Bachelorproef

Studiegebied	**Gezondheidszorg**
Bachelor	**Biomedische Laboratoriumtechnologie**
Afstudeerrichting	**Farmaceutische en biologische laboratoriumtechnologie**
Keuzetraject	**Bio-informatica**
Academiejaar	**2011-2012**
Student	**Svenn D'Hert**

Thema

Ontwikkeling van Illumina data analyse pipeline voor resequencing

Stageplaats

Studiecentrum voor kernenergie (SCK-CEN)

Howest-de Hogeschool West-Vlaanderen
Rijselstraat 5 – 8200 Brugge – T 050 38 12 77 – F 050 38 11 71
brugge@howest.be – www.howest.be

Bachelorproef

Studiegebied	**Gezondheidszorg**
Bachelor	**Biomedische Laboratoriumtechnologie**
Afstudeerrichting	**Farmaceutische en biologische laboratoriumtechnologie**
Keuzetraject	**Bio-informatica**
Academiejaar	**2011-2012**
Student	**Svenn D'Hert**

Thema

Ontwikkeling van Illumina data analyse pipeline voor resequencing

Stageplaats

Studiecentrum voor kernenergie (SCK-CEN)

Howest-de Hogeschool West-Vlaanderen
Rijselstraat 5 – 8200 Brugge – T 050 38 12 77 – F 050 38 11 71
brugge@howest.be – www.howest.be

Woord vooraf

De basis voor deze bachelorproef werd gelegd gedurende een praktische stage aan SCK-CEN, tijdens de maanden februari tot en met mei. De stage is een ervaring om als stagiair in een echt onderzoek mee te werken, alsook verdiepen in het onderwerp.

Dit alles is niet mogelijk zonder begeleiding, het is dan ook met vol lof dat ik dhr. Pieter Monsieurs wens te bedanken voor de constante support. De organisatie en planning van deze periode. Ook de meer onzichtbare krachten die mijn stage mogelijk maakten in de expertisegroep moleculaire biologie wens ik te danken. Dankzij hen is mijn stage en dit eindwerk mogelijk. Ook wens ik mijn promotor dhr. Raphael Kiekens danken voor de begeleiding in het schrijven van dit document.

Mijn dank gaat ook uit naar mijn ouders die mij altijd ondersteunen en die mij de kans hebben gegeven om de studies te doen.

Svenn D'Hert Kortrijk, maandag 28 mei 2012

Samenvatting

Aan SCK-CEN (studiecentrum voor kernenergie), afdeling microbiologie werkt men mee aan het MELGEN project (*Melissa Genetic Stability study*), die deel uitmaken van het MELiSSA ecosysteem onderzoek (*Micro-Ecological Life Support System Alternative*). Hierin bestudeert men onder andere de *Rhodospirillum rubrum* bacterie. Het huidige doel is, het onderzoeken van de genomische / DNA verschillen tussen de *Rhodospirillum rubrum* stam S1 ten opzichte van de S1H stam.

In deze context was er nood aan een uniforme pipeline die de trage, niet-objectieve manuele zoektocht naar variaties in resequencing data kon vervangen en verbeteren. Door de constante daling van de sequenerings prijs per genoom is de vraag naar een dergelijke tool nog sterker toegenomen.

Door optimaal gebruik te maken van de *paired-end read* technologie kunnen we voorspellingen doen over deleties en inserties in de mutant. Door gebruik te maken van SAMtools kunnen we ook SNP's (*single-nucleotide polymorphism*) identificeren.

Wij hebben een robuuste pipeline gebouwd die deleties en inserties groter dan 200bp kan detecteren met een aanvaardbare kwaliteit. De pipeline is nog niet volledig gevalideerd, maar de eerste resultaten zijn positief.

Lijst met afkortingen en symbolen

ANS	Advanced Nuclear Science
ATCC	American type culture collection [1] [2]
BGI	Beijing Genomics Institute
bp	Basenparen
BWA	Burrows-Wheeler Aligner
CH34	Stam van *Cupriavidus metallidurans*
EHS	Environment, Health and Safety
ESA	European Space Agency
INDELS	Insertie en deletie, mutatie waarbij één of meerdere nucleotiden verwijderd worden of één of meerdere ingevoegd worden
ISS	International Space Station
ITER	Vroeger : *International Thermonuclear Experimental Reactor*, deze afkorting wordt nu niet meer in die zin gebruikt
kb	1000 basenparen
MELGEN	Melissa Genetic Stability study
MELiSSA	Micro-Ecological Life Support System Alternative
MP	Mate paired
MYRRHA	*Multi-purpose hybrid research reactor for high-tech application*
PE	Paired End reads
S1/S1H	*Rhodospirillum rubrum* stammen
SAM	sequence alignment/Map
SCK	Studiecentrum voor kernenergie
SNP	*Single nucleotide polymorphisme*, Enkelnucleotide polymorfisme

Verklarende woordenlijst

Contigs	Consensus DNA segmenten, opgebouwd uit reads
Denaturatie	Breken van de waterstofbruggen tussen AT en CG, met verlies van helix structuur tot gevolg
Denovo assembly	Bepalen van de sequentie zonder gebruik te maken van bestaande referentiesequenties
Genomics	Studie van het genoom
Genotyping	Het proces om de genetische "make-up" verschillen tussen individuen of populaties te bepalen.
Mapping	Een strenge vorm van lokale alignering, waarbij een read op een overeenkomende referentiesequentie wordt "geplakt"
Multifasta	Meerdere fasta sequenties in één bestand
Paired-end[1]	Beide einden van een kort stuk DNA worden gesequeneerd. (zie ook 1.4.1)
Phred quality score	Kwaliteit score die voor individueel base is bepaald tijdens sequenering
proteobacterie	bacteriën waarvan vele belangrijke stikstofvastleggers zijn. Ze kunnen de stikstof uit de lucht vastleggen. Er zijn ook ziekteverwekkers die behoren tot de proteobacteriën [3]
Proteomics	Studie van alle eiwitten in een organisme
Read	de korte nucleotide sequentie – term specifiek gebruikt bij *Next-Generation Sequencing technologies*
SAM	Uniform bestandsformaat om gealigneerde reads op te slaan
Transcriptomics	Studie van alle RNA moleculen in een organisme
T-test	Een parametrische statische test om na te gaan of een gemiddelde van een normaalverdeelde grootheid afwijkt van een bepaalde waarde.

[1] Dit is de definitie volgens Illumina, Inc.

Lijst van tabellen en figuren

Inhoudsopgave

1 Inleiding, probleemstelling en situatieschets

Sequeneren heeft de afgelopen decennia een exponentiële groei gekend en de kostprijs is enorm gedaald. Daardoor krijgen meer onderzoekscentra toegang tot de sequenties van de organismen waarmee ze werken. Veel organismen zijn evenwel reeds gesequeneerd. Deze informatie kan gebruikt worden, om nauw verwante organismen te gaan hersequeneren. Dit is het sequeneren van een genoom, gen of regio in het genoom war reeds een referentie voor bestaat. Het doel; de zoektocht naar variaties ten opzichte van de referentie (het verwante organisme) en onrechtstreeks de sequentie, dit vereist evenwel ruime kennis en ervaring van de onderzoeker, daarbij zijn resultaten niet altijd even objectief en is deze zoektocht arbeidsintensief. Dit samen met de snelle productie van sequencing data (*next generation sequencing*) zorgt dat er nood is aan een uniform en geautomatiseerde systeem. Een tool dat *Illumina sequencing data* samen met een referentie analyseert en herleid tot biologische relevantie.

In SCK•CEN[2], afdeling microbiologie werkt men mee aan de MELGEN-1 en -2 (*Melissa Genetic Stability study*) projecten, die als doel hebben; de effecten van ruimtereizen[3] en compartiment omstandigheden[4] te onderzoeken op cellulair, *proteomic*, *transcriptomic* en *genomic* niveau van de bacteriën uit het MELiSSA ecosysteem te bepalen (*Micro-Ecological Life Support System Alternative*) [4].

In deze context werd een pipeline ontwikkeld, onder andere om de resequentieanalyse van de bacterie *Rhodospirillum rubrum* (S1, S1H) te versnellen. De bacterie werd gehersequeneerd omdat ze belangrijk is tijdens de 2de stap in het MELiSSA ecosysteem. In dit compartiment wordt koolstof getransformeerd onder anaerobe omstandigheden onder invloed van lichtenergie [5].

[2] Studiecentrum voor Kernenergie – Centre d'Etude de l'énergie Nucléaire, in Mol
[3] straling, microzwaartekracht, vibratie en magnetisme
[4] pH, temperatuur, licht, invloed van vorige compartimenten

1.1 SCK-CEN

Het studiecentrum voor kernenergie (SCK-CEN), gelegen in Mol, is één van de grootste onderzoekscentra in België en is opgericht in 1952, tien jaar nadat de eerste kernreactor (*Chicago Pile-1*) kritisch werd. Dit is de toestand wanneer evenveel neutronen worden geproduceerd als er worden geabsorbeerd door de regelstaven [6] [7].

Het onderzoek in SCK-CEN is voornamelijk gericht naar maatschappelijk belangrijke thema's, zoals veiligheid van nucleaire installaties, stralingsbescherming, veilig beheer en opslag van radioactief afval. Er zijn drie onderzoeksdomeinen waar SCK-CEN in actief is, namelijk EHS (*Environment, Health and Safety*), NMS (*Nuclear Materials Science)* en ANS (*Advanced Nuclear Systems*) [8] [9].

1.1.1 ***Nuclear Materials Science*** **(NMS)**

Nucleaire Materiaalwetenschappen (NMS, Nuclear Materials Science), doet onderzoek naar structurele en functionele materialen die deel uitmaken van vreedzame nucleaire systemen. Daarnaast produceert het NMS ook radio-isotopen voor medische beeldvorming en therapie. De radio-isotopen worden bijvoorbeeld gebruikt bij (bij-) schildklierscans [10]. Daarnaast wordt ook silicium bestraald voor de industriële halfgeleiderproductie. Hiervoor gebruikt men de "Belgische reactor 2" (BR2) [11].

1.1.2 **Advanced Nuclear Systems (ANS)**

Geavanceerde Nucleaire Systemen (ANS, *Advanced Nuclear Systems*), ontwikkelt en test nieuwe technologieën en instrumentatie voor innovatieve reactorconcepten zoals de MYRRHA (*Multi-purpose hybrid research reactor for high-tech application*) een snelle loodgekoelde reactor. MYRRHA moet tegen 2023 volledig operationeel zijn. Daarnaast werkt het mee aan de ITER, een internationale testfusie reactor in Frankrijk. Het doet ook experimenten op de BR1 (Belgische Reactor 1) en VENUS (*Vulcain Experimental Nuclear Study)*. [12]

1.1.3 ***Environment, Health and Safety*** **(EHS)**

Milieu, Gezondheid en Veiligheid (EHS, Environment, Health and Safety), doet onderzoek naar nucleaire veiligheid, afvalbeheer, bescherming van mens en milieu, beheer van splijtstoffen, … Zo bestudeert het de biologische effecten van lage stralingsdosissen. Daarnaast bestudeert het ook aanpassingen van bacteriën in extreme omstandigheden, zoals de ruimte, Antarctisch platform, vervuilde bodems en onder straling. Ook gebeurt er onderzoek in medische toepassingen, voornamelijk procedures met hoge dosissen straling en toepassingen van straling in de kindergeneeskunde. Ook bestudeert het EHS de manier waarop radioactieve stoffen zich verspreiden via de lucht, de biosfeer en de geosfeer en evalueert de impact van ioniserende straling op mens en omgeving. [13]

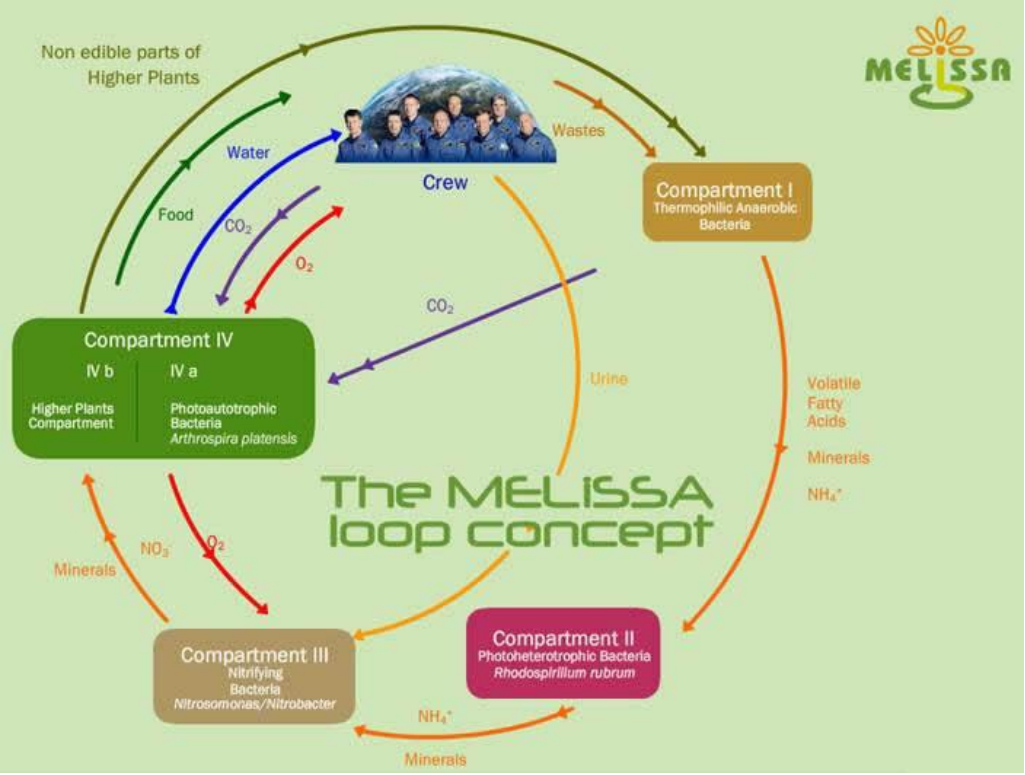
Een van de expertisegroepen van het EHS is de afdeling microbiologie, een van de projecten waaraan men aan meewerkt, is het onderzoek naar het MELiSSA ecosysteem (*Micro-Ecological Life Support System Alternative)*. Het doel van het MELiSSA project is het opzetten van een micro-ecosysteem waar afval en CO_2 verwerkt word tot voedsel en zuurstof, om langdurige ruimtereizen mogelijk te maken. Binnen MELiSSA zit het MELGEN -1 en -2 projecten *(Melissa Genetic Stability study)*. *“Deze studie onderzoekt de effecten van compartiment omstandigheden (T°C, pH, licht, invloed van vorige compartiment substanties), externe fysische omstandigheden (ruimte ioniserende straling, micro zwaartekracht, trillingen, magnetisme) en evolutionaire processen op de bacteriën. Hiervoor worden de bacteriën onderzocht op cellulair, proteomic, transcriptomic en genomic niveau”* [14] [15] [16].

