Norges miljø- og biovitenskapelige universitet

Masteroppgave 2019 30 stp
Fakultet for kjemi, bioteknologi og matvitenskap

Effekt av hybridassembly på genomer med shuffloner og repeterte områder

The effect of hybrid assembly on genomes with shufflons and repetitive regions

Maren-Helene Høie Degnes
Bioinformatikk, mikrobiell genomikk

Forord

Denne masteroppgaven var gjennomført for Norges miljø- og biovitenskapelige universitet ved faktultet for kjemi, bioteknologi og matvitenskap. Min hovedveileder har vært Lars Snipen og biveilederen min har vært Knut Rudi.

Jeg vil benytte anledningen til å takke min veileder Lars Snipen for svært god veiledning, hans tålmodighet, nyttige møter og rask respons på mail. Denne veiledningen har vært avgjørende for gjennomføringen av lab-arbeidet på pc-en og masterskrivingen.

Takk til Knut Rudi, Inga Leena Angell og Mari Hagbø for rå-data og veiledning med mikrobiologi-delen av arbeidet.

Takk til familie, venner og kjæresten min for fine stunder, motivasjon og hjelp i løpet av de fem åra jeg har studert.

Sammendrag

Antibiotikaresistens spres mest effektivt mellom bakterier via konjugasjon, men konjugasjon forutsetter at bakteriene binder seg godt nok til hverandre. IncI1-plasmidet er et konjugativt plasmid og inneholder et område kalt shufflon som består av flere deler. Rekkefølgen på de ulike delene er med på å bestemme hvilke forbindelser bakterien kan binde seg til og dette er grunnen til at det er interessant å studere rekkefølgen av innholdet i shufflonet. De ulike delene er høyt konservert, men utfordringen er at rekkefølgen på delene varierer mellom bakterier fra samme kultur.

Assemblering av parvise reads fra plasmider med ulike versjoner av shufflonet vil være utfordrende, fordi det kun er shufflon-sekvensen som varierer mellom plasmidene. En annen utfordring ved assemblering er repeterte områder som blir utfordrende dersom det repeterte området er lengre enn fragmentlengden på fragmentene de parvise readene er sekvensert fra.

For å studere disse to utfordringene gjøres en systematisk studie av simulerte reads fra 1) konstruerte plasmider med repetert område av ulike lengder og 2) konstruerte plasmider med hver sin utgave av shufflonet. Først simuleres korte parvise Illumina-read som assembleres med SPAdes-assembleren for å undersøke om utfordringene nevnt over faktisk gir problemer for assemblering. Deretter blir lange Nanopore-reads simulert og HybridSPAdes hybridassemblerer Illumina- og Nanopore-readene for å se i hvilken grad hybridassembly forbedrer assembly. I tillegg blir FLASH brukt til å lage forlengede reads av de delvis overlappende parvise readene. De forlengede readene skal vistnok forbedre assembly dersom disse blir brukt i tillegg til de parvise. MetaSPAdes som er beregnet for assemblering av metagenomer kjøres også. Etter simuleringen assembleres også reelle Illumina- og Nanopore-reads sekvensert fra *E. coli* med shufflon-område.

Resultatene fra simuleringen viser at assembly av kun korte parvise reads blir ufullstendig når repetert område er lengre enn fragmentlengde. Grunnen til dette er at assembleren utnytter at hvert par av de parvise readene har en gitt avstand til hverandre, og at enkelte par overlapper delvis. Dette er også grunnen til at forlengede reads ikke forbedrer assembly, men derimot kan gi assembly med mer feil, fordi FLASH gjør feil ved skjøting av readene. HybridSPAdes løste opp ufullstendige assemblyer av parvise reads fra det repeterte området, men ikke fra shufflonene. Både SPAdes, HybridSPAdes og MetaSPAdes ga ufullstendig assemblering der korte contiger besto kun av deler av shufflonet og de fullstendige sekvensene av alle shufflon-variantene ikke var mulig å finne.

MetaSPAdes var assembleren som fungerte best på shufflon-dataene, samtidig som den ga aller færrest contiger ved assemblering av de reelle readene. Grunnen til at MetaSPAdes fungerer godt på shufflon-readene kan være at de minner om et metagenom. Assemblering av de reelle readene ga heller ikke fullstendige sekvenser av shufflon-variantene. Det foreslås videre å bruke korte reads til å korrigere de lange dersom det kun er shufflon-sekvensen som er interessant.

Abstract

Antibiotic resistance is spread most efficient among bacterias through conjugation, but that requires sufficient binding between the bacterias. One type of conjugative plasmid called IncI1-plasmid contains a sequence called shufflon which consists of multiple parts. The parts' order in the shufflon decides which molecules the bacteria can bind. This makes it interesting to investigate these orders. The parts are highly conserved, but the challenge with the investigation is that the order of the parts varies between bacterias within the same culture.

Assembly of reads from plasmids with different versions of the shufflon will be challenging because the shufflon is the only difference between the plasmids. Another challenge with assembling are repeated regions if they are longer than the length of the fragments the paired-end reads are sequenced from.

To study these two challenges it was done a systematic study of simulated data from 1) constructed plasmids containing repeated regions of different lengths and 2) constructed plasmids with different variant of the shufflon. First, short Illumina-reads were simulated and assembled by SPAdes to investigate if the challenges mentioned above really was challenging for the assembler. Then long Nanopore-reads were simulated and HybridSPAdes assembled both Illumina- and Nanopore-reads to investigate if hybridassembly improves assembly. In addition the FLASH-software link partly overlapping paired-end reads prior to assembling. The assembler used these linked reads along with R1- and R2-reads to improve the assembly. MetaSPAdes was also runned on the simulated reads. After the simulation real Illumina- and Nanopore-reads from sequenced *E. coli* with shufflon were also assembled.

The results show that assembly of only short Illumina-reads are challenging when the repeated region is longer than fragment length. The reason for this is the assembler utilizes that each pair of paired reads have a fixed distance from each other and that some paired reads partly overlaps. This is also the reason why linked reads from FLASH doesn't improve assemblies. However, the linked reads can give assembly with more errors, due to mistakes done by FLASH during linking. HybridSPAdes solved the challenged assembly of reads from repeated regions, but not from the shufflons. Both SPAdes, HybridSPAdes and MetaSPAdes gave uncomplete assemblies where the short contigs consists of parts of the shufflon, and it was not possible to detect all the shufflon-variants.

MetaSPAdes was the best working assembler on the shufflon-data, and it also gave least contigs with the real reads. The reason MetaSPAdes worked best is possibly because shufflon-data and the real data are somewhat simulair to a metagenome. Assembly of the real reads did not either give the complete sequences of the shufflon variants. It is further suggested to use short reads to correct the long reads if the sequences of the shufflons is the only focus.

Innhold

Ordforklaringer

DBG De-Bruijn-graf

DNA Deoksyribonukleinsyre

dNTP deoksynukleotidtrifosfat

IncI1 Incompatibility I1

OLC Overlap-layout-consensus

PCR polymerase-chain-reaction

RO Repetert område

RS Repetert sekvens

Tabell 1: Beskrivelser av (engelske) begreper innen bioinformatikk og genomikk

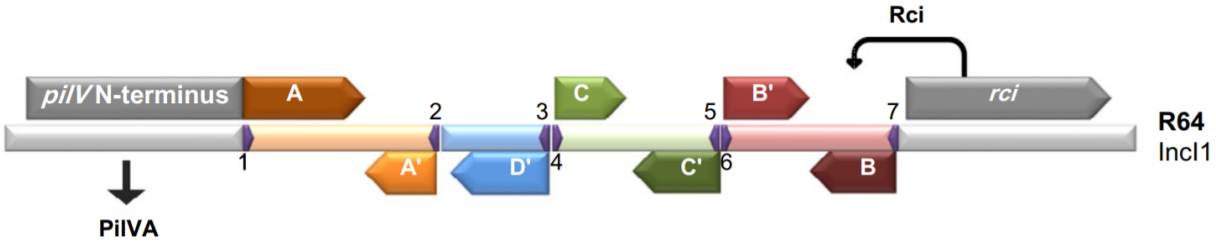
Begrep	**Beskrivelse**
Reads	Resultat av sekvensering [1]. Reads eksisterer i form av ord i datafiler. Ordene består i hovedsak av bokstavene som representerer de unike basene i DNA; A, C, G og T. Under sekvensering analyseres fragmenter av DNA. Analysen genererer signaler om hvilke baser som finnes i hvilke fragmenter og hvor sikkert det er at det er akkurat denne basen i form av en kvalitet. Datafiler som inneholder reads har for hver read en overskrift, readen og en kvalitetslinje med alle kvalitetene for hver posisjon i readen.
Fragmentlengde/Insert size	Dersom fragmenter av DNA sekvenseres fra hver ende av fragmentet vil dette gi to reads som defineres som par og som kalles R1 og R2 [2]. Lengden på hele fragmentet som et R1-og R2-par sekvenseres fra kalles insert size.
Assembly	Direkte oversatt betyr assembly montering på norsk. Et assembly er i bioinformatikk overlappende reads som er skjøtet sammen til én eller flere sammenhengende sekvenser med eller uten mapping mot referanse-genom.
Å assemblere	Prosessen med å sette sammen readene til et assembly.
De-novo-assemblering	Assemblering uten aligning mot referanse-genom.
Contig	En contig er en sammenhengede sekvens av assemblerte reads. Dersom et assembly består av flere contiger betyr det at disse contigene av en eller annen grunn ikke kunne skjøtes sammen under assemblering.
Alignment	Et alignment i bioinformatikk viser sammenligning av to eller flere hele sekvenser. Likheter og forskjeller mellom sekvensene er uthevet med tegn.
Å aligne	Prosessen med å lage et alignment.
Å mappe	Sammenligne kortere sekvenser mot hverandre eller mot én lang.

1 Introduksjon

1.1 IncI1-plasmidet

I sin masteroppgave fra 2017 [3] identifiserte Mari Hagbø antibiotikaresistente *Esherischia coli* fra tarmen hos et spansk fortidligfødt tvillingpar. Disse tvillingene hadde enda ikke mottatt noe antibiotika og det er usikkert hva som er kilden til disse bakteriene. *E. colienes* gener for antibiotikaresistens ble funnet i et plasmid av typen Incompatibility I1 (IncI1). IncI1-plasmidet er et konjugativt plasmid [4][5] og konjugasjon er den mest effektive formen for horisontal genoverføring [6]. For å undersøke mekanismene for spredningen av IncI1 nærmere ble en prøve fra en kultur av disse *E. coli*-cellene sekvensert og assemblert. Siden *E. coliene* kommer fra samme kultur vil kromosomene være identiske, men dette gjelder ikke IncI1-plasmidene.

Grunnen til at IncI1-plasmider fra samme bakterie-kultur er ulike hverandre er på grunn av et spesielt sekvensområde i plasmidet som kalles shufflon. Shufflonet ble funnet og beskrevet i IncI1-plasmidet R64 i 1986 [7]. Flere varianter av shufflonet er også funnet i andre IncI1-plasmider [8]. Et shufflon består av opptil 7 deler fordelt på 4 segmenter, se eksempel i figur 1. Enzymet fra *rci*-genet kalles rekombinase og kan flippe og flytte rundt på (invertere) delene inni shufflonet [9]. Felles for alle shuffloner er at de ligger nedstrøms for *pilV*-genet og oppstrøms for *rci*-genet. Disse tre sekvensområdene er en del av overføringsregionen i IncI1 [10].



Figur 1: Figuren er hentet fra [8] og viser henholdsvis *pilV*-genet, shufflonet og rekombinase-genet *rci* fra IncI1-plasmidet R64. Dette shufflonet består av 7 deler adskilt av korte repeterte sekvenser som er indikert her i lilla. De delene med samme bokstav, for eksempel A og A', tilhører samme segment. Figuren viser 7 deler fordelt på 4 segmenter.

