Kumar and Blaxter *BMC Genomics* 2010, **11**:571
http://www.biomedcentral.com/1471-2164/11/571

BMC Genomics

RESEARCH ARTICLE Open Access

Comparing *de novo* assemblers for 454 transcriptome data

Sujai Kumar[†], Mark L Blaxter[*†]

* Correspondence: mark.blaxter@ed.ac.uk
† Contributed equally
Institute of Evolutionary Biology, University of Edinburgh, West Mains Road, Edinburgh EH9 3JT, UK

Age Tats
JClub, 30th May 2011

Transcriptome sequencing

- Traditionally Sanger
- For organisms with genomic sequence available → Illumina SOLEXA and ABI SOLiD
- Non-model organisms → Roche 454 (longer reads for assembly and annotation)

Complications in assembly

- Errors and polymorphisms in individual reads complicate recognition of overlaps
- Individual transcripts in non-normalised data can have several orders of magnitude variation in abundance

Table 1 Assemblers previously used for *de novo* assembly of 454 pyrosequencing transcriptome projects

Assembler	Organism
Newbler	*Arabidopsis thaliana* [1,2]; *Eucalyptus grandis* [3]; *Castanea dentata* and *Castanea mollissima* [4]; *Sarcophaga crassipalpis* [5]; *Acropora millepora* [6]; *Palomero toluqueño* [7]; *Eschscholzia californica* and *Persea americana* [2]; *Vitis vinifera* [8]; *Rhagoletis pomonella* [9]; *Heliconius spp.* [10]; *Euphydryas aurinia, Manduca sexta, Chrysomela tremulae, Papilio dardanus, Heliconius melpomene, Heliconius erato,* and *Melitaea cinxia* [11]; *Panax quinquefolius* [12]; *Sclerotium rolfsii* [13]; *Laternula elliptica* [14]
CAP3	*Zea mays* [15,16]; *Arabidopsis thaliana* [1]; *Ambystoma mexicanum* [17]; Human breast cancer [18]; *Artemisia annua* [19]; *Solanum arcanum* [20]; *Epimedium sagittatum* [21]; *Haemonchus contortus* [22]; *Laodelphax striatellus* [23]; *Coleochaete orbicularis* and *Spirogyra pratensis* [24]; *Bugula neritina* [25];
MIRA	*Centaurea solstitialis* [26]; *Chrysomela tremulae* [27,11]; *Pandinus imperator* [28]; *Zygaena filipendulae* [29]; *Manduca sexta* [30,11]; *Euphydryas aurinia, Papilio dardanus, Heliconius melpomene, Heliconius erato,* and *Melitaea cinxia* [11];
TGICL	*Pythium ultimum* [31]; *Zoarces viviparous* [32]; *Medicago truncatula* [33]
SeqMan	*Melitaea cinxia* [34]; *Cochliomyia hominivorax* [35]; *Pinus contorta* [36]
stackPACK	*Arabidopsis thaliana* [1]

Table 2 Features of assembly programmes compared in this study

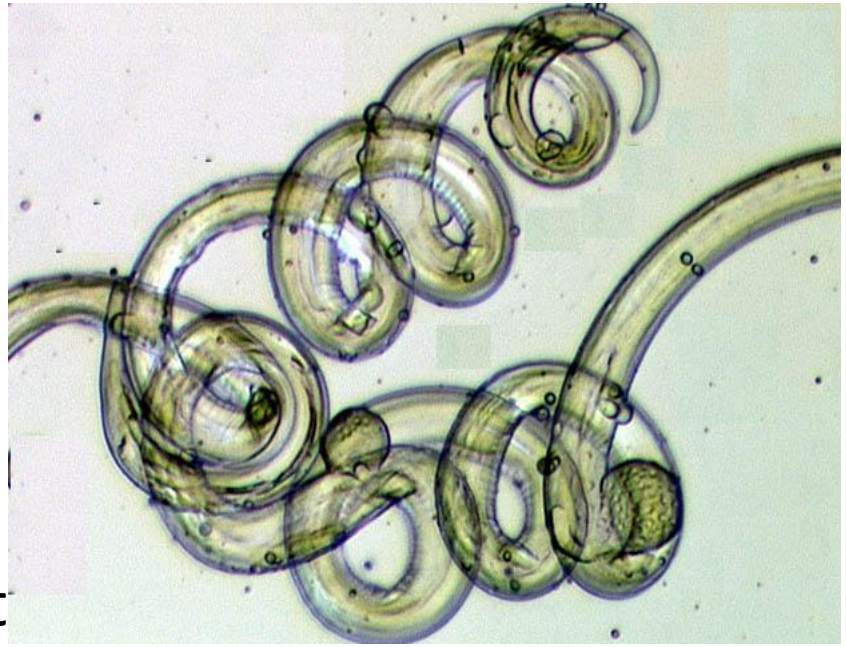
Assembler	Type†	Splits reads*	Author	Cost	Source available	URL
CAP3	OLC†	No	X Huang and A Madan [38]	Free for use at non-profit organizations	No	http://seq.cs.iastate.edu/
CLC Assembly Cell 3.0	de Bruijn graph	Yes	CLC	Request quote or trial license	No	http://www.clcbio.com/
MIRA 3.0	OLC	No	Bastien Chevreux [40]	Free	Yes, GPL	http://sourceforge.net/projects/mira-assembler/
Newbler 2.3 and Newbler 2.5	OLC	Yes	Roche 454 [37]	Free for academic use	No	http://454.com/products-solutions/analysis-tools/gs-de-novo-assembler.asp
SeqMan NGen 2.1	OLC	No	DNAStar [41]	Request quote or trial license	No	http://www.dnastar.com/t-products-seqman-ngen.aspx

