Capstone Project

Diemer Harbers & Douwe Metselaar

16-2-2021

Contents

---

1 Achtergrond

1.1 Inleiding

Een tekort aan ijzer kan de gewasopbrengst sterk beïnvloeden. Daarnaast speelt ijzer een belangrijke rol in het metabolisme en is het nodig voor plantengroei. IJzer fungeert onder andere als cofactor van veel enzymen. Verder is het betrokken bij de elektronentransportketen en fotosynthese. In eerdere onderzoeken is al onderzoek gedaan naar ijzertekorten in verschillende planten, echter is het achterliggend moleculaire mechanisme betreffende ijzertekorten niet bekend voor tarwe. Met behulp van RNA-seq zal getracht worden dit mechanisme te ontrafelen. (Wang et al., 2020)

De gebruikte cultivar is Triticum aestivum cv. Bobwhite S26 s. Zaden werden in gelijke omstandigheden gekiemd met voldoende ijzer aanwezig. Na 7 dagen werd de helft van de zaden overgebracht naar een ijzerarme omgeving. Vervolgens werden de planten gegroeid gedurende 90 dagen. Vervolgens werden van de wortels en bladeren van de planten in de twee condities RNA geïsoleerd. Per sample werden drie biologische replicaten genomen. Na een zuiveringsstap werden de sequencing libraries gemaakt. Na de library preparation werden de samples met behulp van de Illumina HiSeq gesequenced. In totaal werden 12 samples gesequenced. (Wang et al., 2020)

Na de transcriptoom analyse is gebleken dat 5969 genen differentieel tot expressie kwamen in bladeren. In de wortels kwamen minder genen differentieel tot expressie tussen de controle en behandelgroep, namelijk 2591. In figuur 1 zijn de hoeveelheden DEG's in samples samengevat. Een Gene ontology (GO) enrichment analysis is uitegevoerd op de data om inzicht te krijgen welke pathways in respons op ijzertekort verhoogd of verlaagd worden. Uit de resultaten is gebleken dat verschillende genfamilies, MFS, ABC transmporters, OPT en NRAMP, differentieel gereguleerd werden. (Wang et al., 2020)

1.2 Methoden

De reads in het artikel die zijn verkregen door RNA-seq zijn voor verdere verwerking gecontroleerd met FastQC, getrimmd met Trimmomatic en gemapt tegen het genoom doormiddel van Star. Om iets over genexpressie te kunnen zeggen is een kwantificatie gedaan in R met de functie featureCounts van het package Rsubread. De functie featureCounts wordt gebruikt voor een telling van reads die zijn gegenereerd op basis van RNA of DNA-sequencing. Het voordeel is dat het een snelle methode is en weinig computergeheugen vereist. Omdat het genoom van gewone tarwe hexaploid is het nodig om ook de subgenomen te identificeren, hiervoor is HomeoRoq gebruikt. Om een differentiële genexpressie uit te voeren is de functie DESeq2 nodig in R. Dit is een functie die verpakt is in het pakket van Bioconducter. Met de functie DESeq2 worden onbewerkte tellingen aan een NB-model toegekend, daar worden ook statistische test voor differentieel tot expressie gebrachte genen uitgevoerd. In deze stap wordt dus bepaald of de gemiddelde expressieniveaus van verschillende steekproefgroepen significant verschillen. Voor de GO-verrijkings analyse die ook is uitgevoerd in R is gebruik gemaakt van het R-pakket "TopGo". Om de significante GO-termen te berekenen is een gekeken naar de WeightFisher-algoritme. Na alle voorgaande stappen zijn de resultaten gevisualiseerd met ggplot en ggnetwork. Hierin kunnen grafieken en plots gemaakt worden ter visuele assistentie van de theoretische informatie. Daarnaast zijn ook visualisaties gemaakt met VennDiagram en Pheatmap., met deze functies kan met Venndiagrammen maken en headmaps. (Wang et al., 2020)

1.3 Data

De dataset van het artikel is opgeslagen in de GEO database op NCBI in de vorm van een Excel bestand. Dit Excel bestand is opgedeeld in twee onderdelen, de twee onderdelen bestaand uit samples van de wortelen en samples van de bladeren. Van elk onderdeel zijn zes samples, drie controle samples en drie samples met een laag ijzergehalte. De dataset is zo opgebouwd dat in de eerste kolom de reads zijn weergegeven. Deze zijn te herkennen aan de TreasCS naam. In de kolommen 2 tot en met 7 staat de ruwe data van de samples weergeven in de vorm van counts. Naast de ruwe data zijn in de kolommen 8 tot en met 23 nog extra data te vinden. Hierin zijn bijvoorbeeld log2 Foldchanges al weergegeven, Wald Test p-waarden en adjusted p-waarden te vinden, functie omschrijving van de genen en de locatie in de PFAM, GO en Interpro databases.

