Marine Genomics
August 2019, Volume 46 Pages 58-61
https://doi.org/10.1016/j.margen.2018.12.001
https://archimer.ifremer.fr/doc/00615/72737/

De novo transcriptome assembly for *Tracheliastes polycolpus*, an invasive ectoparasite of freshwater fish in western Europe

Mathieu-Begne Eglantine [1, 2, 3, *], Loot Geraldine [1], Blanchet Simon [1, 2], Toulza Eve [3],
Genthon Clemence [4], Rey Olivier [3]

[1] Univ Paul Sabatier, CNRS, Ecole Natl Format Agron, Lab Evolut & Diversite Biol,UMR5174, 118 Route Narbonne, F-31062 Toulouse, France.
[2] Univ Paul Sabatier, CNRS, Stn Ecol Ther & Expt, UMR5321, 2 Route CNRS, F-09200 Moulis, France.
[3] Univ Perpignan Via Domitia, Lab Interact Hote Parasite Environm, UMR5244, 58 Ave Paul Alduy, F-66860 Perpignan, France.
[4] INRA, US 1426, GeT PlaGe, Genotoul, Castanet Tolosan, France.

* Corresponding author : Eglantine Mathieu-Bégné, email address : eglantine.mb@gmail.com

geraldine.loot@univ-tlse3.fr ; simon.blanchet@sete.cnrs.fr ; eve.toulza@univ-perp.fr ; clemence.genthon@inra.fr

Abstract :

Tracheliastes polycolpus is an ectoparasitic copepod that recently emerged in western Europe and that infects several freshwater fish species. Its recent successful spread might be due to its fascinating ability to shift to new host populations and/or species. Here, we present the first non-redundant and protein-coding de novo transcriptome assembly for T. polycolpus along with a quality assessment and reliable transcript annotations. This assembly was built from fifteen adult female parasites sampled from three different host species from a single river in southwestern France. Overall, 17,157 non-redundant contigs likely corresponding to protein-coding transcripts were identified, of which 13,093 (i.e., 76%) were successfully annotated. This assembly displayed good representativeness since 65.8% of the raw reads properly aligned back to the assembly. Similarly, 90.5% of the single copies of orthologues conserved across arthropods were retrieved in this assembly, which reflects a very good completeness. Finally, this transcriptome assembly gathered 7979 T. polycolpus specific transcripts when compared with the two closest referenced species (Lepeophtheirus salmonis and Caligus rogercresseyi), thus constituting an original genomic resource. This high-quality transcriptome is an important genomic resource for deciphering the molecular bases underlying host shifts in generalist parasites, and for studying the evolutionary biology of copepods that are major biological components of both freshwater and marine ecosystems.

Keywords : Copepods, Monoxenous parasite, Parasite specificity, Curation

1. Introduction

Tracheliastes polycolpus (Copepoda, Neocopepoda, Podoplea, Siphonostomatoida, Lerneaopodidae) is a monoxene ectoparasitic copepod of freshwater fish [1]. Only adult females have a parasitic lifestyle, whereas males are microscopic and free-living organisms [2]. Adult females anchor on fish fins and feed on their epidermal cells. This causes partial to total fin degradation, which generally favours secondary infections by bacteria or/and viruses [1], and ultimately reduces the fitness of their hosts [3,4].

In the 1920s, *T. polycolpus* was introduced from eastern Eurasia to western Europe through fish trades, and has since spread over several watersheds in England, France and Spain [5]. In France, *T. polycolpus* is primarily associated with the common dace and the rostrum dace (*Leuciscus leuciscus* and *L. burdigalensis*), but is also commonly observed on several alternative cyprinid species [6]. The recent invasion history of *T. polycolpus*, together with its switch to new alternative host species, constitutes an unprecedented opportunity to study the mechanisms underlying parasite specificity and the rapid adaptive processes associated with host shifts [6–8]. Developing“-omics” resources is a key step towards a better understanding of the molecular bases of such processes, notably the role of plasticity and/or selection associated with host shift based on gene expression profiles [8,9].

