P02 Informed and automated k-mer size selection for genome assembly. Chikhi et al. Bioinformatics 2014, 30(1):31-7

Romayssae Boudouassel, Lahem Asres Ioab

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:

R.Chikhi and P.Medvedev

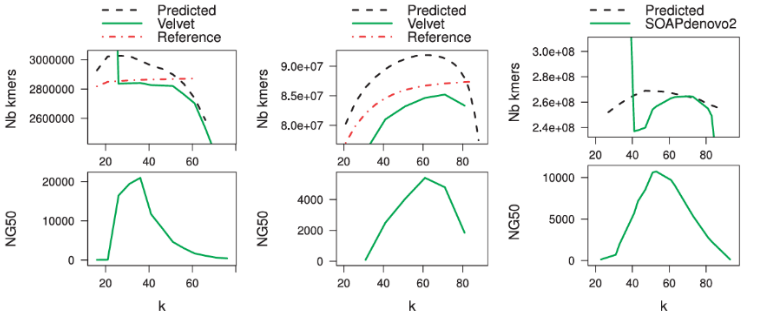


Fig. 4. Relation of the number of distinct genomic *k*-mers to assembly quality. We show the results for the three datasets: *S.aureus* (left), *chr14* (middle) and *B.impatiens* (right). We plot the number of distinct genomic *k*-mers predicted from the histogram from our model, the number present in 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 from histogram, the reference and the assembly by Velvet. It also shows the NG50 of the assembly. The number of distinct genomic k-mers in the assembly of *S.aureus* (left) and *chr14* (middle) closely matches the numbers predicted by KMERGENIE, while the number of the assembly for *B.impatiens* (right) shows variations to the KMERGENIE's predictions. These variations may be due to heterozygosity. 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. This can either be due to heterozygosity or the need of optimization in the statistical model.

P03 Assembly of long, error-prone reads using repeat graphs. Kolmogorov et al. 2019 Nature Biotech 73:540-546

Wanying Deng Dilara Sarach

M-ASA-S

Dilara Sarach (dilara@stud.uni-frankfurt.de)

May 25, 2025

Summary

Repetitive regions are the Achilles heel of any known genome assembly algorithm. The de Bruijn graph introduced great improvement for short-read-based genome reconstructions: Construct graphs of k-mers, connect these in an all-encompassing assembly graph, collapse the repeat-denoting bulges down to simple paths, find a Eulerian walk. This process remains inapplicable to long-read data, however, because the latter does not fulfill a key assumption of the de Bruijn approach - that most k-mers are present in multiple reads.

Kolmogorov et al. propose Flye, an assembler tailored for long, error-prone reads that employs repeat characterization to effectively handle repetitive regions. Instead of contigs, it utilizes rapidly generated strung-together disjointigs – concatenations of multiple disjoint genomic segments – to arrive at an accurate assembly graph. Flye then resolves bridged repeats by aligning all reads to said graph, and unbridged repeats by identifying repeat copy variations, matching every read to a copy based on the variations, then using the reads to derive a consensus sequence for each copy.

Benchmarked against five state-of-the-art assemblers (Canu, Falcon, HINGE, Miniasm, MaSuRCA) with six datasets (BACTERIA, METAGENOME, YEAST, WORM, HUMAN, HUMAN+), Flye produced comparable or higher-quality assemblies based on seven metrics (assembly length, contig number, NG50, reference coverage, reference percentage identity, misassembly number, NGA50). Flye placed on par with HINGE on the BACTERIA dataset – the two reached near-consensus on the structure of the resulting assembly graphs and outperformed the other tools – and with Canu on the WORM and METAGENOME dataset in terms of best assembly contiguity and sequence identity, respectively. The novel program generated the most accurate YEAST and WORM assemblies and often showed significantly better running times – up to an order of magnitude. Most notably, it nearly doubled the contiguity of the human genome in comparison with the first and second runner-ups, MaSuRCA and Canu, with NGA values equal to 6.35Mb, 3.81Mb and 2.87Mb.

The authors see further room for improvement over existing technology: More sophisticated algorithms for unbridged repeats handling could be designed, more segmental duplications – resolved, and higher NGA50 – reached.

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.

Summary of the paper 3 (Kolmogorov et al., 2019)

Kolmogorov et al. address the critical challenge of assembling long, error-prone single-molecule sequencing (SMS) reads, such as those produced by Pacific Biosciences (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, measured by the NGA50 meric. Notably, Flye effectively utilized ultra-long ONT reads to resolve complex mosaic repeats that were unresolved by standard ONT reads alone. Moreover, Flye's runtime was significantly faster, sometimes by an order of magnitude, making it both efficient and scalable.

