Chapter 5

Identification of Common Carp Innate Immune Genes with Whole-Genome Sequencing and RNA-Seq Data

Based on

Yanju Zhang, Elia Stupka, Christiaan V. Henkel, Hans J. Jansen, Herman P. Spaink and Fons J. Verbeek. (2011). Identification of Common Carp Innate Immune Genes with Whole-Genome Sequencing and RNA-Seq Data. Journal of Integrative Bioinformatics, 8(2):169.

Summary

The common carp is a candidate model system for immunology research. Using next-generation sequencing technology, we have generated a huge amount of sequence reads from the carp genome and transcriptome. Currently, our aim is to identify carp genes, particularly genes involved in the development of the innate immune response, given the circumstance that the carp genome assembly is not yet completed. To achieve this, we developed a comprehensive genome annotation pipeline. This analysis allowed us to estimate that the carp has approximately 39 TIR domain-containing transcript isoforms and genes.

1 Introduction

Common carp (*Cyprinus carpio*) is one of the most important freshwater cultured fish species that has been widely used in fish biology research [3]. A single female is capable of producing up to a few hundred thousand eggs that can be efficiently fertilized *in vitro*. Since the innate immune response is already active in developing embryos, common carp can be a relevant model for studying its mechanisms. The innate immune response is the first line of defence against infectious disease and cancer by identifying and killing pathogens and detrimental cells. This innate immune response relies heavily on signaling by pattern recognition receptors. The best-studied pattern recognition receptors of the vertebrate innate immune system are the Toll-like receptors (TLRs). All the TLRs, some Interleukin receptors (IL-Rs) and downstream adaptor proteins contain a Toll/Interleukin-1 receptors (TIR) domain which is a highly conserved functional unit mediating the protein-protein interactions between the receptors and the adaptors thus relaying the signal.

TIR domain-containing genes therefore play important roles in immunity signalling pathways. In zebrafish (*Danio rerio*), this family has been studied using microarray technology [23]. However, microarrays have a number of shortcomings, i.e. low sensitivity and specificity, low consistency across platforms, and, above all, they rely on a fixed definition of the transcriptome for their design.

Next-generation sequencing (NGS) is a recently developed, high-throughput sequencing technology, with which one can produce millions of sequence reads in a few days at a low cost and without the need for a priori knowledge of the sequences [19]. Applying such technology to the entire genome of a particular organism is referred to as whole-genome sequencing. Another application, RNA-Seq, is to sequence cDNA for transcriptome profiling. In comparison to microarrays, RNA-Seq has a much higher dynamic range, base-level resolution, richer splicing information and the ability to detect previously unknown transcripts.

Our ultimate goal is to study how the transcriptome, especially the expression of the innate immune response genes, changes upon pathogen infection using NGS. Since common carp is not a model system and no reference genome assembly is available, both the whole carp genome and several transcriptomes are sequenced. Currently, as a pilot study, we focus on discovering the innate immune response genes, especially TIR domain-containing genes.

In this Chapter, we present a gene identification strategy that integrates whole genome sequencing data, RNA-Seq data and relevant data obtained from public databases in order to identify TIR domain-containing genes and transcripts in carp. With limited data available, different data sources and methods are compared and integrated in order to maximize the likelihood of detecting the target sequences.

2 Background

2.1 Common carp

The common carp is a serious candidate model system for very high throughput screens of pharmaceutical compound libraries. The closely related cyprinid zebrafish has been developed into a medium throughput capacity model for drugs related to cancer and immune-related diseases. The advantage of carp is that a single female is capable of producing up to a few hundred thousand eggs that can be efficiently fertilized in vitro. Therefore, the genomic homogeneity of carp eggs is easier to control than for zebrafish that provide smaller clutches of 150 to 200 eggs. Thus, large clutches of embryos derived from a small group of common carps enable hundreds of thousands of pharmaceutical drug candidates to be tested with less genetic diversity in the screening model.

2.2 TLR pathway and TIR domain containing genes

The innate immune system is the first line of defence that protects the host against infectious disease and cancer [23]. This system relies heavily on pattern recognition receptors mediating immune responses to pathogenic microorganisms. The Toll-like receptors (TLRs) and interleukin-1 receptors (IL-1Rs) are probably the most essential pattern recognition receptors of the vertebrate immune system.

Stimulation of TLRs by their ligands leads to the recruitment of adaptor proteins to the receptors. Differential utilization of the adaptor molecules by the TLRs causes specific activation of a range of transcription factors such as NF-kB, activator protein 1 (AP-1), and IFN regulatory factors (IRF) 3, 5, and 7 through distinct signalling pathways, eventually leading to the downstream activation of proinflammatory cytokines [13]. The details of the TLR signalling pathway are depicted in Fig. 1.

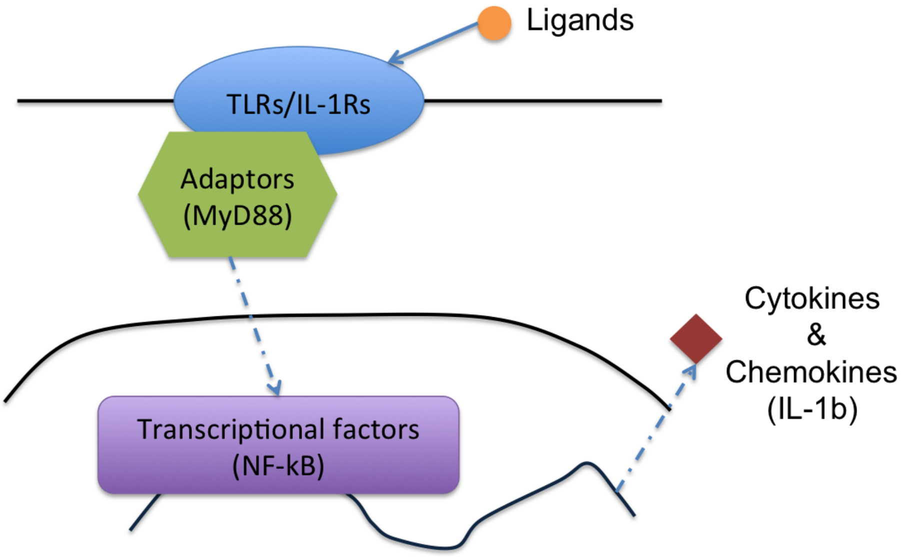

Ligands
TLRs/IL-1Rs
Adaptors (MyD88)
Transcriptional factors (NF-kB)
Cytokines & Chemokines (IL-1b)

Figure 1: TLR signalling pathway.