Here, we built a curated *de novo* transcriptome assembly for this species. More specifically, we present the first non-redundant and protein-coding transcriptome assembly together with functional annotations for *T. polycolpus* obtained from fifteen adult female parasites sampled on three different host species. Only a few assemblies exist for copepods and notably for parasitic copepods, although some have led to novel discoveries (including mechanisms for toxin resistance, gene expression patterns associated with molting or development) [10–12]. Consequently, this database constitutes an important genomic resource for an emerging parasite in wild populations and for the diverse group of Crustacean copepods.

2. Data description

2.1. Sampling

Fifteen adult females of *T. polycolpus* were sampled on three different fish host species at a single location in the Salat River in southwestern France on a single day to limit confounding environmental effects (Table 1). The three host species were the rostrum dace (*L. burdigalensis*) and two alternative hosts; the Occitan gudgeon (*Gobio occitaniae*) and the European minnow (*Phoxinus phoxinus*). Fish were caught by electric fishing using a DEKA® 7000, anaesthetized according to a standardized protocol, and inspected for the presence of parasites. Parasites were collected using sterile forceps as follows: five parasites were sampled on five dace, five parasites were sampled on four gudgeons, and five parasites were sampled on four minnows. The fifteen parasites were immediately stored in RNAlater for 24 hours and then conditioned at -80°C (to stabilize and protect cellular RNAs) until RNA extraction.

Table 1: MIxS specifications, Assembly description and Annotation statistics for *T. polycolpus de novo* transcriptome assembly.

MixS descriptors	
Investigation type	Eukaryote
Project name	*Tracheliastes polycolpus* Transcriptome
Organism	*Tracheliastes polycolpus*
Geographical location name	The Salat River
Geographical location	43°04'43.0"N; 0°57'29.0"E
Collection date	11/07/2013
Environment (biome)	Aquatic biome
Environment (feature)	River
Environment (material)	Freshwater
Host-associated	*Leuciscus budigalensis, Gobio occitaniae* and *Phoxinus phoxinus*
Tissue type	Trunk
Developmental stage	Adult
Sex	Female
Assembly description	
Sequencing method	Illumina HiSeq 2000
Assembly	Trinity (version 2014-07-17)
Coding-protein transcript selection	TransDecoder
Assembly redundancy reduction	CD-HIT-EST (version 4.6)
Transcript number	17 157
Total assembled bases	20 355 465
N25	2 676
N50	1 599
N75	954
Average contig length	1 186
Longest contig length	23 889
% GC	38.55
Annotation (total)	
Predicted ORF	17 157
Complete protein (including start and stop codons)	10 112
Swiss-Prot top BLASTP hit	12 343
Gene ontology in blastp	12 098
Pfam	11 751
KEGG	10 924
eggNOG	10 400
Gene ontology in Pfam	7 641

2.2. Sequencing

Total RNA from each parasite was extracted using the RNeasy Plus Mini Kit (Qiagen) following the manufacturer's instructions from the parasite trunk only—to minimize possible contamination with host RNA—with a final elution volume of 40 µL RNAse-free water. The quantity and quality of RNA extractions were assessed using a nanodrop ND-8000 (Thermo Scientific) and a BioAnalyser (Agilent Technologies), respectively. The individual RNA-seq libraries were prepared on a Tecan EVO2000 using the IlluminaTruSeq Stranded mRNA protocol. The individual libraries obtained were multiplexed then pooled and sequenced on two lanes of an Illumina HiSeq 2000 (High Throughput mode) using a paired-end read length of 2x100 bp. The reads were demultiplexed, those that did not pass the chastity filter (i.e., internal filtering procedure from Illumina sequencers) were filtered out (about 7% of the total reads), and the adapters were trimmed automatically at this stage. The sequencing resulted in approximately 420 million 2 x 100 bp paired-end reads, with an average of 28 million paired-end reads per sample. The library preparation, sequencing and pre-processing automatic filtering steps were performed at the GeT-PlaGe core facility (Toulouse, France).

2.3. De novo transcriptome assembly and curation

Based on the reads obtained from all individuals, we first assembled a raw *de novo* transcriptome using the Trinity pipeline (version 2014-07-17) [13]. Reads were trimmed using Trimmomatic [14] with default parameters to discard low-quality and/or poorly informative reads. Then, each sample was parsed separately as Trinity input, and the option "--SS_lib_type" was selected to keep strand specificity information. Other parameters were set to the default values. At this step the assembly contained a total of 101 636 contigs.

