

METHOD

Open Access



SuperTranscripts: a data driven reference for analysis and visualisation of transcriptomes

Nadia M. Davidson^{1,2*}, Anthony D. K. Hawkins¹ and Alicia Oshlack^{1,2*} 

Abstract

Numerous methods have been developed to analyse RNA sequencing (RNA-seq) data, but most rely on the availability of a reference genome, making them unsuitable for non-model organisms. Here we present superTranscripts, a substitute for a reference genome, where each gene with multiple transcripts is represented by a single sequence. The Lace software is provided to construct superTranscripts from any set of transcripts, including de novo assemblies. We demonstrate how superTranscripts enable visualisation, variant detection and differential isoform detection in non-model organisms. We further use Lace to combine reference and assembled transcriptomes for chicken and recover hundreds of gaps in the reference genome.

Background

High throughput sequencing has revolutionised transcriptomics because it allows complementary DNA (cDNA) sequences to be read and expression levels quantified using a single, affordable assay [1, 2]. RNA sequencing (RNA-seq) can examine expression at the gene level as well as infer transcript abundances and differential isoform usage. Alternative splicing can alter gene function and contributes to the overall transcriptional diversity in eukaryotes [3, 4]. In addition, RNA-seq has the power to detect variation in expressed sequence, such as single nucleotide variants [5], post-transcriptional editing [6] and fusion genes [7]. Our knowledge of the transcriptome of model organisms has matured through projects such as ENCODE [4] and we now have robust and well established methods for RNA-seq data analysis [8]. Most of these methods use the accurate reference genomes and annotations now available for model organisms.

For non-model organisms, however, reference genomes are generally not available. Instead, an experiment specific transcriptome can be built from RNA-seq data through de novo transcriptome assembly [9], a process designed to reconstruct the full-length sequence of each expressed transcript. However downstream analysis of the data

remains challenging. With the exception of a few recent methods, such as RNA quantification with Kallisto [10], Salmon [11] and RSEM [12], most analytical approaches for RNA-seq are designed to work with a reference genome rather than transcriptome. Those methods compatible with a reference transcriptome often rely on accurate gene models, which are not necessarily produced through de novo assembly of short read data. In addition, visualisation of read coverage across a gene, which is common for the exploration, curation and communication of the analysis results, is impossible using a reference transcriptome. At best, reads may be mapped and visualised against a representative transcript from each gene, such as the longest isoform, but a significant proportion of the gene sequence can be missed (Additional file 1: Table S1, Figure S1).

Here we propose an alternative representation for each gene, which we refer to as a superTranscript. SuperTranscripts contain the sequence of all exons of a gene without redundancy (Fig. 1a). They can be constructed from any set of transcripts including de novo assemblies and we have developed a python program to build them called Lace (available from <https://github.com/Oshlack/Lace/wiki>). Lace works by building a splice graph [13] for each gene, then topologically sorting the graph using Kahn's algorithm [14] (Fig. 1b). Building superTranscripts is a simple post-assembly step that promises to unlock

* Correspondence: nadia.davidson@mcri.edu.au; alicia.oshlack@mcri.edu.au
¹Murdoch Children's Research Institute, Royal Children's Hospital, Melbourne, VIC, Australia