Het MELiSSA ecosysteem is een project in samenwerking met onder andere VITO, UGent, Universiteit van Clermont Ferrand, Universiteit van Barcelona en andere. Het is gecoördineerd door ESA (*European Space Agency*) [17].

1.2 Onderzoek & test cases

1.2.1 Rhodospirillum rubrum S1/S1H

De Gramnegatieve, facultatief anaerobe, –onder anaerobe omstandigheden- paarse proteobacterie, *Rhodospirillum rubrum* word gebruikt in compartiment twee van het MELiSSA ecosysteem (*Micro-Ecological Life Support System Alternative,* zie figuur 1.1).



Figuur 1.1 : MELiSSA ecosysteem

In dit compartiment worden voornamelijk “vluchtige” vetzuren, CO_2, H_2 en H_2S van het eerste compartiment verwerkt. *Rhodospirillum rubrum* is extra interessant ten opzichte van andere leden van de familie omdat hij geen toxines produceert, en hierdoor als complementair voedsel kan gebruikt worden [15]. Onlangs is de *Rhodospirillum rubrum* S1 (ATCC 11170) volledig gesequeneerd (2011) als onderdeel van het DOE *Joint Genome Institute,* programma DOEM 2002 door Munk *et al* [18]. Een van de doelen van de MELGEN projecten is de verschillen tussen de S1 stam en de *Rhodospirillum rubrum* S1H (ATCC 25903) bepalen op genomic en proteomic niveau [19]. De sequenering gebeurde door *Beijing Genomics Institute*

(BGI), op een HiSeq 2000, in 2012 [20] [21] [22]. Daarnaast bestaat er een reeds gesequeneerde mutant van *Rhodospirillum rubrum* namelijk de F11 stam, dit is een geïsoleerde mutant van de S1 stam (ATCC 11170). De sequencing gebeurde op een *Illumina GAii* door *Lonjers ZT, et al* (september, 2011) [23]. We gebruiken deze als alternatieve referentie, de F11 stam heeft evenwel geen plasmide [24].

Tabel 1.1 : genoom van referentie S1 (NCBI)

GeneBank ID	*Lengte*	*Type*
NC_007643	4,352,825 bp	Chromosoom
NC_007641	53,732 bp	Plasmide

1.2.2 ***Cupriavidus metallidurans* CH34**

De proteobacterie *Cupriavidus metallidurans* stam CH34, vroeger gekend onder de namen *Ralstonia metallidurans*, *Ralstonia eutropha* en *Alcaligenes eutrophus* is een niet sporenvormende, staafvormige bacterie. De bacterie is voornamelijk bekend voor de resistentie tegen zware metalen. Opmerkelijk bij de *Cupriavidus metallidurans* is dat het genoom van deze bacterie uit twee chromosomen bestaat, namelijk NC_007973, chromosoom 1 (3.928 kb) en NC_007974, chromosoom 2 (2.580 kb). Naast deze chromosomen zijn ook twee megaplasmiden aanwezig namelijk NC_007971, pMOL30 (233 kb) en NC_007972, pMOL28 (170 kb) [25]. Megaplasmiden zijn plasmiden die een lengte hebben die groter is dan 100 kb.

1.2.2.1 Zilver resistentie

In een experiment heeft men aan de voedingsbodem van gekweekte *Cupriavidus metallidurans* (*Minimal Inhibitory Concentration* (MIC) van 0.5 µM) een lethale dosis zilver ($AgNO_3$) toegevoegd. Na zeven dagen heeft men twee onafhankelijke mutanten van de *Cupriavidus metallidurans* geïsoleerd (AE2720, AE2722). Deze mutanten hadden een hogere zilver resistentie (tot 20 maal hoger, respectievelijk 40µM en 80µM). Met behulp van microarray analyse heeft de genexpressie van beide mutanten ten opzichte van de *Cupriavidus metallidurans,* stam CH34 vergeleken. De expressiewaarden van de mutanten ten opzichte van de CH34 waren zeer sterk verschillend. Vreemd genoeg waren expressiewaarden die genen

waarvoor gekend was dat ze van belang zijn voor zilverresistentie binnen de *Cupriavidus metallidurans* niet differentieel tot expressie kwamen ten opzichte van de wild-type stam CH34. Er is van beide mutanten een volledige genoom *sequencing* gebeurt op een Illumina GAIIx platform, met 50bp *paired-end reads* (PE) en 300bp *insert size* door *BaseClear NV* [26] [27].

1.2.2.2 NASAIV

De bacterie, *Cupriavidus metallidurans* werd in het verleden meerdere malen in de ruimte geïsoleerd uit het drink- en koelwater van de spaceshuttle, het Mir ruimtestation en het ISS (*International Space Station*). Dat deze bacterie hierin kan overleven is een opmerkelijke eigenschap, omdat in dergelijke watervoorraden zilver als biocidaal product gebruikt wordt. Behalve deze praktische toepassingen, is het ook vanuit evolutionair oogpunt belangrijk om te achterhalen hoe deze bacterie zijn verhoogde zilverresistentie verkrijgt. De *sequencing* van een van de overlevende stammen in het ISS gebeurde op GS FLX+ system (454-sequencing Roche) en de geproduceerde contigs werden gebruikt als referentie voor recent gesequeneerde mutant, waarvan de resistentie nog verhoogd was.

Tabel 1.2 : genoom *Cupriavidus metallidurans stam CH34*

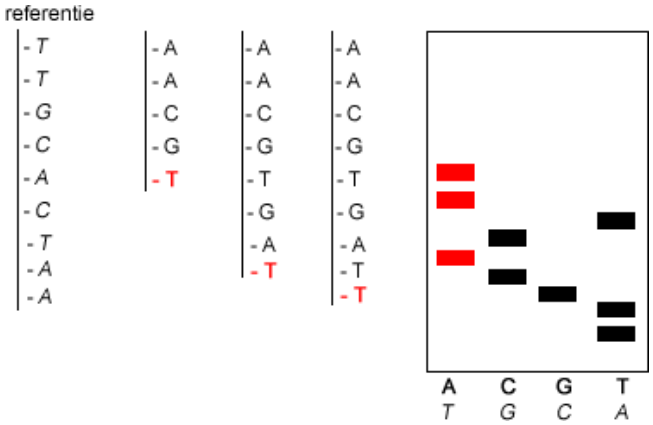
GeneBank ID	*Lengte*	*Type*
NC_007973	3,928,089 bp	Chromosoom
NC_007974	2,580,084 bp	Chromosoom
NC_007971	233,720 bp	Megaplasmide
NC_007972	171,459 bp	Megaplasmide

1.3 Sequeneren

Sequeneren, in een biologische context, is het bepalen van de volgorde van de basen van erfelijk materiaal (DNA/RNA). Al in 1976 is het eerste erfelijke materiaal van een organisme gesequeneerd, namelijk de bacteriofaag *MS2*, toen op RNA niveau [28]. Een jaar later was het volledige genoom van de bacteriofaag *phi X 174* op DNA niveau gesequeneerd [29]. Het eerste gesequeneerd bacteriële genoom was dit van de *Haemophilus influenzae* in 1995 [30].

1.3.1 ***Dideoxy chain termination* methode**

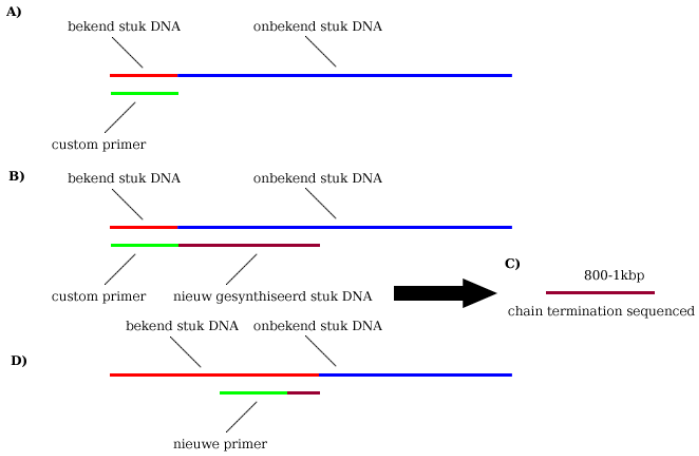
De dideoxy chain termination methode, (ook gekend als Sanger sequencing) beschreven in 1975 was tot rond de eeuw wisseling de norm. De methode gaat in zijn werk door vier aparte reacties op te zetten met gekloneerd single *stranded* DNA met voor iedere base een apart reactie medium. In dit medium zijn de vier standaard basen aanwezig plus één base die gemodificeerd is zodat er geen verdere synthese meer kan gebeuren eens deze gebonden is op de originele strand (de zogenaamde dideoxynucleotiden). Hierdoor bekomt men verschillende stukken die steeds eindigen op de gemodificeerde base. Wanneer men deze nu met behulp van elektroforese op moleculaire grote scheidt kan men aan de hand van de afgelegde lengte de basen aflezen van de gel. Nadien is deze methode verder geoptimaliseerd door het gebruik van fluorescent gelabelde dideoxynucleotiden, wat toelaat om de sequeneringsreactie door te voeren in één medium met een mengsel van deze vier nucleotiden apart. Deze methode wordt de dye-termination methode genoemd, en wordt in een high-througput manier toegepast bij sommige next-generation sequencing technologieën. De methode kan DNA stukken accuraat lezen tot ongeveer 800bp lang (zie figuur 1.2). Om toch grotere stukken te kunnen lezen zijn er twee methoden gekend, *primer walking* of *chromosome walking* en de veel bekendere -door het *Human Genome Project* - de *shotgun* methode (zie figuur 1.3 en 1.4) [31] [32].



Figuur 1.2 : schematische voorstelling van *chain termination*, hier rood aangeduid voor A (referentie, of T als complementaire base)

1.3.1.1 *Primer walking* methode

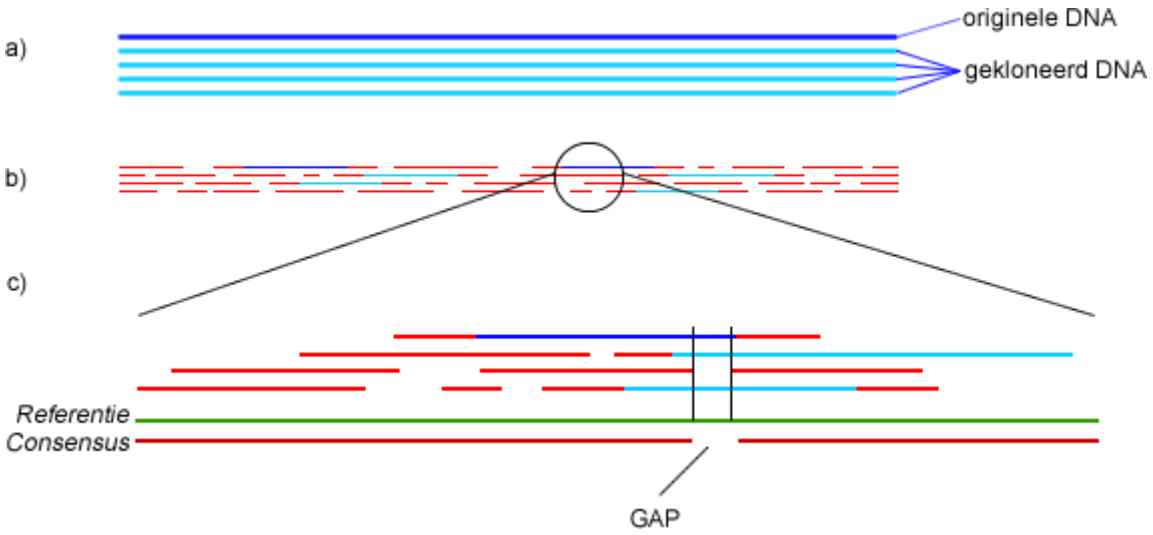
Bij *primer walking* ook *genome walking* genoemd, wordt gebruik gemaakt van aangepaste primers, de primer gaat binden op het bekende stuk DNA (zie figuur 1.3, A) door toevoegen van basen, zal de gebonden primer groeien (zie figuur 1.3, B). Door *denaturatie* bekomt men een stuk DNA van die complementair is aan het onbekende stuk. Dit kan met behulp van *dideoxy chain termination* methode gaan sequeneren. Door in het nieuw gesequeneerd stuk DNA een nieuwe primer te gaan zoeken kan men de hele procedure overdoen om het volgende stuk te bepalen [33] [34]. *Primer walking* wordt –zelfs bij het gebruik van next-generation sequencing technologie- nog vaak gebruikt om *gaps* te sluiten. [35].



