Reference-based *de novo* assembly of two *Brassica napus* inbred lines from short read data

Bachelor thesis

for attainment of the academic degree of

Bachelor of Science

presented by

Thomas Lordick

written in the Institute of Quantitative Genetics

Supervisor: Dr. David Ries
First reviewer: Prof. Dr. Benjamin Stich
Second reviewer: Prof. Dr. Bernd Weisshaar

Declaration of Authorship

I hereby confirm that I have written the accompanying thesis by myself, without contributions from any sources other than those cited in the text and acknowledgements.
This applies also to all graphics and images included in the thesis.[1]

…… ……

Place and date Signature

[1] template from https://www.ent.wi.tum.de/fileadmin/w00bcx/www/Ehrenwoertliche_Erklaerung_deutsch_und_englisch.pdf

Table of Contents

1. Abstract

In this project, I present a newly developed multiple-reference-based assembly pipeline which was used to generate four different *de novo* assemblies referring to two parental *Brassica napus* lines. This assembly pipeline relies on multiple reference genomes and short read data only but still provides an exhaustive process of scaffolding and is less expensive than using long PacBio reads for example. Introducing the high quality genomes of *B. napus, Brassica rapa* and *Brassica oleracea* as references, I assembled large scaffolds with a largest scaffold of of 11,5 Mbp, highest N50 of 2 Mbp and an alignment percentage of not less than 99,9 % to the reference. I also shed light on the benefits of using merged assemblies of the same line compared to using the separate assemblies for the pipeline. Through a genome-wide alignment, a further structural variations (SVs) analysis between the assemblies as well as between assembly and *B. napus* reference were carried out. On the one hand, it reveals the amount of differences in genetic material between individuals of the same species. On the other hand, it also shows up the possibility of mixing up real SVs and simple misassembly events corroborated by a comparison to the assembly-to-reference alignment. The assembly pipeline also reveals the current limitations in handling large eukaryotic genomes, especially in the scaffolding process using multiple reference genomes.

2. Introduction

With the availability of a high quality reference genome, assembling of a species' genotypes followed by a genome-wide alignment is a way to call different kind of variations between two or more assemblies and excels calling of variants by mapping of reads to a reference genome (1). Even individuals of the same species often reveal a large number of variations in genetic material, not only by small scale differences like single nucleotide polymorphisms (SNPs) or short insertion and deletions (indels). Copy number variants (CNVs), long indels, inversion and translocations often draw the bigger distinction in genetic material within and between closely related species (2). In the last few years, different projects have been dealing with the reference-based assembling of species, annotating of genes and detecting structural variants (SVs) (2,3,4,5). In 2011, a first reference-based assembly of four different *Arabidopsis thaliana* strains was released. By means of Illumina short-read data and the single reference genome of *A. thaliana*, scaffolds longer than 260 kb (maximum 2,2 Mb) which covered over 96 % of the reference genome were created (2). Five years later, an assembly of the *A. thaliana* Landsberg *erecta* genome was carried out and resulted in 5 chromosome-equivalent sequences of 117 Mb by using s a combination of short Illumina reads, long PacBio reads and linkage information (5). Both project revealed a wide array of large and small scale differences between the *A. thaliana* individuals.

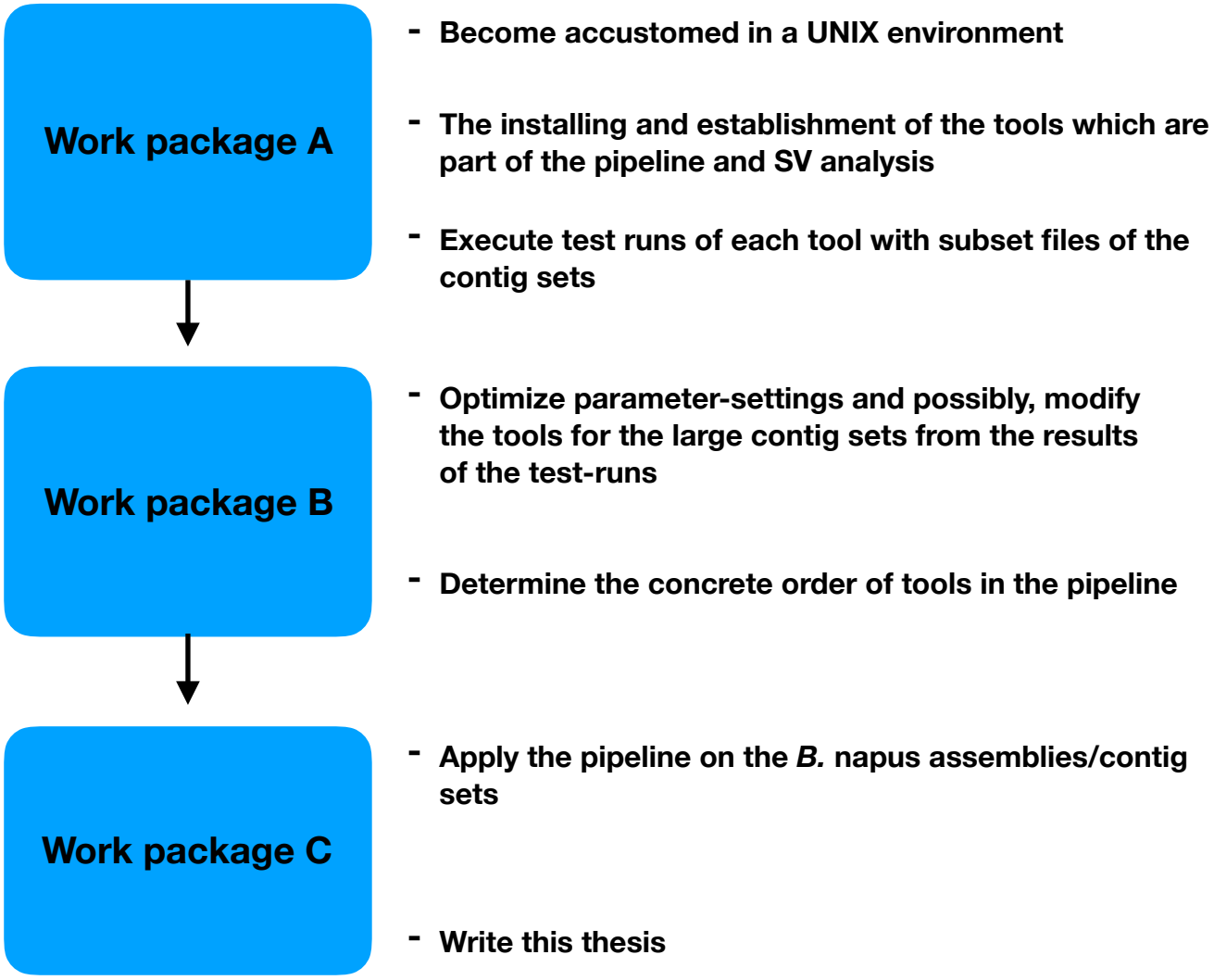
In this project, I firstly present a new multi-reference-based assembly of two *Brassica napus* lines and secondly draw a genome-wide comparison in order to identify SVs. For two lines, short paired-end read data and contigs of de-novo assemblies were available in cooperation with „AG Genomforschung" from the University of Bielefeld. I created a pipeline of tools in order to merge

those contig-sets referring to the same line and generate a second de-novo assembly and especially build scaffolds by means of short read data and multiple high quality reference genomes. As distinguished from other projects, this method is out to be less expensive because it relies on short short read data only and should provide a more exhaustive scaffolding process, because it is guided by using multiple related high quality reference sequences. By this, scaffolds up to chromosome-arm length can be assembled, without the need to apply expensive long read sequencing techniques, mate-pair libraries or the generation of a mapping population. Additionally, I validated and evaluated the assemblies after each step of the pipeline. As the main reference, I made use of the homozygous *B. napus* genome of European winter oilseed cultivar ‚Darmor-*bzh*'. The genome has been assembled with long-read 454 GS-FLX+ Titanium and Sanger sequences in 2015 (6). It was formed ~ 7.500 years ago by hybridization between *Brassica rapa* and *Brassica oleracea*, followed by chromosome doubling, a process known as allopolyploidy (6). These facts moved me to make use of the *B. rapa* and *B. oleracea* genome as additional reference genomes, especially for the scaffolding process (7,8),. In this way, more possibilities are created to build accurate scaffolds when contigs are being sorted by syntheny to multiple references which share the majority of its genetic material, in comparison to mapping to a single reference. Additionally, it is possible, that the lines provide genetic material that occurs in *B.rapa* and/or *B.oleracea,* but doesn't occur in Darmor-*bzh*. For the final comparison, I carry out an genome-wide alignment. In this way, genes will be predicted and SVs identified.

3. Objective, work plan and workflow

The objective of this project is the establishment of a functional pipeline which carries out a multiple-reference-based assembly of four *Brassica napus* contig sets. This pipeline should consist of tools which provide the merging, scaffolding and gapclosing of large, eukaryotic contig sets only by means of short read data and multiple, high quality reference genomes. After applying the assembly pipeline, a genome-wide alignment with a following SVs analysis will be carried out. Thereby, differences in genetic material between the assemblies as well as between the assemblies and the *B. napus* should be detected.

The following flow chart illustrates the intended **work plan** for this project:



The following table presents the intended workflow of this project by weeks. Twelve weeks are planed to realize the previously shown work packages plus the whole reproduction in this theses.

Table 0: Workflow

	Week											
Work Package		1	2	3	4	5	6	7	8	9	10	11	12
	A	■	■	■	■	■	■						
	B					■	■	■	■	■			
	C									■	■	■	■

4. Results

4.1. Source material

For the reference based assembly I used Hiseq 1500 generated paired-end reads of 70 - 142 bp length with an an insert size distribution of 397 - 740 bp as well as Hiseq 2500 generated paired-end reads of 80 - 251 bp with an insert length distribution of 190 - 594 bp (Table 1). The reads and four initial CLC (9) *de novo* assemblies by Daniela Holtgräwe from the AG Genomforschung (University of Bielefeld) were available (all CLC reports are provided in the supplement). Working names were: S2, S3, BnLorenz and BnJanetzki. By two of those assemblies had been made for one parental *B. napus* line. A commen metric to assess assembly quality is N50 which is the length for which the collection of all contigs of that length or longer covers at least half the assembly (10). Those values reached from 1.980 bp to 13.678 bp. The contig count of S3 (437.184) substantially differs from the other assemblies which count a total of less than 140.000 for each of them. The GC content of BnJanetzki reveals a relatively lower value of 31,5 % than the other assemblies which come to values of 34,5 - 35,6 %. The number of reads for S2 comes to a three times higher value (~ 300.000.000) than for the other assemblies which count reads of ~ 100.000.000 by number for each of them. S2 also reveals no differences in read length (all reads scale 251 bp) compared to the other assemblies from which read lengths are typically normally distributed (CLC Reports, Supplement). In the following table (Table 1) the read and CLC *de novo* assembly statistics are listed.

Table 1. De-novo assembly statistics before

	S2	S3	BnLorenz	BnJanetzki
Contigs				
Count	89.306	437.184	119.729	137.972
av. Length (bp)	7.426	1.145	4.132	3.503
N50 (bp)	13.678	1.980	6.796	5.726
GC content (%)	35,6	35,5	34,8	31,5
Reads				
Count	303.379.689	92.527.457	103.171.816	94.378.940
Distribution of read length (bp)	251	80-241	70-142	70-142
Insert size distribution (bp)	190-594	-*	440-740	397-687

*The insert size distribution for the S3 assembly was not available from the CLC report.

Prior to the work of this project, two of the assemblies with the working names BnLorenz and BnJanetzki had been assembled with CLC 7.5 with default parameters. To ensure, that only high quality reads were used, *trimmomatic* (11) had been applied for adapter trimming of the Illumina reads before. The contigs were mapped on the B.napus reference genome using *blat* (12) and ordered by their matching position. S2 and S3 had been assembled with CLC 7.5 (default parameters) from PCR free reads. These assemblies were downloaded from the server of the Center for Biotechnology in Bielefeld and were then submitted to the pipeline developed in this work.

4.2. Multi-reference based assembly

BnLorenz and BnJanetzki were merged with the referring assemblies S2 and S3 using *minimus2-blat* from the *AMOS* open source package (13) : S2 + BnLorenz which originate from the line „Lorenz“ and S3 + BnJanetzki which originate from the line „Janetzkis Schlesischer“ were joined together,

respectively. The working names were S2_BnLorenz and S3_BnJanetzki. Through the merging, I tried to reduce the number of redundant fragments which are present in each assembly as well as concatenate overlapping fragments (Fragments are defined both as contigs and/or scaffolds.). *Minimus2-blat* normally uses a *blat*-based overlap detector which I modified to use *pblat* (14) because It works multi-threaded and therefore much faster than *blat.* This was necessary because of long run time by using *blat* (Material & Methods). From this point on, the merges and both of the assemblies from PCR-free sequencing, working names S2 and S3, were further processed and evaluated in parallel.

In order to reduce the number of amigious bases (Ns) within the contigs and scaffolds, I used *sealer* (15) as a gapclosing tool. It uses latent information in the raw reads to close stretches of Ns within the contigs or scaffolds (Material & Methods). It also extends them by means of the reads. I executed this tool two times per assembly: one time after merging and a second time right after scaffolding with *sspace* with different Kmer sizes, because *sspace* is likely to introduce new Ns during the scaffolding process (SI Material & Methods). In the next step, I executed *sspace* as a scaffolding tool (16). It uses short read data to join 2 or more fragments together to scaffolds. For the two sets of merged assemblies, both sets of corresponding reads were concatenated and used as input for *sealer* and *sspace*.

The final step was the use of the multi-reference based scaffolding tool *medusa* (17). For this purpose, I introduced the genomes of *B. oleracea* and *B. brapa* beside *B. napus* as reference genomes. *Medusa* links a non-contiguous series of fragments together in a graph-based approach. Fragments that correspond to each other will be separated by gaps (stretches of Ns) of known length are estimated by mapping them on the reference

genomes and joined together as scaffolds. Because of a relatively long run time, I had to filter our draft assemblies right before using *medusa* and additionally modify *medusa* (Material & Methods). For instance, for the S2 assembly, I removed all fragments <= 10.000 bp. In this way, 67.527 from 89.296 initial fragments were filtered out. Those remaining 21.769 fragments which are less than 25 % of the initial set still remain ~ 61 % of all bases. The filtered assemblies plus the *medusa* results are provided in Table 3.

4.3. Assembly Validation & Statistics

To asses the assembly quality, I executed *quast* (10, Material & Methods) after each pipeline step. I made usage of the *B. napus* genome as athe reference genome. The draft genomes were aligned against the reference genome after each step to compute various metrics and statistics (Table 2 & 3). That is how I could evaluate in which way the different tools of the pipeline worked on our draft assemblies. Different to previous *quast* executions, the gene-finding option in the last *quast* execution was included right after the *medusa* multiple-reference scaffolding process. Furthermore, read data to call SVs between the assemblies and the *B. napus* reference genome was included. Important information of the *quast* report for each assembly after each pipeline step until the usage of *quast* is provided in the four tables (Table 2.1 - 2.4). Fragment distributions are illustrated in the referring figures (Fig 2.1 - 2.4). The whole report can be found in the supplement. The following two tables (Table 2.1 - 2.2) and their referring figures (Fig 1.1 - 1.2) provide information about the merged assemblies S2_Lorenz and S3_Janetzki.

Table 2.1 Genome statistics for the S3_Janetzki merge and its origins S3 and BnJanetzki

Genome statistics for S3_Janetzki	**S3 raw**	**BnJanetzki raw**	**Merge after MINIMUS2-BLAT**	**Merge after SEALER**	**Merge after SSPACE**	**Merge after SEALER (2nd run)**
Genome fraction (%)	52.754	55.945	39.41	41.06	41.051	41.069
Total length (bp)	500.619.028	483.452.794	477.161.055	477.314.148	477.293.031	477.296.369
N50 (bp)	2.277	5.726	7.263	7.269	7.321	7.321
L50	57.284	24.886	20.009	20.009	19.862	19.862
NG50	604	1.756	2.444	2.449	2.462	2.462
LG50	65.070	80.769	62.807	62.746	62.351	62.350
Unaligned						
Fully unaligned fragments	23.518	3.441	4.218	3.935	3.924	3.924
Fully unaligned length (bp)	22.193.808	3.090.466	6.222.574	5.747.045	5.719.395	5.719.407
Mismatches						
N's	0	9.662.496	32.420.365	28.766.780	28.768.966	28.729.850
N's per 100 kbp	0	1998.64	6794.47	6026.88	6027.53	6019.29
Statistics without reference						
Fragments (>= 0 bp)	437.184	137.972	97.742	97.742	97.176	97.176
Fragments (>= 1000 bp)	157.093	108.001	89.977	89.968	89.524	89.525
Fragments (>= 5000 bp)	10.262	30.341	33.765	33.771	33.766	33.766
Fragments (>= 10000 bp)	673	8.204	10.851	10.862	10.943	10.943
Fragments (>= 25000 bp)	5	278	530	530	548	548
Fragments (>= 50000 bp)	1	2	7	7	8	8

L50 is the number of contigs equal or longer than N50. NG50 is the length for which the collection of all contigs of that length or longer covers at least half the reference genome. According to that, LG50 is the number of contigs equal or longer than NG50 (Quast Manual 4.6.0).