Upon binding to Interleukin-1 ligand, IL-1R interacts with IL-1R accessory proteins. With the aid of adaptor proteins, this signal is transduced and leads to the activation of NF- kB. The whole process is very similar to the TLR pathway.

The Toll/Interleukin-1 receptor (TIR) domain is the conserved intracellular signalling domain shared by TLR and IL-1R families. It is also found in some of the adaptors, e.g. MyD88, which connect the Toll-like or Interleukin receptors with downstream transcriptional factors. The TIR-domain containing proteins in the human genome comprise ten members of the TLR family, eight members of the IL-1/IL-18 receptor group (IL-1R/IL-18R) and five adaptor proteins. Zebrafish has one or more counterparts for the human TLR1, TLR2, TLR3, TLR4, TLR5, TLR7, TLR8, TLR9, IL-1R and IL-18R genes and one copy of the adaptor genes MyD88, MAL, TRIF and SARM [13].

2.3 Next generation sequencing

With the advances of second generation sequencing technologies, including Solexa/Illumina Genome Analyzer, ABI SOLiD, Roche/454 and Helicos, massive genome sequencing projects ranging from prokaryote to eukaryote have been carried out. These technologies are now widely applied in advanced research such as genome sequencing, transcriptome sequencing, exome sequencing, microRNA expression profiling and DNA methylation studies [10]. Unlike traditional the Sanger sequencing technology [18], these new

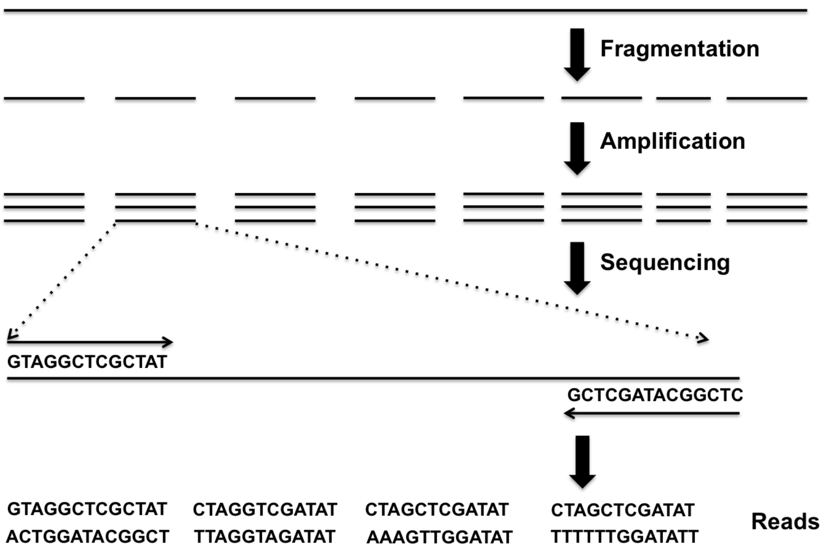


Figure 2: Using next-generation sequencing technology to sequence genome (the whole genome sequencing).

sequencing technologies produce shorter reads with higher genome coverage at very low cost in, and moreover, a short period of time [15]. In the mean time, some of the advantages, i.e. fast and high-throughput data generation, also lead to computational challenges in the genome assembly as well as the subsequent genome annotation analysis.

2.3.1 Whole genome sequencing and RNA Sequencing (RNA-Seq)

Completed genome sequences are immensely useful in biological researches. A systematic transcriptome analysis reveals the dynamic activities in the cell. Currently, the whole genome sequencing and RNA-Seq are the promising technologies which apply next generation sequencing to sequence the whole genome and transcriptome. Fig. 2 illustrates the principle of the whole genome sequencing technology. First DNA molecules are extracted and then sheared into short fragments. Later on adaptors are attached to one or both ends. With or without amplification, each fragment is then sequenced by the sequencer to obtain short sequences from one end or both ends resulting single-end or paired-end reads respectively. In principle, RNA-Seq technology is similar to whole genome sequencing. The main difference is at the sample preparation stage. Instead of extracting DNA, RNA-Seq extracts RNAs which are characterized with polyA tails and converts them to cDNAs. The output of next generation sequencing experiments are short reads which are typically 30 to 400 bp depending on the sequencing technology.

2.3.2 Assembly

Genome assembly is the process that constructs the original continuous DNA sequences from millions of short DNA reads. It can be accomplished from two directions: a *de novo* approach which constructs genomes from scratch and comparative approach that uses a closely related organism as a guide to produce contiguous genome sequences [15]. Currently, the genomes of several species, e.g microbes, yeast, worm, fruitfly and human have been completed sequenced and assembled.

A *de novo* genome assembly approach is to reconstruct genomes without using any previously sequenced organisms. This method is capable of capturing the sequence diversity for each individual, however, it is extremely hard, complicated and memory consuming. The main computational strategies applied in *de novo* assembly are overlap computation and Eulerian path. Overlap computation calculates all pair-wise alignments between a set of reads and constructs the contigs according to the read overlap. It includes the greedy algorithm based tools such as TIGR assembler [24] and overlap-layout-consensus (OLC) based assemblers such as Arachne [1]. Eulerian path utilizes graph theory and the de Bruijn graph [14] based data structures for assembling a genome. Most recently developed second-generation assemblers applied the Eulerian strategy such as SOAPdenovo [11], Velvet [27], ABySS [20] and Allpaths [4].

The principle of comparative assembly is different. First, the reads are aligned to a highly related genome called reference genome. This alignment is then used directly to compute the consensus sequence of the new genome. In practice, comparative assembly is easier since it utilizes alignment instead of expensive overlapping step. An example in this category is AMOScmp [16].

2.4 Gene annotation

Genome annotation is an important downstream analysis after an organisms genome has been completely sequenced and assembled. The main goal of genome annotation is to add the biological context to the sequence, to identify the key features of the genome in particular genes and their products. One of the analyses is gene finding that predicts the precise start and stop position of a gene and the splicing pattern of its exons [22]. Similar to genome assembly, this analysis can be approached using *ab initio* gene prediction and comparative prediction.