Hvordan delene i shufflonet er organisert avgjør hvilke forbindelser en bakteries pilus kan binde stabilt til. En pilus er et hår/utvekst på utsiden av bakterieveggen bestående av proteinet pilin [11] og bakterier kan både ha tynn og tjukk pilus [12]. *PilV*-genet koder for den delen av den tynne pilusen til *E. coli* som stabiliserer konjugasjonen mellom celler [13][14]. Stabiliseringen skjer ved at enden av *pilV*-proteinet binder til spesifikke karbohydratstrukturer på overflaten av andre celler [15]. Denne enden på *pilV* består av proteinet fra sekvensen av den delen av shufflonet som forekommer nærmest *pilV*-genet [16][17]. I eksempelet i figur 1 ville enden av *pilV* som stabiliserer binding være delen av shufflonet som kalles "B". Dette er grunnen til at rekkefølgen av delene i shufflonet er interessant å undersøke. For å bedre forstå mekanismene bak antibiotikaresistens, og binding mellom pilus og andre forbindelser, studeres sekvensen til IncI1-plasmidet. Prøvene med *E. coli* sekvenseres etterfulgt av assemblering.

1.2 Sekvensering

For å undersøke IncI1-plasmider og deres shuffloner sekvenseres *E. coli*. Sekvensering er en prosess der basene i DNA-molekylene via ulike metoder genererer signaler som oversettes til reads. "Base-kaller" er et type program som oversetter signalene til reads. Readene fra sekvensering eksisterer i form av tekst i et standard fil-format som kalles fastq-fil [18]. I en fastq-fil er det en overskrift, sekvensen til readen og en linje med tegn som indikerer kvaliteten til hver base i sekvensen. Med kvaliteten menes et Q-poeng på hvor sannsynlig det er at den aktuelle basen er riktig sekvensert. Q-poengene er angitt som tegn fra ASCII-tabellen.

1.2.1 Illumina-sekvensering

Illumina er en sekvenseringsteknologi som genererer mange korte reads med lite sekvenseringsfeil til en lav pris sammenlignet med andre sekvenseringsteknologier [19]. En av sekvenseringsmaskinene fra Illumina er MiSeq v3 og den genererer parvise-reads med lengde 250 til 300 basepar (bp) [20]. Readene er parvise fordi de sekvenseres fra hver ende av samme fragment og utgjør et read-par kalt R1 og R2.

Før sekvensering må DNA-molekylene prepareres [21]. Første del av prepareringen er fragmentering og denaturering av DNA til korte enkelt-trådede DNA-fragmenter. Deretter ligeres adaptersekvenser på begge endene av fragmentene. Sekvenseringen av DNA-et foregår i kanaler i en flytecelle (les: flowcell) [22]. Adaptersekvensene på endene av fragmentene binder til sekvenser som er festet i flytecella. Fragmentene blir så PCR-amplifisert slik at det dannes mange kopier av hvert fragment i flytecella [21].

Sekvenseringen skjer ved syntese [22] der byggeklossene til DNA, deoxynukleotidtrifosfater (dNTP-er) bindes til fragmentene i flytecella og til hverandre. DNTP-ene flyter fritt i flytecella og binder seg til den komplementære basen på DNA-fragmentene fra den ene enden. Hver dNTP er merket med fluoriserende forbindelse som avgir en lysbølge med spesifikk bølgelengde når dNTP-en binder seg. Lyssignalene fanges opp av en sensor og brukes til å generere reads med en base-kaller. I første omgang lages R1-readene. Når DNA-fragmentet er sekvensert fra den ene siden, vaskes den syntetiserte DNA-tråden bort og DNA-fragmentet blir amplifisert for å syntetisere den komplementære tråden. Deretter blir DNA-fragmentet vasket bort og den komplementære tråden sekvenseres fra den andre enden og genererer i andre omgang R2-reads [22]. Flytecella inneholder like mange av hver utgave av dNTP samtidig, slik at den utgaven som binder aller best er den som skal binde seg under sekvensering [23].

1.2.2 Nanopore-sekvensering

Oxford Nanopore og Pacific Biosciences er de to dominerende sekvenseringsteknologiene som gir lange reads med gjennomsnittlig read-lengde på henholdsvis 10 000 til 60 000 bp [24][25][26]. Den negative siden ved de lange readene er deres store andel sekvenseringsfeil [27].

Nanopore tilbyr flere typer sekvenseringsmaskiner. En av de mindre sekvenseringsmaskinene ligner en USB-penn, veier 90 gram og er lett å bruke [28] [25]. Denne USB-pennen heter MinION og kan kobles direkte til en datamaskin. MinION har opp til 2048 nanoporer der sekvensering skjer [28]. En nanopore er en kanal på samme måte som ionekanaler eller andre proteinkanaler i celler [29].

Før sekvensering ligeres adaptersekvenser på endene av DNA-et som skal sekvenseres [26]. Den ene adaptersekvensen sørger for at sekvenseringen skjer fra én ende til den andre på DNA-molekylet. 5'enden på adapteren sendes først gjennom nanoporen og resten av DNA-et følger etter. Den andre adaptersekvensen sørger for at begge trådene i DNA-molekylet er bundet sammen slik at de kan sekvenseres sammenhengende. Så dras DNA-et gjennom nano-poren og en sensor detekterer ladningsforskjeller mellom det ulike innholdet i DNA-tråden. Disse signalene samles i fast5-filer. Etter sekvensering blir signalene i fast5-filene omgjort til reads i fastq-filer med et base-kaller-program.

1.3 Assemblering

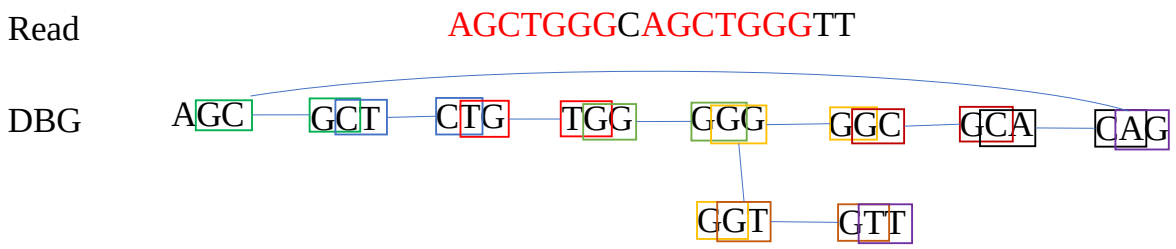
For å kunne undersøke spredningen av antibiotikaresistens er det behov for å finne rekkefølgen av shufflon-delene ved å studere shufflonenes fullstendige sekvenser. Sekvensering gir stykkevis informasjon om DNA-et i form av reads, og readene kan settes sammen til lengre sammenhengende sekvenser, contiger, med assemblering [1]. Ved assemblering kan readene enten mappes til et referanse-genom eller de overlappende readene kan settes sammen til lengre sekvenser med de-novo-assemblering uten referanse-genom. Antall reads som dekker én posisjon i genomet som blir sekvensert kalles sekvenseringsdybde. Utregning av gjennomsnittlig sekvenseringsdybde for sekvensering er vist i likning (1) som er hentet fra [30]. Etter assemblering kan dybden av readene på hver contig regnes ut og sammenlignes med sekvenseringsdybden, vist i likning (2).

$$\text{Gjennomsnittlig sekvenseringsdybde} = \frac{\text{antall reads} \cdot \text{mean(read-lengde)}}{\text{genomlengde}} \tag{1}$$

$$\text{Gjennomsnittlig read-dybde på contig} = \frac{\text{antall mappende reads på contigen} \cdot \text{mean(read-lengde)}}{\text{contig-lengde}} \tag{2}$$

1.3.1 Assemblering av korte reads

De fleste de-novo-assemblere for korte reads er de-Bruijn-grafen-assemblere (DBG) [31]. DBG brukes til å finne delvise overlapp mellom enkle og parvise reads, [32] og er et nettverk av noder og koblinger mellom de nodene som har overlappende innhold. Readene blir delt opp i k-merer som brukes som noder i DBG, se figur 2. To noder i grafen bindes sammen dersom k-1 av suffixen til en node tilsvarer k-1 av prefixen av den andre noden. Noder som er koblet sammen omgjøres til sammenhengende sekvenser i en assembly-graf. To av utfordringene med assemblering av korte reads er lange repeterte områder [33] og høyt varierende områder som shufflonet [8].



Figur 2: Denne figuren er en forenklet visualisering av de-Bruijn-grafen der 3-merer av en read kobles sammen. Rød sekvens i readen indikerer repetert sekvens. Bokstavene med samme farge på rammen rundt er identiske. Figuren viser at repetert område lager koblinger tilbake til eksisterende noder, såkalte bobler. Dersom RO er mye lengre vil det være utfordrende for assembleren å finne den sanne rekkefølgen av noder på grunn av komplekse boble-strukturer.

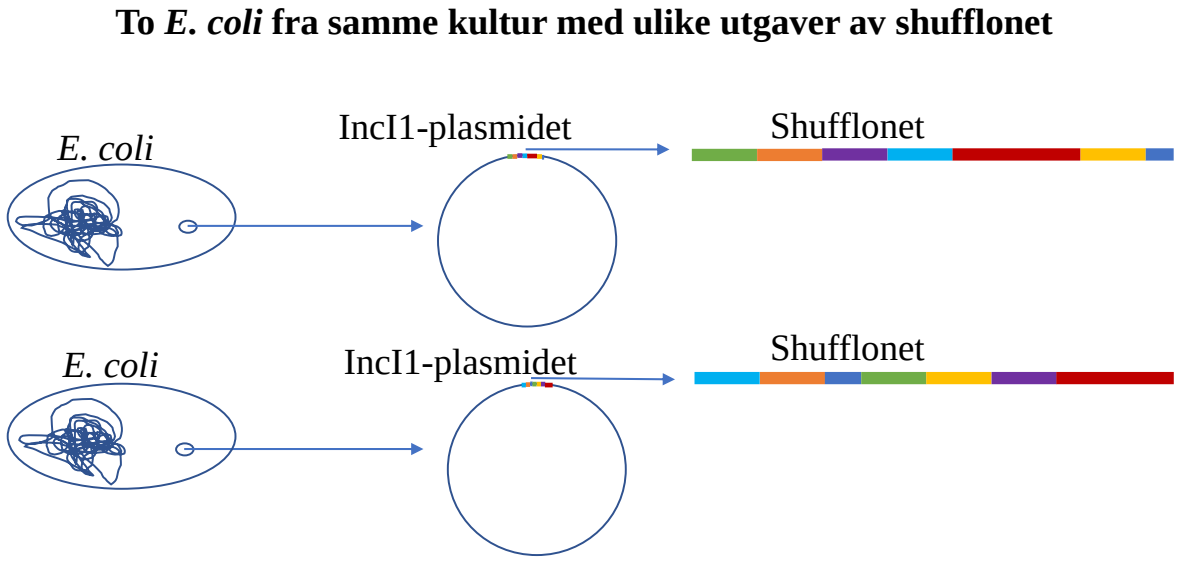
Et repetert område er en lengre sekvens som forekommer flere ganger i et genom. Et eksempel på et repetert område er det 1500 bp lange 16S-genet i genomet hos bakterier [34]. Dersom et repetert område er lengre enn read-lengde vil ikke readene alene gi nok informasjon til at assembleren klarer å plassere kopiene av det repeterte området [35], se figur 3 på neste side. Parvise reads kommer fra sekvensering av hver ende av hvert DNA-fragment, og disse inneholder derfor informasjon om lengre strekker av DNA-fragmentene enn enkle reads [36]. Assemblere kan bruke parvise reads til å løse opp repeterte områder sålenge den parvise distansen (les: fragmentlengden) er lengre enn det repeterte området [37]. Repeterte områder kan være for lange til at parvise reads kan strekke over dem [38]. Derfor kan assembly med lengre repeterte områder bli utfordrende selv med parvise reads [33]. Repetert område blir heretter referert til som RO og sekvensen av den, altså den repeterte sekvensen blir heretter referert til som RS.

