Insights from the first genome assembly of Onion (Allium cepa)

Richard Finkers,1,* Martijn van Kaauwen1,1 Kai Ament,2 Karin Burger-Meijer,1 Raymond Egging,3 Henk Huits,2 Linda Kodde,1 Laurens Kroon,4 Masayoshi Shigyo,4 Shusei Sato4, Ben Vosman,1 Wilbert van Workum,6 and Olga Scholten1

1Plant Breeding, Wageningen University and Research Centre, 6700 AA Wageningen, The Netherlands
2Bejo Zaden B.V., 1749 CZ Warmerhuizen, The Netherlands
3GenomeScan, 2333 EZ Leiden, The Netherlands
4Laboratory of Vegetable Crop Science, College of Agriculture, Graduate School of Sciences and Technology for Innovation, Yamaguchi University Yamaguchi City, Yamaguchi 753-8515, Japan
5Graduate School of Life Sciences, Tohoku University, Sendai 980-8577, Japan
6Limes Innovations B.V., 2351 SX Leiderdorp, The Netherlands

*Corresponding author: Email: richard.finkers@wur.nl

Abstract

Onion is an important vegetable crop with an estimated genome size of 16 Gb. We describe the de novo assembly and ab initio annotation of the genome of a doubled haploid onion line DHCU066619, which resulted in a final assembly of 14.9 Gb with an N50 of 464 Kbp. Of this, 2.4 Gb was ordered into eight pseudomolecules using four genetic linkage maps. The remainder of the genome is available in 89.6 K scaffolds. Only 72.4% of the genome could be identified as repetitive sequences and consist, to a large extent, of (retro) transposons. In addition, an estimated 20% of the putative (retro) transposons had accumulated a large number of mutations, hampering their identification, but facilitating their assembly. These elements are probably already quite old. The ab initio gene prediction indicated 540,925 putative gene models, which is far more than expected, possibly due to the presence of pseudogenes. Of these models, 47,066 showed RNASeq support. No gene rich regions were found, genes are uniformly distributed over the genome. Analysis of synteny with Allium sativum (garlic) showed collinearity but also major rearrangements between both species. This assembly is the first high-quality genome sequence available for the study of onion and will be a valuable resource for further research.

Keywords: large genome; repeats; gene space; DHCU066619

Introduction

More than just a tasty culinary sensation, onion (Allium cepa L.) is one of the most important vegetable crops worldwide. In terms of global production value, onion ranks second after tomato (http://www.fao.org/faostat/en/#home). Onion is a diploid (2n = 2x = 16) species with a genome size of approx. 16,400 Mb/1C (Van’T Hof et al. 2014; Harkess et al. 2013). Due to the size of the genome and of a size comparable to the allo-hexaploid bread wheat (Brenchley et al. 2012; Marcusen et al. 2014). A large genome size is often associated with repeat accumulation (Kelly and Leitch 2011). The C2T reannealing kinetics indicate that about 40% of the onion genome is highly repetitive (>1000X copies) and 40% has 100-1000 copies and is thus middle to low repetitive (Stack and Comings 1979). Overall, at least 95% of the A. cepa genome consists of repetitive sequences (Flavell et al. 1974), most of which are dispersed repeats (Shibata and Hizume 2002) and LTR retrotransposons of the Ty1/copia and Ty3/gypsy type (Pearce et al. 1996; Kumar et al. 1997; Pich and Schubert 1998; Vitte et al. 2013). Due to the size of the genome and the repetitive nature, developing an onion reference genome assembly is challenging (Havey and McCallum 2012).

For onion, molecular breeding strategies are currently limited to the use of molecular markers and genetic linkage maps (Martin et al. 2005; McCallum et al. 2012; Baldwin et al. 2012; Scholten et al. 2016). Knowledge of the genome of onion and related species is scarce compared to other crop plants, with only transcriptome sequences available (Kim et al. 2014; Kamenetsky et al. 2015; Sohn et al. 2016; Abdelrahman et al. 2017). While the availability of a reference genome has greatly stimulated research and led to accelerated breeding in many other crops (Kuhl et al. 2005; Vitte et al. 2013; Sun et al. 2020), onion has not yet had this benefit. Garlic (Allium sativum; 16.2 Gb) and asparagus (Asparagus officinalis L.; 1.1 Gb) are the most closely related species with reference genomes available (Harkess et al. 2017; Sun et al. 2020). Though, useful for gene discovery, the lack of insight into detailed synteny relationship between these crops and onion is still limiting the utilization in onion (breeding) research.

