THE UNIVERSITY *of* EDINBURGH
School of Biological Sciences

Quality Assessment of a Newly Sequenced Eukaryotic Genome

Next Generation Genomics

Quality Assessment of the *Solanum stenotomun* Genome

The rate of de novo genome assembly has drastically increased along with the various sequencing technologies and genome assembling methods. Here I will discuss the approach and quality of *Solanum stenotomun's* genome assembly by Yan et al. (2021). Sst will then be compared to (1) *Solanum chacoense* (Sch), a related diploid potato (Leisner et al., 2018), and (2*) Sinonovacula constricta* (Sct), commonly known as the razor clam (Dong et al., 2020). A summary of the assembly and assembly statistics of all three organisms is depicted in Table 1.

Genomic studies of potato varieties, like the diploid plant *Solanum stenotomun* (Sst), is a reserve of gene and allelic diversity which may improve the yield, yield quality, disease and climate change resistance of the cultivated potato, *Solanum tuberosum* (Yan et al., 2021). As the world's third most important food crop, feeding over 1 billion people, it is important to find gene and allelic variation to assure our future food security as potato production is threatened by disease and climate change (International Potato Center, n.d.).

S. stenotomun Whole Genome Assembly

Sst read library consisted of paired-end and mate-end reads of various sizes (200bp – 20kb). Larger reads can span over repeating sequences while short reads can fill the gap between larger reads, allowing for longer contigs and higher consensus sequence accuracy. The prepared library was sequenced using the Illumina HiSeq 250 instrument, it produced 198.03 Gb of raw Illumina reads and a sequence coverage of 237.76x.

Illumina's sequencing errors were corrected by the BFC, a high-performance tool that uses a non-greedy spectrum alignments algorithm, specifically designed for high-coverage whole genome data. Not only does BFC remove error, but it does not overcorrect either, increasing the reliability of the reads processed (Li, 2015).

SOAP-denovo2 assembled the corrected short reads with a k-mer size of 8bp. The relatively small k-mer could be an attempt to decrease sequencing errors while assembling at low coverage depth (Luo et al., 2012). However, you risk increasing the association of unrelated reads. It is also of note that SOAP-denovo2 only supports k-mers between 12- 63/127bp (Luo, 2022). The authors' rationale for the small k-mer was not explained in the paper, nor was I able to find other assemblies using such small k-mers. Platanus GapCloser and SSPACE were used to for additional gap filling and scaffolding to improve the assembly's quality. As a result, a 769.59Mb assembly was produced with 39,308 scaffolds, a contig N50 of 41kb, and a scaffold N50 of 1.49Mb. The predicted genome size was estimated to be 847Mb, the preliminary assembly was ~80Mb bellow this value and has a percentage difference of 9.58%. This could be due to the k-mer size and SOAPdenovo2's sensitivity for repeated sequences.

Further genome manipulation and quality analysis was conducted. Self-to-self BLAS removed redundancy by summarising hits with large identity and coverage values. Short reads were mapped to the assembly, producing a mapping rate of 99.4% and a coverage of 99.66%. CEGMA checked the completeness of core embryophyta genes and BUSCO checked for the orthologous –single copy genes completeness within the assembly; they produced values of 97.18% and 95%, respectively. BWA (burrows-wheeler alignment) mapped the reads against the assembly to estimate the assembly consensus quality value, estimated to be 34.1. Free-Bayes conducted error detection, as it can finds SNP, indels, multi-nucleotide polymorphisms and complex events (Garrison and Marth, 2012).

Next generation mapping was performed on high molecular weight Sst DNA with the BioNano Irys system. This system labels specific motifs of high molecular weight genomic DNA, then creates an optical genome map (Lewin et al., 2009). BioNano's proprietary assembly tools produced a 358.9 Gb map with a 424-fold coverage. The map was compared to the assembled sequence to construct super scaffolds and consequently produce a higher quality de novo assembly. The hybrid assembly produced 852.85 Mb from 39,554 scaffolds and an improved N50 of 3.7Mb. The final assembly size is ~5Mb greater than the estimated genome size and has a percentage difference of 0.68%, the hybrid approach increased the accuracy of the assembly.

