BIOINFORMATION
Discovery at the interface of physical and biological sciences

BIOMEDICAL INFORMATICS

Open access

www.bioinformation.net
Volume 12(2)

Prediction Model

Optimizing k-mer size using a variant grid search to enhance *de novo* genome assembly

Soyeon Cha & David McK Bird*

Bioinformatics Research Center and Department of Plant Pathology, NC State University, Raleigh, NC, USA; David McK Bird - Email: bird@ncsu.edu;*Corresponding author

Received March 18, 2016; Revised April 6, 2016; Accepted April 6, 2016; Published April 10, 2016

Abstract:
Largely driven by huge reductions in per-base costs, sequencing nucleic acids has become a near-ubiquitous technique in laboratories performing biological and biomedical research. Most of the effort goes to re-sequencing, but assembly of *de novo*-generated, raw sequence reads into contigs that span as much of the genome as possible is central to many projects. Although truly complete coverage is not realistically attainable, maximizing the amount of sequence that can be correctly assembled into contigs contributes to coverage. Here we compare three commonly used assembly algorithms (ABySS, Velvet and SOAPdenovo2), and show that empirical optimization of k-mer values has a disproportionate influence on *de novo* assembly of a eukaryotic genome, the nematode parasite *Meloidogynechitwoodi*. Each assembler was challenged with ~40 million Iluumina II paired-end reads, and assemblies performed under a range of k-mer sizes. In each instance, the optimal k-mer was 127, although based on N50 values,ABySS was more efficient than the others. That the assembly was not spurious was established using the "Core Eukaryotic Gene Mapping Approach", which indicated that 98.79% of the *M. chitwoodi* genome was accounted for by the assembly. Subsequent gene finding and annotation are consistent with this and suggest that k-mer optimization contributes to the robustness of assembly.

Keywords:ABySS, CEGMA, contigs, KmerGenie, N50, next-generation sequencing, SOAPdonovo, Velvet

Background:
The progression of technology from Sanger sequencing to the current "next-generation" platforms has heralded striking reductions in the cost of generating data. Sequencing nucleic acids has become a near-ubiquitous technique in laboratories performing biological and bio-medical research. Sequencing comes in two forms, distinguished by their needs for assembly into a contiguous reconstruction of a larger molecule. Most prevalent are various forms of "re-sequencing" in which the sequencing reads are aligned with a reference genome to reveal bases polymorphic between samples. Computationally, this is not a difficult undertaking. The other mode is the assembly of *de novo*-generated, raw sequence reads into contigs that are, as close as possible a full accounting of the genome of the organism in question. In practice, except for the smallest of genomes, complete coverage is neither attainable nor usually needed. None-the-less, maximizing the amount of sequence that can be correctly assembled into contigs is desirable. Reference-free assembly is based on stacking overlapping sequences of genomic fragments of a defined size (the k-mer), generated by breaking each read into k-mer size. Here we examined three commonly used assembly platforms, and showed that optimization of k-mer values has a disproportionate influence on *de novo* assembly of a eukaryotic genome.

Genome assembly algorithms permit adjustment of k-mer size, and also of the related feature coverage (or depth) of the k-mer assembly. The k-mer optimizing tool "Velvetadvisor"**[1],** for example, estimates a theoretically optimal k-mer size as follows:

$$\text{k-mer} = 1 + \text{read length} - \frac{\text{k-mer coverage} * \text{read length}}{\text{Genome coverage}}$$

where,

$$\text{Genome coverage} = \frac{\text{A total number of reads} * \text{read length}}{\text{Estimated genome size}}$$

Thus,k-mer size and k-mer coverage approximatean inverse relationship.Because k-mer size and coverage impact the assembly, methods to predict optimal k-mer size have been proposed.

