Fish genomes : a powerful tool to uncover new functional elements in vertebrates

Stupka, E.

Chapter 5: Assembly of the carp genome

Elia Stupka[1,2] , Yanju Zhang[3], Chrstian Henkel[4], Hans Jansen[4], Geert Wiegertjes[5], Maria Forlenza[5], Ron Dirks[4], Herman P Spaink[6], Fons J Verbeek[3]

1 UCL Cancer Institute, University College London, Gower Street, London, WC1E 6BT, United Kingdom

2 Institute of Cell and Molecular Science, Barts and The London School of Medicine and Dentistry, 4 Newark Street, Whitechapel, London, E1 2AT, United Kingdom

3 Leiden Institute of Advanced Computer Science (LIACS), Universiteit Leiden, Niels Bohrweg 1, 2333 CA Leiden, The Netherlands

4 ZF-Screens BV, Bio Partner Center, Niels Bohrweg 11, 2333 CA Leiden, The Netherlands

5 Cell Biology and Immunology Group, Wageningen Institute of Animal Sciences, Wageningen University, Zodiac, Marijkeweg 40, 6700 AH Wageningen, The Netherlands

6 Department of Molecular Cell Biology, Universiteit Leiden, Einthovenweg 20, 2333 ZC, Leiden, The Netherlands

In preparation for publication

Abstract

In this study we present the assembly of the common carp (*Cyprinus carpio*) genome. We utilized only next-generation sequencing technologies applied to a combination of standard short DNA libraries as well as mate-pair libraries. We assessed several de novo assemblers (Abyss, SOAPdenovo, CLCBio) and parameters. Our final assembly was obtained by using CLCBio for contig assembly, and SOAPdenovo for scaffold assembly. We were thus able to assemble a genome of 1.647Gb, well in line with the estimated genome size for this organism, of which ~300Mbs which are found in gaps. The final scaffold N50 produced is of ~8Kb, thus presenting many scaffolds with complete gene structures. The largest scaffolds present very good collinearity with the zebrafish genome, and the mitochondrial genome has been completely covered in a single scaffold. Utilizing over 2,000 carp sequences available in Genbank (mostly gene fragments) we were able to show that the majority of sequences had 100% coverage in the current assembly, and most of them were recovered in a single scaffold. Based on Genbank carp sequences the assembly shos 99.5% coverage of existing data. Using our own contig obtained by assembling RNA-Seq data we were able to confirm that the assembly provides very good coverage of carp gene content (RNA-Seq contigs had median coverage of 98.7% and average coverage of 92.47%). However this data, which presents more complete gene structures than current Genbank datasets, indicates fragmentation of gene models in the current assembly (median number of hits=3), suggesting we ought to obtain further mate-pair library data to improve scaffold assembly and lead to less fragmentation.

Introduction

Cyprinus carpio (common carp) is one of the most important freshwater cultured fish species. It has been widely used in fish biology research[1]. A single female is capable of producing up to a few hundred thousand eggs that can be efficiently fertilized in vitro, which enables hundreds of thousands of pharmaceutical drug candidates to be tested with less genetic diversity. Thus, common carp is a relevant model system for high throughput screens of pharmaceutical compound libraries.

Microarray technologies have been widely used and have been remarkably successful in identifying genome-wide gene expression patterns. However, there are a number of shortcomings e.g. low sensitivity and specificity and low consistency across platforms, and, above all, they rely on a very accurate definition of the transcriptome for their design. Next generation sequencing is a high-throughput sequencing technology which can produce millions of sequence reads from DNA and cDNA in a few days at a low cost, without the need for a priori knowledge of the full transcriptome. In terms of expression profiling, in comparison to microarrays, NGS has much higher dynamic range, base-level resolution, richer splicing information and ability to detect previously unknown genes, as long as adequate sequencing depth is obtained.

The aim of this project is to obtain an assembly of the carp genome using de novo sequencing of carp DNA and an assessment of the transcriptome by deep sequencing of cDNA as well using existing data from other species (e.g. zebrafish). This chapter is structured as follows:

- Introduction

- Results:
    - Assessment of a preliminary assembly obtained from pseudo-tetraploid DNA comparing ABYSS [4] and CLCBio [5]
    - In-depth evaluation of different assembly strategies using sequence from haploid DNA, comparing SOAPdenovo and CLCBio, as well as testing several parameters for SOAPdenovo assembly.
    - Quality assessment of the assembly produced, based on alignment to the assembly of carp BAC sequences, carp Genbank sequences, and carp RNA-Seq contigs and analysis of percentage of query sequence aligned as well as number of hits obtained.
    -

Results

Initial Dataset: pseudo-tetraploid material

Initially a partially inbred carp strain, pseudo-tetraploid, was available and used to produce and sequence the following Illumina genomic libraries:

- Standard DNA library of 200bp DNA fragments, sequenced with 10 lanes of Illumina GAIIx 51bp paired-end reads and 2 lanes of single-end 51bp reads
- Mate-Pair 5Kb library sequenced using 7 lanes of Illumina GAIIx 36bp paired-end reads

The following additional data was used to QC and improve the assembly was downloaded from Genbank:

- 2,136 Genbank DNA records including the mitochondrial genome
- Zebrafish genome mapping

Furthermore two Illumina GAIIx 51bp paired-end lanes were used to sequence RNA samples using RNA-Seq, and the reads were de novo assembled into 10,197 contigs of length greater than 500bp, used to identify potential genes (and/or gene fragments) in genomic data. See [2] for more details of RNA-Seq assembly and initial contig assembly of genomic data.

Preliminary Genome Assembly

Using ABYSS and a varying K parameter from 20 to 30, several genome assemblies were generated. Owing to the low coverage. Using the N50 (i.e. 50% of the contigs are at least N50 long) we could assess the genome assembly that was most contiguous. Given the low coverage of the initial dataset (approximately 10X) we did not expect a very contiguous assembly. Indeed, as shown in Figure 1, the best N50 was obtained using K=23, and was 231bp, i.e. not significantly longer than the original DNA fragments, indicating that a large part of the genome was not assembled beyond the original DNA fragments. Given the low coverage obtained, increasing the N50 also has a significant effect in the overall portion of the genome that is found within the assembly, i.e. significant removal of redundant fragments in sub-optimal assemblies at lower K parameters. As shown in Figure 1 the overall size of the genome assembled into contigs of at least 100bps decreases significantly as K (and N50) increase.

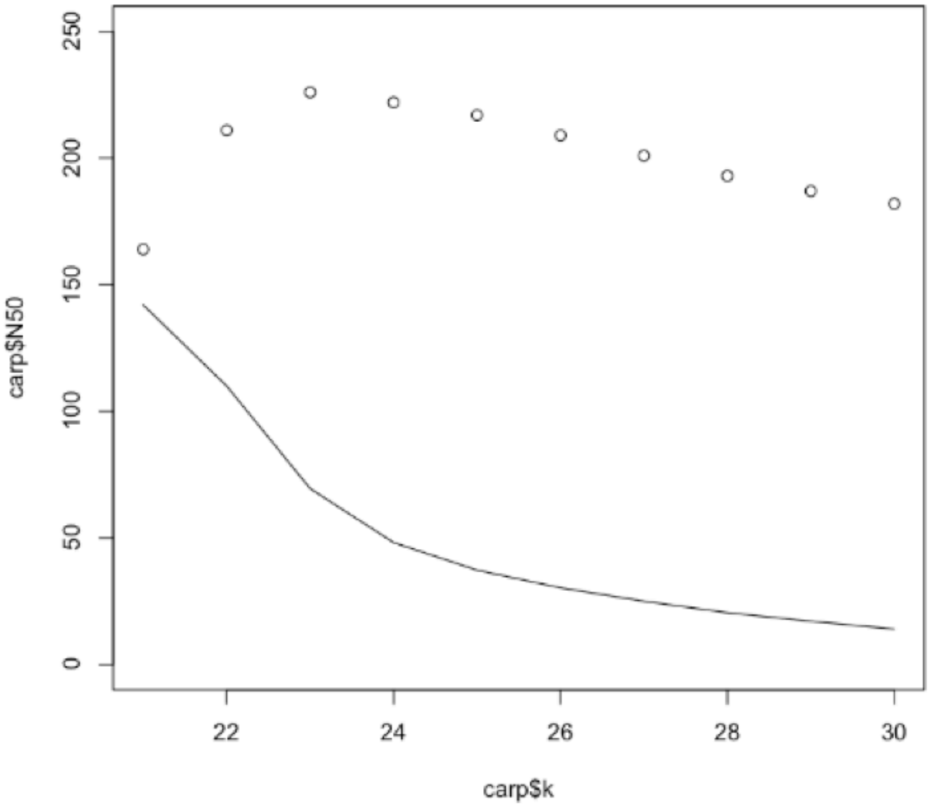


Figure 1 Preliminary assembly of the carp genome. Continuous line indicates the N50 of the assembly. Circles indicate overall assembly size (i.e. sum of the contigs longer than 100bp)

Subsequently further ABYSS assemblies were performed on this dataset introducing the scaffolding options available in ABYSS, but due to the low coverage, no significant differences were found in the final assemblies.

Haploid material assembly

Given the poor results obtained from the initial assembly, we then moved onto working from new material, i.e. deriving sequence reads from haploid DNA. Also, we changed assembly strategy, evaluating the commercial CLC Bio assembler, for contig assembly, and the SOAPdenovo tools (for both contig assembly and scaffolding). We also applied some pre-processing of the data prior to CLC Bio assembly (based on adaptor removal, trimming of low quality bases, and bridging of paired-end reads). The utilization of the CLC Bio assembler yielded a much improved contig assembly, with an N50 of 1,409bp. The additional pre-filtering steps improved the contig assembly further to an N50 of 2,260bp, as shown below.

