University of Natural Resources and Life Sciences Vienna

Department of Biotechnology
Extremophile Center

Transcriptome Analysis of *Cladophialophora immunda*

Masterthesis

submitted by
Christina Kustor, BSc.
Vienna, 2015

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

Acknowledgement

Abstract

The black yeast *Cladophialophora immunda* is known for the capability to grow on polyaromatic hydrocarbons and moreover for the ability to degrate hydrocarbons. The thesis will exermine the way in which the transcriptomes of *Cladophilaphora immunda* can be determined and which genes are transcribed under different conditions. The transcriptome data were obtained with RNA-seq and the analysis were realised with bioinformatics 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.

....

Keywords: Extremophile, Black Yeast, Transcriptomes

Zusammenfassung

...

Stichwörter:

Contents

Introduction

1.1 Extremophilic fungi

An extremophile is an organism that thrives in an extreme environment. The habitants are quite different including the physical extremes for instance temperature, UV radiaton or pressure and the geochemical extremes such as desiccation, salinity, pH or redox potential. [22] In the last decade the interest of extreme environments and the previously unknown extremophilic microorganisms in such region has increased. The enthusiasm to isolate them in pure culture and profile their metabolites make them so attractive for research. Their products have potential to be valuable resources for the development of a bio-based economy through their application to different branches in biotechnology. [20]
The diversity of fungi to occur in stressful environments that are hostile to most eukaryotes is one of the greatest in microorganisms. Comparative results of evolutionary studies have attempted to explain adaptability of extremophiles. Fungi from extremely cold and salty habitats share patterns in morphology, phylogeny and population. [5]

1.2 Black yeasts

"Black yeasts" is a *terminus technicus* subscribing a group of fungi that conquer extreme environements characterized by oligothrophic nutrient conditions, elevated temperatures, UV radiation, matrix and osmotic stress and combinations of these factors. Other terms are "meristematic fungi" and "MCF" Meristematic fungi was first mentioned by de Hoog and Hermanides Nijhof in 1977, for fungi

that form aggregates of thick-walled, melanized cells enlarging and reproducing by isodiametrical division. MCF refers to growth pattern of meristematic fungi and some black yeasts. These fungi grow in mineral substrates like rock but also glass or metal.[25]
The combined influence of the mentioned stress factors exerts a high selective pressure on the microbial community and as a consequence black yeasts are rarely found in complex microbial populations. Black yeasts are quite heterogeneous from taxonomic and phylogenetic point of view but they have in common melanized cell walls. The production of melanins and the incrustation of the cell walls with this high-molecular substances are the most important factors in stress restistance of black yeasts. The black yeasts include species are also found in human environments and have a human-pathogenic potential. Beside the protection factor, melanin also affect the penetration of host tissue in plants, animal and human tissue. *Exophiala dermatitidis* and *Cryptococcus neoformans* are such human pathogenic yeasts where melanin is one of the virulence factors. [10, 25] Other black fungi that associated with humans are represented by the typical black yeasts belonging to the genera *Fonsecaea, Capronia, Phaeococcomyces and Cladophialophora.* [3]
The group of black fungi was chosen for the purpose of bioremediation, as they are found in extreme environments and therefore resistant to stress conditions. [18]

1.3 Bioremediation

Toluene and other related aromatic hydrocarbons are abutant environmental pollutants. Bioremediation of pollutants is an attractive method because of its environmental and economical advantages. The use of bacteria for biodegradation of toluene to CO_2, water and biomass has been studied widely. However the benefit of fungi for this purpose becomes very important, chiefly because of the advantages under conditons of reduced water activity and low pH, which often preveal in biofilters. [16] The process of using fungi to degrade contaminants in the environment is also termed mycoremediation. This biological treatment can be differentiated into bio-augmentation, biosparging, bioventing, composting and several other less frequently applied methods. The method of the precending study was bioaug-

mentation, a bioremediation option for hydrocarbon-contaminated, oily- sludge restoration. Bioavailability of the contaminent for the microorganisms, the degradation to less toxic compounds and the opportunity for optimization of biological activity are critical points in bioremediation. [18]