In general, the *ab initio* gene prediction algorithms, e.g. Augustus [21], utilize the existing gene structures as training set and then build up statistical models to predict the gene structures for the functionally unknown sequences. In contrast, comparative gene prediction induces a gene structure from close related species. A high similarity between a sequence in one species and a gene in another species is good evidence that this sequence contains a gene with similar structure. BLAST [12] and BLAT [9] are two fast sequence similarity search algorithms often applied in this area.

3 Methodology

The genome of a fully homozygous common carp, obtained in a single generation without inbreeding [3] and a heterozygous carp genome were sequenced using Illumina Genome Analyzer IIx sequencing technology. For the homozygous carp, we generated a paired-end sequencing library with insert sizes of about 200 base pairs (bp), from which we obtained approximately 40 Gbp of usable sequences with a read length of 76 bp. For the heterozygous carp genome, three DNA libraries were constructed: one single-end library with read length 51 bp; one paired-end library with 200 bp insert size and 51 bp read length; one mate-pair library with insert sizes of 5 Kbp and 36 bp read length. In total, we generated about 10 Gbp sequences for this strain.

We also sequenced the total mature messenger RNAs of common carp at different developing stages and conditions. Four mRNAs samples, wild-type carp and carp infected with the *Mycobacterium marinum* pathogen both at embryonic and adult stages, were extracted and then sequenced using the Illumina Genome Analyzer IIx. For each sample, an RNA-Seq sequencing library was constructed from which single 51 bp reads were sequenced. Details of all the data sets are listed in Table 1.

3.1 Genome assembly strategy

In the absence of a carp reference genome, the first task is to generate a carp genome assembly. Considering the fact that the homozygous carp genome sequencing data is approximately 4 times abundance than that of the heterozygous carp genome and is about 13x genome coverage, we chose to assemble the homozygous carp genome and considered to use 5 Kb mate-pair library from the heterozygous carp only for the purpose of scaffolding.

Dataset 1: gDNA, heterozygous carp

	Library size (bp)	Lanes	Read length (bp)	Size (GB)
Single-end	-	2	51	0.6
Paired-end	200	10	51	5.1
Mate-pair	5 K	7	36	3.7

Dataset 2: gDNA, homozygous carp

	Library size	Lanes	Read length (bp)	Size (GB)
Paired-end	200 bp	6	76	40

Dataset 3: RNA-Seq

		Lanes	Read length (bp)	Size (GB)
Single-end	Embryo, wt	1	51	2.6
Single-end	Embryo, infected	1	51	2.6
Single-end	Adult, wt	1	51	2.7
Single-end	Adult, infected	1	51	3.1

Table 1: The genomic reads and RNA reads generated from homozygous carp using Illumina sequencing.

The strategy of the carp genome assembly is illustrated in Fig. 3. First all the raw reads are filtered by quality control criteria and further pre-assembled in the pre-processing stage. Having the high quality reads, three *de novo* assemblers are applied and compared in order to achieve the best assembly.

The raw reads generated from the Illumina sequencer included base-calling errors and adapter contamination. It has been found that Illumina sequencing quality decreases specially at the end of the read [17]. Adapter contamination is mainly caused by insert sizes smaller than the read length. Therefore the reads were filtered at a threshold if either Read 1 or Read 2 contained an adapter sequence longer than 6 bp and the low quality reads were eliminated.

We then merged the remaining paired-end reads into a longer single-end read if they had 7 overlapping nucleotides. Pairs were not collapsed into a longer read if repetitive sequences within them tended to create ambiguous connections. This preassembly procedure not only produces long reads which will potentially improve the efficiency and quality of the assembly, but also provides confirmation for the quality of the 3' end of the reads. After the pre-processing, 3.5% of nucleotides are discarded and 69.9% of pairs are merged.

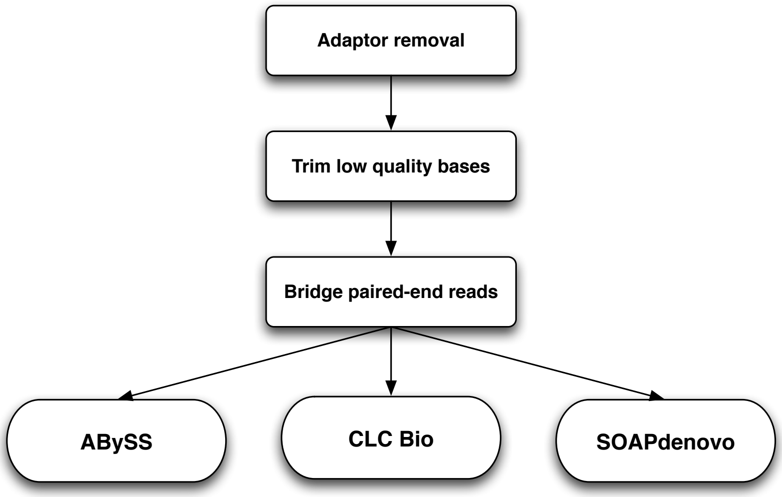


Figure 3: Genome assembly pipeline

Next, we assembled the high quality and merged genomic DNA reads using three *de novo* assemblers: ABySS, CLCBio and SOAPdenovo. All of them are de Bruijn graph-based assemblers. A de Bruijn graph is a directed graph representing overlaps between sequences of symbols [14]. In the graph, basically, short reads are broken into smaller sequences of DNA, called *k*-mers. Each node represents a *k*-mer nucleotides; an edge exists only if the adjacent nodes are overlapped by *k-1* nucleotides. Extracted contiguous sequences are represented by unambiguous paths through the nodes.

Both ABySS and SOAPdenovo are free software; whereas CLCBio is a commercial product. Each of the three assemblers has its own innovations. ABySS improves the memory efficiency by using a distributed de Bruijn graph and has been successfully applied in assembling of an African male human genome [20]. SOAPdenovo performs error correction before the de Bruijn graph is built and specifies different libraries for assembly and scaffolding. It was successfully applied for giant panda genome assembly [10]. CLCBio has strength in handling repeats using the information from paired-end reads. Unlike ABySS and SOAPdenovo, which require manually optimization of the parameter *k*, CLCBio tunes and optimizes it automatically [2].

