MaGuS: a tool for map-guided scaffolding and quality assessment of genome assemblies

Mohammed-Amin Madoui[1§], Carole Dossat[1], Léo d'Agata[1], Jan van Oeveren[2], Edwin van der Vossen[2], Jean-Marc Aury[1]

[1]CEA, DSV, Institut de Génomique, Genoscope, 2 rue Gaston Crémieux, CP5706, 91057 Evry, France

[2]Keygene NV, Agro Business Park 90, 6708 PW, Wageningen, The Netherlands

[§]Corresponding author

Email addresses:

MAM: amadoui@genoscope.cns.fr

CD: cdossat@genoscope.cns.fr

LA: ldataga@genoscope.cns.fr

EVDV: edwin.van-der-vossen@keygene.com

JVO: jan.van-oeveren@keygene.com

JMA: jmaury@genoscope.cns.fr

Abstract

Background

Scaffolding is a crucial step in the genome assembly process. Current methods based on large fragment paired-end reads or long reads allow an increase in continuity but often lack consistency in repetitive regions, resulting in fragmented assemblies. Here, we describe a novel tool to link assemblies to a genome map to aid complex genome reconstruction by detecting assembly errors and allowing scaffold ordering and anchoring.

Results

We present MaGuS (map-guided scaffolding), a modular tool that uses a draft genome assembly, a genome map, and high-throughput paired-end sequencing data to estimate the quality and to enhance the continuity of an assembly. We generated several assemblies of the Arabidopsis genome using different scaffolding programs and applied MaGuS to select the best assembly using quality metrics. Then, we used MaGuS to perform map-guided scaffolding to increase continuity by creating new scaffold links in low-covered and highly repetitive regions where other commonly used scaffolding methods lack consistency.

Conclusions

MaGuS is a powerful reference-free evaluator of assembly quality and a map-guided scaffolder that is freely available at https://github.com/institut-de-genomique/MaGuS. Its use can be extended to other high-throughput sequencing data (e.g., long-read data) and also to other map data (e.g., genetic maps) to improve the quality and the continuity of large and complex genome assemblies.

Keywords

Scaffolding, genome map, anchoring, whole genome profiling, Arabidopsis

Background

Technical advances and cost reduction in genome sequencing have allowed the completion of numerous genome sequencing projects based on whole-genome shotgun fragments using high-throughput sequencing data and the assembly of these data. The genome assembly process usually involves four main steps: reads assembly into contiguous sequences (contigs), linking of contigs into larger gap-containing sequences (scaffolds), gap closing to fill gaps generated by the scaffolding, and anchoring onto a genetic map to build the final pseudo-molecules. During the second step, end sequences of large fragments (>1 kb) or long reads are aligned to the contigs and the alignment information is used to link contigs into scaffolds. Several commonly used scaffolding programs have been published in the last decade [1]. The efficiency of the scaffolding depends mainly on the diversity and fragment size of the input reads libraries and on the size and quality of the long reads. Typically, 1 to 20 kb libraries are used consecutively during the scaffolding step, which allows repetitive regions of various sizes to be spanned [2]. However, during the alignment step, the presence of repeated sequences creates multiple assembly solutions, which generally causes ambiguities that scaffolder programs cannot untangle. This is often the case in large and complex genomes where repetitive elements are large and cover a large fraction of the genome [3]. To decrease the number of false links, scaffolder programs require a cutoff for the minimum number of read pairs (or long reads) that validate a contigs junction; as a consequence, low-covered contigs are overlooked for scaffold building.

Access to a genome map is a great advantage in obtaining a high-quality genome assembly [4]. Genome maps can also help in detecting assembly errors by revealing discrepancies between the map and the assembly [5] and can provide independent information for evaluating genome assembly quality. Currently, three different types of genome maps can be produced to drive or improve assemblies: physical maps, optical maps, and genetic maps.

Historically, physical maps have been used for large genome sequencing projects to order clones and perform clone-by-clone sequencing, which reduces the complexity of the assembly by sequencing single or pooled clones [6, 7]. Although, this strategy is time consuming and expensive, it remains the best option for high quality genome sequencing of large and complex (polyploid) genomes such as the wheat genome [8]. Recently, the Whole Genome Profiling (WGP™) approach was developed by Keygene NV (Wageningen, The Netherlands) to create an accurate sequence-based physical map starting from a bacterial artificial chromosome (BAC) library [9]. In the WGP method, pooled BAC DNA is digested by a restriction enzyme and after amplification, Illumina technology is used to obtain sequence tags (typically 50 nucleotide sequences flanking the restriction sites). WGP has been used successfully to build physical maps of several plant genomes such as those of wheat [10] and tobacco [11].