It is concluded that Flye advances genome assembly methods by combining innovative algorithmic strategies and the use of ultra-long reads. This leads to significantly more accurate and comprehensive genomic reconstructions, particularly of complex and repetitive regions, positioning Flye as a critical tool for future genomic research and clinical applications.

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.

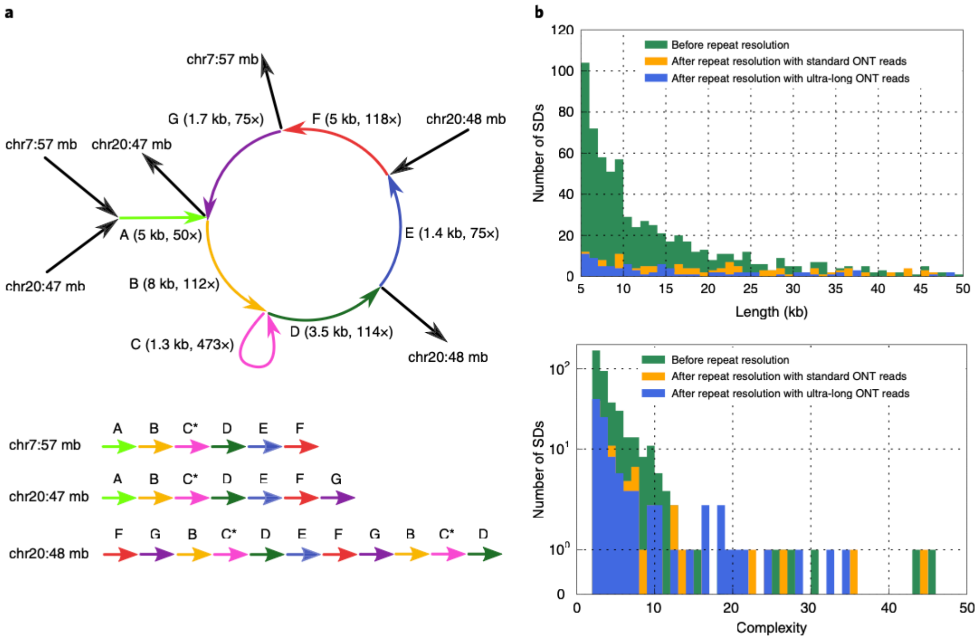


Fig. 4 | An SD from the Flye assembly of the HUMAN dataset and the distribution of the lengths and complexities of all SDs from the same assembly. **a**, A mosaic SD of complexity 7 represented as a connected component formed by repeat edges (7 colored edges of total length 25.7 kb) in the assembly graph 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.

P04 BRAKER3: Fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS, and TSEBRA. Gabriel et al. 2024. Genome Res 34(5):769-777.

Lucie Biesecker, Emre Inciler

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

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.

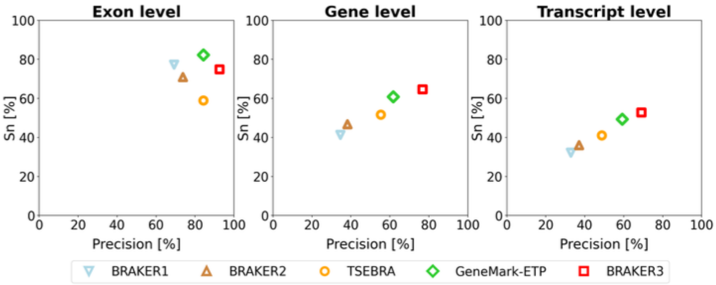


Figure 2: Average precision and sensitivity of gene predictions made by BRAKER1, BRAKER2, TSEBRA, GeneMark-ETP, and BRAKER3 for the genomes of 11 different species (listed in Supplemental Table S1). Inputs were the genomic sequences, short-read RNA-seq libraries, and protein databases (order excluded).

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.