To curate the raw transcriptome assembly from non-protein-coding contigs, we discarded contigs lacking an open reading frame (ORF) or having an ORF shorter than 100 amino acids using the Trinity plugin TransDecoder [13]. This filtering step resulted in 33 984 contigs enriched with only messenger RNAs (putative protein-coding genes). Finally, to reduce redundancy, we ran CD-HIT-EST (Version 4.6, [15]) to cluster the resulting transcripts according to their similarity with a minimum sequence identity set to 95%, (options -c 0.95 –n 8, where c and n are "similarity threshold" and "word size" parameters respectively). These filtering steps resulted in a final protein-coding and non-redundant assembly containing17 157 transcripts corresponding to as many putative unique protein-coding genes.

2.4. Annotation

Transcriptome functional annotation was performed using the software suite Trinotate (version 3.1.0, [13]). Translated sequences from the final assembly were used to perform a blastp search on the Swiss-Prot database and a protein domain search on the Pfam database (both uploaded from Trinotate version 3.1.10 on 07/03/2016). Results were then integrated in Trinotate to retrieve functional annotations leveraging eggNOG, GO and Kegg databases [16–18]. Only highly significant matches were reported in the annotation file (i.e., those with an e-value < 10^{-5} for the blast hits and the domain noise cut-off for the Pfam domains). Overall, 13093 contigs were annotated from at least one database. Details about the number of contigs annotated on each database are given in Table 1. Transcript nucleotide and amino acid sequences are also provided in the annotation file (Supplementary S1).

2.5. Data validation and quality control

Quality of our final transcriptome assembly was evaluated based on three main criterions (Table 2 and Fig. 1, [13,19]). 1. The representativeness criterion refers to raw read content represented by the assembly. 2. The completeness criterion is based on evolutionary expectations of near-universal gene content that the assembly includes. 3. Specificity refers to the number of genes that are specific to *T. polycolpus* when compared to closely related species.

2.5.1. Representativeness

Representativeness was evaluated by measuring the percentage of quality filtered reads that were properly re-aligned against the assembly. To do so, the software Bowtie 2 (version 2.1.0, [20]) was used with default "end-to-end" alignment parameters. We estimated representativeness as the percentage of paired reads that aligned back to the assembly exactly once and with respect to the forward and reverse directions [20]. We also reported the percentage of reads that mapped back more than once (as an indication of redundancy or repeated sequences) and the percentage of reads that did not map back to the final assembly (i.e., the proportion of non-represented reads within the assembly).

2.5.2. Completeness

Completeness was assessed by counting the percentage of orthologues conserved across arthropods that are present in the assembly using the software BUSCO [21]. However, if more than one transcript was found to be aligned with the same orthologous gene, then the assembly was therefore considered partially redundant. Thus, to account for both the number of orthologues and redundancy, we quantified the number of orthologues conserved across arthropods that were of a single copy within the assembly (i.e., BUSCOs: "Benchmarking Universal Single-Copy Orthologs", [21]).

2.5.3. Specificity

Specificity was assessed by quantifying the number of non-homologous transcripts obtained for *T. polycolpus* when compared to *Lepeophtheirus salmonis* and *Caligus rogercresseyi,* two parasitic copepod species sharing the same order with *T. polycolpus* and for which transcriptomic resources exist (available on Genbank under Bioproject numbers PRJEB1804 and PRJNA234316 respectively). These transcriptomes were obtained with comparable technologies to those we used in this study and from different development stages including adult females [11,12], which make the comparison relevant. We expect to find common orthologs between the three species, but also a large amount of development stage specific genes in the two sea lice species. We searched for reciprocal blast hit (RBH) between the assembly of *T. polycolpus* (after *in silico* RNA translation) and that of each of the two copepods species using “tblastx”. The e-value threshold was set to 0.001 in order to select the most significant matches [22], and transcripts not having a reciprocal match according to this criteria were considered specific to *T. polycolpus*.

