Genomics applications of nanopore long-read sequencing for small to large sized genomes

Liem, M.

— Addendum

Nederlandse samenvatting - Dutch summary

Introductie

In dit proefschrift focus ik op de toepassingen van Oxford Nanopore Technologies (ONT) sequencing. Deze techniek is een relatief nieuwe benadering in het sequencing-veld, waarbij nanoporiën zijn ingebed in een membraan, DNA-moleculen door nanoporiën worden getrokken en een elektrische stroom dient als het sequencing-signaal. Deze techniek levert sequenties ("reads") van >10Kbp op en heeft theoretisch geen bovengrens voor de lengte van reads. De positieve impact op de datakwaliteit als gevolg van verbeterde chemie is uitgelicht, verbeterende chemie leidt tot minder sequentiefouten en een meer homogene verdeling van reads over complexe genomische architecturen. De voordelen van langere read lengtes zijn beoordeeld voor het oplossen van genoomassemblages die gefragmenteerd blijven met gebruik van uitsluitend korte-read-sequentiedata. Vervolgens is de assemblage van een groot genoom met ONT-data beschreven, wat laat zien dat ONT een geschikte kandidaat is voor het oplossen van extreem grote genomen met geavanceerde assemblagesoftware. En tot slot komt het potentieel van ONT sequencing naar voren voor in-het-veld sequencing, waarbij gebruik wordt gemaakt van de eenvoud, mobiliteit en de datakwaliteit die worden geboden door deze nieuwe techniek.

De centrale hypothese van dit proefschrift is dat Oxford Nanopore Technologies data waardevol kunnen zijn voor gevestigde genomics toepassingen, zoals volledige genoom sequencing (hoofdstukken 2–4) en het karakteriseren van metagenomen voor microbiële gemeenschappen (hoofdstuk 5). Hier evalueer ik deze algemene stelling in het kader van de beschreven resultaten van de voorgaande hoofdstukken. Daarnaast bespreek ik de vooruitzichten voor opkomende en toekomstige genomics-toepassingen op basis van de mogelijkheden die worden geboden door ONT data.

De kwaliteit van long-read sequencing en assemblages

ONT-sequencing verschilt van traditionele sequencing-methoden doordat nucleotiden rechtstreeks worden gemeten met behulp van elektrische signalen in plaats van synthetische kopieën of markers zoals fluorescerende labels. Meerdere nucleotiden (5-mers) bezetten tegelijkertijd een porie, daarom is het de set van nucleotiden die de elektrische interferentie veroorzaken. Dit profielsignaal moet via algoritmes worden ontrafeld om een enkele base te identificeren. Het zijn dus de algoritmes die de uiteindelijke reads aanleveren en door deze algoritmes te verbeteren kan de kwaliteit van sequentie data zelfs verbeteren voor eerder geanalyseerde projecten[1]. ONT liet aanvankelijk 30 nucleotiden per seconde door de nanoporie passeren. De snelheid waarmee het aantal nucleotiden door een enkele porie werden gehaald werd gelimiteerd omdat algoritmes moeite hadden om nucleotiden te onderscheiden uit een set van nucleotiden die tegelijk de porie bezetten als deze te snel door de porie bewegen. Dit resulteerde in een bijzonder lage sequentiekwaliteit. Het beperken van de snelheid tot 30 basen per seconde leverde een nauwkeurigheid van ~70% op. Momenteel kan ONT ~450 basen per seconde doorlaten, wat reads oplevert van >10Kbp met een nauwkeurigheid tussen ~90-99%. Hoofdstuk 2 benadrukt het effect van verbeterde sequentiesnelheid, verbeterde algoritmes en chemie voor een zeer heterogene giststam.