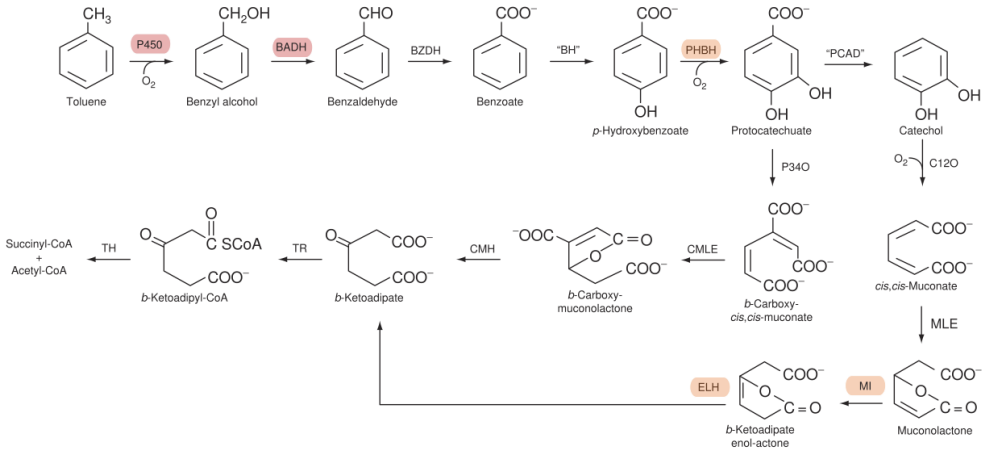

Figure 1: Toluen degradation pathway

Toluene (methylbenzene) is an aromatic hydrocarbon natural product of diagenic origin and an important commercial chemical. The BTEX mixtures referred to in bioremediation applications contain benzene, ethylbenzene, toluene and xylenes. These toxic compounds are ussually difficult to remove to their wide dispersal in ecosystem. A toluen degradation pathway in fungi was first proposed for *Cladophialophora saturnica* [1] and in an other previous study the connection between toluen metabolism and cytochrom P450 was established [16]. The exerminations for this presence of genes belonging to the pathway for the toluen degradation (figure 1) were done in comparison with the genome of the model yeast *Saccharomyces cerevisiae*. [2]

1.4 Study of *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. The genus *Cladophialophora* belongs to the ascomycetes and forms two phylocenetic glades (*Carrionii* and *Bantiana*) in the family of *Herpotrichiellaceae*, order *Chaetothyriales*. *Cladophialophora* is morphologically characterized by one-celled, ellipsoidal to fusiform, dry conidia arising through blastic, acropetal conidiogenesis and arranged in branched chains. The genus includes species causing skin infections and other human diseases. *C. immunda* is of special medical and biotehcnological interest because it is frequently isolated both from humans and from contaminated soil. [26, 1]

PICTURE BLACK YEAST

Cladophialophora immunda was isolated from a gasoline station, in a hydrocarbon rich environment, therefore the fungi is an important candidate for bioremediation and for application in biofilters. [19]
To better understand the mechanisms of black yeasts and their effect to degrate hydrocarbons, a preceding study on bioremediation was done. Therefore a collection of 163 strains of black yeast-like fungi from the CBS Fungal Biodiversity Center (Utrecht, The Netherlands) has been screened for the ability to grow on hexadecane, toluene and polychlorinated biphenyl 126 (PCB126) as the sole carbon and energy source. These compounds were chosen as representatives of relevant environmental pollutants. The results indicated that the two strains of *C. immunda* and *Exophiala mesophila* are able to grow on toluene, confirmed by an increase in CO_2 and a toluene decrease shown in figure 2.[2, 18]

