GOETHE UNIVERSITÄT FRANKFURT AM MAIN

APPLIED BIOINFORMATICS FRANKFURT

Aktuelle Themen der Sequenzanalyse

Seminar ASA-S

Sommersemester 2025

Informed and automated k-mer size selection for genome assembly

Written by:

Rayan Chikhi and Paul Medvedev

Published:

Bioinformatics, Volume 30, Issue 1, 2013, Pages 31–37

Presented by:

Romayssae Boudouassal and Lahem Asres Ioab

Abstract 1 written by Romysaae Boudouassal:

De Bruijn graph-based assembler depend on selecting an optimal k-mer size. The De Bruijn graph is a data structure used in genome assembly, where reads are divided into k-mer (substrings of length k). Nodes in the graph represent (k-1)mer, and edges represent k-mer present in the reads. The choice of k represents a trade-off between several effects that are critical to the quality of the assembly. However, there are currently no automated tools to determine the optimal k value and efficiently generate its abundance histogram. The paper presents a tool called KMERGENIE that estimates the optimal k-mer size for genome assembly quality. KMERGENIE provides a fast and accurate sampling method for building an approximate abundance histogram. The tool fits a generative model to each histogram to estimate the number of distinct genomic k-mer. It then selects the k value that maximizes the number of genomic k-mer for improved assembly quality. This tool was benchmarked using datasets from *S.aureus*, human chromosome 14, and *B.impatiens*, and the optimal k values were predicted. KMERGENIE can be integrated into assembly pipeline to choose k without user intervention, thereby simplifying sequencing data analysis.

Abstract written by Lahem Asres loab:

De Bruijn graph-based assemblers require k-mers to reconstruct genomes. First reads are split into k-mers, and the graph is then constructed with (k-1)-mers as nodes and k-mers present in the reads as edges. The size of k-mers represents a trade-off between several effects. However, there is a lack of tools to estimate the optimum k automatically and to efficiently generate its abundance histograms. Here we present KMERGENIE, a method designed to find the optimal value of k. KMERGENIE uses a sampling-based approach instead of a time-consuming exact counting approach implemented in other tools. It then chooses the value of k that gives the maximum number of genomic k-mers by fitting a generative abundance model to each histogram. To validate its accuracy, we compared the assemblies of different datasets using a k value chosen by KMERGENIE to other assemblies. Our results demonstrate that KMERGENIE selects an informed k-mer size that leads to a genome assembly with high NG50 and low error rate. We anticipate that this method can be integrated into assembly pipelines so that the choice of k can be made automatically without user intervention. The method of automatically choosing k from non-uniform coverage could be tested, and the accuracy of our statistical model can be improved, ensuring efficient and precise sequencing data analysis.

Summary:

The Paper titled “Informed and automated k-mer size selection for genome assembly” by Rayan Chikhi and Paul Medvedev, published in 2013, addresses a key challenge on estimating the optimum size of k-mer for assembling a genome using de Bruijn graph. The authors identified a lack of tools to estimate the optimum k and to efficiently generate its abundances histograms. Therefore, the research question was whether there is a method that automates this process and reduces the time-consuming parameter selection.

To answer this question, they generated an abundance histogram for several assumed values of k using a sampling-based approach, reducing the computation time compared to counting algorithms. After building the abundance histograms, a generative model, that they adapted from Kelley et al. (2010), was fitted to each histogram, to evaluate the unique genomic k-mers. Ultimately the value of k that gives the maximum distinct genomic k-mers was chosen. The method was implemented in the tool KMERGENIE and tested on three datasets from the Genome Assembly Gold-standard Evaluation (GAGE) datasets: *S.aures* (2.8 Mb), human chromosome 14 (88 Mb) and *B.impatiens* (250 Mb) using seven potential k values ranging from 21 to 81. KMERGENIE predicted the optimal k values of 31, 71 and 51 respectively. The authors assembled each genome using the predicted optimal and other reasonable k values, evaluated the resulting assemblies based on contig NG50 length, assembly size and number of errors. KMERGENIE’s predictions yielded best assemblies for *S.auereus* and *B.impatiens* in terms of NG50 and size, while for human *chr14*, the selected k yielded lower NG50 and size but fewer errors. They then compared KMERGENIE to two methods: VelvetOptimizer(VO) and VelvetAdvisor. Unlike these tools, KMERGENIE is designed to select the optimal k-mer size while being faster and applicable on larger datasets. The number of distinct genomic K-mer correlated with assembly quality, but discrepancies at low k values and in heterozygous genomes revealed limitations in the model and assembler.

Nevertheless, the Paper presented that KMERGENIE provides an effective and efficient way of selecting the optimal value of k despite acknowledging its limitations when sequencing coverage is non-uniform. In the future, the authors aim to work towards expanding the applicability of KMERGENIE and improving its accuracy.

Key figure:

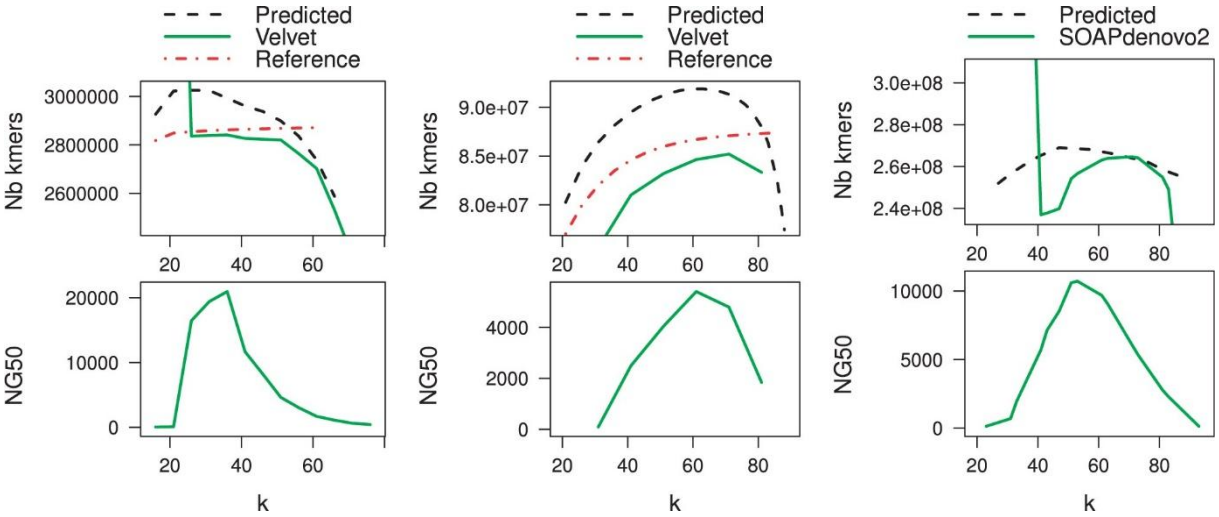


Fig. 4. Relation of the number of distinct genomic k-mers to assembly quality. We show the results for the three datasets: S.aureus (left), chr14 (middle) and B.impatiens (right). We plot the number of distinct genomic k-mers predicted from the histogram from our model, the number present in the reference and the number present in the assembly. We also show the NG50 of the assembly

Figure 4 shows the number of distinct genomic k-mers for the three datasets predicted by KMERGENIE, the reference and the assembly by Velvet and SOAPdenovo2 along with NG50 of the assembly. The number of distinct genomic k-mers in the assembly of *S.aureus* and *chr14* closely resembles the numbers predicted by KMERGENIE, while for *B.impatiens* shows variations to the KMERGENIE's predictions. Despite this, the figure represents a correlation between the number of distinct genomic k-mers predicted and NG50 for all datasets.

Overall, figure 4 assesses the tool presented in the paper, by showing that the predictions of KEMERGENIE lead to the best assemblies while also showing the limitations of KMERGENIE of overestimating the number of genomic k-mers when compared with the reference genome.

Impact:

1. List of the five most relevant references

Reference	Authors (F, C, L)	Title	Journal	Pub.	Citat.	Justification
Pevzner,P.A. et al. (2001)	Pavel A. Pevzner Pavel A. Pevzner Michael S. Waterman	An Eulerian path approach to DNA fragment assembly	PNAS 98 (17) 9748-9753	2001	1915	implemention of k-mers for assembling genomes using de Bruijn graph
Salzberg et al., 2011	Steven L. Salzberg Steven L. Salzberg James A. Yorke	GAGE: a critical evaluation of genome assembly & assembly algorithms	Genome Research 22, 557-567	2011	905	provides GAGE datasets and metrics for qualifying genome assembly
Kelley et al., 2010	David R Kelley David R Kelley Steven L Salzberg	Quake: quality-aware detection & correction of sequencing errors	Genome Biology 11, R116	2010	751	adoption of generative models for estimating the no. of distinct genomic k-mers
Rizk et al., 2013	Guillaume Rizk Rayan Chikhi Rayan Chikhi	DSK: *k*-mer counting with very low memory usage	Bioinformatics Vol. 29, no. 5, 652-653	2013	396	comparison to an existing time-consuming method for counting k-mers
Cormode et al. 2005	Graham Cormode Graham Cormode Irina Rozenbaum	Summarizing & Mining Inverse Distribution Data Streams via Dynamic Inverse Sampling	VLDB Endowment, pp. 25-36	2005	116	sampling k-mers in a time efficient way

2. Google Scholar: 864 and PubMed: 394
2025: 18, 2024: 35, 2023: 37, 2022:3, 2021: 43, 2020: 39, 2019: 47, 2018: 41, 2017: 49, 2016: 43, 2015: 21, 2014:13, 2013: 1

Influential articles:

1. Páll Melsted, Bjarni V. Halldórsson, KmerStream: streaming algorithms for *k* -mer abundance estimation
2. Al-Qurainy, F.; Gaafar, A.-R.Z.; Estimation of Genome Size in the Endemic Species Reseda pentagyna and the Locally Rare Species Reseda lutea Using comparative Analyses of Flow Cytometry and K-Mer Approaches
3. Zhang Q, Pell J, Canino-Koning R, Brown CT; These Are Not the K-mers You Are Looking For: Efficient Online K-mer Counting Using a Probabilistic Data Structure
4. Swati C. Manekar, Shailesh R. Sathe, Estimating the k-mer Coverage Frequencies in Genomic Datasets: A Comparative Assessment of the State-of-the-art
5. Cha S, Bird DM. Optimizing k-mer size using a variant grid search to enhance *de novo* genome assembly

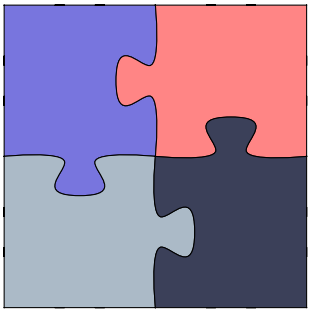
3. The five relevant publications by Rayan Chikhi

Titel	**Journal (Impact Factor)**	**Citation**	**Position**	**Justification**
Towards complete and error-free genome assemblies of all vertebrate species	Nature (69.5 in 2021)	2458	Co-author	His highest citated publication
The complete sequence of a human Y chromosome	Nature (50.5 in 2023)	311	Co-author	Contributed in the first complete sequence of a human genome after the initial draft in 2000
Computability of models for sequence assembly	International Workshop on Algorithms in Bioinformatics (WABI)	239	First-author	Academic conference
Compacting de Bruijn graphs from sequencing data quickly and in low memory	Bioinformatics (7.3 in 2016)	264	Last author	Last author position
Computational methods for discovering structural variation with next-generation sequencing	Nature Methods (9.69 in 2009)	699	First author	First author position

NATURE BIOTECHNOLOGY, 2019

ASSEMBLY OF LONG, ERROR-PRONE READS USING REPEAT GRAPHS

Mikhail Kolmogorov, Jeffrey Yuan, Yu Lin and Pavel A. Pevzner



PRESENTED BY
DILARA SARACH AND WANYING DENG

Abstract by Wanying Deng

Genome assembly is the process of reconstructing complete genomes from DNA sequencing reads. While long single-molecule sequencing (SMS) reads offer improved resolution over short-read data, assembling repetitive genomic regions, especially complex segmental duplications (SDs), is still challenging. Traditional assemblers (such as PacBio and ONT) often fail to fully resolve these repeats, limiting assembly quality. Previous methods struggle particularly with unbridged repeats and mosaic repeat structures, which are common in large genomes like humans. The key problem addressed by this study is how to improve the assembly of complex, highly repetitive genomic regions, using error-prone long reads. Here, the authors present Flye, a novel long-read assembly algorithm, which builds a repeat graph, and allowing it to resolve the repeats. Flye produces more contiguous and accurate assemblies compared to five state-of-the-art assemblers (Canu, Falcon, HINGE, Miniasm, MaSuRCA), nearly doubling the NGA50 for the human genome. Furthermore, Flye effectively reconstructs detailed mosaic SD structures of repeats. These results suggest that incorporating repeat graph approaches into long-read assembly has transformative potential, improving not just basic assembly metrics but also enabling deeper insights into genomic structures and variations. Flye's approach could significantly reduce the need for additional finishing experiments and enhance studies of genome evolution, disease, and structural variation. It provides a major step forward in leveraging long-read data for high-quality genome assemblies, paving the way for more comprehensive and accurate genomic research.

Legend: ● Basic introduction ● Detailed background ● General problem ● "Here we show" ● Main problem ● General context ● Broader perspective

Abstract by Dilara Sarach

Have we de-mystified repeat untangling in genome assembly? Yes and no – short-read data relies on tried-and-true approaches to reconstruct repetitive regions as well as conceivably possible given the inherent limitations in working with 300bp segments; long-read data, while indispensable for better repeat resolution, are a different – and yet undefeated – beast. Firstly, single-molecule sequencing reads fall short of being tractable to the classic de Bruijn graph method by one unfulfilled requirement – most k-mers in the genome must be preserved in multiple reads. Secondly, error-prone reads complicate the distinguishing of repeat copies with divergence below 10%. Both challenges torpedo attempts at resolving bridged and unbridged repeats. Here we show that long-read genome assembly can be significantly improved in quality and runtime by exploiting one of its main flaws – repeat copy variants. Counter-intuitively, we assemble reads sloppily and concatenate the resulting "disjointigs" arbitrarily, but arrive at a repeat graph guaranteed to be same as if derived from the complete genome. Finally, we use errors in repeats to find a Eulerian tour trough the graph – an assembly that is on par with or better than those produced by five state-of-the-art assemblers. Our algorithm doubled the contiguity of the human genome assembly. The potential for improvement in genome reconstruction seems far from exhausted, and our results show a direction worth further exploring. Algorithms tackling mosaic segmental duplications are needed; so are approaches resolving repeats with divergence below 3%. Once achieved, these milestones could elevate assembly contiguity, as measured by NGA50, by an order of magnitude.

Summary

Kolmogorov et al. address the critical challenge of assembling long, error-prone single-molecule sequencing (SMS) reads, such as those produced by PacBio and Oxford Nanopore Technologies (ONT), into accurate and contiguous genome assemblies. A major bottleneck in assembling such reads has been the presence of repetitive and complex genomic regions, particularly segmental duplications (SDs) and unbridged repeats, which often prevent standard assemblers from producing complete or correct assemblies. The study's central research questions focus on whether these technical limitations can be overcome with new algorithmic approaches, and whether it is possible to achieve high-quality assemblies without relying on auxiliary data like Hi-C, optical maps, or mate-pair libraries.