Lucie Biesecker, 7518074 & Emre Inciler, 8675150

P05 MetaEuk—sensitive, high-throughput gene discovery, and annotation for large-scale eukaryotic metagenomics Levy et al. *Microbiome* volume 8, Article number: 48 (2020)

Valeriya Kolos, Ischa Tahir

P06 A long-read RNA-seq approach to identify novel transcripts of very large genes. Uapinyoying et al. 2020 Genome Res 30: 885-897

Jörn Fischer, Tao Le

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

Prech Uapinyoying, Jeremy Goecks, Susan M. Knoblach, et al. 2020

Genome research, 30(6), 885-897

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

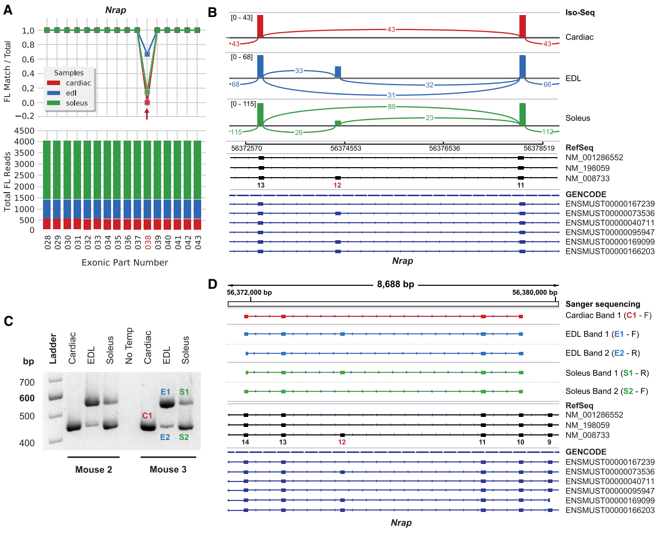
A finding of the present of nrap (5kb) isoforms excluding exon 12, previously believed to only exist in cardiac muscles (Lu et a. 2008), in skeletal muscles of mouse. On the other hand, the exclusive expression of the isoform excluding exon 12 in cardiac muscles is verified. Additionally, a rare transcript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

P07 PureCLIP: capturing target-specific protein–RNA interaction footprints from single-nucleotide CLIP-seq data. Krakau et al. Genome Biology 2017, 18:240

Marie Elise Barié, Anna Tabea Dieter

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.

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.

P08 A Pan-plant Protein Complex Map Reveals Deep Conservation and Novel Assemblies McWhite et al. *Cell* Volume 181, Issue 2, 16 April 2020, Pages 460-474.e14

Leonie Bernshausen, Saaruky Chanthirakanthan

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:

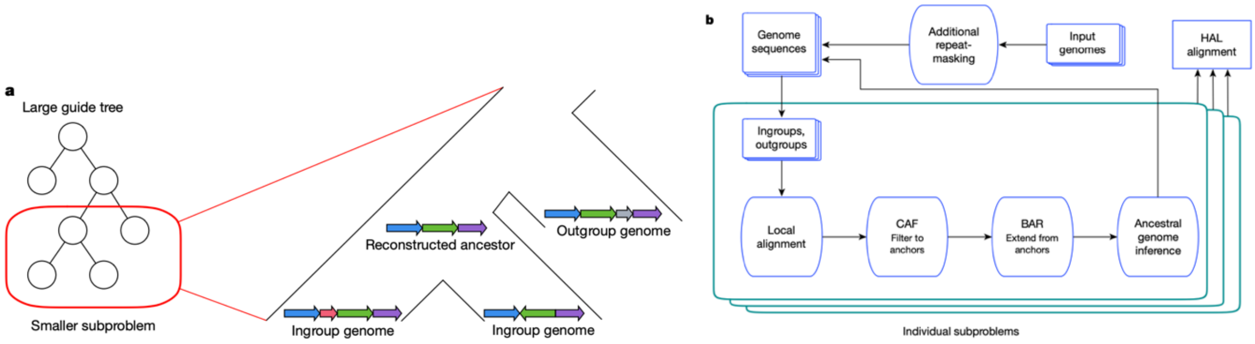
- Progressive Cactus enables linear runtime scaling with increasing genome numbers, outperforming previous implementations with quadradic scaling.
- Demonstrated the largest known alignment of over 600 vertebrate genomes, confirming feasibility and accuracy at scale.
- Achieved top alignment accuracy 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