Om echter nauwkeurige haplotypes voor dit genoom te genereren, is extra sequentiekwaliteit vereist. Uit een BUSCO-analyse bleek dat uit onze beste assemblageresultaten er nog steeds genen ongeïdentificeerd bleven. De impact van deze sequencing fouten wordt benadrukt door de vergelijking van geïdentificeerde genen vóór en na foutcorrectie. Waarbij meer genen worden geïdentificeerd wanneer de sequentienauwkeurigheid wordt verhoogd. Het volume van datasets word voor sequencing uitgedrukt in het aantal kopieën van het genoom ("coverage"). Vergeleken met andere studies, die coverage gebruiken variërend van 70x tot 1000x, heeft onze dataset relatief lage coverage. Daarom kan een toename van data helpen bij het oplossen van eventuele resterende assemblageproblemen, evenals het verhogen van de sequentiekwaliteit door meer bewijs te leveren voor de foutcorrectieprocedure[2-4].

Aangezien assemblagealgoritmen moeite hebben om de uiteinden van circulair DNA te definiëren is het assembleren van circulaire constructen voor complexe genomen een uitdagende taak. In deze studie hebben we de architectuur van mitochondriaal DNA of circulaire plasmiden niet onderzocht. Een logische volgende stap zou dus zijn om de assemblageresultaten te onderwerpen aan software dat specifiek is ontworpen voor het sluiten van circulaire contigs voor long-read data[2,5].

In hoofdstuk 2 en 3 hebben we een veelvoud aan assemblage-, consensus en correctietools geëvalueerd, die variëren van middelmatig tot veelbelovend. De meeste assemblagestrategieën zijn vergelijkbaar en resulteren in relatief kleine verschillen. Het benadrukken van de oorsprong van die kleine discrepanties en het beslissen over de uiteindelijke assemblage is een tijdrovende en arbeidsintensieve aangelegenheid. De momenteel beschikbare tools bieden ruimte voor verbetering van gebruiksvriendelijke workflows, inclusief visualisaties van base niveau tot genoomwijd. Deze workflows zouden voortgang moeten rapporteren op het niveau van alignment, assemblage, consensus en correctie om besluitvorming voor downstream analyse te faciliteren. Kant-en-klare assemblageworkflows zouden de snelheid waarmee genoomanalyse wordt uitgevoerd verhogen en onderzoeken verlichten voor grote sequentie-datasets. De huidige standaard is het uitvoeren van meerdere assemblagestrategieën en doorgaan op een resultaatgerichte manier. Analysetools voor genomen van klein tot middelgroot tonen vergelijkbare maar niet identieke assemblageresultaten waardoor vergelijking tussen analyses uiterst moeilijk is[7,9].

De novo assemblageresultaten op basis van long-read data voor kleine genomen laten veelbelovende reconstructies zien. Assemblages voor medium-grote genoomgroottes van vergelijkbare kwaliteit, zoals onderzocht in hoofdstuk 4, zijn steeds vaker openbaar beschikbaar. Echter, afzonderlijke haplotypen van dergelijke organismen moeten nog worden gepubliceerd, deze inhaalslag word nu pas gemaakt omdat de kwaliteit pas recentelijk van voldoende kwaliteit is geworden om chromosomale kopieën nauwkeurig te faseren[10].

Ondanks een verhoogde capaciteit om een gelijkmatige coverage te bereiken, langere reads te genereren en verbetering naar low-complexity regio's, zijn voor ultra-grote genomen aanvullende ontwikkeling vereist[6]. Het routinematig sequencen van ultra-grote genomen vereist een aanvullende ontwikkelingsupdate die met name gericht op de snelheid van sequencen en de kosten. Bijvoorbeeld, het sequencen van het genoom van *Paris japonica*, een plantensoort met een genoomgrootte van ongekende omvang, geschatte genoomgrootte ~150 Gbp voor een enkel genoomkopie[8]. Het sequencen van een genoom van deze omvang duurt iets minder dan een uur op een volledig geladen PromethION (dat wil zeggen 48 flowcellen, elk ~$2.000 en gebruikmakend van ~ 2.500 poriën bij 450 basen per seconde) voor een enkele genoomkopie. Daarom, hoewel haalbaar, duurt het sequencen op de vereiste sequentiediepte voor dergelijke genomen nog steeds dagen en is het zeer duur. Voor wat betreft de verbeteringen van read-lengte, read-kwaliteit en schaalbaarheid is ONT een pionier die het onderzoeksgebied van echt grote genomen mogelijk maakt[6].