Optical maps were used to assemble the *Amborella* [12] and goat genomes [13]. For *Amborella*, this allowed the reordering and super-scaffolding of the draft assemblies and increased their continuity (N50 increased from 4.9 to 9.3 Mb). More recently, the release of the Irys system from BioNano Genomics provided new opportunities to improve the quality and the continuity of genome assemblies [14].

Genetic maps allow the construction of pseudo-molecules by anchoring the assembly on linkage groups that correspond to the chromosomes [15]. Genetic map construction takes advantage of sequence-based genotyping (SBG) [16], genotyping-by-sequencing, and RAD-seq libraries [17] to obtain ultra-dense genetic linkage maps [18]. However, missing data or genotyping errors cause map inaccuracies [19]. Moreover, the physical distance between markers can be very high in genomic regions where the recombination rate is low, which makes it difficult to anchor or orientate scaffolds located in those regions.

Methods used to anchor whole-genome shotgun (WGS) assemblies on genomes have been investigated using several genetic maps to estimate assembly quality, as implemented in MetaMap [5]. The ability of these methods to produce pseudo-molecules also was tested, as reported in Popseq [20] and Allmaps [21]. Allmaps infers the sizes of gaps using the relation between the local recombination rate and the physical distance between two adjacent genetic markers; however, the estimations can be inconsistent considering the inaccuracy of the recombination rate.

Hybrid strategies, combining WGS and genome map data, are likely to help increase the quality of the assembled genome sequence. With this in mind, we developed MaGuS, a modular program that combines a genome map and WGS data. MaGuS can anchor a draft assembly onto a genome map for two applications: quality assessment of a draft assembly by calculating novel metrics, and improvement of the continuity of a draft assembly based on evidence provided by a genome map and high-throughput screening (HTS) data. Here, we detail the MaGuS pipeline and provide an example of its applications using the Arabidopsis TAIR10 genome assembly.

Methods

Arabidopsis thaliana genome assembly

One 350-bp paired-end (PE) (ERX372154) and two 5.35-kb mate-pair (MP) (ERX372148, ERX372150) Illumina sequence libraries from *A. thaliana* were downloaded from the European Nucleotide Archive (ENA). A total of 95.3 Gb of data were obtained representing a coverage depth of 562X of PE and 170X of MP reads. Adapters and primers were removed from the reads, and low quality nucleotides were trimmed from both ends (quality values lower than 20). Reads were also trimmed from their second N to the end and reads longer than 30 nucleotides were kept. Reads that mapped onto run quality control sequences (i.e., the PhiX genome that is used in Illumina sequencing as quality control) were removed. To decrease the number of sequencing errors present in the paired-end (PE) reads, we applied Musket v1.1 [22] with a k-mer size of 26 '-k 26'. We ran Kmergenie v1.5692 [23] on the PE reads to find the best k-mer size for the contig construction step and obtained an optimal k-mer size of 91 bp. For the SOAPdenovo2 [24] assembly, a *de Bruijn* graph was constructed with parameters '-K 91 –R'. We selected contigs that were longer than 500 bp.

We used the PE and MP reads in five different scaffolding programs: SOAPdenovo2, SSPACE [25], SGA [26], BESST [27], and OPERA-LG [28]**.** For SOAPdenovo2, we ran the *map* command with parameter '-k 31', the *scaf* command with parameter '–L 500', and set the minimum number of links in the configuration file as 'pair_num_cutoff=5'. For SSPACE, we manually set the bowtie k-mer size '-l 31' and ran the program with parameter '-k 5'. For SGA and BESST, we first aligned the MP reads onto the contigs using BWA aln [29] with parameter '-l 31'. For SGA, the links file was created using the *sga-bam2de* command with parameters '-n 5 -m 500 --

mina 31 –k 31'. The *astat* file was generated setting '–m 500'. The *scaf* file and the corresponding FASTA file were both created with parameters '–m 500'. For BESST, we chose the optimal k-mer size used for the contig assembly as '-K 91' and ran the program with parameter '-e 5'. For each program, we selected the scaffolds that were over 2 kb in length. For OPERA-LG, we set the k-mer size for scaffolding with option 'kmer=91'. The minimum contig size required for the scaffolding step was fixed as 500 bp with the parameter 'contig_size_threshold=500'. Finally, the number of links to validate a connection between two contigs was assigned with the parameter 'cluster_threshold=5'.

The source code of QUAST was modified to avoid, as much as possible, the detection of misassemblies (relocation, translocation, and inversion) that correspond to false positives. Because Nucmer generated numerous spurious alignments lower than 5 kb in highly repetitive regions, the minimum alignment length in both parts of a misassembly was fixed as 5 kb. Moreover, the gap or overlap size threshold length was increased to 5 kb to detect relocations. By default, QUAST reports misassemblies found within a scaffold only if at least 50% of the scaffold is aligned. We modified this parameter to report all misassemblies regardless of the aligned fraction of a scaffold.

