Genomics en Transcriptomics analyse muis B-cellen

Floris Menninga

2024-09-26

```
# Libraries
library(tidyverse)
library(ggplot2)
```

0.0.1 Inleiding:

Het gekozen artikel: (https://pubmed.ncbi.nlm.nih.gov/38866970/) heeft betrekking tot de verandering in 3d conformatie van chromatine die leiden tot verminderde expressie van het gen ebf1. Uit het onderzoek bleek de reden voor deze vermindering migratie van genen naar het B compartiment te zijn, hier liggen stukken chromatine die niet op dat moment tot transcriptie hoeven te komen. Het onderzoek is uitgevoerd op voorloper- B-cellen van muizen. Deze cellen hebben het gen Ebf1 nodig om te differentiëren van hematopoetische stamcellen naar uitgerijpte B cellen. Als dit gen minder tot expressie komt kunnen deze B cellen meer eigenschappen van de voorloper cel hebben.

0.0.2 Workflow (Genomics):

Ondanks dat DNA data onderdeel van de dataset was, staat er in de methode/artikel niet wat ze hiermee gedaan hebben. Daarom hebben we besloten een variant analyse hierop uit te voeren. Het referentiegenoom (muis) zal vergeleken worden met DNA seq data. In het onderzoek was een muis genoom versie uit 2012 gebruikt (mm10), deze gaan we vervangen door de nieuwste versie (GRCm39). De eerste stap is het alignen van de genen van de DNA seq data met het referentiegenoom. Daarna worden de verschillen gevisualiseerd en mits deze er zijn.

0.0.3 Workflow (Transcriptomics):

Net zoals in het originele artikel beschreven was wordt read mapping uitgevoerd met het FASTQ bestand van elk van de samples. Met het .SAM bestand dat verkregen is werd daarna met FeatureCounts en RSEM de gen expressie gekwantificeerd. Deze read mapping was met STAR uitgevoerd maar wij gaan hier HISAT2 voor gebruiken omdat STAR verouderd is volgens het github repo. FeatureCounts gebruikt namelijk het .SAM (of .BAM) bestand en telt hoeveel reads bij elke “feature” (gen/exon) horen. RSEM kan hier ook voor gebruikt worden maar is complexer om te gebruiken dan FeatureCounts. Dit kan daarna gevisualiseerd worden met R in een box-plot, viool-plot, heatmap, MA-plot etc. Ook was er in het artikel gebruik gemaakt

van “Cufflinks”, deze gaan we vervangen door “StringTie”. Volgens de website van Cufflinks is StringTie accurater en veel efficiënter.

De volgende R libraries en commandline tools zijn gebruikt voor de analyse: Samtools: 1.16.1

0.0.4 Log:

02-09-2024:
Het gekozen artikel: https://pubmed.ncbi.nlm.nih.gov/38866970 Gene expression omnibus: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE211975

0.0.5 SRA’s downloaden (Genomics):

De eerste stap om een variant calling analyse uit te voeren is het downloaden van de genetische data die verkregen was door het sequencen van de samples. Met behulp van de SRA run selector van NCBI is een selectie van SRA’s gemaakt die gedownload moeten worden. Een SRA (Sequence Read Archive) is een gecomprimeerd archief dat de sequencing reads bevat. Om deze bestanden te downloaden wordt gebruikt gemaakt van “prefetch”, deze commandline tool is onderdeel van de SRA toolkit.

In de volgende stap worden deze bestanden uitgepakt. Het onderstaande stuk code is op mijn computer uitgevoerd en de data is op een externe HDD opgeslagen. Dit zelfde is gedaan op de assemblix computer waar de andere groepsleden ook bij kunnen. export om de “prefetch” binairy (tijdelijk) toe te voegen aan het PATH.

```
export PATH="/home/floris/Documenten/Applicaties/sratoolkit.3.1.1-ubuntu64/bin/:$PATH"

cd /run/media/floris/FLORIS_3/DATA_SETS/3D_Chromatine_Conformatie/DNA/SRA/

prefetch $(</run/media/floris/FLORIS_3/DATA_SETS/3D_Chromatine_Conformatie/DNA/SRA/SraAccList.csv) \
--output-directory "/run/media/floris/FLORIS_3/DATA_SETS/3D_Chromatine_Conformatie/DNA/SRA/" --max-size
```

.SRA bestanden naar .FASTQ omzetten:
Met behulp van fasterq-dump worden de .SRA bestanden uitgepakt.

fasterq-dump maakt ook onderdeel uit van de SRA toolkit en haalt de data uit het .SRA archief naar het FASTQ format.