2.5.4. Quality control results & discussion

Regarding representativeness, we obtained a mapping rate of 65.8%, which is lower than expected for a Trinity assemblage (i.e., approximately 90% [23]). Importantly, a large proportion of raw reads (27.3%) could not be mapped back to the non-redundant and protein-coding assembly (Table 2) likely because some unique contigs were also lost during the filtering steps. Given that the raw assembly size was reduced by almost a factor of six in the non-redundant and protein-coding assembly, this assembly still displays a good compromise between its informative content (representativeness) and its complexity reduction. Indeed, when limiting redundancy and removing chimaera contigs, usually up to 47% of reference genes are retrieved in *de novo* assemblies conducted with several assemblers [24]. Furthermore, high completeness was achieved with overall 90.5% of BUSCOs retrieved in the assembly and only 2.6% of conserved orthologues across arthropods having being missed (Table 2). Interestingly, the BUSCO analysis also confirmed that we significantly limited redundancy during the curation process, as 37.7% of BUSCOs were duplicated in the raw assembly and only 5.7% of duplicated BUSCOs were found in the final assembly (Table 2). Finally, we found 6 910 potentially specific transcripts in our *T. polycolpus* assembly compared with the two most closely related species currently available (Fig. 1). This high number of specific transcripts might be partly explained by the fact that the three species are from the same order but different families, except for *L. salmonis* and *C. rogercresseyi* that share the same family, which may explain their higher number of shared genes (see also Supplementary Fig. S2 for similarity histograms of the shared genes between the three species). Overall, this highlights that this assembly constitutes an original specific resource of interest. Our quality assessment thus clearly revealed that the resulting non-redundant and protein-coding assembly fulfilled the three criteria we evaluated (i.e., representativeness, completeness and specificity). As such, we propose this non-redundant and protein-coding transcriptome assembly as a reference transcriptome assembly for *T. polycolpus*.

Table 2: Quality scores for the raw assembly and the final assembly of *Tracheliastes polycopus*.

		Raw Assembly[a]	Final Assembly[b]
Representativeness	Re-mapped reads (1 time)	71.00%	65.77%
	Re-mapped reads (>1 times)	16.40%	6.89%
	Re-mapped reads (0 time)	12.60%	27.34%
Completeness	Single-copy BUSCOs	59.10%	90.50%
	Duplicated BUSCOs	37.70%	5.70%
	Missing BUSCOs	2.20%	2.60%
	Fragmented BUSCOs	1.00%	1.20%
	Total BUSCOs groups searched	1 066	1 066

[a] *Tracheliastes polycolpus* assembly generated from Trinity only
[b] *Tracheliastes polycolpus* reference assembly, resulting from curation steps