The genome fraction which is considered to be the percentage of aligned bases in the reference genome, decreased significantly through the merging. The raw S3 assembly reveals a genome fraction of 52,57 %, while BnJanetzki reveals ~ 55,95 %. Through the merging, this value sank down to 39,41 %. Over the rest of the pipeline, it could stagnate at ~ 41,1 % after the second *sealer* execution. Basically, both N50 and NG50 values increased, while L50 and LG50 values decreased over the pipeline. The most significant increase occurred after *minimus2-blat* merging of the origin assemblies S3 and BnJanetzki. Similar changes in value occurred for L50 and LG50 in the opposite direction. The total length decreased through the merging from ~ 500 Mbp for S3 and 483 Mbp for BnJanetzki down to 477 Mbp after merging. Across the pipeline steps, the length could have been further decreased and stagnated at ~ 477 Mbp which is a reduction of ~ 135,3 kbp after merging.

S3 reveals a number of 23.518 fully unaligned contigs while BnJanetzki reveals a much smaller amount of 3.441. Through the merging, this value came to 4.218 and could have been decreased down to 3.924. A similar tendency occurred for the fully unaligned length of contigs. After the second run of *sealer*, ~ 5,8 Mbp which are ~ 1,2 % of the whole assembly was unaligned. The amount of ambiguous bases increased significantly through the merging. Zero Ns in the raw S3 assembly and 10.000.000 Ns in the raw BnJanetzki have been raised up to 324.203.65 after *minimus2-blat* and could get decreased over the pipeline down to 287.298.50 which corresponds to a reduction of ~ 11,4 % after merging. S*ealer* could close 64.279 gaps (stretches of Ns within contigs) after the first execution and further 359 gaps after the second execution (Sealer report, Supplement).

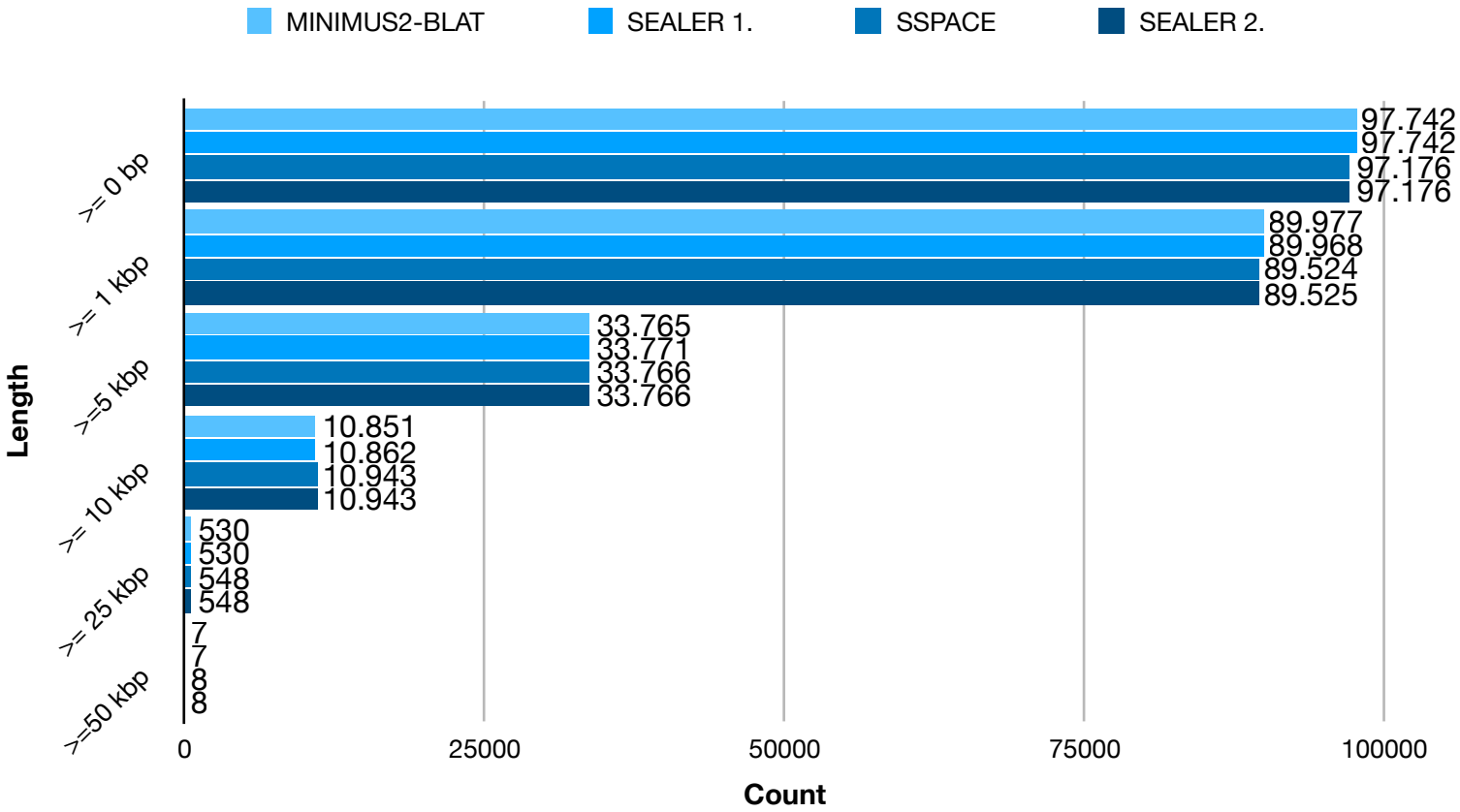


Fig 1.1 Fragment size distribution for S3_Janetzki after *minimus2-blat*.

The whole number of fragments decreased from 437.184 (S3) and 137.972 (BnJanetzki) down to 97.742 through the merging and after that, to 97.176 through the scaffolding by *sspace*. An increase of fragments larger than 10 kbp, 25 kbp and 50 kbp mainly after merging and scaffolding through *sspace* was also observed. The most significant enlargement of fragments >= 10 kbp, occurred after merging. For instance, there are 530 fragments >= 25 kbp after the *minimus2-blat* execution while the raw assemblies S3 and BnJanetzki reveal a count of 5 and 278 (Fig 1.1).

Table 2.2 Genome statistics for the S2_Lorenz merge

Genome statistics	S2 raw	BnLorenz raw	Merge after MINIMUS2-BLAT	Merge after SEALER	Merge after SSPACE	Merge after SEALER (2nd run)
Genome fraction (%)	77.798	58.461	50.773	51.004	50.978	50.985
Total length (bp)	66.319.548	494.826.375	571.560.462	571.455.891	571.416.003	571.429.654
N50 (bp)	13.678	6.796	18.779	18.782	19.219	19.219
L50	14.325	21.671	9.627	9.623	9.402	9.402
NG50	9.714	2.377	11.227	11.224	11.450	11.450
LG50	22.429	64.722	19.204	19.203	18.782	18.782
Unaligned						
Fully unaligned fragments	4.739	2.118	528	517	508	508
Fully unaligned length (bp)	9.191.217	1.934.413	1.573.011	1.529.644	1.501.443	1.501.443
Mismatches						
N's	2.511.929	199.013	30.260.258	29.093.066	29.094.805	29.077.973
N's per 100 kbp	378.76	237.541	5294.32	5091.04	5091.7	5088.64
Statistics without reference						
Fragments (>= 0 bp)	89.306	119.729	44.767	44.767	43.908	43.908
Fragments (>= 1000 bp)	87.868	98.111	44.755	44.749	43.890	43.890
Fragments (>= 5000 bp)	39.782	33.021	33.237	33.236	32.801	32.801
Fragments (>= 10000 bp)	21.748	10.524	21.494	21.488	21.385	21.386
Fragments (>= 25000 bp)	4.343	517	5.411	5.411	5.563	5.563
Fragments (>= 50000 bp)	328	5	552	552	596	596

For the S2_Lorenz merge, similar values and tendencies to S3_Janetzki occurred. However, there are also clear differences in value between the two of them. After merging, genome fraction decreased from 77.8 % (S2) and 58.5 % (BnLorenz) down to 50.8% and ended up at ~ 51 % after the second *sealer* execution. N50 values reveal its biggest increased after merging and the scaffolding by *sspace*. At the beginning, it started with 13.678 bp (for S2) and 6.796 bp (for BnLorenz) and increased to a value of 18.779 bp after merging. After scaffolding with *sspace*, the value increased to 19.219 bp again.

NG50 values reveal a similar development. L50 and LG50 values reveal the same development in the opposite direction. The number of fully unaligned fragments significantly decreased through the merging from 4739 (S2) and 2118 (BnLorenz) to 528 and 517 after the first sealer execution. Through *sspace*, it was increased to 508 fragments. The same tendency applies to the fully unaligned length. Nearly 0,2 % of all bases keep being unaligned.

Ambiguous Bases increased from 2.511.929 (S2) and 199.013 (BnLorenz) up to 302.602.58 through the merging and were further increased by *sspace.* By using *sealer,* ~ 7 % of all N's could get reduced after merging. Due to the *sealer* report, *sealer* could close 35.987 gaps after the first execution and further 463 gaps after the second one.

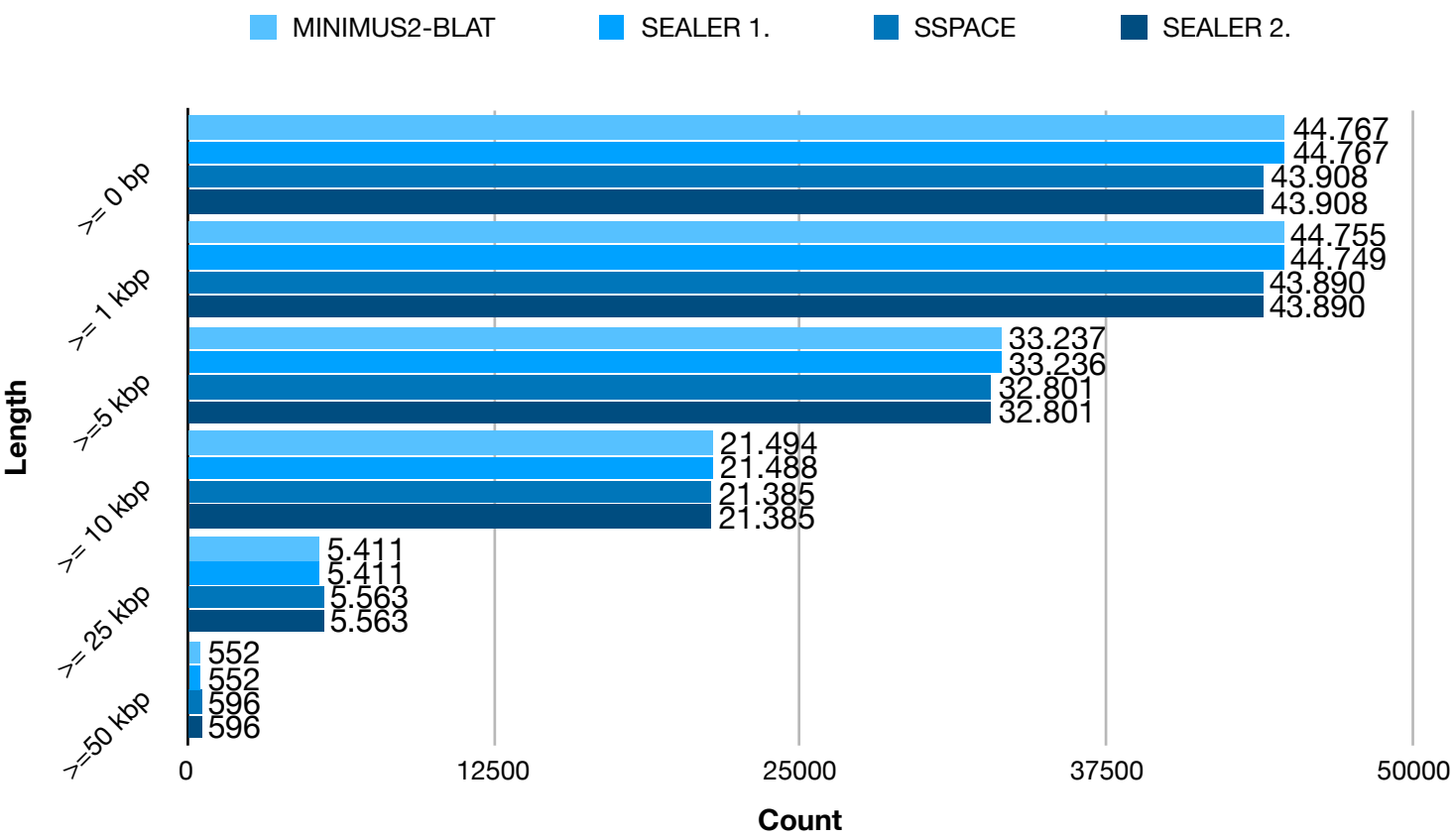


Fig 1.2 Fragment size distribution for **S2_Lorenz** after *minimus2-blat*.

The fragment size distribution (Fig 1.2) reveals expected values due to the N50 and NG50 development. The whole number of fragments decreased from 89.306 (S2) and 119.729 (BnLorenz) to 44.767 through the merging while especially fragments >= 50 kbp gained in count. Through *sspace*, fragments >= 25 kbp, probably scaffolds, gained in count with a resulting reduction of the total fragment count.

The following two tables reveal information about the un-merged (no *minimus2-blat*) assemblies S2 and S3.

Table 2.3 Genome statistics for the S2 assembly

Genome statistics	S2 raw	S2 after Sealer	S2 after SSPACE	S2 after Sealer (2nd run)
Genome fraction (%)	77.798	78.017	78.016	78.062
Total length (bp)	663.192.277	663.747.878	663.747.458	663.872.741
N50 (bp)	13.678	13.680	13.680	13.677
L50	14.325	14.334	14.333	14.335
NG50	9.714	9.728	9.728	9.732
LG50	22.429	22.407	22.406	22.401
Unaligned				
Fully unaligned fragments	4.739	4.658	4.657	4.641
Fully unaligned length (bp)	9.191.217	9.030.608	9.025.306	8.975.562
Mismatches				
N's	2.511.929	1.974.398	1.974.399	1.864.986
N's per 100 kbp	378.76	297.46	297.46	280.93
Statistics without reference				
Fragments (>= 0 bp)	89.306	89.306	89.296	89.296
Fragments (>= 1000 bp)	87.868	87.903	87.900	87.912
Fragments (>= 5000 bp)	39.782	39.816	39.815	39.821
Fragments (>= 10000 bp)	21.748	21.767	21.766	21.769
Fragments (>= 25000 bp)	4.343	4.348	4.347	4.347
Fragments (>= 50000 bp)	328	329	329	330

Basically, there is except for a few parameters no big change in value through the pipeline for the S2 assembly. The genome fraction could get increased by ~ 0,26 % from 77,798 to 78,062 %, mainly caused by the two *sealer* executions. The total length also went up in value from 663.192.277 bp to 663.872.741 bp which is an increase of 680.464 bp. N50 and NG50 values

didn't reveal any significant change in value through the pipeline. It started with 13.678 bp for N50 and 9.714 bp for NG50 and finished at 13.677 bp (N50) and 9732 bp (NG50). Similar non-developements apply to L50 and LG50.

The amount of fully unaligned bases and and length decreased. It started with a value of 4739 fragments and sank down to 4641 after the second *sealer* execution. Thereby, the fully unaligned length also reveals a loss in value. It started with 9.191.217 bp and finished with 8.975.562 bp before applying *medusa*. That's a reduction of 215.655 bp and a final amount of ~ 1,3 unaligned bases of the whole assembly. The number of ambiguous bases, started with 2.511.929 by number, has been reduced by ~ 26 % to 1.864.986. Due to the *sealer* report, *sealer* could close 15.612 gaps and further 2352 gaps through both executions.

