WGGC Autumn School 2022

CBI

Exported on 09/22/2022

Table of Contents

Project: Maize gene expression in four genotypes under drought and control conditions

1 Part 1: Get raw read files

Go to one of the internationally coordinated primary databases for sequence information:

–EMBL (ENA, UniProt) http://www.ebi.ac.uk/ena http://uniprot.org

–NCBI (GenBank, nr) http://www.ncbi.nlm.nih.gov

–DDBJ www.ddbj.nig.ac.jp[1]

and search for the maize data with accession **SRP058750**

You should find 32 short read .fastq files. You do not need to download all of them, you can limit your analysis to one genotype, e.g. B73.

1 http://www.ddbj.nig.ac.jp

2 Part 2: FastQC and Trimming

1) Check your sequence files

Your sequence files should contain the following number of sequences:

B73_con_1.fastq 8085432
B73_con_2.fastq 22252697
B73_con_3.fastq 30180272
B73_con_4.fastq 25459356
B73_drought_1.fastq 12832759
B73_drought_2.fastq 19119496
B73_drought_3.fastq 23522007
B73_drought_4.fastq 23629579
BxM_con_1.fastq 20020741
BxM_con_2.fastq 38101382
BxM_con_3.fastq 36301348
BxM_con_4.fastq 29114503
BxM_drought_1.fastq 30529098
BxM_drought_2.fastq 20047868
BxM_drought_3.fastq 15611343
BxM_drought_4.fastq 36086599
Mo17_con_1.fastq 25510816
Mo17_con_2.fastq 47712454
Mo17_con_3.fastq 21797918
Mo17_con_4.fastq 16207101
Mo17_drought_1.fastq 26595536
Mo17_drought_2.fastq 20519875
Mo17_drought_3.fastq 15943346
Mo17_drought_4.fastq 35003386
MxB_con_1.fastq 24528086
MxB_con_2.fastq 23173809
MxB_con_3.fastq 36593846
MxB_con_4.fastq 16845158
MxB_drought_1.fastq 19716300
MxB_drought_2.fastq 24501823
MxB_drought_3.fastq 10448405
MxB_drought_4.fastq 30966840

You can check this using UNIX, or using the FastQC output, see below.

2) FastQC

Create FastQC reports using commands like:

/path/to/fastqc -o /path/to/yourOutputDirectory/ /path/to/rawreads/B73_con_1.fastq

View them by opening the .html outputs.

Are any problems detected (contaminations, low quality ends, remaining adapter sequences...)?

Consider important statistics (Total number of sequences, filtered sequences, sequence length) and the result of the QC.

Use the FastQC tutorial video and documentation if you need help.

https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/

3) *Trimming*

Read the Trimmomatic manual; (http://www.usadellab.org/cms/uploads/supplementary/Trimmomatic/TrimmomaticManual_V0.32.pdf) to understand the options in the following command line.

Change the command line, if you want to change the default settings to fit your dataset.

java -jar /path/to/Trimmomatic-0.38/trimmomatic-0.38.jar SE -threads 6 -trimlog B73_drought_1.trimlog B73_drought_1.fastq B73_drought_1.trim.fastq ILLUMINACLIP: LEADING:15 TRAILING:15 MAXINFO:30:0.8 MINLEN:36

This will take a LONG time due to data being read from disk.

4) Compare QC before and after trimming

- Rerun FastQC on the output of the trimming step and compare the results with the original FastQC output.

- What happens to sequence length distribution after trimming?

Advanced:

Redo the trimming with other parameters.

Why could that be interesting? What are the important parameters?

3 Part 3: Alignment and IGV

1) Index the genome

Download the maize genome sequence .fasta and annotation .gtf files from https://plants.ensembl.org/info/data/ftp/index.html or https://gramene.org and ungzip it. The file with all chromosomes is called "toplevel", see the readme.

Copy the maize reference genome to your project directory (replace *NAME* with your folder):
cp /path/to/sequences/Zea_mays.Zm-B73-REFERENCE-NAM-5.0.dna.toplevel.fa path/to/workingdirectory

In order to use HISAT2, you first have to build an index from the genome sequences using hisat2-build. This again is quite time consuming.

List the sequence (most importantly the chromosome-) names in the index and make sure that the names match the values in the first column of the GTF file.

2) Alignment using HISAT2

Start the alignment to the maize genome by adapting this to your directories and dataset:

hisat2 -p 4 -x /path/to/base-name-of-genomeindex /path/to/inputfastqfile -S /path/to/hisatoutput.sam

The log for each alignment (how many reads did and didn't align, etc.) is printed to standard error in hisat2, so be sure to save this for reference..