De kosten van genoomsequencing

Het evalueren van de kosten van genoomsequencing met behulp van de Wet van Moore heeft duidelijk gemaakt dat ongelooflijke hoeveelheden sequentiegegevens worden en zullen worden gegenereerd. Deze datavolumes geven de noodzaak aan van efficiënte software voor downstream analyse. Momenteel is sequencing-data betaalbaarder geworden in tegenstelling tot de kosten voor het analyseren van grote datasets met behulp van computerclusters.
Het voordeel van verminderde kosten, verhoogde sequencing-snelheid en schaalbaarheid gaat verloren wanneer gegevensanalyse duizenden CPU-uren vereist op dure toegewijde clusters.
We moeten daarom de wetenschappelijke gemeenschap voorzien van meer geavanceerde tools voor het verwerken van grote datasets, die minder rekenintensief zijn, minder geheugen vereisen, sneller zijn en gebruiksvriendelijker zijn.

Alles en overal sequencen

Standaard laboratoriumtechnici hebben geen ervaring met commandline tools en beschikken niet over de vaardigheden om zich adequaat aan te passen aan alternatieve resultaten.
Dit duidelijk aanwezige hiaat kan worden overbrugd door gestandaardiseerde eenheden en formaten te gebruiken, gemakkelijk toegankelijke, gratis maar geavanceerde software die wordt ondersteund met logische visuele representaties.

Voor het idee alles overal kunnen sequencen is de omvang van sequencingmachines belangrijk, momenteel is het kleinste sequencingapparaat slechts zo groot als een grote USB-stick en biedt mobiliteit om sequencing in het veld mogelijk te maken, dit wordt besproken in hoofdstuk 5. Echter, veld gegenereerde gegevens moeten worden verwerkt door computerclusters of op zijn minst een high-end laptop met voldoende energievoorziening. Het volledig benutten van dit mobiliteitskenmerk vereist afgeschaalde verwerkingskracht, geheugen- en energieverbruik.