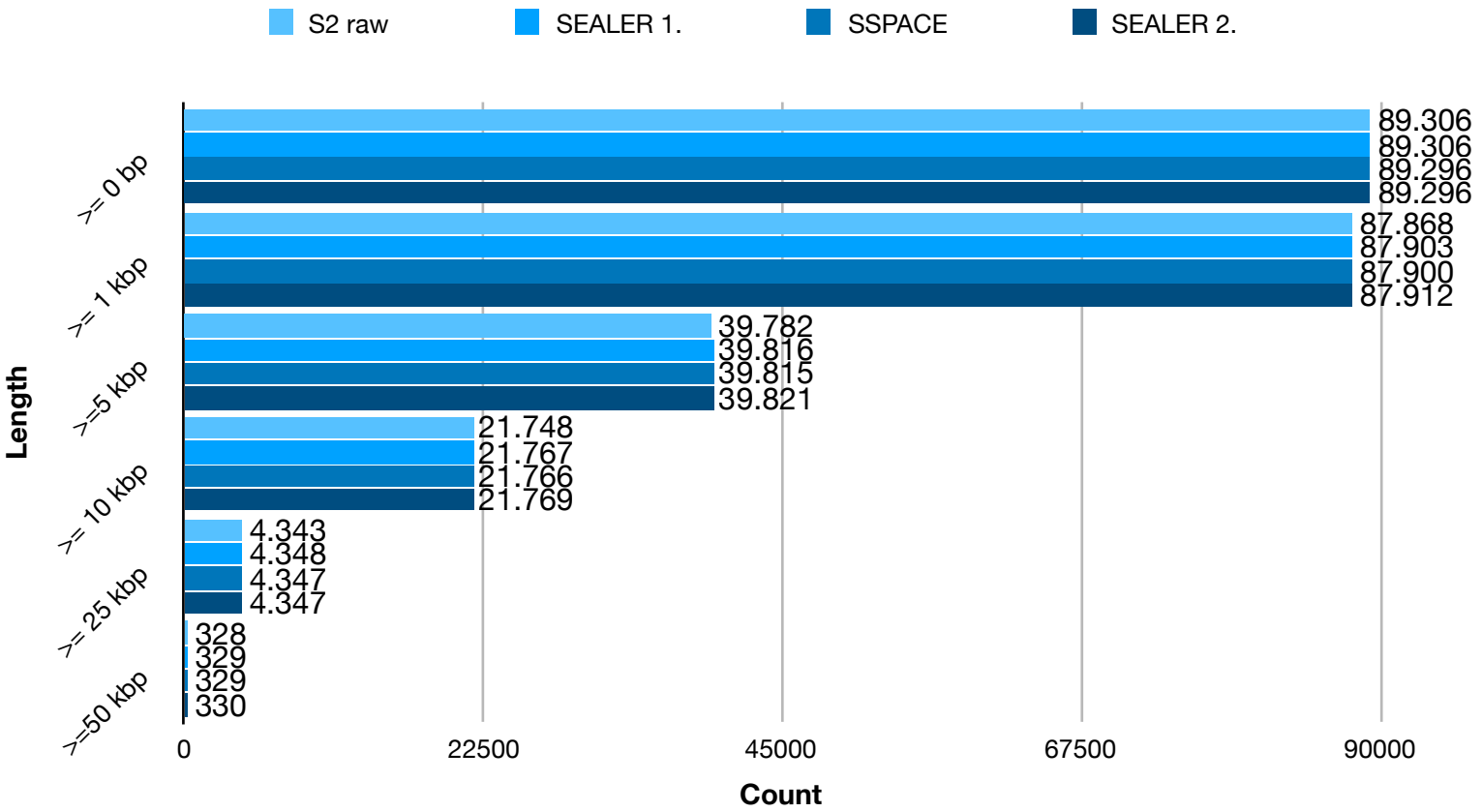


Fig 1.3 Fragment size distribution for S2.

The fragments size distribution didn't reveal any big change in value. For instance, starting with 89.306 fragments and finishing with 89.296 is a reduction of 10 fragments. Thereby, a single fragments >= 50 kbp and a small amount of fragments between 1 and 50 kbp could get gained.

Table 2.4 Genome statistics for the S3 assembly

Genome statistics	**S3 raw**	**S3 after Sealer**	**S3 after SSPACE**
Genome fraction (%)	52.754	52.754	52.752
Total length (bp)	442.815.485	442.815.485	442.815.485
N50 (bp)	2.277	2.277	2.277
L50	57.284	57.284	57.284
NG50	604	604	604
LG50	225.623	225.623	225.623
Unaligned			
Fully unaligned fragments	23.518	23.518	23.519
Fully unaligned length (bp)	22.193.808	22.193.808	22.194.394
Mismatches			
N's	0	0	0
N's per 100 kbp	0	0	0
Statistics without reference			
Fragments (>= 0 bp)	437.184	437.184	437.184
Fragments (>= 1000 bp)	157.093	157.093	157.093
Fragments (>= 5000 bp)	10.262	10.262	10.262
Fragments (>= 10000 bp)	673	673	673
Fragments (>= 25000 bp)	5	5	5
Fragments (>= 50000 bp)	1	1	1

The S3 assembly doesn't reveal mismatches in form of ambiguous bases. That's why *sealer* couldn't find N's (*). Thus, a second execution of *sealer* revealed to needless.

The fewest changes in value occurred for the S3 assembly. The genome fraction reveals an overall decrease of 0,002 %. Total length plus N and L values (N50, NG50, L50, LG50) didn't change. The same applies to the whole fragment size distribution (Fig 1.4). Because the S3 assembly doesn't have Ns, there were no Ns and gaps to be filled by *sealer*. The number of fully unaligned fragments was increased by a single fragment from 23.518 to 23.519 by *sspace*.

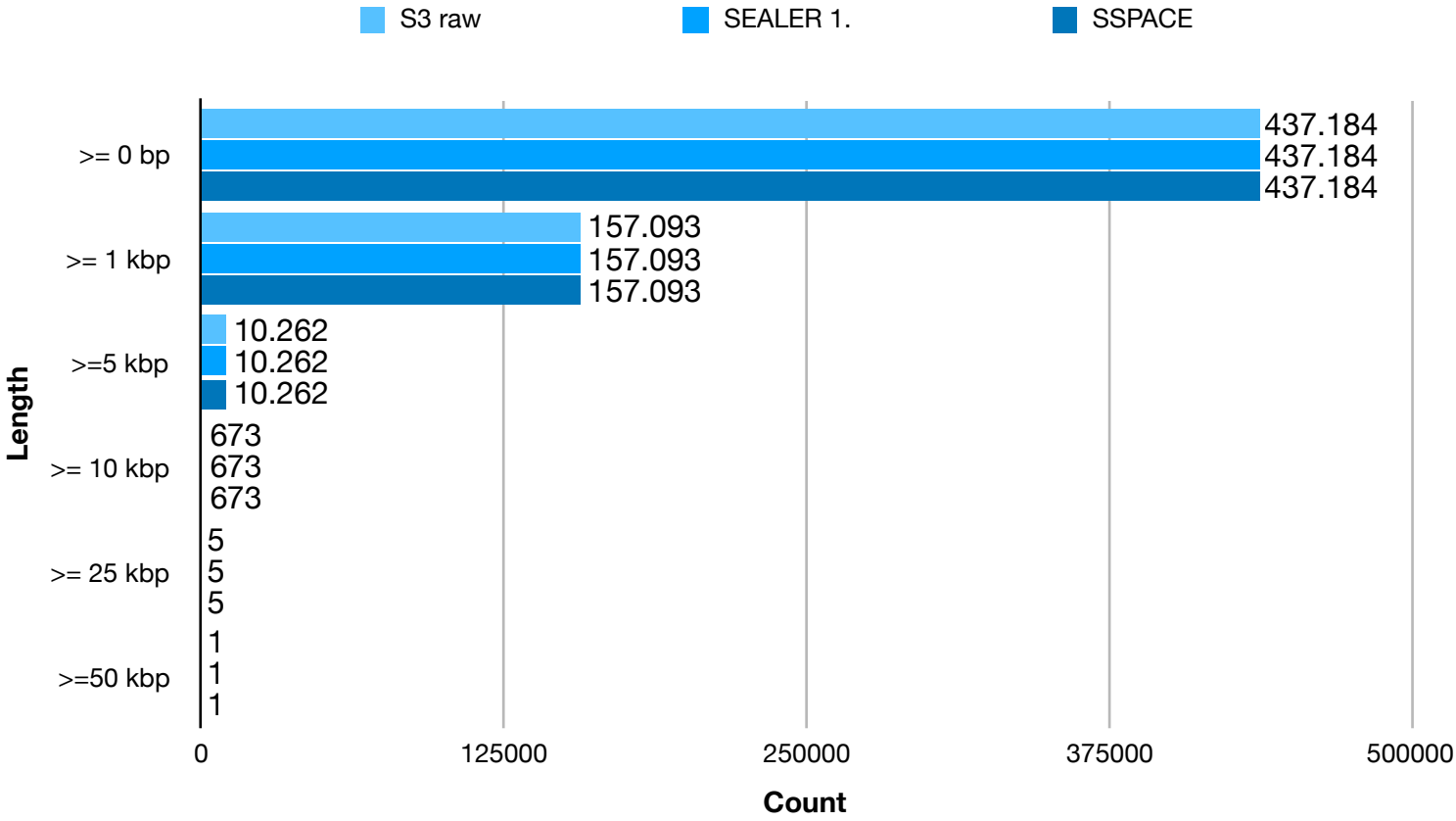


Fig 1.4 Fragment size distribution for S3.

After applying the pipeline on the assemblies so far, the fragments were filtered and *medusa* was executed on the filtered sets. Statistics after filtering and the q*uast* report after scaffolding are listed in the following table (Table 3) (whole report see supplement). Predicted genes are also included.

Table 3 Genome statistics after filtering & MEDUSA

Genome Statistic	S2	S3	S2_Lorenz	S3_Janetzki
Filtering statistics				
Initial fragment number	89.296	437.184	53.420	97.176
Filtered fragment number	21.769	18.672	15.281	21.314
Minimum fragment size (bp)	10.000	4.000	13.736	7.000
Remaining fragments (%)	~ 24,4	~ 4	~ 28,6	~ 21,9
Remaining bases (%)	~ 63,1	~ 24,2	~ 66,9	~ 52,1
Scaffold statistics & distribution				
Genome fraction (%)	50.79	13.108	33.985	20.891
Total length (bp)	420.813.103	108.804.404	382.607.675	250.789.876
Scaffolds (>= 0 bp)	2.165	2.281	386	3.026
Scaffolds (>= 5000 bp)	2.165	2.149	386	3.026
Scaffolds (>= 10000 bp)	2.165	1.910	386	2.882
Scaffolds (>= 25000 bp)	2.016	1.270	372	2.364
Scaffolds (>= 50000 bp)	1.712	747	350	1.640
Largest scaffold (bp)	1.861.219	478.452	11.481.033	1.015.252
N50 (bp)	324.492	81.858	1.999.167	132.918
L50	395	398	54	566
NA50 (bp)	25.204	6.037	14.533	5.761
LA50	4.896	5.487	6.621	10.786
Unaligned				
Fully unaligned scaffolds	3	28	0	3
Fully unaligned length (bp)	42.381 (0,01 %)	148.125 (0,1 %)	0	28.222 (0,01 %)
Mismatches				
N's	2.349.027 (0,6 %)	1.556.800 (1 %)	21.374.672 (5 %)	17.347.006 (7 %)

Genome Statistic	S2	S3	S2_Lorenz	S3_Janetzki
N's per 100 kbp	558.21	1430.82	5586.58	6916.95
Predicted genes				
Genes	99.926	27.265	68.835	44.083

NA50 is N50, where the length of aligned blocks are counted instead of aligned fragments (in this case scaffolds). Aligned blocks are obtained by breaking scaffolds at misassembly events and removing all unaligned bases (Quast Manual 4.6.0)

I reduced the initial fragment number for percentages of not more than 29 %. S3 was reduced to 4 % of the initial fragment number which is the strongest reduction compared to the rest of the assemblies (21,9 - 28,6 %). Compared to S2, S2_Lorenz and S3_Janetzki, S3 also provided by far the highest count of 437.184 initial fragments. S2, S2_Lorenz and S3_Janetzki had initial fragment numbers of less than 100.000. After filtering, the fragment numbers range from ~ 15.000 to ~ 22.000 with a maximum of 21.769 counts for S2 and a minimum of 15.281 counts S2_Lorenz. For S2 and S2_Lorenz, the remaining bases are not less than 63 % with 63,1 % for S2 and 66.9% for S2_Lorenz. Because of a more strict filtering, which is caused by a bigger amount of relatively small fragments <= 5000 bp (See Tables 2.1 & 2.4), S3 and S3_Janetzki only remain 24,1% and 52,1 % of initial bases after filtering.

After scaffolding, the genome fraction was significantly reduced for all of the assemblies caused by filtering before. Compared to the unfiltered assemblies, S3 diminished in value the most. From ~ 52,8 % to ~ 13,1 % is a loss of 39,7% . S2_Lorenz and S3_Janetzki reveal a genome fraction of ~ 33,99 % and 20,89 %, respectively, which is a reduction of not more than ~ 20 % after filtering for both of them. The S2 assembly lost ~ 27,3 % in value from initial ~ 70,1 % to ~ 50,8 %. As a further consequence of filtering, the total lengths of the scaffolded assemblies were strongly reduced compared to the initial lengths.

The majority of the scaffolds is mainly distributed over values not smaller than 25 kbp for each assembly. For S2, there are overall 2.165 scaffolds >= 10 Kb from which 1.712 scaffolds (~ 79 %) are as big as or bigger than 50 kb. S2_Lorenz reveals a set of overall 386 scaffolds >= 10 kbp from which 350 scaffolds (~ 90 %) are not smaller than 50 kbp. S3 and S3_Janetzki reveal more evenly distributed scaffold sets than S2 and S2_Lorenz. Overall, S3 reveals a set of 2.281 scaffolds from which 2.149 scaffolds (~ 94 %) are not smaller than 5 kbp, 1.910 scaffolds not smaller than 10 kbp, 1.270 scaffolds not smaller than 25 kbp and 747 scaffolds which is a portion of up to 32 % not smaller than 50 kbp. A similar distribution shows S3_Janetzki which reveals overall 3.026 scaffolds >= 5 kbp from which 1.640 scaffolds (~ 54 %) are scaled not less than 50 kbp. The largest scaffold of ~ 11,5 Mbp occurs for S2_Lorenz. S3_Janetzki reveals its largest scaffold of ~ 1 Mbp. For S2 and S3, the largest scaffolds scale ~ 1,8 and ~ 0,5 Mbp. Caused by filtering and further scaffolding, N50 values strongly increased compared to the unfiltered assemblies and range from ~ 81 kbp up to ~ 2 Mbp. Same goes in the opposite direction for the L50 values. NA50 and LA50 are basically smaller than the referring N50 and L50 values. They are similar defined as NG50 and LG50 with the difference that aligned blocks (to the reference) are counted instead of whole aligned fragments which would be scaffolds in this case. Aligned blocks are obtained by breaking scaffolds at SVs or misassembly events and removing all unaligned bases. NA50 values range from ~ 5 kbp to ~ 25 kbp with a maximum of 25,2 Mbp for S2 and a minimum of 5,7 Mbp for S3_Janetzki. LA50 values are basically smaller than their referring NA50 values with an exception for S3_Janetzki which scales an almost doubled LA50 value of 10.786 than the NA50 of 5761 bp. Relative high LA50 values often indicate high counts of misassembly events and/or SVs.

The numbers of unaligned scaffolds are low. S2 reveals zero unaligned scaffolds. Both S2 and S3_Janetzki reveal three unaligned scaffolds while S2_Lorenz comes up with 28 unaligned scaffolds as the maximum value compared to the rest. For all scaffold sets, not less than 99,9 % of all bases aligned against the *B. napus* reference. Considering the fact that the assemblies diminished in small contigs through the filtering, they reveal a relatively high count of ambiguous bases which is has been further enhanced by the scaffolding process. S2 and S3 have an even higher count of Ns while S2_Lorenz and S3_Janetzki reveal slightly less Ns than their unfiltered versions of fragments. For all assemblies the amount of Ns doesn't measures percentages over 7 %. Finally, up to 99.926 genes were predicted for S2 as the maximum value. For the rest, gene numbers between 27.265 and 44.083 were predicted.

In the following figures (Fig 4.1 - 4.4), fragment and scaffold distribution are compared to each other. The fragment distribution refers to the state after the second *sealer* execution.

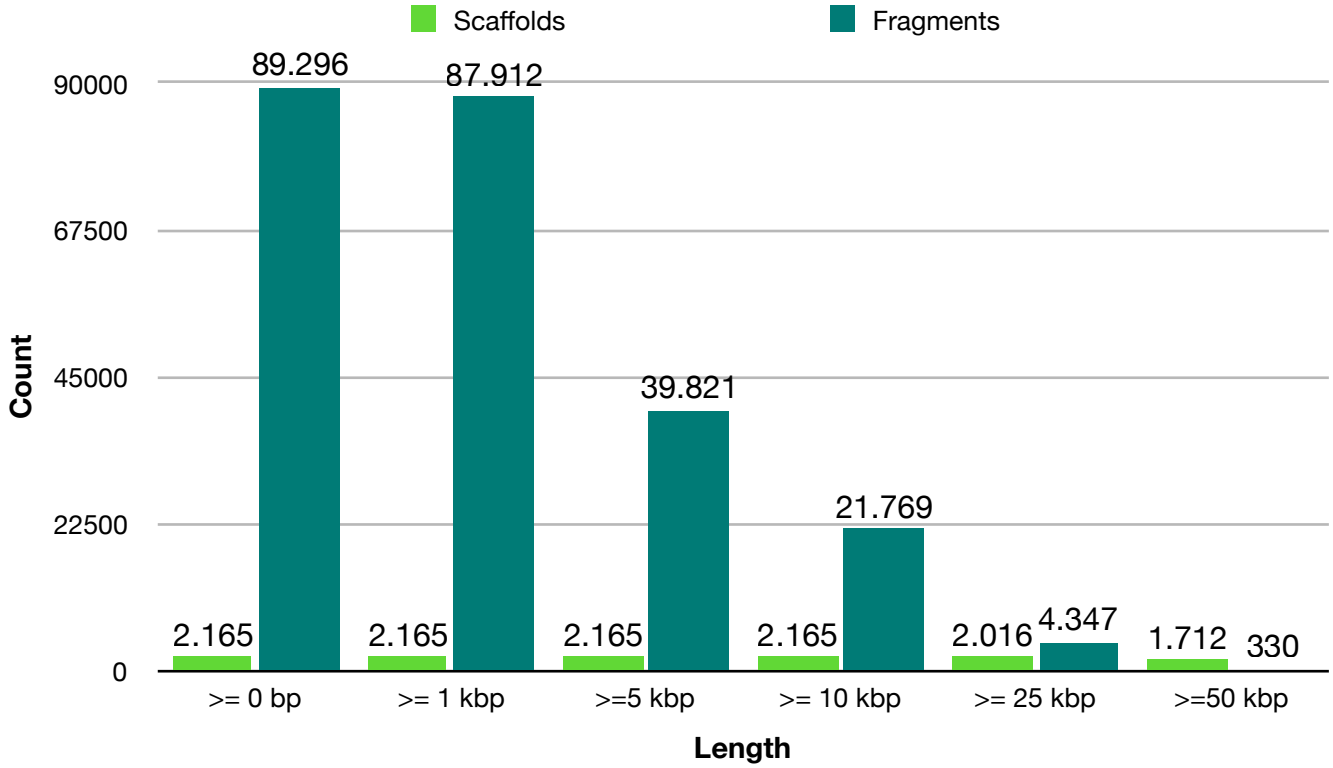


Fig 4.1 Scaffold and fragment distribution for **S2** by comparison. All scaffolds remain ~ 63,1 % bases and 24,4 % sequences of the unfiltered set of fragments. Differences in count between scaffolds and fragments continuously diminish by length.