Een FASTQ bestand bevat tekst met de sequence data, de reads en ook een bijbehorende kwaliteitsscore. Deze score wordt aangegeven met een ASCII karakter.

Een FASTQ bestand heeft de volgende indeling:

Een header die met “@” begint gevolgd door een sequentie ID en optioneel een beschrijving wat het bestand bevat. Daar onder zitten de sequentie letters.

En daar onder een regel met een “+” karakter. Hier onder staat de kwaliteitsscore, deze gaat van ASCII 33 (“!” teken), laagste kwaliteit tot ASCII 126 (“~” teken).

De kwaliteitsscore wordt ook wel Phred score genoemd. Deze score is logaritmisch gerelateerd aan de waarschijnlijkheid dat de "base call" verkeerd is. Q = -log(E) waarbij Q de phred score is en E de waarschijnlijkheid van verkeerde base call.

Phred-33 is het meest gebruikt maar Phred-64 bestaat ook. Het verschil is dat Phred-33 ge-encodeerd is met met ASCII 33 (!) tot ASCII 126 (~) terwijl Phred-64 van ASCII 64 (@) tot ASCII 126 gaat.

Bron: https://gatk.broadinstitute.org/hc/en-us/articles/360035531872-Phred-scaled-quality-scores

```
# Eerst nieuwe map maken voor fastq
# Locale test: (niet op de assemblix server)
cd /run/media/floris/FLORIS_3/DATA_SETS/3D_Chromatine_Conformatie/DNA/SRA


find /run/media/floris/FLORIS_3/DATA_SETS/3D_Chromatine_Conformatie/DNA/SRA -name "*.sra" | \
  parallel fasterq-dump -O /run/media/floris/FLORIS_3/DATA_SETS/3D_Chromatine_Conformatie/DNA/SRA/FASTQ


# parallel gebruiken: cat accessionlist.txt | parallel ls -lah SRA/{}/{}.sra
```

Onderzoeksvraag Genomics De volgende onderzoeksvraag is geformuleerd voor het genomics onderzoek:

Zijn er varianten aanwezig van PAX5 en Ebf1 in het genoom van de rag2(-/-) muizen die gebruikt zijn?

Deze vraag is relevant omdat voor het transcriptomics gedeelte van het onderzoek gekeken wordt naar verschillen in expressie van deze genen. Als het blijkt dat er variaten van deze genen aanwezig zijn kan het verschil in expressie tussen de muizen die gebruikt zijn in het onderzoek niet alleen toegewezen worden aan de factoren waar naar getest wordt, de invloed van veroudering op de expressie. Dan zou het ook kunnen komen door mutaties van deze genen.

Er is DNA sequentie data beschikbaar van rag2(-/-) muizen. Daarom kan er niet gekeken worden naar mutaties in het rag2 gen omdat deze niet meer aanwezig is.

Rag2 knockout muizen zijn niet in staat om uitgerijpte T en B lymfocyten te maken. Bron: https://pubmed.ncbi.nlm.nih.gov/1547487/

Het gen Ebf1 (Early B-cell factor 1) is een gen dat bijdraagt aan de differentiatie van voorloper b cellen. bron: https://www.nature.com/articles/ni.2641

De reden dat gekozen is voor dit gen is dat het resultaat van een verminderde expressie zichtbaar gemaakt kan worden door functieverlies van B cellen. Samen met PAX5 werken deze genen aan de uitrijping van hematopoetische stam cellen. Functieverlies van PAX5 leidt tot accumulatie van snel proliferende lymfoblasten die niet meer normale differentiatie kunnen ondergaan. Ook is PAX5 een tumorsuppressor gen maar dat is niet van invloed op dit onderzoek.

bron: https://www.sciencedirect.com/topics/neuroscience/pax5

Test data genereren 16-09-2024

Om te voorkomen dat een parallel commando na het uitvoeren van een lange bewerking een fout of onverwacht resultaat geeft moet er eerst een subset van de te gebruiken data gemaakt worden. Hiermee kunnen de volgende commando's eerst uitgevoerd worden om te verifieren dat alles goed werkt.