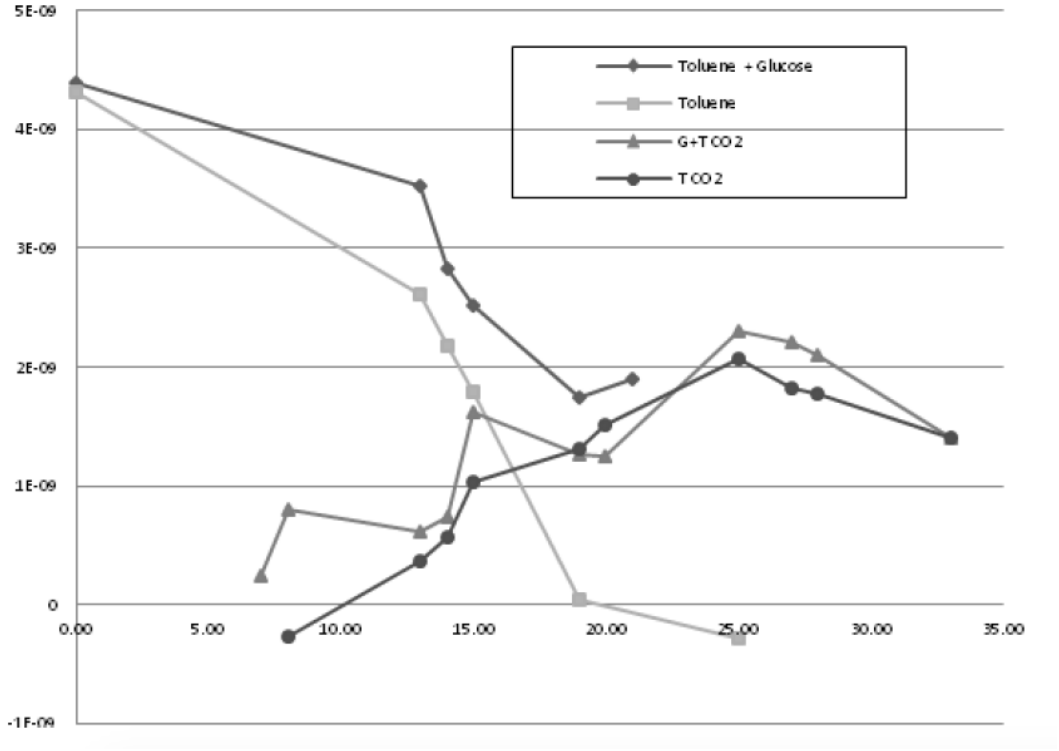

Figure 2: Toluen and CO_2 values for *C. immunda*: Carbon equivalence [mol] of the two molecules was plotted against runtime [days]. [18]

Together all the recent studies indicate that maybe the analysis of transcriptome will answer all the question about that special fungi. Accordingly the aim of the thesis is to exermine the way in which the transcriptomes of *Cladophilaphora immunda* can be determined and identify which genes are transcribed under different conditions. The major issue is to determine the pathways that allow *C. immunda* to use toluene as carbon source.
The transcriptome data were obtained with RNA-seq performed by Ion Torrent technology coupled with the Ion Proton sequencer (Life Technologies, Carlsbad, CA). [2]

1.5 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. [30]
RNAs in total or fractioned are first converted into a library of cDNA fragments. Sequencing adaptors are added to each cDNA fragment to one or both ends. Each molecule is then sequenced in a high-throughput manner to obtain short sequences from one end (single-end sequencing) or both ends (pair-end sequencing).The reads are typically 30–400 bp, depending on the sequencing technology used. If the resulting sequence reads have been obtained, the first task of data analysis is to map the short reads from RNA-seq to the reference genome or to assemble them into contigs before aligning them to the genomic sequence to reveal transcription structure. The way of analysing transcriptomic data differs considering short reads also contain reads that span exon junctions or that contain poly(A) ends. Alignment can be also complicated for large transcriptomes as a result of matching multiple locations in the genome with a significant portion of sequence reads. [29]
An advantage 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. [11]

