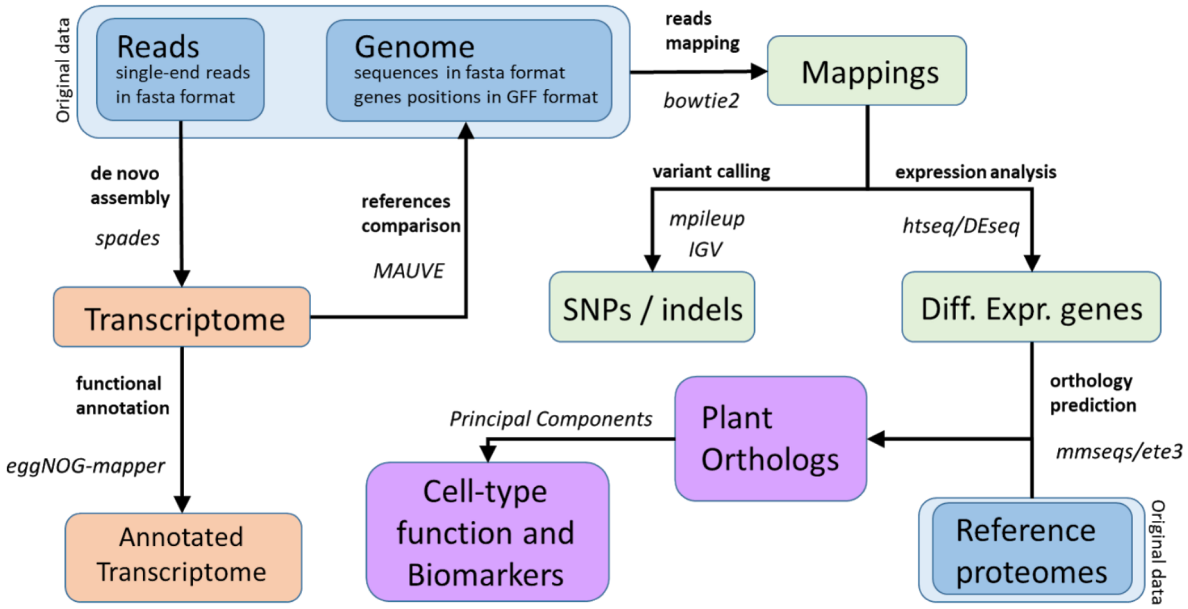
Final project

Genomic Data Analysis and Visualization

"Your laboratory has just received a new cyanobacterial strain that shows strange behaviour when its growth condition is switched from light to dark. This cyanobacterium, isolated from a German lake where cyanobacterial blooms happen very often, is still misclassified. Your supervisor is very excited with this organism and she wants to find out which genes could be responsible for such a phenotype. For that purpose, the lab sent for sequencing four RNA samples, including two biological replicates from cultures of this cyanobacterium grown in either light or dark conditions. The sequencing facility will perform quality checking to provide only high quality reads in fasta format. A few days later, the RNAseq data have just arrived… That means work to do for the bioinformaticians!"

This is an overview of your planned work:



E-mail for questions: gdav18@googlegroups.com

Exam Report. Answer all questions using this form:
https://docs.google.com/forms/d/e/1FAIpQLSfIYEWL-ZVNz-h9dvf1aO8D3AEOIdAGqYWdgFgaYY7M6Aj1iA/viewform
Deadline: Jan 25th 2019, 18:00 CET

Exercise description

Data provided for the practice. You should have a compressed file containing the following files:

- Sequencing reads: these are located within the `fastq` folder. There should be 4 fasta files.
- Reference genome and proteome: comprises 3 files:
  - Fasta sequence of the reference genome in `reference_genome.fna`
  - Fasta sequence of the reference proteome in `reference_proteome.faa`
  - Gene annotation, i.e. what are the gene positions on the genome sequence, in `reference_genome.gff`.
- Proteomes from other species to be used in the orthology analysis are within the `reference_proteomes` folder.
- Some useful scripts and program binaries are provided in the `bin` folder.

1. Preliminary inspection of your data

"You sit in front of your computer. Your coffee cup is still smoking and everybody is silently hitting the keyboard in the lab. You open a bash terminal and `ls` the directory where you have previously downloaded your reads from the FTP server of the sequencing facility. There they are. First things first… you want to be sure what you are dealing with before starting to perform any analysis."

