2017-12-01

Assembly, Annotation and Optical Mapping of the A Subgenome of Avena

Rebekah Ann Lee

Assembly, Annotation and Optical Mapping of the

A Subgenome of *Avena*

Rebekah Ann Lee

A thesis submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Master of Science

Peter J. Maughan, Chair
Eric N. Jellen
David Jarvis

Department of Plant and Wildlife Sciences

Brigham Young University

ABSTRACT

Assembly, Annotation and Optical Mapping of the A Subgenome of *Avena*

Rebekah Ann Lee
Department of Plant and Wildlife Sciences, BYU
Master of Science

Common oat (*Avena*) has held a significant place within the global crop community for centuries; although its cultivation has decreased over the past century, its nutritional benefits have recently garnered increased interest for human consumption. No published reference sequences are available for any of the three oat subgenomes. Here we report a quality sequence assembly, annotation and hybrid optical map of the A-genome diploid *Avena atlantica* Baum and Fedak. The assembly is composed of a total of 3,417 contigs with an N_{50} of 11.86 Mb and an estimated completeness of 97.6%. This genome sequence will be a valuable research tool within the oat community.

Keywords: Avena, oats, annotation, genome sequence assembly, BioNano physical mapping

ACKNOWLEDGEMENTS

I would like to thank my committee chair, Dr. Peter J. Maughan for the support, guidance, aide, time and patience he has provided throughout this project. I am also especially grateful to Dr. Eric N. Jellen; without his encouragement I would not have had this wonderful opportunity. Thank you to Tim Langdon for providing seeds, Illumina sequence, RNA-Seq data and support for this project. Thank you to the members of the Josh Udall Lab for collaboration and advisement as I learned the BioNano protocols and bioinformatics. I would also like to acknowledge BYU, the Department of Plant and Wildlife Sciences and the National Science Foundation for generous financial support for the project. Lastly, I would like to express gratitude to my parents and husband for the patience and support they have provided throughout the length of this project.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

INTRODUCTION

Since early human history oats and other cereals have served as a major component of human diets and livestock consumption worldwide. Archeological evidence indicates that the ancestral forms of *Avena* originated in the Fertile Crescent and were first domesticated by humans in Europe during the late Bronze Age (Brouwer et al., 1999). Common oat (*Avena sativa* L., $2n = 6x = 42$, AACCDD genome) belongs to a polyploid complex of grasses (family Poaceae) native to the Mediterranean and Near East, with its center of greatest diversity along the Atlantic littoral of Northwest Africa. Besides *A. sativa*, other domesticated forms include the red oat (*A. byzantina* C. Koch, $2n = 6x = 42$, AACCDD); the endemic Ethiopian oat (*A. abyssinica* Hochst., $2n = 4x = 28$, AABB); and a complex of A_sA_s diploids including slender oat (*A. strigosa* Schreb.) and naked oat (*A. nuda* L.). Most oat varieties have a fibrous root system, upright stems (Fig. 1A) and subgenomes of 14 chromosomes (Fig. 1B). The flower head has multiple branches, or racemes, arranged in a panicle containing 20-150 spikelets; each containing two to five florets (Fig. 1C). Usually two kernels, or seeds, are produced per flower. These kernels are rolled or crushed to make oatmeal or ground to make oat flour that can be used to make oat cakes or bread. Presently, 75% of commercially grown oats are used for livestock feed (Brouwer et al., 1999; Cereal Disease Laboratory, 2015).

Nutritional Components

Although they have decreased in agricultural use, owing largely to the demise of the agrarian horse culture, relatively recent interest in the use of oats for human nutrition has led to an increase in their cultivation for food. The current global production of oats for trade year-end 2016 was 22,168 thousand metric tons. Within the United States there has also been an increase in both exports and imports. The imports for the U.S. went from 1,355 thousand metric tons for trade year 2012/2013 to 1,600 thousand metric tons for trade year 2015/2016. The same is true for U.S. exports, which increased from 18 to 25 thousand metric tons for trade years 2012/2013 and 2015/2016, respectively (United Stated Department of Agriculture, 2016).

Much of oat's acclaim to exceptional nutritional benefits comes from its status as a 'whole grain.' Whole grains have been defined by the U. S. Food and Drug Administration, the Whole Grains Council and other research organizations as limited processed grains with the whole seed's nutritional value and relative proportions of the endosperm, germ and bran preserved. Conserving the natural ratios found in these whole grains preserves their numerous nutritional benefits (Fardet et al., 2010; United States Department of Agriculture, 2016). Among the nutritional benefits of oats are a high level of dietary fiber, antioxidants and anti-carcinogenic qualities. Due to oat's low-glycemic nature, including it in a diet plan can lead to elevated carbohydrate tolerance, or in other words, a higher metabolic rate (Fardet et al., 2010).

Other health promoting compounds of whole grain oats include β-glucan and other soluble hemicellulose fibers, avenanthramides and saponins (Fardet et al., 2010; Peterson et al., 2001). β-glucans are mixed-linkage (1-3, 1-4) β-glucose polymers deposited in the walls of oat endosperm cells. They are known to lower cholesterol and slow gut nutrient absorption, increasing satiety and appetite control due to their characteristic hydration and viscosity properties (Coon, 2012; Rebello et al., 2001; Rebello et al., 2013). Oats are designated by the FDA as a qualified source of soluble fiber that can reduce the risk of heart disease (Andon et al., 2008; Jenkins et al., 2002; Queenan et al., 2007; U. S. Food and Drug Administration, 2016; Whitehead et al., 2014). Studies have also shown that these polysaccharides likely have a preventative effect on cancer (Shen et al., 2016) as well as immunostimulating effects

(Akramienė et al., 2007). Avenanthramides are phenols found in oats that are known antioxidants with anti-inflammatory effects (Yang et al., 2017). They can also provide protection from the formation of arterial fatty plaques by inhibiting the oxidation of low-density lipoproteins (Whitehead et al., 2014).

Due to the many nutritional benefits and their gluten free status, interest in oats as a food product is growing. Although barley has many of the same nutritional benefits, oats have less input requirements for cultivation than both barley and wheat. Oats also have a larger level of genetic diversity, due to their extensive cultivation throughout Europe and North America. This genetic diversity will be invaluable for breeders to further improve oat's nutrient levels, yield and increase their environmental range for cultivation (Brouwer et al., 1999).

Saponins

Saponins, which can be found throughout the oat plant, can be divided into two separate classes: avenacosides and avenacins. Avenacins, produced in the plant root, are antimicrobial triterpene glycoside compounds that provide resistance to a wide range of soil borne pathogens (Mylona et al., 2008). Avenacosides, produced in the plant leaves, are steroidal compounds that undergo a chemical conversion through cleavage by a specific glucose hydrolase into an antifungal compound (Wang et al., 2017). Both types of saponins are enclosed within the oat bran and are important components for disease resistance (Coon, 2012; Moses et al., 2014).

Oat Genome Composition

The hexaploid oats are now known to have arisen due to hybridization between a CCDD allotetraploid closely related to the modern *A. insularis* Ladiz. and an AA diploid (Yan et al., 2016). The A-genome group consists of diploids with the A_c, A_d, A_l, A_p, and A_s genome variants. The A-genome species harbor several genetic features of significance. Among these are a major crown rust resistance gene at the *Pc94* locus in *A. strigosa* (A_sA_s). Crown rust is caused by the basidiomycete fungus *Puccinia coronata*, that is found wherever oat species grow. *P. coronata* is an obligate biotrophic pathogen which must maintain a long term parasitic relationship with its host and can, under heavy infestation, reduce yield up to 20% (Klenová-Jiráková et al., 2010). Plants with a high coverage of rust pustules are prone to excessive water loss from hot dry winds leading to premature ripening, shriveled grain and decreased yield (Integrated Pest Management, 1989).

The A-genome is also part of a major intergenomic translocation (7C-17A). This relatively recent rearrangement in hexapliod oat has been associated with daylight sensitivity and winter hardiness – key elements in oat production. This translocation was probably key in the shift from Mediterranean winter ecology to Eurasian summers (Jellen et al., 2000). *Avena atlantica* Baum et Fedak (Baum and Fedak, 1985) includes genotypes that contain very high levels of groat soluble fiber and protein (Jellen et al., 2000; Welch et al., 2000). The 'groat', also called the caryopsis, is the part of the oat kernel remaining after hull removal and is considered the whole grain portion of the oat (Oats and Health, 2016).

The C-genome contains several genetic features of interest as well. The C-genome chromosomes have a high amount of diffuse heterochromatin along their entirety (Fominaya et al., 1988) This is not the case in other cereal grasses such as barley (*Hordeum vulgare* L.) and wheat (*Triticum* spp.). In these other cereal grasses, as well as the A and D diploid oat species, heterochromatin appears in localized and seemingly concentrated areas around the centromeres, at the telomeres and flanking secondary constrictions where rRNA genes are located. Why this heterochromatin pattern proliferated in the C-, but not the A- or D-genome diploids, is unknown.

The C-genome also contains alleles important to oat improvement. Among these is the C-genome segment at the terminus of the long arm of 21D carrying the putative *CSlF6c* locus that likely has a negative effect on seed soluble fiber content (Coon, 2012; Jellen et al., 1994). Linkage has also been demonstrated (2.1 cM in a wild X domesticated CD *Avena magna* F_2 population) between the chromosome 5C telomeric knob and co-segregating genes controlling awn production and basal abscission layer formation which have been implicated in the domestication of oats (Oliver et al., 2011).

The D-genome, found in tetraploid and hexaploid oat species, is considered to be more homologous with the A-genome (Jellen et al., 1994). Studies using direct hybridization of D-genome specific probes have been unable to identify an extant D-genome progenitor. The inability to identify extant D diploid genome progenitor suggests that the original D-genome progenitor may be extinct (Loskutov et al., 2008). Genetic and cytogenetic data suggests that an unknown D diploid likely hybridized with an AC tetraploid to form *Avena sativa*, hexaploid oats ($2n = 6x = 42$, ACD) (Loskutov et al., 2008; Rajhathy et al., 1959).

