



Internal Report CS Bioinformatics Track 13-04

November 2013

Leiden University

Computer Science

Bioinformatics Track

Development of Quality Assessment Methods for
De Novo Transcriptome Assemblies & Expression Analysis

Jonathan den Boer

MASTER'S THESIS

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands



LEIDEN UNIVERSITY & DELFT UNIVERSITY OF
TECHNOLOGY

THESIS PROJECT

**Development of Quality Assessment Methods for De
Novo Transcriptome Assemblies & Expression
Analysis**

Author:

Jonathan DEN BOER

Master & Specialisation:

Computer Science - Bioinformatics

Supervisors:

Dr. Walter PIROVANO

M.Sc. Marten BOETZER

Dr. Erwin M. BAKKER

November 19, 2013

Abstract

The analysis of expression and regulation of genes and other cellular components, such as RNA and proteins has become a key element in many genetic studies since these may reveal important conclusions in comparative analyses. Understanding regulatory mechanisms that underly differential behaviour between for instance healthy and affected patients, or pathogen-resistant and vulnerable crops is a crucial step in setting up effective therapies and treatments (e.g. drug design). In this light the identification of differential expression between genes or transcripts is often employed to attach a quantitative measure to gene expression and activity.

Whereas until recently most gene expression experiments were performed using microarrays, the introduction of Next Generation Sequencing (NGS) has opened new doors for transcriptome profiling as expression analysis could be performed at low costs (through the sequencing of RNA molecules) eventually without prior sequence knowledge of an organism. Ideally an annotated reference genome and/or transcriptome is available to allow expression profiling through short-read alignment. However often such reference information is not available and consequently a *de novo* transcriptome needs to be assembled from scratch. The assembly of transcriptomes still poses lots of challenges, however. Mechanisms such as alternative splicing introduce a lot of complexity, often resulting in multiple contigs for single transcripts. Finally, regardless of the availability of a reference or not, it is still a challenge to translate alignment coverage into true expression values. In particular the proper normalization of differential expression between samples suffers from different data yields, amplification biases and background noise.

Finally, for both the *de novo* transcriptome assembly and differential expression analysis no accurate validation procedures exist. In this thesis several assemblers, aligners, and tools which test for differential expression are compared, while also testing new validation methods. This has finally led to the development of a *de novo* transcriptome assembly pipeline, which includes some of the researched validation methods.

This thesis aims to provide the reader with more insight in the current state-of-the-art tools for both the *de novo* transcriptome assembly and differential expression analysis, while proposing new methods for the evaluation of these results.

Contents

General Information	2
Introduction	3
Biological Background	3
Next Generation Sequencing using the Illumina platform	5
RNA-Seq	7
Transcriptome Assembly	7
Transcriptome Analysis	9
Baseclear	12
Methods & Materials	13
<i>de novo</i> Transcriptome Assembly	13
RNA-seq Analysis	18
Results	23
<i>de novo</i> Transcriptome Assembly Quality Assessment	23
Reference-guided Evaluation	23
Non-Reference-guided Assembly Evaluation	32
RNA-seq analysis	34
Read alignment & Transcript abundances calculation	34
Differential Expression	36
Discussion	40
<i>de novo</i> Transcriptome Assembly	40
Expression Analysis	41
Conclusion	44
<i>De novo</i> Transcriptome Assembly	44
Expression Analysis	45
References	47
Supplementary Results	52
Assembly Evaluation Results	52
Appendix 1 - Manuals	56
<i>de novo</i> transcriptome assembly pipeline	56

General Information

Student

Name: Jonathan den Boer

E-mail 1: jonathandenboer@gmail.com

E-mail 2: jonathan.denBoer@baseclear.nl

University: Leiden University & Delft University of Technology Master program: Computer Science - Bioinformatics

Supervisors

Name: Dr. Walter Pirovano

E-mail: Walter.Pirovano@baseclear.nl

Company: Baseclear

Department: Bioinformatics and genome analysis

Name: M.Sc. Marten Boetzer

E-mail: Marten.Boetzer@baseclear.nl

Company: Baseclear

Department: Bioinformatics and genome analysis

Name: Dr. Erwin M. Bakker

E-mail: erwin@liacs.nl

Faculty: Leiden Institute for Advanced Computer Science

Adress: Leiden University, Niels Bohrweg 1

Internship details

Thesis Project starting date: February 18, 2013

Thesis Project ending date: October 31, 2013

Company

Company: Baseclear

Department: Genome Analysis and Technology

Adress: Einsteinweg 5, 2333 CC Leiden

45 credits (ECTS)

Introduction

Biological Background

Protein Synthesis

During protein synthesis DNA is transcribed to form a pre-mRNA (pre-messenger RNA) molecule. The DNA template strand is first read in the 3' to 5' direction to form the pre-mRNA; the latter is then transcribed in the 5' to 3' direction. This highly complex process is regulated by a large number of proteins, such as RNA polymerase, transcription factors, and co-activators. The RNA polymerase enzyme is responsible for the transcription of the pre-mRNA chain. The transcriptional process is controlled by transcription factors (alone or in a protein complex with for example co-activators). These transcription factors bind to specific locations on the DNA, where they act as activators or repressors for the recruitment of the RNA polymerase enzyme. The main role of the co-activators is to increase gene expression by binding to a transcription factor. In eukaryotes, directly after or concurrent with transcription non-coding introns are spliced out of the pre-mRNA. The resulting mRNA consist of exons only. Splicing is carried out by the spliceosome (a complex of five small nuclear RNAs and several proteins factors. When combined they form an RNA-protein complex called snRNP)in cooperation with auxiliary proteins which recognize the splice sites. In prokaryotes, splicing is observed only rarely, and mostly results in non-coding RNAs. This is caused by the absence of the complete spliceosomal pathway. The splicing that does occur, is caused by either self-splicing or the tRNA (transfer RNA). Finally the mRNA (also known as a transcript) is translated into a specific amino-acid chain by the ribosome. The ribosome consists of two major subunits, the small ribosomal subunit which reads the mRNA, and the large subunit which joins amino acids to form the amino acid chain. The amino-acid chain will later fold into a protein. The process of transcription and translation is shown in Figure 1.

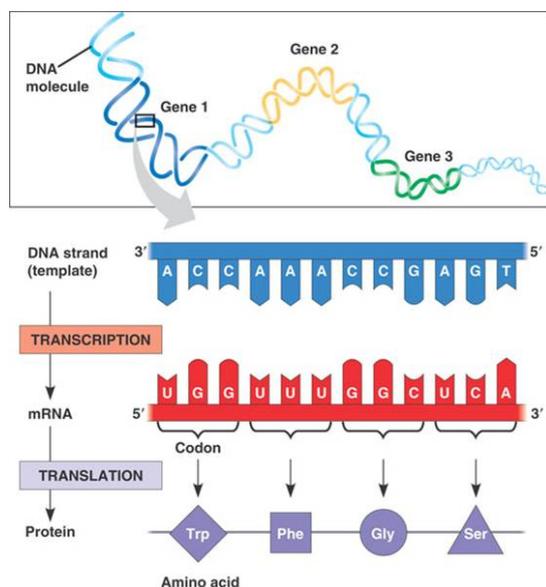


Figure 1: Gene 1 from the DNA molecule is transcribed into mRNA and translated into a sequence of aminoacids. When the transcription and translation process is complete, the aminoacid sequence can fold into a protein. Illustration adapted from [37].

Co-transcriptional modifications & Splicing Mechanisms

Whereas the basic transcription and translation process for eukaryotes and prokaryotes is very similar, in eukaryotes the transcribed mRNA requires co-transcriptional modifications. These modifications allow for the generation of a large number of mRNA and protein isoforms from a relative low number of genes. These modifications are caused by the following mechanisms or molecules: alternative polyadenylation, RNA editing, capping, Small Interfering RNA (siRNA, most notable in the RNA interference pathway), and alternative splicing. Polyadenylation always occurs directly after transcription of a gene is finished. A set of proteins cleaves off part of the 3'-most segment of the mRNA and synthesizes a poly A-tail at the cleavage site, from which the mRNA is later translated. Polyadenylation is important for translation, but also for mRNA stability and nuclear transport. Alternative polyadenylation cleaves a different part of the 3'-most segment. Because of this the mRNA is differently translated. An important molecule responsible for mRNA modification is siRNA, which interferes with the expression of specific genes by binding to the mRNA with a complementary nucleotide sequence, making transcription impossible.

The last mechanism described here is alternative splicing. Alternative splicing generates proteomic diversity by altering the RNA produced from the same genomic information, for example by splicing out a specific exon. The different splicing mechanisms are demonstrated in Figure 2. Alternative splicing is regulated by an intricate protein-RNA network and many different factors. One of these factors is the inhibition of splice site recognition where splicing silencers are located close to splice sites. Through this regulation different proteins are generated which function in diverse cellular processes. This protein-RNA network provides means for a cell to respond differently to changes in the environment. Such changes can be short-term like stress and time (e.g. day and night), but also accounts for long-term changes (e.g. age). While these mechanisms allow multi-cell organisms to carry the same information in each cell and still be able to respond different to the environment, mis-regulation of alternative splicing can also lead to diseases. It has also been shown that, of the human tissues examined, 50% or more of alternative splicing isoforms are differently expressed among tissues[3]. This indicates that most alternative splicing is regulated in a tissue-specific manner. Recent studies tried to estimate the alternative splicing isoform frequencies on sequencing data from mammalian cells. Evidence was found that over 90% of all transcription units are spliced in more than one pattern[2, 3].

Since all of these mechanisms are capable of producing different transcripts, the number of transcripts is much larger than the number of genes in a genome. The total set of mRNA molecules that are produced in one cell or a population of cells of an organism is called the transcriptome[1, 10, 15]. Obtaining a complete transcriptome is a very challenging process. The transcriptome complexity follows from the diversity in transcripts, which can be explained through the combinations of mechanisms mentioned earlier.

Metatranscriptomics

Environmental transcriptomics, or metatranscriptomics is a relatively new branch of transcriptomics that combines the transcriptomes of a group of interacting organisms or species. Metatranscriptomics is used to study (complex) communities. For example, metatranscriptomics can provide insight into the effect of environmental changes to a group of organisms. While metagenomics technologies have been used for many years to reveal the biodiversity of communities in an attempt to understand complex ecosystems (by providing information on genetic content), metatranscriptomics aims at understanding gene expression profiles in (microbial) communities (e.g. metabolic activities of a community at a specific time and place.[27])

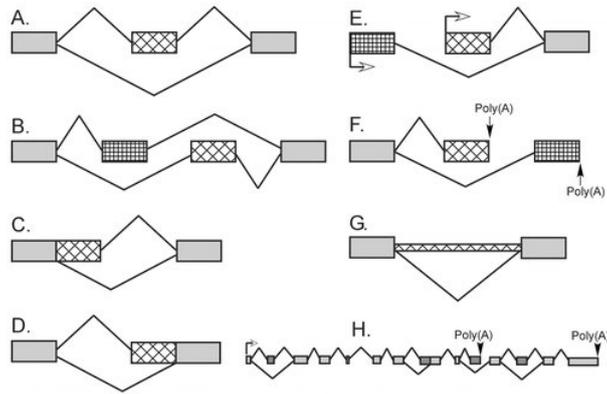


Figure 2: Patterns of alternative splicing. Constitutive sequences present in all final mRNAs are gray boxes. Alternative RNA segments that may or may not be included in the mRNA are hatched boxes. (A) A cassette exon can be either included in the mRNA or excluded. (B) Mutually exclusive exons occur when two or more adjacent cassette exons are spliced such that only one exon in the group is included at a time. (C, D) Alternative 5 and 3 splice sites allow the lengthening or shortening of a particular exon. (E, F) Alternative promoters and alternative poly(A) sites switch the 5- or 3-most exons of a transcript. (G) A retained intron can be excised from the pre-mRNA or can be retained in the translated mRNA. (H) A single pre-mRNA can exhibit multiple sites of alternative splicing using different patterns of inclusion. These are often used in a combinatorial manner to produce many different final mRNAs. Figure adapted from [3]

Next Generation Sequencing using the Illumina platform

For the past 25 years Sanger sequencing has been the most widely-used sequencing technique, and though still in use, it is becoming gradually supplanted by Next Generation Sequencing (NGS) methods such as provided by Illumina, Roche, and Life Technologies. Especially large-scale projects make use of these new NGS platforms. The key element of NGS (or high-throughput sequencing) concerns the parallelization of the sequencing process, resulting in the output of millions of sequences at once. As a consequence the sequencing price per base has dramatically dropped to 2400\$/million bases for Sanger sequencing and 0.05-0.10\$/million bases for Illumina sequencing[38]. Nonetheless a major drawback of the novel high throughput techniques concerns the short read length (typically between 50-250 bases for Illumina versus 400-900 for sanger[38]). Given the large amount of sequences produced at such low costs, NGS has also become one of the most widely used methods for gene expression analyses. Where microarrays require specific designs, NGS does not require any prior knowledge and can be used to detect expression levels through the very deep sequencing coverage.[38]

At present the leading platform in the NGS field is produced by Illumina. These sequencers are capable of producing a lot of reads per run (up to 3 billion, in comparison, other sequencers produce 1 million to 1,4 billion reads per run.), for a very low price[12, 38]. Figure 3 and 4 respectively show clustering of samples and the sequencing process.

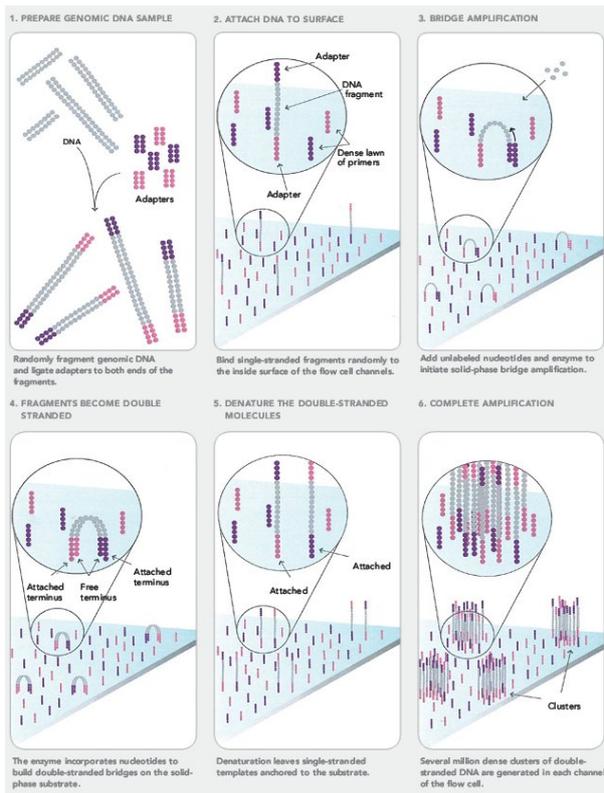


Figure 3: Schematic overview of the Illumina sequencing technology: Clustering. 1) Randomly fragmented DNA sequences are ligated to adapters. 2) the DNA sequences (with adapters) are bound to primers which are attached to the "Flow cell", after which the other end is also bound to a primer, and unlabeled nucleotides and amplification enzyme is added (3). 4) Unlabeled nucleotides are incorporated to create double-stranded DNA.

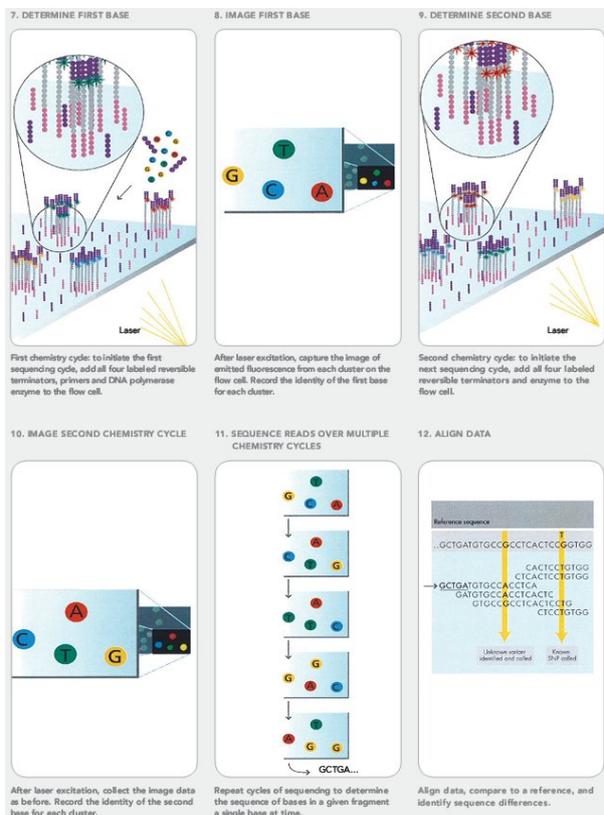


Figure 4: Schematic overview of the Illumina sequencing technology: Sequencing by Synthesis. 7) The synthesis process is initiated by adding fluorescent nucleotides, primers, and DNA polymerase. 8) After incorporation of the first base, the laser-excited fluorescence is captured by a CCD camera from which the base (of all clusters in parallel) is identified. 9-11) steps 7 & 8 are repeated for each base until the complete sequence has been identified.

RNA-Seq

RNA-Seq is a recently developed approach for transcriptome profiling that uses NGS technologies. The main advantage of RNA-Seq over genome sequencing is that the data is both qualitative and quantitative, meaning that gene expression levels can be retrieved. Currently a significant part of expression analysis are performed using microarrays. With NGS becoming more popular, microarrays will most likely be replaced by NGS.

In comparison to the sequencing of DNA, preparing RNA (cDNA) samples requires more and different steps. In general, a population of RNA (total or fractionated, such as poly(A)+) is converted to a library of cDNA fragments with adaptors attached to one or both ends (Figure 5). Each molecule, either amplified or not, is then sequenced in a high-throughput manner to obtain short sequences from one end (single-end sequencing) or both ends (paired-end sequencing). However, dependent on the research question it might be necessary to deviate from the standard protocol (e.g. when removing ribosomal RNA or for single strand sequencing)[38].

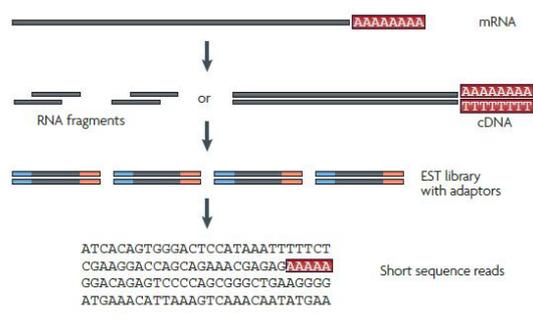


Figure 5: A typical RNA-seq experiment. Briefly, long RNAs are first converted into a library of cDNA fragments through either RNA fragmentation or DNA fragmentation. Sequencing adaptors (blue) are subsequently added to each cDNA fragment and a short sequence is obtained from each cDNA using high-throughput sequencing technology. Illustration adapted from [12].

Two other challenges for expression analysis are the lab-induced or protocol-specific biases caused by for example sequence bias (sequences around the start and end of fragments being non-random), position specific bias (non-uniformity of fragments along a transcript), and normalisation of the sequence data (e.g. when comparing to samples, the total number of reads should be taken into account)[4]. These may cause less accurate results for transcript quantification if not properly normalized.

