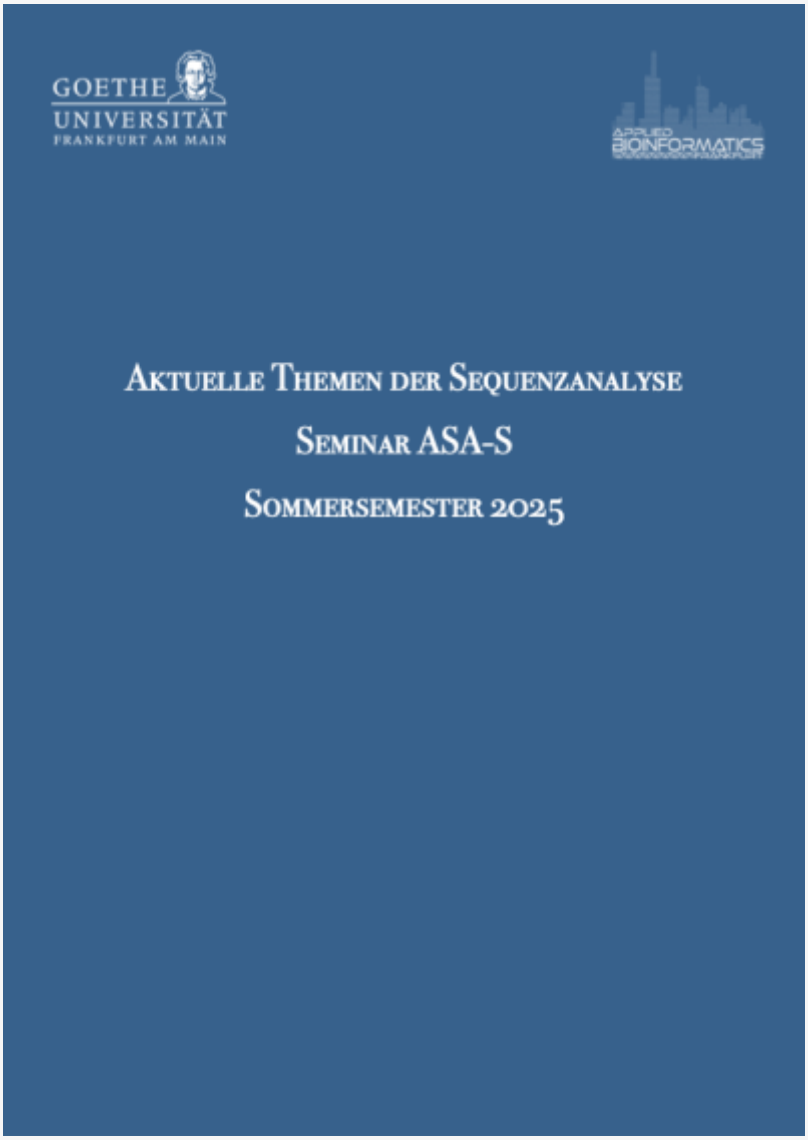
Postersession 2025

Abstract Book



asa-s_2025-abstract_book.pdf

Flash talks 2025

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

Motivation: Genome assembly tools based on the de Bruijn graph framework rely on a parameter

k, which represents a trade-off be- tween several competing effects that are difficult to quantify. There is currently a lack of tools that would automatically estimate the best k to use and/or quickly generate histograms of k-mer abundances that would allow the user to make an informed decision. Results: We develop a fast and accurate sampling method that con- structs approximate abundance histograms with several orders of magnitude performance improvement over traditional methods. We then present a fast heuristic that uses the generated abundance histo- grams for putative k values to estimate the best possible value of k. We test the effectiveness of our tool using diverse sequencing data- sets and find that its choice of k leads to some of the best assemblies. Availability: Our tool KMERGENIE is freely available at: http://kmergenie.bx.psu.edu.