In particular, Chikhi and Medvedev (2013) developed the KmerGenie algorithm **[2]** to guide selection of k-mer size, and demonstrated its utility with the assembly tools Velvet **[3]** and SOAPdenovo2 **[4].**

In our lab, we study plant-parasitic, root-knotnematodes (*Meloidogyne* spp.), which are responsible for annual crop losses approaching USD 80 billion worldwide. These pathogens have genomes in the 50 Mbp to 150 Mbp range, with marked differences in gene number betweenspecies. In cool climates, two species (*M. hapla* and *M. chitwoodi*) predominate and appear to occupy the same niche (i.e., are sympatric). Whole genome comparison would likely shed much light on the basis for sympatry. A well-annotated draft sequence is available for *M. hapla* **[5, 6],** and we recently sequenced the *M. chitwoodi* genome; a comprehensive biological annotation will be published elsewhere.

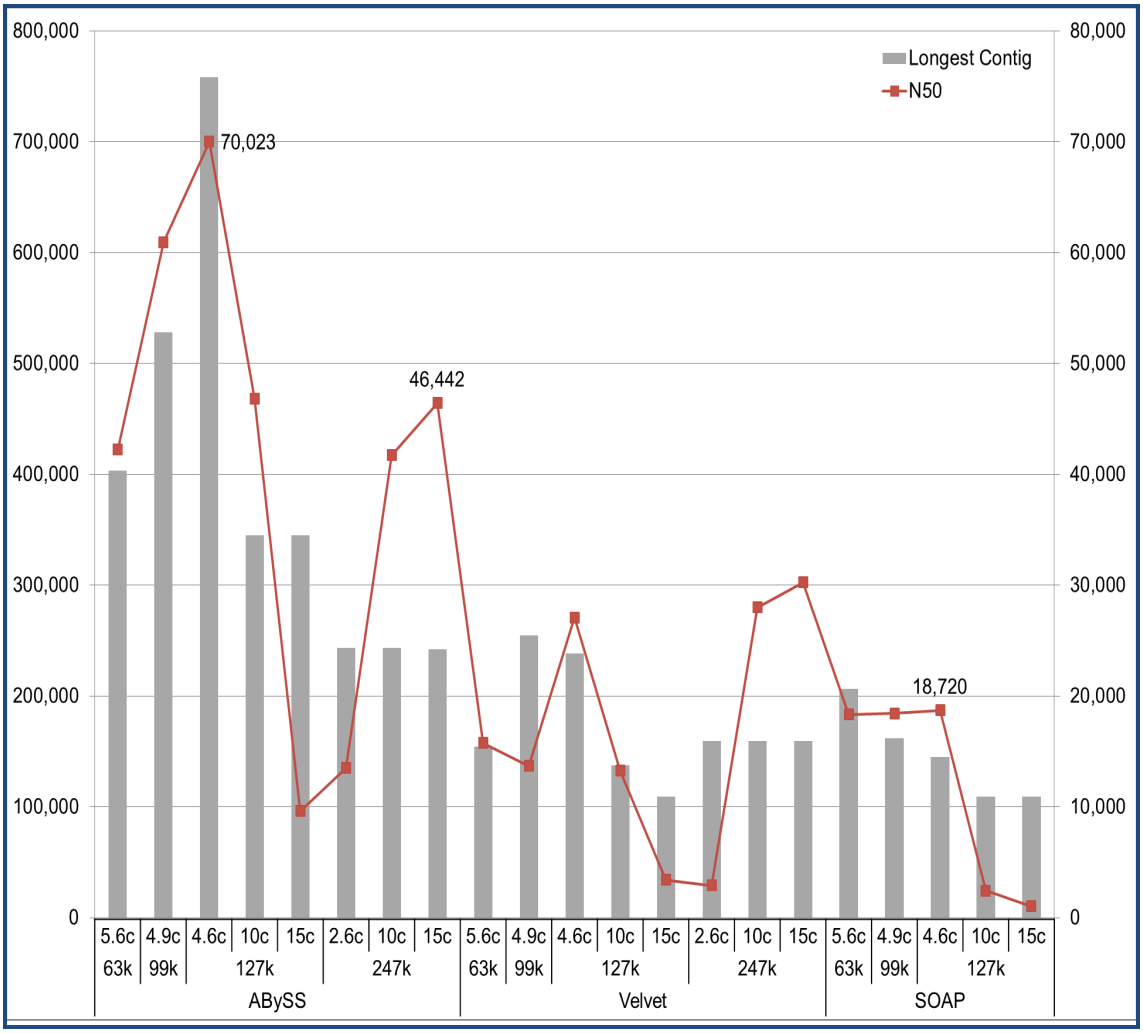

Figure 1: Empirical optimization of k-mer sizes enhances genome assembly across three software platforms (For details, see Table 1-2). X-axis indicates software, k-mer size, and coverage cut off. Y-axis on the left side indicates the length of longest contig (bp) as a function of x-axis, corresponding to grey bars. Y-axis on the right side indicates N50 length (bp), corresponding to red lines. During optimization process, to assess assemblies by N50 (red edges), it is compared of *de novo* assembly of ABySS, Velvet, and SOAPdenovo using different k-mer sizes and coverage cut offs.A more contiguous assembly is obtained for larger N50. At the default coverage thresholds, when k-mer sizes were increased, N50 was overall concave, peaking at 127-mer. When coverage threshold was increased within the same k-mer size, N50 was decreased within 127-mer whereas increased within 247-mer. The length of longest contig (grey bar), though not exactly identical, shows similar pattern as N50. Among the selected k-mers, the largest numbers of N50 and the length of the longest contig were achieved at 127-mer and 4.6 coverage-cut-off by ABySS.

	KmerSize	Cov. Cut Off	Reads OnContigs	# of Contigs	TotalLgth	Reads /Contig	Avg. Lgth	Longest Contig	N50