Although several recent publications have reported high-density linkage maps for avena species (Oliver et al., 2013; Chaffin et al., 2016), there are no quality reference sequences for any of the *Avena* species or subgenomes. The objective of this research was to produce a high-quality, annotated genome assembly of the A-subgenome of *Avena* using PacBio-based single-molecule sequencing, deep transcriptome RNA sequencing and BioNano optical mapping. Our results should provide the basis for future gene discovery experiments as well as shed insight into the evolution of the *Avena* genus.

MATERIALS AND METHODS

Plant Material and DNA Extraction

A. atlantica seeds were kindly provided by Tim Langdon (Aberystwyth University, United Kingdom). Seed was increased in the greenhouses at Brigham Young University (Provo, UT) using Sunshine Mix II (Sun Gro, Bellevue, WA, USA) supplemented with Osmocote fertilizer (Scotts, Marysville, OH, USA) and maintained at 25 °C under broad-spectrum halogen lamps with a 12-h photoperiod. Young leaf tissue was harvested for DNA extraction when plants were between six and ten centimeters in height (~14-21 days post emergence).

Single-Molecule PacBio Sequencing

Plant material was flash frozen and sent to the Arizona Genomics Institute (AGI; Tucson, Arizona) for DNA extraction, library prep and quality control. Large insert (20 kb) libraries were size selected using the BluePippin™ System. Libraries were sequenced using either the RS II sequencer (Pacific Biosciences; Menlo Park, CA) at RTL genomics (Lubbock, TX) or the Sequel sequencer (Pacific Biosciences; Menlo Park, CA) at the DNA Sequencing Center (Brigham Young University, Provo, UT) using P6C4 chemistry.

Genome Assembly and Polishing

Three PacBio-based assemblies were created, including a 30x coverage (Sequel data only), 63x coverage (Sequel data only), and 83x coverage (Sequel and RS II data). Data in all three assemblies was corrected, trimmed and assembled using the program Canu version 1.6 (Released August 2017) using default parameters (specifically, corMhapSensitivity=normal and corOutCoverage=40).

The assemblies were polished by the computer program PILON (v1.22) using Illumina short reads sequenced by BGI Genomics (Shenzhen, Guangdong, China). The short reads came from 500bp insert libraries which were paired-end sequenced (2 X 150 bp). To supplement the short-read sequence coverage, an additional set of Illumina paired-end reads (2 X 100 bp) was obtained from Tim Langdon (Aberystwyth University, Ceredigion, UK) which increased the Illumina short read coverage to 49x (Table 1).

K-mer Analysis

Because of differing genome sizes published for the A genome, a k-mer analysis was used to estimate the size of the *A. atlantica* genome using the program GenomeScope (Vurture et al., 2017). The *k-mer* profile measures how often substrings of length *k* (k-mers) occur in the Illumina sequencing reads as computed by the program Jellyfish (Marcais and Kingsford, 2011). The genome size is then estimated by fitting a mixture model of four evenly spaced negative binomial distributions to the k-mer profile. The final set of parameters removes sequencing errors and higher copy repeats (i.e., ex-nuclear genomes and/or contaminant) and estimates the total genome size by normalizing the observed k-mer frequencies to the average coverage value (Vurture et al., 2017). Genome size estimates were performed for k-mer lengths of 19, 21 and 23.

Repeat Analysis

Repeat analysis was conducted with RepeatModeler v1.0.8 and RepeatMasker v4.0.5 relative to RepBase libraries (20140131; www.girinst.org). RepeatModuler consists of two main subprograms: RECON v1.08 and RepeatScout v1.0.5, that work to find novel repeats in the input genome; that are then characterized using a perl tool created by Bailly-Bechet et al., 2014. A k-mer analysis also provides an estimation of repeat content. RepeatMasker was then used to quantify and classify the RepeatModeler output.

Transcriptome Assembly

A *de novo* transcriptome was assembled from RNA-Seq data kindly provided by Tim Langdon (Aberystwyth University, Ceredigion, UK). The RNA-Seq data consisted of 100 bp paired-end Illumina reads derived from 11 different plant tissue types: stem, mature leaf, stressed mature leaf, seed (2 days old), hypocotyl (4-5 day old), root (4-5 days old), vegetative meristem, green grain, yellow grain, young flower (meiotic) and green anthers. The reads were trimmed using the computer program Trimmomatic-0.35 (Bolger et al., 2014) and assembled and mapped back to the 83x polished reference assembly using Tophat2 v.2.1.1 (Kim et al., 2013) and Cufflinks v.2.2.1 (Trapnell et al., 2010). The quality of the assembled transcriptome was assessed relative to completeness using BLAST comparisons to the reference *brachypodium distachyon* L. (P. Beauv et al., protein data set (ftp://ftp.ensemblgenomes.org/pub/plants/release-37/fasta/brachypodium_distachyon/pep/).

Genome Completeness and Annotation

The polished genome was also run through the BUSCO pipeline v3.0.2 using the flowering plant (embryophyta_odb9) orthologous gene data set and the --long argument. BUSCO tests for conserved orthologous genes (COGs) expected to be found in all flowering plants and is a widely accepted assessment of genome completeness (Simão et al., 2015). The results from this program also make an AUGUSTUS (Stanke et al., 2006) program training set specific to *A. atlantica* which was used for downstream annotation of the assembled genome.

The program MAKER2 (Cantarel et al., 2008) was used to annotate the polished genome. EST evidence for annotation used by the MAKER pipeline included the *A. atlantica* de novo transcriptome (described above) and the cDNA gene models from *B. distachyon* L. (v1.0; (ftp://ftp.ensemblgenomes.org/pub/plants/release-37/fasta/brachypodium_distachyon/cdna/; downloaded 9/13/17). Protein evidence included the uniprot_sprot database (ftp://ftp.ebi.ac.uk/pub/databases/uniprot/knowledgebase/; downloaded 9/13/17) as well as the peptide models from *B. distachyon* (v1.0; ftp://ftp.ensemblgenomes.org/pub/plants/release-37/fasta/brachypodium_distachyon/pep/; downloaded 9/13/17). For repeat masking, MAKER2 was set to Viridiplantae and was given a consensi.fa.classified file, specific for *A. atlantica,* developed by RepeatModuler as well as a te_proteins.fasta file. For *ab initio* gene prediction, an *A. atlantica* specific AUGUSTUS gene prediction model and a rice (*Oryza sativa*) based SNAP model were provided.

Bionano Genomics Optical Map

For hybrid assembly, a *de novo* physical map was developed using BioNano Genomics optical mapping technology (BNG; San Diego, CA). In short, genomic DNA was extracted from one gram of fresh plant leaf material according to the protocols provided by BNG. High molecular DNA was quantified using a Qubit Assay (ThermoFisher Scientific; Waltham, MA) and labeled in accordance with the IrysPrep Reagent Kit (BNG, kit #RE-011-10). The process included four basic steps: Nick, Label, Repair and Stain. Nicking was accomplished by the single-stranded nicking endonuclease Nb.BssSI. Nb.BssSI has a six base pair recognition site which produced an average 8-18 labels per 100 Kb. After DNA nicking, the molecules were labeled with a flourescent-dUTP nucleotide analog by means of *Taq* polymerase. *Taq* DNA ligase was then used to repair the nicked and labeled regions. lastly, the DNA backbone was stained with an intercalating dye and run through a nanofluidic chip where the labeled DNA was linearized and imaged (Fig. 2) using the BNG Irys system.

RESULTS AND DISCUSSION

Genome Size and Heterozygosity

K-mer analysis was conducted using 191 Gb of Illumina short reads to determine the expected genome size for *A. atlantica*. The resulting size estimates were as follows: 3.790 Gb, 3.784 Gb and 3.779 Gb for kmer lengths k=19, k=21 and k=23 respectively, with an average estimate of 3.784 Gb (Fig. 3, Table 2) which is similar to the estimated size provided by Bennett (1976) for *Avena strigosa* (3.912 Gb), a closely related A-genome diploid oat. The slight decrease in predicted genome size is likely a reflection of the high level of repetitive elements found within the *A. atlantica* genome which confounds k-mer based analyses due to confounding effects associated with contaminating sequences that are removed by the model (e.g., extranuclear genomes and/or bacterial contamination). Indeed, high repetitive fractions were observed in both the k-mer analysis as well as the results from the RepeatModeler pipeline (described below). At k = 19, k=21 and k=23, 80.7%, 78.0%, and 75.8%, of the *A. atlantica* genome, respectively, was estimated to be repetitive sequence (Table 2). The average estimate of heterozygosity per k-mer profile was 0.074% which relatively quite low and is reflective of a predominantly inbreeding species (Table 2). Given the similarities between the published *A. strigosa* and the predicted *A. atlantica* genome size using k-mer analyses we use a base genome size of 3.9 Gb for all subsequent analyses.

Whole Genome Assembly

A total of 31,544,396 PacBio reads were generated across 122 cells, including 40 cells generated on the Sequel instrument and 82 cells on the RSII instruments. As expected, the output per cell was significantly greater for the Sequel instrument relative to the RSII (6.16 Gb v. 1.82 Gb). The longest reads came from the Sequel instrument (max length = 194,884 bp). The total length of all reads summed to 325,888,096,473 bases with an N_{50} read length of 18,658 bp, which represents ~83x coverage of the predicted *A. atlantica* genome (Table 1).

