

Genomics en Transcriptomics analyse muis B-cellen

Floris Menninga

2024-10-07

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

0.0.1 Inleiding:

Het gekozen artikel: (Ma F (2024)) (<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 Het onderzoek

Het doel van het originele onderzoek (Ma F (2024)) was om te achterhalen of verschil is in gen expressie van onder andere ebfl tussen oude en jonge muizen. (verminderde interactie tussen ebfl promoter en zijn enhancers)

Met deze genomics analyse trachten we de volgende onderzoeksvraag te beantwoorden: 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 varianten 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. (Shinkai Y (1992))

Het gen Ebf1 (Early B-cell factor 1) is een gen dat bijdraagt aan de differentiatie van voorloper b cellen. (Nechanitzky (2013))

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 prolifererende lymfoblasten die niet meer normale differentiatie kunnen ondergaan. Ook is PAX5 een tumorsuppressor gen maar dat is niet van invloed op dit onderzoek. (Chivukula and Dabbs (2011))

Verschillen tussen het originele onderzoek en dat van ons: In tegenstelling tot de analyses die betrokken zijn bij het transcriptomics gedeelte van de analyses hebben de onderzoekers van het artikel niet gebruik gemaakt van deze data daarom is er niets om mee te vergelijken. Al hebben ze wel een verouderd muis referentiegenoom gebruikt (mm10) deze vervangen we door mm39.

0.0.3 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, gericht op de volgende genen: PAX5, Ebf1 en FOXO1. 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 downloaden van de .SRA archieven die de fastq bestanden bevatten, hier wordt prefetch voor gebruikt. Daarna worden deze .SRA's uitgepakt met behulp van fasterq-dump. De nieuw verkregen fastq bestanden kunnen nu op kwaliteit gecontroleerd worden met fastqc, hierna worden ze getrimmed met trimmomatic en wordt de kwaliteit nogmaals bekeken. Duplicaten kunnen verwijderd worden.

0.0.4 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, viol-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

0.0.5 Log:

02-09-2024:

Het gekozen artikel: Three-dimensional chromatin reorganization regulates B cell development during ageing. (Ma F (2024)) Gene expression omnibus: <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE211975>

0.0.6 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” binary (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 $(cat /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
```

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 verifiëren 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 assemblx 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
```

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+.

0.0.7 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 geïndexeerde muisgenoom “testnaam” genoemd. Daarna worden twee samples (forward read en reverse read) ge-aligned met het zojuist verkregen geïndexeerde referentiegenoom. Deze twee bewerkingen worden beiden met bwa-mem2 uitgevoerd.

```
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 -t 40 testnaam fastq/SRR26980549_1.fastq fastq/SRR26980549_2.fastq

# Met parallel:
cat /students/2024-2025/Thema05/3dconformatieChromatine/SRA/SraAccList.csv | parallel ' ./bwa_mem2/bwa-mem2 mem -t 40 testnaam fastq/SRR26980549_1.fastq fastq/SRR26980549_2.fastq > aln-pe.sam '
```

0.1 Trimmen van reads

```
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.gz ' \
        'ILLUMINACLIP:/students/2024-2025/Thema05/3dconformatieChromatineTrimmomatic/adapters/TruSeq-PE-2.5.1-mi.fa ' \
        '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.


```
# Voor een enkel sample:
```

```
samtools fixmate -m --threads 40 aligned_sorted_SRR26980549.bam \ fixed_mates_aligned_sorted_SRR26980549.bam
```

```
# Met parallel:
```

```
cat /students/2024-2025/Thema05/3dconformatieChromatine/SRA/SraAccList.csv | parallel 'samtools fixmate -m --threads 40 aligned_sorted_SRR26980549.bam \ fixed_mates_aligned_sorted_SRR26980549.bam'
```

Daarna met “samtools sort” op coördinaten gesorteerd.

```
# Voor een enkel sample:
```

```
samtools sort -@80 -o sorted_coordinates_SRR26980549.bam \ fixed_mates_aligned_sorted_SRR26980549.bam
```

```
# Met parallel:
```

```
cat /students/2024-2025/Thema05/3dconformatieChromatine/SRA/SraAccList.csv | parallel 'samtools sort -@80 -o sorted_coordinates_SRR26980549.bam \ fixed_mates_aligned_sorted_SRR26980549.bam'
```

Met samtools markdup worden duplicaten gemarkeerd. Met het “-r” argument worden deze verwijderd en met “-s” worden statistieken over de data en uitgevoerde handelingen.

Bron: <http://www.htslib.org/doc/samtools-markdup.html>

```
# Voor een enkel sample:
```

```
samtools markdup -r -s sorted_coordinates_SRR26980549.bam dedup_SRR26980549.bam
```

```
# Met parallel:
```

```
cat /students/2024-2025/Thema05/3dconformatieChromatine/SRA/SraAccList.csv | parallel 'samtools markdup -r -s sorted_coordinates_SRR26980549.bam dedup_SRR26980549.bam'
```

Met samtools index wordt er een index bestand gemaakt.

