MOLECULAR ECOLOGY

Molecular Ecology (2012) doi: 10.1111/mec.12014

Challenges and strategies in transcriptome assembly and differential gene expression quantification. A comprehensive *in silico* assessment of RNA-seq experiments

NAGARJUN VIJAY,*[1] JELMER W. POELSTRA,*[1] AXEL KÜNSTNER† and JOCHEN B. W. WOLF*

Department of Evolutionary Biology and Science for Life Laboratory, Uppsala University, Norbyvägen 18D, Uppsala, SE-752 36, Sweden, †Department of Molecular Biology, Max Planck Institute for Developmental Biology, Spemannstrasse 37-39, 72076 Tübingen, Germany

Abstract

Transcriptome Shotgun Sequencing (RNA-seq) has been readily embraced by geneticists and molecular ecologists alike. As with all high-throughput technologies, it is critical to understand which analytic strategies are best suited and which parameters may bias the interpretation of the data. Here we use a comprehensive simulation approach to explore how various features of the transcriptome (complexity, degree of polymorphism π, alternative splicing), technological processing (sequencing error ε, library normalization) and bioinformatic workflow (*de novo* vs. mapping assembly, reference genome quality) impact transcriptome quality and inference of differential gene expression (DE). We find that transcriptome assembly and gene expression profiling (*EdgeR* vs. *BaySeq* software) works well even in the absence of a reference genome and is robust across a broad range of parameters. We advise against library normalization and in most situations advocate mapping assemblies to an annotated genome of a divergent sister clade, which generally outperformed *de novo* assembly (Trans-Abyss, Trinity, Soapdenovo-Trans). Transcriptome complexity (size, paralogs, alternative splicing isoforms) negatively affected the assembly and DE profiling, whereas the effects of sequencing error and polymorphism were almost negligible. Finally, we highlight the challenge of gene name assignment for *de novo* assemblies, the importance of mapping strategies and raise awareness of challenges associated with the quality of reference genomes. Overall, our results have significant practical and methodological implications and can provide guidance in the design and analysis of RNA-seq experiments, particularly for organisms where genomic background information is lacking.

Keywords: bioinformatics, comparative genomics, differential gene expression, RNA-seq, simulation, systems biology, transcriptome assembly

Received 21 March 2012; revision received 13 June 2012; accepted 11 July 2012

Introduction

The study of gene expression has traditionally been reserved for genetic model organisms. For organisms like human, *Drosophila* or *Arabidopsis*, rich genomic resources readily allow the design of microarrays to examine global patterns of gene expression and various features of the transcriptome. For genetic nonmodel organisms, gene expression studies were long restricted to qPCR analyses of candidate genes (Axtner & Sommer 2009) or had to rely on cross-species hybridization on microarrays (Naurin *et al.* 2011), which ultimately remains a compromise (Bar-Or *et al.* 2006).

Massively parallelized RNA sequencing technology (RNA-seq) has added a valuable tool: millions of short reads are generated from steady-state RNA and concomitantly provide transcriptome sequence information and a digital measure of gene expression (Ozsolak &

Correspondence: Jochen B. W. Wolf, Fax: +46 (0)18 471 6310; E-mail: jochen.wolf@ebc.uu.se

[1] Equally share first authorship.

Milos 2010). Compared with microarray data, results from a number of studies suggest that RNA-seq is generally more accurate and captures a broader range of expression levels (Marioni *et al.* 2008; Fu *et al.* 2009). It also holds great promise to detect unknown transcripts and unravel previously inaccessible complexities such as allele-specific expression, novel splicing variants or promotors (Ozsolak & Milos 2010). It is therefore not surprising that RNA-seq has become the technology of choice for transcriptome investigation (Deng *et al.* 2011).

For molecular ecologists, RNA-seq has opened the unprecedented opportunity to explore transcriptomes of basically any species in a number of different ways (Ekblom & Galindo 2011). For example, consensus sequences from *de novo* assembled transcriptomes have been used for comparative genomic analyses (Elmer *et al.* 2010; Künstner *et al.* 2010), to obtain resources for SNP genotyping or for the design of custom microarrays to study gene expression (Kvist *et al.* 2012). Gene expression levels have also been directly inferred from the number of sequencing reads themselves, independent of prior genomic knowledge of the species in question. This approach has recently gained considerable momentum in a broad range of research areas, including the role of differential gene expression (DE) in phenotypic divergence and speciation (Lenz *et al.* 2012; Goetz *et al.* 2010; Wolf *et al.* 2010) and areas such as dosage compensation (Wolf & Bryk 2011) and alternative splicing (Harr & Turner 2010) that have been traditionally reserved to genetic model systems under laboratory conditions.

Although an increasing number of empirical studies convincingly portray RNA-seq as a promising tool for genetic nonmodel organisms, little attention has been paid to the parameters bearing on transcriptome quality and RNA-seq based measures of gene expression. This is surprising, as many aspects including transcriptomic features (e.g. size, repetitiveness, normalization), quality (e.g. library preparation, clustering efficiency) and quantity of the short-read data, as well as the bioinformatic processing (e.g. transcriptome assembly, read alignment, statistical modelling of read count data) alter the current state of the data on which all subsequent downstream analyses depend. So far, only isolated aspects such as mapping accuracy (Palmieri & Schlötterer 2009), *de novo* assembly (Grabherr *et al.* 2011) and statistical approaches to differential expression analysis (Kvam *et al.* 2012) have been explored.

We here use extensive in silico computer simulations based on two vertebrate genomes (zebra finch and human) to evaluate the performance of standard RNA-seq pipelines from whole-transcriptome sequencing to its assembly, subsequent measurements of gene expression and statistical inference of differences between treatment groups (Fig. 1).

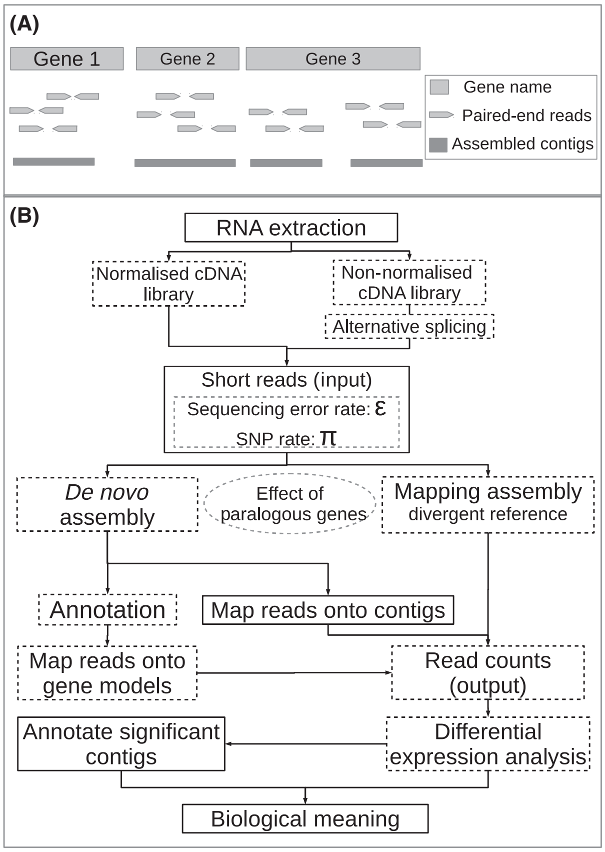

Fig. 1 Overview of an RNA-seq pipeline with special focus on nonmodel organisms lacking a reference genome. (A) Graphical display of how short reads are sequenced from a gene and are then assembled into contigs. (B) Workflow of a typical RNA-seq experiment. Dashed boxes refer to the parameters that are specifically investigated in this study.

In a first step, we explore how the quality of a freshly assembled transcriptome is affected by the following set of parameters:

1. Mode of assembly (three *de novo* assemblers vs. mapping assembly on divergent reference genomes),
2. transcriptome complexity (~4400 unique genes vs. ~17400 genes including paralogs of different age),
3. RNA library normalization (uniform vs. strongly skewed gene expression profile),
4. different levels of sequencing error ε (0, 0.01),
5. polymorphism π (0, 0.001, 0.01),
6. annotation quality of the reference genome and
7. alternative splicing.

In a second step, we assess how well the simulated expression levels are represented after passing through the entire RNA-seq pipeline (input vs. output), and what the influences of the above mentioned parameters are. Last, we investigate their effects on differential expression metrics (false and true positives). We anticipate that this exploration will be useful both in guiding the experimental and bioinformatic design of RNA-seq experiments and interpreting the results against an *in silico* null model.