Van amplicon tot *in situ* metagenoomsequencing en assemblage

In hoofdstuk 5 hebben we metagenomics gebruikt om de microbiële diversiteit te identificeren met behulp van ONT, wat een eerste stap is in het begrijpen van de biocomplexiteit en ecologie van de grote wateren. Echter, het bepalen welke soorten gedijen op welke locaties is slechts het begin van het begrijpen van de ecologie achter de microbiële diversiteit. Om deze diversiteit functioneel te beoordelen zijn volledige genoomassemblages nodig. Deze kennis kan bijvoorbeeld leiden tot een beter begrip van de resistentiemechanismen die door microbiële gemeenschappen worden gebruikt om de harde oceaansomstandigheden te overleven of om de mechanistische eigenschap te onthullen voor het uitwisselen van genetisch materiaal via plasmiden.
Bovendien zou het de diversificatie van soorten op een tijd- en ruimtelijke manier kunnen ontrafelen waardoor de gezondheid van oceanen, zeeën en rivieren die de basis van het leven op het land vormen, kan worden gevolgd. Om in-field monitoring van zeewater adequaat toe te passen moeten DNA-isolatie- en laboratoriummethoden ter plaatse worden uitgevoerd.
In hoofdstuk 5 hebben we het DNA onder laboratoriumomstandigheden geïsoleerd. Hoewel deze procedure een zeer eenvoudige richtlijn volgt, is het verzamelen van lang moleculair DNA van mariene organismen bijzonder uitdagend vanwege overmatige afscheiding van metabolieten die co-precipiteren met DNA[11]. Daarom moet optimalisatie voor isolatie van lang moleculair DNA met betrekking tot sequencing op locatie verder worden ontwikkelt zowel wat betreft sequencingsnelheid alswel wat betreft het gebruiksgemak. Apparatuur voor het voorbereidingen van DNA voor sequencing moet voldoen aan de gewenste eisen om *in situ* te kunnen worden ingezet. Voltrax laboratorium voorbereiding biedt een potentieel oplossing en is in staat om geïsoleerd DNA in een kwestie van minuten klaar te maken, echter, als gevolg van het gebrek aan zuiveringsstappen, zou geïsoleerd hoogmoleculair DNA nogal verontreinigd kunnen zijn.
Zelfs met kleine en gebruiksvriendelijke apparaten zoals Voltrax blijft *in situ* DNA-isolatie en -zuivering uitdagend[11]. Bovendien vereist de chemie die nodig is voor sequencing specifieke opslagbeperkingen; zowel flowcellen als chemie zijn temperatuurgevoelig en de koelkast-capaciteit voor veldexpedities is meestal onregelmatig vanwege het gebrek aan adequate stroomvoorziening[12]. Ten slotte is aanvullende analyse vereist om geïdentificeerde soorten fylogenetisch te positioneren. Onecodex (gebruikt in hoofdstuk 5) is gunstig om organismen snel en gemakkelijk in de context van bestaande databases te plaatsen, dit scheelt tijd en verlicht de arbeidscomplexiteit.
Aan het analyse portaal dat Onecodex biedt ontbreekt echter de phylogenetische afstand tussen soorten, welke naar boven gehaald kan worden door een tijdrovende methode zoals multiple sequence alignment. Bovendien biedt het alleen uitgebreide functionaliteit met een betaalde licenties waardoor kosten toenemen en het voor onderzoekers moeilijk maakt om resultaten te vergelijken. Eerdere studies tonen succesvolle fylogenetische plaatsing onder afgelegen omstandigheden met behulp van JModelTest, daarom zou dit een potentieel kandidaat kunnen zijn voor downstream-analyse van metagenoommonsters uit zeewater[13].

De toekomst van Oxford Nanopore Technologies-sequencing en de toepassingen

Met het gebruik van de huidige beste flowcellen en chemie worden kwaliteiten van Q20 bereikt, wat zich vertaalt naar >99% nauwkeurigheid. Deze methoden maken het mogelijk om van de moleculen die door de nanoporie worden gehaald de basenvolgorde uit te lezen. Hoewel het sequencen van beide gescheiden enkelstrengs DNA al in ~2015 door Oxford Nanopore Technologies werd geïntroduceerd, werd het later vervangen door chemie van slechts één kopie uitleest. Echter, chemieën om beide gescheiden enkelstrengs DNA te sequencen zijn recentelijk opnieuw uitgebracht door Oxford Nanopore Technologies. Hier wordt de informatie van beide enkelstrengs DNA gebruikt om basecalling-fouten te verminderen door de sequentie-signalen te combineren. Zodra het dubbelstrengsmolecuul zijn weg naar de porie heeft gevonden, wordt één van de twee strengen door de porie getrokken, deze streng wordt de templatestreng genoemd. Vervolgens laat na ontvouwen van het dubbelstrengs-DNA het 5'-eind van de complementaire streng in de nabijheid van de porie achter met behulp van een bevestigings-molecuul dat aan het membraan is bevestigd. Naarmate de sequencing het einde van het molecuul bereikt, volgt met enige waarschijnlijkheid de complementaire streng onmiddellijk de templatestreng door dezelfde porie. Vanuit de sequencing signalen worden reads die na elkaar overgaan met vergelijkbare sequentielengtes en complementaire base-samenstelling gedetecteerd als paren, aangeduid als een duplexpaar.

