Comparative genome analysis, identification of biocontrol genes
BIN6002 – Summer 2022

(Supervisor B. Franz LANG)

TASK: Determine in a set of fungal plant symbionts genes that are potentially involved in microbial biocontrol; track the evolutionary origin of these genes, which may have been acquired from other eukaryotes or bacteria.

BACKGROUND: Mycorrhizal fungi (i.e., fungi living in symbiotic associations with host plants) have the capacity to promote plant growth, suppress plant pathogens and microbial competitors in general (referred to as biocontrol), and increase the plant's survival under adverse environmental conditions. The understanding of the molecular and genetic basis of these symbiotic interactions remains largely unknown, in particular in the group of ericacean plants (such as blueberry, cranberry, rhododendron, etc.) that do not associate with the otherwise most widespread type of plant symbionts called the Arbuscular Mycorrhizal Fungi (AMF; taxon Glomeromycotina). Instead, ericoid plants interact with a variety of fungi from a separate taxon, the Ascomycota (most of them belonging to the Helotiales order).

The laboratory of B. F. LANG has isolated a large number of fungal symbionts from cranberry plants, and has sequenced the genomes and transcriptomes of several species to get insight into their coding capacity, in particular with respect to their symbiotic interactions. Among the genes of interest are those involved in biocontrol, the topic of this project. The majority of these genes participate in the synthesis and secretion of secondary metabolites, notably non-ribosomal polypeptides or polyketides. Genes encoding non-ribosomal peptide synthetases (NRPKs) and polyketide-synthetases (PKSs) are known across bacteria and eukaryotes. Currently, their evolutionary origin and distribution is uncertain, probably because of frequent gene losses and acquisitions by horizontal gene transfer.

QUESTIONS: What are the sets of NRPK and PKS genes in the given set of fungal genomes? To what extent are the gene sets conserved across fungal ericoid symbionts? What is the deeper origin of these genes, i.e. which organismal group(s) are the donors, other fungi, bacteria?

OBJECTIVES: 1. Complete the structural and functional nuclear genome annotation of two fungal symbionts (with the help of Matt Sarrasin). 2. Predict NRPK, PKS and related genes involved in biocontrol based on the genome sequence. 3. Group genes by protein-sequence similarity (and if possible, infer phylogeny). Identify potential gene origins by comparing against the sequences in GenBank nr.

RECOMMENDED TOOLS: genome annotation pipeline (developed in-house by Matt Sarrasin); antiSMASH for the identification of NRPK, PKS etc. in annotated genome sequences. BlastP against GenBank nr. If a sufficiently large sets of aligned proteins can be compiled, use Muscle or alternatives for multiple protein alignment; reconstruct phylogenetic trees with RaxML or PhyloBayes (Bayesian phylogenetic inference).

INTRODUCTORY READING: Review explaining functioning of NRPS and PKS: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3131160/

antiSMASH 5.0 and 6.0: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3125804/pdf/gkr466.pdf https://academic.oup.com/nar/article/49/W1/W29/6274535 (as a start for further literature search).

PROVIDED DATA & INFORMATION: genome assemblies from about six fungal symbionts, four of which already annotated. Transcriptome data for high quality structural genome annotation.

PREREQUISITS: ease working with the Linux operating system; scripting skills desirable; understanding of sequence similarity, homology, protein domains, evolution and horizontal gene transfer.

HARDWARE: All tools require a Linux operating system. A computational server with installed software is available, but you can install and run the tools on your own computer as well.

Transcriptomics: temperature-dependent changes in gene expression

(Supervisor Matt Sarrasin; Team:) - version 27 Apr 2022

TASK: Quantitative comparison of the transcript levels observed under different temperature conditions.
BACKGROUND: Climate change, and particularly ocean warming, affects marine ecosystems, which in turn threatens human food security and health, and causes increased extreme weather and the loss of coastal land (IUCN; https://www.iucn.org/resources/issues-briefs/ocean-warming). An organismal group of choice for studying the effect of ocean warming are diplonemids —unicellular, flagellated, heterotrophic eukaryotes. Widely unnoticed until a decade ago, they have been recently recognized as one of the most species-rich groups of marine protists. Diplonemids have an immensely broad ecological distribution colonizing the marine surface layer down to the permanently dark deep ocean, planktonic and sediment habitats, and all geographic regions from the tropics to the poles. Due to their abundance, distribution, and diversity, diplonemids are likely to play an important role in the marine food web. However, we have little data regarding the response of this group to diverse environmental stimuli, and to temperature shifts, in particular.

An international collaboration headed by the laboratory of Gertraud BURGER has recently determined and annotated the genome sequence of *Diplonema papillatum*, the type species of diplonemids. This particular species is broadly distributed in surface waters of coastal regions of the Atlantic and Pacific. While apparently preferring temperate regions (15–22 °C), it displays notable temperature tolerance thriving in waters as cold as 6 °C and as warm as 30 °C. This cold and heat resilience is most likely due to the differential expression of certain genes, but their nature, molecular function, and the biological process they participate in, are yet to be unraveled.