Transcriptome Assembly

Transcriptome reconstruction is an important application of *de novo* RNA-Seq analysis, providing a (subset of a) transcriptome which can be used in both qualitative and quantitative research[26]. The assembly of a transcriptome can be done using two different methods: Reference-guided assembly and *de novo* assembly. The first method makes use of a reference genome to find the location of reads on the genome (taking into account splicing), after which overlapping reads are assembled into transcripts. One tool which makes use of a reference-guided assembly method is Scripture[14]. Scripture uses aligned reads (to the reference genome, including spliced reads). Next, a connectivity graph is constructed to find all transcripts. In contrast, *de novo* transcriptome assembly methods directly reconstruct overlapping reads into transcripts by utilizing the redundancy of sequencing reads themselves[7, 26]. Most *de novo* assembly tools (such as Oasis[39], Trinity[9], and CLC Bio[34]) use the De Bruijn graph approach to efficiently find overlapping regions[25]. While the reference-guided assembly has certain advantages over the *de novo* assembly, a close reference is not always available. Although RNA-Seq offers the potential to reconstruct the complete transcriptome picture, there are still many challenges and hurdles to take. The first issue is a wide range of gene expression levels, which leads to non-uniform sequence coverage. For instance lowly expressed genes may be only partially covered by a few reads and hard to be recovered to their full length. The second critical problem is to

handle pervasive alternative spliced isoforms. One gene may have several isoforms, yet reads are too short to tell which isoform they are from. Thirdly, homologous and repeated sequences, such as similar isoform sequences derived from the same gene may cause ambiguities in the assembly. Fourthly, it is problematic to discriminate exons and introns which may originate from incompletely spliced precursor RNAs[26].

Since the earlier mentioned *de novo* assemblers use the De Bruijn graph method, they have a lot in common. However, they all have unique characteristics to attempt to solve the challenges described above. Especially pervasive alternative spliced isoforms are handled differently by the three assemblers: CLC Bio was originally developed as a *de novo* genome assembler, and as a result, only reports one contig/transcript per region (CLC Bio is capable of detecting ambiguities in a De Bruijn graph, but due to its nature selects the most occurring path). Alternatively, in order to extract all alternative splicing isoforms, the Trinity method first recovers (using greedy and fast approaches) a single (best) representative for a set of alternative variants that have overlap (owing to alternative splicing, gene duplication or allelic variation). In the next step, Trinity constructs a De Bruijn graph for each cluster of related contigs (taking into account read pairs), and reports all plausible transcripts. Oases uses a similar approach but additionally uses a heuristic for complex cases. The method uses the coverage to traverse the graph and reconstruct transcripts.

Apart from the four challenges described above, CLC Bio also developed a duplication removal protocol, which removes technical duplicates before assembling the transcriptome. This should improve the assembly's quality as duplicates may negatively influence the De Bruijn Graph, and reduce the computation time. The duplication removal process on its own is time consuming, leaving no significant advantage timewise.

While certain studies[8, 25, 31, 39] indicate that all three methods are relatively well capable of accurately reconstructing transcriptomes, one of the major challenges of *de novo* work is how to measure the quality of a *de novo* assembly. Especially for alternative splicing isoforms, which are challenging to reconstruct, it is important to evaluate the quality of the assembled transcriptome.

Since a lot of downstream analysis are performed on the assembly, and taking into account that the quality of *de novo* assembled transcriptomes might vary, based on the total complexity, it is of great importance to evaluate the assembly.

Naturally, performing downstream analyses on a low quality assembly might lead to meaningless, or flawed results. For evaluation purposes several metrics have been defined to measure (the quality of) a reference-guided assembly. for example the number of transcripts found, and whether those transcripts were reconstructed for at least a certain percentage are good metrics to define the quality of an assembly[26]. But since generally no reference sequence is available, it is impossible to use those metrics or any other method which compares the reconstructed transcripts to that of a related species. One of the research strategies used to find one or more metrics to measure the quality of a *de novo* assembly is based on the question "What can we measure, and what is expected in an assembly?". For this reason several characteristics and possible markers such as transcript length or the presence of specific genes were studied and evaluated.

A final note from a biological point of view is that due to the lack of the spliceosome mechanism in prokaryotes, the *de novo* assembly of bacterial transcriptomes is much less challenging, as splicing occurs only rarely. Algorithmically this results in a much less complex De Bruijn graph. In contrast, *de novo* metatranscriptome studies on bacterial (and eukaryotic) samples are much more challenging. One of the greatest challenges here is to classify all reads to the right species before an assembly can be made. Often in such environmental samples closely-related species are present. Some methods try to first annotate the reads using some sort of reference, while other methods assemble all reads, and subsequently annotate the contigs. Apart from the biological challenges, environmental samples are often much larger than single transcriptome samples, and often contain much more redundant sequence information, making the assembly process much more computationally intensive. A direct consequence is that either more powerful hardware is required, or that the assembly accuracy is decreased (e.g. by decreasing the k-mer length).

Transcriptome Analysis

A transcriptome offers both qualitative and quantitative information. As mentioned earlier, the coverage distribution of RNA-seq data is non-uniform. Although this increases the complexity of a *de novo* transcriptome assembly, the wide-spread coverage distribution can be almost directly linked to gene expression. High expressed genes produce a lot of transcripts, whereas low expressed genes produce less transcripts. When a sample yields sufficient mRNA, this can be used to estimate the relative transcript abundance. One of the obstacles here is that besides mRNA, a lot of ribosomal RNA (rRNA is essential in the synthesis of proteins) is present in a cell (over 90% of the RNA in a cell is rRNA). As a consequence these samples yield insufficient mRNA to efficiently and accurately perform expression analyses. Until the development of accurate methods to remove the rRNA, over 70% of a sample was comprised of rRNA. Since then several (sample preparation) methods have been developed to either select only mRNA (Poly-A enrichment) or remove rRNA (rRNA depletion) from a sample. Currently, using these methods, samples containing <3,5% of rRNA have been reported[35]. By comparing the data to rRNA databases, most of the remaining rRNA can be filtered out. Transcriptome analysis involves three major steps: estimating transcript abundances, differential expression calculations, and testing for significant differential expressed genes.

While several models exist to estimate the relative transcript abundance, they are all based on the alignment of sequencing reads to a (reference) transcriptome (e.g. CLC Bio or Tophat2[44], which uses the fast aligner Bowtie2[36]). Again, different approaches can be applied to create such an alignment. One of the main challenges here is that often a significant part of the reads align to multiple locations on the transcriptome. While this underlies a natural phenomenon (e.g. when two transcripts share one or more exons, it is not possible to identify the origin), it makes transcript abundance estimations a highly complex exercise. To overcome these limitations a number of algorithms have been developed (HTSeq[43], CLC Bio, and Cufflinks[6]), all having their specific implementation to account for multiple aligned reads.

In addition to the alignment challenges, several other (biological) biases have been pinpointed[4]. For example, short transcripts or specific regions (e.g. GC-rich) on transcripts are sequenced less frequently. Since some of these biases might be library specific, there is no proven "best" method. Consequently different approaches have implemented alternative normalization strategies to estimate the transcript abundance, and their output is highly connected to the specific library preparation protocol used.

Finally, in order to compare expression values within and between samples, two additional normalization steps are required. Since the transcript abundance is expressed in read counts, a fair comparison can only be made when these counts are normalized by the transcript lengths. Secondly, a correction is required when comparing multiple samples with each other. Since the total number of aligned reads is never equal, the transcript abundances have to be normalized

over this number. A popular normalization method is the "RPKM" (Reads Per Kilobase per Million mapped reads) strategy. Gene counts are rescaled to correct for differences in both library sizes and gene length[40]. Another method, "FPKM" is used by cufflinks (Fragments instead of Reads). FPKM is similar to the RPKM normalization but optimized for paired-end reads. The difference between the RPKM and FPKM strategy is that RPKM calculates the number of **reads**, whereas the FPKM strategy calculates the number of fragments. A fragment here is a normal RNA fragment in paired-end RNA-Seq experiments, where the fragment is sequenced from both ends, providing two reads per fragment. An example of calculating the RPKM is given below:

$$\begin{aligned} \text{RPK} &= \text{nr. of Mapped reads} / \text{length of transcript in kb} \text{ (transcript length} / 1000) \\ \text{RPKM} &= \text{RPK} / \text{total nr. of reads in million} \text{ (total nr. of reads} / 1,000,000) \\ \text{nr. of mapped reads} &= 3 \\ \text{length of transcript} &= 300 \text{ bp} \\ \text{Total no. of reads} &= 10,000 \\ \text{RPK} &= 3 / (300 / 1000) = 3 / 0.3 = 10 \\ \text{RPKM} &= 10 / (10,000 / 1,000,000) = 10 / 0.01 = 1000 \end{aligned}$$

A deficiency of this approach is that the proportional representation of each gene is dependent on the expression levels of all other genes. Often a small fraction of genes account for large proportions of the sequenced reads and small expression changes in these highly expressed genes will skew the counts of lowly expressed genes under this scheme. This can lead to erroneous results of the differential expression analysis[41]. Other normalization methods attempt to normalize based on subsets of (stable) genes or by ignoring genes with very high expression values (e.g. the upper 25% quantile of the gene count sum)[42].

The large attention that has been given to both the technology and software aspects involved to properly perform comparative gene expression analysis has resulted in the development of many tools. However, it is often unclear to what extent they differ in practice and only a limited amount of method comparison studies have been performed [9, 41, 46]. Given the rapid developments it is crucial to validate the reliability of novel applications. Some of the tools used in comparative gene expression analysis are Cuffdiff (part of Cufflinks), CLC Bio, and DESeq[42]. The final step, testing differential expressed genes for significance, can be done in several ways, and might depend on the experiment design. While simple T-tests and ANOVAs are available, several algorithms/tests have been specifically developed for this purpose. For example, Cufflinks identifies a transcript as differentially expressed by testing the log-fold-change in its expression against the null hypothesis of no change. The significance is then assessed using a model of variability in the log-fold-change under the null hypothesis. DESeq[42], another popular tool for testing differential expression, calculates a scaling factor to normalize different samples. A ratio is calculated for each gene by dividing its read count by its geometric mean across all samples. The median of all ratios is the scaling factor for that sample[42].

To summarize, for both the *de novo* transcriptome assembly and expression analysis algorithms applies that results are of unknown quality, and that no best-practice protocol is available. The research performed here tries to contribute to some of the critical, and less researched aspects of both the *de novo* assembly of transcriptomes and the expression analysis. Starting with the assembly part, results of the different assembly algorithms are of unknown quality. Moreover, reliable assembly evaluation methods are not available, leaving users with a transcriptome that might contain uncomplete, or faulty transcripts.

For the expression analysis the complexity lies in the normalization of the data. Some of these normalizations are essential for all differential expression experiments, but some are experiment specific or library specific. To make it even more complex, for most normalizations a wide variety of algorithms are available. This research aims to provide a clear overview of the available algorithms and their differences.

The final goal of this project is both to be able to develop a *de novo* transcriptome assembly and expression analysis pipeline which incorporates both assembly quality evaluation methods and robust and accurate normalization methods for the expression analysis. In parallel, this thesis should give readers an objective overview of the available tools and algorithms, from which one can make the right choices, based specifically on their research.

Baseclear

BaseClear is a DNA service provider who provides high quality DNA analysis. This comprises sequencing using traditional methods (Sanger) and brand-new approaches among which Next Generation Sequencing using the illumina and PacBio platform. In addition a wide range of downstream bioinformatics analysis are performed, mostly focusing on the analysis of NGS data. Examples include *de novo* genome assemblies, SNP analysis, and reference alignments. BaseClear also offers a mRNA-sequencing service with corresponding transcriptome assembly and analysis services. Currently no reliable quality evaluation is available for *de novo* transcriptome assemblies. While BaseClear has a great interest in such a evaluation protocol to ensure the quality of their products toward clients, the bioinformatics community is also in need of such protocols for both validation and research purposes.

Methods & Materials

de novo Transcriptome Assembly

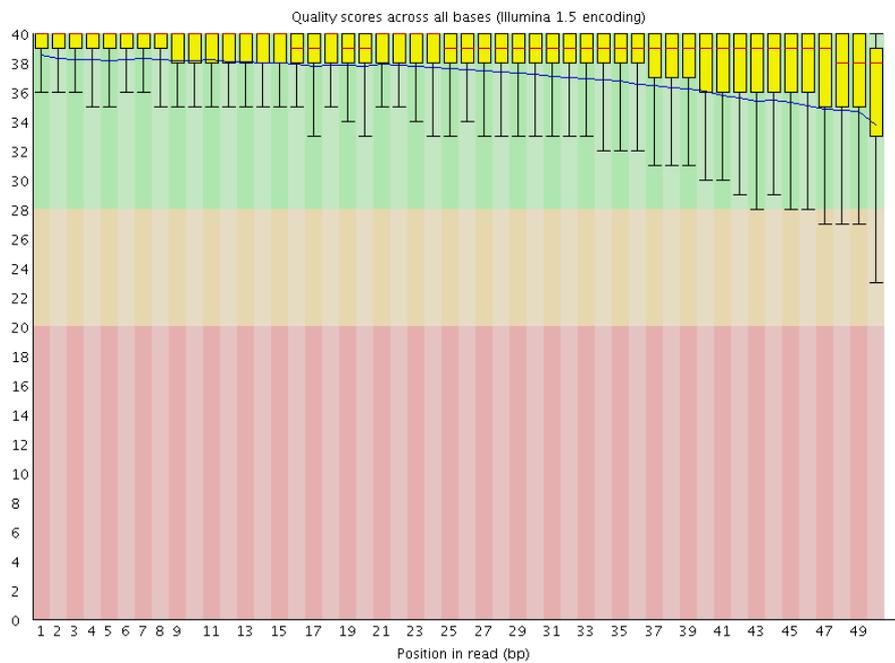
Data used

For the evaluation and optimization of the *de novo* transcriptome assembly pipeline two publicly available datasets of the tomato (SRR570061) and mouse (SRR654833) have been selected based on their characteristics to resemble next generation sequencing data generated by Baseclear. Both datasets were generated with an Illumina HiSeq 2000 platform. A typical transcriptome set generated by Baseclear is a paired-end dataset with reads of 50-150 bp long and 50x coverage. The datasets were downloaded from the NCBI Sequence Read Archive (SRA). Detailed information of the datasets is shown in Table 2.

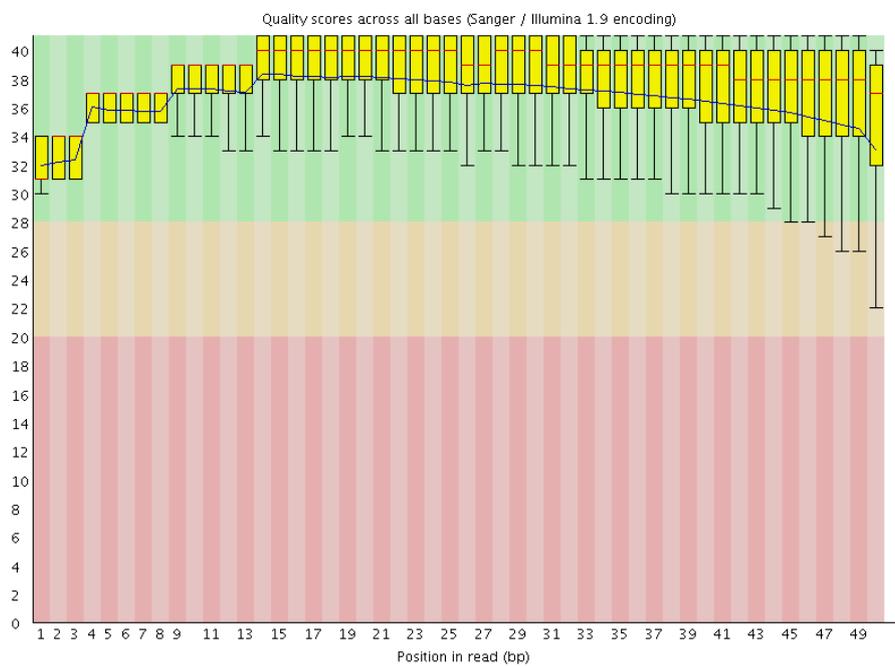
Dataset	SRA Accession	Paired-end	Length (bp)	# of Reads	GC-content
Tomato	SRR570061	Yes	50	11,042,746 (x2)	42%
Mouse	SRR645833	Yes	50	24,074,664 (x2)	49%

Table 1: Characteristics of the datasets used for the *de novo* transcriptome assembly. Both datasets were sequenced using the Illumina HiSeq 2000 platform.

Figure 6 shows the quality scores for the tomato data (6a) and mouse data (6b). These quality scores provide important information about the accuracy of the sequencing run. Phred quality score is the most common metric used to assess sequencing accuracy. Phred scores are generated during base-calling based on a complex model. Parameters relevant to a particular sequencing chemistry are analyzed for a large empirical data set of known accuracy. The resulting quality score lookup tables are used to calculate quality scores for new sequencing runs (in real time). Phred scores follow a log-scale (e.g. Phred score of Q30 is equivalent to the probability of an incorrect base call 1 in 1000 times), but are post-sequencing converted to ASCII characters to be compliant with the FastQ format. The figures (generated with FastQC) show the quality score per base averaged over all reads. The y-axis represents the average score, and the x-axis the position (base) in the reads. The amount of duplicates (not shown here) is 59% in the tomato, and 60% in the mouse data set.



(a) Tomato Dataset



(b) Mouse dataset

Figure 6: Average base quality for tomato and mouse dataset.

Evaluation and quality assessment of the results is partly done using reference transcriptomes. The mouse reference transcriptome (GRCm38/mm10) was obtained from the UCSC Genome Browser and contains 338,551 transcripts. The tomato reference transcriptome was obtained from the International Tomato Annotation Group (ITAG Release 2.3) and contains 34,727 transcripts.

Note that the datasets are not directly derived from the reference transcriptomes and as a consequence there might be a discrepancy between the reference and the data sets used in this research. While the expectation is that a significant part of the data is also in the reference, there might be some discrepancy between the raw data and the reference sets. On the other hand, the raw datasets contain only a part of the reference transcriptome, since not all possible transcripts are present in a single sample.

Assembly Methods

Two state-of-the-art methods for *de novo* transcriptome assembly are Trinity[14] and CLC Bio[34]. In this study, their relative performance is compared and evaluated using different metrics and methods. Other tools such as Oasis and Velvet have been left out as Trinity is often found to perform best [14, 25] (Trinity also is the first NGS transcriptome assembler that does not rely on a genome assembler while also addressing alternative isoform reconstruction). While CLC Bio is also used a lot, it is almost never used in comparative studies due to the license costs (as it is commercial software). CLC Bio was originally developed as a genome assembler, and thus searches for the longest contigs (using the De Bruijn graph approach). Trinity was developed to efficiently *de novo* reconstruct transcriptomes, keeping three critical challenges in mind: efficiently handling large amounts of raw data, defining a suitable algorithm to recover all splice forms, and providing robustness to the noise stemming from sequencing errors[14]. Trinity consist of three modules: Inchworm, Chrysalis, and Butterfly. Inchworm first assembles reads into single transcripts. Next, Chrysalis clusters related contigs (e.g. alternative spliced isoforms), and constructs a De Bruijn graph for each cluster. Finally Butterfly analyzes the paths taken by the reads and reports all plausible transcript sequences.