Full list of author information is available at the end of the article

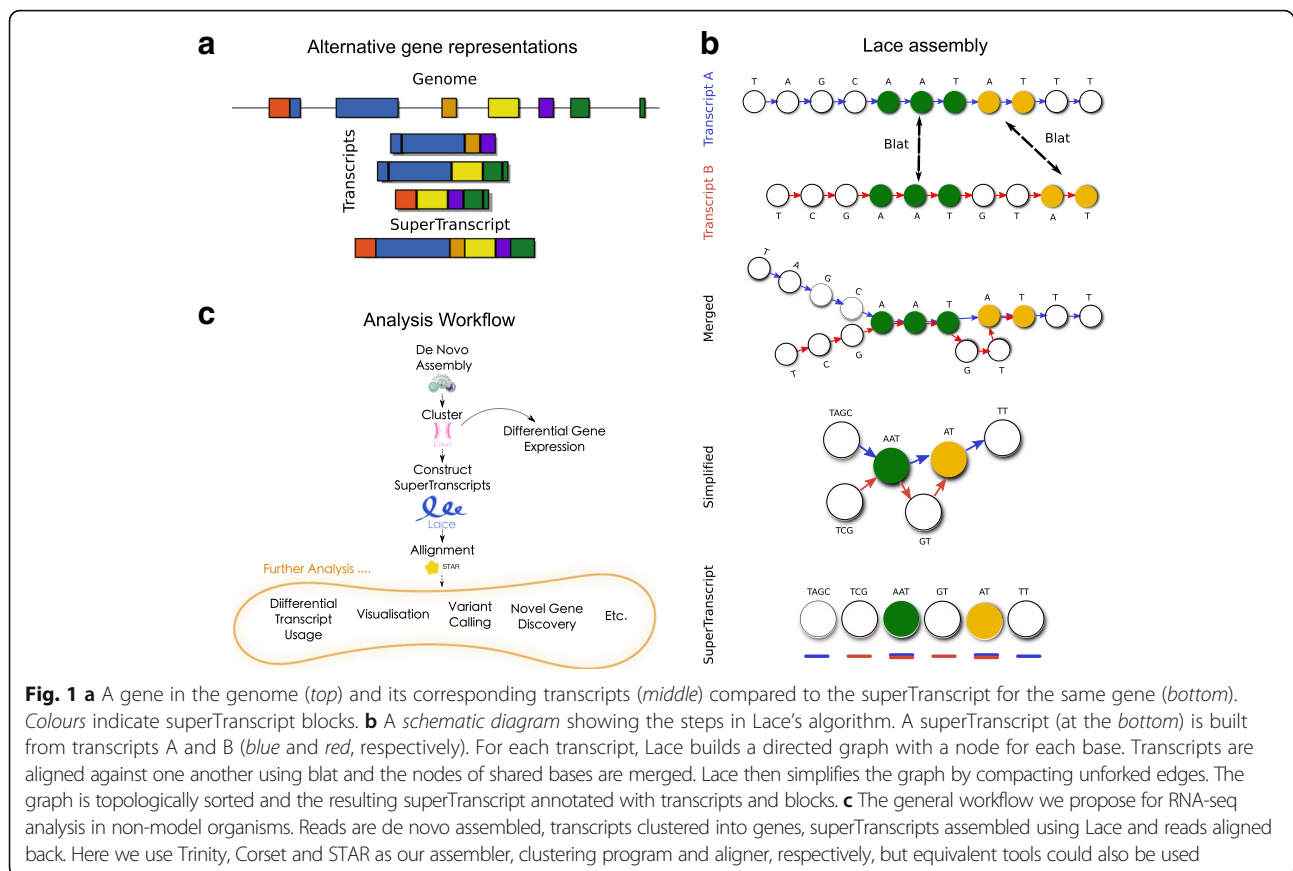


Fig. 1 **a** A gene in the genome (*top*) and its corresponding transcripts (*middle*) compared to the superTranscript for the same gene (*bottom*). Colours indicate superTranscript blocks. **b** A schematic diagram showing the steps in Lace's algorithm. A superTranscript (at the *bottom*) is built from transcripts A and B (blue and red, respectively). For each transcript, Lace builds a directed graph with a node for each base. Transcripts are aligned against one another using blat and the nodes of shared bases are merged. Lace then simplifies the graph by compacting unforked edges. The graph is topologically sorted and the resulting superTranscript annotated with transcripts and blocks. **c** The general workflow we propose for RNA-seq analysis in non-model organisms. Reads are de novo assembled, transcripts clustered into genes, superTranscripts assembled using Lace and reads aligned back. Here we use Trinity, Corset and STAR as our assembler, clustering program and aligner, respectively, but equivalent tools could also be used

numerous analytical approaches for non-model organisms (Fig. 1c).

Although superTranscripts do not necessarily represent any true biological molecule, they provide a practical replacement for a reference genome. For example, reads can be aligned to the superTranscriptome using a splice aware aligner and subsequently visualised using standard tools such as IGV [15]. Quantification can also be performed with existing software by counting the reads that overlap superTranscript features. In non-model organisms, we further demonstrate using superTranscripts to call variants and we show that we can accurately detect differential isoform usage. We also demonstrate applications of superTranscripts to model organisms. Specifically, we combined a reference and de novo assembled transcriptome into a compact superTranscriptome using chicken RNA-seq data, to allow identification of novel transcribed sequence. We found conserved coding sequence in over 500 genes that was missed in the current chicken reference genome, galGal5.

Results and discussion

Lace constructs superTranscripts

SuperTranscripts can be built from any set of transcripts, including de novo assembled transcripts, using an overlap

assembly method. We have implemented this algorithm in an open source Python program called Lace (available from <https://github.com/Oshlack/Lace/wiki>). The Lace algorithm takes two input files: (1) a set of transcript sequences in fasta format; and (2) a text file with the clustering information that groups each transcript into a gene or cluster. Lace outputs a fasta file of superTranscript sequences and a gff file with their annotation. The Lace assembly is conceptually described in Fig. 1b and includes the following steps:

1. For each gene, all pairwise alignments between transcripts are performed using BLAT [16].
2. The BLAT output is parsed to determine the sequences common to each transcript pair.
3. A single directed graph is constructed per gene, where each node is a base in one of the transcripts and the directed edge retains the ordering of the bases in each transcript. Bases from overlapping sequence are merged based on the BLAT output.
4. The graph is simplified by concatenating sequences of nodes along non-diverging paths. Any cycles in the graph are detected, for each cycle the node of the cycle which contains the fewest bases is selected and duplicated. The outgoing edges from the

selected node are re-routed to the duplicate node and the cycle broken (Additional file 1: Figure S2). This method was inspired by that used in Pevzner et al. [17]. This creates a directed acyclic graph.

5. The nodes are topologically sorted (each node becomes a string of bases from the original graph) using Khan's algorithm [14], which gives a non-unique sorting of the nodes.
6. Finally, Lace will annotate each superTranscript with blocks either by using the graph structure itself or alternatively using a script called Mobius (packaged with Lace) which infers the annotation from spliced junctions discovered when mapping reads to the superTranscript.

One of the advantages of Lace is that it can produce superTranscripts from any combination of transcripts and is compatible with any transcriptome assembler. However, Lace relies on information for clustering transcripts into genes. This can be achieved with our previously published method, Corset [18] (Fig. 1c).

Lace's running time is primarily limited by the speed of the BLAT alignments, however, for genes with a large number of transcripts, processing the splicing graph is significantly slower. For this reason, Lace uses only the first 50 transcripts from each gene by default. In practice, this only affects a small number of genes for most assemblies. Typically, constructing superTranscripts for an entire de novo assembly on eight cores takes approximately 0–8 h on a linux cluster and uses up to 4 Gb of RAM, depending on the size of the input transcriptome.

Application of Lace and superTranscripts to non-model organisms

SuperTranscripts allow a broad range of RNA-seq analyses to be performed on non-model organisms using standard software that has been designed to work with reference genomes. To demonstrate, we analysed human RNA-seq data without the use of the reference genome or transcriptome. First, we assembled transcripts with Trinity [19] then clustered transcripts into genes using Corset [18] and subsequently built superTranscripts with Lace. Using these superTranscripts as a reference we then aligned reads back to the superTranscripts using STAR [20]. This approach allowed us to perform a variety of analysis and visualisation (Fig. 1c).

Detecting variants in non-model organisms

In model organisms, variant calling can be performed on RNA-seq data using the established GATK Best Practices workflow for RNA-seq. Here we demonstrate that using superTranscripts as a reference allows variant calling to be performed in non-model organisms using the same pipeline with similar performance. In addition, called

variants can be easily inspected in IGV for the first time (Additional file 1: Figure S3).

In order to demonstrate variant calling from RNA-seq data using the assembled superTranscripts as the reference, we utilised RNA-seq from the Genome in a Bottle cell line (GM12878). We called variants using the GATK RNA-seq variant calling pipeline and compared them to known variants reported by the Genome in a Bottle Consortium [21]. Specifically, we took high quality single-nucleotide polymorphisms (SNPs) with a read coverage of ten or more that were detected in our SuperTranscript analysis (26,367 SNPs, Additional file 1: Table S2). Only heterozygous SNPs, which we defined as those with at least one read supporting the reference allele, were analysed. Reported homozygous SNPs were removed because they are likely to be false positives of the assembly or alignment. True homozygous SNPs should be assembled into the reference and are therefore not detectable. Note that this is a general limitation of using the same sample to create the reference and call variants and is not unique to the superTranscript method. However, homozygous variants could be detected for non-model organisms if multiple samples were available or if superTranscripts were constructed and called with respect to a control.

