Identification of common carp innate immune genes with whole-genome sequencing and RNA-seq data

Zhang, Y.; Stupka, E.; Henkel, C.V.; Jansen, H.J.; Spaink, H.P.; Verbeek, F.J.

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

Yanju Zhang[1]*, Elia Stupka[2], Christiaan V. Henkel[3], Hans J. Jansen[3], Herman P. Spaink[3,4] and Fons J. Verbeek[1]*

[1]Section Imaging and BioInformatics, Leiden Institute of Advanced Computer Science (LIACS), Leiden University, The Netherlands

[2]Bioinformatics group, UCL Cancer Institute, London, UK

[3]ZF-screens B.V., Leiden, The Netherlands

[4]Institute of Biology Leiden, Leiden University, The Netherlands

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 involved in the development of the innate immune response, particularly TIR domain-containing genes, from a preliminary genome assembly. To achieve this, we developed a comprehensive gene identification pipeline. This analysis allowed us to estimate that the carp has 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 [1]. 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 diseases and cancer by identifying and killing pathogens and detrimental cells, and relies heavily on signalling 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, a highly conserved functional unit mediating the protein-protein interactions between the receptors and the adaptors.

TIR domain-containing genes therefore play important roles in immunity signalling pathways. In zebrafish (*Danio rerio*), this gene-family has been studied using microarray technology [2]. 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.

* To whom correspondence should be addressed. Email: {yanju|fverbeek}@liacs.nl

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

Dataset 1: gDNA, homozygous carp

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

Dataset 2: 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

Next-generation sequencing (NGS) is a recently developed, high-throughput sequencing technology, which produce millions of sequence reads in a few days at a low cost and without the need for a *priori* knowledge of the sequences [3]. 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 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. As a pilot study, we focus on discovering the main group of innate immune gene, i.e. the TIR domain-containing genes.

In this paper, 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 Methodology

The genome of a fully homozygous common carp, obtained in a single generation without inbreeding [1], was sequenced using Illumina Genome Analyzer IIx. We generated a paired-end sequencing library with insert sizes of about 200 base pairs (bp), from which approximately 24.5 Gbp of usable sequences with a read length of 76 bp was obtained. 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. 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.

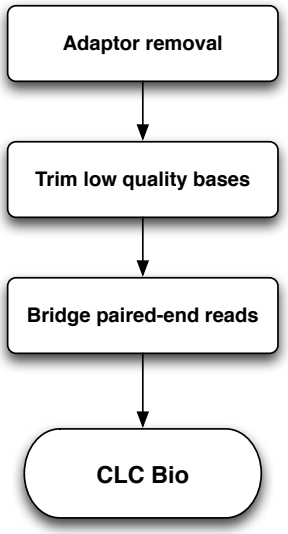

Figure 1: Genome assembly pipeline

2.1 Genome assembly strategy

In the absence of a carp reference genome, the first task is to generate a genome assembly. The strategy of the carp genome assembly is illustrated in Figure 1. First all the raw reads are filtered by quality control criteria and further pre-assembled in the preprocessing stage. Three *de novo* assemblers are applied to the high-quality reads, and the results are evaluated in order to achieve the best assembly. Finally, CLCBio has been chosen as the final tool.

The raw reads generated from the Illumina sequencer included base-calling errors and adapter contamination. It has been found that the Illumina sequencing quality decreases at the end of the read [4]. Adapter contamination is mainly caused by insert sizes smaller than the read length. Therefore, adapter sequence longer than 6 nucleotides were removed from individual reads, and low quality nucleotides or reads were discarded.

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 longer reads if repetitive sequences within them tended to create ambiguous connections. This preassembly procedure produces long reads which will not only potentially improve the efficiency and quality of the assembly, but also provide confirmation for the quality of the 3' end of the reads. After the preprocessing, 3.5% of nucleotides are discarded and 69.9% of pairs are merged.

Subsequently, we assembled the high quality and merged genomic DNA reads using three *de novo* assemblers: ABySS [5], CLCBio [6] and SOAPdenovo [7]. After a series of testing, CLCBio was chosen as the final tool to generate the carp genome assembly (see Supplementary Table 1). CLCBio is a commercial software product in which the assembly strategy is based on the de Bruijn graph theory. A de Bruijn graph is a directed graph representing overlaps between sequences of symbols [8]. Basically, in the graph, 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. An advantage of the CLCBio assembler is that it can tune and optimize the parameter *k* automatically [6].