The *A. atlantica* genome was assembled using the canu assembler which specializes in assembling PacBio data, with minimum genome coverage recommendations > 20x and ideally between 30x and 60x. Given the high repetitive fraction of the *A. atlantica* genome, we evaluated the effect of coverage on the assembly process using three differing levels of read coverage: 30x coverage (Sequel data only), 63x coverage (Sequel data only), and 83x coverage (Sequel and RSII data, Table 1). All assemblies were polished using Illumina short read sequences resulting in small changes, mostly indels, to the three raw assemblies. As expected, the increase in coverage from 30x to 63x resulted in a substantial decrease in the overall number of contigs, (21,329 to 4,616) with a simultaneous increase in N_{50} (Fig. 4A) from 204,301 bp to 3,955,572 bp for the respective assemblies (Table 3). The total lengths of the assemblies were also substantially different, with the 30x coverage genome spanning a total of 3.16 Gb and the 63x coverage based genome spanning 3.66 Gb, a 502 Mb increase in total genome length. The increases were less pronounced when comparing the 63x assembly to the 83x; however, they were still notable. The overall genome assembly increased by about 20 kb to 3.684 Gb but contiguity of the genome, as measured by the N_{50}, increased substantially from 3.955 Mb to 5.55 Mb (Table 3). The cumulative genome length and N_{50} (Fig. 4B) from the different coverage assemblies are instructive (although not unexpected) as they confirm the recommendations by the developers of the Canu assembler that the most contiguous and complete assemblies, even with PacBio long reads, are obtained at coverage depths greater than 50x.

Optical Map Assembly and Hybrid Scaffolding

An optical map of the *A. atlantica* genome was constructed using the BNG Irys platform. A total of 807 Gb of data were produced equaling ~218x coverage of the *A. atlantica* genome. After filtering out low-quality single molecules, a total of 538 Gb of data were included in the final *de novo* physical map assembly. The resulting physical map assembly consisted of 6,707 individual genome maps that spanned 3.362 Gb (86.2% of the predicted genome size), with an N_{50} of 0.629 Mb. To make a hybrid assembly, all next generation sequencing (NGS) contigs for the polished 83x genome assembly greater than 20 Kb with a minimum of five label sites were aligned to the physical map assembly using IrysView (Fig. 5). Of the 6,707 individual physical maps, 6,648 (99%) aligned to the NGS reference assembly, with a unique alignment length of 3.273 Gb, or ~90% of the reference NGS assembly (Table 4). The strong congruence between physical and NGS assemblies is indicative of a high-quality NGS assembly. A total of 1,136 NGS contigs were collapsed into 612 hybrid super-scaffolds, producing a final hybrid assembly consisting of 3,417 scaffolds: 612 hybrid super-scaffolds and 2,805 PacBio contigs (Table 5). Of the 612 hybrid super-scaffolds, 340 (55.6%) were simple 5' or 3' extension of a single NGS contig, whereas 272 (44.4%) joined and oriented two or more NGS contigs. The largest number of NGS contigs joined were in super-scaffold_232 (18.1 Mb) which consisted of 9 NGS contigs. The largest super-scaffold (super-scaffold_193) consisted of 7 NGS contigs and spanned 44.1 Mb. The hybrid assembly spanned a total of 3.70 Gb (94.8% of the predicted genome size) with a substantial increase in the N_{50} value – increasing from 5.5 Mb (83x NGS only assembly) to

11.86 Mb for the hybrid assembly. As a consequence of the hybrid assembly, the %N in the assembly increased from zero to 450 per 100 kb. This was not unexpected, as the hybrid assembly uses non-sequence based physical restriction maps (called molecular maps) to join NGS contigs. Figure 4 shows a QUAST analysis (Quality Assessment Tool; Gurevich, 2013) of the three NGS and the hybrid assembly. The hybrid assembly consisted of a GC content of 44.4% (Fig. 4C) which is similar to the GC content reported for several species within Poaceae, including *B. distachyon* (46.2%), *Oryza sativa* (43.5%) and the more distantly related Sorghum bicolor (43.9%; Singh et al., 2016) and follows the general paradigm that GC content is highest in the grasses followed by the non-grass monocots and then the dicots.

Repetitive Elements

RepeatModeler and RepeatMasker were used to identify and classify the total genomic repeat content of the *A. atlantica* hybrid assembly. Similar, but slightly higher than the k-mer analysis (see above), RepeatModeler estimated the total interspersed repeat fraction of the *A. atlantica* genome to be 82.3% (3.01 Gb). Of the total interspersed repeat fraction, 58.2% and 20.9% were classified as Gypsy and Copia elements, respectively (Fig. 6), both of which are non-long terminal repeat retrotransposons found widely throughout eukaryotic genomes, and particularly in the genomes of grass species. Approximately 12% of the identified repeat elements were categorized as unknown. Given the extensive investigations of repeat elements in the grasses (Bilinski et al., 2017; Feschotte et al., 2003; Minaya et al., 2013), this unknown fraction may represent repeat elements unique to *Avena* (Fig. 6, Table 6) and could be invaluable in distinguishing the closely related A and D subgenomes. In addition to the interspersed repeat elements identified, 1.5% of the genome was classified as low complexity (0.03%), satellite (0.52%), telomeric repeat (0.68%) or microsatellite (0.29%), with the most common di-, tri- and tetranucleotide repeat microsatellite motif identified being (AT)n, (AAC)n, and (TTTA)n, respectively (Table 6). To date, no microsatellites have been derived from an A genome oat diploid – thus these new putative microsatellite loci represent important genetic tools for studying diversity in the A-genome diploids (*A. atlantica*; *A. strigosa*, *A. longiglumis* Durieu, etc.). Repeat-sequence content is known to correlate with genome-size. Indeed, within published plant genomes, repeat content varies widely, ranging from 3% for the minute 82 Mb genome of *Utricularia gibba* L. (Ibarra-Laclette et al., 2013) to 85% for maize (Schnable et al., 2009). Thus, given the large size of the *A. atlantica* genome (3.9 Gb) it is not surprising that only ~20% of the genome is classified as non-repetitive. Repeat content is believed to be an important driver of genome organization and evolution (Michael et al., 2014), thus understanding the repetitive content of the *A. atlantica* genome will be undoubtedly important for understanding the overall evolution of modern day hexaploid oats.

Transcriptome Assembly

The *A. atlantica* genome was annotated using a *de novo* assembled transcriptome based on 11 different plant tissue types using 115 million paired-end reads (PE 2 X 101 bp; ~23 Gb). The transcriptome was assembled using Tophat2, which is a splice-aware mapper that uses Bowtie2 (Langmead et al., 2009) to map RNA-Seq reads to a sequence reference assembly. Overall, 96.2% of the reads were successfully mapped to the finished reference genome, of which 94.2% were in aligned pairs, with 93.1% of paired reads aligning concordantly. The high mapping rate and concordance among paired reads is indicative of a high-quality genome assembly, while the low percentage (5.9%) of multiple alignments for the mapped reads was expected for a diploid species. We then used Cufflinks to assemble 51,222 transcripts, including 12,288 isoforms, from the mapped reads. The mean transcript length was 1,888 bp with an average GC% of 50.3%. The increase in GC% within coding regions is a well-known phenomenon and is hypothesized to be the result of GC-biased gene conversion – a process by which the GC content of DNA increases due to gene conversion during recombination (Duret and Galtier, 2009).

To evaluate the quality of the *A. atlantica* transcripts, we used BLASTX to query the transcripts against the reported peptide sequences from *B. distachyon* (a related model grass species) and identified 35,370 (69.0%) transcripts with e-value hits < 1e^{-20} (Fig. 7A). When the length of the predicted peptides (based on the assembled transcripts) were compared to the lengths of putative orthlogs in *B. distachyon*, 86.2% of the *A. atlantica* transcripts covered > 70% of the ortholog length identified in *B. distrachyon* (Fig. 7B, Table 7). BUSCO was used to assess the completeness of the transcriptome. BUSCO uses a set of 1,440 conserved orthologous genes (COGs) found in a wide range of plant taxa. Of the 1,440 COGs, 1349 (93.7%) were identified in the transcriptome, of which 1096 (76.1%) were identified as complete (alignment > 70% of their sequence) and single copy, 253 (17.6%) were identified as complete and duplicated while 62 (4.3%) identified as fragmented (C:93.7% [S:76.1%, D:17.6%], F:4.3%, M:2.0%, n:1440; Table 7).

Genome Annotation

MAKER was used to annotate the *A. atlantica* genome. The MAKER pipeline annotated 47,070 gene models spanning 65.8 Mb (~1.7% of the total genome size) with a mean and median transcript length of 1,398 bp and 1,137 bp, respectively. Genome annotation quality was then assessed by calculating the Annotation Edit Distance (AED) for each model. AED is a measure of sensitivity, accuracy and specificity (Eilbeck et al., 2009). Greater than 73% of the annotated *A. atlantica* genome had an AED value < 0.25, which is similar to the benchmark gold standard for maize chromosome 4 (Holt and Yandell, 2011) and indicative of high-quality annotation (Fig. 8). A BUSCO analysis of the final genome assembly identified 1,393 (96.7%) complete COGs, of which 1,354 (94.0%) were complete and single-copy and 39 (2.7%) were identified as complete and duplicated. An additional 12 (0.8%) COGs were identified as fragmented (C:96.7% [S:94.0%, D:2.7%], F:0.8%, M:2.5%, n:1440; Table 8). Thus 44 more complete COGs were identified in the assembled genome than in the transcriptome (described above) which further highlights the completeness and quality of the assembled NGS genome. The increased completeness is likely a reflection of the different read technologies used to develop the transcriptome (Illumina short read) and the genome (PacBio long read). A BLASTP search of the annotated gene models produced by the MAKER pipeline against the uniprot-sprot and *B. distrachyon* protein data set identified 29,760 (63%; e-value <1e-6) and 34,852 (74%; e-value <1e-20) hits, respectively. When the lengths of the MAKER predicted peptides were compared to the lengths of their putative orthologs in *B. distachyon*, 88.6% of the *A. atlantica* transcripts covered > 70% of the ortholog identified in *B. distrachyon* (Table 7).