Figur 3: Figuren viser en DNA-sekvens som inneholder to kopier av et repetert område indikert med rød farge og rød R. De korte linjene under DNA-sekvensen er reads og de er plassert under det området på DNA-sekvensen de er sekvensert fra. Readene i samme farge er identiske (gjelder ikke de svarte). Figuren er inspirert av forelesning fra Lars Snipen.

Et ufullstendig assembly på grunn av RO kan resultere i to scenarioer [39]. Første scenarioet er at kopier av RS blir tolka som samme område og blir lagt oppå hverandre. Eller omvendt; at kopier av RS blir tolka som flere sekvenser og blir plassert på flere områder. Denne typen assembly vil ha henholdsvis høyere og lavere read-dybde i RO sammenlignet med ikke-RO. Andre scenario er at kopiene av den repeterte sekvensen blir lagt i feil rekkefølge, slik at områdene mellom dem også ligger i feil rekkefølge [39]. Dette kan ha konsekvenser for den biologiske tolkningen av sekvensen. En løsning på denne utfordringen kan være å bruke lange nok reads som strekker over RO [38][40].

Enda en utfordring for de-novo-assemblering av korte reads er shufflon-området. Delene i shufflonene er høyt konservert, men rekkefølgen av innholdet varierer fra plasmid til plasmid [8], og derfor kan bakterier fra samme kultur ha ulike varianter av shufflonet, se figur 4 på neste side. Altså er genomene fra bakterier av samme kultur identiske utenom shufflonet. Ifølge Brouwer *et. al* [8] gir de-novo-assemblering én contig med bare én variant av shufflonet og/eller contiger bestående av de ulike delene av shufflonene. Med slike resultater vil det ikke være mulig å finne alle variantene av shufflonene fra kulturen. Resultatene Brouwer *et. al* beskriver minner om assembly av sekvensert metagenom, som er utfordrende for nært beslektede genomer med mye felles sekvens [41].

I tillegg vil det være utfordrende å skille mellom reads fra plasmid og kromosom ved sekvensering av hele celler [42]. Et forslag til løsning er å utnytte at plasmider ofte forekommer i mange kopier og kromosomet bare i én [42]. Men ifølge Knut Rudi forekommer IncI1-plasmidet bare én til to ganger og da vil ikke denne løsningen fungere, fordi kromosom og plasmid forekommer omtrent like mange ganger. Brouwer *et. al* foreslår at lengre reads enn fra Illumina kanskje kan løse assembly av shufflonene [8].



Figur 4: Visualisering av to *E. colier* fra samme kultur som har identiske genomer utenom shufflonet. De to *E. coliene* har kromosomalt DNA og et plasmid vist som en sirkel. De forskjellige fargene på linjene helt til høyre indikerer de forskjellige delene i shufflonet.

1.3.2 Assemblering med korte reads og forlengede reads

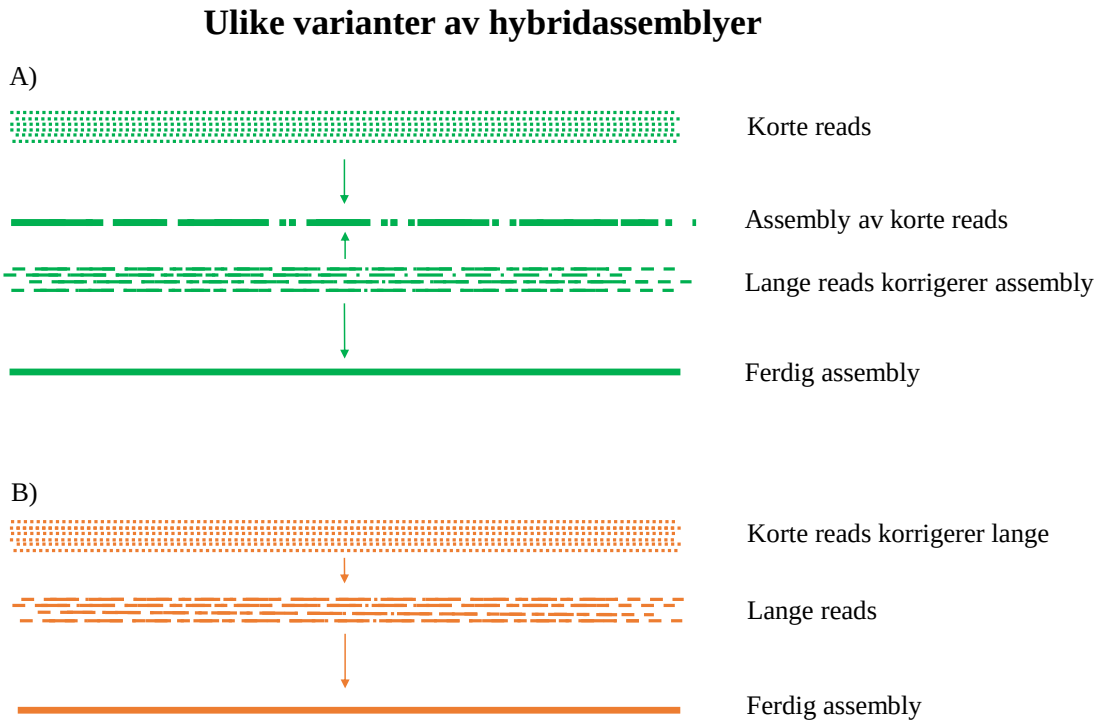
I 2011 utviklet Magoč og Salzberg [43] FLASH-programvaren som skal skjøte delvis overlappende parvise reads før assemblering. Dersom disse readene brukes i tillegg til de parvise readene skal de tette hull i de-novo-assemblyer [43]. Senere har de to forfatterne, sammen med flere andre i 2013 [37] skrevet at assemblere kan løse opp i repeterte områder dersom fragmentlengden for de parvise readene er lengre enn det repeterte området. Dette skrev de uten å nevne FLASH-programvaren.

1.3.3 Assemblering av lange reads

Lange Nanopore-reads har lengde på rundt 10 000 bp og vil strekke over lange RO. Utfordringen er at disse lange readene inneholder en stor andel sekvenseringfeil [40][44]. På grunn av mye sekvenseringsfeil egner ikke konstruksjon av DBG seg som assemblerings-metode for Nanopore-reads [45]. Assemblering gjøres heller ved bruk av "overlap-layout-consensusmetoden (OLC) [45] der konsensusen av readene brukes [46][47].

1.3.4 Hybridassembly

Det går an å kombinere korte Illumina- og lange Nanopore-reads i et hybridassembly [40][25][48]. Det er flere måter å gjøre hybridassembly på og to av disse er visualisert i figur 5 på neste side. I 2017 klarte Sekizuka *et. al* [16] å lage et vellykket hybridassembly fra bakterier med IncI-plasmid som inneholdt shufflon. Disse forskerne korrigerte lange PacBio-reads med Illumina-reads før de ble assemblert. En annen variant av hybridassembly er å assemblere korte Illumina-reads først og så bruke de lange readene til å skjøte sammen assembly [40]. SPAdes er en assembler som kan gjøre hybridassembly [40] og som gjør det sistnevnte; først assemblerer de korte, og så bruker de lange til å kombinere usammenhengende deler av assembly.



Figur 5: Figur som viser to varianter for hybridassembly, altså måter å kombinere lange og korte reads på. A) viser metoden som HybridSPAdes bruker og som assemblerer korte reads og skjøter sammen assembly med lange reads etterpå [40]. B) viser metoden som er brukt i [16] der korte reads korrigerte lange reads. Figuren er inspirert av figur i [16].

1.4 Mål med oppgaven

Utgangspunktet for denne masteroppgaven er et ønske om å assemblere hele *E. coli*-genomer inkludert IncI1-plasmider som inneholder et shufflon-område. Med assembleringen kommer utfordringer med lengre repeterte sekvenser og det variable shufflonet som vil variere for hvert IncI1-plasmid. *E. coli*-genomene er sekvensert med både Illumina og Nanopore i håp om at hybridassembly skal forbedre situasjonen.

Formålet med oppgaven er å forstå hvordan repeterte områder og shuffloner skaper problemer for assemblering ved å gjøre en systematisk studie basert på simulerte data. Først blir kun Illumina-reads assemblert og så gjøres et hybridassembly med både Illumina- og Nanopore-reads. Studien er delt i tre deler der det først undersøkes konsekvensene av repeterte områder med ulike lengder, deretter flere plasmider med hver sin utgave av shufflonet og til slutt assembleres de reelle dataene fra sekvensering av *E. coli*.

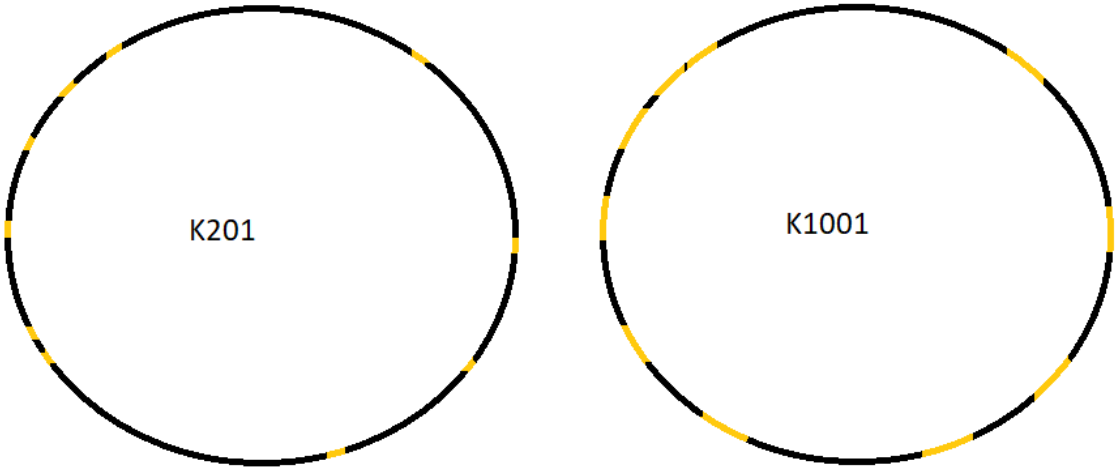
2 Metode

2.1 Data

Målet med oppgaven er å studere hvordan assembly blir påvirket av sekvenser som repetert område av ulik lengde og shuffloner med en systematisk studie av simulerte data. Simulerte data brukes i del 1 og 2 av denne oppgavens 3 deler. I første del konstrueres et repetert område inn i en plasmid-sekvens, så simuleres reads fra denne sekvensen og deretter assembleres disse readene. Lengden på det repeterte området økes systematisk for å undersøke effekten av lengden. Både korte Illumina- og lange Nanopore-reads simuleres og assembleres til assemblyer og hybridassemblyer. I andre del konstrueres fem ulike shufflon-varianter inn i 5 identiske plasmider. Her simuleres også reads på samme måte som i del 1, og disse readene brukes til å assemblere til assemblyer og hybridassemblyer. I tredje og siste del av studien er det fokus på de reelle readene, som assembleres på samme måte som de simulerte readene. De ulike delene av oppgaven er beskrevet i figur 7 på side 17 og programmene som brukes er beskrevet i tabell 2 på side 17.

For å kunne konstruere plasmid-sekvenser til simuleringen trenger vi et utgangspunkt-plasmid der RO og shufflonene kan kontrueres inn. Utgangspunkt-plasmidet vi bruker er av typen pSH4469 som kommer fra *Shigella sonnei* beskrevet i artikkelen [10]. Det er store likheter mellom *S. sonnei* og *E. coli*, og derfor er det ikke i veien for å bruke sekvensen av dette plasmidet som utgangspunkt for simuleringene. Plasmidet ble sjekket for repeterte områder, men det var ingen sekvens over 126 bp som var repetert mer enn én gang. I artikkelen plasmid-sekvensen er hentet fra [10] er det ikke nevnt noe om shufflon, men et shufflon med færre enn 7 deler er annotert i sekvensen på NCBI [49].