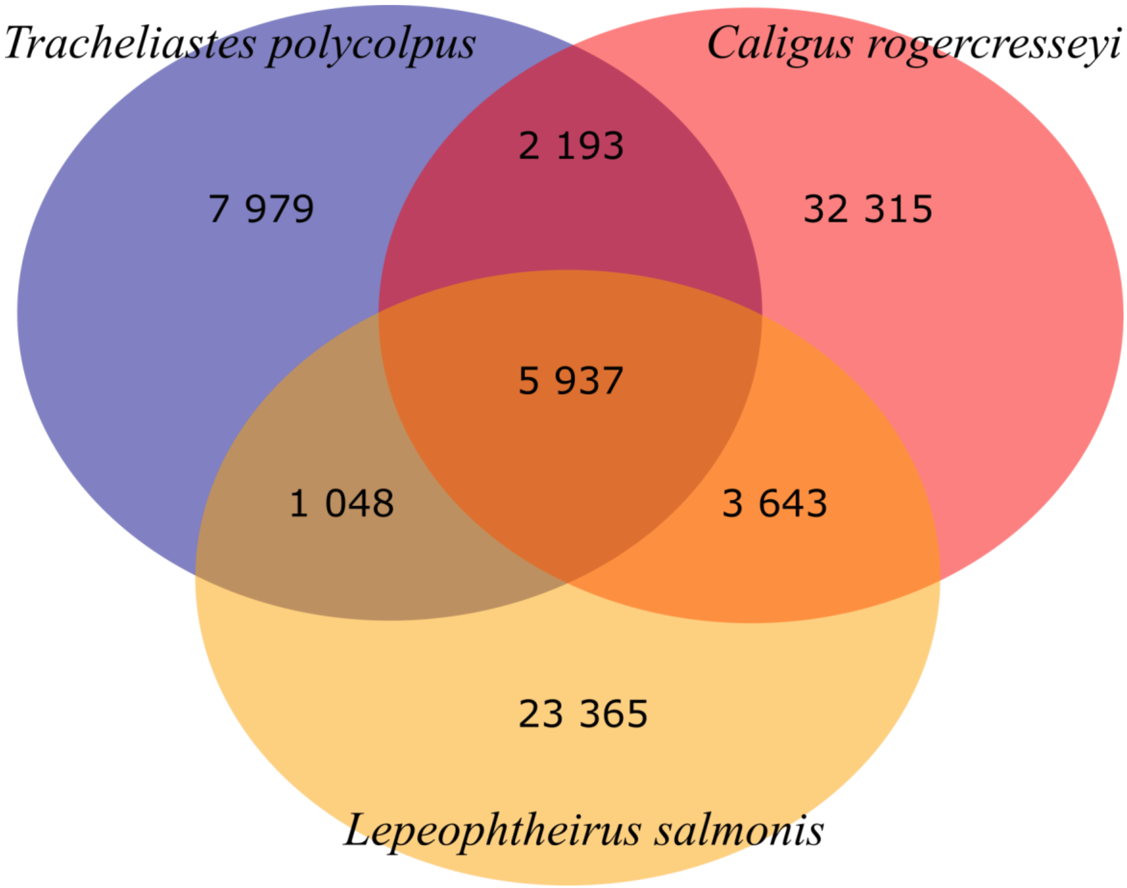

Figure 1: Venn plot showing the number of specific and shared genes between *T. polycolpus, C. rogercresseyi* and *L. salmonis,* resulting from the RBH analysis.

3. Conclusion

The transcriptome assembly reconstructed in this study is the most complete genomic/transcriptomic resource available for *T.polycolpus*, and more generally for freshwater copepods [25]. As such, this paves the way for future studies on this species, notably about the molecular bases of host shift by this parasite. Beyond the studies on this targeted species, this transcriptome will also open new avenues for studying freshwater copepods that constitute cornerstone elements for the biological functioning of freshwater ecosystems.

4. Availability of supporting data

Raw reads are deposited in Sequence Read Archive (GenBank) under project number PRJNA476682. Accession numbers are SRR7411022, SRR7411021, SRR7411024, SRR7411023, SRR7411026, SRR7411025, SRR7411028, SRR7411027, SRR7411030, SRR7411029, SRR7411019, SRR7411018, SRR7411017, SRR7411016, and SRR7411020. This Transcriptome Shotgun Assembly project has been deposited at DDBJ/EMBL/GenBank under the accession GGQW00000000. The version described in this paper is the first version, GGQW01000000.

Ethics statement

Fish carrying the parasites used in this study were caught under prefectoral order and in accordance to current ethical laws in France.

Competing interests

None.

Funding

This work was funded by the Agence Nationale de la Recherche (project INCLIMPAR, grant no. ANR-11-JSV7-0010), BiodivERsA (project PROBIS) and the Institut Universitaire de France. EMB was supported by a PhD grant from the French Ministry for Education and Sciences. This work was performed in collaboration with the GeT core facility, Toulouse, France (http://get.genotoul.fr), and was supported by France Génomique National infrastructure, funded as part of “Investissement d’avenir” program managed by Agence Nationale pour la Recherche (contract ANR-10-INBS-09). This work was undertaken in laboratories being part of the “Laboratoire d’Excellence” (LABEX) entitled TULIP (ANR-10-LABX-41).

Authors' contributions

GL, SB and OR designed the experiment and sampled the specimens. OR and CG performed molecular analyses and sequencing. EMB, OR and ET assembled and evaluated the transcriptome assembly. EMB, GL, ET, OR and SB wrote the paper. All authors read and approved the final version of the manuscript.

Acknowledgements

We warmly thank all the students involved in field sampling.

References

1. Fryer G. The parasitic copepoda and branchiura of British freshwater fishes. A handbook and key. Freshwater Biological Association. Ambleside; 1982.

2. Kabata Z. Redescriptions of and comments on four little-known Lernaeopodidae (Crustacea: Copepoda). Can J Zool. 1986;64:1852–9.

