DNA aflezen met Next Generation Sequencing

Ken Kraaijeveld
Bo Blanckenburg

Eerste druk

Syntax Media – Amersfoort

ISBN: 978 94 91764 64 6

Ontwerp omslag: Lapis Vivus grafisch ontwerp, Oosterbeek
Opmaak binnenwerk: AlphaZet prepress, Bodegraven
Tekstredactie: Redactie & zo, ir. Caroline van der Meulen

Illustratieverantwoording:
Illustratie omslag: Bo Blanckenburg
Figuur 5.8: Lee, J.L. et al.
Figuur 9.9; 9.10: Kerkvliet, J. et al.
Figuur 9.13a: Zhang, D. et al., bewerkt
Figuur 9.13b; 9.13c: Zhang, D. et al.
Figuur 12.1; 12.2: Alfaro, J.A. et al., bewerkt
Overige illustraties: Bo Blanckenburg

Eventuele opmerkingen over deze uitgave kunt u richten aan:
info@syntaxmedia.nl

www.syntaxmedia.nl

Inhoud

Voorwoord 9

1 DNA sequencing 11

1.1 Geschiedenis DNA sequencing 11
1.2 Waarom DNA aflezen? 13
1.2.1 Mogelijkheden NGS 13
1.2.2 Rol van bio-informatici 15
1.3 Hoe lees je DNA af? 16
Box I – Structuur DNA-molecuul 17
Box II – Genen 18

2 Sequencing apparaten 19

2.1 Verschillende NGS-technieken 19
2.2 Short-read sequencers 19
2.2.1 Illumina 19
2.2.2 Ion Torrent 25
2.3 Single-molecule sequencers 28
2.3.1 PacBio 29
2.3.2 Oxford Nanopore 31
2.4 Oefenvragen en opdrachten 34

3 Read kwaliteit 35

3.1 Wat is read kwaliteit? 35
3.2 Belang van read kwaliteit 36
3.3 Platformspecifieke fouten 36
3.3.1 Illumina 36
3.3.2 Ion Torrent 38
3.3.3 PacBio 38
3.3.4 Oxford Nanopore 39
3.4 Weergave van de kwaliteit in de output 39
3.4.1 Phred score 39
3.4.2 Run statistieken 41
3.5 Verbeteren van de kwaliteit 47
3.5.1 Read trimming 47
Box I – Twee trimming algoritmes 48
3.5.2 Lange reads 51
3.6 Uitgewerkt voorbeeld 51
3.7 Veelgebruikte tools 52
3.8 Oefenvragen en opdrachten 53

4 Reads plaatsen op een referentiegenoom 55

4.1 Wat is read alignment? 55
4.2 Belang van read alignment 56
4.3 Hoe werkt read alignment? 57
4.3.1 Referentiegenoom 57
4.3.2 Alignment algoritmes 59
Box I – Alignment penalty's 67
4.3.3 SAM format 67
Box II – Bitwise flag 71
4.3.4 Visualisatie van alignments 73
4.4 Uitgewerkt voorbeeld 75
4.5 Veelgebruikte tools 77
4.6 Oefenvragen en opdrachten 78

5 Varianten detecteren 79

5.1 Wat zijn varianten? 79
5.2 Mogelijkheden van variantdetectie 80
5.3 Hoe werkt variantdetectie? 81
5.3.1 SNPs en korte indels 81
5.3.2 Structurele varianten 83
5.3.3 Varianten opslaan 84
5.3.4 Varianten faseren 86
5.4 Uitgewerkt voorbeeld 87
5.5 Veelgebruikte tools 88
5.6 Oefenvragen en opdrachten 89

6 Een genoom samenstellen zonder voorkennis 91

6.1 Wat is de novo assembly? 91
6.2 Wat kun je ermee? 92
6.3 Hoe werkt de novo assembly? 92
6.3.1 Grafen theorie 94
6.3.2 Onregelmatigheden in de graaf 95
Box I – De Bruijn graaf 101
6.3.3 Scaffolding 101
6.3.4 Assembly kwaliteit 105
6.4 Uitgewerkt voorbeeld 106
6.5 Veelgebruikte tools 107
6.6 Oefenvragen en opdrachten 107

7 Gericht aflezen 109

7.1 Specifieke delen van een genoom aflezen 109
7.2 Mogelijkheden van gericht aflezen 109
7.3 Hoe werkt gericht aflezen? 110
7.3.1 Target capture 110
7.3.2 PCR-amplificatie 112
7.3.3 ChIP-seq 114
7.4 Uitgewerkt voorbeeld 115
7.5 Veelgebruikte tools 116
7.6 Oefenvragen en opdrachten 117

8 Transcriptoom sequencen 119

Box I – RNA-moleculen 119
8.1 RNA sequencen 121
8.2 Mogelijkheden van RNA sequencing 121
8.2.1 Transcriptstructuur 122
8.2.2 De novo assembly 123
8.3 Technieken voor RNA sequencing 123
8.3.1 RNA-isolatie en library prep 123
8.3.2 Transcriptstructuur 124
8.3.3 Transcriptoom assembly 125
8.4 Uitgewerkt voorbeeld 126
8.5 Veelgebruikte tools 126
8.6 Oefenvragen en opdrachten 126

9 Genexpressie bepalen 127

9.1 Genexpressie 127
9.2 Differentiële genexpressie 128
9.3 DGE meten 128
9.3.1 Experimentele opzet 128
9.3.2 mRNA-isolatie 128
9.3.3 Read alignment 129
9.3.4 Kwantificatie 130
9.3.5 Normalisatie 131
9.3.6 Differentiële expressie analyse 135
9.3.7 Single-cell RNAseq 136
9.4 Uitgewerkt voorbeeld 139
9.5 Veelgebruikte tools 141
9.6 Oefenvragen en opdrachten 141

10 Epigenetica en genregulatie 143

10.1 Wat is het? 143
Box I – Transposons 144
10.2 Wat kun je ermee? 146
10.3 Hoe werkt het? 146
10.3.1 Methylatiedetectie 146
10.3.2 Chromatine toegankelijkheid 148
10.3.3 Kleine RNAs sequencen 149
10.3.4 RNA-editing 149
10.4 Uitgewerkt voorbeeld 150
10.5 Veelgebruikte tools 151
10.6 Oefenvragen en opdrachten 151

11 Metagenomics 153

11.1 Meerdere genomen tegelijk sequencen 153
11.2 Mogelijkheden van metagenomics 154
11.3 Metagenomics technieken 155
11.3.1 Meten van biodiversiteit 155
11.3.2 Functionele analyse 159
11.4 Uitgewerkt voorbeeld 159
11.5 Veelgebruikte tools 160
11.6 Oefenvragen en opdrachten 161

12 Toekomstmuziek: eiwit sequencing 163

12.1 Waarom je eiwitten zou willen sequencen 163
12.2 Next Generation Eiwit Sequencing 164

Bijlagen 167

Bijlage I Lijst van de belangrijkste NGS-platforms 168
Bijlage II Verklarende woordenlijst 169

Literatuur 175

Register 179

Voorwoord

Vissenpopulaties in kaart brengen door DNA te analyseren uit emmertjes water. Of terwijl een operatie bezig is, al weten welk tumortype de patiënt heeft! De opkomst van het razendsnelle Next Generation Sequencing (NGS) heeft de laatste twintig jaar biologisch onderzoek fundamenteel veranderd. Hoe je van DNA in een reageerbuis tot lettervolgorde en genetische informatie komt, is essentiële kennis geworden voor iedere biologisch analist. Want steeds meer laboratoria krijgen hun eigen NGS-machines of gebruiken de techniek in het veld.

We hebben *DNA aflezen met Next Generation Sequencing* geschreven om deze complexe materie toegankelijk te maken voor hbo-studenten bio-informatica, biomedisch laboratoriumonderzoek en professionals. Dit boek behandelt hoe verschillende NGS-platforms data generen, hoe deze geanalyseerd worden met bio-informatica en wat veelgebruikte toepassingen zijn. De focus ligt op de dataverwerking en niet op de labtechnieken. Basale voorkennis van veelgebruikte moleculaire labtechnieken, zoals PCR, wordt als bekend verondersteld. We hebben ons best gedaan om alles stap voor stap uit te leggen en zo duidelijk mogelijk te laten zien in illustraties. In elk hoofdstuk worden voorbeelden van toepassingen gegeven. Vaak komen deze uit eigen onderzoek. In de gevallen dat het werk niet in een peer-reviewed tijdschrift is gepubliceerd, zijn de gegevens beschikbaar gemaakt via een online repository.

Tijdens het schrijven hebben we besloten om voor de technische termen het Engelse jargon te handhaven, omdat deze termen ook zo in het Nederlandse werkveld worden gebruikt.
Belangrijke begrippen in het boek hebben een blauwe kleur gekregen en zijn voorzien van een uitleg als ze voor het eerst voorkomen. Achterin vind je een verklarende woordenlijst (bijlage II) met de belangrijkste begrippen. Alle blauwgekleurde termen in het boek vind je terug in het register.
Antwoorden en uitwerkingen op de oefenvragen en opdrachten staan online op www.syntaxmedia.nl op de boekpagina.

Zonder hulp van onze collega's op de Hogeschool Leiden was dit boek niet tot stand gekomen. Dank aan Nicole Almering, Maria Plug en Danny Dukers die ons hebben aangemoedigd dit project aan te gaan. De basis van dit boek wordt gevormd door onderwijsmateriaal dat ontwikkeld is bij de opleiding bio-informatica aan de Hogeschool Leiden. Veel collega's hebben hieraan bijgedragen: Saskia Stahlecker-Vosslamber, Koen Bossers, Marc Besseling, Marlous Hoogstraat, Hilda van Mourik, Paul de Raadt, Said Basmagi, Jan Oliehoek en Jeroen Pijpe.