Link to PDF

Quiz:

p2_quizz.pdf

Team: Boudouassel, Ioan

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

Accurate genome assembly is hampered by repetitive regions. Although long single molecule sequencing reads are better able to resolve genomic repeats than short-read data, most long-read assembly algorithms do not provide the repeat character- ization necessary for producing optimal assemblies. Here, we present Flye, a long-read assembly algorithm that generates arbitrary paths in an unknown repeat graph, called disjointigs, and constructs an accurate repeat graph from these error-rid- dled disjointigs. We benchmark Flye against five state-of-the-art assemblers and show that it generates better or comparable assemblies, while being an order of magnitude faster. Flye nearly doubled the contiguity of the human genome assembly (as

measured by the NGA50 assembly quality metric) compared with existing assemblers.

Link to PDF

Video → separate file

Quiz:

p03_flashtalk_quiz.pdf

Team: Sarach, Deng

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.

Gene prediction has remained an active area of bioinformatics research for a long time. Still, gene prediction in large eukaryotic genomes presents a challenge that must be addressed by new algorithms. The amount and significance of the evidence available from transcriptomes and proteomes vary across genomes, between genes, and even along a single gene. User-friendly and accurate annotation pipelines that can cope with such data heterogeneity are needed. The previously developed annotation pipelines BRAKER1 and BRAKER2 use RNA-seq or protein data, respectively, but not both. A further significant performance improvement integrating all three data types was made by the recently released GeneMark-ETP. We here present the BRAKER3 pipeline that builds on GeneMark-ETP and AUGUSTUS, and further improves accuracy using the TSEBRA combiner. BRAKER3 annotates protein-coding genes in eukaryotic genomes using both short-read RNA-seq and a large protein database, along with statistical models learned iteratively and specifically for the target genome. We benchmarked the new pipeline on genomes of 11 species under an assumed level of relatedness of the target species proteome to available proteomes. BRAKER3 outperforms BRAKER1 and BRAKER2. The average transcript-level F1-score is increased by about 20 percentage points on average, whereas the difference is most pronounced for species with large and complex genomes. BRAKER3 also outperforms other existing tools, MAKER2, Funannotate, and FINDER. The code of BRAKER3 is available on GitHub and as a ready-to-run Docker container for execution with Docker or Singularity. Overall, BRAKER3 is an accurate, easy-to-use tool for eukaryotic genome annotation.

Link to PDF

Team: Biesecker, Inciler

MetaEuk—sensitive, high-throughput gene discovery, and annotation for large-scale eukaryotic

metagenomics Levy et al. Microbiome volume 8, Article number: 48 (2020)

Background Metagenomics is revolutionizing the study of microorganisms and their involvement in biological, biomedical, and geochemical processes, allowing us to investigate by direct sequencing a tremendous diversity of organisms without the need for prior cultivation. Unicellular eukaryotes play essential roles in most microbial communities as chief predators, decomposers, phototrophs, bacterial hosts, symbionts, and parasites to plants and animals. Investigating their roles is therefore of great interest to ecology, biotechnology, human health, and evolution. However, the generally lower sequencing coverage, their more complex gene and genome architectures, and a lack of eukaryote-specific experimental and computational procedures have kept them on the sidelines of metagenomics. Results MetaEuk is a toolkit for high-throughput, reference-based discovery, and annotation of protein-coding genes in eukaryotic metagenomic contigs. It performs fast searches with 6-frame-translated fragments covering all possible exons and optimally combines matches into multi-exon proteins. We used a benchmark of seven diverse, annotated genomes to show that MetaEuk is highly sensitive even under conditions of low sequence similarity to the reference database. To demonstrate MetaEuk's power to discover novel eukaryotic proteins in large-scale metagenomic data, we assembled contigs from 912 samples of the Tara Oceans project. MetaEuk predicted >12,000,000 protein-coding genes in 8 days on ten 16-core servers. Most of the discovered proteins are highly diverged from known proteins and originate from very sparsely sampled eukaryotic supergroups. Conclusion The open-source (GPLv3) MetaEuk software (https://github.com/soedinglab/metaeuk) enables large-scale eukaryotic metagenomics through reference-based, sensitive taxonomic and functional annotation.