3) Prepare IGV

Download IGV to your local computer (Manual: http://www.broadinstitute.org/software/igv)

StartIGV.
Select Genomes → Load Genome from File → choose the maize genome sequence FASTA file
Select File → Load from File → choose the annotation .gtf file

Advanced: Find out what gtf stands for.

4) Prepare your genome alignment for IGV

Sort and index the .sam files for use with IGV → creating sorted BAM files

(see https://www.broadinstitute.org/software/igv/BAM and http://www.htslib.org/doc/samtools.html for samtools manual):
e.g.:
/path/to/samtools sort Mo17_con_1_trimmed.sam -o Mo17_con_1_sorted.bam
/path/to/samtools index Mo17_con_1_sorted.bam

Copy the sorted bam file plus the .bai index file to your local computer.
You will need to do this for all your samples, making sure you name them properly.

Load a control and drought dataset into IGV. Use: File -> Load from file

Answer the following questions:

IGV 1):

Find a gene with sufficient mapped RNAseq reads. Are all exons and introns validated by the RNAseq read alignments? Check the 5' and 3' end of the gene; is there a defined transcription start and stop?

IGV 2):

Find gene GRMZM2G126900

Does it look as if its expression level is the same in control and drought conditions?

Are all introns validated by the RNAseq data?

Specifically what do you think about GRMZM2G126900_T03?

What do you conclude for the gene model in the maize genome annotation?

Bonus question:

Can you show heterozygosity using data from B73, Mo17 and hybrids in gene GRMZM2G140799?

(Tip: have a look at SNPs, especially at 9:143,648, 807)

4 Part 4: Read count quantification (featureCounts)

1) Run featureCounts to summarize read counts on both gene and transcript level. See e.g. scripts/ featureCounts.sh.

```
# gene level
featureCounts -T 4 -O -t exon -g gene_id -a /path/to/zea_mays.protein_coding.gtf -o /path/to/gene-level/
total_file.count \
B73_con_1_trimmed_sorted.bam \
B73_con_2_trimmed_sorted.bam \
B73_con_3_trimmed_sorted.bam \
B73_con_4_trimmed_sorted.bam \
B73_drought_1_trimmed_sorted.bam \
B73_drought_2_trimmed_sorted.bam \
B73_drought_3_trimmed_sorted.bam \
B73_drought_4_trimmed_sorted.bam


# isoform/ (previously established - gtf file based) transcript level
featureCounts -T 4 -O -t exon -g transcript_id -a /path/to/zea_mays.protein_coding.gtf -o /path/to/transcript-level/
total_file.count \
B73_con_1_trimmed_sorted.bam \
B73_con_2_trimmed_sorted.bam \
B73_con_3_trimmed_sorted.bam \
B73_con_4_trimmed_sorted.bam \
B73_drought_1_trimmed_sorted.bam \
B73_drought_2_trimmed_sorted.bam \
B73_drought_3_trimmed_sorted.bam \
B73_drought_4_trimmed_sorted.bam
```

→ the tables will be saved by us at https://github.com for easy usage in the following step

5 Part 5: Differential expression

Analysis:

Open the Jupyter Notebook https://nbviewer.jupyter.org/github/tgstoecker/teaching/blob/master/AppliedBioinformatics/Notebooks/WGGC_diff_exp_edgeR.ipynb

First save your own copy (e.g. to Google Drive).

Follow the notebook and change it according to your own genotype when necessary (the notebook was created using B73 data as an example).

Questions Differential Expression:

At the default level of significance (FDR/q-value < 0.05), how many differentially expressed genes do you find between control and drought treatment for your dataset/genotype?

How many genes are differentially expressed at significance level 0.01?

At the default level of significance, how many genes are upregulated between control and drought? How many are downregulated?

6 Part 6: Mapman - Differential expression in the context of function and pathways

*The basic use of Mapman will be shown in a **short presentation** and is also documented in the following:*

1) Download Mapman 3.6 from http://mapman.gabipd.org and install. The installation instructions are not completely up to date. For Windows and MacOS, the installers usually work, but maybe not on more current systems. Then you will need to download the .jar file for MapMan and install Java yourself.