QUESTIONS: Which genes are specifically up- and down-regulated when the organism is cultivated at different temperatures? Which genes follow a common pattern of differential expression? In which metabolic pathways or biological processes are differentially expressed genes implicated?

STEPS INVOLVED:

- Processing of raw RNA-Seq reads obtained from cultures grown at 4, 15, and 27 °C;
- Mapping of RNA-Seq reads onto the provided function-annotated genome sequence;
- Quantification of expression by gene-based read counting;
- Normalization, transformation of read counts, and differential expression analysis;
- Assignment of Gene-Ontology terms to select gene groups;
- Enrichment analysis of over- and under-represented classes within the GO categories.

TOOLS RECOMMENDED: Bamtools, MUltiQC, DESeq, IGV, etc.
PROVIDED DATA & INFORMATION: raw Illumina RNA-Seq reads (fastq); assembled reference genome sequence (fasta) and annotation (gff3).
INTRODUCTORY READING: http://chagall.med.cornell.edu/RNASEQcourse/Intro2RNAseq.pdf ;
PREREQUISITS: Moderate scripting skills, basics in eukaryotic gene expression, metabolic pathways, and the Gene-Ontology framework.

Proteomics: regulation level of proteome expression

(Supervisor Gertraud Burger, Matus Valach; Team:)

TASK: Determine whether protein levels are regulated at the mRNA level or translationally/post-translationally.

BACKGROUND: Cells regulate meticulously the amounts of their proteins, a process called protein homeostasis or briefly, proteostasis. Proteostasis refers to the balance between biological pathways within cells that control the biogenesis, folding, trafficking, and degradation of proteins. The levels of certain proteins are strictly maintained, while others are modulated according to physiological, developmental or environmental conditions. Proteostatic regulation takes place at the level of both mRNA and protein by modulating (i) gene transcription, and processing and degradation of mRNAs, and then (ii) translation and modification, transport, and degradation of proteins.

The most rigorous approach for finding out at which level protostasis is regulated would be to determine the rates of each process, which, in practice, is rarely feasible. But what can be easily measured is the steady-state amount of transcripts and proteins. For instance, when high mRNA levels coincide with high protein levels, then proteostasis is mainly controlled during transcription or post-transcriptionally. Alternatively, if proteins whose mRNAs have similar steady-state levels occur in very different amounts in the cell, then proteostasis is mainly controlled during or after translation. Knowing at which step regulation occurs builds the foundation for subsequent studies to unravel the interplay of biological processes in proteostasis.

Homeostasis of the eukaryotic proteome is assumed to be predominantly regulated at the RNA level. However, this notion is based on studies of a few model organisms and might not apply to other species. To test the hypothesis that transcripts are the main handle in eukaryotic proteostasis, under-explored microeukaryotes have to be examined. Here we chose *Diplonema papillatum*, the type species of an ecologically important group of marine flagellates, the diplonemids.

QUESTIONS: Is the level of proteins correlated with the steady-state level of their mRNAs in *Diplonema*? Is this correlation maintained across various environmental conditions? Do certain classes of proteins display different patterns of transcript-to-protein ratios? Is there a link between mRNA:protein ratios and particular biological processes or biochemical pathways?

OBJECTIVES: 1. Analyze proteome data to quantify proteins of the organism cultured under normal vs low oxygen conditions, and fed on a regular vs. peptide-rich diet. 2. Map transcriptome reads on a reference transcriptome to quantify individual transcript levels. 3. For each gene, compare protein and transcript steady-state levels. 4. Compare the effects across four different growth conditions.

STEPS INVOLVED:

- Perform peptide identification in raw mass-spectrometry data.
- Quantify proteins by two different measures (spectral counting and ion intensity) and normalize to the number of theoretically observable peptides.
- Map RNA-Seq reads onto a provided reference transcriptome;
- Quantify expression by counting reads per transcript and normalize to the length of the transcript
- Correlate transcript and protein steady-state levels, group genes according to transcript-to-protein ratios;
- If time allows, perform Gene Ontology-term enrichment on interesting bins.

TOOLS RECOMMENDED: E.g. Philosopher, IonQuant, rsem, samtools, bedtools, Blast suite.

INTRODUCTORY READING: https://en.wikipedia.org/wiki/Quantitative_proteomics; http://chagall.med.cornell.edu/RNASEQcourse/Intro2RNAseq.pdf ; Li & Dewey *BMC Bioinformatics* 12, 323 (2011) https://doi.org/10.1186/1471-2105-12-323

PROVIDED DATA & INFORMATION: Raw proteome data from MS/MS experiments, raw Illumina RNA-Seq read files, and reference transcriptome and proteome sequences.

PREREQUISITS: ease working with Linux operating system; knowledge of basic Perl (or Python or R) and scripting skills desirable (but not required); understanding basics of eukaryotic gene expression, regulation of transcription and translation, and the Gene Ontology framework.