To address these challenges, the authors introduce Flye, a novel long-read assembler designed to handle the high error rates of SMS reads and to resolve complex repeat structures. The Flye algorithm begins by constructing disjointigs from the raw reads. It then builds a repeat graph and applies graph simplification and untangling methods to resolve both bridged repeats and unbridged. Flye can also use the subtle sequence differences between repeat copies, which allows it to disentangle even highly similar unbridged repeats, what traditional assemblers often struggle with. This research benchmarked Flye on a diverse set of genomes, including bacteria, yeast, worms, humans, and complex metagenomic datasets, comparing its performance to established assemblers such as Canu, Falcon, HINGE, Miniasm, and MaSuRCA.

The results show that Flye consistently outperforms competitors in terms of both accuracy and contiguity. Notably, Flye produced higher NGA50 values in large and complex genomes, indicating more contiguous assemblies that are closer to the reference. Moreover, Flye's runtime was significantly faster, sometimes by an order of magnitude, making it both efficient and scalable. It demonstrates that Flye can resolve previously inaccessible regions of the human genome, such as mosaic segmental duplications, which are known to play crucial roles in genome evolution and disease. It is concluded that Flye marks a significant advancement in long-read genome assembly.

A key figure in the paper, Figure 4, illustrates how Flye resolves mosaic segmental duplications in the human genome. Panel (a) illustrates a specific example of a mosaic SD of complexity 7, where long repetitive segments are connected across multiple chromosomes. Panel (b) presents the overall statistics of SD length and complexity, comparing results quantitatively before and after resolving repeats with standard and ultra-long ONT reads. We consider Figure 4 a key figure because it visually demonstrates the major advantage of ultra-long reads: they enable the resolution of complex and previously unresolved genomic duplications, which are critical for accurate genome assembly and for matching the reference human genome. This figure underscores Flye's capability to resolve genomic complexities that are otherwise challenging or impossible with standard sequencing methods.

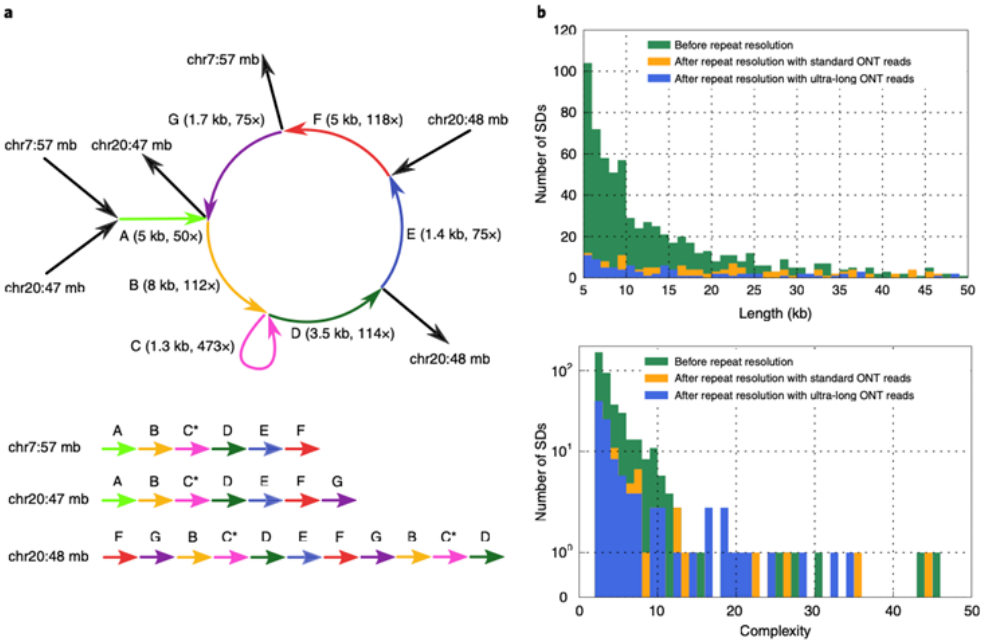


Fig. 4 | An SD from the Flye assembly of the HUMAN dataset and the distribution of the lengths and complexities of all SDs from the same assembly. **a**, A mosaic SD of complexity 7 represented as a connected component formed by repeat edges (7 colored edges of total length 25.7 kb) in the assembly graph of the HUMAN dataset (flanking unique edges shown in black). The loop-edge C with coverage 473× represents a tandem repeat C* with unit length 1.3 kb that is repeated ~19 times. The colored edges of the assembly graph align to a region on chromosome 7 of length 31 kb and two regions on chromosome 20 of lengths 30 kb and 46 kb. These three instances of SDs were not resolved using standard ONT reads but were resolved using ultra-long reads in a way that is consistent with the reference human genome. **b**, Statistics are given before resolving bridged repeats (green), after resolving bridged repeats with standard ONT reads (orange), and with ultra-long ONT reads (blue). Only SDs between 5 kb and 50 kb in length and with complexity between 2 and 50 contributed to the SD length and SD complexity histograms. Only two SDs have complexity exceeding 50 before bridged repeat resolution. Of the 688 SDs between 5 kb and 50 kb, 545 were resolved using the standard ONT reads, and ultra-long reads resolved an additional 58 SDs. There were 1,256 simple SDs before bridged repeat resolution and 143 after bridged repeat resolution with ultra-long reads. Since Flye usually resolves SDs shorter than the typical read length, the SDs identified by Flye do not include many known human SDs.

Citations: 4369

Citations per year: 728

5 most influential papers cited by the paper:

Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation (2017)

Authors: S. Koren[F], A.M. Phillippy [C, L]
Journal: Genome Res. (27: 722–736)
Citations: 6934
Relevance: most cited in the paper (8x)

De novo repeat classification and fragment assembly (2004)

Authors: P.A. Pevzner[F], G. Tesler [C, L]
Journal: Genome Res. (14: 1786-1796)
Citations: 341
Relevance: Flye improved on this approach in constructing repeat graphs

HINGE: long-read assembly achieves optimal repeat resolution (2017)

Authors: G.M. Kamath[F], I. Shomorony[F], F. Xia[F], T.A. Courtade[C], D.N. Tse[C, L]
Journal: Genome Res. (27: 747-756)
Citations: 131
Relevance: Flye is compared with the approach

SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing (2012)

Authors: A. Bankevich [F], M.A. Alekseyev [C], P.A. Pevzner [L]
Journal: J Comput Biol. (19: 455–477)
Citations: 24921
Relevance: high citation count, introduces the foundation for graph simplification

Assembly of long error-prone reads using de Bruijn graphs (2016)

Authors: Y. Lin [F], P.A. Pevzner [C, L]
Journal: Proc. Natl. Acad. Sci. (113: E8396–E8405)
Citations: 365
Relevance: Flye's repeat graph structure is developed from ABruijn

5 most influential papers citing the paper:

Major Impacts of Widespread Structural Variation on Gene Expression and Crop Improvement in Tomato (2020)

Authors: M. Alonge [F], Z.B. Lippman [C, L]
Journal: Cell (IF: 45.6)

Microbial liberation of N-methylserotonin from orange fiber in gnotobiotic mice and humans (2022)

Authors: N.D. Han [F], J.I. Gordon [C, L]
Journal: Cell (IF: 45.6)

The evolution of two transmissible cancers in Tasmanian devils (2023)

Authors: M.R. Stammnitz [F], E.P. Murchison [C, L]
Journal: Science (IF: 44.8)

Variant calling and benchmarking in an era of complete human genome sequences (2023)

Authors: N.D. Olson [F], J.M. Zook [C, L]
Journal: Nat Rev Genet. (IF: 39.1)

Genome assembly in the telomere-to-telomere era (2024)

Authors: H. Li [F, C], R. Durbin [C, L]
Journal: Nat Rev Genet. (IF: 39.1)

5 most influential by the paper's corresponding author (P.A. Pevzner):

SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing (2012)

Authors: A. Bankevich [F], M.A. Alekseyev [C], P.A. Pevzner [L]
Journal: J Comput Biol. (IF: 1.5)
Citations: 24921
Relevance: author's most cited publication with last author credit; author's most cited publication overall; publication introduced a state-of-the-art assembler

Assembly of long, error-prone reads using repeat graphs (2019)

Authors: M. Kolmogorov [F], P.A. Pevzner [C, L]
Journal: Nat Biotechnol (IF: 33.1)
Citations: 4369
Relevance: author's most cited publication with corresponding-author credit; author's most cited publication in a high-impact journal; author's second most cited paper despite the recent publishing year

metaSPAdes: a new versatile metagenomic assembler (2017)

Authors: S. Nurk [F, C], P.A. Pevzner [L]
Journal: Genome Res. (IF: 6.2)
Citations: 3962
Relevance: last author in a paper that introduced a state-of-the-art metagenomic assembler

Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution (2004)

Authors: International Chicken Genome Sequencing Consortium
Journal: Nature (IF: 50.5)
Citations: 2997
Relevance: highest-impact journal the author has published in

An Eulerian path approach to DNA fragment assembly (2001)

Authors: P.A. Pevzner [F], M.S. Waterman [C, L]
Journal: Proc. Natl. Acad. Sci. (IF: 9.4)
Citations: 1915
Relevance: author's most cited publication with first-author credit

[F] first author, [C] corresponding author, [L] last author

Aktuelle Themen der Sequenzanalyse

BRAKER3: Fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS, and TSEBRA

Lars Gabriel, Tomáš Bruna, Katharina J. Hoff, Matthis Ebel, Alexandre Lomsadze, Mark Borodovsky, and Mario Stanke

Journal:
Genome Research, 34: 757-768
Publication Year: 2024
Issue: Vol. 34, No. 5
Pages: 769-–777

Presenting Persons:
Lucie Biesecker, 7518074 & Emre Inciler, 8675150

July 6, 2025

Abstract 1

Gene prediction in eukaryotic genomes is still a difficult task. Many genomes differ in the quality and amount of available data, making it hard to create accurate annotations. Previous tools like BRAKER1 and BRAKER2 could use either RNA-seq or protein data, but not both simultaneously. Some newer technologies try to enhance gene prediction by combining various kinds of data. However, there is still no easy and fully automatic tool that can use both RNA-seq and protein data to give precise gene predictions. Here we show that BRAKER3, a pipeline that runs GeneMark-ETP and AUGUSTUS for gene prediction and then uses TSEBRA to select the best transcripts with joint extrinsic evidence, enables accurate genome annotation. BRAKER3 performs more effectively than prior tools such as MAKER2, Funannotate and FINDER. It finds genes and transcripts more correctly, with F1-scores up to 20% higher. Although BRAKER3 is slower than some other tools, it works well even with large genomes and big protein databases. These results indicate that BRAKER3 is an effective tool for large genome studies such as the Earth BioGenome Project. The pipeline is provided as a Docker container for easy deployment, making it accessible for a wide range of users. This makes BRAKER3 a useful option not only for scientists working on genome research, but also for substantial genome projects that need fast and accurate results.

Abstract 2

Accurate gene annotation is essential for understanding the biological functions encoded in genomes, as well as for supporting research in many areas of the life sciences. As genome sequencing becomes faster and more affordable, the demand for accurate and highly automated annotation tools continues to increase. Gene annotation can be supported by both intrinsic and extrinsic evidence. Intrinsic evidence is based on computational models and statistical features of the genome sequence itself, while extrinsic evidence includes external data sources such as transcriptome data or known protein sequences. The previous annotation tools BRAKER1 and BRAKER2 use either RNA-seq data or protein data as evidence, but not a combination of both. Here we present BREAKER3, a new gene annotation tool for eukaryotic genomes, which integrates GeneMark-ETP, continues the annotation with AUGUSTUS, and combines the results with TSEBRA. With GeneMark-ETP, it is now possible to use both RNA-seq and protein data to improve annotation accuracy compared to previous models. In addition, the integration of the ab initio gene prediction tool AUGUSTUS enables the training of statistical models on high-confidence genes to further optimize the annotation. This approach of BRAKER3 outperforms both its predecessors and other gene annotation tools in benchmark tests. Evaluation on 11 well-annotated genomes showed that BRAKER3 achieved an average F1-score that was 20% higher than that of the compared tools, with the largest advantage observed on large and complex genomes. With its improved accuracy and ability to handle complex genomes, BRAKER3 enables more comprehensive and meaningful gene annotations.

Summary

1. Research Question Addressed
The study addresses how to improve the accuracy and automation of genome annotation in eukaryotes by integrating heterogeneous data types, such as RNA-seq and protein evidence. It investigates if combining various data sources into a single pipeline outperforms existing genome annotation methods.

2. Key Method
BRAKER3 is a fully automated genome annotation pipeline that integrates GeneMark-ETP, AUGUSTUS, and TSEBRA to improve gene prediction accuracy. GeneMark-ETP assembles RNA-seq transcripts, predicts protein-coding genes, and selects high-confidence models for training. AUGUSTUS uses this training set for genome-wide prediction, guided by GeneMark-ETP hints. TSEBRA then scores and filters transcripts based on support from extrinsic evidence, selecting the most reliable models.

3. Relevant Results
BRAKER3 outperforms BRAKER1, BRAKER2, TSEBRA, GeneMark-ETP, MAKER2, Funannotate, and FINDER in precision and sensitivity across 11 species, notably in large or GC-heterogeneous genomes. It exceeds GeneMark-ETP and AUGUSTUS in gene and transcript prediction accuracy but shows limitations with single-exon genes and low RNA-seq coverage. On unannotated genomes, it achieves high BUSCO completeness, emphasizing precision over gene count and reducing false positives. Though slower than some tools, it scales well with increasing genome and protein database sizes.

4. Knowledge Gap
Despite its improved accuracy through integration of RNA-seq and protein evidence, BRAKER3 is limited to predicting protein-coding genes, discarding non-coding transcripts. Additionally, its reliance on RNA-seq data prevents its use in protein-only evidence scenarios.

5. Conclusion
BRAKER3 improves eukaryotic genome annotation by combining RNA-seq and protein evidence using GeneMark-ETP, AUGUSTUS, and TSEBRA. Benchmarking across diverse species demonstrated that BRAKER3 consistently outperforms previous versions and other leading annotation tools in both accuracy and precision.

6. Key Figure
Figure 2 provides the paper's central claim, that BRAKER3 outperforms other genome annotation tools, by visualizing accuracy improvements.

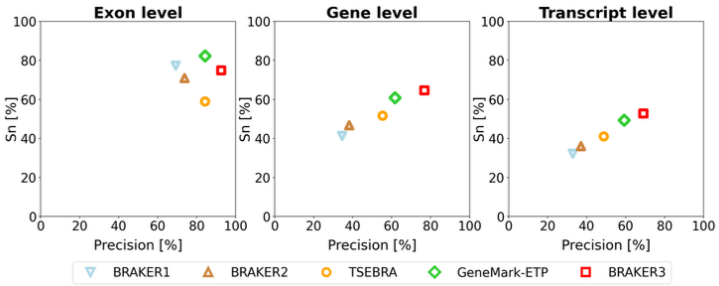


Figure 2: Average precision and sensitivity of gene predictions made by BRAKER1, BRAKER2, TSEBRA, GeneMark-ETP, and BRAKER3 for the genomes of 11 different species (listed in Supplemental Table S1). Inputs were the genomic sequences, short-read RNA-seq libraries, and protein databases (order excluded).

7. Outlook
Future improvements could include support for non-coding RNA prediction, integration of long-read RNA-seq data, and enhanced handling of low-expression genes. As genome sequencing scales up globally, BRAKER3's automation and accuracy make it well-suited to support projects like the Earth BioGenome Project.