Assembly	**n**	**n>100bp**	**n>50bp**	**median**	**mean**	**N50**	**Max**	**SUM**
CLC Bio	1637271	1635617	250045	384	735	1409	17597	1.20E+09
CLC Bio + pre-filtering	1086163	1083124	159656	587	1135	2260	26293	1.23E+09

Assembly strategy

In order to obtain a final carp genome assembly which would be both of good contiguity (i.e. with a high N50) as well as of good representation of the genome (i.e. with a high coverage of currently known carp sequences) we tested several strategies, briefly summarized below:

- Different algorithms: ABYSS (see above), CLCBio de novo assembler, SOAPdenovo, as well as combination of CLCBio and SOAP
- Different SOAPdenovo K paramters: K tested from 33 to 40
- Different SOAPDenovo L (length of contig used for scaffolding) parameter (from 70 to 400)
- Trimming of 200bp sequence reads as well as 5Kb mate-pair reads

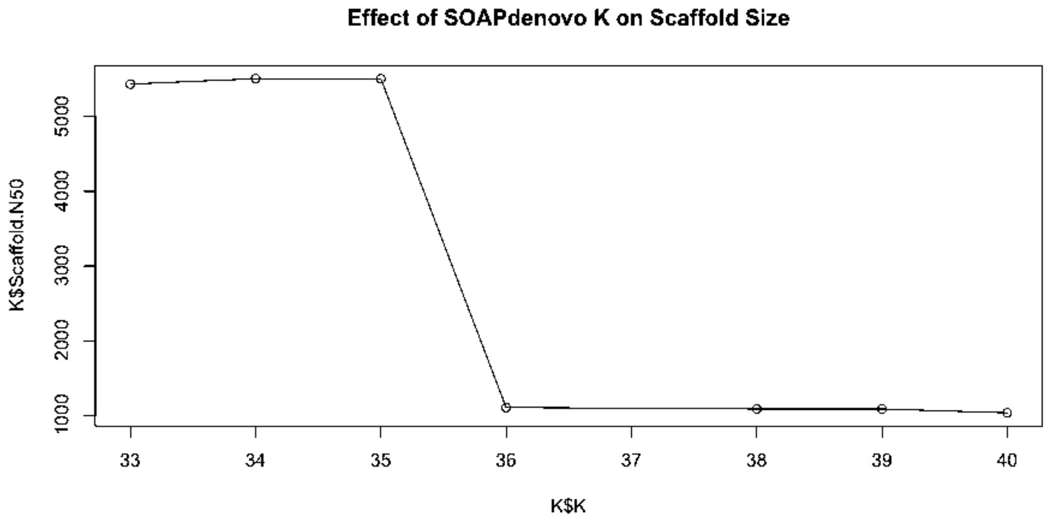
The chapters below detail the results obtained when varying each of these parameters.

Varying the K parameter in SOAPdenovo

De novo assembly algorithms implement an efficient strategy for a task that is computationally very demanding, i.e. the comparison of all short sequence reads against each other. This is performed in order to identify which reads align to each other. This data in turn forms the basis of a graph that is used to identify the optimal path to reconstruct the full genome. Taking into account that for a genome such as the carp genome one is usually starting with 10^9 reads, these tools have to make 10^9 x 10^9 comparisons, i.e. 10^{18} comparisons. If traditional alignment approaches were to be used, the task would be computationally impossible on standard servers. As an example, if each alignment took 1 CPU second, it would require 10^{10} CPU years to obtain all the alignments. In order to drastically reduce the CPU time required, therefore, de novo assembly algorithms implement a K-mer based approach, i.e. before comparing sequences to each other, the sequences are rapidly scanned for all possible K-mers of a user-defined length. All subsequent steps are then performed on a K-mer representation of the original sequences, reducing drastically computational time required. This is a gross approximation of a full alignment strategy, which works well because of the sheer amount of sequences involved in the assembly process. The optimal size of K-mers to be used, however, cannot be easily determined a priori, since each specific length will leads to quite different, and unpredictable, results. Generally speaking K-mers which are approximately half the size of the sequence read tend to produce good results, but specific K-mer lengths (represented as the K parameter in all algorithms) have to be tested. Moreover it is important to note that the "optimal" K might be different depending on what is taken as a measure of success. As shown in our work a specific K-mer length might lead to better N50 but slightly lower overall sequence quality and viceversa.

In the CLCBio package until recently the user was not able to specify the K parameter to be used, and the package did not report the K parameters chosen. In recent releases this has been modified.

Thus, in order to test the effect of the K parameter on the final assembly, we deployed separate SOAPdenovo assemblies for K-mer length ranging from 33 to 40 (since the reads are 76bp, and thus K=38 should be the optimal K). As shown in Figure 2, a K = 35 leads to the most contiguous scaffold assembly (N50 = 5,510bp), while K=39 lead to the most contiguous contig assembly (N50 = 689). Notably, while the scaffold N50 is better than the N50 obtained with the CLC Bio software, the best Contig N50 is still well below that obtained by CLC Bio.



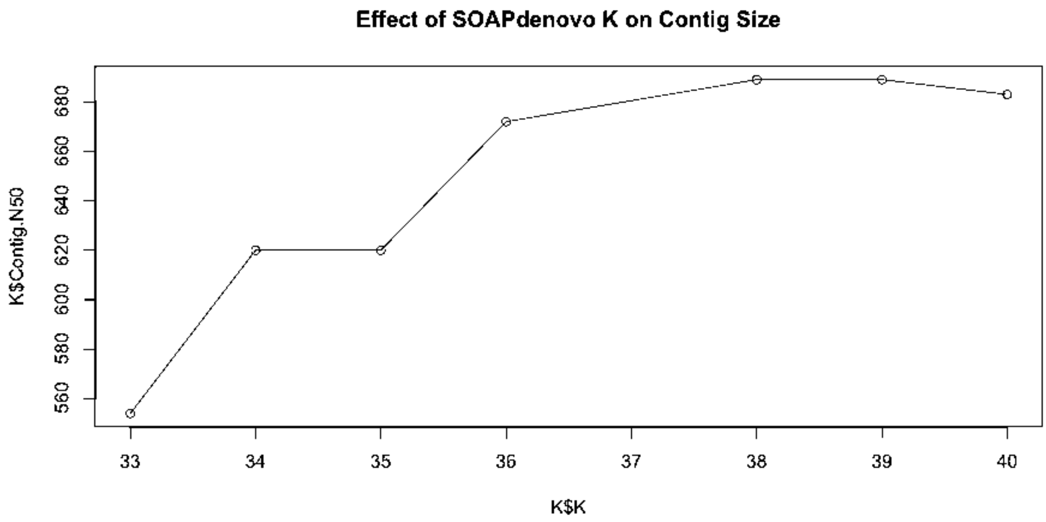


Figure 2 Assembly contiguity (top=scaffold, bottom=contig, based on the N50 statistics) as a function of the K parameters utilized for obtaining the assembly

Varying the L parameter in SOAPdenovo

One of the parameters which can be modified in SOAPdenovo is also the minimum length of contigs which are used for scaffolding. We thus tested from the minimum (2*k, i.e. in our case 70) to 400, and found that in terms of final N50 the optimal L is equal to the minimum, i.e. 70.

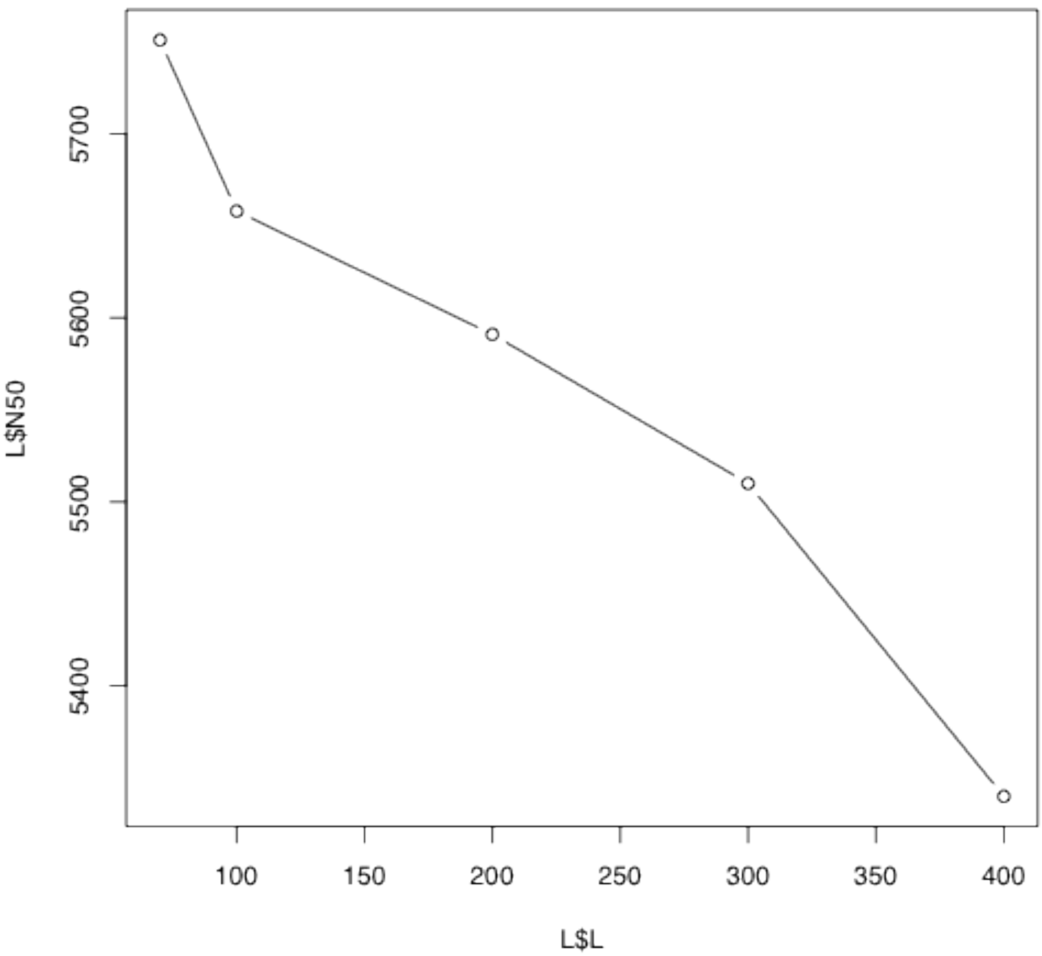


Figure 3 Assembly contiguity (based on the N50 of the scaffold length) as a function of the L parameter utilized for obtaining the assembly, tested for default L=70, as well as for L=100,200,300,400.

