Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels

Marcel H. Schulz\(^1,2,3,\)†, Daniel R. Zerbino\(^3,4,\)†, Martin Vingron\(^1\) and Ewan Birney\(^3\)

\(^1\)Department of Computational Molecular Biology, Max Planck Institute for Molecular Genetics, Ihnesstraße 63-73, D-14195 Berlin, Germany, \(^2\)Lane Center for Computational Biology, Carnegie Mellon University, Pittsburgh, PA 15213, USA, \(^3\)European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK and \(^4\)Center for Biomolecular Science and Engineering, University of California Santa Cruz, Santa Cruz, CA 95064, USA

ABSTRACT

Motivation: High-throughput sequencing has made the analysis of new model organisms more affordable. Although assembling a new genome can still be costly and difficult, it is possible to use RNA-seq to sequence mRNA. In the absence of a known genome, it is necessary to assemble these sequences de novo, taking into account possible alternative isoforms and the dynamic range of expression values.

Results: We present a software package named Oases designed to heuristically assemble RNA-seq reads in the absence of a reference genome, across a broad spectrum of expression values and in presence of alternative isoforms. It achieves this by using an array of hash lengths, a dynamic filtering of noise, a robust resolution of alternative splicing events and the efficient merging of multiple assemblies. It was tested on human and mouse RNA-seq data and is shown to improve significantly on the transAbYYSS and Trinity de novo transcriptome assemblers.

Availability and implementation: Oases is freely available under the GPL license at www.ebi.ac.uk/~zerbino/oases/

Supplementary information: Supplementary data are available at Bioinformatics online.

Received on December 2, 2011; revised on January 20, 2012; accepted on February 17, 2012

1 INTRODUCTION

Next-generation sequencing of expressed mRNAs (RNA-seq) is gradually transforming the field of transcriptomics (Blenrowe et al., 2009; Wang et al., 2009). The first attempts to discover expressed gene isoforms relied on mapping the RNA-seq reads onto the exons and exon–exon junctions of a known annotation (Jiang and Wong 2009; Mortaza\(^i\) et al., 2008; Richard et al., 2010; Sultan et al., 2008; Wang et al., 2008). Consequently, reference-based ab initio methods have been developed to assemble a transcriptome from RNA-seq data using read alignments alone, inferring the underlying annotation (Denoeud et al., 2008; Gutman et al., 2010; Trapnell et al., 2010; Yassouf et al., 2009).

\(^{†}\)To whom correspondence should be addressed.

\(^{\dagger}\)The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

Unfortunately, the use of a reference genome is not always possible. Despite the drop in the cost of sequencing reagents, the complete study of a genome, from sampling to finishing the assembly is still costly and difficult. Sometimes, the model being studied is sufficiently different from the reference because it comes from a different strain or line such that the mappings are not altogether reliable. For these cases, de novo genome assemblers have been employed to create transcript assemblies, or transcriptomes, from the RNA-seq reads in the absence of a reference genome (Birol et al., 2009; Collins et al., 2008; Jackson et al., 2009; Wakaguri et al., 2009).

However, these short read genomic assemblers, based mainly on de Bruijn graph genomic assemblers (Zerbino and Birney, 2008; Simpson et al., 2009), make implicit assumptions regarding the evenness of the coverage and the colinearity of the sequence. Indeed, the coverage depth fluctuates significantly between transcripts, isoforms and regions of the transcript, therefore it cannot be used to determine the uniqueness of regions or to isolate erroneous sequence. In addition, these tools are geared to produce long linear contigs from the given sequence, not to detect the overlapping sequences presented by isoforms of a single gene. This affects a number of steps, including error correction, repeat detection and read pair usage. These methods are therefore not necessarily suited to process transcriptome data which does not conform to either of these assumptions.

More recently, transcriptome assembly pipelines were developed to post-process the output of de novo genome assemblers: Velvet and ABySS (Martin et al., 2010; Robertson et al., 2010; Surget-Groba and Montoya-Burgos, 2010). The common idea shared by these pipelines is to run an assembler at different k-mer lengths and to merge these assemblies into one. The rationale behind this approach is to merge more sensitive (lower values of k) and more specific assemblies (higher values of k).