Eerdere basecalling-methoden gebruiken ofwel signalen van enkelstrengs DNA of gecombineerde signalen van zowel template- als complementaire strengen, 'paired decoding' genoemd. Enerzijds is simplex basecalling (het verwerken van het signaal van een enkele streng individueel) zeer snel maar levert hogere foutpercentages op. Anderzijds, het voeden van beide strengen aan een neuraal netwerk basecalling-algoritme levert nauwkeurige sequenties op ten koste van middelen en tijd. Het decoderen van gecombineerde signalen is een rekenkracht intensief proces, tot wel vijf keer trager vergeleken met simplex basecalling en ontbreekt daardoor aan schaalbaarheid[14]. De noviteit van de kwaliteitsverbetering voor 'stereo duplex basecalling' vindt zijn oorsprong door base informatie, kwaliteitsscores en het sequentiesignaal voor zowel de template- als complementaire streng te voeden aan een 'stereo' basecaller. Deze basecalling-methode is eenvoudig, snel en robuust en maakt betere schaalbaarheid mogelijk om grote hoeveelheden gegevens te genereren over een redelijke tijdsperiode, welke Q30 kwaliteiten kan genereren. Met kwaliteit die de standaard sequencing-platforms benadert, lijkt Oxford Nanopore-technologie een veelbelovende techniek voor analyse die een hoge nauwkeurigheid op base niveau vereisen, zoals SNP-detectie en haplotype-identificatie, met name voor polyploïde genomen.

Hoewel we een overtreffing van de Wet van Moore (Figuur 9 – inleiding) zien wat betreft de kosten van sequencing in het algemeen, blijft long-read sequencing relatief duur. Onder meer kostenefficiënte omstandigheden is long-read sequencing ook een geschikte kandidaat voor functionele genomics-analyse. Het vermogen om samples voor te bereiden zonder amplificatie voorkomt de introductie van biases waarbij sommige moleculen ondervertegenwoordigd zijn en andere overmatig worden versterkt. Zonder deze biases kan nauwkeurige kwantificering mogelijk worden gemaakt. Long-read-sequenties kunnen volledige transcripten in één keer beslaan, waardoor ingewikkelde transcript-assemblages worden vermeden en vereenvoudigde identificatie mogelijk is. Hierdoor is er minder data nodig om hetzelfde aantal genen te identificeren in vergelijking met methoden voor short-read sequencing[15].
Bovendien zijn volledige transcripten die direct worden geregistreerd, uitzonderlijk waardevol voor de karakterisering van structurele variatie zoals isoformen. Isoformen kunnen verschillende functionele eigenschappen en expressieniveaus vertonen, en ze zijn uiterst moeilijk te bepalen met behulp van short-read sequencing. Bovendien wordt structurele variatie gebruikt over een breed spectrum van onderzoeksgebieden die lopen van het begrijpen van kankers in een klinische setting tot aan het coderen van commercieel aantrekkelijke eigenschappen voor de agrarische sector. Structurele variatie strekt zich in veel gevallen uit over Mbp-stukken in het genoom en is onmogelijk vast te leggen met een enkele read vanuit traditionele sequencing-technieken. Daarom worden die regio's, met traditionele data, sequentieel in stukjes gelezen en opnieuw samengesteld om de volledige structurele variatie te onthullen. Voor de standaard sequencing technieken leidt dit tot misassemblages en het ontbreken van regio's die vatbaar zijn voor amplificatie-biases. Bovendien, omdat long-reads een verhoogde aligneringspecificiteit bieden, wordt het aantal onduidelijke alignments aanzienlijk verminderd, in vergelijking met short-read sequencing data.