2 Analyse

2.1 Op basis van count matrix

2.1.1 Data

De data die verkregen is voor dit onderzoek is verkregen uit de NCBI GEO database. Omdat de data niet alle raw gene counts bevatte is actie ondernomen. Met de nieuwe verkregen dataset van de onderzoekers waarin wel alle raw gene counts aanwezig waren is verdere analyse uitgevoerd. Het format van de data is een csv-file. De data ziet er als volgt uit:

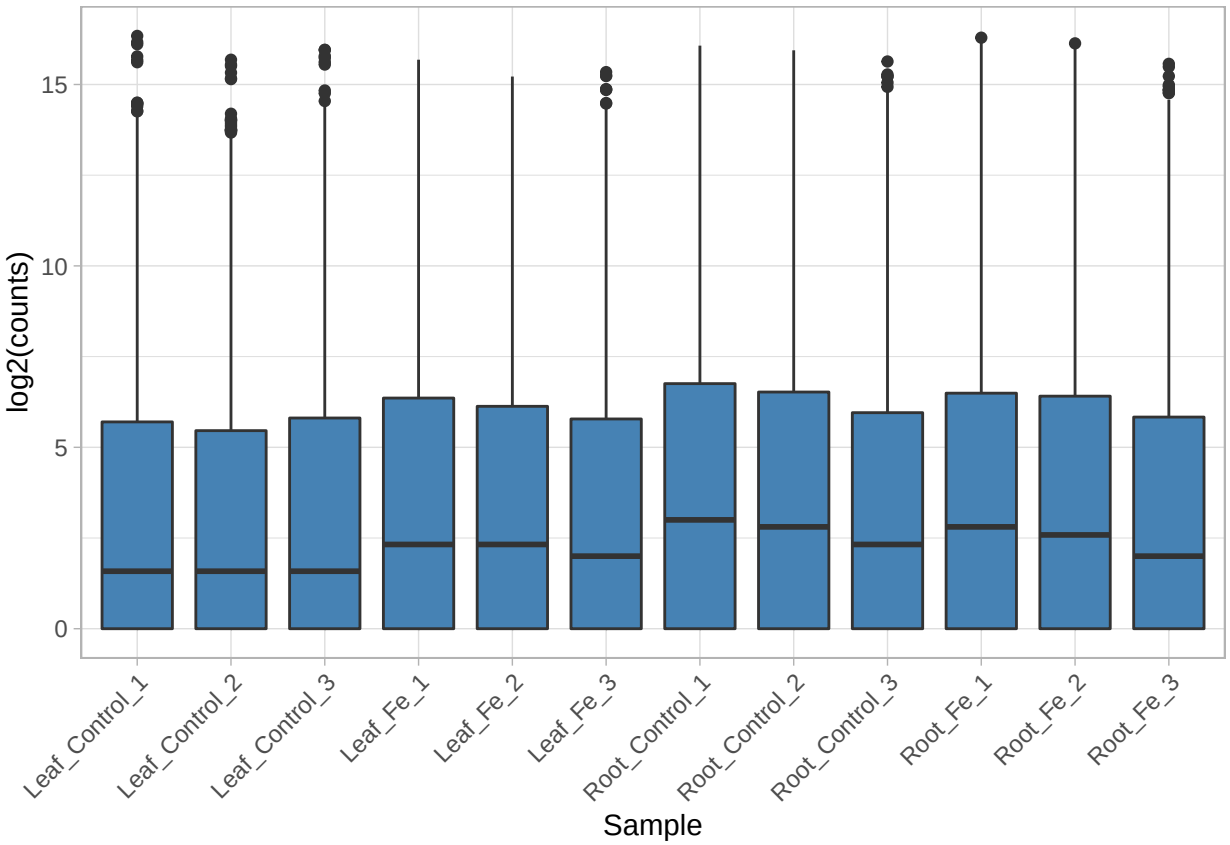
Gene	Leaf_Fe_1	Leaf_Control_1	Root_Fe_1	Root_Control_1
TraesCS1A02G000100	3	4	16	16
TraesCS1A02G000200	12	8	26	20
TraesCS1A02G000300	1	3	4	4
TraesCS1A02G000400	24	20	22	30
TraesCS1A02G000500	4	2	0	1
TraesCS1A02G000600	2	0	0	0