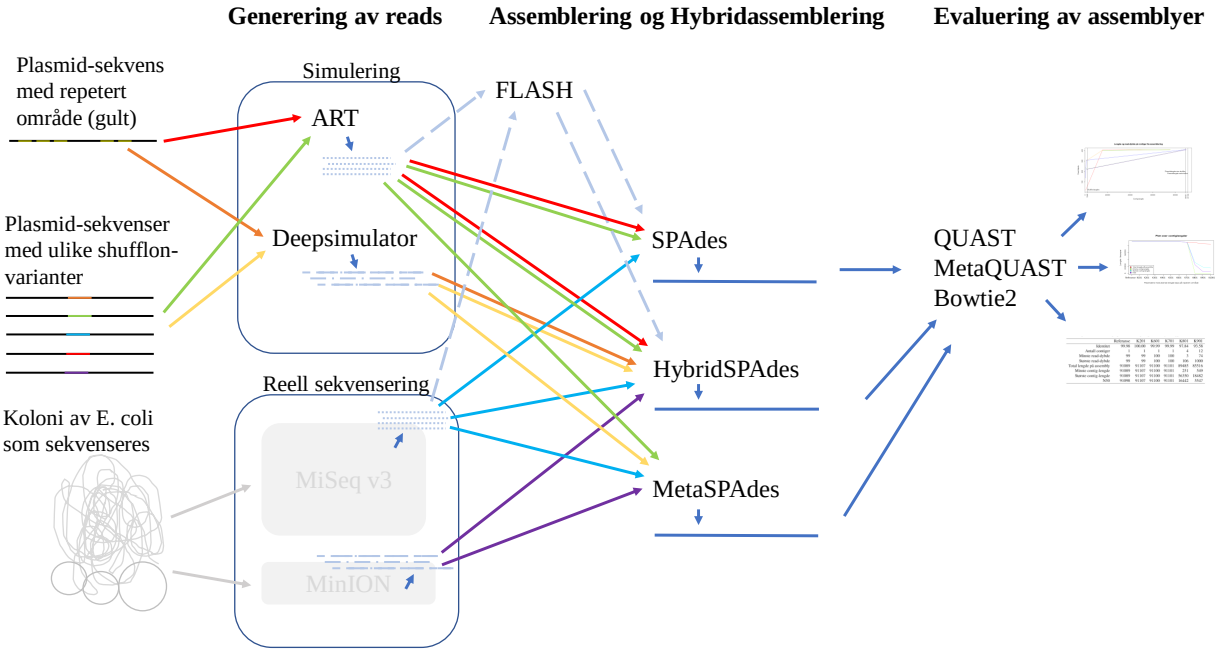
I del 1 av studien ble det konstruert et k langt repetert område som forekom totalt 10 ganger i utgangspunkt-plasmidet. Ifølge Salzberg *et. al* [37] vil et repetert område som er lengre enn fragmentlengden for parvise reads skape problemer for assemblering og resultere i et ufullstendig assembly. For å sjekke om det stemmer ble det lagd flere plasmid-utgaver med systematisk økende lengde på det repeterte området k. Totalt ble det konstruert 10 plasmid-utgaver der plasmid-utgave 1, 2, 3,..,10 inneholdt henholdsvis repetert område med k 101, 201, 301,..,1001 bp. Se figur 6 på neste side der to av disse plasmidene er visualisert. Sekvensen som ble kopiert opp 10 ganger for hvert plasmid og plasseringen av de 10 kopiene i utgangspunkt-sekvensen ble valgt ut tilfeldig. Hver plasmid-utgave ble lagt i hver sin fasta-fil. Grunnen til at det er valgt oddetall på lengdene er fordi ingen av k-merenes revers-komplementære skal kunne være identiske til k-merene. Inkludert utgangspunkt-plasmidet som ble brukt som referanse ble 11 plasmid-sekvenser brukt til å simulere reads, både korte Illumina-reads og lange Nanopore-reads. Simulering av readene er beskrevet i avsnitt 2.1.1 på side 18 og avsnitt 2.1.2 på side 18. Disse readene ble brukt i assemblering og deretter hybridassemblering.



Figur 6: Visualisering av to plasmider med repetert område som forekommer 10 ganger i hvert plasmid. RO i plasmidene er på henholdsvis 201 bp og 1001 bp. Størrelsesforholdene i denne figuren er ikke realistiske.

I del 2 av studien ble assemblering av ulike versjoner av shufflonet undersøkt. For å gjøre en realistisk simulering ble en sekvens av shufflonet funnet på NCBI [50] ved å søke etter shufflon-sekvensen i annoteringene til en fullstendig sekvens av et E.coli-genom. Denne shufflon-sekvensen ble lastet ned og kuttet opp i dets sju deler. Disse delene ble stokket om på og fem versjoner av shufflonet ble konstruert inn inn i hvert sitt utgangspunkt-plasmid. I figur 4 på side 13 er to plasmider med hvert sitt shufflon visualisert. De fem plasmid-sekvensene ble samlet i én felles fasta-fil som det ble simulert Illumina- og Nanopore-reads fra. Som i første delen av denne studien ble det også i denne delen undersøkt hvordan assemblering av kun de simulerte korte Illumina-reads gikk, og så hybridassembly med simulerte Illumina- og Nanopore-reads.

De simulerte Illumina-readene i del 1 og del 2 ble assemblert med SPAdes. Deretter ble også simulerte Nanopore-reads assemblert med Illumina-readene til et hybridassembly. For hvert konstruerte plasmid ble det gjort 10 simuleringer for å hindre at tilfeldige feil skulle dominere resultatet. De ulike programmene som ble brukt er beskrevet i tabell 2 på neste side og hele metoden er visualisert i figur 7 på neste side. I avsnitt 2.2 på side 20 er en beskrivelse av assembleringsprogrammet som blir brukt, SPAdes samt ulike algoritmer av SPAdes.



Figur 7: Graf over simuleringen med programmer og data. Pilene i forskjellige farger følger de enkelte prosessene. Del 1 av oppgaven har røde (simulerte Illumina-reads) og oransje (simulerte Nanopore-reads) piler. Del 2 har grønne (simulerte Illumina-reads) og gule (simulerte Nanopore-reads) piler. Del 3 har lyseblå (reelle Illumina-reads) og lilla (reelle Nanopore-reads) piler. Figuren viser at sekvenser brukes til å simulere reads i del 1 og 2, og hvilke reads som blir gitt til hvilken assembler. De grå delene av figuren viser arbeid som er utført av Mari Hagbø og Inga Leena Angell i forkant av denne masteroppgaven.

Tabell 2: Oversikt over de programmene som er brukt i denne studien og i hvilken del av studien de er brukt. Del 1 av studien fokuserer på repetert område av ulik lengde, del 2 på ulike shuffloner i plasmider og del 3 de reelle dataene. Programmets funksjon er en kort forklaring som fordypes mer i senere i teksten.

Navn på program	**Programmets funksjon i denne studien**	Del		
		1	2	3
ART (v2.5.8)	Program som simulerer Illumina-reads [51].	x	x	
Deepsimulator	Program som simulerer Nanopore-reads [52].	x	x	
SPAdes (v3.13.0)	Assembleringsprogram for Illumina-reads [32].	x	x	x
HybridSPAdes	SPAdes sin algoritme for hybridassembly med både korte og lange reads [40].	x	x	x
MetaSPAdes	SPAdes sin algoritme for assembly av metagenom [53].		x	x
Trimmomatic (v0.36)	trimmer og filtrerer reads for adaptersekvenser og lav kvalitet [54].			x
Canu (v1.8)	Assembler for Nanopore- og PacBio-reads, som kan trimme og filtrere readene uten å assemblere [55].			x
FLASH (v1.2.8)	Program som kan kombinere overlappende R1-og R2-par til sammenhengende lange reads [43].		x	x
Bowtie2 (v2.3.4.1)	Alignerprogram som lager en indeks av en sekvens og mapper andre sekvenser til denne. Kan brukes til å finne ut antall reads som mapper på assembly. Bowtie2 kan også mappe reads mot en referansesekvens for å for eksempel fjerne kontaminering [56].	x	x	x
SAMtools (v1.3.1)	Program som kan konvertere filer mellom formatene sam, bam og fastq [57].	x	x	x
QUAST (v5.0.0)	Program som kan evaluere assembly med eller uten en referanse-sekvens [58].	x	x	x
MetaQUAST (v5.0.0)	Program som kan evaluere assembly mot flere referanse-sekvenser [58].		x	

2.1.1 ART

ART er et simuleringsprogram fra 2012 som syntetiserer reads fra blant annet Illumina-teknologien [51]. Modellene som brukes til å syntetisere reads fra en input-sekvens er trent på å gi samme mengde feilavlesninger som ved sekvensering. Lengden på readene er også styrt av hvilken sekvenseringsteknologi som er spesifisert.

ART (v2.5.8) simulerer Illumina MiSeq v3 reads fra de konstruerte plasmid-sekvensene. Den gjennomsnittlige lengden på disse readene er 250 bp, og antall reads som simuleres tilsvarer en gjennomsnittlig dybde på 100 · genom-lengden; 100x som er normalt for Illumina-reads [59]. Fragmentlengde er gjennomsnittlig lengde på fragmentet R1 og R2 sekvenseres fra og denne er satt til 750 bp med et standardavvik på 100 bp. Den innebygde feilraten for MiSeq v3 reads brukes og som tidligere nevnt simuleres reads 10 ganger for å hindre at tilfeldige feil dominerer resultatet. Dersom opsjonen ”ErrorFree” er spesifisert vil ART i tillegg til de vanlige readene gi ut feilfrie versjoner i ei sekvens-alignment/map-fil (SAM-fil). Slike feilfrie versjoner av readene blir også generert, og ved bruk av SAMtools (v1.3.1) konverteres samfilene til bam-filer og deretter til fastqfiler.

2.1.2 Deepsimulator

Deepsimulator er et simuleringsprogram som syntetiserer Nanopore-reads [52]. Deepsimulator etterligner sekvenseringsprosessen på input-sekvensen med en statistisk modell som forfatterne kaller pore-modell. Denne modellen trenes til å generere signaler fra input-sekvensen på samme måte som ved sekvensering. Deretter genereres reads på grunnlag av disse signalene med en base-kaller som i dette tilfellet kalles Albacore.

For å endre på opsjonene i Deepsimulator må brukeren endre på shellscriptet ’main.sh’ som lastes ned sammen med programvaren fra [60]. Hvilke opsjoner som kan settes i main.sh er beskrevet i tilleggsdata for [52]. Read-lengdene trekkes fra en fordeling og brukeren kan velge mellom beta, eksponensiell, og mixed-gamma fordeling. De reelle nanopore-readene har en gjennomsnittslengde på ca 5000 bp og fordelingen av read-lengdene på disse readene ligner på eksponensiell fordeling. For at de simulerte Nanopore-readene skal ligne på de reelle brukes eksponensiell fordeling i simuleringen. Den kombinasjonen av parametere som gir en fordeling av read-lengde mest lik på fordelingen av de reelle er den forhåndsinnstilte kombinasjonen i sampling.py. Gjennomsnittlig read-lengde på de simulerte Nanopore-readene er rundt 6800 bp, som ikke er altfor langt ifra de reelles gjennomsnittlige read-lengde på 5000 bp.

Ved simulering av nanopore-readene kan antall reads bestemmes av brukeren og dersom sekvenseringsdybden økes vil assembleren ha mer informasjon om sekvensen og assembly kan bli bedre [44]. DeepSimulator (kun én versjon tilgjengelig) simulerer nanopore-reads fra de konstruerte versjonene av utgangspunkt-sekvensen. Antallet reads som simuleres økes systematisk fra en sekvenseringsdybde på 5x og oppover til resultatene slutter å forbedre seg eller endre seg. Det genereres Nanopore-reads med mye sekvenseringsfeil (0.1 på skala fra 0 til 1) for å teste ut verste mulige tilfellet.

