Voor akkoord verklaring

*Dit eindwerk is een examen; eventuele fouten die worden vastgesteld tijdens de eindwerkverdediging of erna werden niet gecorrigeerd. Het gebruik als referentie in publicaties is toegelaten na goedkeuring van de promotor en eindwerkbegeleider (van de stageplaats)..*

Naam Familienaam Naam Familienaam

Stagementor Stagegever

Naam Familienaam

Promotor

Woord vooraf

Samenvatting

Lijst met afkortingen en symbolen

Verklarende woordenlijst

INDEL = insertie & deletie

Pareidolie = is een psychologisch verschijnsel, een vorm van illusie waarbij iemand een zodanige interpretatie van onduidelijke of willekeurige waarnemingen heeft, dat hij hierin herkenbare dingen meent waar te nemen. De naam is afkomstig van het Griekse para (naast) en eidolon (beeld). [[1](#Par12)]

Lijst van tabellen en figuren

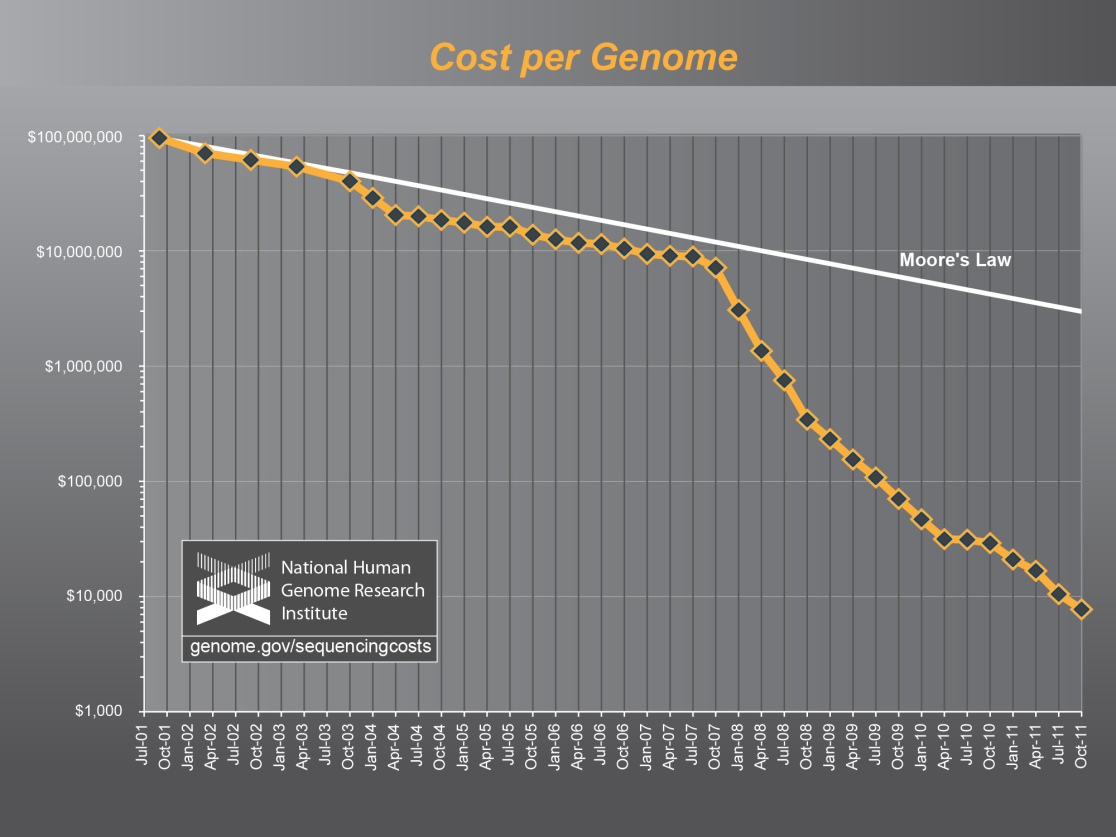
Inhoudsopgave

# Inleiding, probleemstelling en situatieschets

Resequencing is het opnieuw sequeneren van een reeds gesequeneerd genoom, gen of gebied in het genoom. Er zijn twee methoden, enerzijds een beperkte hersequenering van bepaald gebied of gen met doel mutaties te bepalen (*genotyping)* of anderzijds het volledige genoom hersequeneren om mutaties te vinden. Bij een volledige hersequenering kunnen *genome* *rearrangements* en/of mutaties zoals inserties (extra gebieden), deleties (verdwenen gebieden) en *single nucleotide polymorphismen* (SNP) worden gevonden. [[2](#Res12)] De zoektocht naar zulke regio’s is echter tijdsintensief en vereist ruime kennis en ervaring van de onderzoeker. Daarnaast kunnen paradolies[[1]](#footnote-1) gezien worden.