Testing read trimming strategies

Illumina sequencing reads have a quality profile which decreases as a function of read length, especially in the final part of the read. It is therefore often common to trim sequencing read ends in order to improve overall sequence quality and decrease. We therefore tested whether "hard trimming" (i.e. removal of a certain number of bases from the end of the sequence, regardless of its quality) would improve the assembly. We tested trimming of all reads (200bp library and 5Kb library) to 74bp and 72bp. As shown in Figure 4 the trimming of the last 2 and 4 bases had a marginally negative effect on the final N50 (N50 was 10bp less for the 74bp trimmed reads, and 50bp less for 72bp trimmed reads), and we thus did not utilize trimming of all reads for the final assembly. This reflects the fact that the majority of sequence reads were of good quality and thus "hard trimming" (i.e. trimming of all reads to a certain length, regardless of quality) leads to an overall loss of information content because the number of bases of good quality outweighs those of poor quality, which would be beneficial to remove. A better approach would be to trim sequence reads to a variable length, by trimming only low quality base pairs. Variable length reads, however, are not compatible with some softwares and aligners, and thus are not an ideal choice for the overall protocol.

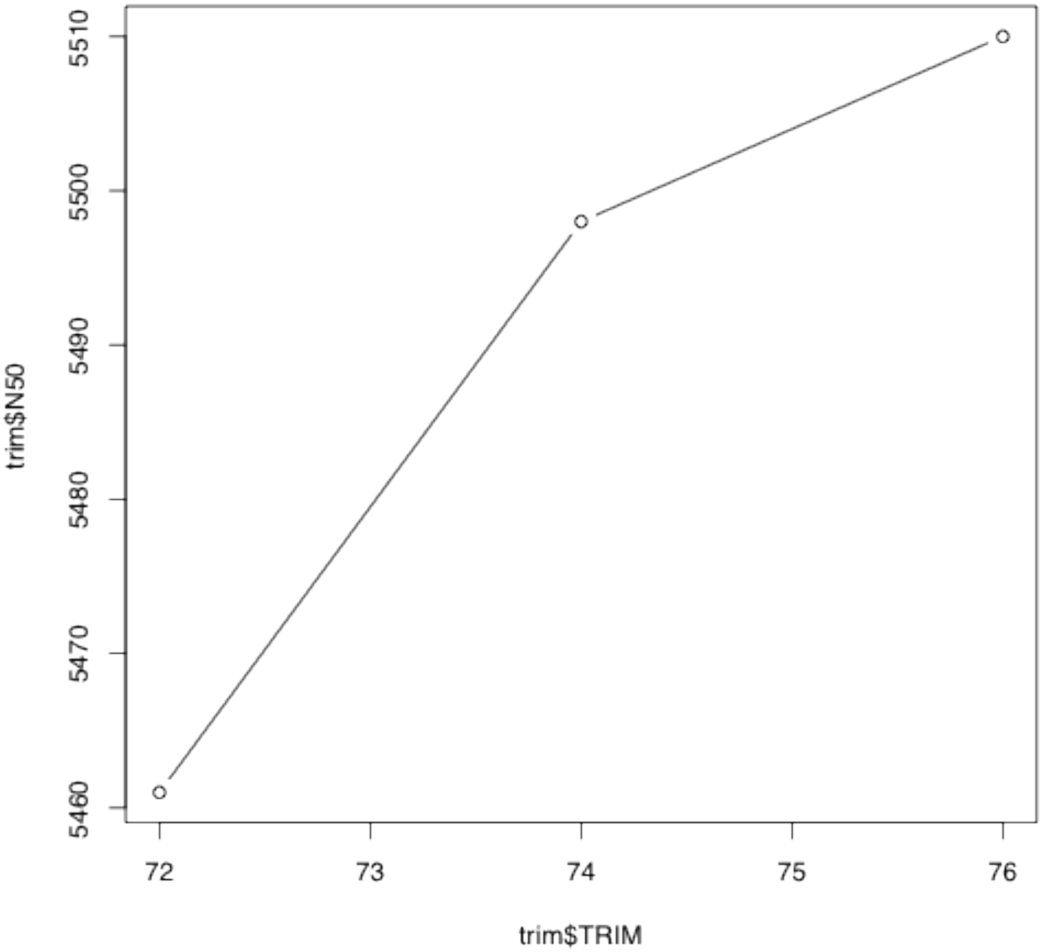


Figure 4 Assembly contiguity (based on the N50 of the scaffold length) as a function of the length to which reads were trimmed (full read = 76bp) Linear fit or Exp fit more realistic ? include datapoints

Mate-pair libraries often present a different issue which affects the final quality of the assembly, referred to as the "read-through" problem, i.e. that some of the fragments generated from the mate pair library might be very proximal to the junction of the mates (i.e. the point where the circuralization of the 5Kb fragment occurs), and thus, when sequencing at longer read lengths (as in our case 76bp), the sequence might "read through" into the opposite mate, which would then cause issues when mapping. We therefore tested the trimming of the 5Kb library reads to 35bps. When K is smaller than the trimmed length (i.e. in our case K=33, shorter than 35bp length) trimming has a minor negative effect (from an N50 of

5,438 to an N50 of 5424). When K >= 35, it actually produces a much poorer assembly, since K is as long (or longer) than the read itself and thus fails to produce correct alignments. Overall the results suggest a similar conclusion to minor trimming of all reads above, i.e. that due to limited read-through issues, the disadvantages of trimming outweight the advantages.

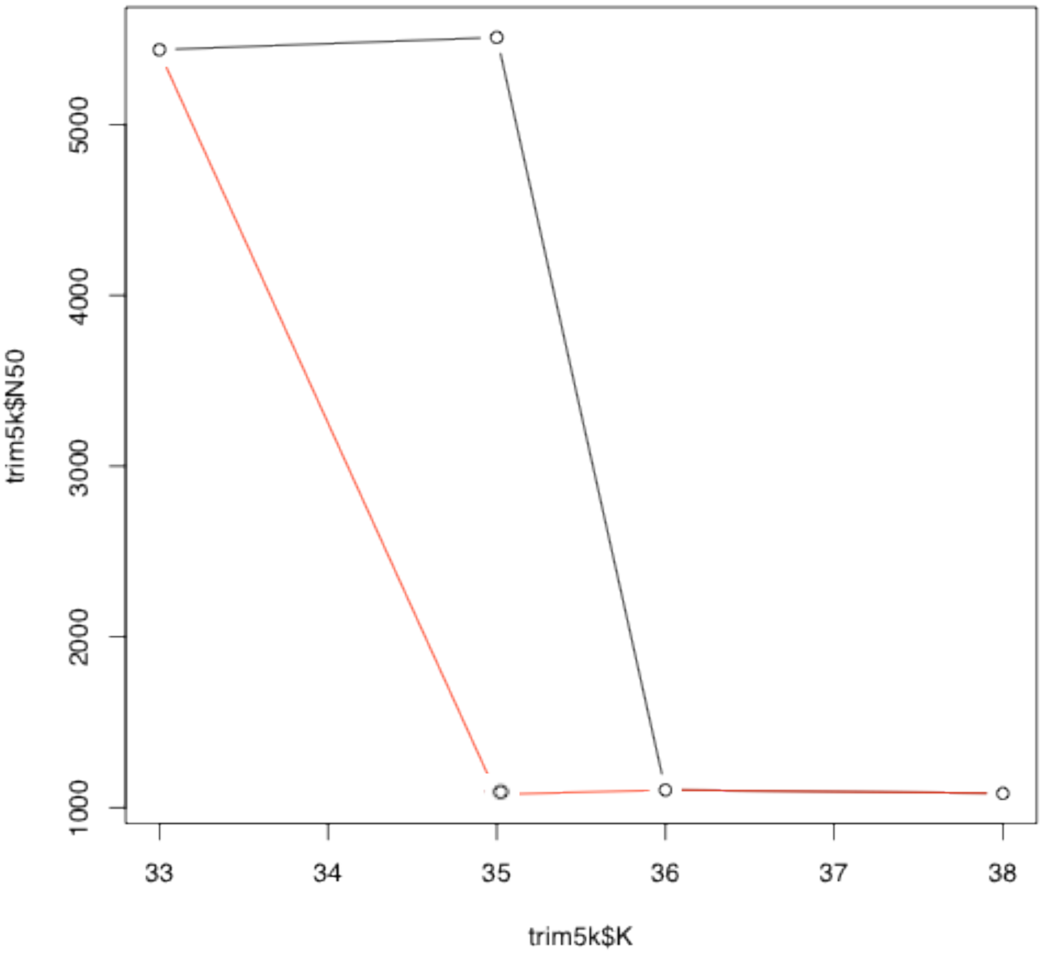


Figure 5 Comparison of assembly scaffold N50 at several K parameters, without trimming reads (black) and trimming 5Kb library reads to 35bps (red)

Testing combination of assembly softwares

CLC Bio yielded a better contig N50 than all the SOAP approaches, but it provides no scaffolding capability, therefore we decided to test a combined approach utilizing the SOAP scaffolding tools on the contigs generated using the CLC Bio software. This yielded the assembly with the best Scaffold N50 (8,043bp), far superior to all other approaches utilized previously. In order to do this we had to contact the developers of SOAPdenovo at BGI to obtain a separate tool which allows to prepare existing contigs for a particular K parameter for the subsequent mapping and scaffolding steps peformed by SOAPdenovo. This was performed with a compiled program named *prepare* kindly provided directly by the SOAPdenovo team, not currently released in the public domain.

