# Technische documentatie

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

## Zoeken naar de eiwitten van de genen uit chromosoom 10

In dbSNP (van NCBI) zijn alle pathogene SNP’s van de eiwitten van alle genen van chromosoom 10 gezocht en opgehaald. Hiermee worden alle SNP’s opgehaald die voorkomen in deze eiwitten, het script slaat alle pathogene SNP’s op in een lijst en schrijft die later weg naar een output TSV-bestand (SNPs\_chr10.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 patiëntdata 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 patiënt 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. Dit Python script geeft een CSV-bestand met daarin de genormaliseerde counts.

## GATK

Voor GATK zijn er 6 commando’s toegepast per patiënt/sample. Het eerste commando (AddOrReplaceReadGroups) zorgt ervoor dat de reads in het opgegeven bestand toe worden gevoegd aan één nieuwe read-group. Het tweede commando (SplitNCigarReads) zorgt ervoor dat een read wordt gesplit op exonen en de intronen wordt weg gefilterd. De actie wordt gebaseerd op de CIGAR-string van de read. Het derde commando (BaseRecalibrator) wat is uitgevoerd, genereerd een recalibration tabel, hierin staan de read groups beschreven, de quality score, machine cycle and nucleotide context. Het vierde commando (ApplyBQSR) doet een herijking van de bases. Dit zijn de bases van de reads die in de recalibration table zijn meegeven door het 3e commando.

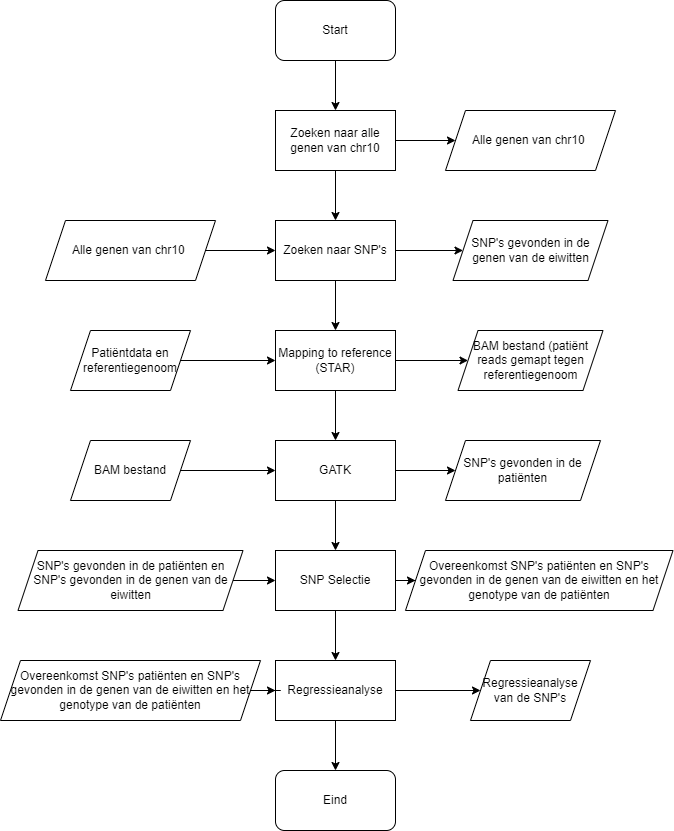
Nu komen de 2 commando's die zijn gebruikt voor de bepaling van de SNP's. Het eerste van deze commando is de HaplotypeCaller, deze zorgt ervoor dat de SNP’s en indels worden bepaald in de reads. Het allerlaatste commando (GenotypeGVCFs) zorgt ervoor dat het genotype bepaald voor HaplotypeCaller wordt weggeschreven in een VCF. De output van het eerste commando wordt gebruikt voor de twee commando en de output van commando twee wordt weer gebruikt voor commando 3 enzovoort.

## SNP-selectie

Nadat de SNP’s van de patiënten verkregen zijn en dat SNP’s die afkomstig zijn van de genen van chromosoom 10, 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 chromosoom 10 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 van het python script regressie\_analyse\_final.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.