The transcriptome of A.crassus: Evaluating a method of combining assemblies

Emanuel Heitlinger

1 Overview

The pre-processed *Anguillicola crassus* data-set consisting of 100491819 bases in 353055 reads (58617 generated using "FLX-chemistry", 294438 using "Titanium-chemistry") was assembled following an approach proposed by Kumar & Blaxter [1]: Two assemblies were generated, one using `newbler v2.5.3` [2], the other using `mira` [3]. The resulting assemblies (refferd to as first-order assemblies) were combinded with `Cap3` [4] into a combinded assembly (refferd to as second-order assembly).

2 The `newbler` first-order assembly

During transcriptome-assemby (with options `-cdna -urt`) `newbler` can split individual reads spanning the breakpoints of alternate isoforms, to assemble e.g. the first portion of the reads in one contig, the second portion in two different contigs. Later multipe so called isotigs would be constructed and reported, one for each putative transcript-variant. While this approach could be helpful during the detection of alternate isoforms, it also produces short contigs (especially at error-prone edges of high-coverage transcripts) when the building of isotigs fails. All reports the programm provides are including short contigs only used during the assembly-process, but not reported in the contigs-file used in transcriptome-assembly projects (`454Isotigs.fna`). Therefore to get all non-assembled reads it was necessary to add all reads appearing only in contigs not reported in the fasta-file to the reported singletons. The number of singletons increased in this step form the 26211 reported to 109052.

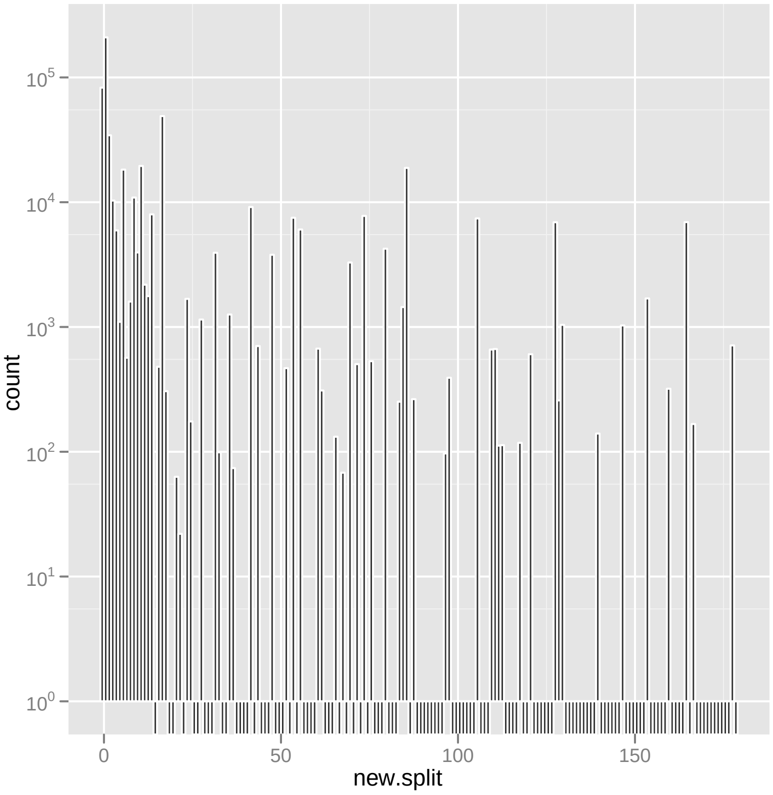

Figure 1: Number of contigs/isotigs newbler splitted one read into

As mentioned above, the splitting of reads in the `newbler` assembly can give useful information on possible isoforms. Figure 1 gives an histogram of the number of contigs/isotigs newbler splitted a single read into. This number and it's usefulness will be discussed later in greater detail.

3 The `mira`-assembly and the second-order assembly

The `mira` assembly (with options `-job=denovo,est,accurate,454`) provided a second estimate of the transcriptome. In this assembly individual reads are not split. The number of reads not used in the `mira`-assembly was 65368.

To combine the two assemblies `cap3` was used with default parameters and including the quality information from first-order assemblies. The reminder of this text deals with the exploratory analysis of how information from both estimates of the transcriptome are integrated into the final second-order assembly.

4 Evaluation of the three assemblies

Table 1 gives basic summary-statistics of the different assemblies. `Mira` clearly produced the biggest assembly (in terms of both number of bases and contigs), the second-order assembly is of comparable size than the `newbler` assembly.