Adding BAC end reads

Recently a research group in China published a small number of BAC end sequences (2,688) for the carp genome [2]. We contacted the authors and obtained the FASTA sequences for these BAC end reads. We then converted them to the appropriate format required by SOAPdenovo and tested whether the addition of these BAC end reads would improve the assembly. However the final assembly obtained by adding the BAC end sequences was actually marginally worse, with an N50 of 7,803 bp. This is probably due to the very low number of BAC end sequences utilized. A further dataset is being produced by the same laboratory of more than 80,000 sequences, which in the future could significantly aid the assembly.

Assembly Statistics

Following all the above attempts we opted for our final assembly for the following strategy:

1. Contigs were obtained by pre-processing reads (checking for overlaps between paired reads, so that they can be merged and discarding low quality nucleotides) and then using the CLC Bio software for de novo assembly, which yielded a set of contigs with N50 of 2,262bp.
2. These contigs were then prepared for SOAPdenovo assembly using a “prepare” script kindly contributed by the BGI SOAPdenovo team.
3. Scaffolding was then performed using SOAPdenovo by using both the 200bp reads and the 5kb reads without trimming, using a K=35, and default L and G parameters (L=K*2, G=50), the R flag (,i.e. use reads to solve tiny repeats) and M=3 (maximum strength for merging similar sequences)

The statistics of the assembly thus obtained are as follows:

Total number of sequences: 779,686

Total Size: 1,647,732,536 (just below the average of the estimated genome sizes published in genomesize.com, which is 1.71Gb)

Contig Statistics:

Minimum:	100
Maximum:	26327
Average:	1137.73
N50:	2262

Scaffold statistics:

Minimum:	100

Maximum: 115,091

Average: 2113.33

N50: 8,043

A comparison with the CLC Bio contig assembly shows that we thus produced a significantly lower number of total sequences, with a total size that is significantly more in line with the estimated genome size for the carp genome, and with a 4x improvement in terms of final N50:

Number of sequences: 1,086,163

Total Size: 1,235,762,671

Sequence lengths:

Maximum: 26,327

Average: 1,137.73

N50: 2,262

The assembly displays a clear log normal distribution of scaffold sizes, with a small increase of scaffolds of length similar to the overall assembly N50 size, i.e. between 5kb and 8kb.

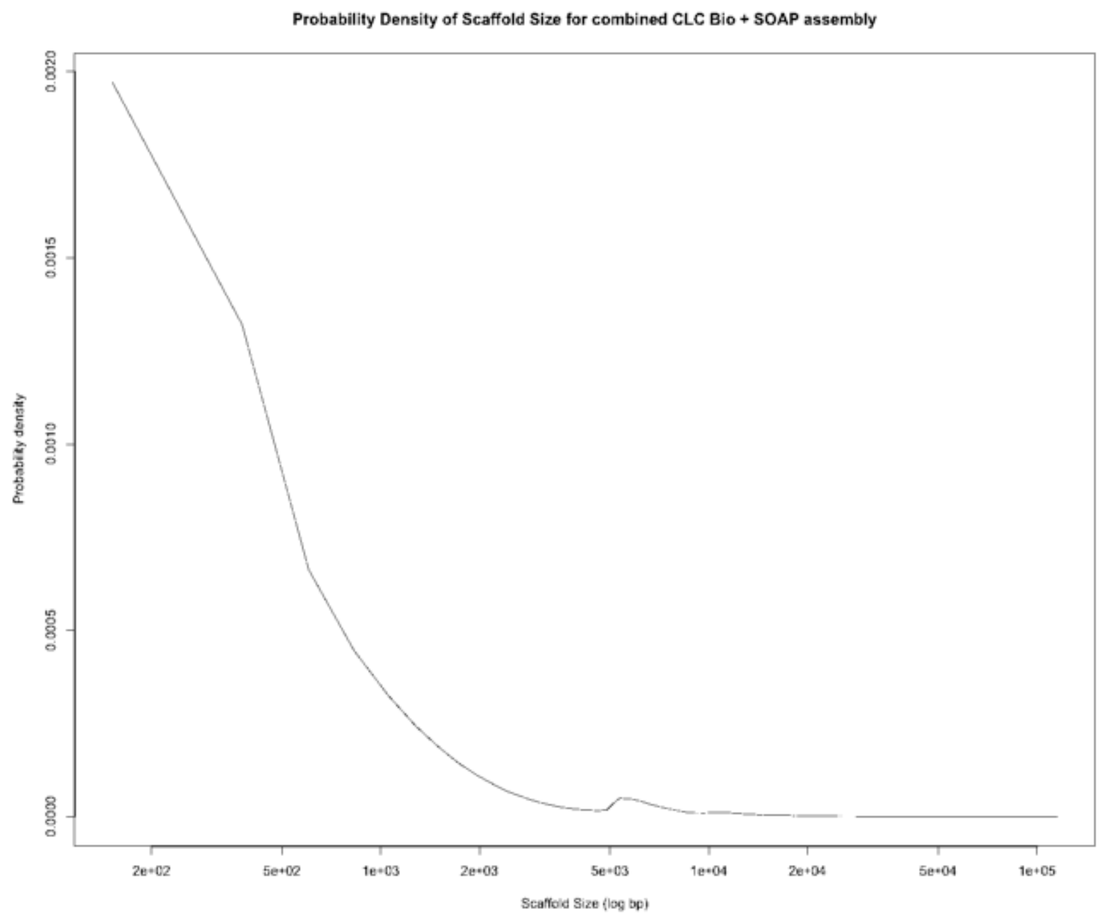


Figure 6 Density distribution of length of scaffolds in final assembly (X axis: length in base pairs). The figure shows that there is a slight preference for scaffolds of approximately 5Kb, i.e. scaffolds of N50 size.

Largest scaffolds

As a preliminary overview of the quality of the final assembly, we used BLAT [6] on the Ensembl genome browser [7], to map the largest scaffolds produced on the zebrafish genome. Owing to the nucleotide BLAT-based search only high similarity hits were found across the scaffold matching to exons and conserved regions within the genome. Reassuringly, all scaffolds were found to match a single location on the zebrafish genome in a collinear manner (see Table below, as well as Figures 7-9), indicating good synteny between these two genomes, and also showing that these largers scaffolds are not due to major assembly artifacts generated during the assembly process. Interestingly, one of the largest scaffolds (scaffold24544) contains the two largest known vertebrate genes, i.e. Titin A and part of Titin B.

Scaffold ID	Length	Zebrafish top match	Gene Names	Gene descriptions
scaffold1889	115,091	chr16:26.44Mb-26.55Mb	cadm4	cell adhesion molecule 4
scaffold8659	106,795	chr16:35.20Mb-35.30Mb	me1	cytosolic malic enzyme 1
scaffold24544	100,347	chr9:43.8Mb-44.01Mb	ttnb,ttna	titin b, titin a

Table 1 Mapping of carp scaffolds larger than 100Kb to the zebrafish genome, indicating zebrafish chromosomal location for main hit, as well as names and descriptions of genes contained in the zebrafish locus.