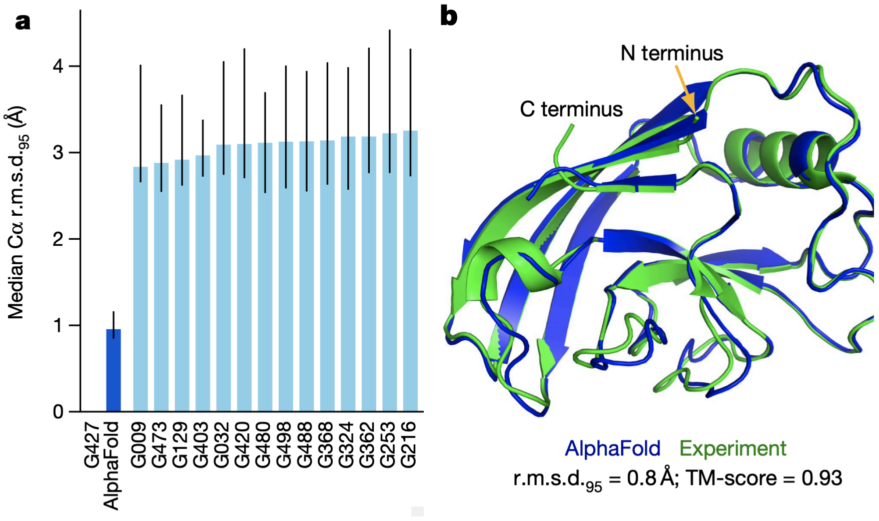
P10 Highly accurate protein structure prediction with AlphaFold Jumper et al. *Nature* volume 596, pages 583–589 (2021)

Eren Alkanat, Piravinth Paraparan

Eren Alkanat Piravinth Paraparan

Summary of “Highly accurate protein structure prediction with AlphaFold by Jumper et al. in Nature (2021)

The paper revolves around a central problem in structural biology known as the protein folding problem by predicting a protein’s three-dimensional structure from its amino acid sequence. Existing computational methods often fall short of atomic-level accuracy, particularly when no similar structures are known. Therefore, the authors introduce ‘AlphaFold’, a novel deep learning system that predicts protein structures with atomic accuracy even without homologous templates which is the main research question of the paper at the same time. The model employs a redesigned architecture that integrates evolutionary information from multiple sequence alignments (MSAs). Its core innovations include the Evoformer module for learning pairwise residue interactions and the ‘Invariant Point Attention (IPA)’ mechanism for building 3D structures. Through a process called ‘recycling’, ‘AlphaFold’ iteratively refines its predictions, improving the structure with each pass. Furthermore, the training procedure incorporates self-distillation, where the model learns from its own high-confidence predictions on unlabeled sequences. ‘AlphaFold’ was evaluated in the 14th Critical Assessment of Structure Prediction (CASP14) and demonstrated backbone root-mean-square deviation (r.m.s.d.) of 0.96 Å at 95%-coverage that approaches the experimental level and exceeds all competing methods. Hence, a central and introductory figure in the paper is Figure 1 that illustrates this case where AlphaFold is shown as very accurate relative to the next-best method in CASP14 (Figure 1a). Figure 1a’s panel shows AlphaFold’s backbone accuracy across 87 protein domains from CASP14. The main metric is Cα root-mean-square deviation at 95%-coverage. The next-best method has just 2.8 Å. For reference, a carbon atom is about 1.4 Å wide, so AlphaFold is achieving near-atomic accuracy. Figure 1b shows a side-by-side comparison of AlphaFold’s prediction (blue) and the experimentally determined structure (green) for a target from CASP14. Finally, Figure 1e illustrates the end-to-end structure prediction pipeline used by AlphaFold. Key components are the input in the form of amino acid sequence, MSAs, and templates, the MSA representation, the Evoformer with 48 blocks processing the MSA and pairwise residue features, the structure module with 8 blocks using the processed representations to predict the 3D atomic structure, and the recycling module where the prediction is refined by re-feeding outputs into the network multiple times. The recycling module of AlphaFold, which is explicitly designed to iteratively refine the predicted protein structure until no significant improvements can be made. The whole network then predicts 3D coordinates for all atoms along with confidence scores like pLDDT, pTM etc.