Voorjaar 2025
Ken Kraaijeveld en Bo Blanckenburg

Hoofdstuk 2
Sequencing apparaten

2.1 Verschillende NGS-technieken

NGS-technieken zijn heel divers. De volgorde van DNA-basen worden bepaald met fluorescent gelabelde nucleotiden, verschillen in pH of met verstoringen in een elektrisch veld. Het omzetten van zo'n signaal naar sequentie heet basecalling. De resulterende stukken sequentie verschillen van 100 basen tot tienduizenden basen. Sommige apparaten hebben de afmetingen van een wandmeubel, andere kun je in je zak steken. Wat al deze technieken gemeen hebben, is dat ze een heleboel DNA-fragmenten tegelijk aflezen. Dit is waarom NGS ook bekend staat onder de naam massive parallel sequencing. In het navolgende worden vier van de meestgebruikte technieken besproken. Twee short-read sequencers (paragraaf 2.2) en twee single-molecule sequencers (paragraaf 2.3). In tabel 2.5 vind je het overzicht van de output van de besproken sequencing technieken. Een volledige lijst van apparaten die in de afgelopen decennia ontwikkeld (en soms ook weer van het toneel verdwenen) zijn, is te vinden in bijlage I.

2.2 Short-read sequencers

NGS is begonnen met het in parallel aflezen van grote aantallen korte DNA-fragmenten, short-read sequencing. Vroege machines van Illumina en ABI lazen zo'n 30 basen per fragment af, terwijl bijvoorbeeld de machine van Roche zo'n 500 basen per fragment las. Inmiddels zijn er technieken waarmee fragmenten van tienduizenden basen afgelezen kunnen worden. Toch is short-read sequencing (ook wel second generation sequencing genoemd) nog steeds een dominante NGS-techniek, omdat het enorme hoeveelheden sequentie van hoge kwaliteit kan produceren.

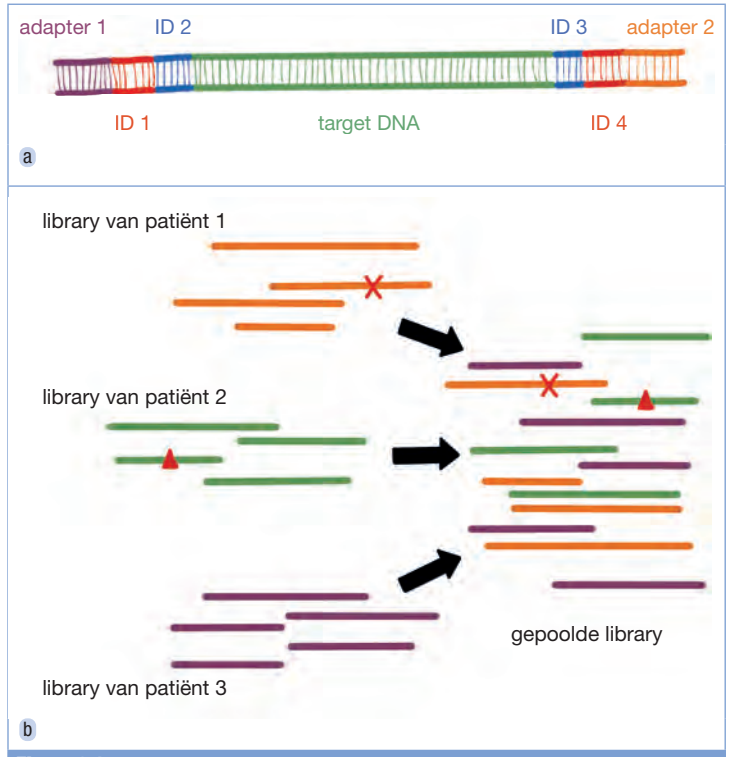
2.2.1 Illumina

De apparaten van Illumina zijn de werkpaarden van de NGS, vooral vanwege de grote capaciteit en de hoge kwaliteit van de sequencing data (tabel 2.1). Je kunt geen moleculairbiologisch artikel openslaan waarin deze techniek niet ergens in het onderzoek is gebruikt. Ook in diagnostiek, forensisch onderzoek, zaadveredeling en vele andere gebieden is het een niet meer weg te denken instrument. Een reeks aan verschillende modellen is op

de markt, van de kleine miniseq, geschikt voor het aflezen van korte PCR-fragmenten, tot de novaseq, waarop 48 menselijke genomen in één run kunnen worden afgelezen.

Illumina library prep

Illumina sequencers lezen korte DNA-fragmenten af. Het DNA dat aangeboden wordt, moet binnen specifieke lengtegrenzen vallen. Dit kan bereikt worden door de DNA-moleculen eerst te fragmenteren (shearing), of door PCR-fragmenten van de juiste lengte te gebruiken. DNA fragmenteren tot de juiste lengte kan op verschillende manieren. Vaak is het belangrijk dat de originele DNA-moleculen op willekeurige plekken gebroken worden, zodat de fragmenten elk op een unieke plaats beginnen en eindigen. Als je hetzelfde stuk genoom meerdere keren hebt afgelezen, zullen de fragmenten elkaar deels overlappen. In hoofdstuk 6 wordt uitgelegd waarom dit belangrijk is. Aan de uiteinden van de DNA-fragmenten worden adapters geplakt (figuur 2.2:1-2). Deze adapters zijn synthetische stukjes dubbelstrengs DNA waarvan de basenvolgorde precies bekend is. Eventueel kunnen



Figuur 2.1

De analyse van meerdere monsters samen in één sequencing run.
(a) Een library-fragment met daaraan de sequencing adapters en verschillende barcodes of indexen.
(b) Het poolen van sequencing libraries van drie patiënten. De ID-sequenties uit a zorgen ervoor, dat je kunt achterhalen van welke patiënt de read afkomstig is. De rode symbolen zijn unieke sequenties voor patiënten.

er ook zogenoemde barcodes of indexen ingebouwd worden (figuur 2.1a). Ook dit zijn stukjes synthetisch dubbelstrengs DNA met een bekende sequentie, die bijvoorbeeld specifiek is voor het monster dat je gaat sequencen. Hierdoor kun je meerdere monsters samen in één sequencing run analyseren, omdat je later met de index-sequenties van elke read kunt achterhalen bij welk monster hij hoort (library pooling of multiplexing; figuur 2.1b).

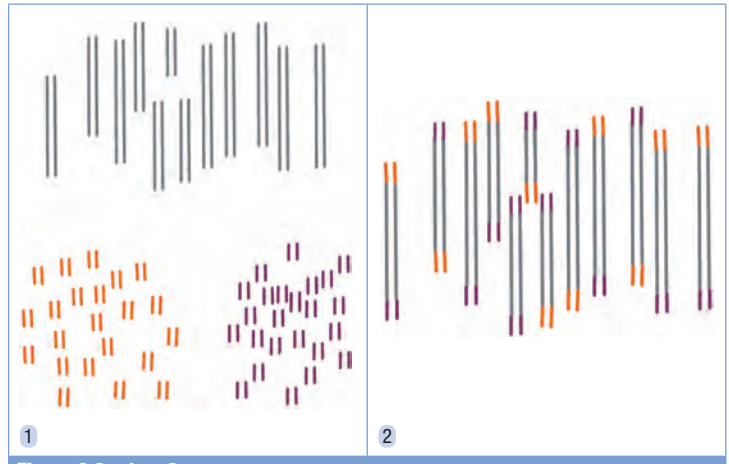
Illumina sequencing

Illumina sequencing vindt plaats in een flowcell. Een flowcell is een glazen plaatje waarbinnen over de lengte holle kanaaltjes (laantjes) lopen. Aan de binnenkant van die laantjes is het glas bedekt met een gazon van korte oligos (korte stukjes synthetisch enkelstrengs DNA; figuur 2.2:3), waarvan de sequentie overeenkomt met die van de adapters die aan de DNA-fragmenten gekoppeld zijn. De library – alle DNA-fragmenten met de daaraan gekoppelde adapters en eventuele indexen – wordt enkelstrengs gemaakt en in een laantje gebracht. Vervolgens hechten de fragmenten uit de library zich aan de complementaire oligos op het glas (figuur 2.2:4.1). Op deze manier worden de library-moleculen geïmmobiliseerd. In principe zou je nu de fragmenten kunnen sequencen, maar het signaal dat van een enkel molecuul afkomt is helaas te zwak om betrouwbaar te detecteren.

Om een sterk genoeg signaal te genereren, wordt elk van de moleculen vermeerderd in een bridge-PCR. Eerst wordt het DNA dubbelstrengs gemaakt door een complementaire streng te synthetiseren (figuur 2.2:4.1-4.2). Het stukje DNA dat aan het glas vastzit, wordt daarbij als beginpunt voor de nieuwe streng (primer) gebruikt. Hierdoor zit de nieuwe streng vast aan de flowcell. De originele streng zit alleen via waterstofbruggen vast aan het stukje DNA dat aan het glas vastzit. Door de temperatuur te verhogen, worden deze waterstofbruggen verbroken en laat de originele streng los (figuur 2.2:4.3). De nieuwe streng is nu enkelstrengs en zit vast aan het glas (figuur 2.2:4.4). Vervolgens wordt de temperatuur verlaagd tot de optimale temperatuur voor het hechten van twee complementaire DNA-strengen (annealing temperatuur). Hierbij zullen de adapters aan het vrije uiteinde van elk library-molecuul zich hechten aan de complementaire oligos in het gazon op het glas. Hierdoor zit het molecuul aan twee kanten aan het glas vast en vormt een boog (bridge; figuur 2.2:5). Opnieuw maakt een polymerase het DNA dubbelstrengs (figuur 2.2:6-7) en opnieuw worden daarna de waterstofbruggen verbroken. Nu zitten er twee kopieën van hetzelfde fragment vast aan het glas (figuur 2.2:8). Dit proces herhaalt zich tot elk fragment uit de library is vermeerderd tot zo'n 1000 kopieën (figuur 2.2:9). Dit cluster van duizend kopieën geeft een sterk genoeg signaal om te kunnen detecteren.