The pipeline presented by Robertson et al. (2010), transABYYSS, also handles alternative splicing variants. It detects them by searching for connected groups of contigs such that they are connected in a characteristic bubble and one of the contigs has a length of exactly (2\(k\)−2). These bubbles are first removed, then added to the final assemblies, to reconstruct alternate variants. A variety of algorithmic researchers have used splicing graphs to represent alternative splicing which have a direct relationship to de Bruijn graphs, as pointed out by Heber et al. (2002). This homology...
between data structures opens the possibility of a de novo short read transcriptome assembler, as illustrated by the Trinity algorithm (Yassour et al., 2011). Trinity starts by extending contigs greedily, connecting them into a de Bruijn graph, then extracting sufficiently covered paths through this graph. Trinity is designed to reconstruct highly expressed transcripts to full length using only one k-mer length.

We present Oases, a de novo transcriptome assembler that combines these advances. Oases merges the use of multiple k-mers presented in (Robertson et al., 2010; Surget-Groba and Montoya-Burgos, 2010) with a topological analysis similar to that presented by Yassour et al. (2011). It uses dynamic error removal adapted to RNA-seq data and implements a robust method to predict full length transfraggs, even in cases where noise perturbs the topology of the graph. The k-mer assemblies are merged to cover genes at different expression levels without redundancy.

We tested the latest version of Oases (0.2.0.1) on experimental datasets and found that Oases produces longer assemblies than previous de novo RNA-seq assemblers. Oases was compared with a reference-based ab initio algorithm, Cufflinks (Trapnell et al., 2010). The latter approach has a considerable advantage in low expression levels. However, the later stage algorithms, Pebble and Rock Band, which resolve non-covered or overlapping reads, are not used because they rely on assumptions related to genomic position, but at high read coverage, Oases’ sensitivity approaches that of reference-based ab initio algorithms. We also examined the effect of coverage depth, hash length, alternative splicing and assembly merging on the quality of assemblies.

2 METHODS

2.1 Overview

The Oases assembly process, explained in detail below and illustrated in Figure 1, consists of independent assemblies, which vary by one important parameter, the hash (or k-mer) length. In each of the assemblies, the reads are used to build a de Bruijn graph, which is then simplified for errors, organized into a scaffold, divided into loci and finally analyzed to extract transcript assemblies or transfraggs. Once all of the individual k-mer assemblies are finished, they are merged into a final assembly.

2.2 Contig assembly

The Oases pipeline receives as input a preliminary assembly produced by the Velvet assembler (Zerbino and Birney, 2008) which was designed to produce scaffolds from genomic reads. Its initial stages, namely hashing and graph construction can be used independently on transcriptome data. We only run these stages of Velvet to produce a preliminary fragmented assembly, containing the mapping of the reads onto a set of contigs.

However, the later stage algorithms, Pebble and Rock Band, which resolve repeats in Velvet, are not used because they rely on assumptions related to genomic sequencing (Zerbino et al., 2009). Namely, the coverage distribution should be roughly uniform across the genome and the genome should not contain any branching point. These conditions prevent those algorithms from being reliable and efficient on RNA-seq data.

2.3 Contig correction

After reading the contigs produced by Velvet, Oases proceeds to correct them again with a set of dynamic and static filters.

The first dynamic correction is a slightly modified version of Velvet’s error correction algorithm. TourBus, TourBus searches through the graph for parallel paths that have the same starting and end node. If their sequences are similar enough, the path with lower coverage is merged into the path with higher coverage, irrespective of their absolute coverage. In this sense,
These loci can be fragmented because of alternative splicing events which cause the de Bruijn graph to have a branch. Oases, therefore, analyses the connectivity of the locus. To avoid this, a limit is imposed on the number of removed connections. If two connections can be reduced to one, the highest one is preserved.

2.8 Extracting transcript assemblies

The sequence information of the transcripts is now contained in the loci. Oases uses a reimplementation of the algorithm described in (Lee, 2003), which efficiently produces a parsimonious set of putative highly expressed transcripts, assuming independence of the alternative splicing events.

This extension of the algorithm is quite intuitive, since there is a direct analogy between the de Bruijn graph built from the transcripts of a gene and its splicing graph, as noted by Heber et al. (2002). Using dynamic programming, it enumerates heavily weighted paths through the locus graph in decreasing order of coverage, until either all the contigs of the locus are covered, or a specified number of transcripts is produced (by default 10).

As in the transitive reduction phase, this algorithm had to be slightly modified to allow for loops in the putative splicing graph of the locus. Loops are problematic because their presence can prevent the propagation of the dynamic programming algorithm to all the contigs of a locus. When a loop is detected, it is broken at a contig which connects the loop to the rest of the locus, so as to leave a minimum number of branch points, as described in the Supplementary Material.