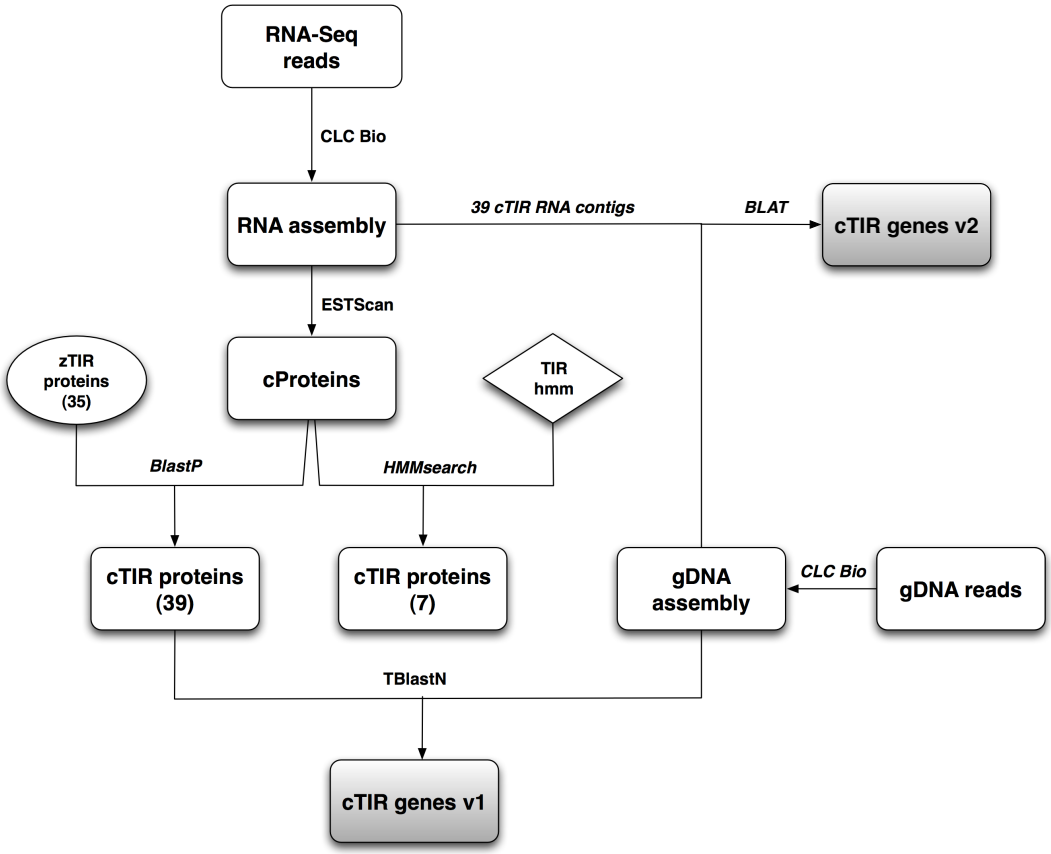

Figure 2: TIR containing gene annotation pipeline. Two version of carp TIR gene sequences (abbreviated as cTIR genes v1 and cTIR genes v2) derived from different methods are compared and further integrated to construct the final TIR genes.

2.2 Annotation strategy

Animal genome assemblies based on NGS data only are generally highly fragmented, and most contigs will not include entire genes. We therefore developed an integrative strategy to maximize the probability of identification of TIR containing genes and transcripts.

Firstly 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 [9]. Thereafter, the protein sequences obtained were searched for the TIR domain found in Interpro [10] using the HMMsearch algorithm [11].

In terms of evolution, the zebrafish is close to the common carp (both are cyprinids) and the zebrafish genome is relatively well covered and annotated in the Ensembl database [12]. Therefore we used this genome to facilitate the annotation of the carp TIR containing genes. The zebrafish TIR containing proteins found in Ensembl are BLASTed against the carp peptides obtained from the RNA-Seq contigs resulting in putative carp TIR proteins, allowing us to identify potentially new carp TIR transcripts. Since the obtained genome assembly is fragmented, the transcript and protein sequences can be used to bridge fragmented DNA contigs. To achieve this, the candidate TIR transcript and protein sequences were further mapped to the carp genomic contigs using TBlastN [13] and BLAT [14]. The DNA contigs identified are

Table 2: List of Carp genome assemblies produced by the CLCBio assembler using different parameters.