Assemblies were made for both datasets with both Trinity and CLC Bio with default settings. Standard assembly statistics such as minimum readlength, maximum readlength, average readlength, and GC and readlength distribution are computed with a Python script.

Assembly Quality Assessment

Although both assembly methods have shown to be accurate in a number of (comparative) studies[8, 14, 31], the quality assessment of *de novo* transcriptome assemblers can still be a challenging task as no standard metrics are defined. Most available metrics are based on the availability of a reference transcriptome[26]. These metrics compute the percentage of transcripts that is recovered, the completeness of the contigs (e.g. if a transcript of length 1200 is covered by a contig of length 600, the contig - transcript overlap is 50%), and the number of contigs that belong to one transcript. Results on the current datasets studied are shown in three different plots (created using a custom made Python script) for each dataset. To validate contigs reconstructed by the *de novo* assemblers, raw reads are mapped to the reference transcriptome, raw reads to the assemblies, and the assemblies to the reference transcriptome. All Alignments are computed with Bowtie2[36], GMAP[30], and CLC Bio. For GMAP alignments an additional plot has been created showing multiple alignment information (by default GMAP returns all valid alignments. Bowtie2 and CLC Bio report the best hit by default).

The results of the alternative alignments are analysed in more detail. For each reconstructed transcript which aligns to more than one reference transcript, the number of unique genes (from which the reference transcripts originate) is computed. These results allow for better evaluation of the aligners (when two splicing isoforms share one or more exons, reads from those exons will

always have alternative alignments). These results are found in the Results section .

These criteria are in principle sufficient for comparison of assembly tools when a reference transcriptome is available, however in most situations this is not the case, and different metrics are required for the assessment of assembly quality. One such method is to identify "core genes" in related species and search for these genes in the assembly. Based on the fact that some highly conserved proteins are encoded in essentially all eukaryotic genomes, a well-conserved set of genes can be identified. In [28], 458 highly conserved genes were selected from the KOGs (euKaryotic clusters of Orthologous Groups) database[29] which are found in all of the following eukaryotic species: *Homo sapiens*, *Drosophila melanogaster*, *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*. While these organisms all share the 458 genes, there is of course natural variation between genes of different eukaryotic species. Therefore it is recommended to select the closest species available (if this information is known).

It is relatively easy to find out whether these core genes are present or not in the assembly by performing a local alignment of the assembled transcripts to the core gene sequences. Because of intronic regions however, it is a much greater challenge to find out how much of such a gene is covered by the (spliced) reconstructed transcripts. Even if all intronic regions are removed, a global alignment for spliced transcripts is impossible. Most of the aligners which are capable of detecting splice sites use statistical models which require a large amount of data in order to accurately predict splice sites. In total, three tools were tested for their capacity of aligning transcripts to core genes. Two of the tools (TopHat2[32] and STAR[33]) use statistical methods to find splice sites. TopHat2 can find splice junctions without a reference annotation. By first mapping RNA-Seq reads to the genome, TopHat identifies potential exons, since many RNA-Seq reads will contiguously align to the genome. Using this initial mapping information, TopHat builds a database of possible splice junctions and then maps the reads against these junctions to confirm them.[32]. STAR uses a different method called Maximum Mappable Prefix search. This method is explained using a simple example of a read that contains a single splice junction and no mismatches(Figure 7). In the first step, the algorithm finds the MMP starting from the first base of the read. Because the read in this example comprises a splice junction, it cannot be mapped contiguously to the genome, and thus the first seed will be mapped to a donor splice site. Next, the MMP search is repeated for the unmapped portion of the read, which, in this case, will be mapped to an acceptor splice site.[33]

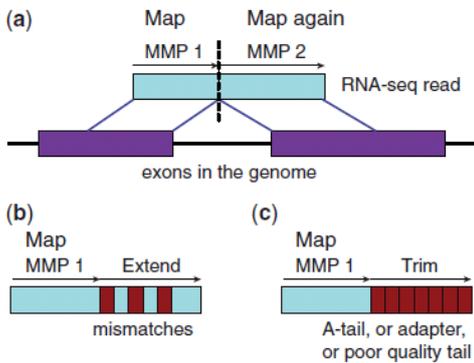


Figure 7: Schematic representation of the Maximum Mappable Prefix search in the STAR algorithm for detecting (a) splice junctions, (b) mismatches and (c) tails. Illustration adapted from [33].

The third tool is the cDNA-genome aligner GMAP[30]. GMAP uses a procedure called 'sandwich DP' (implementation of the Davis-Putnam algorithm) which computes subalignments around introns. Analysis of simple tests with TopHat, STAR, and GMAP showed that the results of the alignment with GMAP are the most accurate. This is also what one would expect, as TopHat and STAR predict splice sites using their statistical model, whereas GMAP "extracts" them from the alignment information.

Using GMAP the reconstructed transcripts were aligned against the DNA sequences of the core genes. The assembled transcripts are also aligned to the transcribed sequences (most abundant isoform) of the core genes. While this is a much simpler task (global alignment) than aligning transcripts to genes because of intronic regions, the former also allows for the identification of splicing isoforms, which might otherwise be overlooked (mainly if the most abundant isoform does not contain all exons). From these alignments the coverage of the core genes can be computed.

In summary: core genes described in [28] have been selected, and the cDNA-genome aligner GMAP was used to find both the number of core genes and the coverage of these genes in the assemblies.

Workflows

Figures 8 and 9 show the data flows in the analysis script and the assembly pipeline, respectively

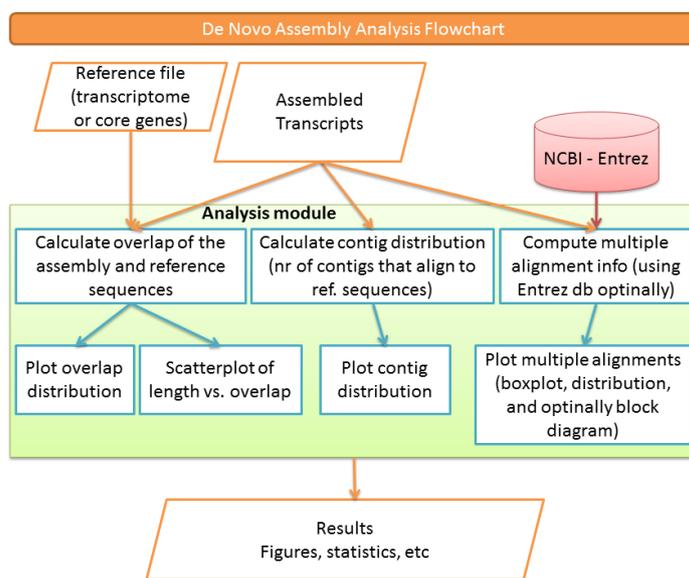


Figure 8: workflow of the script used for the different analysis. Input files are the assembled transcripts in SAM format, the reference transcriptome, and if the reference transcriptome is compatible with the NCBI Entrez database, a connection is established. The analysis module consists of the three analysis blocks which calculate and generate the results and figures.

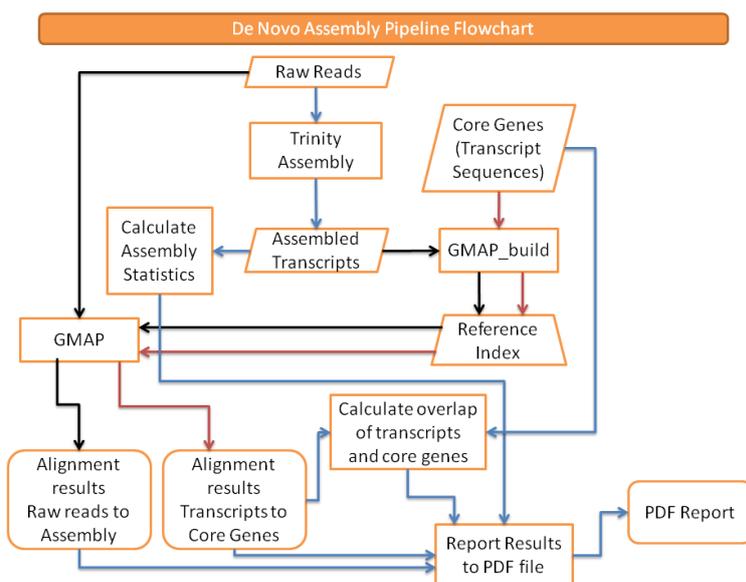


Figure 9: workflow of the *de novo* assembly pipeline. Raw reads are assembled using Trinity and subsequently assembly statistics are calculated. GMAP_build is used to create a 'reference' index for both the core genes and the assembly. GMAP is then used to align raw reads to the assembly, and to align the reconstructed transcripts to the core genes. For the core gene alignments, the overlap is calculated (as in Figure 8). All these statistics and results are then written to a PDF report.

RNA-seq Analysis

RNA-seq expression analysis can roughly be divided into three sections: the alignment of reads against transcripts, the estimation of transcript abundances, and testing for differential expressed transcripts or genes between multiple samples. The RNA-seq analysis section is divided into three subsections. The first paragraph describes the data used, and why these datasets were selected. The second subsections, "Read alignment & Transcript abundance calculation" describes the alignment and counting of reads to transcripts since there are multiple correlated factors that influence the abundance estimation. Finally, statistical methods used in testing the significance of differential expressed transcripts are reviewed.

Data Used

RNA-Seq Read Simulator[45] was used to simulate rna-seq reads from the mouse. The complete Ensembl mouse transcriptome (GRCm38.72) was used to simulate NGS reads at a 50x coverage, resulting in approximately 40m paired-end reads (or 20m pairs). Coverage calculation was done as follows:

```
transcriptome length: 159.352.638 bp
readlength: 100 bp
(transcriptome length * 50 coverage) / 100 = # of reads
(159.352.639 * 50 coverage) / 100 = 39.838.159 reads
```

The mean insert length was set as 300, with a standard deviation of 50. In order to analyse the algorithms and methods used no sequencing errors are introduced. The purpose of the simulated dataset is mainly to validate results from the different expression analysis tools, since significant differential expressed genes are either known, or can be introduced.

Read Alignment & Transcript Abundances Calculation

Read alignment was done using CLC Bio and TopHat2[44], which makes use of Bowtie2[36]. Bowtie2 is an ultrafast and memory-efficient tool for aligning sequencing reads to long reference sequences, and is based on the Burrows-Wheeler transform (BWT). The BWT is a reversible permutation of the characters in a text. Although originally developed within the context of data compression, BWT-based indexing allows large texts to be searched efficiently in a small memory footprint.[36] the alignment algorithms used by Bowtie2 are based on the 'last first' (LF) mapping' property of the Burrows-Wheeler Transformed matrix: The i^{th} occurrence of character X in the last column corresponds to the same text character as the i^{th} occurrence of X in the first column.

Expression analysis has been performed with CLC Bio, cufflinks, and HTSeq[43]. All three tools use different methods to obtain transcript counts/coverage. An overview of these tools is given in Table 2. The methods are described in the following subparagraphs.

CLC Transcript Abundance Estimation

CLC offers three different methods to obtain expression values: unique read count, total read count, and RPKM (explained in the Introduction). "Unique read count" indicates the number of reads that are uniquely assignable to the transcript in the **gene mapping**. Note that these reads can be non-uniquely assignable in the **read mapping**.

"Total read count" is the number of reads that are not uniquely assignable + the unique read count, where the non-unique reads are assigned to transcripts proportionally to the unique read count. The assignment of non-unique mapped reads is also normalized by transcript length.

Package	Expression count	Differential Expression (DE) calculation	Significance testing
Cufflinks	+	+	+
CLC	+	+	+
HTSeq	+	-	-
DESeq	-	+	+

Package	Notes
Cufflinks	Transcript abundance estimation tool, Often used in combination with Tophat
CLC	Commercial NGS data analysis package, focused on user-friendliness
HTSeq	Publicly available python-written tool for counting of expression data
DESeq	Publicly available R-written tool for normalizing counts, and testing for DE.

Table 2: Overview of the RNA-seq analysis tools used in this research. Columns 2-4 show whether the tools are capable of performing the three main tasks of a standard expression analysis. The 2nd part of the table gives a small summary of the tool.

See figure 10 for an example of how unique and total read counts are calculated. CLC requires a set of annotated reference sequences in order to perform the analysis.

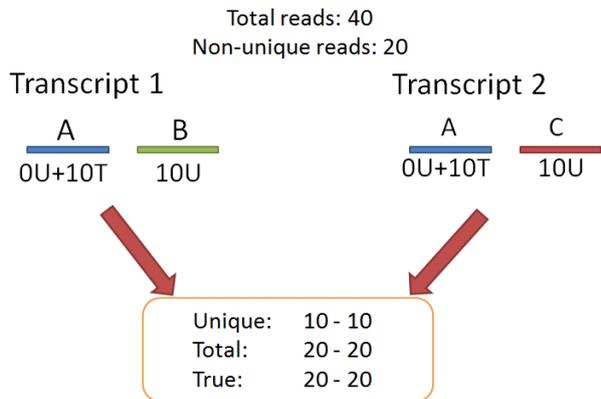


Figure 10: Unique versus Total readcount. when two transcript isoforms sharing one exon (A) are present in a dataset, the reads belonging to exon A can never be assigned uniquely to either transcript, as they are identical. When using the "unique reads" method only the 10 reads which mapped to exon B and C are counted. The total read-count method would result in a 20 - 20 count, proportionally to the unique readcount.

Cufflinks Transcript Abundance Estimation

Cufflinks calculates the Fragments Per Kilobase of transcript per Million mapped reads (FPKM), and estimates transcript coverage. FPKM is analogous to RPKM, but optimized for paired-end data. To solve the problem of reads that align to more than one position in the reference, Cufflinks uniformly divides each multi-mapped read to all the positions it mapped to. Furthermore, Cufflinks uses several normalization steps for more biological or library specific biases. Cufflinks requires a annotation file (General Transfer Format (GTF)/General Feature Format (GFF)) and an alignment file, such as produced by TopHat2.

HTSeq Transcript Abundance Estimation

HTSeq is a Python-written tool developed for the purpose of counting RNA-seq reads. The input is an alignment file in Sequence Alignment/Map (SAM) format (such as created with CLC or TopHat2) and an annotation file with the features of the sequences to which the reads align. HTSeq offers three ways on dealing with reads that overlap more than one feature: *union*, *intersection-strict*, and *intersection-nonempty*. These are defined as follows: for each position i in the read, a set $S(i)$ is defined as the set of all features (transcripts) overlapping position i . Then, consider the set S , which is (with i running through all positions within the read) the

union of all sets, the **intersection of all sets**, or the **intersection of all non-empty sets**. If S contains precisely one feature, the read is counted for this feature. If it contains more than one feature, the read is counted as **ambiguous** (and not counted for any feature), and if S is empty, the read is counted as **no_feature**. How these three modes work in practice is shown in figure 11. In this study the default "Union" mode is used. The "Union" mode always assigns reads to a transcript as long as all mapped bases map to the same transcript. Note that HTSeq cannot handle reads that align to multiple transcripts, or features. These reads are skipped, and reported in the "alignment_not_unique" result section.

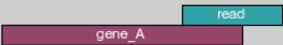
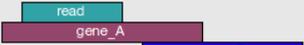
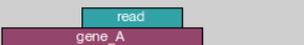
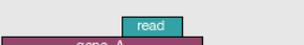
	union	intersection_strict	intersection_nonempty
	gene_A	gene_A	gene_A
	gene_A	no_feature	gene_A
	gene_A	no_feature	gene_A
	gene_A	gene_A	gene_A
	gene_A	gene_A	gene_A
	ambiguous	gene_A	gene_A
	ambiguous	ambiguous	ambiguous

Figure 11: HTSeq modes. Seven examples illustrating the effect of the three modes. Note that none of these modes allow reads that align to multiple transcripts.

Results of CLC Bio, Cufflinks, and HTSeq were compared with each other and the true transcript abundances. Aspects such as coverage distributions, and the number of reads that aligned with the different tools (and whether a relation between the two is observed) were analyzed. Also library-specific normalization methods have been reviewed (found in Cufflinks). For example Cufflinks assumes a sequence bias at the end of sequenced fragments caused by primers used in either PCR or reverse transcription.

Differential Expression Analysis

To simulate differential expressed transcripts, expression values of 12 manually selected transcripts were modified by simulating differing coverage depths. Simulation was done by random removal of a number of reads using a custom Python script. The remaining reads were stored in a new dataset (together with reads from the unmodified transcripts). Selection was based on transcript coverage, transcript length, and the absence of alternative splicing isoforms. Six transcripts with a coverage lower than the average coverage were selected, and six with higher than the average coverage were selected. To avoid getting biased results due to erroneous mapped reads, only transcripts that do not have any alternative isoforms were selected (for validation purposes). For each set of selected transcripts (2x6), the read removal percentages are: 50%, 60%, 70%, 80%, 90%, and 95%.

Differential expression (DE) calculations were performed with CLC Bio, cuffdiff (part of cufflinks), and DESeq. Table 4 shows the different methods used in the three tools.

Package	Significance test	Normalization procedure
Cuffdiff	fold-change vs. null hypothesis testing	FPKM
CLC	Kal's test & Baggerly's test	RPKM
DESeq	conditioned test	scaling factor (see description)

Table 3: The differential expression methods used in cufflinks, CLC Bio, and DESeq. Cuffdiff identifies differentially expressed transcripts by testing the log-fold-change of the expression against the null hypothesis of no change. The significance is then assessed using a model of variability in the log-fold-change under the null hypothesis. CLC offers two options: Kal's test (Z-test) for the comparison of two single samples, and Baggerly's test (beta-binomial) for either multi-group experiments or experiments with replicate samples. DESeq calculates a scaling factor to normalize different samples. A ratio is calculated for each gene by dividing its read count by its geometric mean across all samples. The median of all ratio's is the scaling factor for that sample. Testing for differential expression is then done using a test analogous to other conditioned tests, such as fisher's exact test.[42]

A number of different methods and parameters were used for the evaluation of Cuffdiff, CLC Bio and DESeq. These are summarized in Table 4.

Experiment	Alignment	Count/Abundance estimate	Differential expression testing
1	Tophat2	Cufflinks	Cuffdiff
2	CLC	Cufflinks	Cuffdiff
3	CLC	CLC	CLC
4	Tophat2	raw alignment counts	DESeq
5	CLC	raw alignment counts	DESeq

Table 4: Performed experiments. Experiment were designed to compare the default processing pipelines of CLC Bio and Cufflinks (experiment 1 and 3), and to compare different parts of (independent) tools (experiment 2, 4, and 5). For example, analyzing results of experiment 1 and 2 show a possible influence of the alignment on the differential expression test results. In CLC Bio it is only possible to perform a differential expression analysis on a CLC alignment, hence the TopHat2 - CLC Bio - CLC Bio experiment could not be performed. Experiment 4 and 5 have been performed to test and evaluate DESeq. Note that the counts in experiment 4 and 5 are raw read counts obtained from the alignment using a Python script.

