MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2012) **12**, 834–845 doi: 10.1111/j.1755-0998.2012.03148.x

Reference-free transcriptome assembly in non-model animals from next-generation sequencing data

V. CAHAIS,* P. GAYRAL,* G. TSAGKOGEORGA,*† J. MELO-FERREIRA,‡ M. BALLENGHIEN,* L. WEINERT,*§ Y. CHIARI,*‡ K. BELKHIR,* V. RANWEZ* and N. GALTIER*
*CNRS UMR 5554, Institut des Sciences de l'Evolution de Montpellier, Université Montpellier 2, Place E. Bataillon, 34095 Montpellier, France, †School of Biological and Chemical Sciences, Queen Mary University of London, Mile End Road, London, E1 4NS, UK, ‡CIBIO, Centro de Investigação em Biodiversidade e Recursos Genéticos, Universidade do Porto, Campus Agrário de Vairão, 4485-661 Vairão, Portugal, §Medical Research Council (MRC), Centre for Outbreak Analysis and Modelling, Imperial College Faculty of Medicine, London, UK

Abstract

Next-generation sequencing (NGS) technologies offer the opportunity for population genomic study of non-model organisms sampled in the wild. The transcriptome is a convenient and popular target for such purposes. However, designing genetic markers from NGS transcriptome data requires assembling gene-coding sequences out of short reads. This is a complex task owing to gene duplications, genetic polymorphism, alternative splicing and transcription noise. Typical assembling programmes return thousands of predicted contigs, whose connection to the species true gene content is unclear, and from which SNP definition is uneasy. Here, the transcriptomes of five diverse non-model animal species (hare, turtle, ant, oyster and tunicate) were assembled from newly generated 454 and Illumina sequence reads. In two species for which a reference genome is available, a new procedure was introduced to annotate each predicted contig as either a full-length cDNA, fragment, chimera, allele, paralogue, genomic sequence or other, based on the number of, and overlap between, BLAST hits to the appropriate reference. Analyses showed that (i) the highest quality assemblies are obtained when 454 and Illumina data are combined, (ii) typical *de novo* assemblies include a majority of irrelevant cDNA predictions and (iii) assemblies can be appropriately cleaned by filtering contigs based on length and coverage. We conclude that robust, reference-free assembly of thousands of genes from transcriptomic NGS data is possible, opening promising perspectives for transcriptome-based population genomics in animals. A Galaxy pipeline implementing our best-performing assembling strategy is provided.

Keywords: alleles, next-generation sequencing, paralogues, transcriptomics

Received 20 October 2011; revision received 6 March 2012; accepted 19 March 2012

Introduction

Evolutionary and population genomic research has long been restricted to a relatively small number of taxonomic groups, especially in eukaryotes, in which the sequencing effort was first concentrated on a handful of model organisms. Mammals, *Drosophila* and yeasts, for instance, contribute a disproportionate amount of the existing molecular evolutionary literature, for non-obvious scientific reasons. The recent and ongoing technological advances are modifying the situation: next-generation sequencing technologies (NGS) offer the opportunity to generate genome-wide sequence data sets in non-model organisms at reasonable cost (Hudson 2008). A popular target for NGS is species transcriptome, which offers direct access to the coding sequences of many genes, plus information on their relative expression levels (Vera *et al.* 2008; Elmer *et al.* 2010; Renaut *et al.* 2010; Wolf *et al.* 2010). Transcriptome analysis can be particularly appropriate for species carrying a large, repetitive genome, whose complete sequencing and assembly would be both costly and challenging. NGS transcriptome data are therefore a promising source of genetic markers (SNPs, sets of orthologues genes), potentially applicable to any taxonomic group.

As a first step, NGS transcriptome data analysis requires assembling millions of relatively short reads into predicted cDNAs. On one hand, assembling transcriptomes looks easier than assembling (large) genomes (Paszkiewicz & Studholme 2010), because the amount of repetitive DNA is typically lower in coding than in non-

Correspondence: Nicolas Galtier, Fax: +33 467 14 36 10;
E-mail: nicolas.galtier@univ-montp2.fr

coding regions. On the other hand, as far as gene sequences are concerned, transcriptome data are less informative than genome data because introns are lacking. This is problematic for multigene families, which are presumably more difficult to disentangle when transcriptome data are used. For instance, a recently born, intronless retrogene may be incorrectly assembled with its progenitor gene using transcriptome data, whereas confusion is almost impossible with genome data. Another distinctive feature of transcriptome data is unequal coverage across transcripts. Lowly expressed genes contribute a small fraction of total RNA's, and of NGS reads, and are therefore less easily assembled than highly expressed transcripts (Surget-Groba & Montoya-Burgos 2010). Transcriptome data are also affected by the occurrence of alternative splicing, which obviously complicates the assembling task (Sammeth 2009). A final problem with both data types is coping with natural heterozygosity when applying NGS to wild (diploid) specimen, which, depending on species, can be far from negligible (Small *et al.* 2007; Wang *et al.* 2010a). Gathering the alleles of a gene in a single contig is a challenge for assembling algorithms, especially in highly polymorphic species.

For all these reasons, *de novo* transcriptome assembly from NGS data is likely to produce a substantial fraction of erroneous predictions of various sorts, including chimera, fragmented genes, unassembled alleles and assembled paralogues, which would mislead subsequent population or comparative genomic analyses. Trusting such predictions would mislead gene orthology annotation and SNP detection. The standard way of getting around these problems would be to focus on a subset of predicted cDNAs whose validity is guaranteed by, for example strong reciprocal similarity with a reference genome (Künstner *et al.* 2010). However, this approach means throwing away a fraction of the data, this fraction being largest in the least studied taxa, for which no close reference is available. In such non-model taxa, which represent the vast majority of existing species, defining reliable evolutionary markers from transcriptomic short-read data is far from obvious.

In this study, we investigated the reliability of *de novo* transcriptome assembly from newly generated NGS data in five animal species – *Ciona intestinalis* B (tunicate), *Ostrea edulis* (flat oyster), *Messor barbarus* (harvest ant), *Emys orbicularis* (European pond turtle) and *Lepus granatensis* (hare) – of which two (*C. intestinalis* and *L. granatensis*) have a fully sequenced close relative. We designed a BLAST-based assessment of the quality of *de novo* assemblies by annotating each contig as either a full-length cDNA, fragment, chimera, unassembled allele, assembled paralogue, genomic sequence or other. We compared sequencing technologies and assembling programmes, and identified reference-free filtering strategies optimizing the number and proportion of correctly predicted cDNAs, available for SNP detection and between-species orthologue comparisons. We conclude that reliable *de novo* coding sequence assembly from NGS transcriptome data for marker discovery and population genomic analyses is possible even in the absence of a reference genome.