Synteny and Sequence Comparison

Synteny between the *A. atlantica* genome and the recently published barley reference genome (2n = 14; Von Wettstein-Knowles et al., 1990) was analyzed using the *SynMap* tool on the CoGe platform (genomevolution.org/coge). The synteny was analyzed for coding sequences (CDS) regions in barley genome relative to the *A. atlantica* genomic sequence. The shared ancestry between the two grass species can be seen in the significant level of synteny observed across all seven barley chromosomes (Fig. 9). Interestingly, large non-syntenic blocks in each of the barley chromosomes was identified – which likely correspond to centromeric regions of the barley chromosome that are known to be devoid of coding sequences and thus share no homology to with the *A. atlantica* scaffolds.

CONCLUSION

The *A. atlantica* hybrid optical map and sequence assembly reported in this paper is composed of a total of 3,417 contigs with an N_{50} of 11.86 Mb. Annotation of the sequence assembly shows a high level of completeness, as 97.6% of conserved orthologs were assigned predicted locations. This is the first reported reference assembly for any of the three oat subgenomes. Analysis of the genome assembly classified much of the genome as repetitive sequence (~80 – 83%) and confirmed its expected large size (~3.9 Gb). Annotation of the genome, using a deeply sequenced transcriptome identified 51,222 gene models. This sequence assembly expands the available genomic resources for the oat research community and has the potential to positively impact *Avena* research in numerous areas, particularly in the identification of resistance genes as well as the development of complete biosynthethic pathways involved in nutritionally favorable traits (e.g., β-glucan and avenanthramides), which are often complex and need complete genome assembly to be properly deciphered. This genome assembly will also provide a unique tool for the eventual assembly of the domesticated hexaploid oat genome, as it will help resolve homoelogous relationships among the hexaploid subgenome.

On-going efforts to improve this assembly are being made using Hi-C (chromatin contact maps) technology. Hi-C methods detect *in vivo* chromatin interactions through crosslinking to determine physical closeness of DNA fragments. Such information can be used to statistically cluster, order and orient scaffolds into pseudo chromosomes (Lightfoot et al., 2017). These efforts will likely be complicated due to the large, highly repetitive fraction of the genome.

Additionally, we are making efforts to construct a high-density linkage map based on genotyping by sequencing between A-genome species which would provide a backbone for clustering and ordering our NGS scaffolds into a chromosome based assembly.

LITERATURE CITED

Akramienė D, Kondrotas A, Didziapetriene J, Kevelaitis E (2007) Effects of Beta-glucans on the Immune System. Medicina (Kaunas) 43:597-606

Andon MB and Anderson JW (2008) State of the Art Reviews: The Oatmeal-Cholesterol Connection: 10 Years Later. American Journal of Lifestyle Medicine 2:51-57

Bailly-Bechet M, Haudry A and Lerat E (2014) "One code to find them all": A perl tool to conveniently parse RepeatMasker output files. Mob. DNA 5:13. doi:10.1186/1759-8753-5-13

Baum BR and Fedak G (1985) *Avena atlantica*, a new diploid species of the oat genus from Morocco. Can. J. Bot. 63:1057-1060

Bennett MD, Smith JB (1976) Nuclear DNA amounts in angiosperms. Philosophical Transactions of the Royal Society of London Series B-Biological Sciences 274: 227-274

Bilinski P, Han Y, Hufford MB, Lorant A, Zhang P, Estep MC, Jiang J, Ross-Ibarra J (2017) Genomic abundance is not predictive of tandem repeat localization in grass genomes. PLOS ONE 12(6):e0177896. doi:10.1371/journal.pone.0177896

BioNano Genomics. (2016) Irys Technology. bionanogenomics.com

Bolger AM, Lohse M and Usadel B (2014) Trimmomatic: A Flexible Trimmer for Illumina Sequence Data. Bioinformatics 30:2114–2120. doi: 10.1093/bioinformatics/btu170

Brouwer J and Flood RG (1999) The Oat Crop: World Crop Series, Aspects of Oat Physiology 177-222. (ISBN 0412373106)

Cantarel BL, Korf I, Robb SMC, et al. (2008) MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Research 18:188-196. doi:10.1101/gr.6743907

Cereal Disease Laboratory (2015) Unites States Department of Agriculture | Agricultural Research Service. Retrieved 19 October 2015

Chaffin A, Huang Y, Smith S, Bekele W, Babiker E, Gnanesh B, Foresman B, Blanchard S, Jay J, Reid R, Wight C, Chao S, Oliver R, Islamovic E, Kolb F, McCartney C, Fetch JM, Beattie A, Bjornstad A, Bonman J, Langdon T, Howarth C, Brouwer C, Jellen E, Klos KE, Poland J, Hsieh T-F, Brown R, Jackson E, Schlueter J, Tinker N (2016) A consensus map in cultivated hexaploid oat reveals conserved grass synteny with substantial sub-genome rearrangement. The Plant Genome 9:1-21

Clouse JW, Adhikary D, Page JT, Ramaraj T, Deyholos MK, Udall JA, Fairbanks DJ, Jellen EN, Maughan PJ (2016) The Amaranth Genome: Genome, Transcriptome, and Physical Map Assembly. The Plant Genome doi: 10.3835/plantgenome2015.07.0062

Coon, Melissa A. (2012) Characterization and Variable Expression of the CslF6 Homologs in Oat (Avena sp.). All Theses and Dissertations: 3750. https://scholarsarchive.byu.edu/etd/3750

Doležel J, Greilhuber J, Lucretti S, Meister A, Lysák MA, Nardi L, Obermayer R (1998) Plant genome size estimation by flow cytometry: Inter laboratory comparisons. Ann Bot 82:17– 26. doi:10.1093/oxfordjournals.aob.a010312

Duret L and Galtier N (2009) Biased gene conversion and the evolution of mammalian genomic Landscapes. Annu Rev Genomics Hum Genet 10:285–31

Emmons CL and Peterson DM (2001) Antioxidant Activity and Phenolic Content of Oat as Affected by Cultivar and Location. Journal of Cereal Science 33: 115-129

Fardet A (2010) New hypotheses for the health-protective mechanisms of whole-grain cereals: what is beyond fibre? Nutrition Research Reviews 23:65–134

Feschotte C, Swamy L, Wessler SR (2003) Genome-wide analysis of mariner-like transposable elements in rice reveals complex relationships with stowaway miniature inverted repeat transposable elements (MITEs). Genetics 163:747-58

Fominaya A, Vega C and Ferrer E (1988) Giemsa C-banded karyotypes of *Avena* species. *Genome* 30:627-632

Gurevich A, Saveliev V, Vyahhi N, and Tesler G (2013) QUAST: quality assessment tool for genome assemblies. Bioinformatics 29:1072-1075

Ibarra-Laclette E, Lyons E, Hernández-Guzmán G, Pérez-Torres CA, Carretero-Paulet L, Chang TH and Herrera-Estrella L (2013) Architecture and evolution of a minute plant genome. Nature 498:94–98. doi:10.1038/nature12132

Integrated Pest Management (1989) RPD No. 109 Crown Rust of Oats. Reports on Plant Diseases. ipm.illinois.edu/diseases/series100/rpd109/

Jellen EN and Beard J (2000) Geographical Distribution of a Chromosome 7C and 17 Intergenomic Translocation in Cultivated Oat. Crop Science 40:256-263

Jellen EN, Gill BS and Cox TS (1994) Genomic in situ hybridization differentiates between A/D- and C-genome chromatin and detects intergenomic translocations in polyploid oat species (Genus Avena). Genome 37:613– 618

Jenkins DJ, Kendall CW, Vuksan V, Vidgen E, Parker T, Faulkner D, Mehling CC, Garsetti M, Testolin G, Cunnane SC, Ryan MA, Corey PN (2002) Soluble fiber intake at a dose approved by the US Food and Drug Administration for a claim of health benefits: serum lipid risk factors for cardiovascular disease assessed in a randomized controlled crossover trial. Am J Clin Nutr 75:834-839

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biology 14:R36. doi:10.1186/gb-2013-14-4-r36

Klenová-Jiráková H, Leisova-Svobodova L, Hanzalova A, Kucera L (2010) Diversity of Oat Crown Rust (Puccinia coronata f.sp. avenae) Isolates Detected by Virulence and AFLP Analyses. Plant Protect Sci 46:98–106

Koren S, Walenz BP, Berlin K, Miller JR, Phillippy AM (2017) Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res doi:10.1101/gr.215087.116

Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology 10:25. doi:10.1186/gb-2009-10-3-r25

Li F, Fan G, Wang K, Sun F, Yuan Y, Song G, Li Q, Ma Z (2014) Genome sequence of the cultivated cotton Gossypium arboreum. Nature Genetics 46:567–572. doi:10.1038/ng.2987

Lightfoot DJ, Jarvis DE, Ramaraj T, Lee R, Jellen EN, Maughan PJ (2017) Single molecule sequencing and Hi-C based proximity-guided assembly of amaranth (Amaranthus hypochondriacus) chromosomes provides insights into genome evolution. BMC Biology 15:74. doi:10.1186/s12915-017-0412-4

Loskutov IG (2008) On evolutionary pathways of Avena species. Genet Resour Crop Evol 55:211–220

Luo X, Zhang H, Kang H, Fan X, Wang Y, Sha L, Zhou Y (2014) Exploring the Origin of the D Genome of Oat by Fluorescence in Situ Hybridization. Genome 57:469-472

Marcais G and Kingsford C (2011) A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics 27: 764-770. doi:10.1093/bioinformatics/btr011

Manly BF, McDonald L, Thomas DL, McDonald TL, Erickson W (2002) Resource selection by animals: statistical design and analysis for field studies. Boston, Massachusetts, USA: Kluwer Academic. (ISBN 978-0-306-48151-2)

Michael TP (2014) Plant genome size variation: Bloating and purging DNA. Briefings Funct. Genomics 13:308–317. doi:10.1093/bfgp/elu005

Mikheenko A, Valin G, Prjibelski A, Saveliev V, Gurevich A (2016) Icarus: visualizer for *de novo* assembly evaluation. Bioinformatics 32:3321-3323