* i.e. data from one read can appear in multiple contigs.

† OLC: Overlap-Layout-Consensus.

Target organism

- *Litomosoides sigmodontis*
  - model filarial nematode
  - Originally derives from cotton rat hosts, but can be maintained in laboratory rodents
  - System for investigating the dynamics of immune response induction and modulation, test vaccine and drug candidates



Expected number of transcripts ~30 000 with mean length ~1.2 kb

Table 3 The *Litomosoides sigmodontis* transcriptome dataset read statistics

L. sigmodontis lifecycle stage	Technology	Number of reads	Number of raw bases	Number of trimmed reads	Number of trimmed bases	Mean length of trimmed reads	Median length of trimmed reads
Microfilaria (first stage larvae)	Titanium	366,813	203,227,223	351,387	118,039,337	335.92	374
Adult female	Standard	180,271	48,434,306	176,454	38,352,888	217.35	236
Adult male	Standard	216,940	59,231,575	213,546	48,673,441	227.93	245
Total	**Titanium + Standard**	**764,024**	**310,893,104**	**741,387**	**205,065,666**	**276.60**	**257**

Default parameters recommended for transcriptome assembly.

The optimal assembler (1/2)

- Uses all the reads given.
- Unambiguous mappings of reads to contigs.
- The longest summed length of contigs.
- Avoids over-assembly of reads into *in silico* chimaeras.
- Avoids the production of near-identical, largely overlapping contigs from allelic copies or error-rich data.

The optimal assembler (2/2)

- Produces a transcriptome estimate with a mean and variance in contig length similar to that expected from the whole transcriptome.
- Completes analysis in a short period of time.
- In comparison to other assemblies, it includes the largest proportion of the unique bases present in the sum of all assemblies.
- Returns contigs that match well previously determined sequences for the target species.
- Delivers a high coverage of the conserved proteome of related taxa.

	CAP3	CLC	MIRA	Newbler 2.3	Newbler 2.5	SeqMan
Number of contigs (>100 bp)	24727	22746	35827	12019	21734	29969
Total Bases	16733217	14875522	21339704	14456476	20066883	21355682
Number of contigs (>= 1 kbp)	4403	4174	4770	6320	7661	6082
Total bases (in contigs >= 1 kbp)	6461079	6255785	7027775	10810962	13691429	9296011
Max contig length	4011	4368	5784	5872	6228	6263
Mean contig length	677	654	596	1203	923	713
N50	806	850	708	1487	1448	880
Number of contigs in N50	6533	5459	9148	3406	4649	7555

	CAP3	CLC	MIRA	Newbler 2.3	Newbler 2.5	SeqMan
Reads used (SSAHA2)	670425	679152	672036	616672	667597	681974
Multi-hit reads (SSAHA2)	271648	118334	392884	249210	352887	322409
Reads used (CLC)	690889	691818	696527	600132	681831	711726
Multi-hit reads (CLC)	91951	24485	162365	213670	262178	128631
Time taken	1 day*	4 minutes*	3 days*	2 hours*	45 minutes*	6 hours**

* on a dual quad-core 3 GHz Xeon workstation with 32 GB RAM
** on a dual core 2.53 GHz Mac mini server with 4 GB RAM