Al in 1976 is het eerste organisme gesequeneerd, namelijk de bacteriofaag *MS2*, toen nog op RNA niveau [[3](#WFI76)]. Een jaar later was het volledige genoom van de bacteriofaag *phi X 174* op DNA niveau gesequeneerd. [[4](#San77)] De snelheid van sequeneren is exponentieel gestegen, door nieuwe generatie sequencing systemen (*Next Gen Sequencing* toestellen) en de dalende prijs per gesequeneerd genoom. (zie figuur 1). [[5](#Wet12)]

Door deze trend komt manueel verwerken van (re-)sequenerings data en analyse van zulke data heel duur. Om aan dit probleem een oplossing te bieden moeten objectieve systemen ontworpen en geautomatiseerd worden. Bij resequencing is reeds informatie beschikbaar, door gebruik te maken van comperatieve genoom analyse kunnen bepaalde stappen verkort worden bij de analyse en kunnen extra validatie voor de data worden gevonden, door gebruik te maken van relevante gekende genomen.



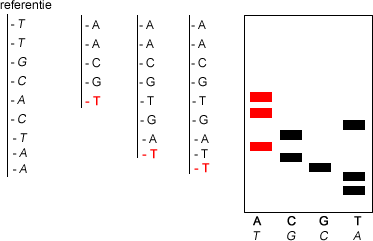
Figuur 1 : kost per genoom daalt enorm, de Y-as is logaritmisch. In deze prijs zitten werk, administratie, management, verbruik van producten, sequeneer toestellen en andere grote toestellen (over een periode van drie jaar), informatica management systemen en analyse, shotgun bibliotheek constructie en indirecte kosten. [[5](#Wet12)]

## Sequeneren

Sequeneren is het bepalen van de volgorde van de basen van erfelijk materiaal. Men kent verschillende methoden, om een inzicht te krijgen in de werking van onze oplossing is een klein overzicht van de meest gebruikte methoden nodig.

### *Dideoxy chain termination* methode

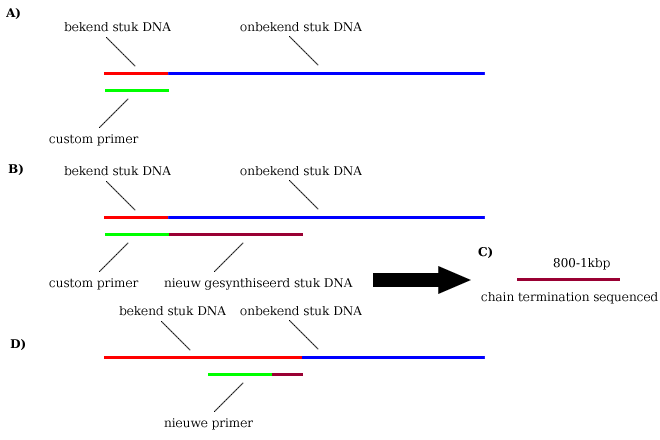
De dideoxy chain termination methode, (ook gekend als Sanger sequencing) beschreven in 1975 was tot rond de eeuw wisseling de norm. De methode gaat in zijn werk door vier aparte reacties op te zetten met gekloneerd single *stranded* DNA, namelijk voor iedere base één reactie medium. In dit medium zijn de vier basen aanwezig en van één van die basen is gemodificeerd zodat er geen verdere synthese meer kan gebeuren eens die gebonden is op de originele strand. Hierdoor bekomt men verschillende stukken die steeds eindigen op de gemodificeerde base. Wanneer men deze nu met behulp van elektroforese op grote scheidt kan men de basen aflezen van de gel. De methode kan DNA stukken accuraat lezen tot ongeveer 800 baseparen (bp) lang. (*reads*) Om toch grotere stukken te kunnen lezen zijn er twee methoden gekend, *primer walking*  of *chromosome walking* en de veel bekendere -door het gebruik in het *Human Genome Project* - de *shotgun* methode.(zie figuur 2) [[6](#Mik12)] [[7](#Seq12)]



Figuur 2 : schematische voorstelling van chain termination, hier rood aangeduid voor A (referentie, of T als complementaire base)