```
# Voor een enkel sample:
```

```
samtools index -@80 dedup_SRR26980549.bam dedup_SRR26980549.bai &
```

```
# Met parallel:
```

```
cat /students/2024-2025/Thema05/3dconformatieChromatine/SRA/SraAccList.csv | parallel 'samtools index -@80 dedup_SRR26980549.bam dedup_SRR26980549.bai &'
```

0.2 Variant calling en filteren van varianten:

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.

Bron: <https://github.com/freebayes/freebayes> Na het uitvoeren van het variant calling met freebayes worden de output bestanden, de .vcf's gefilterd door ze in vcfilter te pipen. In dit onderstaande stuk code worden

de resultaten van de freebayes variant calling direct gefilterd op kwaliteit met vcffilter. Met “QUAL = 20” worden enkel de variants geselecteerd die 99% kans hebben dat er een variant zit op die plaats. De reden dat op kwaliteit gefilterd moet worden is... (nog bedenken.)

```
# Voor een bestand:
/students/2024-2025/Thema05/3dconformatieChromatine/freebayes/freebayes-1.3.6-linux-amd64-static \
-f /students/2024-2025/Thema05/3dconformatieChromatine/Mapping_ref/ncbi_dataset/ncbi_dataset/data/GCA_

# Met parallel:
cat /students/2024-2025/Thema05/3dconformatieChromatine/SRA/SraAccList.csv | parallel '/students/2024-2025/Thema05/3dconformatieChromatine/Mapping_ref/ncbi_dataset/ncbi_dataset/data/GCA_

# Test zonder filter stap:
cat /students/2024-2025/Thema05/3dconformatieChromatine/SRA/SraAccList.csv | parallel '/students/2024-2025/Thema05/3dconformatieChromatine/Mapping_ref/ncbi_dataset/ncbi_dataset/data/GCA_
```

Filteren: Met `-minQualScore 30` wordt gefilterd op de waarschijnlijkheid dat de variant echt aanwezig is. Met de formule $P = 10^{-(Q/10)}$ kan de kans uitgerekend worden dat de variant niet echt aanwezig is. Met een score van 30 is dit: $P = 10^{-(30/10)} = 0.001$. In andere woorden: De kans is $100 - 0.001 = 99.999\%$ dat het wel klopt.

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

De .vcf's zijn nu gefilterd, voorbeeldbestandsnaam: `filtered_variant_SRR26980549.vcf`

18-09-2024

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/2024-2025/Thema05/3dconformatieChromatine/Mapping_ref/ncbi_dataset/ncbi_dataset/data/GCA_
```

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 SLIDING-WINDOW: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

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.

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_000001635.9_GRCm39_genomic.fna
```

0.3 Annoteren van de .vcf bestanden:

We hebben nu een serie grote bestanden die het verschil beschrijven tussen onze samples en het referentiegenoom van de muis (mm39). Om vragen te kunnen beantwoorden zoals: zit de mutatie in een gen, in een exon, veranderd het eiwit er door of maakt de variant een stopcodon. (nonsense mutatie). Om dit achterhalen moeten de .vcf bestanden geannoteerd worden, hier wordt SnpEff (Cingolani et al. (2012)) voor gebruikt. Aan de hand van een referentiedatabase die gedownload kan worden met SnpEff kan een .vcf geannoteerd worden.

Deze databases zijn voor 38000 genomen gemaakt maar het is ook mogelijk om deze zelf te maken. Met een bekend model organisme zoals de muis (*Mus musculus*) zou dit niet hoeven maar tijdens het uitvoeren van

SnEff bleek dat de nieuwste mm39 database niet gedownload kon worden. Tenzij we een oude versie zoals mm10 willen gebruiken moeten we zelf een database hiervoor maken.

De genen worden ook geannoteerd zodat ze geen namen hebben zoals SRR26980549.46263152/2 maar bijvoorbeeld: EBF1.

Overzicht van enkele mutaties die met SnEff gedetecteerd kunnen worden: (afkomstig van de SnEff website)

Soort mutatie:	Betekenis:	Voorbeeld:
SNP	Single-Nucleotide Polymorphism	Reference = 'A', Sample = 'C'
Ins	Insertion	Reference = 'A', Sample = 'AGT'
Del	Deletion	Reference = 'AC', Sample = 'C'
MNP	Multiple-nucleotide polymorphism	Reference = 'ATA', Sample = 'GTC'
MIXED	Multiple-nucleotide and an InDel	Reference = 'ATA', Sample = 'GTCAGT'

Converteren van .bam naar .bed

Eerst is er een .bed bestand van het niet door mensen leesbare .bam bestand gemaakt, de leesbare tegenhanger hiervan is .sam maar voor sommige analyses in R is een .bed bestand handiger.

Met behulp van bamToBed wordt het .bam bestand omgezet naar een .bed (Browser extensible data). De eerste kolom van een .bed bevat de naam van het chromosoom, de tweede kolom het start coördinaat van de feature, de derde kolom het eind coördinaat van de feature, de 4e kolom de naam, daarna optioneel een score en dan de streng (-/+). bron: <https://genome.ucsc.edu/FAQ/FAQformat.html>

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

Voordat bcftools isec uitgevoerd kan worden moeten de vcf bestanden gecomprimeerd worden met bgzip.

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

Daarna, om de vervolg processen sneller te maken (en omdat isec anders een foutmelding geeft) moet een index bestand gemaakt worden. Hier wordt “tabix” voor gebruikt. Het onderstaande commando gebruikt de net gemaakte bgzip gecomprimeerde bestanden van de .vcf als input.