Paper Impact

Key Supporting References

1. **Authors:** First: Tomáš Brůna and Alexandre Lomsadze (contributed equally), Corresponding: Mark Borodovsky, Last: Mark Borodovsky
   Title: *GeneMark-ETP significantly improves the accuracy of automatic annotation of large eukaryotic genomes*
   Journal: Genome Research, 34: 757-768
   Year: 2024
   Citations: 63
   Relevance to the Paper: GeneMark-ETP is part of the BRAKER3 pipeline.

2. **Authors:** First: Mario Stanke, Corresponding: Mario Stanke, Last: Burkhard Morgenstern
   Title: *AUGUSTUS: ab initio prediction of alternative transcripts*
   Journal: Nucleic Acids Research, 34: W435–W439
   Year: 2006
   Citations: 2503
   Relevance to the Paper: AUGUSTUS is part of the BRAKER3 pipeline.

3. **Authors:** First: Lars Gabriel, Corresponding: Mario Stanke, Last: Mario Stanke
   Title: *TSEBRA: transcript selector for BRAKER*
   Journal: BMC Bioinformatics, 22: 566
   Year: 2021
   Citations: 239
   Relevance to the Paper: TSEBRA is part of the BRAKER3 pipeline.

4. **Authors:** First: Alexandre Lomsadze, Corresponding: Mark Borodovsky, Last: Mark Borodovsky
   Title: *Gene identification in novel eukaryotic genomes by self-training algorithm*
   Journal: Nucleic Acids Research, 33: 6494–6506
   Year:
   Citations: 2005
   Relevance to the Paper: GeneMark.hmm is fundamental to BRAKER3, as it introduced the HMM-based gene prediction framework that later evolved into GeneMark-ETP.

5. **Authors:** First: Tomáš Brůna and Katharina J. Hoff (contributed equally), Corresponding: Mark Borodovsky, Last: Mario Stanke and Mark Borodovsky
   Title: *BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database*
   Journal: NAR genomics and bioinformatics,
   Year: 2021
   Citations: 1515
   Relevance to the Paper: BRAKER2 is the previous version that only uses protein data as extrinsic evidence.

Citation Analysis

Total Citations: 314
Citations per Year: 104

1. **Title:** *Genome assembly in the telomere-to-telomere era*
   Journal: Nature Reviews Genetics
   5 Year Impact Factor: 52.3

2. **Title:** *The European Reference Genome Atlas: piloting a decentralised approach to equitable biodiversity genomics*
   Journal: Nature Partner Journals Biodiversity
   5 Year Impact Factor: not available yet
3. **Title:** *GeneMark-ETP significantly improves the accuracy of automatic annotation of large eukaryotic genomes*
   Journal: Genome Research
   5 Year Impact Factor: 8.4
4. **Title:** *Adaptation repeatedly uses complex structural genomic variation*
   Journal: Science
   5 Year Impact Factor: 62.3
5. **Title:** *Multiple Horizontal Transfers of Immune Genes Between Distantly Related Teleost Fishes*
   Journal: Molecular Biology and Evolution
   5 Year Impact Factor: 14.5

Corresponding Author: Katharina J. Hoff

1. **Title:** *BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database*
   Author Position: First
   Journal: NAR genomics and bioinformatics
   5 Year Impact Factor: 4.1
   Citations: 1515
   Reason for Selection: The journal does not have a high impact factor but it is her most cited paper and she is a first author.
2. **Title:** *Butterfly genome reveals promiscuous exchange of mimicry adaptations among species*
   Author Position: Middle (as part of the Heliconius Genome Consortium)
   Journal: Nature
   5 Year Impact Factor: 54,4
   Citations: 1257
   Reason for Selection: Her second most cited paper and Nature has a high impact factor.
3. **Title:** *Standards recommendations for the Earth BioGenome Project*
   Author Position: Middle
   Journal: Proceedings of the National Academy of Sciences of the United States of America
   5 Year Impact Factor: 10.8
   Citations: 87
   Reason for Selection: The paper establishes key standards to ensure high-quality, consistent genome data for the global Earth BioGenome Project, which aims to sequence all known eukaryotic species.
4. **Title:** *WebAUGUSTUS—a web service for training AUGUSTUS and predicting genes in eukaryotes*
   Author Position: First and corresponding
   Journal: Nucleic Acids Research
   5 Year Impact Factor: 16.1
   Citations: 309
   Reason for Selection: This was published in a high-impact journal and enables easy browser-based gene prediction for all users; we also used WebAUGUSTUS in a comparative genomics exercise.
5. **Title:** *Tiberius: end-to-end deep learning with an HMM for gene prediction*
   Author Position: Middle
   Journal: Bioinformatics
   5 Year Impact Factor: 7.6
   Citations: 4
   Reason for Selection: Tiberius is relevant becaus it improves gene prediction by combining deep learning with HMMs, enhancing accuracy in genome annotation.

Abstract book

“MetaEuk – a sensitive, high-throughput gene discovery and annotation for large-scale eukaryotic metagenomics”

by E. Levy Karin, M. Mirdita and J. Soeding

Published in Microbiome 8, number of article 48, in 2020.

Presented by Valeriya Kolos and Ischa Tahir

Goethe University Frankfurt

29th June of 2025

Unicellular eukaryotes inhabit every ecological niche, playing vital roles in global ecosystems. 18S ribosomal DNA metabarcoding has revealed huge diversity among them, uncovering new taxons with significant potential for biotechnology and biomedicine. Prediction of eukaryotic protein-coding genes in metagenomic assemblies is often complicated by their exon-intron structure, typically large genome sizes and low content in samples. Existing methods usually require taxonomic binning or species-specific training data, limiting their use to poorly characterized or *de novo* discovered species with very few references. Therefore, there is a need for tools that can predict and reconstruct eukaryotic genes without prior knowledge of the organisms present in metagenomes.

Here we introduce MetaEuk, a scalable, reference-based tool for sensitive identification and annotation of eukaryotic protein-coding genes in metagenomes. MetaEuk accurately recovers complex gene structures within different eukaryotic clades, predicting over 12 million genes from 1,35 million contigs of Tara Oceans metagenomic datasets. It uses 6-frame translation, rapid homology searches and exon-chaining to reconstruct gene models, achieving over 90% sensitivity in benchmarked genomes, even with low similarity to references. This approach outperforms traditional tools, dependent on genome binning or closely related training data.

These results show that MetaEuk enables efficient, large-scale discovery of eukaryotic genes from metagenomic data, making it particularly valuable for studying uncultured and diverse eukaryotes in natural environments. As environmental sequencing is becoming more common, it has a potential for uncovering new protein families, improving reference databases and advance functional and evolutionary analyses of eukaryotic microbial life.

244 words

Abstract of the Paper "MetaEuk - sensitive, high-throughput gene discovery and annotation for large-scale eukaryotic metagenomics" by Eli Levy Karin, Milot Mirdita and Johannes Söding (2020)

Metagenomics analyses all DNA from different microorganisms in a given environmental sample. By analysing them, we gain a wide range of knowledge about their function and metabolism which provide more biomedical usage. The eucaryotes have been an important part in this due to their evolutionary variability and their still undetected functions and gene parts. In past, their metabolism have provided many ways for therapy in biological and biomedical fields. Their intron-exon structure, increased genome sizes and fewer reference genomes makes it difficult to analyse, annotate and predict proteins of such genomes. The large amounts of contigs in usual methods leave a sampling sign that makes the analysis unclear.

Here we introduce the MetaEuk algorithm that provides an annotation and gene discovery tool for large-scale eukaryotic metagenomics to identify single- and multi-exon protein coding genes. By making no assumptions about splicing signals it does not rely on binning steps and identifies putative exons within the fragments. It was benchmarked using annotated genomes and proteins of different organisms from different parts of the eucaryotic phylogenetic tree to ensure it works for distantly related species.

With this new tool, the eucaryotic metagenomics can be analysed faster, clearer and with better results. By matching with referenced databases, it ensures safety and transparency. MetaEuk provides a future tool to analyse and explore new protein and gene regions in the eucaryotic metagenomics field and opens a wide range of possible hidden secrets of knowledge on variability and possibilities for biotechnology and biomedicine.

246 words

Summary of the Paper "MetaEuk - Sensitive, High-Throughput Gene Discovery and Annotation for Large-Scale Eukaryotic Metagenomics" by Eli Levy Karin, Milot Mirdita and Johannes Söding

1. **Research questions treated in the paper**
   - How can metagenome data from eucaryotes be analysed easily, fast and in a secured/referenced way?
   - How can protein-coding gene parts from eucaryotic metagenomic data efficiently be annotated?
   - How can the genetic variability of eucaryotic microorganisms be analysed?
     ➔ The analysis of eucaryotic organisms in metagenomics was not given yet because of difficulties in complex intron-exon-structures and less advanced algorithms/technologies
     ➔ MetaEuk as a given solution from the authors to solve this problem

2. **Relevant methodic attempts**
   - MetaEuk algorithm: recognizing protein-coding genes in eucaryotic gene parts based on data banks → Reference data banks usage to ensure quality and right information
   - Protein parts of Contigs are translated into six reading parts
   - Spliced alignment and dynamic programming: exons being combined into multi-exon proteins
   - Clustering: more and often called protein gene parts are put into cluster to find them easily
     ➔ MetaEuk is very fast
     ➔ Can be used on different sequences (using MMseqs2)

3. **Relevant results**
   - Strong benchmarking tool: 90 % sensitivity on nine organisms
   - Less than 0,01% false-positive in test results
   - More than 12 million protein-coding genes were predicted in 8 days on 16 core servers
   - MetaEuk finds high diverse protein coding genes that were unknown before

4. Conclusion

MetaEuk is a fast and secure annotation tool that uses data banks as a secure reference and is capable in finding new protein-coding genes in eucaryotic microorganisms in metagenomics.

5. Key illustration

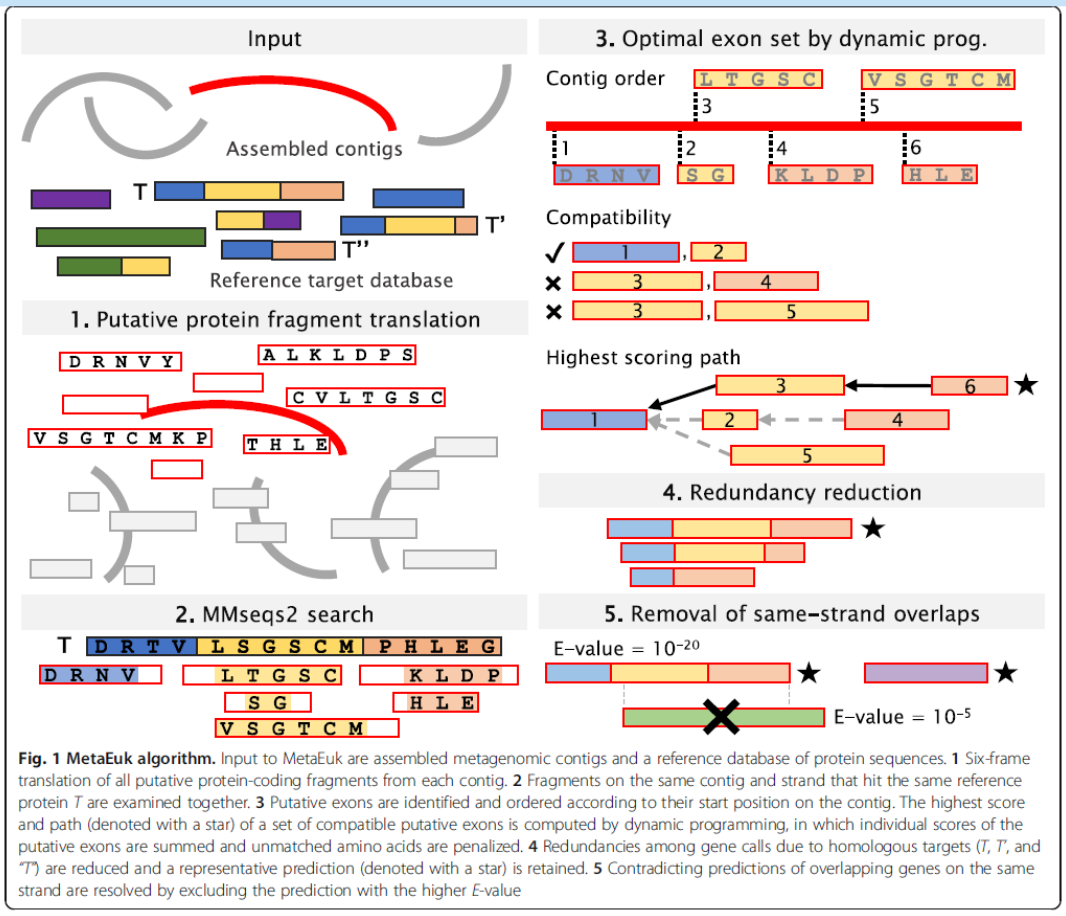


Fig. 1 MetaEuk algorithm. Input to MetaEuk are assembled metagenomic contigs and a reference database of protein sequences. **1** Six-frame translation of all putative protein-coding fragments from each contig. **2** Fragments on the same contig and strand that hit the same reference protein *T* are examined together. **3** Putative exons are identified and ordered according to their start position on the contig. The highest score and path (denoted with a star) of a set of compatible putative exons is computed by dynamic programming, in which individual scores of the putative exons are summed and unmatched amino acids are penalized. **4** Redundancies among gene calls due to homologous targets (*T*, *T'*, and *"T"*) are reduced and a representative prediction (denoted with a star) is retained. **5** Contradicting predictions of overlapping genes on the same strand are resolved by excluding the prediction with the higher *E*-value

This illustration is considered to be the key illustration of the paper because it emphasizes the key aspects of how the actual MetaEuk algorithm works with all of its steps in between, given from the input, through the MMseqs2 search, the dynamic programming approach in exon processing, the reduction of redundancy which makes the algorithm very effective and useful and the last step of removing the same-strand overlaps by excluding the prediction with a higher E-value (Karin et al., 2020).

References:

Levy Karin, E., Mirdita, M. & Söding, J. MetaEuk—sensitive, high-throughput gene discovery, and annotation for large-scale eukaryotic metagenomics. *Microbiome* **8**, 48 (2020). https://doi.org/10.1186/s40168-020-00808-x

Paper impact of paper 5: MetaEuk

References are not sorted by importance but simply listed in the order they appear in the text.

Number of citations are given in curly brackets { }.

1) **First reference:**

   Escobar-Zepeda A, Vera-Ponce de León A, Sanchez-Flores A. The road to metagenomics: from microbiology to DNA sequencing technologies and bioinformatics. Front Genet. 2015;6:348. {324}

   This reference gives a **huge overview on the bioinformatical analysis of microorganisms**. It generally describes approaches to research of eukaryotic genomes and their challenges, and it confirms the need for the tool.

2) **Second reference:**

   Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol. 2017;35:1026-8. {1973}

   MMseqs2 is a very important search tool which is involved in the workflow of MetaEuk. It is used to identify putative exons and calculate Smith-Waterman scores for sensitivity and reliability evaluation.

3) **Third reference:**

   Carradec Q, Pelletier E, Da Silva C, Alberti A, Seeleuthner Y, Blanc-Mathieu R, et al. A global ocean atlas of eukaryotic genes. Nat Commun. 2018;9:373. {236}

   The authors tested the whole toolkit on **Tara Oceans** datasets.

