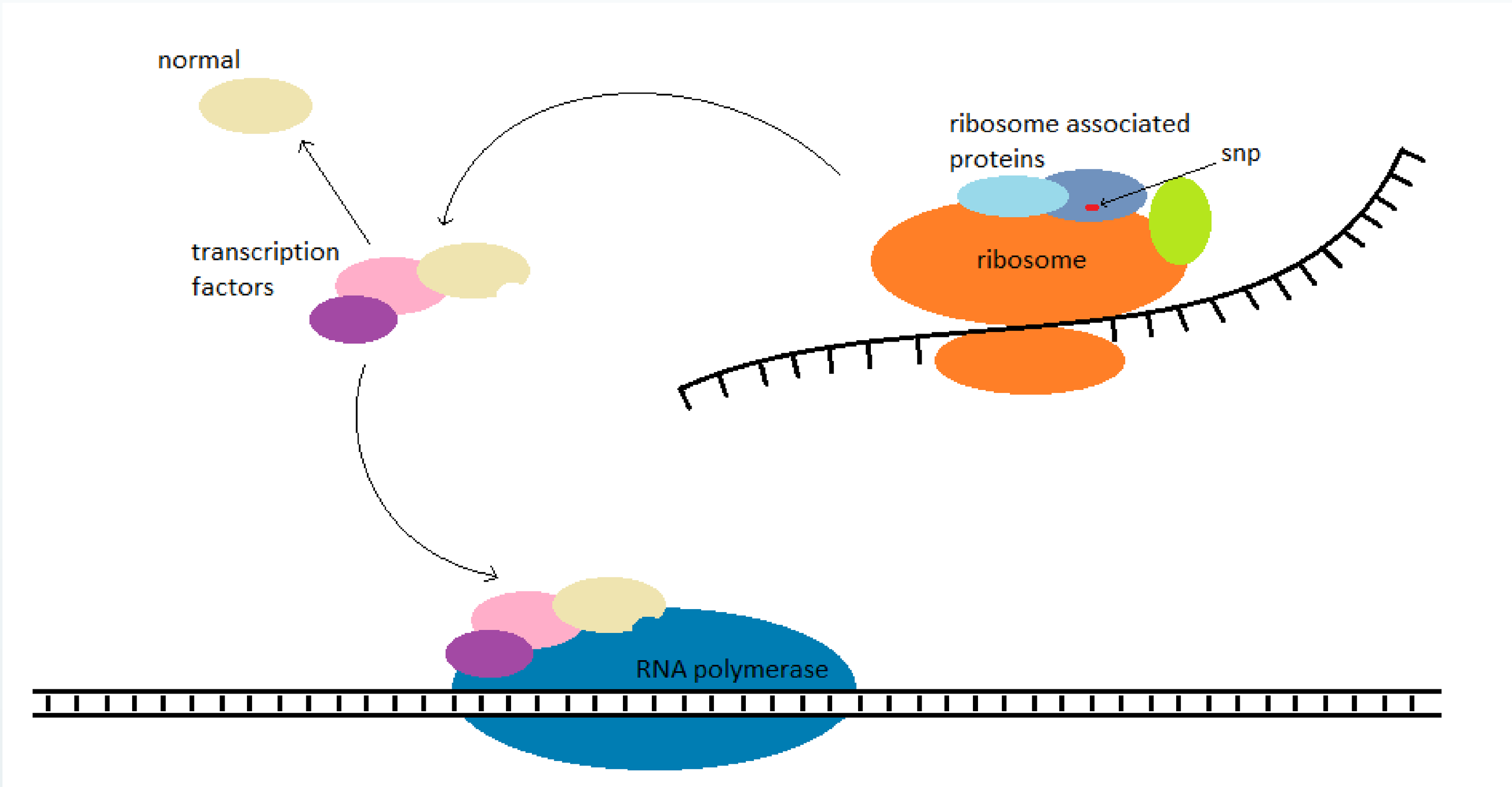
Genomische polymorfismen in ribo-interactoom liggen mogelijk ten grondslag aan variatie in genexpressie in mensen met kanker tot gevolg

Gabe van den Hoeven, Gijsbert Keja, Moshtach Ismail, Michelle Memelink & Christel van Haren

Introductie

Kanker is een van de meest voorkomende doodsoorzaken ter wereld (Zaimy, et al., 2017). Ondanks de vooruitgang die gemaakt is, blijft het door de diversiteit van kanker lastig om op tijd de juiste diagnoses te stellen. Het is van belang dat er meer bekend wordt over de ontwikkeling van kanker. Een verstoring van genexpressie is al meerdere keren in verband gebracht met de ontwikkeling van verschillende soorten kanker (Bradner, Hnisz, & Young, 2017), (Sato, Fukushima, Chang, Matsubayashi, & Goggins, 2006).