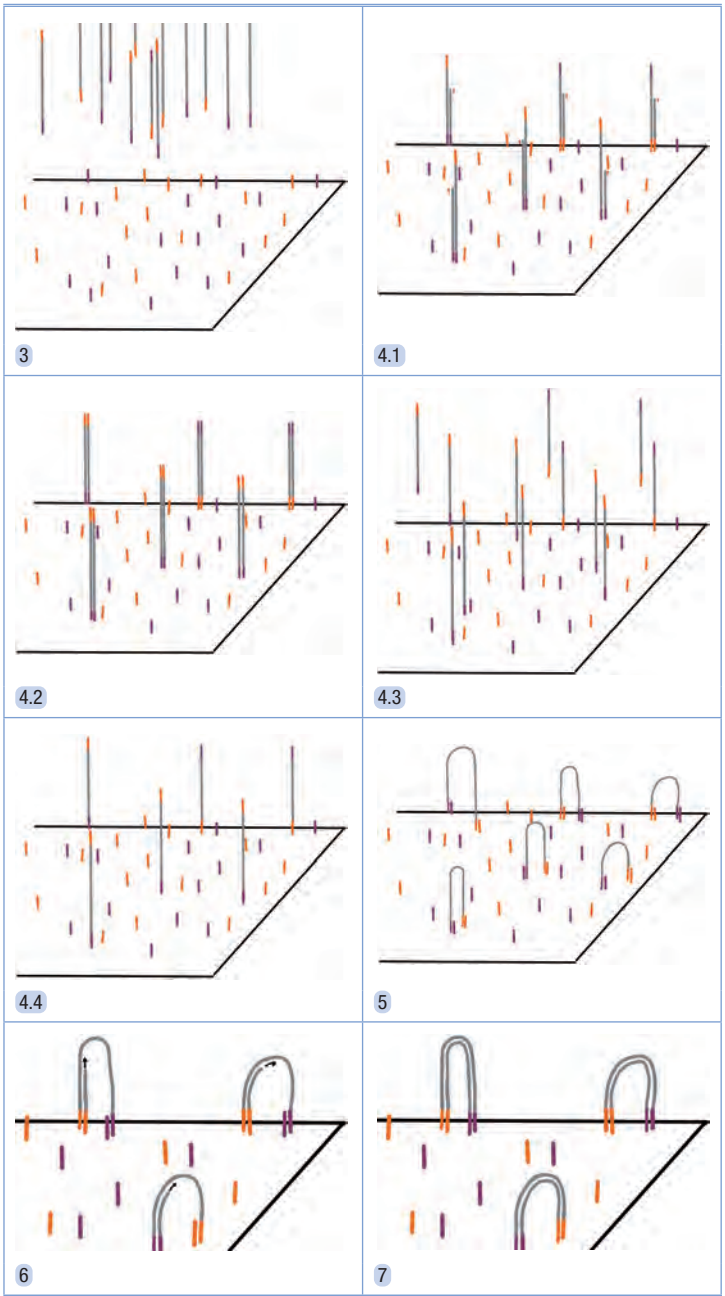
Als de clusters gevormd zijn, kan de sequencing beginnen (figuur 2.2:10.1). Een primer hecht zich aan de adapter-sequentie aan het vrije uiteinde van de moleculen van het cluster (figuur 2.2:10.2). Deze primer vormt het startpunt vanwaar de polymerase een nieuwe streng kan gaan synthetiseren. De nucleotiden die het daarvoor nodig heeft, bevinden zich in de oplossing en zijn voorzien van een fluorescent label (één kleur voor elk van de vier basen). Als de eerste base van het fragment een G is, zal de polymerase de nieuwe streng van elk van de 1000 kopieën laten beginnen met een C. De C-nucleotiden zijn voorzien van een blauw label, waardoor het cluster blauw oplicht (figuur 2.2:11.1). Tegelijkertijd zorgt het label ervoor dat de polymerase niet verder kan. De volgende nucleotide kan nog niet worden ingebouwd. Dat geeft het systeem de tijd om een foto te maken waarop het blauwe lichtpuntje is te zien. Andere clusters zullen rood, geel, of groen oplichten, afhankelijk van welke nucleotide is ingebouwd. De foto vertelt dus welke nucleotide is ingebouwd op de eerste positie van elk library-cluster. Nadat de foto genomen is, wordt het label verwijderd en kan de polymerase de volgende nucleotide inbouwen (figuur 2.2:11.2-11.3). Opnieuw lichten de clusters geel, rood, blauw, of groen op afhankelijk van welke base zich op positie twee van het fragment bevond (figuur 2.2:12-14). Opnieuw kan de polymerase daarna niet verder en opnieuw wordt er een foto gemaakt.

Dit proces herhaalt zich 100 tot 250 keer. Het eindresultaat is een stapel van 100 tot 250 foto's vol met gele, rode, blauwe en groene puntjes. De software kan voor elk puntje noteren welke kleur het bij elke cyclus had en zo de basenvolgorde van het originele fragment reconstrueren. Afhankelijk van hoeveel keer dit proces is herhaald, zal de resulterende sequentie 100 tot 250 basen lang zijn.



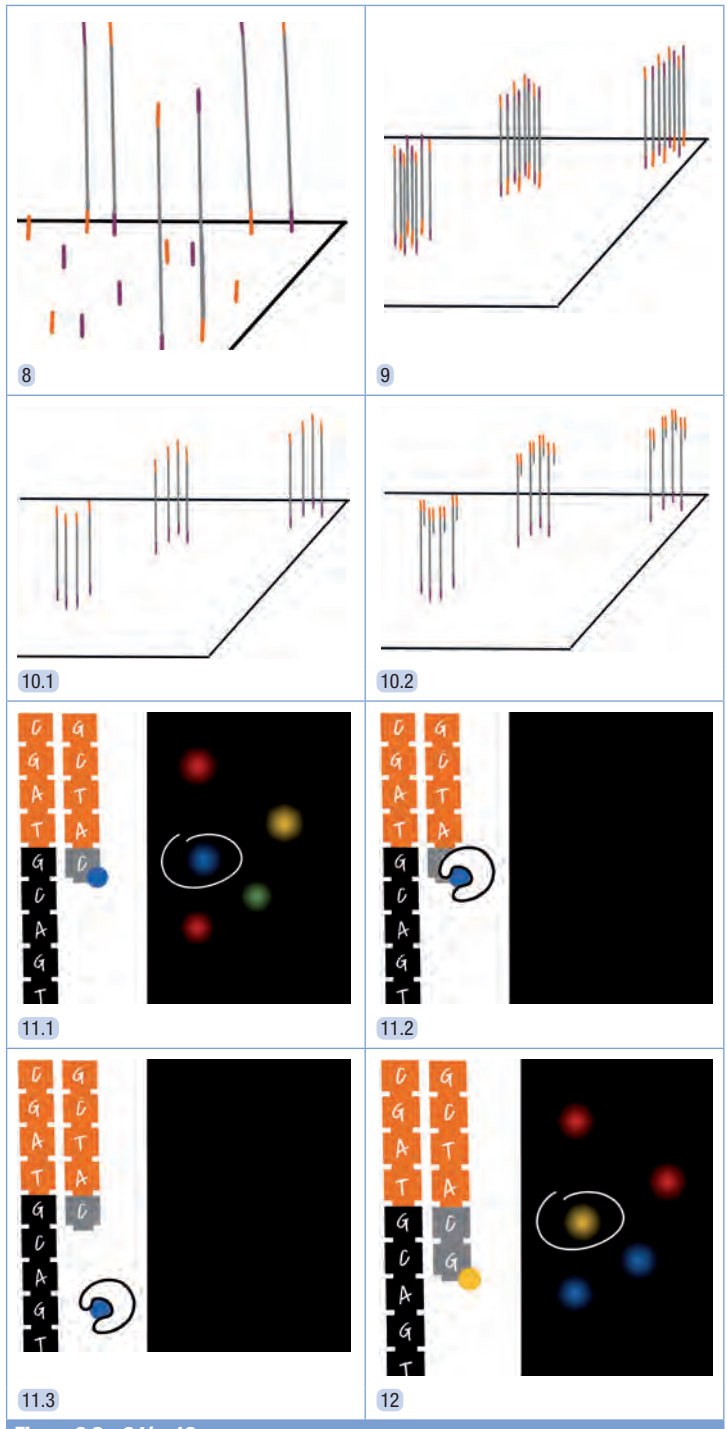
Figuur 2.2 – 1 en 2

Het principe van Illumina sequencing.



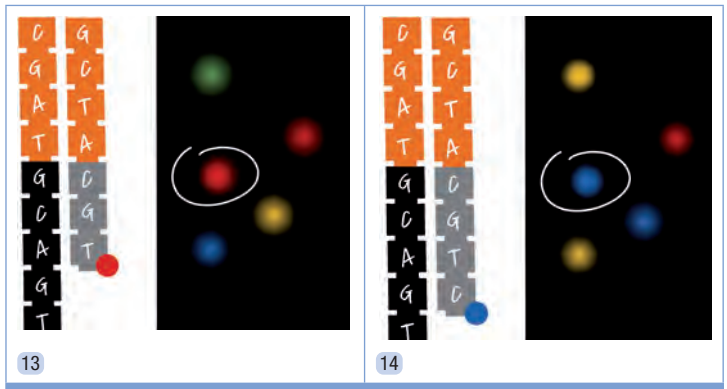
Figuur 2.2 – 3 t/m 7

Het principe van Illumina sequencing.



Figuur 2.2 – 8 t/m 12

Het principe van Illumina sequencing.



Figuur 2.2 – 13 en 14

Het principe van Illumina sequencing.