4) **Fourth reference:**

   Bateman A, Martin MJ, O'Donovan C, Magrane M, Alpi E, Antunes R, et al. UniProt: the universal protein knowledgebase. Nucleic Acids Res. 2017;45:D158–69. {4370}

   The authors used **UniRef90** (protein database) a lot to evaluate MetaEuk's performance on benchmark data, particularly sensitivity with simulated 'evolutionary distant' relatives in the reference database and annotation of individual exons.

5) **Fifth reference:**

   Ovchinnikov S, Park H, Varghese N, Huang P-S, Pavlopoulos GA, Kim DE, et al. Protein structure determination using metagenome sequence data. Science. 2017;355:294–8. {383}

   This reference shows **perspectives and the potential role** of MetaEuk for the future.

The article "MetaEuk–sensitive, high-throughput gene discovery and annotation for large-scale eukaryotic metagenomics" by E. Levy Karin, M. Mirdita and J. Soeding was published in 2020. The number of citations is 145. Hence, the number of citations per year are ~24,4 (including 2020 and 2025).

Five most influential paper that cite this paper:

From Google Scholar, sorted by number of citations:

1) Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., & Steinegger, M. (2022). ColabFold: making protein folding accessible to all. *Nature methods*, *19*(6), 679-682.

2) Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A., & Zdobnov, E. M. (2021). BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Molecular biology and evolution*, *38*(10), 4647-4654.

3) Manni, M., Berkeley, M. R., Seppey, M., & Zdobnov, E. M. (2021). BUSCO: assessing genomic data quality and beyond. *Current Protocols*, *1*(12), e323.

4) Li, H. (2023). Protein-to-genome alignment with miniprot. *Bioinformatics*, *39*(1), btad014.

5) Delmont, T. O., Gaia, M., Hinsinger, D. D., Frémont, P., Vanni, C., Fernandez-Guerra, A., ... & Pelletier, E. (2022). Functional repertoire convergence of distantly related eukaryotic plankton lineages abundant in the sunlit ocean. *Cell Genomics*, *2*(5).

Resource from Google Scholar, URL: https://scholar.google.com/scholar?start=0&hl=de&as_sdt=2005&sciodt=0,5&cites=14207659728442917523&scipsc=. Accessed 06-29-2025.

Corresponding author – Eli Levy Karin

Most important publications:

1) **MetaEuk–sensitive, high-throughput gene discovery and annotation for large-scale eukaryotic metagenomics**, by E Levy Karin, M Mirdita, J Soeding. Microbiome. (2020).
   Citations: 245. Position of author in author's list: 1 (first author).

2) **Fast and sensitive taxonomic assignment to metagenomic contigs**. M Mirdita, M Steinegger, F Breitwieser, J Söding, E Levy Karin. (2021). Bioinformatics 37 (18), 3029-3031.
   Citations: 219. Position of author in author's list: 5 (last author).

3) **SpacePHARER: sensitive identification of phages from CRISPR spacers in prokaryotic hosts**.
   R Zhang, M Mirdita, E Levy Karin, C Norroy, C Galiez, J Söding. (2021). Bioinformatics 37 (19), 3364-3366.
   Citations: 85. Position of author in author's list: 3.

4) **Easy and accurate protein structure prediction using ColabFold**.
   G Kim, S Lee, E Levy Karin, H Kim, Y Moriwaki, S Ovchinnikov and more (2024). Nature Protocols, 1-23
   Citations: 53. Position of author in author's list: 3.

5) **An integrated model of phenotypic trait changes and site-specific sequence evolution**.
   E Levy Karin, S Wicke, T Pupko, I Mayrose. (2017).
   Systematic biology.
   Citations: 39. Position of author in author's list: 1 (first author).

The references are sorted by the number of citations and represent the impact factor sorted by Google scholar.

Reference:

Google Scholar site of the author. URL: https://scholar.google.com/citations?user=DBhNkjQAAAAJ&hl=de&oi=ao. Accessed 06-08-2025.

A long-read RNA-seq approach to identify novel transcripts of very large genes

Prech Uapinyoying, Jeremy Geckos, Susan M. Knoblach, et al.

Genome Research, 30(6), 885-897 (2020)

Tao Le

and

Jörn Fischer

Abstract by Tao Le:

Alternative splicing allows individual genes to generate multiple mRNAs. Many of these mRNAs encode functionally distinct protein isoforms, thereby bridging the gap between genome and proteome. Short-read sequencing struggles with alternative splicing analysis because individual reads typically cover no more than two exon junctions. This limits the ability to resolve full exon compositions and their phasing within transcripts. Long-read sequencing addresses this by capturing entire transcripts, but it lacks the precision needed for accurate quantification, making it less suitable for differential expression analysis. Here we show that combining PacBio long-read isoform sequencing with a novel analysis approach enables comparison of alternative splicing in large, repetitive structural genes in muscle. We found that Nrap isoforms excluding exon 12, previously considered cardiac-specific, are also expressed in skeletal muscle, and a rare isoform lacking both exons 2 and 12 is present in cardiac and soleus muscle. In Nebulin, we identified a novel exon (u-002), mutually exclusive splicing of exons 127 and 128, and several unannotated phased isoforms showing fiber-type–specific patterns. In Titin, we discovered exon 191 as a likely unannotated cassette exon present in all skeletal but only a subset of cardiac transcripts. Our quantitative analysis with full-length reads enabled isoform-level comparisons across tissues, demonstrating how long-read sequencing reveals complex, tissue-specific splicing and uncovers unannotated isoforms. Improved transcript identification and quantification from our approach removes prior barriers to quantitative differential expression of ultralong transcripts. Unannotated exons and splicing patterns may directly impact clinical sequencing and interpretation of muscle disease variants.

Abstract by Jörn Fischer:

Alternative Splicing in eukaryotes is the ability to express different versions of the same gene to adapt to different circumstances. Albeit short reads have a high quality, they lack the ability to cover multiple splice junctions. This limitation makes it hard to discover rare isoforms. On the other hand, long reads exceed in this filed but have lower quality and coverage leading to limited usability for differential expressions analysis. Here we show a novel analysis pipeline using long reads generated by PacBios Hifi Iso-Seq technology to not only discover new isoforms but to open the door for differential expression analysis using long reads. As an example, we are able to find a Nrap isoform, previously thought to only be expressed in cardiac tissue, to be present in all muscle tissue types. Additionally, we have a high enough coverage to compute meaningful PSI values for all tissue types. This information enables us to identify the isoforms using Sanger sequencing. Our results demonstrate how long read sequencing can provide not only information of novel splice events but also correct older believes based on insufficient data. The increasing accessibility of high quality long read data gives us more and more opportunity not only make new discoveries of splicing events but also to challenge our knowledge thus far. Thereby, we are not limited to the identification of splice events, but can perform quantitative analysis for a wast number of splicing events.

Introduction

For studying alternative splicing, short reads technology leads to limitations due to its inability to span more than two exon junctions per read. This makes it difficult to accurately determine the composition and phasing of exons within transcripts. Although long-read sequencing improves this issue, it is not amenable to precise quantitation, which limits its utility for differential expression studies. The goal of this study is to use PacBios long-read isoform sequencing combined with a novel analysis approach to compare alternative splicing of large, repetitive structural genes in muscles.

Methodes

This study uses mRNA from cardiac, soleus and EDL maus muscle tissue, subjected to long read-read sequencing using PacBio Iso-Seq method and HiFi protocol (max 10kb) to produce consensus reads with an error rate of less then 0.12%. GENCODE (release M10) of mouse is used as reference for the alignment. The short-read data for the comparison are produced by Singh et al. 2018. For analyzing of the read data, two pipelines were developed, exCOVator, used to identify unannotated exons and differential exon usage and exPhaser to quantify and annotate splicing patterns of larger transcript structures for given exons. Filtering of the data with cutoffs of 30 times consensus read coverage and 20% difference in PSI as well as further manual inspection leads to 285 unannotated exons/exonic parts with alternative splicing. The results are confirmed with endpoint PCR and Sanger sequencing.

Results

A finding of the present of nrap (5kb) isoforms excluding exon 12, previously believed to only exist in cardiac muscles (Lu et a. 2008), in skeletal muscles of mouse. On the other hand, the exclusive expression of the isoform excluding exon 12 in cardiac muscles is verified. Additionally, a rare transcript excluding exon 2 in addition to exon 12 is present in cardiac and soleus muscles.

In the Z-disk region of nebulin (22kb), a novel exon (u-002) was identified, and most exons were skipped in fast-twitch EDL but retained in slow-twitch soleus, correlating with known Z-disk width differences. The study also detected mutually exclusive splicing of exons 127 and 128, located at the super repeat–Z-disk boundary. Using exPhaser, multiple novel phased isoforms in those regions were also discovered, none of which matched existing annotations in RefSeq or GENCODE (release M10).

For titin (106kb), they compared two titin isoforms *N2-A* (skeletal) versus *N2-B* (cardiac) between three muscles. Cardiac muscle transcripts were missing exons 47 and 167, but included exons 45∗, 46, 168, and 169 aligning with titin isoform *N2-B*. Exon 191 is retained in all skeletal muscle but spliced out in 68% of cardiac transcripts. Exon 312 is 100% included in cardiac but EDL and soleus muscles splice out exon 312 in 25% and 1.2%. Exon 45‡ (Alternative 3’ exon) is expressed more in skeletal muscle (13.9% in EDL, 35.5% in soleus) than in cardiac tissue (3.6%). Exon 11 is exclusive to cardiac muscle, while exons 12 and 13 are co-included in soleus but nearly absent in EDL, suggesting these domains may contribute to Z-disk specialization in slow and cardiac fibers.

Conclusion

While Iso-Seq is traditionally used for transcript isoform discovery rather than quantification, this study demonstrates that full-length reads from consensus sequences, combined with custom analysis and internal priming, enables reliable relative quantification and resolves complex splicing in ultralong transcripts. Limitations include sensitivity to internal oligo(dT) priming—which not all genes support—and the inability to equate isoform proportions with RNA abundance.

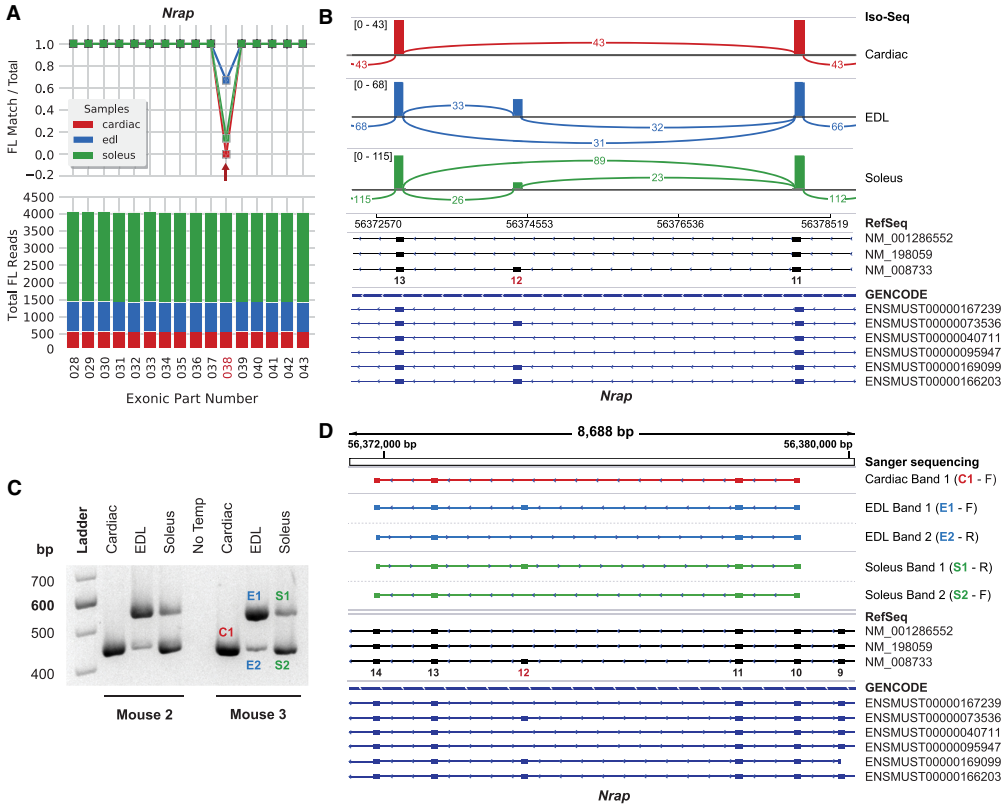
Graphical Abstract

This figure shows the differential usage of nrap exon 12 between cardiac, soleus and EDL, (A) coverage of the splice junctions (B) the expression of exon 12 in the different tissues (C) Agarose gel showing RT-PCR of exon 12 over two replicates (D) Sanger sequencing of the replicates from C.

This figure represents the steps from quality checks over the finding of new information to their verification.

Lu et al. 2008. Expression and alternative splicing of NRAP during mouse skeletal muscle development. Cell Motil Cytoskeleton 65: 945–954. doi:10.1002/cm.20317

Singh et al. 2018. Rbfox-splicing factors maintain skeletal muscle mass by regulating Calpain3 and proteostasis. Cell Rep 24: 197–208. doi:10.1016/j.celrep.2018.06.017


A
Nrap
FL Match / Total
Samples
cardiac
edl
soleus
Total FL Reads
Exonic Part Number
B
Iso-Seq
Cardiac
EDL
Soleus
RefSeq
NM_001286552
NM_198059
NM_008733
GENCODE
ENSMUST00000167239
ENSMUST00000073536
ENSMUST00000040711
ENSMUST00000095947
ENSMUST00000169099
ENSMUST00000166203
Nrap
C
bp
Ladder
Cardiac
EDL
Soleus
No Temp
Mouse 2
Mouse 3
D
8,688 bp
Sanger sequencing
Cardiac Band 1 (C1 - F)
EDL Band 1 (E1 - F)
EDL Band 2 (E2 - R)
Soleus Band 1 (S1 - R)
Soleus Band 2 (S2 - F)
Nrap

Key Citing Publications

Title	Journal	JIF	Citations
A guide for the diagnosis of rare and undiagnosed diseases: beyond the exome	Genome Medicine (2022), Article 23	10.4	316
Beyond the exome: What's next in diagnostic testing for Mendelian conditions	Am. J. Hum. Genet. (2023), Vol. 110(8), pp. 1229–1248	8.1	84
Expectations and blind spots for structural variation detection from long-read assemblies and short-read genome sequencing technologies	Am. J. Hum. Genet. (2021), Vol. 108(5), pp. 919–928	8.1	116
Isoform Age – Splice Isoform Profiling Using Long-Read Technologies	Frontiers in Molecular Biosciences (2021), Vol. 8	3.9	58
Panorama of the distal myopathies	Acta Myologica (2020), Vol. 39(4), pp. 245–265	1.35	57

Corresponding Author

Prech Uapinyoying is the corresponding author of the following publication. Below are five key publications with their full titles.