Figuur 1.3 : schematische voorstelling van *primer walking* methode, A) eerst bindt de custom primer aan het bekende stuk. B) men voegt nucleotiden aan toe, die binden aan de primer. C) chain termination sequencing. D) nieuwe primer in het nieuw gesequeneerde stuk.

1.3.1.2 *Shotgun* methode

Bij de *shotgun* methode wordt het DNA meerdere keren geamplificeerd (figuur 1.4, A) en vervolgens in willekeurige stukken geknipt (figuur 1.4, B). Deze worden dan allemaal gesequeneerd, door overlappende stukken te gaan bepalen kan men de originele sequentie achterhalen (figuur 3, D). Dit is computationeel erg intensief en vereist voor grotere genomen, zoals het humane, een gebalanceerde load over meerdere parallelle servers [31] [36]. Er worden miljoenen stukken geproduceerd. Daarom is het onmogelijk om de originele DNA streng samen te stellen zonder gebruik van tools.



Figuur 1.4 : Schematisch overzicht *shotgun* methode; A) de originele DNA streng wordt gekloneerd B) het DNA word in willekeurige stukken geknipt C) de originele streng wordt berekent door overlappende stukken DNA te mappen. Eventuele *gaps* kunnen voorkomen (zie verder). Het bekomen resultaat noemt men de consensus sequentie. (contigs)

Om de volledige originele sequentie te achterhalen, moet men de streng amplificiren om te zorgen dat iedere nucleotide minstens enkele keren word gesequeneerd. Het aantal keren dat een nucleotide in een te sequeneren stuk zit noemt men de coverage (C). Om deze te berekenen heeft men de hoeveelheid gesequeneerde reads nodig (R), de gemiddelde lengte van de read (L) en de totale lengte van het genoom (G) [37].

$$C = \frac{RL}{G}$$

Formule 1.1 : berekening *coverage*

Wanneer de steekproef perfect en willekeurig uniform is, dan gelden de volgende uitspraken [37].

- De kans dat een base niet gesequeneerd is e^{-C}
- Deel van genoom dat bedekt is $1 - e^{-C}$
- Totale lengte van alle kloven (gaps) Ge^{-C}
- Totale hoeveelheid van kloven Re^{-C}
- De gemiddelde lengte van iedere kloof $\frac{Ge^{-C}}{Re^{-C}} = \frac{L}{C}$
- De gemiddelde lengte van iedere contig $\frac{G}{Re^{-C}} = \frac{L}{C}e^{C}$

1.4 *Next Generation Sequencing*

Sinds de sequenering van eerste bacteriële genoom, in 1995 is er een enorme groei geweest zowel in snelheid van het sequeneren, als in de kostprijs per gesequeneerde base, en dit door de constante innovatie van de markt. Op de huidige markt zijn een aantal zogenaamde *Next Generation Sequencing* toestellen beschikbaar. [38]

De nieuwe generatie *Next-Generation Sequencing* toestellen genereren zogenoemde “short” reads. Dit zijn stukken DNA met een lengte van die varieert tussen de 25 en 500bp. In geval van de *Rhodospirillum rubrum* Illumina sequencing bedraagt de read lengte, 90bp. Dit betekent dat d*enovo assembly* een complexe taak is, zowel op het vlak van de ontwikkeling van algoritmen, als op het technologische vlak (CPU en geheugen gebruik). Indien mogelijk gebruikt men daarom reeds bestaande *assembly’s* of referenties. Wanneer deze beschikbaar zijn, spreekt men over hersequeneren of *resequencing*. Bij *resequencing* gebruikt men *mappers*, dit zijn tools die de short reads op een bestaande referentie passen.
Bij resequencing zijn twee methoden mogelijk, enerzijds een beperkte hersequenering van bepaald gebied of gen (*genotyping)* of anderzijds het volledige genoom hersequeneren om mutaties te vinden. Het verschil bestaat erin dat bij een volledige hersequenering genome rearrangements kunnen worden gevonden zoals inserties (extra gebieden) en deleties (verdwenen gebieden), terwijl dit bij beperkte hersequenering niet mogelijk is. Bij beide methoden van hersequeneringen kunnen ook INDELS (kleine insertie & deletie) en *single nucleotide polymorphismen* (SNP) worden gevonden [39].

Bij *Illumina sequencers* gebruikt men de *dye termination* methode, de detectie van de base gebeurt met een gelabelde base. Zodra de gelabelde base bindt op de enkele DNA streng zal deze licht uitzenden, de frequentie van het licht is voor de vier verschillende basen verschillend, waardoor men kan bepalen welke base gebonden is op de positie. Deze methode is veel sneller doordat alles “real-time” kan gebeuren in één reactie, in tegenstelling tot bijvoorbeeld *dideoxy chain terminati-*

on. Een van de vele innovaties van de sequence technologie is zogenoemde "*paired reads*" dit zijn reads die naast de sequentie ook informatie geven van hun relatieve positie. Er zijn twee gangbare formaten naast de originele *single read*, namelijk *paired end* en *mate pair* reads.

Het grote verschil tussen *mate pair reads* en *paired end reads* zit in het circulair maken van het DNA, dit is een inefficiënt proces en zorgt ervoor dat tot vier keer meer DNA nodig is. Daarnaast is er voor *mate pair reads* een hogere fout marge, dit komt door de lengte selectie [40]. Ook is bij *paired end* de variatie op de *insert size* kleiner, maar is bij mate pair de *insert size* groter.

1.4.1 ***Paired end reads* (PE)**

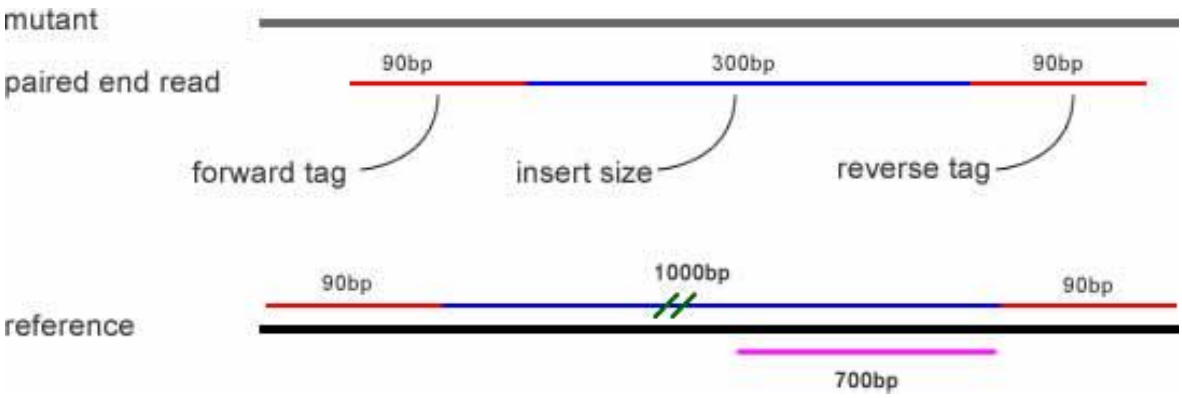
Bij *paired end reads* gaat men fragmenten maken van een bepaalde lengte (200-500bp), die men vervolgens gaat sequeneren aan beide uiteinden. Hierdoor krijgt men drie informatiebronnen die nuttig zijn om de originele sequentie te achterhalen, namelijk :

- De voorwaartse sequentie (*forward tag*)
- De achterwaartse sequentie (*reverse tag*)
- De lengte tussen twee sequenties (*insert size*)

Uit deze informatie kan men eenvoudiger mappen en *genome rearrangements* zoals inserties, deleties en inversies vinden. Bovendien is *denovo assembly* eenvoudiger omdat men eenvoudiger repetitieve regio's kan identificeren en overbruggen, ten opzichte van single reads [41] [40].

1.4.1.1 Verdwenen regio's identificeren

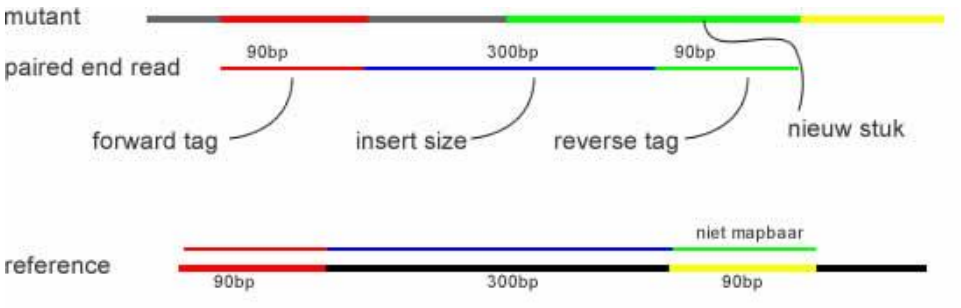
Wij gebruiken *paired-end* data. Zo werd voor de *Rhodospirillum rubrum* (S1/S1H) een *insert size* van 500bp gebruikt en voor *Cupriavidus metallidurans* stam CH34 300bp. Deze informatie maakt het eenvoudiger om variaties ("genome rearrangements") zoals extra regio's of verdwenen regio's te gaan bepalen. Gezien de afstand op de gesequeneerde streng bv. 300bp is zouden de twee gemapte reads, op de referentie een afstand van 300bp moeten hebben. Wanneer dit echter niet het geval is en de afstand op de referentie groter is, is een stuk DNA die tussen de twee reads zit op de mutant verdwenen (deletie). (zie figuur 1.5)



Figuur 1.5 : De afstand op de mutant is 300bp, op de referentie is het evenwel 1kb. Er is dus een regio van 700bp (paars gebied) op de referentie die verdwenen zijn op de mutant.

1.4.1.2 Nieuwe regio's identificeren

Een nieuwe regio op de mutant kan men eveneens identificeren, door gebruik te maken van de paired-end eigenschap. De *forward tag* zit 300 bp verwijderd van de *reverse tag* hierdoor zal een insertie die groter is dan 300bp enkel maar de forward tag mappen, de reverse tag kan niet worden gemapt omdat deze in het nieuwe gebied valt. Over een regio van de lengte van de *forward tag* plus de *insert size* zullen enkel de forward tags kunnen mappen. Na de insertie zullen enkel de *reverse tags* mappen voor een gebied van dezelfde lengte. Deze eigenschap kan worden gebruikt om nieuwe regio's te identificeren. Tussen de regio voor de insertie (waar enkel *forward tags* gemapt kunnen worden) en na de insertie (waar enkel *reverse tags* gemapt kunnen worden) kan geen enkele *read* correct gemapt worden, tenzij het over een repititief element gaat zoals een transposon of insertiesequentie (zie figuur 1.6).



Figuur 1.6 : door dat we zeker weten dat de forward en reverse tag ongeveer 300bp van elkaar liggen verwachten we dat één tag zal mappen en één niet zal mappen.

1.4.2 *Mate pair reads*

Bij *mate pair reads* gaat men grotere fragmenten (bijvoorbeeld voor Illumina 2-5kb) gaan selecteren. Daarna gaat men deze fragmenten op de uiteinden met een *biotine* label labelen en maakt men de sequentie circulair. Niet circulaire fragmenten worden verwijderd door enzymatisch digestie. Vervolgens gaat men deze opnieuw fragmenteren (400-600bp) gevolgd door de zuivering en amplificatie van de *biotine* gelabelde fragmenten. Daarna gaat men net zoals bij *paired end reads* beide uiteinden sequeneren. Hierdoor krijgt men opnieuw drie informatiebronnen [40] : de voorwaartse sequentie (*forward tag*), achterwaartse sequentie (*reverse tag*), lengte tussen de twee tags (*insert size*). Het grote voordeel van mate pair reads is dat met repetitieve regio's tot 5kb kan overbruggen, en dus eenvoudiger de volgorde van verschillende contigs bij *denovo assembly* kan bepalen.

1.5 Formaten

We gebruiken tijdens de run van de pipeline enkele uniforme formaten. Het is belangrijk te weten dat deze meestal kunnen omgezet worden van meer exotische extensies.

1.5.1 FASTQ

FASTQ formaat is een tekst gebaseerd bestandsformaat. Wij gebruiken het om DNA sequenties mee op te slaan. Het formaat is opgebouwd uit vier lijnen. Bij Illumina *paired-read* of *mate-pair* sequencing zijn er steeds twee fastQ's namelijk met forward reads en reverse reads (zie verder) [42].

Tabel 1.3 : waardes in FASTQ formaat

#	**Informatie**
1	Sequentie indentificatie
2	Sequentie
3	"+" teken
4	*Phred score*

1.5.2 SAM formaat

Het SAM (Sequence Alignment/Map) formaat is een CSV file formaat (*Comma-seperated values)*, als *delimiter* is evenwel een tab gekozen, in plaats van de komma. Er zijn meerdere kolommen, en er is ook ruimte voorzien voor extra variabelen (zie tabel 1.4). Er is ook een Bitwise flag, dat waarden bevat (zie tabel 1.5). Het is de bedoeling dat het 1000 genomes project, een project dat 1000 humane genomen wenst te sequeneren, ook in dit formaat zal worden vrijgegeven [43].