Workflow

Figure 12 shows the different parts of the RNA-seq expression analysis

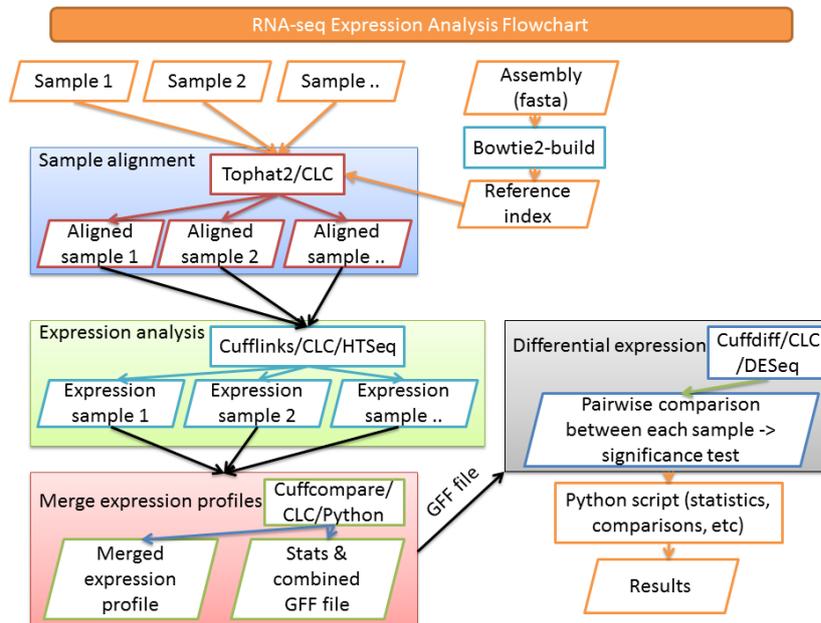


Figure 12: Workflow of the RNA-seq expression analysis. A number of samples (minimum of two) is analyzed for differential expressed genes, by first aligning reads and estimating the transcript abundances. Subsequently the expression values (transcript abundances) are compared, after which the differential expressed transcripts or genes are computed through statistical analyses. Output of the different tools is further analyzed and visualized using custom Python scripts.

Results

de novo Transcriptome Assembly Quality Assessment

Reference-guided Evaluation

Assembly Statistics

Assembly statistics have been calculated and are shown in Table 5. Column two and three show respectively the number of transcripts that are assembled, and the total number of basepairs. Columns four to six show the minimum, maximum and average transcript length. Column seven shows the N50[47], which is the length of the transcript where the cumulative size reaches 50% of the total transcriptome size. The N50 metric is often used for the evaluation of genome assemblies, but has also been adopted for the evaluation of transcriptome assemblies[48]. The N50 is found by calculating the total length of all transcripts, sorting all transcripts on their length in a descending order, and identifying the transcript that is located at 50% of the total size. These statistics have also been computed for the reference transcriptomes. Comparison of these numbers show that the amount of transcripts in the tomato assembly is relatively close to that of the reference (90.6% and 79.9% for respectively the Trinity and CLC assemblies). However, the total number of basepairs and average transcript length in the reference is almost twice the number and length of the assemblies. The mouse assemblies show completely different numbers: only 10.4% (Trinity) and 9.4% (CLC) of the transcripts in the reference have been reconstructed. The average transcript length in the assemblies is 80.3% (Trinity) and 77.9% (CLC) of the average length in the reference however. As a sidenote: it should be stressed however that the sequence reads of the mouse and tomato set only represent a subset of the complete transcriptome (i.e. the reads do not cover the complete transcriptome). It can therefore be expected that the final assemblies are incomplete. However it is still of great importance to compare the different results yielded by CLC and Trinity in order to validate which method can best generate an as complete as possible transcriptome. In order to make a fair comparison the amount of bases in the reference that are covered by each assembly is calculated and shown in Table 6. These results show again that the mouse sample contains only a small subset of the total mouse transcriptome. The reference coverage of the tomato is 41.4% 43.3% for Trinity and CLC, respectively, which is in accordance with the assembly statistics when taking into account the amount of transcripts that were reconstructed (90.6% and 79.9%) and their length (slightly more than 50%). The last column in Table 6 shows the reference coverage by the reads. From both the read coverage and assembly coverage an objective comparison between Trinity and CLC is made. In both cases (tomato and mouse samples) the reference is covered less by the assemblies than the raw reads (65% versus 41.4% (Trinity) and 43.4% (CLC) for the tomato, and 17.2% versus 12% (Trinity) and 11.1% (CLC). No significant difference is observed between Trinity and CLC for this comparison.

Evaluation of all assembly results suggest that Trinity outperforms CLC in terms of the number of transcripts that are reconstructed. This is also what is expected from the comparison of both algorithms (see methods section and the CLC paper on the assembly algorithm[34]). Whereas CLC reconstructs one transcript per gene, Trinity aims to also reconstruct all transcript isoforms that are present in the dataset. The length of the reconstructed transcripts is in all four cases significantly lower than the reference transcripts, and results of Trinity and CLC do not suggest that one performs better than the other, taking only transcript length into account.

Duplication Removal

In Table 5, assemblies denoted with "DR" are the assemblies build from the same data though after removal of duplicate reads. Duplicates have been removed using a custom script. Reads are considered duplicates if they are 100% identical. Differences between assemblies with and without duplicates are negligible for the assemblies made with Trinity, while the assemblies made with CLC show significantly more variation. The number of reconstructed transcripts decreased dramatically (roughly 26-30%), whereas the average transcript length increased by 24% in both cases. In summary, while the total length of the transcriptome (sum of base pairs) decreases after duplicate removal, the average transcript length increases.

Assembly	Transcripts	Sum of bp	GC %	Min	Max	Avg	N50	Gaps	Gap-size
Tomato - Reference	34.727	41.982.942	40,61	63	219.418	1.208	-	-	-
Tomato - Trinity	31.479	22.511.384	41,30	201	8.692	715	1.082	0	0
Tomato (DR) - Trinity	31.436	22.457.025	41,30	201	8.692	714	1.078	0	0
Tomato - CLC	27.769	21.054.901	41,38	189	8.692	758	1.141	5	6
Tomato (DR) - CLC	20.270	20.044.127	41,44	249	8.692	988	1.334		
Mouse - Reference	338.551	388.366.964	48,53	6	23.414	1.147	-	-	-
Mouse - Trinity	35.257	34.220.416	49,27	201	15.441	970	1.872	0	0
Mouse (DR) - Trinity	35.227	34.142.420	49,27	201	15.441	969	1.869	0	0
Mouse - CLC	31.956	30.094.660	49,28	181	16.577	941	1.720	35	143
Mouse (DR) - CLC	22.649	28.934.358	49,26	271	15.190	1.277	2.057		

Table 5: Statistics of the *de novo* assemblies for the Tomato and Mouse datasets and the reference transcriptomes.

sample	unique bases in assembly	bases in reference	coverage in %
Tomato reads	27.272.842	41.982.942	65,0%
Tomato - Trinity	17.399.262	41.982.942	41,4%
Tomato - CLC	18.179.447	41.982.942	43,3%
Mouse reads	66.798.906	388.366.964	17,2%
Mouse - Trinity	46.607.576	388.366.964	12,0%
Mouse - CLC	43.102.825	388.366.964	11,1%

Table 6: Reference coverage information. For each assembly the unique amount of bases that cover the reference transcript is calculated.

Alignment Evaluation

Several alignments have been constructed between raw data, reference data, and assemblies. Table 7 shows these results for the tomato and mouse dataset. Alignments are performed with Bowtie2, GMAP, and CLC. These results show that GMAP is capable of aligning significantly more reads than Bowtie2 in all cases (3-16% more reads are aligned). Between the two datasets, higher alignment rates are found for the mouse set. GMAP was developed as a mRNA to genome aligner, and most likely has the best results given the following alignment properties: capability of handling SNPs and sequencing errors, splice site detection (without the use of probabilistic models), and microexon identification[30].

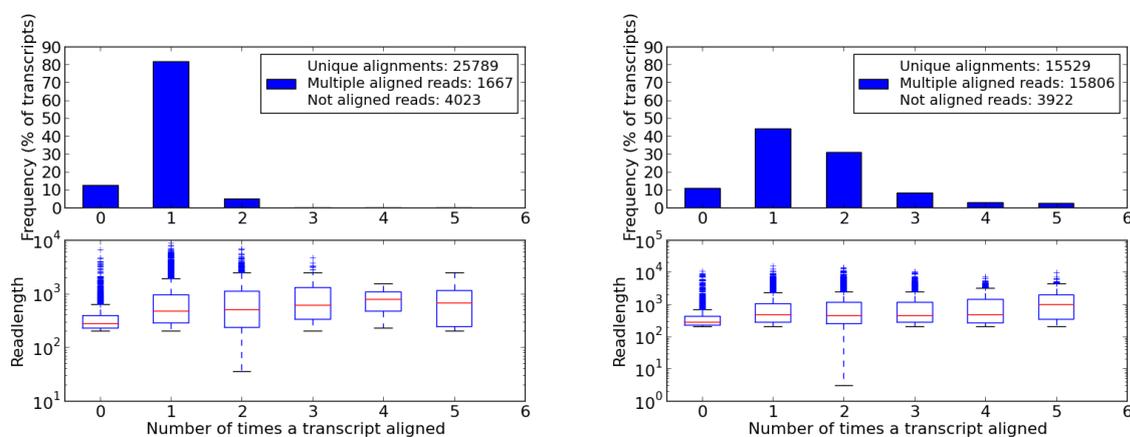
Alignment Type	Bowtie2	GMAP	CLC
Tomato: 11,042,746 reads	-	-	-
Raw to reference	83.32%	88.86%	82,14%
Raw to Trinity Assembly	86.97%	91.51%	85,33%
Raw to CLC Assembly	83.80%	88.38%	82,76%
Trinity assembly to reference	71.05%	90.81%	83,03%
CLC assembly to reference	72.87%	88.69%	84,34%
Mouse 24,074,664 reads	-	-	-
Raw to reference	92.63%	98.11%	91,19%
Raw to Trinity Assembly	90.76%	93.80%	89,44%
Raw to CLC Assembly	86.60%	90.57%	85,65%
Trinity assembly to reference	82.44%	93,32%	87,00%
CLC assembly to reference	86.37%	94.44%	89,45%

Table 7: Overall results of the following alignments for both datasets with Bowtie2, GMAP, and CLC: Raw data to reference, raw data to assembly, trinity assembly to reference, and CLC assembly to reference. GMAP obtains better alignment scores than the other two methods.

Finally, combining the assembly statistics and alignment rates, the percentage of bases in the reference that are covered by the reads is calculated to show how they compare to the assemblies. For the tomato which contains roughly 42 million bases, 65,0% of the bases are covered (at least once) by the sequencing reads (11,458,444/41,982,942), following the GMAP alignment. The Trinity and CLC alignments cover the reference by 75,8% and %, respectively. On the results presented in Table 5, GMAP is selected as the preferred alignment routine. For downstream analyses that require alignments GMAP is used. Results of the downstream analyses for Bowtie2 and CLC are found in the supplementary results chapter.

Multiple Aligned Transcripts

In certain cases, contigs (reconstructed transcripts) map to more than one transcript of the reference transcriptome. Often this concerns transcripts that represent alternative splicings (which originated from a single gene). In such scenarios the contig either aligns to isoforms which are very similar, or the contig might be as short as one exon. In the latter case alignments are created between the contig and all transcripts that share the particular exon. In the upper half of Figure 13a and 13b, distributions of the multiple aligned transcripts are shown for the tomato and mouse assemblies made with Trinity (results are similar for CLC Bio, data not shown). The lower half of the figure shows a boxplot of the length of transcripts, where each boxplot is made up of the samples in the distribution above the boxplot. The tomato assemblies show almost all unique alignments, transcripts in the mouse assembly often align to multiple reference transcripts. While one would expect the average transcript length to decrease when multiple alignments to the reference occur, as longer transcripts often contain more exons, which reduces the amount of "false" alignments to splicing isoforms, the opposite is observed.



(a) Trinity assembly of the tomato transcriptome (b) Trinity assembly of the mouse transcriptome

Figure 13: The number of times transcripts in the tomato and mouse assemblies (made with Trinity) align is shown in distributions in the upper half of the figure. The lower half of the figure shows statistics of the contiglength (boxplots) for each set of transcripts that align 0-5 times.

One of the challenges in *de novo* transcriptome assembly is to correctly distinguish alternative splicing isoforms produced by a single gene. It is very informative to find out whether the reference transcripts, to which a reconstructed transcript aligns, originates from a single gene. If that is the case the alternative alignments at least belong to the same gene (e.g. they are alternatively aligned transcripts). For the transcripts that align more than one time, the corresponding Entrez gene entries were retrieved, after which for each reconstructed transcript the number of unique genes is computed to find out whether a transcript aligns to alternative splicing isoforms from one gene, or to transcripts from different genes. Figure 14 shows whether multiple aligned transcripts aligned to alternative splicing isoforms or transcripts from different genes for the mouse data. The figure shows that in all cases over 80% of the reference transcripts to which the reconstructed transcript aligns, the reference transcripts originate from the same gene.

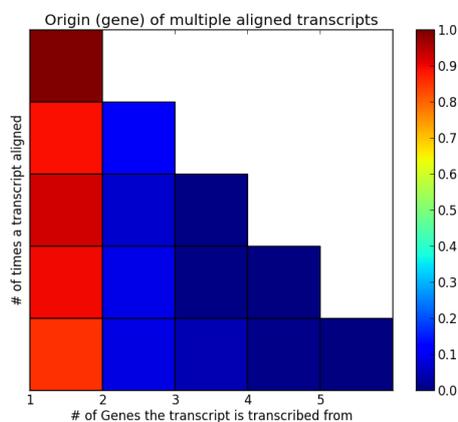


Figure 14: Graph showing the distribution of the number of unique genes that the alignments (reference transcripts) originated from, for the mouse dataset. Each row represents the set of reconstructed transcripts which aligned 1 (top row) - 5 (bottom row) times to the reference. The columns represent the number of unique genes. Each row of course has a maximum number of unique genes identical to the number of alignments (e.g. transcripts which align twice can only come from two different genes). The color shows the percentage of transcripts that originated from the number of genes defined by the column the block is in (e.g. Row 2 contains all contigs that aligned to two alternative reference transcripts. If the reference transcripts from that those transcripts are transcribed by the same gene, they are added to column one. If they are transcribed by different genes (max 2), they are added to column 2) colors, or percentages in each row thus sum up to 100%.

Tomato Assembly Evaluation

In Figure 15 the sequence overlap between the contigs (de novo transcripts) and the reference transcripts is illustrated for the tomato dataset. Figure 15a shows the findings for the Trinity assembly, 15b shows the CLC Bio results. In all cases, a lot of transcripts are found that share only 10-30% sequence identity with the corresponding reference transcript. Interestingly, only very few alignments were made between 30-90% sequence identity, whereas a second peak is observed between 90-100% sequence identity.

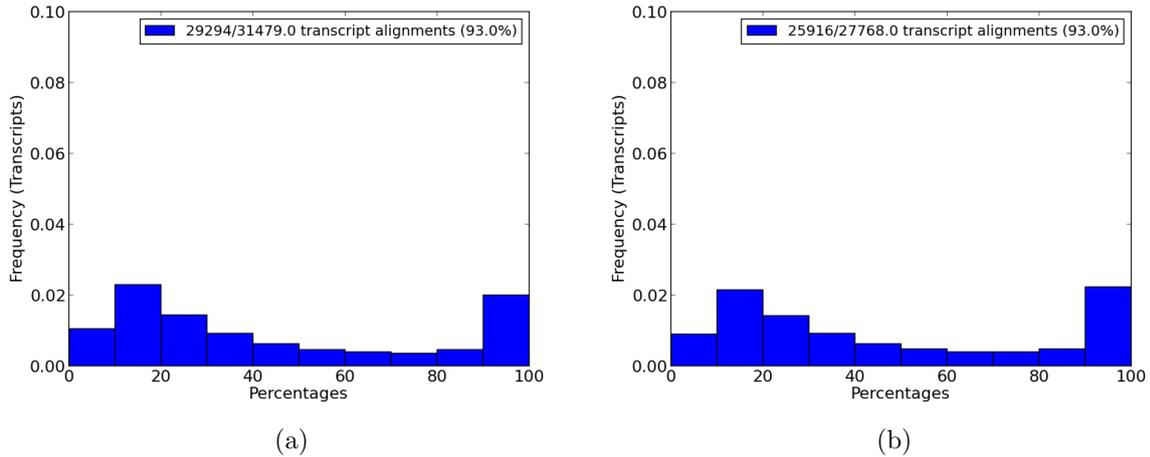


Figure 15: Transcriptome distribution illustrating the percentage of overlap between the reconstructed transcripts and reference transcripts. The x-axis shows the percentage of overlap between reconstructed and reference transcripts. The y-axis shows the relative amount of transcripts in percentages. (a) and (b) show the results for the Trinity and CLC assemblies, respectively.

As short transcripts are easier to assemble in general (since they often consist of less exons, or are incompletely reconstructed), one would expect these to have more overlap with the reference transcript than larger transcripts. Figures 16a (Trinity) and 16b (CLC Bio) show the relation between the length of the reconstructed transcript and the percentage of overlap with the reference transcript. The regression has been computed and is shown as a red dotted line. The regression shows that a positive correlation is found between transcript length and the relative overlap with the reference transcript, meaning that transcripts are covered relatively more as they increase in length. The plot also shows that almost all transcripts with length 200 and smaller only have 0-20% overlap with transcripts in the reference. These short transcripts might as well be junk introduced by e.g. noise in the sequencing data.

Finally, Figure 17a (Trinity) and 17b (CLC Bio) show the number of assembled transcripts that validly align (using the default threshold in GMAP) to one reference transcript. Under ideal circumstances each reference transcript is covered (almost) fully by a reconstructed transcript. When transcripts are not completely covered in the sequence data however, assemblers cannot correctly reconstruct the complete transcript. As a result several transcript fragments are reconstructed. This can also be the result of complex sequences/transcripts (e.g. repeats or through alternative splicing). The result of this is that multiple reconstructed transcripts align to one reference transcript. When taking into account that a lot of reconstructed transcripts only have 0-20% overlap with the reference transcripts (Figure 15), the result shown here is not surprising. Nonetheless in both cases over 50% of the reference transcripts are covered by one reconstructed transcript.

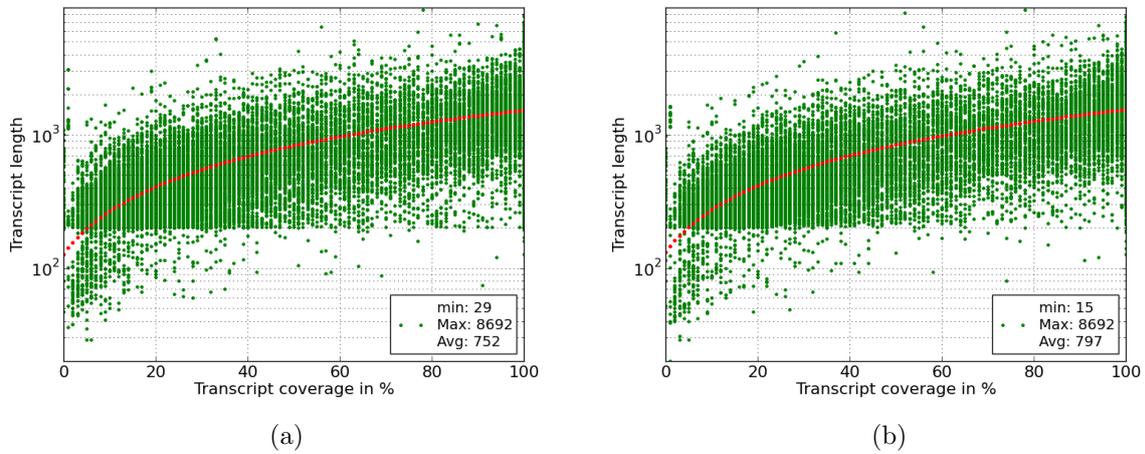


Figure 16: Scatterplot of the transcript length plotted against their relative coverage in the reference transcriptome. (a) and (b) show the overlap of the Trinity and CLC assemblies, respectively.