1.6 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. [11]

1.6.1 Workflow System

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, outputs and parameters 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. [12]
In figure 3 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 3: Directed acyclic graph (DAG) of a Snakemakefile

1.6.2 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. [4]

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. [11]

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. [11]

Methods

2.1 Growth of *Cladophialophora immunda*

For the experimental conditions, the cultivation of *Cladophilaphora immunda* in glucose and toluen was chosen.
The *Cladophialophora immunda* strain (CBS 110551) was isolated from a toluene-charged air biofilter inoculated with gasoline-polluted soil. [19]
Cladophialophora immunda (CBS 110551) was cultured in malt extract agarose media (2 % malt extract, 2 % D-glucose, 0.1 % bacto-peptone and 2 % agar).
For the RNA-seq experiments, *C. immunda* was grown in liquid culture in a modified Hartmans‘ mineral media with 2 % glucose or 1.35 mM toluene as carbon sources. [9] The toluene was supplied through the air of a sealed flask in a 5 % solution in dibutylphtalate. The experiments duration was 90 days (growth with toluene) and one week (growth with glucose) at the temperature of 22 °C at 100 rpm on an orbital shaker. Both experiments were performed in 3 biological replicates. At the end of the experiments the biomass was collected by centrifugation (5000 g per 15 minutes at 4 °C), washed with RNAse free water, frozen in liquid nitrogen and stored at -80 °C until use.

The medium (1 litre of demineralized water) is compose by:
Solution A: 10 mL
Solution B: 25 mL
+ 0.02 % yeast extract as nitrogen source

Solution A (1 litre demineralized water)		Solution B (1 litre demineralized water)	
$(NH_4)_2SO_4$	200 g	K_2HPO_4	155 g
$MgCl_2.6H_2O$	10 g	$NaH_2PO_4.2H_2O$	85 g
EDTA	1 g		
$ZnSO_4.7H_2O$	0.2g		
$CaCl_22H_2O$	0.1 g		
$FeSO_4.7H_2O$	0.5 g		
$Na_2MoO_4.2H_2O$	0.02 g		
$CuSO_4.5H_2O$	0.02 g		
$CoCl_2.6H_2O$	0.04 g		
$MnCl_2.2H_2O$	0.1 g		

Table 1: Hartmans'mineral media

2.2 RNA-seq library preparation

The work in the laboratory included the extraction of total RNA from 100 mg of fungal biomass with FastRNA PRO™ RED KIT (MP Biomedicals) according to the instructions of the manufacturer. The mRNA was isolated with the Dynabeads® mRNA DIRECT™ Micro Kit (Ambion by Life Technologies) and the following transcriptome library preparation was performed with the Ion Total RNA-Seq Kit v2 (Life Technologies). Total RNA, isolated mRNA and the final cDNA library were all qualitatively and quantitatively evaluated by mean of Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA). The RNA-seq was performed by the Ion Proton™ sequencer (Life Technologies) with the Ion PI Chip v2 (Life Technologies). The average read length of the six cDNA libraries was 175 bp for all five samples. Total reads generated per sample varied between 57,611,573 and 99,965,344.

2.3 Mapping Algorithm

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. [30] The

input files in this case were the preprocessed reads and in addition the reference sequence.

