# Analysis of Eudiplozoon nipponicum genome

*Eudiplozoon nipponicum* is a parasitic worm of *Monogenea* class. We try to assemble the whole genome of this parasite.

We have two types of sequencing libraries – Illumina and Roche 454. Illumina libraries (HiSeq, MiSeq) are paired-end and we have 5 type of this library, and only one Roche 454 library. This library is single-end. In next table (Tab. 1) are summarized initial information about all libraries. These results were obtained with FastQC tool.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| File name | Instrument | Type | Sequence length | Number of reads | GC (%) |
| A72DD | MiSeq | 250 | 35-251 | 13,274,453 | 38 |
| A7KL0 | MiSeq | 250 | 251 | 10,783,010 | 37 |
| C5KL9ANXX | HiSeq | 125 | 125 | 39,443,850 | 33 |
| C4VFYACXX | HiSeq | 100 | 101 | 34,425,978 | 34 |
| C841DACXX | HiSeq | 100 | 101 | 66,703,004 | 33 |
| IDBBUAV01 | Roche 454 | 40-808 | 40-808 | 143,667 | 36 |

Tab. Information about all libraries

After this quality check the adapters were from reads removed. Also at almost all Roche 454 reads were the first eleven bases same, so this bases were cut from. Adapters were trimmed with **Trimmomatic** tool. After this preprocessing, the quality of reads was checked again (Tab. 2).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| File name | Sequence length after trimming | Number of reads before preprocessing | Number of reads after preprocessing | GC (%) |
| A72DD | 35-250 | 13,274,453 | 12,337,026 | 38/37 |
| A7KL0 | 35-250 | 10,783,010 | 10,365,367 | 36 |
| C5KL9ANXX | 35-125 | 39,443,850 | 37,976,023 | 33 |
| C4VFYACXX | 35-101 | 34,425,978 | 31,995,548 | 33 |
| C841DACXX | 35-100 | 66,703,004 | 65,453,250 | 33 |
| IDBBUAV01 | 35-552 | 143,667 | 98,495 | 38 |

Tab. Information about Illumina libraries after preprocessing

Our samples *E. nipponicum* were gained from *Cyprinus carpio*, so we assumed that in our reads could be contamination with carp blood. Cause of that, the next step was mapping our reads set on *C. carpio* genome. The reads that were mapped on carp genome were discharged according to some adjusted criteria (mapq quality 30 for Illumina reads, mapq quality 20 for Roche 454 reads).

In table below are summarized data from mapping on carp genome.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| File name | Number of reads before preprocessing | Number of reads mapped on carp genome | Number of reads remained after filtration | Number of reads mapped on carp genome (%) |
| A72DD | 13,274,453 | 3,337,564 | 8,999,462 | 27.05 |
| A7KL0 | 10,783,010 | 1,849,514 | 8,515,853 | 17.84 |
| C5KL9ANXX | 39,443,850 | 10,245,138 | 27,730,885 | 26.98 |
| C4VFYACXX | 34,425,978 | 9,261,565 | 22,733,983 | 28.95 |
| C841DACXX | 66,703,004 | 2,785,761 | 62,667,489 | 4.26 |
| IDBBUAV01 | 143,667 | 4,926 | 93,569 | 5.00 |

After pre-processing our data, the next step was the assembly. In first attempt, we try to use SOAPdenovo and SPAdes. Both SOAPdenovo and SPAdes did not end successfully. SPAdes have too much request for memory, but also this tool was not made for larger genomes. For this reason (we assumed that this parasite has large genome), this tool was excluded from possible used programs. SOAPdenovo was success only at one library, but if we try assemble all libraries simultaneously, results were same as at SOAPdenovo. For this reason, we select as next possible assembler MaSuRCA tool.

Into MaSuRCA were gave as input the raw reads. So the previous step were at this assembler useless. MaSuRCA was the only assembler that end successfully with all Illumina libraries as input. The Roche 454 will be used later. As output we get fasta file with scaffolds and contigs. The following statistics is summarization about these output files (Tab. 3).

|  |  |  |
| --- | --- | --- |
|  | Scaffolds | Contigs |
| Scaffolds/contigs (>= 0 bp) | 1,348,459 | 1,523,686 |
| Scaffolds/contigs (>= 1000 bp) | 463,988 | 434,268 |
| Scaffolds/contigs (>= 5000 bp) | 53,532 | 46,175 |
| Scaffolds/contigs (>= 10000 bp) | 9,451 | 6,523 |
| Scaffolds/contigs (>= 25000 bp) | 187 | 107 |
| Scaffolds/contigs (>= 50000 bp) | 30 | 27 |
| Total length (>= 0 bp) | 1,746,161,452 | 1,739,454,547 |
| Total length (>= 1000 bp) | 1,234,120,448 | 1,125,506,874 |
| Total length (>= 5000 bp) | 426,028,184 | 351,725,551 |
| Total length (>= 10000 bp) | 128,264,684 | 86,979,923 |
| Total length (>= 25000 bp) | 73,797,49 | 4,882,386 |
| Total length (>= 50000 bp) | 25,946,27 | 2,348,668 |
| Scaffolds/contigs[[1]](#footnote-1) | 989,953 | 1,051,515 |
| Largest scaffold/contig | 192,299 | 192,299 |
| Total length | 1,604,366,768 | 1,554,896,522 |
| GC (%) | 35.67 | 35.64 |
| N50[[2]](#footnote-2) | 2,308 | 2,062 |
| N75 | 1,055 | 933 |
| L50[[3]](#footnote-3) | 166,393 | 181,656 |
| L75 | 433,949 | 476,392 |
| N’s per 100 kbp[[4]](#footnote-4) | 418,04 | 0.00 |

Tab. Statistic information about scaffolds and contigs from MaSuRCa

In next step, we blast these scaffolds on carp genome and filter some scaffolds that will blasted on carp genome according adjusted criteria. After this step, we try to assemble these scaffolds with SOAPdenovo to improve the assembly and also we try to improve this assembly with transcriptome.

At the time when assembly was running, we try to estimate a genome size. The method that is mostly used is k-mer size estimation. A k-mer is a sequence of length *k* that is observed more than once in a genomic or sequenced sequence. Firstly, k-mer frequency was counted with Jellyfish tool (used k-mer sizes were 17, 21, 25, 29, 33, 37, 41). As input files were used files that were filtered from carp reads. At larger k-mer sized, precisely 29 and more, the estimation was not possible. This could mean that data are contaminated or that the genome is huge. Also, the larger k-mer we have, the more unique k-mer are and coverage is than smaller, and estimation is not possible. The genome size is than estimated as the total number of k-mers divided by the expected depth (coverage). Also some k-mers, mostly k-mer whose coverage was for example 0, 1, etc. (it was derived from histogram) were considered as errors. So for this filtered data, the estimation of genome size could be done only for k-mers of size – 17, 21 and 25. The estimated size was about 1,8 Gb.

We try to confirm this estimation on error corrected reads. Here we could estimate genomes size only from k-mer 17, but probably genome size was estimate on 1,77 Gb.

When I used reads that were corrected in MaSuRCa tool, the estimation was about 3x larger, it could be cause of carp contamination.