Title & Journal	Year	Citations	JIF	Author Role	Reasoning
A long-read RNA-seq approach to identify novel transcripts of very large genes Genome Research	2020	51	6.2	First Author	**Core publication** – Prech leads the study and introduces methodological innovation in RNA-seq
MSTO1 mutations cause mtDNA depletion, manifesting as muscular dystrophy with cerebellar involvement Acta Neuropathologica	2019	40	9.3	Middle (17th)	Highly relevant topic: studies genetic causes of muscular disease. Similar to the muscle gene focus (e.g. Ttn, Neb) in RNA-seq paper.
The pediatric cell atlas: defining the growth phase of human development at single-cell resolution Developmental Cell	2019	69	10.7	Middle (43rd)	Applies cutting-edge single-cell technologies — complements long-read RNA-seq in transcriptomic profiling. Builds understanding of gene regulation and alternative splicing.
Toll/Interleukin-1 Receptor Domain-Containing Adapter Inducing Interferon-β Journal of Neuroscience	2012	126	4.4	Fourth	Early, high-impact work — demonstrates molecular biology expertise in neuroinflammatory mechanisms. Reflects strong foundation in signaling and data analysis.
A selective thyroid hormone β receptor agonist enhances human and rodent oligodendrocyte differentiation Glia	2014	81	5.4	Sixth	Solid citation count and biological relevance — involvement in differentiation processes is important for interpreting transcript variation in muscle models.

Key Referenced Studies

Title	Journal	Citations	Authors	Use in This Study
Widespread Polycistronic Transcripts in Fungi Revealed by Single-Molecule mRNA Sequencing	PLoS One (2015), Vol. 10.7, e0132628	393	First: Gordon, Tseng Last/CA: Wang	Early example of long-read sequencing to detect novel splice events
Expression and alternative splicing of N-RAP during mouse skeletal muscle development	Cell Motility and the Cytoskeleton (2008), Vol.65(12), pp. 945–954	34	First: Lu Last/CA: Horowits	Reference for analysis of Nrap expression and splicing
Complete genomic structure of the human nebulin gene and identification of alternatively spliced transcripts	Eur. J. Hum. Genet. (2004), Vol. 12, pp. 744–751	126	First: Donner Last/CA: Pelin	Used for Nebulin gene structure and isoform reference
The complexity of titin splicing pattern in human adult skeletal muscles	Skeletal Muscle (2018), Vol. 8, Article 11	83	First/CA: Savarese Last: Hackman	Reference for Titin splicing complexity and isoform catalog
Repetitive DNA and next-generation sequencing: computational challenges and solutions	Nat. Rev. Genet. (2012), Vol. 13, pp. 36–46	215	First: Treangen Co-First: Salzberg	Justification for long-read methods due to NGS limitations in repetitive regions

PureCLIP: capturing target-specific protein-RNA interaction footprints from single-nucleotide CLIP-seq data

Sabrina Krakau, Hugues Richard and Annalisa Marsico

Krakau et al. Genome Biology, 2017 18:240

Marie Barié und Anna Dieter

Abstract 1

During the last few years, techniques like iCLIP and eCLIP make it possible to study protein–RNA interactions with high resolution. We use specific patterns in the sequencing data, like truncation events, to find where proteins bind to RNA. However, many existing tools don't fully take into account effects like background noise or how often a transcript is present. This can make the results less reliable or lead to too many false positives. A main challenge is that, even with these high-resolution methods, it is still difficult to clearly identify the real binding sites of RNA-binding proteins, especially in noisy or biased data. Here we show PureCLIP, a program based on Hidden Markov Models that tries to improve the detection of crosslink sites by combining different types of signals in the data. Compared to other tools like CITS, CLIPper, or Piranha, PureCLIP seems to do better in many cases. In the paper, it's shown that PureCLIP finds binding sites more precisely, is more reproducible between experiments, and works even when the protein doesn't bind very strongly. Because of that, PureCLIP could be a useful method for future studies that want to look at RNA-binding proteins across the whole transcriptome. It can also be adapted to work with other similar kinds of data. In general, it might help to study such as long non-coding RNAs or weak interactions that are harder to detect. This could give new insights into how RNA and proteins work together inside cells.

Abstract 2

RNA-binding proteins (RBPs) are central to gene regulation, influencing various cellular processes. Understanding where RBPs bind to RNA is crucial for deciphering the complexities of post-transcriptional control. CLIP-seq methods are used to map these protein-RNA interactions at high resolution. However, existing methods don't fully account for biases and truncation paKerns specific to iCLIP and eCLIP data, potentially leading to inaccurate identification of binding sites. This poses a challenge for researchers aiming to precisely define RBP binding landscapes. This study addresses this challenge by developing a computational method that improved the accuracy of CLIP-seq analysis. Here we show that PureCLIP, a hidden Markov model-based approach, simultaneously addresses peak-calling and individual crosslink site detection with increased accuracy. Unlike previous methods, PureCLIP explicitly models non-specific background signals and sequence biases, significantly reducing false positives and improving the precision of crosslink site identification. This advance offers a more reliable means of exploring protein-RNA interaction networks. Accurate identification of RBP binding sites is fundamental to understanding gene regulatory networks and will likely advance our knowledge of many biological processes. The algorithm and model presented can be used in other genomic applications.

Zusammenfassung

The study introduces *PureCLIP*, a computational framework developed to accurately detect protein–RNA interaction sites with single-nucleotide resolution from iCLIP and eCLIP datasets. These high-throughput techniques capture truncation events at protein–RNA crosslink sites, but previous analysis tools often failed to fully account for protocol-specific biases and truncation patterns.

At its core, *PureCLIP* leverages a non-homogeneous Hidden Markov Model (HMM) that integrates two essential signals: the density of pulled-down RNA fragments (derived from smoothed read start counts) and the read start positions themselves, which mark truncation events. The model classifies each nucleotide position into one of four hidden states (combinations of enriched/non-enriched and crosslink/non-crosslink) and identifies sites most likely to represent specific protein–RNA interactions.

A significant advancement of *PureCLIP* lies in its ability to correct for major sources of bias:
• **Background noise** from non-specific crosslinking,
• **Transcript abundance** (using input control experiments),
• **Crosslinking sequence preferences**, modeled through data-driven identification of CL (crosslink-associated) motifs.

These biases are incorporated into the HMM via generalized linear models, allowing for flexible and precise modeling.

The tool was benchmarked extensively against existing methods such as CITS, Piranha, and CLIPper. On both simulated datasets and real-world iCLIP/eCLIP datasets (e.g., for PUM2, RBFOX2, and U2AF2), *PureCLIP* consistently outperformed alternatives in terms of:
• **Precision** in identifying bona fide binding sites,
• **Reproducibility** across experimental replicates (showing up to 20% improvement),
• **Robustness** to parameter changes like bandwidth in kernel density estimation.

Moreover, the integration of input signals and CL motif data significantly improved the precision of binding site detection, especially for low-abundance RNAs or RBPs with weak binding affinity.

PureCLIP is implemented as a command-line tool and is openly available to the research community. It is also adaptable to future extensions, including other CLIP-seq variants (e.g., irCLIP, miCLIP) and additional diagnostic event types beyond truncations.

In conclusion, *PureCLIP* offers a robust, accurate, and bias-aware approach for high-resolution analysis of protein–RNA interactions, making it a valuable tool for transcriptome-wide studies of RNA-binding proteins.

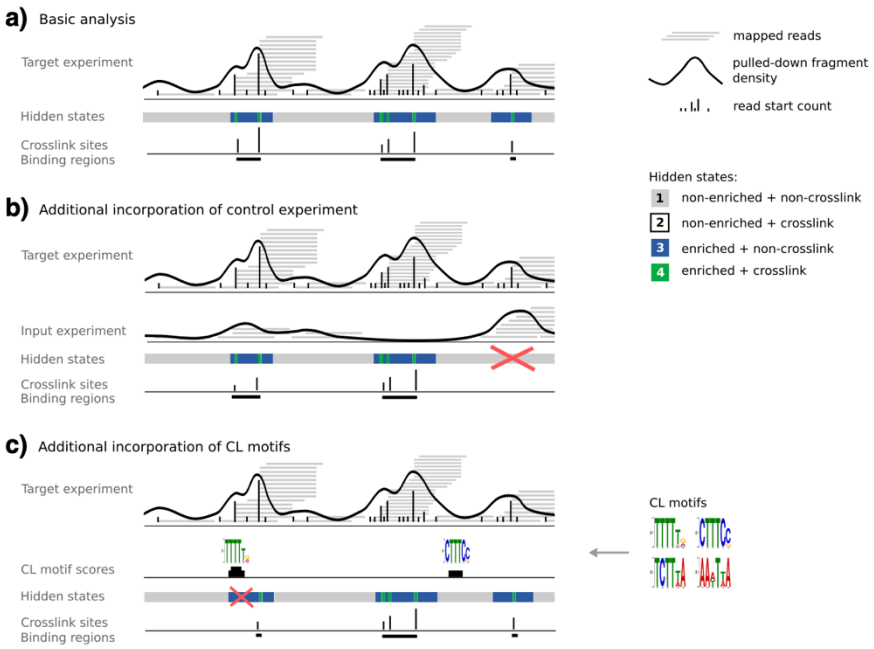


Fig. 1 Overview of the PureCLIP approach. **a** PureCLIP starts with mapped reads from a target iCLIP/eCLIP experiment and derives two signals: the pulled-down fragment density and individual read start counts. Based on these two observed signals, it infers for each position the most likely hidden state. The goal is to identify all sites with an *enriched + crosslinked* state. Individual crosslink sites can then be merged to binding regions. **b** Additionally, information from input control experiments can be incorporated. Its fragment density is used to correct for a non-specific background signal, which reduces the number of false calls. **c** Furthermore, PureCLIP can incorporate information about CL motifs to reduce false calls caused by non-specific crosslinks. CL crosslink-associated

Figure 1 is the key illustration as it summarizes the core methodology of PureCLIP. It shows how read-start counts and fragment density are integrated in a Hidden Markov Model to detect protein–RNA crosslink sites, while also incorporating input controls and sequence motif biases to reduce false positives. This figure visualizes the comprehensive design of the approach.

Paper Impact

Autor(en)	Titel der Arbeit	Journal	Jahr	Zitationen	Begründung
Van Nostrand, E. L., Pratt, G. A., & Shishkin, A. A. et al. (Erstautor, Letztautor unbekannt, korrespondierender Autor unbekannt)	Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP)	Nat Methods.	2016	1435	Describes eCLIP, an enhanced version of iCLIP, which is used in PureCLIP as one of the most important data foundations.
Sugimoto, Y., König, J., & Hussain, S. et al. (Erstautor, Letzt- und korrespondierender Autor unbekannt)	Analysis of CLIP and iCLIP methods for nucleotide-resolution studies of protein-RNA interactions	Genome Biology	2012	269	Compares different CLIP and iCLIP variants and highlights the strengths and weaknesses of each method.
König, J., Zarnack, K., & Rot, G. et al. (Erstautor, Letztautor unbekannt, korrespondierender Autor unbekannt)	iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution	Nat Struct Mol Biol.	2010	1307	Introduces iCLIP, a method that enables single-nucleotide resolution in the mapping of RNA-protein interactions and forms the basis for PureCLIP.
Timothy L. Bailey	DREME: motif discovery in transcription factor ChIP-seq data	Bioinformatics Pages 1653–1659	2011	1231	DREME i spart oft he MEME Suite and is used within the PureCLIP workflow for motif discovery to enhance binding site specificity.
Matthew B Friedersdorf, Jack D Keene	Advancing the functional utility of PAR-CLIP by quantifying background binding to mRNAs and lncRNAs	Genome Biology	2014	136	This study deals with background noise in CLIP data, which is also an important problem that PureCLIP tries to solve

Gesamtzahl der Zitationen: 158

Progressive Cactus is a multiple-genome aligner for the thousand-genome era

Joel Armstrong, Glenn Hickey, Mark Diekhans, Ian T. Fiddes, Adam M. Novak, Alden Deran, Qi Fang, Duo Xie, Shaohong Feng, Josefin Stiller, Diane Genereux, Jeremy Johnson, Voichita Dana Marinescu, Jessica Alföldi, Robert S. Harris, Kerstin Lindblad-Toh, David Haussler, Elinor Karlsson, Erich D. Jarvis, Guojie Zhang & Benedict Paten

Nature | Vol 587, 246-251 | 11 November 2020

Presented by Leonie Bernshausen, Saaruky Chanthirakanthan

Progressive Cactus is a multiple-genome aligner for the thousand-genome era

Joel Armstrong, Glenn Hickey, Mark Diekhans, Ian T. Fiddes, Adam M. Novak, Alden Deran, Qi Fang, Duo Xie, Shaohong Feng, Josefin Stiller, Diane Genereux, Jeremy Johnson, Voichita Dana Marinescu, Jessica Alföldi, Robert S. Harris, Kerstin Lindblad-Toh, David Haussler, Elinor Karlsson, Erich D. Jarvis, Guojie Zhang & Benedict Paten
2020
Nature, Vol. 587, 12 November 2020, Seiten 246–251
https://doi.org/10.1038/s41586-020-2871-y

More cost-effective genome sequencing technologies are leading to a rapid increase in available genomic data, which poses new challenges for alignment tools in terms of their performance and accuracy. What is needed now are tools for precise alignment in large and complex datasets.
However, current alignment tools often perform poorly with increasing numbers of genomes and often experience reference bias. In addition, they lack the ability to represent multiple orthologous regions correctly, especially in regions with duplications or rearrangements.
This raises the question of how multiple-genome alignments can be efficiently and accurately performed for tens to thousands of large vertebrate genones.
Considering this, Progressive Cactus was developed.
It is a scalable, reference-free whole-genome aligner designed to meet the demands of large-scale comparative genomics. It uses a progressive alignment strategy, breaking down the problem into smaller sub-alignments using phylogenetic guide trees. These sub-alignments are then combined step by step, while reconstructed ancestral genomes serve as bridges between related species, allowing for a better representation of evolutionary history.
Advantages over previous aligner tools include higher accuracy in benchmark tests (e.g. Alignathon) and good tolerance to poor assemblies. It is suitable for large-scale comparative studies and supports downstream applications such as gene annotation or variant discovery. It is also suitable for pangenomics and population comparisons.
In the era of high-throughput genomics, the ability to align many genomes precisely and without reference bias is key to understanding genome evolution, function, and diversity.
Furthermore, opens progressive Cactus new perspectives for future studies such as pangenome construction or population-scale genome comparison.

Saaruky Chanthirakanthan

Abstract: **Progressive Cactus – improved and better - a multiple-genome aligner for the thousand-genome era**

Multiple-genome Alignment is a necessary method to find to find differences and similarities within DNA sequences. It is an essential tool to identify structural variations. The number of genome assemblies is growing more and more. Therefore, there is still the need to improve multiple-genome alignments to make it more efficient and accurately, because of the large amount of available data. Here we present progressive Catus, an improved version of the Cactus aligner, which enables a reference free multiple genome alignment with high accuracy, in context of vertebrate genomes. Existing aligners created reference bias or can not scale beyond a limited number of genomes because of computational constrains. Progressive Cactus overcomes these problems, using a progressive alignment strategy with ancestorial reconstruction with a linear runtime scale. This strategy improves the alignment accuracy and flexibility significant. This way, Progressive Cactus enables a new era of comparative genomics. It lets researchers analyze genome evolution and variation across whole clades. This study provides a foundation for better comparisons of vertebrate genomes and other species, which highlights its flexibility.