#### *Primer walking* methode

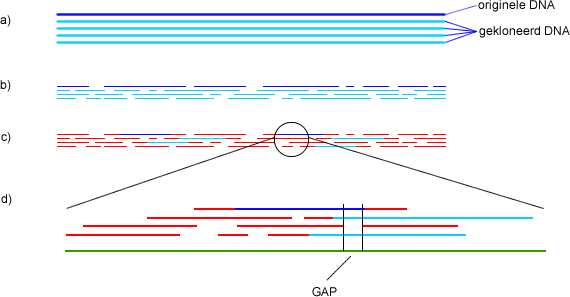
Bij *primer walking* ook *genome walking* genoemd, wordt gebruik gemaakt van aangepaste primers, de primer gaat *hybdridiseren* op het bekende stuk DNA (zie figuur 3, A) tijdens de *elongatie* (groei van DNA ,zie figuur 3, B) zal een nieuw stuk DNA geproduceerd worden, door *denaturatie* bekomt men een complementair stuk DNA van het onbekende stuk. Dit kan met m.b.v. *dideoxy chain termination* methode gaan sequeneren (zie figuur 3, C). Door in het nieuw gesequeneerd stuk DNA een nieuwe primer te gaan produceren kan men dezelfde stappen herlopen. (zie figuur 3, D) [[8](#pri12)] [[9](#Pri12)] [[10](#Pri121)]



Figuur 3 : schematische voorstelling van *primer walking* methode. A) hybridisatie van primer aan DNA. B) synthese van nieuw stuk DNA C) sequencing van nieuw stuk DNA D) productie van nieuwe primer om te hybridiseren verder in het onbekende stuk DNA.

#### *Shotgun* methode

Bij de *shotgun* methode wordt het DNA meerdere keren geamplificeerd (figuur 3, A) en vervolgend in willekeurige stukken geknipt (figuur 3, B). Deze worden dan allemaal gesequeneerd (figuur 3, C), door overlappende stukken te gaan bepalen kan men de originele sequentie achterhalen (figuur 3, D). Dit is erg computationeel intensief en vereist voor grotere genomen, -zoals het onze- gebalanceerde load over meerdere parallelle servers. [[6](#Mik12)] [[11](#Sho12)]



Figuur 4 : Schematisch overzicht *shotgun* methode; A) het originele DNA wordt gekloneerd B) het DNA word in willekeurige stukken geknipt C) de sequenties worden gesequeneerd met de dideoxy chain termination methode. D) de originele streng word achterhaalt door overlappende stukken DNA te matchen. Eventuele gaps kunnen voorkomen (zie VERDER, lader E & waterman MS, bespreking gaps + gaps length!!!!!!!!!!####).

### Pyrosequencing

Pyrosequencing is een alternatieve sequencing methode, die in 1996 ontwikkeld is [[12](#Ron96)] en commercieel beschikbaar is gemaakt door 454 Life Sciences (nu onderdeel van Roche) in 2004. [[13](#Seq121)] De methode maakt gebruik van de reactie die gebeurt wanneer pyrofosfaat (P2O74-) vrijkomt bij de hydrolyse van adenosinetrifosfaat (ATP) tot adenosinemonofosfaat (AMP). Deze methode produceert reads tot 100 bp. Nieuwere versies kan men reeds reads van 400 tot 500 bp bekomen met deze methode. Toch veel korter dan met de *dideoxy chain termination* methode. Door de veel kortere looptijden en de lagere kostprijs dan *dideoxy chain termination* methode werd deze techniek snel geadopteerd door grote sequencing centrums. [[6](#Mik12)]

### *Dye termination* methode

Bij de *dye termination* methode gebeurt de detectie van de base, met een gelabelde base. Zodra de gelabelde base bindt zal deze een licht uitzenden, de frequentie van het licht is van de vier verschillende basen verschillend, hierdoor kan men bepalen welke base op de positie aanwezig is. Deze methode wordt ook gebruikt bij *illumina sequencers*. Het grote voordeel ten opzichte van *dideoxy chain termination* is dat het in één reactie gebeurt en dat de snelheid hoger is gezien geen nood is aan elektroforese.

Door deze trend van grotere en goedkopere sequeneringsdata gecombineerd met veel kortere reads en bijgevolg veel grotere datafiles, is de nood aan een objectief en geautomatiseerd systeem erg hoog. Om een oplossing te bieden aan deze problemen hebben wij een pipeline gebouwd.

# Materiaal en methoden

## Hardware

### Server

De server die werd getest werkte op *centOS*, een *redhad* distributie.