We aligned each superTranscript back to the human genome with BLAT [16] to determine the SNP position in the genome. We then examined SNPs in the high confidence call region for Genome in a Bottle (17,035 SNPs). Eighty percent (13,639 SNPs) were true positives reported by the Genome in a Bottle Consortium. The precision lifted to 90% if repeat regions of the genome were excluded (Additional file 1: Table S2).

Next, we assessed how well superTranscripts compared against using the genome as a reference for calling variants. On the same dataset with the same filtering, we found that the genome-based approach gave similar results to using superTranscripts derived from de novo assembly. The genome-based approach was more sensitive with more true positives reported (15,925 compared to 13,639), but the precision was similar (81% compared to 80%). Ninety-one percent of true positives and 37% of false positives that were reported by the superTranscript approach were also reported for the genome-based approach. These results suggest that the accuracy of detecting heterozygous SNPs in non-model organisms using superTranscripts is similar to the detection accuracy of heterozygous SNPs from RNA-seq in model organisms.

Finally, we validated our approach against KisSplice [22, 23], an alternative method for SNP and indel detection in non-model organisms. KisSplice performs local assembly of RNA-seq reads and detects variants directly from the De Bruijn graph. KisSplice results were also filtered for a read coverage of ten or more and by default the method only reports SNPs with at least one read

supporting each path (or heterozygous SNPs). KisSplice reported 28,447 SNPs, of which 20,922 were located in the high confidence call region for Genome in a Bottle and 11,269 were true positives (Additional file 1: Table S2). We found that KisSplice detected fewer true positives than superTranscripts (11,269 compared to 13,639) and had a lower precision (54% compared to 80%).

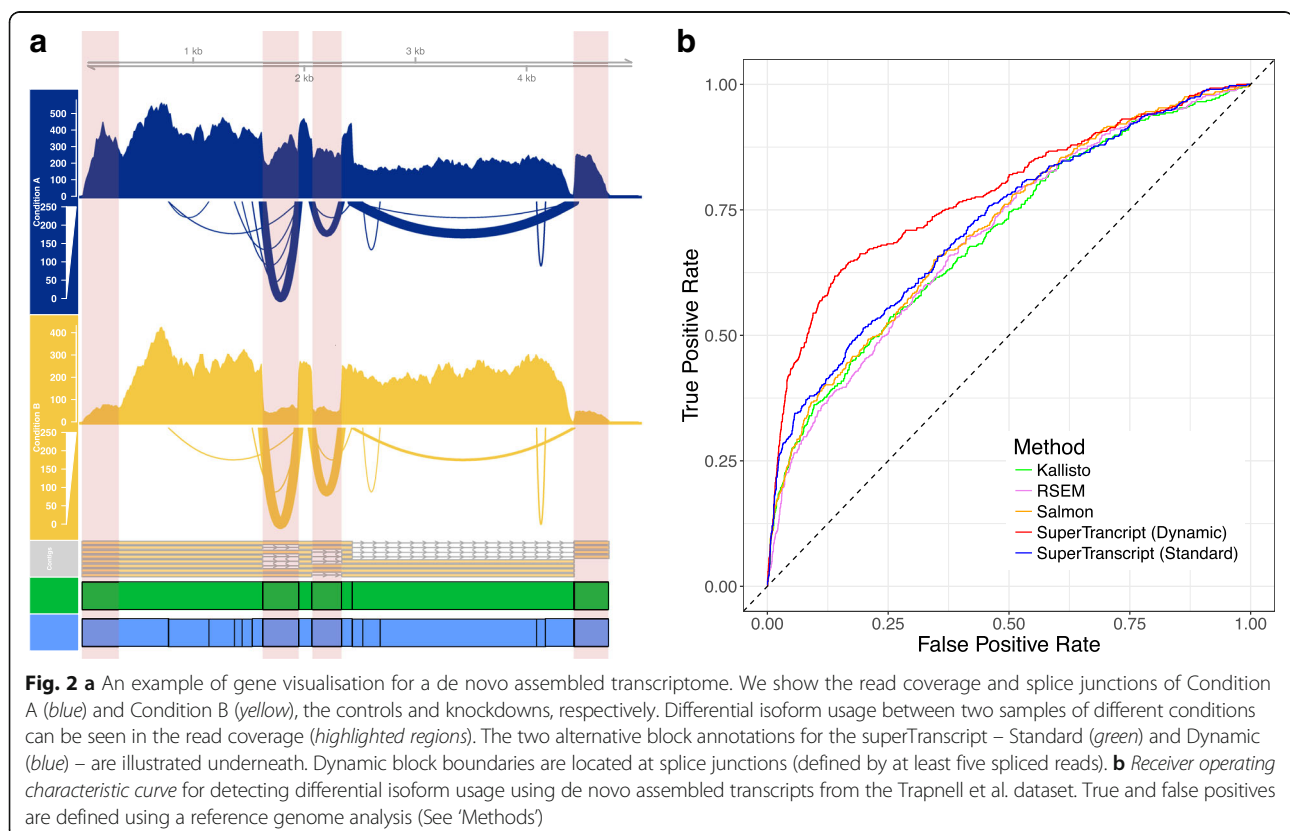
Differential isoform usage in non-model organisms

While methods exist for detecting differential gene expression in non-model organisms (such as Corset [18]), the procedure for detecting differential isoform usage is less well defined. When a reference genome is available, a common method for detecting differential isoform usage is to first align reads to the genome then count the number of reads that overlap exons for each sample (for example, using featureCounts [24]) and finally perform statistical testing of the count data looking for differential exon usage using methods such as DEXSeq [25]. An alternative approach is to use estimates of transcript abundances from inference methods such as Kallisto and Salmon and subsequently perform a similar statistical testing method for differential isoform usage [26].

