Universität für Bodenkultur Wien
University of Natural Resources and Life Sciences Vienna

Department of Biotechnology
Extremophile Center

Transcriptome Analysis of *Cladophialophora immunda*

Master Thesis

submitted by
Christina Kustor, BSc.
Vienna, 2015

Supervisor: Katja Sterflinger-Gleixner, Assoc. Prof. Dr.
Co-Supervisor: Hakim Tafer, Dr.

Acknowledgement

Abstract

...

Keywords:

Zusammenfassung

...
Stichwörter:

Contents

Background **4**
1.1 Black yeasts' reaction on different environmental conditions 4
1.2 Cladophialophora immunda 5
1.3 Transcriptome Analysis using RNA-seq 6
1.4 Bioinformatics methods 7
1.4.1 Workflow Systems 7
1.4.2 Algorithms 9

Methods **10**
2.1 Alignment 12
2.2 Transcriptome Assembly 12
2.3 Gene Prediction 13
2.4 Differential Expression Analysis 15
2.5 Functional Annotation 15
2.6 Enrichment Analysis 15

Results **16**

Discussion **17**

Conclusion **18**

Bibliography **21**

List of Tables **22**

List of Figures **23**

Background

1.1 Black yeasts' reaction on different environmental conditions

Black yeasts differ from the most fungis because they can grow in extreme environments and are resistant against high levels of stresses like Ozon, UV and radioactivity. [4] ....
The aim of the thesis was to determine the transcriptomes of *Cladophilaphora immunda* and to understand which genes are transcribed under different conditions. The transcriptome data were obtained with RNA-seq performed by Ion Torrent technology coupled with the Ion Proton sequencer (Life Technologies, Carlsbad, CA) [1] and the analysis were realised with bioinfromatics methods. A workflow system for genome annotation with RNA-seq data was developed and followed by performing statistics to obtain a list of genes and their expression level differences between two groups. Furthermore functional annotation and additonal enrichment analysis were performed.

1.2 Cladophialophora immunda

The black yeast *Cladophialophora immunda* is known for the capability to grow on polyaromatic hydrocarbons and moreover for the ability to degrate hydrocarbons. [14] *Cladophialophora immunda* was isolated from a gasoline station, in a hydrocarbon rich environment. Previous studies with bioredimiation spotted that *Cladophialophora immunda* is able to grow on toluen and also can degrade toluen. [11] [10]
For the experimental conditions, the cultivation of *Cladophilaphora immunda* in glucose and toluen was chosen.

1.3 Transcriptome Analysis using RNA-seq

The powerful technology RNA-seq and the analysis of transcriptomes, to know which gene is expressed, became an important method in science. Transcriptome analysis enablies the understanding of how sets of genes work together to form metabolic, regulatory and signaling pathways within a cell. [16] A distinction of RNA-seq from earlier methods, such as microarrays, is the high throughput of current RNA-seq platforms, the sensitivity afforded by newer technologies and the ability to discover novel transcripts, gene models and noncoding RNA species. [3]

The work in the laboratory included the extraction of total RNA with FastRNA PRO RED KIT (MP Biomedicals, Santa Ana, CA), furthermore the isolation by means of Dynabeads mRNA DIRECT Micro Kit (Ambion by Life Technologies, Carlsbad, CA), transcriptome library preparation based on the Ion Total RNA-Seq Kit v2 (Life Technologies, Carlsbad, CA) and the key part, the RNA-Seq, implemented by Ion Torrent technology coupled with the Ion Proton sequencer (Life Technologies, Carlsbad, CA). [1]

1.4 Bioinformatics methods

The increasing use of next-generation sequencing methods is related to a large amount of produced data, which has to be analysed and interpreted. Therefore bioinformatics methods are a necessary step in analysing transcriptomes. As RNA-seq is an active field of research producing new approaches and tools at a rapid pace, many alternative programs exist for each analysis step. [3]