Materials and methods

Data simulation

We chose the zebra finch transcriptome as the main backbone of our simulations (Warren *et al.* 2010), which has around 17 475 protein coding genes (ENSEMBL 61, www.biomart.org, mean length of 1843 bp, range: 560–26 831 bp) of which 13 752 have ENSEMBL gene status 'known', 3235 have status 'novel', 131 have status 'putative' and 357 have status 'known by projection'. Parts of the assembly cannot be precisely placed along the chromosome and are located on a 'random' version of the chromosome (878 genes) or are not localized to any chromosome (3113 'Un' genes) in ENSEMBL. Similar annotation uncertainties will be found in most other moderately well curated genomes (Church *et al.* 2011). We are confident that the results presented below are comparable with other species, but it may still be useful to follow some of the steps before starting an RNA-seq experiment on any particular species (all scripts are accessable via Dryad doi:10.5061/dryad.3t3n7).

Complete annotated Coding DNA Sequence (CDS) were downloaded from BioMart (ENSEMBL 61, www.biomart.org) resulting in a total of 17 475 genes of which 75 were randomly removed to achieve equal binning of expression levels (see below). As untranslated regions (UTRs) are not well annotated in the zebra finch genome (~8% of the genes have annotated 5′ UTRs and ~21% have annotated 3′ UTRs, respectively), we added 100bp 5′ and 400bp 3′ (median length of annotated 5′/3′ UTRs: 100/330, 1st quantile 56/198, 3rd quantile 167/473) to mimic the situation encountered when sequencing 'real' EST libraries. The zebra finch transcriptome contains a considerable number of gaps (239 786 Ns) which were replaced with bases chosen at random using average CDS nucleotide frequencies. Paired-end sequencing (100 bp read length, 300 bp insert size) was simulated using DWGSIM (version 0.1.2 http://sourceforge.net/apps/mediawiki/dnaa/index.php?title=Whole_Genome_Simulation, accessed on 12th January 2012) to mimic data from a standard Illumina based RNA-seq experiment (for coverage see below).

To address various crucial issues inherent in RNA-seq experiments, the following data sets were simulated (see Fig. 2a and Introduction):

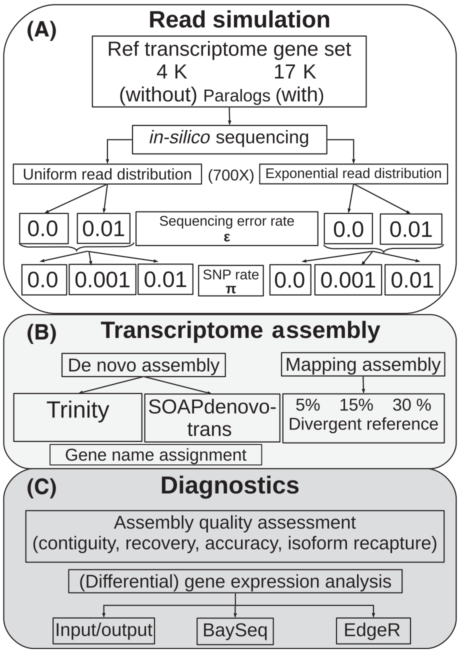

Fig. 2 Overview of this study. (A) Description of the 24 raw data sets that were simulated to address aspects of transcriptome complexity (17 400 [17K] vs. 4400 [4K] genes), library normalization (uniform vs. exponential), sequencing error ε and polymorphism levels π. (B) These data sets consisting of short reads were then assembled in different ways by two *de novo* assembly approaches and a mapping assembly strategy to reference genomes of 'closely related species' that differed in the degree of sequence divergence (5%, 15%, 30%). (C) Evaluation of how parameters in (A) and assembly strategies in (B) influence transcriptome quality and measures of (differential) gene expression. In this step gene name assignment and the influence of gene annotation quality of the reference genome were also explored.

1 Transcriptome complexity (size, presence of paralogs) will likely adversely affect the assembly and may also influence estimation of DE. We simulated reads based on a simple transcriptome reference including only unique genes (4400 genes = 4K) without paralogs, and a more realistic large complex transcriptome including paralogs (full 17 400 gene set = 17K). Paralog status was inferred from the homolog filter in BioMart (ENSEMBL Version 61).

2 Data generated in an RNA-seq experiment will have large variation in gene coverage corresponding to different levels of expression. While it is possible to normalize mRNA extracts to obtain a more even distribution of gene transcripts, such a procedure compromises the subsequent use of differential expression analyses (but see Ekblom *et al*. 2012). We simulated both a scenario where reads are uniformly distributed across all genes ('normalized library') and a realistic scenario of highly skewed read coverage per gene.

While expression profiles across genes are usually approximated with a gamma-distribution [e.g. (Kvam *et al.* 2012)], the exponential distribution provides a good compromise. It captures the strong right-skew of expression data (Figs S1 and S2, Supporting information) and only uses one parameter λ (=mean coverage) which makes it suitable for simulation. To mimic skewed gene expression, genes were randomly distributed into 100 bins corresponding to read coverage given by the percentiles of the exponential distribution.

3 In order to evaluate the tolerance of assembly and mapping tools with respect to sequencing errors and polymorphism rates, we set sequencing error rate (ε) to 0 and 0.01 [ε for Illumina data ~0.05 for raw data (Pareek *et al.* 2011)] and the polymorphism rate (π) to 0, 0.001 and 0.01 (thus covering a realistic range).

In total, the combination of these scenarios results in 24 different data sets (Fig. 2a). Sequencing-bias in CG-content was not simulated to keep the parameter space to a manageable size. Alternative splicing was investigated in separate simulations on the human genome (see below).

For all scenarios, we simulated ~100 million paired-end 100 bp reads for the 17K data set and ~20 million paired-end reads for the 4K data, respectively which translates to a ~700× coverage per base varying between 2× and 4000× for the skewed distribution. The number of reads chosen was based on the hallmark output currently obtained from one lane of the Illumina HiSeq 2000 technology after adapter and quality trimming (~100 million paired-end reads). This amount of data seems realistic for a broad range of RNA-seq projects and should capture most of the genes present in an RNA sample (Wang *et al.* 2011).

Transcriptome assembly

In order to examine the parameters in questions, we first needed to assemble the simulated reads into contigs. In an ideal world, every contig would correspond to exactly one transcript from which reads were simulated (Fig. 1) which in reality is, however, rarely the case (see recovery and accuracy below). Each of the 24 data sets was assembled using both *de novo* and mapping based strategies. *De novo* assemblies were performed using TRINITY version r2011-08-20 (Grabherr *et al.* 2011) and SOAPDENOVO-TRANS version 1.01 (http://SOAP.genomics.org.cn/SOAPdenovo-Trans.html, accessed on December 12th 2011). In TRINITY the assemblies were performed with default settings. SOAPDENOVO-TRANS assemblies were performed at all odd k-mers between 21 and 99. The k-mer with the best recovery (definition see below) was selected for each data set (see Fig. S3, Table S4, Supporting information). TRANS-ABYSS was tested for the simplest data set and found to have significantly lower recoveries than both TRINITY and SOAPDENOVO-TRANS and was therefore excluded from further analyses (results not shown).

To mimic a reference-based mapping assembly, we generated three differently diverged reference transcriptomes (sequence divergence 5%, 15%, 30%) using DAWG (Cartwright 2005) with a Jukes-Cantor model of nucleotide substitution. Mapping assemblies were then performed on the in silico reference transcriptomes using STAMPY version 1.0.13 (Lunter & Goodson 2011) with the *bwa* option to speed up the mapping process. STAMPY was chosen to make use of its hybrid mapping algorithm which has been shown to be more sensitive and efficient in mapping divergent reads (Lunter & Goodson 2011). For each of the data sets, the consensus sequence was called based on the STAMPY mapping output *sam* file using SAMTOOLS version 0.1.12 with *pileup* with default options (Li *et al.* 2009).

Assessment of assembly quality

Recovery and accuracy. Efficiency of the different strategies and tools was evaluated by calculating the *recovery* and *accuracy* of the assemblies. Recovery refers to the proportion of bases from the reference transcriptome recovered in the assembled transcriptome, and accuracy refers to the proportion of bases that correctly matched the orthologous position in the reference genes. Recovery was calculated separately for fully and partially assembled genes (Fig. S4, Supporting information). Contigs that contained more than one gene were designated as chimera and their recovery was calculated separately, considering only the gene that was covered to a larger extent.

For mapping assemblies, recovery and accuracy can be directly calculated by comparing the reference transcript to the assembled consensus sequence for a given transcript. For *de novo* assemblies, contigs first need to be aligned to the reference transcriptome. We used the *NUCmer* alignment tool from the MUMMER package version 3.22 (Kurtz *et al.* 2004) to align all contigs generated by the assemblers to the respective reference transcriptomes (4K, 17K). The contigs were assigned to gene names based on the gene to which they best aligned (for details, see Fig. S5, Supporting information). For partially or fully overlapping contigs that aligned to the same gene, we used the *coverageBed* command from BEDTOOLS version 2.11.2 (Quinlan & Hall 2010) on the alignments reported by *NUCmer* to calculate the maximum total recovery. Only those contigs that aligned to reference genes were used to calculate recovery and accuracy.