De eerste 100 tot 250 basen van miljoenen library-fragmenten geven een schat aan informatie, maar korte reads brengen in bepaalde toepassingen beperkingen met zich mee (zie bijvoorbeeld hoofdstuk 6). Langere sequenties bieden veel voordelen. Om meer sequentie van elk fragment af te lezen, heeft Illumina het zogenoemde paired-end sequencing ontwikkeld. Nadat de eerste read is afgelezen, wordt via een nieuwe bridge-PCR-stap het library-molecuul als het ware omgedraaid en de sequencing hervat vanaf de andere kant van het molecuul. Op deze manier verkrijg je twee sequenties van elk library-molecuul: 100 basen van de linkerkant en 100 basen van de rechterkant. Als het totale fragment bijvoorbeeld 500 basen lang was, zit er tussen de twee sequenties een stuk van 300 (500 − 100 − 100) posities waarvan je de sequentie niet weet. Uit de library prep weet je tot welke lengte je het DNA gefragmenteerd hebt en dus bij benadering hoe lang elk library-fragment is. Net als bij de eerste read bepaalt de software welke base op elke positie is ingebouwd aan de hand van de kleur van elk cluster op de foto's. Aangezien de clusters nog steeds op dezelfde plek zitten, kan de software ook noteren welke read 2 bij welke read 1 hoort.

Tabel 2.1
De voor- en nadelen van Illumina sequencing

Voordelen	*Nadelen*
hoge capaciteit (veel data per run)	korte reads
hoge kwaliteit (weinig fouten)	PCR-stappen nodig

2.2.2 Ion Torrent

In 2010 kwam de Ion Torrent op de markt. De Ion Torrent was het eerste sequencing apparaat dat compact genoeg was om op een labtafel te passen. De sequencing op Ion Torrent verloopt relatief snel en de goedkope reagentia maken het apparaat aan-

trekkelijk voor allerlei toepassingen zoals het sequencen van PCR-fragmenten (amplicon sequencing). De opbrengst is echter te klein om bijvoorbeeld een compleet genoom van een mens af te lezen (tabel 2.2). De grote versie van de Ion Torrent, de Ion Proton, is hier wel geschikt voor.

Ion Torrent library prep

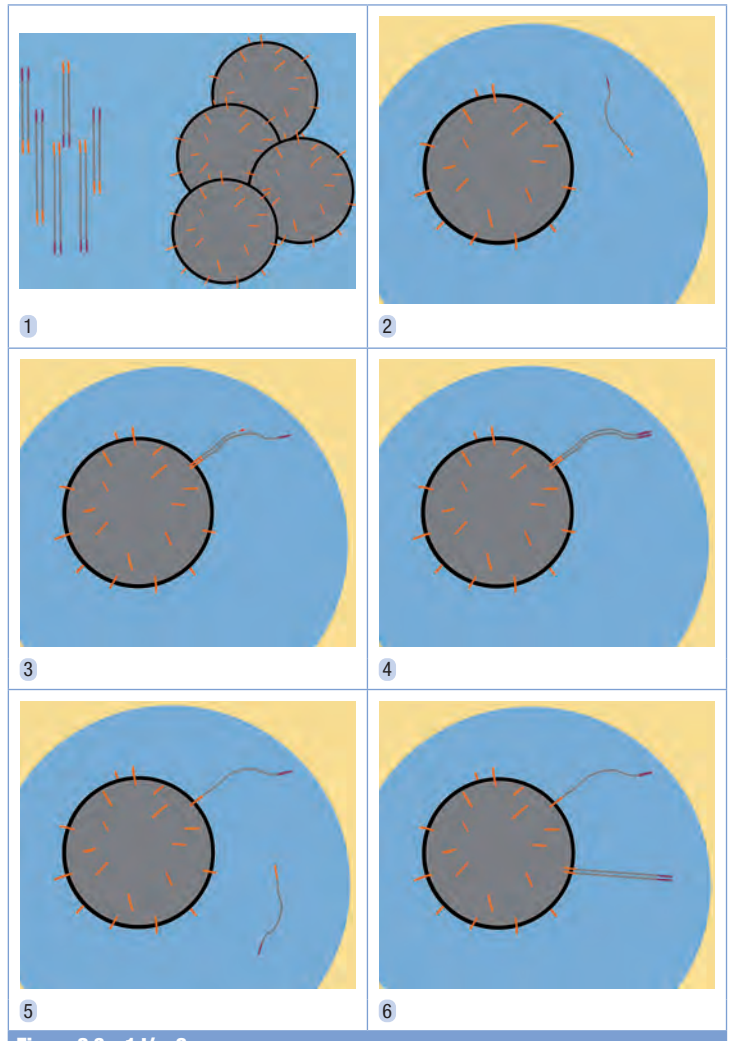
De library prep voor Ion Torrent maakt gebruik van emulsie-PCR (figuur 2.3). Het DNA wordt gefragmenteerd tot een lengte van enkele honderden basen en voorzien van adapters (figuur 2.3:1). Deze DNA-library wordt gemengd met PCR-reagentia en magnetische balletjes (beads). Die balletjes zijn bedekt met synthetische stukjes DNA waarvan de sequentie complementair is aan die van de adapters (figuur 2.3:1). Dit alles wordt in een emulsie van water in olie gebracht. Water mengt niet met olie, dus na flink schudden ontstaan er waterbelletjes in de olie. Als alles in de juiste verhoudingen is gecombineerd, zal elk van deze belletjes één balletje, één DNA-fragment en alle PCR-reagentia bevatten (figuur 2.3:2). Op deze emulsie wordt een PCR uitgevoerd, waarbij de stukjes DNA op de balletjes als primer werken (figuur 2.3:3-7). Het gevolg is dat elk van de balletjes bedekt raakt met kopieën van het DNA-fragment uit de library dat in hetzelfde waterbelletje zat (figuur 2.3:8). Na afloop van de PCR worden de balletjes uit de emulsie gehaald en zijn de DNA-fragmenten die eraan vastzitten klaar om afgelezen te worden.

Ion Torrent Sequencing

Als de library klaar is om te worden gesequencet, wordt deze op een chip gebracht Zo'n chip is een klein plaatje met op de bodem kleine gaatjes die precies groot genoeg zijn voor een balletje. Elk van de met DNA bedekte balletjes wordt in een gaatje van de chip gebracht. Een primer waarvan de sequentie complementair is aan de buitenste adapter (de kant waarmee het fragment uit de library niet vastzit aan het balletje), dient als startpunt voor een polymerase. Nu worden één voor één de nucleotiden toegevoegd, zodat de polymerase de nieuwe DNA-streng kan synthetiseren (figuur 2.3:9). Belangrijk hierbij is dat er maar één type nucleotide tegelijk wordt toegevoegd. Als op de positie in het DNA-fragment waar de polymerase gebleven is, een C aanwezig is en er worden G's aangeboden, dan zal de polymerase een G inbouwen. Worden er T's aangeboden, dan gebeurt er niets. Elke keer dat er een nucleotide wordt ingebouwd, komt er een waterstofion vrij (figuur 2.3:10). Dit verandert de pH van de oplossing in het gaatje van de chip en dat is meetbaar. In zekere zin is de Ion Torrent dus een geavanceerde pH-meter. Uit het patroon van aan- of afwezigheid van pH-verandering bij elke cyclus kun je afleiden in welke volgorde de DNA-basen werden ingebouwd. Doordat je weet in welke volgorde de verschillende nucleotiden

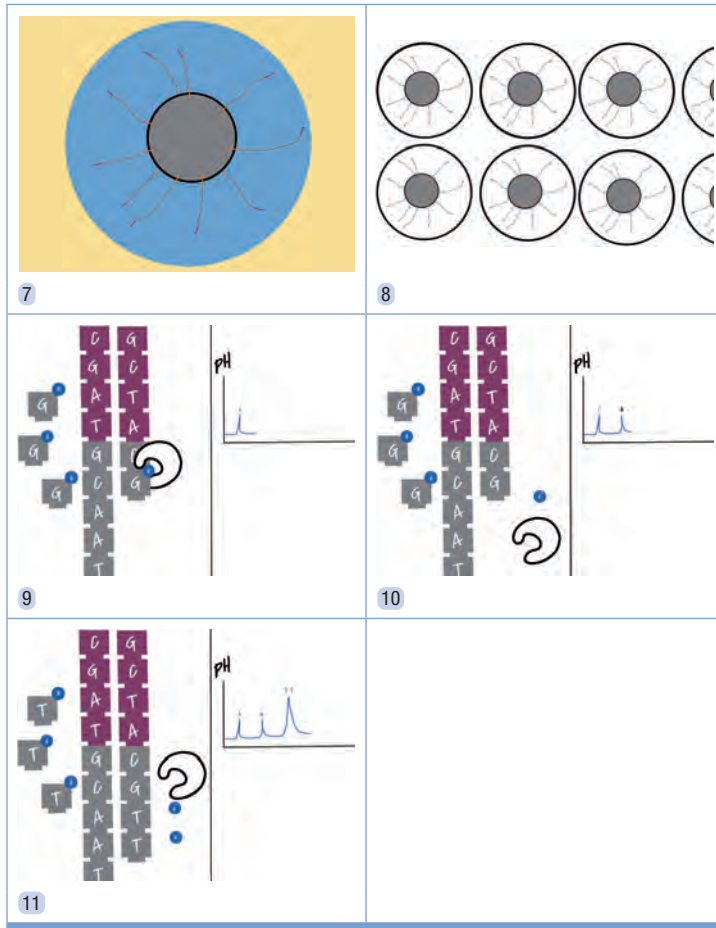
zijn aangeboden, kom je achter de oorspronkelijke sequentie van het afgelezen DNA-fragment.

Soms bevat het af te lezen DNA-fragment een rijtje van dezelfde basen, een homopolymeer zoals twee A's (sequentie AA). Als er in zo'n geval T-nucleotiden worden aangeboden, zullen alle twee de T's worden ingebouwd (figuur 2.3:11). Daarbij komen dan twee waterstofionen vrij, wat een signaal geeft dat twee keer zo sterk is dan dat van een enkel waterstofion. Hiermee kunnen dus homopolymeren toch afgelezen worden. De software reconstrueert vervolgens de basenvolgorde van de DNA-fragmenten uit het patroon van aan- of afwezigheid van een piek tijdens elke cyclus en de hoogte van die pieken.



Figuur 2.3 – 1 t/m 6

Het principe van Ion Torrent sequencing.