Tabel 1.4 : opdeling variabelen uit sam bestanden

#	**Variabele**	**Betekenis**
1	Qname	Naam van read
2	Flag	Bitwise flag (zie verder)
3	Rname	Naam van de referentie
4	Pos	Positie van meest linkse base
5	Mapq	*Mapping quality*
6	Cigar	*Cigar string*
7	NRNM	*Ref. name of the mate /next segment*
8	Mpos	Positive van meest linkse base van paar
9	Isize	Lengte tussen 2 reads
10	SEQ	Sequentie
11	QUAL	Pred-scaled base quality

Tabel 1.5 : informatie uit *bitwise flag*

Bit	Variabele	Betekenis
0x1	paired_seq	Reads zijn paired
0x2	proper_mapped	Reads zijn goed gemapt
0x4	quer_unmapped	Read is unmapped
0x8	mate_unmapped	Piar read is unmapped
0x10	strand_quer	Strand of query (ligging van query)
0x20	strand_mate	Strand of mate (ligging van mate)
0x40	first_read	Dit is de eerste read
0x80	second_read	Dit is de tweede read

2 Materiaal en methoden

2.1 Materiaal

Er waren meerdere computers beschikbaar, de software is op allen geïnstalleerd en getest (zie tabel 2.1).

Tabel 2.1 : overzicht server operating systemen

Operating system	*Versies*	machine	*Architecture*	*Based*
Ubuntu	10.04, 10.10, 12.10	Server 1	32 bit	Debian
CrunchBang	10	Server 1	32 bit	Debian
CentOS	5	Server 2	64 bit	Red Hat

Dit zijn de specificaties.

Tabel 2.2 : overzicht beschikbare rekenkracht

Machine	*Processorkracht*	*RAM*
Server 1	8 x 2.83 Ghz	16 GB
Server 2	2 x 2.40 Ghz	4 GB (32 bit)

2.2 Simulatie

2.2.1 dwgsim

Dwgsim was een onderdeel van DNAA (DNA Analysis Package) maar is nu een *standalone* tool om next-generation sequencing data te simuleren uitgaande van een multifasta bestand. Het is gemodificeerd om ABI SOLiD data te produceren, maar produceert naadloos Illumina data vanuit een al dan niet gemodificeerde referentie. Simulaties werden gebruikt om de pipeline te valideren. Wij werkten met versie 0.1.10 [44].

2.3 Mapping

2.3.1 Burrows-Wheeler Aligner

Burrows-Wheeler Aligner (BWA) is een snelle, accurate, geheugenefficiënte implementatie van het *Burrow-Wheeler transformation* algoritme, dat korte queries (*reads)* mapped op een referentie. In het programma zijn twee aparte implementaties beschikbaar, namelijk bwa-short (IS) en bwa-sw. De eerste implementatie is gemaakt om *reads* tot 200bp met lage foutenfrequentie (<3%) te aligneren, de tweede implementatie is gemaakt om langere *reads* met meer foutenfrequentie te gaan aligneren. [45] [46] Tijdens de ontwikkeling van de pipeline hebben we gebruik gemaakt van versie 0.6.1, deze versie *mapped* niet-unieke *reads* op een willekeurige positie. Als alternatief hebben wij een gepatchte versie gebruikt die niet-*unieke reads* allemaal *mapped* en deze reads allemaal in de resultaten opneemt in tegenstelling tot de huidige versie van BWA (0.6.1), deze versie was 0.5.7.

2.3.2 Bowtie

Bowtie is een snelle, geheugenefficiënte, korte reads *mapped*, die ook gebruik maakt van het Burrow-Wheeler transformation algoritme. Wij gebruikten tijdens de ontwikkeling van de pipeline Bowtie *2.0.0-beta2*, als alternatief op BWA, om geen specifieke resultaten van BWA te gebruiken die enkel door BWA zouden ondersteund worden. Op deze manier kon de uniformiteit van de pipeline aangetoond worden [47].

2.4 Variantie calling

2.4.1 Samtools

SAMtools is een pakket dat verschillende tools combineert om informatie te filteren uit het SAM formaat. Zo kan het SAM omzetten naar een binaire vorm, BAM. Ook biedt het een rudimentaire SAM verkenner aan. Het komt samen met BCFtools, dit is een pakket die werkt met BCF en VCF bestanden, dit zijn bestanden die gebruikt worden voor SNP calling. Wij gebruiken versie 0.1.8 [48].

2.5 PipeLine

2.5.1 perl

Perl is een programmeertaal ontworpen door Larry Wall en beschikbaar gesteld in 1987 op een nieuwsgroep. Het is een uitbreiding op de toenmalige op de *Unix-shell scripts*, *Bourne Again-shell*. Het is een *Swiss army knife of programming languages*, een eenvoudige, *high-end* krachtige programmeertaal. Het is een ideale taal om een pipeline mee te bouwen omdat het cross-platform is. Daarnaast is het ook mogelijk meerdere oplossingen te vinden voor één probleem. Het richt zich ook op tekst manipulatie waardoor het voor tekst analyse zoals sequentie data uitermate geschikt is. Perl zelf kan worden uitgebreid met modules, die vrij beschikbaar zijn via *Comprehensive Perl Archive Network* (CPAN). Daarnaast is er ook een *toolkit* beschikbaar voor bio-informatici ontwikkeld genaamd BioPerl. Wij werken met versie perl 5.14 [49] [50] [51] [52].

2.5.2 R

R is een programmeertaal voor statistische doeleinden. Het is ook handig voor visualisatie van grafieken en daarenboven multiplatform waardoor het als data-manipulatie en analyse tool erg handig blijkt voor grote hoeveelheden data, waarmee bio-informatici in contact komen. Daarnaast hebben we tijdens de development gebruik gemaakt van R-studio een *graphical user interface* (GUI) voor R. R is eenvoudig uitbreidbaar via *Comprehensive R Archive Network* (CRAN) een bibliotheek van R modules naar het voorbeeld van CPAN. Een bekend voorbeeld is het Bioconductor package die in R kan worden gebruikt. Wij gebruikten de versie 2.15.0 van R. [53]

2.6 Methoden

2.6.1 Kwaliteitsbepaling

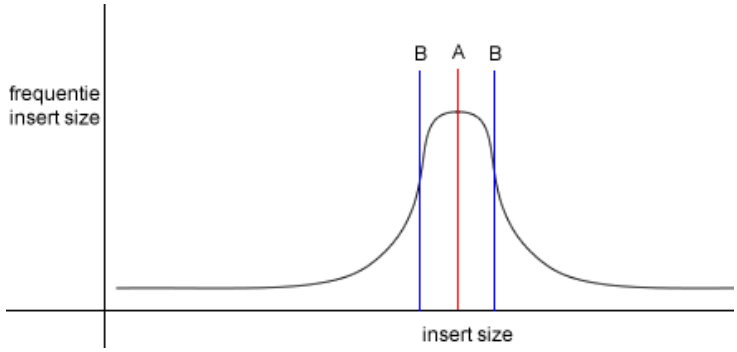
Het is belangrijk om voor een analyse, te controleren of de ontvangen data voldoet aan enkele basis kwaliteitseisen. Er zijn meerdere manieren om dit te doen, wij hebben er voor gekozen om er twee te implementeren. Onze kwaliteitsbepaling is enkel een waarschuwing als onze resultaten erg afwijkend zijn van de verwachtingen.

2.6.1.1 *Pred quality score* controle

Iedere base die bepaald is door de sequencer heeft een score gekregen, deze score noemt men *Phred quality score*. Wij geven een waarschuwing als deze kwaliteitsscore gemiddeld voor een basepositie te laag is, de drempel staat standaard op 30, dit betekent dat er 1 op 1000 kans is dat de base foutief is. De waarschuwing gebeurt in een logbestand. Optioneel kan gekozen worden voor visualisatie in R.

2.6.1.2 *Insert size* variatie controle

Wanneer we de frequentie uitzetten van alle *insert size* waarden dan krijgen verwachten we één grote piek te zien, gelegen rond de verwachte *insert size*. Het resultaat van deze controle is een plot door R geproduceerd die de gebruiker kan valideren, onderstaand een voorbeeld plot van de verwachte *insert size* (zie figuur 2.6).



Figuur 2.1 : distributie plot; A: voorspelde insert size (target insert size), B: van B tot A variaties van de insert size die men mag verwachten, buiten B, links inserties, rechts deleties.

2.6.2 Verdwenen regio's

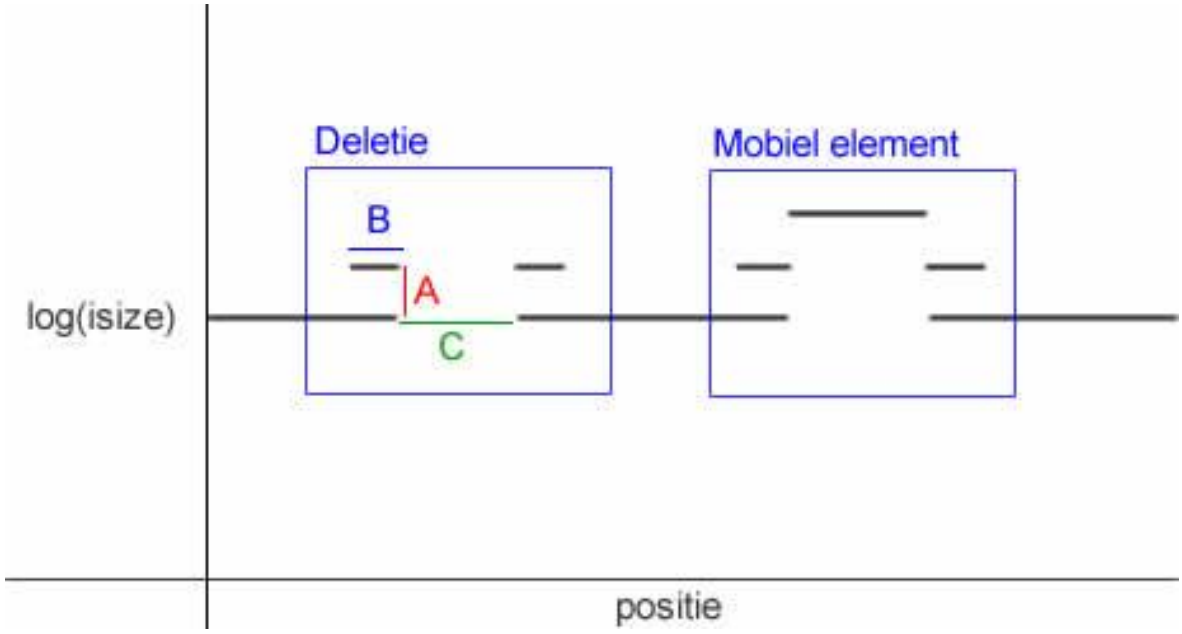
Om verdwenen regio's te ontdekken hebben we vier methodes ontwikkeld en geïmplementeerd in de pipeline.

2.6.2.1 Methode 1

In deze methode zoeken we deleties door de afstand (logaritmisch), tussen de *forward* en *reverse read* op het referentie genoom (*insert size*), uit te zetten per positie. Vervolgens berekenen we de gemiddelde waarde van de *insert size* te bepalen over een klein gebied (window van ±100 bp), om te bepalen of deze boven een bepaalde drempel komt of niet. Wanneer deze regio boven de drempel komt, slaan we het beginpunt op als "vermiste/verdwenen" regio. Daarnaast wordt met behulp van R voor iedere zone (1kb) waar tenminste één punt boven een drempel aanwezig is het gebied visueel weergegeven ter controle (zie figuur 2.1 en figuur 2.2).

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

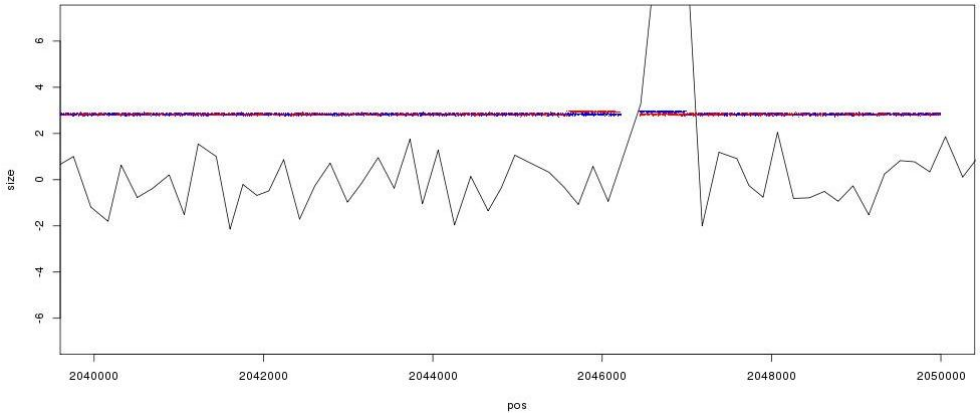
Formule 2.1 : berekening gemiddelde, hier n = 100



Figuur 2.2 : schematische voorstelling van deletie plot, A : de logaritmische grootte van de deletie, B : de *insert size*, C : de grootte van de deletie

2.6.2.2 Methode 2

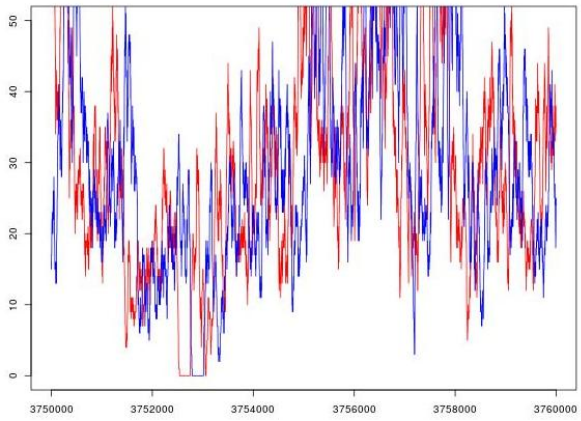
Een andere methode is de variantie van de *insert size* waarden te bepalen over een klein gebied, bij een deletie zullen de waarden hoger liggen. Wanneer de variantie boven een drempel gaat wordt het aanzien als een verwijderde regio. Ook hier worden met behulp van R grafieken gemaakt van de zones die al interessant worden gemerkt.