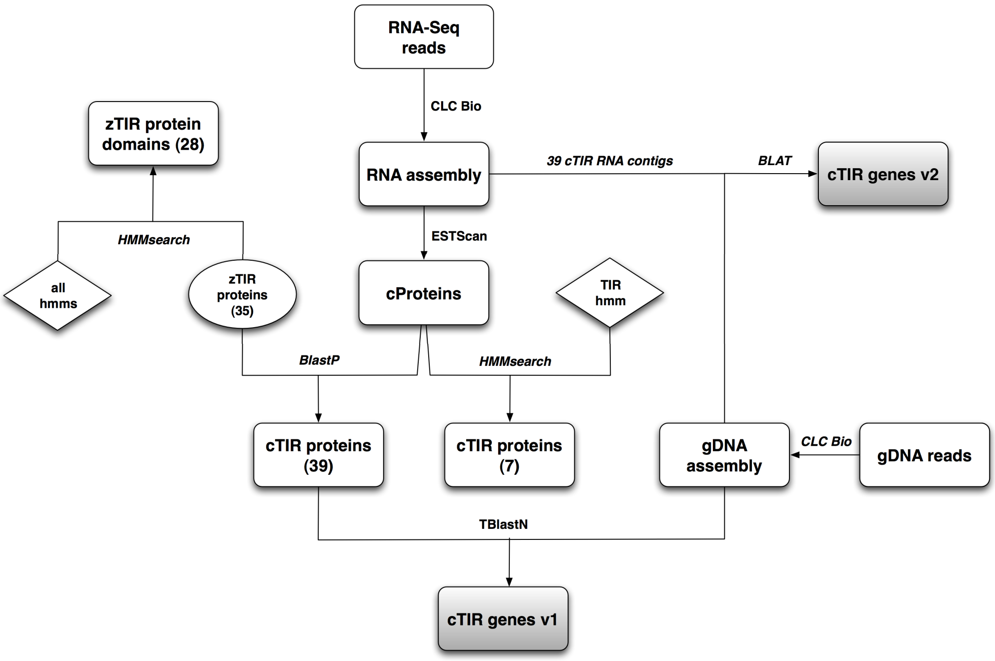


Figure 4: TIR containing gene annotation pipeline. The output TIR gene sequences (cTIR genes v1 and cTIR genes v2 depicted in shade) derived from different methods are compared and further integrated to construct the final TIR genes

3.2 Annotation strategy

Animal genome assemblies based on NGS data only are generally highly fragmented. We therefore developed an integrative strategy to maximize the probability of identification of carp genes; specifically TIR containing genes and transcripts.

Firstly all the RNA-Seq reads from all the samples were pooled and assembled using the CLCBio *de novo* assembler. The resulting RNA contigs were used as potential gene product fragments. These sequences were then translated to protein sequences using the ESTScan algorithm [8]. After that, the protein sequences obtained were searched for the TIR domain found in Interpro [7] using the HMMsearch algorithm [5].

The zebrafish is evolutionarily close to the common carp (both are cyprinids) and the zebrafish genome is relatively well covered and annotated in the Ensembl database [6]. Therefore we used this genome to facilitate the annotation of the carp TIR containing genes. In this manner, we performed a comparative genomic analysis using zebrafish resources: the zebrafish TIR containing proteins found in Ensembl were BLASTed against the carp peptides obtained from the RNA-Seq contigs resulting in the putative carp TIR

Assembly	k	n	n:N50	median	mean	N50	max	sum
Without preprocessing	25	1637271	250045	384	735	1409	17597	1.20E+09
Without preprocessing	27	1847118	-	-	680	1389	18684	1.26E+09
After preprocessing	**25**	**1086163**	**159656**	**587**	**1135**	**2260**	**26293**	**1.23E+09**

Table 2: Performance of CLCBio

proteins allowing us to identify potentially new carp TIR transcripts. Considering the fact that the obtained assembly is fragmented, the transcript and protein sequences are used to bridge fragmented DNA contigs. To achieve this, the candidate TIR transcript and protein sequences were further mapped to the carp genomic contigs by using TBlastN [12] and BLAT [9]. The DNA contig hits are connected using a number of 'N' as gap sequences and finally result in an alternative set of the carp TIR genes for comparison. The entire pipeline is shown in the Fig. 4.

4 Results

4.1 Carp genome assembly

We ran the CLCBio *de novo* assembler on both the raw genomic reads and the reads after preprocessing. Parameter *k* was optimized to 25, as we manually changed it to 27, the N50 value slightly dropped as showed in Table 2. N50 is defined as the length N for which 50% of all nucleotides in the contigs are in a length of at least N nucleotides long. It is a useful heuristic for measuring the quality of an assembly. A higher N50 represent a longer average contig length. As illustrated in the table, the N50 increased from 1409 to 2260 after the preprocessing step, a 60% improvement compared to the assembly derived from the raw data. This result shows that the preprocessing is a crucial step that makes a huge difference in the final assembly.

ABySS and SOAPdenovo were applied to the preprocessed data. Parameter k for *k*-mer is tuned from 17 to 55 in order to achieve the best performance. ABySS produced the longest N50 contig length (N50 = 716 bp) when the *k*-mer length is 50. As for SOAPdenovo, it restricts the value of *k* to an odd number. The best assembly with N50 of 729 bp was achieved when *k* is 45. The assemblies generated by three assemblers are summarized in Table 2, 3, 4.