In this article, we describe the first de novo assembly of the genome of a doubled haploid A. cepa accession through a combination of strategies and the development of an initial set of pseudomolecules. Synteny between onion and garlic was studied and provides a first insight into the similarities and differences in...
genome organization. This onion assembly will be an important tool, facilitating onion breeding and research.

Materials and methods
Plant material
Seeds of the doubled haploid (DH) Allium cepa line DHCU066619 (Hyde et al. 2012) were kindly provided by Dr. M. Mutschler (Cornell University). This DH accession was selected for whole-genome shotgun sequencing (WGS), as it is a homozygous, vigorously growing genotype.

DNA and RNA isolation
Two grams of leaf tips from young and rapid-growing onion plants were harvested and pooled for nuclei DNA isolation according to the protocol described by Bernatzky and Tanksley, (1986).

For RNA extraction, tissues from bulb and basal plate, as well as leaf and flowers from mature plants were harvested, frozen in liquid nitrogen and ground. Approximately 100 mg of each tissue was transferred to a 2 ml screwcap tube, followed by adding 800 μl of Trizol and vortexing for 1 min. Subsequently, 160 μl of Chloroform was added, and gently mixed for 15 s. After incubating the samples for 3 min at RT, they were centrifuged at 10,000 rpm in an Eppendorf centrifuge for 3 min, followed by transfer of the water-phase to a clean tube. A 350 μl of RLT buffer (Qiagen) containing 10 μg/μl β-mercaptoethanol, was added to each 100 μl of water-phase, followed by a RNeasy (Qiagen) column extraction.

Sequencing methods, preparation details, and data processing
Three TruSeq sequencing libraries (with median insert sizes of 230, 350, and 500) were made according to the manufacturer’s recommendations and sequenced during a single Illumina HiSeq 2500 run (across 16 lanes) by GenomeScan (Leiden, The Netherlands). Quality of the data was assessed using FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and by evaluating the k-mer profile using JELLYFISH (Marçais & Kingsford 2011).

PacBio Sample preparation was performed according to the PacBio protocol "20 kb Template Preparation Using BluePippin™ Size-Selection System." In short, 8 μg of sample was fragmented using a Covaris g-Tube at 4800 rpm for 1 min. The PacBio SMRTbell™ Template Prep Kit 1.0 was used for the DNA library preparation of the primer annealed SMRTbells. The SMRTbells were size selected using the BluePippin set for 10 kb–50 kb long reads. The SMRTbells were size selected by evaluating the k-mer profile using JELLYFISH (Marçais & Kingsford 2011).

The PacBio reads were assembled with the Illumina assembly as a backbone using DBG2OLC (Ye et al. 2016) and varying settings for kmerSize, KmerCov, MinOverlap, and AdaptTH (Supplementary Files S2 and S3). Final DBG2OLC assembly was performed with kmerSize = 21, KmerCov = 2, MinOverlap = 20, and AdaptThr = 0.05.

Genome scaffolding
Dovetail scaffolding
Five grams of young leaf tissue (lyophilized) was sent to Dovetail Genomics (USA) for High Molecular Weight (HMW) DNA isolation, together with the PacBio/Illumina hybrid assembly, the unplaced Illumina scaffolds >3 Kb and a published onion chloroplast sequence (von Kohn et al. 2013) to produce a scaffolded assembly according to the Chicago protocol (Putnam et al. 2016). For this, the generated DOVETAIL sequencing libraries were analyzed with a modified version of the HiRise algorithm (Dovetail Genomics, USA) to accommodate the large genome size.

Genome annotation
Completeness
Completeness of the onion genome assembly was assessed by BUSCO v4.1.1; database version embryophyte_obd10 2019-11-27 (Simao et al. 2015).