Assembly	k	n	n:N50	median	mean	N50	max	sum
CLCBio - preprocessing	25	1637271	250045	384	735	1409	17597	1.20E+09
CLCBio - preprocessing	27	1847118	-	-	680	1389	18684	1.26E+09
CLCBio + preprocessing	25	1086163	159656	587	1135	2260	26293	1.23E+09

connected using a number of 'N' as gap sequences and finally result an alternative set of carp TIR genes for comparison. The entire pipeline is shown in the Figure 2[1].

3 Results

3.1 Carp genome assembly

We ran the CLCBio *de novo* assembler on both the raw genomic reads and the reads after preprocessing. The optimal value for parameter *k* was automatically determined to be 25. An assembly with a manually adjusted *k-mer* of 27 yielded shorter contigs (lower N50 value 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 corresponds to a more contiguous assembly. 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 for the final assembly.

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 and total size of contigs on sequencing depth. Lower sequencing depth is simulated by taking subsets of the sequencing data. In Figure 3 (left scale and the solid black line), it shows how the number of contigs depends on 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 in the number of of reads number leads to a decrease in the number of contigs. This trend in general corresponds with our expectations, however, this figure also shows that there are still lots of contigs in the final assembly. The dotted red line in Figure 3 illustrates that the size of the assembly almost reached the saturation status. These observations support the conclusion that the current sequencing data is sufficient to cover the whole carp genome, although the assembly remains fragmented.

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, we obtained an RNA assembly with an N50 contig length of 896 bp and 18.1 Mbp in total size.

[1] c, z and v are the abbreviations of carp, zebrafish and version in the Figure 2

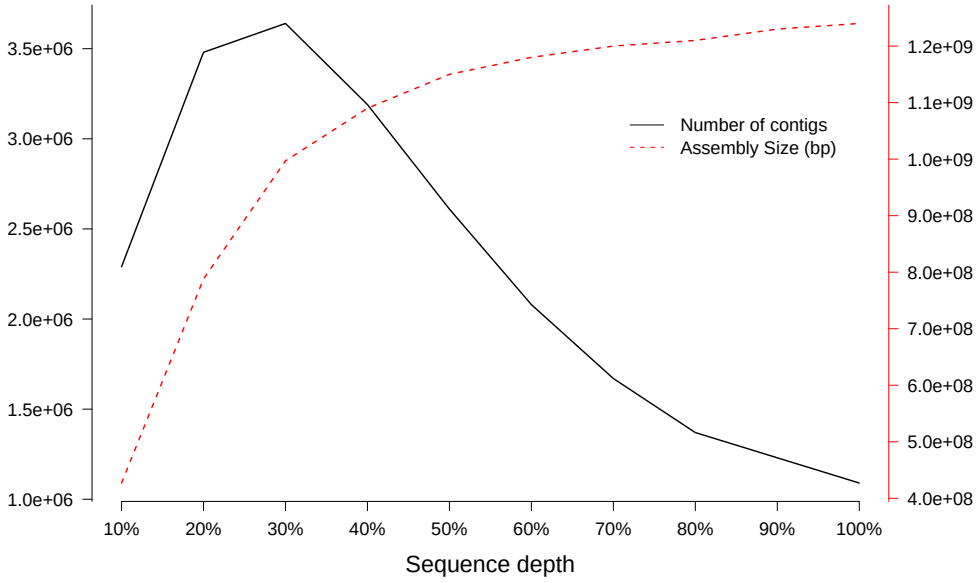

Figure 3: The dependency of the number and total size of contigs on sequencing depth which is represented by different subsets of the read data size. It shows how the number of contigs (solid line and left scale) and assembly size (dotted line and right scale) change with the data size in different assemblers.

3.2 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 retrieved 35 TIR proteins and 33 TIR domain-containing genes in zebrafish from Ensembl and found that out of 33 zebrafish TIR genes, 16 are single-exon genes, 15 genes contain multiple exons and two gene structures are missing from Ensembl. The structures of zebrafish TIR domain-containing genes are displayed in Figure 4.