Figuur 2.3 : een gesimuleerde deletie van 400bp word gedetecteerd door een hoge piek in de variatie.

2.6.2.3 Methode 3

Wanneer we een bepaalde regio verdwenen is op de mutant dan kunnen daar geen reads gemapt worden. Wanneer we dus de coverage van reads over iedere positie zouden bepalen, dan zouden we een dal verwachten op de plaats waar de deletie zich voordoet. Deze methode werk echter alleen als de gedeleteerde regio geen repititief gebied is (bijvoorbeeld insert sequentie of transposon) die ook op andere plaatsten in het genoom voorkomen.



Figuur 2.4 : coverage plot; X-as : positie in het genoom; Y-as : aantal reads die een bepaalde positie overlapt.

2.6.2.4 Methode 4

In deze methode maken we gebruik van de T-test voor steekproeven, namelijk :

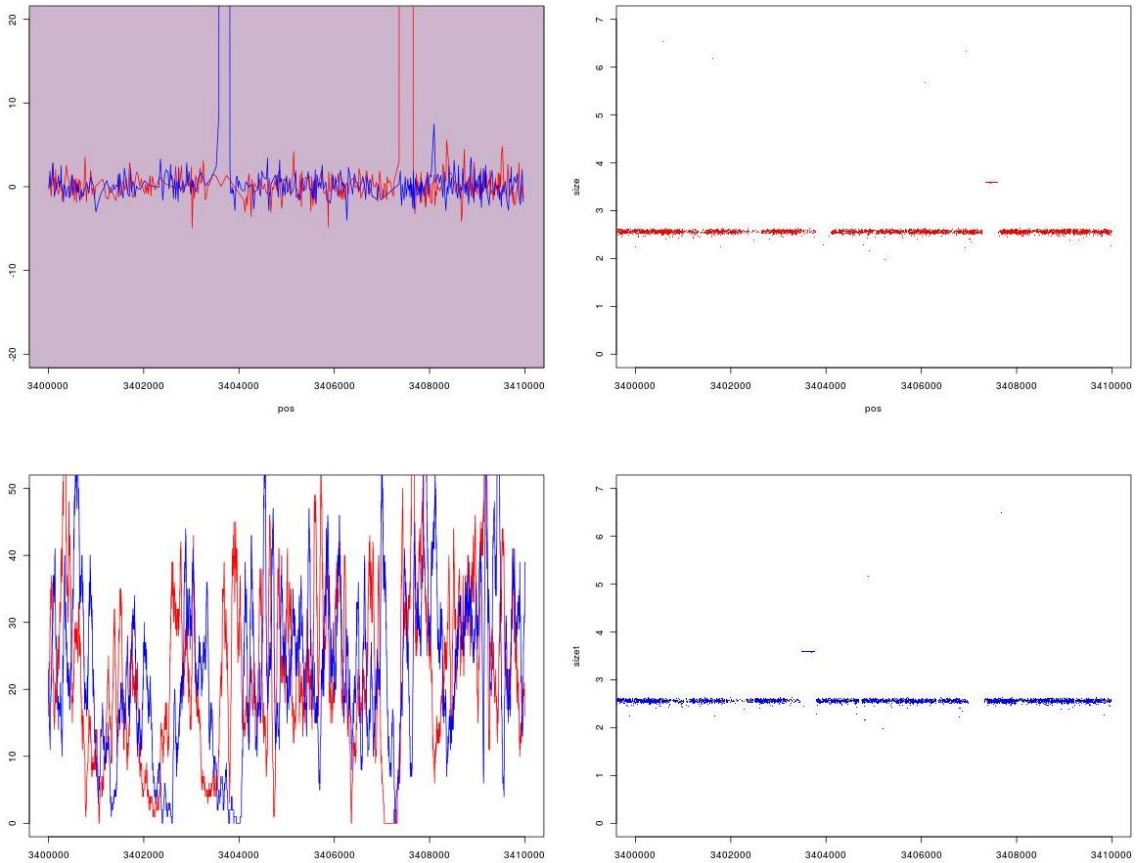
$$T = \frac{X - \mu_0}{S}\sqrt{n}$$

Formule 2.2 : T-test voor steekproef, T = T-toets waarde, X = gemiddelde steekproefwaarde, μ0 = verwachte waarde bij H0, S steekproefstandaard deviatie, n = steekproefgrootte

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

Formule 2.3 : Berekening steekproefstandaard deviatie

Wanneer we H_0 gelijk stellen aan het globale gemiddelde *insert size* kunnen we berekenen of een bepaald gebied merkbaar hoger of lagere waarden heeft. Wanneer we daarna gebruik maken van een lokale piekdetectie en de piek is boven een drempel, dan selecteren we deze.

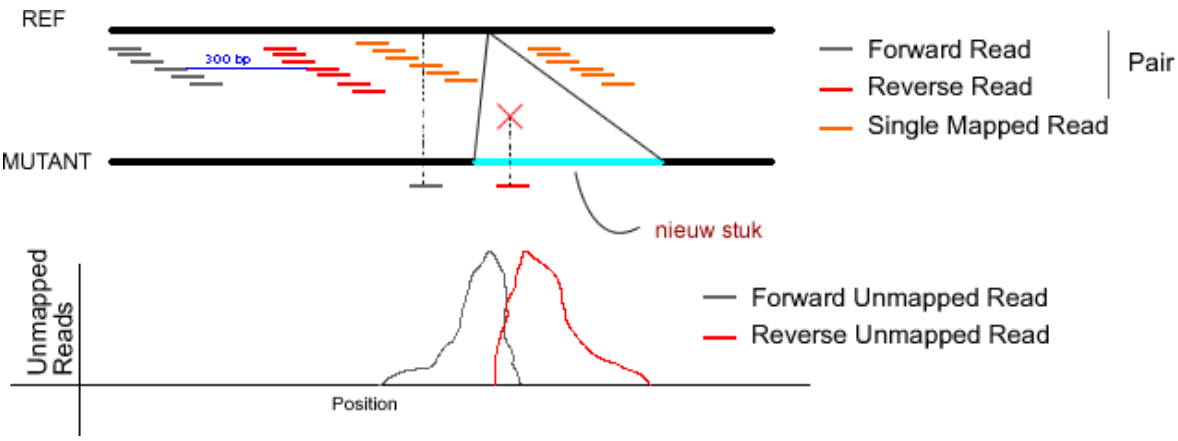


Figuur 2.5 : overzicht resultaat deletie plot; links boven : duidelijke pieken van gemodificeerde T-Test waarden over de gebieden waar het mobiel element zich bevindt (zichtbaar in rechts boven en onder). De coverage (links onder) toont hier de deletie niet aan.

2.6.3 Extra regio's

2.6.3.1 Extra regio's zoeken

Inserties kunnen worden geïdentificeerd met dezelfde informatie zoals deleties, namelijk paired-end data. Wanneer men een eerste read (*forward tag*) kan mappen op een referentie, maar een 2de read (*reverse tag*) niet, dan betekent dit dat de eerste sequentie te vinden is op de referentie, maar de 2de niet. Dit kan veroorzaakt worden doordat de eerste *read* juist is gemapt en dat de 2de in een gebied valt dat niet "beschikbaar" is op het referentie genoom (nieuwe sequentie). Wanneer meerdere *reads* niet gemapt kunnen worden over een gebied, kan dit het gevolg zijn van een insertie. Wij zoeken inserties door de niet gemapte reads verminderd met gemapte reads per positie uit te zetten en dan opzoek te gaan naar pieken van de overgebleven niet gemapte reads. We doen dit via een eigen perl module, dat lokale pieken zoekt en deze wegschrijft naar een file indien ze boven een drempel zitten. Daarnaast wordt de data overlopen in R en wordt iedere piek die voldoet aan bepaalde variabelen getekend en wordt de *forward mapped reads*, *reverse mapped reads*, *forward unmapped reads* en *reverse unmapped reads* getekend.



Figuur 2.6 : extra regio's worden geïdentificeerd door unmapped reads pieken te analyseren.

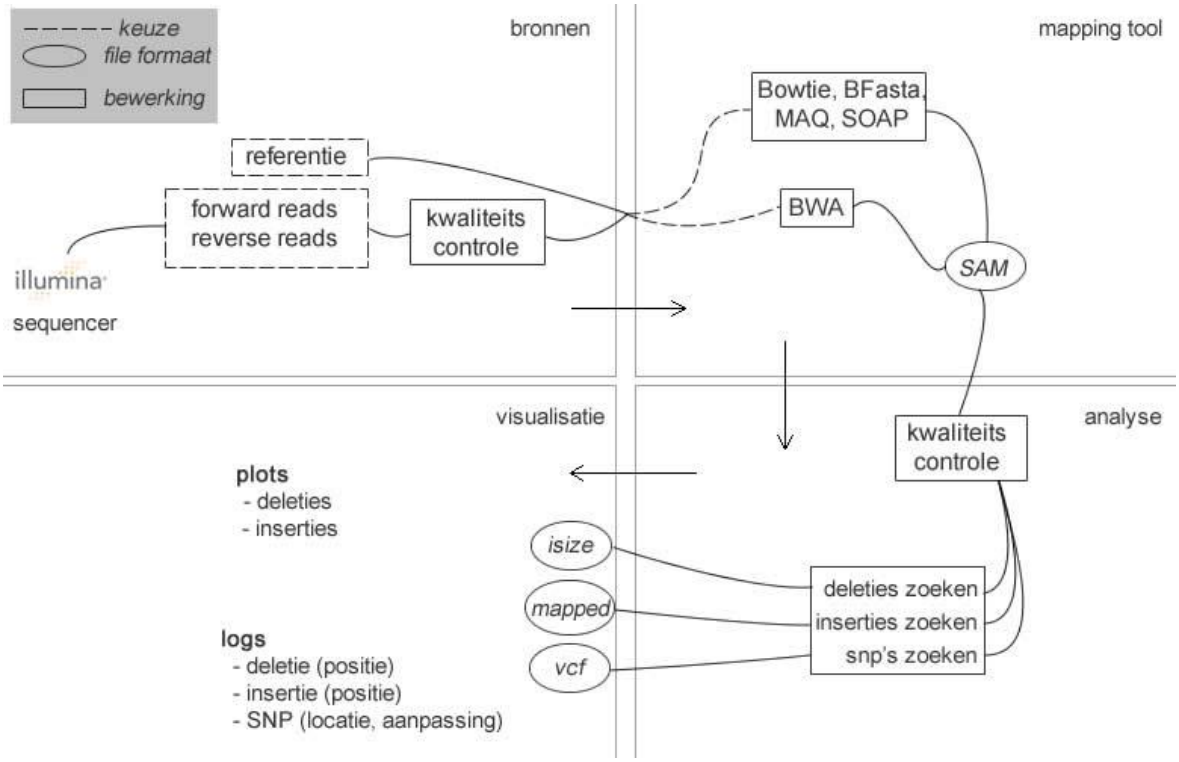
2.6.4 **SNP Calling**

De SNP calling gebeurt met behulp van SAMtools, we gebruiken het mpileup commando. We verwerken de SNP's vervolgens via BCFTools en met het perl script *vcfutils.pl* die deel uitmaakt van SAMtools. Deze methode is een opgesteld protocol van *Tobias Rauch*, *Variant Calling from Genomic Sequencing Data*. (Juli, 2011) [54].

Het oude commando *pileup* wordt ook ondersteund. We filteren de resultaten met een perl script. Dit doen we door te berekenen hoeveel keer de SNP ondersteunt wordt en hoeveel keer de referentie wordt ondersteund, wanneer de SNP's 95% ondersteund wordt door de reads en er meer dan vijf reads de positie mappen, rapporteren we de SNP als geldig.

3 Resultaten

3.1 Werking pipeline



Figuur 3.1 : overzicht werking van pipeline

De pipeline maakt gebruik van R, perl, bash scripts en verschillende tools. Om onderhoud en updates eenvoudiger te maken hebben we de code zo gemaakt dat ze modulair werkt. Dit betekent concreet dat bepaalde stukken onafhankelijk van elkaar kunnen werken. Daardoor kunnen opties in of uitgeschakeld naar gelang de wens. Dit werkt met optionele parameters die kunnen ingesteld worden in de console waarvan, het gebruik steunt op de officiële modules GetOpts.

3.1.1 Voorbereiding

De eerste stap in het proces is het voorbereiden van de data, als input verwachten we een bestaande referentie, in multifasta of gecompreseerd (.gz) en twee fastq files met de reads, namelijk *forward* en *reverse reads*, ook deze mogen gecomprimeerd zijn. Alternatief kan de mapping stap overgeslagen worden en kan een

sequence alignment/Map (SAM [55]) bestand worden gebruikt. Tijdens deze stap kan men optioneel de kwaliteitscontroles uitvoeren.

3.1.2 **Mapping**

De tweede stap is het mappen van de reads, dit gebeurt in de module BWA. We kozen voor deze naam om duidelijk te maken dat we gebruik maken van het programma Burrows-Wheeler Aligner (BWA). In deze module wordt ook gecontroleerd of het programma kan worden opgeroepen en eventueel wordt rekening gehouden met de optie "-- bwa *location*" om de locatie aan te geven. Dit is handig wanneer de pipeline werkt op een machine waar BWA niet op de standaard locatie is of kan geïnstalleerd worden, zoals bij een *shared* server. BWA berekent zelf de maximale waarde die de *insert size* mag bedragen om een *read pair* de vlag *proper mapped* te geven. Wanneer evenwel de *insert size* variatie niet normaal verdeeld is zoals bij S1/S1H sequencing, door bijvoorbeeld slechte *library preparation*, kan deze benadering foutief zijn. In dit geval is een constante waarde een betere oplossing, dan de standaard methode die BWA gebruikt. Hiervoor is een parameter voorzien in de pipeline (-- insert_size *verwachte_insert_size*).