SuperTranscripts can be used in a similar way to the reference genome approach where reads are aligned to the superTranscripts instead of a reference genome. Each superTranscript is segmented into blocks which are used

as the counting bins for statistical testing instead of exons. The blocks that annotate a superTranscript are defined as a contiguous sequence without splice junctions. Different isoforms are therefore represented by different combinations of blocks. Hence, a block may correspond to one exon, multiple exons or part of an exon, depending on the splicing structure of the alternative transcripts within a gene. Blocks are similar in concept to the exon counting bins used by approaches such as DEXSeq; however, adjacent exons which are always spliced together will form a single superTranscript block. Lace provides two different types of block annotation. In one case, block positions are defined by forks or divergences in the splice graph ('Standard Blocks'); in the second case, blocks are defined dynamically using splice junctions detected in the reads that are mapped back to the superTranscript ('Dynamic Blocks') (see Fig. 2a).

We tested the ability of the superTranscript method to call differential isoform usage on de novo assembled data. For this we used human RNA-seq data from Trapnell et al. [27]. It was also analysed using two reference-based approaches for comparison: a standard reference genome approach (reads mapped and counted with featureCounts) and a reference transcriptome approach (Kallisto) with each method then followed by statistical testing using DEXseq. We defined genes as true positives or true negatives for differential isoform usage using the intersection



of results of the genome and transcriptome reference-based approach (see ‘Methods’). We then compared the accuracy of five approaches using the de novo assembly of the same data: (1) superTranscripts with standard block counts; (2) superTranscripts with dynamic block counts; (3) transcript counts from Kallisto; (4) transcript counts from Salmon; and (5) transcript counts from RSEM (see ‘Methods’). We found that using superTranscript methods for defining and counting reads resulted in better performance when testing for differential isoform usage (Fig. 2b). Specifically, a false discovery rate < 0.05 applied to all methods gave a specificity of approximately 97%. The inference methods had a sensitivity of 20% (Kallisto and Salmon) and 17% (RSEM) compared to 22% (superTranscripts with standard block) and 32% (superTranscripts with dynamic blocks). The remarkable increase in sensitivity of dynamic blocks was predominantly due to the ability of dynamic blocks to detect splicing events in genes where only a single transcript was assembled.

SuperTranscripts also provide a means of visualising differential isoform usage in non-model organisms for the first time. In Fig. 2a, we show differential isoform usage in ENSG00000160613 between the *HOXA1* knock-down and control groups from the Trapnell dataset. The superTranscript provides a convenient means of presenting all eight assembled transcripts, their corresponding read coverage and splicing, within a single visual. Although this type of visualisation is often taken for granted in model organisms, it is only made possible in non-model organisms by using superTranscripts as a reference.

Combining reference and de novo assembled transcriptome

Despite the increasing number of species with a reference genome, the quality of many reference genomes and their annotations remain variable. Ideally, an RNA-seq analysis could utilise prior knowledge of the gene-models available in a reference genome and annotation, while also extracting information about the genes from the RNA-seq data itself. Lace has the ability to integrate such information. It can produce superTranscripts from any source, including a combination of reference and de novo assembled transcriptomes. We demonstrate this idea on chicken, a model organism which we know from our previous work on chicken gonads had missing and rearranged sequence in its reference genome [28, 29].

We explored the transcriptome in chicken gonads by using Lace to assemble SuperTranscripts combining four different transcriptomes: the Ensembl annotation, RefSeq annotation, a Cufflinks [30] genome-guided assembly and a Trinity [19] de novo assembly (see ‘Methods’). The Ensembl, RefSeq and Cufflinks transcriptomes were

annotations of the galGal4 genome from November 2011. We used an older version of the reference genome so that we could validate our approach using the most recent version, galGal5 from December 2015. To construct the superTranscriptome, we first combined the genome-based annotations (Ensembl, RefSeq and Cufflinks) by merging exons and concatenating the exon sequence to build a genome-based superTranscriptome. Trinity transcripts were then aligned against the chicken genome-based and human superTranscriptome; they were subsequently assigned to gene clusters. Finally, the genome-based superTranscriptome and Trinity transcripts were assembled together with Lace. See ‘Methods’ for a detailed description.