ABySS	63	5.6	40,630,414	297,634	169,228,606	137	569	403,332	42,265
	99	4.9	38,878,837	185,458	170,576,726	210	920	528,011	60,946
	125	4.6	37,224,400	132,597	169,195,886	280	1,276	758,109	69,778
	127	4.6	37,088,213	128,239	168,988,320	289	1,318	758,111	**70,023**
	129	4.5	36,955,274	126,133	169,001,206	292	1,339	758,113	70,751
	131	4.5	36,815,722	123,000	168,764,165	299	1,372	758,114	69,968
	135	4.4	36,534,686	118,212	168,707,383	309	1,427	733,018	69,506
	161	4.0	34,563,084	92,400	167,205,432	374	1,809	515,211	68,049
	197	3.5	31,610,306	55,721	162,550,195	567	2,917	328,230	55,555
	233	3.0	28,370,399	42,096	159,159,885	673	3,780	243,375	30,450
	247	2.6	27,070,127	44,302	155,691,515	611	3,514	243,375	13,486
	259	2.2	25,107,148	64,765	147,149,476	387	2,272	243,375	4,552
	261	2.2	24,517,414	70,074	143,110,673	349	2,042	243,375	3,690

Table 1: Comparison of de novo assembly over different k-mer sizes, setting other parameters at default. We performed assemblies using k-mer values 63, 99, 161, 197, 233, 247, and 261. The value 247-mer was predicted "Velvet advisor, and 261-mer by "KmerGenie". At the default k-mer coverage-cut-offs, 5.6, 4.9, 4.0, 3.5, 3.0, 2.6, 2.2, and 2.2 respectively,ABySS resulted in gradual increase in N50 from 63-mer to 161-mer and gradual decrease from 161-mer to 261-mer. To investigate more narrow ranges of k-mer, the averaged value of k-mersizes which resulted in two largest N50 (161-mer: 68,049; 99mer: 60,946), 130-mer, was chosen. Surrounding 130-mer, we increased or decreased k-mer size by 2 (125, 127, 129, 131, and 135), resulting in increasing and decreasing N50s (69 778, 70 023, 70 751, 69 968, 69 506).Though 129-mer resulted in a slightly higher N50 (70,751), it wasted about 20 percentages reads from one-end (unaligned reads: 975,453; singleton: 3,104,888; total one-end reads on contigs: 16,925,193 = 21,005,534 - 975,453 - 3,104,888). Thus, to keep more than 80 percentages of reads, we determined to cut at 127-mer which achieved the second largest N50 (70,023) as well as enough amount of information on reads (unaligned reads: 936,337; singleton: 3,050,181; total one-end reads on contigs: 17,019,016 = 21,005,534 - 936,337 - 3,050,181).

