Challenges and solutions for transcriptome assembly in non-model organisms with an application to hybrid specimens

UNGARO Arnaud*, PECH Nicolas*, MARTIN Jean-François#, MCCAIRNS R.J. Scott*§, CHAPPAZ Rémi*, GILLES André*.

* Aix Marseille Université, CNRS, IRD, IMBE – UMR 7263, Équipe Évolution Génome Environnement, Centre Saint-Charles, Case 36, 3 place Victor Hugo, 13331 Marseille Cedex 3, France.

Montpellier SupAgro, UMR CBGP, F-34988 Montferrier-sur-Lez, France

§ Institut National de la Recherche Agronomique, UMR ESE, 35042 Rennes, France

contact@arnaud-ungaro.fr

nicolas.pech@univ-amu.fr

jean-francois.martin@supagro.fr

scott.mc-cairns@supagro.inra.fr

remi.chappaz@univ-amu.fr

andre.gilles@univ-amu.fr

Abstract

Analyses of high-throughput transcriptome sequences of non-model organisms are based on three main approaches: de novo assembly, genome-guided assembly, or direct read to genome mapping (DGM). We describe a flexible DGM pipeline, and demonstrate its performance by using simulated reads of lengths corresponding to those generated by the most common sequencing platforms, and over a realistic range of genetic divergence. We also evaluate the performance of a combined pipeline (*de novo* + DGM) via simulation and empirically, using data from two hybridizing Cyprinid fish species. Finally, we explore the assignation of F1 hybrids reads to their parental species, and discuss the implications of erroneous assignations on gene expression studies. Our DGM pipeline recovers 94.8% of the

genes irrespective of read length at 0% divergence; however, assignation rate of reads is negatively impacted both by increasing divergence level and reducing read lengths. Likewise, our combined *de novo* + DGM pipeline outperforms *de novo* analyses alone at all levels of divergence and the read length.

Keywords

De novo assembly, direct genome mapping, RNA-seq, hybridization, *Danio rerio*

Running title

Transcriptome assembly in non-model organisms

Introduction

Synthesis and maturation of RNAs is an elemental cog in the cellular machinery. Although inherently noisy, transcriptional variation can be associated with basal/fundamental processes such as enzyme activity (Nikinmaa *et al.* 2013) and protein production (Bar-Even *et al.* 2006). Transcriptional variation may also underlie complex morphological differences (Ayroles *et al.* 2009; Alvarado *et al.* 2015), and may itself be a target for natural selection (Leder *et al.* 2015). As such, the quantification of RNA abundance remains an essential link in deciphering the genotype-phenotype map. In this context, transcriptome-based inference is an initial and requisite basis for studying gene expression (Wang *et al.* 2009; Qian *et al.* 2014). After two decades of RNA microarrays (Schena *et al.* 1995), RNA-seq has democratized the analysis of transcriptomes for any non-model organism. This technological innovation has spread to several new uses in multiple domains in the life sciences such as transcript annotation (Grabherr *et al.* 2011; Wolf 2013), *cis* and *trans* regulation of allopolyploid species (Shi *et al.* 2012), speciation (Shi *et al.* 2012; Poelstra *et al.* 2015), heat stress (Pratlong *et al.* 2015), and ecotoxicology (Elran *et al.* 2014).

Although RNA-seq may be applied to non-model organisms, meaningful inference from transcriptome sequences in the absence of a reference genome is not a trivial problem, and one that is particularly perilous for hybrid specimens. An essential step in the RNA-seq pipeline for non-model organisms is the assembly of short reads into longer contigs that facilitate construction of a reference transcriptome and/or direct sequence annotation. The two most common approaches are *de novo* assembly (i.e. assembling reads free of any reference genome) and genome-guided transcriptome assembly (i.e. mapping and assembling reads

using a closely related reference genome). Recently, Ockendon *et al.* (2016) have proposed an alternative method for RNA-seq data, termed direct genome/transcriptome mapping (DGM). In contrast to guided assembly which uses positional information in the reference genome/transcriptome to ultimately reconstruct complete transcript sequences of each gene in the target sample, in DGM sequencing reads themselves are directly annotated via alignment against the closest genome/transcriptome available.

For non-model organisms, Cahais *et al*. (2012) reconstructed transcriptomes through *de novo* assembly combining 454 and Illumina sequence reads, and explored the efficiency of the approach through BLAST search of assembled contigs against the taxonomically closest reference genome. Although they retrieved a great number of transcripts, and opened new opportunities of transcriptome inference for non-model organisms, they also found potential issues, namely: amplification bias in transcript coverage; sensitivity of alignment error due to paralogs and multigene families; production of artefactual chimeras; problems reconstructing transcript length, and; potentially misestimating allelic diversity. These issues were further confirmed in several subsequent articles dealing with the *de novo* strategy (Seehausen *et al.* 2014; Rana *et al.* 2016; Palma-Silva *et al.* 2016), although the extent of sensibility of each vary slightly depending on the software and analytical solutions used.

Vijay *et al*. (2013) compared *de novo* and genome guided transcriptome assemblies, concluding that the guided transcriptome approach is less sensitive to sequencing error rate and polymorphisms. The guided approach is also less sensitive to paralogous loci. However, by simulating polymorphic sequences, they also showed that guided transcriptome assembly is highly sensitive to genetic divergence between the reads and the reference transcriptome, even for divergence levels as low as 1-5% (Vijay *et al.* 2013). Likewise evolved structural differences between reference and target (e.g. indels, inversions) can generate bias/error in contig sequences inferred via guided assembly, due to its dependence on the reference as a template in mapping (Vijay *et al.* 2013; Huang *et al.* 2016).

Jain *et al*. (2013) successfully increased the efficiency of transcriptome assembly through combining *de novo* and genome-guided approaches, demonstrating its efficacy on *Arabidopsis* spp. and *Danio rerio*. This combined method is currently the most widely used to reconstruct transcriptome assemblies for non-model organisms (Jain *et al.* 2013; Huang *et al.* 2016). Recently, Ockendon *et al.* (2016) compared *de novo*, guided transcriptome and direct genome mapping (DGM) approaches using *in silico Drosophila* and primate transcriptomes. They clearly demonstrated that DGM outperforms both *de novo* and guided transcriptome assemblies in both the quantity and accuracy of genes detected. Nevertheless, the full capacities of DGM in reconstructing transcriptomes have not yet been explored, particularly with respect to problematic assemblies.

Reconstructing parental transcriptomes for hybrid specimens and analyzing their gene expression is a real conundrum. When past and present hybridization occur in lineages, the genomes of parental species get mixed (Seehausen *et al.* 2014), and hybrid transcriptomes show emergent properties in gene regulation (Wittkopp *et al.* 2004; Renaut *et al.* 2009; Czypionka *et al.* 2012; Shi *et al.* 2012; Dion-Cote *et al.* 2014). Hybridization impacts inference of transcriptomes through homogenization and reorganization of different parts of the genome (Liu *et al.* 2016). As a result, when hybrids occur in nature, transcriptomes from each parental species appear tangled in those specimens. In this case, the assignation of reads to the right parental species (i.e. avoiding false orthologous gene assignation) and the right gene (avoiding false paralogous gene assignation) is even more challenging and requires prior knowledge of the parental transcriptomes. This is a scenario in which DGM may facilitate the alignment process, but one that has not yet been evaluated formally.

In this paper we describe a flexible DGM pipeline and demonstrate its performance by using simulated Illumina reads generated from a well-known reference transcriptome. We estimate the assignation error rate of a read onto its gene by simulating reads of lengths corresponding to those generated by the most common sequencing platforms (100, 150, 200, 350 bases), and over a realistic range of genetic divergence between the simulated reads and the reference genome (0%, 5%, 15% and 30%). Whatever the approach used, genomic complexity can create difficulties in reconstructing transcriptomes. In particular, genome duplications exacerbate biases in transcriptome assembly because of the higher number of paralogous loci. Teleostean fishes in general, and cyprinids in particular, are good models for studying the effects of paralogs on transcriptome assembly as they display a complex genome with multiple rounds of duplication (Meyer & Van de Peer 2005). We therefore choose *Danio rerio* (Teleostean: Cyprinidae) as a reference transcriptome for this *in silico* analysis.

We also compare the performances of *de novo* transcriptome inference alone, and in combination with DGM. This is done first on the same simulated data as previously described when testing for the DGM method alone, using *Danio rerio* as a reference, then also applied on two divergent cyprinid species, *Parachondrostoma toxostoma* (Vallot, 1837) and *Telestes souffia* (Risso, 1827), for which no transcriptome or genome are available, again keeping *Danio rerio* as the genome reference for DGM. This allows for an empirical comparison of the relative performance of the two methods in the face of high divergence between target transcriptome and reference genome. Finally, to evaluate DGM as a potential solution to the problems associated with hybrids, we infer the transcriptome of F1 hybrids from the same two divergent cyprinid species (*P. toxostoma* and *T. souffia*) analyzed in previous steps.

Materials and Methods

The Direct Genome Mapping pipeline

We implemented a flexible DGM pipeline (Fig. 1A) as follows; note that all custom scripts are available at https://github.com/egeeamu/voskhod. Paired-end sequencing reads are filtered for over-representation (PCR duplicates or over-expression) using a custom Python script and then merged, when possible, with Pear (v 0.9.6; Zhang *et al.* 2014) using the following parameters: minimum overlap -v 8, scoring method -s 2, p-value -p 0.01, minimum assembly length -n 30, maximum assembly length -m 0. After merging – a step in the pipeline that is optional – all available reads are quality filtered using a custom script that trims 5' and 3' ends until the base Phred quality score > 13 (i.e. < 0.05 error rate). A read is rejected (and replaced by a sequence made up by 50 "Ns" to keep the parity between R1 and R2 files when it applies) if it displays at least one of the following conditions:

i) at least one base with a Phred quality score < 5;

ii) more than 10% of bases with a Phred quality score < 13;

iii) a mean Phred quality score for the read < 20;

iv) a length shorter than 30 bases.

Next, Nucleotide-Nucleotide BLAST (blastn; Camacho *et al.* 2009) is used to assign *gene-id(s)* to filtered reads by finding regions of similarity between the read and the reference genome(s)/transcriptome(s), which are stored in a local SQLite database(Hipp & SQLite Development Team 2015) with Word-size = 9, HSP > 70% and similarity > 70%. Each read is assigned the *gene-id* corresponding to the best hit. When identical scores (overall quality of an alignment) are obtained for multiple *gene-ids*, this information is stored in variable *Hit_multigene*. This variable is used to reflect the level of mismatches on paralogous genes or multigene families. When identical scores are obtained for a single *gene-id* from at least two distinct reference genomes/transcriptomes in the local database (when relevant), this information is stored in the variable *Hit_multispecies* along with the associated *gene-id*. This variable is used to reflect the level of mismatches with orthologous genes. The final output is a collection of database entries displaying the number of identified reads for each *gene-id* (see supplementary methods 1 for an example) stored in the local database and exported in a tabular file for further analysis (see supplementary methods 2 for computing time estimates).