Minaya M, Pimentel M, Mason-Gamer R, Catalan P (2013) Distribution and evolutionary dynamics of Stowaway Miniature Inverted repeat Transposable Elements (MITEs) in grasses. Mol Phylogenet Evol 68:106-18. doi: 10.1016/j.ympev.2013.03.005

Moses T, Papadopoulou KK, Osbourn A (2014) Metabolic and functional diversity of saponins, biosynthetic intermediates and semi-synthetic derivatives. Critical Reviews in Biochemistry and Molecular Biology 49:439-462. doi: 10.3109/10409238.2014.953628

Mylona P, Owatworakit A, Papadopoulou K, Jenner H, Qin B, Findlay K, Hill L, Qi X, Bakht S, Melton R, Osbourn A (2008) Sad3 and Sad4 Are Required for Saponin Biosynthesis and Root Development in Oat. The Plant Cell 20, 201–212

Nishiyama T (2016) K-mer Analysis and Genome size Estimate. http://koke.asrc.kanazawa-u.ac.jp/HOWTO/kmer-genomesize.html

Oats and Health (2016). www.oatsandhealth.org

Oliver RE, Jellen EN, Ladizinsky G, Korol AB, Kilian A, Beard JL, Dumlupinar Z, Wisniewski-Morehead NH, Svedin E, Coon M, Redman RR, Maughan PJ, Obert DE, Jackson EW (2011) New Diversity Arrays Technology (DArT) markers for tetraploid oat

(Avena magna Murphy et Terrell) provide the first complete oat linkage map and markers linked to domestication genes from hexaploid A. sativa L. Theor Appl Genet 123:1159-1171

Oliver RE, Tinker NA, Lazo GR, Chao S, Jellen EN, Carson M, Rines HW, Obert DE, Lutz JD, Shackelford I, Korol A, Wight CP, Gardner KM, Hattori J, Beattie AD, Bjørnstad Å, Bonman JM, Jannink J-L, Sorrells M, Brown-Guedira GL, Mitchell Fetch JW, Harrison SA, Howarth CJ, Ibrahim A, Kolb FL, McMullen MS, Murphy JP, Ohm HW, Rossnagel BG, Yan W, Miclaus KJ, Hiller J, Maughan PJ, Redman Hultz RR, Anderson JM, Islamovic E, Jackson EW (2013) SNP discovery and chromosome anchoring provide the first physically-anchored hexaploid oat map and reveal synteny with model species. PLoS One 8:1-12

PacBio (2016) Comprehensive View of Maize Genome Reveals Regulatory and Structural Mechanisms. www.pacb.com/agbio PN: CS114-010516

Pacific Biosciences (n.d.) Retrieved June 20, 2016, from http://www.pacb.com/

Queenan KM Stewart ML, Smith KN, Thomas W, Fulcher RG, Slavin JL (2007) Concentrated oat β-glucan, a fermentable fiber, lowers serum cholesterol in hypercholesterolemic adults in a randomized controlled trial. Nutr J 6:6. doi:10.1186/1475-2891-6-6

Rajhathy T and Morrison JW (1959) Chromosome morphology in the genus Avena. Can J Bot 37: 331–337

RBG Kew DNA C-values query results (2001) data.kew.org

Rebello CJ, Johnson WD, Martin CK, Han H, Chu Y, Bordenave N, van Klinken BJ, O'Shea M, Greenway FL (2016) Instant Oatmeal Increases Satiety and Reduces Energy Intake Compared to a Ready-to-Eat Oat-Based Breakfast Cereal: A Randomized Crossover

Trial. J Am Coll Nutr 35:41-9

Rebello C J, Johnson WD, Martin CK, Xie W, O'Shea M, Kurilich A, Bordenave N, Andler S, van Klinken BJ, Chu YF, Greenway FL (2013) Acute effect of oatmeal on subjective measures of appetite and Satiety compared to a ready-to-eat breakfast cereal: a randomized crossover trial. J Am Coll Nutr 32:272-9

Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Cordes M, et al. (2009) The B73 maize genome: Complexity, diversity, and dynamics. Science 326:1112–1115. doi:10.1126/science.1178534

Shen R, Wang Z, Dong J, Xiang Q, Liu Y (2016) Effects of oat soluble and insoluble β-glucan on 1,2-dimethylhydrazine-induced early colon carcinogenesis in mice. Food and Agricultural Immunology 27:657-666

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31:3210-3212. doi: 10.1093/bioinformatics/btv351

Singh R, Ming R and Yu Q (2016) Comparative Analysis of GC Content Variations in Plant Genomes. Tropical Plant Biol 9:136-149

Smit AF, Hubley R (2008) *RepeatModeler Open-1.0*. http://www.repeatmasker.org

Songa G, Huoa P, Wua B, Zhangab Z (2015) A genetic linkage map of hexaploid naked oat constructed with SSR markers. The Crop Journal 3:353-357

Stanke M, Steinkamp R, Waack S, and Morgenstern B (2004) AUGUSTUS: a web server for gene finding in eukaryotes. Nucleic Acids Research 32:309-312

Stanke M, Tzvetkova A and Morgenstern B (2006) AUGUSTUS at EGASP: using EST, protein and genomic alignments for improved gene prediction in the human genome. Genome Biology doi:10.1186/gb-2006-7-s1-s11

Trapnell C, Williams B, Pertea G, Mortazavi A, Kwan G, Baren J, Salzberg S, Wold B, Pachter L (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature Biotechnology doi:10.1038/nbt.1621

United States Department of Agriculture (2016) Grain: World Markets and Trade. Retrieved June 2016. www.fas.usda.gov/data

Von Wettstein-Knowles P. (1990). Locus maps of complex organisms, Edition: 5, Barley (Hordeum vulgare) 2N = 14. Cold Spring Harbor Laboratory Press 6:125 - 134

Vurture GW, Sedlazeck FJ, Nattestad M, Underwood CJ, Fang H, Gurtowski J, Schatz MC (2017) GenomeScope: fast reference-free Genome profiling from short reads. Bioinformatics 33:2202–2204. https://doi.org/10.1093/bioinformatics/btx153

Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, Earl AM (2014) Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. PLoS ONE 9:e112963. doi:10.1371/journal.pone.0112963

Wang P, Yang J, Yerke A, Sang S. (2017) Avenacosides: Metabolism, and potential use as exposure biomarkers of oat intake. Mol Nutr Food Res 61. doi:10.1002/mnfr.201700196

Welch RW, Brown JCW and Leggett JM (2000) Interspecific and Intraspecific Variation in Grain and Groat Characteristics of Wild Oat (Avena) Species: Very High Groat (1→3),(1→4)-β- D -glucan in an Avena atlantica Genotype J. Cereal Sci 31:273-279

Whitehead A, Beck EJ, Tosh S, Wolever TM (2014) Cholesterol-lowering effects of oat β-glucan: a meta-analysis of randomized controlled trials. Am J Clin Nutr 100:1413-1421

Yan H, Martin SL, Bekele WA, Latta RG, Diederichsen A, Peng Y, Tinker NA (2016) Genome Size Variation in the Genus Avena. Genome 59:209-220. https://doi.org/10.1139/gen-2015-013

Yang J, Ou B, Wise ML, Chu Y (2014) In vitro total antioxidant capacity and anti-inflammatory activity of three common oat-derived avenanthramides. Food Chemistry 160:338–345

FIGURES

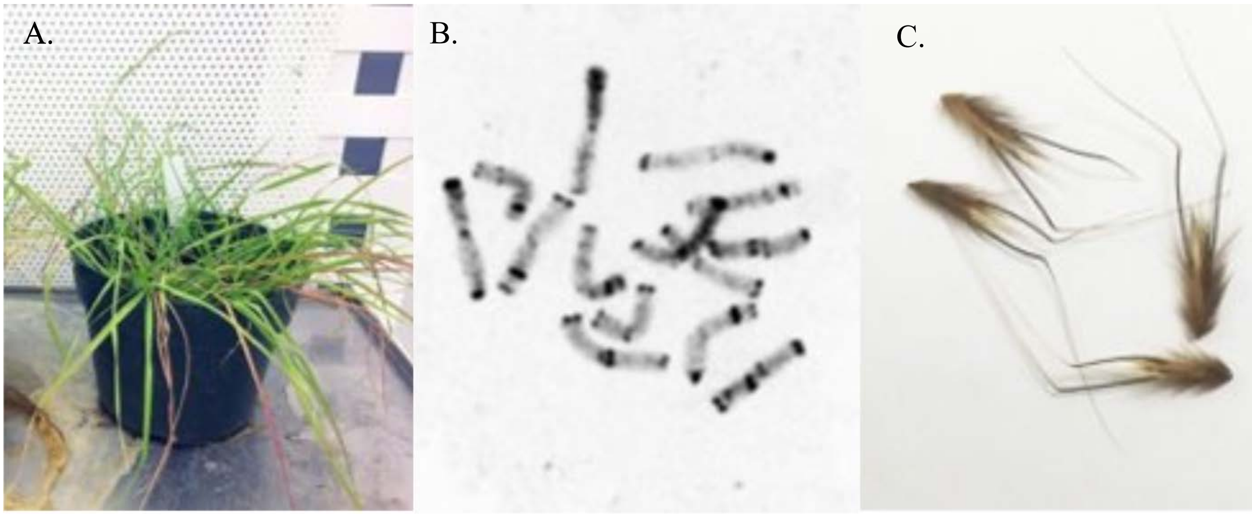


Figure 1. *A. atlantica* plant (A); chromosomes visualized by C-banding method (B) of *A. atlantica* (2n = 14, AA genome); spikelets (C).