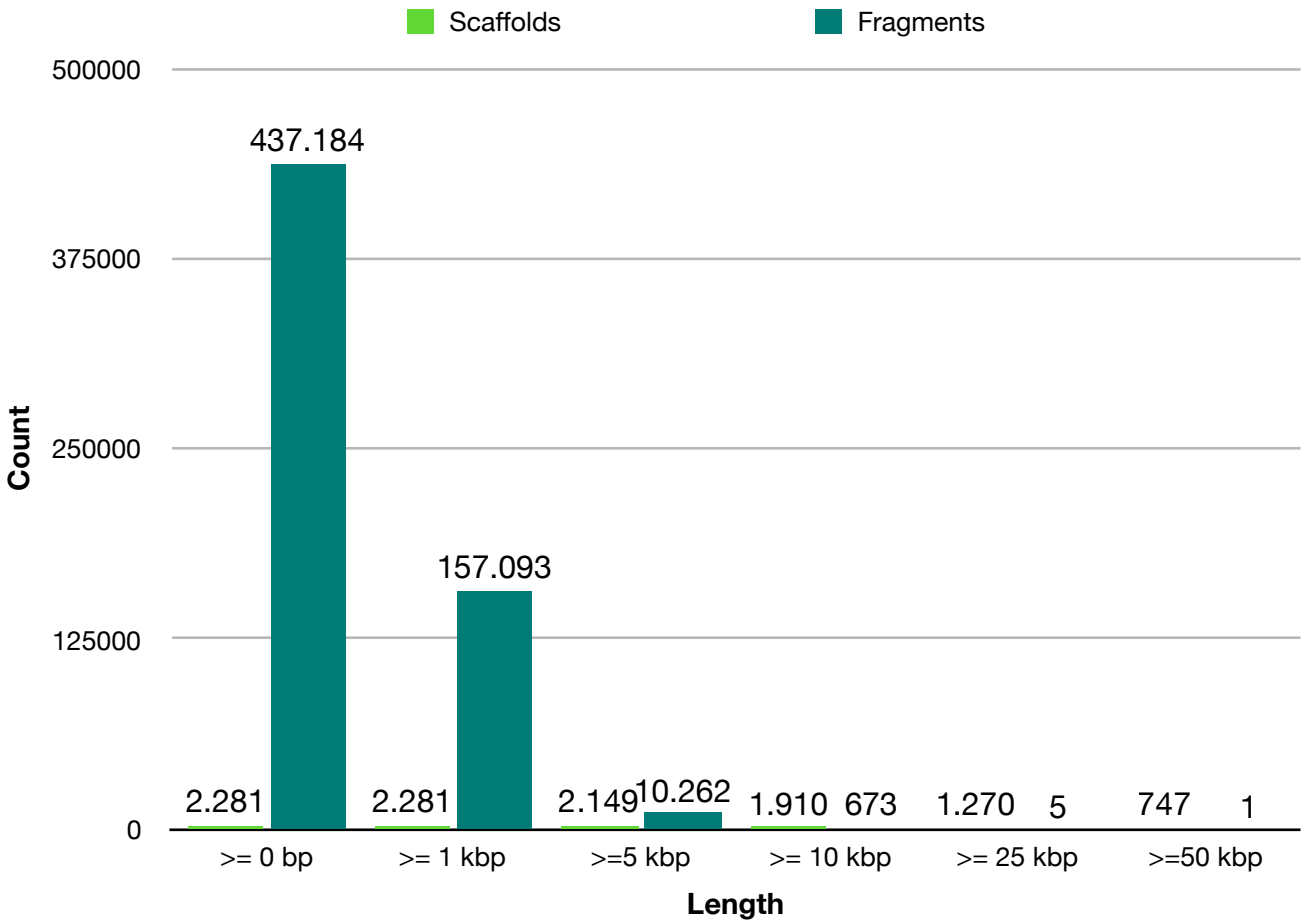


Fig 4.2 Scaffold and fragment distribution for **S3** by comparison. All scaffolds remain ~ 24,2 % bases and 4 % sequences of the unfiltered set of fragments with a comparative higher count in sizes >= 10 kbp.

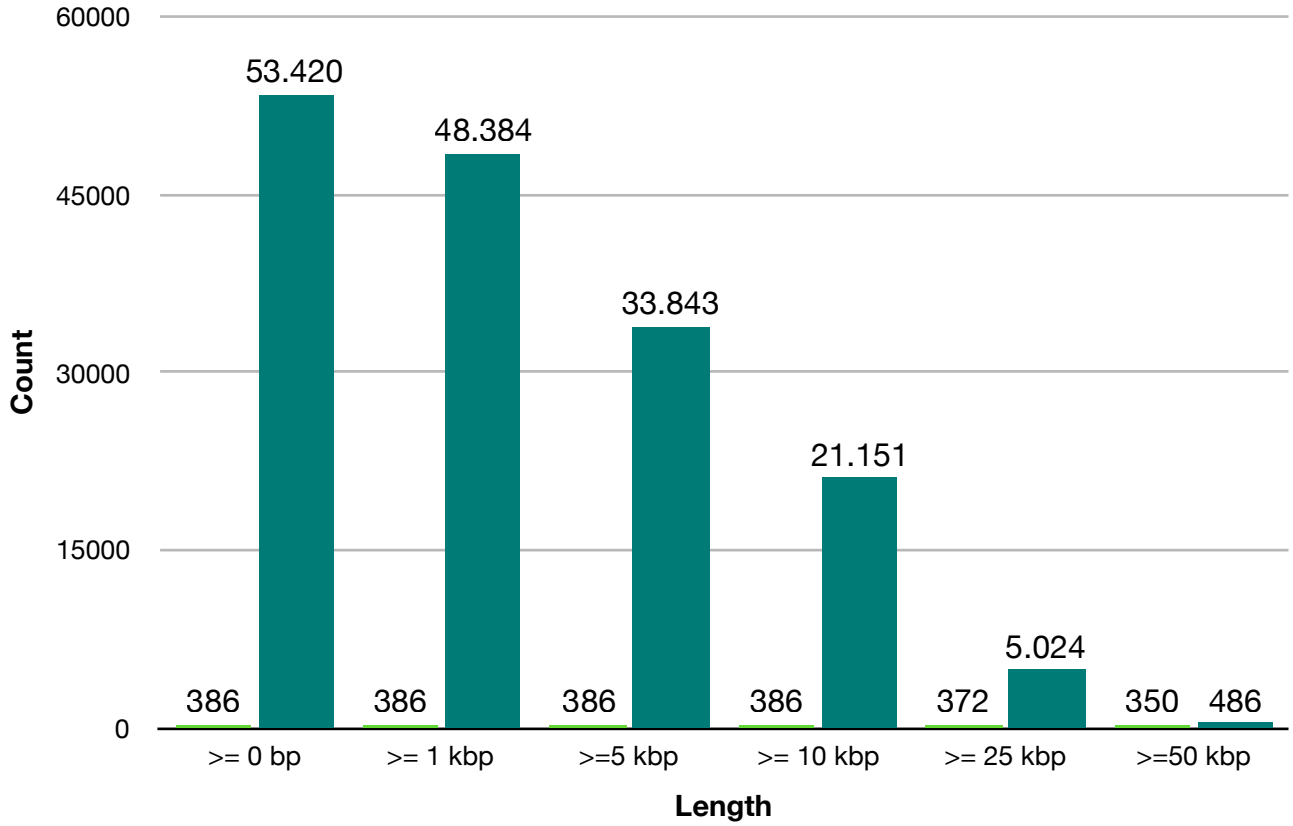


Fig 4.3 Scaffold and fragment distribution for **S2_Lorenz** by comparison. All scaffolds remain ~ 66,9 % bases and 28,6 % sequences of the unfiltered set of fragments.

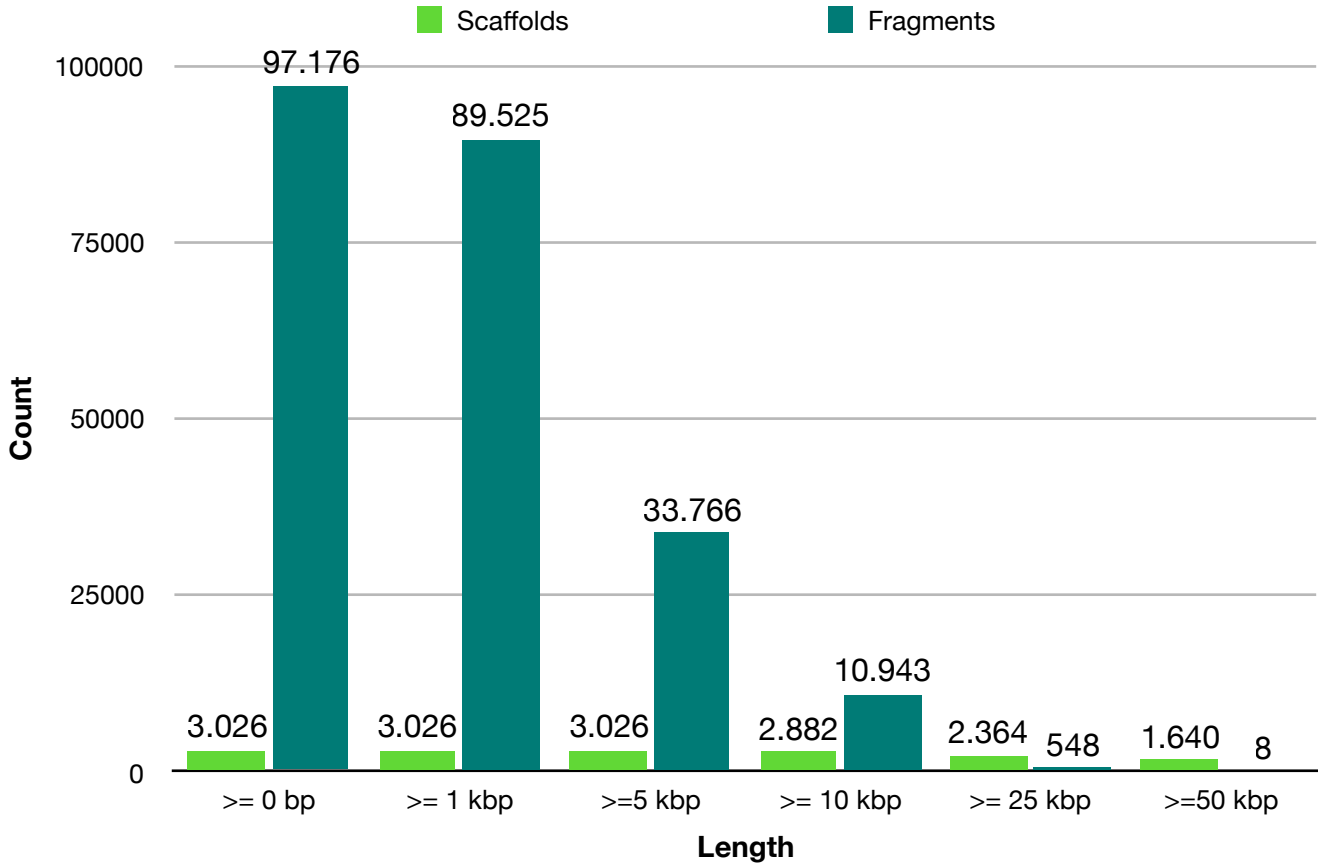


Fig 4.4 Scaffold and fragment distribution for **S3_Janetzki** by comparison. All scaffolds remain ~ 52,1 % bases and 21,9 % sequences of the unfiltered set of fragments. Scaffolds >= 25 kbp reveal a higher count than fragments of that size or larger.

4.4. Alignment & SV Analysis

For the alignment between the parental lines, I focused on the scaffold sets of both the merged assemblies S2_Lorenz and S3_Janetzki by performing an genome-wide alignment between them. It was realized by using tools from the *MUMmer* package (18): The alignment was performed by *nucmer4.0.0* and further filtered by using *delta-filter* (Material & Methods). The alignment was analysed using *dna-diff (SI* Materials & Methods) in order to capture alignment statistics and SVs detections. In the following table, *dna-diff* results are listed (Table 4). The whole report can be found in the supplement. It is important to say that all SV detections based on the *MUMme*r utilities can not be distinguished from misassembly events.

Table 4 Alignment statistics and SV detections between S2_Lorenz and S3_Janetzki

	S2_Lorenz	**S3_Janetzki**
Total scaffolds	386	3.026
Total bases	382.607.675	250.789.876
Alignment statistics		
Aligned scaffolds	385 (99,74 %)	2993 (98,91 %)
Aligned Bases (%)	~ 27,52	~ 38,02
SVs		
Relocations	2.102	2.137
Translocations	23.323	21.526
Inversions	1.666	1.651
Total SNPs	402.781	402.781
G-SNPs	122.244	122.244
Total Indels	300.803	300.803
G-Indels	16.433	16.433

Relocation is an event where adjacent aligned blocks are in the same sequence, but not consistently ordered. Translocations are defined as events where aligned blocks are in different sequences, in this case scaffolds. Inversions are defined as events where aligned blocks are inverted with respect to one another. Those mentioned SVs are calles from the 1-to-1 *nucmer* alignment. SNPs are single nucleotide polymorphisms and indels are defined as insertions and deletions as well.

All scaffolds expect for a singe one of S2_Lorenz were aligned against 2.993 scaffolds of S3_Janetzki which are 98,91 % of the whole scaffold set. By contrast, the fraction of aligned bases comes to percentages with less than 40 % for both of the assemblies. That can indicate a large amount of ambiguous bases within the scaffolds or corroborates the relatively high SV or misassambly frequency between the assemblies. Considering the translocations and inversions, slightly more were detected for S2_Lorenz for each of that kind compared to S3_Janetzki. S3_Janetzki reveals a higher amount of relocations compared to S2_Lorenz. G-SNPs are defined as SNPs bounded by 20 or more base-pair matches on both side. For both the assemblies, 122.244 were detected which are ~ 25 % of all detected SNPs.

G-Indels are similar defined to G-SNPs. A count of 16.433 G-Indels were detected while the count of all detected indels comes to 300.803 which is a nearly 18 times bigger count.

To call SVs between the assemblies and the *B. napus* reference genome, I carried out a second alignment using *quast*'s alignment utility *minimap2*. By including the short read data, *quast* attempts to distinguish between real SVs and misassembly events. I focused on the scaffold sets of S2_Lorenz and S3_Janetzki and their referring read libraries. The results are listed in the following table (Table 5).

Table 5 Misassembly and SVs detection through quast alignment to B.napus reference

	S2_Lorenz	**S3_Janetzki**
Relocations	9.957	6.894
Translocations	8.514	7.977
Inversions	31	33
Real SVs (from relocations, translocation and inversions)	-*	23
SNPs	972.323	542.603
SNPs per 100 Kbp	259.28	351.78
Total indels	170.997	104.889
Indels per 100 kbp	45.6	68
Indels (<= 5 bp)	129.808	83.716
Indels (> 5 bp)	41.189	21.173

By means of the short read data, *quast* attempts to call real SVs from those which *quast* detects as misassembly events. *For S2_Lorenz, *quast* had problems dealing with the read data. Because I couldn't fix this problem in the intended time frame, a SVs analysis couldn't get carried out for this scaffold set.