Task 1.1

Check your samples and answer the following questions:

1. How many samples do you have?
2. Could you identify the biological replicates and the ones from the dark and light growing conditions?
3. How many reads do you have in each of your samples?
4. What kind of reads are they? (e.g. paired-end reads, mate-pair, single-end…)
5. Are all the reads of the same length?
6. Just from the files you have been provided, could you say something about reads orientation (5' to 3', 3' to 5')? And what about DNA strand (forward or reverse strand)?
7. Is there any additional comment you would like to do about your reads?

2. Transcriptome *de novo* assembly and annotation

"After checking the literature and the databases it seems that there is not a good reference genome for your species. In light of this, you decide to set up your own *de novo* transcriptome assembly to use it as a reference in downstream analyses. Fortunately, you already know the SPAdes assembler from you TFM practice lessons..."

Task 2.1

As you have different samples from the same organism, you decide to use them all. Therefore, the first thing you need to do is to **join the 4 samples together.**

Now you are ready to run SPAdes (https://github.com/ablab/spades) (tip: check SPAdes parameters, take a look at the `--rna` parameter). Run SPAdes to **perform a *de novo* transcriptome assembly** and try to answer the next questions:

1. What is the full command line which you used to join your 4 samples together? (tip: you could use `cat`).
2. What is the size of the new file obtained with the reads from the 4 samples? (tip: you could use `ls -l`).
3. How many contigs do you have in your *de novo* assembly? (tip: you could use `grep -c` to count the number of fasta headers in your `contigs.fasta` or `transcripts.fasta` file).
4. What is the length of the 10 longest contigs? What is the length of the 10 shortest contigs? (tip: you could use `grep -v` to avoid the headers and then count the length of each line with `awk` printing the result of the `length($0)` function. Then, you could pipe the `awk` result to a `sort -n` and then either use `head` or `tail` at the end of your command).
5. Could you use your de novo assembled transcriptome as a reference to carry out a ChIP-seq analysis?

Task 2.2

"Once you obtained your new assembly you realized that you ignore what do your sequences encode for. Thus, you decide to carry out a functional annotation of your transcripts: a good assembly with functional annotation is something you would call a transcriptome reference."

Try to obtain **functional annotation** for your first transcripts using eggNOG-mapper, with `Diamond` mapping mode all the other options by default. *Note that eggNOG-mapper might take a few hours to run for all your transcripts.*

1. How many functional annotations did you obtain with eggNOG-mapper?
2. How many of your transcripts are lacking functional annotation? (tip: you could download the eggNOG-mapper results as a table and use the next command line, guessing by yourself what the `XXX` command should be: `cat transcripts.fasta.emapper.annotations | cut -f 1 | sort | XXX | wc -l`).

Task 2.3

"Just when you were about to annotate the whole transcriptome you receive an email from the guy who is sitting 1 meter apart in the bioinformatics lab. He has sent you the *doi* of a new paper which was published yesterday: a new top quality reference genome for your cyanobacterial species. You decide to compare your transcriptome with the new genome to use the best one for further analyses."

First, try to answer the next questions about the new reference genome:

1. How many sequences are in the reference genome file?
2. What is the size of the reference genome?
3. How many genes are annotated in the GFF file included with the genome?
4. What is the total length of the annotated genes?
5. Based on that and the length and number of reads of one of your samples, what would be the average depth of coverage you would obtain after mapping them to the genes? Do you think that you will need more coverage for an RNAseq analysis? And to perform a variant calling analysis?

Task 2.4

Compare your transcriptome to the reference genome of your species using Mauve (http://darlinglab.org/mauve/mauve.html).

- Take a snapshot of your the MAUVE contig alignment

3. Read mapping

"In light of the comparison with your transcriptome, you decide to use the new genome as a reference. You will start by mapping your reads to the reference. You will later use these mappings to perform variant calling and differential expression analysis."

Task 3.1

First, you will need to **create an index of the reference genome** (tip: use the `bowtie2-build` command).

Next, **map each of your samples to the reference genome** using Bowtie2 (http://bowtie-bio.sourceforge.net/bowtie2/). (tip: check the `bowtie2 --help` for a parameter which allows you to use fasta instead of fastq files as input; also make sure to redirect the stderr output of bowtie2 to a file using the `2>` redirection, so that you can collect bowtie2 mapping stats).

1. What is the alignment stats of your 4 samples?
2. How many mapping records are in your mapping files? How many different reads are in your mapping files? How these numbers compare with the number of reads in your samples and with the alignment statistics?
3. Could you use these mappings to perform an analysis of Copy Number Variation?

4. Variant calling

"The next step is performing variant calling to compare the DNA of your strain with that of the reference genome. Maybe some mutation is responsible for the unique behaviour of these German cyanobacterias..."