En tot slot, dankzij de gevoeligheid van sequentiesignalen en ontwikkelingen in kunstmatige intelligentie, kan nanopore-sequencing gemodificeerde basen detecteren. Het epigenoom is een ingewikkeld raamwerk bestaande uit een veelheid van chemische verbindingen die de functionaliteit van DNA dicteren. De hoog over structuur die de genomische functie orkestreert, omvat onder andere CpG-methylatie, nucleosoombezetting, chromatine-toegankelijkheid, histonmodificaties en proteïnebindende gebeurtenissen die helpen bij de juiste segregatie van chromosomen[16,17]. Het meest bekende epigenetische component is CpG-methylatie en is geassocieerd met het onderdrukken van gen-transcriptie onder hypergemethyleerde promotoromstandigheden of transcriptieactivering voor hypo- en hypermethylering van het promotorgebied en een gen, respectievelijk. Een standaard methode om methylering te detecteren is whole genome bisulfite sequencing, waarbij ongemethyleerde cytosines worden vervangen, eerst met uracil en later door thymine nucleotiden, waardoor de methylerings-fingerprint wordt onthuld. Deze methode vereist echter ingewikkelde bisulfietconversiestappen, amplificatie en levert short-read data op. Daarom is deze strategie met name moeilijk toe te passen voor regio's met een lage complexiteit zoals GC-eilanden. Oxford Nanopore Technologies methyleringsidentificatie heeft aangetoond vergelijkbare nauwkeurigheid te behalen in vergelijking met standaard methoden. Bovendien bieden ze het voordeel van langere reads en de afwezigheid van amplificatie, wat betere alignments mogelijk maakt voor regio's met een lage complexiteit, het vermijd ingewikkelde laboratoriumprocedures, en heeft alleen het sequencing signaal nodig en een basecalling-algoritme[18].

Toepassingen voor functionele genomics en epigenetics hebben hun waarde bewezen voor specifieke wetenschappelijke knelpunten en hebben kennislacunes overbrugd van gebieden die onaangeroerd zijn gebleven door traditionele technologieën. Het huidige kostenperspectief maakt Oxford Nanopore Technologies specifiek aantrekkelijk voor gespecialiseerde gevallen, of dat nu is om genen te identificeren die omringd zijn door repetitieve sequenties, splice-varianten met repetitieve inhoud te kwantificeren, methyleringsfingerprints over lange reeksen epigenetische elementen te genereren of assemblage fragmenten te sluiten voor grote en complexe genomen. Wanneer Oxford Nanopore Technologies een kosteneffectieve verhouding bereikt die vergelijkbaar is met standaard methoden, zal het zijn ware potentieel vinden en zal het een nieuw tijdperk openen voor gestandaardiseerd sequencen, waardoor de analyse van "alles door iedereen, overal" mogelijk wordt.

Moedig gaan waar niemand ooit geweest is

Zoals gesuggereerd wordt door de verbeteringen in read lengtes, wordt het realistischer om te hypothetiseren dat toekomstige sequencing zal transformeren van een methode voor het uitlezen van fragmenten naar een telomeer-tot-telomeer-sequencing-mode. Momenteel zijn de maximale leeslengtes die worden gerapporteerd >4 Mbp, in vergelijking met >10 Kbp in 2010, wat aangeeft dat het niet lang zal duren voordat telomeer-tot-telomeer-sequencing de standaard is. Het sequencen van hele chromosomen zou aanzienlijke voordelen met zich meebrengen in vergelijking met huidige sequencing-technologieën, omdat het de assemblage voor hele-genoom-sequencing volledig buitenspel zet. Het verkleinen van de computationele druk zal de wetenschappelijke gemeenschap verlichten van rekenintensieve downstream-analyses en zal wetenschappers bevrijden van toegewijde computerclusters en commandline software.