Het ribo-interactoom bestaat uit ribosomale en ribosoom-geassocieerde eiwitten die de translatie regelen van verschillende eiwitten, waaronder transcriptiefactoren. Ribosoom-geassocieerde eiwitten kunnen, doordat zij de translatie van transcriptiefactoren beïnvloeden, de genexpressie verstoren wat kanker kan veroorzaken. De hypothese is dat single nucleotide polymorphisms (SNP's) binnen het ribo-interactoom de translatie van transcriptiefactoren beïnvloeden, waardoor transcriptiefactoren veranderen en zo indirect de genexpressie wordt verstoord en kankerontwikkeling plaats kan vinden. Dit concept is gevisualiseerd in figuur 1.



Figuur 1: Visualisatie van een SNP in de genen die coderen voor de ribosoom-geassocieerde eiwitten. Dit voor een verandering in de translatie van een transcriptiefactor wat vervolgens de gen expressie kan verstoren.

Het doel van dit onderzoek is door middel van een expression quantative trait loci (eQTL) analyse een verband aan te tonen tussen SNP's in ribosomale en ribosoom-geassocieerde eiwitten, en een verschil in genexpressie tussen kanker- en niet-kanker patiënten. Tijdens de eQTL analyse wordt door middel van regressie de genexpressie vergeleken met SNP's in het ribo-interactoom van individuen. Hierdoor kan het effect van deze SNP's op de genexpressie worden bepaald. Voor dit onderzoek is een pipeline gemaakt waarmee de analyse is uitgevoerd voor test data van het menselijk chromosoom 10.

Materiaal & Methode

Voor dit onderzoek zijn er een aantal stappen verricht om de eQTL analyse uit te voeren, deze stappen zijn te zien in figuur 2. Doordat er geen adequate patiënt data beschikbaar was, is er gebruikt gemaakt van test datasets afkomstig van STAR.

Zoeken naar ribosomale eiwitten en eiwit-eiwit interacties
Door middel van de HUGO Gene Nomenclature (Tweedie S, 2021) is er gezocht naar humane ribosomale eiwitten, dit zijn de 40S en de 60S. De ribosomale eiwitten zijn handmatig opgehaald vanuit de database. Vervolgens is er door middel van een Python (3.8, Biopython (1.77) module) script gezocht naar de SNP's van de genen die deze eiwitten coderen. Door gebruik te maken van NCBI gene database (Eric W Sayers, 2021) is er gekeken of de eiwitten correcte 40S en 60S eiwitten waren.
In hetzelfde script is er een functie ontwikkeld om de eiwit-eiwit interacties te vinden tussen de ribosomale eiwitten en de eiwitten die een interactie aangaan met ribosomale eiwitten. Dit script zoekt interacties in de String database (Szklarczyk D, 2021), hiervoor is er een confidence score (0.9) vastgesteld. Hiermee zijn alle eiwitten die een interactie aan gaan met ribosomale eiwitten en voldoen aan de confidence score verzameld. Hierdoor kan er met 90% zekerheid gezegd worden dat de gevonden eiwitten een interactie aangaan met het ribosoom.

Zoeken naar SNP's
Nadat alle ribosoom geassocieerde eiwitten en ribosomale eiwitten waren verzameld kon er naar SNP's gezocht worden in de genen van deze eiwitten. Hiervoor is een BioPython algoritme ontwikkeld en NCBI dbSNP gebruikt als bron van de SNP's. Het algoritme is verwerkt in het eiwit-eiwit interactie algoritme. Deze zoekt naar alle SNP's verwant aan de ribosoom geassocieerde eiwitten en de ribosomale eiwitten. Ook zijn er naast het type polymorfisme SNP, de types deleties en inserties aangetoond. Echter is er alleen gekeken naar het type SNP en zijn tevens de benign mutaties verwijderd.