Sst's genomic DNA was cross-linked and fragmented into Hi-C libraries which were sequenced on an Illumina HiSeq 2500 instrument. 120Gb of raw Illumina data was produced and filtered to obtain 122.99 million valid interaction pairs for chromosome-level assembly. LACHESIS (Burton et al., 2013) used the pairs for long-range scaffolding. The 1,034 scaffolds produced, containing 788.75Mb, were anchored and orientated into 12 pseudochromosomes with an anchor rate of 92.4%

Yan et al. (2021) were able to generate the de novo genome of Sst by assembling Illumina short-reads with the Soap-denovo2 (v. May 2017) assembly and auxiliary tools, followed by next-generation mapping technology. A final assembly sized 852.85Mb was created with 29,554 scaffolds, the longest being 20.5 Mb, and with a scaffold M50 of 3.7Mb.

Critical Comparison to other Published Genomes

The potato samples (Sst, Sch) Illumina short-reads (refer to Table 1 for their lengths) were assembled by SOAPdenovo2 and ALLPATHS-LG, respectively. The two algorithms utilise the Bruijn approach and are used to assemble large genomes (Chu et al., 2013). The SOAPdenovo2 generated a 769.59 Mb preliminary assembly from 39,308 scaffolds, with the longest being 5.79Mb, with an N50 of 1.49Mb. Whereas ALLPATHS generated an 825.77Mb assembly from 8,260 scaffolds (4.7x fewer scaffolds), with the longest being 7.38Mb, with a 2 fold lower N50 of 713.6 Kb. This a data aligns with Luo et al. (2012)'s paper evaluating the performance of SOAPdenovo2, stating that ALLPATHS has a lower assembly quality than SOAPdenovo2. Optical mapping of Sst increased the longest scaffold to 20.4Mb (2.7x longer than Sch) and the N50 to 3.7Mb (5-fold larger than Sch). Sst assembled genome was closer to its estimated genome (percentage difference= 0.68%), than that of Sch (percentage difference= 6.69%). Finally, the BUSCO score for both were 95% and 96% respectively.

Five observations can be made from the assembly comparison: (1) The longer reads used in Sst may have benefited scaffold formation, (2) SOAPdenovo2 assembly seems to produce higher quality assemblies, (3) the hybrid assembly of Sst greatly improved the assembly, (4) Sst assembly may have benefited from the extensive use of pre- and post- assembly accessory tools, and (5) SOAPdenovo2 and ALLPATH's BUSCO result are similar suggesting that the two assemblies are equally proficient at producing the universal single-copy orthologs.

Sch used genetic mapping using two reference genomes to achieve chromosome level assembly. 508.15 Mb scaffolds were anchored into 12 pseudochromosomes at a 62% anchor rate. Sst anchored 1.5 times more data into 12 pseudochromosomes with an increased anchor rate of 92.4% using Hi-C. The increased accuracy throughout Sst's pipeline, and a better chromosome assembly method allows for a better representation of its genome than that of Sch.

Sct long-reads were sequenced on the PacBio platform, and its short-reads on Illumina. Compared to short-reads, long reads can span across long segments of the genome and reach beyond repetitive sequence regions, detect and fill gaps, identify structural variation and more. However, it is error prone (but improving) and offer lower coverage compared to short-reads. These drawbacks can be remedied by aligning short-reads to the long-read assembly. The WTDBG package was picked for the Sct assembly as it can process very long reads produced by PacBio (Ruan and Li, 2020). To correct long-read assembly errors, the Illumina paired-end reads were aligned to the assembly using BWA. Additional accessory tools were used to improve the assembly's accuracy.

The pipeline produced a 1,331.28 Mb assembly with a contig N50 of 678.9kb, a scaffold N50 of 57.99Mb containing 7,930 scaffolds. This was followed by proximity ligation using Hi-C technology. A total of 154.68 Gb of clean reads were group and anchored into 19 pseudochromosomes using the LACHESIS algorithm.

Sct hybrid-assembly produced a scaffold N50 15.5 times larger than that of Sst with 5 times fewer scaffolds. The larger scaffolds can be explained by the use of long-reads and the inherently larger genome. The larger N50 is expected as it tends to be proportional to the assembly size. Biostar forums state that comparing the N50 of different sized assemblies is meaningless (kbradnam, 2013). I suggest, due to the failure of finding any source proposing a solution, for the Sst and Sct comparison to be based on the ratio between their N50 and assembly length. The N50 : assembly ratio is 0.434 and 0.004 for Sct and Sst respectively. This form of comparison suggests that Sct assembly's N50 value is better relative to its size; this could be due to the use of long-reads and a better and more sensitive pipeline. Organism's whose assemblies are considerably different in size should not be compared.