While measures of recovery and accuracy provide a good comparison of the overall assembly quality, they do not provide any information about the structural errors prevalent in the assemblies. We utilized the scripts available from GAGE (Salzberg *et al.* 2012) to find the presence of various structural errors such as inversions (part of a contig reversed with respect to the reference gene), translocations (chimera), relocations (within a gene), duplications, repeat compression, and insertions and deletions (indels).

Contiguity. Contiguity indices such as N50 can give an indication about how fragmented the recovered transcripts are (N50 is defined as the length of the contig such that half of all bases in the assembly are made of sequences of equal or longer length). However, N50 values can be misleading for transcriptome assemblies as transcript length is highly heterogeneous. We therefore standardize contig N50 by the expected N50 given by the actual transcript length of the reference ($N50_{ratio}$). In case of full recovery of every transcript, the $N50_{ratio}$ will be 1. Chimeras can increase the value and should be treated separately. This index can also be used if a reference genome is only available for a closely related species under the justified assumption that gene length is usually well conserved (Xu *et al.* 2006).

Assessing effects of transcriptome complexity. Paralogous genes constitute a particular challenge for both *de novo* and mapping assembly strategies. To examine the susceptibility of paralogs to biases in mapping and *de novo* assembly, we used the contrast between the paralog-free 4K gene set and the full 17K gene set. Note that transcriptome complexity here refers to a combination of size and presence of paralogs.

Gene annotation. Gene annotation is of special concern in nonmodel organisms lacking reference genomes. In the absence of a reference genome, gene annotation hinges on the availability of transcriptome sequences from the closest available taxon. We compared the performance of suffix-tree-based methods (*NUCmer*, *PROmer*) and slower intensive alignment tools such as BLAST2GO, SATSUMA and PAPAYA for gene annotation across different levels of divergence.

Simulation of differential gene expression

For each of the 24 generated data sets (see Fig. 2a), we simulated 20 biological replicates by randomly sampling reads for each gene. We chose an average per-replicate sequence coverage of 35×, corresponding to on average of 500 read pairs per gene. These 20 libraries were partitioned into two treatment groups (i.e. conditions, phenotypes or populations, in a biological context) of 10 individuals each, between which differential expression was simulated. A $\log_2$-fold change (LFC) in expression levels between these two groups was randomly assigned for each gene, with in total 50% of genes with a LFC of 0, 20% with a LFC of 0.5, 16% with a LFC of 1, 8% with a LFC of 2, 4% with a LFC of 3 and 2% with a LFC of 4. Average expression levels were kept the same across LFC classes. For each gene, read counts for each individual library were generated according to a negative binomial (NB) distribution using the *rnbinom* function in R version 2.14.0. The 'mean' parameter of the NB distribution was set to the previously assigned mean level of expression for a gene across the 10 libraries in each treatment group (*i.e.* using a different mean for each treatment group). The 'size' parameter of the NB distribution represents the reciprocal of the dispersion parameter, which was in turn randomly drawn from a gamma distribution with shape and scale parameters set to 0.85 and 0.5, respectively [as in (Hardcastle & Kelly 2010)].

Assessment of parameter influence on gene expression levels

Comparing the number of reads per gene that was simulated (input ~ RNA concentration) to that which is actually counted for, the same gene after the assembly and mapping steps (output ~ RNA-seq count data) provides a quantitative measure of the bias caused by the bioinformatic processing (Fig. 1b). As the origin for each of the simulated reads was known, we also inferred the proportion of erroneously mapped reads using a random assignment strategy of multi-mapped reads in contrast to only quantifying best mapping or uniquely mapping read pairs.

de novo assembly. In a first step, gene names were assigned to contigs obtained by the SOAPDENOVO-TRANS and *Trinity* assembler using *NUCmer* (see above). These assemblies were then used as a reference to map the paired-end reads originally used for the assembly (input) using BWA version 0.5.9 (Li *et al.* 2009) with default settings. The number of reads aligning to contigs that had been assigned gene names was then used to calculate the number of reads which would in a real case scenario be used as a digital measure of gene expression (output).

Mapping assembly. Reference genome-based transcriptome assemblers like CUFFLINKS (Trapnell *et al.* 2010) directly use read counts from the initial mapping step to infer (differential) gene expression. When mapping reads to a distant genome, a consensus sequence has to be generated prior to the mapping step. This has the additional advantage that intronic regions and pseudogenes are excluded during the mapping step which can reduce the problems of mismapping at the cost of missing out on unannotated genes or mismapping of reads from unannotated genes onto annotated paralogs. We used consensus sequences obtained from SAMTOOLS from mapping assemblies generated using STAMPY (see above) as references to map paired-end reads using *bwa*.

Assessment of parameter influences on differential gene expression (DE)

Differential expression analysis of the mapped read counts was conducted with EDGER version 2.4.3 (Robinson *et al.* 2010) and BAYSEQ version 1.8.2 (Hardcastle & Kelly 2010). Library sizes were normalized using the TMM method in *EDGER*, and the quantile method in *bay-Seq*. In *EDGER*, dispersion was estimated on a tagwise (i.e. gene-by-gene) basis using the dispersion estimation prior determined by the *getPriorN* function. The prior determines the amount of squeezing of the dispersion estimate for each gene towards the estimated mean dispersion. For our comparison of 10 biological replicates for each group, the *getPriorN* function returned a prior of 1.33. This prior performed better than either a common (i.e. nontagwise) dispersion estimate, or tagwise dispersion estimates with priors 0 (no squeezing towards the mean) and 5 (strong squeezing towards the mean) (data not shown). In BAYSEQ, prior parameters were estimated by resampling ($n = 10\,000$ as recommended); otherwise, default estimation parameters were used.

Performance in the inference of differential expression was compared for transcriptome complexity (two levels), polymorphism and error rates (six levels), assembly types (seven levels) and DE software (two levels) (*cf.* Fig. 2). In addition, we compared performance for two levels of gene expression: genes with very low expression (average number of reads <25, 6.5–7.8% of genes), and genes with average to high expression (average number of reads of 100 or more, 72.8–76.5% of genes). The combination of these conditions amounts to a total of 336 data sets.

Performance was measured in several ways. First, we computed false positive rates (FPR: number of genes that were simulated to have no differential expression (LFC = 0), but were inferred to be differentially expressed) as well as true positive rates (TPR) for each LFC category with differential expression (0.5, 1, 2, 3, and 4). Second, we plotted receiver-operator-characteristic (ROC) curves using the ROCR package version 1.0-4 (Sing *et al.* 2005) in R. Third, for each data set, we calculated the correlation between the gene-by-gene LFC among the simulated read counts (input), and the gene-by-gene LFC among the mapped read counts (output). As comparisons between conditions (e.g. 4K and 17K genes) are replicated across many other of the 336 conditions, we could perform paired Mann-Whitney-*U* tests to test for significant differences among LFC correlations, FPRs and TPRs (the latter at each level of LFC).

Influence of the annotation quality of the (distant) reference genome

We finally tested whether the annotation status of a gene ('known' vs. 'novel', 'putative' or 'known by projection') or knowledge of chromosome location ('known' vs. 'random' or 'unlocated') influenced the estimates of (differential) gene expression. We calculated the proportion of mismappped reads/number of simulated reads for each gene and averaged across all genes for each of the parameter combination (see above). We then compared the influence of annotation quality and the interaction with assembly methodology in an ANOVA framework where each parameter combination represents one independent data point.

Alternative splicing

With respect to alternative splicing, we limited ourselves to estimate isoform usage (per gene abundances and frequency distributions) from empirical RNA-seq data, in order to keep our already extensive simulation set-up tractable. We chose to use the human genome, as high annotation quality is vital for this exercise. RNA-seq data from four liver libraries of four different individuals (Perry *et al.* 2012) were pooled and mapped to the human genome (version 37 with ENSEMBL 66 annotation). Read counts for each of those genes [using RSEM (Li & Dewey 2011)] were then used to infer the read distribution per gene and isoform. This information was then used to simulate transcriptome data in the same way as for the zebra finch based analyses (see above), including 73 629 (of 150 465 annotated) isoforms expressed in 21 405 protein coding genes from the liver data set considered. All the isoforms belonging to the same gene were forced to have the same level of LFC between groups. To reduce statistical noise all analyses were restricted to genes with a minimum expression level [FPKM value >0.5, *cf.* (Grabherr *et al.* 2011)]. Overall, this set-up allowed us to analyse performance both at the gene-level and the isoform-level.