De data is verdeeld in 14 kolommen. In de eerste kolom wordt de gene name weergegeven zoals deze in EnsemblPlants database beschreven staat. In de daarop volgende kolommen staan de samples beschreven met de bijbehorende counts per gene name. Er is gebruik gemaakt van 4 samples, elke sample bevat 3 replicaten. De samples die vergeleken gaan worden op basis van DEG's zijn: Leaf_Fe, Leaf_control en Root_Fe en Root_control. In elke rij staat voor elke gene name het aantal getelde reads per sample.

2.2 Exploratory Data Analysis

2.2.1 Boxplot

Om verder onderzoek te kunnen doen moet eerst de ruwe data geanalyseerd en gevisualiseerd worden. Een veel gebruikte plot om naar de verschillen in aantallen te kijken is de boxplot. De boxplot werkt doormiddel van boxen en is een weergave van een dataset waarbij een minimum, eerste kwartiel, mediaan, derde kwartiel en een maximum worden weergegeven. Daarnaast worden de uitschieters als bolletjes aangeduid. Voor de gebruikte dataset is een boxplot gemaakt. De counts zijn eerst omgezet naar de log2 van de counts, dit is gedaan omdat het nummerieke bereik van de counts erg groot kan zijn. De gecreëerde boxplot wordt hieronder weergegeven:



In de box wordt voor elk sample de mediaan weergegeven. Dit is de dikke zwarte lijn die in elke box te zien is. Het minimum is bij elk sample 0. Dit is te verklaren omdat niet tegen elke gene name een count is gemappt. Het eerste kwartiel is het begin van de box bij 0 en het derde kwartiel wordt weergegeven als het einde van de box. Het maximum wordt weergeven door de dunne lijn die vanuit de box omhoog gaat. Boven het maximum zijn de uitschieters weergegeven, deze zijn weergegeven als bolletjes.

Wat opvalt aan de data is dat de Leaf_Control samples een gelijke mediaan hebben. Daarnaast heeft deze groep wel de meeste uitschieters. De andere drie groepen laten bij de samples een afnemend box niveau zien, er zijn geen grootte verschillen waarneembaar. Op dit gebied is dan ook besloten om geen samples weg te laten in het verdere onderzoek.

2.3 Op basis van fastq

2.3.1 Data

Voor de analyse zijn een aantal bestanden met relevante data. Een referentiegenoom van tarwe is beschikbaar via deze **site**. Een annotatie bestand in ggf3 formaat is te vinden op deze **webstite**, daarbij is gekozen voor het `iwgsc_refseqv2.1_gene_annotation_200916.zip` bestand. Vervolgens is het volgende bestanden opgeslagen: `iwgsc_refseqv2.1_annotation_200916_HC.gff3`. Ten derde zijn de data afkomstig van de sequencer nodig. De accession code voor het BioProject op de NCBI website is **PRJNA680330**, vanaf hier zijn links te vinden naar de 12 SRA pagina's van het onderzoek.

2.3.2 Stap 1: SRA bestanden downloaden

De eerste stap in de analyse is het downloaden van de `.fastq` bestanden van SRA. De volgende 12 samples moeten gedownload worden:

- SRR13114670
- SRR13114671

- SRR13114672
- SRR13114673
- SRR13114674
- SRR13114675
- SRR13114676
- SRR13114677
- SRR13114678
- SRR13114679
- SRR13114680
- SRR13114681

Met het onderstaande commando wordt de data van één sample gedownload. Dit resulteert in twee `.fastq` bestanden, één met de forward reads en één met de reverse reads.

```
fasterq-dump SRR13114670
```

2.3.3 Stap 2: QC

Op de `.fastq` bestanden wordt als eerste een kwaliteitscontrole uitgevoerd worden, hiervoor wordt gebruik gemaakt van FASTQC. In figuur 1 is het resultaat van een kwaliteitscontrole op de forward reads van de ruwe data van sample Roots_Low Fe_1 te zien. In de figuur is te zien dat er zich veel spreiding in de kwaliteit van de data aanwezig is. Daar kan voor gecorrigeerd in de volgende stap: trimmen en filteren.