Bovendien is de sequencing snelheid gebaseerd op het aantal nucleotiden dat door de nanopore passeert, om de nauwkeurigheid te beschermen zijn de snelheden momenteel beperkt tot 450 nucleotiden per seconde. Deze snelheid maakt het mogelijk voor moderne deep learning algoritmen om de basenvolgorde met een nauwkeurigheid tot Q30 te bepalen. Het verhogen van de sequentiesnelheid met behulp van die basecalling-modellen zou echter leiden tot een vermindering van de nauwkeurigheid omdat sequentiesignalen te moeilijk worden om te achterhalen. Desalniettemin zouden verbeteringen in deep learning, resulterend in meer geavanceerde neurale netwerk basecallers, de sequentiesnelheid kunnen verhogen tot een theoretisch maximum van $>10^6$ nucleotiden per seconde[19]. Het benutten van de maximale sequencing snelheid zou een enkel kopie van het menselijk genoom in iets minder dan twee uur kunnen worden uitgelezen met behulp van een enkele porie. Een dergelijke verminderde computationele druk en verhoogde sequentiesnelheid zullen de analyse van DNA-inhoud van elk organisme op een middelmatige laptop in een kwestie van minuten mogelijk maken, in plaats van dagen met behulp van toegewijde en dure computerclusters.

Bovendien zouden gestandaardiseerde analyse-werkbanken moeten helpen om de tijdbeperkingen nog verder te verminderen, waardoor wetenschappers snel en gemakkelijk door de data kunnen navigeren op een uitgebreide, gebruiksvriendelijke en visueel aantrekkelijke manier. Hoewel read lengtes de lengtes van chromosomen benaderen, moet er aanvullende vooruitgang worden geboekt met laboratoriumtechnieken om, onder andere, verstrengeling of breken van dergelijke lange moleculen tijdens het isoleren en ontvouwen van het dubbelstrengs-DNA-molecuul te vermijden.

Een andere potentiële toepassing voor toekomstige Oxford Nanopore Technologies die cellysis omzeilt om lang moleculair DNA te verkrijgen, is het vermogen om DNA / RNA rechtstreeks uit de cel te sequencen. Door de kern in de nabijheid van het buitenmembraan te brengen en strategisch een nanopore op zowel de kern envelop als op het buitenmembraan te incorporeren, kan het binnenste van de kern worden verbonden met de sequentieporie.
Door gebruik te maken van het intrinsieke mechanisme dat proliferatie regelt om verstrengeling en vouwing te regelen, kunnen DNA-moleculen de kern envelop verlaten door het buiten-membraan in de sequentieporie. Dit zou op zijn beurt de ingewikkelde verstrengeling van zeer grote moleculen omzeilen en tegelijkertijd het breken van DNA-moleculen vermijden dat vaak voorkomt als gevolg van invasieve laboratoriumprocedures zoals pipetteren of mechanische lysis.

Met een beetje fantasie zou het zelfs mogelijk kunnen zijn om het uitgelezen DNA of RNA terug te voeren via een extra feedbackporie. Het afwikkelen van de DNA-strengen wordt dan gefaciliteerd door shaperone eiwitten die de losgekoppelde eiwitten verzameld en terug plaatst na het sequencen. Dit maakt de uitlezing van de volledige genomische inhoud van een enkele cel mogelijk zonder de noodzaak om de cel op te offeren. En zou onderzoekers in staat stellen om gepaarde datasets te genereren die statistisch enorm waardevol zijn, waarbij biologische variatie op cellulair niveau wordt vermeden.

Literatuurlijst

1. Lee J Kerkhof, (2021), Is Oxford Nanopore sequencing ready for analyzing complex microbiomes?, FEMS Microbiology Ecology, Volume 97, Issue 3, fiab001, https://doi.org/10.1093/femsec/fiab001

2. Giselle C. Martín-Hernández, et. al., (2021), Chromosome-level genome assembly and transcriptome-based annotation of the oleaginous yeast Rhodotorula toruloides CBS 14, Genomics, Volume 113, Issue 6, Pages 4022-4027, ISSN 0888-7543, https://doi.org/10.1016/j.ygeno.2021.10.006.