Analogous to the main simulation described above, a subset of parameter values were explored. We assessed the relative performance of *de novo* vs. mapping assemblies and the effects of different expression levels. When testing the effect of sequencing error (ε) and polymorphism (π), we restricted the simulations to the most extreme scenarios ($\varepsilon/\pi = 0/0$ and $0.01/0.01$, respectively).

We only used the complex, empirically based transcriptome and restricted the DE analysis to *edgeR*. For *de novo* assembly, we used TRINITY with default settings. Assembled transcripts were assigned to gene/isoform names based on alignments to the human transcriptome (Fig. S5, Supporting information). RSEM was used for inference of read counts per transcript (isoform) for each of the 20 libraries. For mapping assemblies we first generated (5%, 15%, 30%) divergent reference transcriptomes using *Dawg* with the Jukes-Cantor model of nucleotide substitution. The simulated reads were mapped onto these transcriptomes using STAMPY allowing for multiple mapping. SAMTOOLS was used to call consensus sequence after converting all the alternate mapping positions into multimapping positions using the script xa2multi.pl from SAMTOOLS so as to accommodate isoforms. RSEM was used to find read counts per transcript for each of the 20 individuals by mapping reads onto the human consensus transcriptome obtained by the mapping assembly in the previous step.

Results

Assembly quality

Assembly type. Mapping assemblies on 5% and 15% divergent reference transcriptomes had on average higher recoveries than *de novo* assemblies across all conditions (Fig. 3 and Fig. S6a,b, Table S1, Supporting information). Mapping assemblies based on a 30% divergent reference fared notably worse. Even though having higher recovery, mapping assemblies had slightly lower accuracy values than de novo assembled contigs (SOAP DENOVO-TRANS: mean accuracy 99.75; TRINITY: mean accuracy 99.65%; mapping to 5% divergent reference: 99.37%; 15%: 98.18%; 30%: 93.98%). Mapping assemblies also outcompeted *de novo* assemblies in the presence of isoforms. The difference in recovery between mapping and *de novo* assembly was here most pronounced for lowly expressed genes and almost vanished for highly expressed genes (Table S1, Supporting information).

Among *de novo* assemblies, those generated by SOAP-DENOVO-TRANS (SOAP hereafter) had a slightly higher recovery than TRINITY (Fig. 3), but comparable accuracies (Table S1, Supporting information). Both assemblers produced relocations in comparable number and only one of the TRINITY assemblies had one inversion (exponential distribution, ε = 0.01, π = 0.01). Indels were more common for data sets with higher sequencing error and polymorphism rates for both TRINITY assemblies and SOAP assemblies. All assemblies had a considerable number of compressed repeats and duplications (Table S2, Supporting information).

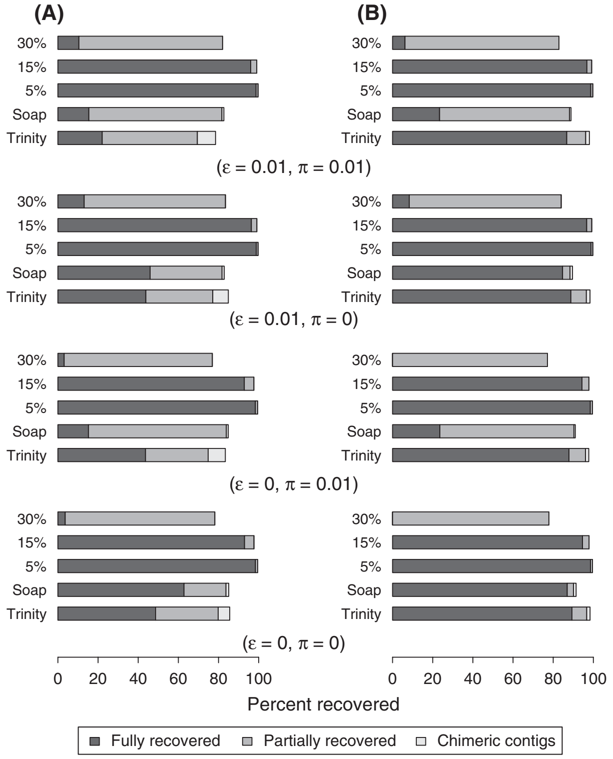

Fig. 3 Recovery of the reference transcriptome for (A) the full set of 17 400 genes and (B) a less complex transcriptome of 4400 genes (both with highly skewed read distribution) displayed in relation to assembly strategy, sequencing error ε and polymorphism levels π.

TRINITY assemblies appeared to be more contiguous than assemblies obtained from SOAP (Fig. 3), particularly in situations with high sequencing error and polymorphism levels (Table S1, Supporting information). While this was partly owing to the presence of chimera which were twice as common for the TRINITY assembler (Fig. 3, Table S3, Supporting information) contiguity of TRINITY assemblies remained higher even after excluding chimeras ($N50_{ratio}$ values: TRINITY mean 0.98; range 0.8–1.04; SOAP mean 0.75; range 0.24–1).

De novo assemblies produced by the TRINITY assembler provide additional information about the contigs generated by grouping alleles, isoforms and paralogs into 'components'. We found that even in the data sets simulated *without* alternative splicing, no sequencing error, no polymorphism and no paralogs for 7.87% of the genes many isoforms were erroneously inferred (ranging from 2 to 335 isoforms per gene, mean/median 100/55 isoforms, Fig. S7, Supporting information). While an increase in transcriptome complexity (4K–17K) strongly increased their number, higher polymorphism and error levels had only slight effects (Fig. S7, Supporting information).

Transcriptome complexity. For *de novo* assemblies, recovery was considerably lower for complex transcriptomes (17K vs. 4K) across all conditions (Fig. 3 and Fig. S6a,b, Table S1, Supporting information). Complex transcriptomes were also less contiguous as indicated by lower $N50_{ratios}$ particularly for the SOAP assembler (mean $N50_{ratios}$ 17K/4K SOAP: 0.71/0.80; TRINITY: 0.99/0.98). For mapping assemblies, recovery was less compromised by transcriptome complexity (Fig. 3).

Read count distribution ('library normalization'). The uniform read distribution data sets had comparable, but in all cases slightly higher recovery than exponential read distributions (*uniform/exponential* mapping assembly: mean 94.40/92.85%, ranges 83.74–98.50/76.91–99.74%; *de novo*: mean 93.15/89.01, 83.74–98.50/78.52–98.35%). While the read distribution slightly influenced the degree of fragmentation and the proportion of chimeric contigs, its overall influence on contiguity was low (Fig. S6a,b, Tables S1–S3, Supporting information).

Polymorphism and error levels. Overall, polymorphism and error level had only a very small influence on assembly quality. While this is counter intuitive at first, an increase in both polymorphism and error levels resulted in a small increase in recovery for mapping assemblies [($\varepsilon = 0$, $\pi = 0$)/($\varepsilon = 0.01$, $\pi = 0.01$): mean 92.36/94.66%, range 77.86–99.80%/82.02–99.97%). The increase was accompanied by a decrease in accuracy indicating the presence of mis-assemblies [($\varepsilon = 0$, $\pi = 0$)/($\varepsilon = 0.01$, $\pi = 0.01$): mean 97.60/96.62%, range 92.84–99.75/89.22–99.46%).

De novo assemblies followed a different pattern: recovery dropped with sequencing error and polymorphism (Fig. 3), but accuracy remained basically constant [recovery ($\varepsilon = 0$, $\pi = 0$)/($\varepsilon = 0.01$, $\pi = 0.01$) mean 91.95/89.98%, range 85.09–98.50/78.52–98.14%; accuracy (ε=0, $\pi = 0$)/($\varepsilon = 0.01$, $\pi = 0.01$) mean 99.99/99.35%, range 99.98–100.00/99.33–99.41%) indicating that 'problematic reads' get purged during the assembly process. At high levels of sequencing error and polymorphism recovery, contiguity was more negatively impacted for SOAP than for TRINITY (Fig. 3, Tables S1–S3, Supporting information).

Data sets simulated from the human transcriptome with alternative splicing showed similar trends. *De novo* assemblies with no sequencing error and no polymorphism ($\varepsilon = 0$, $\pi = 0$) had a recovery of 34% compared to a recovery of 29% with sequencing error and polymorphism ($\varepsilon = 0.01$, $\pi = 0.01$). Mapping assemblies on 5% divergent transcriptome performed better [($\varepsilon = 0$, $\pi = 0$): 67% & ($\varepsilon = 0.01$, $\pi = 0.01$): 64%). The increase in recovery with increasing error and polymorphism rates seen in the zebra finch data set was also observed in the human data set for the mapping assemblies at 15% and 30% divergence. The increase in recovery was again accompanied by a slight decrease in accuracy (Table S1, Supporting information).