2.1.3 De reelle dataene

De tidligere nevnte *E. coliene* som ble identifisert i Mari Hagbø sin masteroppgave ble sekvensert med Illuminas MiSeq v3 og med Oxford Nanopores MinION. Illumina-readene er én R1-fastq-fil og én R2-fastq-fil, begge med 194 075 reads i hver. Gjennomsnittslengden på disse readene er 255.5 bp og en gjennomsnittlig sekvenseringsdybde vises i likning (3). Merk at dette er før Illumina-dataene er filtrert og trimmet.

$$\text{Gjennomsnittlig sekvenseringsdybde} = \frac{194\,075 \cdot 2 \cdot 255.5}{5\,000\,000 + 100\,000} = \underline{19.4x} \tag{3}$$

Nanopore-readene er 338 fastq-filer med ca. 4000 reads i hver, totalt 1 350 595 reads. Den gjennomsnittlige read-lengden ligger på rundt 5000 bp. Base-kalleren som ble brukt til å generere reads fra fast5-filene heter MinKNOWN, og base-kallingen ble utført på forhånd slik at arbeidet mitt kun har dreid seg om readene og assembleringen. Ikke alle Nanopore-readene brukes til assemblering, men et utvalg av de 100 000, 10 000, 5 000, 2 500, og 1 000 lengste. I likning (4) er gjennomsnittlig sekvenseringsdybde basert på alle readene regnet ut. For utvalgene av de lengste read-dybdene vil gjennomsnittlig sekvenseringsdybde være avhengig av gjennomsnittlig read-lengde og antallet.

$$\text{Gjennomsnittlig sekvenseringsdybde} = \frac{1\,350\,595 \cdot 5000}{5\,000\,000 + 100\,000} = \underline{1324.1x} \tag{4}$$

2.1.4 Prosessering av readene før assemblering

De reelle parvise Illumina-readene filtreres, trimmes med Trimmomatic og kontaminering fjernes med Bowtie2 før assemblering. For at Trimmomatic skal kunne fjerne adapterne som satt på fragmentene under sekvensering må brukeren definere adapterne. Trimmomatic har en samling av adaptere som kan brukes og ifølge manualen skal adapteren kalt ”TruSeq3” brukes for Miseq v3-reads [61]. Deretter fjerner Trimmomatic readene med for lav kvalitet og trimmer readene med delvis lav kvalitet. De readene som har mistet partner-readen sin under trimming legges i en egen fil og defineres som enkle reads under assemblering. De reelle readene er sekvensert fra en kultur av transkonjugante *E. coli* av de *E. coliene* som ble hentet fra tarmen til tvillingene. Det kan likevel være rester av menneske-DNA i de sekvenserte prøvene. Derfor mappes readene mot menneske-DNA for å filtrere ut eventuell kontaminering med programvaren Bowtie2. Readene som ikke mapper tas vare på videre. Bowtie2 gir ut alignmentet i en sam-fil, og Samtools konverterer disse til bam-fil og videre til fastq-filer; R1 og R2.

Forlengede Illumina- R1- og R2-reads brukes for å se om assembly blir forbedret, se tabell 2 på side 17. Ifølge Magoč *et. al*[43] blir assembly bedre dersom overlappende parvise reads kombineres med FLASH og assembleres i tillegg til de parvise R1- og R2-readene. FLASH har på forhånd innstilt at overlapp mellom et read-par må være minst 10 bp for at de skal kombineres. Ifølge manualen skal opsjonen –threads settes til 1 hvis readene skal ligge i samme rekkefølge som de originale.

Nanopore-readene inneholder mange feil og de ble derfor korrigert og trimmet før assemblering. Programvaren som gjorde dette var assembleren Canu (v1.8), som kan trimme og filtrere uten å assemblere readene ifølge dokumentasjonen til Canu [62].

2.2 SPAdes

SPAdes er et assembleringsprogram ment for Illumina- eller IonTorrent-reads fra sekvensering av små genomer [63]. Readene blir delt opp i *k*-merer og en DBG konstrueres med *k*-merene som noder og koblinger mellom overlappende noder [32], se figur 2 på side 11. Bankevich *et. al* beskriver at tidligere har ikke parvise reads vært riktig utnyttet under assemblering, men at SPAdes da den kom i 2012 utnytter parvise reads [32]. Noder overlapper dersom *k*-1 av suffixen til en node er lik *k*-1 av prefixen til en annen node.

SPAdes itererer over ulike *k*-mer-lengder under assembleringen og på denne måten konstrueres De-Bruijn-grafene med flere verdier av *k* kalt "Multisized De Bruijn graphs"[32], på norsk: De-Bruijn-grafer med mange størrelser. Grunnen til at SPAdes bruker flere verdier av *k* er at for små verdier vil repeterte sekvenser tolkes som (kollapse til) samme sekvens, mens store verdier vil gjøre det vanskeligere for SPAdes å finne overlapp mellom reads i regioner med lav sekvenseringsdybde. Nodene blir satt sammen til sammenhengende sekvenser med mål om å få sekvensene lengst mulig. De ulike verdiene av *k* som SPAdes itererer over i denne studien er 21, 33, 55, 77, 99, 127 etter anbefalinger fra forfatterne av SPAdes [63]. DBG blir utvidet til lengre strekker med sammenhengende sekvens som blir en assembly-graf.

SPAdes utfører feilkorrigeringer på input-readene. "BayesHammer" kalles en av disse modulene som korrigerer feil i Illumina-reads [64] og som er inkludert i SPAdes-pipelinen hvis ikke annet er spesifisert [63]. En annen modul kalles "MismatchCorrector" og korrigerer for feil samt insersjoner og delesjoner i contigene og scaffoldene på slutten av assembleringen [63]. Denne modulen bruker "Burrows-Wheeler aligner" som mapper korte sekvenser mot assembly med tolleranse for feil og hull [65]. "MismatchCorrector" er i utgangspunktet ikke inkludert i SPAdes-pipelinen, men produsentene anbefaler å aktivere den [63]. Den aktiveres ved å spesifisere opsjonen –careful. SPAdes utfører ikke korrektur på nanopore-reads [63]. MismatchCorrector er aktiverert når SPAdes kjøres i denne studien, dersom ikke MetaSPAdes kjøres.

Det er flere måter å kjøre SPades på [63]. Assemblering av kun R1- og R2-reads fra Illumina gjøres av SPAdes-algoritmen, mens dersom Nanopore-reads gis i tillegg kjøres SPAdes sin algoritme for hybridassemblering kalt HybridSPAdes. Det går også an å sette SPAdes i meta-modus og da kalles assemblerings-algoritmen for MetaSPAdes. Input-reads til SPAdes gis i ulike typer biblioteker [63]. For eksempel gis R1- og R2-reads i et "parvis bibliotek", og dersom enkelte R1 og R2 har mistet par-readen sin under preprosessering gis disse enkle readene i samme parvise bibliotek og spesifiseres som enkle reads med opsjonen -s. Dersom brukeren ønsker å gi SPAdes for eksempel Nanopore-reads for å kjøre HybridSPAdes gis disse inn i et eget type bibliotek kalt "nanopore-bibliotek" som enkle reads. Dersom brukeren vil assemblere med ferdig forlengede reads av R1 og R2 satt sammen av programvaren FLASH, kan disse readene gis i samme bibliotek som R1 og R2 med egen opsjon for denne typen read (opsjonen -m) [63].

2.2.1 HybridSPAdes

Assemblering av både korte og lange reads fra samme genom kalles hybridassemblering [40]. SPAdes sin algoritme for hybridassemblering kalles HybridSPAdes og lange reads kan være enten Nanopore- eller PacBio-reads gitt til SPAdes i et eget type bibliotek med opsjonen –nanopore eller –pacbio. HybridSPAdes assemblerer de korte readene først, og bruker de lange readene til å tette hull og løse opp i repeterte områder [40] [63]. Utviklerne av HybridSPAdes påstår at hybridassembly med HybridSPAdes blir bra selv med lavt antall lange reads [40].

Første steg i HybridSPAdes er å kjøre SPAdes på de korte readene for å konstruere assembly-grafen [40]. Deretter blir *t*-merer av de lange readene mappet til assembly-grafen. Dersom minst 8 (forhåndsinnstilt verdi) *t*-merer fra én bestemt read mapper til et felles strekke på assemblygrafen, mapper denne readen til dette strekket. *t*-merene brukes til å lage en graf som kan sammenlignes med assembly-grafen for å finne områder som kan kobles sammen. Dersom det oppstår komplekse områder i grafen som ikke er entydig kan de lange readene som stikker ut på hver side av dette området koble sammen de riktige nodene. Eksempel på hvordan komplekse områder kan oppstå er visualisert veldig forenklet i figur 2 på side 11. I en større skala vil langt nok repetert område gi komplekse boble-strukturer som SPAdes vil ha problemer med å løse opp i med kun korte reads. Hull i assemblygrafen kan dekkes igjen av de lange readenes konsensuser. ExSPAnder er en modul i SPAdes som brukes for å sette sammen endene i assemblygrafen. ExSPAnder søker etter områder som kan utvides basert på antall reads som dekker en slik utvidelse. Dersom det er flere ender som passer til å utvide med stopper ExSPAnder [40].

2.2.2 MetaSPAdes

MetaSPAdes er en annen algoritme fra SPAdes og denne er ment for assemblering av metagenom [63]. Grunnen til at MetaSPAdes er nevnt her er fordi at assemblering av reads som inneholder ulike versjoner av shufflonet minner om et metagenom. De reelle *E. coli*-bakteriene som sekvenseres og assembleres i denne oppgaven kommer fra samme kultur, altså antas å være identiske *E. coli*. Men på grunn av shufflonet som varierer fra bakterie til bakterie vil disse bakteriene være identiske forutenom rekkefølgen av innholdet i shufflonet, så derfor *nesten* identiske. Det er derfor interessant å teste ut hvordan MetaSPAdes takler shufflon-readene. MetaSPAdes skal også kjøres på de reelle dataene.

MetaSPAdes-algoritmen starter med at SPAdes konstruerer DBG og assembly-graf [53]. Videre jobber MetaSPAdes med å konstruere lange strekker av sekvenser som er felles for alle artene som er representert i metagenomet. Nanopore-reads kan gis slik at HybridSPAdes og MetaSPAdes kjøres samtidig, men forfatterne skriver at de ikke kan garantere optimal assemblering [63]. ”MismatchCorrector” kan ikke aktiveres ved kjøring av MetaSPAdes.

2.3 Evaluering av assembly

Etter assemblering er det interessant å finne ut hvor mange av readene som dekker hver posisjon i hver contig. Illumina-readene brukes for å generere gjennomsnittlig read-dybde for hver contig, fordi de har mindre sekvenseringsfeil enn Nanopore[27] [19]. Den indeks-baserte read-aligneren Bowtie2 [56] mapper readene til contiger med "Burrows-Wheeler-metoden". Først lages en indeks av contig-fila som skal "rapidly narrow the list of candidate alignment locations" [56] oversatt til norsk: raskt forminske mulige steder på contigene der reads alignes. Deretter alignes readene til indeksen og alignmentene utvides til et større alignment. Dette er en rask måte å aligne på, og indeksen tar lite plass å lagre [56].

Bowtie2 kjøres på contigene i studien for å evaluere assemblyene. Illumina-readene mappes til en indeks av assemblyet, og alignmentet lagres i ei sam-fil. SAMtools konverterer filtypen sam til bam for å komprimere fila og deretter sorterer readene etter hvor de mapper til referansen (BIN310). Programmet MetaBAT (v0.26.3) har funksjonen "jgi_summarize_bam_contig_depths" som brukes til å kalkulere dybden av reads på contigene [66]. Denne funksjonen kjøres på alignmentene og genererer read-dybder som brukes til å sammenligne assemblyene.