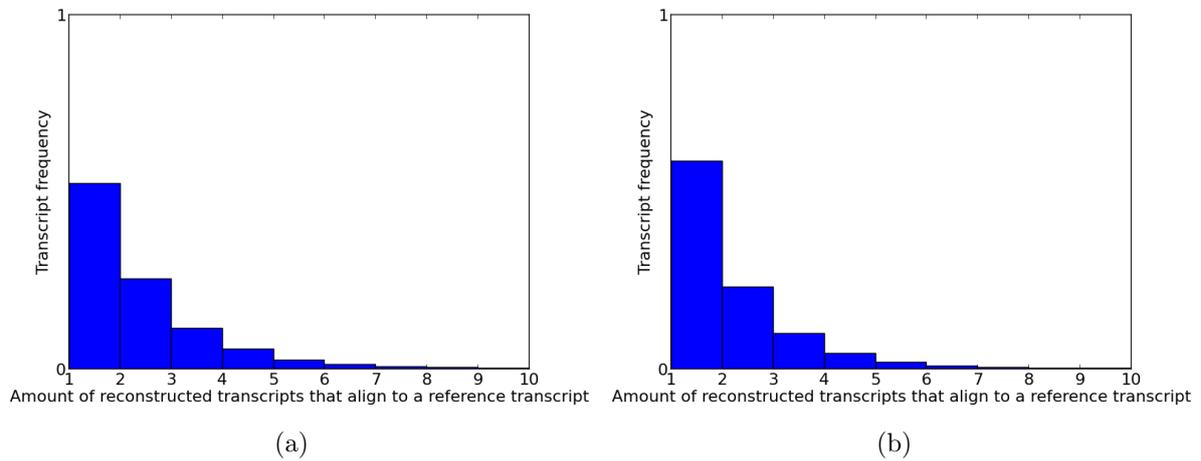


Figure 17: For each reference transcript, the number of reconstructed transcripts that align to that reference transcript is computed and shown as a distribution. Numbers on the x-axis represent the number of reconstructed transcripts that align to a reference transcripts, and the y-axis shows the frequency of those occurrences (0-100%). Figure (a) and (b) show the results for the Trinity and CLC assembly, respectively.

Mouse Assembly Evaluation

The analyses on the tomato assemblies have also been performed on the mouse assemblies. (Figure 18 - 20). The results show similar patterns for all three analysis in the tomato dataset.

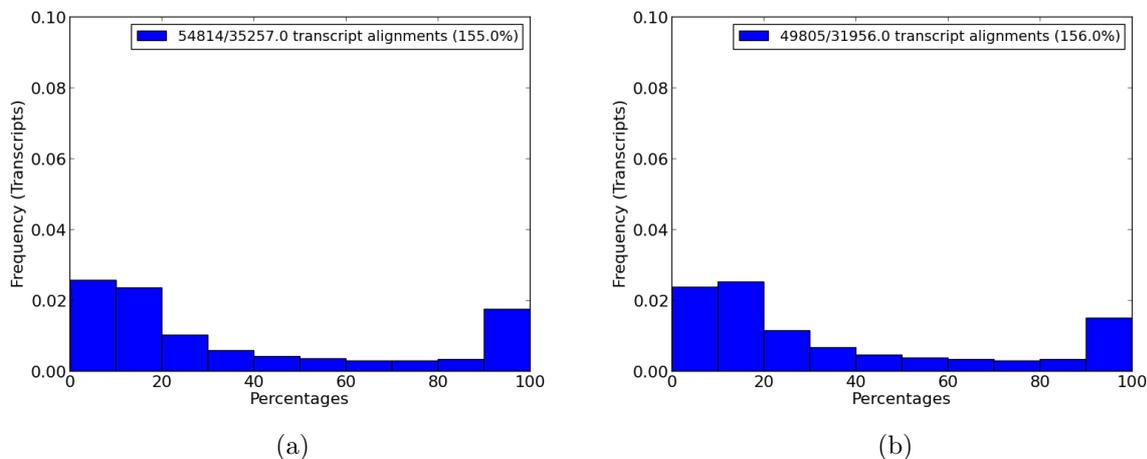


Figure 18: Transcriptome distribution illustrating the percentage of overlap between the reconstructed transcripts and reference transcripts. The x-axis shows the percentage of overlap between reconstructed and reference transcripts. The y-axis shows the relative amount of transcripts in percentages. (a) and (b) show the results for the Trinity and CLC assemblies, respectively.

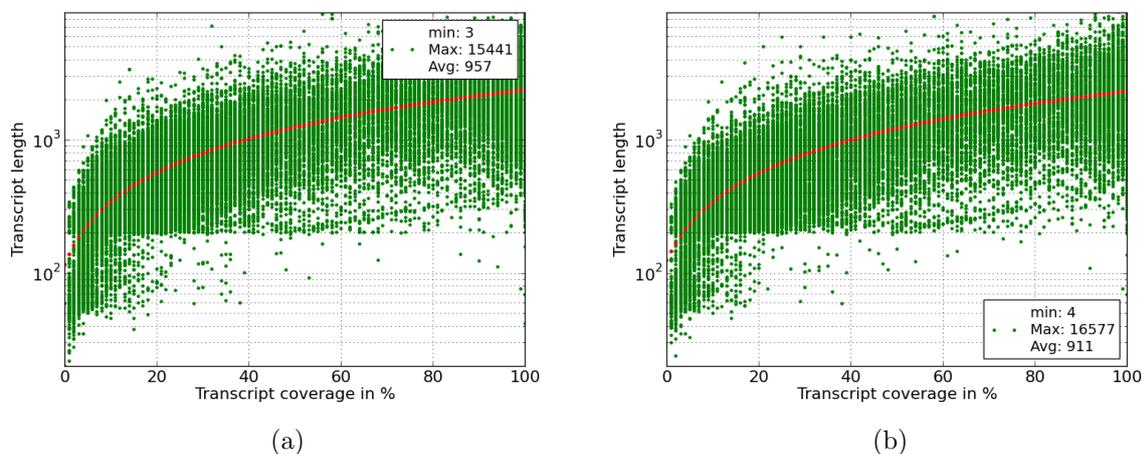


Figure 19: Scatterplot of the transcript length plotted against their relative coverage in the reference transcriptome. (a) and (b) show the overlap of the Trinity and CLC assemblies, respectively.

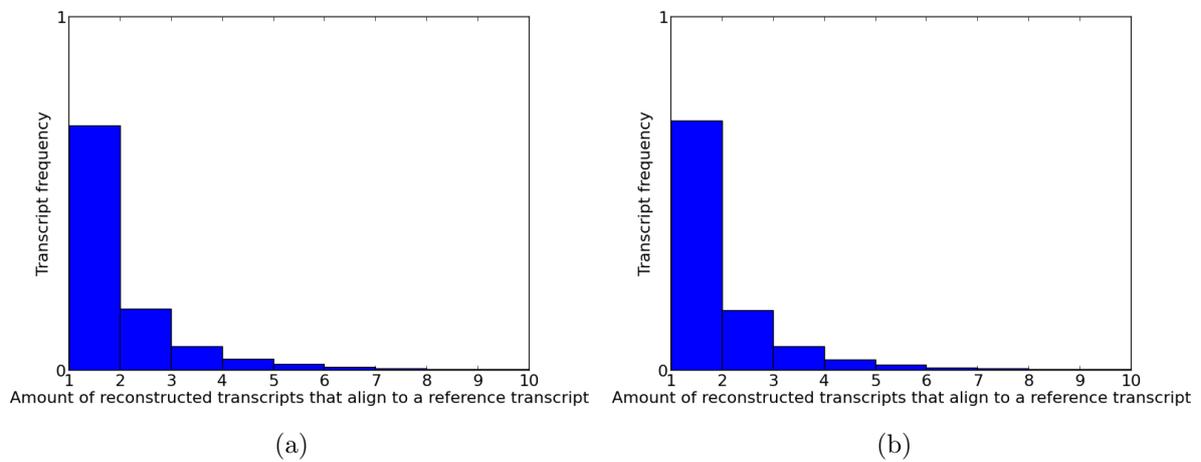


Figure 20: For each reference transcript, the number of reconstructed transcripts that align to that reference transcript is computed and shown as a distribution. Numbers on the x-axis represent the number of reconstructed transcripts that align to a reference transcripts, and the y-axis shows the frequency of those occurrences (0-100%). Figure (a) and (b) show the results for the Trinity and CLC assembly, respectively.

Non-Reference-guided Assembly Evaluation

Core Genes - Local Alignment

As described in the methods, specific (conserved) genes are found in all eukaryotes[28]. Using this information, the quality of a *de novo* transcriptome assembly can be assessed. Because of the presence of alternative isoforms, which are formed through splicing, two methods have been used to identify core **genes** in the assemblies.

In the first method the local alignment option of Bowtie2 is employed to quantify the number of core genes in a reconstructed transcriptome. The local alignment option is used since genes contain introns which are spliced out in the transcripts, making a continuous alignment impossible. In [28] a set of 458 genes were identified that are conserved among eukaryotes. For a number of species, this set of genes is available. The tomato assembly (Trinity) is aligned to core genes of *Arabidopsis thaliana* (both from the plants kingdom) as this is the closest related species, from an evolutionary point of view. The mouse assembly (Trinity) is aligned to core genes of the *Homo sapiens* (kingdom Animals, class Mammals). Table 8 shows the local alignment results. Out of 458 core genes, 73 and 317 unique hits were found with Bowtie2 for respectively the tomato and mouse assemblies. The remaining 9 and 392 alignments found in the tomato and mouse assemblies, respectively, aligned not uniquely. These multiple aligned hits are reconstructed transcripts which validly align to multiple core genes.

Reconstructed transcriptome	Tomato	Mouse
Unique hits	73	317
Multiple hits	9	392

Table 8: Local alignment results of the assemblies against the core genes

While above results give an indication of the number of core genes present in the assembly, and thus an indication of the assembly quality, no information regarding the overlap between the core genes and reconstructed transcripts is available.

For the second approach GMAP[30], a genomic mapper for mRNA was used. GMAP proved to be succesful at mapping the reconstructed transcripts to core genes by identifying splice sites. However, for transcripts which are only partly assembled, the total size of the reference transcript cannot be defined since it might not be clear which transcript it is, and thus which exons it contains. This is solved by using the most abundant transcript of each core gene (finding alternative splicing isoforms is limited this way). Figure 29 shows the overlap of reconstructed transcripts and the core genes, and the number of transcripts that align to a core gene. Results are shown for the assemblies made with Trinity. Results of the CLC assemblies are highly comparable, and can be found in the supplementary results section. The alignment numbers in the figure include multiple aligned transcripts. The number of unique alignments is shown in the caption of the figures. In all cases over 70% of the (transcripts of the) core genes is covered by the reconstructed transcripts for 90-100%. When looking together at figure (a), (b) and (c), (d), one can see a relation between the overlap in (a) and (b), and the number of reconstructed transcripts that align to a core gene in (c) and (d). In other words, if a core gene is covered more by a single reconstructed transcript, naturally less reconstructed transcripts align to that core gene.

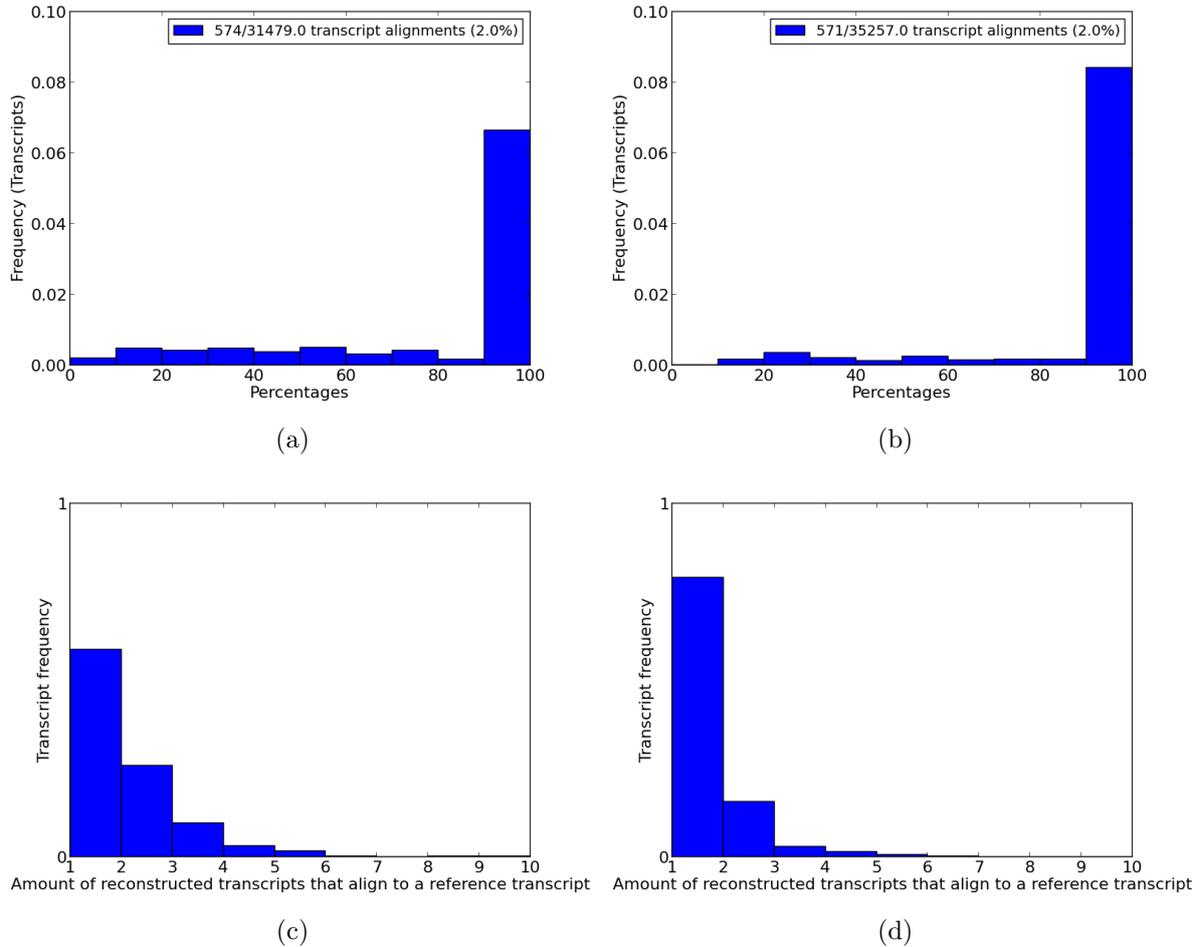


Figure 21: a) and b) Transcriptome distribution illustrating the percentage of overlap between the reconstructed transcripts and core genes. The x-axis shows the percentage of overlap between reconstructed and reference transcripts. The y-axis shows the relative amount of transcripts in percentages. The legend shows the number of core genes found, including multiple aligned reconstructed transcripts. For a) and b), 533 and 565 unique hits are found, respectively. c and d) Number of reconstructed transcripts that align to the transcript of a core gene. Tomato and mouse transcripts were aligned to core transcripts of respectively *Arabidopsis Thaliana* and *Homo Sapiens*. (a), (c) and (b), (d) are the tomato and mouse assemblies, respectively.

RNA-seq analysis

Read alignment & Transcript abundances calculation

Read Alignment

The simulated dataset is aligned to the reference transcriptome (complete mouse transcriptome) using CLC Bio and Tophat2 (Bowtie2). Table 9 shows a summary of the alignment results. The percentage of reads that aligned, and how many reads were aligned in proper pairs, according to the aligner, is shown, as well as the number of reads that align correctly (on the right transcript). Analysis of the alignments of the simulated dataset show that for both Tophat2 and CLC Bio almost all reads align properly in pairs (as expected), but also that only around 60% of the reads align on the correct transcript. However, for all pairs which aligned on a wrong transcript, the wrong transcript is an isoform which originates from the same gene.

Simulated dataset	Tophat2	CLC Bio
Parameter: insert size	0-500	0-500
Parameter: mean inner-distance	200	-
Parameter: standard deviation	50	-
Aligned	37.649.250/37.649.250 (100%)	37.035.368/ 37.649.250 (98.4%)
Aligned in proper pair	37.602.314 (99.9%)	37.035.368 (100%)
Correct Aligned	22.736.328 (60.4%)	22.698.728 (61.3%)

Table 9: Alignment results of the simulated dataset. Alignments were performed with Tophat2 and CLC. row 2-4 are parameter options for alignment optimization.

Transcript Abundance Calculation

Transcript abundances, or readcounts, were calculated according to the methods described in the Methods section using CLC Bio, Cufflinks, and HTSeq. The abundances calculated by CLC Bio and HTSeq are measured in amount of reads per gene (or transcript), and thus are directly comparable to each other and the reference. The output of Cufflinks is, for each transcript, an estimate of the absolute read coverage across that transcript and the corresponding FPKM value. To create a fair comparison the Cufflinks coverage has been calculated back to readcounts (by multiplying the coverage by the transcript length, and dividing that by the readlength). This not only offers the possibility for fair comparisons, but by assuming the Cufflinks expression estimation method is accurate, the results of other normalizations is "left" in the data by calculating the readcounts. Evaluating transcript abundances is done by plotting the calculated transcript abundance against the readcounts from the reference. Figure 22 shows scatterplots of the transcript abundances calculated by CLC Bio (22a), HTSeq (22b), and Cufflinks (22c) versus the truth. These plots show for both CLC Bio and Cufflinks that a correlation is found between the computed abundances and the reference counts. While the CLC results also show a regression of 1, the cufflinks abundances show an elevated regression, meaning that Cufflinks estimates a higher readcount than actually present in the data. The HTSeq results do not show a clear correlation between the HTSeq counts and the reference. Furthermore all HTSeq counts are lower than the reference.

For completeness the total readcount among all transcripts is calculated and shown in Table 10. TopHat2 has also been added as the Cufflinks abundances are based on these alignments. Both CLC Bio and Tophat have a total readcount of around 37 million, which is close to the 40 million of the reference. The readcounts of HTSeq and Cufflinks deviate a lot from the reference, however. The low amount of reads counted by HTSeq is presumably caused by the presence of alternative splicing isoforms, since reads which are mappable to multiple transcripts are

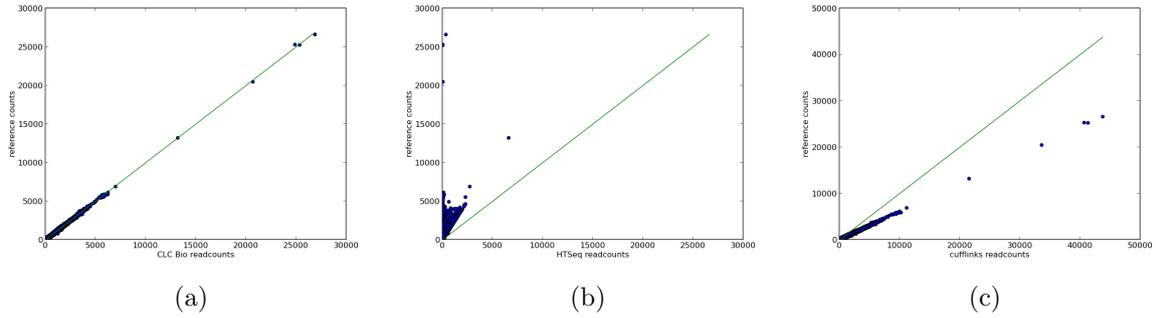


Figure 22: Readcounts calculated by CLC Bio (a), HTSeq (b), and Cufflinks (c). The green diagonal represents a regression of 1.

discarded by HTSeq.