## Analysis of scaffolds

In Tab. 3 we can see statistic information about output scaffolds from MaSuRCa tool. These scaffolds were checked for contamination with *Cyprinus carpio.* The tool used for blast was **Blast-2.6.0+[[5]](#footnote-5)**. The whole process is described in Fig. 1. There were two approaches for blasting. First, gaps in alignment were allowed. Second, gaps were restricted.

From 1,348,459 were only 535,739[[6]](#footnote-6) scaffolds not blasted at all in case of both approaches. Remaining scaffolds were blasted either whole or part of sequence. In output file were 15 columns (Tab. 4) that contains information about these resulting alignments. The scaffolds that were mapped on genome were pull out[[7]](#footnote-7). For sure were also E-values checked[[8]](#footnote-8), it there are some higher, but all E-values was very small.

|  |  |
| --- | --- |
| Column | Content |
| 1 | Query ID |
| 2 | Query length |
| 3 | Subject ID |
| 4 | Subject length |
| 5 | % identity |
| 6 | Alignment length |
| 7 | Mismatches |
| 8 | Gaps open |
| 9 | Query start |
| 10 | Query end |
| 11 | Subject start |
| 12 | Subject end |
| 13 | E-value |
| 14 | Bit score |
| 15 | % query coverage |

Tab. The content of columns in output file from blast

The filtration was based on ratio of alignment length and query length. If this ratio was larger or equal than 0.95 it was taken as significant. The first columns (ID column) were printed into another file and then were compared IDs from both blast files (gaps allowed and gaps restricted). About 270,450 scaffolds were only in gaps allowed blast.

I tried some of these scaffolds blasted manually and check if gaps could have some meaning. In most cases gaps were not too large, and it looks, that also some of these (maybe all) scaffolds could be from carp.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Gaps allowed | Gaps restricted | Number of difference |
| ≥ 0.95 | 574,281 | 303,831 | 270,450 |
| < 0.95 | 703,005 | 777,318 | 74,313 |

Tab. Number of filtrated scaffolds

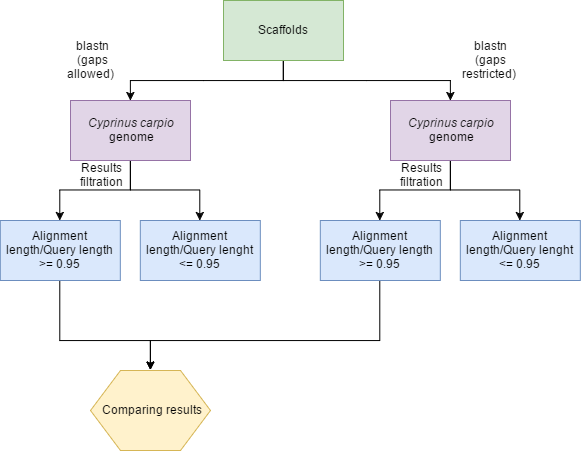
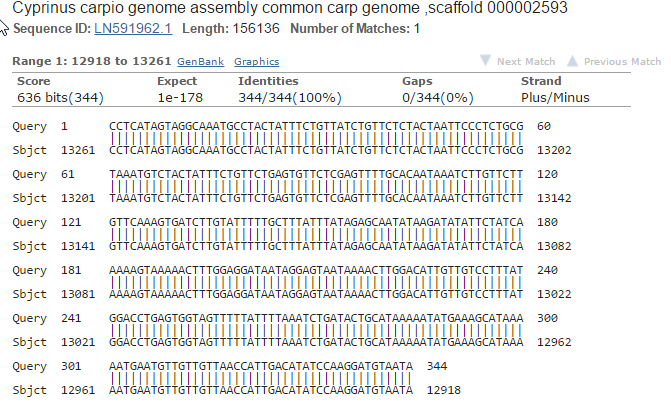


Fig. The process of filtration of blasted scaffolds

For sure I try to blast four scaffolds from gaps allowed if results are approximately same.

>jcf7180012513347

CCTCATAGTAGGCAAATGCCTACTATTTCTGTTATCTGTTCTCTACTAATTCCCTCTGCGTAAATGTCTACTATTTCTGTTCTGAGTGTTCTCGAGTTTTGCACAATAAATCTTGTTCTTGTTCAAAGTGATCTTGTATTTTTGCTTTATTTATAGAGCAATATAAGATATATTCTATCAAAAAGTAAAAACTTTGGAGGATAATAGGAGTAATAAAACTTGGACATTGTTGTCCTTTATGGACCTGAGTGGTAGTTTTTATTTTAAATCTGATACTGCATAAAAATATGAAAGCATAAAAATGAATGTTGTTGTTAACCATTGACATATCCAAGGATGTAATA



>jcf7180012525694

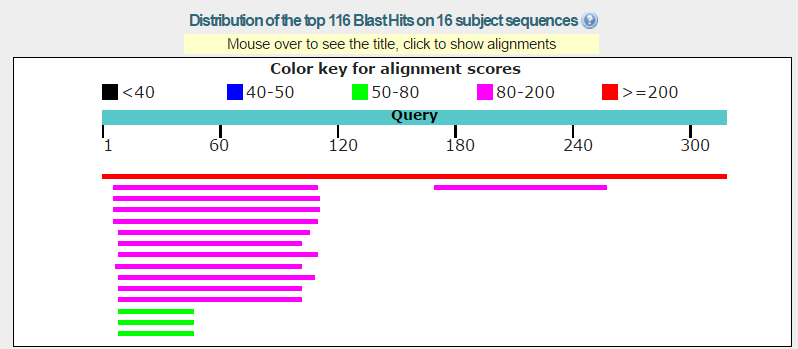
AGTGCAATCGCGAATCTCTCCTCTTCGCTTAAAATAAGCGTGTATGTGCTTAGACTGTGTCCCGTGCGTTCGCGCATGGCCAAGTGCTCTTCTCTTCGATTTCTTTCTCTTTGCTCTTGGCATAAATGCTCTTGGTTATAATCAATCACGCATTTTTATTTTTTTTTATCTAAAAAAAATCTCACAATGGGCATGGCACGTAATGTGAACAATGTGGTTGACAGTCTAGGTAATTTATATGTAAAATAATCCGATCGCCCTCTTCCCAGACATTTAGTTTTAAAACTTAAATGGTTTAAGGCAATCGGTTTCCTCAAACGGTTTGAGTAAACCTAAATTGTCAGGTTTTTAAGGATATCTGTTTACTCAAACGGTTTGAGTTACGTTACTTGTCGGGTTTTACAGTGTAAGATACATTTGAAATATTTAACATCAAAACAAATAAAACTGGTTACCCGAGATCGTCGATTCGGCGGATCCATCAAAGTCCCTCCAGAAAAAAGCGGATTTTTTTTGTGATTGTTGAGGGCAAAAATCCTTGATTTTGCGGCACGTTTTCTTAAAAAATGCAATGGAATATGCGGGATATTTTTGCAATTTTATGCGATGAAATTGTCGGAACTTGCAAAAACTGCGGTTTGATTAAAAAGAGAGAAAAAAAAGTGATTTCCTAGCCTAGGCTACTACCTTAATGTAAAGAGTCATTTCTTATTACTTGCTATGATAAGCAAGCATACTAAATCACAGAATATTTGTTCTCATTCTGTTTTATTTCAGACCTTAATTAACACATGGCTCAAACTTACAAAACATTGAGTCAAACAAGTTCTGTAGCTTTACAAAAGGAATATTCAGCAACTTTCTATAAAAATGTATAAATAAAATATAAAAAAAATAAAATGTGTGCTTCCTGTTTTACAGAGGTAGCTATACAAAACAGACATCTTCTGTTAAATTAAGCGGTTCAAATGCACAAAGTGCAATCTCTTACTGCAACGTAAATTATTTTAAATTAAAATTATTAATTAAATTATTAATTATTAATTCAACGTAAATTTACATCATGGTTTCACCCAGGTTTCACCACGGGGTTTGCAGATGATGTTCACGTCGTGGAATTACGTCACTTCATAAGGTTCCCATGGCAATGGGGGAAAATAATGGGGTTTTACTTTTGTGAAGTAAATGCAACATTTTTCAACTTTCTGCTAAGATATATGTGACTTTTTTGCAACAAAAATACAGGGATTATGAAATCATTTTGCTTGCTGAAATTGGCAATTTATGCGGCGAAAGTGCGGCGTATTTGAAAAAAAAAATCCGACCCCCGCATAAATTTGCGGCCTTTAGCTGATTATGCATGAAATCAGGCGATCGCATAATCGCGTTTTTCTGGAGGGACTGAGTCTTGATTGAACTTCAGAGAACTTATTCCATGTGTTCCATGTGCGGTTCAATATCAACCAATAAGATTCCATGTACGATGTCGCGTTT