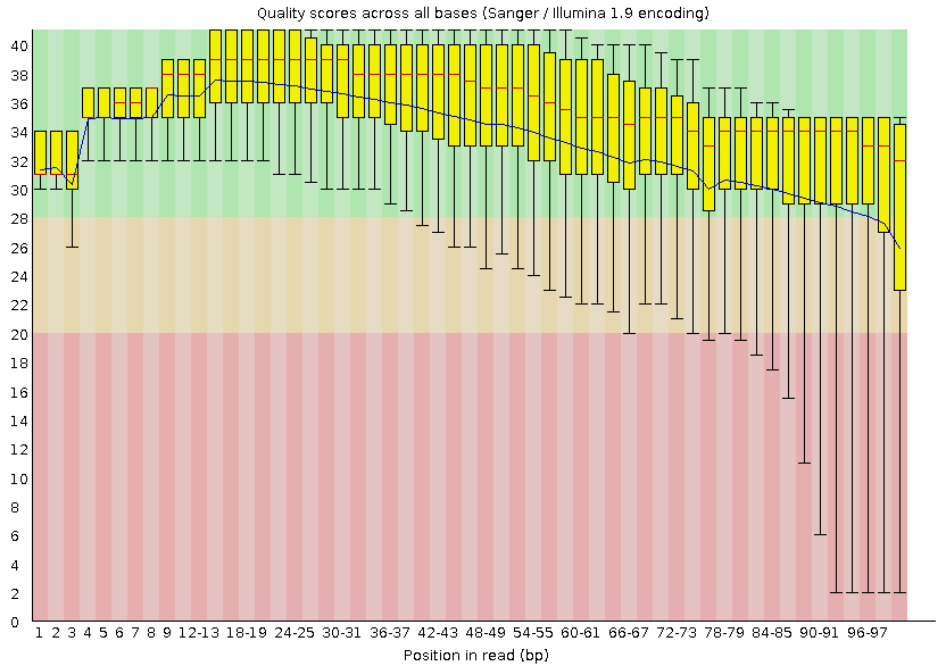


Figure 1: Kwaliteitscontrole op de forward reads van de ruwe data van sample Roots_Low Fe_1

2.3.4 Stap 3: trimmen & filteren

Aangezien de ruwe data niet van voldoende kwaliteit is, zulllen de reads getrimd en gefilterd worden. Daarnaast bevatten de reads nog adaptersequenties die ook verwijderd moeten worden. Voor dit proces zal gebruik gemaakt worden van Trimmomatic (v0.39). Hiermee zijn de adaptersequenties van de reads verwijderd, daarnaast zijn nog de volgende parameters meegegeven: SLIDINGWINDOW met een window van 4 en een kwaliteit van 20, een gemiddelde kwaliteit van minimaal 20 en een minimum lengte van 36 bp. Het onderstaande commando is gebruikt om de forward reads van sample Roots_Low Fe_1 te trimmen en filteren.

```
java -jar ../../Trimmomatic-0.39/trimmomatic-0.39.jar PE
  ../../raw_data/sample_SRR13114670/SRR13114670_1.fastq
  ../../raw_data/sample_SRR13114670/SRR13114670_1.fastq
  trimmed_SRR13114670_1_p.fastq trimmed_SRR13114670_1_u.fastq
  trimmed_SRR13114670_2_p.fastq trimmed_SRR13114670_2_u.fastq
  ILLUMINACLIP:TruSeq3-PE.fa:2:30:1 SLIDINGWINDOW:4:20 AVGQUAL:20 MINLEN:36
```

Met FASTQC is nogmaals een rapport gegenereerd om de kwaliteit van de data te inspecteren. Figuur 2 laat een figuur hieruit zien. In vergelijking tot figuur 1 is de spreiding van de kwaliteit nu een stuk kleiner. Verder zijn korte reads verwijderd, wat niet zichtbaar is in deze figuur. De data is nu van voldoende kwaliteit voor de volgende stap.