Methods

Sampling and sequencing

Eight individuals of the urochordate *Ciona intestinalis* species B (ascidian) were caught in the wild in Norway (North Sea) and Canada (Atlantic Ocean). Siphon and gonads were dissected and preserved in RNA later solution. For each individual, total RNA was extracted from a mix of the two tissues using standard protocols (details in Gayral *et al.* 2011). For one individual, 25 μg total RNA served to build a random-primed cDNA library. After size selection on gel, it was normalized, thus homogenizing the concentrations across the cDNAs. This library was sequenced for half a run using a Genome Sequencer (GS) FLX Titanium Instrument (Roche Diagnostics). Reads were trimmed of low-quality terminal portions using the SeqClean program (http://compbio.dfci.harvard.edu/tgi/). For eight individuals, 3'-primed, non-normalized cDNA libraries were built from 5 μg of total RNA using SMART cDNA library construction kit (Clontech, Mountain View, CA, USA). An oligo(dT)-primed first-strand synthesis and cap-primed second-strand synthesis were performed. Adapters were ligated after cDNA was colligated and nebulized, and sequencing (five tagged individuals per lane) was carried out on a Genome Analyzer II (Illumina, Inc.) to produce 100 bp single-end fragments. Three data sets resulted from these experiments, respectively, called Ciona_454, Ciona_illu and Ciona_mix. The Ciona_mix data set corresponds to the union of reads from 454 and Illumina.

The same strategy was followed in *Lepus granatensis* (Iberian hare), *Ostrea edulis* (European flat oyster), *Messor barbarus* (harvest ant) and *Emys orbicularis* (European pond turtle). Eight individuals of each species were caught in various localities of their natural geographic range. RNA was extracted using adapted protocols, as described in Gayral *et al.* (2011) and Chiari & Galtier (2011). In hare and oyster, several tissues (liver, kidney, spleen and lung in hare; muscle, mantle, gills and digestive gland in oyster) were dissected and an equimolar mix of total RNA from the various tissues was used to build cDNA libraries. In ant, the whole body was used. In turtle, blood samples were used. In all four species, cDNA libraries were prepared as with *C. intestinalis* (see above), with the exception that the ant 454 cDNA library

was not normalized. The resulting data sets were called Lepus_454, Lepus_illu and Lepus_mix, for hare, and similarly for the other species. The analysed data sets therefore reflect the molecular diversity of one (454) or several (Illumina) diploid individuals, at both the sequence and gene expression levels, as expected for typical NGS-based transcriptome data sets from wild specimen.

Assembling protocols

The assembling protocols used in this study were based in the first place on two well-established and robust assembling programmes, namely Cap3 (Version Date: 15 October 2007, Huang & Madan 1999) and Abyss (version 1.2.0, *k*-mer length set to 60, Simpson *et al.* 2009). The former has been primarily designed for relatively long (typically 400 bp) reads, the latter (Abyss) for shorter reads (typically 30–100 bp). Long 454 reads were assembled using Cap3 in a single run, with default parameter settings. Illumina reads were assembled using a combination of Abyss and Cap3: the contigs generated by Abyss were used as an input to Cap3, which was run twice consecutively. Figure 1 shows the 454-only and Illumina-only assembling strategies. Note that all these assembling programmes return two output files, that is, a contig file (sequences obtained by aligning at least two input sequences) and a singleton file (sequences that were not combined with any other sequence), indicated by 'c' and 's' in Fig. 1. In this study, reads returned as singletons by the first assembling run were discarded. Mixed data sets were analysed in two distinct ways (Fig. 2). In the 'merge reads' approach, 454 and Illumina reads were mixed and analysed with Abyss and Cap3, similarly to Illumina data sets. In the 'merge contigs' approach, 454 and Illumina data were first processed independently and the resulting contigs were then mixed and assembled again with Cap3. Contigs shorter than 200 bp were discarded prior to quantitative and qualitative assessment in all the assemblies considered in this study. Besides Cap3 and Abyss, six additional assembling programmes were tried for comparative purposes: Celera (version 6.1, Myers *et al.* 2000; settings as suggested by O'Neil *et al.* 2010), Mira (version 3.0.3, Chevreux *et al.* 2004), Newbler (version 2.6), TransAbyss (Robertson *et al.* 2010), Soap_de_novo_Trans (version 1.01, Li *et al.* 2010) and Trinity (version R2011-10-29, Grabherr *et al.* 2011).

Reference transcriptomes and genomes

cDNA collections of *C. intestinalis* A (14 012 sequences) and *Oryctolagus cuniculus* (rabbit, 24 800 sequences) were downloaded from NCBI (ftp://ftp.ncbi.nih.gov/genomes/) and ENSEMBL (http://www.ensembl.org/info/data/ftp), respectively, and served as reference transcriptomes for *C. intestinalis* B and *L. granatensis* data. The average genomic divergence between the target and the reference species is ~5% for hare vs. rabbit (Carneiro *et al.* 2010) and ~10% for *C. intestinalis* A vs. B (Nydam & Harrison 2010). The contig annotation procedure introduced in this study also requires a reference genome. In *Ciona*, the genome of *C. intestinalis* A (version 12 September 2008) was used. In *Lepus*, the reference genome used was the human non-redundant DNA (human nt), because the rabbit genome is only partially assembled.

In *C. intestinalis*, the existence of two distinct species, currently named A and B, was discovered after the start of the genome sequencing project (Caputi *et al.* 2007; Iannelli *et al.* 2007). To confirm the species assignment of our *C. intestinalis* reference, we relied on seven diagnostic genes, six nuclear and one mitochondrial (*Cox1*), identified by Nydam & Harrison (2010). Using BLAST search, the sequences of these genes were retrieved from the reference and included in Nydam and Harrison's alignments. In the resulting maximum-likelihood phylogenies, with no exception, the GenBank reference sequences were clustered with individuals from species A, to the exclusion of individuals from species B (data not shown). A similar procedure was performed for each of our

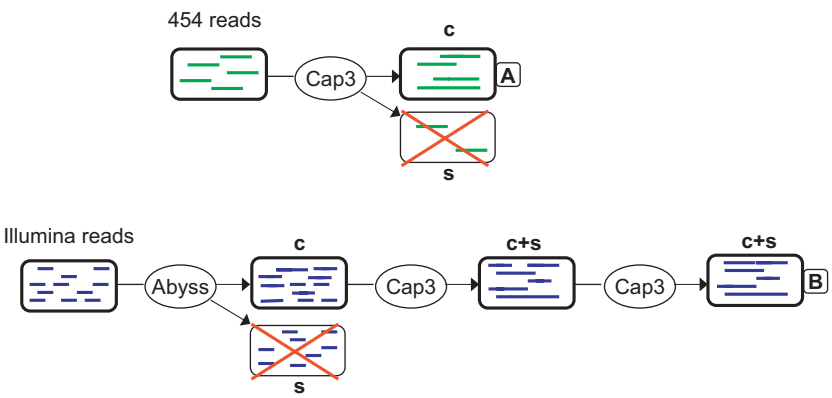

Fig. 1 Separate assembly schemes for 454 and Illumina data. Lower case 'c' and 's' are for 'contigs' and 'singletons', respectively. Upper case 'A' and 'B' correspond to assembly names as listed in Tables 2–4.