Map-guided scaffolding of genome using MaGuS

First, the WGP tags were aligned to scaffolds using BWA aln [29] and tags with multiple locations were filtered out of the BAM file [30]. We used the resultant alignments to anchor the scaffolds on the genome map and created links between adjacent scaffolds (Figure 1a). However, scaffolds located within other scaffolds, according to the anchoring information, were not considered. More formally, let a tag

$t(c,r)$ be defined by its BAC contig origin c and its rank r. Let a scaffold $s((t_1,p_1),(t_2,p_2),\dots,(t_n,p_n))$ be defined by the n-uplet of a (t_j,p_j) couple, where the tag t_j aligns uniquely at position p_j with $p_j \leq p_{j+1}$. We define an anchored scaffold $A_i\left(t_{a_i}(c_{a_i},r_{a_i}),t_{b_i}(c_{b_i},r_{b_i})\right)$ by the origin of the BAC contigs and the ranks of its leftmost and rightmost tags, $t_{a_i}$and t_{b_i}, with $r_a < r_b$. We define a map-link between two adjacent scaffolds $a_i$and a_j only if $\min(r_{a_i},r_{b_i}) \leq \min(r_{a_j},r_{b_j})$ does not include scaffolds located within other scaffolds.

The MP reads were aligned to the assembly using BWA mem [29] and pairs whose mates mapped to different scaffolds were selected. Multiple hits were recorded and mapping possibilities that confirmed a map-link were kept. We estimated the gap size between two adjacent scaffolds from the set of map-anchored scaffolds using the MP fragment size distribution. If the computed gap size was smaller than the maximum expected gap size derived from the MP library, the map-link and the orientation of the two scaffolds were validated. If multiple scaffold orientations were reported by the read mapping, the one supported by the highest number of read pairs was selected. More formally, Let a mapping possibility of a read pair $((scaf_1,orient_1,pos_1),(scaf_2,orient_2,pos_2))$ be defined by its scaffold name, orientation, and location of both reads with $scaf_1 \neq scaf_2$. For each read pair, we calculate the gap size based on the orientation of the two linked scaffolds inferred by each supporting pair, where len_1 and len_2 are the lengths of $scaf_1$ and $scaf_2$ respectively, R is the read length, and μ is the mean of the mate-pair (MP) library fragment size as:

$$\begin{cases} gap_{++} = \mu - pos_1 - pos_2 - 2R \\ gap_{+-} = \mu - (pos_1 + R) - (len_2 - pos_2) \\ gap_{--} = \mu - (len_1 - pos_1) - (len_2 - pos_2) \\ gap_{-+} = \mu - (len_1 - pos_1) - (pos_2 + R) \end{cases}$$

We validate the link if $0 \leq \frac{1}{n}\sum gap_{(orient_1, orient_2)} \leq \mu + 3.\mathrm{sd}$, where μ and sd are the mean size and the standard deviation of the MP library fragment size respectively, and n is the number of supporting pairs for the scaffolds link with the following orientation $(orient_1, orient_2)$. Finally, all validated links were formatted for the SGA program to perform the final scaffolding.

Analysis of *A. thaliana* WGP data

We used the WGP data produced from the *A. thaliana col-0* BAC library by Keygene (Wageningen, The Netherlands) [9]. WGP tags were ordered by an automated procedure that performed the following steps. First, fingerprinted contig data were read with contig and position information per BAC. Then, BACs were sorted on their left and right positions in the contig and assigned a rank number (identical left and right positions lead to identical ranks). Next, tag information from the WGP tag file was read and occurrences of tags per BAC were listed. For a given contig, a tag position was calculated as the mean value of BAC rank numbers on which the tag occurred. If BAC ranks were too far apart, the tag was identified as an outlier and put aside. The remaining tags were ranked according to their mean BAC rank value, possibly with equal rank scores for equal average BAC rank values.

Quality evaluation of genome assembly using MaGuS

We generated new quality assembly metrics from the anchoring based on the commonly used N50 metric (used to evaluate assembly continuity) and the NA50 introduced by the quality assessment tool QUAST (used to evaluate both continuity and quality of assembly using a genome reference [31]). For each scaffold, we defined collinear segments as the fraction of a given scaffold that was correctly

organized, i.e., segments anchored with tags that have the same order in the genome map and in the scaffolds (Figure 1b). For a given assembly, the lengths of all these segments were used to calculate the following metrics: An50 (50% of the anchored assembly contains collinear segments with length over An50 bp), AnA50 (50% of the total assembly contains collinear segments with length over AnA50 bp), and AnG50 (50% of the estimated genome size contains collinear anchored segments with length over AnG50 bp). MaGuS also generates Anx, AnAx, and AnGx graphs (based on the Nx graph [2]) that is a plot of the metrics for x values ranging from 1% to 100%.