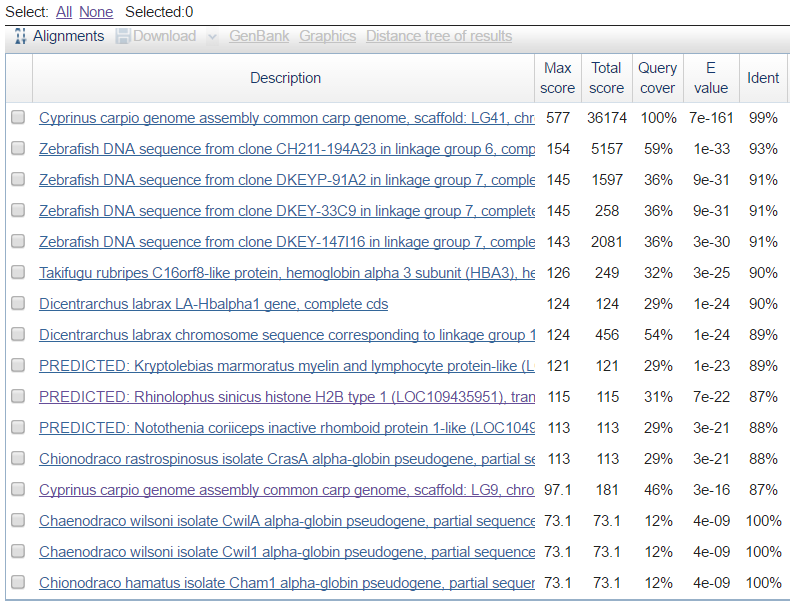


This is part of alignment, not whole. But in alignment is 36 gaps, it is maybe too large, but still it looks that it is from carp.

>jcf7180013851491

CGACATCAGCTCCAGTGGCGCAATCGGTCAGCGCGCGGTACTTATAAGGCAGTACGAAGCAGAGCAATGCCGAGTTTGTGAGTTCGAGCCACACCTGGAGCAAGGTTTTAGAGGATGAGAAGCATTTCAGCAGCTACAGAGTTCATGGGATAATATAAAAGATTTAATATATAAGGCAGTACGCAGCAGAGCAATGCCGAGGTTGTGAGTTTGAGCCCCACCTGGAGCAAGGTTTTAGAGGATGAGAAGCATTTCAGCAGCTACAGAGTTCATGGGATAATGTAAAAGATCCATTGAATATTAGATGGCTGTAATCTT

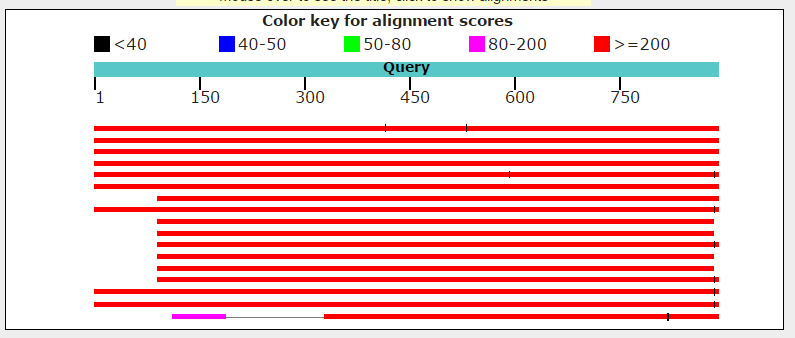


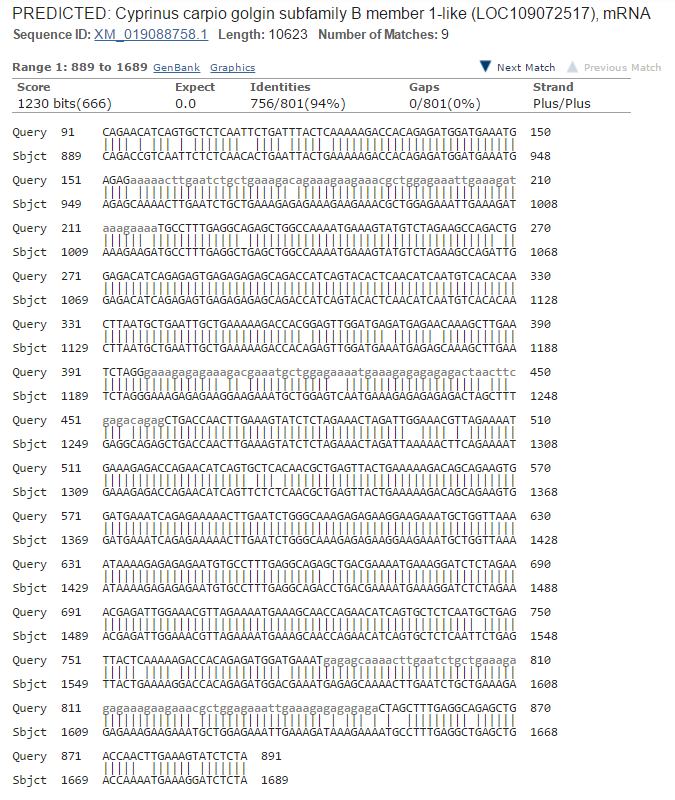


Here is some part of sequence common for more species, but still that’s are mainly fish species.