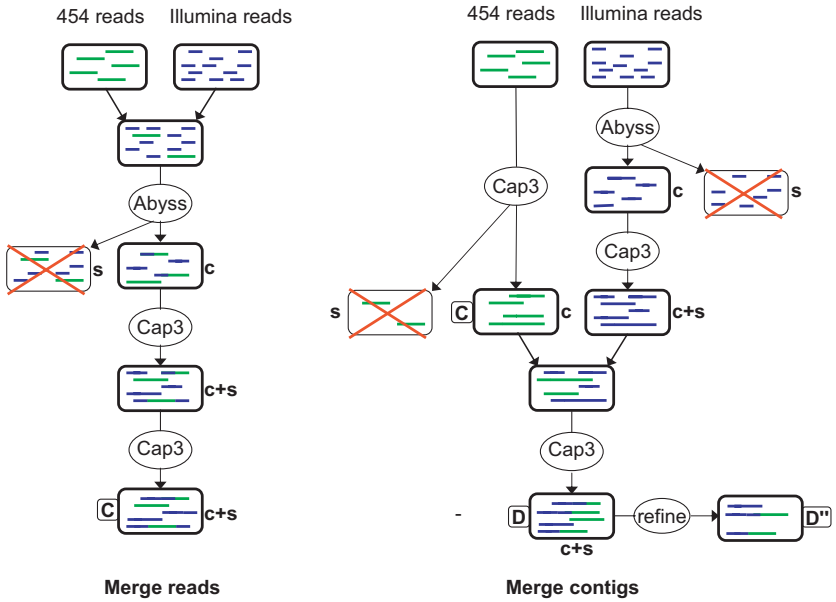

Fig. 2 Two approaches to combine 454 and Illumina data in a single assembly. Lower case 'c' and 's' are for 'contigs' and 'singletons', respectively. Upper case 'C', 'D' and 'D'" correspond to assembly names as listed in Tables 2–4.

samples, in which the mitochondrial *Cox1* gene confirmed assignment to species B.

Assembly quality control

For each contig, a BLASTN search was performed against the appropriate transcriptome reference. Hits were considered significant when (i) the alignment length (merging all high-scoring segments pairs) was at least 50% of the query sequence or at least 50% of the hit sequence, and (ii) sequence identity between query and hit was more than 80% across the aligned portion. Contigs were first classified in five categories, defined from the number and nature of significant BLAST hits (Fig. 3). Contigs with no significant hit were called 'no hit'. Contigs with a single significant hit, *h*, such that no other contig finds *h*, were called '1 → 1'. Contigs with a single significant hit shared by other contigs were called '*m* → 1'. Contigs with several significant hits, all specific to this contig, were called '1 → *n*'. Contigs not included in the above-defined categories, because involved in more complex patterns, were called '*m* → *n*'.

Annotations were then refined based on the characteristics of contig–hit alignments (Fig. 4). Among the 1 → 1, contigs were called 'full' if the contig/hit alignment covered at least 90% of the hit sequence or 'fragments' otherwise. For *m* → 1 contigs, annotation depended on the overlap between contigs – the overlap between two of the *m* contigs was defined as the *a*/*z* ratio, where *a* was the length of hit sequence to which both queries align and *z* the length of the shortest of the two queries. Overlap was considered significant when higher than 0.5. Among

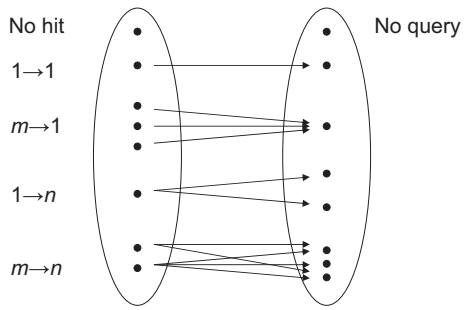

Fig. 3 BLAST-based contig annotation – number of hits.

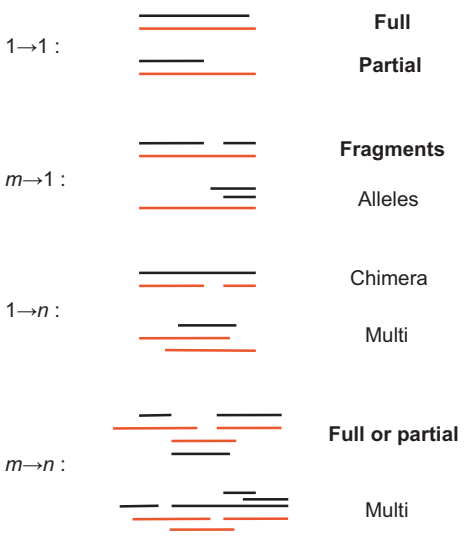

Fig. 4 BLAST-based contig annotation – overlap between hits and/or queries.

$m \rightarrow 1$ contigs, those showing no significant overlap with any other contigs were called 'fragments', whereas those overlapping with at least one other contig were called 'alleles'. Similar overlap calculations were made between hits of the $1 \rightarrow n$ category. Among the $1 \rightarrow n$, contigs whose hits showed no significant overlap were called 'chimera', the others were called 'multi' (for multigene family). Among the $m \rightarrow n$, we distinguished two cases. When m and n were equal, and when the m first hits of the m contigs were distinct from each other, contigs were called 'full' or 'fragments', depending on their alignment length. In all other cases, $m \rightarrow n$ contigs were called 'multi'. Finally, the 'no hit' contigs were BLASTed against the reference genome and called 'DNA' when a significant hit was found or 'other' otherwise. Here, BLAST hits were considered significant when sequence identity between query and hit was more than 80% (70% in *Lepus*) across alignment length >150 base pairs, irrespective of query (and reference) sequence length. Please note that the 'allele' category should include alternatively spliced isoforms of a single gene, as well as true allelic variants.

Results

Data sets

Table 1 lists the main characteristics of the newly generated sequence data sets analysed in this study, which were fairly homogeneous across species. The 454 experiment yielded fewer reads and a lower average read length in *Lepus granatensis* than in the other species, whilst Illumina was most successful in *Emys orbicularis*. In the 'mix' data sets, 454 sequences amounted to 1–1.5% of the total number of reads and 3–5% of the total number of base pairs, depending on the species.

Transcriptome assemblies: quantitative assessment

Various assembling strategies were followed, either using only 454 or only Illumina (Fig. 1), or using both (Fig. 2). Table 2 compares data sets and assembling strategies in terms of contig numbers and contig lengths in *Ciona intestinalis* and *L. granatensis*. Note that only contigs longer than 200 bp were considered here. Regarding the Illumina data sets, applying just Abyss yielded a very large number of short contigs (not shown). Applying Cap3 twice consecutively to Abyss-generated contigs yielded a relatively low number of relatively long contigs. Cap3 could not be directly applied to Illumina data sets because of computational limitations. Here, Abyss essentially compressed reads into a reasonable number of little-assembled contigs to be sent to Cap3, which did most of the assembling task. With these data sets, a single Cap3 run after Abyss assembled substantially less than two, and three consecutive runs did not assemble substantially more than two (data not shown).