Repeat annotation
To structurally annotate repeat sequences in the A. cepa genome de novo, RepeatMasker was applied using the REPBASE v20.5 library. Ab initio prediction of repeats was performed using the TeDenovo pipeline of REPET v2.5 (Quesneville et al. 2005; Flutre et al. 2011), with default parameters, utilizing NCBI-Blat+. To reduce computational time needed to execute the TeDenovo pipeline, 0.52 Gb of the 14.4 Gb (3.6%) of the onion assembly was selected at random. Grouper, Recon, and Filer steps were invoked both with and without structural detection. Repbase v20.05 and Pfam27.0 HMM profiles were used to annotate repeats identified in onion by the pipeline. The output of the TeDenovo pipeline was subsequently used as the reference library to run the TAnnot pipeline on randomized chunks of the whole assembly using default parameters and exported to GFF3 format.

Ab initio gene prediction
RNA from four tissues (bulb, basal plate, leaf, and flower) was isolated and sequenced by GenomeScan (Leiden, The Netherlands) using an Illumina HiSeq 2500 sequencer according to the manufacturer’s recommendations. BRAKER1 (Hoff et al. 2015) was used for unsupervised training of Augustus 3.2.2 (Stanke et al. 2008), using splice junctions identified from mapping RNAseq reads with STAR mapper (Dobin et al. 2013) of all four tissues combined. Subsequently, the training parameters from BRAKER1 were used to annotate the masked genome sequence.
Functional annotation/Blast2go interpro

Blast2Go (Gotz et al. 2008) was used for functional annotation of the predicted protein models with the default settings for the mapping and annotation step. The initial blastp 2.6.0+ step was performed against the Swissprot database (version 4 October 2017) with an e-value cutoff of 1.0e-3, word size of 6, Low Complexity filter on true, and a maximum of 20 blast hits. InterProScan v5.26 (Jones et al. 2014) including panther 12.0 libraries was used to identify protein domains within the predicted protein sets.

Synteny analysis

Sequence of EST-based markers from Fujito et al. (2021) were blasted to the onion genome and to the garlic genome. A total of 4034 markers had a hit to both genome sequences, of which 519 positioned at unanchored onion scaffolds, leaving 3515 markers positioned on the onion pseudomolecules. Based on the top hits, the physical positions of markers with a match to both genomes were plotted against each other in a XY plot using the python matplotlib and seaborn libraries.

Results and discussion

Genome assembly

The doubled haploid Allium cepa accession DHCU066619 (Hyde et al. 2012) was selected for WGS, to facilitate genome assembly especially of a large genome like onion. Three small insert libraries were used for Illumina HiSeq 2500 sequencing (Supplementary Table S1) and yielded 769 Gb sequence data. Analysis of the sequence libraries resulted in ~450 G k-mers (k = 31), and an estimated genome size of approx. 13.6 Gb. Of this, approx. 7.4 Gb (53.8%) is single copy based on k-mer statistics, indicating that an initial assembly from Illumina reads is feasible. The MaSurCa based assembly resulted in 10.8 Gb in 6.2 M contigs with a contig N50 of 2.7 Kb (Table 1). This assembly was further scaffolded using 18.1 M PacBio RS II reads >1 Kb with DBG2OLC. This resulted in an assembly of 14.6 Gb in 316 K contigs and a contig N50 size of 59 Kb (Table 1). In the last step, the hybrid Illumina/PacBio assembly was further improved using Dovetail Chicago and subsequent HiRise scaffolding which indicated that there were no mis-assemblies in the Illumina/PacBio hybrid contigs, showing the high quality of this initial assembly. The combination of these three technologies resulted in an assembly of 14.9 Gb in 92.9 K scaffolds with a scaffold N50 size of 454 Kb. With an estimated genome size of 16,400 Gb/1C, we managed to assemble ~91% of the onion genome. As our assembly is mostly based on short read illumina sequencing, with limited data from third generation PacBio long reads, we hypothesize that the most complex highly repetitive regions are missing from our assembly.