Cufflinks applies several corrections and normalizations on the alignment data, which resulted in a total readcount of roughly 47 million. Figure 23 shows two scatterplots of the coverage calculated by cufflinks (not raw reads) and the CLC Bio transcript abundances. The transcript length is included through the colormap. The longer a transcript is, the lighter the color (blue to red). Figure 23a shows that a certain correlation is found between the two. A regression of 1 (the green line) is of course not found, since the Cufflinks estimates represent the coverage. Figure 23b shows all transcripts with a maximum coverage of 500. A clear correlation is found between the expression level and the transcript length. Shorter transcripts are found to have more coverage than longer transcripts, which is one of the normalizations Cufflinks applies during the analysis. During abundance calculations, RPKM and FPKM values are also calculated for CLC Bio and Cufflinks, respectively.

Abundance estimator	Total reads counted
Reference	39.977.244
CLC Bio	37.035.368
HTSeq	7.594.007
Cufflinks	46.762.191
Tophat	37.649.250

Table 10: Sum of reads counted by CLC Bio, HTSeq, and Cufflinks. The second row shows the total number of reads in the reference. The last row shows the amount of reads found in the Tophat alignment.

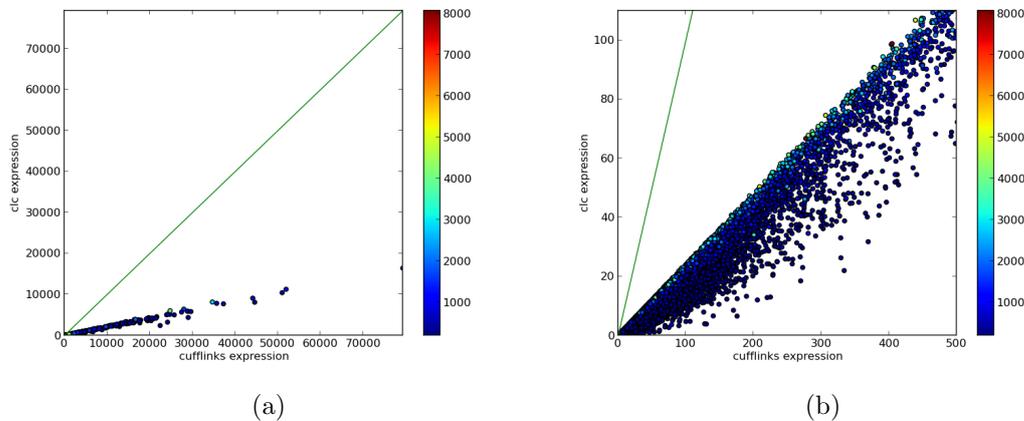


Figure 23: Scatterplots of the readcounts calculated by CLC and Cufflinks.

Differential Expression

After transcript abundances have been estimated, samples can be compared to find differential expressed transcripts. Identifying differential expressed transcripts can be done using several methods. Cuffdiff tests the fold-change of transcripts versus the null hypothesis (that no change is observed), using FPKM values obtained with Cufflinks. CLC Bio offers Kal’s test and Baggerly’s test, depending on the experiment, and uses either RPKM values or readcounts. DESeq first scales the raw data using a model based on the binomial distribution before applying a conditioned test.

To test the three methods for their capability and accuracy of identifying differentially expressed transcripts, 12 transcripts are modified by simulating their expression. Table 11 shows the selected transcripts and their original and modified transcripts readcounts. The transcripts are selected to represent the complete set of transcripts (selection criteria include transcript coverage and length). The only exception made is that only transcripts lacking splicing isoforms are selected to rule out any bias that might be introduced by misaligned reads. Using the same protocol, the modified sample is aligned with Tophat and CLC Bio. Analysis of the alignment results of the unmodified and modified samples show, a part from the modified transcripts, almost identical results. Note that since splicing isoforms often share exons, aligners can not always successfully align reads from these exons to the right transcript. Both CLC Bio and Cufflinks have implemented methods to tackle this problem (e.g. by assigning non-uniquely mapped reads to transcripts, according to the proportions observed in the uniquely mapped reads in CLC Bio).

Transcript ID	transcript length	sample 1 readcount	filter rate	sample 2 readcount
ENSMUST00000000090	690	80	5%	4
ENSMUST00000034226	1.797	193	10%	19
ENSMUST00000067318	545	59	20%	12
ENSMUST00000180279	684	66	50%	33
ENSMUST00000183325	980	123	5%	6
ENSMUST00000183329	981	105	20%	21
ENSMUST00000000001	3.262	371	50%	186
ENSMUST00000000080	4.217	551	40%	220
ENSMUST00000000095	3.626	476	10%	48
ENSMUST00000000122	3.446	450	30%	135
ENSMUST00000000127	3.076	362	40%	145
ENSMUST00000067246	3.178	356	30%	107

Table 11: Modified transcripts in the simulated dataset. two sets of 6 transcripts are manually selected based on their length, coverage, and the absence of alternative splicing isoforms. Reads of these transcripts are filtered according to the percentages in column 3.

Differential Expression in CLC Bio

For a two-sample experiment CLC Bio offers Kal's analysis to identify Differential Expressed genes or transcripts. The test relies on an approximation of the binomial distribution, and considers proportions rather than raw counts. Hence, the test is also suitable in situations where the sum of counts is different between the samples. When applying Kal's test, the 'Proportions difference', 'Fold change', 'Test statistic', and 'P-value' are calculated for each transcript in the samples. The 'Proportions difference' represents the difference between the proportions in sample 1 and the proportions in sample 2. The 'Fold-change' shows how many times bigger the proportions in sample 2 is relative to that of sample 1. The 'Test-statistic' and 'P-value' represent the value of the test statistic and the two-sided p-value for the test. Table 14 shows a selection of the results of Kal's analysis on the two simulated samples, sorted on the Test-statistic (ascending). Comparing these results with the transcript counts of the reference in Table 11 shows that out of the 12 modified transcripts, 11 of those transcripts have identical readcounts. The fold-change, test-statistic and p-values were calculated according these counts. Since the "Test-statistic" is based on the proportions fold-change, one would expect transcripts with the highest filter rate to have the lowest test-statistic (e.g. if only 5% of the reads of a transcript are selected, the fold-change is higher than when 50% of the reads are removed). No clear correlation between the fold-change or test-statistic and the filter rate is observed, however. The p-value indicates the probability that the observed measurement is the result of random chance under the null hypothesis (e.g. that no differential expression is in the samples). For the 11 modified transcripts, CLC Bio calculated p-values varying between 2,44E-8 and 0, which is lower than the often used thresholds of 0,05 or 0,01 to indicate whether an observation is significant. For unmodified transcripts however, p-values as low as 0.00157 are also observed.

Transcript ID	sample 1	sample 2	Proportions Fold-change	Test-statistic	P-value	Filter ratio
ENSMUST00000000095	476	48	-9,92	-18,7	1,11E-16	10%
ENSMUST00000000122	450	135	-3,33	-13,02	0,0	30%
ENSMUST00000034226	193	19	-10,16	-11,95	0,0	10%
ENSMUST00000000080	551	220	-2,5	-11,92	0,0	40%
ENSMUST00000067246	356	107	-3,33	-11,57	0,0	30%
ENSMUST00000183325	123	6	-20,5	-10,3	0,0	5%
ENSMUST00000000127	362	145	-2,5	-9,64	4,44E-16	40%
ENSMUST00000000090	80	4	-20,0	-8,29	5,55E-15	5%
ENSMUST00000000001	371	186	-1,99	-7,84	3,77E-15	50%
ENSMUST00000183329	105	21	-5,0	-7,48	7,26E-14	20%
ENSMUST00000067318	59	12	-4,92	-5,58	2,44E-8	20%
ENSMUST00000146424	30	10	-3,0	-3,16	1,57E-3	-

Table 12: Differential Expression analysis in CLC Bio using Kal's analysis. Results are sorted on the 'Test statistic', afterwhich the first 12 rows are selected. Column 2 and 3 show the transcript readcounts in sample 2 and 3, respectively. Column 4-7 show the results of Kal's analysis, and column 8 shows the filter rate applied to the 12 transcripts in sample 2.

Differential Expression in Cuffdiff

Cuffdiff tests for differential expressed transcripts by testing the observed log-fold-change in its expression against the null hypothesis of no change. Cuffdiff outputs the transcript coverage, 'log2 fold-change', 'test-statistic', 'p-value', and whether the observed change is significant. Table 13 shows a selection of the differential expression analysis of Cuffdiff. Comparing these results with the modified transcripts in Table 11 shows that 8 of the modified transcripts are present in the p-value sorted Cufflinks results (the top hits of the test-statistic sorted results included only 5 modified transcripts). Since Cufflinks calculates transcripts coverage rather than readcounts, these cannot be compared directly to the readcounts in Table 11. To be able to compare the expression values the log2 fold-change of the readcounts of the reference is calculated and shown in Table 13. Comparing these numbers shows that the calculated fold-change is very accurate for all but one transcript. For all eight transcripts, a p-value of 5E-05 is calculated, which suggests that the expression difference is significant. The Significance result of Cufflinks shows that all differential expression found is not significant, however. To emphasize the coverage accuracy, for unmodified transcripts a fold-change of 0 is found, indicating equal coverage for both samples (which results in a p-value of 1).

Transcript ID	sample 1	sample 2	fold-change (Cuff.)	log2 fold-change (ref.)	test-stat.	P-value	Significance
ENSMUST00000183329	7,72031	1,32571	-2,54189	-2.32193	-9,64136	5E-05	no
ENSMUST00000067246	6,58071	1,97791	-1,73427	-1.73427	-8,84006	5E-05	no
ENSMUST00000000080	7,4847	2,98845	-1,32455	-1.32455	-8,17348	5E-05	no
ENSMUST000000000127	6,93736	2,77877	-1,31994	-1.31994	-7,18799	5E-05	no
ENSMUST000000000001	6,6636	3,34078	-0,996116	-0.99612	-5,62523	5E-05	no
ENSMUST000000000095	7,61431	0,76829	-3,30986	-3.30986	-16,4253	5E-05	no
ENSMUST000000000122	7,61005	2,28302	-1,73697	-1.73697	-13,3067	5E-05	no
ENSMUST00000034226	6,85535	0,674879	-3,34453	-3.34453	-11,5326	5E-05	no
ENSMUST00000000033	0,140506	0,140506	6,64E-08	-	0	1	no
...						no	
ENSMUST000000000163	0,0685691	0,0685691	1,46E-07	-	0	1	no

Table 13: Differential Expression analysis with Cuffdiff. Results are sorted on the P-values. Column 2 and 3 show the transcript coverage in sample 1 and 2, respectively. Column 4 and 5 show the log2 fold-change of the expression values calculated by Cufflinks and those in the reference, respectively (log2 of sample 2 divided by sample 1). Column 6 and 7 show the test-statistic and the P-values. The results are sorted on the p-values.

Differential Expression in DESeq

DESeq tests for differentially expressed genes by calculating a scaling factor to normalize the raw counts. The model is based on a binomial distribution. Note that a raw count table is required, and that no normalizations should have been applied to these counts. The CLC Bio count table is used as input for DESeq as it is also possible in CLC Bio to output readcounts while correcting for non-uniquely mapped reads (by assigning these reads to transcripts according to the proportions calculated from uniquely mappable reads) without other normalizations. Comparing the DESeq results with the reference readcounts in Table 11 shows that all readcounts are exactly twice the reference readcounts. This is caused by the fact that, for paired-end samples, CLC Bio counts a pair as one, while DESeq considers its input as raw counts. Since DESeq uses proportions for its calculations, results are not influenced, however. The fold-change reported by DESeq is compared to the fold-change of the reference readcounts. The fold-changes are identical for 11 (like the CLC Bio analysis) transcripts. The p-values indicate that the differential expression observed is significant for these 11 transcripts.

DESeq is also run with the Tophat alignment, but no correction is done for non-uniquely mapped reads. Results are not shown here, but are comparable, but less accurate to the DESeq results on the CLC Bio readcounts.

Transcript ID	sample 1	sample 2	log2 fold-change (DESeq)	log2 fold-change (reference)	P-value
ENSMUST00000000095	952	96	9.917	9.917	1.46E-79
ENSMUST00000034226	386	38	10.158	10.158	3.64E-34
ENSMUST00000000122	900	270	3.33	3.333	6.33E-33
ENSMUST00000183325	246	12	20.5	20.5	1.91E-28
ENSMUST00000067246	712	214	3.327	3.327	1.56E-26
ENSMUST00000000080	1102	440	2.504	2.505	1.87E-26
ENSMUST00000000090	160	8	20	20	7.87E-19
ENSMUST00000000127	724	288	2.514	2.497	2.49E-18
ENSMUST00000183329	206	40	5.15	5	6.3E-13
ENSMUST00000000001	742	372	1.995	1.995	3.64E-12
ENSMUST00000067318	118	24	4.917	4.917	1.07E-7

Table 14: Differential Expression analysis with DESeq. Column 1 and 2 represent the readcounts as obtained from the Tophat alignment. Column 3 and 4 represent the fold-change and p-value as calculated by DESeq. A selection of the p-value sorted results is shown.

Discussion

de novo Transcriptome Assembly

Data

The studied data was selected based on its characteristics to resemble those commonly generated for this type of research. Because of this, and to decrease computation time, relative small datasets were selected. A direct consequence of this, and also what one would expect is that a significant part of the transcripts does not have sufficient coverage, and as a result, the assembly will consist of a lot of fragmented transcripts. After running fastQC (quality evaluation tool) on the data, a rather high duplication rate was observed, and since it is not clear what the consequences of high duplication rates are regarding the assembly accuracy, multiple assemblies have been made with and without technical duplicates.

de novo Assembly

When comparing two or more assemblies generated from different assemblers, basic statistics such as the number of contigs, average contig length, and the N50 already give a vague idea on which assembler performed better. In most cases, the better assembly consists of a larger number of contigs in combination with a higher average contig length. These numbers do not linearly relate to the assemblies accuracy. For example the assembly of transcripts with several splicing isoforms can be challenging and may result in incorrectly assembled transcripts.

While the differences in the normal assembly and the assembly without duplicates made with Trinity are negligible, the CLC assemblies show significant differences, leaving a lot of room for discussion. While the increase in average transcript length is of great importance, the loss in total number of transcripts might indicate that only the short transcripts were not reconstructed, giving a 'false' boost to the average length. The fact that the loss in basepairs divided by the decrease in transcripts ($1,010,774/7499 = 135$) suggests that the average length increased only by 95 ($[988-758]-135 = 95$) instead of 230 ($988-758$). When taking into account that the minimum length also increased by 60 ($249-189$), almost no increase in read length is found, while 30% less transcripts are reconstructed. Note that these are only quick calculations which might show that duplication removal is not as good as these numbers suggest. More research is required to obtain proof of whether duplication removal leads to better results. This however, is not the scope of this project.

Reference-guided Assembly Evaluation

A popular method to assess *de novo* transcriptome assemblies is by aligning raw reads and reconstructed transcripts to the reference transcriptome. While the alignment rates of Bowtie2 are consistently lower than those of GMAP, alignments of assemblies to the reference are significantly lower than the difference between the other alignments. This suggests that Bowtie2 is less sensitive when it comes to aligning longer sequences. Another point of interest is the difference between the two datasets. All alignment rates, for Bowtie2 and GMAP, are higher for the mouse dataset. While both the tomato and mouse are model organisms, the mouse has been researched far more, resulting in a more complete reference transcriptome.

From the multiple alignments reported by GMAP, the hypothesis that smaller reconstructed transcripts align more often to multiple transcripts in the reference was set. However, no significant difference in length was found between transcripts that align one or multiple times. The transcripts to which a reconstructed transcript aligned often come from a single gene, suggesting that the assembler is not always able to make a clear distinction between splicing isoforms, resulting in fragmented contigs.

the observation that a significant part of the reconstructed transcripts is fragmented is enforced by two of the alignment results. 1) from the distribution showing the overlap between the reconstructed and reference transcripts can be seen that a large part of the transcripts cover only 10-40%. 2) In 30-50% of the cases more than one reconstructed transcript is mapped to the reference transcript, vouching for a certain rate of fragmented transcripts.

Core genes

finally, the methods to evaluate a transcriptome without the use a reference transcriptome/genome will be discussed. STAR and Tophat2, which use statistical models to find splice sites, were tested on both the assemblies and the assemblies in combination with the raw data. Even with less stringent parameters hardly any splice sites were identified.

Using the local alignment option in Bowtie2 transcripts were aligned to the core genes. Since the local alignment option was used a large number of hits was expected. For the tomato only 82 hits were found however. Apart from the doubtful results, one of the issues is that it is impossible to calculate the percentage of the core gene that is found (due to the presence of introns). Concurrently GMAP was used to align the reconstructed transcripts to the core genes (most occurring transcript). In all cases over 95% of the core genes were found with an overlap of 90-100%. The other transcripts found are short (fragmented) transcripts which, most likely, are not completely reconstructed splicing isoforms.

Expression Analysis

Read Alignment

Following standard expression analysis protocols the simulated reads are aligned to the transcriptome using both Tophat2 and CLC Bio. Given that the simulated reads are a subset of the transcriptome, and no sequencing errors or indels are introduced, one would expect all reads to align to the reference. Although this is true for both aligners (99.9% and 100% proper pair alignments for Tophat2 and CLC Bio, respectively), solely 60% of the reads aligned to the correct transcript. This may seem low, but as mentioned in the introduction, most transcripts have one or multiple splicing isoforms, thus leading to ambiguous alignments. For example if one gene is transcribed and translated into two different transcripts through alternative splicing, these transcripts will most likely share one or multiple exons. Given that the NGS reads are relatively short compared to the full transcripts and as such often do not comprise splice junctions, for a significant portion of the reads it is impossible to trace back from which transcript they originated. The number of reads that is mapped to an incorrect transcript is thus related to the number of shared exons between splicing isoforms and the total number of isoforms. This hypothesis is enforced by the fact that all reads that did not align to the correct transcript, aligned to a transcript isoform of the same gene. These reads are characterized by the fact that they align to multiple transcripts. In the next paragraph the impact of these incorrect mappings will be discussed.