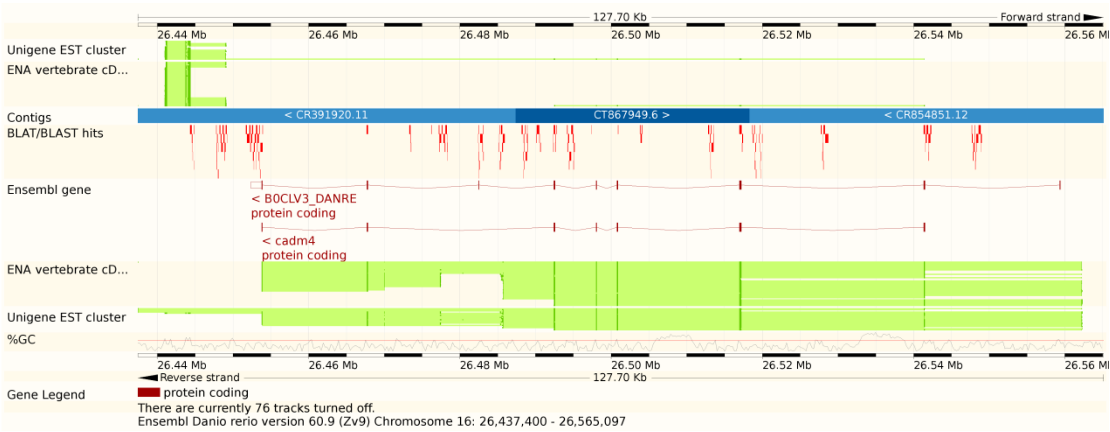


Figure 7: Scaffold1889 mapping to zebrafish chromosome 16

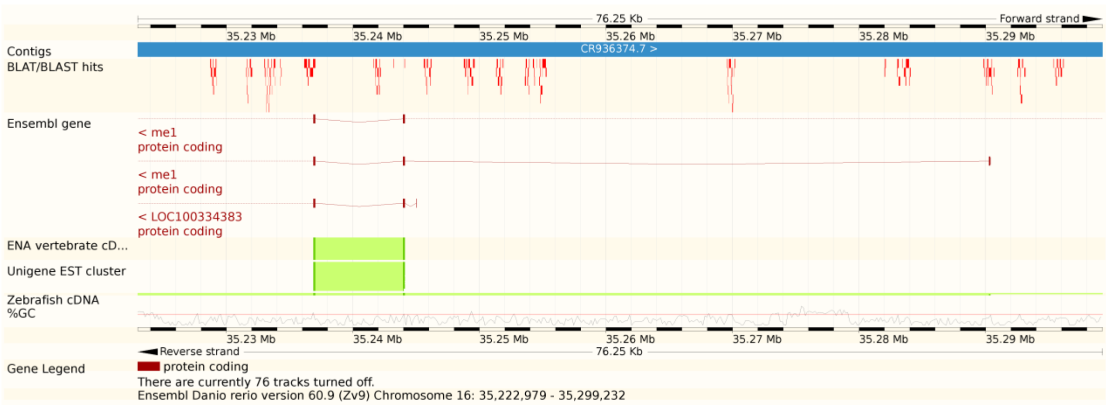


Figure 8 Scaffold8659 mapping to chromosome 16

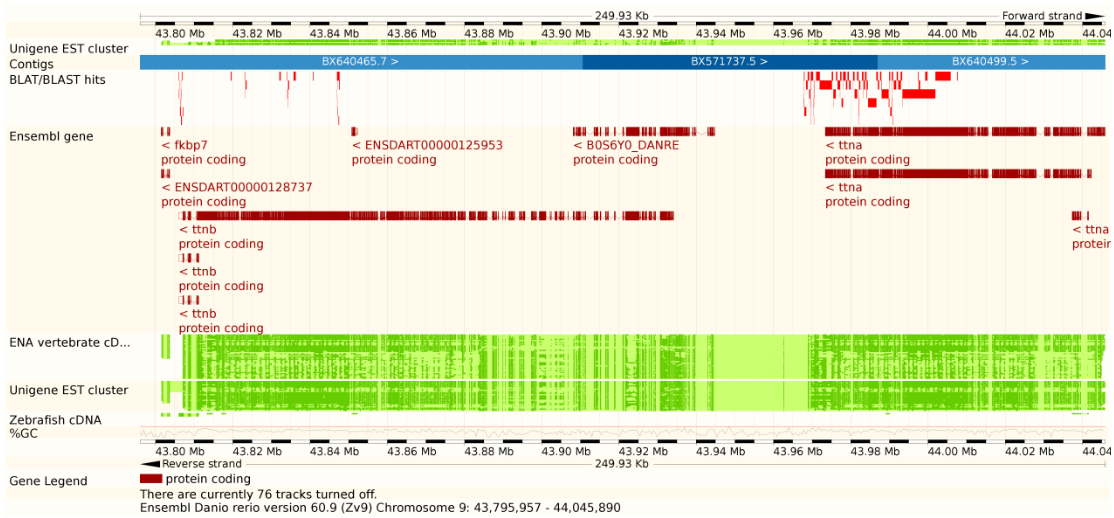


Figure 9 Scaffold24544 mapping to zebrafish chromosome 9 containing the Titin loci

Quality Assessment

In order to assess the quality of the assemblies produced, in terms of their usefulness and representation of known carp genes and BAC sequences, we downloaded all available carp nucleotide sequences from GenBank. This comprises over 2,000 sequences, mostly of known partial or full carp genes, as well the whole mitochondrial genome and two BAC clones. We then proceeded to map stringently (using Blat) all these sequences to each assembly generated.

Coverage of existing BAC clones

The Genbank deposited sequences include two longer BAC sequences, which can be useful to obtain an initial (although biased) measure of quality. The statistics obtained are shown in the table below. While BAC Clone BX571725 shows little variability across assemblies, BAC clone BX571686 indicates clearly that the highest coverage of the clone is achieved with the combined approach CLC + SOAP, which reaches 81% coverage of the clone, a remarkable result considering that 98% identity cutoff was used for this comparison. The coverage increases to 87% when using a 95% percentage identity.

Assembly	BAC BX571686.2 coverage	BAC BX571725.1 coverage
k35_r_m3_l70_g50_BAC_fake6	76.11653646	74.31533749
k35_r_m3_l300_g200	77.36002604	74.59810441
k35_r_m3_l70_g50	75.13020833	73.37278107
k33_r_m3_l300_g200	74.93489583	72.03225638
k39_r_m3_l300_g200	76.12955729	73.82049537
CLC	78.88346354	72.69466408
CLC + SOAP K35	**81.5625**	74.26297324

Table 2 Coverage of two carp BAC clone sequences obtained from different genome assemblies attempted

Coverage of all carp Genbank sequences

While the BAC clones provide interesting evidence of good coverage of existing carp sequence data for longer sequences, this is very anecdotal in nature, because only two clones are available. We thus proceeded to assess the coverage of the entire Genbank dataset containing over 2,000 sequences. As shown below, the average coverage of the query carp DNA sequences increases with improved assembly strategy. Both assemblies based on the CLCBio contigs produce the best coverage possible, with very small differences between the combined CLC Bio + SOAP approach and the CLC Bio approach alone. This is expected since the additional SOAP step is providing mostly “bridging” information across existing sequence, rather than novel sequence information. Among the SOAPdenovo assemblies the one made using K=39 has better coverage than others. Although this assembly has poor scaffold N50 statistics, it has the best Contig N50 statistics, indicating possibly that these are a better measure of final coverage of existing data (at significant expense of scaffold contiguity). In other words owing to the higher Contig N50 it is likely to better represent actual contiguous sequence which affects the final coverage mapping.

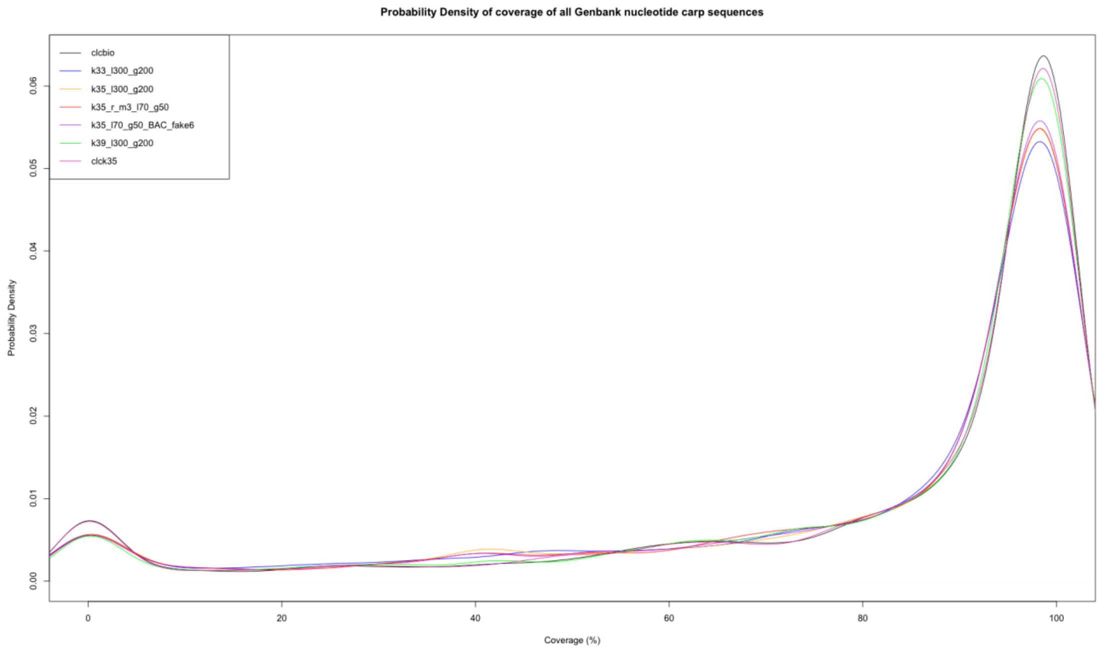


Figure 100 Probability density of coverage of carp Genbank nucleotide sequences for each assembly analyzed. The two assemblies based on the initial CLC Bio Contig assembly (clcbio and clck35 which indicates the CLCBio + SOAP assembly) have the highest fraction at high coverage.

In order to assess fragmentation (i.e. to what extent known carp sequences are fragmented on multiple scaffolds), we verified the number of hits found in each assembly for the entire set of carp nucleotide sequences found in Genbank. The graph in Figure 10 indicates clearly that the CLC Bio based assemblies have lower fragmentation (i.e. the DNA sequences used present fewer hits, despite the increased coverage), and the combination of SOAP and CLC Bio (clck35) has the lowest fragmentation of hits, as expected given its higher N50. Importantly the majority of known carp sequences have a single hit in the assembly, indicating that the majority of the genes (or gene fragments) searched can be found in a single scaffold sequence.

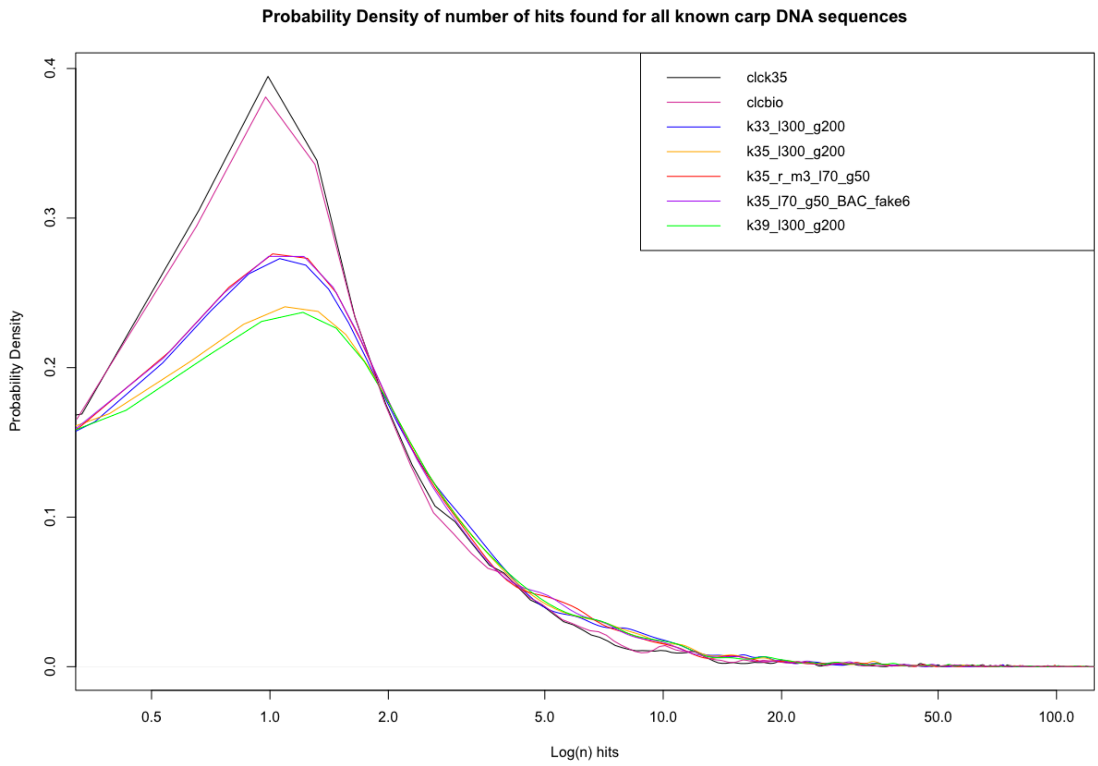


Figure 11 Probability density of number of hits obtained for each carp Genbank query sequence. The combined CLC Bio + SOAPdenovo assembly shows the lowest number of hits per query, with the majority of query sequences having a single hit.