2.3.1 Suffix Tree

The rapid advance of new sequencing technology expand the scale and resolution of many biological applications like quantitative analysis of transcriptome where sequence reads must be compared to a reference. There are different algorithms developed in the last few years. Some of them are based on hash tables and the others are based on suffix trees. The algorithm associates two steps, identifying exact matches and building inexact alignments supported by exact matches. Finding exact matches rely on a representation of suffix tree. The advantage in comparison to a hash table is that by using a tree the alignment to multiple identical copies of substring in the reference has to be done only once. Because the identical copies collapse on a single path in the tree, otherwise in a hash table index an alignment must be performed for each copy. [13]
A suffix tree is a data structure that stores all the suffixes of a string to enabling fast string matching. The time complexity of determining if a query has an exact match against a tree is linear in the length of the query, independent of the length of the reference sequence. Nonetheless a tree takes $O(L^2)$ space, where L is the length of reference. To reduce the space a approach to achieves linear space while allowing linear time searching is used. In theory it is possible to represent a suffix tree $L\ log_2 L + O\ (L)$ bits using rank-selection operations. To solve this space efficient implemantation, an enhanced suffix array was derived that consits of a suffix array and several auxiliary arrays. A suffix array can be regarded as an implicit representation of suffix tree and has an identical time complexity to suffix tree in finding exact matches. [13]

2.3.2 *STAR* specific approach

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 1.6.2. *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. [4, 11]

2.4 Count Algorithm

Read mapping results have to be summarized in terms of read coverage for genomic features of interest. Read counts are required for statistical methods like differential expression analysis.The approach of counting the number of mapped reads to annotation was accomplished by the program *featureCounts*.
featureCounts performs precise read assignment by comparing mapping location of every base in the read or fragment with the genomic region spanned by each feature. It takes account of any gaps like insertions, deletions, exon–exon junctions or fusions) that are found in the read. The input for *featureCounts* consits of aligned reads in Sequence Alignment/Map (sam) or Binary Alignment/Map (bam) format and a list of genomic features in general feature format (gff). The read alignment and the feature annotation should correspond to the same reference genome, which is a set of reference sequences representing chromosomes or contigs. *featureCounts* supports strand-specific read counting with the option "-s".

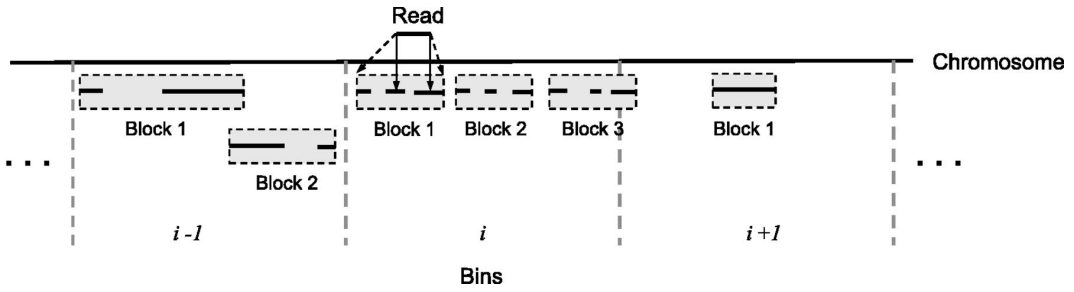

Figure 4: Approach of *featureCounts*: genome bins and feature blocks [14]

The algorithm is starting to generate a hash table for the reference sequence names. This enables that the reference sequence names are found to match in the input of sam/bam-files and gff-annotation quickly. Next the features in each reference sequence are sorted by their start position and a two-level hierarchy is created. Each chromosome is divided into non-overlapping 128kb bins. Features are assigned to bins according to their start postions and grouped into blocks within each bin. The query read is compared first with genomics bins, then with feature blocks within any overlapping bins and then with features in any overlapping blocks. This approach is shown in the illustration 4. [14]

2.5 Differential Expression Algorithm

Differential expression analysis refers to identification of genes or type of genomic features such as transcripts or exons, that are expressed in significantly different quantities in distinct groups of samples. In this study two sets of biological conditons, growth on toluene vs. growth on glucose, was compared by R/BioConductor packages *limma*.

The first step of differential analysis is achieved by mapping RNA-seq data with *STAR* (2.3). The data input to *limma* should be counts such as reads-per-kilobase-per-million (RPKM) produced by *featureCounts* (2.4). *Limma* was then used to implement differential expression.

