Abstracts: Blattodea

Yang *et al.*, 2019, NCBI's Conserved Domain Database andTools for Protein Domain Analysis.

Use Standalone RPS-BLAST and rpsbproc to compute and retrieve domain annotation programmatically.

Hoff Stanke, 2018, Predicting Genes in Single Genomes with AUGUSTUS.

Ab initio gene prediction by AUGUSTUS.

Fierst Murdock, 2017, Decontaminating eukaryotic genome assemblies with machine learning

Decision tree classifier performed well for decontaminating genome assemblies. Mean DNA sequencing coverage and mean RNA sequencing coverage had the highest Gini importances and a model constructed solely with these predictors was able to correctly classify ¿97% of the *C. remanei* dataset. When a third predictor, the percent of the scaffold covered in RNA alignment, was added the model correctly classified ¿98% of the dataset. Model accuracy and sensitivity plateaued above 99.5% when a fourth variable, scaffold GC content, was included but specificity increased slightly as successive predictors were added to the model.

Qu *et al.*, 2020, Effect of sequence depth and length in long-read assembly of the maize inbred NC358.

We have documented how both the completeness and contiguity of assemblies improve with increasing depth and read length. With long-read sequencing (PacBio), the low-copy gene space (including tandem gene arrays) can be well assembled with as low as 30x genomic coverage across a range of read lengths. Complete characterization of transposon elements in complex genomes such as maize will require a greater depth of sequence (40x) and should employ library preparation protocols that maximize read-length N50. Finally, complete assembly of highly repetitive genomic features such as heterochromatic knobs, telomeres, and centromeres will require substantially more data. In fact, complete assembly of these latter highly repetitive sequences will likely require innovations beyond current sequencing technology.

Amarasinghe *et al.*, 2020, Opportunities and challenges in long read sequencing data analysis.

Yandell Ence, 2012, A beginner's guide to eukaryotic genome annotation.

Mudge Harrow, 2016, The state of play in higher eukaryotic gene annotation.

Kapli *et al.*, 2020, Phylogenetic tree building in the genomic age.

Pearson, 2013, An introduction to sequence similarity ("homology") searching.

Howe *et al.*, 2021, Significantly improving the quality of genome assemblies through curation.

Zhang *et al.*, 2020, A comprehensive evaluation of long read error correction methods.
Long read error correction strategies are catagorized as hybrid (requires short accurate reads) or non-hybrid. Hybrid strategies include two methods: alignment-based that aligns short reads to long reads, and assembly-based that aligns long reads to short-read-assembled contigs or de Bruijn graph. Non-hybrid strategies generate consensus sequences using overlap information between long reads.

Hybrid methods aided by short accurate reads can achieve better correction quality, especially when handling low coverage-depth long reads, compared with non-hybrid methods. Within the hybrid methods, assembly-based methods are superior to alignment-based methods in terms of scalability to large data sets. Besides, better performance on correction such as preserving higher proportion of input bases and high alignment identity often leads to better performance. FMLRC outperformed other hybrid methods in almost all the experiments.

Sovic *et al.*, 2016, Evaluation of hybrid and non-hybrid methods for de novo assembly of nanopore reads.
We benchmarked five non-hybrid (in terms of both error correction and scaffolding) assembly pipelines as well as two hybrid assemblers which use third generation sequencing data to scaffold Illumina assemblies. Tests were performed on several publicly available MinION and Illumina datasets of *Escherichia coli* K-12, using several sequencing coverages of nanopore data (20x, 30x, 40x and 50x). Results show that hybrid methods are highly dependent on the quality of NGS data, but much less on the quality and coverage of nanopore data and perform relatively well on lower nanopore coverages. All non-hybrid methods correctly assemble the *E. coli* genome when coverage is above 40x, even the non-hybrid method tailored for Pacific Biosciences reads.

Mascher *et al.*, 2021, Long-read sequence assembly: a technical evaluation in barley.

The recent development of fast and accurate long-read sequencing by circular consensus sequencing (CCS) on the PacBio platform may greatly increase the scope of plant pan-genome projects. Here, we compare current long-read sequencing platforms regarding their ability to rapidly generate contiguous sequence assemblies in pan-genome studies of barley (*Hordeum vulgare*). Most long-read assemblies are clearly superior to the current barley reference sequence based on short-reads. Assemblies derived from accurate long reads excel in most metrics, but the CCS approach was the most cost-effective strategy for assembling tens of barley genomes. A downsampling analysis indicated that 20-fold CCS coverage can yield very good sequence assemblies, while even five-fold CCS data may capture the complete sequence of most genes.

Chakraborty *et al.*, 2016, Contiguous and accurate de novo assembly of metazoan genomes with modest long read coverage.

Regarding assembly, we recommend that researchers obtain between 50x and 100x Illumina sequence. Next, researchers must determine how much long molecule coverage to obtain: between 25x and 35x, or ¿35x. With coverage below 35x, PacBio only methods often fail to assemble and produce low contiguity when they do assemble, and thus, we can only confidently recommend a hybrid assembly. Above 35x, we recommend meta assembly of a hybrid and a PacBio only assembly. In this case, we recommend down-sampling to the 30x longest PacBio reads when generating the hybrid assembly because hybrid assembly contiguity decreases above this coverage level, but this has not been extensively tested. We show that this approach is effective both in Drosophila and human genomes, which differ in size and extent of repetitive regions.

Scalzitti *et al.*, 2020, A benchmark study of ab initio gene prediction methods in diverse eukaryotic organisms.

The experiments showed that ab initio gene structure prediction is a very challenging task, which should be further investigated.