Mapping to reference (STAR)
Het mappen van genen naar het referentie genoom is gedaan door middel van STAR. Ook zorgt de STAR-methode voor het indexen van het genoom, zodat de reads gealigned tegen het genoom konden worden.
Om de bovenstaande stappen uit te voeren zijn de volgende onderdelen van de STAR-handleiding van BioCore CRG (Cozzuto, Ponomarenko, & Bonnin, 2019) gebruikt: Course Data, Mapping With STAR en de Differential Expression Analysis (t/m Raw Count Matrices Option 1)
Voor het normaliseren zijn de raw counts gebruikt die uit de raw counts matrix kwamen. Het normaliseren werd gedaan om redundantie te voorkomen (Code Expert, 2021)

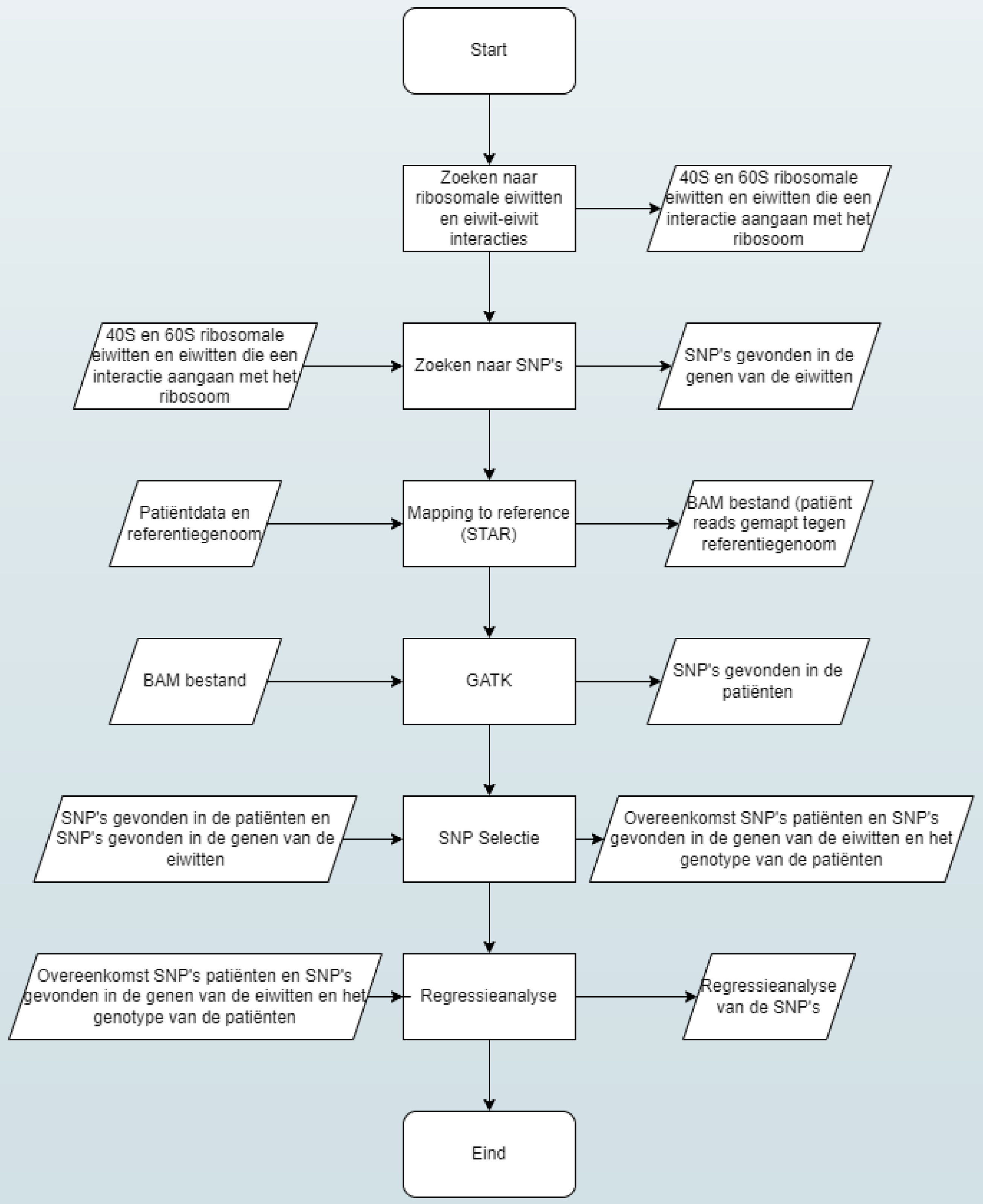
GATK
GATK (van der Auwera & O'Conner, 2020) is een tool wat alom in het onderzoeksveld wordt gebruikt voor het identificeren van indels en SNP's in DNA en RNAseq data. In dit onderzoek wordt er gebruikt gemaakt van RNAseq data, hierom is GATK uiterst toepasselijke tool voor het identificeren van SNP's. De SNP's die geïdentificeerd worden zijn SNP's afkomstig uit de patiënt testdata. Om deze SNP's te identificeren is er gebruikt gemaakt van de volgende GATK-functies: SplitNCigarReads, BaseRecalibrator, HaploTypeCaller, CombineGVCFs en GenotypeGVCFs.