The chromosome assembly method of Sst and Sct is the same. These cannot be compared as the anchorage rate was missing from Sct's paper. Sct's assembly contains 88.7% of the metazoan orthologs single-copy-genes as assessed by BUSCO and 91.51% by CEGMA's assessment. Sst scored >95% for both assessments. The lower Sct scores may be due to the error-prone third generation sequencing technology. Sct's hybrid (Long-read aligned with short-reads) assembly may have benefited by an independent assembly of the Illumina short reads, followed by an alignment of these two assemblies as it would only require some additional computational power and they have already produced these short-reads. I would not recommend using optical mapping, as was done for Sst, as it may be too expensive and laborious.

The rapid advancement of whole genome assembly technology can be appreciated by observing the difference between the three years separating the Sst and Sch genome assembly. The comparison shows that the use of more sensitive, precise, and accurate algorithms, technology and pipelines help increase the quality of the outcome. The Sst vs Sct shows the diversity of methods and tools available. The rapid development of sequencing technology and bioinformatics tools makes the field dynamic and demand constant study to assure the use of the latest and best tools. The field still needs to solve common problems, such as decreasing sequencing errors, sequencing bias, complex genome regions and computational limitations (Liao et al., 2019). At the rate of the field's development, these issues will dissipate, and we will achieve more authentic genomes.

	Solanum stenotomum (Yan et al., 2021)			*Solanum chacoense* (Leisner et al., 2018)			*Sinonovacula constricta* (Dong et al., 2020)		
Estimated genome size	847 Mb			883 Mb			1244.72 Mb		
Sequencing Technology	**Ilumina HiSeq** Paired-end (250 & 450bp) Mate-pair (2,5,10 & 20 kb)	198.03 Gb raw reads Read depth 237.76x		**Ilumina HiSeq** Paired-end (116, 186 & 250bp) Mate-pair (3.6, 6.2, 6.6 & 9.6 kb)			**PacBio** (>10kb)	101.79 Gb clean reads Read depth 82.01X	
							Illumina Paired-end (350bp)	129.73 Gb clean reads Read depth 104.62X	
Assembly	**SOAP *denovo2*** **K-mer = 8**	Assembly	769.59 Mb	**ALLPATHS-LG**	**Assembly**	**825.77 Mb**	**WTDBG** + Illumina read alignment	Assembly	1,331.97 Mb
		Number of Scaffolds	39,308		**Number of Scaffolds**	**8,260**		Number of scaffolds	7,932
		Maximum Length scaffold	5.79 Mb		**Maximum Length scaffold**	**7.38 Mb**		Contig N50	678.9 Kb
		Scaffold N50	1.49 Mb		**N50**	**713.6 Kb**			
		Contig N50	41 kB						
	Optical mapping (BioNano)	**Hybrid assembly**	**852.85 Mb**						
		Scaffolds	**39,554**						
		Maximum length scaffold	**20.4 Mb**						
		Scaffold N50	**3.7Mb**						
	Proximity Ligation (Hi-C + Lachesis)	Scaffolds anchored	788.75 Mb	**Genetic Mapping** (JoinMap + Monte Carlo mapping algorythm)	Scaffolds anchored	508.15 Mb	**Proximity Ligation** (Hi-C + Lachesis)	Scaffolds anchored	154.68 Gb
		Anchor rate	92.4%		Anchor rate	62%		Linkage groups	19
		pseudochromosomes	12		pseudochromosomes	12	FINAL Assembly	**Assembly**	**1,331.28 Mb**
								Scaffold N50	**57.99 Mb**
Completeness assessment	BUSCO	95%		**BUSCO**	96%		**BUSCO**	88.7%	
	CEGMA	97.18%					**CEGMA**	91.51%	
Additional Tools used	BFC – corrected illumine reads before SOAP*denovo2* SSPACE – additional scaffolding Platanus GapCloser – gap closing of illumine corrected reads Self-to-self Blast – identify redundancy Free-Bayes – calls base pair errors			FASTQC – Illumina paired-end read quality assessment NextClip – Illumina mate-pair-end reads correction Cutadapt – cleaning Illumina mate-end and paired-end reads GapCloser – fills gaps with the aid of 3 additon paired-end libraries			WTDBG-cns – reduces sequencing errors Arrow – PacBio sequence polishing Pilon- corrects draft assemblies and calls variants BWA – aligned Hi-C reads		

Table 1 - Genome assembly statistics of the *Solanum stenotomum, Solanum chacoense* and *Sinonovacula constricta* genomes. The sequencing, assembly and assessments are listed in the chronology of their respective papers, except for ‘Additional tools used’. Values in read are the final assembly statistics.

