# Technische documentatie

Voor dit onderzoek is er gebruik gemaakt van Python, Bash en Snakemake. Zie figuur 1 voor de schematische weergave van alle gemaakte stappen binnen het onderzoek.

## Snakemake

Voor dit onderzoek is er gebruik gemaakt van een Snakemake, om alle tools en zelf ontwikkelde Python en Bash scripts aan elkaar vast te maken.

Er zijn 2 manieren om de Snakemake te gebruiken. Ten eerste, als je SnakeMake draait op de server die door de HAN wordt geleverd en wordt gebruikt door BIN-3f, dan hoef je alleen maar naar de map te navigeren waar de Snakefile zich bevindt. In dit geval is dat /home/administrator/final. Dan kun je gewoon het volgende commando gebruiken: Snakemake -s Snakefile, of als je een specifiek bestand wilt genereren: Snakemake -s "outputfile.txt". Let op: Het "outputfile.txt" is een voorbeeld van een willekeurig uitvoerbestand, in de gemaakte Snakefile is er niet zo'n bestand als "outputfile.txt" en je zult in de Snakefile moeten zoeken naar een outputbestand.

De tweede manier om SnakeMake uit te voeren is door Snakemake zelf te downloaden en op Linux te draaien. Het commando blijft hetzelfde, een van de import dingen zijn de locaties van het bestand, deze moeten veranderd worden omdat ze aangepast zijn aan de server waarop deze SnakeMake gedraaid heeft.

## Zoeken naar ribosomale eiwitten en eiwit-eiwit interacties

Met een lijst van de 71 40S en 60S ribosomale eiwitten is een .txt bestand gemaakt dat de gennamen bevat van deze eiwitten (eiwitten\_ribosoom.txt). Met dit bestand is het python script SNP\_data.py uitgevoerd. Het script neemt “ewitten\_ribosoom.txt” samen met een variabele bestandsnaam als input. De bestandsnaam wordt gebruikt om een output TSV-bestand te maken. De eiwitten uit eiwitten\_ribosoom.txt worden door de stringdb database gehaald om interactie eiwitten te vinden.

## Zoeken naar SNP’s

Vervolgens worden alle ribosomale en ribosoom-geassocieerde eiwitten tegen de dbSNP aangezet. Hiermee worden alle SNPs opgehaald die voorkomen in deze eiwitten, het script slaat alle pathogene SNPs op in een lijst en schrijft die later weg naar een output TSV-bestand (All\_SNPs.tsv). Het TSV-bestand bevat voor elke SNP het SNP ID, de clinical\_significance, de gen naam, de mutatie, de soort variatie en de positie.

## Mapping to reference (STAR)

Om te kunnen mappen naar het referentiegenoom is als input gebruik gemaakt van de test patientdata en het referentiegenoom. Met de input data is er een bestand gemaakt met STAR-index. Vervolgens zijn de reads gealigned en zijn de reads geteld. Als output kwam er een BAM-bestand met daarin de patient reads gemapt tegen het referentie genoom.

### Normaliseren van counts

Nadat de reads geteld (counts) waren zijn deze waardes genormaliseerd met behulp van de Python-script normalized\_counts.py. Deze Python script geeft een CSV-bestand met daarin de genormaliseerde counts.

## GATK

## BASH

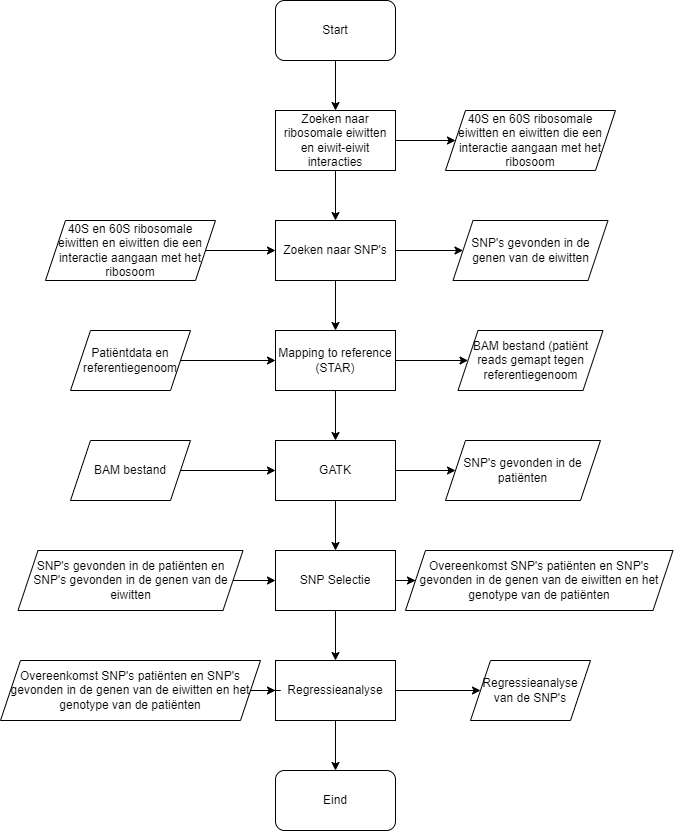
De BASH-script is geschreven om de processen binnen STAR en GATK volledig uit te kunnen voeren.

## SNP-selectie

Nadat de SNP’s van de patiënten verkregen zijn en dat SNP’s die afkomstig zijn van de ribo-interactoom, konden deze vergeleken worden met elkaar. Door middel van het python script snps\_sel.py is dit uitgevoerd. Dit script leest een gezipt vcf bestand in die de SNP’s bevat, die aangetoond zijn bij de patiënten. Ook is er een tsv bestand ingelezen die de SNP’s bevat van de ribo-interactoom. Deze SNP’s worden ook wel de SNP’s van interesse genoemd. Vervolgens zijn alle SNP locaties van de patiënten die ook voorkomen in de lijst van SNP’s van interesse opgeslagen in een nieuw bestand. In dit bestand is voor iedere SNP-locatie het genotype voor iedere patiënt opgeslagen.

## EQTL analyse

De regressieanalyse die is uitgevoerd is een eQTL analyse. Dit is uitgevoerd door middel het python script linear\_regression.py. Er zijn hier 2 input bestanden voor gebruikt, namelijk de genormaliseerde counts en het bestand waar de genotypes in staan. Als output komt er van één SNP die tegen de expressie van één gen is aangezet. En dat voor iedere SNP die een genexpressie heeft.



Figuur 1: flowchart van alle gemaakte stappen binnen dit onderzoek. Rechts staan alle input bestanden beschreven en links alle output bestanden met daar tussen het proces om tot de output bestanden te komen. .