Link to PDF

Quiz:

the_metaeuk_quiz.pdf

Team: Kolos, Tahir

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

RNA-seq is widely used for studying gene expression, but commonly used sequencing platforms produce short reads that only span up to 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. We used long-read isoform sequencing combined with a novel analysis approach to compare alternative splicing of large, repetitive structural genes in muscles. Analysis of muscle structural genes that produce medium (Nrap: 5 kb), large (Neb: 22 kb), and very large (Ttn: 106 kb) transcripts in cardiac muscle, and fast and slow skeletal

muscles identified unannotated exons for each of these ubiquitous muscle genes. This also identified differential exon usage and phasing for these genes between the different muscle types. By mapping the in-phase transcript structures to known annotations, we also identified and quantified previously unannotated transcripts. Results were confirmed by endpoint PCR and Sanger sequencing, which revealed muscle-type-specific differential expression of these novel transcripts. The improved transcript identification and quantification shown by our approach removes previous impediments to studies aimed at quantitative differential expression of ultralong transcripts.

Link to PDF

Team: Le, Fischer

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

The iCLIP and eCLIP techniques facilitate the detection of protein–RNA interaction sites at high resolution, based on diagnostic events at crosslink sites. However, previous methods do not explicitly model the specifics of iCLIP and eCLIP truncation patterns and possible biases. We developed PureCLIP (https://github.com/skrakau/PureCLIP), a hidden Markov model based approach, which simultaneously performs peak-calling and individual crosslink site detection. It explicitly incorporates a non-specific background signal and, for the first time, non-specific sequence biases. On both simulated and real data, PureCLIP is more accurate in calling crosslink sites than other state-of-the-art methods and has a higher agreement across replicates.

Link to PDF

Team: Barie, Dieter

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

New genome assemblies have been arriving at a rapidly increasing pace, thanks to decreases in sequencing costs and improvements in third-generation sequencing technologies1,2,3. For example, the number of vertebrate genome assemblies currently in the NCBI (National Center for Biotechnology Information) database4 increased by more than 50% to 1,485 assemblies in the year from July 2018 to July 2019. In addition to this influx of assemblies from different species, new human de novo assemblies5 are being produced, which enable the analysis of not only small polymorphisms, but also complex, large-scale structural differences between human individuals and haplotypes. This coming era and its unprecedented amount of data offer the

opportunity to uncover many insights into genome evolution but also present challenges in how to adapt current analysis methods to meet the increased scale. Cactus6, a reference-free multiple genome alignment program, has been shown to be highly accurate, but the existing implementation scales poorly with increasing numbers of genomes, and struggles in regions of highly duplicated sequences. Here we describe progressive extensions to Cactus to create Progressive Cactus, which enables the reference-free alignment of tens to thousands of large vertebrate genomes while maintaining high alignment quality. We describe results from an alignment of more than 600 amniote genomes, which is to our knowledge the largest multiple vertebrate genome alignment created so far.

Link to PDF

Team: Bernshausen, Chanthirakanthan

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

Proteins are essential to life, and understanding their structure can facilitate a mechanistic understanding of their function. Through an enormous experimental effort1,2,3,4, the structures of around 100,000 unique proteins have been determined5, but this represents a small fraction of the billions of known protein sequences6,7. Structural coverage is bottlenecked by the months to years of painstaking effort required to determine a single protein structure. Accurate computational approaches are needed to address this gap and to enable large-scale structural bioinformatics. Predicting the three-dimensional structure that a protein will adopt based solely on its amino acid sequence—the structure prediction component of the ‘protein folding problem’8—has been an important open research problem for more than 50 years9. Despite recent progress10,11,12,13,14, existing methods fall far short of atomic accuracy, especially when no homologous structure is available. Here we provide the first computational method that can regularly predict protein structures with atomic accuracy even in cases in which no similar structure is known. We validated an entirely redesigned version of our neural network-based model, AlphaFold, in the challenging 14th Critical Assessment of protein Structure Prediction (CASP14)15, demonstrating accuracy competitive with experimental structures in a majority of cases and greatly outperforming other methods. Underpinning the latest version of AlphaFold is a novel machine learning approach that incorporates physical and biological knowledge about protein structure, leveraging multi-sequence alignments, into the design of the deep learning algorithm.