Figuur 2.3 – 7 t/m 11

Het principe van Ion Torrent sequencing.

Tabel 2.2
De voor- en nadelen van Ion Torrent sequencing

Voordelen	*Nadelen*
compact	korte reads
snel	PCR-stap nodig
relatief goedkoop	homopolymeren onnauwkeurig

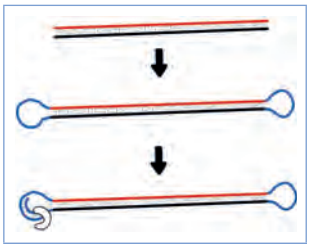
2.3 Single-molecule sequencers

De nieuwe generatie NGS-technieken (long-read sequencers, ook wel third generation sequencers genoemd) onderscheiden zich op een aantal belangrijke punten van de short-read sequencers (paragraaf 2.2).

Ten eerste lezen deze machines direct de moleculen af die worden aangeboden, zonder dat daar een PCR-stap aan vooraf gaat. Daarmee hebben deze technieken geen last van de beperkingen van PCR, zoals GC-percentage en amplificatiefouten. PCR werkt alleen binnen bepaalde grenzen van GC-percentage. Sequenties met een GC-percentage dat daarbuiten valt, zijn met PCR-gebaseerde technieken dus niet af te lezen.
Ten tweede produceren deze technieken vaak (zeer) lange reads. Waar een Illumina sequencer niet verder komt dan 250 basen, zijn read lengtes van tienduizenden basen geen probleem voor single-molecule sequencers. Daarom hoeven de DNA-moleculen niet of nauwelijks gefragmenteerd te worden, voordat ze door één van de long-read technieken gesequencet kunnen worden. Sterker nog, de fragmenten die kunnen worden gesequencet zijn soms zo lang, dat het moeilijk is de DNA-moleculen zo te isoleren dat ze intact blijven.
Daar staat tegenover dat de nieuwe sequencers nog veel fouten maken. Simpelweg omdat het signaal afkomstig is van een enkel molecuul en daardoor moeilijk is te detecteren. Aan oplossingen voor dit probleem wordt hard gewerkt.

2.3.1 PacBio

Het sequencing apparaat van Pacific Biosciences dat in 2011 op de markt kwam, was niet de eerste single-molecule sequencer (dat was de Helicos), maar wel de eerste die een commercieel succes werd. Bovendien was het de eerste techniek waar DNA-fragmenten van meer dan tienduizend basen mee konden worden afgelezen (tabel 2.3).



Figuur 2.4

PacBio SMRTbell library-molecuul. Aan beide kanten van het dubbelstrengs DNA is een lusvormige enkelstrengs adapter gekoppeld.

PacBio library prep

De adapters die aan weerszijden van de DNA-fragmenten worden geplakt, vormen bij PacBio een lus (figuur 2.4). Het 5'-uiteinde van de bovenste streng wordt verbonden met het 3'-uiteinde van de onderste streng en omgekeerd aan de andere kant. Elk fragment in de library is daarmee in feite een enkelstrengse cirkel, een zogenoemde SMRTbell. Hierbij staat SMRT voor single-molecule realtime sequencing.

PacBio sequencing

Voordat een PacBio SMRTbell library gesequencet kan worden, moeten nog de sequencing primer en de polymerase worden gebonden. Deze nemen plaats in de lussen aan weerszijden van een SMRTbell-molecuul (figuur 2.4). Als dat gebeurd is, worden de library-moleculen geladen op een SMRT cell. Een SMRT cell is een plaatje waarvan de bodem bestaat uit gaatjes van 100 nm doorsnede, de zero-mode waveguides. Op de bodem van elk van die gaatjes bevinden zich biotinegroepen die kunnen binden met een streptavidine-modificatie van de polymerase. Op die manier wordt een library-molecuul via de polymerase gebonden aan de

bodem van een gaatje. De bedoeling is dat er één library-molecuul per gaatje wordt gebonden.

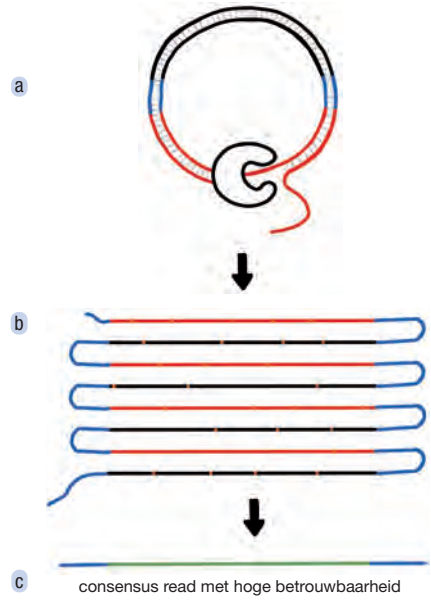
Figuur 2.5

Het principe van PacBio sequencing.

De polymerase kopieert vervolgens het DNA-fragment door een nieuwe streng te synthetiseren (figuur 2.5). Dat het DNA dubbelstrengs is, is niet erg: de DNA-polymerase dwingt het oude complementaire DNA aan de kant, terwijl het langs de streng loopt (strand-displacement; figuur 2.6a). De nucleotiden die de polymerase inbouwt bij het maken van de nieuwe streng zijn voorzien van een fluorescent label aan de fosfaatgroep. In tegenstelling tot de fluorescente labels die Illumina gebruikt, verhinderen de labels van PacBio de polymerase niet om ook de volgende base in te bouwen. Deze gelabelde nucleotiden zweven in de oplossing in de zero-mode waveguide en zijn alle vier tegelijkertijd beschikbaar voor de DNA-polymerase. De bodem van de cell wordt van onderaf belicht door een laser. Het licht van de laser kan maar een klein stukje de gaatjes in waar de polymerase zich heeft gebonden. Op het moment dat er een gelabelde nucleotide wordt ingebouwd, zit deze een fractie van een seconde op zijn plek. Dat is het moment dat een lichtsignaal van het fluorescente label boven de ruis uitstijgt en kan worden gedetecteerd (pulse; figuur 2.5). Onder de oppervlakte van de cell bevindt zich ook een camerasensor die de signalen van elk gaatje filmt. Bouwt de polymerase een A in, dan is er even een groene pulse te zien. Wordt er een G ingebouwd, dan detecteert de sensor een gele pulse, enzovoort. De polymerase is aangepast om het trager te maken, zodat de sensor het proces kan bijhouden. De polymerase gaat hier net zolang mee door totdat het uitgeput raakt, of het filmen stop wordt gezet. Bij sommige polymerasemoleculen duurt dat een hele tijd, waardoor deze zeer lange DNA-fragmenten kunnen aflezen.

Als de polymerase het volledige fragment gekopieerd heeft en nog steeds intact is, komt het bij de adapter aan het andere uiteinde van het fragment. Omdat deze adapter de vorm van een lus heeft, kan de polymerase de bocht om en doorgaan met kopiëren. Ditmaal loopt de polymerase langs de andere streng van het DNA terug naar waar het begon. Vanaf dit moment zal het alleen nog dubbelstrengs DNA tegenkomen waarvan het de eerste helft zelf had ingebouwd. Zolang de polymerase door kan blijven gaan, zal het hetzelfde DNA-fragment meerdere keren aflezen. Elke afgelegd stuk DNA van adapter naar adapter wordt gezien als een subread. Deze subreads bevatten nog steeds willekeurige foutjes. De software legt vervolgens de verschillende subreads van hetzelfde DNA-fragment onder elkaar en berekent een circular consensus sequencing (CCS) (figuur 2.6). Als de verschillende subreads een andere base geven op een bepaalde positie doordat er in één van de subreads een foutje zit, hakt de software de knoop door op basis van de meeste stemmen gelden. Dus als één subread een T zegt, maar drie andere subreads geven

op dezelfde plek een G, dan wordt er een G in de uiteindelijke consensus read geplaatst. Op die manier worden foutjes gecorrigeerd en hebben de uiteindelijke high fidelity (HiFi) reads een zeer lage foutmarge (maximaal 1 foutje op 100 basen of beter).



Figuur 2.6

Circular consensus sequencing met PacBio.
(a) Een polymerase bezig een kopie te maken van een circulair library-molecuul.
(b) De resulterende read waarbij het library-molecuul meerdere keren wordt afgelezen. De oranje punten geven sequencing fouten weer.
(c) Omdat elke positie meerdere keren wordt afgelezen, kan de juiste base voor de consensus sequentie worden bepaald.

Tabel 2.3
De voor- en nadelen van PacBio sequencing

Voordelen	*Nadelen*
lange reads	minder schaalbaar dan sommige andere platforms
hoge kwaliteit (weinig fouten; alleen bij HiFi sequencing)	duur in aanschaf
geen PCR nodig	

2.3.2 Oxford Nanopore

De laatste techniek om DNA af te lezen die hier wordt besproken is die van Oxford Nanopore. Deze techniek gebruikt een fundamenteel andere benadering. Tijdens de sequencing wordt geen nieuwe DNA-streng opgebouwd, maar wordt de sequentie bepaald aan de hand van fysieke eigenschappen van de nucleoti-

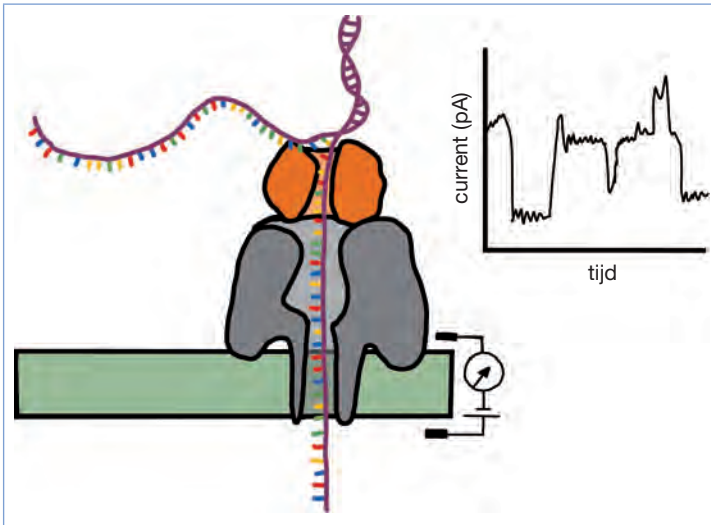
den in het originele molecuul zelf. Geen polymerasen of fluorescent gelabelde nucleotiden, maar kleine poriën in een biologisch membraan waar het DNA-molecuul doorheen wordt geduwd. In de porie wordt de identiteit van de nucleotide bepaald. Het handzame formaat van deze sequencer maakt het mogelijk hem mee het veld in te nemen. Onderzoekers kunnen nu DNA aflezen in tropisch regenwoud of op deinende schepen op de oceaan. Oxford Nanopore technologie is commercieel beschikbaar sinds 2014 en nog volop in ontwikkeling (tabel 2.4). Het belooft allerlei nieuwe toekomstperspectieven, zoals het sequencen van RNA en eiwitten.