Transcript Abundance Estimation

Transcript expression can be estimated using different methods though all of them are based on the number of short reads that align to a transcript. Since the read alignment is the basis of all expression analysis tools, it is needless to point out how important an accurate alignment is. From the alignment results it was shown that a significant part of the reads align to splicing isoforms, which might significantly alter the expression analysis results, depending on the normalization procedure. Note that this is only the case when performing such an analysis on transcript level. On a gene level, reads of all transcript isoforms for a single gene are summed. Both CLC Bio and Cufflinks have implemented methods to deal with these uncorrectly (and non-uniquely) mapped reads. HTSeq does not account for this, and ignores all reads that are non-uniquely mapped. Whether a causal relation exist between the varying abundance estimation results and the manner in which non-uniquely mapped reads are handled is hard to determine, nonetheless results of both tools which implemented methods to deal with these reads show a significant correlation with the reference, while HTSeq does not, implying that one should not just remove these reads from the analysis. Especially the transcript abundances found by CLC Bio look very accurate, as a correlation of roughly 1 is found with the reference. The Cufflinks results are also very interesting in that a correlation of 1 is also observed, but that the regression is drawn toward the cufflinks expression values. In other words, Cufflinks overestimated all transcripts while preserving the ratios. In summary: the reference contains 40 million reads, CLC Bio and the Tophat alignment contain around 37 million reads. Cufflinks (which used the Tophat alignment) contains almost 47 million reads, and HTSeq contains 7.5 million reads. These numbers show that the cufflinks algorithm 'adds' around 10 million reads to the generated dataset, since the Tophat alignment contains 37 million reads, which is close to the amount of reads in the reference data. The addition of this many reads is most likely explained through the normalizations implemented by Cufflinks. One of these is easily observed in the data; During sequencing short transcripts might be sequenced less often than longer transcripts, which would then require a normalization in order to perform valid down-stream analyses. The results show that the normalization is performed (short transcripts are found to have higher readcounts). It is not so easy, if possible at all, to validate these results, however. Since it is not known when, and how much less short transcripts are sequenced.

Differential Expression Testing

Testing for differential expressed features can be performed on two levels: gene level and transcript level. While testing on transcript level yields more information, it is also more complex, and depending on the research, not always favourable. Analyses on transcript level yield more information as often a biological process, or disease is altered or caused by the change in expression of one of many transcript isoforms. Such information can not be gain from gene level transcript expression analyses. One of the aspects making the analysis on transcript level more complex is one already observed during alignment. Since alternatively spliced transcripts share exons, reads from these exons cannot always be correctly mapped back. In contrast, when performing the analysis on a gene level, all reads from alternatively spliced transcripts can be summed up, for each gene.

To test CLC Bio, Cuffdiff, and DESeq the following experiment is set-up: 12 transcripts are manually selected, and their readcount is modified by randomly removing reads from the raw data file. Selection is done to obtain a group of transcripts which is representable for the complete sample. Read removal is done according to predetermined ratios (deletion of 50-95% of the reads). In general, one would not only expect the three tools to identify the 12 modified transcripts as differential expressed, but also to make a distinction between the different ratios used for removal of the reads (e.g. a transcript which had 50% of its reads removed should be less differentially expressed than a transcript which had 95% of its reads removed).

Starting with the evaluation of CLC Bio, from the abundance estimation results one would expect the results of CLC Bio to satisfy above expectations (this of course applies for all tools, but especially for CLC Bio, since the best expression counts are observed for CLC bio). The top 20 differentially expressed transcripts identified by CLC Bio contains 11 modified transcripts. No relation is found between the order in which CLC Bio found transcripts to be differential expressed and the expected order based on the percentage of reads that has been removed, however. Finally, taking a closer look at the fold-change, these match almost completely with the filter ratio of the reference set (e.g. a fold-change of -9,92 matches quite good with a filter ratio of 10%).

For Cuffdiff basically the same expectations apply, since the overestimation should not interfere with the differential expression tests. In the top 20 hits, Cuffdiff identified 8 of the modified transcripts as differentially expressed when sorted on the p-values (sorting on test-statistic resulted in only 5/12 hits). Again no relation is found between the read removal ratio and the order in which Cuffdiff found transcripts to be differentially expressed. That only 8 of the modified transcripts are identified is unexpected, but that all of them are found to be not significant is even more striking, especially when taking into account the corresponding p-values ($5E-05$ for 7/8 transcripts).

DESeq is a stand-alone tool used for testing differential expression in RNA-Seq experiments. Since DESeq has no method implemented for non-uniquely mapped reads, in order to make a fair comparison the CLC Bio alignment is used as a basis for the raw count table required by DESeq. DESeq results are very much comparable to those of CLC Bio; again 11 out of 12 of the modified transcripts are identified, and the results are in the same order (eventhough deviating from the expected order).

while these results show that different methods are capable of identifying differentially expressed transcripts to a greater or less extend, one major question remains: When is the observed differential expression significant, and when is the result random or caused by for example noise? Usually the p-value in combination with a threshold (of e.g. 0.1 and 0.05) is used to assess whether an observation is significant or not. It is not always possible to correctly distinguish differentially expressed transcripts from non-differentially expressed transcripts based on just the p-value. Even in two samples where only 12 transcripts have different expression values, much more transcripts with a low p-value are observed.

Conclusion

De novo Transcriptome Assembly

de novo Assembly

As discussed in the previous chapter, metrics to determine the quality of genome assemblies such as contig lengths and N50 are insufficient to assess the quality of transcriptome assemblies. These statistics can be used, but are certainly not sufficient. Especially when a (rough) estimate of these numbers is available for the studied transcriptome, the number of contigs and their length is informative. It should be noted that when making an estimation of these numbers (e.g. from a closely related species), information regarding the origin of the sample should be included. The mouse dataset is a good example of that the number of transcripts in the assembly and the reference can differ by factor 10. This does not imply that the assembly went wrong, but rather that the sample does not contain all those transcripts. In practice this is less of an issue however, since most experience employ total RNA samples.

Knowing this, one can compare these statistics for different assemblers and already get an idea of how well these performed. For both datasets Trinity showed to be superior in the number of contigs produced. The rest of the statistics show no significant differences.

When comparing the assemblies from the filtered (duplication removal) data, Trinity shows almost no difference, leading to the conclusion that duplication removal does not influence the assembly, while the computation time is decreased. The CLC assemblies however do show great variation. As already discussed more research is required before a sound conclusion can be drawn from these numbers.

Reference-guided Assembly Evaluation

The consistent alignment results clearly show the the difference in algorithms used by the aligners. Bowtie2 and CLC perform similar when aligning raw (short) reads. When aligning longer (transcripts) reads however, Bowtie2 performs significantly worse than both GMAP and CLC. GMAP is found to have the best overall alignment rate, most likely due to the implementation such as SNP and sequence error handling.

The alignment rates show that one can always expect 90-95% of the raw reads to map back to the assembly. From the alignment of the assembled transcripts to the reference transcriptome one would expect these rates to be similar to the alignment of the raw reads to reference transcriptome, given that a significant part of the raw reads is covered by the assembly. this is not the case for both datasets, however. These numbers give an indication of the quality of the sample and the correctness of the reconstructed transcripts but cannot be used to draw conclusions regarding the transcript fragmentation.

The multiple alignment information shows that, because the reference transcripts (to which a reconstructed transcript aligned) come from a single gene, alternative splicing isoforms are not always correctly retrieved. No solid correlation is found between the alignment rate (raw to assembly) and the transcript fragmentation (whether a transcript is fully reconstructed or not). While a certain part of the transcripts were fully reconstructed, a large part of the transcripts were only retrieved for 10-40%. This observation is supported by the contig analysis (number of

contigs that align to a reference transcript). The analysis showed that 40-50% of the transcripts are covered by more than 1 reconstructed transcript.

Core Genes

Finding core genes in the assemblies has been done in multiple ways. STAR and Tophat2 proved not to be suited for this task, as they require large amounts of data to successfully predict splice sites (using statistical methods). Alignments made against the dna sequences of the core genes were not optimal, and could not be used as an accurate method to measure assembly quality. Finally GMAP was used to align the transcripts to the mRNA sequences of the core genes, and successfully 'found' 95% of the core transcripts with 90-100% overlap. GMAP is capable of aligning transcripts to mRNA sequences by searching for splice junctions between exons, since transcripts do not always contain all exons.

Expression Analysis

Read Alignment

As already pointed out in the discussion, the results of the alignment of the simulated reads against the reference transcriptome are as expected. Since no insertions, deletions, and sequencing errors are introduced, the only complex aspect of the alignment are the alternatively spliced isoforms. While these results clearly show that aligners are capable of "correctly" aligning millions of reads, it should be stressed that about 60% of these have 100% identity with multiple transcripts. As a consequence transcript abundance estimations are largely affected by the way alignment routines (randomly) place the reads on an alternative transcript, which directly affects the transcript abundance calculations, if not for corrected.

Transcript Abundance Estimation

From the discussion of the transcript abundance estimation follows that two critical aspects in estimating transcript abundances can be identified. Firstly, the manner in which non-uniquely mapped reads are handled, and secondly, which normalizations are performed, and on what basis.

Starting with how non-uniquely mapped reads are handled, basically three options are available: deletion of all non-uniquely mapped reads, which ensures that no faulty data is included in the analysis, with the drawback of losing a lot of information. The second option is to ignore the fact that they align on multiple transcripts, and continue with the downstream analysis. Finally, one could include all reads, and distribute the non-uniquely mapped reads to transcripts according to some method (as in CLC Bio and Cufflinks). When no such method is available, intuitively one would rather perform an expression analysis without non-uniquely mapped reads than to almost certainly introduce errors by including those reads. When a robust method to distribute these reads is available however, using it is almost always favourable above the other two options.

The second critical point is that many normalizations are available, of which some have been shown to improve the analysis results, but others might only improve the results for certain library preparation methods or sequencing mechanisms. As stated in the introduction, no golden standard is available (except for qPCR verification for a small number of transcripts), and each expression analysis should be performed on well thought through (normalized) samples. One of the normalizations that is necessary in order to make a valid comparison is one that normalizes samples for the total amount of reads in the samples. CLC bio, Cufflinks, and DESeq apply such normalization before the differential expression tests are performed. when comparing different genes or transcript with each other, such a normalization is also required for the readlength.

The RPKM and FPKM normalizations in CLC Bio and Cufflinks, respectively, combine these two.

For the normalizations that have been shown to be a requirement to get meaningful results, several variations exist. For example, normalizing samples for the total number of reads in a sample (sequencing depth) can be done using multiple methods (RPKM, FPKM, and by estimating the total library size, in CLC Bio, Cufflinks, and DESeq, respectively).

Differential Expression Testing

Performing the differential expression analysis and significance test on two or more samples is relatively straightforward, the complexity rather lies in the evaluation and interpretation of the results. The main challenge here is to distinguish real differentially expressed transcripts from transcripts that are not significantly differentially expressed. Whereas the test-statistic and p-value are calculated solely for this purpose, in practice it is still challenging to distinguish significant differentially expressed transcripts from non differentially expressed transcripts, which do not have exactly the same expression values. The challenging aspect here is that transcript expression cannot be measured as "on" or "off", but rather on a continuous scale, which will often result in the question whether the expression difference for transcripts with a p-value close to the threshold is significant or not. First, the threshold has to be defined, however. Often thresholds of 0.01 and 0.05 are used, but the analysis' results also shows transcripts with p-values lower than 0.01 for numerous transcripts that are not modified. While this is explained through the already discussed misassignment of reads due to alternative splicing, this makes it very hard to distinguish the real differentially expressed transcripts from the transcripts that only "look" differentially expressed. A possible solution would be to perform the differential expression test on the unique aligned reads, instead of the total amount of reads. This would introduce other complications, however. For example if splicing isoforms (A and B) have 3 and 2 exons, respectively, where both exons of transcript B are also present in transcript A. When counting uniquely aligned reads, transcript B will also have a count of zero, since these reads also align on transcript A (given that the exons are also in the same order). Do note that this situation is however also not well solved by using the total number of reads, when distributing the reads based on the proportions of the unique aligned reads. Another example is when a gene is present multiple times in a sample. all reads from this gene will not align uniquely, resulting in the removal of the complete gene from the analysis.

To summarize the above: In order to successfully perform a differential expression analysis several normalizations or corrections are available, of which some more essential than others, depending on the sample and experiment design. In the end it always comes down to the question whether the observed change in expression is significant or not, and while the calculated test-statistics and p-value should give the means to answer this question, results have shown that this is not always as accurate as one would expect. Each analysis should be performed while carefully interpreting intermediate results in order to obtain valid expression profiles.

All in all, next generation sequencing positively contributed to new methods for transcriptome profiling by introducing methods to make expression analyses faster, cheaper, and more complete. Nonetheless one should always validate the results through traditional methods such as qPCR.

References

References

- [1] Nicolae M, Mangul S, Mndoiu I I, Zelikovsky A, 2011. *Estimation of alternative splicing isoform frequencies from RNA-Seq data.*
Algorithms for Molecular Biology 2011, 6:9
- [2] Chen M, Manley J L, 2009. *Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches.*
Nature Reviews Molecular Cell Biology 10, 741-754
- [3] Black D L, 2003. *Mechanisms of alternative pre-messenger RNA splicing.*
Annual Review Biochemistry 71, 291-336
- [4] Parchter L, 2011. *Models for transcript quantification from RNA-Seq.*
arXiv:1104.3889 [q-bio.GN]
- [5] Ozsolak F, Platt A R, Jones D R, Reifenger J G, Sass L E, McInerney P, Thompson J F, Bowers J, Jarosz M, Milos P M, 2009. *Direct RNA sequencing*
Nature, Vol 461|8 October 2009| doi:10.1038/nature08390
- [6] Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley D R, Pimentel H, Salzberg S L, Rinn J L, Pachter L, 2012. *Differential gene and transcript expression analysis of RNA-seq experiments with tophat and cufflinks.*
Nature Protocols 2012 Mar 1;7(3), 562-78
- [7] Grant G R, Farkas M H, Pizarro A D, Lahens N F, Schug J, Brunk B P, Stoeckert C J, Hogenesch J B, Pierce E A, 2011. *Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM).*
Bioinformatics (Oxford) 2011 Sep 15;27(18), 2518-28
- [8] Zhao et al., 2011. *Optimizing de novo transcriptome assembly from short-read RNA-Seq data: a comparative study.*
BMC Bioinformatics 2011, 12 (suppl 14), S2
- [9] Grabherr M G, Guttman M, Trapnell C, 2011. *Computational methods for transcriptome annotation and quantification using RNA-seq*
Nature Methods 8, 269-477
- [10] Sharp P A, 2009. *The Centrality of RNA.*
Cell Volume 136 Issue 4, 577-580
- [11] Cooper T A, Wan L, Dreyfuss G, 2009. *RNA and Disease.*
Cell Volume 136 Issue 4, 777-793
- [12] Wang Z, Gerstein M, Snyder M, 2009. *RNA-Seq: a revolutionary tool for transcriptomics*
Nature Reviews Genetics 10, 57-63 (January 2009) | doi:10.1038/nrg2484
- [13] Wilhelm B T, Landry J R, 2009. *RNA-Seq quantitative measurement of expression through massively parallel RNA-sequencing*
Elsevier Methods 48 (2009) 249257
- [14] Grabherr M G, Haas B J, Yassour M, Levin J Z, Thompson D A, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren B W, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A, 2011.

- Full-length transcriptome assembly from RNA-seq data without a reference genome.*
Nature Biotechnology. 2011 May 15;29(7), 644-52
- [15] Licatalosi D D, Darnell R B, 2010. *Resolving RNA complexity to decipher regulatory rules governing biological networks.*
Nature Rev Genet. 2010 January ; 11(1), 7587
- [16] Trapnell C, Hendrickson D G, Sauvageau M, Goff L, Rinn J L, Pachter L, 2013. *Differential analysis of gene regulation at transcript resolution with RNA-seq*
Nature Biotechnology 31, 4653 (2013) doi:10.1038/nbt.2450
- [17] Roberts A, Trapnell C, Donaghey J, Rinn J L, Pachter L, 2011. *Improving RNA-Seq expression estimates by correcting for fragment bias*
Genome Biology 2011, 12:R22 doi:10.1186/gb-2011-12-3-r22
- [18] Roberts A, Pimentel H, Trapnell C, Pachter L, 2011. *Identification of novel transcripts in annotated genomes using RNA-Seq*
Bioinformatics (2011) doi: 10.1093/bioinformatics/btr355
- [19] Birol I, Jackman S D, Nielsen C B, Qian J Q, Varhol R, Stazyk G, Morin R D, Zhao Y, Hirst M, Schein J E, Horsman D E, Connors J M, Gascoyne R D, Marra M A, Jones S J M, 2009. *De novo transcriptome assembly with ABySS*
Bioinformatics (2009) Vol. 25 no. 21 2009, pages 28722877
doi:10.1093/bioinformatics/btp367
- [20] Robertson G, Schein J, Chiu R, Corbett R, Field M, Jackman SD, Mungall K, Lee S, Okada HM, Qian JQ, Griffith M, Raymond A, Thiessen N, Cezard T, Butterfield YS, Newsome R, Chan SK, She R, Varhol R, Kamoh B, Prabhu AL, Tam A, Zhao Y, Moore RA, Hirst M, Marra MA, Jones SJ, Hoodless PA, Birol I, 2010. *De novo assembly and analysis of RNA-seq data*
Nature Methods. 2010 Nov;7(11):909-12. doi: 10.1038/nmeth.1517. Epub 2010 Oct 10.
- [22] Martin J, Bruno V M, Fang Z, Meng X, Blow M, Zhang T, Sherlock G, Snyder M, Wang Z, 2010. *Rnnotator: an automated de novo transcriptome assembly pipeline from stranded RNA-Seq reads*
BMC Genomics 2010, 11:663 doi:10.1186/1471-2164-11-663
- [23] Zhao Q Y, Wang Y, Kong Y, Luo D, Li X, Hao P, 2011. *Optimizing de novo transcriptome assembly from short-read RNA-Seq data: a comparative study*
BMC Bioinformatics 2011, 12(Suppl 14):S2 doi:10.1186/1471-2105-12-S14-S2
- [24] Surget-Groba Y, Montoya-Burgos J I, 2010. *Optimization of de novo transcriptome assembly from next-generation sequencing data*
Genome Res. 2010 20: 1432-1440
- [25] Martin J A, Wang Z, 2011. *Next-generation transcriptome assembly*
Nature Reviews Genetics 12, 671-682 (October 2011) | doi:10.1038/nrg3068
- [26] BingXin L, ZhenBing Z, TieLiu S, 2013. *Comparative study of de novo assembly and genome-guided assembly strategies for transcriptome reconstruction based on RNA-Seq*
Sci China Life Sci. 2013 Feb;56(2):143-55. doi: 10.1007/s11427-013-4442-z. Epub 2013 Feb 8.