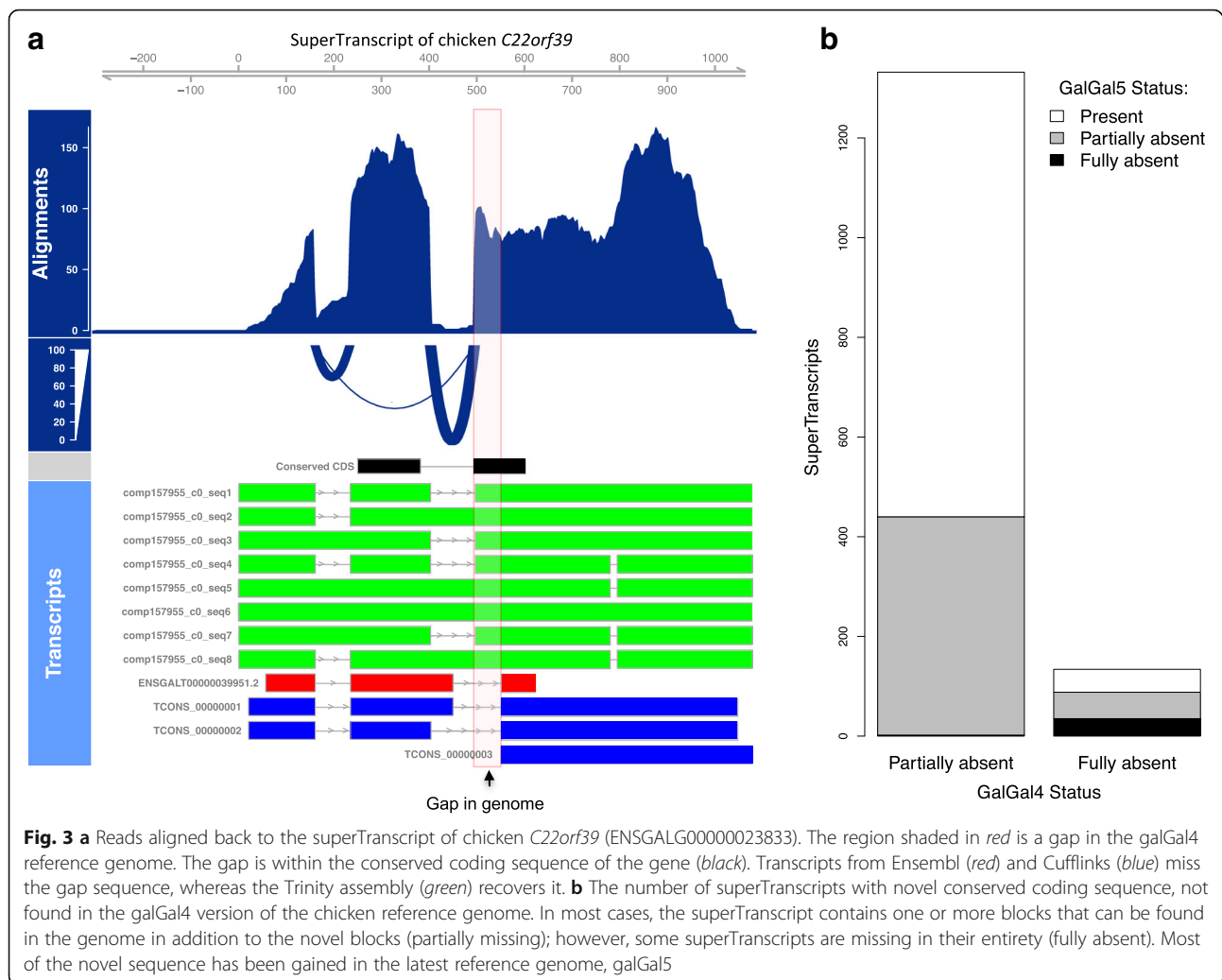
The resulting superTranscriptome was compact, containing just 83 Mbp (compared to almost 550 Mbp for the combined transcriptomes). However, none of the four contributing transcriptomes contained all of the sequence; 88%, 77%, 47% and 17% of bases were covered by Trinity, Cufflinks, Ensembl and Refseq, respectively. Critically, 3% (2.5 Mbp) of the bases in the chicken superTranscriptome could not be found in the galGal4 reference genome. This novel sequence included superTranscripts with protein-coding sequence either completely (134 superTranscripts) or partially (1332 superTranscripts) absent from the galGal4 reference genome. Figure 3a and Additional file 1: Figure S4 show an example of the *C22orf39* gene with a section of novel coding sequence. The novel section coincides with a known gap of approximately 100 bp in the assembly of the reference genome. For most superTranscripts, sections of novel sequence typically coincided with assembly gaps (Additional file 1: Figure S5A).

To validate the novel sequence, we aligned our superTranscripts to galGal5, which contains 183 Mbp more genomic sequence than galGal4 [31]. For 64% of superTranscripts with novel coding sequence in galGal4, the complete superTranscript was found in galGal5 (Fig. 3b). A total of 528 superTranscripts remain with missing sequence, including 35 that are entirely absent. This is likely because the current draft chicken genome is still incomplete [31] and approximately half the superTranscripts with novel sequence in galGal5 could be localised to regions that remain poorly assembled (Additional file 1: Figure S5B).

This analysis demonstrates the utility of superTranscripts and Lace to construct comprehensive transcriptome sequences in an automated way. It also highlights the major benefits of exploiting superTranscripts, even for a reasonably complete genome.

Using SuperTranscripts in model organisms

SuperTranscripts also have a number of uses in well annotated model organisms. When a reference genome is



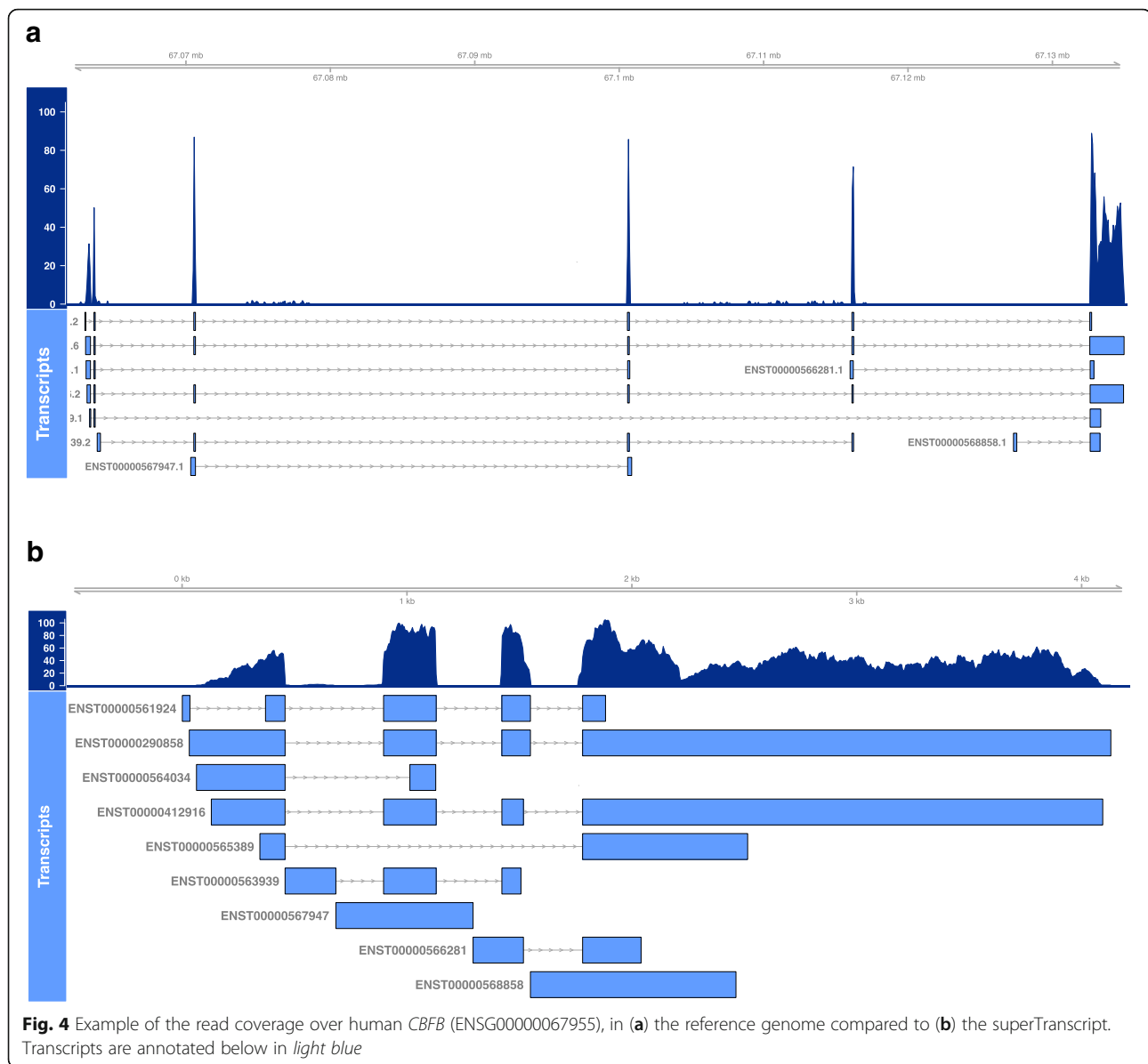
available, we can construct superTranscripts by simply concatenating the exonic sequence of each gene rather than using Lace (we provided the superTranscriptome for human at https://github.com/Oshlack/superTranscript_paper_code). Using this superTranscriptome as a reference drastically improves visualisation because intronic sequence is excluded, giving a compact view of the mapped reads, isoform and splicing structure (Fig. 4). Some genome browsers, for example RNAseqViewer [32], have an option to remove introns, but superTranscripts allow this for any genome browser including IGV. Often this results in the ability to visualise the sequencing data (including exome data) from a whole gene simultaneously in one screen of IGV instead of having to scroll through several screens. For transcriptome data, performing read alignment with the superTranscriptome is simplified compared to a reference genome because there is less sequence and fewer splice junctions (Additional file 1: Table S3). In addition, superTranscripts provide a convenient way of looking at the coverage and expression levels of long-read