Oxford Nanopore library prep

Aan Oxford Nanopore adapters is een complex van drie eiwitten gekoppeld. Eén daarvan is een helicase dat het dubbelstrengs DNA ontwindt, zodat er slechts één streng tegelijk door de porie gaat. Het tweede eiwit is een motoreiwit dat het DNA door de porie duwt en het derde eiwit is een tether dat ervoor zorgt dat het DNA-eiwitcomplex aan de porie bindt. De porie zelf is een transmembraaneiwit, afkomstig uit bijvoorbeeld de bacterie *E. coli*.

Oxford Nanopore sequencing

Over de membraan waar de poriën in liggen, staat een potentiaalverschil. Dit potentiaalverschil zorgt voor een constante ionenstroom door de porie. Terwijl het DNA-molecuul door de porie beweegt, bevinden er zich steeds drie basen tegelijk in de porie. Dit geeft een verstoring van de ionenstroom door de porie

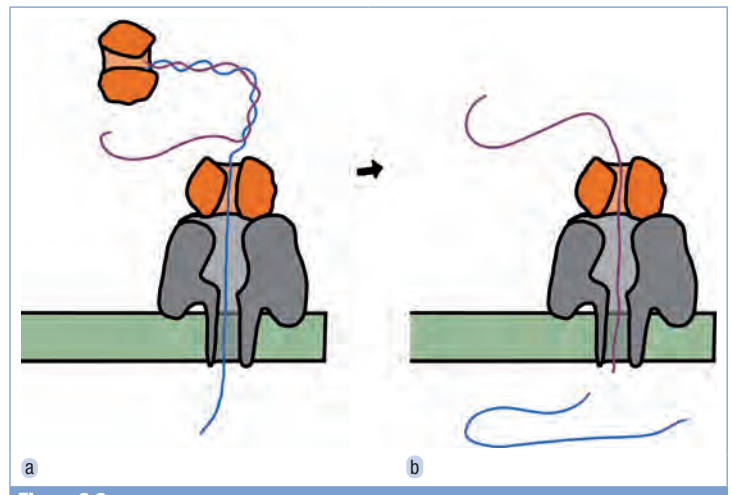


Figuur 2.7

Het principe van Oxford Nanopore sequencing. Het plaatje toont een motoreiwit (oranje) dat een DNA-streng door de nanopore (grijs) duwt. De grafiek toont de afwijkingen in de stroom die dat veroorzaakt.

(figuur 2.7). Het is die verstoring die gemeten wordt. Elke combinatie van drie basen geeft een karakteristieke verstoring. Omdat het DNA-molecuul steeds één base wordt opgeschoven, wordt elke base drie keer gedetecteerd, steeds in combinatie met twee andere basen. Eerst zijn dat de twee basen die voor de bewuste base op het fragment liggen, dan de twee basen aan weerszijden en dan de twee basen erachter. De software kan uit deze serie van verstoringen reconstrueren wat de oorspronkelijke basenvolgorde van het fragment was. Dit gebeurt terwijl het DNA-molecuul door de porie beweegt, zodat de DNA-sequentie al tijdens het sequencen beschikbaar komt.

Als het DNA-fragment in zijn geheel door de porie is gegaan, is de porie vrij voor een tweede molecuul. Het is mogelijk ervoor te zorgen dat dit tweede fragment de complementaire streng is van het fragment dat zojuist door de porie is gegaan. Aan beide uiteinden van het library-molecuul zijn dan motoreiwitten bevestigd (figuur 2.8a). Als de eerste streng door de porie is geduwd door het eerste motoreiwit, blijft het tweede motoreiwit gekoppeld aan het uiteinde van de complementaire streng (die nog niet door de porie is gegaan). Dit tweede motoreiwit duwt vervolgens de complementaire streng door dezelfde porie, zodat ook deze wordt afgelezen (figuur 2.8b). Door de complementaire streng ook af te lezen kan de consensus sequentie worden bepaald en de zekerheid van de eigenlijke sequentie worden verbeterd. Deze techniek staat bekend onder de naam 1D^2 sequencing.



Figuur 2.8

Het principe van 1D^2 sequencing.

De software en gebruiker hebben controle over elke individuele porie en zien de sequentie per porie realtime verschijnen. Het is mogelijk om bij een individuele porie de stroom om te keren, zodat het DNA-molecuul weer uit de porie schiet. Zo kun je

moleculen waarvan je tijdens het sequencen al ziet dat het niet de sequentie heeft waar je naar op zoek bent, weer uit de porie verwijderen. Hierdoor komt de porie direct vrij voor een ander fragment. Dit noem je adaptive sequencing of read until.

Tabel 2.4
De voor- en nadelen van Oxford Nanopore sequencing

Voordelen	*Nadelen*
lange reads	lage capaciteit
geen PCR nodig	veel fouten
compacte apparaten	

Tabel 2.5
Overzicht van de output van de besproken sequencing technieken (stand van zaken 2022 [1])

	Illumina	*Ion Torrent*	*PacBio*	*Oxford Nanopore*
techniek	sequencing by synthesis	sequencing by synthesis	single-molecule realtime sequencing (SMRT)	nanopore sensing
PCR-stap	ja	ja	nee	nee
gemiddelde read lengte	2×300 bp	600 bp	25.000 bp	4 Mbp
data per run	6 Tb	50 Gb	66,5 Gb	14 Tb
max. runtijd	44 uur	12 uur	30 uur	72 uur
fouten	< 1%	1-2%	HiFi < 0,1%	8% $1D^2$ < 1%
kosten per Gb	$ 0,05	$ 4	$ 11	$ 1

* Dit zijn de maximale waarden voor de grootste versies van elke techniek. Van de meeste technieken zijn ook kleine versies beschikbaar voor toepassingen waarbij minder data gegenereerd hoeven worden.

2.4 Oefenvragen en opdrachten

Vraag 2.1 De verschillende NGS-platforms hebben elk hun eigen voor- en nadelen. Bij de keuze van een NGS-techniek voor een specifiek project spelen dus verschillende overwegingen een rol.
Noem drie factoren die meespelen in de keuze voor een NGS-platform.

Vraag 2.2 De NGS-platforms verschillen sterk van elkaar in de manier waarop de sequentie wordt bepaald. Toch hebben ze met elkaar gemeen dat ze vele moleculen tegelijk, in parallel, aflezen.
Leg uit voor elk van de besproken technieken (Illumina, Ion Torrent, PacBio en Oxford Nanopore) hoe dit wordt bereikt.

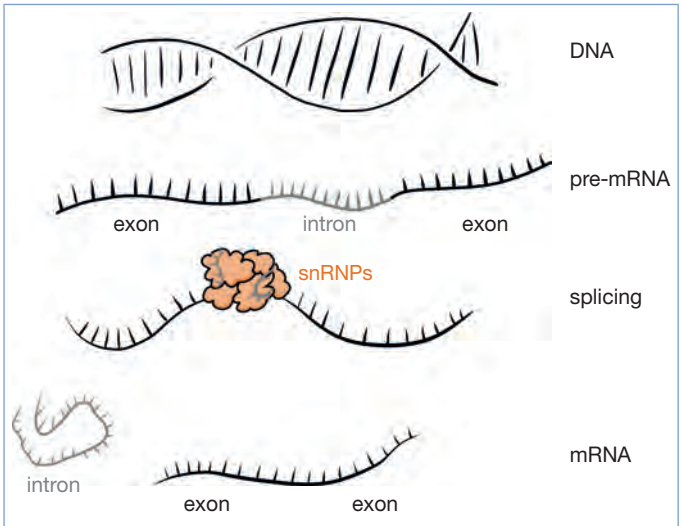
Hoofdstuk 8

Transcriptoom sequencen

Box I – RNA-moleculen

RNA-moleculen verschillen in een aantal opzichten van DNA-moleculen. DNA is een vrij inert molecuul. Het is stabiel, robuust en heeft geen enzymatische activiteit. DNA is dan ook niets anders dan een opslagplaats voor informatie. RNA-moleculen zijn instabieler, maar veelzijdiger dan DNA-moleculen. Ze zorgen ervoor dat de informatie die in het DNA ligt opgeslagen, vertaald wordt naar eiwitmoleculen.

RNA-moleculen bestaan maar uit één streng. Dus één suiker-fosfaatketen met daaraan de nucleobasen. Dubbelstrengs RNA bestaat wel (sommige virussen hebben een genoom dat bestaat uit dubbelstrengs RNA), maar in cellen komt het niet voor. De suikers in de ruggengraat van een RNA-molecuul zijn ribose in plaats van desoxyribose. Verder heeft RNA geen thyminebasen, maar in plaats daarvan uracilbasen.