Illumina-based contigs were slightly shorter, on average, than 454-based contigs. However, mean contig length is not an ideal indicator of assembly success. Figure 5 shows that the distribution of contig lengths is quite different in 454-based and Illumina-based assemblies. Assembly B (Illumina) includes a larger fraction of contigs shorter than 400 bp, which tend to shift the mean downwards, but the right tail of the distribution is long and heavy, which is the indicative of a relatively large number of well-assembled cDNAs. This is expressed by the higher N50 value of assembly B than of assemblies A – N50 is the contig length such that equal or longer contigs amount half of total assembly length. Note that the average cDNA length in the reference transcriptome is 1624 in *O. cuniculus* and 2010 in *C. intestinalis* A.

Table 1 Data sets.

Data set	Species	Technology	Normalized	Kreads	Mbp	Mean lg
Ciona_454	*Ciona intestinalis* B	454	Yes	512	174	340
Ciona_illu	*C. intestinalis* B	Illumina	No	35 839	3280	91
Ciona_mix	*C. intestinalis* B	Both		36 351	3454	
Lepus_454	*Lepus granatensis*	454	Yes	372	99	267
Lepus_illu	*L. granatensis*	Illumina	No	38 283	3304	87
Lepus_mix	*L. granatensis*	Both		38 655	3403	
Ostrea_454	*Ostrea edulis*	454	Yes	557	159	284
Ostrea_illu	*O. edulis*	Illumina	No	37 370	3395	90
Ostrea_mix	*O. edulis*	Both		37 927	3554	
Emys_454	*Emys orbicularis*	454	Yes	570	203	355
Emys_illu	*E. orbicularis*	Illumina	No	47 544	4233	89
Emys_mix	*E. orbicularis*	Both		48 114	4406	
Messor_454	*Messor barbarus*	454	No	678	251	370
Messor_illu	*M. barbarus*	Illumina	No	38 008	3388	89
Messor_mix	*M. barbarus*	Both		38 686	3639	

Table 2 Assemblies: quantitative assessment.

	Data set	Method	Contigs	Mean lg	Median lg	N50	Assembly lg (Mb)	Hit genes
A	Ciona_454	Cap3	24 515	671	540	713	16.5	11 509
B	Ciona_illu	Abyss + Cap3	27 426	574	380	769	15.8	12 915
C	Ciona_mix	Merge reads	29 097	571	399	721	16.6	11 584
D	Ciona_mix	Merge contigs	27 956	726	529	891	20.3	11 966
D''	Ciona_mix	Refined D	4962	1631	1476	1686	8.1	6067
A	Lepus_454	Cap3	34 463	544	440	574	18.8	11 433
B	Lepus_illu	Abyss + Cap3	38 540	574	357	785	22.1	13 791
C	Lepus_mix	Merge reads	43 037	640	391	916	27.5	14 874
D	Lepus_mix	Merge contigs	45 151	657	412	908	29.7	14 999

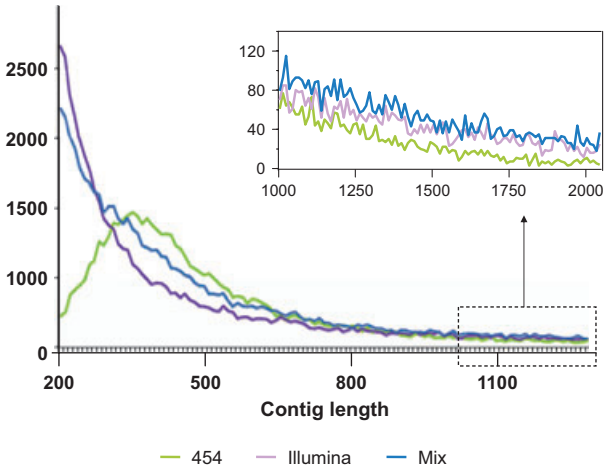

Fig. 5 Distribution of contig lengths for three transcriptome assemblies in *Lepus granatensis*. Contig length distributions are shown for assemblies A (454), B (Illumina) and D (mix). Top-right panel: zoom on the tail of the distribution.

Mixed data sets were analysed in two distinct ways (see Fig. 2). When 454 and Illumina reads were merged and sent to Abyss + Cap3 (assembly C), results were similar to the Illumina-only analyses (assembly B). In *C. intestinalis*, the best assemblies were obtained when 454 and Illumina contigs were combined (assembly D). The 'merge contigs' strategy yielded a smaller number of longer contigs than 454-only and Illumina-only assemblies. The difference between C and D was less clear in *L. granatensis*. The benefit of combining 454 and Illumina data sets was substantial. In both species, contigs from the D assembly were more numerous, and on average 15–25% longer, than those predicted from assembly B (and see Fig. 5).

Transcriptome assemblies: qualitative assessment

Ideally, for comparative genomic purposes, we would like to assemble exactly one consensus sequence per gene of the target species. We developed a BLAST-based procedure to assess the quality of a given set of predicted contigs with respect to this goal. Each contig of each assembly was annotated as either 'full', 'fragment', 'chimera', 'allele', 'multi', 'DNA' or 'other' (Figs 3 and 4, see Methods). Contigs from the 'full' and 'fragment' categories most probably correspond to correct predictions. The 'multi', 'allele' and 'other' categories are made of misleading contigs, which, if trusted, might result in false SNP calling or erroneous orthology annotations. Contigs from the 'chimera' category would be problematic for gene annotation, but not for SNP discovery or comparative purposes. They are likely to be easily detected by BLAST against even a distantly related reference. Finally, the 'DNA' category presumably includes both correct and incorrect predictions.

The relative contributions of each category to the various *Ciona* and *Lepus* assemblies built in this study are listed in Table 3. In both species, the proportion of 'full' + 'fragment' – the unambiguously correct predictions – was quite low: roughly 35% in *C. intestinalis* and 22% in *L. granatensis*. The proportion of 'full' tended to be higher in Illumina-based and especially in mix assemblies. 'Chimera' contigs were uncommon and 'DNA' quite numerous. Split alleles amounted >10% of predicted contigs in *C. Intestinalis*, but <4% in *L. granatensis*. *Ciona intestinalis* is a highly polymorphic species (Nydam & Harrison 2010), and allele splitting is obviously an issue here.

Besides alleles, the main difference between the two species was in the 'other' category, which was very low in *C. intestinalis*, but substantial in *L. granatensis*. When we recalculated the proportion of each category excluding 'other' contigs, then the percentage of 'full' + 'fragment' was similar between the two species. When we BLASTed the 13 457 *L. granatensis* contigs annotated as 'other' onto the nt (non-redundant genomic) database using a less stringent criterion (*e*-value < 10^{-5}), we found that 20% of these sequences had at least one hit, mostly in mammals (94%). The ~11 000 remaining sequences were

Table 3 Assemblies: qualitative assessment.