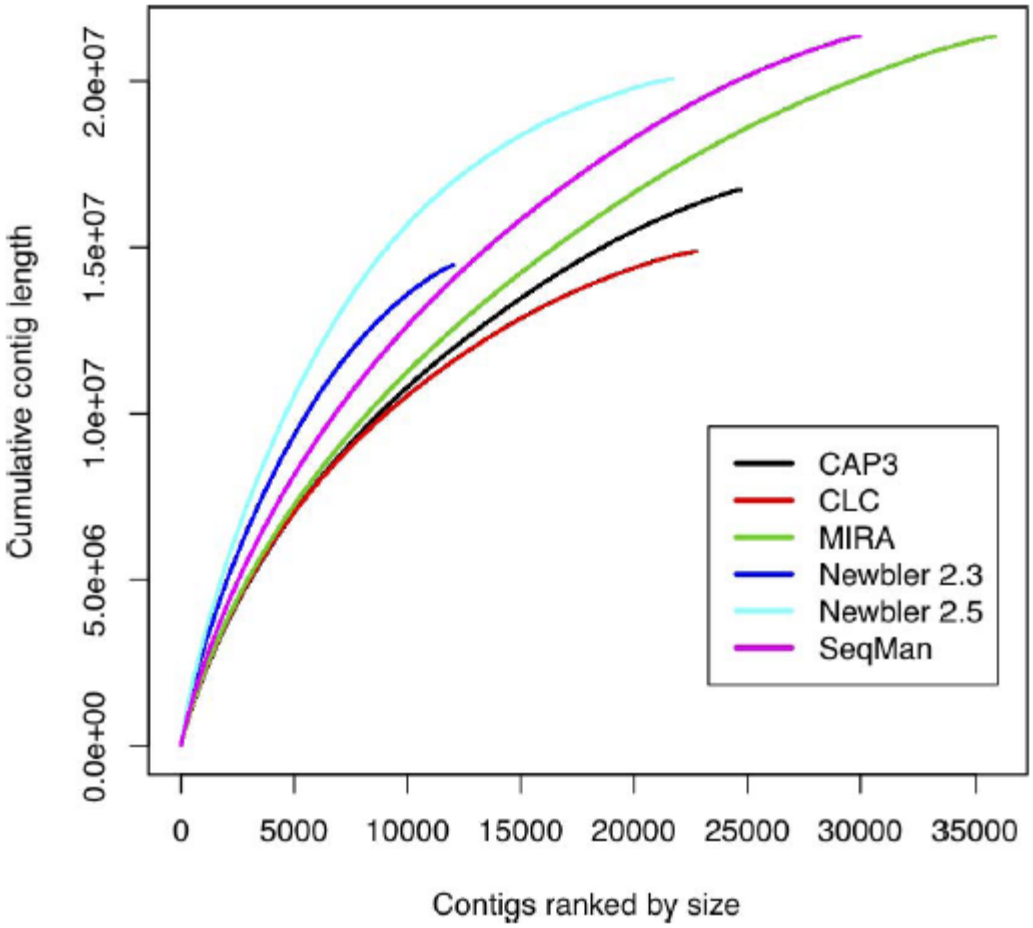


Figure 1 Cumulative contig lengths generated by different assembly programs. For each of six assemblies, contigs longer than 100 bases were ordered by length, and the cumulative length of all contigs shorter than or equal to a given contig was plotted. The total length of the assembly and the number of contigs present in the assembly define the end point of each curve, while the initial slope of each curve reflects the proportion of longer contigs.