Anchoring scaffolds into pseudomolecules using multiple EST-based genetic maps

To further organize our genome assembly, we used three intra-specific genetic linkage maps (Duangjit et al. 2013; Choi et al. 2020; Fujito et al. 2021) and one interspecific genetic linkage map (Scholten et al. 2016) to anchor scaffolds into pseudomolecules using AllMaps (Tang et al. 2015). Except for the GBS-based markers from the study of Choi et al. (2020), all markers were developed from transcriptome sequencing. With this approach, we were able to anchor 3303 scaffolds (3.6% of the total number of scaffolds) into eight pseudomolecules, with an overall length of 2.4 Gb (15.9% of the assembly size; Supplementary File S6). The pseudomolecules were named after their respective chromosome according to the A. cepa monosomic addition lines (Shigyo et al. 2008; Van Heusden et al. 2000; Fujito et al. 2021). A subset of 157 scaffolds (0.15 Gb) could be oriented according to the genetic map order. Overall agreement between the scaffolds and genetic positions was good (Supplementary File S6) with absolute Spearman correlation coefficients over the four maps ranging between 0.86 (BYGxAC) and 0.95 (DHAXDHC). While the overall order of chromosome 8 over all four maps was very consistent (Spearman correlation coefficient of 0.99), chromosome 7 showed the lowest correlation (Spearman rho coefficient of 0.716), mainly because of the interspecific CCxRF map (p = 0.428). Disruption of collinearity in homoeologous chromosomes in Allium species has been previously reported (Khrustaleva et al. 2019a). Not all maps were informative in the Allmaps scaffolding. Markers of the CCxRF map did not contribute to the scaffolding of chromosome 3 while markers of the GBS map did not contribute to the scaffolding of chromosome 3, 5, and 8. ALLMAPS has implemented an algorithm to minimize ambiguity (Tang et al. 2015), and the GBS marker-based sequences showed similarity to more than one scaffold, which were assigned to different linkage groups, and were therefore skipped for the final scaffolding. This resulted in v1.2 of the genome assembly. Overall, the developed pseudomolecules will be very useful in developing additional genetic markers, fine mapping QTL regions and/or candidate gene mining.

Genome annotation

The completeness of the Dovetail scaffolded de novo genome assembly was evaluated using BUSCO genes (embryophyta_odb10; 2020-09-10; Supplementary File S4) and resulted in a completeness score of 87.7%. Other large genome assemblies, such as the loblolly pine v1.01 genome has an CEGMA completeness of 91% (Neale et al. 2014) and the Allium sativum v1.0 genome has an CEGMA completeness score of 92.7% and a BUSCO completeness score of 88.7% (Sun et al. 2020). Of the BUSCO genes, 3.3% were labeled as fragmented, which could be because the length of the gene model did not fall within the expected length distribution of the chosen BUSCO profile. Also technical limitation of the algorithm might increase proportions of fragmented and missing BUSCOs, especially in large genomes (Simao et al. 2015). Still, the slightly lower completeness scores in onion suggests that we miss genes in our assembly.

Initial repeat masking of the genome with repeatmasker, using the REPBASE v2.0.5 database, resulted in 15.1% of the genome to be annotated as repetitive (Supplementary Table S5), which is far below the expected 95% moderate to high repetitive regions (Flavell et al. 1974) though comparable to homology-based repeat annotation in other large genomes, such as loblolly pine (Wegrzyn et al. 2013). Though we expect REPBASE to be missing Allium specific repeats, this result indicates that most repeats in the genome are not recognizable anymore as repeats, have accumulated mutations and/or degraded due to repeated transpositions within a transposon and therefore disturbing its structure, and are probably (very) old. This observation agrees with the observed Kmer statistics, which suggests that 53.8% of the genome is single copy and the observations by Jakše et al. (2008) that onion BAC sequences contain >50% sequences that are like transposons, many of which are degraded. Vitte et al. (2013) showed that current repeat databases contain limited data from onion-related species, hampering homology-based detection of TE’s, which is supported by our finding that <10% of the long terminal repeat (LTR) retrotransposons could be directly identified in such an approach. Using de novo repeat annotation strategies, genomes
Table 1  Statistics of the steps in the onion genome assembly process