1.4.1 Workflow Systems

The data analysing steps were performed by different programs and tools, which may need specific data formats and external files. The multiple steps can be handled through scripting a reusable pipeline with defined inputs, output and parameter for each step.
The software Snakemake was used to evolve the pipelines for genome annotation, functional annotation and enrichment analysis. Snakemake is based in Python language and can be used on a single core workingstation as well as on a cluster without modifying the workflow. Snakemake's option "–dag" creates the directed acyclic graph (DAG) of executed jobs. [7]
In figure 1 exemplified the DAG of the enduced Snakemake file. Jobs (i.e. the execution of a rule) are depicted as nodes, a directed edge between two jobs means that the rule underlying the second job needs the output of the job executed before as an input file. A path represents a sequence of jobs that have to be executed serially.

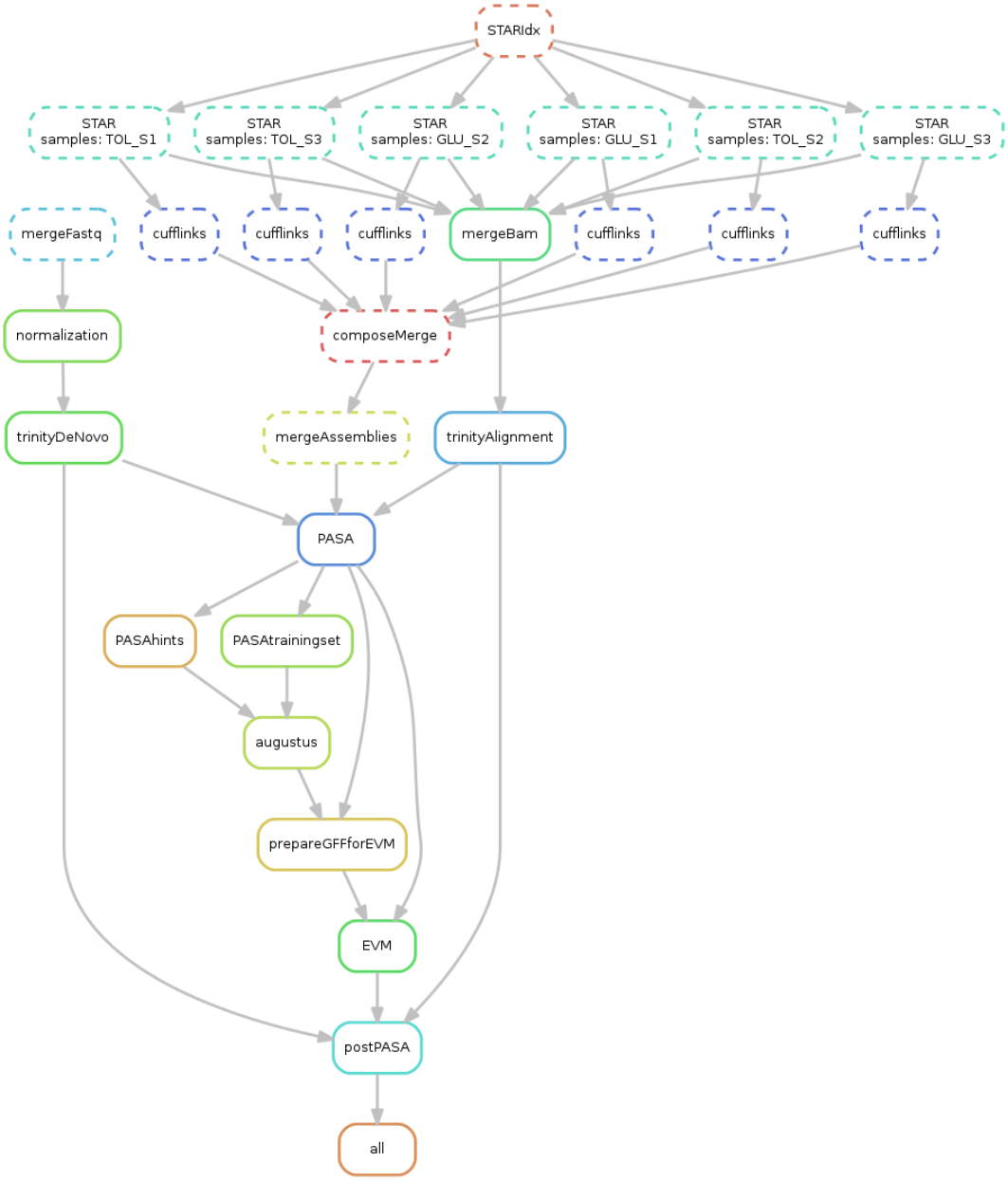

Figure 1: Directed acyclic graph (DAG) of a Snakemakefile

1.4.2 Algorithms

Maximum Mappable Length

The algortihm Maximum Mappable Length (*MML*) was developed to align reads to the genome. The idea of *MML* approach is to search a Maximum Mappable Prefix (*MMP*). Given a read sequence R, read location i and a reference genome sequence G, the *MMP(R,i,G)* is definde as the longest substring

$$[R_i, R_{i+1}, ..., R_{i+MML-I}]$$

which matches exactly one or more substrings of G, where *MML* is the maximum mappable length. Starting from the first base of the read the algorithm finds the *MMP*. If *MMP* can not be mapped contigously to the genome, because of an splice junction, this first seed will be mapped to a donor splice site. The *MMP* search is repeated for the unmapped portion of the read, which will be mapped to an aceptor splice site. The difference to other algorithmn is to align the non-contigous sequences directly to the reference genome, which accelerated the process to map. [2]

Expressed Sequence Tag