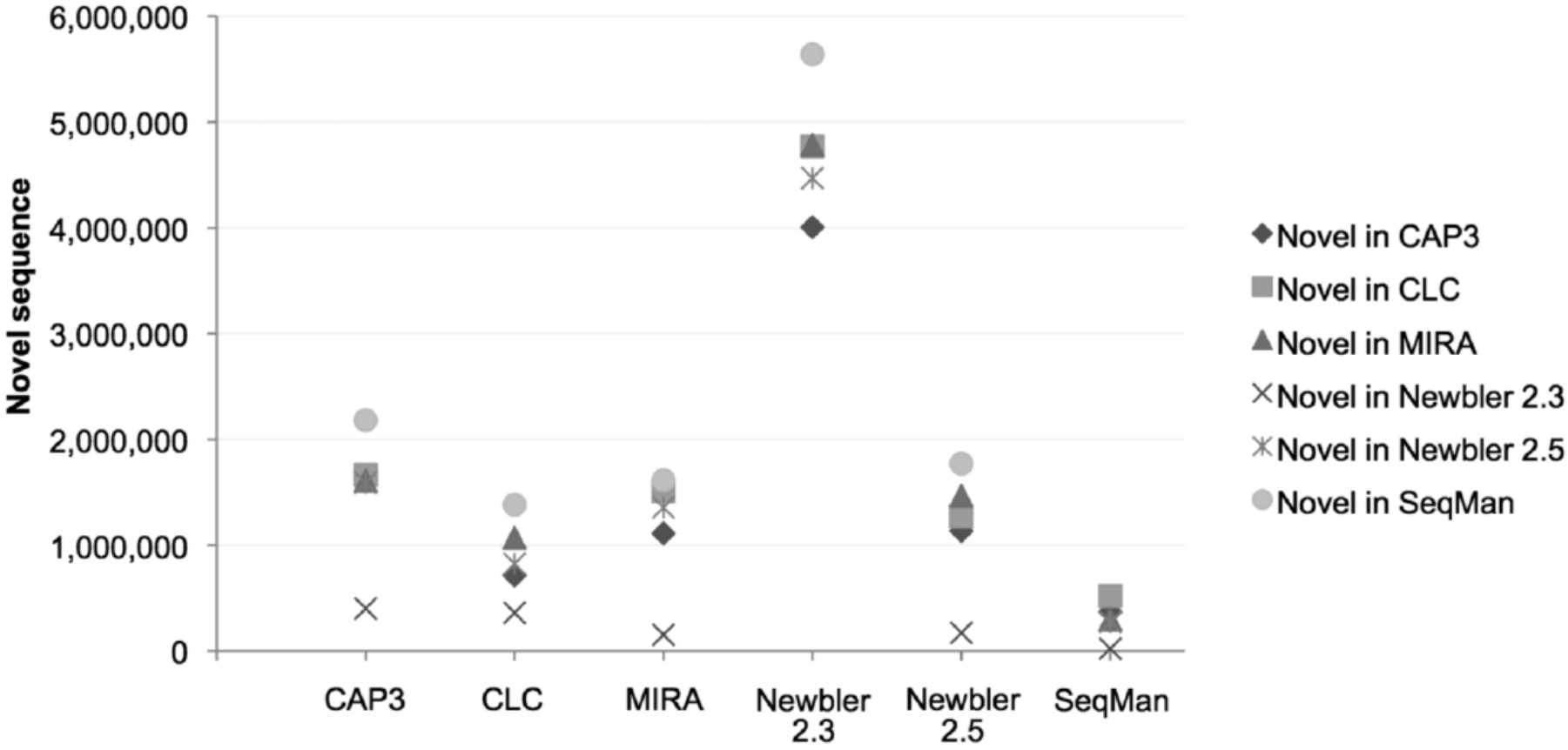


Figure 2 Novel sequence in pair-wise comparisons between assemblies produced by different assemblers. For each assembly, we calculated the number of bases in the other assemblies that were not present in the focal assembly.

Alignments to reference sequence

Table 5 BLAT hits to 1602 *Litomosoides sigmodontis* EST clusters

	CAP3	CLC	MIRA	Newbler 2.3	Newbler 2.5	SeqMan
% of ESTs hit by BLAT	87.8	90.3	89.6	81.6*	89.6	90.8
(% of bases covered)	(78.2)	(80.0)	(80.1)	(71.3)*	(78.7)	(82.0)
% of ESTs hit by BLAT where each hit covered at least 80% of the target EST sequence	59.9	51.9	62.5	59.4	63.9	65.4
(% of bases covered)	(59.1)	(50.5)*	(61.1)	(59.4)	(63.7)	(64.3)

* indicates a value significantly lower than the others, using a Huber M-estimator.

Table 6 BLASTX hits to 11,472 *Brugia malayi* proteins

	CAP3	CLC	MIRA	Newbler 2.3	Newbler 2.5	SeqMan
% of proteins hit	76.7	78.4	77.3	68.9*	77.9	78.6
(% of bases covered)	(60.4)	(62.4)	(59.7)	(51.8)*	(61.5)	(63.0)
% of protein hit by individual HSPs that cover 80% of target protein	27.0	26.0	26.5	29.2	32.4	28.7
(% of bases covered)	(16.8)	(16.1)	(16.4)	(19.9)	(22.3)	(18.0)

Note: E-value cutoff 1e-5.

* indicates a value significantly lower than the others ($p < 0.01$), using a Huber M-estimator.