Compared to ABySS and SOAPdenovo, the CLCBio assembler performs much better.

k	n	n:N50	median	mean	N50	max	sum
23	4.19E+07	1217339	150	181	187	5275	6.97E+08
25	3.23E+07	1124372	166	212	233	10026	8.48E+08
30	2.13E+07	866650	192	275	349	16613	1.04E+09
35	1.56E+07	681428	211	333	479	16601	1.14E+09
40	1.21E+07	566732	227	384	606	16601	1.21E+09
45	9812550	516612	240	421	694	16601	1.26E+09
50	**8151538**	**515595**	**249**	**433**	**716**	**16601**	**1.29E+09**
55	7015100	585470	237	403	653	16601	1.33E+09

Table 3: Performance of ABySS

k	n	n:N50	median	mean	N50	max	sum
17	1.40E+08	41018	110	116	112	338	1.07E+07
21	4.27E+07	1180069	149	179	184	4210	6.62E+08
25	2.81E+07	1007523	177	236	274	7225	9.17E+08
29	2.10E+07	830409	196	284	369	11122	1.05E+09
31	1.85E+07	760064	203	306	418	11124	1.10E+09
35	1.47E+07	647073	215	348	520	10342	1.18E+09
41	1.09E+07	537078	238	409	662	10342	1.24E+09
45	**9054965**	**500758**	**257**	**443**	**729**	**10342**	**1.27E+09**
51	7106660	524543	257	438	724	10141	1.31E+09
55	6201398	570679	280	438	676	8881	1.31E+09

Table 4: Performance of SOAPdenovo

Data	Assembler	n	n:N50	median	mean	N50	max	sum
RNAseq	ABy	1127477	67629	151	195	203	6720	4.66E+07
RNAseq	CLC	315316	76037	160	227	255	7939	7.16E+07
contigs+EST +mRNA	CLC	25157	6621	634	721	896	9919	1.81E+07

Table 5: Transcriptome assembly

The best assembly from CLCBio is the one with N50 of 2260 bp. This number is about two times higher than that from ABySS and SOAPdenovo. According to this, CLCBio is selected as the final assembler.

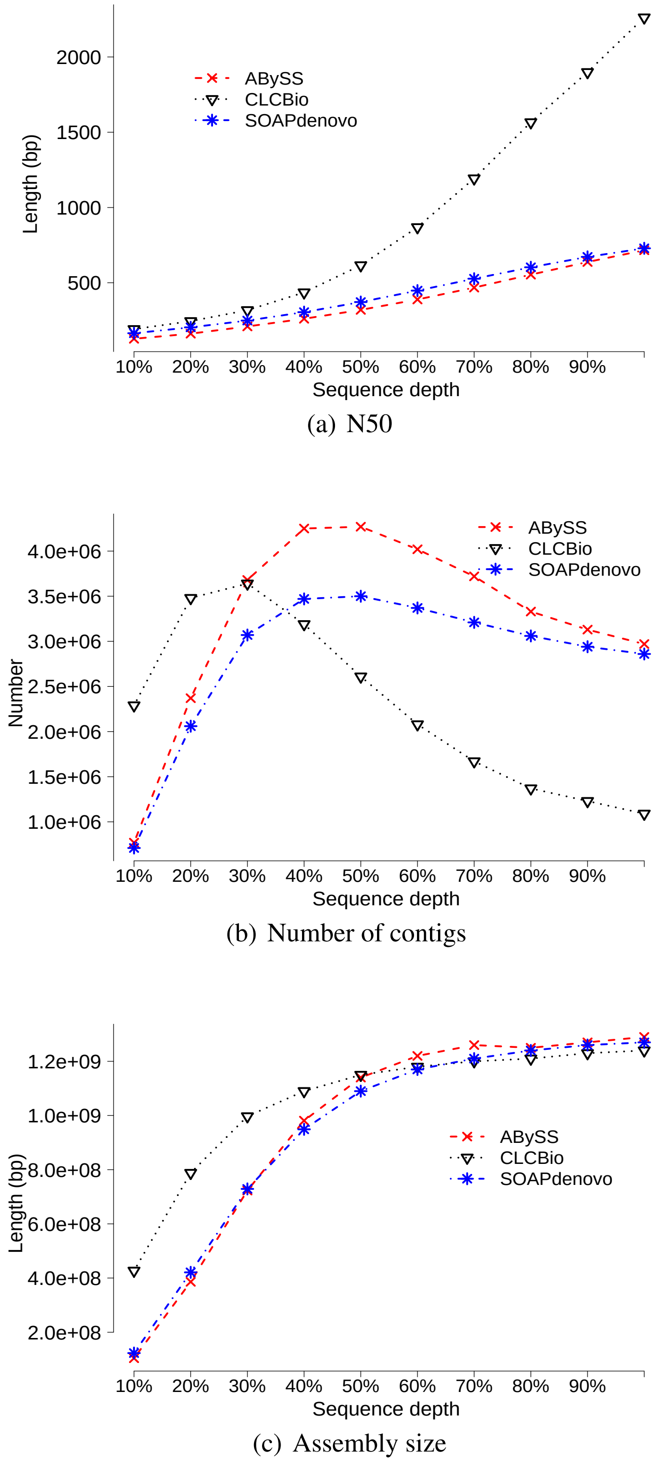
In order to determine whether the current genomic sequencing data is sufficient to generate a reasonably good assembly, we analysed the dependency of the number, length and total size of contigs with sequencing depth. Sequencing depth is represented by different subsets of the read data size. All three assemblers have similar trends for different properties. Fig. 5(a) shows that the average contig length is increasing with sequencing depth in all the assemblers. In Fig. 5(b), it shows how the number of contigs changes along with sequencing depth. The number of contigs first increases as most contigs only consist of one single read. As the number of reads increases, the chance that reads will overlap to form longer contigs increases. After a certain point, the increase of reads number leads to a decrease in the number of contigs. Fig. 5(c) illustrates that the size of the assembly almost reached the saturation status.

From the plots in Fig. 5, we conclude that current data is sufficient to cover the whole carp genome but the assembly is still fragmented. Given more data, CLCBio will generate a better assembly with fewer but longer contigs at higher speed when comparing to the rest.

In order to identify expressed sequences, an assembly of the RNA-Seq data was performed. Using the CLCBio assembler, we were able to achieve RNA contigs from RNA-Seq data with a total size of 71.6 Mbp and an N50 of 255 bp. Integrating RNA contigs with the existing EST and mRNA sequences from GenBank, the RNA assembly reaches to an N50 contig length of 896 bp and 18.1 Mbp in total size. The details are showed in Table 5.

4.2 4.2 TIR containing genes and products

The search for zebrafish TIR genes allowed us to identify characteristics of the gene structures and protein domain structures which will help us to annotate the corresponding genes in the carp genome. We used HMMsearch for mining of the protein domains in 35 zebrafish TIR proteins obtained from Ensembl, and we found that this class of proteins contains 28 domains in total. We also performed a search for all TIR domain-containing genes in zebrafish and found that out of 33 zebrafish TIR genes 16 of them are single-exon genes, 15 genes contain multiple exons and two gene structures are missing from



(a) N50

(b) Number of contigs

(c) Assembly size

Figure 5: Comparison of ABySS and CLCBio and SOAPdenovo.