Measuring DGM efficiency and performance for read assignation

Based on the published transcriptome of *Danio rerio* (version Zv10), we used the longest transcript for each gene i.e. 31,953 transcripts (Flicek *et al.* 2014) and simulated RNA-seq reads of known variability. We simulated reads for four length classes (100, 150, 200 and 350 bases, see supplementary methods 3 for details) representing the whole transcriptome at 10X coverage for each gene. Varying read lengths allows studying erroneous assignations caused by conserved homologous regions (e.g. causing assignation to multiple values for *gene-id*), the expected outcome being that longer reads will decrease ambiguity. Sequence divergence between simulated reads and the reference transcriptome was used to simulate confounding processes such as sequence error rate (estimated to 0.64% for R1 and 1.07% for R2 for the Illumina Miseq sequencer; Schirmer *et al.* 2015), polymorphism (ranging from 0.1% to 1%; Vijay *et al.* 2013), and species divergence. We mimicked sequence divergence by introducing random base errors into simulated reads at rates of 0%, 5%, 15% and 30%, as in Vijay *et al.* (2013) (see supplementary methods 3 for details).

The efficiency (percentage of output to input) was estimated from the number of genes recovered with the DGM pipeline (output), relative to the number of genes within the reference species' genome (input). We assessed the performance (i.e. quality of the output) with two metrics:

i) the recovery rate (*rr*), defined as the proportion of reads simulated from a *gene-id* and correctly assigned to this *gene-id*

ii) the specificity rate (*sr*), defined as the number of reads assigned to a *gene-id* that were simulated from this *gene-id* relative to the total number of reads assigned to that *gene-id*

For example, consider 100 reads simulated from a given *gene-id*. If the DGM pipeline assigns 80 from these reads to the same *gene-id*, then the *rr* is estimated as 80/100 = 0.8. Conversely, if we observe 150 reads assigned to this *gene-id* (80 reads from the same *gene-id* as source and 70 reads from other *gene-id(s)*) the specificity rate *sr* is estimated as 80/150 = 0.53.

We propose a typology of genes (Fig. 2A) based on the assignation of reads considering their recovery rate (*rr*) and specificity rate (*sr*).

i) The category *rr*=1 and *sr*=1 constitutes the 'perfect' genes, all generated reads for a gene are recovered for this gene and no read from another gene is "captured"

ii) The category *rr*=1 and *sr*<1 constitutes the 'recipient' genes, all the generated reads for a gene are recovered for this gene and at least one read from another gene is "captured"

iii) The category *rr*<1 and *sr*=1 constitutes the 'donor' genes, at least one generated read for a gene is not recovered for this gene and no read from another gene is "captured"

iv) The category *rr*<1 and *sr*<1 constitutes the 'mixed' genes, at least one generated read for a gene is not recovered for this gene and at least one read from another gene is "captured"

v) The category *rr*=0 and *sr*=N.A. constitutes the 'undetectable' gene, no assigned reads for this gene

Finally, as gene length is highly heterogeneous in transcriptomes (Grishkevich & Yanai 2014), we tested whether assignation success is impacted by this factor, with each gene's length estimated as the length of its longest transcript. The aforementioned performance metrics (*rr* & *sr*) are modeled as a function of gene length (treated as a continuous variable), read length, sequence divergence (each treated as categorical variables) and all interaction terms amongst factors using a mixed-effect logistic regression implemented in the lme4 package (Bates *et al.* 2014) for R (version 3.3.1; R Core Team 2013). Aforementioned model terms are treated as fixed-effects, and gene identity is included as a random factor (see Supplementary methods 4 for details).

Transcriptome assembly approaches

De novo assembly pipeline

In a recent study, Lu et al. (2013) compared various assembly software, including: Trinityrnaseq_r2012-04-27 (Grabherr *et al.* 2011), Oases (v0.2.06; Schulz *et al.* 2012) and trans-ABySS (v1.3.2; Robertson *et al.* 2010). The authors demonstrated that Trinity produces assemblies with the highest completeness and contiguity (using default parameters); thus, our *de novo* assembly pipeline is built on Trinity (v2.2.0; Grabherr *et al.* 2011; Haas *et al.* 2013; Fig. 1C). When paired-end reads are used, merging the R1 and R2 reads is unnecessary as Trinity uses the information from separate R1 and R2 paired files. The quality-based filtering procedure was the same as for the DGM pipeline, as implemented specifically for non-merged reads. The *de novo* assembly was done using Trinity 2.2.0 with *normalize_reads* and *stranded_library* options using the combination of R1 and R2 when available. The output of the assembly is a collection of generated contigs stored in the local database and exported in a single FASTA file.

Each contig was first identified/annotated by alignment against the transcriptome of the reference species (in our case *D. rerio,* version Zv10) using blastn (Word-size = 9, HSP > 70% and similarity > 70%), and used to populate a reference database. Only the HSP was

conserved and converted to reverse-complement when a strand differed from the reference. When multiple hits occurred (i.e. the same blastn score), the contig was assigned to the corresponding *gene-ids* and denoted with variable *Hit_multigene* in the database. This step is crucial to remove gap effects in some paralogous sequences with conserved regions. The output of this step is a collection of assigned contigs stored in the local database (i.e. the reconstructed *de novo* assembly) and exported as FASTA files, one for each *gene-id*. These files defined an inferred transcriptome of the non-model species that was stored into the local database.

Developing a pipeline combining *de novo* + DGM approaches

First, the *de novo* approach was used as previously described, producing a list of assigned contigs in the local database that was used in subsequent steps (Fig. 1C) . The next step re-uses the raw reads, merged with Pear when working with paired end sequencing. All reads are filtered on quality and aligned using the same parameters as in the DGM pipeline, although alignments are made against the local reference database that combines both the reference transcriptome used to assign contigs during the initial *de novo* step (*D. rerio* in our case) and the newly identified contigs (i.e. the output of the *de novo* assembly). This collection of assigned contigs is of particular interest as it provides the closest available reference, which should improve DGM performance when used for a non-model organism, compared to the use of a more divergent species transcriptome as a reference (Fig. 1B).

Each read (merged or not) is assigned to the *gene-id* corresponding to its best hit as in the DGM pipeline. Each assigned read is stored in the local database and appended to a FASTA file (one file per gene) before assembly. To avoid circularity with the *de novo* assembly, we used successively Spades (v3.7.1.8; *careful* option and *cov-cutoff* = auto; Bankevich *et al.* 2012) with the auxiliary BayesHammer error correction algorithm (Nikolenko *et al.* 2013), and then CAP3 (Huang & Madan 1999; default parameters). We used a combination of these two established software solutions because they rely on different approaches with different behaviors regarding read length, quantity of reads and transcript diversity (Zhao *et al.* 2011; Cahais *et al.* 2012; Lu *et al.* 2013; Bushmanova *et al.* 2016). The output of this step (one SQLite file; output from each assembly program is concatenated into a single file) is a collection of contigs of identified reads for each gene. The output of the assembly is a collection of assigned contigs stored in the local database and exported as *gene-id* FASTA files (one file per gene). A final validation step was used to prevent chimeric or mis-assembled reconstructions generated during the assembly step. All contigs were identified using blastn and the reference transcriptome. Contigs not identified at this step were discarded. Retained

contigs and HSPs were conserved and reverse-complemented to fit the reference orientation when necessary. The final output is a collection of assigned contigs stored in the local database (the reconstructed *de novo* + DGM assembly) and exported as FASTA files (one for each *gene-id*).

Assessing the quality of inferred assemblies

In this section, we compare the performance of *de novo* assembly alone (Fig. 1C) and the *de novo* combined with DGM assemblies (Fig. 1B). This comparison is made first on assemblies based on reads simulated from the *D. rerio* transcriptome, and second on empirical data from two non-model species (two closely related Cyprinid fish species). We test the impact of sequence divergence both over a range of simulated values (0%, 5%, 15% and 30%) and empirically, using extant inter-specific divergence between the reference species (*D. rerio*) and the test species, *T. souffia* and *P. toxostoma*. We use fixed read lengths of 100 and 200 nucleotides at coverage of 10X for simulated data; for empirical data, median read length was 234 bases.

A diversity of metrics should be used when assessing the quality of inferred transcriptomes. Although metrics for systematically assessing the quality of transcriptome assemblies have not been definitively established, accuracy, completeness, contiguity, chimerism and variant resolution capture essential elements of transcriptome quality (Martin & Wang 2011). These metrics are implemented in Rnnotator (Martin *et al.* 2010), and were used in the comparative study of *de novo* and guided assembly by Lu et al. (2013). We have adapted some of these metrics for assessing transcriptome assemblies as we focus on genes and not on transcripts. We define:

i) The number of identified genes (*NIG*). A gene is considered as identified if at least one read is assigned/identified on the basis of the reference transcriptome.

ii) The Completeness (Cp) for a gene *g* corresponds to the proportion of the length for the longest reference transcript (L_g, in number of bases) covered by the whole aligned contigs for the gene g(C_{gj}, j=1…n_g). The completeness is maximal (equal to 1) when the combined length of the aligned contigs (using ProbCons, (Do 2005)) matches the longest reference transcript.

$$Cp_g = \frac{card\{(\cup_{i=1}^{n_g} c_{gi}) \cap L_g\}}{card\ L_g}$$

iii) The Contiguity (Ct_g) for a gene *g* corresponds to the proportion of the longest reference transcript (L_g, in number of bases) covered by the longest contig for this gene (C_{gjmax}). The contiguity is maximal (equal to 1) when there is a perfect match between the longest identified contig and the longest reference transcript.

$$Ct_g = \frac{card\{max_i(C_{gi}) \cap L_g\}}{card\ L_g}$$

These last two metrics are adapted from Martin & Wang (2011) because we focus on gene identification for non-model organisms and not on resolving the whole set of transcripts (i.e. the complete transcriptome). Likewise, given inter-specific differences in orthologue length, our metric of completeness is only relevant to simulated data and should not be applied in assessments of empirical data. We used the longest reference transcript for each gene in the reference genome (*D. rerio*) as an upper bound of transcript length for the contigs from RNA-seq. Variation in contiguity and completeness was analyzed using a linear mixed-effect model with divergence, assembly approach (*de novo* or *de novo* + DGM) and their interaction treated as fixed factors, and random variation attributed to gene. This was done both for simulated and empirical data. For simulated data, we modeled the four divergence levels (0%, 5%, 15%, 30%) as fixed factors. For empirical data, divergence corresponded to the species effect. Finally, biological processes associated with each gene was inferred based on annotations from the Protein ANalysis THrough Evolutionary Relationships database (http://pantherdb.org/about.jsp; Thomas *et al.* 2003; Mi *et al.* 2016).

Biological material & empirical data

Telestes souffia (2 females; 1 male) and *Paracondrostoma toxostoma* (3 males) were collected from two rivers: the Durance (southern France) and Ain (eastern France). For each specimen, eight tissues were sampled (liver, hindgut, midgut, heart, brain, gill, caudal fin, spleen). Additionally, five F1-hybrid specimens were produced by *in vitro* fertilization of 50 eggs and semen from one female *T. souffia* and one male *P. toxostoma* respectively. Free-swimming larvae were first housed in small aquaria (40x25X15 cm); after four months, fish were transferred into bigger aquaria for growing (100x60x40 cm). Fish were fed with *Artemia nauplii* and complemented with commercial fish food. Because hybrids were required for further experimentation, only non-lethal samples from the caudal fin are represented in the empirical dataset.