Prior to assembly of the *M. chitwoodi* reads, we queried "Velvetadvisor" and "KmerGenie" to compute a value for k-mer size (247 and 260 respectively). Although similar, these values are not identical, and led us to explore empirical optimization of k-mer size. In this study, we show that a 'Simple Grid Search', a widely used optimization algorithm, achieves the best k-mer value for assembly. Our proposed method has three steps. Firstly, we explicitly specified an equally-spaced interval including the k-mer size predicted by 'Velvet advisor' or 'KmerGenie'. Those k-mers were evaluated according to N50.Secondly, we selected a next set of k-mers in a more-narrow interval around those k-mers with the largest N50 from the first evaluation. Lastly, we chose the best k-mer by assessingthe second set of k-mers by taking into account N50 as well as other statistics. We found that assembly size is much more sensitive to k-mer size than has been theoretically estimated **(Figure 1**). Importantly, we found that our empirical approach yielded an assembly with an N50 of 70,023, compared to best N50 values of 46,442 (Velvet advisor) or 42,333 (KmerGenie).

Alone, the N50 value provides no information about the quality of the assembly, which needs to be verified by some independent means. One useful metric is to detect the presence or absence of a set of genes encoding proteins established to be crucial for eukaryotes. The "Core Eukaryotic Gene Mapping Approach" (CEGMA) tool performs such an analysis using a defined database of 458 core proteins **[7].** The percent of that protein set identified serves as a surrogate for the percentage genome coverage by the assembly. Additionally, the highly defined CEGMA proteins identify a reference set from which to unambiguously deduce elements of gene structure, including translation start/stop sites and intron/exon boundaries. Such gene models represent a reliable training set for gene prediction algorithms such as AUGUSTUS **[8].** In our project, we further seeded the gene finders with EST data.Finally, genomic features were elucidated using RepeatMasker**[9],** and functional domains were predicted using InterProScan**[10]** and Blast2GO **[11]** as functional annotation.The results we present hereindicate that the assembly based on empirically-determined k-mers yields not just a larger N50, but also a useful genome assembly.

Methodology:

Data Generation and Processing

Total genomic DNA was isolated from *Meloidogynechitwoodi* collected in a potato field in Washington State, and shipped as an ethanol precipitate to NCSU. Libraries with an average insert size of 700 bp were constructed to facilitate 300 bp paired-end reads, and sequences determined on an IlluminaMiSeq II instrument. Low quality reads (Phred values ≤30) were rejected, and the remainder used for assembly. Because it is likely that different assembly algorithms will give different results in a genome-specific (and an *a priori* unpredictable) manner, we performed k-mer optimization on three commonly-used assembly algorithms, *viz.*, ABySS (version 1.3.7), Velvet (version 1.2.10) and SOAPdenovo (version 2.01).

De novo Assembly

In advanceof fine-tuning parameters, we estimated the recommended k-mer size using "KmerGenie" to be a 260-mer.

We trimmed this to 259-mers to suppress palindromes. The "Velvetadvisor" **[1]**recommended an optimal k-mer to be a 247-mer (coverage cut-off 15). We performed assemblies by employing a variant of a 'Simple Grid Search' where we used k-mers ranging from 259-mer to 63-mer, with intervals of 36. Other attributes were set to the default setting of coverage-cut-off for each algorithm. To assess if the largest N50 is observed at one particular k-mer size among the previously tested set, k1 (k-mer with the largest N50) and k2 (k-mer with the second largest N50) were averaged to k3, (k1+k2)/2. More values of k-mer surrounding k3 at intervals of 2 were performed to identify the optimal k-mer value (Table1-1). The total number of reads aligned to contigs was also taken into account. In addition, for further parameter fine-tuning, results from different coverage-cut-offs other than default settings were compared **(Table2).**SOAPdenovo was run under k-mer sizes equal to or less than 127-mer as it is the maximum k-mer size available in this program. To further validate our method, we arbitrarily selected two organisms for evaluation: the bacteria *Neisseriagonorrhoeae* (assembled genome size 2.15 Mbp) and *Camelpox* virus (assembled genome size 0.20 Mbp). FASTQ files were obtained from the European Nucleotide Archive (ENA) and were *de novo* assembled by ABySSusing k-mer sizes chosen by KmerGenie and Velvet advisor as well as by our empirical methods.