QUAST er et program som kan brukes til å evaluere assemblyer både med eller uten referansesekvens [58]. Programmet aligner contigene mot referansesekvensen og resultatet er informasjon om contigene og informasjon om alignmentet med referansen. QUAST gir ut det som i denne oppgaven kalles identitet, men som de i artikkelen kaller "genome fraction" [58]. Direkte oversatt fra artikkelen er genome fraction "totalt antall baser i contigene som aligner på referansen, delt på lengden av referansen. En base i referansen regnes som alignet dersom minst én contig har ett alignment til denne basen". Det står også at repeterte områder kan bidra til at genome fraction øker.

MetaQUAST er en utvidelse av QUAST der assembly alignes mot flere referanse-sekvenser, i utgangspunktet ment for metagenom. Identiteten er fortsatt den samme, men MetaQUAST gir ut én identitet for hvert alignment (hver referanse-sekvens) og summerer resultatene for assembly med et gjennomsnitt av alle identitetene.

3 Resultater

3.1 Assemblering av plasmider med repeterte områder

For å undersøke hvilke lengder av repetert område som skaper utfordringer for assemblering ble plasmider med repetert område av ulike lengder konstruert, og simulerte Illumina-reads fra disse plasmidene ble assemblert. I første assemblering ble det brukt både simulerte feilfrie Illumina-reads og Illumina-reads med feilrate tilsvarende feilraten til MiSeq-v3-reads (kalles heretter feil-reads). Utvalgte resultatene fra evalueringen av assemblyene vises i tabell 3 og tabell 4 på neste side. Figurene er basert på resultater fra assemblering av feil-reads. Illumina-readene ble simulert med dybde 100x. I andre omgang ble også Nanopore-reads fra de samme plasmidene simulert, og med trinnvis økende dybde fra 5x og oppover. Deretter ble både Illumina feil-reads og Nanopore-readene assemblert sammen til hybridassemblyer. Hybridassemblyene er evaluert og resultatene vises i tabell 6 på side 26.

I tillegg ble forlengede reads av de simulerte Illumina feil-readene generert med FLASH og gitt til SPAdes. Disse forlengede readene ble altså gitt i tillegg til Illumina feil-readene. Dette ble kun gjort for resultatene som ga flere enn én contig i tabell 4 og resultatene vises i tabell 5 på neste side. Det samme ble også gjort i Hybridassembly med Nanopore-read-dybde på 15x, med plasmidet med lengst repetert område. Resultatet vises i tabell 7 på side 27.

3.1.1 Assemblering av Illumina-reads

Resultatene fra assemblering av både feilfrie reads og feil-reads viser at assembly av plasmid med repetert område med lengde 801 bp eller lengre gir fler enn én contig, lavere identitet enn 98 % og varierende read-dybde. Dette vises i tabell 3, tabell 4 og i figur 8 på side 25. I tabell 4 er den største gjennomsnittlige read-dybden til plasmid *K901* på 808 som er ca. 10 ganger så høy read-dybde som for plasmidene som ga én contig. Etter å ha undersøkt alle datasettene er resultatet at én contig per simulering hadde gjennomsnittlig read-dybde på rundt 810 og hadde lengde 901 bp.

Plasmidet *K701* i tabell 4 har lavere identitet og kortere total assembly-lengde enn plasmidene med kortere repetert område. Derimot er resultatene fra assemblering av feilfrie reads fra *K701* er like resultatene for de plasmidene med kortere repetert område; *K601* og kortere. Read-dybden til *K701* i tabell 4 er likevel like høy og assembly ga én contig som de andre plasmidene med kortere repetert område. Forskjell mellom resultatene fra *K701* og plasmidene med kortere repetert område vises også i figur 8 og figur 9 på side 25. I figur 9 synker den røde linja litt på 701 bp.

Assemblering med forlengede reads i tillegg til R1- og R2-reads ga like mange eller fler contiger enn med assemblering av kun R1- og R2-reads. De plasmidene som ble assemblert til flere contiger ble assemblert med forlengede reads i tillegg til R1- og R2-readene. Resultatet av evalueringene vises i tabell 5 og utenom antall contiger er resultatene ganske like de i tabell 4.

De fleste resultatene av assemblering av feilfrie reads har høyere identitet og read-dybde enn feil-readene. Dette vises i tabellene 3 og 4.

Tabell 3: Gjennomsnittlige resultater fra assemblering av feilfrie reads som er simulert 10 ganger fra plasmid-sekvenser med økende lengde på konstruert repetert område. Referanse viser til utgangspunkt-sekvensen uten konstruert repetert område. Disse readene ble simulert med sekvenseringsdybde 100x og den totale lengden på utgangspunkt-plasmidet er 91109bp. Identitet er prosent baser i sekvensen som de assemblerte readene er simulert fra, som er alignet av assembly. Gj. snitt read-dybde er gjennomsnittlig antall Illumina-reads som dekker én posisjon i assembly. Dersom alle contigene sorteres etter lengde og legges etter hverandre på én rekke vil lengden av den korteste contigen i midten av denne rekka være N50. Cellene som er farget med rød er de som viser flere enn én contig.

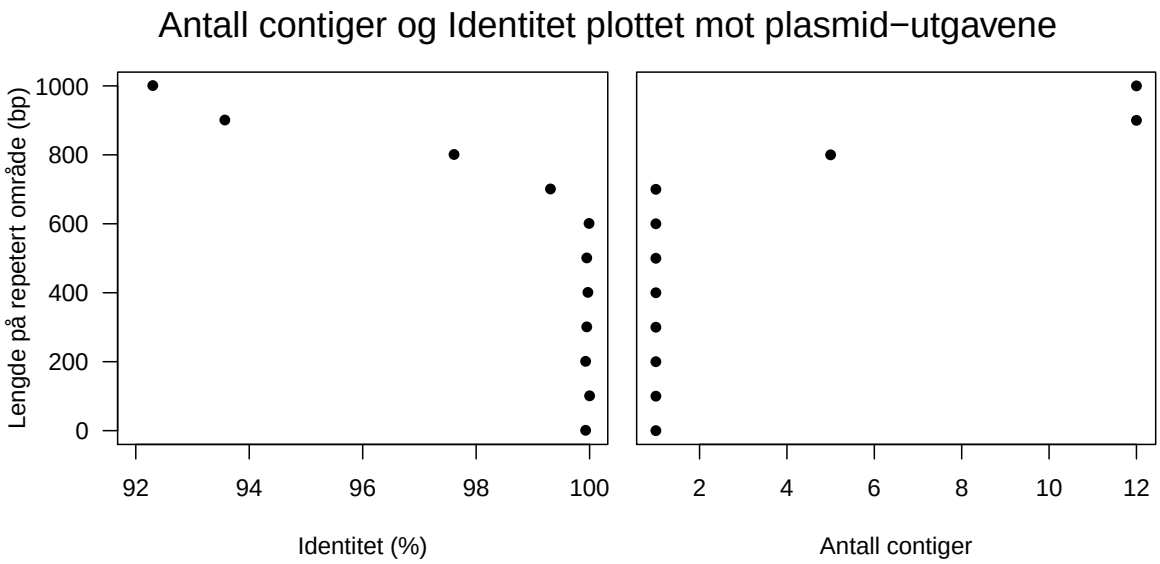
	Referanse	K201	K601	K701	K801	K901
Identitet (%)	99.98	100.00	99.99	99.99	97.84	93.58
Antall contiger	1	1	1	1	4	12
Minste gj. snitt read-dybde	99	99	100	100	3	74
Største gj. snitt read-dybde	99	99	100	100	106	1000
Total assembly-lengde (bp)	91089	91107	91100	91101	89485	85516
Korteste contig (bp)	91089	91107	91100	91101	251	549
Lengste contig (bp)	91089	91107	91100	91101	56350	18482
N50 (bp)	91098	91107	91100	91101	16442	5547

Tabell 4: Resultater fra assemblering av reads med normal feilrate. Tabellen viser gjennomsnittsverdier fra 10 simuleringer for hvert plasmid. Detaljer om tabellen står i tabellteksten til tabell 3.

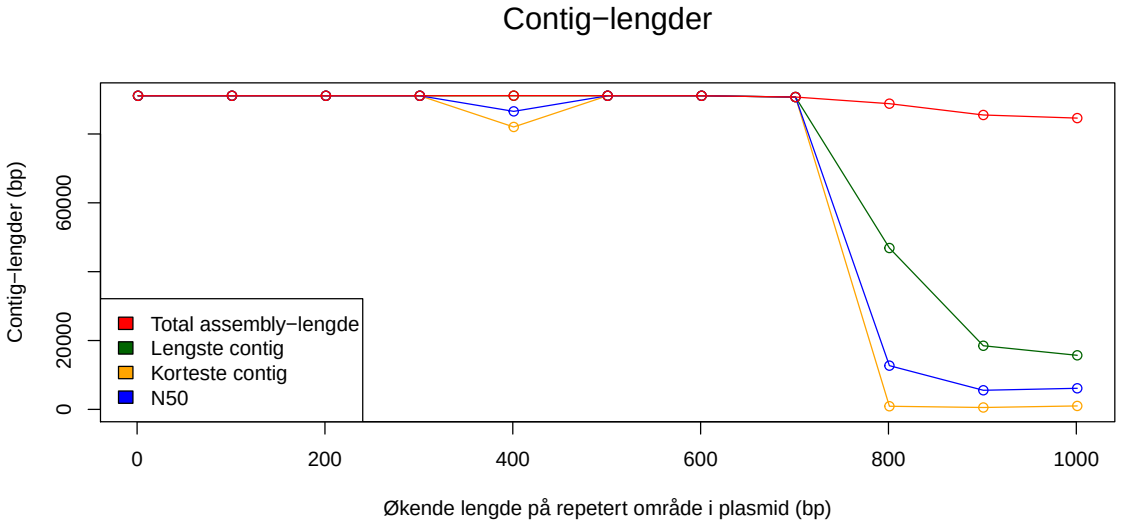
	Referanse	K201	K601	K701	K801	K901
Identitet (%)	99.96	99.98	99.97	99.56	97.30	93.56
Antall contiger	1	1	1	1	5	12
Minste gj. snitt read-dybde	82	82	82	82	47	69
Største gj. snitt read-dybde	82	82	82	82	86	808
Total assembly-lengde (bp)	91095	91096	91098	90725	88801	85511
Korteste contig (bp)	91095	91096	91098	90725	908	549
Lengste contig (bp)	91095	91096	91098	90725	46871	18482
N50 (bp)	91095	91096	91098	90725	12700	5547

Tabell 5: Viser gjennomsnittlige resultater fra assemblering av Illumina-reads og forlengede Illumina-reads. Kun de tre plasmid-variantene med repetert område på 801-1001 bp er tatt med, fordi det var de som resulterte i flere contiger. De forlengede readene ble gitt i eget bibliotek som enkelt-reads. Detaljer som tabellen står i tabellteksten til tabell 3.

	K801 + flash	K901 + flash	K1001 + flash
Identitet (%)	97.29	93.59	92.56
Antall contiger	6	12	12
Minste gj. snitt read-dybde	47	69	68
Største gj. snitt read-dybde	86	808	821
Total assembly-lengde (bp)	89482	85511	84623
Korteste contig (bp)	908	549	914
Lengste contig (bp)	44786	18482	15703
N50 (bp)	12929	5547	6000



Figur 8: Identitet (%) og antall contiger fra assemblering av alle plasmid-utgavene fra referanse til *K1001* vist på y-aksen som 0-1001. Verdiene er gjennomsnittlige fra 10 simuleringer av hver plasmid-utgave.



Figur 9: Total lengde på assembly, korteste og lengste contig og N50 fra alle plasmidene med økende lengde på repetert område langs x-aksen. På *K401* ble 1 av de 10 datasettene assemblert til to contiger. Den ene contigen hadde samme lengde som de andre contigene i de 9 andre datasettene, mens den andre var kort på 450 bp og hadde read-dybde 0.