As shown in Table 2 below only 22 out of 2,100 known carp DNA sequences analyzed could not be located in the assembly, and the majority of them are either haplotype specific, or within regions known to present high variability and difficult assembly such as the olfactory receptors and the MHC complex, which have required years of specific dedicated work in much larger projects such as the human genome project. Only 7 carp gene fragments cannot be located, which, based on existing data in Genbank, would indicate we are missing less than 0.5% of the gene content.

Sequence	Type	Type
FJ198033.1	Microsatellite	Repeat
FJ490421.1	3-beta hydroxysteroid dehydrogenase partial mRNA, 117bp	Gene fragment
FJ490420.1	cytochrome P450 21-hydroxylase partial mRNA, 144bp	Gene fragment
FJ655360.1	haplotype HcII3 cytochrome oxidase subunit II (COII) mitochondrial gene	Haplotype-specific sequence
FJ655287.1	haplotype Hc2 cytochrome b mitochondrial gene	Haplotype-specific sequence
FJ655286.1	haplotype Hc1 cytochrome b gene, partial cds; mitochondrial.	Haplotype-specific sequence
FJ655355.1	haplotype Hd51 tRNA-Pro gene and control region, partial sequence; mitochondrial	Haplotype-specific sequence
EU203669.1	MHC class II antigen beta chain (Cyca-DAB1) gene, Cyca-DAB1*05 allele, exon 2 and partial cds	MHC Complex
X95436.1	mRNA for MHC class II beta chain, D(clc)B	MHC Complex
Z47730.1	Cyca-DXA2*01 gene for MHC class II alpha chain	MHC Complex
EF042096.1	enolase mRNA, partial cds	Gene fragment
EF042095.1	enolase mRNA, partial cds	Gene fragment
BD262014.1	Method of identifying organism by comparative gene analysis and primer and hybridization probe for effecting the method.	Patent sequence
AY505343.1	uncoupling protein 3 (UCP3) mRNA, partial cds	Gene fragment
AB194212.2	OFRE mRNA for olfactory receptor, partial cds, clone: CCOR35	Olfactory Receptor cluster
AB194205.2	OFRE mRNA for olfactory receptor, partial cds, clone: CCOR25	Olfactory Receptor cluster
AB194203.2	OFRE mRNA for olfactory receptor, partial cds, clone: CCOR23	Olfactory Receptor cluster
AB194202.2	OFRE mRNA for olfactory receptor, partial cds, clone: CCOR21	Olfactory Receptor cluster
AJ628728.1	partial mRNA for sialic-acid binding protein-4	Gene fragment
AX084639.1	Sequence 171 from Patent WO0055361	Patent sequence
AF008558.1	brain-derived neurotrophic factor (BDNF) gene, partial cds	Gene fragment

Table 2 List of carp Genbank sequences which could not be mapped to the final genome assembly, indicating Genbank ID, description and type of sequence. Most sequences are either haplotype specific or belong to variable regions of the genome.

Gap Filling

In order to further improve the final assembly generated using the combined CLC Bio + SOAP strategy the GapCloser algorithm (part of the SOAPdenovo package) was used, which aims at filling gaps found within scaffolds. The software was able to remove ~25% of the Ns found in the original assembly, as detailed below:

- Number of Ns before Gap Filling: 421,965,634
- Number of Ns after Gap Filling: 309,918,033

Since gap filling can sometimes lead to lower quality sequences within the gaps, QC was performed again on the new gap-filled assembly by mapping all carp Genbank records to it. The results, shown below, indicate that the gap filled assembly is of higher quality than the assembly produced without gap filling.

- Before Gap Filling:
  - Average Coverage: 91.26%
  - STDEV Coverage: 17.93%
  - Median coverage: 98.06%
- After Gap Filling:
  - Average Coverage: 92.33%
  - STDEV Coverage: 17.38%
  - Median coverage: 98.31%

Mitochondrial genome

A very good example of a high quality scaffold found in the assembly is the mitochondrial genome, which would be found in Scaffold C2172197 and which contains the entire carp mitochondrial genome (as shown from BAC Clone AP009047.1). The alignment indicates complete coverage and only one sequence segment inverted with respect to the BAC clone deposited, as shown in Figure 12.

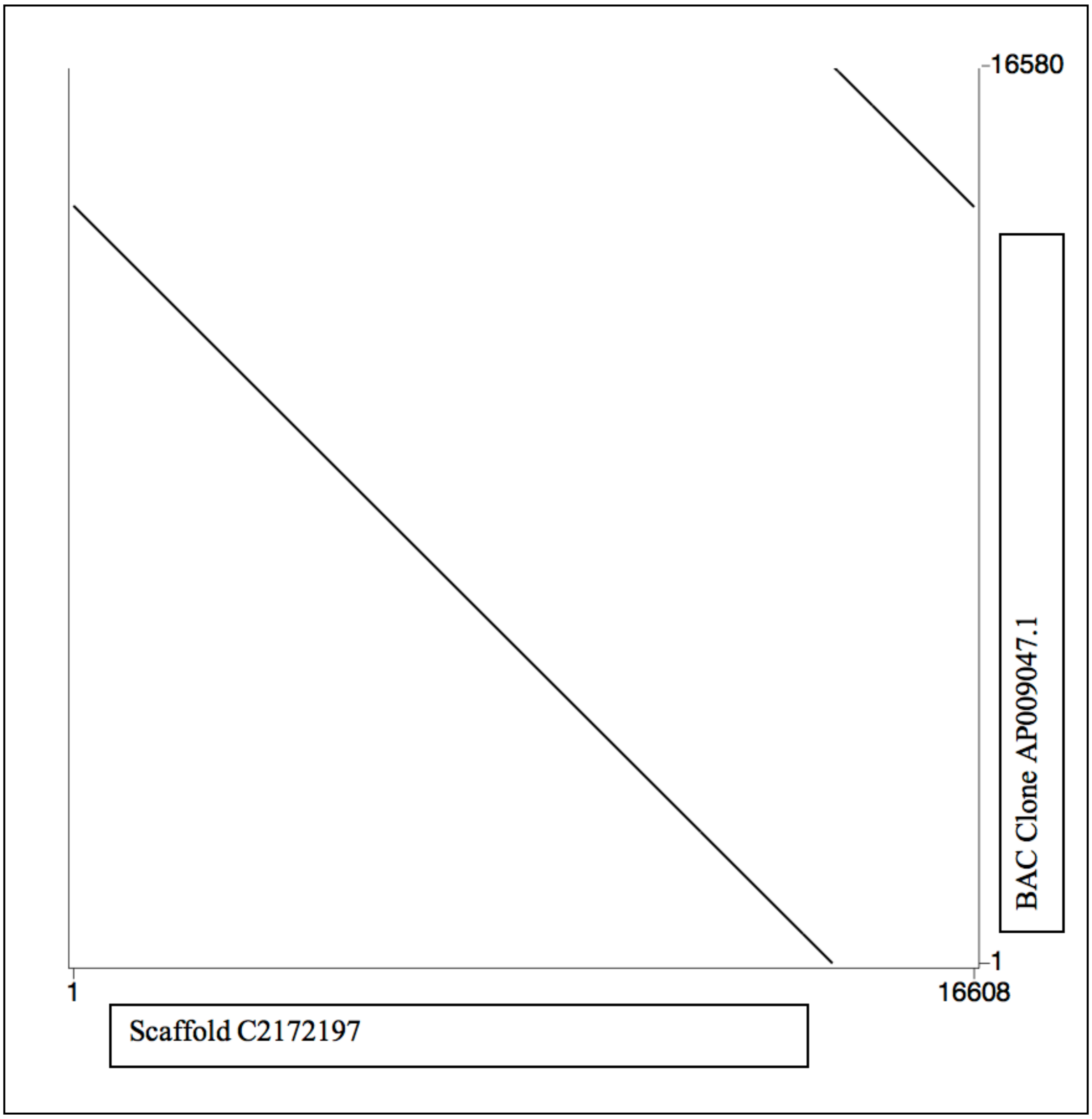


Figure 12 Figure showing the base by base alignment of Scaffold C2172197 and Clone of the mitochondrial genome (BAC Clone AP009047)

RNA-Seq Analysis

The RNA-Seq data, assembled into contigs using CLC Bio, was also mapped to the final genome assembly. In line with the data obtained using Carp Genbank entries, the overall coverage was very high for most contigs (see Figure 13, median coverage = 98.7%, average coverage = 92.47%), although usually spread over a few different scaffolds (see Figure 14, median number of hits = 3, mean number of hits = 9.71)