	Kmer Size	Cov. Cut Off	Reads On Contigs	# of Contigs	Total Lgth	Reads /Contig	Avg. Lgth	Longest Contig	N50
ABySS	63	5.6	40,630,414	297,634	169,228,606	137	569	403,332	42,265
	99	4.9	38,878,837	185,458	170,576,726	210	920	528,011	60,946
	127	4.6	37,088,213	128,239	168,988,320	289	1,318	758,111	**70,023**
		10	36,951,014	66,039	158,776,887	560	2,404	344,946	46,837
		15	35,995,149	64,949	145,747,066	554	2,244	344,995	9,625
	247	2.6	27,070,127	44,302	155,691,515	611	3,514	243,375	13,486
		10	17,222,135	10,028	50,018,879	1,717	4,988	243,375	41,713
		15	16,818,494	6,379	48,832,892	2,637	7,655	241,800	46,442
	259	2.2	25,107,148	64,765	147,149,476	387	2,272	243,375	4,552
		10	16,436,784	7,557	49,225,842	2,175	6,513	197,242	42,333
		15	2,264,807	3,573	20,68,931	633	579	22,094	668
Velvet	63	5.6c	38,385,172	344,938	167,922,256	111	487	154,606	15,745
	99	4.9c	34,464,770	344,938	180,340,024	100	523	254,652	13,703
	127	4.6	31,625,371	193,070	173,230,698	164	897	238,111	27,066
		10	31,925,142	115,193	160,122,474	277	1,390	137,609	13,257
		15	31,155,644	121,249	145,123,478	257	1,197	109,145	3,417
	247	2.6	24,473,404	93,236	159,178,884	262	1,707	159,323	2,917
		10	15,327,828	17,643	49,107,159	869	2,783	159,323	27,978
		15	15,088,514	11,075	45,945,149	1,362	4,149	159,323	30,258
Soap	63	5.6	39,536,184	92,498	150,923,645	427	1,632	206,215	18,340
	99	4.9	37,912,050	265,563	175,094,770	143	659	161,631	18,424
	127	4.6	36,453,910	83,451	157,061,370	437	1,882	144,722	18,720
		10	36,094,908	126,143	151,515,830	286	1,201	109,147	2,419
		15	31,333,799	127,446	103,225,180	246	810	109,147	1,018

Table 2: With the selected k-mer sizes, different coverage-cut-offs were compared acrossthree software tools. Empirical optimization of k-mer sizes enhances genome assembly across different software platforms.

Gene Prediction and Automated annotation

To generate an initial training set, we queried our assemblies using CEGMA (version 2.4). To expand the training set, we incorporated cDNAs as evidence obtained from nematode.net and NCBI. These sets were processed using the AUGUSTUS web server (http://bioinf.uni-greifswald.de/webaugustus/)**[12]**forpredictinggenes in genome *ab initio*.Additionally, gene annotations generated by AUGUSTUS were searched by InterProScan and Blast2GO to identify GO terms and gene families. We investigated DNA elements and repeat regions using RepeatMasker (version 4.0.5; http://repeatmasker.org)**[9],** and GC contents using a tool set of Biopieces (http://www.biopieces.org).

Results &Discussion:

Illumina sequencing yielded a total of 42,011,068 paired-end sequence reads (21,005,534 from each end), occupying 27.5 gigabytes in FASTQ format. The average of quality score is about 34. The reads were empirically optimized for *de novo* assembly.Under default settings of coverage-cut-off, the overall trend of N50 was concave, peaking at 127-mer **(Figure 1 &Table1).**We observed that the decrease of N50 within 127-mer and the increase of N50 within 247-mer across all the software we tested as we increased coverage-cut-offs within each k-mer size**(Figure 1 &Table 2**). The largest N50 within 127-mer was still larger than the largest N50 within 247-mer in ABySS. On the other hand, the largest N50 within 127-mer was smaller than the largest N50 within 247-mer in velvet. When compared across software, the largest N50 of 70,023 was achieved by ABySS tool at optimized k-mer size of 127 at the coverage

threshold of 4.6. Thus, our empirical optimization achieved better assemblies than the commonly-used k-mer predictors. For the following further analysis, we elected to use our strategy to optimize the *M. chitwoodi* genome.The genome size of this selected assembly is 152,604,382 (150Mb).