2.9 Merging assemblies with Oases-M

De Bruijn graph assemblers are very sensitive to the setting of the hash length k. For transcriptome data, this optimization is more complex as transcript expression levels and coverage depths are distributed over a wide range. A way to avoid the dependence on the parameter k is to produce a merged transcriptome assembly of previously generated transfragts from Oases.

Oases is run for a set of k ∈ [kMIN, …, kMAX] values and the output transfragts are stored. All predicted transfragts from runs in the interval are then fed into the second stage of the pipeline, Oases-M, with a user selected kDISTANCE. A de Bruijn graph for kDISTANCE is built from these transfragts. After removing small variants with the Tourbus algorithm, any transfrag in the graph that is identical or included in another transfrag is removed. The final assembly is constructed by following the remaining transfragts through the merged graph.

3 RESULTS

3.1 Datasets

Two datasets were retrieved from the Nucleotide Archive (http://www.ebi.ac.uk/ena/). A human dataset was produced in a study by Heap et al. (2010), where poly(A)-selected RNAs from human primary CD4+ T cells were sequenced. Paired-end reads of length 45 bp with an insert size of 200 bp from one human individual (study ID SRR011545) were downloaded.

A mouse dataset was taken from the study of Trapnell et al. (2010). In a timeseries experiment of C2C12 myoblast mouse cells, paired-end reads of length 75 bp with an insert size of 300 bp were sequenced. Read data from the 24 h timepoint (study id SRR017794) was used.

To reduce the amount of erroneous bases, both paired-end datasets were processed by (i) removing Ns from both ends, (ii) clipping bases with a Sanger quality ≤ 10 and (iii) removing reads with more than six bases with Sanger quality ≤ 10 after steps (i) and (ii), leading to a total of 30 940 088 and 64 441 708 reads for human and mouse, respectively.

3.2 Assemblies and alignments

All experiments were run with Oases version 0.2.01, and Velvet 1.1.06 and the coverage cutoff and the minimum support for connections were set to 3.

TransABySS 1.2.0 was run with ABYSS 1.2.5 through the first two stages of transABySS (assembly and merging, before mapping to a reference genome is required). Instead of just running with
To evaluate the added value of the topology resolution within each k-mer, we compared the Oases contigs from the Velvet assemblies which they are built from. Table 1 shows how the Oases assemblies significantly improve on the Velvet assemblies. This confirms the intuition that in the presence of alternative splicing and dynamic expression levels, the assembly is broken by breaks in the graph, which can be resolved by topological analysis and adapted error correction as described in the Methods section.

As an example, the percentage cutoff for local edge removal was modulated (see Supplementary Table S1). These results show how dynamic filters improve the quality of the assembly.

### 3.5 Impact of k-mer lengths

One of the major parameters in de Bruijn graph assemblers is the hash length, or k-mer length. Comparing single-k assemblies performed by Oases, it is possible to observe that this parameter is crucial in RNA-seq assembly. Figure 2 shows how the k-mer length is closely related to the expression level of the transcripts being assembled. As expected, the assemblies with longer k-values perform best on high expression genes, but poorly on low expression genes. However, short k-mer assemblies have the disadvantage of introducing misassemblies, as shown in Supplementary Table S7.

### 3.6 Impact of merging assemblies

In addition, Figure 2 shows the same statistics for the merged assembly by Oases-M, which is significantly superior to each of the individual assemblies. This result illustrates how the different assemblies do not completely overlap. Further, Supplementary Figure S2 shows how each single k-mer assembly resolved transcripts at different expression levels.

We compared merging different intervals of k-mers (see Supplementary Material). The wider the interval, the better the results. To determine bounds on this interval we arbitrarily bounded on the low values with 19, on the assumption that smaller k-mers are very likely to be unspecific for mammalian genomes (Whiteford et al., 2005). In theory, on the upper end, all the k-mer values (up to read length) could be used. To avoid wasting resources, we measured the added value of each new assembly (see Supplementary Material).

As expected, marginal gains progressively diminish and this metric could be used to determine how large a spectrum of k-mers to use. We also investigated which kMERGE should be used and we found that kMERGE=27 works well with little difference for higher values (see Supplementary Table S4) and is therefore used for all analyses in the article.