Task 4.1

To perform variant calling you can use the mappings from all your samples. To do that, **sort the mappings** of each of your samples (tip: you should use `samtools sort` for that). Then, you can **merge the sorted files** into a single BAM file (tip: you should use `samtools merge` for that). Try to do all these steps so that the header information is present in the final merged BAM file (tip: check the parameters available in the help of the different `samtools` commands).

Finally, **perform the variant calling** using `samtools mpileup` and `bcftools`. Check the parameters used in the practice lessons and apply those. *Important note: if you are using the Virtual Machine of the course you should be safe using the practice lessons, however beware that if you are using a more recent version of samtools/bcftools the commands and parameters can change a lot (e.g. in some versions you don't need to use `samtools mpileup` and you just use instead `bcftools mpileup -Ou -f genome.fna mappings.bam | bcftools call -mv -Ob -o calls.bcf`).*

Transform, to a "tab" separated file, the BCF file you obtained, including the variant ID, its position, the reference allele, the alternative allele, the variant quality and the depth of coverage of the variant (tip: use `bcftools view calls.bcf | grep -v "^#" | cut -f 1,2,4,5,6,8 | sed 's#DP=\([0-9][0-9]*\).*#\1#' > calls.tsv`. You could run the previous command line with and without the `sed` command to check what is its purpose).

1. How many variants did you obtain? How many are SNPs, how many insertions and how many deletions?
2. How many variants have quality greater or equal than 10?
3. How many variants have depth of coverage greater or equal than 10?
4. Find the variant with the best quality. Could this variant be affecting a gene (tip: compare the positions of the variant in your BFC file with the positions of the genes in the genome GFF file until you find one variant within a gene). Which gene did you find if any? Without actually checking it, could you give an example of how the variant could be affecting the protein encoded by such gene?
5. Use the file with all the reads and the fasta file of the genome to locate the variant from the previous question (4.) in IGV and capture the image of the variant in the reads.

5. Differential expression analysis

"As you have only RNAseq reads, any mutation affecting regulatory sequences would be missing from the variant calling. Luckily, you can also compare the expression differences between the samples grown under light and dark conditions, which could provide some additional clue of what might be happening."

Task 5.1

For differential expression analysis (DEF) you need to start using the sorted bam files generated in the task 3.1. First of all, you need to do a "read-count" using `htseq-count`. Some important parameters: you have to use `***-i old_locus_tag***` *and* `***-t gene***`. (tip: you need to use a .gff file to count the reads). Once you have the four count files is necessary to merge all together and

save it as `count.txt` file (tip: explore `join` command). You should finally obtain something like this in your ***count.txt*** file:

```
> cat count.txt
#genes      #counts (totally arbitrary values shown only for this example)
SYNGTS_0001 12 22 2 3
SYNGTS_0002 30 0 22 0
SYNGTS_0003 10 10 10 10
…                  ….
```

Now you can use your counts to perform the DEF analysis. For that use the Bioconductor package DESeq2, using these loading data parameters:

```
> cat my_DESeq2_script.R
# Loading Data in R #

counts = read.table("counts.txt", header=F, row.names=1) # Load the raw counts table
colnames = c("Dark","Dark","light","light") # names for column names
my.design <- data.frame(row.names = colnames( counts ),
                        group = c("dark","dark","light","light")
) # our experiment design for DESeq2 analysis
```

If you inspect in R the data frame of your experiment design (my.design variable) it has look like this:

*Note: this script installs and uses bedtools (https://bedtools.readthedocs.io) to obtain the fasta sequence of a gene, given the genome fasta and the position of genes in the genome (GFF, BED, …). For each gene in the `genes.list` file, a .fna file will be created, using the identifier of the gene and containing its fasta sequence.**Note: this script installs and uses bedtools (https://bedtools.readthedocs.io) to obtain the fasta sequence of a gene, given the genome fasta and the position of genes in the genome (GFF, BED, …). For each gene in the `genes.list` file, a .fna file will be created, using the identifier of the gene and containing its fasta sequence.*

	group
V2	dark
V3	dark
V4	light
V5	light

Important note: *#rlogtranformation for PCA analysis section of the DESeq2.R script used in the practical session have to be silenced or you will have an error message!!.*

Now you are ready to do you DEF analysis:

1. Have a look to the p-adj histogram obtained (`res$padj`), what does this result means?
2. How many genes showed an statistical (p-adj < 0.01) differential expression?. The results has to be justified with a table showing all the altered genes (including, p-val, p-adj, fold change).
3. Taking all this data together, what can you say about the statistical significance of your DEF? Do you feel confident about your DEF result?

6. Orthology prediction

"The differential expression results gave you an idea about the potential roles of those genes in modern plants: are those genes conserved in plants? And, if so, do they conserve their ancestral function?