	Newbler	Mira	Cap3-Second-order
Max length	6300	6352	6377
Number of contigs	15934	22596	14064
Number of Bases	8085922	12010349	8139143
N50	579	579	662
Number of congtigs in N50	4301	6749	3899
non ATGC bases	375	29962	5245
Mean length	508	532	579

Table 1: Bla blub

To further compare the three assemblies (`mira, newber` and second-order `Cap3` assembly) we evaluated the number of bases or proteins their contigs (partially) cover in the related model-nematodes, *Caenorhabditis elegans* (Ce) and *Brugia malayi* (Bm).

The `mira` assembly produced the highest per-base coverage in both species, with the second-order assembly on a close second place and newbler producing the lowest coverage (see Table 2). The same holds for the number of proteins covered at least by 100 bases from the assembly, while considering all hits newbler performed somewhat better and places between `mira` and the combined assembly.

	base coverage in %	prot. hit in %	prot. over 100	sum.CovRed
Bm:mira	19.84	35.44	26.62	1972.05
Bm:newbler	19.41	35.18	26.04	1974.13
Bm:SndO	19.78	34.49	26.23	2004.52
Ce:mira	9.06	18.77	14.30	2070.70
Ce:newbler	8.76	18.42	13.79	2086.18
Ce:SndO	9.06	18.20	14.09	2128.71

Table 2: bla

5 Data-categories in the second-order assembly

Three categories of assembled sequence data can be distinguished from the second-order assembly, each one with different reliability and purpose in downstream applications:

The first category of data obtained are the singletons of the final second-order assembly. It comprises raw sequencing reads that neither of the first-order assemblers used. It is therefore the intersecion of the `newbler`-singletons (as defined in 2) and the `mira`-singletons. 47669 reads fell in this category. This category of data is considered least relieable.

The a second category of sequence contains the first-order contigs, that could not be assembled in the second-order assembly (the singletons in the `cap3`-assembly; M_1 and N_1 in table 3). Furthermore second-order contigs in which first-order contigs from only one assembler are combined (M_n and N_n in table 3) also have to be included in this category. Sequences in this category should be considered only moderately relieable as they are supported by only one assembly algorithm.

Finally the category of contigs considered most reliable contains all second-order contigs with contribution from both first-order assemblies (MN in table 3).

For the last, most relieable category reads contained in the assembly can be categorized depending on whether they enterd the assembly via both or only via one first-order assembly.

	M_1	M_n	MN	N_n	N_1
Snd.o.con		164	13887	13	
Fst.o.con	2347	897	mira=19352/newbler=14410	40	1484
reads	42172	21153	one=269868/both=193308	1538	13100

Table 3: **Number of reads, first-order contigs (Fst.o.con) and second-order contigs (Snd.o.con) for different categories of contigs (M_1 and N_1 = first-order contigs not assembled in second-order assembly, from mira and newbler respectively; M_n and N_n = assembled in second-order contigs only with contigs from the same first-order assembly; MN = assembled in second-order contigs with first order contigs from both first order assemblies**

The numbers reported in table 3 correspond only to a single read for each read splitted in the newbler assembly. Figure 2 gives a more detailed view of the fate of the reads newbler splitted during first-order assembly. Interestingly most reads newbler splitted ended in the high-quality category of the second order assembly.

6 Contribution of first-order assemblies to second-order contigs

Looking at the contribution of contigs from each of the assemblies to one second-order contig in figure 3 a and b it becomes clear, that the `mira`-assembly had a high number of redundant contigs. These were assembled into the same contig by `newbler` and finally also in one second-order contig by `Cap3`.

A different picture emerges from the contribution of reads through each of the first-order assemblies (figure 3 c). Here for most second-order contigs many more reads are contributed through `newbler`-contigs. This is because *newbler* has more reads summed over all contigs caused by the duplication due to splitting of reads.

7 Measures on the second-order assembly and ability to predict `blast`-results from theses

We used the presence of blast results versus *Caenorhabditis elegans* and *Brugia malayi* proteomes as evidence for a contig being likely to reflect biologi-

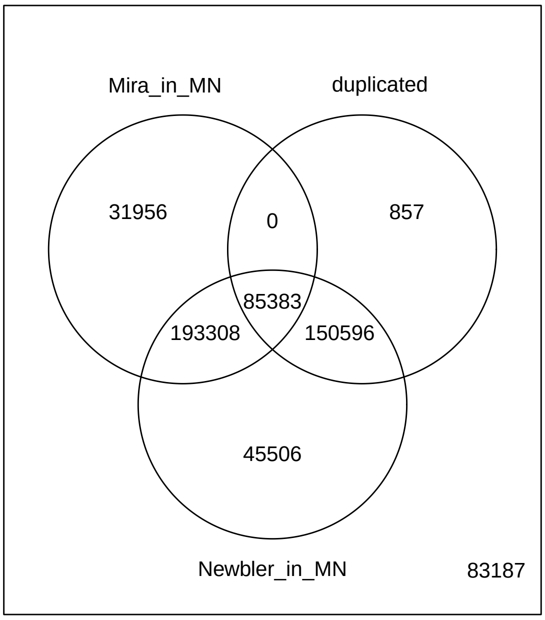

Figure 2: The way of reads into the most relieable (MN) assembly-category

cally real sequence. On this background we compared different measures of the second-order assembly for thier ability to predict `blast`-results.