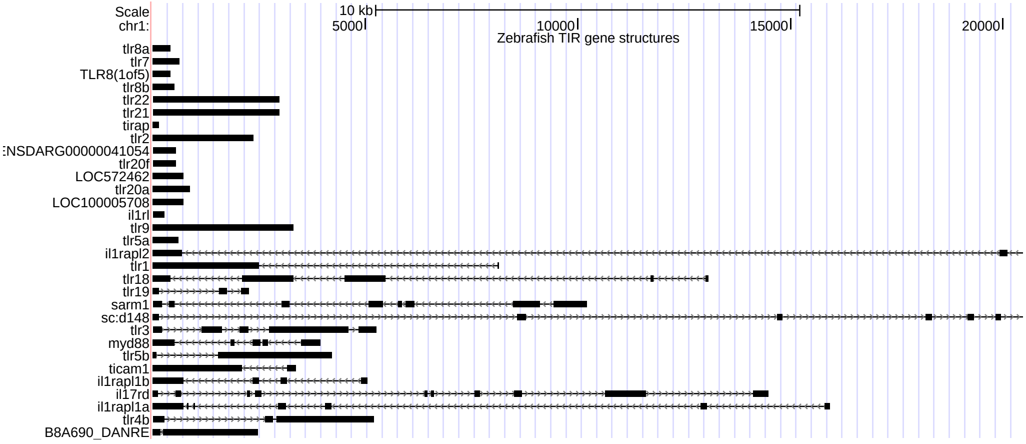


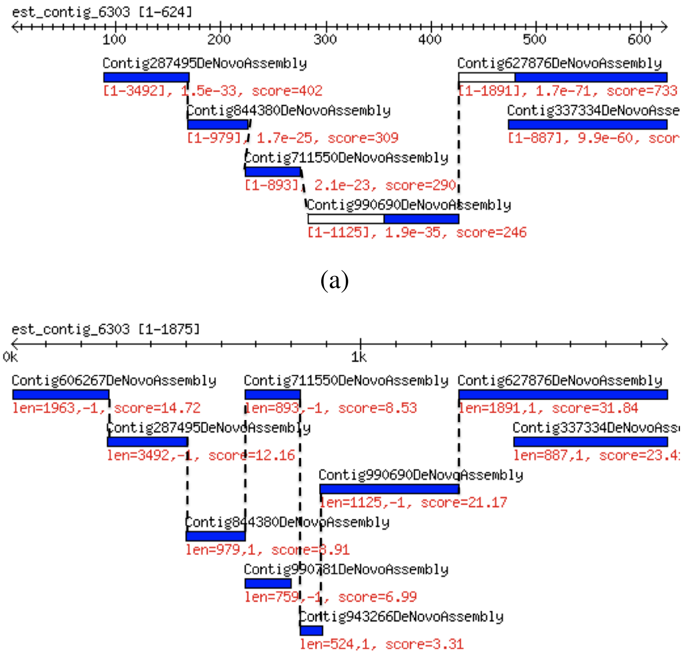
Figure 6: Gene structures of 31 zebrafish TIR domain-containing genes.

Ensembl. The structures of zebrafish TIR domain-containing genes are displayed in Fig. 6.

We were thus able to compare the carp RNA contigs to the zebrafish TIR containing proteins, as well as using their translated sequences to BLAST the carp genome assembly in order to identify potentially fragmented DNA contigs. By setting the cut-off E=1e-05, we discovered 39 carp TIR protein contigs similar to 34 zebrafish TIR RNA sequences as shown in the Supplementary Table 1. Both the protein and derived RNA sequences are further used to discover TIR genomic sequences.

Using protein or RNA contigs as references, we can create longer artificial DNA scaffolds with unknown gap sizes, which can hopefully contribute in obtaining a complete gene model. RNA sequences can be more diverse than DNA due to the alternative splicing, which will increase the chance of connecting the DNA contigs in the wrong order. In this case, we focus on detecting TIR domain-containing genes. In zebrafish, half of these genes contain only one exon which can not lead to alternative splicing. Moreover, there are 35 TIR containing proteins derived from 33 TIR genes. The ratio of gene to transcript is almost 1. Finally, if the alternative splicing events do exist, as long as they do not happen between the joint contigs, the right order of contigs can still be obtained. Therefore, we believe that using RNA contig or peptide sequences to scaffold TIR DNA contigs in carp is a good practical solution.

Using 39 TIR proteins and RNAs, we finally generated two versions of TIR gene sequences, i.e. cTIR genes v1 and cTIR genes v2 depicted in Fig. 4. In Fig. 7, an example is shown of the comparison of the protein sequence est_contig_6303 7(a) and RNA contig


est_contig_6303 [1-624]
100
200
300
400
500
600
Contig287495DeNovoAssembly
[1-3492], 1.5e-33, score=402
Contig844380DeNovoAssembly
[1-979], 1.7e-25, score=309
Contig711550DeNovoAssembly
[1-893], 2.1e-23, score=290
Contig990690DeNovoAssembly
[1-1125], 1.9e-35, score=246
Contig627876DeNovoAssembly
[1-1891], 1.7e-71, score=733
Contig337334DeNovoAsse
[1-887], 9.9e-60, scor
est_contig_6303 [1-1875]
0k
1k
Contig606267DeNovoAssembly
len=1963,-1, score=14.72
Contig287495DeNovoAssembly
len=3492,-1, score=12.16
Contig844380DeNovoAssembly
len=979,1, score=8.91
Contig990781DeNovoAssembly
len=759,-1, score=6.99
Contig943266DeNovoAssembly
len=524,1, score=3.31
Contig711550DeNovoAssembly
len=893,-1, score=8.53
Contig990690DeNovoAssembly
len=1125,-1, score=21.17
Contig627876DeNovoAssembly
len=1891,1, score=31.84
Contig337334DeNovoAss
len=887,1, score=23.4

(a)

(b)

Figure 7: Using RNA and protein contigs to scaffold genomic contigs. Genomic contigs shown in blue are aligned to protein contig 6303 (a) and RNA contig 6303 (b) depicted by the first line in each panel. Hits in the same region are ordered by the mapped scores, with the best matches at the top. Dotted lines indicate that the contigs can be joined with unknown gap size.

est_contig_6303 7(b) mapping to the carp genome. In Fig. 7(a), DNA contigs (No.287495, 843380, 711550, 990690 and 627876) can be joined together as a scaffold; in Fig. 7(b), it shows not only the previous 5 DNA contigs but also DNA contig No.606267 and 943266 can be bridged. We also found that the connected DNA contigs are largely identical between the two versions of TIRs. Scaffolding genomic contigs by BLAT RNA contigs (version 2) against the genome performs better than using protein sequences (version 1), since more DNA contigs can be joined and the RNA sequences can be constructed from the DNA contigs without any missing sequences. In total, 162 genomic DNA contigs are scaffolded using 39 TIR transcript sequences.