3. Loot G, Poulet N, Reyjol Y, Blanchet S, Lek S. The effects of the ectoparasite Tracheliastes polycolpus (Copepoda: Lernaeopodidae) on the fins of rostrum dace (*Leuciscus leuciscus burdigalensis*). Parasitol Res. 2004;94:16–23.

4. Blanchet S, Méjean L, Bourque J-F, Lek S, Thomas F, Marcogliese DJ, et al. Why do parasitized hosts look different? Resolving the "chicken-egg" dilemma. Oecologia. 2009;160:37–47.

5. Rey O, Fourtune L, Paz-Vinas I, Loot G, Veyssière C, Roche B, et al. Elucidating the spatio-temporal dynamics of an emerging wildlife pathogen using approximate Bayesian computation. Mol Ecol. 2015;24:5348–63.

6. Lootvoet A, Blanchet S, Gevrey M, Buisson L, Tudesque L, Loot G. Patterns and processes of alternative host use in a generalist parasite: insights from a natural host-parasite interaction. Mayhew P, editor. Funct Ecol. 2013;27:1403–14.

7. Dybdahl MF, Jenkins CE, Nuismer SL. Identifying the Molecular Basis of Host-Parasite Coevolution: Merging Models and Mechanisms. Am Nat. 2014;184:1–13.

8. De Fine Licht HH. Does pathogen plasticity facilitate host shifts? Round JL, editor. PLOS Pathog. 2018;14:e1006961.

9. Hébert FO, Grambauer S, Barber I, Landry CR, Aubin-Horth N. Major host transitions are modulated through transcriptome-wide reprogramming events in *Schistocephalus solidus* , a threespine stickleback parasite. Mol Ecol. 2017;26:1118–30.

10. Lenz PH, Roncalli V, Hassett RP, Wu L-S, Cieslak MC, Hartline DK, et al. De Novo Assembly of a Transcriptome for Calanus finmarchicus (Crustacea, Copepoda) – The Dominant Zooplankter of the North Atlantic Ocean. Ianora A, editor. PLoS ONE. 2014;9:e88589.

11. Nuñez-Acuña G, Valenzuela-Muñoz V, Gallardo-Escárate C. High-throughput SNP discovery and transcriptome expression profiles from the salmon louse *Caligus rogercresseyi* (Copepoda: Caligidae). Comp Biochem Physiol Part D Genomics Proteomics. 2014;10:9–21.

12. Carmona-Antoñanzas G, Carmichael SN, Heumann J, Taggart JB, Gharbi K, Bron JE, et al. A Survey of the ATP-Binding Cassette (ABC) Gene Superfamily in the Salmon Louse (*Lepeophtheirus salmonis*). Corsi I, editor. PLOS ONE. 2015;10:e0137394.

13. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc. 2013;8:1494–512.

14. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114–20.

15. Huang Y, Niu B, Gao Y, Fu L, Li W. CD-HIT Suite: a web server for clustering and comparing biological sequences. Bioinformatics. 2010;26:680–2.

16. Huerta-Cepas J, Szklarczyk D, Forslund K, Cook H, Heller D, Walter MC, et al. eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. Nucleic Acids Res. 2016;44:D286–93.

17. Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res. 2004;32:258D – 261.

18. Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. :4.

19. Hara Y, Tatsumi K, Yoshida M, Kajikawa E, Kiyonari H, Kuraku S. Optimizing and benchmarking de novo transcriptome sequencing: from library preparation to assembly evaluation. BMC Genomics. 2015;16.

20. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357–9.

21. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31:3210–2.

22. Camacho C, Madden T, Coulouris G, Avagyan V, Ma N, Tao T, et al. Manuel de l'utilisateur des applications BLAST en ligne de commande (BLAST Command Line Applications User Manual). 2009;

23. Clarke K, Yang Y, Marsh R, Xie L, K. ZK. Comparative analysis of de novo transcriptome assembly. Sci China Life Sci. 2013;56:156–62.

24. Yang Y, Smith SA. Optimizing de novo assembly of short-read RNA-seq data for phylogenomics. BMC Genomics. 2013;14:328.

25. Bron J, Boore J, Boxshall G, Bricknell I, Frisch D, Goetze E, et al. Copepod Genome Initiative. 2009.