- You may have a current Java already installed, in that case you can use the .jar file and start MapMan using java -jar MapManInst.jar. You can check for an installed Java using java -version in the MacOS Terminal or Windows console (Pre-windows10: In the start menu, select Run... and type cmd. Windows10: Click start, then type "cmd" in the search box.)
- Else, you need to download Java from https://www.oracle.com/java/technologies/downloads/
- Windows: You need to install Java with administrator privileges, see https://docs.oracle.com/en/java/javase/14/install/installation-jdk-microsoft-windows-platforms.html#GUID-DAF345BA-B3E7-4CF2-B87A-B6662D691840
- Hopefully, your PATH variable will be set so that java will be found. Type java -version in the Terminal or Windows console. If it is not found, you will have to give the full path to the executable. In MacOS, this should be /Library/Java/JavaVirtualMachines/ *versionnumber* .jdk/Contents/Home/bin/java. In Windows, `C:\Program Files\Java\jdk- versionnumber \bin\java`
- Go to the directory where MapManInst.jar was downloaded. Run java -jar MapManInst.jar
- This will create a directory with necessary files for MapMan. In this directory is also a lib folder which contains MapMan_3.6.0RC1.jar
- Go to that directory and start MapMan by running java -jar lib/MapMan_3.6.0RC1.jar (you may have to give the full path for java as above).
- If MapMan runs out of memory (error message e.g. java.lang.OutOfMemoryError: Java heap space or MapMan freezes) you can increase the allocated memory using java -Xmx 2g -jar lib/MapMan_3.6.0RC1.jar This sets the maximum memory allowed for MapMan to 2 Gb. Adjust depending on memory available on your computer, do not assign all available memory or your computer may freeze.

Have the MapMan guide ready http://mapman.gabipd.org/c/document_library/get_file?uuid=0493e69a-d3c2-4278-acad-b58ad1fdede3&groupId=10207 or use the Help in MapMan.

2) Get the Zm_B73_5b_FGS_cds_2012 alignment file. Right-click on the Mappings folder, select New mapping, download, Zea mays/Zm_B73_5b_FGS_cds_2012: 1.1

3) Use the jupyter notebook to create an input table for mapman.

In the notebook you can exchange the input transcript featureCounts table for one you created yourself. The final table you have to download from the filesystem of your virtual google machine to your local computer as shown in the jupyter video.

4) Load your data into Mapman: Right-click the Experiments folder. Select New Folder. Create a folder. Right-click your new folder and select Add data. Find the file with input data you created. Set column 2 (q_value) to Type:derived value (Col:1).

5) Open the triangle/ weird symbol next to your data, and again the triangle next to log2_fold_change. Right-click on Col 2: q_value. Select configure filter. Set e.g. q_value < 0.05.

6) Look at some pathways: Select a pathway, e.g. from Overview. When asked for a mapping, use Zm_B73. When the pathway is displayed, select the Col 1: log2_fold_change from your dataset to display that data. You can select another pathway by double clicking it. Make sure to look at Regulation_Overview and Hormones. Beware of the scale and colours settings!

7) Look for a pathway with upregulated genes, and select one or few upregulated genes. In the InfoTable, you can find the gene id. Find out more information about your genes using the information provided in Mapman or the internet, e.g. using http://bar.utoronto.ca/eplant_maize/ or https://maizegdb.org.

8) Look for a pathway with downregulated genes, and select one or few. Find out more information about your genes using the information provided in Mapman or the internet.

Questions Mapman –

These are guiding objectives; please feel free to investigate whatever genes/transcripts you are interested in, e.g. from your own research.

MM1) Select two pathway images. Discuss where you find significant differential expression.

Are genes up- or downregulated? Can you think of any relation to drought stress?

MM2) Select a strongly differentially expressed gene from any pathway and discuss it: What pathway is it involved in? What can you find out about its function? What can you find out about its expression in other experiments from e.g. http://bar.utoronto.ca/eplant_maize/?

7 FAQ

How to work with R in Google Colab:

Open our Jupyter Notebook.

Expand hidden cells to see some example code.

You need a Google Account to execute code. Click "open in Colab".

The first time you execute any code you will get a warning because the “notebook was not authored by Google”. It was authored by us, so you can just click “RUN ANYWAY”.

Click on the folder symbol on the very left and then click the “Upload” button to copy your counts file to the Jupyter Notebook’s environment.

How to work with R on your workstation:

Download and install R from the R-project’s webpage: https://cran.r-project.org/

Download and install the free version of RStudio from: https://rstudio.com/products/rstudio/download/

Caveat:

Later steps (e.g. the differential expression analysis) require the correct installation of quite a few R packages.

This will be taken care of in the Jupyter Notebooks we will provide to you.

If you want to run R locally, you can try to copy the code to install the packages from the Jupyter Notebooks.