Bash script
Om STAR en GATK aan elkaar vast te knopen is er een Bash script ontwikkeld welke STAR en GATK volledig uitvoert.

SNP Selectie
Het bepalen van de overeenkomstige SNP's van de patiënten en de SNP's uit de genen die voor de eiwitten coderen is er een Python script geschreven. Deze bepaald ook het genotype van de patiënten wat vereist is voor de eQTL analyse. Het genotype wordt alleen bepaald bij patiënten waarbij er een overeenkomst is van de SNP's.

eQTL analyse
Voor de eQTL analyse zijn de genormaliseerde counts en een zelf aangemaakte testset gebruikt. Deze eQTL analyse is uitgevoerd om de SNP's te vergelijken met de genexpressie, elke SNP wordt tegen de gehele genexpressie testset aangezet. Één grafiek geeft één SNP-locatie weer tegen een expressie van één gen. De analyse is uitgevoerd in Python (versie 3.9) met behulp van matplotlib (versie 3.5.2.) en scipy (versie 1.8.1.).

Snakemake
In dit onderzoek werd snakemake toegepast om alle tools en zelf ontwikkelde algoritmes sequentieel aan elkaar te maken, zodat er een pipeline ontstaat (Köster & Rahmann, 2012). De pipeline bevat de volgende tools en scripts: STAR, GATK, de zelf ontwikkelde Python script en BASH script



Figuur 2: Flowchart van het uitgevoerde onderzoek.

Resultaten

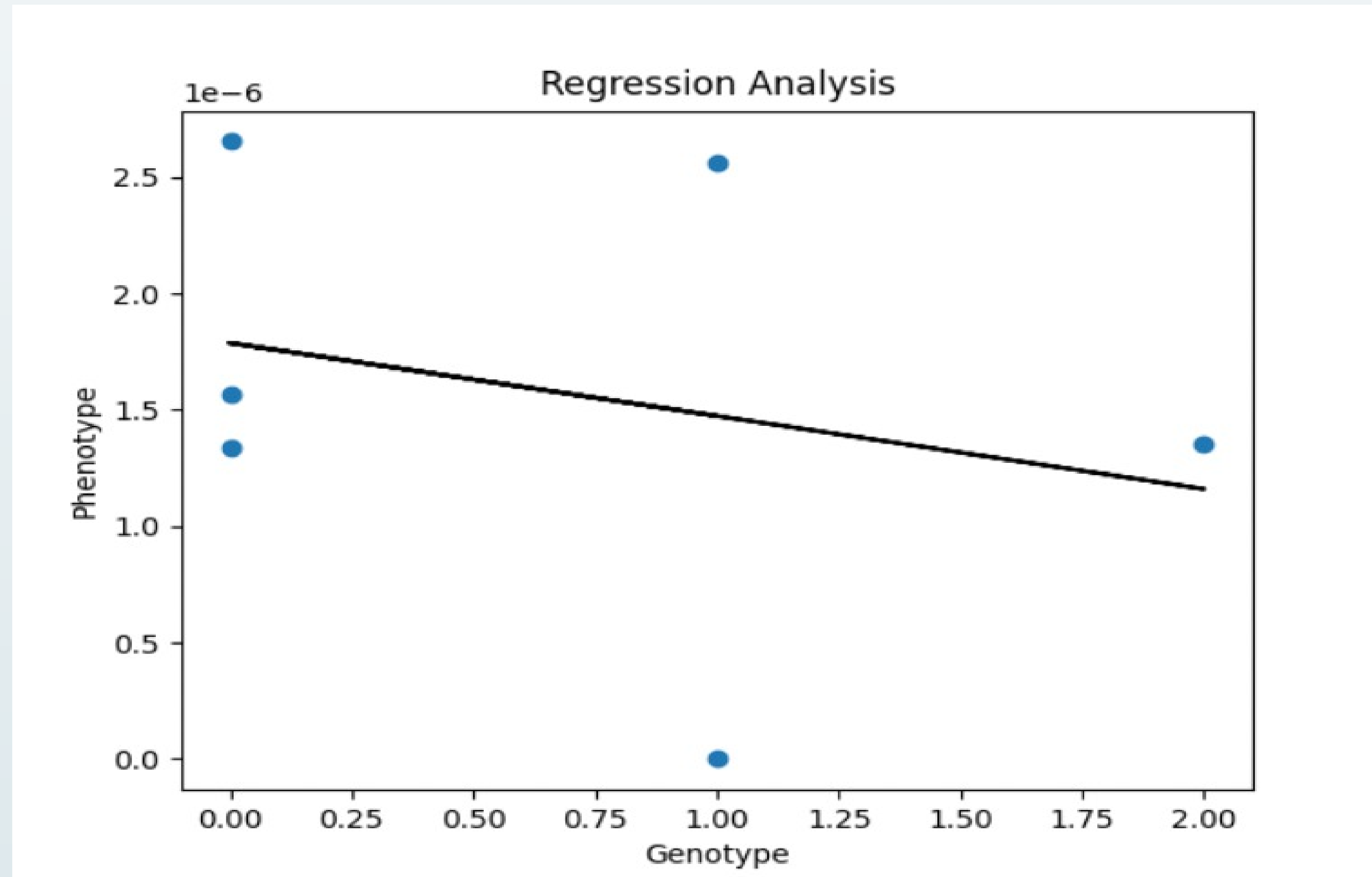
Eiwit-eiwit interacties en bijbehorende SNP's
Er is genotype data nodig voor de eQTL analyse, daarom is het van belang om de eiwitten te achterhalen die een interactie aangaan met het ribosoom en om SNP's te verkrijgen. Er zijn 36 60s ribosomale eiwitten en 35 40s ribosomale eiwitten aangetoond afkomstig van de Homo sapiens, in tabel 1 zijn de gennamen zichtbaar. Door een ontwikkeld python script zijn er 686 eiwitten aangetoond welke een interactie aan gaan met één van de ribosomale eiwitten. Er zijn 2569 SNP's aangetoond welke pathogeen zijn voor zowel de ribosomale eiwitten (71) als de interactie eiwitten (686). Na het uitvoeren van de GATK-tools zijn ervoor de test dataset 9133 SNP's aangetoond. Echter zijn er van deze SNP's geen SNP-locaties die voorkomen in de genen van interesse uit tabel 1.