Implementation of MaGuS

MaGuS was implemented in a Perl program based on five modules: *wgp2map*, which performs the anchoring and creates a MaGuS-format map that contains the anchoring information; *map2qc*, which evaluates the quality of the assembly; *map2link*, which creates the map-links between scaffolds; *pairs2links*, which validates the map-links, orients the scaffolds, estimates the gap size, and creates a link.de file; and *links2scaf*, which runs the SGA scaffolding programs and creates the final assembly.

Results and discussion

Arabidopsis genome assembly and quality evaluation using MaGuS

PE reads were assembled into contigs with SOAPdenovo2. Then we generated five assemblies using five scaffolding programs (BESST, SSPACE, SOAPdenovo2, SGA, and OPERA-LG) with PE and MP reads. The BESST assembly had the highest continuity (N50 = 1.3 Mb) followed by OPERA-LG (N50=1.27 Mb), SSPACE (0.98 Mb), SOAPdenovo2 (N50=0.82 Mb), and SGA (N50=0.28 Mb). To evaluate the assembly quality, we aligned the scaffolds against the Arabidopsis TAIR10 reference genome with Nucmer [32] using the QUAST pipeline [31] (see Additional file 1 for

details). We found that although BESST and OPERA-LG created scaffolds that had longer alignments, they also contained relatively more misassemblies than SOAPdenovo2, SSPACE, and SGA. Based on the QUAST NA50 and NA75 metrics, we ranked the assemblies from the highest to lowest quality as BESST, OPERA-LG, SSPACE, SOAPdenovo2, and SGA.

We used the WGP map to provide a reference-free approach that evaluates the quality of the five assemblies. We applied the *wgp2map* and *map2qc* modules of MaGuS to calculate the length of all collinear segments (Figure1b) and generated Anx values (Table 1, Figure 2a). Considering the MaGuS An50 and the An75 metrics, the ranking of the assemblies was the same as the ranking using the QUAST NA50 and NA75 metrics. The NAx and Anx values were strongly correlated (R^2 >0.96) for the five assemblies (Figure 2c), which allowed us to consider using the MaGuS Anx metrics to compare assembly quality.

Selecting the appropriate bioinformatics tools to perform genome *de novo* assembly is difficult and often depends on the genome complexity and on the sequencing technology used. The absence of a reference sequence leads automatically to the selection of the assembly that has the highest continuity with no regards to the quality. In the present case, access to a genome map and its use with MaGuS allowed the BESST assembly to be selected as being the most continuous and also the most collinear to the WGP map.

Arabidopsis genome map-guided scaffolding using MaGuS

We used the five assemblies produced previously to perform map-guided scaffolding through the MaGuS pipeline (Figure 1c). For each assembly, we first created the map-links (i.e., the links between two adjacent anchored scaffolds) and aligned the MP reads onto the scaffolds to validate the map-links by first determining the scaffolds

orientation (if the scaffold was anchored by only one tag) and then by estimating the new gaps size (see Methods). The validated map-links were used to build the final scaffolds (Table 2). Only a fraction of the map-links (21.2% to 49.9%) was validated by the MP reads. This limitation was clearly due to the MP library size, and a higher fraction of map-links would certainly be validated using larger MP libraries. Although only a fraction of the map-links were used for the scaffolding, the resulting assemblies showed increases in the N50 metrics ranging from 1.13 to 2.24 times higher and increases in N75 from 1.23 to 2.43 times higher (Table 2). To evaluate the accuracy of this scaffolding approach, we aligned the five assemblies generated by MaGuS onto the Arabidopsis TAIR10 reference genome using QUAST (see Additional file 1). MaGuS generated 86% to 97% correct links for the five assemblies and only a limited number of misassemblies (Table2). The quality of the scaffolds also was confirmed by elevated NA50 and NA75 values. The number of read pairs that validated a map-link had a very wide distribution, from 1 to over 1 000 read pairs (Figure 2c), which showed that MaGuS enabled the scaffolding of both low covered and highly covered regions that corresponded to repetitive regions.