data such as PacBio or Nanopore data where each read can originate from a different isoform of the same gene (Additional file 1: Figure S6 shows an example for PacBio data).

Conclusions

Here we have presented the idea of superTranscripts as an alternative reference for RNA-seq. SuperTranscripts are a set of sequences, one for each expressed gene, containing all exons without redundancy. We also introduce Lace, a software program to construct superTranscripts. Lace is unique as it is capable of assembling transcripts from any source, but existing transcriptome assemblers could also be modified to produce superTranscripts as additional output during the assembly. This would simply require the assembly graph to be topologically sorted.

Lace and superTranscripts can potentially be applied in a broad range of scenarios, some of which have been presented herein. Importantly, superTranscripts allow



the visualisation of transcriptome data in non-model organisms for the first time using standard tools such as IGV. Furthermore, superTranscripts allow differential isoform usage to be detected in non-model organisms by defining block (exon-like) structures in the transcripts and then using standard statistical testing methods such as DEX-seq. This is the first time differential isoform usage can be detected in non-model organisms using a count-based approach, rather than inference methods, and we find it to be more accurate. SuperTranscripts can also be used as a reference for reliably calling variants.

A powerful, new application of superTranscripts is merging transcriptomes from a variety of sources. We demonstrate this application using a chicken transcriptome that allowed us to detect hundreds of genes containing

sections of coding sequence that were not contained in the reference genome. We hypothesise that superTranscripts will have many further applications. For example, they are likely to increase power to detect differential isoform usage even in model organisms (Additional file 1: Figure S7). Although the concept of superTranscripts is simple, it has the power to transform how studies of non-model organisms are performed as a multitude of the standard analytical tools and techniques can now be applied across all species.

Methods

Datasets

To demonstrate how superTranscripts can be applied for visualisation and differential isoform usage, we used

the public RNA-seq dataset of human primary lung fibroblasts with a small interfering RNA knock-down of *HOXA1* from Trapnell et al. [27] (GEO accession GSE37704). To demonstrate variant calling, we sequenced RNA from genome in a bottle cell line (GM12878) to a depth of 80 million 150-bp paired-end reads using an Illumina NextSeq (SRA accession SRS2267720). The combined superTranscriptome for chicken was constructed using reads of chicken embryonic gonads from Ayers et al. [28] (SRA accession SRA055442).

De novo transcriptome assembly and clustering

Trinity [19] version r2013-02-25 was used to assemble the Trapnell dataset into contigs using default options and a minimum contig length of 200. Contigs were then clustered by Corset [18] with the test for differential expression turned off ($-D$ 99999999). The corset clustering and de novo assembled contigs were used as inputs to Lace to construct a superTranscript for each cluster. The superTranscript for each cluster was then assigned to a gene by aligning to the human reference superTranscriptome using BLAT with option $-\text{minScore} = 200$ $-\text{minIdentity} = 98$.

Read alignment

Reads were aligned to the genome or superTranscriptomes using the two-pass mode of STAR [20]. We used the STAR option $-\text{outSJfilterOverhangMin}$ 12 12 12 to filter the output splice junctions. Junctions supported by five or more reads were used to define dynamic block positions. The annotation was created using Mobius, a python script in the Lace suite. Read alignments were visualised using Gviz [33]. Our R scripts for Gviz are provided at https://github.com/Oshlack/superTranscript_paper_code.

Variant calling

In order to detect variants using superTranscripts we ran the GATK Best Practice workflow for RNA-seq (<https://software.broadinstitute.org/gatk/best-practices/rnaseq.php>). We then filtered for heterozygous SNPs, which we defined as at least one read supporting the reference allele, and with at least ten reads coverage in total. We used BLAT (options: $-\text{minScore} = 200$ $-\text{minIdentity} = 98$) to align all superTranscripts against the hg38 human genome. For superTranscripts that matched multiple genomic loci, we took the match with the highest score after excluding alternative chromosomes (denoted with ‘_alt’ in the hg38 reference). We then constructed a chain file for liftOver (downloaded from <http://genome.ucsc.edu/>) and converted all superTranscript SNP positions into genome coordinates (scripts are available at https://github.com/Oshlack/superTranscript_paper_code). Indels and other structural variants were excluded because of possible inaccuracy in defining the genomic position

when converting superTranscript to genomic coordinates. Finally, we excluded SNPs outside the Genome in a Bottle high confidence call regions and compared to the true positives reported by the Genome in a Bottle Consortium [21]. We also performed a genome-based approach to variant calling for comparison. For the genome-based analysis, the GATK workflow was followed using hg38 and subsequent steps were identical to the superTranscript analysis apart from translating coordinates to hg38 which was unnecessary. KisSplice version 2.4.0 was run with options, $-\text{max-memory}$ 100GB $-s$ 1 $-k$ 41 $-\text{experimental}$. The genome positions for KisSplice SNPs were found by aligning the output file ending with ‘type_0a.fa’ against hg38 using STAR (options: $-\text{sjdbOverhang}$ 73 $-\text{sjdbGTFfile}$ gencode24.gtf) and then running KisSplice2refgenome. SNPs with fewer than ten reads coverage were filtered out.