Relocations and translocation form the majority of SVs. S2_Lorenz reveals slightly more counts for both of them (9.957 relocations and 8.514 translocations) compared to S3_Janetzki (6.894 reloactions and 7.977 translocations). Inversions form the smallest part of all SVs. With 33 by

number, S3_Janetzki counts two more than S2_Lorenz (31 inversions). *Quast* could identify 23 real SVs from a total of 14.904 SVs which are less than 0,2 % by means of the read data for S3_Janetzki. That means, *quast* evaluates ~ 99,98 % of all SVs as misassembly events. The total SNP number (including indels, and mismatches caused by ambiguous bases) of S2_Lorenz comes up to 972.323 while S3_Janetzki reveals a count of 542.603. The SNPs per 100 kbp rates come to values of ~ 300 for S2_Lorenz and ~ 543 for S3_Janetzki which seems inconsistent on the first view considering S2_Lorenz occurs the almost doubled amount of SNPs than S3_Janetzki. But this is due to the fact that S2_Lorenz occurs more bases and therefore, SNPs are denser located on the sequences (See Table 4). The same tendency applies to the indel per 100 kbp rate. The indel numbers are 170.997 for S2_Lorenz and 104.889 for S3_Janetzki while indels >= 5 bp form the majority of all Indels for both of the assemblies.

5. Discussion

The main objective of building a functioning pipeline which carries out the multi-reference-based assembly could be realized. However, tools like *medusa* and *minimus2-blat* had to be modified in order to handle larger, eukaryotic genomes, for which they initially were not designed for.

Compared to the un-merged assemblies S2 and S3, most of the assembly improvements occurred for the merged sets, especially through the merging by itself. Most of all, we observe fragment size developments in benefit of larger fragment which is normally seen as an assembly improvement. This fragment size enlargement could relate to the strong increase of ambiguous bases during the *minimus2-blat* merge. By introducing Ns, *minimus2-blat* forces the alignment of slightly differing sequences and joins them together.

Thus, we get larger fragments with a higher N frequency. Moreover, the genome fraction covered by fragments should decrease which we observe as a further side effect. With a higher frequency of Ns, a lesser percentage of bases aligns to the reference genome. An improvement would be to use a more stringent alignment conditions by setting *minimus2-blat*'s minimum alignment identity higher and setting the maximum consensus error lower which would result in a lower frequency of Ns and smaller but more accurate fragments.

Comparing results of both *sealer* executions to each other, we observe more gaps getting closed during the first run. This is expected, because the assemblies have shown a larger amount of Ns before applying the *sealer* first time. Even by introducing a few more Ns through the scaffolding by *sspace*, *sealer* didn't close as many gaps as it did the first time. Moreover, *sealer* could remove more gaps in S2_Lorenz than in the single S2 assembly. The main reasons lie probably in the facts that S2_Lorenz generally provides a larger number of gaps to be closed and additionally, was processed with a more extensive read library. By concatenating read libraries (BnLorenz and S2), we basically provide more read information for *sealer* which should result in a more exhaustive process of gap closing. Because both assemblies originate from the same parental line, it is valid to use the concatenated read library (BnLorenz + S2) for each of the origin assembly (S2 or BnLorenz). Including the concatenated read libraries of BnLorenz and S2 into the *sealer* execution for S2 should have enabled more gaps to be closed.

Furthermore, the single S2 read library reveals an increased read count in comparison to the other assemblies with an atypical, constant read length of 251 bp (Table 1). I suspect, this read library was not quality trimmed as read libraries of BnJanetzki or BnLorenz were. Quality trimming is known to

increase the quality and reliability of *de novo* assembly, with concurrent gains in terms of execution time and also computational resources needed (19). Additionally when using programs like *sealer* which only rely on the read data, it probably leads to better results by including read libraries of good quality. Unfortunately, a comparison of *sealer* results for S3 and S3_Janetzki turns out to be irrelevant, because S3 doesn't any reveal ambiguous bases or gaps to be closed. The fact that S3 also shows a strongly increased contig number of lower size compared to the other assemblies (Table 1), indicates that S3 has not been scaffolded in course of the CLC *de novo* assembly.

The scaffolding process using *sspace* reveals similar to *sealer* more changes for both of the merged assemblies S2_Lorenz and S3_Janetzki than for S2 and S3. Reasons could lie in the fact that only smaller read libraries were included for S2 and S3. By using concatenated read libraries, more possibilities for a more exhaustive scaffolding process only relying on read data should be provided. Furthermore, a more relaxed scaffolding process by a less strict parameter-setting could also lead to more scaffolds to be build. On the other hand, the latter would also increase the probability of an incorrect scaffolding process. Considering that, an improvement would be including as much valid read information as possible while setting the parameters to a strict scaffolding process to avoid false positives. Compared to S3, S2 reveals a better scaffolding result by *sspace*, even when probably non-qualilty trimmed reads were used. Reasons for this might lie in an overall worse assembly quality of S3, corroborated by the fact that S3 has probably not been scaffolded at all in course of the CLC assembly.

Applying *medusa* shed light on the strong opportunities multiple-reference based scaffolding provides. Besides the fact that the assemblies had to be filtered before, *medusa* leads to much better scaffolding results than *sspace*.

The reason for that stems from the fact that *medusa* works in a different way than *sspace. Sspace* operates only by means of read data. We technically provide the same, but more valuable information for *medusa*: High quality reference genomes. All relevant and significant read information of high quality are maintained in the reference genomes which were assembled and optimized in quality by previous projects before. With the availability of more high quality genomes in the future, reference-based or even multiple-reference-bases scaffolding tools will replace scaffolding tool which work only by means of read data like *sspace* does. The only condition using programs like *medusa* entails, is the availability of at least one single reference genome of high quality. When this is not the case, programs like *sspace* which relying on read data only, constitute the better option.

In contrast to the single-reference-based assembly of *A. thaliana* in 2011 (2) with its largest scaffold of 2,2 Mbp, the multi-reference-based scaffolding leads to a largest scaffold of ~ 11,5 Mbp. Percentages of Ns come up to nearly similar values while N50 values of the multi-reference-based assembly significantly exceed the single reference-guided assembly in 2011. A strong limitation obviously remains in the difficult handling of large genomes with the necessity to filter out the majority of smaller fragments before applying *medusa*. Filtering always removes data that could technically be relevant for further SVs and general genome investigations. Even by modifying the overlap detector to use multi-threaded working *nucmer4*, *medusa* still shows limitations in run time. For a large genome like *B. napus,* it additionally doesn't work without engaging in steps like building the graph-network file (Materials & Methods). Considering the fact that scaffolding always goes with introducing new Ns into the assembly, a third execution of *sealer* after *medusa* might be an improvement in order increase genome fraction, alignment percentages and overall scaffold quality.

It is difficult to compare alignment results to each other, because firstly, different utilities were used for the alignment and secondly, different assembly combinations were used for the alignment. The high number of SVs detected between the assembled lines S2_Lorenz and S3_Janetzki might corroborate the hypothesis that individuals of the same species reveal large structural differences in genetic material (1). On the other hand, a conspicuous difference between the alignments is the amount of aligned bases which reveals significant higher values for the alignments of assembly to reference than for assembly to assembly. Further considering the fact, that *quast* could detect 23 real SVs from a total of nearly 15.000 SVs from the alignment between S3_Janetzki and the reference, most of the SVs detections (which should apply for the assembly to reference as well as for the assembly to assembly alignment) could probably stem from misassembly events instead of being real differences of genetic material between the assemblies. The high LA50 (Table 3) of the *medusa* scaffolded S3_Janetzki assembly could also mainly stem from misassembly events instead of real SVs. This assumption could pose the problem, that using alignment tools like *nucmer* which can't distinguish between real SVs and misassembly events, might distort serious SVs analyses, when assembly qualities are low and alignments are not further been examined in detail. Considering the facts, that there are many publication which refer SV analyses from *MUMer* utilities (including *nucmer*) (5,20,21) and that the real-SVs detection utility of *quast* is still under development (10, *quast m*anual), further investigations should focus on the question which way of SVs detecting proves to be the best.

A next improvement would be introducing assembly evaluation tools like *reapr* (22), which breaks incorrect fragments at misassembly events. By this, assembly quality would be improved as well as further SVs analyses would gain in reliability.

Summarized, the presented multi-referenced-based assembly demonstrates how effective the usage of multiple reference genomes in course of scaffolding is. I also shed light on the positive effects of joining multiple assemblies and read libraries together before applying further pipeline steps. The usage of short read only additionally provides a cost-efficient way in view of the gapclosing with *sealer* and scaffolding with *sspace.* On the other hand, there are still limitations. Most of all the handling of large, eukaryotic genomes in the process of scaffolding with *medusa*. With further investigation in the detections of real SVs, beside introducing programs that reliably detect and remedy misassembly events before, SV analyses as well as the multiple-reference-based assembly pipeline would be greatly improved.

6. Materials & Methods

6.1. Merging using *minimus2-blat*

For purpose of merging, I first concatenated the corresponding assemblies. After this, the concatenated sets of contigs were converted into Amos message format which has to be the input format for *minimus2-blat* by executing *toAmos* (13). I modified *minimus2-blat* to use *pblat* instead of *blat* as the genome overlap detector. *Pblat* works multi-threaded and proves be a more efficient and faster way than *blat* in order to merge assemblies of large eukaryotic genomes. All parameters were set to default instead of REFCOUNT which is the sequence count of the first assembly in each merging process. For S2_Lorenz, it was set to 89.306 and for S3_Janetzki, it was set to 137.972 respectively (*minimus2-blat,* REFCOUNT=89306 and REFCOUNT=137972).

6.2. Gapclosing using *sealer*

To run *sealer* accurately, PCR free, short paired-end read libraries were integrated into the pipeline. Short read libraries of the assemblies BnLorenz and BnJanetzki were concatenated with the libraries of the corresponding newer assemblies S2 and S3. In concrete terms, the read library of S2 was concatenated with the read library of BnLorenz. The same goes for S3 and BnJanetzki. I applied *sealer* on each of the merges as well as S2 and the S3 assembly. For both of the merges, the concatenated read libraries were included while for the S2 and S3 assembly, only their corresponding library was included. Sealer was executed two times for each assembly: Before using *sspace* and right afterwards. It was applied in both cases with default parameters except for the Kmer sizes (first execution: abyss-sealer, —k 95 -k 90 -k 85 -k 80 -k 75 -k 70 -k 65 -k 60 -k 55 -k 50; second execution: abyss-

sealer, -k 96 -k 95 -k 94 -k 93 -k 92 -k 91 -k 89 -k 88 -k 87) and a bloom filter size of 20 Gb (-b 20G) for both runs. The reason for using different Kmer sizes for the second run was the observation, that *sealer* was able to extract and close by far most of the gaps at kmer size 95 (Sealer reports , supplement). Due to the recommendation of the manual, I therefore decided to run it with kmer sizes distributed among the higher values from 96 to 87.

6.3. Scaffolding using *sspace*

I ran *sspace* using the same read libraries combination that was already used for *sealer*. Different kind of parameters and values had to be set for the „library files“ which *sspace* calls up: As the aligner, I decided to chose bowtie for every assembly. The expected insert size of the reads was set depending on the read libraries. For the S2 assembly, the whole read library is normally distributed with a maximum of ~ 2.500.000 reads at ~ 390 bp insert size. So the expected insert size was set at 390 bp with a minimal allowed error of 0,6 which means that the expected insert size can shift 390 bp * 0,6 = 234 bp in either direction. The insert size, due to the CLC report, is distributed from 190 to 594 bp. By setting the minimal allowed error to 0,6, *sspace* could capture to whole extent of the read library (390 bp - 234 bp = 156 bp < 190 and 390 bp + 234 bp = 624 vp > 594 bp). Under this consideration, I went on for the rest of the assemblies. For S3, the expected insert size was set at 350 bp with a minimum allowed error of 0,7. For both of the merges, the concatenated read libraries were used. They provide different insert size distributions within each library. For instance, the read library that was used to assemble BnLorenz shows an expected insert size distribution of 440-740 bp. Unfortunately, the particular distribution wasn’t shown in the report. That’s why I had to estimate the maximum at (440 bp + 740 bp) / 2 = 590 bp because it should be normally distributed in consideration of the rest of the

assemblies. The maximum for the S2 read library comes to 390 bp. Therefore, most of the reads should reside among 390 bp to 590bp of the concatenated read library. So, I decided to set the expected insert size at (390 bp + 590 bp) / 2 = 490 bp and the minimal allowed error at 0,6 in order to capture most of the reads. For S3_Janetzki, the particular read distribution for BnJanetzki was not available. So I went on calculating estimations like I did for S2_Lorenz. The expected insert size was set at 540 bp and the minimal allowed error was set at 0,7.

Sspace was finally executed with: no extension, minimum contig size for scaffolding = 500 bp, minimum overlap between contigs = 15 bp, minimum number of read pairs between contigs = 30, maximum link ratio = 0,5 and multi-threaded on 15 cores (sspace; -k 30, -a 0,5, -T 15, -x 0, -n 15, -z 500).

6.4. Filtering assemblies & scaffolding using *medusa*

Medusa uses multiple genomes as references. The genomes of *B. oleracea*, *B. brapa* and *B. napus* introduced for this purpose. Before using *medusa*, I had to filter the assemblies for following reason:

Medusa can't compute large assemblies, specifically assemblies with a large fragment number, in a decent time. Run time strongly increases with fragment number and size as well (A first test run lasted ~ 28 days just for the alignment). To optimize run time, I decided to reduce the assembly and reference genome complexity by filtering fragments smaller than a specific size depending on each assembly and reference genome respectively (See Table 3). On that point, *removesmalls.pl* (23) was executed (perl removesmall.pl, <minimum fragment size>). For the reference genomes, all unassigned fragments were removed (fragments < 15 000 bp). Because from several testings, I figured out that *medusa* silently dies after building the .coord and .delta files through the alignment when scaffolding large

genomes. The sub-script called netcon_mummer.py uses the .coord and .delta files but dies while buidling the graph-network. By personal correspondence with the author of *medusa*, I received the information that the size of the graph-network file is the bottleneck and it strongly correlates with fragment number of both the assembly and reference genomes. I also had to execute netcon_mummer.py separately after building .coords and .delta files from filtered assemblies and references. So I first ran *medusa*, also modified using *nucmer4* multi-threaded (to decrease run time) instead of *nucmer3* (java -jar medusa.jar; -f folder_with_reference_genomes/, -i assembly.fasta), with default parameters until all the .coords and .delta files were build and further executed netcon_mummer.py separately (python medusa_scripts/ netcon_mummer.py; -i assembly.fasta, -o network, -d 3, -w).

After this, I ran a second modified version of the whole *medusa* pipeline which skips both the alignment and network process and directly uses the network file as input to complete scaffolding (java -jar medusa_mod.jar; -f folder_with_reference_genomes/, -i assembly.fasta).

6.5. Assessing assembly statistics using *quast*

Quast was used before and after each tool execution for each assembly. Therefore, It was executed 4 times per assembly. The *B. napus* reference was used genome for alignment statistics. Different parameters were set up depending on the assemblies. Until *medusa*, I used to execute *quast* with following parameter setting: Both SNP and SV detection was disabled (—no-snp,—no-sv). For the reference genome option, It was set to „fragmented“ (—fragmented), because chromosomes instead of a single sequence were provided. The draft assembly options were set to eukaryotic and large as well (-e, —large). *Quast* was executed multi-threaded over 15 cores (python

quast.py; -R Bnapusreferencechromosomes.fasta, —no snp, —no-sv, —fragmented, -e, —large, -t 15).

After *medusa*, the gene finding (—gene-finding) was enabled in order to predict genes by using GeneMark-ES (10,*quast* manual). For the merged assemblies S2_Lorenz and S3_Janetzki, SNPs and SVs detecting were enabled by disabling -no-sv and -no-snp. Read libraries were also included for the SV detection.

(python quast.py;—pe1 reads_forward.fastq —pe2 reads_reverse.fastq, -R Brassicanapus_reference_chromosomes.fasta, —gene-finding, —fragmented, -e, —large, —t 15).

6.6. Genome-wide alignment & SV detection

I applied *nucmer —maxmatch* with default parameters defining S2_Lorenz as the reference and S3_Janetzki as the query genome (*nucmer;* —maxmatch, S2_Lorenz_scaffolds.fasta, S3_Janetzki_scaffolds.fasta). The output file (.delta file) was further filtered using *delta-filter*. The minimum alignment length was set to 50 bp (-l 50) and the minimum alignment identity was set to 95 % (-i 95) (*delta-filter*, -l 50, -i 95, output_nucmer.delta).

I applied *dna-diff* on the filtered .delta file using default parameters (*dna-diff*, filtered_output_nucmer.delta).

7. Acknowledgement

First of all, I would like to thank Prof. Dr. Benjamin Stich for giving me the opportunity to write my bachelor thesis in his wonderful institute of quantitative genetics from the University of Düsseldorf. Big thanks to Dr. David Ries who guided me through the whole project and helped me wherever he could with as much as power he could effort. It was a great time. Thanks to Marius Weißweiler, Dr. Amaury de Montaigu (Merci pour la chaise!) and Mara Pfeifer also for helping me, tolerating me and providing me a place in the office whenever I needed to. I'd also like to thank the team AG Genomforschung, especially Dr. Daniela Holtgräwe, from the University of Bielefeld. They made the whole project possible. A final big thank you goes to Roland Dieterich who remarkably helped me in terms of tool installations.