Conclusions

The method presented here and implemented in MaGuS enabled the evaluation of the quality and the scaffolding of a draft genome assembly using a physical map and HTS data. Its application to Arabidopsis with a WGP map provides a first example of its efficiency in reconstructing a eukaryotic genome. Evaluating the quality of a genome assembly is necessary in order to increase the accuracy of downstream analyses, such as genome annotation or comparative genomic analyses. *De novo* assembly projects

often lack a genome reference and different ways to assess the assembly quality have been investigated [2, 33] using either the HTS data used for the assembly or a genome map. The latter remains a very good independent source of information for this task. From this perspective, we developed the *map2qc* module of MaGuS to provide assembly quality metrics. Its application to five Arabidopsis genome assemblies showed that the new quality metrics based on the correctly anchored segments of the assembly gave the same assembly ranking as if a reference genome was available. Existing scaffolder tools encounter issues when dealing with repeat-rich regions. The use of a map overcomes this problem if a contig or scaffold can be anchored onto the map. For large genomes, the sequencing depth of an MP library may result in low covered regions. Users of scaffolding programs often set a minimum cut-off for read pairs required to validate a link between contigs, to avoid assembly errors. The use of a map to guide the assembly allows this cut-off to be lowered without loss of accuracy.

Availability of supporting data

Arabidopsis Illumina reads can be downloaded from the European Nucleotide Archive (ENA) with the following IDs: ERX372154, ERX372148, ERX372150. The WGP data and MaGuS can be accessed through GitHub at https://github.com/institut-de-genomique/MaGuS.

Competing interests

The SBG and WGP™ technologies are protected by patents and patent applications owned by Keygene NV (Wageningen, The Netherlands). WGP™ is a trademark of Keygene NV.

Authors' contributions

MAM designed the method. MAM and CD implemented the method. MAM, LA, CD, and JVO performed the bioinformatics analyses. EVDV and JVO provided the WGP data. MAM and JMA wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgments

This work was supported by Genoscope (Évry, France), the Commissariat à l'Energie Atomique et aux Energies Alternatives (CEA), France Génomique (ANR-10-INBS-09-08), and KeyGene NV.

Additional files

Additional file 1 – supplementary methods

References

1. Hunt M, Newbold C, Berriman M, Otto TD: **A comprehensive evaluation of assembly scaffolding tools**. *Genome biology* 2014, **15**(3):R42.
2. Bradnam KR, Fass JN, Alexandrov A, Baranay P, Bechner M, Birol I, Boisvert S, Chapman JA, Chapuis G, Chikhi R *et al*: **Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species**. *GigaScience* 2013, **2**(1):10.
3. Bennetzen JL: **Patterns in grass genome evolution**. *Current opinion in plant biology* 2007, **10**(2):176-181.
4. Lewin HA, Larkin DM, Pontius J, O'Brien SJ: **Every genome sequence needs a good map**. *Genome research* 2009, **19**(11):1925-1928.
5. Servin B, de Givry S, Faraut T: **Statistical confidence measures for genome maps: application to the validation of genome assemblies**. *Bioinformatics* 2010, **26**(24):3035-3042.

6. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W *et al*: **Initial sequencing and analysis of the human genome**. *Nature* 2001, **409**(6822):860-921.
7. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P *et al*: **Initial sequencing and comparative analysis of the mouse genome**. *Nature* 2002, **420**(6915):520-562.
8. Choulet F, Alberti A, Theil S, Glover N, Barbe V, Daron J, Pingault L, Sourdille P, Couloux A, Paux E *et al*: **Structural and functional partitioning of bread wheat chromosome 3B**. *Science* 2014, **345**(6194):1249721.
9. van Oeveren J, de Ruiter M, Jesse T, van der Poel H, Tang J, Yalcin F, Janssen A, Volpin H, Stormo KE, Bogden R *et al*: **Sequence-based physical mapping of complex genomes by whole genome profiling**. *Genome research* 2011, **21**(4):618-625.
10. Philippe R, Choulet F, Paux E, van Oeveren J, Tang J, Wittenberg AH, Janssen A, van Eijk MJ, Stormo K, Alberti A *et al*: **Whole Genome Profiling provides a robust framework for physical mapping and sequencing in the highly complex and repetitive wheat genome**. *BMC genomics* 2012, **13**:47.
11. Sierro N, van Oeveren J, van Eijk MJ, Martin F, Stormo KE, Peitsch MC, Ivanov NV: **Whole genome profiling physical map and ancestral annotation of tobacco Hicks Broadleaf**. *The Plant journal : for cell and molecular biology* 2013, **75**(5):880-889.
12. Chamala S, Chanderbali AS, Der JP, Lan T, Walts B, Albert VA, dePamphilis CW, Leebens-Mack J, Rounsley S, Schuster SC *et al*: **Assembly and validation of the genome of the nonmodel basal angiosperm Amborella**. *Science* 2013, **342**(6165):1516-1517.
13. Dong Y, Xie M, Jiang Y, Xiao N, Du X, Zhang W, Tosser-Klopp G, Wang J, Yang S, Liang J *et al*: **Sequencing and automated whole-genome optical mapping of the genome of a domestic goat (Capra hircus)**. *Nature biotechnology* 2013, **31**(2):135-141.
14. Lam ET, Hastie A, Lin C, Ehrlich D, Das SK, Austin MD, Deshpande P, Cao H, Nagarajan N, Xiao M *et al*: **Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly**. *Nature biotechnology* 2012, **30**(8):771-776.
15. Mascher M, Stein N: **Genetic anchoring of whole-genome shotgun assemblies**. *Frontiers in genetics* 2014, **5**:208.
16. Truong HT, Ramos AM, Yalcin F, de Ruiter M, van der Poel HJ, Huvenaars KH, Hogers RC, van Enckevort LJ, Janssen A, van Orsouw NJ *et al*: **Sequence-based genotyping for marker discovery and co-dominant scoring in germplasm and populations**. *PloS one* 2012, **7**(5):e37565.
17. Davey JW, Blaxter ML: **RADSeq: next-generation population genetics**. *Briefings in functional genomics* 2010, **9**(5-6):416-423.
18. Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML: **Genome-wide genetic marker discovery and genotyping using next-generation sequencing**. *Nature reviews Genetics* 2011, **12**(7):499-510.
19. Cheema J, Dicks J: **Computational approaches and software tools for genetic linkage map estimation in plants**. *Briefings in bioinformatics* 2009, **10**(6):595-608.
20. Mascher M, Muehlbauer GJ, Rokhsar DS, Chapman J, Schmutz J, Barry K, Munoz-Amatriain M, Close TJ, Wise RP, Schulman AH *et al*: **Anchoring and**