Samples were ground in liquid nitrogen and total cellular RNA was extracted using the RNeasy Plus Universal kit (Qiagen). The TruSeq Stranded mRNA Library Preparation kit (Illumina Inc., USA) was used according to the manufacturer's protocol, with a few modifications (see Supplementary methods 5 for details). Samples were uniquely barcoded for individuals and tissues, and pooled cDNA libraries were sequenced using a Miseq Illumina sequencer with the 2 x 250 paired-end cycle protocol (see Supplementary methods 5 for details). In total, 16,216,379 and 25,541,756 reads were used to reconstruct the transcriptomes of *P. toxostoma* and *T. souffia*, respectively. The number of reads for F1 hybrid

specimens after a filtering step, ranged from 724,021 to 1,128,878. For the five F1 hybrid specimens, we observed 4,449,055 (among 5 449,772) when removing reads of length < 200 for genes of length >200. All data acquired for this study are available as an SRA archive at https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?study=SRP091996.

Comparing divergent transcriptomes and analysis of hybrid specimens

To investigate the extent of overlap between two divergent transcriptomes, we applied the combined method (*de novo* + DGM) to empirical data from *T. souffia* and *P. toxostoma* – in practice this approach could be extended to RNA-seq from any source (tissues, individuals, species; hereafter named A and B). The procedure involves first performing two separate *de novo* + DGM assemblies (one each for reads of source A and B, respectively), using a reference transcriptome 'X' for gene assignation Next, we generate Transcriptionally Active Regions (TARs or simulated reads) from A using a custom Python script to fragment the previously generated transcriptome for A to the equivalent of 10X coverage for each gene; we use TARs (hereafter named fragments) equal to 100 and 200 nucleotides to test for the impact of fragment length on overlap detection. The *gene-id* for each corresponding fragment is named *gene-id-A.* Generated fragments from A are then queried for similarity, using blastn against the local database including the transcriptomes of B and X, and assigned using the best hit. We then compare the *gene-id* between the query (A) and the database (B and X). Three potential outcomes are possible:

i) If a fragment from A is only assigned to B (i.e. it is not assigned to X) with the same *gene-id* (*gene-id-A* = *gene-id-B*), then the assignation is non-ambiguous. This is similar to the recovery rate (*rr*), as defined previously. For the entire set of fragments generated from a specific *gene-id-A*, we can define the recovery rate denoted *rr*A|*BX*. This corresponds to the proportion of fragments generated from this *gene-id* for A that was exclusively recovered for that same *gene-id* in B. A value of *rr*A|*BX* =1 would mean that all fragments for *gene-id-A* are assigned to the same *gene-id* in B.

ii) If a fragment from A is assigned with the same *gene-id* to X (i.e. it is not assigned to B) or to B and X (i.e. it is assigned to both to B and X with the same best hit), we consider this fragment as an ambiguous assignation, caused by the similarity of sequences between orthologous genes of the respective transcriptomes. For the whole set of fragments from a specific *gene-id-A*, we can define the orthologous recovery rate *rro*A|BX. This corresponds to the proportion of fragments generated

from the gene of A that are assigned either to X or to multiple transcriptomes (B and X). A value of *rro*A|*BX* =1 would mean that all fragments from A can equally be assigned to X; *rro*A|*BX* > 0 means that part of the gene is not exclusively found in B, potentially because of an incomplete assembly for B, or due to non-overlapping contigs between A and B. Alternatively, *rro*A|*BX* > 0 could also be due to homology between B and X.

iii) If a fragment from A is assigned to a different *gene-id*, whatever its identity in B and/or X, this fragment is also an ambiguous assignation (multiple hits including a different *gene-id* or a unique best hit assigned to a different *gene-id*). For the entire set of fragments derived from a specific *gene-id-A*, we define the paralogous recovery rate*, rrpA*|*BX.* This corresponds to the proportion of fragments for a gene from A that are assigned to paralogous genes in B and/or X. This could be caused by the similarity of sequences between paralogous genes from those transcriptomes, or due to common domains from multigene families.

For a given *gene-id-A*, the combination of false assignations due to orthologous genes and paralogous genes decreases the recovery rate, each contributing to a unique proportion of the total assigned transcripts: *rr*A|BX + *rro*A|BX + *rrp*A|BX=1. Similarly, for each *gene-id-A*, cross-completeness is defined by calculating *rr*A|BX on the subset of fragments that uniquely capture "correct" *gene-id* assignations from the references (i.e. *sr*A|BX=1; Fig. 2B); here *sr*A|BX is defined as the specificity rate (i.e. the proportion of fragments assigned to *gene-id-B* coming from the same *gene-id* on A).

This process is reiterated reciprocally for the second transcriptome (i.e. B|AX) using the same metrics. Once completed, we can define the perfect-cross-completeness, that is the intersection of the subset of genes from A displaying *rrA*|*BX*=1 and *srA*|*BX*=1 with the subset of genes from B displaying *rrB*|*AX*=1 and *srB*|*AX*=1. This pool of genes presents a complete overlap between the two transcriptomes A and B (Fig. 2B).

Extending the combined method for analysis of hybrid samples

Characterizing discriminant genes between two transcriptomes yields information that is especially useful when analyzing hybrids. In this context, if we have two transcriptomes, A and B, and a reference transcriptome X in a local database, a discriminant gene for a given transcriptome (A or B) is a gene for which all reads can be correctly assigned only uniquely (i.e. to this gene from this particular transcriptome). We tested this for *T. souffia* (A), *P. toxostoma* (B) and *D. rerio* (X), considering two simulated TARs, read lengths as before (100 and 200 bases). This process is similar to the previous transcriptome comparison, but further includes the source transcriptome (A) in the list of transcriptomes to which reads can be

assigned. All the fragments from a *gene-id-A* display a true assignation proportion (recovery rate) *rrA*|*ABX*≤1, and false assignations occupy unique proportions due to assignation to orthologous (*rroA*|*ABX*) or paralogous (*rrpA*|*ABX*) genes such that *rrA*|*ABX* + *rroA*|*ABX* + *rrpA*|*ABX*=1. For each *gene-id-A*, the discriminant score is estimated by *rr*A|ABX on the subset of genes for which *srA*|*ABX*=1, where *srA*|*ABX* is defined as the specificity rate (i.e. the probability of a read assigned to *gene-id-A* is issued from the same *gene-id* on A).

This process is reiterated reciprocally for B|ABX, calculating the same metrics. We then can compute the cross-discriminant proportion of discriminant genes, that is the intersection of the genes from A displaying *rrA*|*ABX*=1 for the subset *srA*|*ABX*=1 with the genes from B displaying *rrB*|*ABX*=1 for the subset *srB*|*ABX*=1. This pool of genes is fully discriminant between the two transcriptomes. The intersection between perfect-cross-completeness and the cross-discriminant genes corresponds to the perfect discriminant gene set for the two transcriptomes (Fig. 2B). These represent the fraction of genes that can be mapped uniquely to a given species' transcriptome and for which cross-completeness is unity (i.e. gene lengths are the same).

Finally, we analyze the five F1 hybrid specimens between *T. souffia* (A) and *P. chondrostoma* (B) using the DGM pipeline and *D. rerio* (X) as a reference transcriptome. The reads are assigned to a *gene-id* and a *transcriptome-id* (A, B or X) using the best hit. Three potential outcomes can be found:

i) a read can be assigned to a single *gene-id* and a single *transcriptome-id*. This non-ambiguous read is noted as a perfect-read. A gene for which all reads display this behavior is called a perfect gene (*rrF1*|*ABX*=1).

ii) a read can be assigned to multiple *transcriptome-ids* with a unique *gene-id*. This ambiguous assignation corresponds to the effect of orthologous genes. If this condition is true for at least one read for an imperfect gene, *rroF1*|*ABX*>0.

iii) a read can be assigned to multiple *gene-ids* whatever the *transcriptome-id*. This ambiguous assignation corresponds to the effect of paralogous genes. If this condition is true for at least one read for an imperfect gene, *rrpF1*|*ABX*>0.

We finally establish a list of perfect-genes and a list of imperfect genes that are compared with the list of perfect-discriminant genes established *in silico* (see supplementary methods 6 for details).

Results

Evaluation of the DGM pipeline with simulated data

The mixed logistic model describing recovery rate (*rr*) was highly significant (X^2=27,605,272, df=23, P< 2.2e^{-16}), with each term of the model being significant (supplementary results 1). Recovery rate (*rr*) increased as a function of gene length irrespective of read length and divergence (Fig. 3); however, the effect was most pronounced at 30% divergence. At 30% divergence between target and reference, fewer than 50% of reads from genes smaller than 2kb (350 bases reads; Fig. 3G) to 4kb (100 bases; Fig. 3A) – a substantial fraction of the genome (Fig. 4A) – were recovered. Likewise, high divergence was associated with the greatest variability in *rr* (dashed lines on Fig. 3). Conversely, *rr* was generally greater than 0.9 under all other scenarios of divergence. Read length influenced the shape of the curve describing the gene length at which *rr* approached unity, with longer reads yielding perfect scores for smaller genes (Fig. 3G & 3H). As before, this effect was substantially reduced under the high divergence scenario.

The mixed logistic model for specificity rate (*sr*) was also highly significant (X^2=1,625,516, df=8, P< 2.2e-16). Specificity rate presents median values near to one (supplementary results 1), with the random effect being greater (standard deviation estimate=4.00) compared that for *rr* (standard deviation estimate=2.341). Each term of the model appeared significant (supplementary results 1). Relations between factors and *sr* are the same as for *rr*: higher are read length and gene length, higher is the *sr*, higher is the divergence lower is the *sr*.

The dataset of simulated reads used to evaluate the DGM pipeline consisted of 31,944 genes of varying length (Figs. 4A & 4B). In general, the proportion of ’perfect’ genes (*rr*=1, *sr*=1) increased with read length. For example, at 0% divergence between reference transcriptome and simulated reads 21,913 genes belonged to the perfect category for read length of 100 bases (Fig. 4C) whereas 27,289 were so classified for read lengths of 350 bases (X^2=2554.4, df=1; P< 2.2e^{-16}, Fig. 4E). This improvement in performance coincided with a decrease in the proportion of ‘donor’ genes (*rr*<1, *sr*=1), but also with slight increases in the number of ‘recipient’ (*rr*=1, *sr*<1) and ‘mixed’ (*rr*<1, *sr*<1) gene classes (Fig. 4E). Surprisingly, gene length also had a modest impact on improving DGM performance, with longer genes accumulating a slightly higher proportion of ‘perfect’ genes; however, this trend was only observed in the absence of divergence between reference transcriptome and simulated reads. Sequence divergence had a drastic impact on the proportion of genes that could be classed

as perfect: at 30% divergence only 700 genes for reads of 100 bases length (Fig. 4D; 2.19% of all genes, X^2=979.5, df=1; P< $2.2e^{-16}$) and 2,401 for read lengths of 350 bases (Fig. 4F; 7.52%).

Increasing divergence also increased the 'donor' gene category (Figs. 4D & 4F), detrimental to the 'perfect' gene category. Here increasing read length also appeared to slightly increase the proportion of 'donor' genes in general, with the maximum proportion of 'donor' genes observed for reads of 30% divergence and a length of 350 bases (81.04%, i.e. 25,888 genes); this also corresponded to a low proportion for the 'perfect' gene category (7.52%, i.e. 2,401 genes). The percentage of 'mixed' genes also increased with the divergence, although this increase was most pronounced for short reads (Fig. 4D) and is maximal for 100 bases read lengths (7,370 genes; 23.07%).