>jcf7180012549109

AAAGAGAGAGAGACTAACTTCGAGACAGAGCTGACCAACTTGAAAGTATCTCTAGAAACTAGATTGGAAATGTTAGAAAATGAAAGCAACCAGAACATCAGTGCTCTCAATTCTGATTTACTCAAAAAGACCACAGAGATGGATGAAATGAGAGAAAAACTTGAATCTGCTGAAAGACAGAAAGAAGAAACGCTGGAGAAATTGAAAGATAAAGAAAATGCCTTTGAGGCAGAGCTGGCCAAAATGAAAGTATGTCTAGAAGCCAGACTGGAGACATCAGAGAGTGAGAGAGAGCAGACCATCAGTACACTCAACATCAATGTCACACAACTTAATGCTGAATTGCTGAAAAAGACCACGGAGTTGGATGAGATGAGAACAAAGCTTGAATCTAGGGAAAGAGAGAAAGACGAAATGCTGGAGAAAATGAAAGAGAGAGAGACTAACTTCGAGACAGAGCTGACCAACTTGAAAGTATCTCTAGAAACTAGATTGGAAACGTTAGAAAATGAAAGAGACCAGAACATCAGTGCTCACAACGCTGAGTTACTGAAAAAGACAGCAGAAGTGGATGAAATCAGAGAAAAACTTGAATCTGGGCAAAGAGAGAAGGAAGAAATGCTGGTTAAAATAAAAGAGAGAGAATGTGCCTTTGAGGCAGAGCTGACGAAAATGAAAGGATCTCTAGAAACGAGATTGGAAACGTTAGAAAATGAAAGCAACCAGAACATCAGTGCTCTCAATGCTGAGTTACTCAAAAAGACCACAGAGATGGATGAAATGAGAGCAAAACTTGAATCTGCTGAAAGAGAGAAAGAAGAAACGCTGGAGAAATTGAAAGAGAGAGAGACTAGCTTTGAGGCAGAGCTGACCAACTTGAAAGTATCTCTA

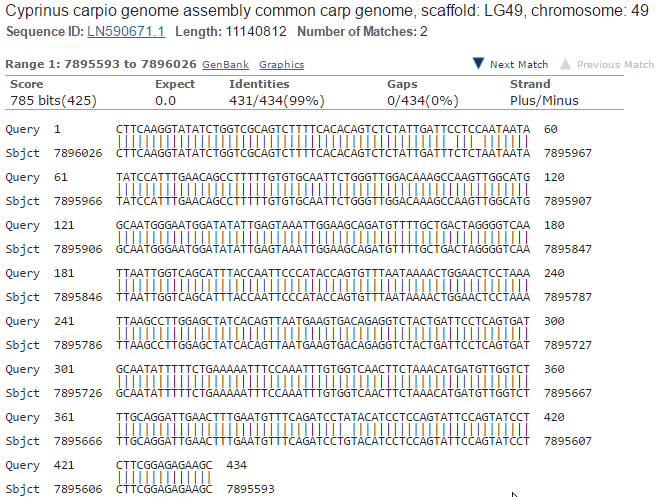




Also, for gaps restricted I try to blast four scaffolds to see resulting alignment.

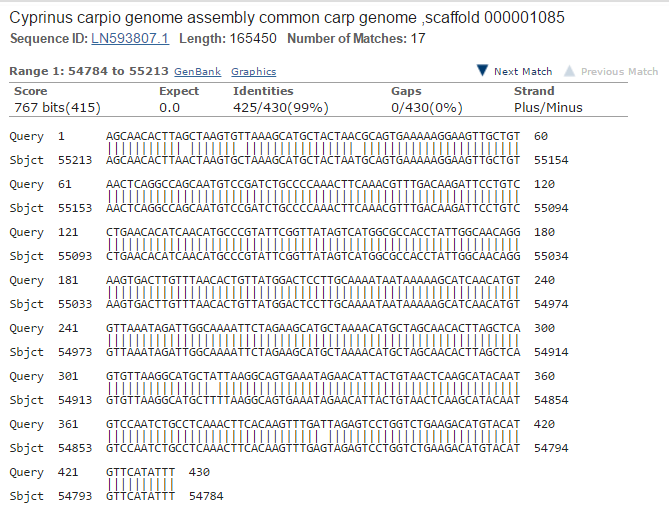
>jcf7180012514337

CTTCAAGGTATATCTGGTCGCAGTCTTTTCACACAGTCTCTATTGATTCCTCCAATAATATATCCATTTGAACAGCCTTTTTGTGTGCAATTCTGGGTTGGACAAAGCCAAGTTGGCATGGCAATGGGAATGGATATATTGAGTAAATTGGAAGCAGATGTTTTGCTGACTAGGGGTCAATTAATTGGTCAGCATTTACCAATTCCCATACCAGTGTTTAATAAAACTGGAACTCCTAAATTAAGCCTTGGAGCTATCACAGTTAATGAAGTGACAGAGGTCTACTGATTCCTCAGTGATGCAATATTTTTCTGAAAAATTTCCAAATTTGTGGTCAACTTCTAAACATGATGTTGGTCTTTGCAGGATTGAACTTTGAATGTTTCAGATCCTATACATCCTCCAGTATTCCAGTATCCTCTTCGGAGAGAAGC



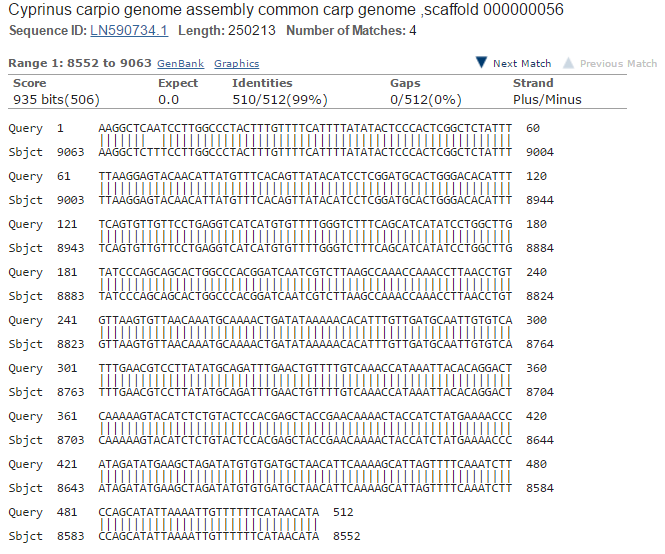
>jcf7180012531422

AGCAACACTTAGCTAAGTGTTAAAGCATGCTACTAACGCAGTGAAAAAGGAAGTTGCTGTAACTCAGGCCAGCAATGTCCGATCTGCCCCAAACTTCAAACGTTTGACAAGATTCCTGTCCTGAACACATCAACATGCCCGTATTCGGTTATAGTCATGGCGCCACCTATTGGCAACAGGAAGTGACTTGTTTAACACTGTTATGGACTCCTTGCAAAATAATAAAAAGCATCAACATGTGTTAAATAGATTGGCAAAATTCTAGAAGCATGCTAAAACATGCTAGCAACACTTAGCTCAGTGTTAAGGCATGCTATTAAGGCAGTGAAATAGAACATTACTGTAACTCAAGCATACAATGTCCAATCTGCCTCAAACTTCACAAGTTTGATTAGAGTCCTGGTCTGAAGACATGTACATGTTCATATTT