Gene name assignment

In mapping assemblies, direct inference of gene names is in principle possible. However, in a realistic situation where a tissue-specific subset of the genome (here 4K) is annotated with the full CDS (17K, based on *NUCmer* alignments), genes from the 4K set were incorrectly assigned in a number of cases even when divergence was 0%. Mis-assignment increased with divergence (e.g. exponential distribution ($\varepsilon = 0$, $\pi = 0$)/($\varepsilon = 0.01$, $\pi = 0.01$): 5% 299/358 genes, 15%: 361/407 genes, 30%: 375/417 genes).

For *de novo* assembly, contigs need to be cross-linked to the annotated reference. In the majority of the genes, gene names could be confidently assigned by aligning contigs to 0% (mean 97.37%, range 97.04–97.78%), 5% (mean 96.42%, range 96.00–96.90%) and 15% (mean 88.21%, range 85.09–90.16%) divergent reference genomes. Genes from the 4K set were incorrectly assigned when using the 17K reference in a number of cases (but less than for mapping assemblies) even when divergence was 0% (mean: 99, range: 55–147). With increasing divergence (5%, 15%) the number of incorrect assignments increased (5%: mean 156, range 147–165, 15%: mean 158, range 147–165, see Tables S5–S6, Supporting information). As the divergence to the reference increased to 30% a large number of genes could not be aligned to the same contigs using faster tools like NUCMER (mean 0.24%, range 0.14–0.29%, see Table S6, Supporting information) and PROMER (see Table S6, Supporting information). More sensitive tools like BLAST, BLAT, BLASTZ, BLAST2GO or the SPINES package, on the contrary, were able to assign a higher proportion of contigs, but were struggling with apparent mis-assignment (Appendix S1, Supporting information).

Gene expression

We assessed how well the original number of simulated reads (input ~ sequenced steady-state mRNA) was still reflected after passing through the bioinformatic pipeline (=output, Figs 1 and 2). We also assessed the number of mismapped reads that were simulated for gene X, but were incorrectly assigned to gene Y. In the following, we report how the different parameters impact on read recovery (correlation $\rho_{Spearman}$ and output/input proportion = recovery) and the proportion of mismappings.

Overall, read counts after assembly and/or mapping were in good accordance with simulated expression values for all assemblies under all conditions (ρ_{Spearman}: mean 0.91, range 0.75–1.00). Recoveries were generally also high. This was, however, partly due to a considerable amount of mismapped reads that for some genes would lead to an overestimation of gene expression [Fig. 4, recovery: mean 98.43% (80.2% correct), range 62% (39.88% correct)–111% (96.16% correct)].

Assembly type. Read counts were slightly closer to the input values for mapping assemblies than for *de novo* assemblies across all other conditions [*mapping*/*de novo* ρ_{Spearman}: mean 0.9528/0.9466, range 0.9133–0.9700/0.8825–0.9925; recovery: mean 94.04% (75.96% correct)/105% (87.12% correct)]. With increasing divergence, the number of reads was overestimated for a number of genes and underestimated for others due to incorrect mapping (Fig. S8, Supporting information). Incorrect mapping was less of a problem in *de novo* assemblies, where instead the number of reads was systematically underestimated (Fig. S8, Supporting information). Among *de novo* assemblies, SOAP on average outperformed TRINITY [ρ_{Spearman}: SOAP/TRINITY mean 0.96/0.93, range 0.91–1.00/0.84–0.99; recovery mean 103% (87.36% correct)/107% (86.88% correct)]. Contigs available in *de novo* assemblies were separated into those with and without chimera. Excluding the chimeric contigs did not produce any drastic differences suggesting

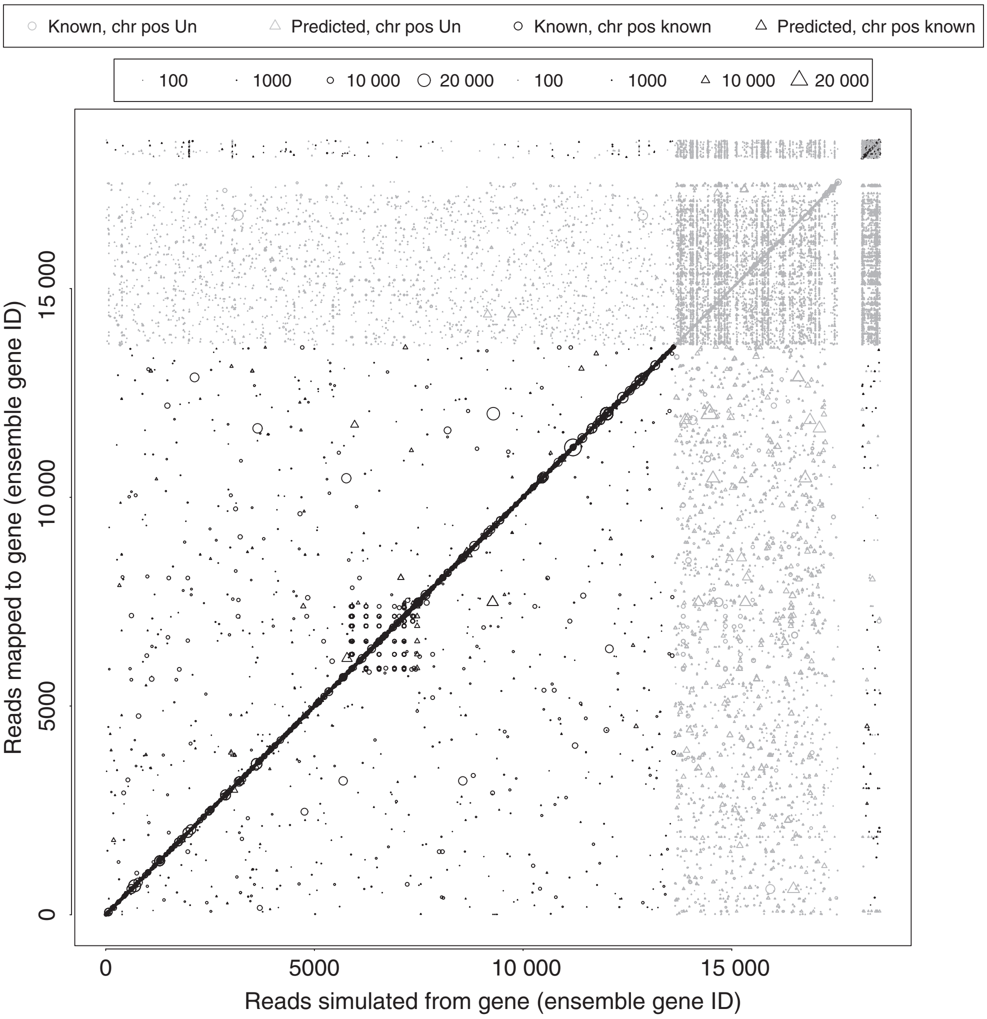

Fig. 4 Example of mapping profile for reads mapped onto TRINITY *de novo* assembled contigs shown for the 17K exponential data set (ε = 0.01, π = 0.01). Reads that get correctly mapped to the same gene from which they were simulated come to lie on the bisecting line. Reads that were simulated from gene X (abscissa) and are erroneously mapped to a different gene Y (ordinate) are distributed around the bisecting line and indicate mismapping problems generated by the bioinformatic pipeline (see Figs 1 and 2). Combinations of colour and symbol indicate chromosomal location of the gene in the reference genome (black = 'known', grey = 'random, unlocated') and its annotation status (circle = 'known', triangle = 'novel, putative known by projection'). Gene expression levels (reads/gene) are indicated by symbol size.

that they do not strongly bias gene expression estimates (ρ_{Spearman}: mean 0.95/0.95, range 0.89–1.00/0.88–0.99).

Transcriptome complexity. The simpler transcriptome data set had a better correlation between the number of input and output reads, but the effect was moderate (4K/17K ρ_{Spearman}: mean 0.97/0.93, range 0.95–0.98/ 0.92–0.95).

Polymorphism and error levels. An increase in polymorphism and error rates resulted in reduced overall proportions and correlations between the number of input and output reads [e.g. ($\varepsilon = 0$, $\pi = 0$)/($\varepsilon = 0.01$, $\pi = 0.01$): proportion mean 0.95/94, range 0.93–0.98/0.92–0.97].