<table>
<thead>
<tr>
<th></th>
<th>Illumina</th>
<th>Illumina and Pacbio hybrid</th>
<th>Illumina, Pacbio, and Dovetail</th>
<th>Illumina, Pacbio, Dovetail, and Allmaps—A. cepa v1.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total assembly size (GB)</td>
<td>10.80</td>
<td>14.59</td>
<td>14.94</td>
<td>14.94</td>
</tr>
<tr>
<td>N50 (bp)</td>
<td>2776</td>
<td>59,009</td>
<td>436,203</td>
<td>464,272</td>
</tr>
<tr>
<td>L50</td>
<td>91,028,083</td>
<td>76,344</td>
<td>9266</td>
<td>6486</td>
</tr>
<tr>
<td>NR of contigs (K)</td>
<td>6200</td>
<td>316</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>NR of scaffolds (K)</td>
<td>—</td>
<td>—</td>
<td>92.9</td>
<td>89.6</td>
</tr>
<tr>
<td>NR of scaffolds incorporated in pseudomolecules (K)</td>
<td>—</td>
<td>—</td>
<td>2.4</td>
<td>3.3</td>
</tr>
<tr>
<td>Assembly size in pseudomolecules (Gb)</td>
<td>33.2</td>
<td>33.5</td>
<td>33.5</td>
<td>33.5</td>
</tr>
<tr>
<td>G + C content (%)</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Estimated TE content (%)</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Number of gene models</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Number of gene models with &gt;90% RNAseq support</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>47,066</td>
</tr>
<tr>
<td>Median/average/max exon length (bp)a</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>137/202/14,836</td>
</tr>
<tr>
<td>Median/average/max intron length (bp)a</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>178/1035/213,043</td>
</tr>
<tr>
<td>N'S per 100 Kbp</td>
<td>0</td>
<td>0</td>
<td>2344</td>
<td>2346</td>
</tr>
</tbody>
</table>

*Gene statistics based on 86k models with blast support.

comparable in size, such as garlic (Sun et al. 2020), bread wheat (Mayer et al. 2014), and loblolly pine (Wegrzyn et al. 2014) were annotated as containing approx. 91%, 80%, and 82% repetitive DNA sequences, respectively. Therefore, we used the “all vs all comparison methods,” as implemented in the REPET pipeline (Flutre et al. 2011) to develop an onion-specific repeat database. Using this de novo developed repeat database, 72.4% of the genome sequence could be classified as repetitive and was subsequently masked for downstream analysis. Still, this percentage is lower than the expected 95%, suggesting that the remaining ~20% of the “repetitive” sequences were too diverged to be recognized. This touches upon a standing debate within the Onion community on the age of onion-LTR’s. Jakse et al. (2008) suggests that most LTRs are old and inactive while Vitte et al. (2013) suggests that young, intact and nested LTRs are abundant in the onion genome. Our results indicate that nesting of elements of LTR sequences occurs frequently, in line with the study of Vitte et al. (2013). Our genome assembly can serve as a template to further investigate this question.

LTR retrotransposon are the major contributors to the size of the onion genome (Supplementary Table S5), which is in line with previous studies (Jakse et al. 2008; Vitte et al. 2013). The REPBASE annotation indicates that the majority of the young LTRs are of the Gypsy type; followed by Copia. This has also been observed in other large plant genomes such as spruce (Nystedt et al. 2013), garlic (Sun et al. 2020), and wheat (Appels et al. 2018). LTR retrotransposons have been described in relation to genome size increase (Kumar and Bennett 1999) and have been suggested to play a significant role in adaptive response of the genome to environmental challenges (McClintock 1984).