(Figure 1a and Figure 1b)

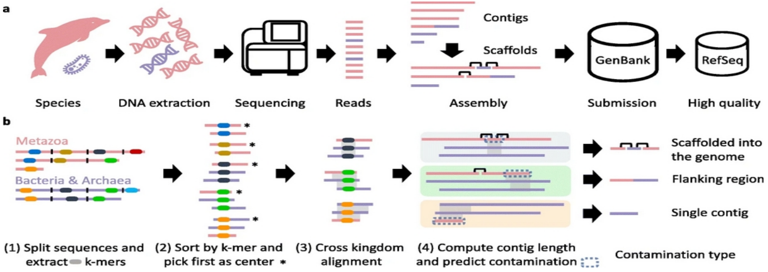
Moreover, the authors acknowledge limitations and factors that affect AlphaFold's accuracy even though it performs very well. The key factor, MSA depth, is critical for accurate predictions. When a protein has fewer than 30 aligned sequences, AlphaFold's accuracy drops significantly. There are more factors listed in the paper. The study concludes that AlphaFold represents a major breakthrough with the potential to transform biological research by enabling proteome-scale structure prediction. Looking forward, the authors note that AlphaFold's framework could be extended to predict multi-protein complexes and dynamical conformational states. As shown in companion work, the method has already been applied to the entire human proteome, signaling a new era in computational biology where structural information becomes broadly accessible.

P11 Terminating contamination: large-scale search identifies more than 2,000,000 contaminated entries in GenBank. Steinegger and Salzberg *Genome Biology* Vol. 21 115 (2020)

Philipp Berger, Timothy Voss

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
Contigs
Scaffolds
GenBank
RefSeq
Species
DNA extraction
Sequencing
Reads
Assembly
Submission
High quality
b
Metazoa
Bacteria & Archaea
*
Scaffolded into the genome
Flanking region
Single contig
(1) Split sequences and extract k-mers
(2) Sort by k-mer and pick first as center *
(3) Cross kingdom alignment
(4) Compute contig length and predict contamination
Contamination type

***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.

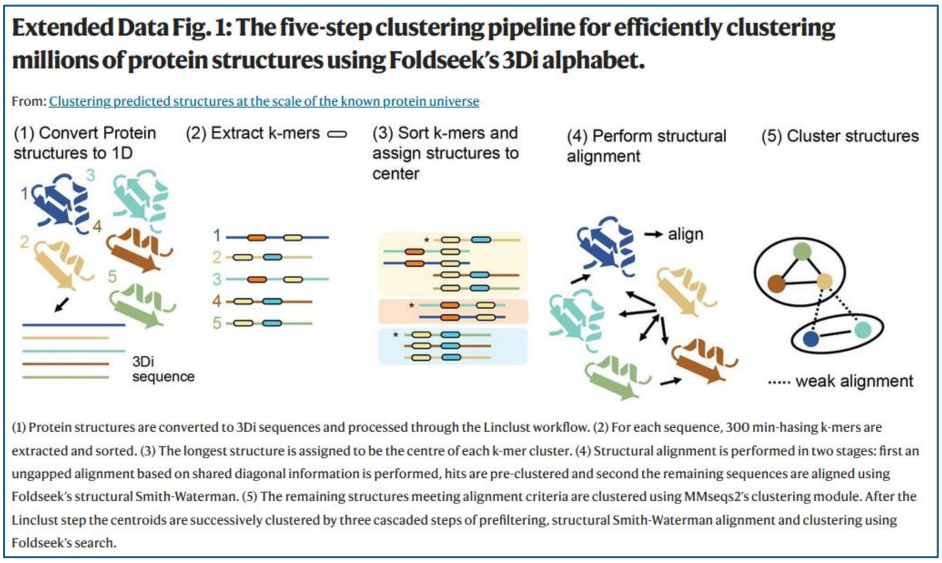
P13 Clustering predicted structures at the scale of the known protein universe. Barrio-Hernandez et al. Nature 622: 637–645 (2023)

Mehmet Gün Kutalmis Batman, Haoxuan Zeng

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) Convert Protein structures to 1D
(2) Extract k-mers
(3) Sort k-mers and assign structures to center
(4) Perform structural alignment
(5) Cluster structures
1
2
3
4
5
3Di sequence
1
2
3
4
5
*
*
*
align
..... weak alignment

(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“.
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.

P14 Fast and sensitive taxonomic assignment to metagenomic contigs. Mirdita et al. Bioinformatics 37(18):3029–3031 (2021)

Daulet Ashirov, Jiamei Qin

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 their total weights (Fig.1.).