Influence of the annotation quality of the (distant) reference genome

The annotation quality of a gene significantly determined the degree of mismapping (Fig. 4, Figs S9 and S10, Appendix S1, Supporting information). While the proportion of correctly assigned reads was not affected by annotation status ('known' vs. 'novel', 'putative' or 'known by projection'; $F_{1,233} = 0.9597$ $P > 0.05$), mismapping proportions were significantly increased for genes placed on 'random' chromosomes or with unknown chromosomal location 'Un' (ANOVA: $F_{1,233} = 64.9$, $P < 0.001$). Statistical interaction between assembly type and chromosome status (ANOVA: $F_{4,225} = 7.2$, $P < 0.001$) indicated that mapping assemblies up to 5% divergence were significantly less sensitive to mismapping errors than both *de novo* assemblies and mapping assemblies of higher divergence (posthoc Bernoulli test: $P < 0.001$, Fig. S9, Supporting information). Genes with unknown chromosome status also clearly performed worse with respect to inference of DE (Fig. S11, Supporting information, Table S7).

Differential gene expression

We compared the performance of inferring differential gene expression (DE hereafter) in relation to our parameters of interest (cf. Figs 1 and 2) as well as two differential expression software types, EDGER and BAYSEQ. Performance assessment is primarily presented using false and true positive rates (FPR and TPR) at different levels of log fold change (LFC). Two other metrics of performance, ROC curves and the correlation in LFC between simulation output and pipeline output (hereafter LFC correlation), generally gave very similar results. They are presented in the Supplementary Materials and only mentioned here when contributing contrasting or additional information. Finally, we also present the effects of absolute levels of expression on the relative performance of all conditions.

Assembly type. Mapping assemblies based on 5% divergence outperformed all types of *de novo* assemblies (Fig. 5a and Fig. S12a, Table S8, Supporting information). Mapping assemblies generally had higher (false and true) positive rates than *de novo* assemblies. FPRs were higher especially at 15% and 30% divergence levels, and at very low levels of expression. As expected, the quality of mapping assemblies decreased with increasing divergence level. At 15% divergence, mapping assembly performance was very similar to, but overall still slightly better than that of *de novo* assemblies (Table S8, Supporting information). At 30% divergence, mapping assemblies were outperformed by *de novo* assemblies: while TPRs are similar, FPRs are higher and ROC curves and LFC correlations lower in the mapping assemblies (Table S8, Supporting information).

Overall, SOAP and TRINITY performed similarly. However, relative performance differed between expression levels: at very low expression levels, SOAP performed much worse than TRINITY, while at high expression levels, SOAP outperformed TRINITY (Fig. 5a; plotted are assemblies with chimeric contigs, the same pattern is seen without chimeric contigs, in Fig. S13c, Supporting information.) This is mainly due to low recovery by SOAP at low expression levels: ROC plots only assess performance for isoforms that are actually detected, and these are similar or even better for SOAP (Figs S12a and S14c, Supporting information). *De novo* assemblies with chimeric contigs performed only slightly worse than assemblies without chimeric contigs, with higher FPRs and TPRs (Fig. S13a,b and Table S8, Supporting information), and lower ROC curves and LFC correlations (Fig. S14a,b and Table S8, Supporting information).

Patterns were similar for the data sets incorporating the effect of alternatively splicing. Overall, 5% mapping assemblies again performed best, 15% mapping assemblies and *de novo* assemblies performed similarly and 30% mapping assemblies performed worst (Figs S15c,d and S16c,d, Supporting information). With mapping being carried out at the isoform level, however, de novo assemblies performed worst of all assemblies, at least in terms of absolute positive rates (Fig. S15c, Supporting information). This poor performance is largely due to many isoforms not being present in the *de novo* assemblies to begin with: among isoforms that were actually retrieved, *de novo* assemblies in fact performed about as well as 5% mapping assemblies (see ROC curves in Fig. S16d, Supporting information, and LFC-correlations in Table S8, Supporting information).

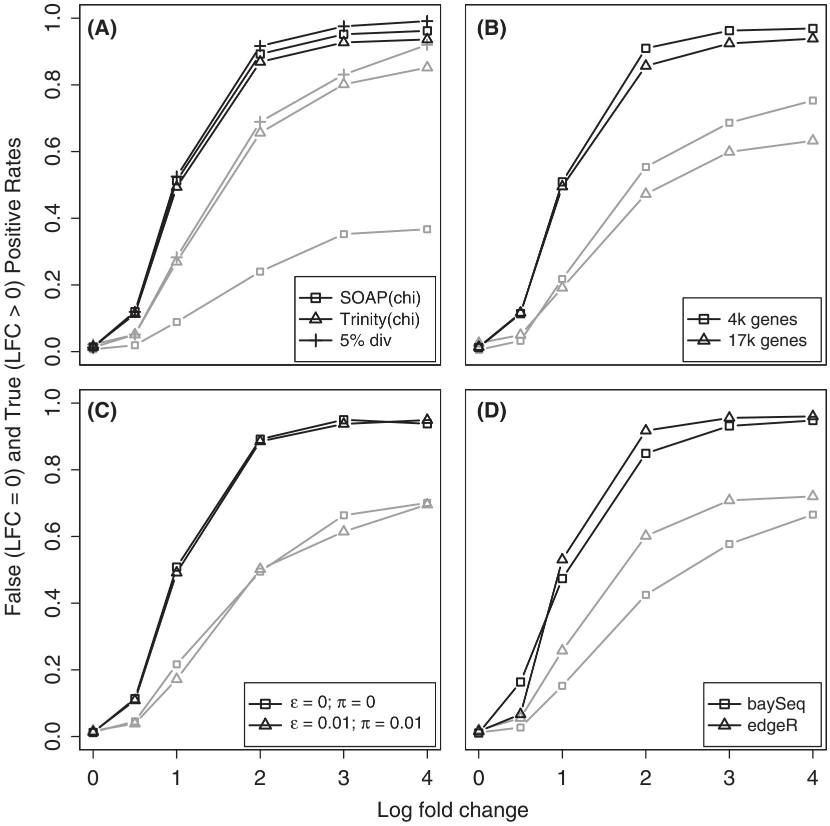

Fig. 5 Statistical inference of differentially expressed genes at different levels of simulated log-fold-change (LFC), averaged over the most realistic conditions (see main text). At a LFC of 0, all inferred positives are false positives; at LFC levels of 0.5–4, the true positive rate is depicted. Black lines represent positive rates for genes with an average per-library expression level of 100 reads/gene or more, whereas grey lines represent average positive rates for very lowly expressed genes (<25 reads/gene). (A) Comparison of assemblies: '*SOAP*(chi)' and '*Trinity*(chi)' refer to SOAPDENOVO-TRANS and *TRINITY de novo* assemblies with chimeras, respectively; '5% div.' refers to mapping assemblies with 5% divergent reference genome; (B) Comparison across different levels of transcriptome complexity (4400 [4K] vs. 17 400 [17K] genes); (C) Comparison of best and worst case error and polymorphism parameter values; (D) Comparison of differential expression software.

Transcriptome complexity. Performance of differential expression inference was slightly worse in the more complex data sets (17K vs. 4K). Overall, in the 17K data sets, the FPR is consistently higher, while TPRs at all levels of LFC are consistently lower, as is the LFC correlation (Table S8, Supporting information). These differences are more pronounced at higher levels of LFC and for very lowly expressed genes (Fig. 5b and Fig. S12b, Supporting information).

Polymorphism and error levels. Levels of polymorphism and sequencing error had very little effect on performance in inferring DE, although, as expected, there was a tendency of decreased performance for higher levels (Fig. 5c and Fig. S12c, Table S8, Supporting information). When analysing rates of polymorphism and error separately, small differences between the highest and lowest level of polymorphism could be detected, while our magnitude of variation in error rate did not appear to have any effect on performance (Table S8, Supporting information). In the isoform data sets as well, error and polymorphism levels had very little effect on performance (Figs S15a and S16a, Table S11, Supporting information).

DE software. Overall, EDGER outperformed BAYSEQ for all metrics (Fig. 5d and Fig. S12d, Table S8, Supporting information). Looking at positive rates, the differences were most pronounced for the lowest levels of expression (Fig. 5d). At higher levels of expression, BAYSEQ in fact had a lower FPR, and also a higher TPR at a LFC level of 0.5 (Fig. 5d, Table S8, Supporting information), yet EDGER had a higher TPR rate from LFC of 1 and higher.

Isoforms. Overall, the TPR of differential expression were only slightly lower when mapping was performed at the isoform level as compared to the level of the gene (Fig. S15c,d, Table S11, Supporting information). This was true both for lowly and highly expressed transcripts, but note that the difference was much larger for *de novo* assemblies (see above). The lower performance at the isoform-level is mostly due to a proportion of isoforms not being present in the assembly to begin with (as ROC plots are similar for gene- and isoform-based analyses, see Fig. S16c,d, Supporting information).