### 3.7 Comparing Oases to other RNA-seq de novo assemblers

Oases-M was compared with existing RNA-seq de novo assemblers, transABySS (Robertson et al., 2010) and Trinity (Yassour et al., 2011). The previous human dataset and a mouse dataset were used.
Fig. 3. Reconstruction efficiency of Ensembl transcripts for different RNA-seq de novo assembly methods (Oases-M, Trinity, and transADySS) on human and mouse datasets. Reference-based assembly results using Cufflinks are provided on the mouse dataset. All annotated genes have been grouped into quantiles by RPKM expression values of 1464 (resp. 1078) genes for the human data (resp. mouse).

for the comparison. The datasets have different read lengths and sequencing depth, as detailed in Methods. Both transADySS and Oases were run for k-mer length 19–35 bp on the human dataset. Because the mouse reads are longer, these two assemblers were run for k-mers 21–35 on that dataset. The highest value of k was determined by an approach similar to that used on the human data (see Supplementary Material for data). Trinity is fixed by implementation at k=25 bp.

Figure 3 shows the number of reconstructed Ensembl transcripts for each assembler on both datasets separated by expression quantiles. The main observation is that all assemblers do not behave equally with respect to expression level. Trinity appears to perform best on high expression genes, whereas transADySS performs best on low expression genes. Oases performs comparatively well throughout the spectrum of expression levels, hence the greater overall success (Table 2).

Regarding correctness, we computed the number of misassemblies and the qualities of the different assemblers are comparable (see Supplementary Material). Transfrags mapped with high confidence to the genome occasionally differ from the known annotation. For example, Oases produced 237 (resp. 390) transfrags longer than 300 bp which mapped to the reference genome, but did not overlap with the human (resp. mouse) annotation.

In Figure 4, the overlap of full length mouse transcripts reconstructed by the three methods is shown. It is interesting to note that although the results greatly overlap, the different assemblers succeeded in assembling different transcripts.

3.8 Comparing de novo and reference-based assemblers

Oases and the other de novo assemblers were finally compared on the mouse data to a reference-based assembly algorithm, Cufflinks (Trapnell et al., 2010), on the mouse dataset. As could be expected, Cufflinks generally outperforms the de novo assembly algorithms,
We have shown that merging different single
memory on a 48 core AMD Opteron machine with 265 GB RAM.
uses multithreading but Oases currently does not. The complete
merged assembly for human took ∼3.2 h and 6.1 GB of peak memory
on a 48 core AMD Opteron machine with 265 GB RAM. The merged assembly for mouse took ∼10.3 h and 15.1 GB peak memory.

3.9 Runtime and memory

A de novo transcriptome assembly is a resource intensive task. Velvet
uses multithreading but Oases currently does not. The complete
merged assembly for human took ∼3.2 h and 6.1 GB of peak memory
on a 48 core AMD Opteron machine with 265 GB RAM. The merged assembly for mouse took ∼10.3 h and 15.1 GB peak memory.

4 DISCUSSION

We have shown that merging different single k assemblies is
beneficial, in concordance with previous work (Sergei-Groba and Montoya-Burgos, 2010; Robertson et al., 2010). Oases employs
dynamic cutoffs, where possible, to allow for a robust reconstruction
with different k-values. However, detailed parameter optimization
for Oases and trans-AbYSS may lead to further improvements.

Overall, the de novo methods produced large numbers of
misassemblies. Given the dynamic ranges involved, the exact parameter settings of these programs define a trade-off between
sensitivity and accuracy. In these experiments, Oases tends to be
more sensitive, Trinity more accurate. The correlation of small k-mer
assemblies and misassembly rates suggests that homologies between
genes are the main source of errors. As reads get longer, and coverage
depths greater, sensitivity will only increase and users will probably
avoid the shorter k-mer lengths for greater accuracy. Short k-mers
will only be necessary to retrieve the very rare transcripts.

An independent but significant factor to these assemblies is read
preprocessing, as read error removal has already been shown to have
a significant impact in the context of de novo genome assembly
(Smeds and Künster, 2011).

Interestingly, the comparison of reconstructed transcripts for
the three de novo methods in Figure 4 reveals that each method
outperforms the others on a separate set of transcripts. These
differences in performance are probably due to the different
strategies employed to remove errors. A more aggressive method,
which discards more data, would presumably end up with many gaps
on low expression data, whereas a more lenient algorithm would
leave too many ambiguities at high coverage.

In particular, it appears that the performance of all the assemblers
sometimes drops at very high coverage depths. This is probably
linked to increased noise. Indeed, this drop is especially marked
for trans-AbYSS, which, to our knowledge, is the only of the three
de novo assemblers not to integrate dynamic filters which adapt with
coverage depth.