Om test data te genereren op basis van twee van de fastq bestanden is de volgende code gebruikt: Per sample bestand zitten er 1000000 reads in. Dit commando is ook uitgevoerd op de assemblix server maar dan met een ander path naar de data. Voor het maken van deze test data is seqkit (versie 2.3.0) gebruikt.

Seqkit kan gebruikt worden voor meerdere bewerkingen zoals het filteren van sequenties op lengte / kwaliteit, DNA/RNA translateren naar een aminozuur sequentie. bron: https://bioinf.shenwei.me/seqkit/

```
cd /run/media/floris/FLORIS_3/DATA_SETS/3D_Chromatine_Conformatie/DNA/SRA

seqkit head -n 1000000 SRR26980527_1.fastq > subset_SRR26980527_1.fastq
seqkit head -n 1000000 SRR26980527_2.fastq > subset_SRR26980527_2.fastq
```

```
fastqc

multiqc

seqkit (voor maken van subset van test data)
```

Gebruikte samples 17-09-2024

Na het controle van de kwaliteit van de fastq bestanden bleek het dat de samples met read lengtes van 250 bp getrimd moeten worden. De volgende samples zijn wildtype: (WT C57BL/6J) SRR26980527, SRR26980528, SRR26980529, SRR26980530. En de volgende samples zijn Rag2(-/-) knockout: SRR26980549, SRR26980550, SRR26980551, SRR26980552 Al deze samples zijn paired end sequenced Primary pro-B cells geselecteerd op CD19+.

0.0.6 Indexeren en alignen:

Met enkel een fastq bestand met reads is nog niet duidelijk waar in het genoom van het organisme deze reads kwamen. Met read mapping worden de reads vergeleken met een bekend genoom, in dit geval het mm39 muis referentiegenoom en worden de reads er tegen aan gelegd zodat hun locatie in het genoom bekend wordt. Aangezien er geannoteerde versies van het referentiegenoom zijn, met de namen van de genen kan dan ook achterhaald worden van welke genen de reads deel uitmaken.

https://www.ebi.ac.uk/training/online/courses/functional-genomics-ii-common-technologies-and-data-analysis-methods/rna-sequencing/performing-a-rna-seq-experiment/data-analysis/read-mapping-or-alignment/

bron: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3375638/

Omdat het programma dat gebruikt gaat worden voor het alignen, BWA-mem2, een geïndexeerde versie van het referentiegenoom nodig heeft moet deze eerst gemaakt worden. Dit kan gedaan worden met bwa-mem2 index. indexeren maakt het align proces veel sneller.

De samples SRR26980528_1 en _2 zijn gebruikt na het referentiegenoom te indexeren met BWA_MEM2. Het gebruikte referentie genoom is mm39 (GCF_000001635.27).

In de onderstaande code chunk wordt het geindexeerde muisgenoom "testnaam" genoemd. Daarna worden twee samples (forward read en reverse read) ge-aligned met het zojuist verkregen geïndexeerde referentiegenoom.

```
cd /run/media/floris/FLORIS_3/DATA_SETS/3D_Chromatine_Conformatie/DNA/SRA


# Het indexeren van het referentiegenoom met bwa-mem2.
bwa-mem2 index -p testnaam referentiegenoom

# Voorbeeld aligning:
bwa-mem2 mem ref.fa read1.fq read2.fq > aln-pe.sam

# Align sample SRR26980549_1.fastq (deel 1 en 2) met het referentiegenoom mm39 muis.
./bwa_mem2/bwa-mem2-2.2.1_x64-linux/bwa-mem2 mem -p 60 testnaam fastq/SRR26980549_1.fastq fastq/SRR2698

# find /run/media/floris/FLORIS_3/DATA_SETS/3D_Chromatine_Conformatie/DNA/SRA -name "*.sra" | \
#   parallel fasterq-dump -O /run/media/floris/FLORIS_3/DATA_SETS/3D_Chromatine_Conformatie/DNA/SRA/FAS
```