	Data set	Method	Full (%)	Fragment (%)	Chimera (%)	Allele (%)	Multi (%)	DNA (%)	Other (%)
A	Ciona_454	Cap3	4.0	33.5	0.3	13.4	16.2	30.4	2.2
B	Ciona_illu	Abyss + Cap3	4.8	31.1	0.4	11.5	17.3	31.8	3.2
C	Ciona_mix	Merge reads	4.6	31.3	0.3	10.1	16.8	33.7	3.3
D	Ciona_mix	Merge contigs	5.9	27.6	0.5	10.8	17.6	34.7	3.0
D''	Ciona_mix	Refined D	34.0	27.1	4.8	4.9	16.5	12.3	0.4
A	Lepus_454	Cap3	1.9	19.0	0.7	3.3	14.9	38.6	21.5
B	Lepus_illu	Abyss + Cap3	5.5	18.6	0.7	0.3	15.7	42.1	17.2
C	Lepus_mix	Merge reads	6.0	17.1	0.5	0.3	14.9	32.5	28.7
D	Lepus_mix	Merge contigs	5.4	15.4	0.5	1.9	14.7	32.2	29.8

much shorter (mean length: 347 bp; assembly average: 657 bp) and less covered (mean coverage: 10.8×; assembly average: 75.4×) than the average. This suggests that the 'other' category in *L. granatensis* mainly comprise misassembled short reads, as can be expected in a highly repetitive genome. This BLAST search does not suggest that *L. granatensis* samples have been strongly affected by viral or bacterial contamination.

Assembly filtering

For each contig of the Ciona_mix and Lepus_mix F assemblies, contig length and contig coverage were recorded. Coverage was obtained by mapping Illumina reads to contigs using Bowtie 0.12.5 (Langmead *et al.* 2009; default parameters), and for each contig, dividing the sum of mapped read lengths by contig length. Assemblies were filtered using various thresholds for minimal length and minimal coverage, and contigs were re-annotated. Figure 6 shows the numbers of contigs in each category for three filtering thresholds in *C. intestinalis*. The relative proportion of the 'full' and 'chimera' categories was substantially increased when short and low-expressed contigs were discarded, consistent with O'Neil *et al.* (2010). The 'multi' and 'DNA' classes were considerably depleted, and 'alleles' essentially disappeared. This occurred at the cost of a marked decrease in the size of the 'fragment' class and in the total assembly size. The statistics of the bottom assembly in Fig. 6 (length > 1000 bp, coverage > 12×), called D'', are given in Tables 2 and 3 for comparison.

Figure 7 shows the relationship between the total number of contigs and the proportion of correct predictions, here defined as the union of 'full', 'fragment' and 'chimera' categories, for a variety of filtering conditions in *C. intestinalis*. This figure reveals the existence of a trade-off between the quantity and quality of contigs when such filters are applied. A similar picture was obtained in *L. granatensis*, in which the percentage of 'DNA' and 'other' categories was dramatically reduced when short, low-coverage contigs were filtered out. We

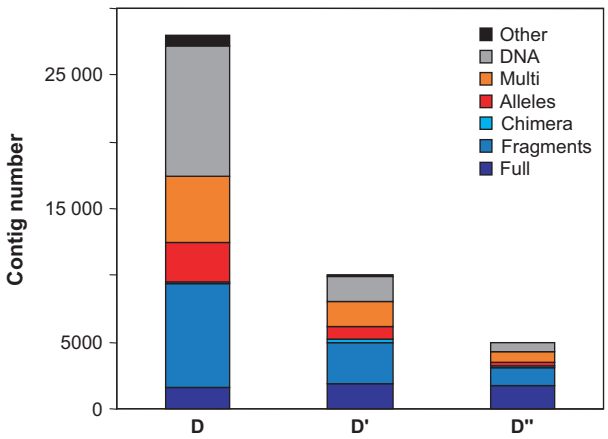

Fig. 6 Relative proportions of contig classes for three filtering stringency levels in *Ciona intestinalis*. D: no filtering; D': contig length > 600 bp, contig coverage > 4×; D'': contig length > 1000 bp, contig coverage > 12×.

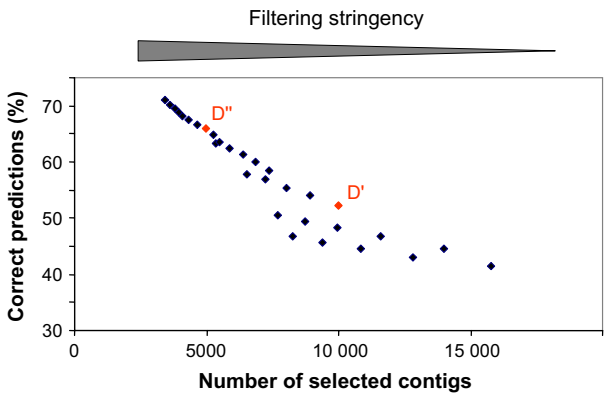

Fig. 7 Relationship between contig number and proportion of correct predictions across filtering conditions in *Ciona intestinalis*. Each dot stands for a filtering condition, that is, a pair of length (>200, >400, >600, >800, >1000, >1200) and coverage (>4×, >8×, >12×, >16×, >20×) thresholds. Red dots indicate the two conditions that are represented in Fig. 6. This figure is available in colour online at wileyonlinelibrary.com.

applied the ORF-Predictor program (http://proteomics.ysu.edu/tools/OrfPredictor.html) to identify the longest open reading frames (ORF) in each of the 611 long, high-coverage *C. intestinalis* contigs annotated as 'DNA' in filtered assembly D''. A total of 85% of these sequences contained an ORF longer than 200 bp and 58% contained an ORF longer than 400 bp. This suggests that a substantial proportion of these sequences are true coding RNA absent from the reference transcriptome, and that the *Y*-axis in Fig. 7 underestimates the true percentage of correct predictions in filtered assemblies.

Additional taxa

We applied the same four (A–D) assembling methods to similar data sets generated in *E. orbicularis* (turtle), *Ostrea edulis* (oyster) and *Messor barbarus* (ant). Results were largely consistent with the *C. intestinalis* and *L. granatensis* analyses, confirming the superiority of Illumina-only over 454-only, and of mix over Illumina-only in this study (Table 4). In each of *O. edulis*, *E. orbicularis* and *M. barbarus*, the 'merge contig' approach performed better than the 'merge reads' one, confirming the *C. intestinalis* result. Interestingly, the ant 454 data set, which was obtained from non-normalized cDNA library, yielded the lowest number of contigs, as compared to normalized data sets. This did not negatively affect the mix assemblies in ant, which showed the highest N50 value across the five considered species.

Alternative assembling programmes

Besides Cap3 and Abyss, we tried a number of publicly available, published assembling programmes with the *C. intestinalis* data set (Table 5). As far as 454 data were concerned, both Celera and Mira returned more numerous, shorter contigs than Cap3. Qualitatively, Celera and especially Mira returned a much higher (21% and 28%, respectively) proportion of split alleles than Cap3 (13.4%). Overall, we found that the main results obtained with Cap3 were still valid when Celera or Mira was used, and that Cap3 was the most robust of the three programmes for these data sets. The commercial Newbler program was also tried. It returned a smaller number of longer contigs than Cap3, which collectively corresponded to 8837 *C. intestinalis* genes, as compared to 11 509 with Cap3 – a 23% decrease. This suggests that Newbler tends to assemble long cDNA and discard short ones. Despite this apparent, implicit filtering, the quality of Newbler contigs was not improved, as compared to Cap3 contigs. To compare programmes in a perhaps fairer way, we selected from each assembly the 10 000 longest contigs and re-calculated our quantitative and qualitative assessment criteria (Table 5, bottom). In this comparison, Cap3, Mira and Newbler yielded a similar N50 and Cap3 outperformed the other programmes in terms of contig reliability.