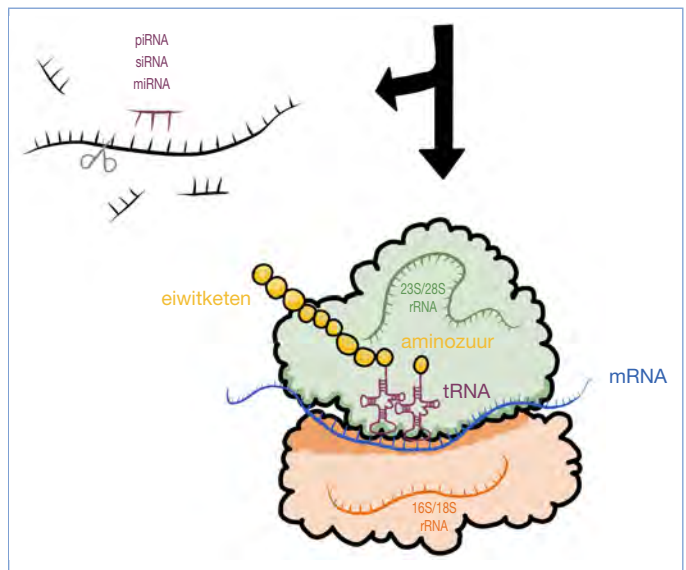


Figuur 8.1

Het proces van transcriptie en splicing. Van een stuk DNA (een gen) wordt een RNA-kopie gemaakt, het pre-mRNA. Hier worden vervolgens de intronen uit verwijderd door middel van splicing.

Er zijn verschillende typen RNA-moleculen. Ten eerste messenger-RNA (mRNA). Dit type RNA wordt geproduceerd door middel van *transcriptie* (figuur 8.1). In de celkern wordt van de eiwit-coderende genen op het DNA een RNA-kopie gemaakt, het zogenoemde pre-mRNA. Een complex van RNA en eiwit (small nuclear ribonucleoprotein, ofwel snRNP) verwijdert vervolgens de intronen uit de pre-mRNA-moleculen. Dit proces heet *splicing*. Het resultaat zijn de mRNA-moleculen.

De mRNA-moleculen verlaten de celkern om in het cytoplasma of het endoplasmatisch reticulum te binden aan *ribosomen*. In de ribosomen worden de instructies op het mRNA vertaald naar eiwit (*translatie*, figuur 8.2). Elke drie basen op het mRNA vormen een *codon*, dat correspondeert met drie complementaire basen, het *anticodon*, op een ander RNA-molecuul, het transfer-RNA (tRNA). Zo'n tRNA is een klein RNA-molecuul waarvan rijen basen met elkaar paren om zo een soort klaverbladstructuur te vormen. Op de middelste lus liggen de drie basen van het anticodon die paren met het codon op het mRNA. Aan het andere uiteinde is het tRNA-molecuul gebonden aan een aminozuur. Welk aminozuur dat is, is specifiek voor elk anticodon. Een serie codons op een rij zal dus paren met een specifieke serie anticodons. Daarmee zal een specifieke rij aminozuren naast elkaar komen te liggen. In het ribosoom worden deze met elkaar verbonden tot een eiwit. Dat ribosoom bestaat zelf voor een deel uit RNA-moleculen. Dit zijn de ribosomale RNAs (rRNA). In tegenstelling tot DNA heeft RNA wel enzymatische activiteit. In het ribosoom katalyseren rRNAs de vorming van eiwitten.



Figuur 8.2

In het ribosoom worden mRNA-moleculen vertaald naar eiwit met behulp van rRNAs en tRNAs. Sommige mRNA-moleculen worden afgebroken met behulp van verschillende typen kleine RNAs, voordat ze het ribosoom bereiken.

Naast mRNA, tRNA en rRNA zijn er nog vele andere, meestal korte, RNA-moleculen die betrokken zijn bij allerlei activiteiten in de cel (tabel 8.1). Sommige typen kleine RNAs (miRNA en siRNA) paren aan mRNAs waar de sequentie complementair is. Dit resulteert in stukjes dubbelstrengs RNA, wat in de cel gelijk wordt afgebroken. Al gemaakte mRNA-moleculen worden op deze manier weer afgebroken, voordat ze naar eiwit vertaald kunnen worden. Op dezelfde manier zorgen andere kleine RNAs (piRNA) ervoor, dat RNA afkomstig van retrotransposons wordt afgebroken. Er zijn ook langere (> 200 bp) RNA-moleculen die niet voor eiwit coderen. Deze moleculen worden long non-coding RNA (lncRNA) genoemd en hebben een belangrijke rol in allerlei cellulaire processen, zoals transcriptie, translatie en metabolisme.

Tabel 8.1
De belangrijkste kleine RNAs

Type RNA	*Afkorting*	*Functie in de cel*
microRNA	miRNA	beïnvloeden van genexpressie door mRNA af te breken na transcriptie
piwi-interacting RNA	piRNA	afbreken van RNA afkomstig van retrotransposons
small interfering RNA	siRNA	functie vergelijkbaar met miRNA
small nuclear RNA	snRNA	van belang bij RNA splicing
small nucleolar RNA	snoRNA	betrokken bij het gebruiksklaar maken van rRNAs

8.1 RNA sequencen

Net als DNA- kunnen RNA-moleculen worden afgelezen met NGS. Omdat RNA minder stabiel is dan DNA, is het sequencen ervan lastiger dan van DNA. Meestal wordt het eerst omgezet naar DNA (copyDNA of complementDNA, ofwel cDNA), voordat het wordt afgelezen. Met PacBio en Oxford Nanopore is het echter ook mogelijk om de RNA-moleculen af te lezen zonder ze eerst naar cDNA om te zetten.

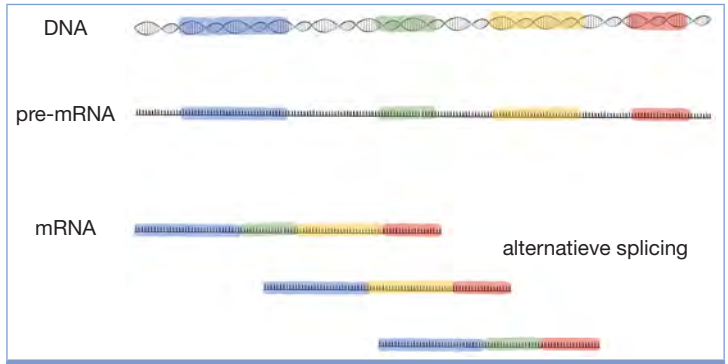
8.2 Mogelijkheden van RNA sequencing

Al de verschillende typen RNA kunnen worden afgelezen met NGS. Er zijn veel verschillende toepassingen en RNA-moleculen komen op verschillende plekken in dit boek aan bod. Allereerst kun je al het mRNA van een organisme sequencen. Dit heet het transcriptoom en is waar dit hoofdstuk over gaat. De meestgebruikte toepassing is het meten van genexpressie. Daaraan is hoofdstuk 9 gewijd. In hoofdstuk 10 komen ook nog de kleine

RNAs aan bod, die onder andere een rol spelen bij het reguleren van genexpressie. rRNA wordt veel gebruikt bij het identificeren van bacteriesoorten, wat in hoofdstuk 11 aan bod komt. Als laatste is RNA sequencing een belangrijk onderdeel van het voorspellen waar de genen liggen op een genoomsequentie. Het proces van geautomatiseerde genoomannotatie is complex en valt buiten de scope van dit boek.

8.2.1 Transcriptstructuur

Tijdens transcriptie wordt in de celkern een RNA-kopie gemaakt van een gen (box I). Voordat deze de celkern verlaat, worden de intronen uit het pre-mRNA gespliced. Het mRNA-molecuul wordt ook voorzien van een poly-A-staart (via proces polyadenylatie) en een 5'-cap, waarmee het stabieler blijft buiten de celkern. Tijdens de splicing worden niet alle exonen in elk mRNA-molecuul ingebouwd (figuur 8.3). Een bepaald exon zit dus wel in het ene transcript, maar niet in het andere. Dit heet alternatieve splicing.



Figuur 8.3

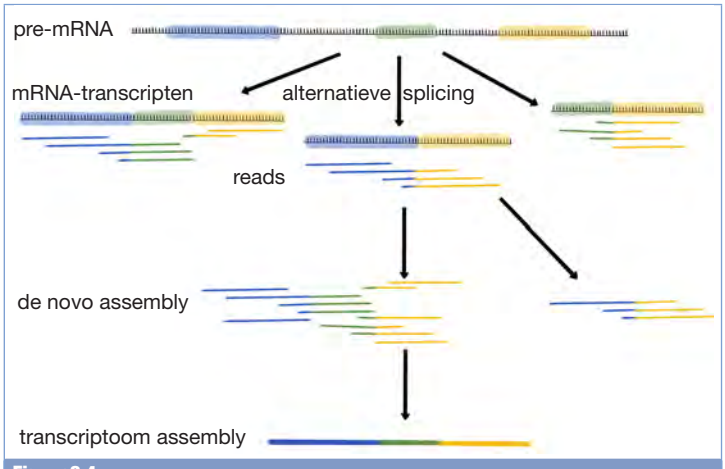
Alternatieve splicing. Het wel of niet inbouwen van verschillende exonen tijdens het maken van mRNA resulteert in verschillende mRNA-moleculen afkomstig van hetzelfde gen.

Alternatieve splicing genereert dus verschillende transcripten van hetzelfde gen. Deze verschillende transcripten kunnen andere functies vervullen in de cel. Het is dus belangrijk te bepalen welke van deze splice varianten voorkomen in de cel. Met korte reads kun je nauwkeurig bepalen welke exonen voorkomen in welke verhoudingen. Maar bepalen uit welke exonen een bepaald transcript is opgebouwd, is met korte reads lastig. Daarvoor heb je lange reads nodig. Sequencing platforms als PacBio en Oxford Nanopore genereren reads die lang genoeg zijn om mRNA-moleculen van begin tot einde af te lezen. Alle onderdelen van een mRNA-molecuul zijn dan aanwezig: de transcriptie startpositie, één of meer exonen en de polyadenylatie positie. Bij PacBio wordt deze techniek Iso-Seq genoemd, omdat het de onderzoekers in staat stelt verschillende mRNA-isovormen in kaart te brengen.