Intriguingly, trans-AbYSS outperformed Trinity in our
experiments, contrary to the observation of Yassour et al.,
(2011). This could not be due to the parameterization of Trinity,
which cannot be parameterized apart from the insert length. Instead,
the larger k range used for trans-AbYSS and the lower sequencing
depth in our analyzed data sets may explain this discrepancy, as
trans-AbYSS was shown to perform especially well for low to
medium expressed genes.

Similarly, our experiments on mouse data show a bigger gap
between Cufflinks and the de novo assemblers than observed by
Yassour et al. (2011). In their work, the comparison was focused
on the set of ‘oracle’ transcripts, which show sufficient coverage of
each k-mer in the reads. However, no such restriction was applied
here and Cufflinks surpasses the de novo methods for low to medium
expression ranges, where coverage is sparse.

In this study, we did not analyze strand-specific RNA-seq datasets.
However, as these datasets become more available (Levin, 2010)
Oases already supports this data. During the hashing phase, reverse
complement sequences can be stored separately instead of being
joined as the two strands of the same sequence.

5 CONCLUSION

Oases provides users with a robust pipeline to assemble unmapped
RNA-seq reads into full length transcripts. Oases was designed to
deal with the conditions of RNA-seq, namely uneven coverage and
alternative splicing events.

Our results show how crucial it is to explore and understand the
relevant conditions. Alternative splicing can significantly confound

---

**Table 2. Overall comparison of the different RNA-seq assembly methods**
on human and mouse datasets

| Data    | Method   | Tfrags > 100bp | Sens (%) | Spec (%) | Full lgth | 80% lgth |
|---------|----------|----------------|----------|----------|-----------|----------|
| Human   | Oases-M  | 174 469        | 21.44    | 92.35    | 1463      | 11 169   |
|         | AbYSS    | 100 127        | 19.65    | 92.16    | 1358      | 10 992   |
|         | Trinity  | 76 232         | 19.99    | 86.63    | 953       | 7129     |
| Mouse   | Oases-M  | 175 914        | 30.83    | 89.08    | 1324      | 9880     |
|         | AbYSS    | 174 744        | 30.66    | 92.79    | 1149      | 9376     |
|         | Trinity  | 92 410         | 31.57    | 87.14    | 1085      | 7028     |

---

**Fig. 4. Venn Diagramm that compares mouse Ensembl transcripts**
reconstructed to full length by Trinity, Trans-AbYSS and Oases-M for the
mouse RNA-seq data.

as it benefits from using the reference genome to anchor its
assemblies (Fig. 3). Nonetheless, it is interesting to note that as
expression level and therefore coverage depth go up, the gap
narrow.

Beyond assembling more transcripts, it is also important to
recover multiple isoforms for each gene. For each assembled
transcript, the average number of additionally assembled transcripts
from the same gene are, respectively, 1.21, 1.25, 1.01 and 1.56 for
Oases, trans-AbYSS, Trinity and Cufflinks. Cufflinks performs better
in that respect, whereas Trinity is less sensitive.
the assembly and has to be specifically addressed. Gene expression levels are a major factor determining the sensitivity of an algorithm. High coverage genes require more selective methods, whereas low coverage genes favor more sensitive algorithms. This is why exploring a range of k-mer lengths is key to success.

In the light of these results, Oases was designed to perform well overall by adapting to these varying conditions and succeeded in obtaining superior overall results compared to previously published RNA-seq de novo assemblers. Nonetheless, it also appears that merging assemblies from a diversity of algorithms could be beneficial. This is probably due to the dynamic range of all the variables, which prevent any single method from being systematically superior.

Finally, we examined the difference between de novo and reference-assisted assembly. In the presence of a well-assembled genome (typically human or mouse), the latter methods are generally systematically superior. In the absence of an assembled genome (typically human or mouse), the former methods are generally beneficial. This is probably due to the dynamic range of all the variables, which prevent any single method from being systematically superior.

ACKNOWLEDGEMENTS

Thank you to Cole Trapnell, Ali Mortazavi and Diane Trout for their help in obtaining the C2C12 dataset. Many thanks to Brian Haas for his patient testing of Oases, and to all the other users for their feedback. Thank you to Saket Navlakha and Benedict Paten for their proofreading and commentaries.

REFERENCES

Birgner,J.L. et al. (2009) De novo transcriptome assembly with ABlySS. Bioinformatics, 25, 2872–2877.