Ab initio prediction of gene models on the repeat masked genome sequence with Augustus resulted in the identification of 540,925 gene models, from which 47,066 showed >90% coverage with reads from the RNAseq dataset (Table 1). The number of gene models is way beyond the average number of genes of 36,795 reported for plant genomes (Ramirez-Sanchez et al. 2016). A proper annotation to identify only functional genes would require extensive manual curation (Hosmani et al. 2019b; Tello-Ruiz et al. 2019; Athanasouli et al. 2020). We decided to make the extended set available for the community, rather than restricting ourselves to making only models available with additional RNAseq support. The abundance of gene models may most likely be explained by the presence of pseudogenes (Zhang and Gerstein 2004; Xiao et al. 2016). Pseudogenes are non-functional copies of genes that were once active in the ancestral genome. For example, in Arabidopsis 924 pseudogenes are known (Xiao et al. 2016) while in wheat 288,839 pseudogenes were identified (Appels et al. 2018). A pseudogene still has characteristics of a gene and will be detected using an ab initio gene model prediction algorithm, but not be annotated by blast against curated protein databases, such as TrEMBL, due to partial matches. Blast analysis against TrEMBL resulted in hits for 86,073 gene models (15.9% of the ab initio predicted models). Using Blast2Go (Conesa et al. 2005), 88,259 gene models were functionally annotated, of which 49,918 models were annotated using data from both Blast and InterPro. Of these, 17,457 models were annotated exclusively by InterPro, while 20,884 models had a blast hit only. For subsequent analysis, we focus on the subset of 86,073 models with similarity to genes from TrEMBL, of which 25,344 showed >90% coverage with reads from the RNAseq dataset. The average coding sequence length of onion genes is 879 bp and is spread over 4.4 exons. This is shorter than the average plant gene length of 1308 bp (Ramirez-Sanchez et al. 2016), though still within the observed variation and larger than the average gene length of 797 bp reported for garlic (Sun et al. 2020). Average intron length is 1029 bp (Table 1), though the largest predicted intron is 213 Kb. Although a positive relationship between intron size and genome size has been observed (Stival Sena et al. 2014), due to large variations that occur in intron sizes, it is not a good predictor for genome size (Wendel et al. 2002). With a median and average intron size of 178 and 1035 bp, respectively, the majority of introns in onion is short, while a limited number of introns is (very) long. This is probably because of the energy required for transcribing the long genes.

Organization of the onion gene space

In plants, we have seen two scenario’s for the distribution of genes: genes mainly located in actively recombining euchromatin regions, while large non-recombining regions (centromeres) have a low gene content, such as in tomato (Hosmani et al. 2019a), or genes equally distributed over the genome, such as in garlic (Sun et al. 2020). If onion would show a similar pattern as tomato, then most gene models will be included in the current set of pseudogenes. Based on the set of 86K models with a match to the TrEMBL database, we calculated a rate of 8108 and 5352 gene models/Gb for anchored and unanchored scaffolds, respectively. This shows that gene density in the anchored scaffolds is approximately 1.5 x higher than in the unanchored scaffolds. In tomato, we calculated gene density in euchromatin and heterochromatin
Sim et al. 2012; Viquez-Zamora et al. 2014) to be 97,715 and 20,010 gene models/Gb respectively, a difference of approximately 4.9x. Interpretation of this data and the meaning for onion must be treated with caution as it is influenced by two factors; (1) the difference in genome size between onion (16 Gb) and tomato (850 Mb) and (2) the fact that only 2.4 Gb (out of 14.9 Gb) of the onion scaffolds are yet organized into pseudomolecules. Having said that, like Jakse et al. (2008) who previously sequenced two onion BACs, we hypothesize that genes in onion are more equally distributed over the genome, and reside in an ocean of repetitive elements, as the ratio between scaffolds incorporated in pseudomolecules and the scaffolds not incorporated in pseudomolecules is much closer to 1, than the ratio observed in tomato. This hypothesis is further supported by the results of the synteny analysis with garlic. Not only the 3515 of the 5339 onion unigene derived EST markers (Fujito et al. 2021) showed a uniform distribution on the garlic and onion pseudomolecules (Figure 1), also the distribution of transposable elements is uniform over the garlic and onion chromosomes (Sun et al. 2020; Supplementary Figure S7). Our hypothesis is in line with previous results in which the mapping of genes on physical chromosomes using molecular cytogenetic methods showed that

Figure 1 Physical positions of the marker on the garlic pseudomolecules (x-axis) plotted against the physical position of the marker on the onion pseudomolecules (y-axis). Color of the points represent the original linkage group assignment of the marker on the onion genetic map. Synteny between the garlic and onion genome can be observed in several chromosomes (e.g., garlic chromosome 2, 3, and 7 and its onion counterparts chromosome 8, 3, chr6, respectively) though signals of translocation can also be observed (e.g., garlic chr5 and parts of onion chromosome 1 and 4). The data suggest inversions within a chromosome (e.g., garlic chromosome 7 and its onion counterpart chromosome 6), but this cannot be estimated with certainty as not all our contigs could be oriented using Allmaps.
the genes in Allium are localized in all three regions of the chromosome arms: proximal, interstitial, and distal (Scholten et al. 2007; Masamura et al. 2012; Khrustaleva et al. 2016, 2019a,b).