We calculated the following measures for each contig in the second-order assembly:

- number of `mira` and `newbler` first-order contigs
- number of reads through `mira` and reads through `newbler`
- number of reads being split by `newbler` in first-order assembly
- maximal number of first-order contigs a read has been split in during `newbler`-assembly
- the number of reads beeing merged back into one contig from different first-order contigs
- Number of other second order contigs containing the same read (size of the "second-order cluster")

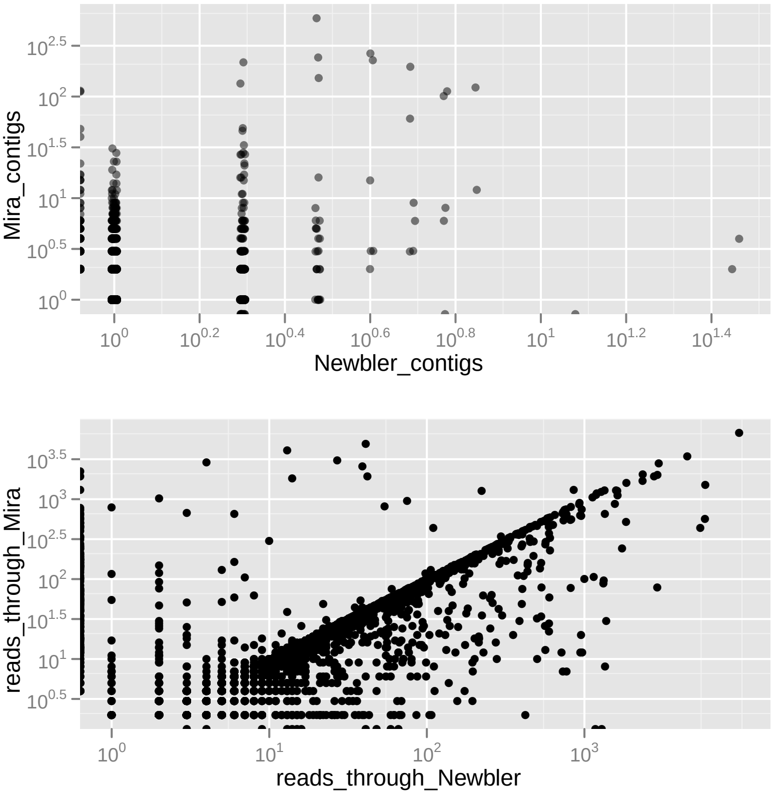

Figure 3: Number of first-order contigs from both first-order assemblies for each second order contig (a) number of reads through newbler and mira for each second-order contig (b)

We used all these meaures in generalizid linear models of the binominal family (logit link) to test how the they predict a hit. Staring from the

8 Validity of `newbler's` isoform information and representation of isoforms in second-order assembly

To test the validity of isoform information in both the `newbler` and second-order assembly of alternate isoforms was inferred for each protein in the *C. elegans* database. It was then tested whether measures the measures outlined above were able to predict the probability of blast-hits in general and to alternatively spliced genes.

References

[1] Kumar S, Blaxter ML: **Comparing de novo assemblers for 454 transcriptome data**. *BMC Genomics* 2010, **11**:571, [`http://dx.doi.org/10.1186/1471-2164-11-571`].

[2] Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Ho CH, Irzyk GP, Jando SC, Alenquer ML, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM: **Genome sequencing in microfabricated high-density picolitre reactors**. *Nature* 2005, **437**:376–380, [`http://dx.doi.org/10.1038/nature03959`].

[3] Chevreux B, Pfisterer T, Drescher B, Driesel AJ, Muller WE, Wetter T, Suhai S: **Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs**. *Genome Res.* 2004, **14**:1147–1159, [`http://www.ncbi.nlm.nih.gov/pmc/articles/PMC419793`].

[4] Huang X, Madan A: **CAP3: A DNA sequence assembly program**. *Genome Res.* 1999, **9**:868–877, [`http://www.ncbi.nlm.nih.gov/pmc/articles/PMC310812`].