ordering NGS contig assemblies by population sequencing (POPSEQ). *The Plant journal : for cell and molecular biology* 2013, **76**(4):718-727.

21. Tang H, Zhang X, Miao C, Zhang J, Ming R, Schnable JC, Schnable PS, Lyons E, Lu J: **ALLMAPS: robust scaffold ordering based on multiple maps**. *Genome biology* 2015, **16**(1):3.
22. Liu Y, Schroder J, Schmidt B: **Musket: a multistage k-mer spectrum-based error corrector for Illumina sequence data**. *Bioinformatics* 2013, **29**(3):308-315.
23. Chikhi R, Medvedev P: **Informed and automated k-mer size selection for genome assembly**. *Bioinformatics* 2014, **30**(1):31-37.
24. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y *et al*: **SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler**. *GigaScience* 2012, **1**(1):18.
25. Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W: **Scaffolding pre-assembled contigs using SSPACE**. *Bioinformatics* 2011, **27**(4):578-579.
26. Simpson JT, Durbin R: **Efficient de novo assembly of large genomes using compressed data structures**. *Genome research* 2012, **22**(3):549-556.
27. Sahlin K, Vezzi F, Nystedt B, Lundeberg J, Arvestad L: **BESST--efficient scaffolding of large fragmented assemblies**. *BMC bioinformatics* 2014, **15**:281.
28. Gao S, Bertrand D, Nagarajan N: **OPERA-LG: Efficient and exact scaffolding of large, repeat-rich eukaryotic genomes with performance guarantees**. *biorxiv* 2015.
29. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform**. *Bioinformatics* 2009, **25**(14):1754-1760.
30. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R: **The Sequence Alignment/Map format and SAMtools**. *Bioinformatics* 2009, **25**(16):2078-2079.
31. Gurevich A, Saveliev V, Vyahhi N, Tesler G: **QUAST: quality assessment tool for genome assemblies**. *Bioinformatics* 2013, **29**(8):1072-1075.
32. Delcher AL, Salzberg SL, Phillippy AM: **Using MUMmer to identify similar regions in large sequence sets**. *Current protocols in bioinformatics / editoral board, Andreas D Baxevanis [et al]* 2003, **Chapter 10**:Unit 10 13.
33. Earl D, Bradnam K, St John J, Darling A, Lin D, Fass J, Yu HO, Buffalo V, Zerbino DR, Diekhans M *et al*: **Assemblathon 1: a competitive assessment of de novo short read assembly methods**. *Genome research* 2011, **21**(12):2224-2241.

Figure legends

Figure 1 - MaGuS pipeline

a Flowchart of the MaGuS pipeline**. b** Comparison of the QUAST and MaGuS metrics. **c**

Application of MaGuS to WGP data**.**

Figure 2 - Comparison of MaGuS and QUAST quality metrics for the five assemblies

a MaGuS Anx plot. **b** QUAST NAx plot. **c** Correlation between Anx and NAx values.

Figure 3 - Distribution of the number of mate-pairs that validates map-links for the five assemblies

Tables

Table 1 - QUAST and MaGuS quality metrics for the five assemblies

The R^2 values indicate the Pearson correlation coefficients between the QUAST NAx and MaGuS Anx values.