Regarding the Illumina data set, three assembling methods were used in combination with Cap3 and compared with our main Abyss + Cap3 analysis. The Soap_de_novo_Trans program, when substituted for Abyss, performed less efficiently, both quantitatively and qualitatively. Trinity yielded results similar to Abyss, but appeared computationally much more demanding, especially RAM wise. Then, we varied *k*-mer length in Abyss, using $k = 60$ (our main analysis), 50, 40 or 30. Decreasing *k*-mer length yielded higher numbers of contigs of reduced average length (not shown). Then, the four assemblies were combined in a single one using the TransAbyss strategy (Robertson *et al.* 2010), and Cap3 was applied. The final assembly, called TAbyss + Cap3 in Table 5, appeared quite similar to the one we obtained with a single *k*-mer length of 60 bp, although the N50 was a bit lower. Similar comparisons were performed with the mix data set, and the same trends appeared, with Abyss + Cap3 overperforming Soap_de_novo_Trans + Cap3, and being more efficient than Trinity + Cap3 (not shown). Importantly, these re-analyses in *C. intestinalis* confirmed the relative performances of our A, B, C and D strategies, irrespective of the programmes used. So the main conclusions of our Abyss + Cap3-based analyses are not dependent on software choice.

Discussion

In this study, we designed a BLAST-based quality assessment of transcriptome assembly from NGS data and applied it to newly generated data sets combining 454

Table 4 Transcriptome assemblies in *Emys orbicularis* (turtle), *Ostrea edulis* (oyster) and *Messor barbarus* (ant).

	Emys			*Ostrea*			*Messor*		
	Contigs	Mean lg	N50	Contigs	Mean lg	N50	Contigs	Mean lg	N50
A	38 949	590	564	54 033	562	577	17 695	612	688
B	44 383	549	740	64 853	547	717	26 845	592	846
C	59 628	559	673	80 001	568	724	29 565	651	945
D	59 883	615	719	85 093	612	746	31 522	670	944

Table 5 Software comparisons in *Ciona intestinalis*.

Data	Software	Contigs	N50	Assembly lg (Mb)	Hit genes	Full (%)	Fragment (%)	Allele (%)	Multi (%)	DNA (%)
454	Cap3	24 515	713	16.5	11 509	4.0	33.5	13.4	16.2	30.4
454	Mira	33 196	650	21.1	10 977	0.9	20.9	27.8	18.5	29.6
454	Celera	25 669	495	12.6	10 085	1.3	29.4	20.6	16.6	28.9
454	Newbler	14 243	997	12.3	8837	5.2	31.3	20.7	17.7	23.6
illu	Abyss + Cap3	27 426	769	15.8	12 915	4.8	31.1	11.5	17.3	31.8
illu	TAbyss + Cap3	28 098	683	15.3	10 399	3.4	33.0	11.1	16.9	25.5
illu	Soap + Cap3	18 523	731	10.2	8621	3.8	34.8	10.4	17.3	29.4
illu	Trinity + Cap3	33 490	623	16.9	11 424	2.5	39.3	8.4	13.7	31.6
454	Cap3	10 000	1010	9.9	8326	12.6	40.3	7.7	15.7	21.8
454	Mira	10 000	1011	10.2	7529	7.4	29.5	24.4	17.8	19.7
454	Celera	10 000	682	6.8	6754	3.9	33.1	21.4	16.2	23.6
454	Newbler	10 000	1114	10.4	7540	7.5	30.0	23.5	17.8	19.8
illu	Abyss + Cap3	10 000	1108	9.8	8180	15.0	41.1	4.2	16.3	21.1
illu	TAbyss + Cap3	10 000	1025	5.7	8150	12.2	48.2	2.8	14.7	20.2
illu	Soap + Cap3	10 000	917	7.7	7082	8.1	40.4	7.6	14.7	26.6
illu	Trinity + Cap3	10 000	1048	9.9	8087	9.5	52.8	4.2	11.0	19.8

and Illumina data in two non-model animal species, with the goal of assessing the reliability of *de novo* transcriptome assemblies in the absence of a closely related reference. Contigs were annotated as 'full', 'fragment', 'chimera', 'allele', 'multi' 'DNA' or 'other', depending on the number of and overlap between BLAST hits against a reference transcriptome and genome. For comparative and population genomic purposes, the ideal assembly includes one sequence per gene of the target genome, not one sequence per transcript of the target tissue: we want distinct transcripts of a given gene to be merged into a single cDNA-like sequence. Consequently, our quality criteria were more stringent, and arguably more informative in the context of population genomics, than just the percentage of contigs-to-reference or reference-to-contigs BLAST hits, typically used in existing benchmarks (e.g. Papanicolaou *et al.* 2009; Kumar & Blaxter 2010).

Noisy de novo *transcriptome assemblies*

The newly introduced approach obviously depends on the reliability of the references used and of the criterion used to define significant BLAST hits. The reference transcriptomes used here were 5–10% divergent from target species (Carneiro *et al.* 2010; Nydam & Harrison 2010) when 80% sequence similarity, over 50% of the query or hit length, was sufficient to consider a hit significant. We believe that these criteria are such that we would fail to link query to an available reference orthologue in only a tiny fraction of rapidly diverging sequences.

It might be, however, that the reference and target transcriptomes differ in terms of gene content. Some of our contigs might be true cDNAs missing from the reference because of recent duplication (in the target genome), or recent gene loss (in the reference genome), or annotation problems. Reciprocally, some genes or transcripts might be missing from the target species, but present in the reference, for similar reasons. Such situations would typically result in calling a 'multi', 'allele' or 'DNA' annotation of truly trustable contigs. Our annotation procedure is therefore probably a bit conservative or pessimistic. Some of the contigs which passed stringent filters regarding coverage and length, and were yet annotated as 'multi' or 'DNA', could be correctly assembled contigs, whose annotation results from a true difference between target and reference genomes, as revealed by our identification of long ORFs in 'DNA' contigs of the D'' assembly. The surprisingly large proportion of contigs annotated as 'DNA' in our BLAST-based assays presumably results in part from this effect, DNA contamination being another possible explanation. We suggest, however, that the problem of reference incompleteness in cDNA annotation is primarily relevant to multi-gene families – which, at any rate, are not good candidates for SNP definition and population genomic analyses. We note that this problem is general to any attempt to annotate a new sample from existing genomic resources, even when a fully sequenced genome from the same species is available, owing to within-species variations in gene copy number (e.g. Maydan *et al.* 2010; Scavetta & Tautz 2010; Wang *et al.* 2010b).