3.1.2 Hybridassemblering av Illumina- og Nanopore-reads

Resultatene i tabell 6 viser gjennomsnittsverdier fra hybridassemblering av referansen og de plasmidene som ga flere enn én contig ved assemblering av kun Illumina-reads vist i tabell 4 på side 24. Illumina-readene ble simulert med dybde 100x, mens Nanopore-readene er simulert med systematisk økende dybde fra 5x. HybridSPAdes assemblerte også Nanopore-reads med dybde opp til og med 200x, men resultatene forble som ved dybde 15x.

Tabell 6: Gjennomsnittlige resultater fra 10 simulerte hybridassemblyer av plasmidene; Referanse, *801*, *901* og *1001*. Referanse er plasmid uten konstruert repetert område. Dybden av Illumina-reads var 100, og dybden av Nanopore-reads er 5, 10 og 15x. Identitet er prosent baser i sekvensen som de assemblerte readene er simulert fra, som er alignet av assembly. Read-dybde er gjennomsnittlig antall Illumina-reads som dekker én posisjon i assembly. Dersom alle contigene sorteres etter lengde og legges etter hverandre på én rekke vil lengden av den korteste contigen i midten av denne rekka være N50. De røde cellene skal indikere at assembly har flere enn én contig.

	Referanse	K801	K901	K1001
Nanopore-read-dybde 5				
Identitet (%)	99.96	99.85	99.98	99.57
Antall contiger	1	1.40	1.60	1.50
Minste gj. snitt read-dybde	82	81	80	82
Største gj. snitt read-dybde	82	82	82	82
Total assembly-lengde (bp)	91095	91676	92085	91298
Korteste contig (bp)	91095	65389	65151	65732
Lengste contig (bp)	91095	81435	77989	77894
N50 (bp)	91095	73412	68570	71165
Nanopore-read-dybde 10				
Identitet (%)	99.96	99.73	99.99	99.57
Antall contiger	1	1.60	1	1
Minste gj. snitt read-dybde	82	78	81	82
Største gj. snitt read-dybde	82	82	81	82
Total assembly-lengde (bp)	91095	91635	92010	91235
Korteste contig (bp)	91095	82820	92010	91235
Lengste contig (bp)	91095	85378	92010	91235
N50 (bp)	91095	83630	92010	91235
Nanopore-read-dybde 15				
Identitet (%)	99.98	99.99	99.98	99.57
Antall contiger	1	1	1	1
Minste gj. snitt read-dybde	82	81	81	82
Største gj. snitt read-dybde	82	81	81	82
Total assembly-lengde (bp)	91100	91911	92010	91235
Korteste contig (bp)	910100	91911	92010	91235
Lengste contig (bp)	91100	91911	92010	91235
N50 (bp)	91100	91911	92010	91235

Hybridassembly med forlengede Illumina-reads ga flere contiger enn med kun Illumina- og Nanopore-reads ved dybde 15x. Kun plasmidet med repetert område på 1001 bp ble assemblert med HybridSPAdes og med forlengede Illumina-reads generert fra FLASH. Gjennomsnittsresultater fra denne assembleringen vises i tabell 7.

Tabell 7: Gjennomsnitt-resultat av hybridassembly for plasmid-utgaven K1001 med Nanopore-reads med dybde 15x og med forlengede Illumina-reads generert av flash. De forlengede readene ble gitt i et eget bibliotek som enkelt-reads.

	Hybridassembly av K1001 + flash
Identitet (%)	99.00
Antall contiger	2
Minste gj. snitt read-dybde	82
Største gj. snitt read-dybde	84
Total assembly-lengde (bp)	90598
Korteste contig (bp)	49213
Lengste contig (bp)	68065
N50 (bp)	59122

3.2 Assemblering av plasmider med ulike shufflon-varianter

For å undersøke assembleringen av reads fra plasmider med hver sin utgave av shufflonet ble slike plasmider konstruert. 5 plasmider med 5 ulike shuffloner i ble konstruert og simulerte reads fra disse plasmidene ble assemblert. I tillegg ble 5 identiske plasmider brukt som referanse. Shufflon-sekvensen er fra IncI1-plasmidet R64 som består av 7 ulike deler som for hver versjon av shuffloner har ny rekkefølge, se avsnitt 1.1 på side 8 for mer beskrivelse. I avsnitt 3.2.1 beskrives resultatene fra assemblering av Illumina-reads og i avsnitt 3.2.2 på side 31 beskrives hybridassembly med både Nanopore-reads og Illumina-reads. Fordi plasmidene med shufflon-variantene minner om metagenom kjøres i tillegg MetaSPAdes, også samtidig som HybridSPAdes. Verdiene i resultatene er produsert med Bowtie2 og MetaQUAST.

Det er også generert forlengede reads av de overlappende parvise Illumina-readene. De forlengede Illumina-readene ble gitt i tillegg til R1- og R2-readene. I manualen til SPAdes står det at forlengede reads skal gis med opsjonen -m, men det ble prøvd ut flere måter.

3.2.1 Assemblering av Illumina-reads

I tabell 8 på neste side vises gjennomsnittlige resultater fra assemblering av plasmidene med SPAdes og MetaSPAdes. I tabell 9 på neste side vises gjennomsnittlige resultater fra assemblering med de forlengede readene i tillegg til R1 og R2. Gjennomsnittlig read-dybde og contig-lengde for hver contig vises i figur 10 på side 30.

Figur 10 viser at MetaSPAdes ga lengre korte contiger med lavere read-dybde enn SPAdes ved assemblering av simulerte Illumina-reads fra plasmidene med ulike versjoner av shufflonet. Ifølge tabell 8 gir MetaSPAdes flere og kortere contiger enn SPAdes, og det er stor forskjell mellom N50-verdiene, der SPAdes ga mye lengre N50-contig. Men første vindu i figur 10 viser at mange av de korte contigene fra MetaSPAdes har read-dybde 0, at ingen Illumina-reads mapper til disse contigene. Forutenom disse contigene med read-dybde 0 har flere av de korte contigene fra MetaSPAdes lengde nærmere lengden av shufflonet.

Resultatene i figur 10 viser at assemblyene fra alle assemblerne utenom MetaSPAdes ga contiger på lengde med plasmidet, med eller uten shufflonet. De lange contigene fra MetaSPAdes har lengder som spenner fra 15 000 til ca. 75 000. Likevel er den gjennomsnittlige identiteten til assemblyene fra MetaSPAdes høyest.

Contigene med read-dybde over 0 ble sammenlignet med sekvensen av shufflonet som ble brukt til å konstruere shufflon-variantene. Den web-baserte Blast-søkeren på NCBI sine nettsider gjorde sammenligningen. Resultatet for SPAdes' contiger av kun R1- og R2-reads var at de korte contigene bestod av deler av shufflonet, mens de lange contigene inneholdt deler av shufflonet. Resultatet for MetaSPAdes' contiger var også at de korte contigene bestod av deler av shufflonet og én av contigene per simulering var lengre enn shufflonet og inneholdt hele shufflon-sekvensen. De lange contigene inneholdt ikke deler av shufflonet og lengste contig er kortere for MetaSPAdes enn SPAdes i tabell 8.

Det er små forskjeller mellom assemblyer der forlengede reads blir gitt i samme bibliotek som R1- og R2-reads og de forlengede readene blir gitt i eget bibliotek. Tabell 9 viser resultater fra assemblering av simulerte Illumina-reads og forlengede Illumina-reads fra 5 plasmider med hvert sitt shufflon. Resultatene fra assemblering der de forlengede readene ble gitt med enten -m (*merged*) eller -s (*single*) i samme par-bibliotek som R1- og R2-readene er helt like. Der de forlengede readene ble gitt i et eget par-bibliotek som enkle reads (-s) ble resultatet litt annerledes. Men av 10 simuleringer var det 2/10 fra assemblering av forlengede reads i eget par-bibliotek som var annerledes enn forlengede reads i samme par-bibliotek som R1 og R2.

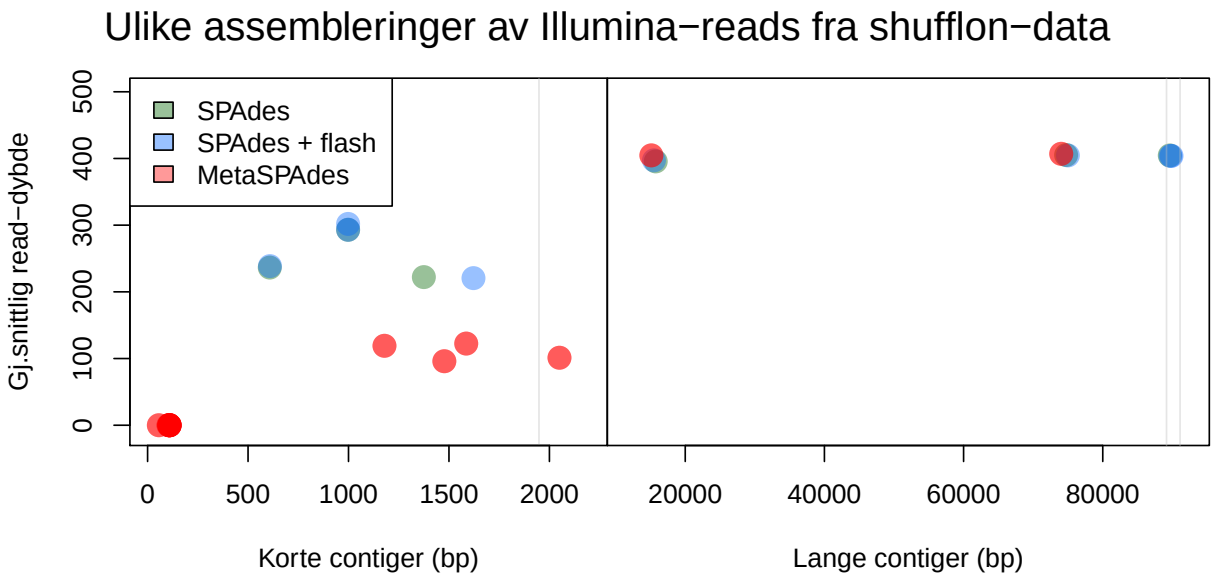
Read-dybden ligger på maksimalt 400 og minimum 0. Figur 10 viser at de lange contigene har read-dybde på rundt 400, og at de korteste contigene fra MetaSPAdes har read-dybde 0.

Tabell 8: Gjennomsnittlige resultater fra assemblering av kun Illumina-reads fra 5 plasmider. Referanse S er SPAdes på R1- og R2-reads fra fem identiske plasmider, Shuffloner S betyr SPAdes og Shuffloner M betyr MetaSPAdes på de 5 plasmidene med ulike shufflon-utgaver. Identitet er prosent i sekvensen som de assemblerte readene er simulert fra, som er alignet av assembly. Identiteten som vises i tabellen er gjennomsnittlig identitet fra alle gjennomsnittlige identitene fra hver aligning mellom assembly og de 5 plasmid-sekvensene med MetaQUAST. Gjennomsnittlig read-dybde er gjennomsnittlig antall Illumina-reads som dekker én posisjon i assembly. Dersom alle contigene sorteres etter lengde og legges etter hverandre på én rekke vil lengden av den korteste contigen i midten av denne rekka være N50. Skriftfargene samsvarer med fargene i figur 10 på neste side

	Referanse S	Shuffloner S	Shuffloner M
Identitet (%)	98.32	93.86	99.36
Antall contiger	1	3	12
Minste gj.snitt read-dybde	401	249	0
Største gj.snitt read-dybde	401	409	412
Total assembly-lengde (bp)	91361	91475	96305
Korteste contig (bp)	91361	2421	56
Lengste contig (bp)	91361	79512	74033
N50	91361	18678	684

Tabell 9: Gjennomsnittlige resultater fra assemblering av Illumina-reads fra shufflon-variantene med forlengede reads av disse gitt på ulike måter til SPAdes. -m eller -s viser til om de forlengede readene ble gitt som henholdsvis merged/forlengede reads eller single/enkle reads. Samme bib vil si samme bibliotek som R1- og R2-readene ble gitt i, og eget bib vil si at de forlengede readene ble gitt i et annet bibliotek enn R1 og R2.