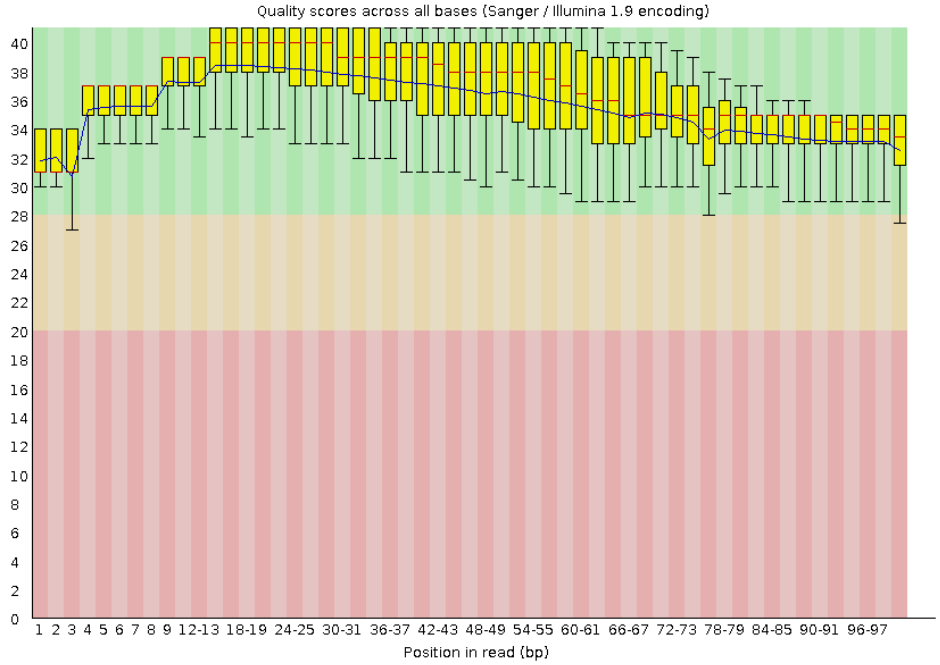


Figure 2: Kwaliteitscontrole op de forward reads van de getrimde en gefilterde data van sample Roots_Low Fe_1

2.3.5 Stap 4.1: genoom indexeren met STAR

Voordat reads tegen het referentiegenoom gemapt kunnen worden, moet het genoom eerst geïndexeerd worden. Aangezien STAR gebruikt wordt voor het mappen, zal hiermee ook het indexeren uitgevoerd worden. Hiervoor is een `.fasta` bestand nodig met het referentiegenoom en een annotatie bestand. Met het onderstaande commando wordt het referentiegenoom gedownload.

```
wget "https://urgi.versailles.inrae.fr/download/iwgsc/IWGSC_RefSeq_Assemblies/v2.1/iwgsc_refseqv2.1_ass
```

Daarna kan het geïndexeerd worden met het commando hieronder. Deze indexeringsstap zal eenmaal uitgevoerd worden, waarna dit telkens overnieuw gebruikt kan worden in het mappen.

```
STAR --runThreadN 80 \
--runMode genomeGenerate \
--genomeDir wheat_genome/genome_indices \
--genomeFastaFiles wheat_genome/iwgsc_refseqv2.1_assembly.fa\
--sjdbGTFfile iwgsc_refseqv2.1_annotation_200916_HC.gff3 \
--sjdbOverhang 100
--sjdbGTFtagExonParentTranscript Parent
```

2.3.6 Stap 4.2: reads mappen tegen genoom met STAR

Nu het genoom geïndexeerd is kunnen de reads tegen het referentiegenoom gemapt worden. Met het onderstaande commando wordt dit gedaan. Dit resulteert in een `.SAM` file, waarin 16.598.504 (72.41%) reads gemapt zijn tegen het genoom.

```
STAR --genomeDir wheat_genome/genome_indices \
     --readFilesIn trimmed_data/sample_SRR13114670/test/trimmed_SRR13114670_1_p.fastq \
       trimmed_data/sample_SRR13114670/test/trimmed_SRR13114670_2_p.fastq \
     --runThreadN 80
```

2.3.7 Stap 5: featureCounts

https://hbctraining.github.io/Intro-to-rnaseq-hpc-O2/lessons/05_counting_reads.html