Progressive Cactus is a multiple-genome aligner for the thousand-genome era

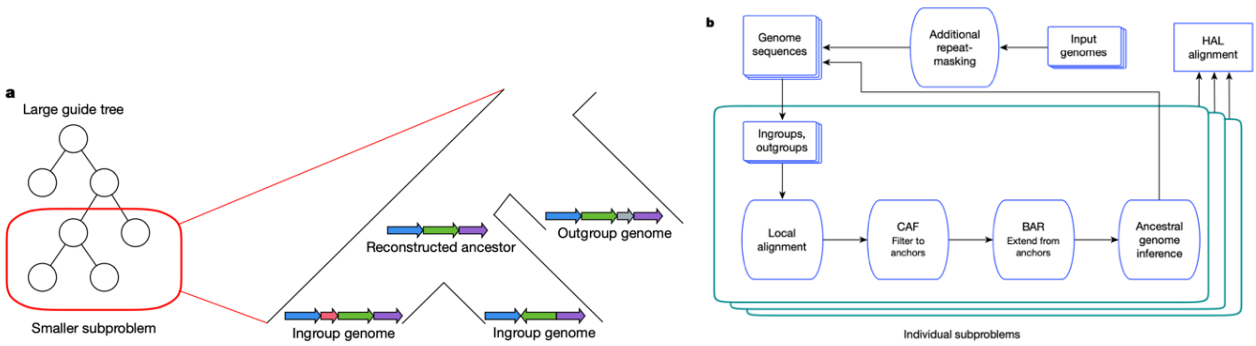
Research Questions:
- How can multiple-genome alignments be performed efficiently and accurately for hundreds to thousands of large vertebrate genomes?

Relevant Methodological Approaches:
- Introduction of Progressive Cactus, an improved version of the Cactus aligner that implements a progressive alignment strategy.
- Use of a guide tree to recursively split a large alignment problem into smaller subproblems.
- Construction of ancestral genome reconstructions at internal nodes of the tree is required for the purpose of combining sub-alignments.
- Utilization of tools such as LASTZ for sensitive pairwise alignment and Toil for distributed computing across clusters or cloud.
- Enable incremental addition and/or removal of genomes without full re-alignment, via HAL toolkit.

Relevant Results:
- Progressive Cactus enables linear runtime scaling with increasing genome numbers, outperforming previous implementations with quadradic scaling.
- Demonstrated the largest known alignment of over 600 vertebrate genomes, confirming feasibility and accuracy at scale.
- Achieved top alignment accuracy of benchmark datasets, with F1 scores of 0.0989 (primates) and 0.795 (mammals).
- Maintained high coverage across evolutionary distances, e.g. 86% of human coding bases retained in ancestral mammal reconstructions.
- Robust to variations in guide tree structure and genome assembly quality, minimizing alignment bias.
- Enabled detection of evolutionary events (e.g. indels, rearrangements) and reconstruction of ancestral genome states.

Conclusion:
- Progressive Cactus addresses the core research challenge by enabling reference-free, scalable, and high-fidelity alignment of large genome sets.
- The approach supports future large-scale comparative genomics and pangenome studies by making genome-wide evolutionary analysis computationally tractable.
- Its flexibility in alignment updates (adding/ removing genomes) is particularly valuable in the rapidly evolving landscape of genomic data.

Key Figure: Figure 1: The alignment process within Progressive Cactus
- Explains how scalability and reference-free alignments are achieved
- Fundamental concept of the method: progressive decomposition with ancestral reconstruction



Paper Impact: Progressive Cactus is a multiple-genome aligner for the thousand-genome era

1. Five relevant References:

1.
   a. Authors: Benedict Paten, Dent Earl, ..., David Haussler
   b. Title: Cactus: algorithms for genome multiple sequence alignment
   c. Journal: Genome research
   d. Year: 2011
   e. Citations: 306
   f. Relevance: Describes the first-generation Cactus aligner.
2.
   a. Authors: Dent Earl, Ngan Nguyen,..., Benedict Paten
   b. Title: Alignathon: a competitive assessment of whole-genome alignment methods
   c. Journal: Genome research
   d. Year: 2014
   e. Citations: 133
   f. Relevance: Describes the problem of different aligners.
3.
   a. Authors: Da-Fei Feng, Russel F. Doolittle
   b. Title: Progressive sequence alignment as a prerequisite to correct phylogenetic trees.
   c. Journal: Journal of molecular evolution
   d. Year: 1987
   e. Citations: 2855
   f. Relevance: Describes the main strategy of Cactus, that changes the runtime (progressive alignment strategy)
4.
   a. Authors: John Vivian, Arjun Akal Rao, ..., Benedict Paten
   b. Title: Toil enables reproducible, open source, big biomedical data analyses
   c. Journal: Nature biotechnology
   d. Year: 2017
   e. Citations: 1197
   f. Relevance: Describes the workflow tool which is used by Cactus
5.
   a. Authors: Erich D. Javaris, Siavash Mirarab,..., Guojie Zhang
   b. Title: Whole-genome analyses resolve early branches in the tree of life of modern birds
   c. Journal: Science
   d. Year: 2014
   e. Citations: 2050
   f. Relevance: Describes why birds are a suitable testcase for guide trees

2.

Number of Citations: 468
Citations per year: ~ 171
5 most influential studies, which cite the paper:

1. Genetic load: genomic estimates and applications in non-model animals. https://doi.org/10.1038/s41576-022-00448-x
2. Towards population-scale long-read sequencing. https://doi.org/10.1038/s41576-021-00367-3
3. The UCSC Genome Browser database: 2023 update, https://doi.org/10.1093/nar/gkac1072
4. Pangenome graph construction from genome alignments with Minigraph-Cactus, https://doi.org/10.1038/s41587-023-01793-w
5. A draft human pangenome reference, https://doi.org/10.1038/s41586-023-05896-x

3. The Corresponding Author: Glenn Hickey

5 most relevant Publications:

1. The UCSC Genome Browser database: 2015 update; cited 1108 times; Journal: Nucleic acids research; Position of Author: equal contribution
2. Progressive Cactus is a multiple-genome aligner for the thousand-genome era; cited 468 times, Journal: Nature; Position of the Author: Corresponding Author
3. Genotyping structural variants in pangenome graphs using the vg toolkit; cited 269 times; Journal: Genome biology; Author: First Author
4. HAL: a hierarchical format for storing and analyzing multiple genome alignments; cited 226 times; Journal: Bioinformatics; Author: First Author
5. Pangenome graph construction from genome alignments with Minigraph-Cactus, cited 197 times; Journal: Nature biotechnology; Author: First Author

Terminating contamination: large-scale search identifies more than 2,000,000 contaminated entries in GenBank

Martin Steinegger & Steven L. Salzberg

Genome Biology volume 21, Article number: 115 (2020)

by Phillip Berger & Timothy Voß

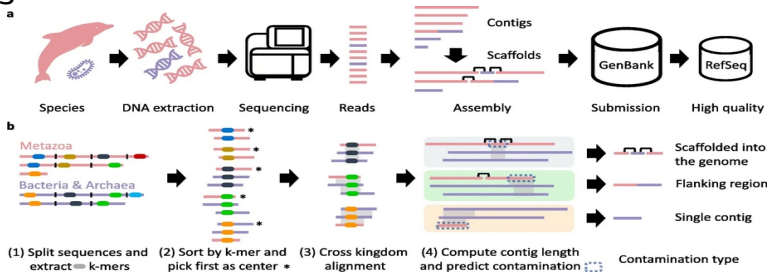
Written by Phillip Berger

Biological sequence databases are growing immensely every year. Finding contamination, i.e. falsely labeled data, is difficult, because of the problem size. **Solving this problem can help alleviate downstream analysis issues. Databases like GenBank from NCBI do have contamination filters. But exactly how much contamination is still present in big sequence databases like RefSeq, GenBank, and NR protein is not known. Here we describe Conterminator, an efficient method of detecting contamination in sequence databases. The analysis using Conterminator shows that the contaminated sequences of RefSeq, GenBank, and NR protein amount to 2,161,746, 114,035, and 14,148 respectively and that the search can be done efficiently. The results show the possibility of searching for contamination even in big databases. The preliminary filters of the databases is not enough to filter out all contamination. The contamination found here is probably not all contamination in these databases. New contamination is added every day. In the future methods like the one shown here could be used to regularly check sequence databases for contamination to increase the quality of data. The increasing size of sequence databases is not an insurmountable problem.**

Public sequence repositories have grown rapidly, but the extent of mislabeled or contaminant sequences remains largely unknown. Existing quality controls like VecScreen and BLAST against known vectors can detect common synthetic contaminants, but they do not manage to find cross-kingdom mislabeling. Without modern approaches, contaminated sequence entries can remain undetected, potentially misleading future analyses in genomic research. In this paper, we aim to find out how much are the nucleotide and protein sequence databases contaminated by incorrect sequence labeling. Here we show that, in the sequence databases GenBank and RefSeq we detected over 2.16 million and 114000 contaminated nucleotide entries. The strongest source for contamination we found out to be human nucleotide sequences. Also we found 14148 protein contaminants in NR. This result shows that new methods for detecting cross kingdom contamination in nucleotide and protein databases are needed. Additionally the findings show a fundamental challenge in modern biology, as vast amounts of data are generated, ensuring its accuracy becomes a critical challenge. Contamination in sequences can not only affect genomic studies but can also mislead research in other fields like medicine, ecology and biotechnology. Therefore cooperation between different fields is needed to develop and implement robust strategies and standards for contamination detection, to safeguard the validity of scientific research.

Summary by Phillip Berger and Timothy Voß

The paper Terminating contamination: large-scale search identifies more then 2,000,000 contaminated entries in GenBank, from Martin Steinegger and Steven L. Salzberg is about the contamination of public sequence databanks with wrongly labled sequences. Despite the explosive growth of public sequence repositories, e.g. GenBank doubling roughly every 18 months, the extent to which these resources are compromised by mislabeled sequences remains poorly quantified. Existing quality controls (e.g., VecScreen, BLAST against known vectors) are used to remove common synthetic or well known contaminants . This raises the central question: How much are the nucleotide and protein sequence databases contaminated by incorrect sequence labeling?



a *How contamination occurs and* ***b*** *how Conterminator detects it.*

To address this, the authors developed Conterminator, a two-stage pipeline combining rapid prefiltering and exhaustive alignment. First, Linclust-based k-mer clustering and ungapped alignments quickly highlight candidate fragments (≥100 nt at ≥90 % identity) that may originate from a different kingdom. Second, MMseqs2 performs sensitive, full alignments of these candidates against a reference set spanning multiple kingdoms. To avoid conflating genuine horizontal gene transfer with contamination, only short contigs (<20 kb) mapping to long reference contigs (>20 kb) trigger contamination flags. This design achieves near-linear runtime, processing over 3 TB of GenBank entries in 12 days on a 32-core server.

Applying Conterminator to GenBank and RefSeq identified the following contamination: 2.16 million nucleotide entries (0.54 %) in GenBank and 114 000 entries (0.34 %) in RefSeq, along with 14 148 proteins in the non-redundant (NR) database. Even high-quality assemblies, including the human and *C. elegans* reference genomes, contained bacterial inserts, such as an 18 kb *Acidithiobacillus thiooxidans* fragment in an alternative human scaffold of chromosome 10 and a 4 kb *E. Coli* segment in C**.** *elegans*.

Conterminator demonstrates that a non-negligible fraction of nucleotide and protein entries in major public databases are mislabeled. By combining speed with sensitivity and guarding against misclassifying true horizontal gene transfers, it provides a practical framework for systematic quality control. The findings not only quantify the problem but also show that even flagship reference genomes may harbor undocumented contaminants. Using Conterminator regular re-screening of growing databases can flag new intrusions. Ultimately, automated reporting and removal of suspect fragments can ensure that public sequence repositories remain reliable foundations for genomics research.

References by Phillip Berger and Timothy Voß

Five most relevant References

1.

- Martin Steinegger (first)
- Johannes Söding (corresponding, last)
- MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets
- Nature Biotechnology 35, 1026-1028
- 2017
- 1973 citations (by CrossRef)
- MMseqs2 is used and is one of the most important factors that enables the efficiency of Conterminator. Used for fast alignment.

2.

- Martin Steinegger (first, corresponding)
- Johannes Söding (last, corresponding)
- Clustering huge protein sequence sets in linear time
- Nature Communications 9, Article number: 2542
- 2018
- 856 citations (Google Scholar);
- Reduces computational cost of comparisons from quadratic (naive all-against-all) to a linear number of comparisons.

3.

- Eric W Sayers (first, corresponding)
- Ilene Karsch-Mizrachi (last)
- GenBank
- Nucleic Acids Research, Volume 47, Issue D1, Pages D94-D99
- 2019
- 441 citations (by Dimensions)
- Is important because GenBank is the biggest sequence database, of the three, that are searched for contamination in this study.

4.

- Christiam Camacho (first)
- Thomas L Madden (corresponding, last)
- BLAST+: architecture and applications
- BMC Bioinformatics, volume 10, article number 421
- 2009
- 19992 citations (Google Scholar)
- Is the method the databases use to check for contamination. This is the base the new method from the paper is compared to.

5.

- Florian P. Breitwieser (first, corresponding)
- Steven L. Salzberg (last, corresponding)
- Human contamination in bacterial genomes has created thousands of spurious proteins
- Genome Res. 29, 954 – 960
- 2019
- One of the examples of proven contamination in databases. Underlines the reason for the research in this paper.

Paper citations

Citations by year	Google Scholar	PubMed
2020	12	6
2021	44	22
2022	41	19
2023	45	24
2024	43	20
2025	32	17
All time	217	108

Five most relevant papers that cite the paper

1.

- The complete sequence of a human Y chromosome
- Nature (2023) (IF 50.5), 311 citations

2.

- A complete reference genome improves analysis for human genetic variation
- Science (2022) (IF 44.7), 341 citations

3.

- Clustering predicted structures at the scale of the known protein universe
- Nature (2023) (IF 50.5), 234 citations

4.

- Metagenome analysis using the Kraken software suite
- Nature protocols (2022) (IF 13.1), 428 citations

5.

- ContScout: sensitive detection and removal of contamination from annotated genomes
- Nature communications (2024) (IF 14.7), 10 citations

Corresponding author Martin Steinegger

- H-Index: 32
- i10-Index: 54
- Citations all time: 59,131

The five most relevant publications:

1.

- Highly accurate protein structure prediction with AlphaFold
- Although M. Steinegger is not the first, last or corresponding author of this paper, it is still a very relevant paper, because it is the foundational AlphaFold paper. It has 36k citations at the moment and was published in Nature. This paper is more than half of all citations of M. Steinegger.

2.

- ColabFold: making protein folding accessible to all
- His second most cited paper with 7k citations. Published in Nature Methods (IF 36.1) and he is the last and corresponding author.

3.

- MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets.
- Has 3k citations and was published in Nature biotechnology (IF 33.1) with him as the first author.

4.

- Fast and accurate protein structure search with Foldseek
- Was also published in Nature biotechnology (IF 33.1) with him as the corresponding and last author. It has 1509 citations.

5.

- Clustering huge protein sequence sets in linear time
- This paper was published in Nature communications (IF 14.7) with M. Steinegger as its first and corresponding author. It was cited 857 times.