Expressed Sequence Tag (EST) was evolved in early days of genome assembly. First the same assemblers where used for transcriptomes, but there a fundamental differences between genome assembly and transcriptome assembly. In the genome assembly the ideal output is a linear sequence representing each genomic region, whereas in the transcriptome assembly the gene is most naturally described as a graph. [3]

de Bruijn Graph

Each node of a de Bruijn Graph is associated with a (k -1)-mer. Two nodes A and B are connencted if there is a k-mer whose prefix is the (k -1)-mer of the node A and the suffix is (k - 1)-mer of the node B. The k-mers create edges in the Buijn graph. [3]

Methods

Before implementing the workflow system a comparison of publications, different programs and tools were done. The choice of a suitable program depended on the kind of analysis and the amount of time is required. Another important issue was the option available of the used program.
Six samples of reads were downloaded from the sequencer, three of them were the transcriptomes of *Cladophialophora immunda* cultivated with glucose and the other three with toluen. In the beginning the reads' origin was identified by aligning them to a reference genome with STAR. Mapped reads were used to perform transcript assembly with the softwares Cufflinks and Trinity Genome Guided. Trinity also was employed for mapping-free transcript assembly. With the produced assemblies from Cufflinks and Trinity, the Program to Assembly Spliced Alignments (PASA) performed the spliced alignment mapping. PASA assemblies were utilized to create hints for the location of introns and exons. Those hints were ingrated in a hintbased run of Augustus. Evidence Modeler (EVM) was used to combine weighted predictions. The last step was a annotation update implemented by PASA, where the output of EVM compared to PASA alignment assemblies and generated a updated gene set. [6] [13] [8]
Figure 2 shows the draft of the entire workflow system for genome annotation with transcriptomic data. The inputs are the transcriptomes from RNA-Seq in bam-format, the genome in fasta-format and the de novo annotation without RNA-Seq data in gff3-format. After runs of the different steps the output is an annotation file in gff3-format.