8. References

1. **Saxena RK, Edwards D, Varshney RK**. Structural variations in plant genomes. *Briefings in Functional Genomics*. 2014;13(4):296-307.
2. **Schneeberger, K., Ossowski, S., Ott, F., Klein, J. D., Wang, X., Lanz, C., et al. (2011).** Reference-guided assembly of four diverse Arabidopsis thaliana genomes. *Proceedings of the National Academy of Sciences of the United States of America*, 108(25), 10249-10254.
3. **Nam K, Jeong H, Nam J-W.** Pseudo-Reference-Based Assembly of Vertebrate Transcriptomes. Corominas M, ed. *Genes*. 2016;7(3):10.
4. **EZZI, Francesco; CATTONARO, Federica; POLICRITI, Alberto. e-RGA.** Enhanced Reference Guided Assembly of Complex Genomes. *EMBnet.journal*, [S.l.], v. 17, n. 1, p. pp. 46-54, aug. 2011. ISSN 2226-6089.
5. **Luis Zapata, Jia Ding, Eva-Maria Willing, Benjamin Hartwig, Daniela Bezdan, Wen-Biao Jiao, Vipul Patel, Geo Velikkakam James, Maarten Koornneef, Stephan Ossowski, and Korbinian Schneeberger.**
   Chromosome-level assembly of *Arabidopsis thaliana* L*er* reveals the extent of translocation and inversion polymorphisms. *PNAS* 2016 113 (28) E4052-E4060; published ahead of print June 27, 2016
6. **Chalhoub, Boulos & Denoeud, France & Liu, Shengyi & Parkin, Isobel & Tang, Haibao & Wang, Xiyin & Chiquet, Julien & Belcram, Harry & Tong, Chaobo & Samans, Birgit & Correa, Margot & Da Silva, Corinne & Just, Jérémy & Falentin, Cyril & Shin Koh, Chu & Le Clainche, Isabelle & Bernard, Maria & Bento, Pascal & Noel, Benjamin & Wincker, Patrick (2014)**. Plant genetics. Early allopolyploid evolution in the post-Neolithic Brassica napus oilseed genome. *Science* (New York, N.Y.). 345. 950-3. 10.1126/science.1253435.

7. **Wang, Xiaowu & Wang, Hanzhong & Wang, Jianxin & Sun, Rifei & Wu, Jian & Liu, Shengyi & Bai, Yinqi & Mun, Jeong-Hwan & Bancroft, Ian & Cheng, Feng & Huang, Sanwen & Li, Xixiang & Hua, Wei & Wang, Junyi & Wang, Xiyin & Freeling, Michael & Pires, J & H Paterson, Andrew & Chalhoub, Boulos & Zhang, Zhonghua (2011).** The genome of the mesopolyploid crop species Brassica rapa. N*ature genetics.* 43. 1035-9. 10.1038/ng.919
   FASTA download from: http://plants.ensembl.org/Brassica_rapa/Info/Index
8. **Parkin IA, Koh C, Tang H, et al**. Transcriptome and methylome profiling reveals relics of genome dominance in the mesopolyploid *Brassica oleracea*. *Genome Biology*. 2014;15(6):R77.
   FASTA download from http://plants.ensembl.org/Brassica_oleracea/Info/Annotation/#assembly
9. CLC Assembly (https://www.qiagenbioinformatics.com/)
10. **Iexey Gurevich, Vladislav Saveliev, Nikolay Vyahhi and Glenn Tesler.** QUAST: quality assessment tool for genome assemblies, *Bioinformatics* (2013) 29 (8): 1072-1075
    Quast Manual: http://quast.bioinf.spbau.ru/manual.html
11. **Bolger, A. M., Lohse, M., & Usadel, B. (2014)**. Trimmomatic: A flexible trimmer for Illumina Sequence Data. *Bioinformatics*, btu170.
12. **Kent WJ.** BLAT—The BLAST-Like Alignment Tool. *Genome Research*. 2002;12(4):656-664. doi:10.1101/gr.229202.
13. **Treangen TJ, Sommer DD, Angly FE, Koren S, Pop M**. Next Generation Sequence Assembly with AMOS. *Current protocols in bioinformatics / editoral board, Andreas D Baxevanis . [et al]*. 2011;CHAPTER:Unit11.8. doi:10.1002/0471250953.bi1108s33.
14. https://github.com/icebert/pblat

15. **Daniel Paulino, René L. Warren, Benjamin P. Vandervalk, Anthony Raymond, Shaun D. Jackman and Inanç Birol**. Sealer: a scalable gap-closing application for finishing draft genomes, *BMC Bioinformatics (2015),* 16:230
16. **Marten Boetzer, Christiaan V. Henkel, Hans J. Jansen, Derek Butler, Walter Pirovano**; Scaffolding pre-assembled contigs using SSPACE, *Bioinformatics*, Volume 27, Issue 4, 15 February 2011, Pages 578–579,
17. **Emanuele Bosi, Beatrice Donati, Marco Galardini, Sara Brunetti, Marie-France Sagot, Pietro Lió, Pierluigi Crescenzi, Renato Fani, Marco Fondi.** MEDUSA: a multi-draft based scaffolder, *Bioinformatics*, Volume 31, Issue 15, 1 August 2015, Pages 2443–2451
18. **Kurtz S, Phillippy A, Delcher AL, et al.** Versatile and open software for comparing large genomes. *Genome Biology*. 2004;5(2):R12. doi:10.1186/gb-2004-5-2-r12.

    MUMmer4 was used. https://github.com/mummer4/mumme
19. **Del Fabbro C, Scalabrin S, Morgante M, Giorgi FM (2013).** An Extensive Evaluation of Read Trimming Effects on Illumina NGS Data Analysis. *PLOS ONE 8*(12): e85024.
20. **Tan JL, Ng KP, Ong CS, Ngeow YF.** Genomic Comparisons Reveal Microevolutionary Differences in *Mycobacterium abscessus* Subspecies. *Frontiers in Microbiology*. 2017;8:2042. doi:10.3389/fmicb.2017.02042.
21. **Beghini F, Pasolli E, Truong TD, Putignani L, Cacciò SM, Segata N.** Large-scale comparative metagenomics of *Blastocystis*, a common member of the human gut microbiome. *The ISME Journal*. 2017;11(12): 2848-2863. doi:10.1038/ismej.2017.139.
22. **Hunt M, Kikuchi T, Sanders M, Newbold C, Berriman M, Otto TD.** REAPR: a universal tool for genome assembly evaluation. *Genome Biology*. 2013;14(5):R47. doi:10.1186/gb-2013-14-5-r47.

23. https://github.com/drtamermansour/p_asteroides/blob/master/scripts/removesmalls.pl)

9. List of Abbreviations

bp - **b**ase **p**airs

k - 1000 (number), e.g 1 kbp = 1000 bp

M - 1000.000 (number), e.g 1 Mbp = 1.000.000 bp

GC (%) - The total number of **G** and **C** nucleotides in the assembly, divided by the total length of the assembly (10).

N50 - The length for which the collection of all fragments of that length or longer covers at least half an assembly (10).

NG50 - The length for which the collection of all fragments of that length or longer covers at least half the reference genome (10).

L50(LG50) - The number of contigs equal to or longer than N50 (NG50).In other words, L50, for example, is the minimal number of fragments that cover half the assembly (10).

NA50, LA50 - ("A" stands for "aligned") are similar to the corresponding metrics without "A", but in this case aligned blocks instead of fragments are considered (10).

Ns - The number of ambiguous bases (10).

SVs - **S**tructural **V**ariations (10).

SNPs - **S**ingle **N**ucleotide **P**olymorphisms (10).

Indels - A **In**sertion or **Del**etion of bases.

Supplement

Quast Reports untill *medusa*

Genome statistics	S3 raw	Janetzki raw	S3_Janetzki – merge raw	S3_Janetzki after Sealer	S3_Janetzki after SSPACE	S3_Janetzki after Sealer, 2nd run
Genome fraction (%)	52.754	55.945	39.41	41.06	41.051	41.069
Duplication ratio	1.046	1.049	1.067	1.068	1.068	1.068
Largest alignment	33193	58899	54074	54074	54074	54074
Total aligned length	39878	423881	304412	317737(	3176810	3178236
NG50	604	1756	2444	2449	2462	2462
NA50	1942	4198	1928	2297	2308	2312
NA75	990	1637	–	–	–	–
NGA50	–	–	–	–	–	–
LG50	65070	80769	62807	62746	62351	62350
LA50	14461	31463	50190	45909	45665	45613
LA75	65070	76681	–	–	–	–
LGA50	–	–	–	–	–	–
Misassemblies						
# misassemblies	11386	13159	12808	13451	13488	13492
# relocations	1261	1860	1383	1457	1467	1467
# translocations	10084	11250	10347	10779	10810	10813
# inversions	41	49	1078	1215	1211	1212

# misassembled contigs	10847	11892	11461	12048	12064	12069
Misassembled contigs length	32329	706362	757542	802955	8130799	8135826
# local misassemblies	10952	31129	38830	37626	37722	37701
# scaffold gap size misassemblies	0	13	5	2	2	2
# unaligned mis. contigs	686	3967	13633	11014	10956	10922
Unaligned						
# fully unaligned contigs	23518	3441	4218	3935	3924	3924
Fully unaligned length	22193	309046	622257	574704	5719395	5719407
# partially unaligned contigs	10485	28984	56580	53044	52924	52887
Partially unaligned length	13002	469388	160570	147887	1479939	1478569
Mismatches						
# mismatches	16714	171865	118363	123250	1233007	1233233
# indels	26634	28410	213455	227507	227646	227812

Indels length	13887	227030	143995	1553399	1568699	1571580
# mismatches per 100 kbp	429.13	416.07	406.77	406.55	406.8	406.7
# indels per 100 kbp	68.38	68.78	73.36	75.04	75.11	75.13
# indels (<= 5 bp)	21635	217254	168394	183204	183250	183397
# indels (> 5 bp)	49985	66846	45061	44303	44396	44415
# N's	0	966249	324203	2876678	2876896	2872985
# N's per 100 kbp	0	1998.64	6794.47	6026.88	6027.53	6019.29
Statistics without reference						
# contigs	25776	137972	97736	97729	97176	97176
# contigs (>= 0 bp)	43718	137972	97742	97742	97176	97176
# contigs (>= 1000 bp)	15709	108001	89977	89968	89524	89525
# contigs (>= 5000 bp)	10262	30341	33765	33771	33766	33766

# contigs (>= 10000 bp)	673	8204	10851	10862	10943	10943
# contigs (>= 25000 bp)	5	278	530	530	548	548
# contigs (>= 50000 bp)	1	2	7	7	8	8
Largest contig	57494	102115	117006	117006	117006	117006
Total length	50061	483452	477161	477314	4772930	4772963
Total length (>= 0 bp)	50061	483452	477161	477314	4772930	4772963
Total length (>= 1000 bp)	37204	461864	471035	471184	4712534	4712578
Total length (>= 5000 bp)	69876	270905	321307	321445	3225787	3225804
Total length (>= 10000 bp)	82871	116971	160768	160925	1625912	1625921
Total length (>= 25000 bp)	16867	828308	161854	161894	1682619	1682632

Total length (>= 50000 bp)	57494	161153	448672	448683	509250	509250
N50	2277	5726	7263	7269	7321	7321
N75	1292	3006	4104	4105	4129	4129
L50	57284	24886	20009	20009	19862	19862
L75	122055	53874	41836	41835	41554	41554
GC (%)	35.32	35.44	35.42	35.41	35.41	35.41
Unique 101-mers						
Unique k-mers completeness	46.15	52.87	34.77	35.55	35.55	35.56
Similarity statistics						
# similar correct contigs	0	1931	454	658	657	659
# similar misassembled blocks	0	453	480	687	703	707

Genome statistics	S2 raw	Lorenz raw	Lorenz_plus_S2	S2_Lorenz after Sealer	S2_Lorenz after SSPACE	S2_Lorenz after Sealer, 2nd run
Genome fraction (%)	77.798	58.461	50.773	51.004	50.978	50.985
Duplication ratio	1.048	1.037	1.059	1.06	1.06	1.06
Largest alignment	105789	58899	105792	105792	105792	105792
Total aligned length	594973	4400710	3948366	39748314	3972575	39731916
NG50	9714	2377	11227	11224	11450	11450
NG75	1697		–	–	–	–
NA50	10051	5114	5951	6070	6163	6164
NA75	3609	2130	–	–	–	–
NGA50	6162	798	–	–	–	–
LG50	22429	64722	19204	19203	18782	18782
LG75	69499		–	–	–	–
LA50	18004	26771	20242	20051	19778	19776
LA75	45081	63604	–	–	-	-
LGA50	29840	102526	–	–	–	–
Misassemblies						
# misassemblies	19636	12241	12283	12508	12528	12536
# relocations	3496	1693	1988	2024	2038	2040
# translocations	16076	10498	10110	10277	10283	10288
# inversions	64	50	185	207	207	208
# misassembled contigs	15698	11003	9430	9608	9533	9539

Misassembled contigs length	157822	7387235	1342978	13637599	1399722	14005917
# local misassemblies	27520	25536	24441	23509	23691	23678
# scaffold gap size misassemblies	4	11	3	3	3	3
# unaligned mis. contigs	2103	3159	8493	8253	8123	8119
Unaligned						
# fully unaligned contigs	4739	2118	528	517	508	508
Fully unaligned length	919121	1934413	1573011	1529644	1501443	1501443
# partially unaligned contigs	21567	26355	28371	27855	27548	27535
Partially unaligned length	519830	4533950	1731723	17061572	1708887	17084039
Mismatches						
# mismatches	189682	237541	1151961	1169399	1169127	1169353
# indels	320328	2156860	209446	208757	208766	208708
Indels length	496734	320.49	3156525	3089025	3137098	3133649
# mismatches per 100 kbp	330.22	55.03	307.29	310.52	310.61	310.63
# indels per 100 kbp	55.77	180420	55.87	55.43	55.46	55.44

# indels (<= 5 bp)	236673	57121	156849	160044	159969	160039
# indels (> 5 bp)	83655	9847686	52597	48713	48797	48669
# N's	251192	199013	3026025	29093066	2909480	29077973
# N's per 100 kbp	378.76	237541	5294.32	5091.04	5091.7	5088.64
Statistics without reference						
# contigs	89299	119729	44767	44767	43908	43908
# contigs (>= 0 bp)	89306	119729	44767	44767	43908	43908
# contigs (>= 1000 bp)	87868	98111	44755	44749	43890	43890
# contigs (>= 5000 bp)	39782	33021	33237	33236	32801	32801
# contigs (>= 10000 bp)	21748	10524	21494	21488	21385	21386
# contigs (>= 25000 bp)	4343	517	5411	5411	5563	5563
# contigs (>= 50000 bp)	328	5	552	552	596	596
Largest contig	105789	101991	111785	111785	131328	131328
Total length	663195	4948263	5715604	57145589	5714160	57142965
Total length (>= 0 bp)	663195	4948263	5715604	57145589	5714160	57142965

Total length (>= 1000 bp)	661942	4791017	5715490	57143938	5713995	57141315
Total length (>= 5000 bp)	547874	3135142	5355846	53554103	5368015	53681405
Total length (>= 10000 bp)	418439	1559255	4494472	44938881	4530442	45306289
Total length (>= 25000 bp)	150507	1568185	1947011	19472595	2017943	20179759
Total length (>= 50000 bp)	195930	326388	3369471	33699969	3682076	36821186
N50	13678	6796	18779	18782	19219	19219
N75	6899	3633	11046	11053	11273	11273
L50	14325	21671	9627	9623	9402	9402
L75	31221	46499	19521	19513	19083	19083
GC (%)	35.65	35.48	35.35	35.35	35.35	35.35
Unique 101-mers						
Unique k-mers completeness	75.11	57.88	46.7	46.99	46.99	47
Similarity statistics						
# similar correct contigs	8542	2690	3762	4162	4064	4074

# similar misassembled blocks	4651	636	4150	4640	4763	4780

Genome statistics	S2 raw	S2 after Sealer (gapclosing)	S2 after SSPACE (scaffolder)	S2 after Sealer 2nd run
Genome fraction (%)	77.798	78.017	78.016	78.062
Duplication ratio	1.048	1.049	1.049	1.049
Largest alignment	105789	105789	105789	105789
Total aligned length	594973799	597232670	597223871	597670932
NG50	9714	9728	9728	9732
NG75	1697	1717	1717	1721
NA50	10051	10062	10062	10066
NA75	3609	3641	3641	3645
NGA50	6162	6194	6194	6201
LG50	22429	22407	22406	22401
LG75	69499	69218	69215	69153
LA50	18004	18001	18001	18002
LA75	45081	44986	44987	44975
LGA50	29840	29763	29763	29747
Misassemblies				
# misassemblies	19636	19643	19646	19634
# relocations	3496	3493	3492	3489
# translocations	16076	16087	16091	16081
# inversions	64	63	63	64
# misassembled contigs	15698	15695	15700	15688
Misassembled contigs length	157822122	158272810	158304246	158441780
# local misassemblies	27520	25944	25946	25719
# scaffold gap size misassemblies	4	4	3	3