```
cat /students/2024-2025/Thema05/3dconformatieChromatine/SRA/SraAccList.csv | parallel 'tabix -p vcf fil
```

```
## cat: /students/2024-2025/Thema05/3dconformatieChromatine/SRA/SraAccList.csv: Bestand of map bestaat n
```

0.4 Isec

Er zijn vier biologische replicaten, met behulp van “bcftools isec -c all” worden enkel de varianten die in alle vier de muizen aanwezig zijn geselecteerd. Hierdoor worden veel “willekeurige” mutaties er uit gefilterd

die niet in alle vier de samples aanwezig waren. Bron: <https://samtools.github.io/bcftools/bcftools.html> In het onderstaande code chunk wordt bcftools isec uitgevoerd met argumenten: “-c all” dit betekend dat alle varianten die overeen komen tussen de vier samples in het output bestand opgeslagen moeten worden.

```
bcftools isec -c all filterd_variant_SRR26980549.vcf.gz filterd_variant_SRR26980550.vcf.gz filterd_vari
```

0.5

Het onderstaande script was eerst uitgevoerd met mm39 als referentiegenoom voor de annotatie maar dit gaf een foutmelding (Op de github pagina van het programma snpEff waren er meer meldingen van dit gedrag. <https://github.com/pcingola/SnpEff/issues/536>) dus is het opnieuw uitgevoerd met het oude mm10 referentiegenoom. Dit is het zelfde referentiegenoom dat ook het artikel gebruikt was.

```
# Test:
/students/2024-2025/Thema05/3dconformatieChromatine/snpEff/snpEff.jar mm39 filtered_variants.vcf > anno

# Voor alle .vcf bestanden:
cat /students/2024-2025/Thema05/3dconformatieChromatine/SRA/SraAccList.csv | parallel 'java -jar /studen

# Alleen het isec .vcf bestand:
java -jar /students/2024-2025/Thema05/3dconformatieChromatine/snpEff/snpEff/snpEff.jar mm39 /students/2
```

0.6 Kwaliteitscontrole vcf bestanden:

0.7 Visualisatie van de verkregen data:

0.8 Verschillen tussen onze resultaten en die van het originele onderzoek

0.9 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/sratookit.3.1.1-ubuntu64/bin/:$PATH"
```

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

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

- Chivukula, Mamatha, and David J. Dabbs. 2011. "Chapter 21 - Immunocytology." In *Diagnostic Immunohistochemistry (Third Edition)*, edited by David J. Dabbs, Third Edition, 890–918. Philadelphia: W.B. Saunders. <https://doi.org/https://doi.org/10.1016/B978-1-4160-5766-6.00025-X>.
- Cingolani, P., A. Platts, M. Coon, T. Nguyen, L. Wang, S. J. Land, X. Lu, and D. M. Ruden. 2012. "A Program for Annotating and Predicting the Effects of Single Nucleotide Polymorphisms, SnpEff: SNPs in the Genome of *Drosophila Melanogaster* Strain W1118; Iso-2; Iso-3." *Fly* 6 (2): 80–92.
- Ma F, Du H, Cao Y. 2024. "Three-Dimensional Chromatin Reorganization Regulates b Cell Development During Ageing." *Nat Cell Biol.* Jun;26(6) (26): 991–1002. <https://doi.org/10.1038/s41556-024-01424-9>.
- Nechanitzky, Akbas, R. 2013. "Transcription Factor EBF1 Is Essential for the Maintenance of b Cell Identity and Prevention of Alternative Fates in Committed Cells." *Nat Immunol* 14 (June): 867–75. <https://doi.org/https://doi.org/10.1038/ni.2641>.
- Shinkai Y, Lam KP, Rathbun G. 1992. "RAG-2-Deficient Mice Lack Mature Lymphocytes Owing to Inability to Initiate v(d)j Rearrangement." *Cell* 6 (March): 855–67. [https://doi.org/10.1016/0092-8674\(92\)90029-c](https://doi.org/10.1016/0092-8674(92)90029-c).