16**, 387 (2015).
6. Shannon, P. *et al. Genome Res.* **13**, 2498–2504 (2003).

nature research

Corresponding author(s): Lihua Julie Zhu, Michael Brodsky

☐ Initial submission ☒ Revised version ☐ Final submission

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Experimental design

1. Sample size

Describe how sample size was determined.

n/a

2. Data exclusions

Describe any data exclusions.

Yes. There were two categories of potentially mammalian-specific clusters that we have excluded from our final set of species-specific motifs (Supplementary Notes).

3. Replication

Describe whether the experimental findings were reliably reproduced.

Yes. We consistently observed that the difference in motif generating methods (computational/experimental) can result in differences in information content or in the recovery of dimeric binding sites, which can strongly influence the apparent biochemical similarity or difference between TF motifs (Supplementary Notes). Species-specific binding motifs are identified using two different motif alignment methods MotIV and MatAlign (Supplementary Notes).

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

n/a

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

n/a

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a	Confirmed
☒	☐ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☒	☐ A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐ A statement indicating how many times each experiment was replicated
☒	☐ The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
☒	☐ A description of any assumptions or corrections, such as an adjustment for multiple comparisons
☒	☐ The test results (e.g. *P* values) given as exact values whenever possible and with confidence intervals noted
☒	☐ A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
☒	☐ Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.	No commercial code has been used. We describe here a bioconductor package motifStack, freely available at http://bioconductor.org/packages/release/bioc/html/motifStack.html. The code and dataset used to demonstrate its utilities and biological findings are freely available as a docker container, described in detail in the Supplementary Notes.

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.	n/a

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).	n/a

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.	n/a
b. Describe the method of cell line authentication used.	n/a
c. Report whether the cell lines were tested for mycoplasma contamination.	n/a
d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.	n/a

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.	n/a

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.	n/a