The software is based on the concept of linear models with the idea to model the expression of each gene as a linear combination of some different explanatory factors. For the experiment on *C. immunda* the linear model for each gene assembled by the measured gene expression (y) is equal to the intercept (a) representing the average expression level of the gene when the factor ($condition$) is in its reference state plus the error term (e):

$$y = a + b * condition + e$$

A generalized linear model (GLM) is a more flexible version of a standard linear model that allows the distribution of the response variable to be different from the normal distribution used in standard linear regression. GLM assume that the read

counts are distributed according to the negative binominal distribution. Accordingly the linear model should be written in matrix form, where the expression level (y) is equal to experimental factor (X) multiplied by the vector of parameters to be estimated from the data (β) plus the error term (ε):

$$\mathrm{y} = \mathrm{X} * \beta + \varepsilon$$

Hence a contrast matrix has to be set up that describes which comparisons wants to be done. The adavantages of *limma* is that it can account for more than one varying experimental factor using a generalized linear modeling framework, it is very fast and memory-intensive. [21]

2.6 Enrichment Approach

Trough Enrichment Analysis a list of over-represented molecular funtions, biological processes and cell locations, that can then be used to test whether genes are regulating biochemical or cellular pathway, is provided. Enrichment analysis is a necessary step after data analysis for more detailed annotations about differentially expressed genes, so that a biological meaning can be postulated.
Functional annotation elements are identified with R-packages Gene Set Enrichment Analysis (GSEA) [27] and Gene Ontology (GO) and the metabolic pathways by Kyoto Encyclopedia of Genes and Genomes (KEGG). A custom script was used to summarize the list of significantly overrepresented GO terms and data are graphically represented with R.

wird mit results ergänzt ....
Funtional Annotation: BLAST INTERPROSCAN CAZY MEROPS TCDB (Transproter Classification database) DVFV antiSmash KOBAS (KEGG Orthology Based Annotation System)

2.7 Pipeline implementation

Six samples of reads were downloaded from the sequencer-server, three of them were the transcriptomes of *Cladophialophora immunda* cultivated with glucose and the other three with toluene.

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.

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 hint-based 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. [8, 24, 15]

Figure 5 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 *deNovo* annotation data in gff3-format. After runs of the different steps the output is an annotation file in gff3-format.