We were thus able to compare the carp RNA contigs to the zebrafish TIR containing proteins, as well as employ 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 2. 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 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 probability of connecting the DNA contigs in the wrong order. However, considering the zebrafish TIR containing genes, this is unlikely to present a major problem in assembling this class of genes: half of these genes contain only one exon which cannot lead to alternative splicing, and the ratio of genes to known proteins is close to one (35 TIR-containing proteins derived from 33 TIR genes). Even if alternative splicing events do exist, as long as they do not happen between the joined contigs, the right order of contigs can still be obtained.

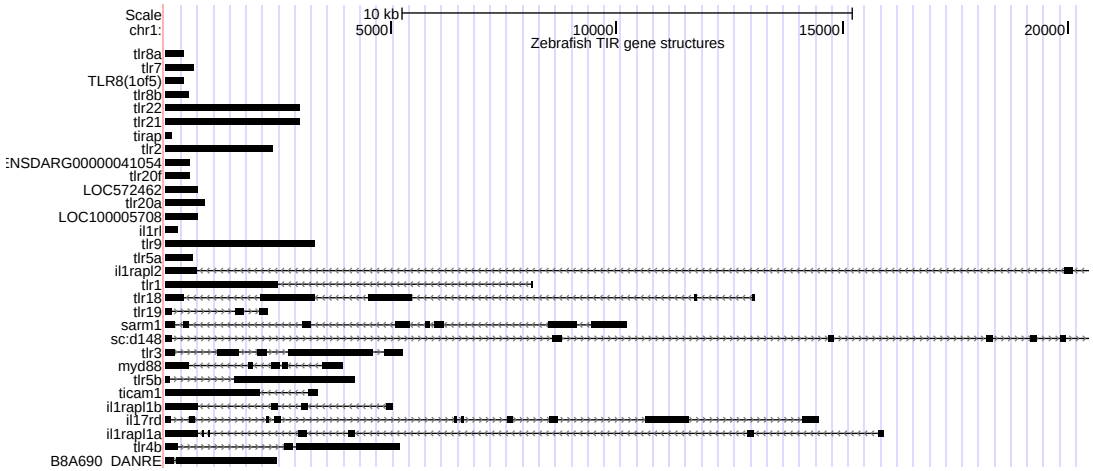

Figure 4: Gene structures of 31 zebrafish TIR domain-containing genes.

Using 39 TIR proteins and RNAs, we finally generated two versions of TIR gene sequences noted as cTIR genes v1 and cTIR genes v2 in Figure 2. All the BLAT and BLAST results are visualised using Bioperl scripts. In Figure 5, the comparison of the protein sequence est_contig_6303 (5a) and RNA contig est_contig_6303 (5b) mapping to the carp genome is shown as an example. In the figures, the first line with scale represents the reference and the bars represent the hits which are the DNA contigs in this case. When having the multiple hits, the hits are sorted by the alignment scores descend from up to down (the contigs in the uppermost within a region resemble the reference the most). Only the most likely contigs are used to construct the reference. In Figure 5a, DNA contigs (No. 287495, 843380, 711550, 990690 and 627876) can be joined together as a scaffold; in Figure 5b, 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.

4 Conclusions

We have generated a draft assembly for common carp with an 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 methods 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 together in 39 DNA scaffolds. The extended genomic scaffolds are likely to contain entire coding sequence of genes. 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 assumption. We therefore propose that the common carp genome contains approximately 39 TIR containing genes. The wet-lab experiments to validate this result are currently on the way.