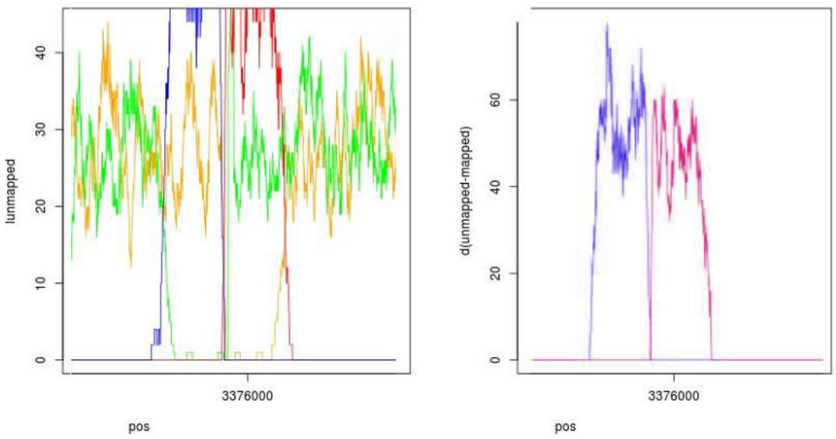
3.1.3 **Variaties onderzoeken**

In deze stap gaan we opzoek naar de drie variaties, namelijk *single nucleotide polymorphism* (SNP's) en *Indels*, nieuwe regio's en verdwenen regio's. De SNP's en *Indels* worden gezocht met behulp van de mpileup methode, hoewel de -niet meer ondersteunde- pileup methode nog steeds in de code aanwezig is en werkt, indien aangeroepen. Daarna splitsen we het SAM bestand in de gemapte referentienamen die voorkomen in het SAM bestand. De originele SAM file word dus gesorteerd in *ssam files* (splitted SAM). Daarnaast krijgen we een bestand "bSAM" (bad SAM) die niet gemapte reads bewaard. Dit kunnen moeilijk te mappen regio's zijn, maar dit kunnen ook inserties zijn of plasmiden die niet in de referentie beschikbaar zijn. De volgende stap is de verwerking van de ssam files, deze stap word uitgevoerd door de PipeLine module. Iedere ssam file wordt doorlopen. Iedere lijn wordt opgesplitst in variabelen. Ook uit de *bitwise flag* wordt belangrijke informatie.

De opdeling in ssam files gebeurt met behulp van de *rname* variabele. In deze stap gebruikt de pipeline, de bitwise flag, pos en size. Deze variabelen zijn nodig om de *forward proper mapped reads*, *reverse proper mapped reads, forward unmapped reads, reverse unmapped reads* en de *size* per *pos* te berekenen. Deze variabelen worden gebruikt om inserties mee te identificeren, de resultaten van de berekeningen worden opgeslagen in *mapped* bestanden. De *isize* per *pos* word gebruikt om deleties mee te bepalen, deze worden opgeslagen in *isize* bestanden.

De volgende stap is het sorteren van de *isize* bestanden, zodat de positie oplopend is. Deze worden oplopend numeriek gesorteerd met behulp van het unix *sort* commando, opgeslagen met *.sort* extensie. Hierdoor kan R en perl de resultaten eenvoudiger selecteren.

De volgende stap is de zoektocht naar inserties, in de vorige stap hebben we de benodigde variabelen reeds berekend. Door de *(forward + reverse) unmapped reads* te verminderen met *mapped reads* bekomen we pieken die een extra regio aanduiden ten opzichte van de referentie. Wanneer de pieken dicht bij elkaar voorkomen duidt dit meestal op een insertie, wanneer ze verder van elkaar voorkomen, duidt dit meestal op een deletie.



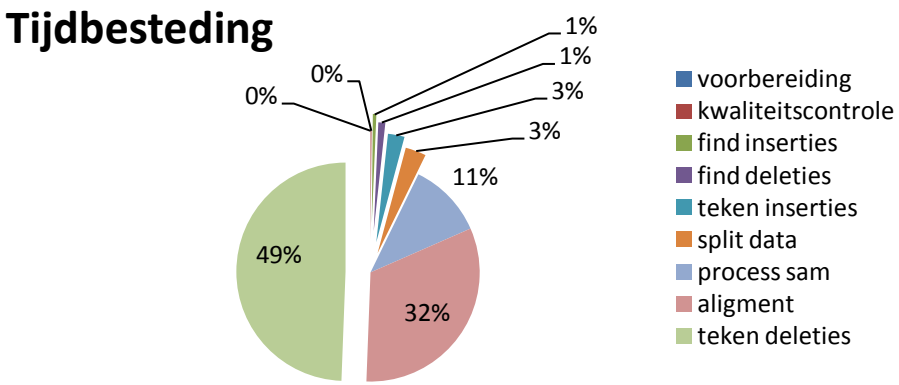
Figuur 3.2 : links, op grafiek uitgezet forward unmapped reads (blauw), forward mapped reads (groen), reverse unmapped reads (rood) en reversed mapped reads (oranje). Rechts, op de grafiek uitgezet forward unmapped reads – forward mapped reads (blauw) en reverse unmapped reads – reverse mapped reads (rood)

3.2 Code Pipeline

De pipeline is opgebouwd uit 12 perl modules die samen ongeveer 2000 lijnen code vormen. De modules zorgen ervoor dat de pipeline snel kan worden aangepast aan alternatieve verwachtingen. Eveneens kunnen kleine wijzigingen hierdoor snel doorgevoerd worden en kunnen fouten eenvoudiger opgespoord worden. Het zou bijvoorbeeld mogelijk zijn om geen gebruik te maken van R om grafieken te visualiseren, maar een alternatieve perl module. De pipeline zelf vraagt behalve R en Perl geen eisen, BWA is nodig indien geen SAM formaat wordt gegeven en SAMTOOLS indien SNP/Indel calling gevraagd wordt.

3.3 Tijdsbesteding pipeline

We hebben een simulatie run gedaan met drie deleties en drie inserties in de megaplasmide pMOL30 en vervolgens de pipeline de deleties en inserties laten identificeren en iedere stap in de pipeline een tijdsmeting gedaan.



Figuur 3.3 : Tijdsbesteding tijdens een test run, exacte waarden zie bijlage 6.

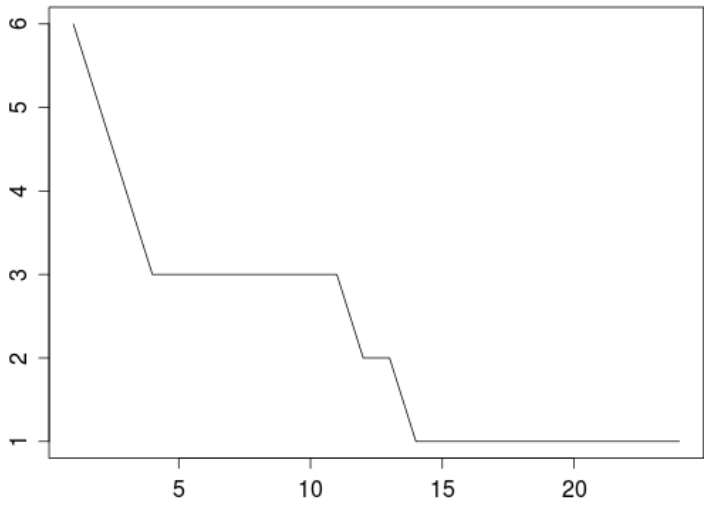
3.4 Test cases

3.4.1 Simulaties

Om de betrouwbaarheid van voorspellingen na te gaan met oog op variatie in *insert size* hebben we een simulatie gelopen met gesimuleerde *genome rearrangements* (zie tabel 3.3) in *Cupriavidus metallidurans* (CH34). De variatie op de *insert size* werd iedere run met 10 verhoogd, tot een maximum van 650.

Tabel 3.1 : gesimuleerde variaties

Deleties	Inserties
- 70bp	+ 70 bp
- 210bp	+ 210 bp
- 700bp	+ 700 bp



Figuur 3.4 : uitgezette resultaten met identificaties voor de insertie detectie methode van variaties. X as is variatie/10 en Y as zijn de aantal getekende variaties. (deze grafiek is gelimiteerd tot variatie 250, de waarde één blijft evenwel aangehouden tot variatie 650.)

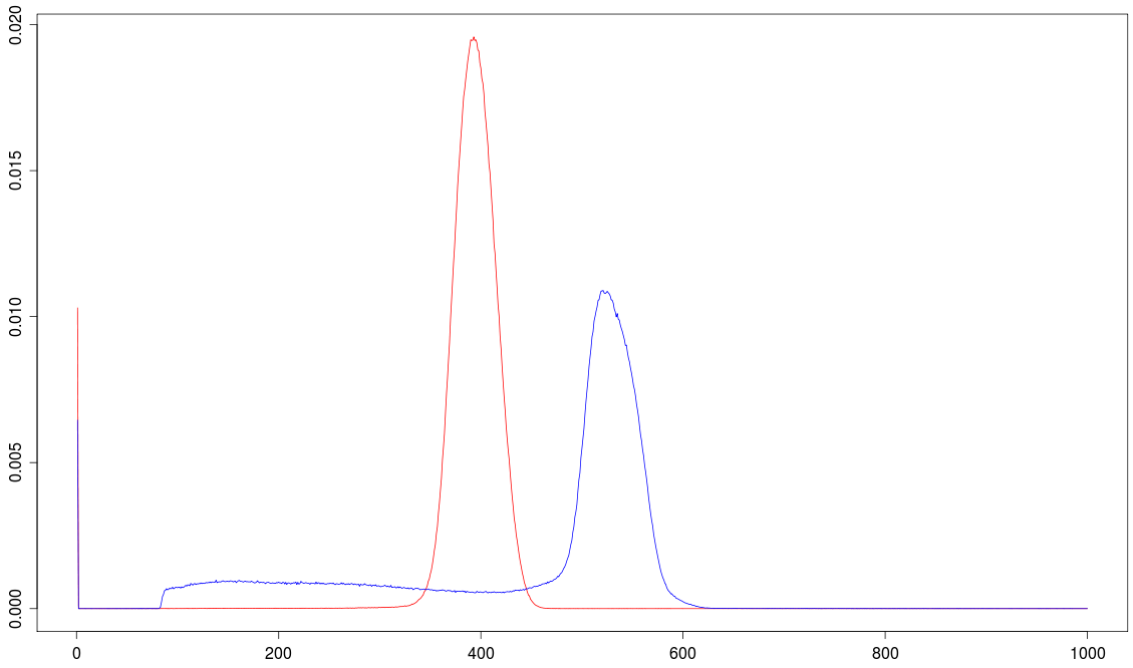
3.4.2 NASAIV

De kwaliteitscontrole werd uitgevoerd met de standaard opties en er werd een kwaliteit per base plot geproduceerd door R (zie bijlage 3). De resultaten van de pipeline zijn toegevoegd als tabel (zie bijlage 2), na manuele controle. Heel wat contigs hadden geen enkele gemapte positie en deze werden uit de pipeline gehaald. Deze contigs zijn vermoedelijk artefacten van de assembly op basis van de 454 sequencing resultaten.

3.4.3 *Cupriavidus metallidurans*

De kwaliteitscontroles werden uitgevoerd en gevisualiseerd. De eerste kwaliteitscontrole, de kwaliteit per positie. Was voor beide stammen (AE2720, AE2722) positief. De offset was 66, de kwaliteitsbepaling door Illumina was een oude methode gebruikt op de GAII sequencers (zie bijlage 4). Daarnaast is ook de *insert size* distributie getekend in R, voor zowel mutant AE2720 als AE2722 was de distributie gelijkaardig (enkel AE2720 is opgenomen, zie figuur 3.5).

Tabel 3.2 : variaties in AE2720 mutant



Figuur 3.5 : *insert size* frequentie plot, in rood het frequentie plot van *Cupriavidus metallidurans* (AE2720), in blauw het frequentieplot van *Rhodospirillum rubrum (S1).*

3.4.4 Rhodospirillum rubrum

Ook voor de *Rhodospirillum rubrum* S1 en S1H stammen werden kwaliteitsscores bepaald (zie bijlage 4). Ook de distributie van de *insert size* waarden werden gemaakt (zie figuur 3.5). Zowel deleties als insertie voorspellingen vormden een probleem bij deze data. Toch is op basis van deze preliminaire sequencing resultaten een poging gewaagd om deze variaties (inserties, zie tabel 3.3) te voorspellen gebruik makend van twee referenties, namelijk de S1 stam (NCBI) en de F11 stam.

Tabel 3.3 : overzicht inserties; algemene referentie fout (grijs), verschil tussen S1 en F11 (oranje), betrouwbare insertie voorspellingen (groen), onbekende voorspellingen (rood)

Ref <-> S1	Ref <-> S1H	F11 <-> S1	F11 <-> S1H
2.444.000 bp	414.000 bp	424.000 bp	414.000 bp
2.646.000 bp	2.444.000 bp	2.444.000 bp	424.000 bp
	2.810.000 bp	2.646.000 bp	1.368.000 bp
	2.946.000 bp	2.984.000 bp	2.444.000 bp
			2.810.000 bp
			2.946.000 bp
			2.984.000 bp

4 Discussie

4.1 Mapping software

Wanneer vijf miljoen reads moeten gemapped worden op een referentie is dit in geen geval een triviaal probleem. Daarom zijn er voor de mapping tools verschillende implementaties beschikbaar elk met zijn voor- en nadelen. Er is gekozen om BWA als standaard tool te implementeren maar als alternatief eveneens support te bieden voor het SAM formaat. Dit SAM formaat is een opkomend uniform systeem om mapped reads te beschrijven. Alternatieven zoals bowtie, Bfast, novoalign en SHRiMP bieden allen de mogelijkheid om in SAM formaat hun resultaten te bewaren.