Interestingly the proportion of 'recipient' genes (*rr*=1, *sr*<1) was low whatever the factor studied. The minimum value was 0% (no gene) at 0% divergence and for read lengths of 100 and 200 bases. The greatest number of "recipient" genes (391 genes; 1.22%) was observed for reads of 15% divergence and a read length of 100 bases (supplementary results 2). Likewise, the 'undetectable' gene category (*rr*=0, *sr*=N.A.) presented low proportions overall, ranging from 5.08% to 7.40% (i.e. 1,623 to 2,365 genes), with the greatest proportions being observed for short genes.

Comparison between *de novo* and *de novo* + DGM approaches

For reads of 100 bases, *de novo* assembly alone recovered a marginally higher number of genes (*NIG*) than the combined *de novo* + DGM pipeline, irrespective of transcriptomic divergence (Fig. 5A): numbers for *de novo* ranged from 31,930 to 31,938 and from 31,147 to 31,944 for *de novo* +DGM. Conversely, measures of the quality of transcript assembly were higher for *de novo* + DGM than for *de novo* alone. The completeness (*Cp*) of *de novo* assembled 100 bases reads ranged from 0.74 to 0.89, with *Cp* generally increasing with sequence divergence (Fig. 5B); *Cp* for *de novo* + DGM ranged from 0.97 to 0.99. Contiguity (*Ct*) exhibited similar patterns, increasing with divergence for *de novo* assemblies (Fig. 5C; 0.67 to 0.87), yet was relatively stable, but higher for *de novo* + DGM (0.97 to 0.99). Increasing read length to 200 bases resulted in an overall decrease in the number of genes recovered, as well as a reversal in the patterns of pipeline performance, with *de novo* + DGM generally identifying more genes (28,339 to 28,532); *de novo* alone at the same levels of divergence tended to recover fewer genes (27,597 on average), with the exception of 28,604 genes at 15% divergence (Fig. 5A). *Cp* and *Ct* were improved in both pipelines when read length was increased to 200 bases, although *de novo* + DGM continually outperformed *de*

novo alone at most levels of divergence (average *Cp* of 0.95 *vs* 0.99; average *Ct* of 0.89 *vs* 0.99), except at 30% where metrics were similar for each pipeline.

Fig. 6 shows a non-parametric estimation of the bivariate distribution of contiguity and completeness for 0% divergence and read lengths 100 and 200 bases – note that for clarity, only 'imperfect' genes are displayed (the full range of factors is described in supplementary results 3). For 100bp reads of 0% divergence, the *de novo* approach alone showed only 13,590 genes (46.1%) with a contiguity and a completeness equal to 1, while a large cluster of genes was observed with quality metrics inferior to 0.1 (Fig. 6A). In contrast, 94.8% (27,924) of genes assembled using the combined *de novo* + DGM approach showed perfect contiguity and completeness (Fig. 6B). Increasing read length to 200 bases appeared to increase the number of genes with high *Cp* scores in the *de novo* only approach (Fig. 6C); however, only 14,252 genes (53%) in total had contiguity and completeness equal to 1. A much greater fraction of genes (98%; 27,861) with perfect contiguity and completeness were also observed for the *de novo* + DGM approach using reads of 200 bases (Fig. 6D).

Increasing divergence led to a drastic reduction in the number of genes with perfect *Ct* and *Cp* for both pipelines, decreasing at 30% divergence to only 5.3% of *de novo* assembled genes and 5.7% of genes from the combined *de novo* + DGM approach using 100 bases reads. However, approximately 80% of genes present contiguity and completeness above 0.95, for both methods (supplementary results 3). Increasing read length did little to improve performance of either method, with only 5.5% of genes assembled from 200bp reads scoring perfect contiguity and completeness irrespective of pipeline used. Nevertheless, 85% of genes did present contiguity and completeness above 0.95 for both methods in this scenario (supplementary results 3).

Assembling and comparing Parachondrostoma and *Telestes* transcriptomes

A total of 16,374 genes were common to both species and also detected by each of the two methods (supplementary results 4). Nevertheless, we observed a clear difference in the number of detected genes between *de novo* and *de novo* + DGM assemblies: significantly higher proportion of genes were detected for both *P. toxostoma* ($t=4.65$, $P<10^{-5}$) and *T. souffia* ($t=3.58$, $P<10^{-3}$) when using *de novo* + DGM (20,605 and 20,368 genes, respectively; 18,989 common to the two transcriptomes) than for the *de novo* approach alone (20,032 and 19,931 genes, respectively; 17,984 shared). It should also be noted that the biological processes inferred from the list of genes identified using *de novo* + DGM did not differ significantly

between the two species (X^2=2.46, P=0.99; supplementary results 5). The performance of the combined *de novo* + *DGM* approach was also higher than *de novo* alone as measured by both contiguity (supplementary results 6; F=1109.486, P< 2.2e-16) and completeness (supplementary results 6; F=28.47, P=9.56e-08) and further analyses are only conducted based on *de novo* + DGM assemblies.

When we compared the perfect cross-completeness parameter (i.e. the reciprocal completeness between the two species; Fig. 2B) for all genes, we observed that contig assignations generally overlapped between species, irrespective of the length of TARs used. Indeed, 87% and 88% of the genes (for 100bp and 200bp reads, respectively) exhibited a cross-completeness between 0.95 and 1, indicating that the assembled contigs for the two species were highly similar.

Of the 18,989 genes identified using *de novo* + DGM, we detected 13,988 that were common to both species and free of 'recipient' genes.

In contrast to cross-completeness, read length had a dramatic effect on the detection of discriminant genes: the number of 'perfect' genes for *P. toxostoma* increased from 1,727 to 4,712 when fragments (TARs) lengths increased from 100 to 200 bases, and from 1,486 to 4,487 for *T. souffia*. It should be noted that all the discriminant genes identified using 100bp fragments were also included in the set of discriminant genes identified with 200bp fragments. As with discriminant genes, the number of cross-discriminant genes (i.e. the shared *T. souffia* and *P. toxostoma* discriminant genes; Fig. 2B) also increased with increased fragment length (n=357, 100 bases; n=2,381; 200 bases).

Likewise, the cross-completeness scores for the cross-discriminant genes increased as a function of fragment length. Whereas 44% of the cross-discriminant genes (157 of the 357) identified from 100bp fragments had a cross-completeness between 0.95 and 1, 77.5% of the cross-discriminant genes identified from 200bp fragments (1,846 of the 2,381) showed similar levels of cross-completeness. The number of perfect discriminant genes within the cross-discriminant set also increased from 54 to 1,028 when contrasting the 100bp and 200bp fragments, respectively (supplementary results 7).

Application to hybrid specimens

For the five F1 hybrid specimens, we detected 8,000 genes, assigned based on contigs constructed from a minimum of 50 reads – here a minimum read criterion was adopted to reflect that typically used to achieve acceptable statistical power for the quantification of mRNA expression from RNA-seq data. A total of 424 genes were 'perfect' genes (*rrF1|ABX* =1) and 7,576 were 'imperfect' genes (*rrF1|ABX* <1). Although only 41.3% of 'perfect' genes (175 of 424) could be classified as 'perfect discriminant' genes (supplementary results 7), this fraction

could be increased to 62.0% (263 genes) by applying lower thresholds (from 1 to 0.95) for cross-completeness and/or cross-discriminance (supplementary results 7). Relatively fewer 'imperfect' genes (323 of 7,576 genes; 4.26%) could be classed as 'perfect discriminant' genes, although 3,823 were categorized with 'perfect cross-completeness' (supplementary results 7). These genes with perfect cross-completeness by definition are fully overlapping, including both identical and divergent sequences (supplementary results 7), and as such could be used as partial discriminant genes. Indeed, these genes display a proportion of reads that are not discriminant between the two species (reads assigned to multiple transcriptomes or multiple genes) and other reads that are discriminant between the transcriptomes. In the case of those five F1 hybrids, this represents 4,392 genes with a perfect cross-completeness that could ultimately be used for the analysis of gene expression among hybrids. When we use 95% overlap (for cross completeness) this provides 6,020 genes (287 perfect and 5,733 imperfect genes,)

Discussion

Direct Genome mapping and characterizing gene categories

We developed a DGM pipeline for flexible filtering of data and incorporation of a reference transcriptome for read assignation. We tested this pipeline with simulated data over a range of read lengths and sequence divergence relative to the reference. Our analyses provide a much-needed empirical evaluation of the utility and limits of the DGM approach, demonstrating that both read length and sequence divergence impact the efficiency and performance of read annotation and transcriptome assembly. Irrespective of read length and with 0% sequence divergence, 1,663 of the 31,944 simulated genes (5.21%) could not be retrieved when using DGM alone. This result was surprising given that reads were simulated directly from the corresponding reference transcriptome. Yet in the analysis, multiple assignations were retained as equal best hits. We observed that when a read equally matches two different genes, the blast score is highest for the gene that displays a length closest to the query read length, hence this gene acts as an attractor for those reads.

High sequence divergence (30%) and short read lengths (100 and 150 bases) had a negative impact on recovery rate, substantially reducing the number of genes with a recovery rate *rr*=1 (i.e. all the generated reads being correctly assigned). The effect of sequence divergence is congruent with studies on *Drosophila* and primates (Ockendon *et al.* 2016), suggesting that 30% divergence may represent an upper threshold for the recovery of orthologous sequences. When sequence divergence is non-null, increasing read length (200

and 350 bases) limits the read assignation error towards paralogous genes. Genes with a specificity rate *sr*=1 (i.e. no assigned read belong to another gene-id), were less sensitive to either divergence or read length. Although few genes were classified as “recipient’ genes (*sr*<1), these could represent an important source of bias in quantitative analyses of RNA-seq data, appearing as over-expressed via the erroneous assignations from “donor” genes. It should also be noted that a “recipient gene” can act as an attractor for multiple donors, as evidenced by the high fraction of donors relative to recipients. Such genes would appear as highly “over-expressed”. Although these would obviously not lead to erroneous inference of over-expression in our pipeline (i.e. tey would be removed/filtered prior to analyses), they could nevertheless underestimate the true levels of expression for donors. As such, we highly recommend identifying such problematic genes as a means of distinguishing biologically meaningful signals from artefact in the interpretation of differential expression profiles from RNA-seq data. To this end, we have made all scripts used for simulating reads from a reference genome available upon request to facilitate similar analyses in other species; these scripts can be modified for use on any other reference transcriptome used to annotate RNA-seq data, with read length and divergence parameters adjusted to match those under investigation.

Our analyses also highlight the difficulties that can arise when working with a reference transcriptome that is highly divergent from one’s focal species – this was facilitated by our *in silico* generation of reads that in turn enabled development of a matrix of gene categories (given the length of the TARs) based on recovery and specificity rates (parameters *rr* and *sr*). Specifically, we demonstrated that when sequence divergence increases to 30%, the proportion of ‘donor’ genes (*rr*<1, *sr*=1) and ‘mixed’ genes (*rr*<1, *sr*<1) increases, and the proportion of ‘perfect’ genes (*rr*=1, *sr*=1) decreases. Surprisingly, the proportion of ‘recipient’ genes (*rr*=1, *sr*<1) does not increase; however, their distribution within the genome likely does, particularly if analyses are based on short sequence reads. For example, when analyses are based on longer reads, recipient genes are largely derived from the smaller genes (i.e. <1,000 bases); however, short genes appear sensitive to the accumulation of “recipients” of various size (contrast Fig. 4E & 4F). In this instance gene length alone could not be used as a quality filter for excluding genes in downstream analyses of expression. Additionally, some genes previously classed as perfect become ‘donor’ genes at high divergence. Likewise, the percentage of ‘perfect’ genes decreased from 85.43% to 68.60% when read lengths where reduced from 350 to 100 bases. This result can be caused by i) a recently duplicated gene found at multiple loci in the genome and/or ii) a conserved domain (or repetitive domain) in gene families. Irrespective of the underlying causes, this result highlights the advantages in terms of increased confidence of assignation when working with longer reads, a point not to be neglected when planning RNA-seq experiments for non-model species.