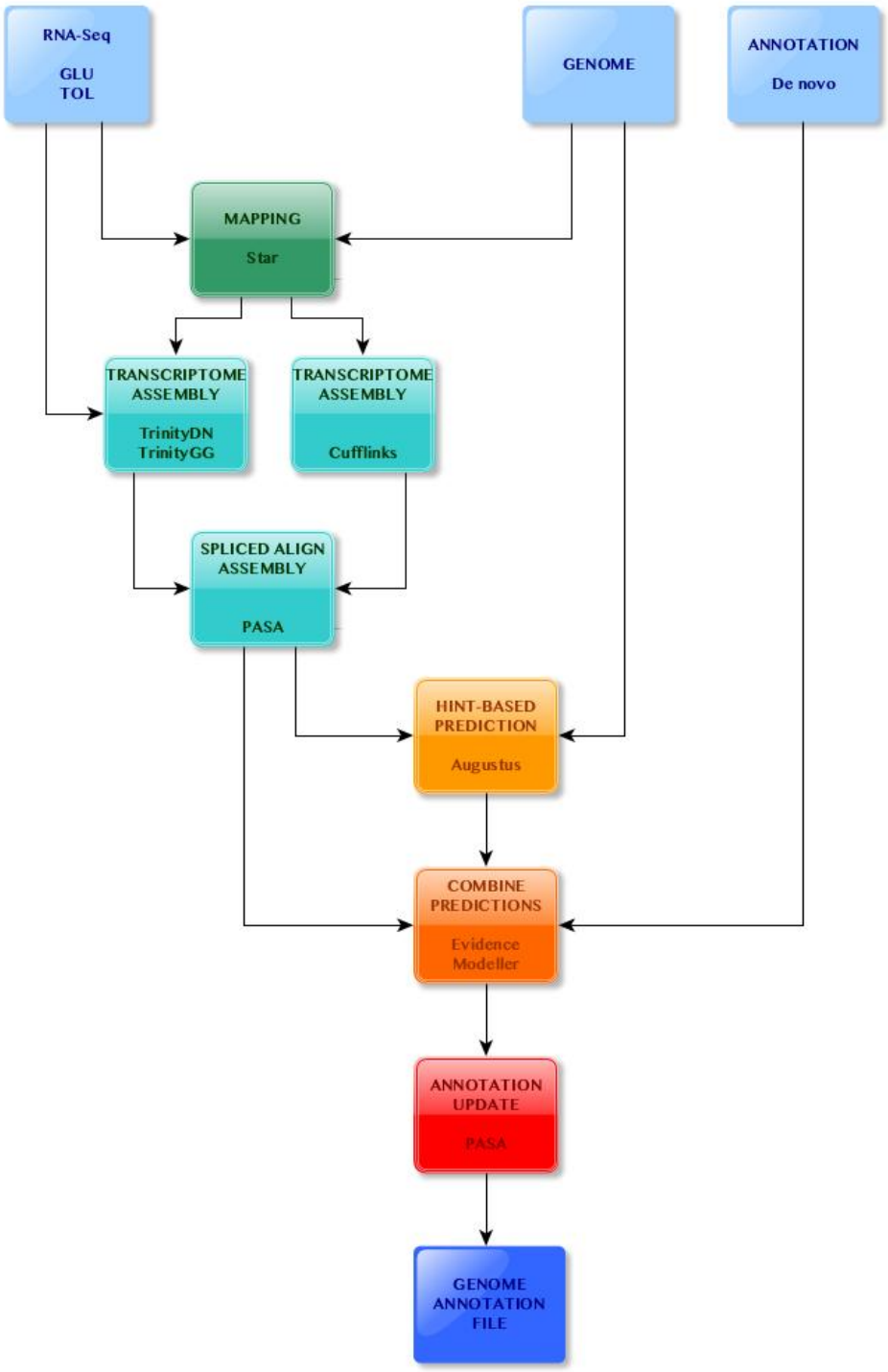

Figure 5: Genome annotation workflow system

Mapping

STAR

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*.

Transcriptome Assembly

The goal of RNA-seq assembly is to reconstruct full-length transcripts 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 *deNovo* assembly. [11]

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. [28]
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

Trinity consits of three seperate programs: *Inchworm*, which construct initial contigs, *Chrysalis* which clusters the contigs produced by *Inchworm* and creates a de Bruijn Graph (1.6.2) for each locus and *Butterfly*, which extracts the isoforms within each de Bruijn Graph. The k-mers is fixed to be 25 in the version 2.0.2. *Jellyfish* is the software, which calculates the k-mers and therefore maximum memory has to be defined. [17]
Trinity was executed genome-guided and *deNovo*. For genome-guided *Trinity* the mapped reads in bam-format and for the *deNovo* approach the raw reads in fastq-format were used. In both the option for strand-specific RNA-Seq read orientation for single end forward was required.

PASA

Program to Assembly Spliced Alignments (*PASA*) annotates protein-coding genes and alternatively splicing isoforms automatically. The goal of *PASA* is to find for each alignment the largest assembly, which is used to create gene models or to modify existing gene models. [6] Further reconstructed assemblies are the input into *PASA* tool. Then *PASA* aligns these newly assembled transcripts to the genome using the software *GMAP*. Next *PASA* filters invalid alignments and those transcripts more likely resulting as artifacts of the RNA-seq assembly process and reconstructs more complete transcripts using its alignment assembly algorithm. [8]
The recontructed inputs for the *PASA* run were *Trinity deNovo* assemblies, *Trinity genome-guided* assemblies and *Cufflinks* transcript structures.