At the protein level, CEGMA predicted, in the *M. chitwoodi* genome, 245 (98.79%) of the 248 core proteins, implying near 100% genome coverage. In addition, it identified 2.23 average number of orthologs per CEG and 94.29% had more than one potential ortholog. This was supported by blasting CEGMA proteins as a query against the assembled contigs as a database, resulting in one protein hit with more than two contigs. This would imply genome duplication or a genome with high heterozygosity, as has been established for *M. incognita***[13]** but not for *M. hapla***[5].**The broad applicability of our approach was demonstrated in diverse species, including a bacterium and a virus. For *Neisseriagonorrhoeae*, the k-mers predicted by Velvet advisor and KmerGenie were 275-mer and 198-mer, respectively, yielding N50s of 28,848 and 44,552. In contrast, our method returned an N50 of 48,678 using a 155-mer. On *Camelpox*, the Velvet advisor k-mer of 301 resulted in a failed assembly. KmerGenie recommended a k-mer of 58, resulting in an assembly with an N50 of 179,206. By contrast, our method yielded an assembly with an N50 of 190,481.

Conclusion:

In assembling a whole genome, it is desirable to achieve a balance between computational costs and the trade-off relationships between k-mer size and its coverage; namely large k-mer size with low coverage or a small k-mer size with deep coverage. Tools "Velvetadvisor" and "KmerGenie" were developed to resolve these problems. As seen in our study, however, those tools cannot be directly applied to the experimental data. Their predicted k-mer sizes gave *de novo* assembly quite different from our empirically optimized assembly of *M. chitwoodi*. This was confirmed by our experiments with two other organisms of bacteria and virus.To overcome this, we showed that our approach, using a variant of a 'Simple Grid Search' to identifyoptimal k-mer size and coverage, led to a more complete assembly. The quality of assembly was confirmed by CEGMA, predicting 98.79% core proteins in the *M. chitwoodi*genome.By integrating different tools of CEGMA and AUGUSTUS, more reliable gene models could be generated. This could also improve the completeness of subsequent analyses, for example, functional analysis or comparative genomics approach. In future studies, we aim to examine the evolutionary history of the genus *Meloidogyne* and how that relates to, or is derived from, attributes germane to parasitism. For example, because *M. chitwoodi* and *M. hapla* are sympatric, they presumably have similar gene compliments.

References:

[1] http: //www. vicbioinformatics.com/velvetk.pl.
[2] Chikhi R &Medvedev P, *Bioinformatics* 2014 **30:**31 [PMID: 23732276]
[3] Zerbino DR &Birney E,*GenomeRes.* 2008 **18:**821 [PMID: 18349386]
[4] Luo R *et al*. *Gigascience*. 2012 **1:**18 [PMID: 23587118]
[5] OppermanCH *et al*. *Proc NatlAcadSci USA*. 2008 **105:**14802 [PMID: 18809916]
[6] Guo Y *et al*. *Worm* 2014 **3:**e29158 [PMID: 25254153]
[7] Parra G *et al. Bioinformatics* 2007 **23:**1061 [PMID: 17332020]
[8] Stanke M *et al. NucleicAcids Res.* 2004 **32:**W309 [PMID: 15215400]
[9] Smit A *et al*.RepeatMasker at http://repeatmasker.org.
[10] Quevillon E *et al. NucleicAcids Res.* 2005 **33:**W116 [PMID:15980438]
[11] Conesa A *et al. Bioinformatics* 2005 **21:**3674 [PMID: 16081474]
[12] Hoff KJ &Stanke M, *NucleicAcids Res.* 2013**41:**W123 [PMID: 23700307]
[13] Abad P *et al*. *Nat Biotechnol*. 2008 **26:**909 [PMID: 18660804]

Edited by P Kangueane
Citation: **Cha & Bird**, Bioinformation 12(2): 36-40 (2016)