Comparing *de novo* and *de novo* + DGM assemblies

Our comparison of the efficiency and performance of *de novo* vs a combined *de novo* + DGM approach was based on three key metrics: the number of genes identified, contiguity and completeness. Although short reads yielded a higher number of genes identified, the quality of genes was low with respect to contiguity and completeness (i.e. an increased proportion of segmented and/or fragmented transcripts). Moreover, the observed increase of these two metrics with increasing sequence divergence – a trend most pronounced in the *de novo* only approach – was certainly counter-intuitive, but explained by the fact that several *D. rerio* genes share identical sequences (paralogous genes and repetitive conserved domains). To some extent this might be expected in species that have undergone extensive genome duplication, such as cyprinids; the extent to which this trend is evident in species with more conserved genomes is a topic of potential future interest. Nevertheless, as we generated divergent sequences (up to 30%) with a random process, this artificially decreased the similarity between paralogous genes or conserved domains, and artificially enhanced assembly metrics. At 0% divergence, the efficiency (i.e. percentage of genes identified) of *de novo* assembly is less impacted than its performance (contiguity and completeness). Several strategies are available to optimize d*e novo* assembly, for example using multiple K-mer values in a de Bruijn graph to handle both over- and under-expressed transcripts (Robertson *et al.* 2010; Zhao *et al.* 2011). Evaluating various optimization strategies is beyond the scope of this article, but as we propose an assembly solution using the combination of *de novo* + DGM, an improvement in any one of the constituent parts of the pipeline would benefit the entire assembly.

As a guideline, we would recommend using the combination of *de novo* assembly with DGM, as this systematically increases the number of identified genes for divergences up to 15%, relative to *de novo* alone; although there was a slight decrease at 30% divergence. Even when target and reference are identical (i.e. 0% divergence), the combined approach yields advantages, with 98% of identified genes displaying contiguity and completeness equal to 1. Moreover, a higher percentage of genes are successfully retrieved using the *de novo* + DGM approach implemented here (88.1%) than for the *de novo* approach alone (81.5%). As far as performance is concerned, *de novo* + DGM outperforms *de novo* alone both for contiguity and completeness. When divergence increases beyond 15%, both approaches converge to the same distribution. This greater efficiency and performance of *de novo* + DGM allows identifying more genes with a higher degree of confidence in associated assignations.

Deciphering cyprinid transcriptomes and application of DGM to hybrid specimens

Typically, studies on hybrid transcriptomes using RNA-seq have focused on model organisms such *Arabidopsis spp*. (Shi *et al.* 2012), *Drosophila spp*. (McManus *et al.* 2010) and *Mus spp*. (Mack *et al.* 2016). All of them have well annotated genomes and transcriptomes that facilitate deciphering hybrid expression patterns, i.e. assigning genes from the hybrid specimens to the parental species. In the absence of a reference genome for the two parental species, one must generate at least one transcriptome for each species to identify discriminant genes. The most useful discriminant genes should also display perfect completeness (perfect discriminant genes), particularly if the objective is to compare true gene expression differences amongst the two species and their hybrids. With the *de novo* + DGM approach applied on *T. souffia* and *P. toxostoma*, we obtain 18,989 identified genes common to each species. 13,988 of these genes were never identified from reads originating from other genes or species (*sr*=1 whatever the value of *rr*), representing genes that are free from reads erroneously assigned from donors. Consequently, these genes should be considered the most reliable for expression analysis given this lack of ambiguity. A false identification of reads in hybrid specimens caused by orthologous and/or paralogous genes could induce important bias in expression analysis. Likewise, only the fraction of discriminant genes with a cross-completeness between 0.95 and 1 for both species is useful for unambiguous hybrid gene expression analysis (1,846 genes). This is because discriminant genes could be due to non-overlapping regions of the contig that mimic divergent regions between the two species. If non-overlapping regions are caused by different transcript lengths (e.g. gaps in the constructed library of biological diversity in transcript length), it induces a bias in the expression proportional to the length of the non-overlapping zone. In this particular case-study, this category corresponds to 22% of the discriminant genes for fractionated contigs of 200 bases (up to 56% for 100 bases). Moreover, cross-discriminant genes were more abundant when assigned fractionated contigs (i.e. simulated TARs) of 200 bases than 100 bases (2,381 *vs* 357 genes). Together, this suggests that sequencing reads as long as possible should be used when analyzing hybrid transcriptomes and gene expression. This is especially crucial for future studies working on hybridization to decipher the patterns of gene expression related to each parental species.

Sequencing reads from F1 hybrid specimens identified only 25.04% of genes based on *D. rerio.* Furthermore, among the 1,028 perfect genes from the parental species, we observed only 175 in the F1 hybrids (17%), and an additional 323 imperfect genes (31.4%). There are three main reasons for these results: i) analyses were restricted to only the caudal

fin for hybrid specimens, whereas the transcriptome database of each parental species was constructed from eight tissues; ii) the number of reads generated for the different F1 hybrids was lower than the number of reads used to build the transcriptome database for each parental species; and, iii) a possible underestimation of polymorphism in the reference database. Despite these limitations, the high number of genes being partially discriminant and displaying a complete overlap i.e. 4,637 genes (6,020 genes for a 95% overlap) is a striking result. These genes can indeed be used to identify hybrid patterns. This is particularly promising in comparison to the 9,966 genes studied for *Drosophila spp*. (McManus *et al.* 2010), 14,713 genes for *Arabidopsis spp*. (Shi *et al.* 2012) and 9,851 genes for *Mus spp*. (Mack *et al.* 2016).

Conclusion

In conclusion, this analysis emphasizes the high potential of a combined *de novo* + DGM approach, implemented in our pipeline to generate transcriptome assemblies from multiple non-model species. Although sequence divergence and genome duplication represent important potential pitfalls, the method nonetheless shows promise for the analysis of divergent transcriptomes. Identification of the unique contribution of parental transcriptomes to hybrid specimens is one area where the approach may have great potential, opening up opportunities for future application to hybrid zones and/or in the context of hybrid speciation.

Acknowledgment

We specially thank Bernard Barascud for his great involvement with the common garden experiments. AU was supported by a PhD grant from EDF (Electricité de France). We are grateful to the different departments of Electricité de France for the financial support of the present study: EDF -Recherche et Développement, Clamart especially Dr Mathieu Le Brun and Laurence Tissot, EDF- Unité of Production Méditerranée especially Dr Julie Mosseri and EDF Centre d'Ingénierie Hydraulique Technolac – Chambéry especially Dr Agnès Barillier and Frédéric Jacob.

Authors' contributions

AG conceived and designed the experiments. AU conducted bioinformatics analyses and developed the software used in the pipelines. AU, JFM and AG acquired molecular data. NP conducted statistical analyses. AG, AU and NP interpreted the data. AG, JFM, NP and

RJSM wrote the manuscript. RC contributed to revising the manuscript. All authors read and approved the final manuscript.

References

Alvarado S, Rajakumar R, Abouheif E, Szyf M (2015) Epigenetic variation in the Egfr gene generates quantitative variation in a complex trait in ants. *Nature communications*, **6**, 6513.

Ayroles JF, Carbone MA, Stone EA *et al.* (2009) Systems genetics of complex traits in Drosophila melanogaster. *Nature genetics*, **41**, 299–307.

Bankevich A, Nurk S, Antipov D *et al.* (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of computational biology: a journal of computational molecular cell biology*, **19**, 455–477.

Bar-Even A, Paulsson J, Maheshri N *et al.* (2006) Noise in protein expression scales with natural protein abundance. *Nature genetics*, **38**, 636–643.

Bates D, Maechler M, Bolker B, Walker S (2014) lme4: Linear mixed-effects models using Eigen and S4. *R package version*.

Bushmanova E, Antipov D, Lapidus A, Suvorov V, Prjibelski AD (2016) rnaQUAST: a quality assessment tool for de novo transcriptome assemblies. *Bioinformatics* , **32**, 2210–2212.

Cahais V, Gayral P, Tsagkogeorga G *et al.* (2012) Reference-free transcriptome assembly in non-model animals from next-generation sequencing data. *Molecular ecology resources*, **12**, 834–845.

Camacho C, Coulouris G, Avagyan V *et al.* (2009) BLAST+: architecture and applications. *BMC bioinformatics*, **10**, 421.

Czypionka T, Cheng J, Pozhitkov A, Nolte AW (2012) Transcriptome changes after genome-wide admixture in invasive sculpins (Cottus). *Molecular ecology*, **21**, 4797–4810.

Dion-Cote AM, Renaut S, Normandeau E, Bernatchez L (2014) RNA-seq reveals transcriptomic shock involving transposable elements reactivation in hybrids of young

lake whitefish species. *Molecular biology and evolution*, **31**, 1188–1199.

Do CB (2005) ProbCons: Probabilistic consistency-based multiple sequence alignment. *Genome research*, **15**, 330–340.

Elran R, Raam M, Kraus R *et al.* (2014) Early and late response of Nematostella vectensis transcriptome to heavy metals. *Molecular ecology*, **23**, 4722–4736.

Flicek P, Amode MR, Barrell D *et al.* (2014) Ensembl 2014. *Nucleic acids research*, **42**, D749–55.

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–U130.

Grishkevich V, Yanai I (2014) Gene length and expression level shape genomic novelties. *Genome research*, **24**, 1497–1503.

Haas BJ, Papanicolaou A, Yassour M *et al.* (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature protocols*, **8**, 1494–1512.

Hipp R, Team, SQLite Development (2015) *SQLite*.

Huang X, Chen XG, Armbruster PA (2016) Comparative performance of transcriptome assembly methods for non-model organisms. *BMC genomics*, **17**, 523.

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

Jain P, Krishnan NM, Panda B (2013) Augmenting transcriptome assembly by combining de novo and genome-guided tools. *PeerJ*, **1**, e133.

Leder EH, McCairns RJS, Leinonen T *et al.* (2015) The evolution and adaptive potential of transcriptional variation in sticklebacks--signatures of selection and widespread heritability. *Molecular biology and evolution*, **32**, 674–689.

Liu S, Luo J, Chai J *et al.* (2016) Genomic incompatibilities in the diploid and tetraploid offspring of the goldfish x common carp cross. *Proceedings of the National Academy of Sciences of the United States of America*, **113**, 1327–1332.

Lu B, Zeng Z, Shi T (2013) Comparative study of de novo assembly and genome-guided

assembly strategies for transcriptome reconstruction based on RNA-Seq. *Science China. Life sciences*, **56**, 143–155.

Mack KL, Campbell P, Nachman MW (2016) Gene regulation and speciation in house mice. *Genome research*, **26**, 451–461.

Martin J, Bruno VM, Fang Z *et al.* (2010) Rnnotator: an automated de novo transcriptome assembly pipeline from stranded RNA-Seq reads. *BMC genomics*, **11**, 663.

Martin JA, Wang Z (2011) Next-generation transcriptome assembly. *Nature reviews. Genetics*, **12**, 671–682.