Alignments to reference sequence

Table 7 BLASTX hits to 3,681 tribes containing 120,926 conserved nematode proteins

	CAP3	CLC	MIRA	Newbler 2.3	Newbler 2.5	SeqMan
% of unique tribes hit	91.7	92.2	91.4	87.0*	92.0	92.4
% of unique tribes hit where individual HSPs covered 80% of target protein	81.7	81.1	78.9	77.1*	82.1	81.6

Note: E-value cutoff 1e-5.

* indicates a value significantly lower than the others ($p < 0.01$), using a Huber M-estimator.

Table 8 BLASTX hits to 3,731 KOGs containing 9,782 *C. elegans* proteins

	CAP3	CLC	MIRA	Newbler 2.3	Newbler 2.5	SeqMan
% of unique KOGs hit	89.2	89.7	88.6	83.7*	89.7	90.1
% of unique KOGs hit where individual HSPs covered 80% of target protein	30.2	27.1	28.6	33.0	35.7	30.4

Note: E-value cutoff 1e-5.

* indicates a value significantly lower than the others ($p < 0.01$), using a Huber M-estimator.

Merging assemblies to improve credibility

Table 9 Secondary assemblies by merging pairs of initial assemblies using CAP3 with default settings

Assembly 1	Assembly 2	Number of "Reads" (contigs) in Assembly 1	Bases in Assembly 1	Number of "Reads" (contigs) in Assembly 2	Bases in Assembly 2	Number of second-order contigs with "reads" from both assemblies	Bases in second-order contigs with "reads" from both assemblies
MIRA	SeqMan	35827	21339704	29969	21355682	18068	16293192
MIRA	Newbler 2.5	35827	21339704	21734	20066883	15951	15866051
Newbler 2.5	SeqMan	21734	20066883	29969	21355682	15783	15701053
CLC	Newbler 2.5	22746	14875522	21734	20066883	15778	15825663
CAP3	MIRA	24727	16733217	35827	21339704	15688	14243534
CLC	SeqMan	22746	14875522	29969	21355682	15504	14679975
CAP3	SeqMan	24727	16733217	29969	21355682	15387	14824287
CLC	MIRA	22746	14875522	35827	21339704	15334	14357031
CAP3	Newbler 2.5	24727	16733217	21734	20066883	14275	14830304
CAP3	CLC	24727	16733217	22746	14875522	14149	13753398
Newbler 2.3	Newbler 2.5	12019	14456476	21734	20066883	9733	13252303
MIRA	Newbler 2.3	35827	21339704	12019	14456476	9380	11731374
CLC	Newbler 2.3	22746	14875522	12019	14456476	8884	12318589
CAP3	Newbler 2.3	24727	16733217	12019	14456476	8484	11426423
Newbler 2.3	SeqMan	12019	14456476	29969	21355682	8274	11452990

Table 10 Alignments to 1602 EST clusters where > 80% of the EST was covered by a match, by pairs of assemblies merged using an OLC assembler

Assembly pair		% EST clusters hit	% EST bases covered
CLC	MIRA	65.4	65.1
MIRA	Newbler 2.5	65.4	65.3
MIRA	SeqMan	64.9	64.6
CLC	Newbler 2.5	64.7	64.6
Newbler 2.5	SeqMan	63.2	62.9
CAP3	Newbler 2.5	63.0	62.6
CAP3	CLC	62.9	62.4
CLC	SeqMan	62.9	62.5
CLC	Newbler 2.3	62.4	62.5
CAP3	MIRA	62.2	61.9
CAP3	SeqMan	61.7	61.3
MIRA	Newbler 2.3	61.7	61.9
Newbler 2.3	Newbler 2.5	61.4	61.5
Newbler 2.3	SeqMan	60.0	60.2
CAP3	Newbler 2.3	59.7	59.6

	+	-
CLC	*Fast *Least amount of memory *Least redundant contigs	*No read tracking *Only one user definable parameter *Lower quality in longer contigs *Cannot create alternative transcripts
MIRA	*Highest number of user-definable parameters	*Slow *Sequencing errors interfere heavily with coverage > 80x (recommended for normalized datasets) *Redundant contigs *Shorter contigs
Newbler	*Designed to deal with transcriptome data *Creates isotigs (=alternative transcripts) *Best alignments with reference ESTs *Highest number of contigs > 1kbp	*Version 2.3 -> redo with 2.5! *Redundant contigs
SeqMan	*Fast *Largest assembly *Most reads remappable to assembly *Best alignments with reference ESTs	*Shorter (and more) contigs