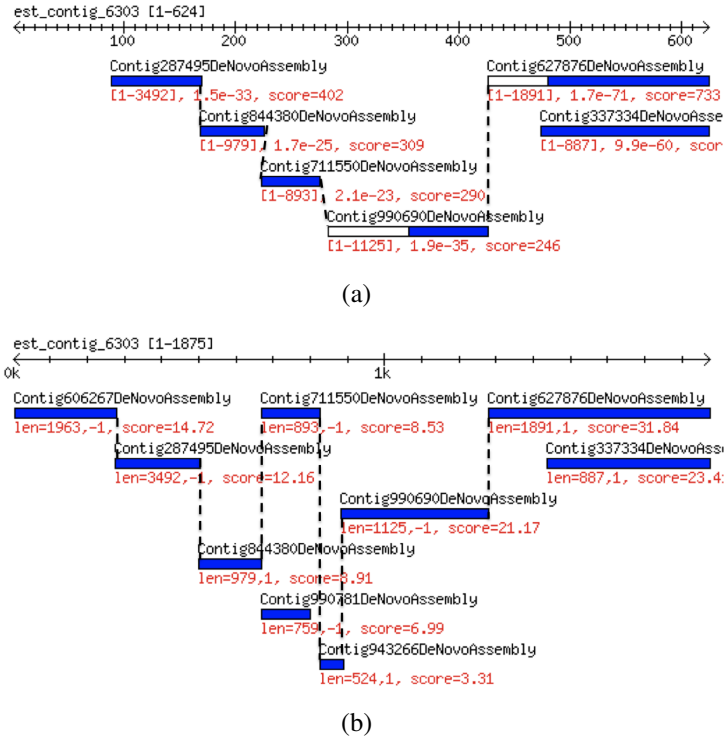

Figure 5: Using RNA and protein contigs to scaffold genomic contigs. Genomic contigs shown in blue are aligned to protein contig 6303 (a) and RNA contig 6303 (b) depicted by the first line in each panel. Hits in the same region are ordered by the mapped scores, with the best matches at the top. Dotted lines indicate that the contigs can be joined with unknown gap size.

We found that when the data is limited, gene identification analysis is not straightforward. A standard analysis usually consists of gene assembly (if necessary) and *ab initio* gene prediction, often in combination with mapping of RNA reads to a 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 complete genome and transcriptome assemblies independently, these data are related and can be used as a complementary resource to support each other's assembly. Therefore, we developed a sophisticated gene identification analysis that integrates different data sources and types to maximize the probability of detecting the target genes. We first assemble the carp genome. Lacking long libraries for scaffolding, RNA-Seq data is used for scaffolding the DNA contigs. Finally, the TIR domain-containing gene sequences in carp are captured by a comparative genomics analysis using zebrafish resources.

However, only knowing the TIR containing gene sequences is not enough. An *ab initio* gene prediction algorithm, e.g. AUGUSTUS [15], 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 that, the TIR containing gene expression in different samples can be measured by mapping the RNA reads to the carp genome using tools such as TopHat [16] and/or Cufflinks [17]. In the future, with more and larger size libraries available, the carp genome assembly will be further improved. The integration of more high-quality and heterogeneous data will therefore facilitate gene identification process.

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 the BioRange program of the Netherlands Bioinformatics Centre (NBIC, BSIK grant).

References

[1] 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. *Laboratory Animals*, 32(4):349–363, 1998.

[2] Oliver W. Stockhammer, Anna Zakrzewska, Zoltán Hegedus, Herman P. Spaink, and Annemarie H. Meijer. Transcriptome Profiling and Functional Analyses of the Zebrafish Embryonic Innate Immune Response to Salmonella Infection. *The Journal of Immunology*, 182(9):5641–5653, 2009.

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

[4] 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, 2009.

[5] 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, 2009.

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

[7] 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, 2009.

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

[9] Christian Iseli, C. Victor Jongeneel, and Philipp 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.

[10] Sarah Hunter, Rolf Apweiler, Teresa K. Attwood, Amos Bairoch, Alex Bateman, David Binns, Peer Bork, Ujjwal Das, Louise Daugherty, Lauranne Duquenne, et al. InterPro: the integrative protein signature database. *Nucleic Acids Research*, 37(Database Issue):D211–D215, 2009.

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

[12] Paul Flicek, M. Ridwan Amode, Daniel Barrell, Kathryn Beal, Simon Brent, Yuan Chen, Peter Clapham, Guy Coates, Susan Fairley, Stephen Fitzgerald, Leo Gordon, et al. Ensembl 2011. *Nucleic Acids Research*, 39(Database Issue), 2011.

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

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

[15] Mario Stanke, Oliver Keller, Irfan Gunduz, Alec Hayes, Stephan Waack, and Burkhard Morgenstern. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Research*, 34(suppl_2):W435–W439, 2006.

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

[17] 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, 2010.

Supplementary results

Supplementary Table 1: Carp genome assemblies produced by CLCBio, ABySS and SOAPdenovo.

	n:N50	median(bp)	mean(bp)	N50(bp)	max(bp)	sum(bp)
CLCBio	**159656**	**587**	**1135**	**2260**	**26293**	**1.23G**
ABySS	515595	249	433	716	16601	1.29G
SOAPdenovo	500758	257	443	729	10342	1.27G

Supplementary Table 2: 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 il1r1 gene which is highlighted and showed in the last of the table.

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)