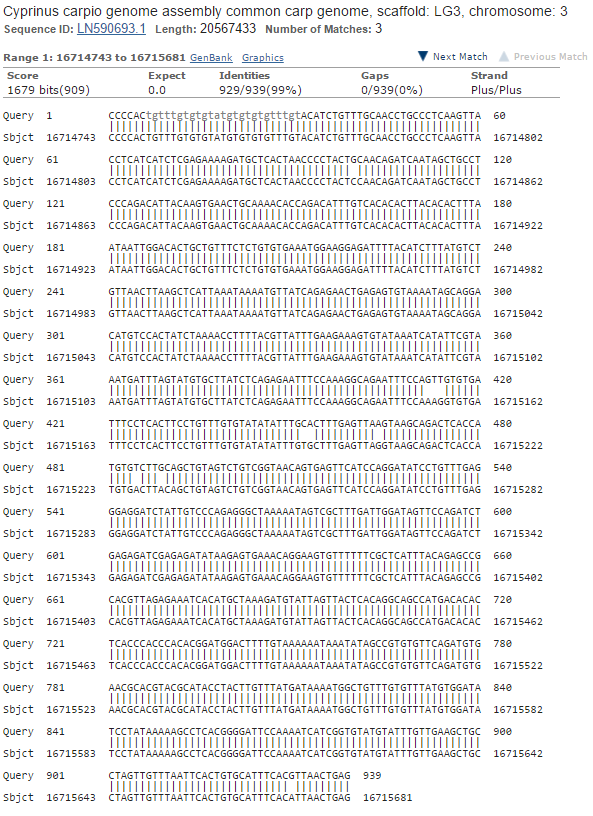
>jcf7180012664840

AAGGCTCAATCCTTGGCCCTACTTTGTTTTCATTTTATATACTCCCACTCGGCTCTATTTTTAAGGAGTACAACATTATGTTTCACAGTTATACATCCTCGGATGCACTGGGACACATTTTCAGTGTTGTTCCTGAGGTCATCATGTGTTTTGGGTCTTTCAGCATCATATCCTGGCTTGTATCCCAGCAGCACTGGCCCACGGATCAATCGTCTTAAGCCAAACCAAACCTTAACCTGTGTTAAGTGTTAACAAATGCAAAACTGATATAAAAACACATTTGTTGATGCAATTGTGTCATTTGAACGTCCTTATATGCAGATTTGAACTGTTTTGTCAAACCATAAATTACACAGGACTCAAAAAGTACATCTCTGTACTCCACGAGCTACCGAACAAAACTACCATCTATGAAAACCCATAGATATGAAGCTAGATATGTGTGATGCTAACATTCAAAAGCATTAGTTTTCAAATCTTCCAGCATATTAAAATTGTTTTTTCATAACATA



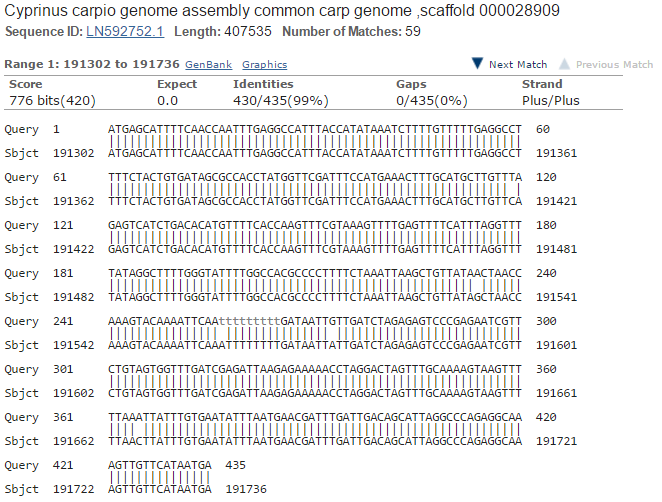
>jcf7180013843365

CCCCACTGTTTGTGTGTATGTGTGTGTTTGTACATCTGTTTGCAACCTGCCCTCAAGTTACCTCATCATCTCGAGAAAAGATGCTCACTAACCCCTACTGCAACAGATCAATAGCTGCCTCCCAGACATTACAAGTGAACTGCAAAACACCAGACATTTGTCACACACTTACACACTTTAATAATTGGACACTGCTGTTTCTCTGTGTGAAATGGAAGGAGATTTTACATCTTTATGTCTGTTAACTTAAGCTCATTAAATAAAATGTTATCAGAGAACTGAGAGTGTAAAATAGCAGGACATGTCCACTATCTAAAACCTTTTACGTTATTTGAAGAAAGTGTATAAATCATATTCGTAAATGATTTAGTATGTGCTTATCTCAGAGAATTTCCAAAGGCAGAATTTCCAGTTGTGTGATTTCCTCACTTCCTGTTTGTGTATATATTTGCACTTTGAGTTAAGTAAGCAGACTCACCATGTGTCTTGCAGCTGTAGTCTGTCGGTAACAGTGAGTTCATCCAGGATATCCTGTTTGAGGGAGGATCTATTGTCCCAGAGGGCTAAAAATAGTCGCTTTGATTGGATAGTTCCAGATCTGAGAGATCGAGAGATATAAGAGTGAAACAGGAAGTGTTTTTTCGCTCATTTACAGAGCCGCACGTTAGAGAAATCACATGCTAAAGATGTATTAGTTACTCACAGGCAGCCATGACACACTCACCCACCCACACGGATGGACTTTTGTAAAAAATAAATATAGCCGTGTGTTCAGATGTGAACGCACGTACGCATACCTACTTGTTTATGATAAAATGGCTGTTTGTGTTTATGTGGATATCCTATAAAAAGCCTCACGGGGATTCCAAAATCATCGGTGTATGTATTTGTTGAAGCTGCCTAGTTGTTTAATTCACTGTGCATTTCACGTTAACTGAG