Discussion

Using extensive simulated RNA-seq data, we here explore how transcriptomic features (complexity, polymorphism level, alternative splicing), common technological challenges (sequencing error, library normalization) and elements of the bioinformatic workflow (mapping and *de novo* assembly, gene annotation, inference of isoforms) affect transcriptome assembly quality and inference of (differential) gene expression. Despite some oversimplification inherent in all simulation approaches, we have attempted to mimic a realistic range of RNA-seq experiments. We are thus confident that this approach allows insight into how transcriptome assembly quality and gene expression profiling are impacted by various factors and can provide guidance to practitioners. In the following, we will highlight the major findings on the basis of which we will assess the most promising strategies and discuss areas that may need more attention in the future.

Computing resources

Computing resources are still a limiting factor in smaller laboratories when it comes to the analyses of RNA-seq data. Our simulations generated ~14 TB (Terabyte) of data and consumed a total of 300 000 h of computing resources on 8–1000 cores (CPUs) using on average 128 GB (Gigabytes) of memory (up to 2 TB). To facilitate *de novo* transcriptome assembly in a reasonable time frame, a computer cluster should contain at least 8 cores and 256 GB of RAM and a fast storage system. The mapping approach is computationally less demanding and an 8-core cluster with 32 GB of RAM should generally be sufficient. Downstream analyses like DE can be performed on a desktop computer.

Transcriptome assembly

Sequencing strategy and general success. Overall, assembly success was reassuringly high and robust across most conditions. Translated to real situations, this suggests that most of the expressed genes can indeed be recovered in an experiment when an adequate number of reads is used (Wang *et al.* 2011). As a rough point of reference, we suggest sequence coverage of 500–800× for most transcriptomes which currently corresponds to one lane of sequencing on an Illumina HiSeq2000 [>100 million reads per lane, (Goldfeder *et al.* 2011)]. Sufficient read coverage also removes the necessity of costly library normalization which has been suggested to increase yields for low coverage data sets, mostly produced with the 454 technology [(Ekblom *et al.* 2012), but see (Künstner *et al.* 2010)]. Our results clearly advise against normalization.

In theory, a single highly inbred individual should ideally be used for transcriptome assembly. However, the detrimental effects of polymorphic sites were comparatively small in our simulations and were only seen for extreme combinations of polymorphism and sequencing error ($\pi = 0.01$; $\varepsilon = 0.01$). This opens the promising prospect that basically any wild caught individual can be used to produce a reference transcriptome, with good success across most realistic polymorphism levels.

Mapping vs. de novo assembly. A major and rather unexpected finding was that mapping assemblies outperformed *de novo* assembly approaches across a broad range of conditions. In general, mapping assemblies recovered a larger proportion of the transcriptome, although in more difficult cases with complex transcriptomes, high error and polymorphism rates suffered from lower accuracies. The drop in accuracy was, however, moderate and outweighed by the fact that mapping assemblies were more robust to mismappings caused by badly annotated genes (chromosome status 'unlocated' or 'random'). Moreover, direct mapping on distant references has the advantage that each assembled contig has a 1:1 orthologous gene name genuinely assigned, which is vital for downstream biological inference (e.g. GO term analysis) and alleviates the problem of gene name assignment (see below). Another advantage of mapping assemblies is the reduced need for computing power and time. The overall good performance of mapping assemblies of up to 15% sequence divergence opens the exciting possibility to use reference transcriptomes as distant as human-rhesus macaque or mouse-rat, which corresponds to tens of million years of independent evolution (Miller *et al.* 2007). With the ever increasing availability of genomes across a diverse array of taxa (Genome 10K Community of Scientists 2009), mapping assemblies will soon be a realistic option for many organisms.

de novo assembly software. Among all three *de novo* assembly softwares compared (TRANS-ABYSS, TRINITY, and SOAPDENOVO-TRANS), the SOAPDENOVO-TRANS assembler performed best across the entire range of conditions. TRINITY assemblies appeared to be slightly more contiguous, which, however, was largely owing to its tendency of producing chimeric contigs spanning more than one gene. One noticeable shortcoming of SOAPDENOVO-TRANS was that under high values of sequencing error and polymorphism levels, assemblies were more fragmented. Also SOAP had very low TPR at low expression levels (Fig. S13c, Supporting information).

In contrast to SOAPDENOVO-TRANS, TRINITY in principle provides additional information about isoform/paralog/allele structure of the transcriptome (Grabherr *et al.* 2011). Different isoforms of one gene are problematic for gene annotations as well as gene expression analysis. Information on isoforms is thus highly valuable, as for most genes, the true number of isoforms is not known, even in model organisms (Wang *et al.* 2008). To address this problem, we simulated data where each gene had only one isoform and still many isoforms were erroneously inferred. This result admonishes to caution in the interpretation of real data.

Gene name assignment

Gene name assignment is crucial for drawing biologically meaningful conclusions from RNA-seq experiments and for comparing results among different studies. As already mentioned, gene name assignment comes for free in mapping approaches. In the case of *de novo* assemblies, contigs provide no information about the sequenced gene and need to be assigned to orthologous genes from (distantly) related genomes. Our results suggest that faster suffix-tree based methods such as *NUCmer* and *PROmer* work well for closely related species, but are not sensitive enough to detect orthologs as divergence increases (see Table S6, Supporting information). For more distant references, BLAST-based orthology detection seems to be a popular alternative and has been widely used in genetic nonmodel organisms, e.g. to annotate genes in fish (Elmer *et al.* 2010) or passerine birds (Künstner *et al.* 2010; Wolf *et al.* 2010). In the case of our simulation, *BLAST2GO* had higher assignment success than *NUCmer* or *PROmer*, but also produced a relatively high number of false assignments. Stringent filtering on blast scores, alignment length and reciprocal-best-hits are thus crucial to guard against false detection of orthologous genes (Chen *et al.* 2007). Other faster yet sensitive alignment programmes like SATSUMA and SPINES (Grabherr *et al.* 2010), which in our case yielded comparable results, can provide viable alternatives (Kristensen *et al.* 2011).

Differential gene expression

As RNA-seq has arisen as a powerful and accurate alternative to microarrays, it will increasingly be used in ecological and evolutionary studies to infer differential expression among phenotypes or experimental conditions (Lenz *et al.* 2012; Wolf *et al.* 2010). Our results demonstrate that across a fairly wide range of conditions, gene expression estimates were robust to most of the parameters under investigation and power to detect differentially expressed (DE) genes was generally high while the amount of false positives was limited. Even inference of isoform-specific (differential) expression appears feasible, though, as expected, the presence of isoforms had a negative effect on overall performance.

Similar to what we observed for transcriptome quality, mapping assemblies even to rather distant references genomes (15% sequence divergence) provided more accurate gene expression levels and outperformed DE inference of *de novo* assemblies. In contrast to the results of other simulation studies (Hardcastle & Kelly 2010; Kvam *et al.* 2012), we found that EDGER robustly outperformed BAYSEQ which may be due to an optimal choice of the dispersion prior with a recently implemented EDGER function.

By far the most important factors influencing performance were expression levels and fold-change levels of differential expression. In our simulations, LFC levels of 2 or higher were necessary to detect the large majority of differentially expressed genes for average and highly expressed genes. As expected, DE inference for lowly expressed genes (<25 reads per library) performed considerably worse, particularly so for SOAPDENOVO-TRANS. This strongly suggests that obtaining high overall coverage is vital for successful RNA-seq experiments.

Acknowledgements

The simulations were performed on resources provided by the Swedish National Infrastructure for Computing (SNIC) at the Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX) under the projects b2010060 and b2010059. We would like to thank Imtiyaz Syed who contributed to the initial exploration of the project and Manfred G. Grabherr who provided advice with *Trinity* and with whom we had fruitful discussions on gene name assignment.

References

Axtner J, Sommer S (2009) Validation of internal reference genes for quantitative real-time PCR in a non-model organism, the yellow-necked mouse, *Apodemus flavicollis. BMC Research Notes*, **2**, 264.

Bar-Or C, Bar-Eyal M, Gal TZ, Kapulnik Y, Czosnek H, Koltai H (2006) Derivation of species-specific hybridization-like knowledge out of cross-species hybridization results. *BMC Genomics*, **7**, 110.

Cartwright RA (2005) DNA assembly with gaps (Dawg): simulating sequence evolution. *Bioinformatics*, **21**, iii31–iii38.

Chen F, Mackey AJ, Vermunt JK, Roos DS (2007) Assessing performance of orthology detection strategies applied to eukaryotic genomes. *PLoS ONE*, **2**, e383.

Church DM, Schneider VA, Graves T *et al.* (2011) Modernizing reference genome assemblies. *PLoS Biology*, **9**, e1001091.

Deng X, Hiatt JB, Nguyen DK *et al.* (2011) Evidence for compensatory upregulation of expressed X-linked genes in mammals, *Caenorhabditis elegans* and *Drosophila melanogaster*. *Nature Genetics*, **43**, 1179–1185.

Ekblom R, Galindo J (2011) Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity*, **107**, 1–15.