Besides this issue, one lesson to be drawn from this analysis is that the set of contigs obtained from typical NGS-based transcriptome data sets and standard assembling techniques includes a minority of unambig-

uously correct cDNA predictions – typically 20–35%, depending on the data and methods used. This was true even when the number of assembled contigs was similar to the number of genes in the target species (~15 000 in *Ciona intestinalis* and ~25 000 in *Lepus granatensis*), which might give the false impression of exhaustive transcriptome prediction. Half of these sets of contigs (at most) consisted of trustable, highly expressed (partial) cDNA, whilst the other half consisted of dubious predictions, owing to alternative splicing, transcriptional noise, low expression level or insufficient coverage, whose relevance for comparative or functional analysis is questionable. The high proportion of contigs annotated as 'DNA' is illustrative of the noisy nature of transcription in animals. Caution should therefore be exercised before interpreting predicted contigs as gene sequences when the transcriptome of a non-model species is newly assembled.

The proportion of unassembled alleles was high in the genetically diverse *C. intestinalis*, but low in *L. granatensis*. One could think of decreasing the stringency of assembling methods in species known to be highly polymorphic. Decreasing stringency, however, would probably increase the number of falsely assembled paralogues – that is, the 'multi' category. When the distribution of sequence divergence between paralogues, on one hand, and between alleles, on the other hand, truly overlap, errorless cDNA assembly is essentially impossible in the absence of additional information.

Comparing technologies and assembling strategies

Our quality assessment strategy was applied to various data sets, and various assembling methods, offering the opportunity to compare the different approaches. Regarding NGS technology, we found that 454-based and Illumina-based sets of assembled contigs have distinct properties (Fig. 5) and are optimally combined – please note that roughly the same amount of money was invested in the two technologies in this study. Illumina alone can lead to acceptable transcriptome assembly, but the benefit of adding even a small fraction of long reads can be substantial, as revealed by the comparison of assemblies A, B and D in Tables 2, 3 and 4. A similar conclusion was reached by Wall *et al.* (2009) and Boisvert *et al.* (2010) from simulations. Most of the *de novo* transcriptome analyses of non-model species published so far have been achieved with 454 only (e.g. Vera *et al.* 2008; Schwarz *et al.* 2009; Guo *et al.* 2010; Parchman *et al.* 2010; Schwartz *et al.* 2010), Chen *et al.* (2010) being a remarkable exception. Our analyses suggest that this popular 454-only strategy is far from optimal.

The difference between the two technologies was reinforced here by the fact that in four species, cDNA libraries were normalized before 454 sequencing, further reducing the per transcript coverage level. This study suggests that normalization is not necessarily a good idea, unless deep coverage is to be expected. The one species we sequenced without normalizing cDNA libraries (*Messor barbarus*) yielded the largest N50. Normalizing leads to more fragmented assemblies, as suggested by the comparison between the non-normalized ant and the other species. If the goal is to assemble the maximal number of correctly predicted full-length or partial cDNAs for comparative or population genomic purposes, our advice would be to not normalize libraries, although more data would be required to formally address this question.

Our conclusions are mostly based on the analyses performed using a combination of Cap3 and Abyss, which were selected because they are well-established references in the field (and see Chen *et al.* 2010). We assessed a number of alternative assembling programmes and found that Abyss and Cap3 were among the best-performing ones, both quantitatively and qualitatively, whilst requiring reasonable amounts of computing resources. This is perhaps a bit surprising, knowing that these two programmes were designed for genomic data. The transcriptome-specific Soap_de_novo_Trans, Trinity and TransAbyss did not improve significantly over Abyss, and Mira performed worse than Cap3. We note, however, that these programmes were primarily designed to identify and separate splicing variants, whereas our goal here was to gather the various transcripts of a given gene in a single contig. Obviously, this analysis does not imply that Abyss + Cap3 is the best-performing solution for all possible applications of NGS transcriptomic data.

Remarkably, the relative level of performance of the various approaches tested was similar in the five species used in this study – two vertebrates (mammal and sauropsid) and three invertebrates (tunicate, mollusc and insect). These species were chosen as the representatives of the phylogenetic, ecological and genomic diversity of metazoans. Heterozygosity is typically high in marine invertebrates such as *Ostrea edulis* and *C. intestinalis* (Small *et al.* 2007), but typically low in relatively large terrestrial vertebrates such as *Emys orbicularis*. The genome of mammals is large (~3.5 Gb) and highly repetitive, in contrast to the much more compact *C. intestinalis* (0.2 Gb) and *M. barbarus* (0.26 Gb) genomes. This diversified panel of target species thus provides a variety of conditions under which assembling strategies were assessed. Quantitative (contig number and length) and qualitative (proportion of correct predictions) criteria were quite consistent across species in their ranking of the various methods.

Reference-free assembly: filtering strategy

The high proportion of incorrect cDNA predictions we uncovered could suggest that high-throughput transcriptome analysis is not a sensible approach to safely identify SNPs, or orthologues, in the absence of a reference genome. However, our analysis reveals that correct and incorrect predictions tend to differ in length and coverage (and see O'Neil *et al.* 2010). By filtering assemblies based on these two criteria, one can substantially increase the proportion of reliable predictions. In *C. intestinalis*, the ~5000 longest and most covered contigs included ~70% of unambiguously correct predictions and a majority of full-length cDNA. This figure, furthermore, is most probably an underestimation of the actual proportion of correctly assembled contigs, as discussed above. This result, together with the remarkably similar behaviour of the different species used in this study, opens promising perspectives for transcriptome-based comparative genomics in animals and the exploration of the molecular diversity of non-model taxa.

Acknowledgements

This work was supported by a European Research Council (ERC) grant to Nicolas Galtier (ERC PopPhyl 232971). Ylenia Chiari was partially supported by a FCT postdoctoral grant SFRH/BPD/73515/2010. Sampling of hares was funded by Portuguese Science and Technology Foundation (FCT) project with reference PTDC/BIA-EVF/111931/2009. JMF benefits of a post-doc grant from FCT with reference SFRH/BPD/43264/2008. We are grateful to P. Alves, C. Ayres, G. Ballantyne, N. Bierne, M. Cantou, L. Fast Jensen, D. Jiang, S. Lapègue, J. Lourenco, A. Lugagne, V. Molnar, I. Nodet, J. Romiguier, M. Vamberger, B. Vercaemer, O. Verneau and M. Zuffi for their help with animal sampling and to G. Dugas for his help with computational issues. This is publication number ISEM 2012–0034.

References

Boisvert S, Laviolette F, Corbeil J (2010) Ray: simultaneous assembly of reads from a mix of high-throughput sequencing technologies. *Journal of Computational Biology*, **17**, 1519–1533.

Caputi L, Andreakis N, Mastrototaro F, Cirino P, Vassillo M, Sordino P (2007) Cryptic speciation in a model invertebrate chordate. *Proceedings of National Academy of Sciences of the United States of America*, **104**, 9364–9369.

Carneiro M, Blanco-Aguiar JA, Villafuerte R, Ferrand N, Nachman MW (2010) Speciation in the European rabbit (*Oryctolagus cuniculus*): islands of differentiation on the X chromosome and autosomes. *Evolution*, **64**, 3443–3460.

Chen S, Yang P, Jiang F, Wei Y, Ma Z, Kang L (2010) De novo analysis of transcriptome dynamics in the migratory locust during the development of phase trait. *PLoS ONE*, **5**, e15633.