4.2 Methode ontwikkeling

Verdwenen regio's werden initieel bepaald door de gemiddelde *insert size* te bepalen over een regio. Bij een normale deletie verwachten we in de *insert size* plots een sprong naar boven, waarvan de hoogte van de sprong overeenkomt met de verdwenen regio. Indien de verdwenen regio echter een mobiel element is, dan is een duidelijke band te zien van *reads* die wel gemapt kunnen worden. Wanneer evenwel een grotere variantie van de waarden aanwezig is, kunnen niet-mobiele elementen verloren gaan. Een tweede benadering was de variatie te gaan bepalen van *insert size* op een kleine regio, en op deze waarden piekdetectie uit te voeren. Dit bleek het probleem met de niet mobiele elementen op te lossen, doordat aan begin en het einde van deleties, een hoeveelheid reads voor en na de deletie mappen (een eigenschap van paired end). De variatie bleek echter niet om te kunnen met een grote *insert size.* De implementatie van een steekproef t-test loste dit probleem deels op na wat modificatie werkt deze methode naast de coverage methode. De coverage methode berekend voor iedere positie hoeveel reads de positie overlappen. Indien dit voor een groot aantal een laag getal is, (instelbaar, standaard 400 posities x 0 coverage) identificeren we het gebied als deletie. Beide methoden vormen samen een basis voor zowel detectie van normale deleties als deleties met een mobiel element. Verdere verfijning is evenwel nog nodig.

Modificaties aan de “t-test” :

- Het globale gemiddelde wordt berekend enkel met waarden die tussen bepaalde grenzen vallen. Deze grenzen zijn instelbaar en zijn sterk afhankelijk van de variantie van de *insert size*.
- Om de t-test waarde te bepalen worden waarden die hoger zijn dan een bepaalde waarde veranderd naar het globale gemiddelde. De waarde wordt niet genegeerd om het programma niet extra complex te maken.

Om extra regio’s te bepalen, gebruiken we de verhouding aan mapped en unmapped reads in een bepaalde regio. De insertie heeft een typische afdruk en is relatief goed te herkennen, namelijk een piek van unmapped forward reads vlak naast een piek van unmapped reverse reads. Het is mogelijk om zowel grote deleties als inserties via deze methode te vinden, deze methode is dan ook een aanvullende informatiebron voor deletiedetectie. Toch is het belangrijk te bemerken dat ook hier nog kan aan verbeterd worden. Voornamelijk inserties die kleiner zijn dan de *insert* size zijn heel moeilijk te bepalen. Zo blijkt ook uit onze insertie benchmark. Grote inserties kunnen evenwel blijven gevonden worden tot variaties van 650. (zie figuur 3.5) Een verdere optimalisatie die hier nog vereist is, is de identificatie van de ingebouwde regio. Indien het om een mobiel genetisch element gaat, kan dit eventueel gebeuren door *pattern match* van de *unmapped reads* met het volledige genoom. Indien het om een unieke regio gaat kan men eventueel terugvallen op assembly algoritmen.

4.3 Tijdsbesteding

De tijd die de pipeline nodig heeft om volledig te lopen kan heel variabel zijn. Deze hangt af van drie grote factoren, namelijk:

1. De hoeveelheid data (grootte van genoom die gesequeneerd is, de coverage, …)?
2. De variantie van de *insert size*, tijdens onze testen bleek de tijd die BWA erover deed om de reads te mappen sterk afhankelijk van de variatie in *insert size*, tot wel 6 keer de volledige looptijd.

3. De ingestelde parameters. Het tekenen van de deleties is het zwaarste proces, deze afbeeldingen zijn ook de meest complexe, daarbij komt dat *subset* een heel traag commando is binnen R. Het is duidelijk dat de optimalisatie stap nog niet is gedaan, maar we de nadruk duidelijk op resultaten gelegd hebben en niet op snelheid.

4.4 Optimalisatie

De pipeline is getest met verschillende artificiële en niet-artificiële sequeneringsdata, maar de lengte van het genoom was nooit groter dan bacterieel genoom. Wanneer we dit vergelijken met humane genomen zouden enkele grote optimalisatie stappen nodig zijn om een zelfde pipeline te gebruiken. De pipeline maakt momenteel geen gebruikt van meerdere processoren, dit betekend dat perl maar 1 CPU 100% kan gebruiken, bij de huidige generatie computers is 2 tot 4 CPU's geen uitzondering. Bij een server kan dat al tot 8 CPU's gaan, wanneer servers in clusters werken ('cloud'), zijn de hoeveelheid CPU's virtueel oneindig. Het zou dus een grote stap voorwaarts zijn om meerdere CPU's in parallel te gebruiken, door de code multicore te maken.

4.5 Zilverresistentie in *C. metallidurans* CH34

De analyse van de mutanten AE2720 en AE2722 bevestigen de resultaten die reeds werden bekomen door de zoektocht naar variaties door per *window* manueel te bekijken. Daarbij moet opgemerkt worden dat enkele resultaten aan de lijst werden toegevoegd namelijk deleties die kleiner dan 1000bp zijn (zie bijlage 1).

4.6 Zilverresistentie in *C. metallidurans* NASAIV

Heel veel *contigs* hadden geen gemapte reads. We vermoeden dat het tool die contigs voorspelt, Newbler (dit is een algoritme dat door Roche zelf ontwikkeld werd voor assembly op basis van 454 sequencing) goed werkt en we vermoeden dat het *artefacts* van de sequencing zijn die de fouten veroorzaken. Er werden 14 deleties geïdentificeerd en 11 inserties, in de overige contigs.

4.7 Rhodospirillum rubrum

De analyse van de Illumina sequencing van *R. rubrum* stam S1 en S1H resulteerden in een opmerkelijk hoog aantal reads dat niet correct gemapped kon worden. Hierdoor voorspelt onze analyse pipeline een groot aantal inserties zichtbaar en zien we op de plot van de *insert sizes* versus de positie in het genoom een grotere variatie op de inset size dan bij andere sequencing projecten. De frequentie distributie van de *insert size* waarden werd tijdens onze kwaliteitscontrole bepaald en deze bleek niet normaal verdeeld te zijn. Er was een duidelijke sloop zichtbaar op het frequentieplot (zie figuur 3.5). Om dit probleem op te lossen zullen de data opnieuw gesequeneerd worden. Door aanpassingen aan de pipeline en parameters anders in te stellen konden we toch preliminaire resultaten bekomen. Om onze voorspellingen beter te staven hebben we de stam F11 gebruikt om een vergelijkend onderzoek te doen. Uit dit onderzoek kwam dat de meeste voorspellingen correct waren ook zijn enkele fouten in de referentie sequentie van NCBI hiermee mogelijks geïdentificeerd. Dit zijn echter nog steeds preliminaire resultaten, die pas kunnen geverifieerd worden nadat een nieuwe *Illumina* paired-end run gelopen is op het genomisch DNA van stam S1 en S1H.

5 Besluit

We hebben gedurende ongeveer drie maanden gebouwd aan een pipeline die sequencing data afkomstig van *Illumina* sequencing machines automatisch kan analyseren via mapping ten opzichte van een evolutionair gerelateerd referentie genoom. Onze pipeline kan deleties, inserties, *Single Nucleotide Polymorphism* en kleine indels detecteren. Wel moet opgemerkt worden dat er geen volledige validatie is gebeurd van de pipeline. Zo is nog geen test run gedaan om de nauwkeurigheid van de “verdwenen regio methode” te gaan valideren. De validatie moet zeker gebeuren voor het gebruik van deze pipeline op een routinematige manier. Ook zijn er enkele kleine bugs onopgelost, deze vormen evenwel geen gevaar voor de resultaten.

Door gebruik van te maken van modulair programmeren in perl, kan de pipeline eenvoudig worden uitgebreid met nieuwe modules. Door de keuze voor perl als scripting language kan ook iedereen de werking tot in detail bekijken en wijzigen indien gewenst. Daarnaast werkt perl op vrijwel elk modern platform, waardoor geen speciale eisen gesteld worden aan de werkomgeving. *Single Nucleotide Polymorphisms* en kleine indels worden gedetecteerd met behulp van SAMtools, voor deleties en inserties zijn behalve perl en R geen extra vereisten nodig.

We hebben de pipeline getest en geoptimaliseerd op real-life data afkomstig uit verschillende, *C. metallidurans* stammen die een verhoogde zilverresistentie vertonen (AE2720, AE2722, NASAIV). Daarnaast hebben we gesimuleerde data gebruikt om de uiterste limieten van onze data analyse pipeline te bepalen. Als proof-concept hebben we onze methodologie ook toegepst op sequencing data afkomstig van *Rhodospirullum rubrum S1* en *S1H* . Hoewel de kwaliteit van de *insert size* niet voldoende was om betrouwbare resultaten te produceren, hebben we toch een preliminaire resultaten voor de deleties en inserties kunnen verkrijgen, onder meer dankzij een vergelijking met een bijkomende referentie, namelijk de *Rhodospirullum rubrum F11* stam. De *Single Nucleoide Polymorphismen* werden

bepaald door de pipeline en toonden hoge overeenkomst met de detectie, geleverd door *Beijing Genomic Institute*.

We kunnen stellen dat we een tool hebben ontwikkeld, die indien het geoptimaliseerd wordt. Een volledige resequencing analyse van enkele dagen kan reduceren tot enkele uren. Daarbij dient de onderzoeker weinig technische achtergrond te hebben om de resultaten te verifiëren en kan deze terug vallen op de handleiding die voorhanden is.

6 Literatuurlijst

[1] wiki ATCC. (2012, maart) wikipedia. [Online]. http://en.wikipedia.org/wiki/American_Type_Culture_Collection

[2] atcc. (2012, maart) lgcsstandards-atcc. [Online]. http://www.lgcstandards-atcc.org/

[3] wikipedia proteobacterien. (2012, maart) wikipedia. [Online]. http://nl.wikipedia.org/wiki/Proteobacteri%C3%ABn

[4] (2012, Maart) SCKCEN. [Online]. http://www.sckcen.be/nl/Ons-Onderzoek/Research-projects/ESA-projects/MELGEN-2

[5] Esa. (2012, Mei) ESA. [Online]. http://ecls.esa.int/ecls/?p=melissa

[6] wiki. (2012, maart) wikipedia early reactor. [Online]. http://en.wikipedia.org/wiki/Nuclear_reactor#Early_reactors

[7] wiki. (2012, maart) wiki nuclear power plant. [Online]. http://en.wikipedia.org/wiki/Nuclear_power_plant

[8] SCKCEN. (2012, maart) SCKCEN. [Online]. http://www.sckcen.be/nl/Over-SCK-CEN

[9] SCKCEN. (2012, maart) SCK•CEN bedrijfssynopsis. pdf document.

[10] wiki. (2012, maart) wiki nucleaire geneeskunde. [Online]. http://nl.wikipedia.org/wiki/Nucleaire_geneeskunde

[11] SCKCEN. (2012, maart) Instituut voor Nucleaire Materiaalwetenschappen. pdf document.

[12] SCKCEN. (2012, maart) Instituut voor Geavanceerde Nucleaire Systemen. pdf document.

[13] SCKCEN. (2012, maart) Instituut voor Mileu, Gezondheid en Veiligheid. pdf document.

[14] SCKCEN. (2012, Mei) SCKCEN MELiSSA. [Online]. http://www.sckcen.be/en/Our-Research/Scientific-Institutes-Expert-Groups/Environment-Health-and-Safety/Molecular-and-Cellular-Biology/Microbiology/Life-support-system-for-space-flights-based-on-bacterial-waste-conversion

[15] "Microbial ecology of the closed artificial ecosystem MELiSSA, Reinventing and compartmentalizing the Earth's food and oxygen regeneration system for long-haul space exploration missions," *Research in Microbiology*, vol. 157, pp. 77-86, April 2006.

[16] Leys Natalie @ SCK. (2012, maart) SCKCEN. [Online]. http://www.sckcen.be/nl/Ons-Onderzoek/Research-projects/ESA-projects/MELGEN-2

[17] DOE Joint Genome Institute. (2012, maart) JGI. [Online]. http://genome.jgi-psf.org/rhoru/rhoru.home.html

[18] Munk et al., "Complete genome sequence of Rhodospirillum rubrum type strain (S1T)," *Genomic Standard Consortium*, vol. 4, no. 3, pp. 128-132, Juni 2011.

[19] (2012, maart) wikipedia JGI. [Online]. http://en.wikipedia.org/wiki/Joint_Genome_Institute

[20] SCKCEN. (2012, maart) SCKCEN. [Online]. http://www.sckcen.be/en/Our-Research/Scientific-Institutes-Expert-Groups/Environment-Health-and-Safety/Molecular-and-Cellular-Biology/Microbiology/Life-support-system-for-space-flights-based-on-bacterial-waste-conversion

[21] straininfo. (2012, maart) straininfo. [Online]. http://www.straininfo.net/taxa/1480;jsessionid=363564D2AB173DC51590AF7F888CD690

[22] lgcstandards-atcc. (2012, maart) lgcstandards-atcc. [Online]. http://www.lgcstandards-atcc.org/

[23] et al. Lonjers ZT, "Identification of a new gene required for the biosynthesis of rhodoquinone in Rhodospirillum rubrum.," *J Bacteriol*, vol. 194, no. 965, p. 71, Juli 2012.