5 Conclusions

We have generated a draft assembly for common carp with N50 of 2260 bp and genome size of 1.23 Gbp. Due to the fact that the assembly still contains many fragments, we could not directly apply *ab initio* gene prediction method for gene discovery. Therefore, we developed an annotation pipeline which integrates whole genome sequencing, RNA-Seq data and available zebrafish data to detect the TIR containing genes in carp. We identified 39 TIR domain-containing transcripts. Using these transcripts as references, 162 DNA contigs are stitched to 39 DNA scaffolds. Potentially, the extended genome contigs will stand a high probability of containing the entire gene. Considering the facts that the ratio of known TIR genes and proteins in zebrafish is 33:35, and that half of these genes contains only a single exon (ruling out alternative splicing), a 1:1 ratio between TIR transcripts and genes in carp is a reasonable approximation. Therefore we established that there are around 39 TIR containing genes and transcripts in common carp.

Based on the performance of ABySS, SOAPdenovo and CLCBio, we decided to apply CLCBio as the final genomic sequence assembler since it produced relatively long contigs covering a similar amount of genome as the other two approaches. Without performing scaffolding, the initial carp genome assembly has reached an N50 of 2260 bp. The N50 contig length is longer than the initial assembly of giant panda genome of which the N50 contig length is 1483 bp [10]. In the giant panda project, the assembly was further improved to contig length N50 of 39886 bp by iterating scaffolding using extra 500 bp, 2 Kbp, 5 Kbp, and 10 Kbp libraries. Compared to this case, the limitation of our experimental design is that we currently did not have different sequence libraries for the haploid fish strain without which scaffolding or the finishing step is difficult to carry out. In the future, with more and larger size libraries available, the assembly will be further improved.

We found that when the data is limited, gene identification analysis is not as straightforward as the standard analysis which usually consists of gene assembly (if necessary) and *ab initio* gene prediction analysis or mapping RNA reads to the well-assembled genome to discover expressed sequences. In our study, we noticed that although the sequencing depth of genomic and transcriptomic data was not sufficient to produce the complete genome and transcriptome assemblies independently, these data are related and can be used as a complementary resource to support each other. Therefore, we developed a complicated gene identification analysis that integrated different data sources and types to maximize the probability of detecting the target genes. We first assemble the carp genome. Lacking long library for scaffolding, RNA-Seq data is not only used for measuring gene expression level but also for scaffolding the DNA contigs. Finally, the TIR domain-containing gene sequences in carp are captured by a comparative genomics analysis using zebrafish resources, since TIR domain is highly conserved and zebrafish genome is well annotated.

In the end, we scaffolded 162 DNA contigs to 39 TIR domain-containing gene sequences. However, only knowing the sequences is not enough. In the future, an âb initio gene prediction algorithm, e.g. AUGUSTUS [21], will be applied on these TIR sequences to further define the gene structures such as the precise start and stop position of a gene and its exons. After having a more complete genome assembly and the gene structures, the TIR containing gene expression in different samples can be measured by mapping the RNA reads to the carp genome using the tools such as TopHat [25] and/or Cufflinks [26].

Acknowledgements

We thank M. Forlenza and G. Wiegertjes from the Cell Biology and Immunology Group, Wageningen University for providing the zebrafish immune gene sequences. This project is partially supported by European Science Foundation FFG Exchange grant No.2952 and BioRange program of the Netherlands Bioinformatics Centre (NBIC, BSIK grant).

References

[1] Serafim Batzoglou, David B. Jaffe, Ken Stanley, Jonathan Butler, Sante Gnerre, Evan Mauceli, Bonnie Berger, Jill P. Mesirov, and Eric S. Lander. ARACHNE: a whole-genome shotgun assembler. *Genome research*, 12(1):177–189, January 2002.

[2] CLC Bio. http://www.clcbio.com/.

[3] A. B. J. Bongers, M. Sukkel, G. Gort, J. Komen, and C. J. J. Richter. Development and use of genetically uniform strains of common carp in experimental animal research. *Lab Anim*, 32(4):349–363, 1998.

[4] Jonathan Butler, Iain MacCallum, Michael Kleber, Ilya A. Shlyakhter, Matthew K. Belmonte, Eric S. Lander, Chad Nusbaum, and David B. Jaffe. ALLPATHS: de novo assembly of whole-genome shotgun microreads. *Genome research*, 18(5):810–820, May 2008.

[5] Richard Durbin, Sean R. Eddy, Anders Krogh, and Graeme Mitchison. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, July 1999.

[6] Paul Flicek, M. Ridwan Amode, Daniel Barrell, Kathryn Beal, Simon Brent, Yuan Chen, Peter Clapham, Guy Coates, Susan Fairley, and et. al. Ensembl 2011. *Nucleic acids research*, 39(Database issue), January 2011.

[7] Sarah Hunter, Rolf Apweiler, Teresa K. Attwood, Amos Bairoch, Alex Bateman, David Binns, Peer Bork, Ujjwal Das, Louise Daugherty, and et. al. InterPro:

the integrative protein signature database. *Nucleic acids research*, 37(Database issue):D211–D215, January 2009.

[8] C. Iseli, C. V. Jongeneel, and P Bucher. Estscan: a program for detecting, evaluating, and reconstructing potential coding regions in est sequences. *Proc Int Conf Intell Syst Mol Biol*, pages 138–148, 1999.

[9] W. James Kent. BLAT–the BLAST-like alignment tool. *Genome Research*, 12(4):656–664, April 2002.