Assembly metrics	SOAP	SSPACE	SGA	BESST	OPERA_LG
Assembly size (bp)	115 319 220	116 017 208	114 956 386	114 996 281	116 406 702
N50 (bp)	821 817	982 887	284 070	**1 299 606**	1 272 891
L50	39	31	115	**22**	26
N75 (bp)	306 051	340 070	118 727	**643 037**	566 836
L75	96	81	270	**54**	60
QUAST metrics					
Number of N's per 100 kbp	3851.60	3000.11	4251.16	2845.19	3139.94
Misassemblies	9	9	**3**	23	51
Largest alignment (bp)	3 482 036	4 678 885	1 680 656	**6 501 653**	5 259 610
NA50 (bp)	757 250	926 429	276 557	**1 210 586**	945 419
NA75 (bp)	268 694	291 099	100 235	**516 026**	351 844
MaGuS metrics					
An50 (bp)	31 217	32 028	23 164	**35 466**	33 908
An75 (bp)	11 887	12 052	6 981	**14 315**	13 113
R^2	0.99	0.98	0.96	0.99	0.96

Table 2 - Assembly metrics after MaGuS scaffolding for the five assemblies

	SOAP	SSPACE	SGA	BESST	OPERA-LG
Assembly size (bp)	115 563 956	116 414 299	115 703 449	115 174 685	116 556 828
N50 (bp)	1 350 715	1 680 424	635 106	1 751 177	1 442 963
N50 fold change	1.64	1.74	2.24	1.35	1.13
L50	23	18	47	18	22
N75 (bp)	509 384	646 442	288 240	787 050	695 198
N75 fold change	1.66	1.9	2.43	1.22	1.23
L75	58	48	110	42	51
Number of N's per 100 kbp	4 055.34	3 331.14	4 869.38	2 995.68	3 264.70
Largest alignment	5 012 555	7 708 756	3 361 051	6 902 343	5 597 743
NA50	1 187 620	1 455 792	579 394	1 407 579	1 258 868
NA50 fold change	1.57	1.57	2.1	1.16	1.18
NA75	354 088	508 625	215 751	609 320	560 902
NA75 fold change	1.32	1.75	2.15	1.18	1.59
Total misassemblies	23	19	19	36	62
Magus misassemblies	14	10	16	13	5
Number of map-links	534	481	1 034	371	368
Number of MP-validated links	209 (39.14%)	214 (44.49%)	516 (49.9%)	93 (25.07%)	78 (21.2%)
Number of correct MP-validated links	195 (36.51%)	204 (42.41%)	500 (48.53%)	80 (21.56%)	73 (19.83%)
False positive rate	6.7	4.7	3.1	14	6.4