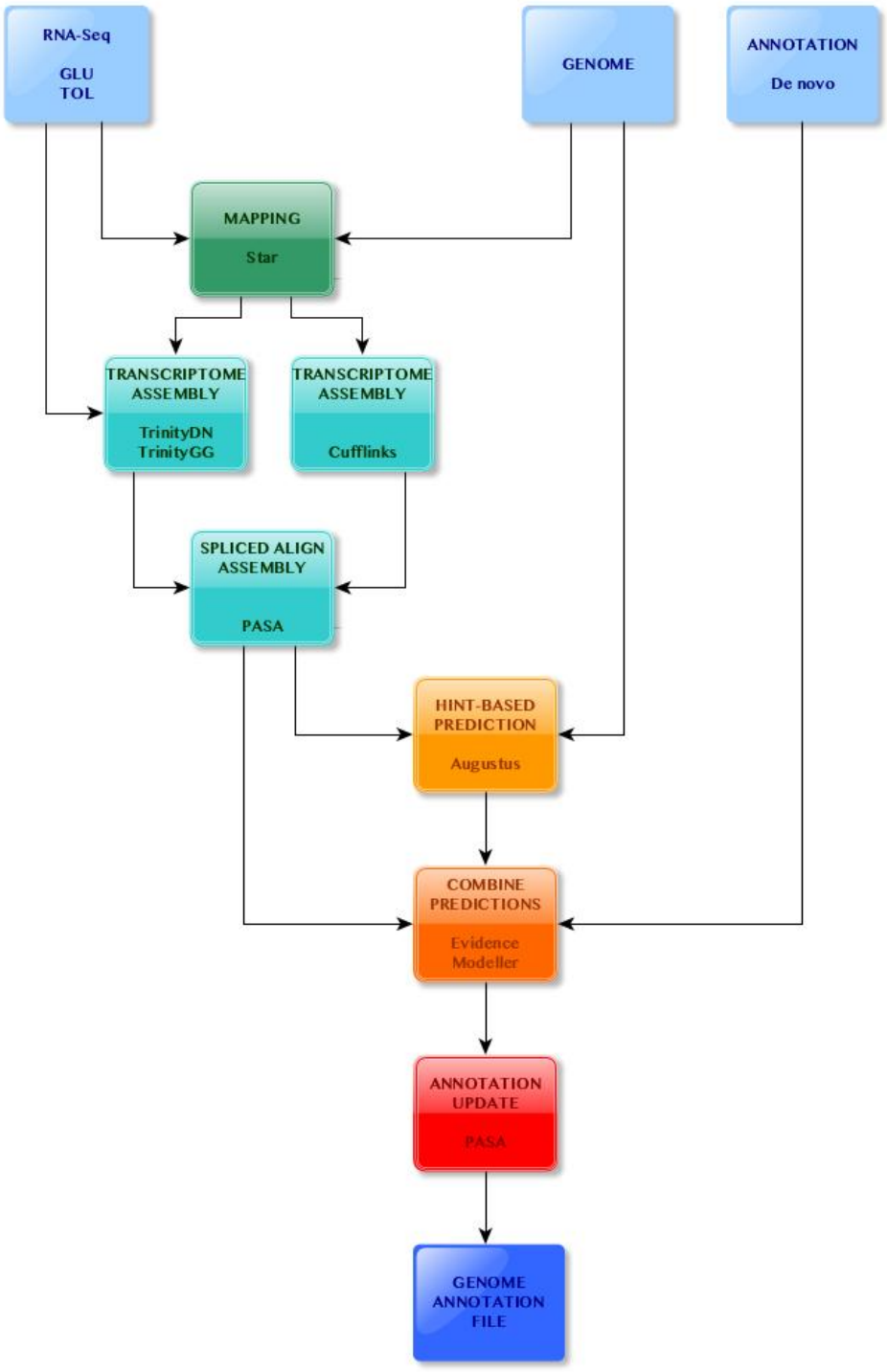

Figure 2: Genome annotation workflow system

2.1 Alignment

The first step to understanding a genome structure is through genome mapping, which is a process of identifying relative locations of genes on a chromosome. When a read is mapped to reference genome, a sequence alignment is created. [16] The input files in this case were the preprocessed reads and in addition the reference sequence.