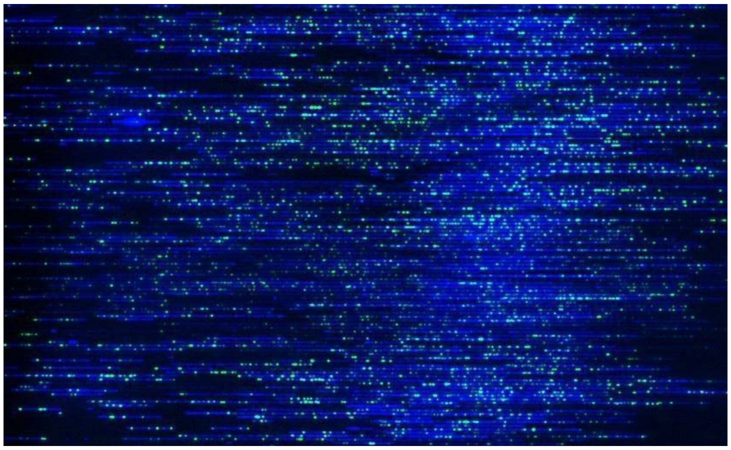


Figure 2. Raw Image of labeled DNA on a flowcell in the Irys imaging machine.

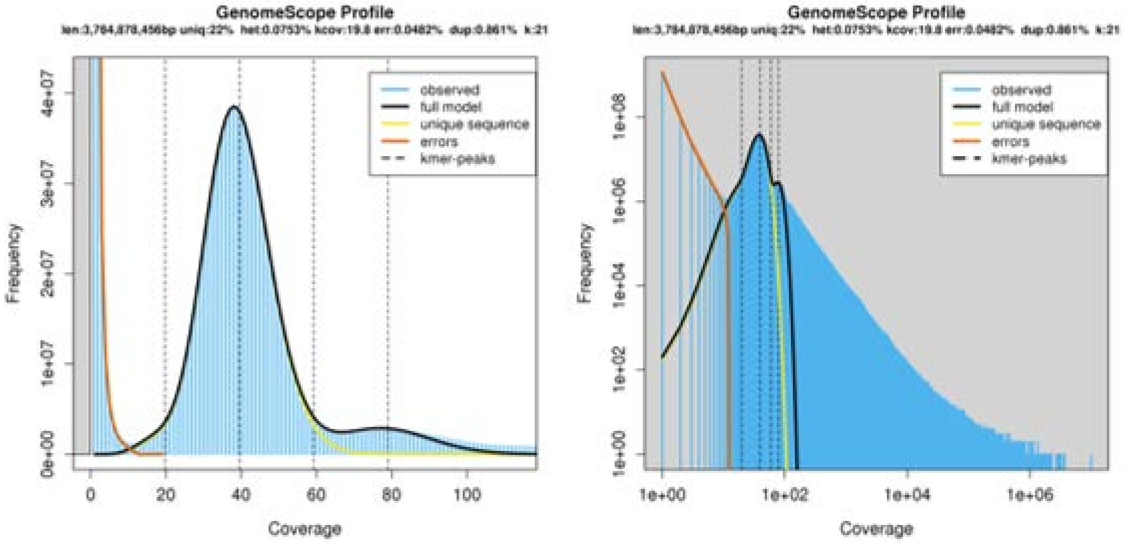
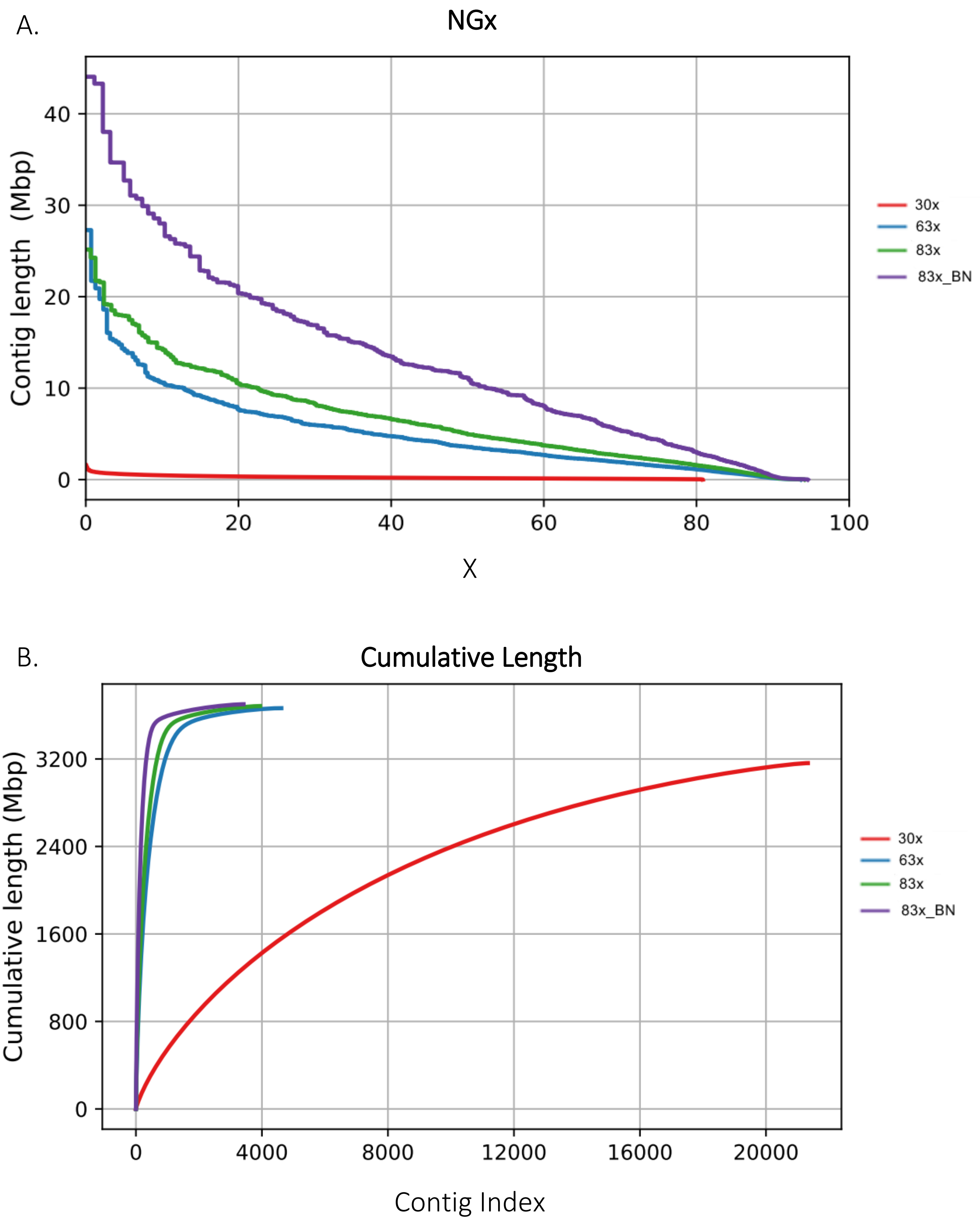


Figure 3. *A. atlantica* (Accession: CC7277) K-mer profile at k=21.


A.
NGx
Contig length (Mbp)
0
10
20
30
40
X
0
20
40
60
80
100
30x
63x
83x
83x_BN
B.
Cumulative Length
Cumulative length (Mbp)
0
800
1600
2400
3200
Contig Index
0
4000
8000
12000
16000
20000
30x
63x
83x
83x_BN

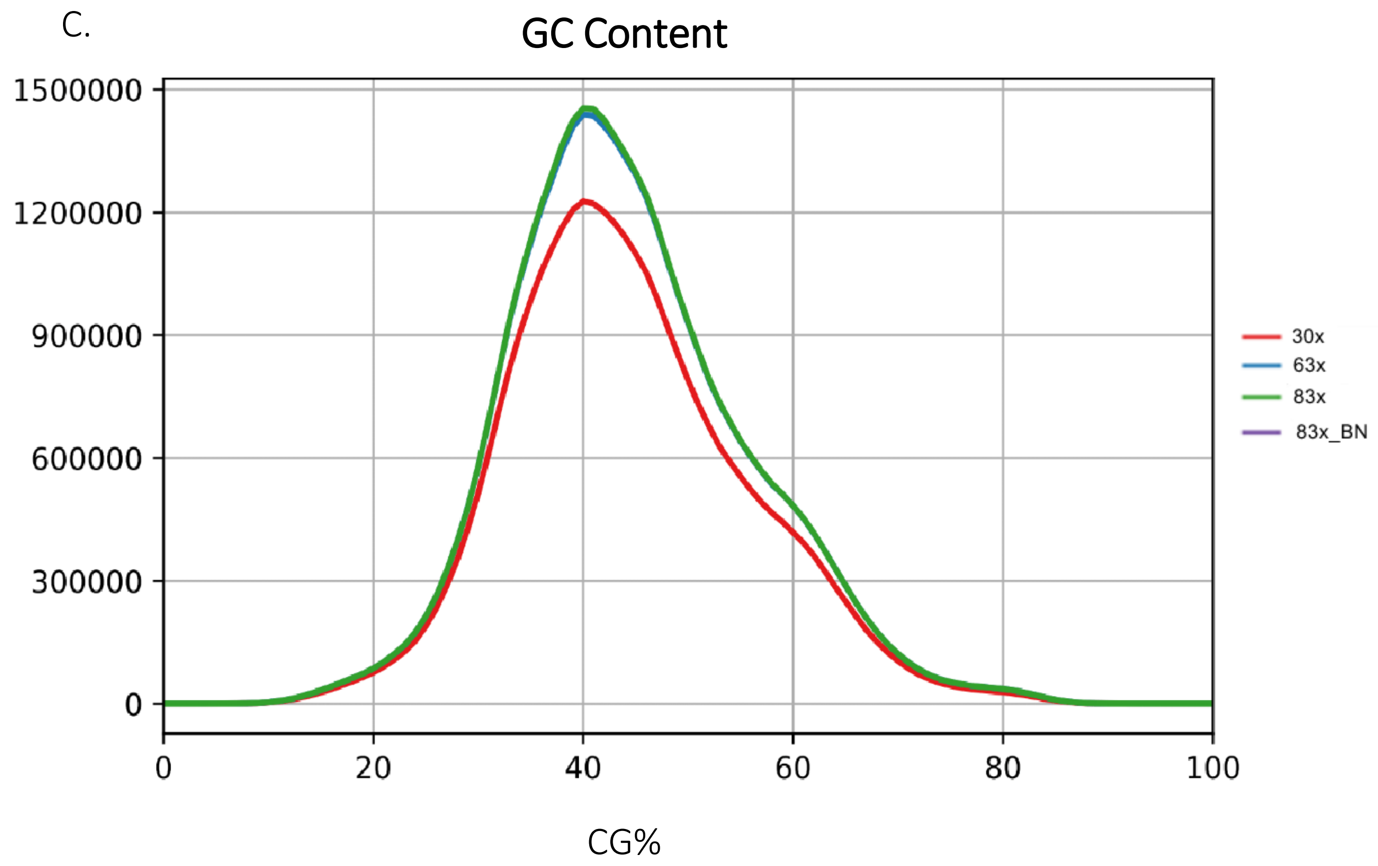


Figure 4. QUAST analysis showing NGx (A) (eg. N0 - N100) as x varies from 0 -100% as well as cumulative length (B) GC content (C) for the Hybrid assembly vs the 27x.