Counting reads per bin

The featureCounts [24] function from Rsubread R package (v 1.5.0) was used to summarise the number of reads falling in a given genomic region. Reads were counted in paired-end mode ($-p$) requiring that both ends map ($-B$). We allowed for reads overlapping multiple features ($-O$) assigning a fractional number of counts, $1/n$, depending on how many features, n , the read overlapped ($-\text{fraction}$). The same summarisation procedure was used in all cases.

Differential isoform usage comparison

The ‘truth’ for differential isoform usage was defined using DEXSeq testing with two different quantification methods. First, by mapping the reads to the human genome and counting reads in exon bins with featureCounts; and second, by quantifying transcript expression with kallisto. In both cases, differential isoform usage was detected with the R package DEXSeq [25] (v1.22.0) based on the reference annotation. DEXSeq takes a table of block level counts from featureCounts and, by default, removes blocks with fewer than ten counts summed across all samples. DEXSeq then produces a per gene q-value as the probability that for a given gene there is at least one exon used differentially between conditions, controlling for multiple testing. All genes that had a q-value < 0.05 in both the featureCounts and kallisto methods were considered true positives while all genes with a q-value > 0.9 in both methods were considered true negatives. For the de novo assembly, where multiple clusters were mapped to the same gene, the cluster with the lowest q-value was chosen and the others discarded. Where a single cluster mapped to multiple genes, the gene with the lowest q-value in the truth analysis was chosen as truth. For the de novo assembly analysis, the use of superTranscripts with standard blocks and dynamic blocks was contrasted with Kallisto [10], RSEM [12] and Salmon [11]. Kallisto, RSEM and Salmon were run on the

de novo assembled contigs using the default settings. Estimated counts per contig were then grouped into clusters using the same Corset clustering as was used by the superTranscripts. The count table was processed by DEXseq in the same way as the block counts table for superTranscripts.

Constructing a comprehensive chicken superTranscriptome

Ensembl and RefSeq references were downloaded for the chicken genome version galGal4 from UCSC on 24 August 2016. Cufflinks transcripts were assembled using the gonad reads from Ayers et al. [28], mapped to galGal4 using TopHat [34] version 2.0.6. The reference and cufflinks assembled transcripts were then merged into loci based on genomic positions using the cuffmerge command. The resulting annotation was flattened and exonic sequence concatenated to create a genome-based superTranscriptome, similar to that described below for the human. To supplement these superTranscripts with de novo assembly, we first assembled all reads using Trinity. Trinity contigs were aligned against the genome-based chicken superTranscriptome using blat with options `-minScore = 200 -minIdentity = 98`. Contigs that aligned to a single genome-based superTranscript were clustered with it. Contigs matching two or more genome-based superTranscripts were discarded (to remove false chimeric transcripts [35]). Remaining contigs were clustered into genes based on their homology to human superTranscripts (using BLAT with options `-t = dnax -q = dnax -minScore = 200`). Contigs that did not align to a gene, or those that aligned to multiple genes, were removed. Lace was then run on the sequence in each cluster, containing genome-based superTranscripts and Trinity contigs.

In analysing the chicken superTranscriptome, we assessed the coverage from each constituent transcriptome, Ensembl, RefSeq, Cufflink and Trinity, by aligning their sequence against the superTranscripts using BLAT with options `-minScore = 200 -minIdentity = 98`. We determined regions which were not present in the genomes galGal4 and galGal5 by aligning the superTranscripts against the chicken reference genome using BLAT with options `-minScore = 100 -minIdentity = 98`. Finally, we looked for regions with homology to human-coding sequence by aligning the superTranscriptome against the Ensembl GRCh38 human protein sequence using BLAST [36] with options `-evalue 0.00001 -num_alignments 20`. For a superTranscript region to be identified as novel protein-coding sequence, we required it to be absent from the chicken genome, match a human protein sequence with BLAST $e\text{-value} < 10^{-5}$, only be annotated by a Trinity transcript and be 30 bp or longer. Scripts used in the

chicken superTranscript analysis are provided at https://github.com/Oshlack/superTranscript_paper_code.

SuperTranscript construction for model organisms

When a reference genome was available we constructed superTranscripts by concatenating exonic sequence rather than using Lace. Doing so is more accurate as it does not rely on BLAT alignment or resolving loops in a splicing graph. The genome and annotation we used for human were taken from https://github.com/markrobinsonuzh/diff_splice_paper. As in Sonesson et al. [26], the genome annotation was flattened, such that transcripts were merged into disjoint blocks. The sequence of each block was then extracted and concatenated for each flattened gene using the `gffread -w` command from the cufflinks suite. To annotate, we projected the genomic coordinates of transcripts onto the superTranscripts, then flattened the transcripts into blocks. The resulting human superTranscriptome, its annotation and the scripts used to create them are provided at https://github.com/Oshlack/superTranscript_paper_code.

Additional file

Additional file 1: Contains supplementary figures and tables. (PDF 817 kb)

Acknowledgements

We would like to thank Paul Ekert and Stefanie Eggers for preparing and providing RNA-seq from the Genome in a Bottle cell line. In addition, we want to thank Ian Majewski, Jovana Maksimovic and Harriet Dashnow for feedback on the manuscript and Michael McLellan for his preliminary contribution.