Blenkow,B.J. et al. (2009) Current-generation high-throughput sequencing: deepening insights into mammalian transcriptomes. Genome Dev., 23, 1379–1386.

Butler,J. et al. (2008) ALLELESH: de novo assembly of whole-genome shotgun microarrays. Genome Res., 18, 830–840.

Collins,L.J. et al. (2008) An approach to transcriptome analysis of non-model organisms using short-read sequences. Genome Inform., 21, 3–14.

Dernovoi,P. et al. (2008) Annotating genomes with massive-scale RNA sequencing. Genome Biol., 9, R175.

Gattman,M. et al. (2010) Ab initio reconstruction of cell-type-specific transcriptomes in mouse reveals the conserved multi-ethnic structure of lincRNAs. Nat. Biotechnol., 28, 503–510.

Hepp,G.A. et al. (2010) Genome-wide analysis of allelic expression imbalance in human primary cells by high-throughput transcriptome sequencing. Hum. Mol. Genet., 19, 122–134.

Hob MOCKR et al. (2002) Splicing graphs and EST assembly problem. Bioinformatics, 18 (Suppl. 1), S181–S188.

Jackson,B.G et al. (2009) Parallel short sequence assembly of transcriptomes. BMC Bioinformatics, 10 (Suppl. 1), S14.

Jiang,H. and Wong,W.H. (2009) Statistical inferences for isoform expression in RNA-Seq. Bioinformatics, 25, 1026–1032.

Kent,W.J. (2002) BLAT–the BLAST-like alignment tool. Genome Res., 12, 656–664.

Lee,C. (2003) Generating consensus sequences from partial order multiple sequence alignment graphs. Bioinformatics, 19, 999–1008.

Levin,J.Z. (2010) Comprehensive comparative analysis of strand-specific RNA sequencing methods. Nat. Methods, 7, 709–715.

Martin,J. et al. (2010) Bowtie: a fast exact short-read aligner. Bioinformatics, 26, 519–520.

Mortazavi,A. et al. (2008) Mapping and quantifying mammalian transcriptomes by RNA-seq. Nat. Methods, 5, 621–628.

Myers,E.W. (2005) The fragment assembly string graph. Bioinformatics, 21 (Suppl. 2), ii79–ii85.

Rizk,P. et al. (2009) EMBISS: the European molecular biology open software suite. Trends Genet., 26, 277–277.

Richard,B. et al. (2010) Prediction of alternative isoforms from cDNA expression levels in RNA-Seq experiments. Nucleic Acids Res., 38, e112.

Robertson,G. et al. (2010) De novo assembly and analysis of RNA-seq data. Nat. Methods, 7, 909–912.

Suzuki,T. et al. (2008) ABySS: a parallel assembler for short read sequence data. Genome Res., 19, 1117–1123.

Smeds,L. and Kito,T. (2010) ConDetri – a content dependent read trimmer for illumina data. PLoS One, 6, e26314.

Sultan,M. et al. (2008) A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. Science, 321, 956–960.

Suret-Grova,Y. and Mortaza-Burges,J.I. (2010) Optimization de de novo transcriptome assembly from next-generation sequencing data. Genome Res., 20, 1432–1440.

Trapnell,C. et al. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol., 28, 511–515.

Wakaguri,H. et al. (2009) Full-malaria/parasites and full-arthropods: databases of full-length cDNAs of parasites and arthropods, update 2008. Nucleic Acids Res., 37, D520–D525.

Wang,E.T. et al. (2008) Alternative isoform regulation in human tissue transcriptomes. Natzen, 456, 470–476.

Weng,Z. et al. (2009) RNA-Seq: A revolution tool for transcriptomics. Nat. Rev. Genet., 10, 57–65.

Wexler,D.B. et al. (2009) RazerS–fast read mapping with sensitivity control. Genome Res., 19, 1646–1654.

Whiteford,N. et al. (2005) An analysis of the feasibility of short read sequencing. Nucleic Acid Res., 33, e131.

Yassour,M. et al. (2009) Ab initio construction of a eukaryotic transcriptome by massively parallel short read sequencing. Proc. Natl Acad. Sci., USA, 106, 3264–3269.

Yassour,M. et al. (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol., 29, 644–652.

Zerbino,D.R. and Birney,E. (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res., 18, 821–829.

Zerbino,D.R. et al. (2009) PEBble and rock band: heuristic resolution of repeats and scaffolding in the velvet short-read de novo assemblers. PLoS One, 4, e4907.