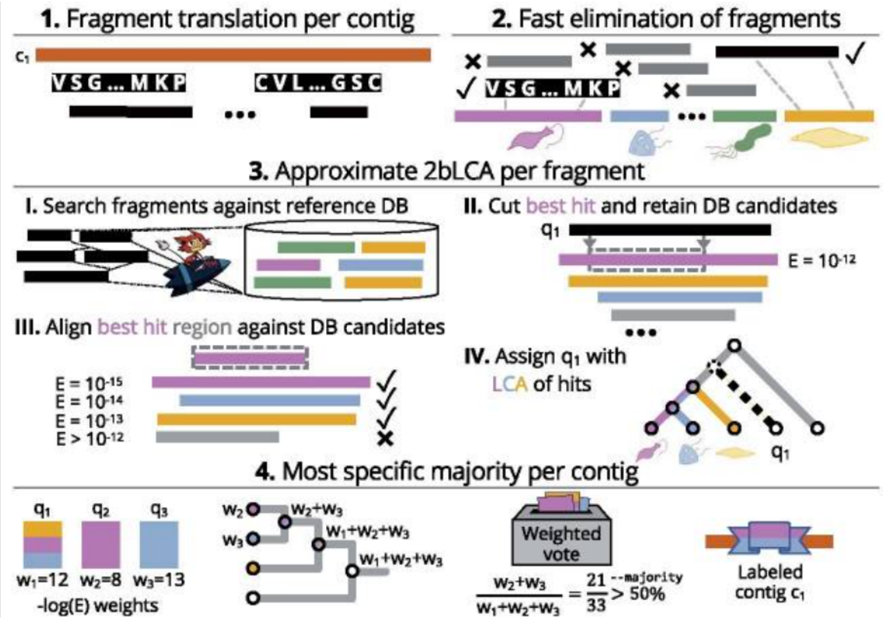

1. Fragment translation per contig
c₁
VSG…MKP
CVL…GSC
2. Fast elimination of fragments
VSG…MKP
3. Approximate 2bLCA per fragment
I. Search fragments against reference DB
II. Cut best hit and retain DB candidates
q₁
E = 10-12
III. Align best hit region against DB candidates
E = 10-15
E = 10-14
E = 10-13
E > 10-12
IV. Assign q₁ with LCA of hits
q₁
4. Most specific majority per contig
q₁ q₂ q₃
w₁=12 w₂=8 w₃=13
-log(E) weights
w₂ w₂+w₃
w₃ w₁+w₂+w₃
w₁+w₂+w₃
Weighted vote
w₂+w₃ / w₁+w₂+w₃ = 21/33 > 50% --majority
Labeled contig c₁

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.

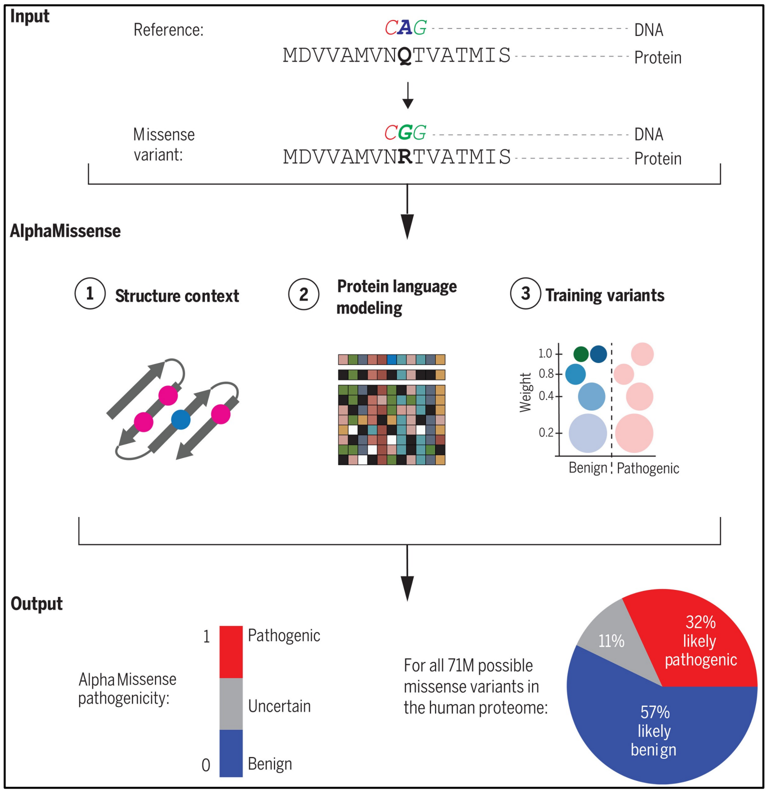
P15 Accurate proteome-wide missense variant effect prediction with AlphaMissense. Cheng et al. Science 381:eadg7492 (2023)

Yakun Li, Fatih Sahin

Summary: Accurate proteome-wide missense variant effect prediction with AlphaMissense

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 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.



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.