McManus CJ, Coolon JD, Duff MO *et al.* (2010) Regulatory divergence in Drosophila revealed by mRNA-seq. *Genome research*, **20**, 816–825.

Meyer A, Van de Peer Y (2005) From 2R to 3R: evidence for a fish-specific genome duplication (FSGD). *BioEssays: news and reviews in molecular, cellular and developmental biology*, **27**, 937–945.

Mi H, Poudel S, Muruganujan A, Casagrande JT, Thomas PD (2016) PANTHER version 10: expanded protein families and functions, and analysis tools. *Nucleic acids research*, **44**, D336–42.

Nikinmaa M, McCairns RJS, Nikinmaa MW *et al.* (2013) Transcription and redox enzyme activities: comparison of equilibrium and disequilibrium levels in the three-spined stickleback. *Proceedings of the Royal Society B: Biological Sciences*, **280**, 20122974–20122974.

Nikolenko SI, Korobeynikov AI, Alekseyev MA (2013) BayesHammer: Bayesian clustering for error correction in single-cell sequencing. *BMC genomics*, **14 Suppl 1**, S7.

Ockendon NF, O'Connell LA, Bush SJ *et al.* (2016) Optimization of next-generation sequencing transcriptome annotation for species lacking sequenced genomes. *Molecular ecology resources*, **16**, 446–458.

Palma-Silva C, Ferro M, Bacci M, Turchetto-Zolet AC (2016) De novo assembly and characterization of leaf and floral transcriptomes of the hybridizing bromeliad species (Pitcairnia spp.) adapted to Neotropical Inselbergs. *Molecular ecology resources*, **16**,

1012–1022.

Poelstra JW, Vijay N, Hoeppner MP, Wolf JB (2015) Transcriptomics of colour patterning and coloration shifts in crows. *Molecular ecology*, **24**, 4617–4628.

Pratlong M, Haguenauer A, Chabrol O *et al.* (2015) The red coral (Corallium rubrum) transcriptome: a new resource for population genetics and local adaptation studies. *Molecular ecology resources*, **15**, 1205–1215.

Qian X, Ba Y, Zhuang Q, Zhong G (2014) RNA-Seq technology and its application in fish transcriptomics. *Omics: a journal of integrative biology*, **18**, 98–110.

Rana SB, Zadlock FJ, Zhang ZP, Murphy WR, Bentivegna CS (2016) Comparison of De Novo Transcriptome Assemblers and k-mer Strategies Using the Killifish, Fundulus heteroclitus. *PloS one*, **11**, 16.

R Core Team (2013) R: A language and environment for statistical computing.

Renaut S, Nolte AW, Bernatchez L (2009) Gene expression divergence and hybrid misexpression between lake whitefish species pairs (Coregonus spp. Salmonidae). *Molecular biology and evolution*, **26**, 925–936.

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

Schena M, Shalon D, Davis RW, Brown PO (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, **270**, 467–470.

Schirmer M, Ijaz UZ, D'Amore R *et al.* (2015) Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic acids research*, **43**, e37.

Schulz MH, Zerbino DR, Vingron M, Birney E (2012) Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics* , **28**, 1086–1092.

Seehausen O, Butlin RK, Keller I *et al.* (2014) Genomics and the origin of species. *Nature reviews. Genetics*, **15**, 176–192.

Shi XL, Ng DWK, Zhang CQ *et al.* (2012) Cis- and trans-regulatory divergence between progenitor species determines gene-expression novelty in Arabidopsis allopolyploids.

Nature communications, **3**, 9.

Thomas PD, Kejariwal A, Campbell MJ *et al.* (2003) PANTHER: a browsable database of gene products organized by biological function, using curated protein family and subfamily classification. *Nucleic acids research*, **31**, 334–341.

Vijay N, Poelstra JW, Kunstner A, Wolf JBW (2013) Challenges and strategies in transcriptome assembly and differential gene expression quantification. A comprehensive in silico assessment of RNA-seq experiments. *Molecular ecology*, **22**, 620–634.

Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews. Genetics*, **10**, 57–63.

Wittkopp PJ, Haerum BK, Clark AG (2004) Evolutionary changes in cis and trans gene regulation. *Nature*, **430**, 85–88.

Wolf JBW (2013) Principles of transcriptome analysis and gene expression quantification: an RNA-seq tutorial. *Molecular ecology resources*, **13**, 559–572.

Zhang J, Kobert K, Flouri T, Stamatakis A (2014) PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* , **30**, 614–620.

Zhao QY, Wang Y, Kong YM *et al.* (2011) Optimizing de novo transcriptome assembly from short-read RNA-Seq data: a comparative study. *BMC bioinformatics*, **12 Suppl 14**, S2.

Figure Legends

Figure 1. Transcriptome Assembly Pipelines. We present three pipelines for transcriptome assembly and contig assignation. (A) The Direct Genome Mapping (DGM) approach (Blue background) with a **read assignation step** and a **read count step** that produces a hit-table for each gene. (B) The *de novo* + DGM approach (green background) combines elements of each, making use of a local database comprising both the reference transcriptome and *de novo* assembly of the focal transcriptome(s). (C) The *de novo* approach (red background) with a blastn identification of generated contigs to obtain a *de novo* assembly for the focal

transcriptome. Cylinders represent the database, rectangles represent analytical steps in the processes, and parallelograms correspond to results. Note that the dashed lines represent optional steps in the pipelines.

Figure 2. **Key concepts**. (A) Measures used to evaluate pipelines, and used for gene classification. Classification is based on the proportion of reads correctly assigned to their gene of origin (recovery rate) and the proportion of reads that are correctly assigned to a given gene (specificity). (B) Combining pipeline metrics for the evaluation of comparative transcriptomics. Cross-completeness is defined by the recovery rate on the subset of genes with perfect specificity rate for a given gene; the overlap/intersection of perfect genes defines perfect cross-completeness, genes of the same length that are common to two distinct transcriptomes. Genes may also be discriminant between species, that is perfect genes within a given transcriptome, but which show no or only partial overlap with a second transcriptome. The intersection of these two larger categories, perfect cross-completeness and cross-discriminance, defines genes that are perfectly discriminant between transcriptomes.

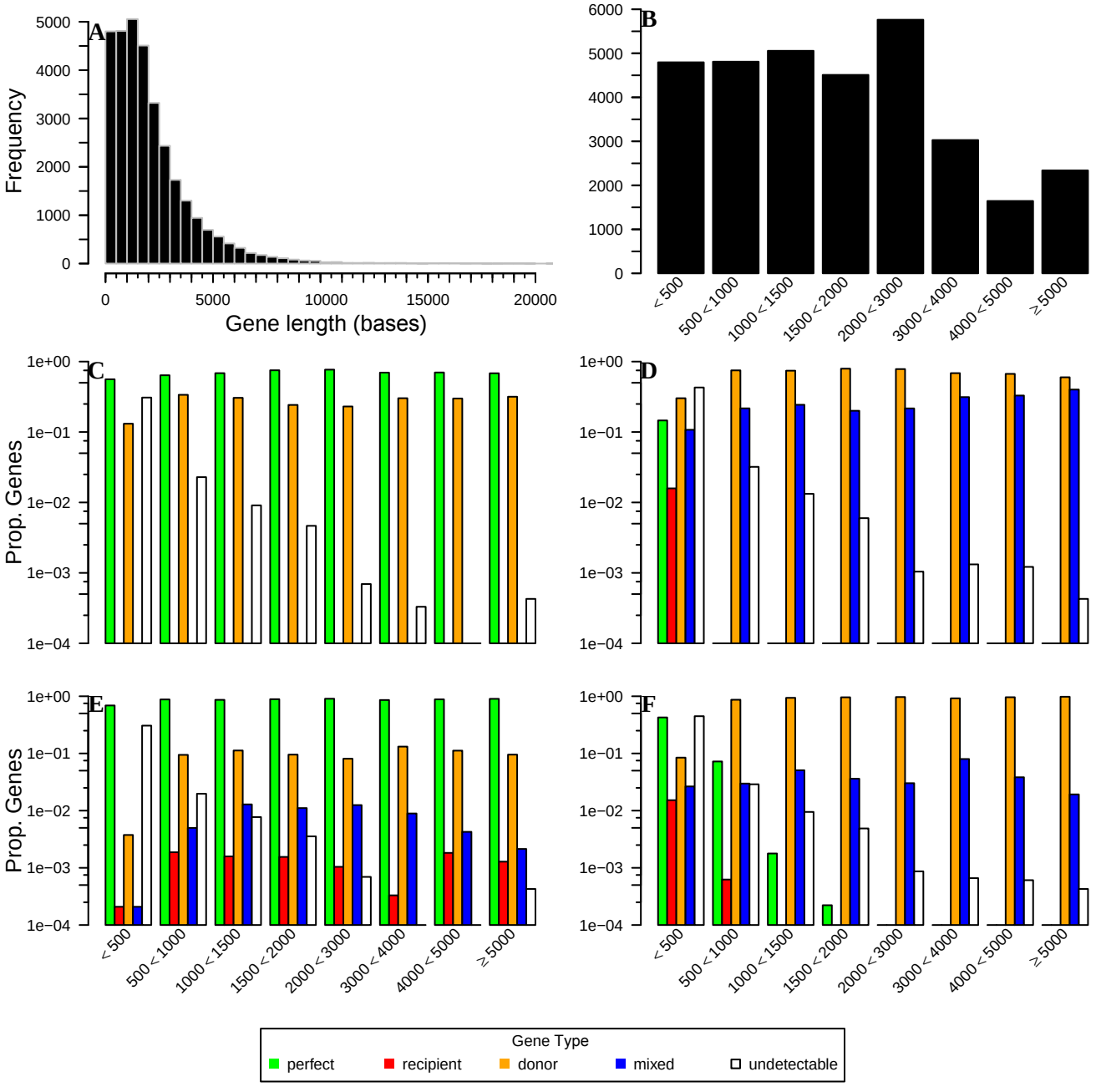
Figure 3. Predicted recovery rate (rr) using the mixed logistic modelization as a function of gene length (*rr*), read length and divergence between target and reference transcriptomes: red lines denote 0% divergence; green 5%; blue 15% & purple 30%. Solid lines correspond to the median of predictions, conditioned on random variation among genes, with 80% prediction intervals indicated by dashed lines. Read length increases downward and panels to the right represent a magnified view of the upper 5th quantile of *rr* scores to better visualize differences between low divergent sequences: 100 bases reads (A & B); 150 bases (C & D); 200 bases (E & F); 350 bases (G & H).