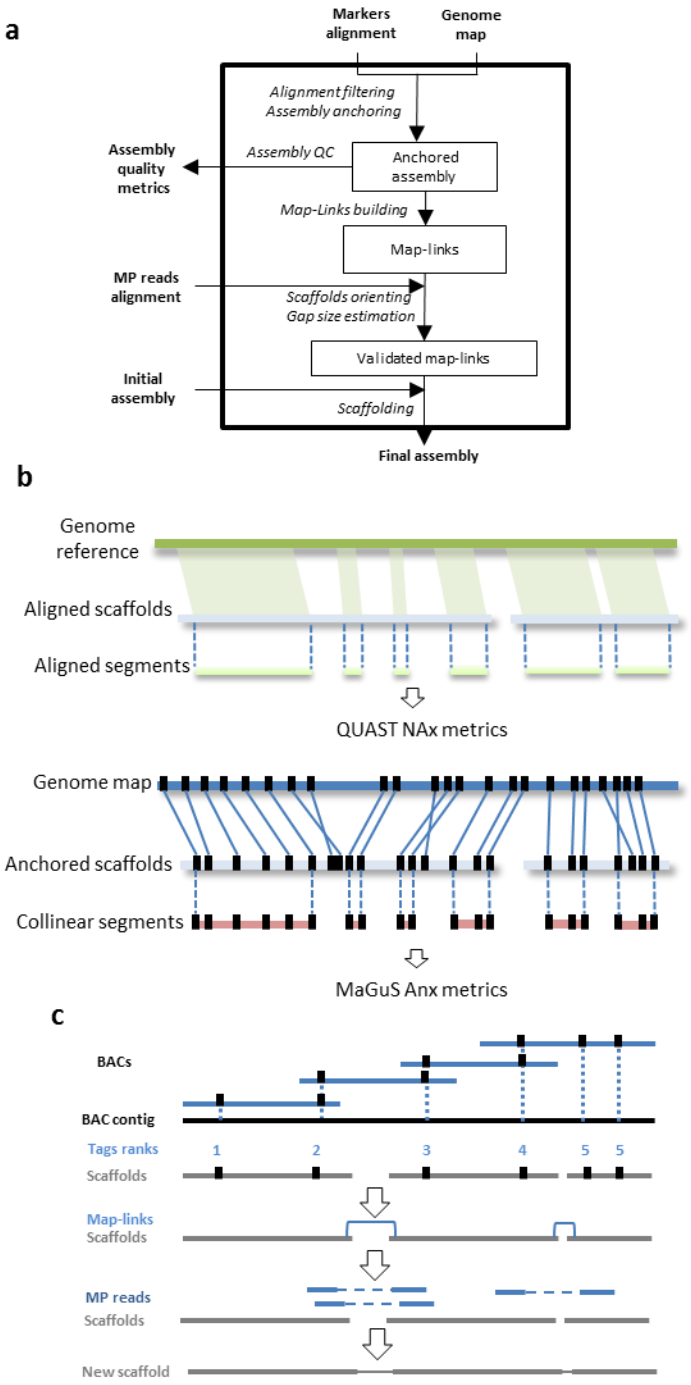

Figure 1

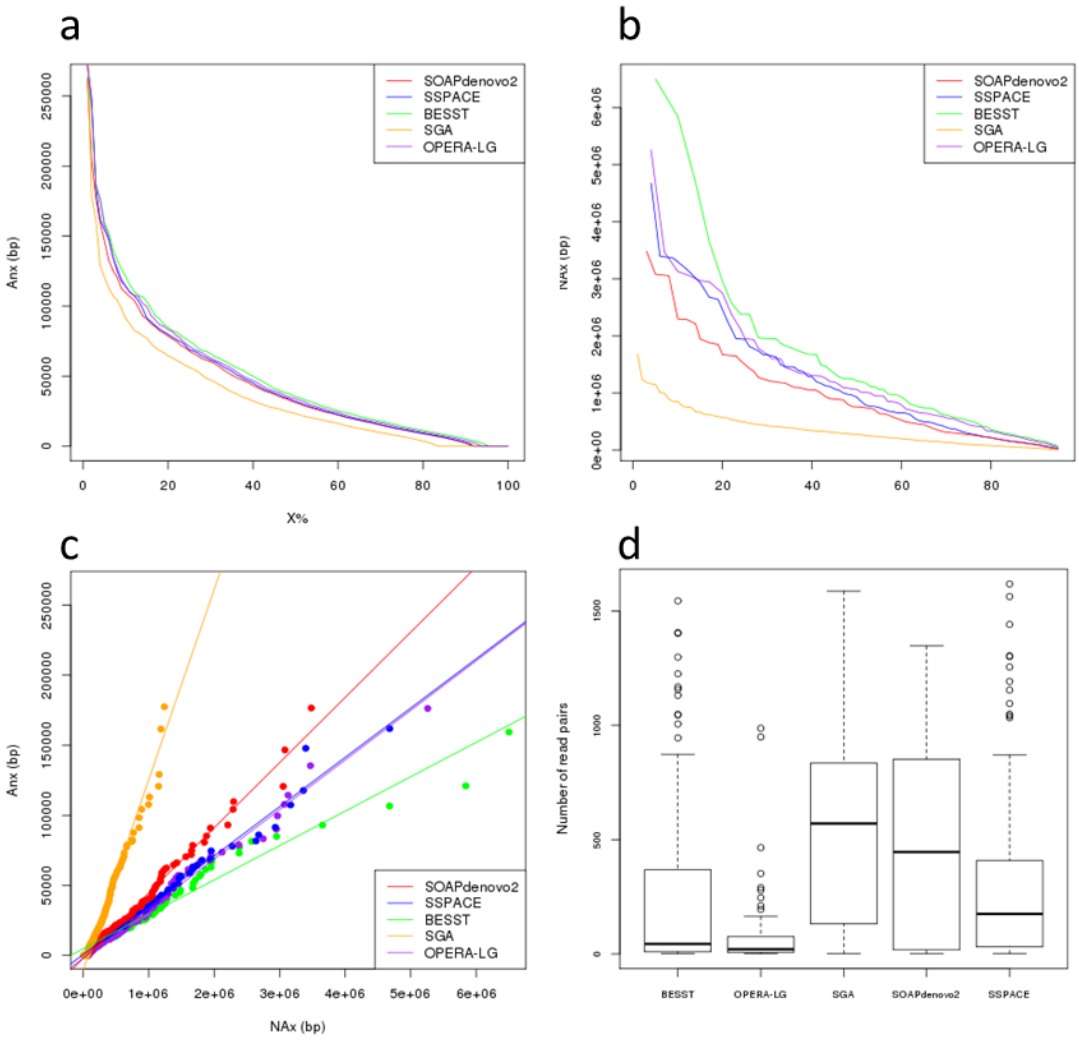

Figure 2