Funding

AO is funded by an NHMRC Career Development Fellowship APP1051481. MCRI is supported by the Victorian Government's Operational Infrastructure Support Program

Availability of data and materials

All data used in this manuscript are available in public repositories as stated in the 'Datasets' section of 'Materials and Methods'. The new RNA-seq from the Genome in a Bottle cell line (GM12878) is found in SRA (accession SRS2267720). Analysis scripts can be found at https://github.com/Oshlack/superTranscript_paper_code. Lace is released under the software license GNU GPL version 3. Lace Version 1.0 is used in this paper DOI:10.5281/zenodo.830594.

Authors' contributions

NMD and AO conceived the idea of superTranscripts and Lace. ADKH developed Lace and performed the differential isoform usage analysis. NMD performed the SNP analysis, differential isoform usage analysis and chicken superTranscriptome analysis. All authors contributed to the writing of the manuscript.

Ethics approval

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

¹Murdoch Children's Research Institute, Royal Children's Hospital, Melbourne, VIC, Australia. ²School of BioSciences, University of Melbourne, Melbourne, VIC, Australia.

Received: 13 April 2017 Accepted: 21 July 2017

Published online: 04 August 2017

References

- Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet.* 2009;10:57–63.
- Ozsolak F, Milos PM. RNA sequencing: advances, challenges and opportunities. *Nat Rev Genet.* 2011;12:87–98.
- Matlin AJ, Clark F, Smith CWJ. Understanding alternative splicing: towards a cellular code. *Nat Rev Mol Cell Biol.* 2005;6:386–98.
- Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, et al. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* 2012;22:1760–74.
- Chepelev I, Wei G, Tang Q, Zhao K. Detection of single nucleotide variations in expressed exons of the human genome using RNA-Seq. *Nucleic Acids Res.* 2009;37:e106.
- Bahn JH, Lee J-H, Li G, Greer C, Peng G, Xiao X. Accurate identification of A-to-I RNA editing in human by transcriptome sequencing. *Genome Res.* 2012;22:142–50.
- Maher CA, Kumar-Sinha C, Cao X, Kalyana-Sundaram S, Han B, Jing X, et al. Transcriptome sequencing to detect gene fusions in cancer. *Nature.* 2009;458:97–101.
- Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, et al. A survey of best practices for RNA-seq data analysis. *Genome Biol.* 2016;17:13.
- Martin J, Wang Z. Next-generation transcriptome assembly. *Nat Rev Genet.* 2011;12:671–82.
- Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol.* 2016;34:525–7.
- Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods.* 2017;14:417–9.
- Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics.* 2011;12:323.
- Heber S, Alekseyev M, Sze S-H, Tang H, Pevzner PA. Splicing graphs and EST assembly problem. *Bioinformatics.* 2002;18 Suppl 1:S181–8.
- Kahn AB. Topological sorting of large networks. *Commun ACM.* 1962;5:558–62.
- Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. *Nat Biotechnol.* 2011;29:24–6.
- Kent WJ. BLAT—the BLAST-like alignment tool. *Genome Res.* 2002;12:656–64.
- Pevzner PA, Pevzner PA, Tang H, Tesler G. De novo repeat classification and fragment assembly. *Genome Res.* 2004;14:1786–96.
- Davidson NM, Oshlack A. Corset: enabling differential gene expression analysis for de novo assembled transcriptomes. *Genome Biol.* 2014;15:410.
- Grabherr M, Haas B, Yassour M, Levin J, Thompson D, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol.* 2011;29:644–52.
- Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics.* 2013;29:15–21.
- Zook C, Chapman B, Wang J, Mittelman D, Hofmann O, Hide W, et al. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nat Biotechnol.* 2014;32:246–51.
- Sacomoto GAT, Kielbassa J, Chikhi R, Uricaru R, Antoniou P, Sagot M-F, et al. KISSPLICE: de-novo calling alternative splicing events from RNA-seq data. *BMC Bioinformatics.* 2012;13 Suppl 6:S5.
- Lopez-Maestre H, Brinza L, Marchet C, Kielbassa J, Bastien S, Boutigny M, et al. SNP calling from RNA-seq data without a reference genome: identification, quantification, differential analysis and impact on the protein sequence. *Nucleic Acids Res.* 2016;7 Suppl 6:gw655.
- Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics.* 2014;30:923–30.
- Anders S, Reyes A, Huber W. Detecting differential usage of exons from RNA-seq data. *Genome Res.* 2012;22:2008–17.
- Soneson C, Matthes KL, Nowicka M, Law CW, Robinson MD. Isoform prefiltering improves performance of count-based methods for analysis of differential transcript usage. *Genome Biol.* 2016;17:12.
- Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L. Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat Biotechnol.* 2013;31:46–53.
- Ayers KL, Davidson NM, Demiyah D, Roeszler KN, Grutzner F, Sinclair AH, et al. RNA sequencing reveals sexually dimorphic gene expression before gonadal differentiation in chicken embryos and allows comprehensive annotation of W-chromosome genes. *Genome Biol.* 2013;14:R26.
- Cutting AD, Ayers K, Davidson N, Oshlack A, Doran T, Sinclair AH, et al. Identification, expression, and regulation of anti-mullerian hormone type-II receptor in the embryonic chicken gonad. *Biol Reprod.* 2014;90:106.
- Trapnell C, Williams B, Pertea G, Mortazavi A, Kwan G, van Baren M, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol.* 2010;28:511–5.
- Warren WC, Hillier LW, Tomlinson C, Minx P, Kremitzki M, Graves T, et al. A new chicken genome assembly provides insight into avian genome structure. *G3 (Bethesda).* 2017;7:109–17.
- Rogé X, Zhang X. RNAseqViewer: visualization tool for RNA-Seq data. *Bioinformatics.* 2014;30:891–2.
- Hahne F, Ivanek R. Visualizing genomic data using Gviz and Bioconductor. *Methods Mol Biol.* 2016;1418:335–51.
- Trapnell C, Pachter L, Salzberg S. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics.* 2009;25:1105–11.
- Yang Y, Smith SA. Optimizing de novo assembly of short-read RNA-seq data for phylogenomics. *BMC Genomics.* 2013;14:328.
- Altschul S, Gish W, Miller W, Myers E, Lipman D. Basic local alignment search tool. *J Mol Biol.* 1990;215:403–10.

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at
www.biomedcentral.com/submit