Processorkracht : octa core, @ 2.83 GHz

Geheugen : 16 GB

### Client

Tijdens het ontwikkelen van de pipeline werd voornamelijk op *clientside* gewerkt. De volgende *operating systems* zijn gebruikt en getest (allen *debian*) Ubuntu 10.04 LTS, Ubuntu 10.10 en Ubuntu 12.10 LTS en CrunchBang Linux 10.

Processorkracht : dual core, @ 2.40 GHz

Geheugen : 2,96 GB beschikbaar.

## Software

### Mapping

#### Burrows-Wheeler Aligner

#### *Burrows-Wheeler Aligner* (BWA) is een snelle, accurate, geheugenefficiënte implementatie van het *Burrow-Wheeler transformation* algoritme, die korte queries (*reads)* aligneerd (*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. [[14](#BWA12)] [[15](#seq12)] Tijdens de ontwikkeling van de pipeline hebben we gebruik gemaakt van versie 0.6.1, deze versie aligneerd *reads* die op meerdere posities kunnen worden gealigneerd (*mapped*) op een random mogelijkheid. Als alternatief hebben wij een gepatchte versie gebruikt die *multimapped reads* wel in de resultaten opneemt, deze versie was 0.5.7.

#### Bowtie

*Bowtie* is een snelle, geheugenefficiënte, korte queries (*reads)* aligneerder, die ook gebruik maakt van het *Burrow-Wheeler transformation* algoritme. Wij gebruikten tijdens de ontwikkeling van de pipeline *Bowtie 2.0.0-beta2*. [[16](#Bow12)]

#### BFast

### Assembly

#### Velvet

Velvet is een *denovo assembler* voor korte *reads*, het werkt zowel met *illumina* data als *454* data. Velvet neemt de *reads*, verwijderd fouten en produceert vervolgens contigs. Wij gebruiken versie 1.2.03. [[17](#Vel12)]

#### SSPACE

SSPACE is een perl script ontwikkeld door *BaseClear BV* die met behulp van *paired-end* data en contigs, scaffolds maakt. Het maakt gebruik van een gemodificeerd SSAKE algoritme. Het SSAKE algoritme is een *denovo assembler* algoritme gemaakt voor extreem korte *reads* (~25 bp). Die volledige assembly oplevert voor kleine genomen (virussen) en voor grotere genomen (bacteriële) kan met niet-repetitieve contigs bepalen. [[18](#Mar10)] [[19](#Ass07)] Het algoritme gebruikt *Bowtie* om de reads (paired-reads) te gaan aligneren op de contigs. [[20](#SSP12)] [[21](#seq121)] Er is zowel een basis als premium versie van de software beschikbaar, wij gebruikten de basis versie 2.0.

### PipeLine

#### 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*, het is een eenvoudige, *high-end* en krachtige programmeer taal. 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 sequencie 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. [[22](#per12)] [[23](#bio12)] [[24](#wik12)] [[25](#wik121)]

#### R

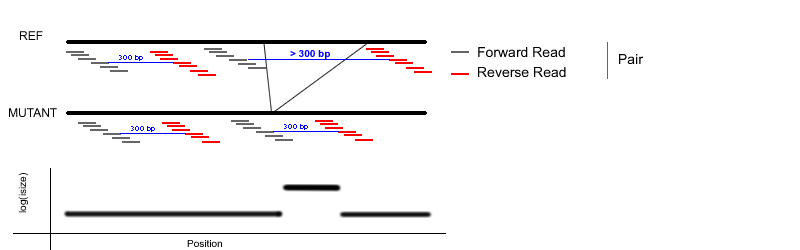
R is een programmeertaal voor statistische doeleinden. Het is ook handig voor visualisatie van grafieken en daarenboven multiplatform waardoor het als data-manipulatie en analyse tool erg handig blijkt voor grote hoeveelheden data, waarmee bio-informatici in contact komen. Daarnaast hebben we tijdens 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 bibeliotheek van R modules naar het voorbeeld van CPAN. Wij gebruikten de versie 2.15.0 van R. [[26](#wik122)]