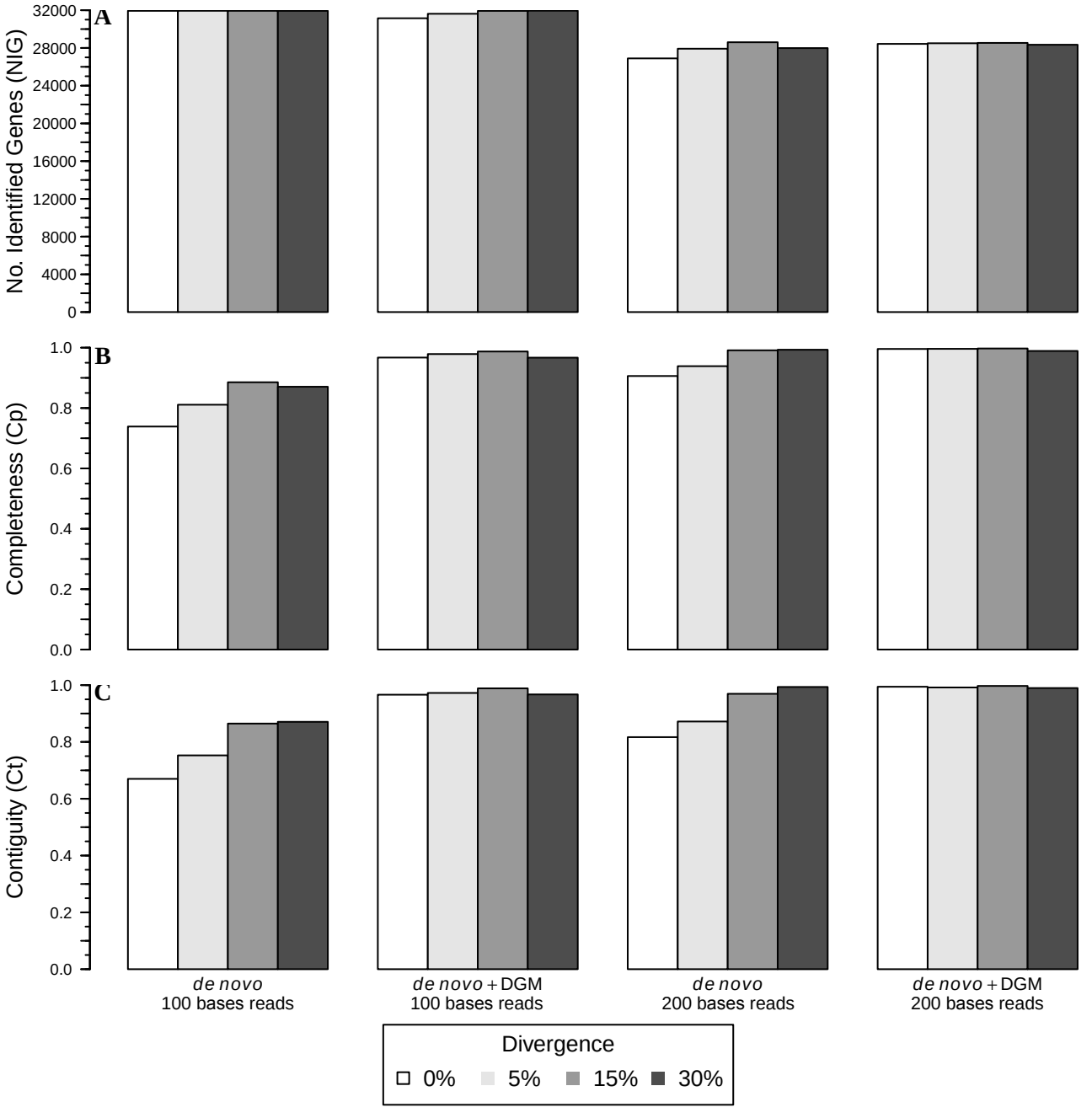
Figure 4. Proportion of gene types recovered in divergent simulations by size-class of gene. (A) Histogram of gene lengths for the *Danio rerio* transcriptome used for simulating RNA-seq reads. (B) Number of genes from A, by size-class in subsequent windows. 100

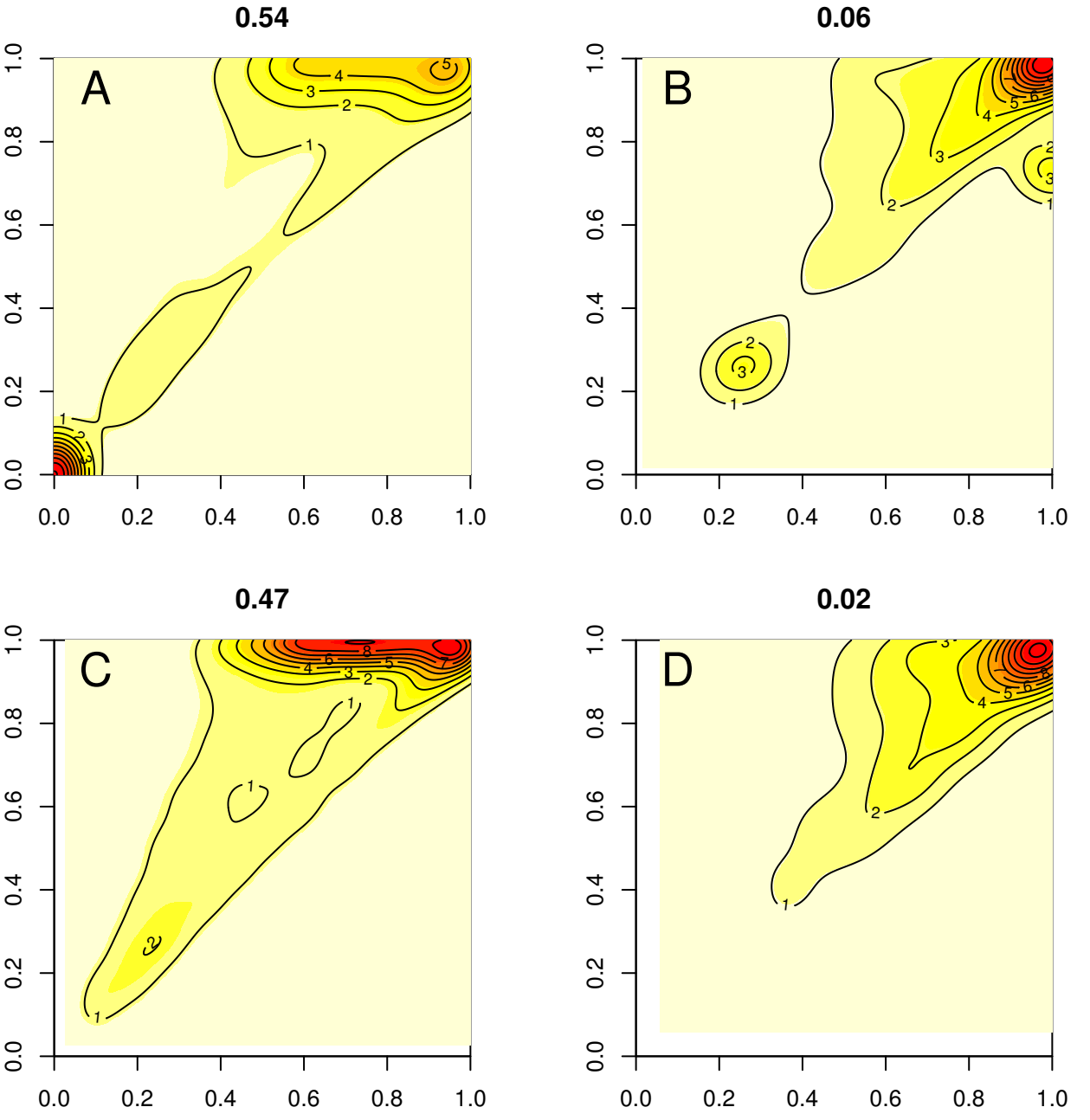
bases reads are plotted with 0% divergence (C) and 30% divergence (D); 350 bases reads with 0% divergence (E) and 30% divergence (F). Gene types are described in the legend.

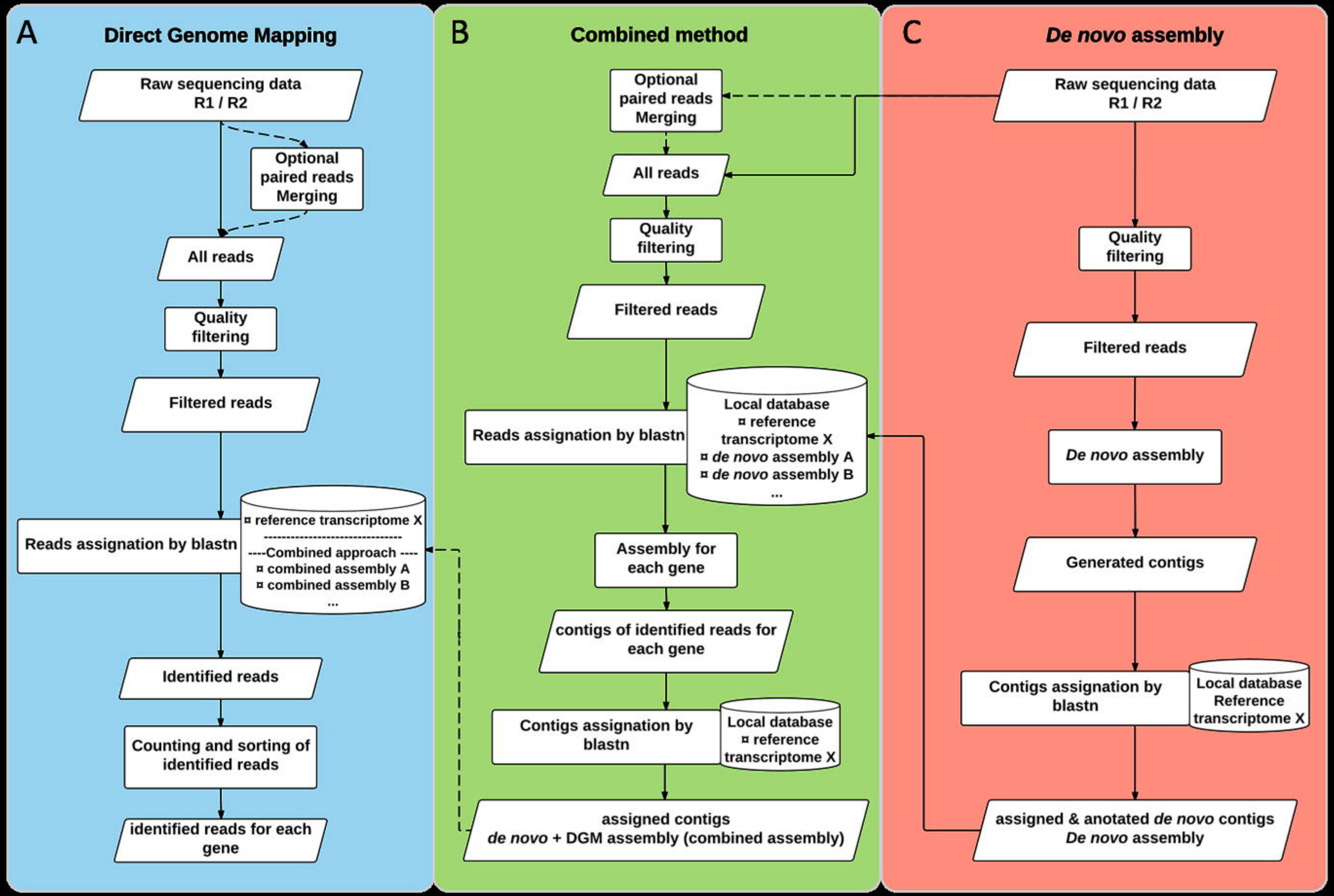
Figure 5. Efficiency and performance of *de novo* and *de novo* + DGM pipelines by read length and divergence. (A) Number of identified genes; (B) completeness; (C) contiguity. Divergence between target and reference transcriptomes are described in the figure legend.

