Genome Biology

MEETING REPORT

Strategies towards sequencing complex crop genomes

James C Abbott* and Sarah A Butcher

Abstract

A report on the Strategies for *de novo* assemblies of complex crop genomes workshop held at The Genome Analysis Centre, Norwich, UK, 8-10 October 2012.

Keywords *de novo* assembly, crop genome, polyploidy, repetitive DNA, next-generation sequencing

Introduction

Many economically important crop species have large and complex genomes that pose significant challenges in producing high-quality genome assemblies. The genomes tend to be significantly larger than mammalian genomes and frequently contain a high proportion of repetitive DNA; additionally, many centuries of selective breeding and hybridization have resulted in agriculturally important species that have highly heterozygous alloploid genomes. Recent advances in sequencing technologies and associated reductions in costs have led to species previously considered intractable to be candidates for genomic sequencing. However, the shorter reads resulting from current technologies typically result in a lower quality assembly than conventional approaches. Thus, considerable technical obstacles must be overcome before such complex genomes can be fully assembled using such methodologies.

This European Science Foundation (ESF) funded workshop brought together scientists from a range of disciplines with an interest in assembly of complex crop genomes. The main focus was on providing a discussion forum for developing strategies to identify and address some of the issues that will be encountered when assembling crop genomes.

*Correspondence: j.abbott@imperial.ac.uk
Centre for Investigative Systems Biology and Bioinformatics, Division of Molecular Biosciences, Faculty of Natural Sciences, Imperial College London, London SW7 2AZ, UK

Size matters...

Keynote speaker Michael Schatz (Cold Spring Harbor Laboratory, USA) described the basic ingredients of a recipe for achieving a good quality assembly and showed results emphasizing the impact of long-range mate-pairs and long-read technologies. The benefit of long-range mate-pairs was demonstrated in the assembly of the sacred lotus (*Nelumbo nucifera*) genome. A combination of Illumina libraries were sequenced with a range of insert sizes up to 8 kb, producing an ALLPATHS assembly with an N50 of 600 bp (N50 is defined as the length for which half of all bases in the assembly are in a contig of more than that length). Including additional reads from a 20 kb mate-paired 454 library resulted in a dramatic N50 increase to 16 Mb, although this approach was less successful when applied to highly heterozygous genomes. Schatz reported obtaining median read-lengths of roughly 4 kb (with maximum read length of up to 20 kb) from Pacific Biosciences single-molecule, real-time sequencing technology, although with an accuracy of 85%. Various strategies for overcoming this error rate were described, alongside modifications to the Celera Assembler supporting its use with long reads.

Memory utilization by de Bruijn graph-based assemblers is a significant obstacle to their use for large and complex genomes. These assemblers represent subsequences of length k (k-mers) as nodes of a directed acyclic graph, with the edges representing the overlaps of k-1 to the adjacent k-mers in the genome, requiring many more k-mers than the number of sequence reads to be stored in memory. Rayan Chikhi (ENS Cachan Brittany, France) described an efficient method for storing *k*-mers during the assembly process through the use of Bloom filters: a bit array stores a hash value for each *k*-mer, which reduces the memory requirement five-fold (where $k = 25$). Remaining on the assembly theme, Zemin Ning (Wellcome Trust Sanger Institute, UK) described large plant genome assemblies generated using Phusion2, a pipeline that initially pre-clusters reads to reduce the size and complexity of the contig assembly problem. Multiple assemblies are produced using a number of different assembly algorithms, before merging the resulting contigs and scaffolds. Moving onto genome scaffolding, Ning

described Spinner, a string-graph-based standalone scaffolder, and Andrea Telatin (University of Padua, Italy) reported on the use of bacterial artificial chromosome pools to sub-sample the genome, allowing long-range scaffolds to be ordered against known pieces of the genome.

Evaluating the outputs of a *de novo* assembly can be difficult in the absence of a reference genome for comparison, or without additional genetic resources. Commonly used statistics such as N50 only provide sizing information, with no indication of correctness. The use of feature response curves to provide a straightforward metric without the requirement of a reference genome was reported by Francesco Vezzi (SciLifeLab, Sweden). The technique depends on monitoring features such as high single nucleotide polymorphism numbers and outlying *k*-mer distributions to identify misassembled regions within contigs.

Complex genome assemblies

The bread wheat (*Triticum aestivum*) sequencing project received attention in several talks. This 17 Gb, allohexaploid genome with an 85% repetitive sequence content undoubtedly makes for the most complex assembly project undertaken so far. Frederic Choulet (INRA, France) described the production of the first 18 Mb of contiguous sequence from the 3B chromosome, which contained at least 50% more genes than would be expected, including many tandem duplications of genes, and a high proportion of pseudogenes. Difficulties with *T. aestivum* genome assembly are exacerbated by an extremely high percentage of non-unique *k*-mers, with 35% non-unique when $k = 100$. Klaus Mayer (MIPS Munich, Germany) reported the use of gene sequences from *Brachypodium*, *Sorghum* and rice as *in silico* exon capture baits, which enabled a 'genes only' approach to wheat assembly, where the genes were first grouped into families based on orthologous groups and used as assembly templates.