Ekblom R, Slate J, Horsburgh GJ, Birkhead T, Burke T (2012) Comparison between normalised and unnormalised 454-sequencing libraries for small-scale RNA-Seq studies. *Comparative and Functional Genomics*, **2012**, 281693.

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

Fu X, Fu N, Guo S *et al.* (2009) Estimating accuracy of RNA-Seq and microarrays with proteomics. *BMC Genomics*, **10**, 161.

Genome 10K Community of Scientists. (2009) Genome 10K: a proposal to obtain whole-genome sequence for 10 000 vertebrate species. *Journal of Heredity*, **100**, 659–674.

Goetz F, Rosauer D, Sitar S *et al.* (2010) A genetic basis for the phenotypic differentiation between siscowet and lean lake trout (*Salvelinus namaycush*). *Molecular Ecology*, **19**, 176–196.

Goldfeder RL, Parker SCJ, Ajay SS, Ozel Abaan H, Margulies EH (2011) A bioinformatics approach for determining sample identity from different lanes of high-throughput sequencing data. *PLoS ONE*, **6**, e23683.

Grabherr MG, Russell P, Meyer M *et al.* (2010) Genome-wide synteny through highly sensitive sequence alignment: Satsuma. *Bioinformatics*, **26**, 1145–1151.

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.

Hardcastle TJ, Kelly KA (2010) baySeq: Empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics*, **11**, 422.

Harr B, Turner LM (2010) Genome-wide analysis of alternative splicing evolution among Mus subspecies. *Molecular Ecology*, **19**, 228–239.

Kristensen DM, Wolf YI, Mushegian AR, Koonin EV (2011) Computational methods for gene orthology inference. *Briefings in Bioinformatics*, **12**, 379–391.

Künstner A, Wolf JBW, Backstrom 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.

Kurtz S, Phillippy A, Delcher AL *et al.* (2004) Versatile and open software for comparing large genomes. *Genome Biology*, **5**, R12.

Kvam VM, Liu P, Si Y (2012) A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data. *American Journal of Botany*, **99**, 248–256.

Kvist J, Wheat CW, Kallioniemi E, Saastamoinen M, Hanski I, Frilander MJ (2012) Temperature treatments during larval development reveal extensive heritable and plastic variation in gene expression and life history traits. *Molecular Ecology*, doi: 10.1111/j.1365-294X.2012.05521.x.

Lenz TL, Eizaguirre C, Rotter B, Kalbe M, Milinski M (2012) Exploring local immunological adaptation of two stickleback ecotypes by experimental infection and transcriptome-wide digital gene expression analysis. *Molecular Ecology*, doi: 10.1111/j.1365-294X.2012.05756.x.

Li B, Dewey CN (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, **12**, 323.

Li H, Handsaker B, Wysoker A *et al.* (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Lunter G, Goodson M (2011) Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Research*, **21**, 936–939.

Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research*, **18**, 1509–1517.

Miller W, Rosenbloom K, Hardison RC *et al.* (2007) 28-Way vertebrate alignment and conservation track in the UCSC Genome Browser. *Genome Research*, **17**, 1797–1808.

Naurin S, Hansson B, Hasselquist D, Kim Y-H, Bensch S (2011) The sex-biased brain: sexual dimorphism in gene expression in two species of songbirds. *BMC Genomics*, **12**, 37.

Ozsolak F, Milos PM (2010) RNA sequencing: advances, challenges and opportunities. *Nature Reviews Genetics*, **12**, 87–98.

Palmieri N, Schlötterer C (2009) Mapping accuracy of short reads from massively parallel sequencing and the implications for quantitative expression profiling. *PLoS ONE*, **4**, e6323.

Pareek CS, Smoczynski R, Tretyn A (2011) Sequencing technologies and genome sequencing. *Journal of applied genetics*, **52**, 413–435.

Perry GH, Melsted P, Marioni JC *et al.* (2012) Comparative RNA sequencing reveals substantial genetic variation in endangered primates. *Genome Research*, **22**, 602–610.

Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.

Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.

Salzberg SL, Phillippy AM, Zimin A *et al.* (2012) GAGE: a critical evaluation of genome assemblies and assembly algorithms. *Genome Research*, **22**, 557–567.

Sing T, Sander O, Beerenwinkel N, Lengauer T (2005) ROCR: visualizing classifier performance in R. *Bioinformatics*, **21**, 3940–3941.

Trapnell C, Williams BA, Pertea G *et al.* (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, **28**, 511–515.

Wang ET, Sandberg R, Luo S *et al.* (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature*, **456**, 470–476.

Wang Y, Ghaffari N, Johnson CD *et al.* (2011) Evaluation of the coverage and depth of transcriptome by RNA-Seq in chickens. *BMC Bioinformatics*, **12**, S5.

Warren WC, Clayton DF, Ellegren H *et al.* (2010) The genome of a songbird. *Nature*, **464**, 757–762.

Wolf JBW, Bryk J (2011) General lack of global dosage compensation in ZZ/ZW systems? Broadening the perspective with RNA-seq. *BMC Genomics*, **12**, 91.

Wolf JBW, Bayer T, Haubold B *et al.* (2010) Nucleotide divergence versus 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.

Xu L, Chen H, Hu X *et al.* (2006) Average gene length Is highly conserved in prokaryotes and eukaryotes and diverges only between the two kingdoms. *Molecular Biology and Evolution*, **23**, 1107–1108.

N.V. uses bioinformatic approaches to understand questions in evolutionary genetics. J.W.P. is an evolutionary biologist with a strong interest in the genetics of speciation including the role of gene expression. A.K. has a strong bioinformatic background which he applies in a comparative genomic framework. J.B.W.W. works at the interface of molecular ecology and evolutionary genetics. He uses genomic approaches to questions in speciation genetics and molecular evolution.

Data accessibility

All scripts and sequence data used in this study are accessible via Dryad doi:10.5061/dryad.3t3n7.

Supporting information

Additional Supporting Information may be found in the online version of this article.

Table S1 Accuracy and recoveries of different assemblies.

Table S2 GAGE structural accuracy measurements.

Table S3 Contiguity metrices for de novo assemblies.

Table S4 Performance of *SOAPdenovoTrans* at different kmer values.

Table S5 Input vs. output correlation coeffecients.

Table S6 Nucmer based annotation of 4K genes at 0%, 5%, 15% and 30% divergence using 17K reference transcriptome.

Table S7 Positive Rate and LFC correlation statistics: comparing gene statuses.

Table S8 Differential expression inference: pairwise comparisons of FPRs, TPRs, and LFC correlations, and accompanying Mann Whitney tests (*P* and *V* values).

Table S9 Positive Rate and LFC correlation statistics for each dataset.

Table S10 Positive Rate and LFC correlation statistics for isoform datasets.

Fig. S1 Comparison between fit of gamma-distribution and exponential-distribution for empirical data.

Fig. S2 Comparison between fit of gamma-distribution and exponential-distribution for Perry *et al.* data.

Fig. S3 Recoveries of SOAPdenovo-trans at different K-mer values for the 17K exponential dataset with 0.01 sequencing error ∊ and 0.01 polymorphism π levels.

Fig. S4 Assembly nomenclature and recovery calculation.

Fig. S5 Flowchart of gene name assignment steps.

Fig. S6 Recovery of the reference transcriptome for different (a) 4400, (b) 17400 genes datasets displayed in relation to assembly strategy, sequencing error ∊ and polymorphism π levels.

Fig. S7 Number of variant sequences per component (i.e. isoforms) predicted by Trin-ity assembler shown for the complex transcriptome (17K) above, and the less complex transcriptome with on unique genes (4K) below.

Fig. S8 Mapping correlations for reads mapped onto 5%, 15%, 30% divergent reference genomes as well as SOAP denovo assembled contigs shown for the 17K exponential dataset (π = 0.01, ∊ = 0.01).

Fig. S9 Analysis of effects of annotation quality and assembly method on the proportion of mismapped reads/number of simulated reads.

Fig. S10 Example of mapping profile for reads mapped onto (a) 5% (b) 15% divergent transcriptome shown for the 17K exponential dataset (∊ = 0.01, π = 0.01).

Fig. S11 Positive rates for the inference of differential expression at several levels of simulated LFC, for genes with known gene status vs. predicted genes.

Fig. S12 Receiver-Operator curves for the inference of differential expression.

Fig. S13 Positive rates for the inference of differential expression at several levels of simulated LFC.

Fig. S14 Receiver-Operator curves for the inference of differential expression.

Fig. S15 Positive rates for the inference of differential expression at several levels of simulated LFC, for the isoform datasets based on Perry *et al.*

Fig. S16 Receiver-Operator curves for the inference of differential expression, for the isoform datasets based on Perry *et al.*, as in Fig. S14.

Appendix S1 Details on gene name assignment.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.