# unaligned mis. contigs	2103	1867	1868	1835
Unaligned				
# fully unaligned contigs	4739	4658	4657	4641
Fully unaligned length	9191217	9030608	9025306	8975562
# partially unaligned contigs	21567	20530	20530	20352
Partially unaligned length	51983099	50693622	50707586	50488356
Mismatches				
# mismatches	1896828	1905969	1905934	1909419
# indels	320328	316052	316054	315658
Indels length	4967348	4773706	4773172	4747014
# mismatches per 100 kbp	330.22	330.88	330.88	331.28
# indels per 100 kbp	55.77	54.87	54.87	54.77
# indels (<= 5 bp)	236673	239858	239858	240381
# indels (> 5 bp)	83655	76194	76196	75277
# N's	2511929	1974398	1974399	1864986
# N's per 100 kbp	378.76	297.46	297.46	280.93
Statistics without reference				
# contigs	89306	89306	89296	89296
# contigs (>= 0 bp)	89306	89306	89296	89296
# contigs (>= 1000 bp)	87868	87903	87900	87912
# contigs (>= 5000 bp)	39782	39816	39815	39821
# contigs (>= 10000 bp)	21748	21767	21766	21769

# contigs (>= 25000 bp)	4343	4348	4347	4347
# contigs (>= 50000 bp)	328	329	329	330
Largest contig	105789	105789	109669	109669
Total length	663192277	663747878	663747458	663872741
Total length (>= 0 bp)	663195481	663751082	663747458	663872741
Total length (>= 1000 bp)	661942969	662527188	662526768	662662261
Total length (>= 5000 bp)	547874732	548366985	548366776	548471457
Total length (>= 10000 bp)	418439801	418841711	418841502	418918403
Total length (>= 25000 bp)	150507084	150694655	150694446	150707079
Total length (>= 50000 bp)	19593056	19648248	19674566	19725250
N50	13678	13680	13680	13677
N75	6899	6899	6899	6900
L50	14325	14334	14333	14335
L75	31221	31242	31241	31246
GC (%)	35.65	35.66	35.66	35.65
Unique 101-mers				
Unique k-mers completeness	75.11	75.3	75.3	75.32
Similarity statistics				
# similar correct contigs	8542	9222	9221	9344
# similar misassembled blocks	4651	4991	4991	5066

Genome statistics	S3 raw	S3 after Sealer (gapclosing)	S3 after SSPACE (scaffolder)
Genome fraction (%)	52.754	52.754	52.752
Duplication ratio	1.046	1.046	1.047
Largest alignment	33193	33193	33193
Total aligned length	398785520	398785520	398783877
NG50	604	604	604
NA50	1942	1942	1942
NA75	990	990	990
NGA50	-	-	-
LG50	225623	225623	225623
LA50	65070	65070	65070
LA75	144611	144611	144611
Misassemblies			
# misassemblies	11386	11386	11383
# relocations	1261	1261	1260
# translocations	10084	10084	10082
# inversions	41	41	41
# misassembled contigs	10847	10847	10845
Misassembled contigs length	32329529	32329529	32332147
# local misassemblies	10952	10952	10953
# scaffold gap size misassemblies	0	0	0
# unaligned mis. contigs	686	686	686
Unaligned			
# fully unaligned contigs	23518	23518	23519
Fully unaligned length	22193808	22193808	22194394
# partially unaligned contigs	10485	10485	10485
Partially unaligned length	13002723	13002723	13004485
Mismatches			
# mismatches	1671497	1671497	1671388
# indels	266343	266343	266342

Indels length	1388771	1388771	1388754
# mismatches per 100 kbp	429.13	429.13	429.12
# indels per 100 kbp	68.38	68.38	68.38
# indels (<= 5 bp)	216358	216358	216358
# indels (> 5 bp)	49985	49985	49984
# N's	0	0	0
# N's per 100 kbp	0	0	0
Statistics without reference			
# contigs	437184	437184	437184
# contigs (>= 0 bp)	437184	437184	437184
# contigs (>= 1000 bp)	157093	157093	157093
# contigs (>= 5000 bp)	10262	10262	10262
# contigs (>= 10000 bp)	673	673	673
# contigs (>= 25000 bp)	5	5	5
# contigs (>= 50000 bp)	1	1	1
Largest contig	57494	57494	57494
Total length	442815485	442815485	442815485
Total length (>= 0 bp)	442815485	442815485	442815485
Total length (>= 1000 bp)	372045584	372045584	372045584
Total length (>= 5000 bp)	69876028	69876028	69876028
Total length (>= 10000 bp)	8287130	8287130	8287130
Total length (>= 25000 bp)	168677	168677	168677
Total length (>= 50000 bp)	57494	57494	57494
N50	2277	2277	2277
N75	1292	1292	1292
L50	57284	57284	57284
L75	122053	122053	122053

GC (%)	35.32	35.32	35.32
Unique 101-mers			
Unique k-mers completeness	46.15	46.15	46.15
Similarity statistics			
# similar correct contigs	0	0	0
# similar misassembled blocks	0	0	0

Quast Reports after filtering & *medusa*

Genome statistics	**S3**	**S2_Lorenz**	**S3_Janetzki**	**S2**
Genome fraction (%)	13.108	33.985	20.891	50.79
Duplication ratio	1.008	1.034	1.02	1.011
Largest alignment	72093	177048	65466	232423
Total aligned length	97315149	259531275	157416137	379096904
NA50	6037	14533	5761	25204
NA75	4337	–	–	12527
NGA50	–	–	–	–
LA50	5487	6621	10786	4896
LA75	10875	–	–	10848
Reads mapping				
# mapped	123940415	503904517	283019496	
Mapped (%)	71.25	92.55	77.41	
# properly paired	85842932	350354140	171548670	
Properly paired (%)	49.35	64.35	46.92	
# singletons	7537558	16901642	23876718	
Singletons (%)	4.33	3.1	6.53	
# misjoint mates	13936960	86739326	56428652	
Misjoint mates (%)	8.01	15.93	15.43	
LAP score	15.528	14.853	15.21	
Reference LAP score	14.989	13.519	13.217	
Avg. coverage depth	159	215	153	
Coverage >= 1x (%)	98.63	83.09	81.7	
Coverage >= 5x (%)	98.59	81.51	79.4	
Coverage >= 10x (%)	98.53	80.8	78.2	
Misassemblies				
# misassemblies	13322	16383	14904	18502
# relocations	7743	8736	6894	9957

# translocations	5574	7613	7977	8514
# inversions	5	34	33	31
# misassembled contigs	1978	334	2221	1982
Misassembled contigs length	106588512	375406819	224784163	414506744
# local misassemblies	7587	26321	29577	26582
# scaffold gap size misassemblies	195	138	174	112
# structural variations	12	136	23	–
# unaligned mis. contigs	29	48	614	51
Unaligned				
# fully unaligned contigs	28	0	3	3
Fully unaligned length	148125	0	28222	42381
# partially unaligned contigs	1587	386	3006	2038
Partially unaligned length	11128910	123093150	93357344	41627718
Mismatches				
# mismatches	364369	679702	542603	972323
# indels	54926	127393	104889	170997
Indels length	569605	2563693	1009803	3876496
# mismatches per 100 kbp	376.47	270.87	351.78	259.28
# indels per 100 kbp	56.75	50.77	68	45.6
# indels (<= 5 bp)	43386	97158	83716	129808
# indels (> 5 bp)	11540	30235	21173	41189
# N's	1556800	21374672	17347006	2349027
# N's per 100 kbp	1430.82	5586.58	6916.95	558.21
Statistics without reference				
# contigs	2281	386	3026	2165
# contigs (>= 0 bp)	2281	386	3026	2165

# contigs (>= 1000 bp)	2281	386	3026	2165
# contigs (>= 5000 bp)	2149	386	3026	2165
# contigs (>= 10000 bp)	1910	386	2882	2165
# contigs (>= 25000 bp)	1270	372	2364	2016
# contigs (>= 50000 bp)	747	350	1640	1712
Largest contig	478452	11481033	1015252	1861219
Total length	108804404	382607675	250789876	420813103
Total length (>= 0 bp)	108804404	382607675	250789876	420813103
Total length (>= 1000 bp)	108804404	382607675	250789876	420813103
Total length (>= 5000 bp)	108212614	382607675	250789876	420813103
Total length (>= 10000 bp)	106420098	382607675	249605061	420813103
Total length (>= 25000 bp)	95690967	382350058	240269416	418313246
Total length (>= 50000 bp)	76697890	381561994	213822024	407026647
N50	81858	1999167	132918	324492
N75	43920	1009837	71326	176684
L50	398	54	566	395
L75	852	119	1211	825
GC (%)	36.23	35.48	35.72	35.56
Unique 101-mers				
Unique k-mers completeness	12.87	32.67	19.7	51.26
Length assigned to one chromosome (%)	0	0	0	0
Length assigned to multiple chromosomes (%)	1.1	51.57	2.54	22.76
Length assigned to no chromosome (%)	98.9	48.43	97.46	77.24

Similarity statistics				
# similar correct contigs	2	27	7	109
# similar misassembled blocks	81	4037	697	3880
Predicted genes				
# predicted genes (unique)	27265	68835	44083	99926
# predicted genes (>= 0 bp)	155044 + 0 part	380452 + 0 part	236363 + 0 part	579631 + 0 part
# predicted genes (>= 300 bp)	30164 + 0 part	73206 + 0 part	43943 + 0 part	108861 + 0 part
# predicted genes (>= 1500 bp)	2415 + 0 part	4122 + 0 part	2266 + 0 part	6805 + 0 part
# predicted genes (>= 3000 bp)	418 + 0 part	507 + 0 part	238 + 0 part	858 + 0 part

CLC Reports for BnJanetzki and BnLorenz

##
ASSEMBLIES OF 2 BRASSICA NAPUS GENOTYPES
##

Genotypes

P1 = B.napus Lorenz, low GSL content
P2 = B.napus Janetzki, high GSL content

Reads

Illumina reads

sample_id	total_Bases	mean_Coverage	%_bases_above_coverage_15
GSL-P1-Lorenz	26204488554	35.49	83.1
GSl-P2-Janetzki	23953244027	32.44	77.7

Estimated paired distance range:
  P1: 440-740 bp
  P2: 397-687 bp

All sequencing samples were pooled equimolar and sequenced on a HiSeq1500 in rapid mode and high output mode.
Clusters on the flowcell for a rapid run were generated by on-board Cluster generation using the TruSeq Rapid PE Cluster kit and sequenced 2x151 BP using the TruSeq Rapid SBS chemistry. Clustergeneration for a high output run was done on a cBot using the TruSeq PE Cluster Kitv3 and the flowcell was sequenced 2x101 BP using the TruSeq SBS Kit v3.

Assembly

quality trimming

To ensure, that only high quality reads and bases were used, we applied Trimmomatic for adapter trimming, filtering of reads with stretches of four concesutive bases with a mean quality value smaller than 30, and cutting bases of the read heads and tails with qualtity values smaller than 25. After trimming,the data was quality controlled using FastQC

assembly parameters

Assembled with CLC 7.5 with default parameters
  Automatic word size => 25
  Automatic bubble size => 50
  Mismatch cost 2
  Insertion / Deletion cost 3
  Length fraction 0.5
  Similarity fraction 0.8

assembly results

P1 reads assembled to 195,133 contigs and an assembly length of 545 Mb (N50 length: 6388).
P2 reads assembled to 239,586 contigs and an assembly length of 546 Mb (N50 length: 5285).
The coverage (or better: the alignment depth) was 51 for P1 and 46 for P2.

Fasta files with assembled contigs were splitted up into two files for each genotype:

LC = long contigs >= 500 nt
SC = short contigs < 500 nt
The SC were discarded. Many of them show very low values for coverage. By this we lose 61,680 P1-contigs and 84,678 P2-contigs which equals to 31.6 and 35.3 % of all contigs, but this is equivalent to only 3.5 and 4.7 % of the bases of the complete assembly.

statistics of the LC-assemblies

- BnLorenz.v1.fa (P1)

Sequences	133453
Total bases	526162018
including N's	12038399
Sequence length (avg)	3942
Sequence length (max)	101991
Sequence length (min)	500
N50 length	6649
N50 contigs	23408
GC content	34.6%

- BnJanetzki.v1.fa (P2)

Sequences	154908
Total bases	520250591
including N's	11649564
Sequence length (avg)	3358
Sequence length (max)	102115
Sequence length (min)	500
N50 length	5591
N50 contigs	27251
GC content	34.6%

Mapping on the reference sequence

The LC were mapped on Brassica_napus_v4.1.chromosomes.fa (Chalhoub 2014) with the program blat using the arguments "-minIdentity=95 -fastMap -extendThroughN". The longest match of adjacent blocks was used if the similarity fraction for the joint matches was >= 0.5 (using the in-house script blt). Afterwards the matches are ordered by their mapping position on the reference. Output format is gff3.

The assembled contigs (the FASTA files) were ordered as well by their mapping positions. Additional data as length of the contig, average coverage and mapping position was added to the description line of each FASTA entry. The unmapped fraction might contain plastid sequences and contaminations.

statistics of the mapped part of the LC-assemblies

- BnLorenz.v1.mapped.fa (P1)

Sequences	119729
Total bases	494826375
including N's	9847686
Sequence length (avg)	4132
Sequence length (max)	101991
Sequence length (min)	500
N50 length	6796

N50 contigs 21671
GC content 34.8%

- BnJanetzki.v1.fa (P2)

Sequences 137972
Total bases 483452794
including N's 9662496

Sequence length (avg) 3503
Sequence length (max) 102115
Sequence length (min) 500
N50 length 5726
N50 contigs 24886

statistics of the unmapped part of the LC-assemblies

- BnLorenz.v1.unmapped.fa (P1)

Sequences 13724
Total bases 31335643
including N's 2190713

Sequence length (avg) 2283
Sequence length (max) 83420
Sequence length (min) 500
N50 length 4198
N50 contigs 1943
GC content 31.5%

- BnJanetzki.v1.unmapped.fa (P2)

Sequences 16936
Total bases 36797797
including N's 1987068

Sequence length (avg) 2172
Sequence length (max) 83486
Sequence length (min) 500
N50 length 3726
N50 contigs 2590
GC content 32.2%


CLC bio
A QIAGEN® Company

Table of contents

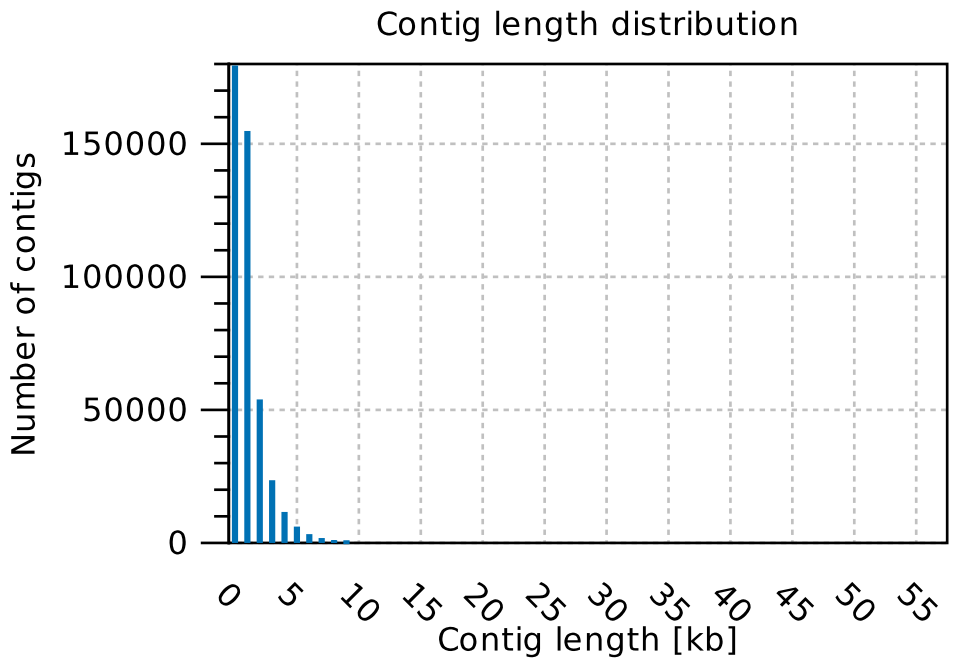
1. IC-2105_BNAP_S3_1_lib99656_4362_1_R2.paired assembly summary report

1.1 Nucleotide distribution

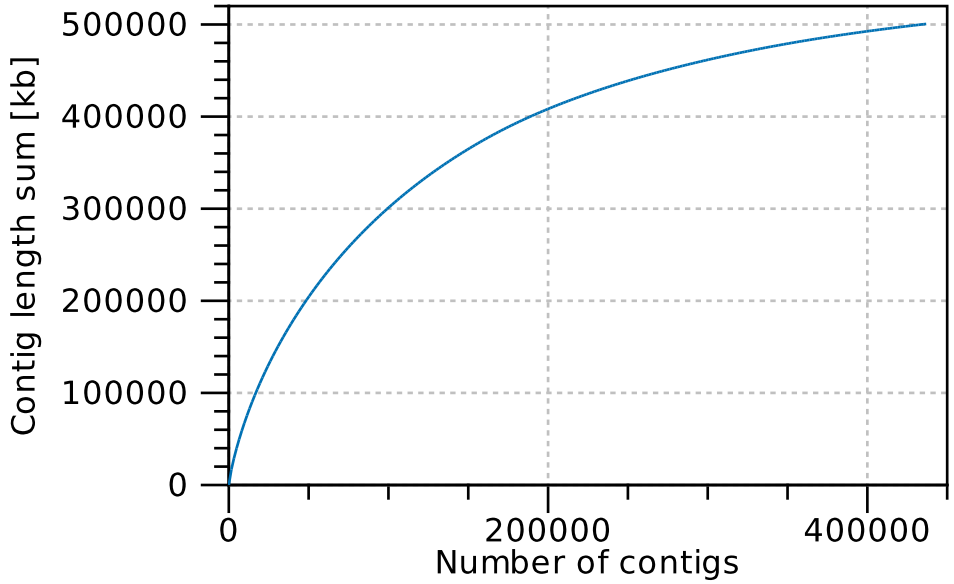
Nucleotide	Count	Frequency
Adenine (A)	161.780.387	32,3%
Cytosine (C)	88.920.438	17,8%
Guanine (G)	88.372.939	17,7%
Thymine (T)	161.545.264	32,3%