STAR

Mapping is computationally intensive, that is why the program STAR was chosen. STAR (Spliced Transcripts Alignment to a Reference) is a spliced alignment program which runs very fast. The benefits of STAR are largely based on the "maximum mappable length" approach described in 1.4.2. STAR splits a read and find the best portion that can be mapped for each piece. Then it maps the remaining portion, which can be far away in the case of a splice junction. The maximum mappable seed search looks for exact matches and uses the genome in the form of uncompressed suffix arrays. The second step of STAR stitches the seeds together within a given genomic window and allows mismatches, indels and splice junctions. The seeds from read pairs are handled concurrently at this step in order to increas sensitivity. [2] [3]
The execution of STAR allocated in two parts, building a reference index and mapping. Before running STAR mapping it was obtained a reference index of the genome. Mapping of the reads was executed without the soft clip aligning at reference ends. The mapped reads needed to be sorted by Samtools.

2.2 Transcriptome Assembly

The goal of RNA-seq assembly is to reconstruct full-length tanscripts based on sequenced reads. Transcript assembly answers the questions about exon regions and splice site. Several transcripts may overlap at different regions or there may be multiple copies of the same transcript.
There are two ways of performing transcriptome assembly. If there is a reference genome, it can be realised to guide the assembly, where the assembly task consits of solving which mapped reads correspond to which transcript. The second possibilty is to perform de novo assembly. [3]

Cufflinks

The software packages Cufflinks can be used for *ab intio* reconstruction. The program within in this packages Cufflinks assembles transcriptomes from RNA-Seq data.
Cufflinks reports the smallest possible set of isoforms.
The program Cuffmerge was used to merge the multiple assembled transcriptomes into a master transcriptome. [15]
The rule Cufflinks was implemented with the following option and values, the maximum intron length(2000) and the minimum intron length (30), the maximum number of fragment a locus may have before skipped(10000) and the library-type ff-firststrand.

Trinity

[9]
Trinity genome guided
Trinity de novo

PASA

2.3 Gene Prediction

Augustus

Augustus is an *ab intio* gene predictor, where only a genomic sequence is needed as input information. In addition it is possible to use hints of various information. For the prediction Augustus combines genomic sequence alignments and alignments of expressed sequenced tags (EST), described in 1.4.2. The model underlying the program is generalized hidden Markov model (GHMM). HMMs and GHMMs for gene prediction define a probabilty for each pair (φ,s) of a sequence s and a gene structure φ. Before starting the program Augustus it is necessary to train the model. [12]
The trainingset was created from PASA assemblies by executed the program "pasa_asmbls_to_training_set.dbi" included in PASA package. For the additional file of hints the location of introns were filtered from PASA assemblies.

Subsequently Augustus was started and after more than 24 hours, the output was a complete gene prediction in ggf-format.

Evidence Modeler

The Evidence Modeler (EVM) is The assembly of PASA (weight = 10), the Augustus prediction (weight = 5) and the de Novo Annotation (weight = 1) were combined. [5]

Pasa Annotaion Update

2.4 Differential Expression Analysis

featureCounts limma

2.5 Functional Annotation

2.6 Enrichment Analysis

Trough Enrichment Analysis a list of over-represented molecular funtions, biological process and cell locations, that can then be used to test whether genes are regulating biochemical or cellular pathways, genes is provided. Gene set enrichment analysis provides a means by which the genes in a data set can be grouped compared to a background such as all genes. The most commin annotation used for grouping genes is the Gene Ontology. ...

Results

Annotation
DIFF
The computational results presented have been achieved in part using the Vienna Scientific Cluster (VSC).

Discussion

text

Conclusion

text

Bibliography

[1] Barbara Blasi, Hakim Tafer, Donatella Tesei, and Katja Sterflinger. From Glacier to Sauna: RNA-Seq of the Human Pathogen Black Fungus Exophiala dermatitidis under Varying Temperature Conditions Exhibits Common and Novel Fungal Response. *Plos One*, 10(6):e0127103, 2015.