14-09-2024

In tegenstelling tot het artikel gaan we gebruik maken van Muis (*Mus musculus*) referentiegenoom: GRCm39 (GCF_000001635.27) gebruik maken. Deze is nieuwer dan versie mm10 die gebruikt was. Aangezien het onderzoek dit jaar gepubliceerd is en voltooid was in 2022 vraag ik mij af waarom ze voor een versie uit 2012 gekozen hebben terwijl er een nieuwere beschikbaar was.

.sam bestanden sorteren: 17-09-2024

Na de vorige read mapping stap zijn er .SAM (Sequence alignment map) bestanden verkregen, dit is een tekst gebaseerde manier om sequenties die aligned zijn tegen een refrentie sequentie op te slaan. Een .sam bestand bestaat uit een header en een alignment deel.

Omdat de .SAM bestanden niet georderd zijn op positie in het genoom moeten ze eerst met samtools sort omgezet worden naar een .BAM bestand, dit is nu wel gesorteerd door samtools. Daarnaast kan het opvragen van data sneller gemaakt worden als het geïndexeerd is.

Na de index stap uit gevoerd te hebben is er ook een .bai bestand, dit een index voor het .bam bestand.

bron: https://gatk.broadinstitute.org/hc/en-us/articles/360035890791-SAM-or-BAM-or-CRAM-Mapped-sequence-data-formats

Om al deze bewerkingen uit te voeren wordt gebruik gemaakt van samtools (versie 1.16.1). Dit is een collectie tools om met high-throughput sequence data te werken. Zo kan het bijvoorbeeld fastq omzetten naar .bam / .cram bestanden, WGS/WES mapping naar variant calls en het filteren van .vcf bestanden.

```
samtools sort aligned_SRR26980549.sam > aligned_sorted_SRR26980549.bam &


samtools index aligned_sorted_SRR26980549.bam &
```

```
freebayes -f ref.fa aln.bam >var.vcf
```

18-09-2024

```
library(fastqcr)

trimomaticPE?
```

Om alle paired read samples te mappen is het onderstaande bash script gebruikt. Hier is weer het parallel commando voor gebruikt. De lijst met sample namen wordt aan bwa-mem2 gegeven zodat het programma naar input bestanden met deze namen kan kijken en de gegenereerde .sam bestanden ook deze naam te geven.

```
cat /students/2024-2025/Thema05/3dconformatieChromatine/SRA/SraAccList.csv | parallel '/students/2024-2
```

23-09-2024
De fastq bestanden zijn met fastqc gecontroleerd en vervolgens getrimmed met Trimmomatic met het onderstaande stuk code. Voor Trimmomatic zijn de volgende instellingen gebruikt: MINLEN:40 en SLIDINGWINDOW:4:20.

Ook zijn de volgende samples zijn niet gebruikt:
B220+CD43+IgM- sorted primary pro-B cells.
Deze samples zijn verkregen met een Illumina MiSeq.

SRR26980527
SRR26980528
SRR26980529
SRR26980530

De reden hiervoor is dat het samples van WT muizen zijn, hier gaan we het referentiegenoom voor gebruiken. Ook was de kwaliteit van deze sequenties volgens fastqc veel slechter dan die van de onderstaande samples. Mogelijk komt dit omdat de reads te lang waren en de kwaliteit daardoor te snel naar beneden ging.