Chevreux B, Pfisterer T, Drescher B *et al.* (2004) Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. *Genome Research*, **14**, 1147–1159.

Chiari Y, Galtier N (2011) RNA extraction from sauropsids blood: evaluation and improvement of methods. *Amphibia-Reptilia*, **32**, 136–139.

Elmer KR, Fan S, Gunter HM *et al.* (2010) Rapid evolution and selection inferred from the transcriptomes of sympatric crater lake cichlid fishes. *Molecular Ecology*, **19**, 197–211.

Gayral P, Weinert L, Chiari Y, Tsagkogeorga G, Ballenghien M, Galtier N (2011) Next-generation sequencing of transcriptomes: a guide to RNA isolation in non-model animals. *Molecular Ecology Resources*, **11**, 650–661.

Grabherr MG, Haas BJ, Yassour M *et al.* (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, **29**, 644–652.

Guo S, Zheng Y, Joung JG *et al.* (2010) Transcriptome sequencing and comparative analysis of cucumber flowers with different sex types. *BMC Genomics*, **11**, 384.

Huang X, Madan A (1999) CAP3: a DNA sequence assembly program. *Genome Research*, **9**, 868–877.

Hudson ME (2008) Sequencing breakthroughs for genomic ecology and evolutionary biology. *Molecular Ecology Resources*, **8**, 3–17.

Iannelli F, Pesole G, Sordino P, Gissi C (2007) Mitogenomics reveals two cryptic species in *Ciona intestinalis*. *Trends in Genetics*, **9**, 419–422.

Kumar S, Blaxter ML (2010) Comparing de novo assemblers for 454 transcriptome data. *BMC Genomics*, **11**, 571.

Künstner A, Wolf JB, Backström N *et al.* (2010) Comparative genomics based on massive parallel transcriptome sequencing reveals patterns of substitution and selection across 10 bird species. *Molecular Ecology*, **19**, 266–276.

Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, **10**, R25.

Li R, Zhu H, Ruan J *et al.* (2010) De novo assembly of human genomes with massively parallel short read sequencing. *Genome Research*, **20**, 265–272.

Maydan JS, Lorch A, Edgley ML, Flibotte S, Moerman DG (2010) Copy number variation in the genomes of twelve natural isolates of *Caenorhabditis elegans*. *BMC Genomics*, **11**, 62.

Myers EW, Sutton GG, Delcher AL *et al.* (2000) A whole-genome assembly of Drosophila. *Science*, **287**, 2196–2204.

Nydam ML, Harrison RG (2010) Polymorphism and divergence within the ascidian genus *Ciona*. *Molecular Phylogenetics and Evolution*, **56**, 718–726.

O'Neil ST, Dzurisin JD, Carmichael RD, Lobo NF, Emrich SJ, Hellmann JJ (2010) Population-level transcriptome sequencing of nonmodel organisms *Erynnis propertius* and *Papilio zelicaon*. *BMC Genomics*, **11**, 310.

Papanicolaou A, Stierli R, Ffrench-Constant RH, Heckel DG (2009) Next generation transcriptomes for next generation genomes using est2assembly. *BMC Bioinformatics*, **10**, 447.

Parchman TL, Geist KS, Grahnen JA, Benkman CW, Buerkle CA (2010) Transcriptome sequencing in an ecologically important tree species: assembly, annotation, and marker discovery. *BMC Genomics*, **11**, 180.

Paszkiewicz K, Studholme DJ (2010) De novo assembly of short sequence reads. *Briefings in Bioinformatics*, **11**, 457–472.

Renaut S, Nolte AW, Bernatchez L (2010) Mining transcriptome sequences towards identifying adaptive single nucleotide polymorphisms in lake whitefish species pairs (*Coregonus* spp. Salmonidae). *Molecular Ecology*, **19**, 115–131.

Robertson G, Schein J, Chiu R *et al.* (2010) De novo assembly and analysis of RNA-seq data. *Nature Methods*, **7**, 909–912.

Sammeth M (2009) Complete alternative splicing events are bubbles in splicing graphs. *Journal of Computational Biology*, **16**, 1117–1140.

Scavetta RJ, Tautz D (2010) Copy number changes of CNV regions in intersubspecific crosses of the house mouse. *Molecular Biology and Evolution*, **27**, 1845–1856.

Schwartz TS, Tae H, Yang Y *et al.* (2010) A garter snake transcriptome: pyrosequencing, de novo assembly, and sex-specific differences. *BMC Genomics*, **11**, 694.

Schwarz D, Robertson HM, Feder JL *et al.* (2009) Sympatric ecological speciation meets pyrosequencing: sampling the transcriptome of the apple maggot *Rhagoletis pomonella*. *BMC Genomics*, **10**, 633.

Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I (2009) ABySS: a parallel assembler for short read sequence data. *Genome Research*, **19**, 1117–1123.

Small KS, Brudno M, Hill MM, Sidow A (2007) Extreme genomic variation in a natural population. *Proceedings of National Academy of Sciences of the United States of America*, **104**, 5698–5703.

Surget-Groba Y, Montoya-Burgos J (2010) Optimization of *de novo* transcriptome assembly from next-generation sequencing data. *Genome Research*, **20**, 1432–1440.

Vera JC, Wheat CW, Fescemyer HW *et al*. (2008) Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. *Molecular Ecology*, **17**, 1636–1647.

Wall PK, Leebens-Mack J, Chanderbali AS *et al*. (2009) Comparison of next generation sequencing technologies for transcriptome characterization. *BMC Genomics*, **10**, 347.

Wang GX, Ren S, Ren Y, Ai H, Cutter AD (2010a) Extremely high molecular diversity within the East Asian nematode *Caenorhabditis* sp. 5. *Molecular Ecology*, **19**, 5022–5029.

Wang X, Nahashon S, Feaster TK, Bohannon-Stewart A, Adefope N (2010b) An initial map of chromosomal segmental copy number variations in the chicken. *BMC Genomics*, **11**, 351.

Wolf JB, Bayer T, Haubold B, Schilhabel M, Rosenstiel P, Tautz D (2010) Nucleotide divergence vs. gene expression differentiation: comparative transcriptome sequencing in natural isolates from the carrion crow and its hybrid zone with the hooded crow. *Molecular Ecology*, **19**, 162–175.

P.G., L.W., M.B., Y.C., G.T., and J.M.-F. obtained samples from the wild and achieved RNA extractions. V.C., P.G., G.T., and K.B. performed the bioinformatic developments and analyses. V.C., N.G., and V.R. designed the BLAST-based annotation strategy. N.G. wrote the manuscript..

Data accessibility

Raw data sets can be downloaded from http://kimura.univ-montp2.fr/PopPhyl. Assemblies are available from the Dryad data repository (doi: 10.5061/dryad.5g32f94b). The successive steps required to achieve transcriptome assemblies B (illumina) and D (mix) under the Galaxy platform are distributed as a workflow file, together with appropriate Perl/Python/xml wrappers. This resource is available from http://kimura.univ-montp2.fr/PopPhyl/resources/datasets/popphyl-galaxy.tar.gz.