Het sequencen van complete mRNA-moleculen levert interessante informatie. Zo blijken sommige exonen alleen voor te komen in transcripten die ook een bepaalde poly-A-site gebruiken. Of exon X komt alleen voor in transcripten waar exon Y afwezig is. Ook blijken verschillende celpopulaties te verschillen in welke mRNA-isovormen ze bevatten. Aan de biologische betekenis van al deze variatie wordt nog veel onderzoek gedaan.

8.2.2 De novo assembly

Voor een organisme waarvan geen genoomsequentie beschikbaar is, kan het transcriptoom een relatief snelle en goedkope manier zijn om veel informatie op genniveau te verkrijgen. Het transcriptoom is de verzameling van alle mRNA-sequenties die in een organisme voorkomen. Om de juiste sequentie van alle transcripten te bepalen, moeten de reads geassembleerd worden. Zoals dat ook gaat bij het bepalen van het genoom vanuit DNA reads (hoofdstuk 6). In het resultaat is elk transcript een contig (figuur 8.4). Een geassembleerd transcriptoom zal dus uit veel meer, en veel kortere contigs bestaan dan de assembly van een genoom. Met long-read sequencing technieken kunnen complete transcripten in een read worden afgelezen (paragraaf 8.2.1). In dat geval is de novo assembly niet veel meer dan het verwijderen van sequencing fouten.



Figuur 8.4

Transcriptoom assembly.

8.3 Technieken voor RNA sequencing

8.3.1 RNA-isolatie en library prep

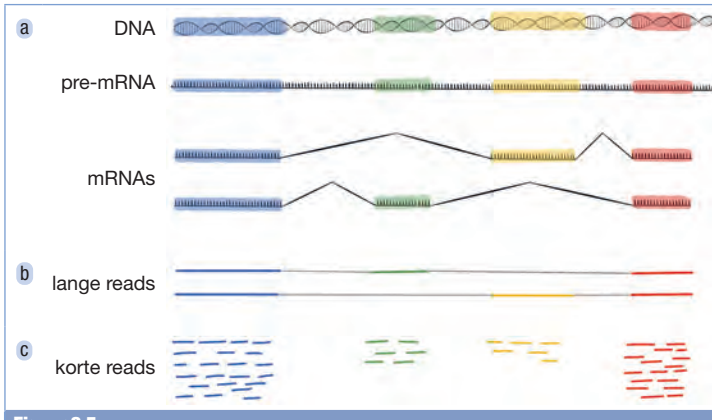
Wanneer RNA geïsoleerd is uit een biologisch monster, zitten alle typen RNA door elkaar. Meestal zijn onderzoekers specifiek geïnteresseerd in een bepaald type. Voorkomen moet worden

dat grote aantallen sequencing reads verspild worden aan het sequencen van het verkeerde type RNA. Daarom wordt het RNA-extract meestal verrijkt voor het gezochte type. Omdat het produceren van eiwitten een van de belangrijkste activiteiten van een cel is, bevat elke cel grote aantallen ribosomen. En daarmee grote hoeveelheden rRNA. Als je dus simpelweg het RNA dat je uit een cel haalt zou gaan aflezen, lees je bijna alleen maar rRNA af. Meestal ben je juist in mRNA geïnteresseerd, maar dat is slechts 1-4% van al het RNA in de cel. Je moet dus iets ondernemen om van al dat rRNA af te komen. Voor het verrijken van mRNA kun je bijvoorbeeld selecteren op fragmenten met een poly-A-staart, want rRNA heeft die niet. Voor het sequencen van kleine RNAs kun je selecteren op RNA-moleculen van de gezochte lengte.

Nadat het RNA is geïsoleerd en verrijkt voor het juiste type, kan er een sequencing library van worden gemaakt. Meestal wordt van het RNA eerst een cDNA-kopie gemaakt door middel van het enzym reverse transcriptase. Dit cDNA wordt vervolgens gesequencet.

8.3.2 Transcriptstructuur

Met long-read sequencing kan een heel transcript met een read worden afgelezen. Door deze Iso-Seq reads te alignen tegen het referentiegenoom, zie je direct welke exonen wel en welke niet zijn meegenomen in elk transcript. Figuur 8.5a toont bijvoorbeeld een denkbeeldig gen waarvan twee transcripten worden geproduceerd. De twee transcripten verschillen in de exonen die ze bevatten. Lange reads geven deze sequenties in één keer (figuur 8.5b).



Figuur 8.5

Het bepalen van transcriptstructuur met lange en korte reads.
(a) Een gen waarvan twee transcripten worden geproduceerd die verschillen in welke exonen (de blokken) worden meegenomen. De dunne streepjes zijn de intronen die uit het RNA worden verwijderd.
(b) Lange reads geven de sequentie van elk transcript.
(c) Korte reads bestrijken elk maar een klein stukje van het transcript.

8.3.3 Transcriptoom assembly

Korte reads zijn niet lang genoeg om complete transcripten in hun geheel te vangen (figuur 8.5c). Om de sequentie van transcripten te bepalen, moeten de korte reads geassembleerd worden op basis van onderlinge overlap. Hiermee kunnen de exonen nauwkeurig worden bepaald. Omdat de reads van de verschillende transcripten door elkaar zitten in de library, is het bepalen van de juiste transcriptstructuur veel lastiger. Hiervoor is extra informatie nodig die toont welke exonen aan elkaar gekoppeld zijn.

Het voordeel van korte reads is dat het veel makkelijker is om er heel veel van te produceren. Met korte reads kun je dus een hogere coverage behalen dan met lange reads. Dat betekent dat het met korte reads makkelijker is om zeldzamere transcripten op te pikken. Voor het bepalen van een transcriptoom kan dus toch vaak voor korte reads worden gekozen. Ondanks dat de verschillende RNA-isovormen dan moeilijker zijn te bepalen.

Het de novo assembleren van korte RNA-reads gebeurt op dezelfde manier als bij DNA reads (hoofdstuk 6). Hierbij moet de software wel rekening houden met een aantal belangrijke verschillen tussen transcriptoom- en genoomdata, namelijk:

- Bij transcriptoomdata varieert de coverage sterk tussen transcripten. Dit komt doordat genen sterk verschillen in de mate waarin ze tot expressie komen. Van een gen dat sterk tot expressie komt, zullen veel meer reads beschikbaar zijn dan van genen die maar weinig worden afgeschreven. Bij DNA-data is dit niet zo en verwacht je een min of meer constante coverage over het genoom.
- De coverage varieert zelfs binnen transcripten. Sommige exonen komen in alle splice varianten voor, terwijl andere veel zeldzamer zijn. Die laatste zullen dus lagere coverage hebben. Ook kunnen er technische artefacten zijn, bijvoorbeeld omdat is geselecteerd op poly-A-staarten en de 3'-kant van het transcript daarom vaker wordt gesequencet.
- Transcripten van genen die dicht bij elkaar liggen, kunnen overlappen en leiden tot samengestelde contigs.
- Het proces van alternatieve splicing zorgt ervoor dat transcripten deels hetzelfde zijn, maar niet helemaal.

Net als de gangbare genoomassemblers voor korte reads maken RNA-assemblers gebruik van De Bruijn grafen. Een genoomassembler zoekt naar zo lang mogelijk grafen, die idealiter hele chromosomen representeren. Transcriptoomassemblers delen de data op in grote aantallen korte grafen. Elke graaf wordt dan geanalyseerd om zo volledige splice varianten te karakteriseren.

8.4 Uitgewerkt voorbeeld

Winterslaap zoals die van bruine beren zorgt voor een dramatische omslag in fysiologie. Beren in winterslaap verlagen hun hartslag en worden insulineresistent. Bij ons zou dat tot ernstige gezondheidsproblemen leiden, maar beren blijven niet alleen gezond, ze behouden zelfs hun spiersterkte. Kennis over hoe ze dat klaarspelen kan aanknopingspunten geven voor de behandeling van bijvoorbeeld obesitas en diabetes type 2 bij mensen. Tseng et al. 2022 [1] gebruikten lange PacBio reads om RNA uit spierweefsel van drie beren te sequencen tijdens de winterslaap en tijdens de actieve periode. Daarmee konden ze in detail splice varianten karakteriseren van ongeveer 13.000 genen. Meer dan de helft van de splice varianten in deze data waren nog niet bekend. Vervolgens werd RNA van nog eens zes beren gesequencet met korte reads. Hiermee konden de onderzoekers kwantificeren welke van de splice varianten het meest werden gebruikt in elk van de seizoenen. Met name in vetweefsel bleken andere splice varianten tot expressie te komen tijdens de winterslaap dan tijdens de actieve periode. Van het gen ITGB3BP komt tijdens de winterslaap een splice variant tot expressie die een exon aan het einde van het gen mist. Dit zal resulteren in een korter eiwit dan het eiwit dat tijdens de actieve periode wordt geproduceerd. ITGB3BP komt ook bij andere zoogdieren voor, inclusief de mens, en is betrokken bij de regulatie van verschillende hormoonreceptoren. De alternatieve splicing van dit gen lijkt dus grote gevolgen te hebben voor de hormoonregulatie bij beren in winterslaap.

8.5 Veelgebruikte tools

- PacBio heeft een set aan software tools ontwikkeld voor de analyse van long-read RNA-data, genaamd Iso-Seq.
- Trinity [2] is een veelgebruikte assembler voor RNA-data.

8.6 Oefenvragen en opdrachten

Vraag 8.1 Korte reads zijn niet lang genoeg om hele transcripten met een read te dekken.
Bedenk hoe je met korte reads toch meer te weten kunt komen over de structuur van transcripten. Met andere woorden: hoe kunnen korte reads je iets vertellen over welke exonen elkaar opvolgen binnen een mRNA-molecuul?

Vraag 8.2 Het genoom van een denkbeeldig organisme bestaat uit één chromosoom met daarop 25 genen. Elk van de genen heeft twee splice varianten.

a. Uit hoeveel contigs zou een succesvolle transcriptoom assembly bestaan?
b. En uit hoeveel contigs een genoom assembly?