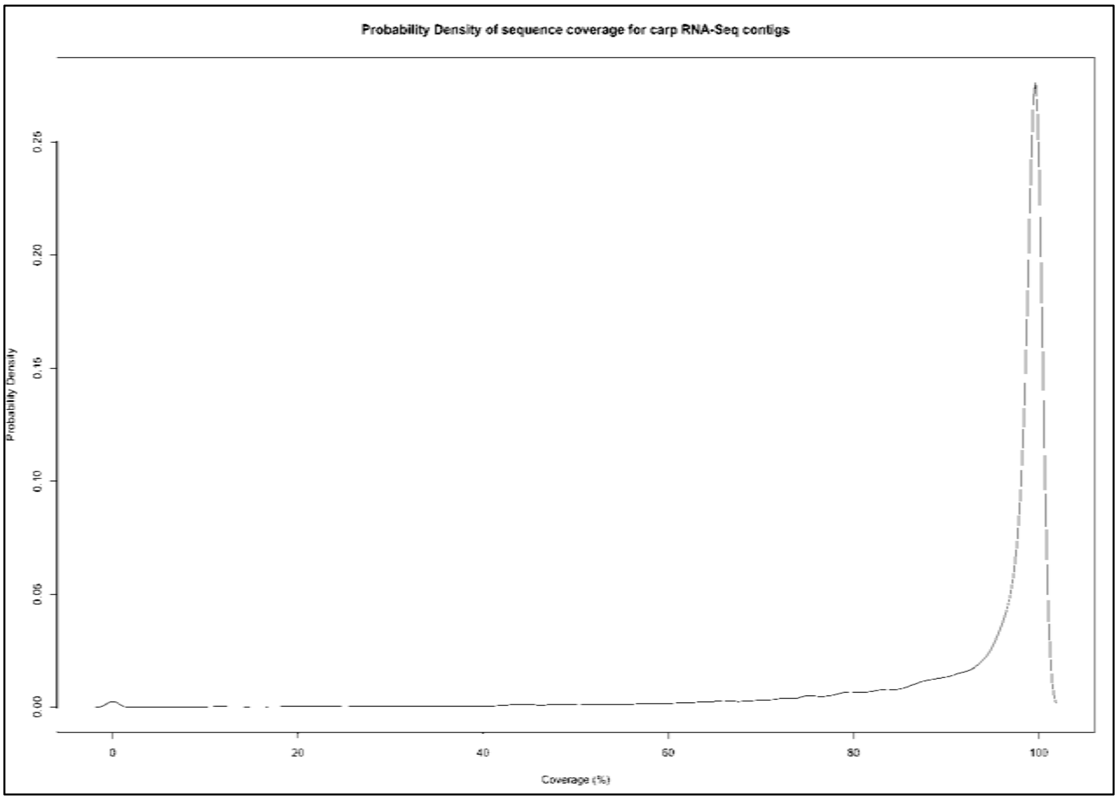


Figure 13 Proability Density of Sequence Coverage (%) for Carp RNA-Seq Contigs.

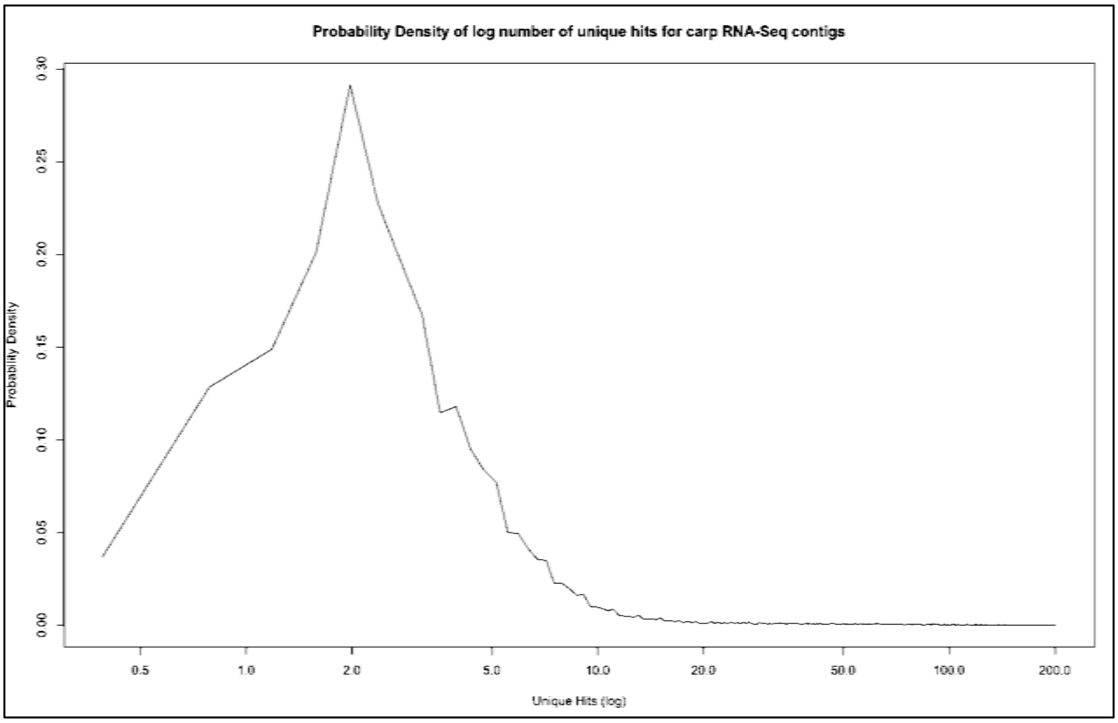


Figure 14 Probability Density of number of unique hits (log scale) for Carp RNA-Seq Contigs

As an example, one of the longest RNA-Seq contigs, contig_1067 (6,347bp) mapped fully to scaffold50890. This scaffold and RNA-Seq contig have their best mapping to chromosome 19 in the zebrafish genome (as well as a few other lower identity copies in other chromosomes, such as chr16, chr14 and chr25). All of the mappings are unspliced and are not annotated. A further analysis, using BLASTX on the NR database indicates that this is likely to be a processed pseudogene, as it contains an RT_LTR region, i.e. a reverse transcriptase domain.

Methods

Genome Assembly

In this project, initially the ABYSS de novo assembler was used due to lower memory requirements as compared to the very first algorithm published, Velvet (40-50Gbs of memory needed by ABYSS as compared to several hundred gigabytes required by Velvet). In a second phase most of the assemblies were performed using SOAPdenovo which not only uses similar amounts of memory but supports multi-threading, thus allowing rapid assembly of large datasets on a multi-CPU machine. The assemblies were performed on a 32 CPU Opteron server with 512GB RAM, allowing several assemblies to be run in parallel, exploring parameter space (such as modifying the K parameter to identify the optimal K and modifying other SOAPdenovo parameters such as L, G and read trimming options described). ABYSS was run in paired-end mode (i.e. using the command abyss-pe) and the following parameters: k=n (with n varying from 20 to 30) l=51 c=2 n=10 name=fish lib='lib200 lib5kb' lib200=pe200.fa lib5kb=pe5kb.fa se=se.fa. The first runs were performed as above. Subsequently ABYSS, from version 1.0.15, supported scaffolding, so the following option was added, to scaffold the contigs generated, including scaffolds which are likely to contain repeats: OVERLAP= --scaffold --mask_repeats.

SOAPdenovo was run with the following parameters: SOAPdenovo all -K n (ranging from 33 to 39) -p 24 (number of CPUs to use) -R -M3 -s soap.config (config file for run) -o name_of_assembly. The SOAPdenovo config file varied depending on the approach taken, the following is an example in which we used the 200bp reads, 5kb reads and BAC end reads, for both building contigs and scaffolds in successive steps:

```
#maximal read length
max_rd_len=76
[LIB]
#average insert size
avg_ins=200
#if sequence needs to be reversed
reverse_seq=0
#in which part(s) the reads are used, 1=contig, 2=scaffold,3=both
asm_flags=3
#in which order the reads are used while scaffolding
rank=1
#fastq file for read 1
q1=/data_n1/stupka/carp/CARPDATA/newdata/soap_elia/all_1.txt
#fastq file for read 2 always follows fastq file for read 1
q2=/data_n1/stupka/carp/CARPDATA/newdata/soap_elia/all_2.txt
[LIB]
avg_ins=5000
asm_flags=2
rank=2
reverse_seq=1
q1=/data_n1/stupka/carp/CARPDATA/5kb_reads/mp5k_1_sequence.txt
q2=/data_n1/stupka/carp/CARPDATA/5kb_reads/mp5k_2_sequence.txt
```

Gap Filling was performed using the GapCloser algorithm part of the SOAPdenovo package and was run as follows: GapCloser -o name (name of filled assembly) -b soap.config (config file similar to the above indicating where reads are located) -a name.scafSeq (file to be used for Gap Filling) -t 24 (number of threads to use).

QC Analysis

The QC analysis was performed using the server/client version of BLAT version 34, which is called gfServer and gfClient respectively. The binaries were downloaded from the UCSC website. Each assembly produced was transformed into 2Bit format (a compressed sequence format used by BLAT), and then a server was loaded with each assembly as follows:

```
gfServer start localhost port name_of_assembly.2bit
```

Each assembly was loaded on a different local port as required by the server.

The gfClient program was then used to search all Genbank nucleotide records specifying the assembly to be searched and the minimum identity cutoff. Furthermore we chose the “pseudoblast” output format, which is easily parsable using BioPerl. Thus the final command-line was as follows:

```
gfClient localhost port assembly_location query.fasta
output_name -minIdentity=95 -out=blast
```

The output was then parsed with a Perl script which uses the BioPerl BLAST parser Bio::SearchIO to parse the results. All hits with e-value < 1e-05 and score < 100 were removed. The number of hits was recorded. Then, since many small, overlapping hits are found, we used a temporary MySQL database to load all hits and quickly find the minimal subset of features which encompass all hits with no repetitions (this is in a module name Bio::MCE::Range which was kindly provided by Remo Sanges). This set of features was used to calculate final coverage of the query sequence.