3. Min-Seung Jeon et al., (2023), Life Science Alliance, 6 (4) e202201744; DOI: 10.26508/lsa.202201744

4. Yury A Barbitoff et al., (2021), Chromosome-level genome assembly and structural variant analysis of two laboratory yeast strains from the Peterhof Genetic Collection lineage, G3 Genes|Genomes|Genetics, Volume 11, Issue 4, jkab029, https://doi.org/10.1093/g3journal/jkab029

5. Hunt, M. et al, (2015), Circlator: automated circularization of genome assemblies using long sequencing reads. Genome Biol 16, 294. https://doi.org/10.1186/s13059-015-0849-0

6. Kathryn Dumschott et al., (2020), Oxford Nanopore sequencing: new opportunities for plant genomics?, Journal of Experimental Botany, Volume 71, Issue 18, Pages 5313–5322, https://doi.org/10.1093/jxb/eraa263

7. De Maio N et al., (2019), Comparison of long-read sequencing technologies in the hybrid assembly of complex bacterial genomes. Microb Genom. 5(9):e000294. doi: 10.1099/mgen.0.000294.

8. Jaume Pellicer et al., (2010), The largest eukaryotic genome of them all?, Botanical Journal of the Linnean Society, Volume 164, Issue 1, Pages 10–15, https://doi.org/10.1111/j.1095-8339.2010.01072.x

9. Segerman B (2020) The Most Frequently Used Sequencing Technologies and Assembly Methods in Different Time Segments of the Bacterial Surveillance and RefSeq Genome Databases. Front. Cell. Infect. Microbiol. 10:527102. doi: 10.3389/fcimb.2020.527102

10. Duan, H., Jones, A.W., Hewitt, T. et al., (2022), Physical separation of haplotypes in dikaryons allows benchmarking of phasing accuracy in Nanopore and HiFi assemblies with Hi-C data. Genome Biol 23, 84. https://doi.org/10.1186/s13059-022-02658-2

11. Sonia Boughattas et al., (2021), Whole genome sequencing of marine organisms by Oxford Nanopore Technologies: Assessment and optimization of HMW-DNA extraction protocols, Ecology and Evolution, https://doi.org/10.1002/ece3.8447

12. Aaron Pomerantz et al., (2018), Real-time DNA barcoding in a rainforest using nanopore sequencing: opportunities for rapid biodiversity assessments and local capacity building, GigaScience, Volume 7, Issue 4, giy033, https://doi.org/10.1093/gigascience/giy033

13. Darriba, D., Taboada, G., Doallo, R. et al. , (2012), jModelTest 2: more models, new heuristics and parallel computing. Nat Methods 9, 772 https://doi.org/10.1038/nmeth.2109

14. Silvestre-Ryan, J., Holmes, I., (2021), Pair consensus decoding improves accuracy of neural network basecallers for nanopore sequencing. Genome Biol 22, 38. https://doi.org/10.1186/s13059-020-02255-1

15.Bayega, A., Oikonomopoulos, S., Gregoriou, ME. et al., (2021), Nanopore long-read RNA-seq and absolute quantification delineate transcription dynamics in early embryo development of an insect pest. Sci Rep 11, 7878. https://doi.org/10.1038/s41598-021-86753-7

16. Nicolas Altemose et al., (2022),Complete genomic and epigenetic maps of human centromeres. Science 376, eabl4178. DOI:10.1126/science.abl4178

17. Lee, I., Razaghi, R., Gilpatrick, T. et al., (2020), Simultaneous profiling of chromatin accessibility and methylation on human cell lines with nanopore sequencing. Nat Methods 17, 1191–1199. https://doi.org/10.1038/s41592-020-01000-7

18. Mitchell R. Vollger et al., (2022), Segmental duplications and their variation in a complete human genome. Science376,eabj6965.DOI:10.1126/science.abj6965

19. Wang Y, Yang Q and Wang Z (2015) The evolution of nanopore sequencing. Front. Genet. 5:449. doi: 10.3389/fgene.2014.00449