Clustering predicted structures at the scale of the known protein universe

Nature volume 622, pages 637–645 (2023)

Inigo Barrio-Hernandez, Jingi Yeo, Jürgen Jänes, Milot Mirdita, Cameron L. M. Gilchrist, Tanita Wein, Mihaly Varadi, Sameer Velankar, Pedro Beltrao & Martin Steinegger

Presenting persons:

Haoxuan Zeng
Mehmet Batman

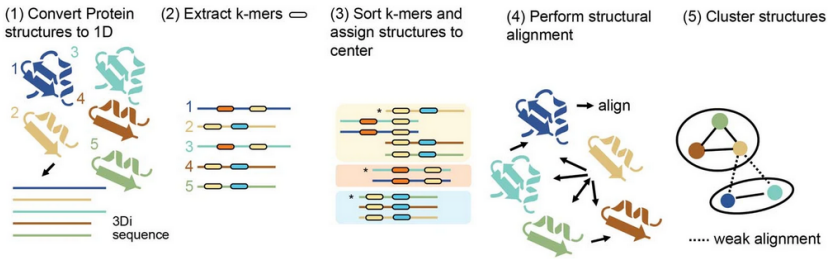
Summary of the article **Clustering predicted structures at the scale of the known protein universe**

The aim of this scientific work was to develop a computational approach to compare and align more than 214 million predicted protein structures from AlphaFold's Protein Structure Database (AFDB) by their structure and group similar proteins into clusters to get **better insights into proteins evolution and function** at an improved speed.

At the core of all methods the **structural clustering algortihm** using Foldseek has been applied.
By appliying this method, structural information of proteins will be translated into Foldseeks internal sequence called 3Di, which describes the structure by its own alphabet. These 3Di sequences will then be used in alignments for detection of protein stucture similarities resulting in clusters.
In the course of this, Foldseek makes use of the existing clustering algorithms MMseqs2 and Linclust.
The following figure illustrates the five steps of the structural clustering algortihm in more detail:

Extended Data Fig. 1: The five-step clustering pipeline for efficiently clustering millions of protein structures using Foldseek's 3Di alphabet.

From: Clustering predicted structures at the scale of the known protein universe



(1) Protein structures are converted to 3Di sequences and processed through the Linclust workflow. (2) For each sequence, 300 min-hasing k-mers are extracted and sorted. (3) The longest structure is assigned to be the centre of each k-mer cluster. (4) Structural alignment is performed in two stages: first an ungapped alignment based on shared diagonal information is performed, hits are pre-clustered and second the remaining sequences are aligned using Foldseek's structural Smith-Waterman. (5) The remaining structures meeting alignment criteria are clustered using MMseqs2's clustering module. After the Linclust step the centroids are successively clustered by three cascaded steps of prefiltering, structural Smith-Waterman alignment and clustering using Foldseek's search.

As result of the Foldseek processing non-singleton clusters (with at least two structures) ended up in a number of ~2,3 million clusters, the remaining singleton clusters were about ~13 million. The top 3 most often predicted molecular functions are related to "transporter activity" in clusters lacking annotations.
From taxonomic point of view the mapping of cluster-members in the tree of life provided information with following distribution: Cellular organism (23%), bacterial (16.1%), Eukaryota (13.5%) and Archaea (0.5%). Human-related cluster analysis did not provide any indication of the emergence of new human-specific structural clusters. One more important finding is that human immunity related proteins are present in clusters which have representatives in bacterial species (e.g. the CD4 like protein B4E1T0 and bacterial protein A0A1F4ZDN5). Two domain families with structural similarity to gasdermin domain could be identified.

Limiting factors will be mentioned regarding domain prediction as only representatives of FoldSeek clusters have been taken into account. It is pointed out that multiple observations on protein regions and larger set of structures will be required for that. Stuctural clustering in general can be inaccurate due to (1) 90% alignment overlap requirement, (2) strict E-value theshold of 0.01 and (3) incompleteness of current AFDB.

As an outlook the authors see great potential that AFDB can help identifying remote homology due to the fact that protein structures have longer conservation periods compared to protein sequences. In future it is also to be expected that further findings will be detected from the cluster analysis as observed in the example of the CD4 like protein and its functional acquistion from bacterial protein.

Abstract of the article **Clustering predicted structures at the scale of the known protein universe**

Proteins are essential for life and in order to better understand cell processes, it is important to have knowledge of proteins structures as this can provide important insights into their function and evolution.

In the past decades thousands of protein strcutures have been experimentally determined by biologists. Since the introduction of AlphaFold in 2021 it is possible to computationally predict highly accurate three-dimensional protein structures from protein sequences. The AlphaFold database contains nowadays over 214 million predicted protein structures.

In turn the research and analysis of such a big number of protein structures requires a comprehensive and efficient method.

For the purpose of this we developed a structural clustering algorithm named FoldSeek which can compare hundred millions of protein structures at once and group them into clusters based on their similarity at an improved speed.

We identified ~2,3 million clusters consisting of non-singleton structures of which 31% have no similarity to previously known structures. Distribution after mapping of cluster-members in the tree of life: Cellular organism (23%), bacterial (16.1%), Eukaryota (13.5%) and Archaea (0.5%).

“Transporter activity“ detected as the most widespread protein function in clusters without annotations. Human immunity-related proteins with structural similarty in bacterial proteins have been found.

Assuming that the number of records in publicly accessible protein strcuture databases is constantly increasing a highly scalable structure-based clustering algorithm is needed to be able to explore billions of structures in future.

Clustering of known and unknown protein structures may help scientists studying proteins function and evolution more precisely and can pave the way for the development of new drugs.

Abstract

Proteins have a wide range of crucial functions in the body, including structural support, acting as enzymes and hormones, and transporting molecules. They also play a vital role in the immune system, fluid balance, and energy production. In recent years, the explosive growth in protein structure predictions, especially via AlphaFold, has led to the availability of over 214 million predicted protein structures. To make sense of this vast structural space, this study developed a structure-based clustering approach called Foldseek cluster to identify millions of proteins by structural similarity. Using this method, they clustered the entire AlphaFold database into 2.30 million non-singleton structural clusters, 31% of which lack known annotations.

This study highlights that while most clusters are conserved and likely ancient, 4% of them appears species-specific, indicating possible de novo gene birth or prediction artifacts. This study also inferred that putative domain families play a important role in evolutionary links across species. Furthermore, remote structural homologies were identified between human immunity proteins and bacterial counterparts, suggesting ancient shared mechanisms. Additionally, thousands of unannotated clusters with confident structural predictions were discovered, some of which may function as enzymes or membrane-bound transporters.

This work provides a crucial framework for exploring protein structure space at scale and deepens our understanding of evolutionary relationships, immune system evolution, and functional diversity across life.

1. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, Steinegger M, Pacholska M, Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior AW, Kavukcuoglu K, Kohli P, Hassabis D. Highly accurate protein structure prediction with AlphaFold. Nature. 2021 Aug;596(7873):583-589. doi: 10.1038/s41586-021-03819-2. Epub 2021 Jul 15. PMID: 34265844; PMCID: PMC8371605.
Citation: 35949

2. Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, Smetanin N, Verkuil R, Kabeli O, Shmueli Y, Dos Santos Costa A, Fazel-Zarandi M, Sercu T, Candido S, Rives A. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science. 2023 Mar 17;379(6637):1123-1130. doi: 10.1126/science.ade2574. Epub 2023 Mar 16. PMID: 36927031.
Citation: 3089

3. van Kempen M, Kim SS, Tumescheit C, Mirdita M, Lee J, Gilchrist CLM, Söding J, Steinegger M. Fast and accurate protein structure search with Foldseek. Nat Biotechnol. 2024 Feb;42(2):243-246. doi: 10.1038/s41587-023-01773-0. Epub 2023 May 8. PMID: 37156916; PMCID: PMC10869269.
Citation: 1286

4. Steinegger M, Söding J. Clustering huge protein sequence sets in linear time. Nat Commun. 2018 Jun 29;9(1):2542. doi: 10.1038/s41467-018-04964-5. PMID: 29959318; PMCID: PMC6026198.
Citation: 853

5. Abramson, J., Adler, J., Dunger, J. et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature 630, 493–500 (2024). https://doi.org/10.1038/s41586-024-07487-w
Citation: 2538

These references are closely related to the computational tools used, the mechanisms of protein structure prediction, and evolutionary aspects. Some were authored by the corresponding authors themselves, while others are highly relevant to the central topics of the study.

Perez-Lopez, R., Ghaffari Laleh, N., Mahmood, F. et al. A guide to artificial intelligence for cancer researchers. Nat Rev Cancer 24, 427–441 (2024). IF72.5

AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences. M Varadi, D Bertoni, P Magana, U Paramval... - Nucleic acids ..., 2024 -academic.oup.com. IF 16.6

Prediction-powered inference. ANASTASIOS N. ANGELOPOULOS HTTPS://ORCID.ORG/0000-0001-9787-0579, SCIENCE. DOI: 10.1126/science.adi600. IF 44.7

Durairaj, J., Waterhouse, A.M., Mets, T. et al. Uncovering new families and folds in the natural protein universe. Nature 622, 646–653 (2023). https://doi.org/10.1038/s41586-023-06622-3. IF 50.5

Bilingual language model for protein sequence and structure. Michael Heinzinger. NAR Genom Bioinform.IF 4.07

2025	27
2024	60
2023	6

Beispielhafte relevante Publikationen von Martin Steinegger:

MMseqs2 (Nat Commun, 2018), Foldseek (Nat Biotechnol, 2023), DeepUnfold (Bioinformatics, 2022), AlphaFold DB Expansion (NAR, 2022), HH-suite3 (Bioinformatics, 2019)

Fast and sensitive taxonomic assignment to metagenomic contigs

M Mirdita, M Steinegger, F Breitwieser, J Söding, E Levy Karin

Bioinformatics 37(18): 3029–3031 (2021)

Jiamei Qin, Daulet Ashirov

The research question of this paper is: State-of-the-art tools for taxonomic annotation of metagenomic contigs have limitations, highlighting the need for a faster, more broadly applicable, efficient, and automatic method.

The key method: It translates the nucleotide reads to protein fragments and retains those with the highest frequency of occurrence of similar k-mer matches. The query sequence is then searched against a reference database to quickly identify the hit with the lowest E-value. The aligned region of the best hit is reused with a slower mode to identify a list of homologs for the read whose E-values are smaller than that of the best hit. The taxonomic classifications of the homologs are simplified to their LCA (lowest common ancestor), which is assigned as the read's taxonomic annotation. Weighted votes from all assigned fragments are used to determine taxonomic classifications by selecting the most specific taxonomic label with > 50% support of thei total weights (Fig.1.).

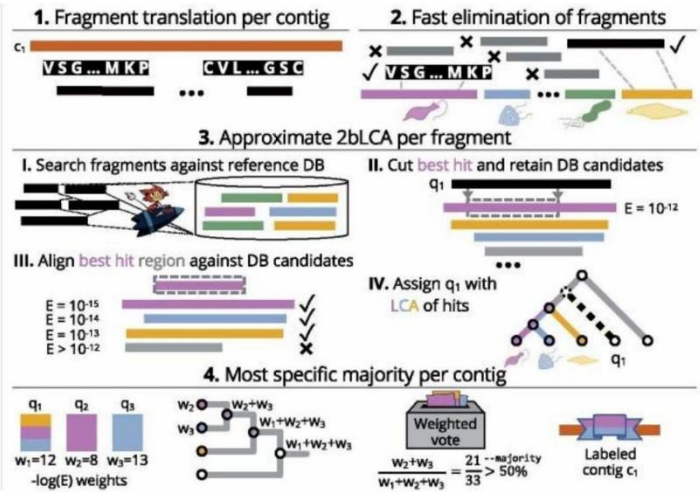


Fig.1. The procedure of the taxonomy assignment algorithm.

The key result: The new tool is faster, more automatic, and similar accurate as the previous tool on bacterial and eukaryotic datasets.

Knowledge gap: Existing tool for taxonomic annotation of contigs has drawbacks: unsuitable for eukaryotic reads, limited by single-threaded performance in metagenomic applications, and require manual selection of a key parameter.

Conclusion: This new tool for taxonomic annotation of metagenomic contigs overcomes the limitations, performs the procedure faster without reducing accuracy across all domains of life, supporting multithreading, eliminating manual parameter selection, thus representing an advance over the previous tool.

Outlook: The original paper does not mention outlook.

14.1

Abstract for Fast and sensitive taxonomic assignment to metagenomic contigs

Metagenomic studies have revolutionized our understanding of microbial diversity, yet accurately assigning taxonomic labels to assembled contigs remains a computational challenge.Current tools like CAT use prokaryote designed gene prediction (Prodigal) and require manual parameter tuning, limiting their performance on eukaryotic contigs and scalability. While short read classification is well established, contig level assignment should improve accuracy but demands faster, domain covering methods.To address these limitations, the authors developed MMseqs2 taxonomy, a novel protein-search-based tool that overcomes prokaryotic bias and automation challenges. The key question was whether a more comprehensive approach could achieve faster, more accurate taxonomic assignment across all domains of life.MMseqs2 taxonomy extracts all possible protein fragments from contigs (suitable for all domains of life), retains fragments with database similarity, and implements an approximate 2bLCA strategy for taxonomic assignment. Benchmarking showed it classifies bacterial contigs 18× faster than CAT with equal accuracy (CAMI-I dataset). For eukaryotic SAR contigs, it achieved 62% assignment (vs CAT's 47%) with 46% species-level precision (vs 28%). The tool automatically determines search parameters, eliminating manual tuning.This work presents MMseqs2 taxonomy as a fast, sensitive solution for metagenomic contig annotation that outperforms existing methods in both bacterial and eukaryotic datasets. Its open source implementation, compatibility with standard databases (UniProt, GTDB…), and included utility modules make it a versatile tool for advancing metagenomic research across the tree of life.

14.2

Metagenomics enables the identification and analysis of uncultivated microorganisms in natural environments, leading to the discovery of novel microbial species. This is of great significance for understanding the functional potential of microbial ecosystems. Metagenomic sequencing of environmental samples generates large numbers of contigs derived from complex microbial communities. However, simply assembling these contigs does not reveal which organisms they originate from. Accurate taxonomic annotation of contigs is therefore essential to assign fragments to specific taxa. Yet, the state-of-the-art tools for taxonomic annotation of metagenomic contigs have limitations. Here the authors show that the new tool, MMseqs2, is faster, more automatic, and similarly accurate to the previous tool on bacterial and eukaryotic datasets. They found that this new tool is suitable for eukaryotic reads, multithreaded performance in metagenomic applications, and does not require manual selection of a key parameter. Moreover, it is ~ 18x faster on a bacterial dataset and ~ 4x faster than existing tools on a eukaryotic dataset. MMseqs2 supports microbial ecology, metagenomics, and epidemiology by making it easier to analyze microbial diversity and taxonomy. For example, MMseqs2 promotes the discovery of new microbial species and their functions. It can detect pathogenic microorganisms in the human intestinal tract, advancing the understanding of microbial diversity and enabling the development of new therapies. Moreover, it identifies antibiotic- and antifungal-producing microbes in soil, which could lead to the development of novel treatments.

Question 1:

Paper 01 **Authors**: first author: Patrick T. West, corresponding author: Jillian F. Banfield, and last author: Jillian F. Banfield.**Title of the work:** Genome-reconstruction for eukaryotes from complex natural microbial communities. **Journal**: Genome research, 28(4), 569-580. **Year of publication**: 2018. **Number of citations**: 233. **Brief justification**: This paper provides a tool for the comparison of performance of the 14th paper that was read.

Paper 02 **Authors**: first author: Pascal Hingamp, corresponding author: Hiroyuki Ogata, and last author: Hiroyuki Ogata. **Title of the work:** Exploring nucleo-cytoplasmic large DNA viruses in Tara Oceans microbial metagenomes. **Journal**: The ISME journal, 7(9), 1678–1695. **Year of publication:** 2013. **Number of citations:** 247. **Brief justification:** Because the assignment of taxonomic nodes is processed using the method introducing by this paper.