>jcf7180013860895

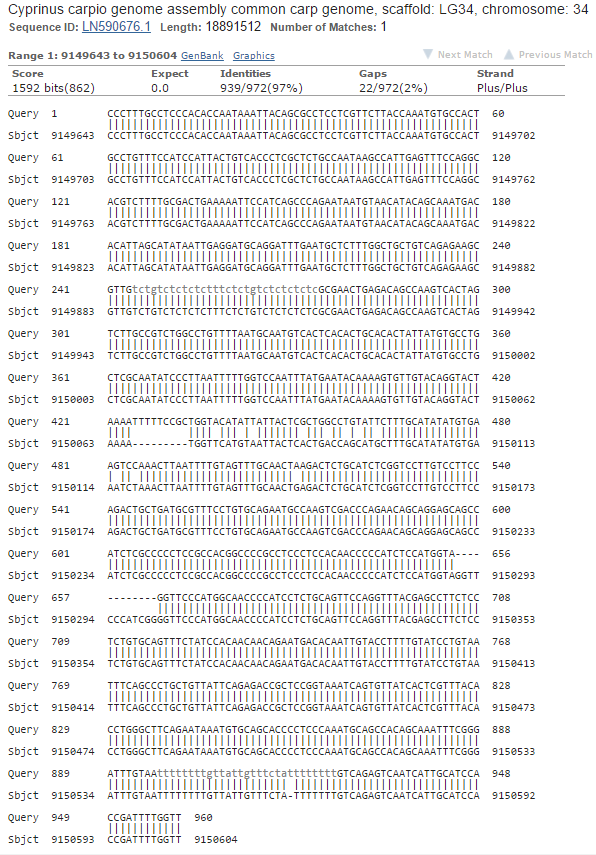
ATGAGCATTTTCAACCAATTTGAGGCCATTTACCATATAAATCTTTTGTTTTTGAGGCCTTTTCTACTGTGATAGCGCCACCTATGGTTCGATTTCCATGAAACTTTGCATGCTTGTTTAGAGTCATCTGACACATGTTTTCACCAAGTTTCGTAAAGTTTTGAGTTTTCATTTAGGTTTTATAGGCTTTTGGGTATTTTGGCCACGCCCCTTTTCTAAATTAAGCTGTTATAACTAACCAAAGTACAAAATTCAATTTTTTTTTGATAATTGTTGATCTAGAGAGTCCCGAGAATCGTTCTGTAGTGGTTTGATCGAGATTAAGAGAAAAACCTAGGACTAGTTTGCAAAAGTAAGTTTTTAAATTATTTGTGAATATTTAATGAACGATTTGATTGACAGCATTAGGCCCAGAGGCAAAGTTGTTCATAATGA



The scaffolds that were blasted in gaps allowed, but also in gaps restricted will be filtered absolutely, but also some scaffolds that were only in gaps allowed should be filtered. Some scaffolds that were blasted only with gaps allowed, were also blasted manually.

>jcf7180013798926

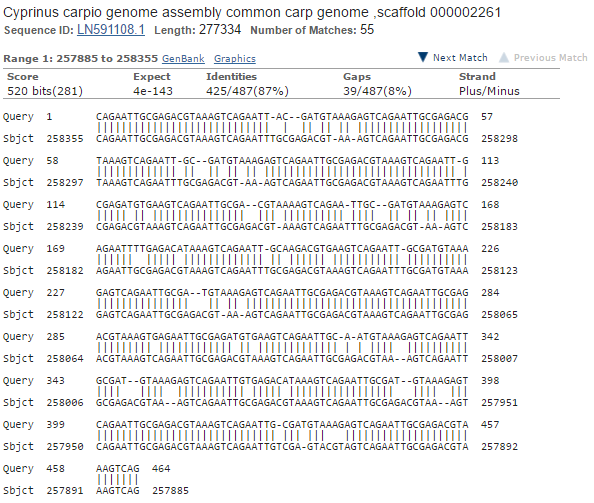
CCCTTTGCCTCCCACACCAATAAATTACAGCGCCTCCTCGTTCTTACCAAATGTGCCACTGCCTGTTTCCATCCATTACTGTCACCCTCGCTCTGCCAATAAGCCATTGAGTTTCCAGGCACGTCTTTTGCGACTGAAAAATTCCATCAGCCCAGAATAATGTAACATACAGCAAATGACACATTAGCATATAATTGAGGATGCAGGATTTGAATGCTCTTTGGCTGCTGTCAGAGAAGCGTTGTCTGTCTCTCTCTTTCTCTGTCTCTCTCTCGCGAACTGAGACAGCCAAGTCACTAGTCTTGCCGTCTGGCCTGTTTTAATGCAATGTCACTCACACTGCACACTATTATGTGCCTGCTCGCAATATCCCTTAATTTTTGGTCCAATTTATGAATACAAAAGTGTTGTACAGGTACTAAAATTTTTCCGCTGGTACATATTATTACTCGCTGGCCTGTATTCTTTGCATATATGTGAAGTCCAAACTTAATTTTGTAGTTTGCAACTAAGACTCTGCATCTCGGTCCTTGTCCTTCCAGACTGCTGATGCGTTTCCTGTGCAGAATGCCAAGTCGACCCAGAACAGCAGGAGCAGCCATCTCGCCCCCTCCGCCACGGCCCCGCCTCCCTCCACAACCCCCATCTCCATGGTAGGTTCCCATGGCAACCCCATCCTCTGCAGTTCCAGGTTTACGAGCCTTCTCCTCTGTGCAGTTTCTATCCACAACAACAGAATGACACAATTGTACCTTTTGTATCCTGTAATTTCAGCCCTGCTGTTATTCAGAGACCGCTCCGGTAAATCAGTGTTATCACTCGTTTACACCTGGGCTTCAGAATAAATGTGCAGCACCCCTCCCAAATGCAGCCACAGCAAATTTCGGGATTTGTAATTTTTTTTGTTATTGTTTCTATTTTTTTTGTCAGAGTCAATCATTGCATCCACCGATTTTGGTT



22 gaps is in this alignment, but still it looks that it is from carp genome.

>jcf7180013799133

CAGAATTGCGAGACGTAAAGTCAGAATTACGATGTAAAGAGTCAGAATTGCGAGACGTAAAGTCAGAATTGCGATGTAAAGAGTCAGAATTGCGAGACGTAAAGTCAGAATTGCGAGATGTGAAGTCAGAATTGCGACGTAAAAGTCAGAATTGCGATGTAAAGAGTCAGAATTTTGAGACATAAAGTCAGAATTGCAAGACGTGAAGTCAGAATTGCGATGTAAAGAGTCAGAATTGCGATGTAAAGAGTCAGAATTGCGAGACGTAAAGTCAGAATTGCGAGACGTAAAGTGAGAATTGCGAGATGTGAAGTCAGAATTGCAATGTAAAGAGTCAGAATTGCGATGTAAAGAGTCAGAATTGTGAGACATAAAGTCAGAATTGCGATGTAAAGAGTCAGAATTGCGAGACGTAAAGTCAGAATTGCGATGTAAAGAGTCAGAATTGCGAGACGTAAAGTCAG