Tabel 1: 60s en 40s ribosomale eiwitten afkomstig van de Homo Sapiens

60s:	40s:
RPL0	RPSA
RPLP1	RPS2
RPLP2	RPS4X
RPLP3	RPS4Y1
RPLP4	RPS4Y2
RPL6	RPS5
RPL7	RPS6
RPL7A	RPS8
RPL71	RPS9
RPL8	RPS10
RPL9	RPS11
RPL11	RPS12
RPL12	RPS14
RPL14	RPS15
RpL15	RPS16
RPL18	RPS17
RPL22	RPS18
RPL23	RPS19
RPL23A	RPS20
RPL24	RPS21
RPL26	RPS23
UBA5	RPS24
RPL29	RPS25
RPL30	RPS26
RPL31	RPS27
RPL32	RPS27A
RPL34	RPS27L
RPL35	RPS28
RPL35A	RPS29
RPL36	FAU
RPL36AL	RPS3
RPL37	RPS3A
RPL37A	RPS7
RPL38	RPS13
RPL39	RPS15A
RPL39L	

Mapping to reference (STAR)
Door de data te indexen en alignen zijn de counts per gen bepaald en opgeslagen in een .txt bestand. Deze bestand bestaat uit vier kolommen: de eerste kolom bevat de gen ID, de tweede kolom counts voor unstranded RNA, derde kolom bevat counts voor de eerste read dat gealigned is en de vierde kolom de counts voor de tweede read dat gealigned is. Vervolgens zijn deze counts genormaliseerd, zodat er met behulp van counts een eQTL-analyse gemaakt kon worden.

EQTL-analyse
Om te kunnen achterhalen of een SNP-effect heeft op de genexpressie en mogelijk kanker zou kunnen veroorzaken is er een eQTL analyse uitgevoerd. In figuur 3 is een eQTL analyse zichtbaar, die met testdata is verkregen. De genexpressie is het hoogste bij homozygoot wild type en het laagste voor homozygoot mutatie type. Dit wordt door middel van de zwarte lijn aangegeven. Hoe verder het genotype afwijkt van het wild type hoe lager de genexpressie is.