1.2 Contig measurements

N75	971
N50	1.980
N25	3.637
Minimum	61
Maximum	57.494
Average	1.145
Count	437.184
Total	500.619.028



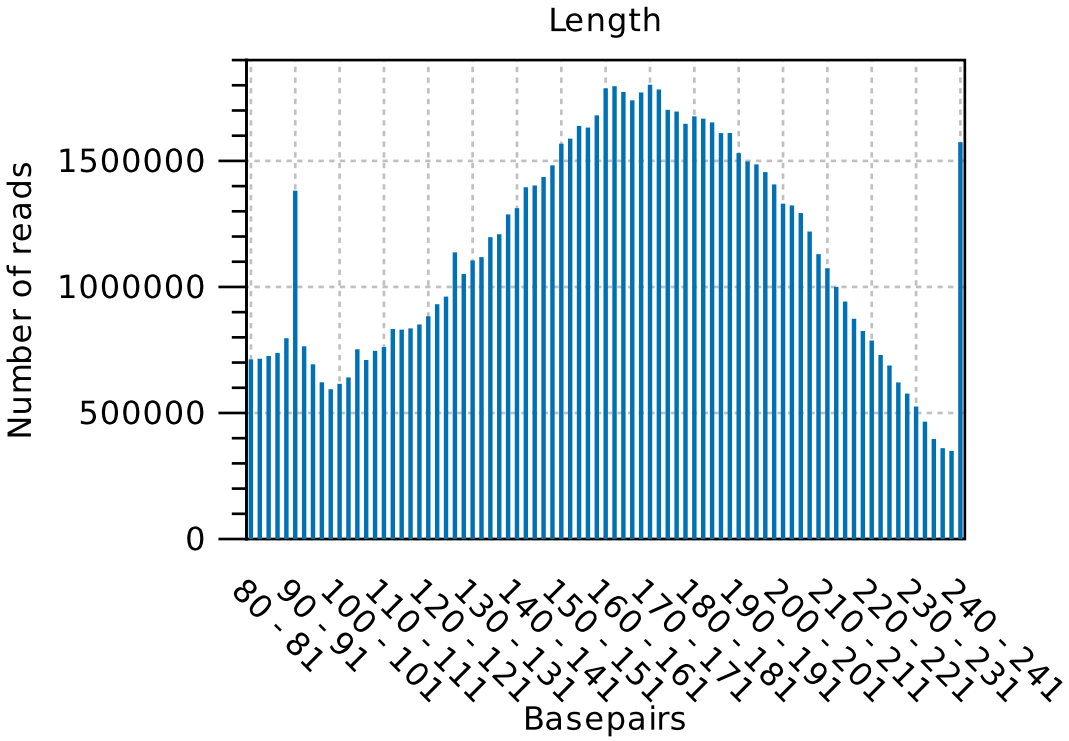
1.3 Accumulated contig lengths



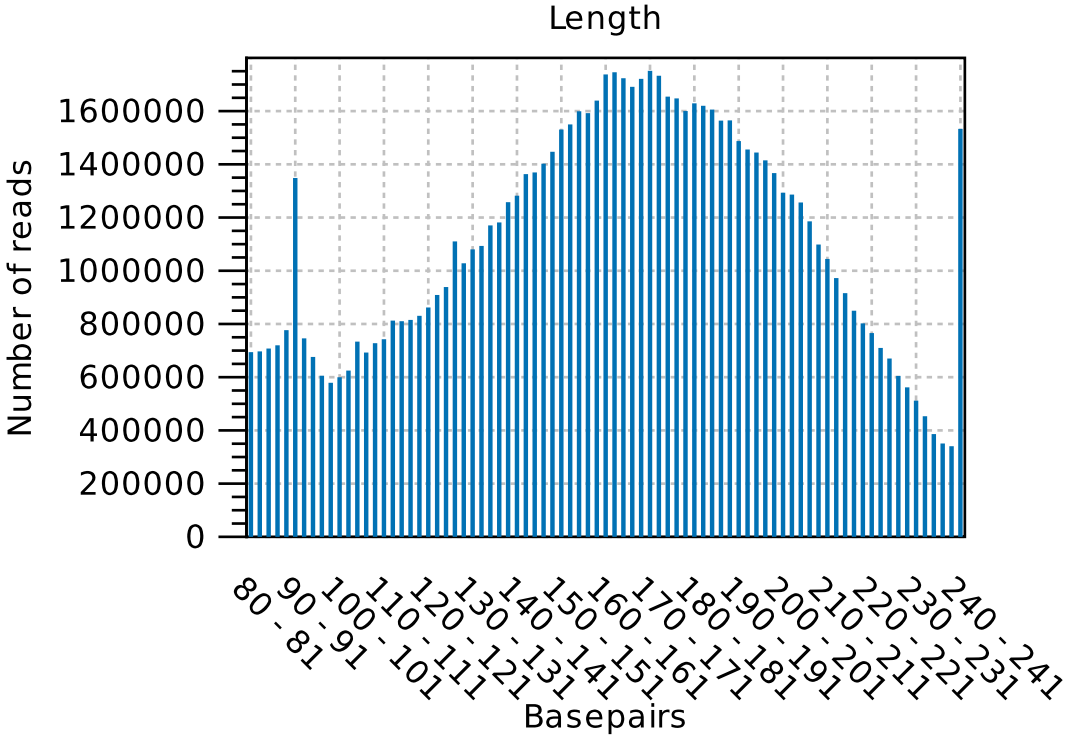
1.4 Summary statistics

	Count	Average length	Total bases
Reads	92.527.457	162,72	15.055.761.295
Matched	90.101.826	162,66	14.656.210.517
Not matched	2.425.631	164,72	399.550.778
Contigs	437.184	1.145	500.619.028

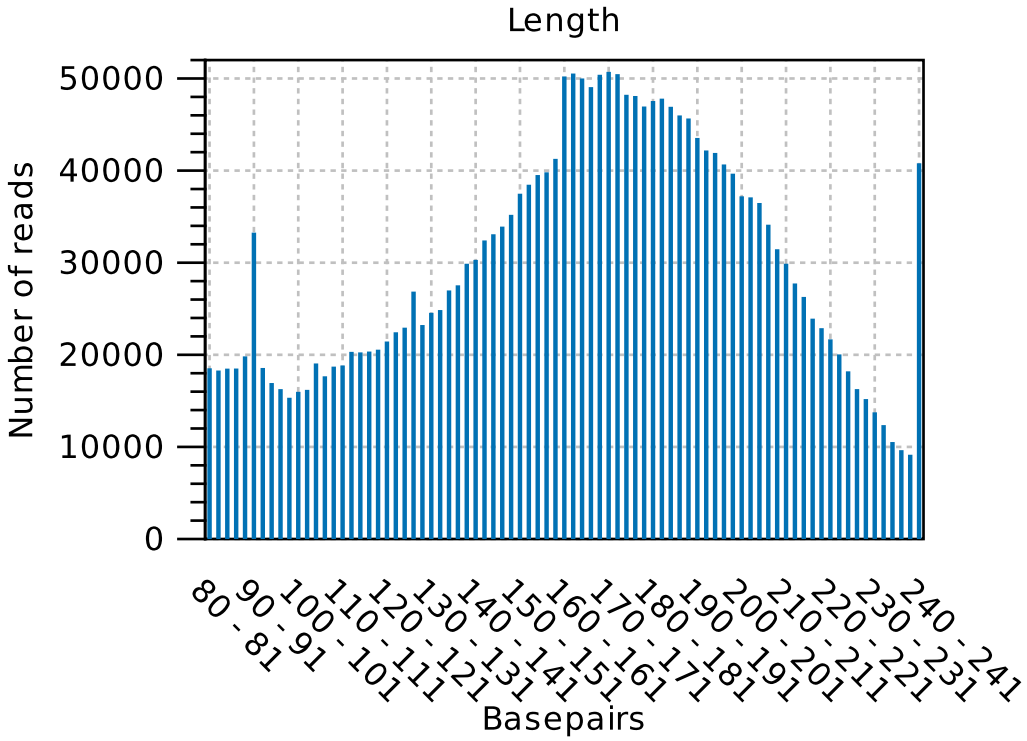
1.5 Distribution of read length



1.6 Distribution of matched read length



1.7 Distribution of non-matched read length




CLC bio
A QIAGEN® Company

Table of contents

1. BNAP_S2_R1 (paired) assembly summary report

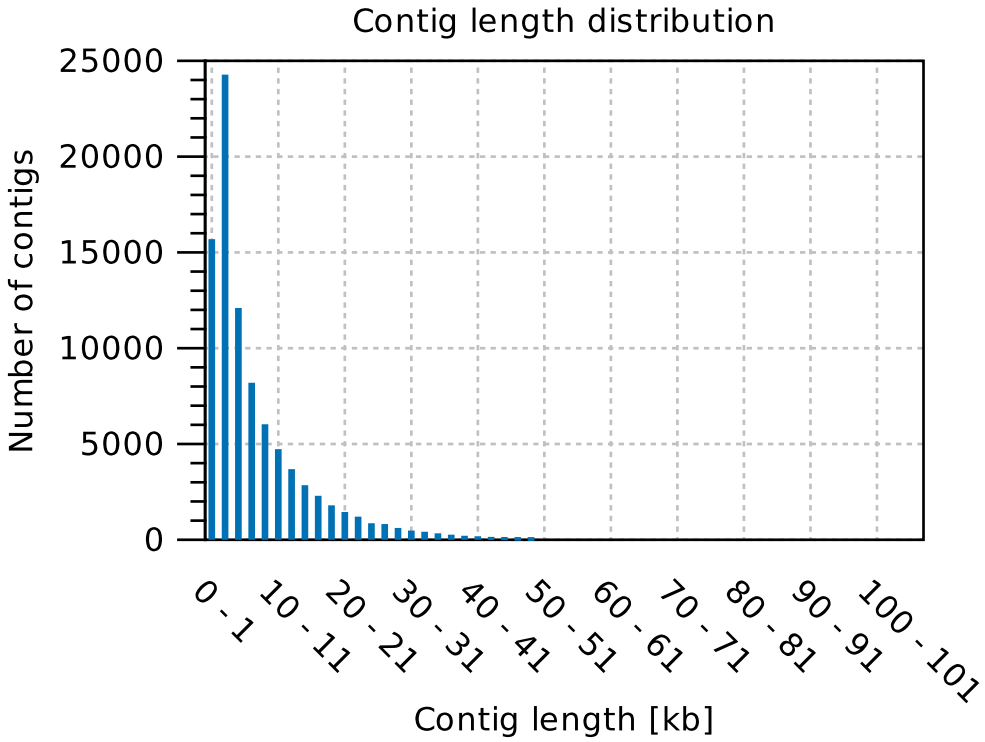
1.1 Nucleotide distribution

Nucleotide	Count	Frequency
Adenine (A)	212,564,947	32.1%
Cytosine (C)	117,807,482	17.8%
Guanine (G)	117,749,495	17.8%
Thymine (T)	212,560,106	32.1%
Any nucleotide (N)	2,513,451	0.4%

1.2 Contig measurements (including scaffolded regions)

N75	6,899
N50	13,678
N25	23,541
Minimum	428
Maximum	105,789
Average	7,426
Count	89,306

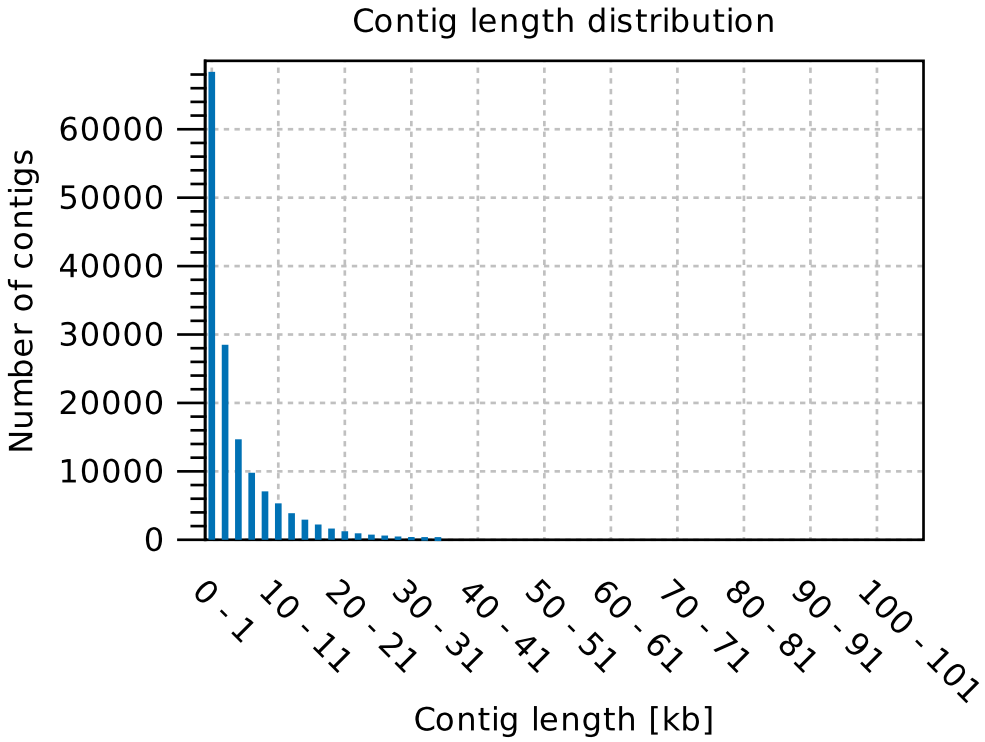
Total	663,195,481



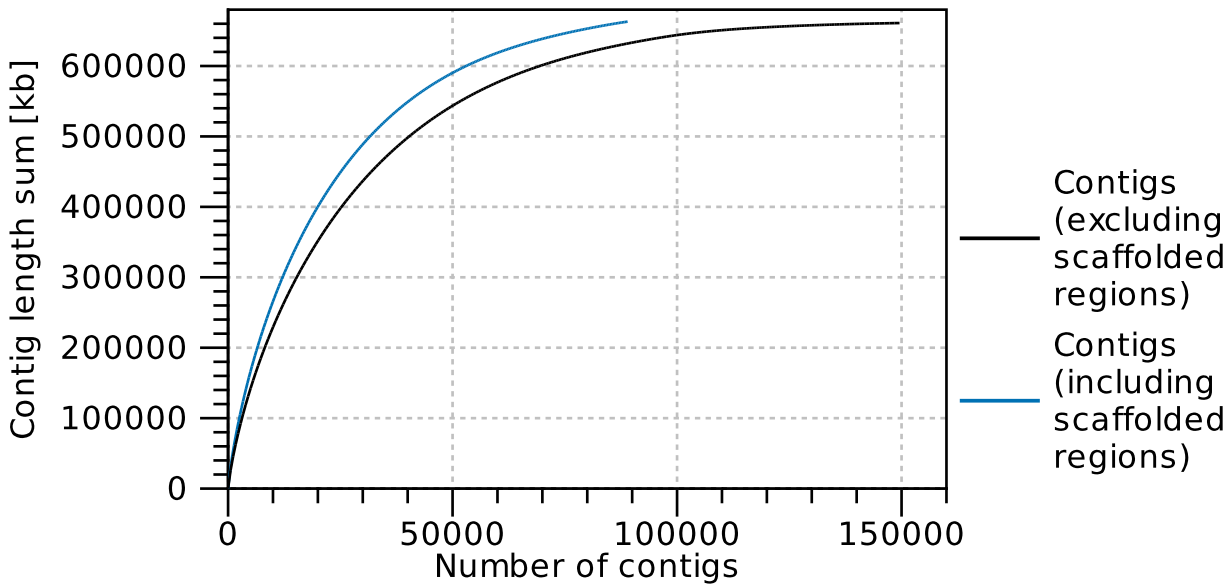
1.3 Contig measurements (excluding scaffolded regions)

N75	5,334
N50	10,791
N25	18,723
Minimum	1
Maximum	105,789
Average	4,405
Count	150,056

Total	660,959,526



1.4 Accumulated contig lengths



1.5 Summary statistics

	Count	Average length	Total bases
Reads	303,379,698	251	76,148,304,198
Matched	278,852,180	251	69,991,897,180
Not matched	24,527,518	251	6,156,407,018
Contigs	89,306	7,426	663,195,481
Reads in pairs	213,112,910	392.55	
Broken paired reads	65,739,270	251	

1.6 Distribution of read length

Length	Count
251	303,379,698

1.7 Distribution of matched read length

Length	Count
251	278,852,180

1.8 Distribution of non-matched read length

Length	Count
251	24,527,518

1.9 Paired reads distance distribution