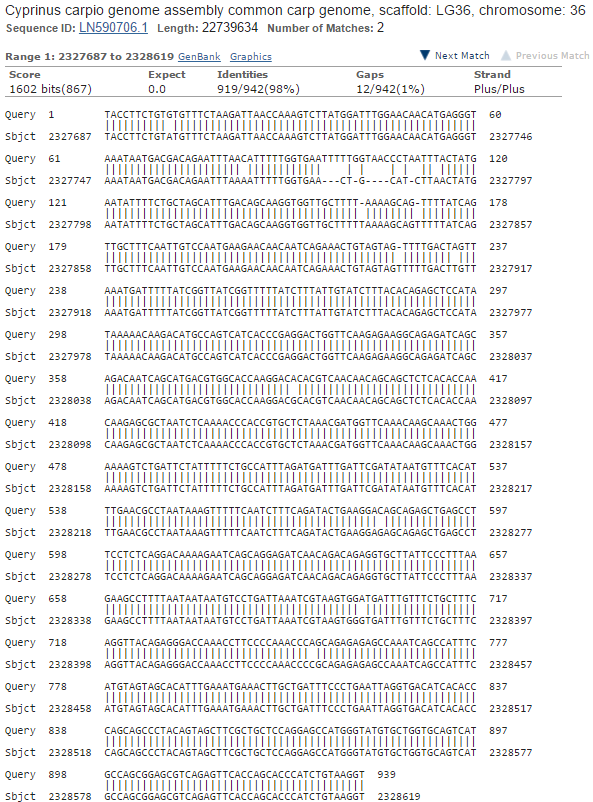
>jcf7180013793945

ACGTTCCAATAATATCATGCATCCCTAATGAATTTTCTGTCATCCTGCAGGCGGTGTGTTGTGAAGTGGATCAAAGTGAACGAGTGAGGGAGAAAATCCAAGAAGAAGTGAACAAACTCAAAGAGGAAATCGCTCTCAGGAAACGTAACAGCGCTGAGGAAGAGAGAAGTATGTACTGATGACTCTGTGCATTTAATGCTTGTTTAACAAAAAAGGAAACAACTTCTTTGAATATCTTCTCTATTTTCCTCACAGGACTCCAAGAGGCAGCCGCTGTCACGCTTCCTGATAACTCTGCCCTTCCCGACTCCTCACACTCTTCGCTGACTCGAGTCATGCTCTCAGGTCCCTCGGTGGAGCCCAGTTTAGAGAGACCCGTGAACTCGTCATCTCCACCAGTCTACAGCACCACACTGCCTCCAGATTCACCAAGCGCTATACTGCTGCCACGGCCCCAAGCCGTAGGTTCGCCCGGCCACCCGTCTCCCAGTCTGGGCCTGGGCAATGGAGACGGTTCGAGCTCGGATTCCCTGAACCGCCAGGCCACCGGCCAGGAAGACCAGGAGGATGACGATGACGAGGACAGTAGCGAGTACGAAAATTTAGTTAGCGAGCAATCGCAAGCACCTTCCTTAACAGAAACTGTTACATTAGGACGACCGGCTACGCTGTGGCCCGACG



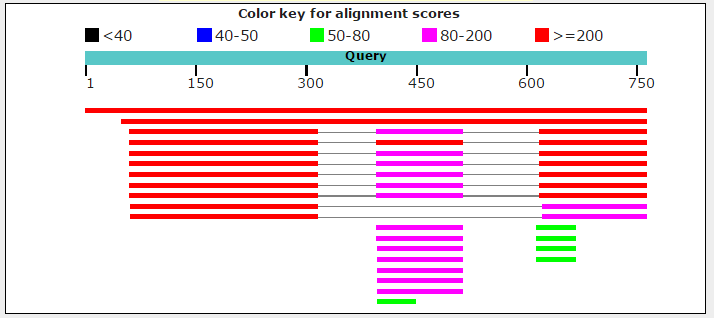
>jcf7180013814197

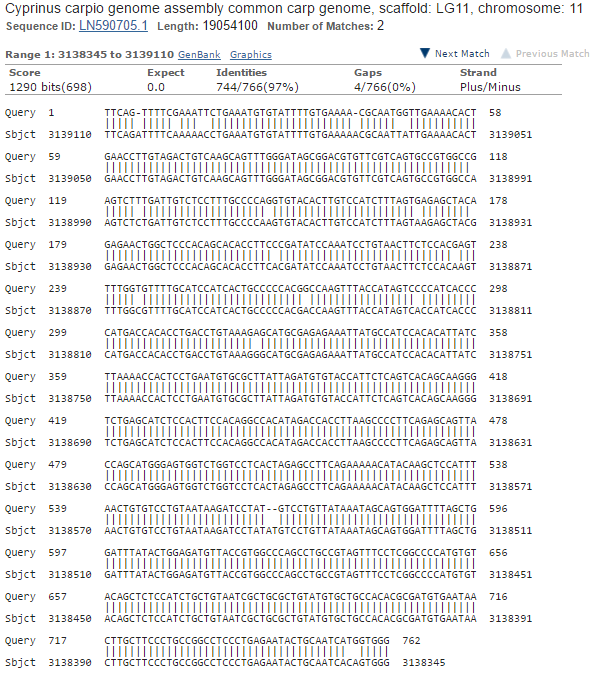
TACCTTCTGTGTGTTTCTAAGATTAACCAAAGTCTTATGGATTTGGAACAACATGAGGGTAAATAATGACGACAGAATTTAACATTTTTGGTGAATTTTTGGTAACCCTAATTTACTATGAATATTTTCTGCTAGCATTTGACAGCAAGGTGGTTGCTTTTAAAAGCAGTTTTATCAGTTGCTTTCAATTGTCCAATGAAGAACAACAATCAGAAACTGTAGTAGTTTTGACTAGTTAAATGATTTTTATCGGTTATCGGTTTTTATCTTTATTGTATCTTTACACAGAGCTCCATATAAAAACAAGACATGCCAGTCATCACCCGAGGACTGGTTCAAGAGAAGGCAGAGATCAGCAGACAATCAGCATGACGTGGCACCAAGGACACACGTCAACAACAGCAGCTCTCACACCAACAAGAGCGCTAATCTCAAAACCCACCGTGCTCTAAACGATGGTTCAAACAAGCAAACTGGAAAAGTCTGATTCTATTTTTCTGCCATTTAGATGATTTGATTCGATATAATGTTTCACATTTGAACGCCTAATAAAGTTTTTCAATCTTTCAGATACTGAAGGACAGCAGAGCTGAGCCTTCCTCTCAGGACAAAAGAATCAGCAGGAGATCAACAGACAGAGGTGCTTATTCCCTTTAAGAAGCCTTTTAATAATAATGTCCTGATTAAATCGTAAGTGGATGATTTGTTTCTGCTTTCAGGTTACAGAGGGACCAAACCTTCCCCAAACCCAGCAGAGAGAGCCAAATCAGCCATTTCATGTAGTAGCACATTTGAAATGAAACTTGCTGATTTCCCTGAATTAGGTGACATCACACCCAGCAGCCCTACAGTAGCTTCGCTGCTCCAGGAGCCATGGGTATGTGCTGGTGCAGTCATGCCAGCGGAGCGTCAGAGTTCACCAGCACCCATCTGTAAGGT