Bibliography

Burton, J.N., Adey, A., Patwardhan, R.P., Qiu, R., Kitzman, J.O. & Shendure, J. (2013) Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nature Biotechnology*. 31 (12), 1119–1125. doi:10.1038/nbt.2727.

Chu, T.-C., Lu, C.-H., Liu, T., Lee, G.C., Li, W.-H. & Shih, A.C.-C. (2013) Assembler for de novo assembly of large genomes. *Proceedings of the National Academy of Sciences*. 110 (36). doi:10.1073/pnas.1314090110.

Dong, Y., Zeng, Q., Ren, J., Yao, H., Lv, L., He, L., Ruan, W., Xue, Q., Bao, Z., Wang, S. & Lin, Z. (2020) The Chromosome-Level Genome Assembly and Comprehensive Transcriptomes of the Razor Clam (Sinonovacula constricta). *Frontiers in Genetics*. 11, 664. doi:10.3389/fgene.2020.00664.

Garrison, E. & Marth, G. (2012) Haplotype-based variant detection from short-read sequencing. *arXiv:1207.3907 [q-bio]*. http://arxiv.org/abs/1207.3907.

International Potato Center (n.d.) *Potato Facts and Figures*. n.d. International Potato Center. https://cipotato.org/potato/potato-facts-and-figures/ [Accessed: 15 March 2022].

kbradnam (2013) *De-Novo Genome Assemblies Comparison*. 2013. https://www.biostars.org/p/113401/ [Accessed: 20 March 2022].

Leisner, C.P., Hamilton, J.P., Crisovan, E., Manrique-Carpintero, N.C., Marand, A.P., Newton, L., Pham, G.M., Jiang, J., Douches, D.S., Jansky, S.H. & Buell, C.R. (2018a) Genome sequence of M6, a diploid inbred clone of the high-glycoalkaloid-producing tuber-bearing potato species *Solanum chacoense* , reveals residual heterozygosity. *The Plant Journal*. 94 (3), 562–570. doi:10.1111/tpj.13857.

Lewin, H.A., Larkin, D.M., Pontius, J. & O'Brien, S.J. (2009) Every genome sequence needs a good map. *Genome Research*. 19 (11), 1925–1928. doi:10.1101/gr.094557.109.

Li, H. (2015) BFC: correcting Illumina sequencing errors. *Bioinformatics*. 31 (17), 2885–2887. doi:10.1093/bioinformatics/btv290.

Liao, X., Li, M., Zou, Y., Wu, F.-X., Yi-Pan & Wang, J. (2019) Current challenges and solutions of de novo assembly. *Quantitative Biology*. 7 (2), 90–109. doi:10.1007/s40484-019-0166-9.

Luo, R. (2022) *Manual of SOAPdenovo2*. https://github.com/aquaskyline/SOAPdenovo2.

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., et al. (2012a) SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience*. 1, 18. doi:10.1186/2047-217X-1-18.

Parra, G., Bradnam, K. & Korf, I. (2007) CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics*. 23 (9), 1061–1067. doi:10.1093/bioinformatics/btm071.

Ruan, J. & Li, H. (2020) Fast and accurate long-read assembly with wtdbg2. *Nature Methods*. 17 (2), 155–158. doi:10.1038/s41592-019-0669-3.

Yan, L., Zhang, Y., Cai, G., Qing, Y., Song, J., Wang, H., Tan, X., Liu, C., Yang, M., Fang, Z. & Lai, X. (2021) Genome assembly of primitive cultivated potato *Solanum stenotomum* provides insights into potato evolution A. Paterson (ed.). *G3 Genes|Genomes|Genetics*. 11 (10), jkab262. doi:10.1093/g3journal/jkab262.