Björn Nystedt (SciLifeLab, Sweden) discussed the approach being taken to sequence the 20 Gb genome of Norway spruce (*Picea abies*). Although the spruce is diploid, its chromosomes are not amenable to flow-sorting since they are very similar in size, and so a whole genome shotgun sequencing approach is required. Haploid tissue can be obtained from the megagametophyte (also known as the embryo sac), but the restricted amount of DNA that can be isolated from this source has produced only fragmented assemblies. These are improved by merging them with assemblies from pooled fosmid sequences derived from diploid tissues, which provide limitless quantities of DNA but are too heterozygous to readily assemble alone.

Focus groups

Breakout groups were tasked with discussing some of the issues around the areas of repeats and scaffolding, *de novo* transcriptome assembly, handling complex genome structures, assembly validation and hybrid assembly approaches, and the outputs of these were reported to the meeting.

On the subject of repeats and scaffolding, after exploring mechanisms of how repeats interrupt assembly, discussion focused on the relative importance of correctly assembling repeats and whether this was strictly necessary. It was generally accepted that there was in many cases little to be gained from correctly resolving the structure of highly repeated transposable elements within a genome, and instead it would be more prudent to provide a consensus sequence and an indication of the genomic span of the repeat. However, in cases that may affect mechanisms underlying phenotypes, such as copy number variations or transposable elements potentially regulating adjacent genes, correct repeat resolution was considered to be of critical importance. There was also an acceptance that more could be done to improve assembly of such regions, with suggestions including making use of repeat borders for scaffolding and using techniques such as whole genome profiling and nanochannel mapping.

De novo transcript assembly was reported as remaining an unsolved problem in complex genomes, being confounded by large heterozygous gene families, although it remains a valuable tool in assembling the genic sequences of intractable genomes. The best approaches currently available to improve transcript assembly include combining the outputs of different algorithms or *k*-mer sizes. Although long-read sequencing instruments offer much to assist *de novo* genome assembly, highly abundant transcripts are likely to dominate these platforms when sequencing transcriptomes, with rarer transcripts being lost. Alternative library preparation methods, such as strand-specific RNA libraries and 5' cap RNA sequencing libraries (enriching for low abundance genes), may help with improving the output of these processes.

A sequencing strategy for complex genome structures was suggested, starting with capturing the genic sequences using a combination of RNA sequencing and whole genome shotgun sequencing approaches to build a good catalog of genes and variation, followed by applying additional whole genome sequencing data and other resources to derive the structure. Assembly techniques for such genomes were still considered immature, and it was suggested that methodological improvements could be helped by using a test case genome from an appropriate organism with a low repetitive content, allowing the issue of polyploid assembly to be addressed in isolation. Visualization was identified as being a difficulty, with current genome browsers unable to effectively

visualize polyploid genomes. Additional technologies were also considered useful in some circumstances but were not always beneficial. For example, there are limitations on the maximum obtainable fragment size from chromosome sorting, and mapping techniques (including optical mapping) can be problematic in polyploid organisms.

The group tasked with discussing assembly validation agreed that size-based metrics of genome assemblies do not provide a reliable indication of the underlying quality. In the absence of a reference genome, it is only really possible to determine internal consistency. The use of additional data not included in the assembly (such as additional sequencing libraries, expressed sequence tags, synteny and gene content) to help determine the validity of an assembly was considered a sound approach. There is, however, no standard way of representing this information, so it was suggested that the commonly used AGP ('A Golden Path') format could be extended to allow representation of these types of additional evidence and supporting decisions made during assembly.

The final group to report, on hybrid technologies, looked at combining different technologies in an iterative manner to correct existing errors and produce a superior result. There is currently a considerable trade-off between sequencing cost, accuracy and read length, with each platform offering a different balance. The new long-read platforms may offer greater benefits for determining the large-scale organization of the genome than in directly contributing to the sequence. Improvements to library preparation methodologies (such as reducing the amount of DNA required for creating long-insert libraries) were considered highly desirable. Maintaining a good understanding of the error models of the different platforms is essential in creating hybrid assemblies, but these rapidly change with continual iterations of instruments and chemistries. It was considered that the routine provision by the instrument vendors of the error model of a platform in a standardized format would be highly advantageous.

Conclusions

Overall, the workshop highlighted important advances that have been made in sequencing complex crop genomes, and also identified key areas that should be targeted to improve the quality of crop genome assemblies. What is clear is that exciting times are afoot for crop genome research.

Competing interests

The authors declare that they have no competing interests.

Published: 19 November 2012

doi:10.1186/gb-2012-13-11-322
Cite this article as: Abbott JC, Butcher SA: **Strategies towards sequencing complex crop genomes.** *Genome Biology* 2012, **13**:322.