#### PHP & HTML & CSS & JQuery & MySQL

### Methoden

#### Deleties zoeken

Wij gebruiken *paired-end* data, deze term slaat op de *library* voorbereiding. Wanneer de originele sequentie in stukken word geknipt zorgt men ervoor dat de afstand tussen twee stukken ongeveer een bepaalde lengte is. Zo is voor de *Rhodospirillum rubrum* (S1) gekozen om die lengte 500bp te maken en voor Cupriavidus metallidurans (CH34) 300bp. Deze lengte word de *Insert length* genoemd. Deze informatie maakt het eenvoudiger om variaties zoals inserties of deleties te gaan bepalen. Gezien de afstand op de gesequeneerde streng 300 bp is (bepaald door *library preparation*) zouden de twee gemaptte reads, op de referentie een afstand van 300bp moeten hebben. Wanneer dit echter niet het geval is, is een stuk DNA die tussen de twee reads zit die onderling 300bp van elkaar liggen op de mutant (gesequeneerde streng) verdwenen (deletie). (zie figuur 5) We zoeken deleties door de lengte (logaritmisch) tussen deze reads op het referentie genoom uit te zetten per positie en de gemiddelde waarde te bepalen over een klein gebied, om te bepalen of deze interessant is of niet. Daarnaast word met behulp van R voor iedere gebied waar tenminste één punt boven een treshold aanwezig is het gebied getekend ter controle.

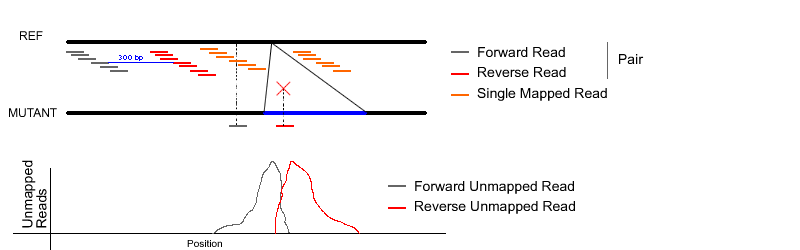


Figuur 5 : op de referentie is het eerste voorbeeld 300 bp van elkaar gescheiden, dit betekend dat er geen DNA is verdwenen uit de mutant, de 2de lijst van reads evenwel zijn op de referentie verder dan 300bp uit elkaar gelegen, dit bewijst dat een stuk DNA op de mutant niet beschikbaar is. Wanneer men logaritmisch de insert size uitzet per positie dan bekomt men duidelijk een plaatselijke verhoging van de insert size. Dit kan wijzen op een deletie op de mutant (tov de referentie)

#### Inserties

##### Inserties zoeken

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



##### Inserties bepalen

Naast de positie van inserties is ook interessant om te bepalen welke sequentie dit nieuw stuk DNA heeft. Dit kan door de reads die enkel eenmaal mappen te nemen, en lokale assembly gebruiken met beide reads om de sequentie proberen te bepalen.

#### SNP / INDEL Calling

# Resultaten

# Discussie

# Besluit

Literatuurlijst

# Geciteerde werken

x

|  |  |
| --- | --- |
| [1] | wikipedia. [Online]. <http://nl.wikipedia.org/wiki/Pareidolie> |
| [2] | NCBI. [Online]. <http://www.ncbi.nlm.nih.gov/projects/genome/probe/doc/TechResequencing.shtml> |
| [3] | 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. |
| [4] | 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. |
| [5] | Wetterstrand KA. genomes.gov. [Online]. <http://www.genome.gov/sequencingcosts/> |
| [6] | 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> |
| [7] | wikipedia. [Online]. <http://en.wikipedia.org/wiki/Sequence_assembly> |
| [8] | fishersequencing. [Online]. <http://www.fishersequencing.com/primerwalking.htm> |
| [9] | wikipedia. [Online]. <http://en.wikipedia.org/wiki/Primer_walking> |
| [10] | umich.edu. [Online]. <http://seqcore.brcf.med.umich.edu/doc/dnaseq/primerwalking.html> |
| [11] | wikipedia. [Online]. <http://en.wikipedia.org/wiki/Shotgun_sequencing> |
| [12] | 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. |
| [13] | (2012) wikipedia. [Online]. <http://en.wikipedia.org/wiki/Sequence_assembly> |
| [14] | (2012, maart) BWA. [Online]. <http://bio-bwa.sourceforge.net/bwa.shtml> |
| [15] | (2012, maart) seqanswers. [Online]. <http://seqanswers.com/wiki/BWA> |
| [16] | (2012, maart) Bowtie. [Online]. <http://bowtie-bio.sourceforge.net> |
| [17] | (2012, maart) Velvet. [Online]. <http://www.ebi.ac.uk/~zerbino/velvet/> |
| [18] | 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. |
| [19] | "Assembling millions of short DNA sequences using SSAKE," *bioinformatics*, vol. 23, no. 4, pp. 500–501, december 2007. |

x

1. Type optische illusies, de hersenen zien zaken die er niet zijn door vlaspositieve patroon herkenning. [↑](#footnote-ref-1)