- [27] Mitra S, Rupek P, Richter D C, Urich T, Gilbert J A, Meyer F, Wilke A, Huson D H, 2011. *Functional analysis of metagenomes and metatranscriptomes using SEED and KEGG*
BMC Bioinformatics. 2011 Feb 15; 12 Suppl 1:S21
- [28] Parra G, Bradnam K, Korf I, 2007. *CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes*
Bioinformatics (2007) 23 (9): 1061-1067
- [29] Tatusov, R.L. et al. 2003. *The COG database: an updated version includes eukaryotes.*
BMC Bioinformatics, 4, 41.
- [30] Wu T D, Watanabe C K, 2005. *GMAP: a genomic mapping and alignment program for mRNA and EST sequences*
Bioinformatics, Vol. 21 no. 9 2005, pages 1859-1875
- [31] CLC Genomics Workbench. *Application note: De novo assembly of paired-end plant transcriptome data*
http://www.clcbio.com/files/appnotes/CLC_bio_De_novo_Assembly.pdf
- [32] Trapnell C, Pachter L, Salzberg S L, 2009. *TopHat: discovering splice junctions with RNA-Seq*
Bioinformatics Vol. 25 no. 9 2009, pages 1105-1111,
doi:10.1093/bioinformatics/btp120
- [33] Dobin A, Davis C A, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M and Gingeras T R, 2012. *STAR: ultrafast universal RNA-seq aligner*
Bioinformatics (2012) doi: 10.1093/bioinformatics/bts635
- [34] CLC Bio *White paper on de novo assembly in CLC Assembly Cell 4.0*
February 6, 2012
- [35] Tariq M A, Kim H J, Jejelowo O, Pourmand N, 2011 *Whole-transcriptome RNAseq analysis from minute amount of total RNA*
Nucleic Acids Res. 2011 October; 39(18): e120.
- [36] Langmead B, Salzberg S L, 2012. *Fast gapped-read alignment with Bowtie 2*
Nature Methods 9, 357-359 (2012)
- [37] <http://substansigenetika.blogspot.nl/2010/04/sintesis-protein.html>
- [38] http://en.wikipedia.org/wiki/DNA_sequencing#Next-generation_methods
- [39] Schulz M H, Zerbino D R, Vingron M, Birney E, 2012. *Oases: Robust de novo RNA-seq assembly across the dynamic range of expression levels*
Bioinformatics Advance Access published February 24, 2012
- [40] Dillies M, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, Keime C, Marot G, Castel D, Estelle J, Guernec G, Jagla B, Jouneau L, Lalo D, Le Gall C, Schaffer B, Le Crom S, Guedj M, Jaffrzic F, 2012. *A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis*
Brief Bioinform (2012) doi: 10.1093/bib/bbs046

- [41] Rapaport F, Khanin R, Liang Y, Krek Z, Zumbo P, Mason C E, Socci N D, Betel D, 2013. *Comprehensive evaluation of differential expression analysis methods for RNA-seq data*
arXiv:1301.5277 [q-bio.GN]
- [42] Anders S, Huber W *Differential expression analysis for sequence count data*
Genome Biology 2010, 11:R106 doi:10.1186/gb-2010-11-10-r106
- [43] Anders S, 2010 *HTSeq: Analysing high-throughput sequencing data with Python*.
<http://www-huber.embl.de/users/anders/HTSeq/doc/overview.html>
- [44] Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg S L, 2013. *TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions*
Genome Biology 2013, 14:R36
- [45] Li W *RNASeqReadSimulator: A Simple RNA-Seq Read Simulator*
<http://alumni.cs.ucr.edu/liw/rnaseqreadsimulator.html>
- [46] Sonesson C, Delorenzi M, 2013. *A comparison of methods for differential expression analysis of RNA-seq data*
BMC Bioinformatics 2013, 14:91
- [47] Miller J R, Koren S, Sutton G, 2010. *Assembly algorithms for next-generation sequencing data*
Genomics Volume 95, Issue 6, June 2010, Pages 315327
- [48] Riesgo A, Andrade S C S, Sharma P P, Novo M, Prez-Porro A R, Vahtera V, Gonzalez V L, Kawauchi G L, Giribet G, 2012. *Comparative description of ten transcriptomes of newly sequenced invertebrates and efficiency estimation of genomic sampling in non-model taxa*
Frontiers in Zoology 2012, 9:33 doi:10.1186/1742-9994-9-33

Supplementary Results

Assembly Evaluation Results

Multiple Aligned Transcripts

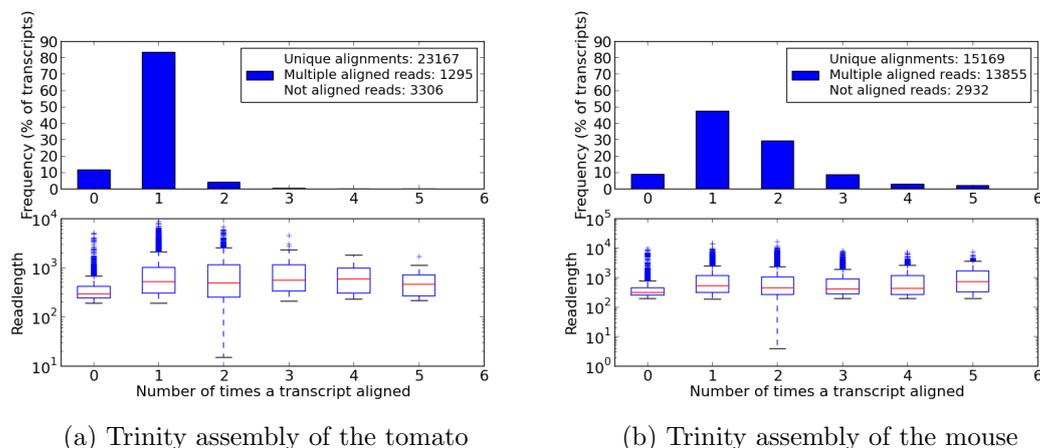


Figure 24: Distributions of the multiple alignments in the tomato and mouse dataset for the assemblies made with CLC (upper half), and readlength statistics (boxplots) for each set of multiple aligned transcripts (lower half). Aligner used: GMAP.

Tomato Assembly Results for Bowtie2 Alignments

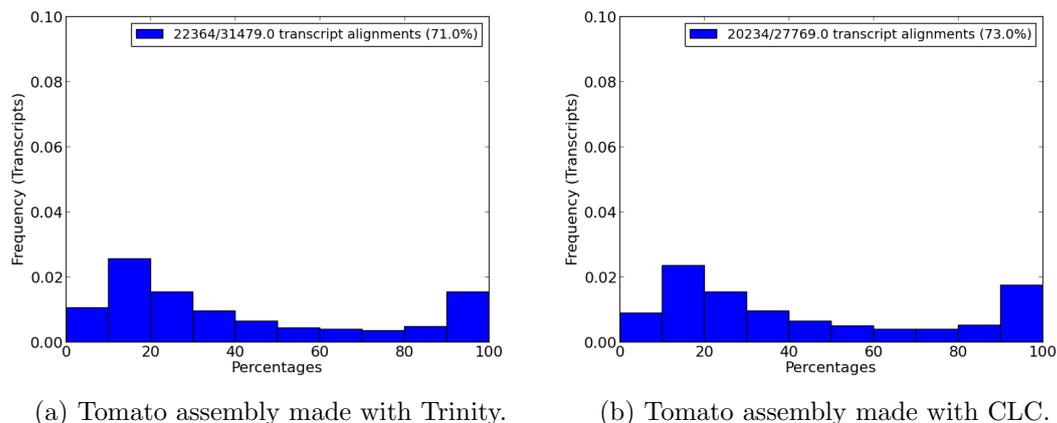
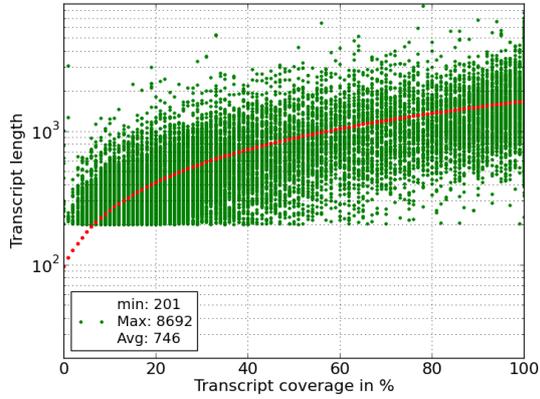
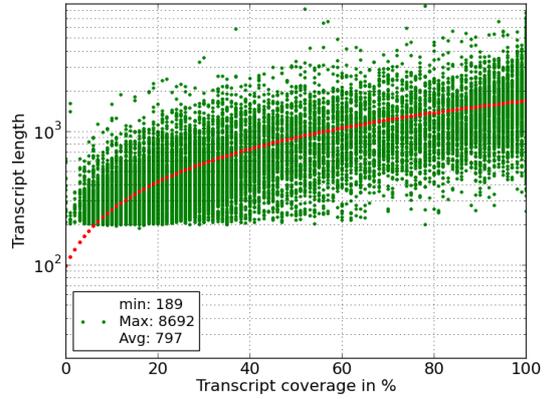


Figure 25: distribution of the normalized overlap of the transcripts assembled by Trinity and CLC and the transcripts in the reference. Alignments were performed with Bowtie2 for these figures.

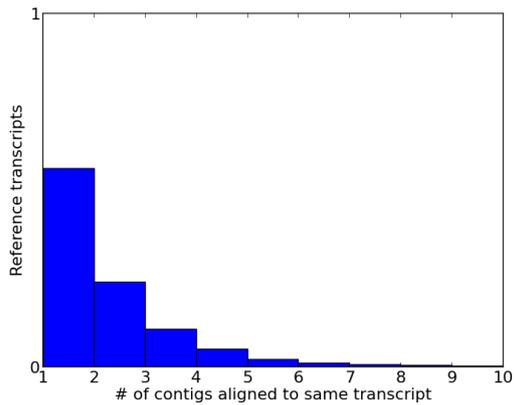


(a) Tomato assembly made with Trinity.

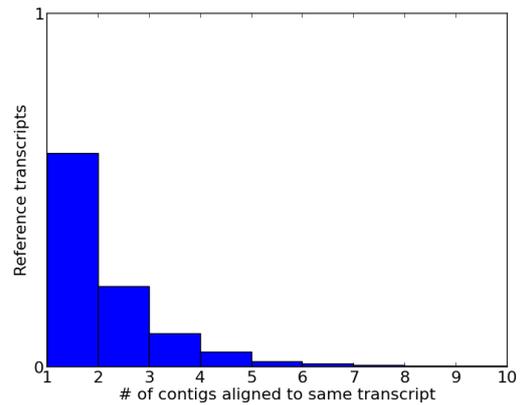


(b) Tomato assembly made with CLC.

Figure 26: per transcript (assembled by Trinity), overlap with the reference transcript plotted against its length(in %). Aligned with Bowtie2.



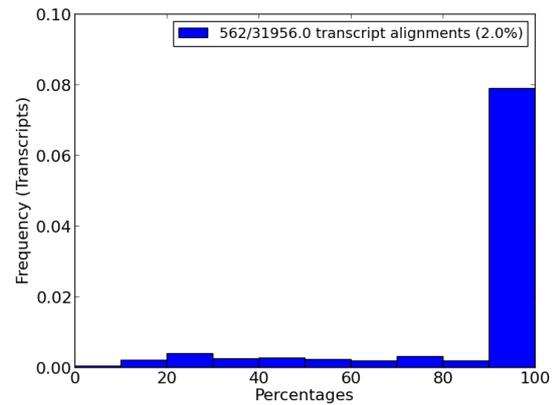
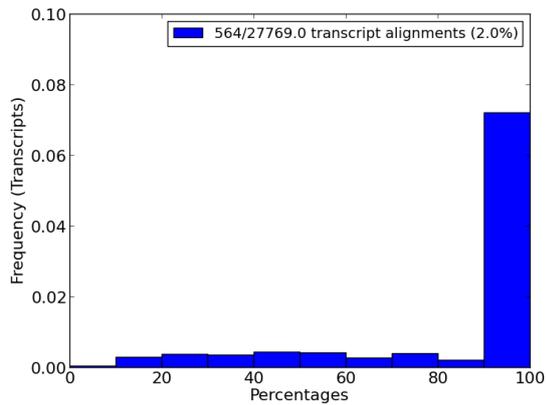
(a) Tomato assembly made with Trinity.



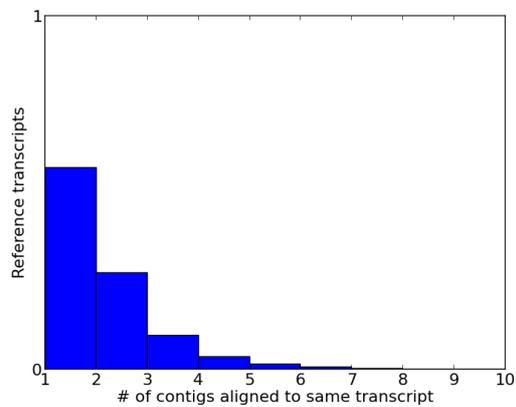
(b) Tomato assembly made with CLC.

Figure 27: For each reference transcript, the number of assembled transcripts that align to that reference transcript. Aligned with Bowtie2.

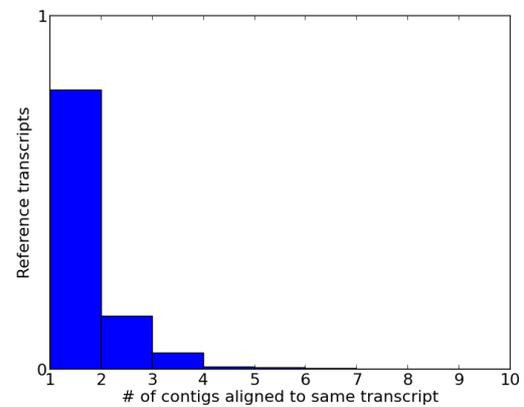
Core Genes



(a) Trinity assembly of tomato data (514 unique hits) (b) Trinity assembly of mouse data (554 unique hits)



(c) Trinity assembly of tomato data



(d) Trinity assembly of mouse data

Figure 28: a and b) Distribution of overlap of reconstructed transcripts and core genes for tomato and mouse assemblies made with CLC. c and d) Number of reconstructed transcripts that align to the transcript of a core gene. Tomato and mouse transcripts were aligned to core transcripts of respectively Arabidopsis Thaliana and Homo Sapiens.

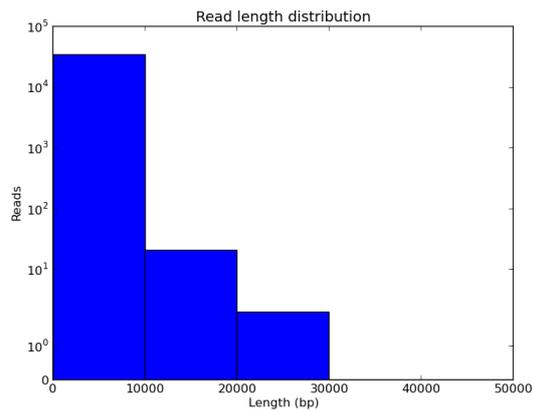
Splice Site Prediction

Splice site prediction with STAR

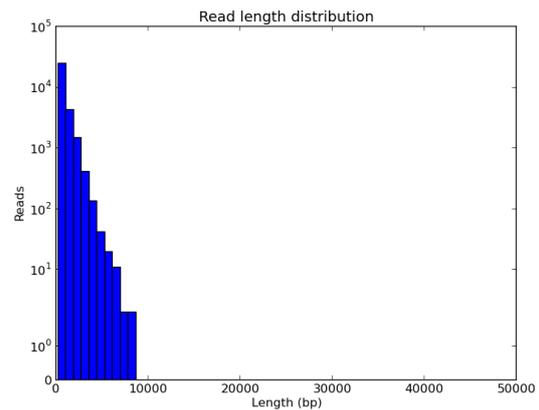
Tomato Number of input reads: 11,042,746 Number of splice sites found: 1,132 Uniquely mapped reads %: 0,16%

Mouse Number of input reads: 24,074,664 Number of splice sites found: 136,765 Uniquely mapped reads %: 3,5%

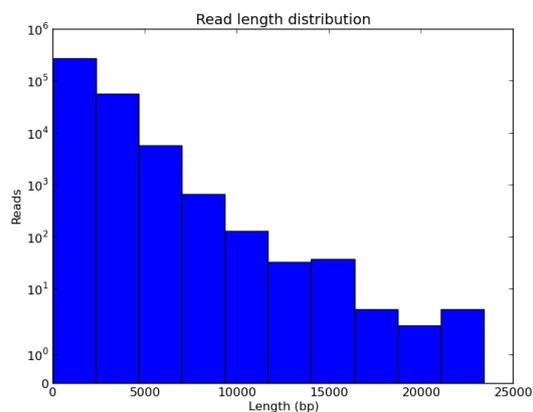
Transcript Length Distribution



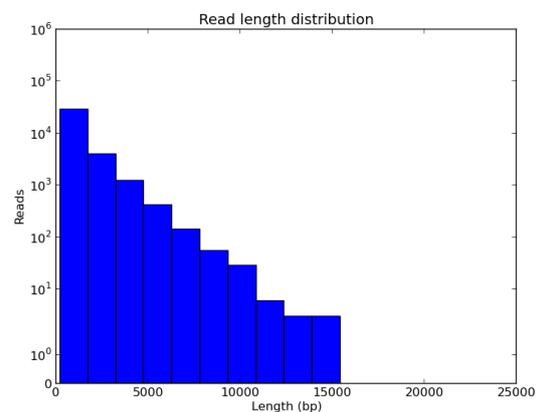
(a) Tomato reference transcriptome



(b) Tomato Trinity assembly



(c) Mouse reference transcriptome



(d) Mouse Trinity assembly

Figure 29: Transcript length distribution for the mouse and tomato assemblies made with Trinity and the reference transcriptomes.

Pipelines & scripts

Digitally Available, please contact the author.

Appendix 1 - Manuals

de novo transcriptome assembly pipeline

The assembly pipeline is fully written in Python 2.7, and makes use of the following tools and packages:

Tools:

- Trinity (+ Bowtie)
- GMAP
- Samtools

Python packages:

- Biopython
- numpy
- pylab
- matplotlib
- reportlab

The pipeline is run via the script: `denovo_transcriptome_assembly.py`

Usage is as follows:

Required parameters:

`-c <string>` `--configfile:` config text file which allows for changes of tool parameters

if paired-end reads:

`-l <string>` `--left:` left reads
`-r <string>` `--right:` right reads

if unpaired reads:

`-s <string>` `--single:` single reads

`-g <string>` `--core_genes:` core gene (transcripts) fasta file (all in subdirectory)
`-p <string>` `--project:` project number (required for PDF report)
`-n <string>` `--name:` client name (required for PDF report)

Optional parameters:

`-o <string>` `--output:` directory for output files (default: current working dir.)

run `denovo_transcriptome_assembly.py` with the `"-h"` or `"-help"` option to print the available parameters.

A typical assembly command would be like this:

```
python2.7 denovo_transcriptome_assembly.py -c config.txt -l left.fq -r right.fq
-g core_gene_transcripts/H.sapiens.fa -n "John" -p "0001" -o project_0001
```

The config file specified with the `"-c"` argument is a text file containing some necessary parameters (such as the path of the different tools), as well as some optional parameters for the tools. If a optional parameter is not specified, the default is used. All parameters available in the different tools can be defined in the config file. The config file format is:

```
# parameter description
parameter name = parameter value/string
or just: parameter name, if no value is required.
```

For example:

```
# Trinity directory
PATH = /data/tools/trinity
```

```
# Inputfile type
seqType = fq
```

```
# Number of cores
CPU = 2
```

Input & Output

Input files can be either in `fasta/fa` or `fastq/fq` format.

Output files are the assembly `fasta` file (`Trinity.fasta`) in the `"trinity"` subfolder, alignment files (raw to assembly and assembly to core genes) in both `sam` and `bam` format. Summaries of the results (statistics and analyses) are found in the generated PDF report. A log file is generated in the main output directory.