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 1.6.2. 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. [23]
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 an automated eukaryotic gene structure annotation tool. *EVM* reports eukaryotic gene structures by using weighted evidence combining technique. The evidence utilized by *EVM* corresponds to *ab initio* gene predictions and transcript alignment. [7]
The assembly of *PASA* (weight = 10), the *Augustus* prediction (weight = 5) and the *deNovo* Annotation (weight = 1) were combined by *EVM*.

Pasa Annotaion Update

Again Program to Assembly Spliced Alignments (*PASA*) was used to update the *EVM* consensus prediction by adding UTR annotations and models for alternatively spliced isoforms.

Results

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

3.1 Up and Down regulated genes

3.2 Corresponding functional enrichment

3.3 Special Pathways

Discussion

text

Conclusion

text

Bibliography

[1] H. Badali, C. Gueidan, M. J. Najafzadeh, a. Bonifaz, a. H G Gerrits van den Ende, and G. S. de Hoog. Biodiversity of the genus Cladophialophora. *Studies in Mycology*, 61:175–191, 2008.

[2] Barbara Blasi, Caroline Poyntner, Tamara Rudavsky, Francesc X. Prenafeta-Boldú, Sybren de Hoog, Hakim Tafer, and Katja Sterflinger. Pathogenic yet environmentally friendly? Black fungal candidates for bioremediation of pollutants. *Geomicrobiology Journal; inPress*, 2015.

[3] 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.

[4] 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.

[5] 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.

[6] B. J. Haas. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Research*, 31(19):5654–5666, October 2003.

[7] 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.

[8] 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.

[9] S. Hartmans and J. Tramper. Dichloromethane removal from waste gases with a trickle-bed bioreactor. *Bioprocess Engineering*, 6(3):83–92, February 1991.

[10] G S De Hoog, V Vicente, R B Caligiorne, S Kantarcioglu, K Tintelnot, and G Haase. Species Diversity and Polymorphism in the. 41(10):4767–4778, 2003.

[11] Eija Korpelainen, Jarno Tuimala, Panu Somervuo, Mikael Huss, and Garry Wong. *RNA-seq Data Analysis: A Practical Approach*. 2015.

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

[13] Heng Li and Nils Homer. A survey of sequence alignment algorithms for next-generation sequencing. *Briefings in bioinformatics*, 11(5):473–83, September 2010.

[14] Yang Liao, Gordon K. Smyth, and Wei Shi. FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features, 2014.

[15] 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.

[16] Dion M.a.M Luykx, Francesc X Prenafeta-Boldú, and Jan a.M de Bont. Toluene monooxygenase from the fungus Cladosporium sphaerospermum.

Biochemical and Biophysical Research Communications, 312(2):373–379, 2003.

[17] 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.

[18] Caroline Poyntner. *Bioremediation of waste gas and soil by black extremotolerant fungi*. PhD thesis, 2014.

[19] 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.

[20] Noura Raddadi, Ameur Cherif, Daniele Daffonchio, Mohamed Neifar, and Fabio Fava. Biotechnological applications of extremophiles, extremozymes and extremolytes. *Applied Microbiology and Biotechnology*, pages 7907–7913, 2015.

[21] M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47–e47, 2015.

[22] L J Rothschild and R L Mancinelli. Life in extreme environments. *Nature*, 409(September 2000):1092–1101, 2001.

[23] 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.

[24] 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.

[25] Katja Sterflinger. Black Yeasts and Meristematic Fungi: Ecology, Diversity and Identification. In Gábor Péter and Carlos Rosa, editors, *Biodiversity and Ecophysiology of Yeasts SE - 20*, The Yeast Handbook, pages 501–514. Springer Berlin Heidelberg, 2006.

[26] 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.

[27] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael a Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, and Jill P Mesirov. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545–50, 2005.

[28] 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.

[29] Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-Seq : a revolutionary tool for transcriptomics. *Nature reviews. Genetics*, 10(1):57–63, 2010.

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

List of Figures

List of Tables