Paper 03 **Authors**: first author: Alexander Sczyrba, corresponding author: Alexander Sczyrba and Alice C McHardy, and last author: Alice C McHardy. **Title of the work:** Critical assessment of metagenome interpretation—a benchmark of metagenomics software. **Journal**: Nat. Methods, 14, 1063–1071. **Year of publication:** 2017. **Number of citations:** 840. **Brief justification:** providing reference database.

Paper 04 **Authors**: first author: Fernando Meyer, corresponding author: Alice C McHardy, and last author: Alice C McHardy. **Title of the work:** AMBER: Assessment of Metagenome BinnERs. **Journal** (issue and page numbers): Gigascience 7 (6): giy069. **Year of publication:** 2018. **Number of citations:** 99. **Brief justification:** was used to assess the taxonomic assignment.

Paper 05 **Authors**: first author: Quentin Carradec, corresponding author: Eric Pelletier, Chris Bowler and Patrick Wincker, and last author: Patrick Wincker. **Title of the work:** A global ocean atlas of eukaryotic genes. **Journal**: Nat. Commun., 9, 373. **Year of publication:** 2018. **Number of citations:** 373. **Brief justification:** the 14th paper followed the length distribution of contigs assembled for a sample in this paper.

Question 2: Total number of citations of the 14th paper: 219. The citations per year: 73. The five studies that are the most influential among those citing my paper:

Pavlopoulos, G. A., Baltoumas, F. A., Liu, S., Selvitopi, O., Camargo, A. P., Nayfach, S., ... & Kyrpides, N. C. (2023). Unraveling the functional dark matter through global metagenomics. Nature, 622(7983), 594-602 (116 citations, IF 64.8).

Barrio-Hernandez, I., Yeo, J., Jänes, J., Mirdita, M., Gilchrist, C. L., Wein, T., ... & Steinegger, M. (2023). Clustering predicted structures at the scale of the known protein universe. Nature, 622(7983), 637-645 (234 citations, IF 64.8).

Santos-Júnior, C. D., Torres, M. D., Duan, Y., Del Río, Á. R., Schmidt, T. S., Chong, H., ... & Coelho, L. P. (2024). Discovery of antimicrobial peptides in the global microbiome with machine learning. Cell, 187(14), 3761-3778 (131 citations, IF 64.5).

Klapper, M., Hübner, A., Ibrahim, A., Wasmuth, I., Borry, M., Haensch, V. G., ... & Stallforth, P. (2023). Natural products from reconstructed bacterial genomes of the Middle and Upper Paleolithic. Science, 380(6645), 619-624 (44 citations, IF 56.9).

Madani, A., Krause, B., Greene, E. R., Subramanian, S., Mohr, B. P., Holton, J. M., ... & Naik, N. (2023). Large language models generate functional protein sequences across diverse families. Nature biotechnology, 41(8), 1099-1106 (859 citations, IF 46.9).

Question 3: Corresponding authors: J Söding (Johannes Söding) and E Levy Karin

1. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Journal: *Molecular Systems Biology*. IF 9.9 (not very high. However, considering the paper was published in 2011, a decrease in impact factor over time is expected). 17065 Citations (the paper is highly citated, which can make us ignore the low value of IF). Not the first author, however, it was stated that all listed authors contributed equally to this work.

2. The HHpred interactive server for protein homology detection and structure prediction. Journal: *Nucleic Acids Research*. IF 14.9. 4198 Citations (published in 2005. Despite the low value of IF, the paper is highly citated). First author and corresponding author.

3. MMseqs2 enables sensitive protein sequence searching for analysis of massive data sets. Journal: *Nature Biotechnology*. IF 46.9. 3356 Citations (highly citated). Last/corresponding author, the last author is usually the man, who conceives the paper.

4. Protein homology detection by HMM–HMM comparison. Journal: *Bioinformatics*. IF 5.8. 2988 Citations (published in 2005. Despite the low value of IF, the paper is highly citated). The only author of the paper.

5. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Journal: *Nature Methods*. IF 48. 2512 Citations (highly cited). Last/corresponding author.

Accurate proteome-wide missense variant effect prediction with AlphaMissense

Jun Cheng, Guido Novati, Joshua Pan, Clare Bycroft, et al.

Science, 2023, Vol. 381, Issue 6664, 11 pages

Yakun Li & Fatih Sahin

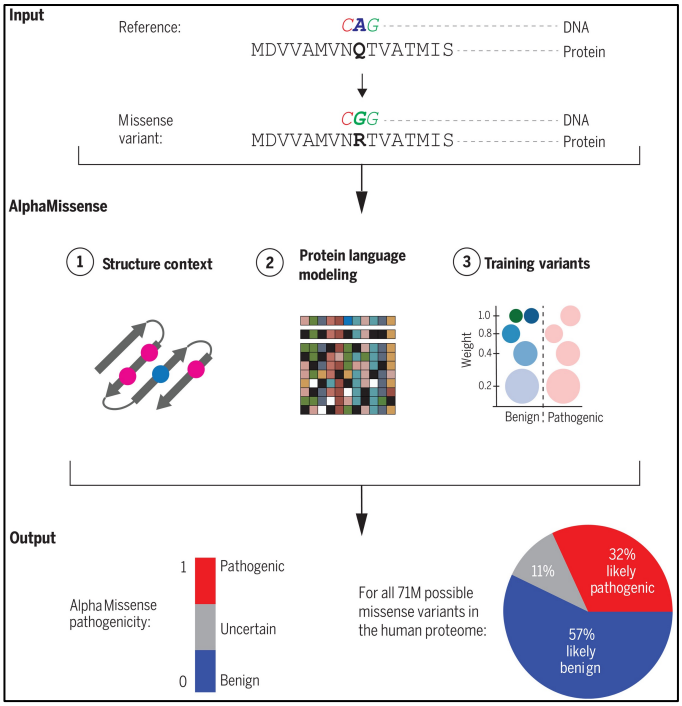
1. Abstracts

a) A key challenge is identifying the functional impact of the large number of missense variants found in human genome sequencing. Most of these variants are still clinically unclassified. This complicates rare disease diagnostics and limits progress in personalized medicine. To show this, the study introduces AlphaMissense, a machine learning tool based on AlphaFold, finetuned using human and primate variant population frequencies. The model predicts pathogenicity for all possible single amino acid substitutions across the human proteome. AlphaMissense integrates structural context and evolutionary conservation to assign a pathogenicity score for each variant. Here we show that it performs well on diverse genetic and experimental benchmarks like ClinVar and MAVE datasets, This indicates strong generalization and helps reduce bias from known variant databases. A notable finding is that the average pathogenicity score per gene correlates with gene essentiality, particularly benefiting short essential genes. The study fills an important knowledge gap in variant interpretation, especially for regions which are hard to classify. Most of the possible human missense variants are categorized as benign (57%) or likely pathogenic (32%), based on thresholds set for 90% accuracy using the ClinVar dataset. This accessible resource is anticipated to accelerate research in molecular biology and human genetics, support clinicians in disease diagnosis, and offer valuable insights into complex human traits.

b) Missense variants change single amino acids in proteins. Many cause genetic diseases. Over 4 million human missense variants exist, but 98% lack clinical classification (VUS). Current tools like PolyPhen-2 rely heavily on limited data. This gap delays rare disease diagnosis and therapy development. Here we introduce AlphaMissense, a new prediction model. We show this model classifies 89% of 71 million variants across 19,233 human proteins as pathogenic or benign with 90% precision. It outperforms tools like EVE by 3.2% (AUC-ROC 0.940 vs. 0.911). Our method combines protein evolution patterns and human population data, avoiding biased clinical annotations. It works well even for challenging proteins like transmembrane domains. These results resolve key limits in VUS interpretation. All predictions are freely available online. They provide immediate help for diagnosing rare diseases. This framework also offers new tools for developing precision treatments.

2. Summary

Missense variants alter a single amino acid in proteins and play a critical role in genetic diseases. Over 4 million such variants exist in humans, yet 98% lack clear clinical classification (often termed “variants of unknown significance” or VUS), hindering rare disease diagnosis. AlphaMissense aims to resolve this by combining evolutionary patterns and population data in a two-step machine learning approach. First, it uses AlphaFold-based structural pretraining to analyze evolutionary relationships in protein sequences. Second, it refines predictions using human and primate population frequencies, assuming common variants are benign, and rare ones are disease-linked. Unlike traditional tools that predict structural changes, AlphaMissense assigns a simple pathogenicity score, since harmful variants often occur in stable protein regions. It also filters out benign variants to remove noise during training. The model provides predictions for ~71 million missense variants across 19,233 human proteins, achieving about 90% precision. It outperforms top methods like EVE and REVEL (AUC-ROC 0.94 vs 0.91), with especially strong performance in challenging regions like transmembrane domains. Predictions are publicly available through the Ensembl Variant Effect Predictor, offering a practical tool to improve rare disease diagnosis and precision medicine.



AlphaMissense workflow overview. The model processes a missense variant using structural context, protein language modeling, and population data, outputting a pathogenicity score that classifies each variant as likely benign, likely pathogenic, or uncertain.

AlphaMissense was evaluated on multiple independent datasets and demonstrated robust results consistently. The study underscores how the predictions correspond to biologically meaningful patterns and clinically significant distinctions. In order to ensure that its predictions were clinically interpretable, a threshold was defined corresponding to 90% precision. Based on this, AlphaMissense classified 57% of variants as likely benign, 32% as likely pathogenic and 11% as uncertain. The predictions aligned with ACMG clinical classification guidelines. Over 90% of the variants marked as likely pathogenic by AlphaMissense were in line with existing clinical interpretations.

This supports the model's practical usefulness, especially when dealing with variants of uncertain significance. One of the key contributions of the study is the scale of the resource. Unlike previous tools that only cover a limited subset of variants, AlphaMissense provides predictions for over 71 million missense variants across the human proteome. This makes it a valuable tool not just for variant prioritization in diagnostics, but also for identifying new disease-associated genes and guiding further research in clinical genomics.

3. Paper Impact

a) 5 relevantesten Referenzen

Authors:	Konrad J. Karczewski (First author), Laurent C. Francioli et al.
Title:	The mutational constraint spectrum quantified from variation in 141,456 humans
Journal:	Nature, pp. 434-455
Year:	2020
Number of citations:	8834 (from Google Scholar, June 2025)
Justification:	Introduction to LOEUF (Loss-of-function Observed/Expected Upper bound Fraction) Metric

Authors:	Laksshman Sundaram (First author), Hong Gao et al.
Title:	Protein fitness prediction with autoregressive transformers and inference-time retrieval
Journal:	Preprint on arXiv, 19 pages
Year:	2022
Number of citations:	239 (from Google Scholar, June 2025)
Justification:	Introduction to Tranception – Proteinfitness prediction

Authors:	Pascal Niotin (First author), Mafalda Dias et al.
Title:	Predicting the clinical impact of human mutation with deep neural networks
Journal:	Nature, pp. 1161-1170
Year:	2018
Number of citations:	478 (from Google Scholar, June 2025)
Justification:	Introduction to PrimateAI, a reference model compared by AlphaMissense

Authors:	John Jumper (First author), Richard Evans et al.
Title:	Highly accurate protein structure prediction with AlphaFold
Journal:	Nature, pp. 583-592
Year:	2021
Number of citations:	36153 (from Google Scholar, June 2025)
Justification:	Introduction to AlphaFold, the foundation of AlphaMissense

Authors:	Jonathan Frazer (First author), Pascal Niotin et al.
Title:	Disease variant prediction with deep generative models of evolutionary data.
Journal:	Nature, pp. 91-108
Year:	2021
Number of citations:	654 (from Google Scholar, June 2025)
Justification:	Introduction to EVE (Evolutionary model of Variant Effect)

b) Citataions:

1. Total Citations of das Paper: 1088 (from Google Scholar, June 2025)
2. Citations per Year: 620 (2025: 437 citations, 2024: 593 citations)
3. Top 5 Influential Citing Studies:
    a) Perez-Lopez, Raquel, et al. "A guide to artificial intelligence for cancer researchers." Nature Reviews Cancer 24.6 (2024): 427-441. (IF: 72.5)
    b) Xie, Xiao, et al. "Oxidative cyclization reagents reveal tryptophan cation–π interactions." Nature 627.8004 (2024): 680-687. (IF: 50.5)
    c) Khan, Artem, et al. "Metabolic gene function discovery platform GeneMAP identifies SLC25A48 as necessary for mitochondrial choline import." Nature Genetics 56.8 (2024): 1614-1623. (IF: 31.8)
    d) Marsh, Joseph A., and Sarah A. Teichmann. "Predicting pathogenic protein variants." Science 381.6664 (2023): 1284-1285. (IF: 44.7)
    e) Zhang, Qiang, et al. "Scientific large language models: A survey on biological & chemical domains." ACM Computing Surveys 57.6 (2025): 1-38. (IF: 23.8)

c) Corresponding author and their publications:

1. Corresponding author: Pushmeet Kohli, Žiga Avsec.

2. Top 5 publications from Pushmeet Kohli:
   a) Kohli, Pushmeet, L'ubor Ladický, and Philip HS Torr. "Robust higher order potentials for enforcing label consistency." International journal of computer vision 82 (2009): 302-324.
      (IF: 11.6) (As 1. author)
   b) Kohli, Pushmeet, M. Pawan Kumar, and Philip HS Torr. "P3 & beyond: Solving energies with higher order cliques." 2007 IEEE conference on computer vision and pattern recognition. IEEE, 2007.
      (IF: 20.8) (As 1. author)
   c) Jumper, John, et al. "Highly accurate protein structure prediction with AlphaFold." nature 596.7873 (2021): 583-589.
      (IF: 50.5) (36153 citations)
   d) Silberman N, Hoiem D, Kohli P, et al. Indoor segmentation and support inference from rgbd images[C]//Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V 12. Springer Berlin Heidelberg, 2012: 746-760.
      (IF: 33.2) (As 3. author, 7239 citations)
   e) Tunyasuvunakool, Kathryn, et al. "Highly accurate protein structure prediction for the human proteome." Nature 596.7873 (2021): 590-596.
      (IF: 50.5) (First coverage of the entire human proteome)

3. Top 5 publications from Žiga Avsec:
   a) Avsec, Žiga, et al. "Effective gene expression prediction from sequence by integrating long-range interactions." Nature methods 18.10 (2021): 1196-1203.
      (IF: 36.1) (As 1. author)
   b) Avsec, Žiga, et al. "Base-resolution models of transcription-factor binding reveal soft motif syntax." Nature genetics 53.3 (2021): 354-366.
      (IF: 31.8) (As 1. author)
   c) Avsec, Žiga, et al. "The Kipoi repository accelerates community exchange and reuse of predictive models for genomics." Nature biotechnology 37.6 (2019): 592-600.
      (IF: 33.1) (As 1. author)
   d) Eraslan G, Avsec Ž, Gagneur J, et al. Deep learning: new computational modelling techniques for genomics[J]. Nature reviews genetics, 2019, 20(7): 389-403.
      (IF: 39.1) (As 2. author)
   e) Cheng, Jun, et al. "Accurate proteome-wide missense variant effect prediction with AlphaMissense." Science 381.6664 (2023): eadg7492.
      (IF: 44.7) (1088 citations)