[10] Ruiqiang Li, Wei Fan, Geng Tian, Hongmei Zhu, Lin He, Jing Cai, Quanfei Huang, Qingle Cai, Bo Li, and et. al. The sequence and de novo assembly of the giant panda genome. *Nature*, 463(7279):311–317, January 2010.

[11] Ruiqiang Li, Hongmei Zhu, Jue Ruan, Wubin Qian, Xiaodong Fang, Zhongbin Shi, Yingrui Li, Shengting Li, Gao Shan, Karsten Kristiansen, Songgang Li, Huanming Yang, Jian Wang, and Jun Wang. De novo assembly of human genomes with massively parallel short read sequencing. *Genome Research*, 20(2):265–272, December 2009.

[12] Scott Mcginnis and Thomas L. Madden. Blast: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res*, 32:20–25, 2004.

[13] A.H. Meijer, S.F. Gabby Krens, I.A. Medina Rodriguez, S. He, W. Bitter, B. Ewa Snaar-Jagalska, and H.P. Spaink. Expression analysis of the toll-like receptor and tir domain adaptor families of zebrafish. *Mol Immunol*, 40(11):773–783, 2004.

[14] Eugene W. Myers. The fragment assembly string graph. *Bioinformatics*, 21(suppl 2):ii79–ii85, September 2005.

[15] Mihai Pop. Genome assembly reborn: recent computational challenges. *Briefings in bioinformatics*, 10(4):354–366, July 2009.

[16] Mihai Pop, Adam Phillippy, Arthur L. Delcher, and Steven L. Salzberg. Comparative genome assembly. *Briefings in Bioinformatics*, 5(3):237–248, September 2004.

[17] Wei Qu, Shin-ichi Hashimoto, and Shinichi Morishita. Efficient frequency-based de novo short-read clustering for error trimming in next-generation sequencing. *Genome Research*, 19(7):1309–1315, July 2009.

[18] F. Sanger, S. Nicklen, and A. R. Coulson. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 74(12):5463–5467, December 1977.

[19] Stephan C. Schuster. Next-generation sequencing transforms today's biology. *Nature Methods*, 5(1):16–18, December 2007.

[20] Jared T. Simpson, Kim Wong, Shaun D. Jackman, Jacqueline E. Schein, Steven J. Jones, and Inanç Birol. ABySS: a parallel assembler for short read sequence data. *Genome research*, 19(6):1117–1123, June 2009.

[21] Mario Stanke, Oliver Keller, Irfan Gunduz, Alec Hayes, Stephan Waack, and Burkhard Morgenstern. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucl. Acids Res.*, 34(suppl_2):W435–439, July 2006.

[22] L. Stein. Genome annotation: from sequence to biology. *Nature reviews. Genetics*, 2(7):493–503, July 2001.

[23] O. W. Stockhammer, A. Zakrzewska, Z. Hegedus, H. P. Spaink, and A. H Meijer. Transcriptome profiling and functional analyses of the zebrafish embryonic innate immune response to salmonella infection. *J Immunol*, 182(9):5641–5653, 2009.

[24] Granger G. Sutton, Owen White, Mark D. Admas, and Anthony R. Kerlavage. TIGR assembler: A new tool for assembling large shotgun sequencing projects. *Genome Science and Technology*, 1(1), 1995.

[25] Cole Trapnell, Lior Pachter, and Steven L. Salzberg. TopHat: discovering splice junctions with RNA-seq. *Bioinformatics (Oxford, England)*, 25(9):1105–1111, May 2009.

[26] Cole Trapnell, Brian A. Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J. van Baren, Steven L. Salzberg, Barbara J. Wold, and Lior Pachter. Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology*, 28(5):511–515, May 2010.

[27] Daniel Zerbino and Ewan Birney. Velvet: Algorithms for de novo short read assembly using de bruijn graphs. *Genome Research*, 18(5):821–829, January 2008.

Supplementary results

Gene	Transcript	Name
ENSDARG00000010610	ENSDART00000004200	sarm1
ENSDARG00000010169	ENSDART00000011143	myd88
ENSDARG00000016065	ENSDART00000013021	tlr3
ENSDARG00000040249	ENSDART00000014310	-
ENSDARG00000022048	ENSDART00000034852	tlr4bb
ENSDARG00000026663	ENSDART00000036422	tlr19
ENSDARG00000069592	ENSDART00000044482	CR392351.1
ENSDARG00000037553	ENSDART00000054687	il1rapl2
ENSDARG00000037758	ENSDART00000055006	tlr2
ENSDARG00000058045	ENSDART00000060142	tlr21
ENSDARG00000041054	ENSDART00000060155	-
ENSDARG00000041164	ENSDART00000060337	si:dkey-193n17.7
ENSDARG00000042714	ENSDART00000062685	ticam1
ENSDARG00000043032	ENSDART00000063176	tlr1
ENSDARG00000044415	ENSDART00000065229	TLR5 (1 of 2)
ENSDARG00000044490	ENSDART00000065340	-
ENSDARG00000052322	ENSDART00000074153	tlr5b
ENSDARG00000062045	ENSDART00000089206	il1rapl1a
ENSDARG00000062204	ENSDART00000089680	sigirr
ENSDARG00000038843	ENSDART00000098676	-
ENSDARG00000038843	ENSDART00000098677	il17rd
ENSDARG00000068812	ENSDART00000099649	-
ENSDARG00000062045	ENSDART00000101171	il1rapl1a
ENSDARG00000031859	ENSDART00000101407	-
ENSDARG00000069593	ENSDART00000101409	si:dkey-100n23.2
ENSDARG00000070392	ENSDART00000103242	tlr19
ENSDARG00000075479	ENSDART00000108837	-
ENSDARG00000074371	ENSDART00000109673	tirap
ENSDARG00000078496	ENSDART00000110194	tlr8a
ENSDARG00000079621	ENSDART00000112641	-
ENSDARG00000078740	ENSDART00000112871	il1rapl1b
ENSDARG00000079737	ENSDART00000113028	-
ENSDARG00000075671	ENSDART00000113952	tlr4al
ENSDARG00000076245	ENSDART00000114883	-
ENSDARG00000068609	ENSDART00000099295	il1rl (no hits in carp)

Supplementary Table 1: 35 TIR domain containing peptides in zebrafish. It is found that 34 out of 35 have homologous sequences in carp. We did not find the homologous for i1lr1 gene which is highlighted and showed in the last of the table.