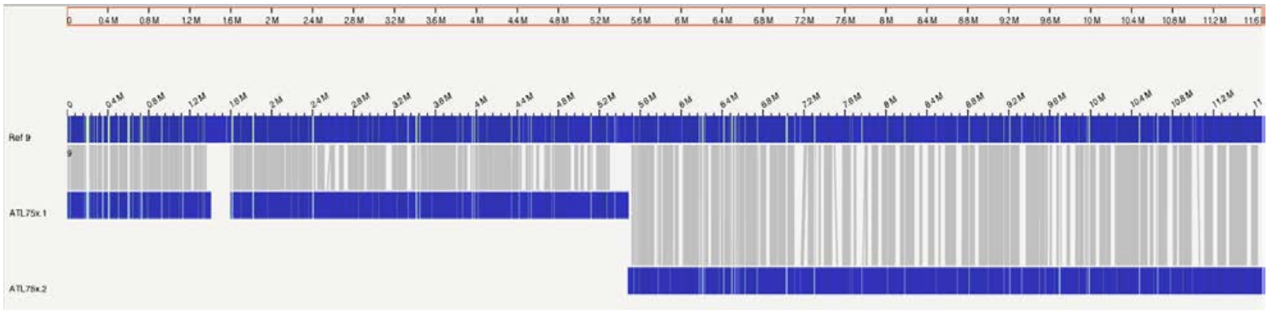


Figure 5. Hybridization of BioNano pseudomolecule (top molecule) and three sequence contigs.

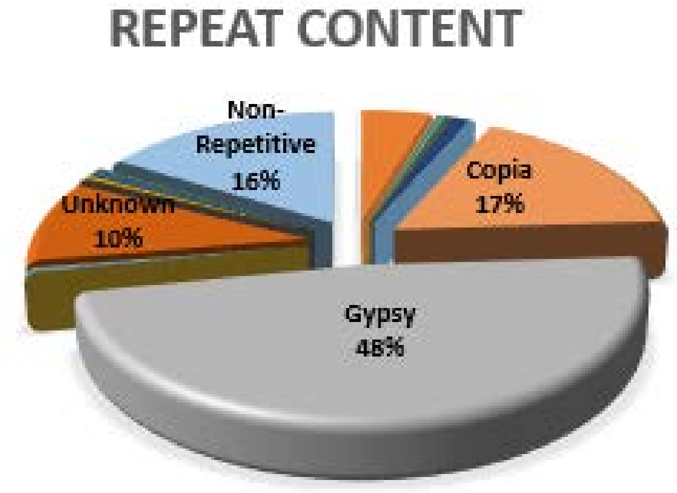


Figure 6. Results of RepeatModler pipeline showing repeat content of the *A. atlantica* assembly. It is estimated that 84% of the genome is repetitive.

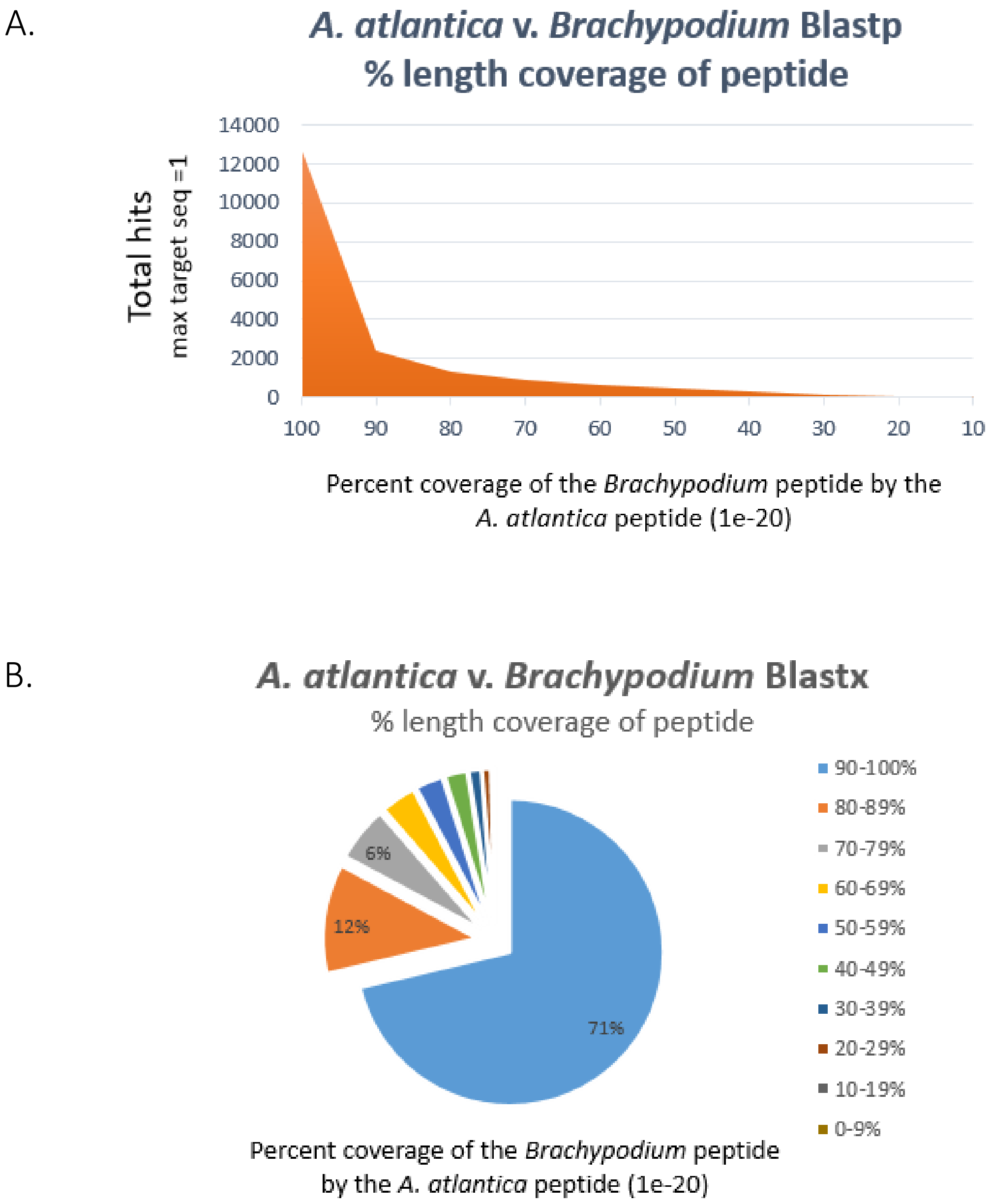


Figure 7. Coverage of *A. atlantica* transcriptome in *Brachypodium*.

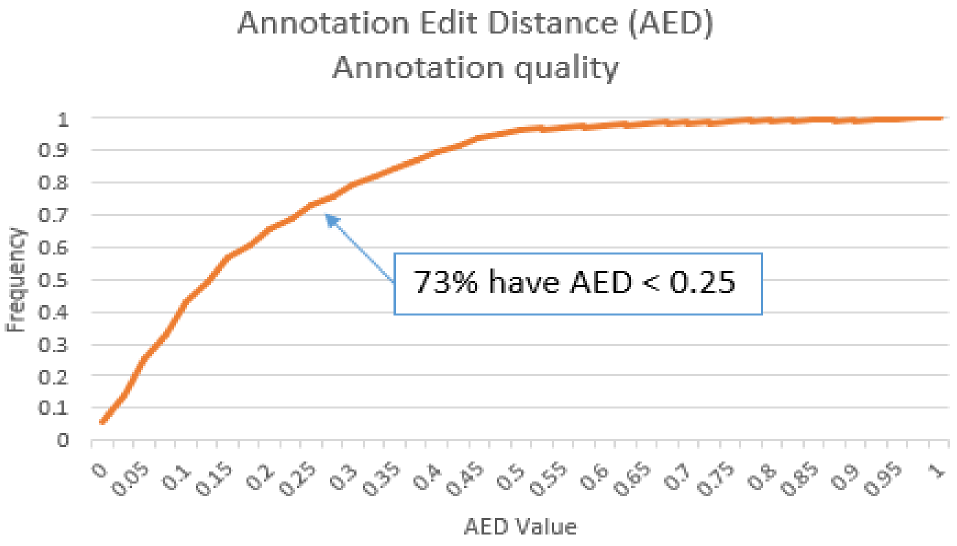


Figure 8. AED calculation to assess annotation quality. 73% of the annotated 83x genome had AED values less than 0.25, indicating a quality annotation.

