De novo Transcriptome Assembly and Annotation

Study the transcriptome of any organism
Discover differential gene expression without a reference genome

Introduction
RNA sequencing by next generation sequencing (NGS) has become a standard tool for studying the transcriptome of an organism and related gene expression profiles. A mandatory prerequisite however, is the availability of a well-annotated reference genome or transcriptome. This reference is used to map the sequencing reads, to count the reads mapped to genetic features (e.g., genes, transcripts) and to perform statistical analysis by comparing sample replicate groups from different conditions to each other. If such a reference is missing, no differential gene expression analysis will be possible unless a reference transcriptome is built. Microsynth’s *de novo* transcriptome assembly and annotation service is designed to fill this gap and allow the study of yet unknown transcriptomes.

Microsynth’s Competences and Services
With more than ten years of experience in the field of next generation sequencing, one of Microsynth’s core competences is to provide high-quality one-stop services from experimental design to bioinformatics analysis. You may either outsource the entire analysis or only single steps to us as illustrated in Figure 1.

Experimental Design
Microsynth’s NGS specialists will help you define suitable experimental setups for your *de novo* transcriptome project and discuss possible strategies to address your associated research questions.

RNA Isolation
You may either perform the RNA extraction yourself or outsource this critical step to Microsynth, which has long-standing experience in processing various sample matrices and RNA sources.

Library Preparation and Sequencing
To be able to cover the transcriptome of an organism well enough, RNA from different tissues, stages of growth and reactions to treatments need to be pooled and sequenced together. This recommendation is in stark contrast to the standard differential expression analysis in which sample groups must not be pooled together to be able to detect variations in gene expression patterns. From the extracted RNA, a normalized cDNA library is created to represent every transcript as uniformly as possible for the future reference transcriptome. Sequencing is performed...
on one of our Illumina MiSeq platforms which may produce paired-end reads with up to 2x300 base pairs.

**Bioinformatics Analysis**

Sequenced reads are quality filtered, trimmed and in silico normalized before being subjected to de novo transcriptome assembly. The resulting assembly is further refined, and transcripts are predicted from the assembled contigs. The transcripts are then searched against the Swissprot and PfamA databases for homologies and are annotated with any available supplementary information such as GO and KEGG terms. A detailed description of the data- and workflow used by Microsynth to create reference transcriptomes can be found in the following publication [1].

### Example Results

**Table 1.** The detected transcripts are described in detail in a GFF3 formatted file along with their untranslated (UTR) and coding (CDS) regions as depicted in this cutout.

<table>
<thead>
<tr>
<th>Transcrip</th>
<th>id</th>
<th>name</th>
<th>score</th>
<th>identity</th>
<th>alignment length</th>
<th>mismatches</th>
<th>gap opens</th>
<th>evalue</th>
<th>bit score</th>
<th>name</th>
<th>pftarget</th>
<th>descr</th>
</tr>
</thead>
<tbody>
<tr>
<td>TRINITY_DN16040_c1_g2_i13-transdecoder_gene</td>
<td>1</td>
<td>1484</td>
<td>62.8</td>
<td>86</td>
<td>28</td>
<td>1</td>
<td>6.4e-27</td>
<td>131.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TRINITY_DN16040_c1_g2_i13-transdecoder_mRNA</td>
<td>1</td>
<td>1484</td>
<td>52.3</td>
<td>102</td>
<td>31</td>
<td>0</td>
<td>3.2e-49</td>
<td>173.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TRINITY_DN16040_c1_g2_i13-transdecoder_five_prime_UTR</td>
<td>660</td>
<td>1484</td>
<td>72.3</td>
<td>152</td>
<td>31</td>
<td>0</td>
<td>2.1e-47</td>
<td>161.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 2.** A cutout of one of the produced annotations detailing homologies of the assembled transcripts to sequences found in various publicly available databases:

<table>
<thead>
<tr>
<th>Transcript</th>
<th>UniProtKB-ID</th>
<th>UPKB-ID</th>
<th>description</th>
<th>identity</th>
<th>alignment length</th>
<th>mismatches</th>
<th>gap opens</th>
<th>evalue</th>
<th>bit score</th>
<th>PfamAC</th>
<th>pftarget</th>
<th>descr</th>
</tr>
</thead>
<tbody>
<tr>
<td>TRINITY_DN16040_c1_g2_i13_p1</td>
<td>O22977</td>
<td></td>
<td>Probable lysophospholipase hydrolase (A10365); Arabidopsis thaliana (2565772)</td>
<td>62.8</td>
<td>86</td>
<td>28</td>
<td>1</td>
<td>6.4e-27</td>
<td>131.3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TRINITY_DN16040_c1_g2_i13_p1</td>
<td>Q9ZPV7</td>
<td></td>
<td>Alpha/beta hydrolase fold</td>
<td>72.3</td>
<td>102</td>
<td>31</td>
<td>0</td>
<td>3.2e-49</td>
<td>173.3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TRINITY_DN22587_c1_g1_i3_p1</td>
<td>PE29481.19</td>
<td></td>
<td>Alpha/beta hydrolase fold</td>
<td>72.3</td>
<td>102</td>
<td>31</td>
<td>0</td>
<td>3.2e-49</td>
<td>173.3</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Related Topics**

1) For Microsynth’s standard RNA sequencing, differential gene expression and pathway analysis services, please visit the related website and download the application note.

2) For Microsynth’s gDNA de novo assembly service, please visit the following website and download the application note.

**References**