Graphical Reporting

All graphical reporting was produced using R statistical functions. The following functions were among those most commonly used:

- read.table was used to import TAB delimited from files generated from Perl scripts or from the command-line into R datasets

- The density function was used to obtain probability densities where needed

- The plot and lines functions were used to produce final plots

Discussion

Initial pseudo-tetraploid ABYSS based assembly

The initial assembly of the pseudo-tetraploid carp genome was performed at a time when genome assembly had just started becoming a possibility for smaller labs as compared to large genome centres, but library making strategies, sequencing protocols and assembly algorithms were heavily under development. The first assembly obtained from the pseudo-tetraploid genome, although clearly not a good basis for future work was very useful to learn some of the key aspects that enable a successful genome assembly, summarized below:

- Obtaining very good coverage (e.g. 30x or more) of standard (e.g.200bp) genomic DNA libraries with long (e.g. 100bp) paired-end sequencing is fundamental to obtain a reliable initial contig dataset where each fragment is ideally sequenced completely with overlapping paired-end reads

- Obtaining mate-pair libraries sequenced at high depth with short reads has a significant effect on the ability to scaffold. This is in order to avoid "read-through", i.e. sequencing through to the other side of the mate-pair.
- Preparing mate-pair libraries of different insert sizes (e.g. 600bp, 5Kb, 10Kb) provides a range of "scaffolding opportunities" (e.g. 600bp for bridging over most repeat elements, 5Kb and 10Kb to provide long-range bridging) and also has significant effect on final scaffolding ability
- Availability of a BAC library for BAC end sequencing is of great benefit for obtaining very long range scaffolding, i.e. in the range of 100s of Kbs
- Recent chemistry upgrades made by Illumina (in particular v5 or TruSeq chemistry) have a significant impact on assembly of large genomes, due to overall higher quality base calling as well as longer overall read length

Evaluation of ABYSS

While clearly ABYSS provided the first "affordable" approach to de novo assembly, by utilizing on average 40Gb of memory for each carp assembly, it still had several caveats, which impacted the final results:

- Since ABYSS does not yet support multi-threading it took several days for each assembly to complete, limiting the range of options we could test
- Despite trying in several ways, the scaffolding options did not yield any results, i.e. no improvement to the scaffold N50, no clear scaffolding.

Haploid DNA CLC Bio and SOAP de novo based assembly

The next dataset of 200bp sequence reads obtained from a haploid genome, combined with an improved assembly strategy, allowed us to produce a radically improved assembly, on which most of the optimization and analysis work was performed.

CLC Bio Contig Assembly

Although CLC Bio was not initially our choice of software due to its commercial nature and costs involved, evaluations of its de novo assembly algorithm lead us to decide to make use of it, since it produced far better contigs than any other approach tested. The assemblies produced "out of the box", i.e. without tweaking the data or parameters, were already superior to those we had produced. Once the data was pre-processed and merged, it lead to even better contig assembly, which was subsequently used for scaffolding with SOAPdenovo. The limitations of using the CLC Bio software are:

- It is commercial software and, unlike other bioinformatics commercial software there is no academic free availability in any form, and a specific license (not the one for the CLC Genomics Workbench) needs to be purchased for the command-line tools
- It is a "black box" approach, i.e. we do not have any details of how the assembly process works, and we cannot understand why it is so much superior to other approaches
- It does not support scaffolding, which was the main reason why obtained the best final assembly by combining this tool with SOAPdenovo

Based on all the QC performed we are quite confident that the superior N50 obtained does not come at a cost in terms of sequence quality, since the assemblies based on CLC Bio contigs consistently show the best coverage of existing sequences.

The K parameter

Using SOAPdenovo we evaluated extensively the best K parameter to use. The analysis of the contiguity of the assemblies highlighted an interesting aspect: the K parameter which produces the best Contig N50 (K=39) is not the same which produces the best Scaffold N50 (K=35). In fact the Scaffold N50 drops radically when using K=39, which is why for the final assembly (which is based on pre-made CLCBio contigs) we decided to use K=35. We discussed this aspect with the authors of SOAPdenovo, and they recommended not varying the K parameter within the different steps of the SOAPdenovo assembly (e.g. using K=39 for contig assembly and K=35 for scaffold assembly), because this would lead to overall decreased sequence quality and inconsistencies in the assembly. From a point of view of QC, the assembly produced with K=39 actually presents slightly higher coverage of carp DNA sequences, indicating that in terms of sequence quality contig N50 is potentially a better measure, at a very significant cost in terms of contiguity. Taking into consideration that we perform a final Gap Filling step in the scaffolds produced, we prefer to have higher scaffold contiguity (i.e. many gaps to fill) if the sequence quality is not overly compromised. The issues above highlight the limits of K-mer based approaches, which strongly limit "comprehensive" assembly of sequences.

Other SOAPdenovo parameters

Although we explored quite extensively other SOAPdenovo parameters, any modifications to defaults seemed to impact negatively the final N50. The minimum length of the contigs used for scaffolding did not improve the final N50, although there were reports recently in the Phallusia genome project

(Patrick Lemaire, personal communication) that increasing L lead to a better assembly. This could be due to the specific nature of our assembly.

Often trimming reads leads to improved results in the context of a variety of next-generation sequencing projects, due to the usually poorer quality of the last bases of each sequence. Our trimming attempts, however, did not improve the N50. This indicates that although base qualities decrease towards the end of the read, overall those positions have a higher ratio of information vs. noise for the overall assembly process. The ideal approach is to perform a variable length quality-based trimming (i.e. remove bases below a certain quality) but since many tools/aligners do not deal with variable length FASTQ sequences well, we did not try that approach. Another approach is simply to replace low quality bases with Ns, and this was done in the pre-processing of the reads for the CLC Bio based contig assembly. In fact read trimming can produce a radical artifact, as was shown in Figure 5, i.e. that if the K parameter is equal or larger than the length of the trimmed reads, the assembly is reduced drastically, because SOAPdenovo is unable to compare K-mers appropriately in the assembly process.

BAC end reads

BAC end Sanger reads are clearly extremely beneficial for obtaining very long-range scaffolding. We obtained a very limited dataset of 2,200 BAC end reads. This set, unfortunately, was not large enough to see any improvement in the assembly. This is probably also because there is no option to assign a "weight" in SOAPdenovo to a piece of evidence. Thus even though the BAC end reads are from Sanger sequencing and are thus likely to provide stronger evidence than a single read from a FASTQ file, there is no way to encode this in the process. As a test it was tried to simply repeat 6 times the same BAC end reads, thus “faking” a stronger weight for the BAC ends and indeed, some longer scaffolds were produced, because of additional linking information. This assembly was not used further, however, because it could contain potentially some erroneous assemblies due to the forced duplication that was introduced. The laboratory which has produced this first set of 2,200 reads has communicated to us that a further, much larger set of approximately 80,000 reads will be available soon and thus we will able to attempt another assembly with this new dataset when it is available.

Assembly Assessment and QC

Given the initial data obtained the final assembly obtained is of very good quality, as assessed through a variety of approaches:

- Mapping the largest scaffolds produced to the zebrafish genome identifies collinear mappings of similar size, indicating that no major artifacts were produced in the process of assembling larger scaffolds
- The coverage of the only two available BAC clones is reasonably high (81% and 74%) indicating good coverage of existing genomic sequences
- The coverage of known Carp nucleotide sequences is remarkably good. The median coverage is 98%, and the median number of hits is just 1, indicating that the vast majority of known carp nucleotide sequences is very well covered and is usually found in a single scaffold
- Very few sequences are not mapped at all in the current assembly, and the majority of those (15 out of 22) are unlikely to be mapped in our genome

assembly because they belong to highly variable regions, recombinant regions and haplotype-specific regions.

- The mitochondrial genome was found completely in a single Scaffold, with only one potential segment in the wrong assembly location
- The RNA-Seq Contigs have very good coverage in the current assembly (median = 98%), confirming that most genes should be found in this assembly.

On the other hand there are clearly some caveats:

- The majority of the sequences deposited in Genbank are fairly short and it is thus unsurprising that they are mapped in one or very few scaffolds (is this referring to DR or CC)
- The BAC coverage is not as good as the coverage of genomic sequences. While two BAC clones are probably too few to make any conclusive statements the assembly probably lacks good coverage of repeat-rich regions.
- While the contiguity is good taking into account the starting data, it is still not sufficient to recover many multi-gene loci, thus strongly limiting any studies aimed at understanding synteny among fishes and/or vertebrates, as well as the search for promoter regions, enhancer regions, etc. which all require long contiguous assemblies spanning several gene loci
- As highlighted by the mapping of the RNA-Seq contigs, on average genes are still fragmented over several scaffolds, thus it is paramount to improve the contiguity of the scaffold assembly with further sequencing of longer libraries and hopefully more BAC end data.

References

1. Hulata G. A review of genetic improvement of the common carp (*Cyprinus carpio L.*) and other cyprinids by crossbreeding, hybridization and selection. Aquaculture. 1995;129:143-155. doi: 10.1016/0044-8486(94)00244-I.
2. Yanju Zhang, Functional annotation of microRNAs and de novo genome sequences through heterogeneous data analysis, PhD Thesis, Leiden University, The Netherlands (2011) In Press.
3. Yan Li, Peng Xu, Zixia Zhao, Jian Wang, Yan Zhang and Xiao-Wen Sun Construction and Characterization of the BAC Library for Common Carp Cyprinus Carpio L. And Establishment of Microsynteny with Zebrafish Danio Rerio *Marine Biotechnology* DOI: 10.1007/s10126-010-9332-9
4. Jared T. Simpson, Kim Wong, Shaun D. Jackman, Jacqueline E. Schein, Steven J.M. Jones and İnanç Birol, ABySS: A parallel assembler for short read sequence data, *Genome Research* (2009) 19: 1117-1123
5. http://www.clcbio.com/
6. W. James Kent BLAT—The BLAST-Like Alignment Tool, *Genome Research* 2002. 12: 656-664
7. P.Flicek et al, Ensembl 2011, NAR (2011) 39 (suppl 1): D800-D806.