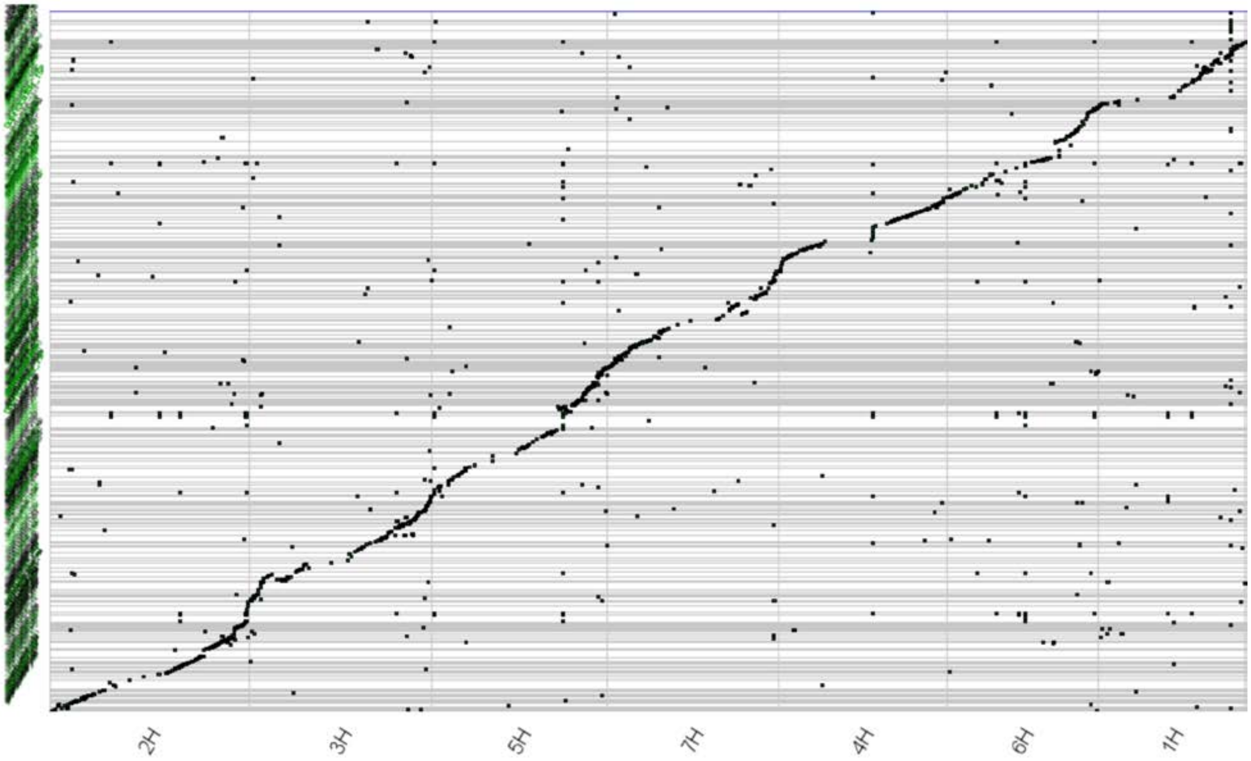


Figure 9. Synteny Analysis: *A. atlantica* vs *Hordium vulgare* (Barley).

TABLES

Table 1. PacBio and Illumina Sequencing Statistics.

PacBio:

Source	**Technology**	**Number of Cells**	**Number of reads**	**Total length of reads**	**Longest read**	**Mean read size**	**Median read size**	**N_{50} read length**	**Genome Coverage**
DNASC	Sequel	40	23,475,393	246,554,592,835	194,884	10706	8,161	18,242	63.2
RTL	RSII	78	7,737,947	75,297,173,119	76,481	9432	6,438	17,316	19.3
AGI	RSII	4	331,056	75,297,173,119	74,575	12373	9,880	20,414	19.3
Total:		**122**	**31,544,396**	**397,148,939,073**	**115,313**	**10,837**	**8,160**	**18,658**	**101.8**

Ilumina:

Source	**Technology**	**Read Length Average (bp)**	**Number of reads (bp)**	**Total length of reads (bp)**	**% N**	**Coverage**
BGI	HiSeq	284	183,480,412	26,116,814,272	0.0003	6.7
Aberystwyth Univ.	HiSeq X	570	1,156,806,386	165,681,504,195	0.0138	42.5
Total:		**142**	**1,340,286,798**	**191,798,000,000**	**0.0024**	**49.2**

Table 2. Average results for k= 19, k=21 and k=23.

Kmer Genome Assessment

Genome Size	3.784 Gb
Unique sequence	21.8%
Repetitive Sequence	78.0%
Heterozygosity	0.0749%

Table 3. Summary statistics for the three sequence assemblies (30x, 63x and 83x) as well as the statistics for the sequence assembly and optical map hybrid (83x_BioNano).

Assembly:	30X	63X	83X	83X_BioNano
# contigs	21.329	4,616	3,941	3,417
# contigs (>= 0 bp)	21,329	4,616	3,941	3,417
# contigs (>= 1000 bp)	21,329	4,616	3,941	3,417
# contigs (>= 5000 bp)	21,309	4,492	3,882	3,358
# contigs (>= 10000 bp)	21,281	4,396	3,840	3,316
# contigs (>= 25000 bp)	21,028	3,840	3,448	2,924
# contigs (>= 50000 bp)	18,193	2,737	2,383	1,859
Largest contig	1,628.267	27,286,170	25,153,855	44,053,509
Total length	3,162,204,854	3,664,452,930	3,683,804,291	3,700,476,325
Total length (>= 0 bp)	3,162,204,854	3,664,452,930	3,683,804,291	3,700,476,325
Total length (>= 1000 bp)	3,162,204,854	3,664,452,930	3,683,804,291	3,700,476,325
Total length (>= 5000 bp)	3,162,138,600	3,664,175,965	3,683,673,652	3,700,345,686
Total length (>= 10000 bp)	3,161,911,534	3,663,444,832	3,683,347,065	3,700,019.099
Total length (>= 25000 bp)	3,156,920,493	3,652,958,992	3,675,687,820	3,692,359,854
Total length (>= 50000 bp)	3,042,858,706	3,612,886,633	3,636,169,914	3,652,841,948
N50	204,301	3,955,572	5,545,214	11,857,933
N75	117,973	1,900,823	2,549,256	5,270,934
L50	4,723	262	196	99
L75	9,802	600	441	215
GC (%)	44.3	44.4	44.38	44.38
# N's	6	1	1	16,655,643
# N's per 100 kbp	0	0	0	450.09
NG50	161,066	3,597,894	4,978,600	111,22,910
NG75	62,186	1,486,583	2,144,290	4,358,836
LG50	6792	295	217	108
LG75	16,246	711	515	249

Table 4. Comparison of hybrid scaffold to original 83x sequence assembly.

Comparison of Hybrid Scaffold to Original Sequence Assembly

Number Genome Maps	6707
Total Genome Map Length (Mbp)	3361.97 (86.2%)
Mean Genome Map Length (Mbp)	0.501
Median Genome Map Length (Mbp)	0.389
Genome Map N_{50} (Mbp)	0.629
Total Reference Length (Mbp)	3661.74
Total Genome Map Length / Reference Length	0.918
Total number of aligned Genome Maps	6648 (0.99)
Total Aligned Length (Mbp)	4571.711
Total Aligned Length / Reference Length	1.249
Total Unique Aligned Length (Gbp)	3.273 (~90%)
Total Unique Aligned Length / Reference Length	0.894

Table 5. Finished hybrid assembly molecule breakdown.

Statistics of Hybrid Scaffold Plus Not-Scaffolded NGS

Count	3417
Hybrid Scaffolds	612
PacBio Contigs	2805
N_{50} length (Mb)	11.86
Total length (Gb)	3.70 (94.8%)

Table 6. Results of RepeatModler pipeline showing categorized repeat content of the *A. atlantica* assembly. It is estimated that 84% of the genome is repetitive.

Repeat Identification:

Total Sequences	4,609
Total Length (bp)	3664271512

Class	Count	bpMasked	%Masked
DNA	4,107	783903	0.02%
CMC-EnSpm	177,304	169248671	4.62%
MULE-MuDR	6,167	1558202	0.04%
Maverick	2,490	1591420	0.04%
MuLE-MuDR	5,319	6591266	0.18%
PIF-Harbinger	31,381	10278851	0.28%
TcMar-Stowaway	82,470	10741899	0.29%
hAT-Ac	2,834	1292254	0.04%
hAT-Tag1	941	419806	0.01%
Line	--	--	--
Jockey	1,526	1215160	0.03%
L1	45,606	41078731	1.12%
RTE-BovB	437	131388	0.00%
LTR	7,714	3921502	0.11%
Copia	263,477	630729638	17.21%
Gypsy	668,071	1753772185	47.86%
RC	--	--	--
Helitron	1,033	451913	0.01%
SINE	--	--	--
Alu	13,761	12259089	0.33%
L1	11,768	5405392	0.15%
tRNA	2,041	420887	0.01%
Unknown	546,143	363725972	9.93%
***Total Interspersed*:**	**1874590**	**3015618120**	**82.30%**
Low_complexity	21942	1167935	0.03%
Satelite	8005	19088160	0.52%
telo	1768	24780179	0.68%
Simple_repeat	175552	10711071	0.29%
***Total*:**	**2081857**	**3071365465**	**83.82%**

Table 7. Percent total putative orthologs displaying synteny to the genome assembly.

Percent Total Orthologs Displaying Mapping Synteny

BUSCO COGs	97%
Uniprot	63%
B. distachyon	88.60%
Long Rice	96.70%
Arabidopsis	93.80%

Table 8. Results of training Augustus using Single-Copy Complete BUSCO. 97% of COGs were identified, indicative of complete and high-quality genome assembly.

Training Augustus using Single-Copy Complete BUSCOs:

Complete BUSCOs	1393
Complete and single-copy BUSCOs	1355
Complete and duplicated BUSCOs	38
Fragmented BUSCOs	12
Missing BUSCOs	35
Total BUSCO groups searched	1440