Synteny between Allium cepa (onion) and Allium sativum (garlic)

The sequence of 5399 Onion unigenes (Fujito et al. 2021) were also used to study the synteny between onion and garlic (Sun et al. 2020). For 3515 unigenes, physical positions were determined on both the garlic and onion pseudomolecules. Overall synteny between some chromosomes is strong, such as for garlic chromosomes 2, 3, and 7 and their onion counterparts’ chromosome 8, 3, and 6, respectively (Figure 1). Signals for translocations between chromosomes were also observed, as garlic chromosome 5 seems to be split over onion chromosome 1 and 4. In addition, signals for inversions were observed, for instance for garlic chromosome 7 (onion chromosome 6). Where the signal for syntenic relationships between garlic and onion genomes are high, the garlic genome sequence may be used to further organize the onion scaffolds into pseudomolecules.

Improving usability of the onion genome assembly

The onion genome assembly in its current form is already a powerful tool for research and practical breeding. The annotation will be a good starting point for mining the genome for candidate genes while the set of pseudomolecules will facilitate the development of new markers for targeted regions. As syntenic relationships between garlic and onion genomes are high, insight in the overall synteny between garlic and onion can be used to develop hypothesis and assign unplaced scaffolds to approximate positions on the onion pseudomolecules, further facilitating discovery of novel insights. However, real improvements should come from additional lab work. Our current assembly is primarily based on Illumina short read sequencing and covers approx. 91% of the expected genome size and has a BUSCO completeness value 87.7% indicating, this assembly still needs improvement. Current third-generation sequencing technologies, such as ONT long read (Jain et al. 2016) and PacBio HiFi (Wenger et al. 2019), have shown to deliver larger continues genome assemblies (Michael and VanBuren 2020). Data from one or both platforms, combined with, for example, Hi-C scaffolding data, would lead to a more continuous assembly, as shown for garlic (Sun et al. 2020), a genome with a size similar to onion.

Perspective for addressing evolutionary questions regarding genome size expansion

With the development of genome assemblies in the Alliaceae, but also in other groups of plant species with large genomes, for example, gymnosperms (Zimin et al. 2014), data are becoming available to study genome size expansion. Earlier studies in onion, suggests that tandem duplications plays a role (King et al. 1998). The garlic genome assembly shows the presence of intra and inter chromosome syntenic blocks (Sun et al. 2020). The onion pseudogenes will likely contain signals supporting either tandem and/or segmental duplication. We have developed a partly annotated genome-specific repeat database, with the purpose to mask the genome for ab initio gene prediction. The study of Vitte et al. (2013) has already shown that homology-based annotation using reference libraries will likely miss intact LTRs, such as onion-specific Gypsy and Copia, and that further annotation is needed using the know domain structure of retroelements (Wicker et al. 2007). Such analysis will give further insights in the history of repeats involved in genome expansion. The comparison of (pseudo)gene and ancient retroelements to, for example, Garlic and Lobolly pine will be a next step in understanding genome size expansion in onion, but also in other plant species with large genomes.

Conclusion

We have produced the first de novo genome sequence of onion. The sequence provides insights into the distribution of genes and repeats in this important cp species. This assembly is the first high-quality genome sequence and will be a valuable resource for both research and breeding.

Acknowledgments

The authors would like to thank Dr Marta Mutschler for providing the DH line and Dr Alexey Zimin for his help in debugging the MaSurCa assembly pipeline. This work was, in part, carried out on the Dutch national e-infrastructure with the support of SURF Cooperative.

Funding

This research was supported by a grant from the Top Sector Horticulture & Propagation Materials (H279-SEQUON) and by the companies Bejo Zaden B.V., De Groot en Slot B.V., and Genomescan. Within the Top Sector, the business community, knowledge institutions and the government work together on innovations in the field of sustainable production of healthy and safe food and the development of a healthy, green living environment.

Data availability

The onion DH is deposited to the U.S. National Plant Germplasm System and is available under the number: G 32985. The final v1.2 assembly and annotation files are available on www.oniongenome.wur.nl for download and in a Genome browser (Skinner et al. 2009). All data, including the genome sequence and raw sequencing reads have been deposited to EBI under the BioProject ID: PRJEB29505. Supplemental Material available at figshare: https://doi.org/10.25387/g3.14778735.

Conflicts of interest

None declared.

Literature cited