Figuur 3: toont een voorbeeld van een eQTL analyse. Op de x-as wordt het genotype aangegeven (0= wild/wild, 1= wild/mutatie en 2= mutatie/mutatie). Op de y-as worden de genormaliseerde genexpressie counts aangegeven.

Discussie & Conclusie

Voor het mappen van de data is er alleen gebruik gemaakt van reads dat zijn gemapt tegen chromosoom 10. Dit had als reden dat het erg lastig bleek te zijn om een goede publieke dataset te verkrijgen met genexpressie data van kanker en niet-kanker patiënten, waardoor er nu een test dataset gebruikt is afkomstig van STAR. Door alleen chromosoom 10 te gebruiken, zijn de resultaten ook beperkt. Aangezien de interessante SNP's nu niet gevonden konden worden bij de patiënten. Voor een vervolgonderzoek is het belangrijk om goede data te gebruiken met reads van het volledige genoom. Dit zal namelijk zorgen voor betere en betrouwbaardere resultaten. Ook is het van belang om een 32 GB aan RAM op de server te hebben, zodat bij een vervolgonderzoek het gehele genoom gebruikt kan worden.
De tools die zijn gebruikt hebben ook voor veel problemen gezorgd tijdens dit project. Mede doordat de test data van chromosoom 10 waarschijnlijk niet representatief is voor echte data, bleken een aantal tools niet goed te werken, wat meer vertraging opleverde.
Het doel van dit onderzoek was om een verband aan te tonen tussen SNP's in ribosomale en ribosoom-geassocieerde eiwitten, en een verschil in genexpressie tussen kanker- en niet-kanker patiënten door middel van een eQTL analyse. Vanwege de kleine dataset, kan dit verband niet aangetoond worden. Echter is er wel een pipeline ontwikkelt voor het verkrijgen van genexpressie en genotype data om vervolgens een eQTL analyse uit te kunnen voeren. Deze zou dan ook gebruikt kunnen worden in een vergelijkbaar onderzoek of bij een eventueel vervolgonderzoek.

Referenties

Baylin, S. B., Esteller, M., Rountree, M. R., Bachman, K. E., Schuebel, K., & Herman, J. G. (2001). Aberrant patterns of DNA methylation, chromatin formation and gene expression in cancer. *Human Molecular Genetics, 10*(7), 687-692. doi:https://doi.org/10.1093/hmg/10.7.687
Bradner, J. E., Hnisz, D., & Young, R. A. (2017). Transcriptional Addiction in Cancer. *Cell, 168*(4), 629-643. doi:https://doi.org/10.1016/j.cell.2016.12.013
Code Expert. (2021, December 8). *Kolommen van panda's dataframe normaliseren*. Retrieved from Code Expert: https://actmp2018.com/kolommen-van-pandas-dataframe-normaliseren.codeexpert
Cozzuto, L., Ponomarenko, J., & Bonnin, S. (2019). *RNAseq course 2019.* Retrieved from BioCore CRG: https://biocorecrg.github.io/RNAseq_course_2019/alnpractical.html
Eric W Sayers, E. E.-B.-N. (2021). Database resources of the national center for biotechnology information.
Nucleic Acids Research, 20-26.
Harm, & Lerick, J. S. (2022, April 06). *Project_ARC*. Retrieved from Github: https://github.com/yumsi/Project_ARC
Köster, J., & Rahmann, S. (2012). Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, 2520-2522.
Sato, N., Fukushima, N., Chang, R., Matsubayashi, H., & Goggins, M. (2006). Differential and Epigenetic Gene Expression Profiling Identifies Frequent Disruption of the RELN Pathway in Pancreatic Cancers. *Gastroenterology, 130*(2), 548-565. doi:https://doi.org/10.1016/j.cell.2016.12.013
Szklarczyk D, G. A. (2021). The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets . *Nucleic Acids Research*, 605-612.
Tweedie S, B. B. (2021). Genenames.org: the HGNC and VGNC resources in 2021. *Nucleic Acids Research*, 939-946.
van der Auwera, G., & O'Conner, B. (2020). *Genomics in the Cloud.* O'Reilly Media, Inc.
Zaimy, M. A., Saffarzadeh, N., Mohammadi, A., Pourghadamyari, H., Izadi, P., Sarli, A., . . . Tavakkoly-Bazzaz, J. (2017). New methods in the diagnosis of cancer and gene therapy of cancer based on nanoparticles. *Cancer Gene Therapy*, 233-243.