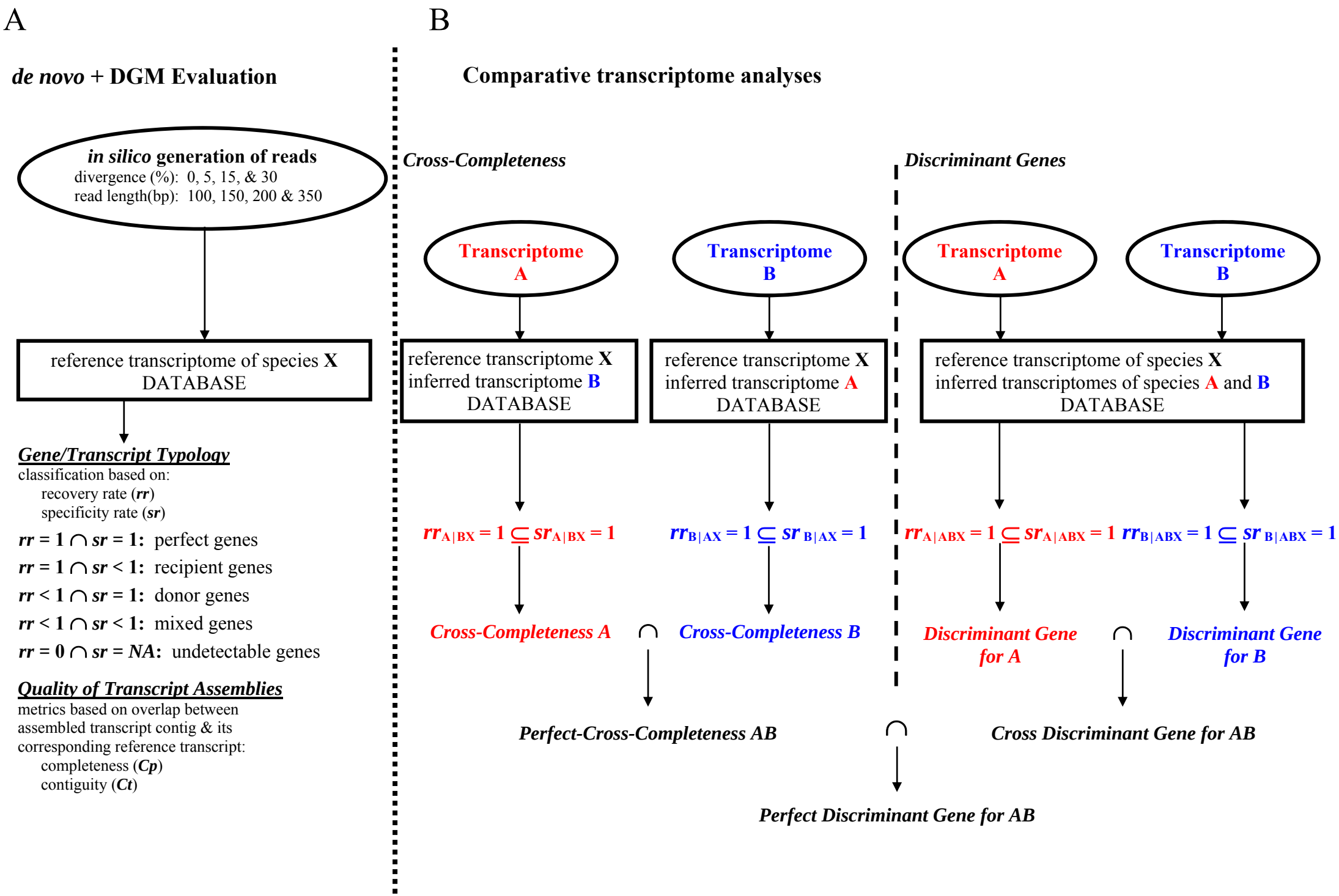
Figure 6. Nonparametric estimation of the contiguity (x-axis) and completeness score (y-axis) obtained from *de novo* and *de novo* + DGM pipelines for 0% divergence. Colors increasing from yellow to dark red denote increasing gene densities. Note also that for clarity of visualization, only the non-perfect fraction of genes is displayed. 100 bases reads assembled *de novo* (A) and with *de novo* + DGM (B); 200 bases reads *de novo* (C) and *de novo* + DGM (D). The proportion of non-perfect genes is indicated at the top of each item.

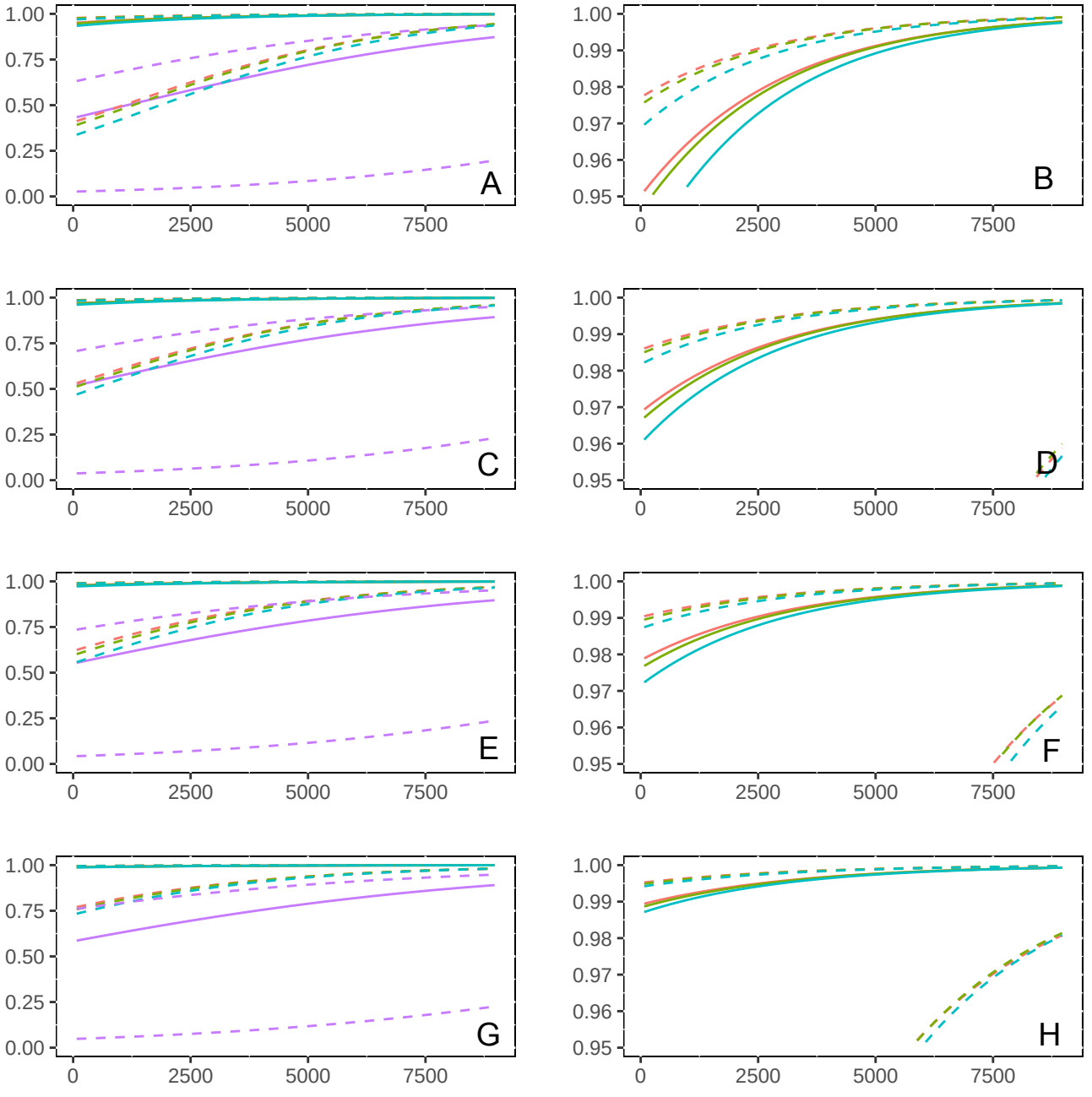

A
Frequency
Gene length (bases)
B
C
D
E
F
Prop. Genes
<500
500<1000
1000<1500
1500<2000
2000<3000
3000<4000
4000<5000
≥5000
Gene Type
perfect
recipient
donor
mixed
undetectable


A
No. Identified Genes (NIG)
32000
28000
24000
20000
16000
12000
8000
4000
0
B
Completeness (Cp)
1.0
0.8
0.6
0.4
0.2
0.0
C
Contiguity (Ct)
1.0
0.8
0.6
0.4
0.2
0.0
de novo
100 bases reads
de novo + DGM
100 bases reads
de novo
200 bases reads
de novo + DGM
200 bases reads
Divergence
0%
5%
15%
30%


0.54
A
0.06
B
0.47
C
0.02
D
0.0
0.2
0.4
0.6
0.8
1.0


A
Direct Genome Mapping
Raw sequencing data
R1 / R2
Optional
paired reads
Merging
All reads
Quality
filtering
Filtered reads
Reads assignation by blastn
¤ reference transcriptome X
----Combined approach ----
¤ combined assembly A
¤ combined assembly B
...
Identified reads
Counting and sorting of
identified reads
identified reads for each
gene
B
Combined method
Optional
paired reads
Merging
All reads
Quality
filtering
Filtered reads
Reads assignation by blastn
Local database
¤ reference
transcriptome X
¤ de novo assembly A
¤ de novo assembly B
...
Assembly for
each gene
contigs of identified reads for
each gene
Contigs assignation by
blastn
Local database
¤ reference
transcriptome X
assigned contigs
de novo + DGM assembly (combined assembly)
C
De novo assembly
Raw sequencing data
R1 / R2
Quality
filtering
Filtered reads
De novo assembly
Generated contigs
Contigs assignation by
blastn
Local database
Reference
transcriptome X
assigned & anotated de novo contigs
De novo assembly


A
B
de novo + DGM Evaluation
in silico generation of reads
divergence (%): 0, 5, 15, & 30
read length(bp): 100, 150, 200 & 350
reference transcriptome of species X
DATABASE
Gene/Transcript Typology
classification based on:
recovery rate (rr)
specificity rate (sr)
rr = 1 ∩ sr = 1: perfect genes
rr = 1 ∩ sr < 1: recipient genes
rr < 1 ∩ sr = 1: donor genes
rr < 1 ∩ sr < 1: mixed genes
rr = 0 ∩ sr = NA: undetectable genes
Quality of Transcript Assemblies
metrics based on overlap between
assembled transcript contig & its
corresponding reference transcript:
completeness (Cp)
contiguity (Ct)
Comparative transcriptome analyses
Cross-Completeness
Discriminant Genes
Transcriptome A
Transcriptome B
Transcriptome A
Transcriptome B
reference transcriptome X
inferred transcriptome B
DATABASE
reference transcriptome X
inferred transcriptome A
DATABASE
reference transcriptome of species X
inferred transcriptomes of species A and B
DATABASE
rrA|BX = 1 ⊆ srA|BX = 1
rrB|AX = 1 ⊆ srB|AX = 1
rrA|ABX = 1 ⊆ srA|ABX = 1
rrB|ABX = 1 ⊆ srB|ABX = 1
Cross-Completeness A
∩
Cross-Completeness B
Discriminant Gene for A
∩
Discriminant Gene for B
Perfect-Cross-Completeness AB
∩
Cross Discriminant Gene for AB
Perfect Discriminant Gene for AB


1.00
0.75
0.50
0.25
0.00
0
2500
5000
7500
A
1.00
0.99
0.98
0.97
0.96
0.95
B
C
D
E
F
G
H