[24] genomes online. (2012, mei) genomesonline. [Online]. http://genomesonline.org/cgi-bin/GOLD/bin/GOLDCards.cgi?goldstamp=Gc01977&collapse=true

[25] SCKCEN. (2012, maart) sckcen. [Online]. http://publications.sckcen.be/dspace/handle/10038/7061

[26] Pieter Monsieurs, Max Mergeay, Natalie Leys, Jacque Mahillon, Rob Van Houdt Kristel

Mijnendonckx. (2012) Silver resistance in Cupriavidus metallidurans CH34 is affected by endogenous insertion sequence elemnts and cross regulation. poster.

[27] Mijnendonckx K., Leys N., Van Houdt R Monsieurs P. (2012) Transcriptional cross-regulation as survival mechanism in bacteria. poster.

[28] R. CONTRERAS, F. DUERINCK, G. HAEGEMAN, D. ISERENTANT, J. MERREGAERT, W. MIN JOU, F. MOLEMANS, A. RAEYMAEKERS, A. VAN DEN BERGHE, G. VOLCKAERT & M. YSEBAERT W. FIERS, "Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene," *Nature*, no. 260, pp. 500-507, April 1976.

[29] Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes CA, Hutchison CA, Slocombe PM, Smith M Sanger F, "Nucleotide sequence of bacteriophage phi X174 DNA," *Nature*, no. 265, pp. 687-695, Februari 1977.

[30] "Whole-genome random sequencing and assembly of Haemophilus influenzae Rd," *Science*, vol. 269, no. 5223, pp. 496-512, Juli 1995.

[31] Mike Gilchrist. medical institute for medical research. [Online]. http://www.nimr.mrc.ac.uk/mill-hill-essays/bringing-it-all-back-home-next-generation-sequencing-technology-and-you

[32] wikipedia. [Online]. http://en.wikipedia.org/wiki/Sequence_assembly

[33] fishersequencing. [Online]. http://www.fishersequencing.com/primerwalking.htm

[34] wikipedia. [Online]. http://en.wikipedia.org/wiki/Primer_walking

[35] umich.edu. [Online]. http://seqcore.brcf.med.umich.edu/doc/dnaseq/primerwalking.html

[36] wikipedia. [Online]. http://en.wikipedia.org/wiki/Shotgun_sequencing

[37] Waterman MS Lander E, "Genomic Mapping by Fingerprinting Random Clones : A Mathematical Analysis," *Genomics*, no. 2, pp. 231-239, januari 1988.

[38] Wetterstrand KA. genomes.gov. [Online]. http://www.genome.gov/sequencingcosts/

[39] NCBI. [Online]. http://www.ncbi.nlm.nih.gov/projects/genome/probe/doc/TechResequencing.shtml

[40] (2012, mei) from millions to one : theoretical and concrete approaches to denovo assembly using short read DNA sequences. digitaal boek.

[41] illumina. (2012, maart) illumina paired end sequencing assay. [Online]. http://www.illumina.com/technology/paired_end_sequencing_assay.ilmn

[42] wikipedia. (2012, mei) wikipedia, fastQ. [Online]. http://en.wikipedia.org/wiki/FASTQ_format

[43] SAM format. (2012, Februari) SAM Format Specifiactions. pdf. [Online]. samtools.sourceforge.net/SAM1.pdf

[44] Nils Homer. (2012, maart) Whole Genome Simulation. [Online]. http://sourceforge.net/apps/mediawiki/dnaa/index.php?title=Whole_Genome_Simulation

[45] (2012, maart) BWA. [Online]. http://bio-bwa.sourceforge.net/bwa.shtml

[46] (2012, maart) seqanswers. [Online]. http://seqanswers.com/wiki/BWA

[47] (2012, maart) Bowtie. [Online]. http://bowtie-bio.sourceforge.net

[48] samtools. (2012, maart) samtools. [Online]. http://samtools.sourceforge.net

[49] (2012, maart) perl. [Online]. perl.org

[50] (2012, maart) bioperl. [Online]. http://www.bioperl.org/wiki/Main_Page

[51] (2012, maart) wikipedia shell. [Online]. http://nl.wikipedia.org/wiki/Shell_(informatica)

[52] (2012, maart) wikipedia Perl. [Online]. http://en.wikipedia.org/wiki/Perl

[53] (2012, maart) wikipedia R. [Online]. http://www.r-project.org/

[54] Tobias Rauch. (2011, Juli) Variant Calling from Genomic Sequencing Data. [Online]. http://www.embl.de/~rausch/

[55] Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan Heng Li, "The Sequence Alignment/Map (SAM) Format and SAMtools," *Bioinformatics advance access*, pp. 1-2, Juni 2009.

[56] Karamohamed S, Pettersson B, Uhlén M, Nyrén P. Ronaghi M, "Real-time DNA sequencing using detection of pyrophosphate release.," *Anal Biochemie*, vol. 242, no. 9, p. 84, November 1996.

[57] (2012) wikipedia. [Online]. http://en.wikipedia.org/wiki/Sequence_assembly

[58] (2012, maart) Velvet. [Online]. http://www.ebi.ac.uk/~zerbino/velvet/

[59] "Assembling millions of short DNA sequences using SSAKE," *bioinformatics*, vol. 23, no. 4, pp. 500–501, december 2007.

[60] Christiaan V. Henkel, Hans J. Jansen, Derek Butler Marten Boetzer, "Scaffolding pre-assembled contigs using SSPACE," *bioinformatics*, vol. 27, no. 4, pp. 578–579, December 2010.

[61] (2012, maart) SSPACE basic. [Online]. http://www.baseclear.com/landingpages/sspacev12/

[62] (2012, maart) seqanswers. [Online]. http://seqanswers.com/wiki/SSPACE

7 Bijlagen

7.1 Bijlage 1 : resultaten AE2720 en AE2722

Tabel 7.1 : resultaten van pipeline, AE2720 en AE2722

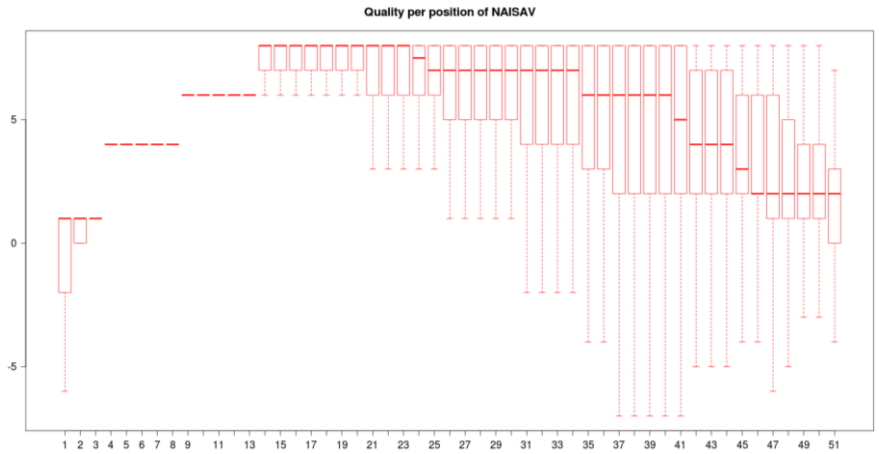
AE2720		**AE2722**	
Deleties	**Inserties**	**Deleties**	**Inserties**
NC_007973			
180.000		180.000	
260.000			5.260.000
300.000		560.000	
560.000		1.000.000	
730.000		1.320.000	
1.000.000		1.680.000	
1.140.000			1.910.000
1.320.000		2.120.000	
1.410.000		2.200.000	
1.460.000		3.410.000	3.410.000
1.680.000		3.760.000	
1.910.000	1.910.000		
2.120.000			
2.200.000			
	2.380.000		
2.540.000			
2.690.000			
	3.260.000		
3.360.000			
3.410.000	3.410.000		
3.760.000			
3.820.000	3.820.000		
NC_007974			
50.000		150.000	
200.000		200.000	
230.000		550.000	
540.000			
550.000			
1.700.000			
2.430.000			
pMOL30			
190.000		190.000	
pMOL28			
10.000		10.000	

7.2 Bijlage 2 : resultaten NASAIV

Tabel 7.2 : resultaten pipeline van het NASAIV project

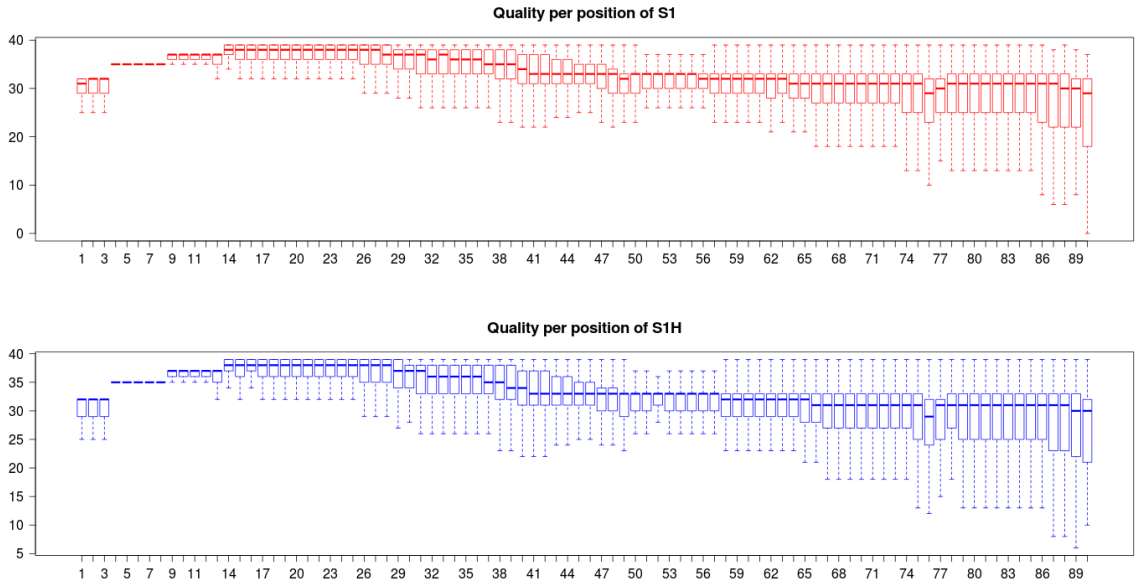
	Deleties	inserties
contig0004	90.000	-
contig0005	320.000	-
contig0008	10.000	10.000
contig0009	10.000	10.000
contig0010	-	50.000
contig0014	10.000	10.000
contig0017	10.000	10.000
contig0019	-	10.000
contig0021	10.000	-
contig0024	10.000	-
contig0029	40.000	-
contig0040	-	50.000
contig0041	180.000	-
contig0042	10.000	-
contig0048	550.000	620.000
contig0050	-	10.000
contig0107	-	10.000
contig0136	190.000, 250.000	250.000

7.3 Bijlage 3 : kwaliteit plot voor NASAIV



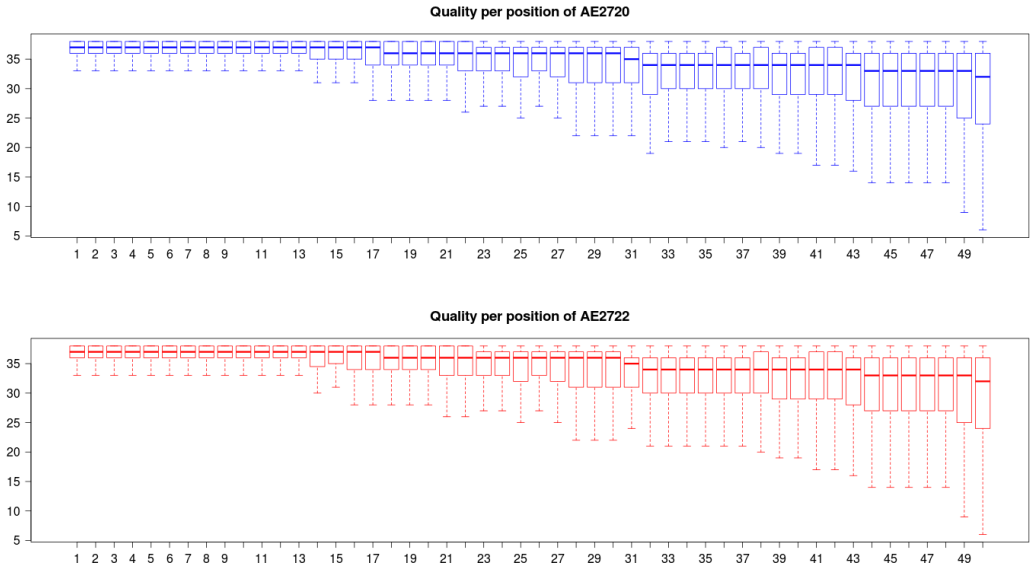
Figuur 7.1 : kwaliteitscontrole; kwaliteitscore per base plot voor de NAISAV stam, een duidelijke foute offset is gekozen. (66 in plaats van 99)

7.4 Bijlage 4 : kwaliteit plot voor S1 en S1H



Figuur 7.2 : kwaliteitscontrole; kwaliteitscore per base plot voor S1 en S1H stam.

7.5 Bijlage 5 : kwaliteit plot voor AE2722 en AE2720



Figuur 7.3 : kwaliteit per base positie plot voor zowel mutant AE2720 als AE2722.

7.6 Bijlage 6 : tijdbesteding voor run

	Gemiddelde tijd (seconden)	Stdev
Voorbereiding	0,05	0,23
Kwaliteitscontrole	2,45	0,74
find inserties	8,68	0,57
find deleties	17,18	0,50
teken inserties	40,82	6,63
split data	50,95	5,61
process sam	185,68	2,42
Aligment	534,05	53,38
teken deleties	819,95	5,63

Figuur 7.4 : tijdsbesteding tijdens een test run