Link to PDF

Team: Alkanat, Paraparan

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

Genomic analyses are sensitive to contamination in public databases caused by incorrectly labeled reference sequences. Here, we describe Conterminator, an efficient method to detect and remove incorrectly labeled sequences by an exhaustive all-against-all sequence comparison. Our analysis reports contamination of 2,161,746, 114,035, and 14,148 sequences in the RefSeq, GenBank, and NR databases, respectively, spanning the whole range from draft to “complete” model organism genomes. Our method scales linearly with input size and can process 3.3 TB in 12 days on a 32-core computer. Conterminator can help ensure the quality of reference databases. Source code (GPLv3): https://github.com/martin-steinegger/conterminator

Link to PDF

Team: Berger, Voss

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

Proteins are key to all cellular processes and their structure is important in understanding their function and evolution. Sequence-based predictions of protein structures have increased in accuracy1, and over 214 million predicted structures are available in the AlphaFold database2. However, studying protein structures at this scale requires highly efficient methods. Here, we developed a structural-alignment-based clustering algorithm—Foldseek cluster—that can cluster hundreds of millions of structures. Using this method, we have clustered all of the structures in the AlphaFold database, identifying 2.30 million non-singleton structural clusters, of which 31% lack annotations representing probable previously undescribed structures. Clusters without annotation tend to have few representatives covering only 4% of all proteins in the AlphaFold database. Evolutionary analysis suggests that most clusters are ancient in origin but 4% seem to be species specific, representing lower-quality predictions or examples of de novo gene birth. We also show how structural comparisons can be used to predict domain families and their relationships, identifying examples of remote structural similarity. On the basis of these analyses, we identify several examples of human immune-related proteins with putative remote homology in prokaryotic species, illustrating the value of this resource for studying protein function and evolution across the tree of life.

Link to PDF

Team: Batman, Zeng

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

Summary MMseqs2 taxonomy is a new tool to assign taxonomic labels to metagenomic contigs. It extracts all possible protein fragments from each contig, quickly retains those that can contribute to taxonomic annotation, assigns them with robust labels and determines the contig's taxonomic identity by weighted voting. Its fragment extraction step is suitable for the analysis of all domains of life. MMseqs2 taxonomy is 2–18× faster than state-of-the-art tools and also contains new modules for creating and manipulating taxonomic reference databases as well as reporting and visualizing taxonomic assignments. **Availability and implementation** MMseqs2 taxonomy is part of the MMseqs2 free open-source software package available for Linux, macOS and Windows at https://mmseqs.com.

Link to PDF

Quiz:

quiz_p14_2025.pdf

Team: Ashirov, Qin

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

The vast majority of missense variants observed in the human genome are of unknown clinical significance. We present AlphaMissense, an adaptation of AlphaFold fine-tuned on human and primate variant population frequency databases to predict missense variant pathogenicity. By combining structural context and evolutionary conservation, our model achieves state-of-the-art results across a wide range of genetic and experimental benchmarks, all without explicitly training on such data. The average pathogenicity score of genes is also predictive for their cell essentiality, capable of identifying short essential genes that existing statistical approaches are underpowered to detect. As a resource to the community, we provide a database of predictions for all possible human single amino acid substitutions and classify 89% of missense variants as either likely benign or likely pathogenic.

Link to PDF

Team: Li, Sahin