>jcf7180013849023

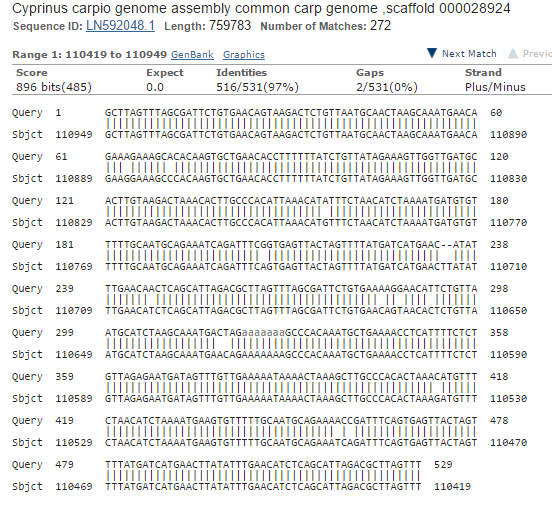
TTCAGTTTTCGAAATTCTGAAATGTGTATTTTGTGAAAACGCAATGGTTGAAAACACTGAACCTTGTAGACTGTCAAGCAGTTTGGGATAGCGGACGTGTTCGTCAGTGCCGTGGCCGAGTCTTTGATTGTCTCCTTTGCCCCAGGTGTACACTTGTCCATCTTTAGTGAGAGCTACAGAGAACTGGCTCCCACAGCACACCTTCCCGATATCCAAATCCTGTAACTTCTCCACGAGTTTTGGTGTTTTGCATCCATCACTGCCCCCACGGCCAAGTTTACCATAGTCCCCATCACCCCATGACCACACCTGACCTGTAAAGAGCATGCGAGAGAAATTATGCCATCCACACATTATCTTAAAACCACTCCTGAATGTGCGCTTATTAGATGTGTACCATTCTCAGTCACAGCAAGGGTCTGAGCATCTCCACTTCCACAGGCCACATAGACCACCTTAAGCCCCTTCAGAGCAGTTACCAGCATGGGAGTGGTCTGGTCCTCACTAGAGCCTTCAGAAAAACATACAAGCTCCATTTAACTGTGTCCTGTAATAAGATCCTATGTCCTGTTATAAATAGCAGTGGATTTTAGCTGGATTTATACTGGAGATGTTACCGTGGCCCAGCCTGCCGTAGTTTCCTCGGCCCCATGTGTACAGCTCTCCATCTGCTGTAATCGCTGCGCTGTATGTGCTGCCACACGCGATGTGAATAACTTGCTTCCCTGCCGGCCTCCCTGAGAATACTGCAATCATGGTGGG





>jcf7180013860884

GCTTAGTTTAGCGATTCTGTGAACAGTAAGACTCTGTTAATGCAACTAAGCAAATGAACAGAcAAGAAAGCACACAAGTGCTGAACACCTTTTTTATCTGTTATAGAAAGTTGGTTGATGCACTTGTAAGACTAAACACTTGCCCACATTAAACATATTTCTAACATCTAAAATGATGTGTTTTTGCAATGCAGAAATCAGATTTCGGTGAGTTACTAGTTTTATGATCATGAACATATTTGAACAACTCAGCATTAGACGCTTAGTTTAGCGATTCTGTGAAAAGGAACATTCTGTTAATGCATCTAAGCAAATGACTAGAAAAAAAGCCCACAAATGCTGAAAACCTCATTTTCTCTGTTAGAGAATGATAGTTTGTTGAAAAATAAAACTAAAGCTTGCCCACACTAAACATGTTTCTAACATCTAAAATGAAGTGTTTTTGCAATGCAGAAAACCGATTTCAGTGAGTTACTAGTTTTATGATCATGAACTTATATTTGAACATCTCAGCATTAGACGCTTAGTTT



Also I try to blast scaffolds that were below 0.95. The next restriction was, that length of scaffold should be higher than 1000 bases and identity higher than 95%. I get next 7144 scaffolds, that could be from carp genome. Some of these scaffolds are already in list. Per manually blasted results, is identity higher. After merging the ids and remove duplicates it looks, that these 7144 scaffold were not in previous list.

After this I join all IDs together, sorted them and remove duplicates and I have 582297 scaffolds that will be removed.

In Tab. 6 below are summarized information about scaffolds after filtration.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Scaffolds | Scaffolds 2 | Contigs | Contigs 2 |
| Scaffolds/contigs (>= 0 bp) | 766,162 | 766,162 | --- | -- |
| Scaffolds/contigs (>= 1000 bp) | 396,298 | 396,301 | 371,849 | 370,466 |
| Scaffolds/contigs (>= 5000 bp) | 53,487 | 53,497 | 46,135 | 46,888 |
| Scaffolds/contigs (>= 10000 bp) | 9,451 | 9,456 | 6,523 | 6,824 |
| Scaffolds/contigs (>= 25000 bp) | 187 | 187 | 107 | 119 |
| Scaffolds/contigs (>= 50000 bp) | 30 | 30 | 27 | 27 |
| Total length (>= 0 bp) | 1,372,595,688 | 1,372,644,634 | -- | --- |
| Total length (>= 1000 bp) | 1,145,341,032 | 1,145,392,589 | 1,044,509,990 | 1,048,378,403 |
| Total length (>= 5000 bp) | 425,742,422 | 425,826,718 | 351,470,293 | 358,912,590 |
| Total length (>= 10000 bp) | 128,264,684 | 128,326,442 | 86,979,923 | 91,138,132 |
| Total length (>= 25000 bp) | 7,379,749 | 7,379,776 | 4,882,386 | 5,511,265 |
| Total length (>= 50000 bp) | 2,594,627 | 2,594,627 | 2,348,668 | 2,348,668 |
| Scaffolds/contigs[[9]](#footnote-9) | 640,175 | 640,175 | 701,164 | 696,269 |
| Largest scaffold/contig | 192,299 | 192,299 | 192,299 | 192,299 |
| Total length | 1,322,576,402 | 1,322,625,348 | 1,277,549,923 | 1,278,971,366 |
| GC (%) | 35.13 | 35.13 | 35.09 | 35.10 |
| N50[[10]](#footnote-10) | 3,102 | 3,102 | 2,792 | 2,832 |
| N75 | 1,483 | 1,483 | 1,270 | 1,283 |
| L50[[11]](#footnote-11) | 113,717 | 113,711 | 123,940 | 122,283 |
| L75 | 270,346 | 270,338 | 295,057 | 291,565 |
| N’s per 100 kbp[[12]](#footnote-12) | 489.47 | 472.39 | 0.00 | 1.04 |

Tab. Information about filtrated scaffolds (scaffolds 2 and contigs 2 are after gap closer with roche 454 reads)

1. Total number of scaffolds/contigs in the assembly [↑](#footnote-ref-1)
2. The length for which the collection of all scaffolds/contigs of that length or longer covers at least half and assembly [↑](#footnote-ref-2)
3. The number of scaffolds/contigs equal to or longer than N50 [↑](#footnote-ref-3)
4. The average number of uncalled bases (N’s) per 100 000 assembly bases. [↑](#footnote-ref-4)
5. This tool was downloaded from <ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/> and installed on MetaCentrum in /storage/brno8/home/jankujova/diplomka/programs/ncbi-blast-2.6.0+-src/. [↑](#footnote-ref-5)
6. This number is same for both types – gaps allowed and gaps restricted. [↑](#footnote-ref-6)
7. grep “^jcf” [↑](#footnote-ref-7)
8. awk '{ if ($13 >= 10) print $0}' [↑](#footnote-ref-8)
9. Total number of scaffolds/contigs in the assembly [↑](#footnote-ref-9)
10. The length for which the collection of all scaffolds/contigs of that length or longer covers at least half and assembly [↑](#footnote-ref-10)
11. The number of scaffolds/contigs equal to or longer than N50 [↑](#footnote-ref-11)
12. The average number of uncalled bases (N’s) per 100 000 assembly bases. [↑](#footnote-ref-12)