De volgende zijn wel gebruikt:
Primary pro-B cells by CD19+ selection (Rag2-/-) Deze samples zijn verkregen met een Illumina NovaSeq 6000. SRR26980549.sam
SRR26980550.sam
SRR26980551.sam
SRR26980552.sam

De onnodige samples zijn ook verwijderd uit de lijst (SraAccList.csv) omdat deze met de pipe operator aan de parallel opdracht gegeven wordt. Als de namen in de accession

```
cat data/GSE149995_Sra_RunInfo.csv | \
  parallel 'TrimmomaticPE -threads 16 ' \
                  '/students/2024-2025/Thema05/3dconformatieChromatine/fastq/{}.fastq.gz ' \
                  '/students/2024-2025/Thema05/3dconformatieChromatine/fastq-trimmed/{}.trimmed.fastq.g
                  'ILLUMINACLIP:/students/2024-2025/Thema05/3dconformatieChromatineTrimmomatic/adapters
                  'MINLEN:40 ' \
                  'SLIDINGWINDOW:4:20'
```

24-09-2024
Het mappen met behulp van BWA-mem2 is opnieuw uitgevoerd na het trimmen. Het referentiegenoom dat in de vorige stap geïndexeerd was wordt weer opnieuw gebruikt.

```
cat /students/2024-2025/Thema05/3dconformatieChromatine/SRA/SraAccList.csv | parallel '/students/2024-2
```

Feedback: Veel meer uitleg toevoegen, redenen waarom keuzes gemaakt zijn. Versies toevoegen van alle gebruikte software. Overal nog bronnen toevoegen. redenen waarom voor specifieke software gekozen is toevoegen.

Variant calling met Freebayes:

Nu er gemapte reads zijn kan er variant calling op uitgevoerd worden. Hiermee kunnen mutaties / varianten gedetecteerd en mits ze statistisch significant zijn gerapporteerd worden. Dit zou ook grafisch met IGV (Genome browser) kunnen maar zou veel meer tijd kosten.

Voor variant calling is freebayes uitgekozen. Dit is een genetische variant detector gemaakt om kleine polymorfismen zoals: SNPs, inserties, deleties en MNPs (multi-nucleotide polymorfismes) te herkennen.

Voordat freebayes uitgevoerd kan worden moet het referentiegenoom weer geïndexeerd worden, net als met BWA-mem2 index maar dan met samtools faidx zoals in de onderstaande code chunk beschreven is.

```
samtools faidx GCA_000001635.9_GRCm39_genomic.fna
```

Met het geïndexeerde referentiegenoom (GCA_000001635.9_GRCm39) was vervolgens de freebayes variant calling uitgevoerd.

```
/students/2024-2025/Thema05/3dconformatieChromatine/freebayes/freebayes-1.3.6-linux-amd64-static -f GCA
```

Gen anotatie Eerst is er een .bed bestand van het niet door mensen leesbare .bam bestand gemaakt, daarna zijn de genen geannoteerd zodat ze geen namen hebben zoals: SRR26980549.46263152/2 maar ....

```
bamToBed -i aligned_sorted_SRR26980549.bam > aligned_sorted_SRR26980549.bed
```

Sequence depth vaststellen: Om te achterhalen wat de coverage is, is het volgende bash commando uitgevoerd:

```
samtools depth -a aligned_sorted_SRR26980549.bam > depth_output.txt
```

0.1 Visualisatie van de verkregen data:

0.2 Statistische analyse

0.3 Verschillen tussen onze resultaten en die van het originele onderzoek

0.4 referenties

Artikel 3d conformatie chromatine: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10329473/ Samtools: http://www.htslib.org/ BWA-MEM2 https://github.com/bwa-mem2/bwa-mem2 Trimmomatic: http://www.usadellab.org/cms/?page=trimmomatic FastQC: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

1 Transcriptomics:

1.0.1 SRA's downloaden (Transcriptomics):

De SRA's van de RNA seq worden gedownload met het onderstaande stuk code. De gebruikte accention list staat in de github repository.

```
export PATH="/home/floris/Documenten/Applicaties/sratoolkit.3.1.1-ubuntu64/bin/:$PATH"



cd /run/media/floris/FLORIS_3/DATA_SETS/3D_Chromatine_Conformatie/Transcriptomics/

prefetch $(</run/media/floris/FLORIS_3/DATA_SETS/3D_Chromatine_Conformatie/Transcriptomics/SRR_Acc_List
--output-directory "/run/media/floris/FLORIS_3/DATA_SETS/3D_Chromatine_Conformatie/Transcriptomics/" --
```