To check your hypothesis you decide to predict orthology relationships between your target genes and the model species *Arabidopsis thaliana*. You already know about the importance of good orthology predictions, so you decide to use a phylogenetic approach including proteomes from several other organisms."

Task 6.1

Identify orthologs of your cyanobacterial differentially expressed genes in the *Arabidopsis thaliana* genome.

Suggested phylogenomic workflow:

1. **Concatenate the proteomes** under the folder `reference_proteomes` plus your target reference proteome "`reference_proteome.faa` into a **single** FASTA file.
2. Run MMseqs2 in clustering mode on the concatenated fasta file to **detect gene family clusters** including your reference proteins. (tip: use `-c 0.1` option in mmseqs cluster to increase sensitivity)
3. Extract each cluster into a FASTA file using the `bin/split_clusters.py`.
4. Identify what clusters contain your protein sequences, and **build a phylogenetic tree** for it (tip: you will see that more than 27,000 clusters are detected. Note that only 3 or 4 will contain your target proteins, so you will only need to identify those (e.g. using grep) and build 3 o 4 trees).
5. Identify duplication and speciation events in the resulting gene trees using `ete3`.
6. **Extract the orthologs of your target proteins sequences** in *Arabidopsis thaliana*.

Questions

1. Report the list of orthologs you found (tip: you should find at least 10 in total, otherwise check your pipeline or ask us for advice).
2. Are the Arabidopsis orthologs involved in any specific function? Are those functions related to the ones you observed for the cyanobacterial genes? (tip: You can look up Arabidopsis gene identifiers at Uniprot).
3. Visualize the trees you obtained (tip: you can use http://etetoolkit.org/treeview, or http://itol.embl.de).
4. Is there any arabidopsis-specific gene duplication event? (tip: duplications involving Arabidopsis thaliana but not other species) .
5. What is the closest species of your sample proteins?

7. Cell-type functions and identification of biomarkers

"Your colleague is telling you of this latest paper in 'Science' about these new molecules and compounds just identified in few plant cell types which appear to have

anticancer properties. You wonder if you could also save the world and decide to check if the Arabidopsis orthologs of your cyanobacterial genes might have specific cell type expression, and thus, if they could be used as tissue biomarkers. Mutations in the Retinoblastoma gene cause cancer, and this gene function is as repressor of cell cycle and stem cell function. You are dying to know if your biomarkers might also work as biomarkers for stem cell function, and thus if they potentially could be used for cancer diagnosis."

Task 7.1

Suggested analyses

1. Visualize expression of Arabidopsis orthologs in the cell-type dataset ("TableCellType.csv") through a heatmap representation. Check dispersion in gene expression using boxplot.
2. Check cell type transcriptome separation using PCA and Arabidopsis orthologs as "variables"/parameters. Check component contribution to cell-type variability. Represent the loadings in a plot and identify isolated cell types. Represent the scores and identify contribution of each gene to sample separation (tip: you do not need to run all transcriptomes together, but you need to relate one another. Additionally, you may want to consider that some transcriptomes have some redundancy, such as -CO2 and COR-, -SCR and E30-, -S18, S4 and APL-, -S17 and S32- and -GL2 and WER-, thus some of these transcriptomes might be perhaps removed or combined. For more information check slide 27 in "Tema5.pdf".
3. Analyze stem cell regulator mutant (*shr*) and complemented transcriptomes in SHR/stem-cell dataset ("TableStemCell.csv") using PCA. Check component contribution to variability in *shr* mutant and its complemented versions. Represent the loadings in a plot and identify if Arabidopsis orthologs' expression recapitulate stem cell recovery, i.e they behave as biomarkers for stem cell function. Represent the scores and identify contribution of each gene to sample separation.

Questions

1. Do these Arabidopsis orthologs have specific cell type expression?
2. Do these genes separate any cell type?
3. Do they separate any of the cell type(s) in which they show enriched expression?
4. Do these genes show dispersed expression in the cell type(s) they separate?
5. Could all of them be used as biomarkers? Which one(s) is(are) potentially the best(s) tissue biomarker(s)?
6. Do you think biomarker gene function(s) (or that of their cyanobacterial orthologs) might associate with one specific cell type?
7. Do(es) the cell type(s) they separate associate with stem cells?
8. Might these Arabidopsis orthologs be used as stem cell function biomarker(s)?
9. Which one(s) is(are) potentially the best(s) stem cell function biomarker(s)?
10. Might any of these genes be used as cancer biomarker? (tip: our hypothesis is that cancer might be devoid of stem cell function).