[2] Alexander Dobin, Carrie a. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, 2013.

[3] Garry Wong Eija Korpelainen, Jarno Tuimala, Panu Somervuo, Mikael Huss. *RNA-seq Data Analysis - A Practical Approach.* 2014.

[4] Cene Gostinčar, Martin Grube, Sybren De Hoog, Polona Zalar, and Nina Gunde-Cimerman. Extremotolerance in fungi: Evolution on the edge. *FEMS Microbiology Ecology*, 71(1):2–11, 2010.

[5] Brian J Haas, Steven L Salzberg, Wei Zhu, Mihaela Pertea, Jonathan E Allen, Joshua Orvis, Owen White, C Robin Buell, and Jennifer R Wortman. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome biology*, 9(1):R7, 2008.

[6] Brian J Haas, Qiandong Zeng, Matthew D Pearson, Christina a Cuomo, and Jennifer R Wortman. Approaches to Fungal Genome Annotation. *Mycology*, 2(3):118–141, 2011.

[7] Johannes Köster and Sven Rahmann. Snakemake-a scalable bioinformatics workflow engine. *Bioinformatics*, 28(19):2520–2522, 2012.

[8] J. Linde, S. Duggan, M. Weber, F. Horn, P. Sieber, D. Hellwig, K. Riege, M. Marz, R. Martin, R. Guthke, and O. Kurzai. Defining the transcriptomic landscape of Candida glabrata by RNA-Seq. *Nucleic Acids Research*, 43(3):1392–1406, 2015.

[9] Nir Manfred G. Grabherr, Brian J. Haas, Moran Yassour, Joshua Z. Levin, Dawn A. Thompson, Ido Amit, Xian Adiconis, Lin Fan, Raktima Raychowdhury, Qiandong Zeng, Zehua Chen, Evan Mauceli, Nir Hacohen, Andreas Gnirke, Nicholas Rhind, Federica di Palma, Bruce W., Friedman, and Aviv Regev. Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nature Biotechnology*, 29(7):644–652, 2013.

[10] Caroline Poyntner. Bioremediation of waste gas and soil by black extremotolerant fungi. pages 1–30, 2014.

[11] Francesc X. Prenafeta-Boldú, Andrea Kuhn, Dion M.A.M. Luykx, Heidrun Anke, Johan W. van Groenestijn, and Jan A.M. de Bont. Isolation and characterisation of fungi growing on volatile aromatic hydrocarbons as their sole carbon and energy source. *Mycological Research*, 105(4):477–484, April 2001.

[12] Mario Stanke, Ana Tzvetkova, and Burkhard Morgenstern. AUGUSTUS at EGASP: using EST, protein and genomic alignments for improved gene prediction in the human genome. *Genome biology*, 7 Suppl 1(May 2005):S11.1–8, 2006.

[13] Tamara Steijger, Josep F. Abril, Pär G. Engström, Felix Kokocinski, The RGASP Consortium, Tim J. Hubbard, Roderic Guigó, Jennifer Harrow, and Paul Bertone. Assessment of transcript reconstruction methods for RNA-seq. *Nature Methods*, 10(12):1177–1184, 2013.

[14] Katja Sterflinger, Ksenija Lopandic, Barbara Blasi, Caroline Poynter, and Sybren De Hoog. Draft Genome of Cladophialophora immunda , a Black Yeast and Efficient Degrader of Polyaromatic Hydrocarbons. 3(1):5–6, 2015.

[15] Cole Trapnell, Brian A Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J van Baren, Steven L Salzberg, Barbara J Wold, and Lior Pachter. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology*, 28(5):511–515, 2010.

[16] Jin Xiong. *Essential Bioinformatics*. 2006.

List of Tables

List of Figures

1 Directed acyclic graph (DAG) 8
2 Genome Annotation Workflow 11