	-m samme bib	-s samme bib	-s eget bib
Identitet (%)	92.70	92.70	96.05
Antall contiger	3	3	3
Minste gj. snitt read-dybde	229	229	195
Største gj. snitt read-dybde	409	409	409
Total assembly-lengde (bp)	91475	91475	91500
Korteste contig (bp)	2411	2411	931
Lengste contig (bp)	78109	78109	82541
N50 (bp)	20091	20091	17164



Figur 10: Hvert punkt i figuren er én contig og figuren viser resultat fra ulike typer assembly av Illumina reads fra de fem plasmidene med ulike varianter av shufflonet. De grønne punktene er contiger fra assemblering av SPAdes, de blå fra SPAdes med forlengede reads som ble gitt i et eget bibliotek, se tabell 9 på side 29. De røde punktene fra MetaSPAdes. For hver assemblering er like resultater gruppert, og ett resultat fra hver gruppe er vist i plottet. Det er 2-3 grupper i disse dataene. De grå vertikale linjene viser henholdsvis lengdene i basepar for shufflonet i det første vindu, og utgangspunkt-plasmidet uten shufflon og utgangspunkt-plasmidet med shufflon i det andre vindu av figuren. I vinduet med de lange contigene er alle punktene med alle fargene nesten oppå hverandre utenom det siste punktet der det ikke er et rødt punkt. Resultatene fra assemblering av plasmidene med ulike varianter av shufflonet gir korte contiger ikke lengre enn 2000 bp og lange contiger på 15 000 bp til 92 000 bp.

3.2.2 Hybridassemblering av Illumina- og Nanopore-reads

Både simulerte Illumina-reads med dybde 100x og Nanopore-reads med økende dybde ble assemblert med HybridSPAdes til et hybridassembly og resultatene vises i tabell 10. Det ble også assemblert med Nanopore-read-dybde 100x, men det assembly ga samme resultater som ved 20x. Nanopore-readene kan også filtreres og trimmes før assemblering, og det ble gjort et forsøk med assembleren Canu, men som ikke utgjorde forskjell for assembly av plasmidene med ulike utgaver av shufflonet. Dette resultatet er ikke tatt med.

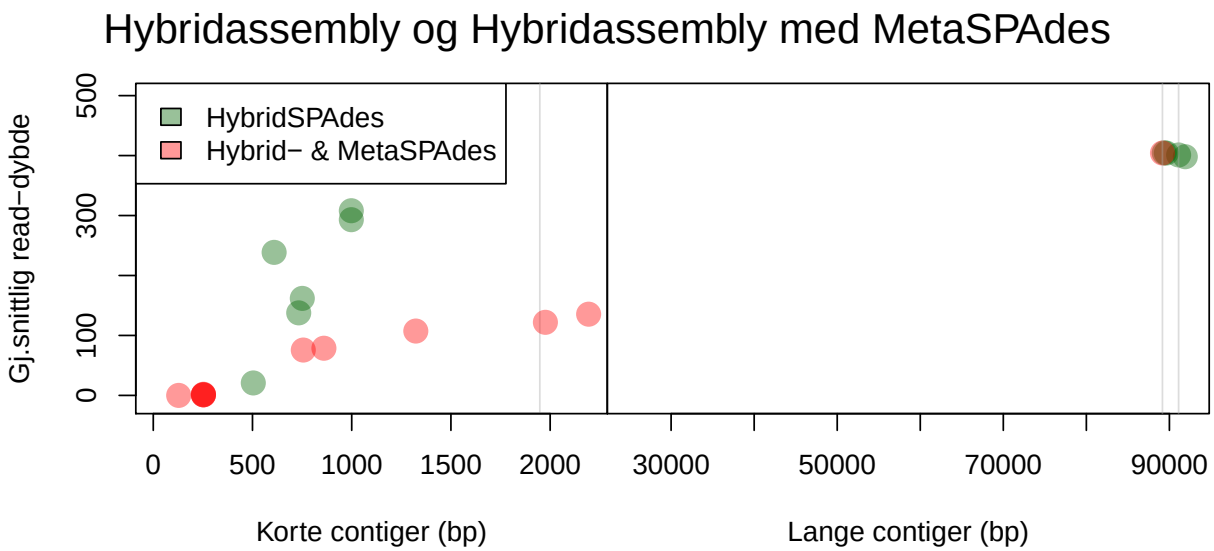
Hybridassembly med HybridSPAdes består av lange contiger med lengde som plasmidet, men ellers gir ikke hybridassembly bedre assembly enn SPAdes på kun Illumina-reads. Figur 11 på neste side viser at hybridassembly gir lange contiger på lengde med plasmid-sekvensen både med og uten shufflon. Sammenlignet de blå punktene i figur 11 med de grønne og blå punktene i figur 10 på side 30 er assembly veldig likt. De korte contigene ble sammenlignet med shufflon-sekvensen med BLAST og de besto av kun deler av shufflonet.

Tabell 10: Gjennomsnittlige resultater fra hybridassembly med HybridSPAdes. 5x, 10x, 15x, og 20x er read-dybdene for Nanopore-readene som ble simulert. Illumina-readene ble simulert med en dybde på 100x som tidligere. Identitet er prosent baser i sekvensen som de assemblerte readene er simulert fra, som er alignet av assembly. Gjennomsnittlig read-dybde er gjennomsnittlig antall Illumina-reads som dekker én posisjon i assembly. Dersom alle contigene sorteres etter lengde og legges etter hverandre på én rekke vil lengden av den korteste contigen i midten av denne rekka være N50.

	5x	10x	15x	20x
Identitet (%)	99.45	99.38	99.38	99.48
Antall contiger	3.10	2.90	2.90	2.80
Minste gj. snitt read-dybde	148	161	161	179
Største gj. snitt read-dybde	408	407	407	407
Total assembly-lengde (bp)	91669	91644	91644	91632
Korteste contig (bp)	18730	18842	18842	27817
Lengste contig (bp)	90063	90111	90112	90262
N50 (bp)	23560	28041	28041	32604

Tabell 11: Den første kolonnen med resultater er fra hybridassembly metagenom-modus med dybde på Nanopore-readene på 5x. Det ble også forsøkt med dybde på Nanopore-readene på 20x, 100x og 200x, men resultatene var de samme som ved 5x. De to siste kolonnene er fra hybridassembly med forlengede reads gitt i samme bibliotek som R1 og R2, eller i eget bibliotek som enkelt-reads.

	Meta 5x	Samme -m	Eget bib -s
Identitet (%)	99.53	99.48	99.48
Antall contiger	12.4	2.90	2.90
Minste gj. snitt read-dybde	0	141	141
Største gj. snitt read-dybde	409	407	407
Total assembly-lengde (bp)	97546	91627	91627
Minste contig (bp)	128	18676	18676
Største contig (bp)	89166	90206	90206
N50 (bp)	404	28005	28005



Figur 11: Figuren viser resultater fra to typer assembly. De grønne punktene er contiger fra HybridSPAdes på Illumina- og Nanopore-reads. De røde punktene er contiger fra hybridassembly i metagenom-modus, altså HybridSPAdes og MetaSPAdes samtidig. Dybden av Nanopore-reads er 20x i begge assemblyene. De tre grå, vertikale linjene indikerer henholdsvis shufflon-lengde, plasmid-lengde uten shufflon og plasmid-lengde med shufflon.

Figur 11 viser at enkelte korte contiger fra MetaSPAdes sammen med HybridSPAdes var lengre og nærmere lengden av shufflonet enn contiger fra kun HybridSPAdes. De korte contigene har veldig varierende lengde både kortere og lengre enn shufflonet, men enkelte er omtrent like lange som shufflonet. De korte contigene er sammenlignet med shufflon-sekvensen. Sammenligningen viser at de contigene som er litt lengder enn shufflonet inneholder hele shufflon-sekvensen, mens de som er kortere inneholder kun deler av shufflonet. Det er kun 1 eller 2 contiger som er lengre enn shufflon-lengden per simulering. Shufflon-sekvensen de ble sammenlignet med er sekvensen som ble brukt til å konstruere shufflon-sekvensene inn i plasmid-sekvensen. Resultatene fra HybridSPAdes med MetaSPAdes i figur 11 viser at de korte contigene, utenom de med read-dybde 0, for det meste er nærmere lengden av shufflonet enn for kun HybridSPAdes.

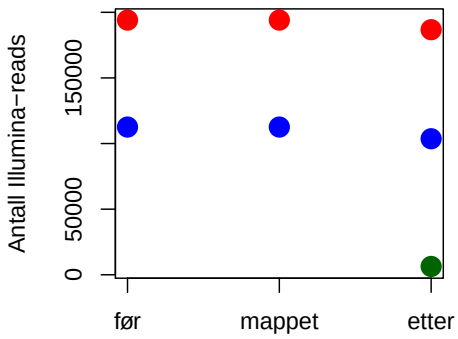
I de to siste kolonnene i tabell 11 på side 31 vises resultatene fra hybridassembly med forlengede Illumina-reads gitt på to måter. Resultatene fra disse to måtene er identiske, og de er samtidig veldig like resultatene i tabell 10 på side 31. N50 er lavere for assembly med de forlengede readene.

3.3 Assemblering av reelle data

De reelle dataene er reads sekvensert fra en prøve fra en kultur av *E. coli*. Fordi dette er reelle data som ikke er simulert kan vi ikke være helt sikre på innholdet i genomene som har brukt til å generere readene. Disse readene assembleres på flere måter med SPAdes, Hybrid- og MetaSPAdes. De reelle Illumina-readene filtreres og trimmes før assemblering, og denne prosessen kan gjøre at par-readene mister partneren sin og blir enkelt-read. Nanopore-readene kan også filtreres og trimmes før assemblering. Resultatene fra de ulike assembliene av Illumina-reads er beskrevet i avsnitt 3.3.1 og ulike hybridassemblyer av Illumina- og Nanopore-reads er beskrevet i avsnitt 3.3.2 på side 35.

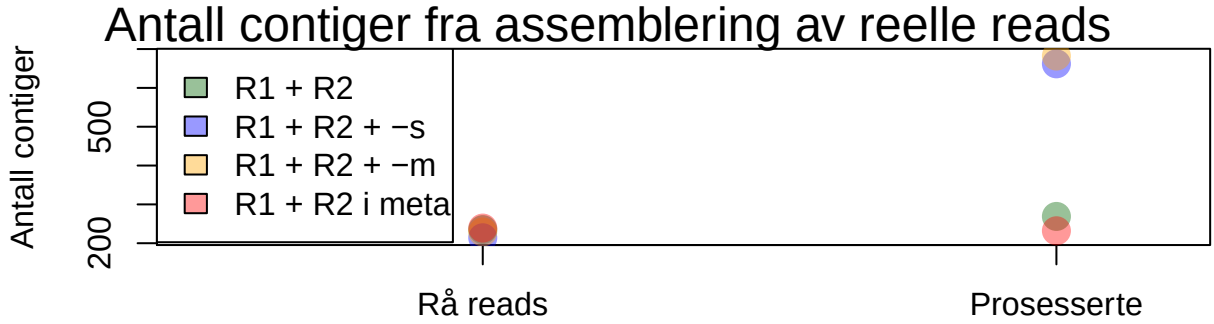
3.3.1 Assemblering av reelle Illumina-reads

Alle figurene er basert på tabellene 12 og 13 i vedlegg i avsnitt 7 på side 58. Figur 12 viser at antall reads før og etter mapping er likt, som betyr at ingen av Illumina-readene mappet til menneske-genomet. Antall reads er mindre etter korrigering. Figur 13 viser at antall contiger fra assemblering med forlengede reads øker mye fra rå til prosesserte Illumina-reads, mens assemblering med R1 og R2 ikke økte like mye. MetaSPAdes var den eneste som assemblerte til færre contiger for prosesserte reads.

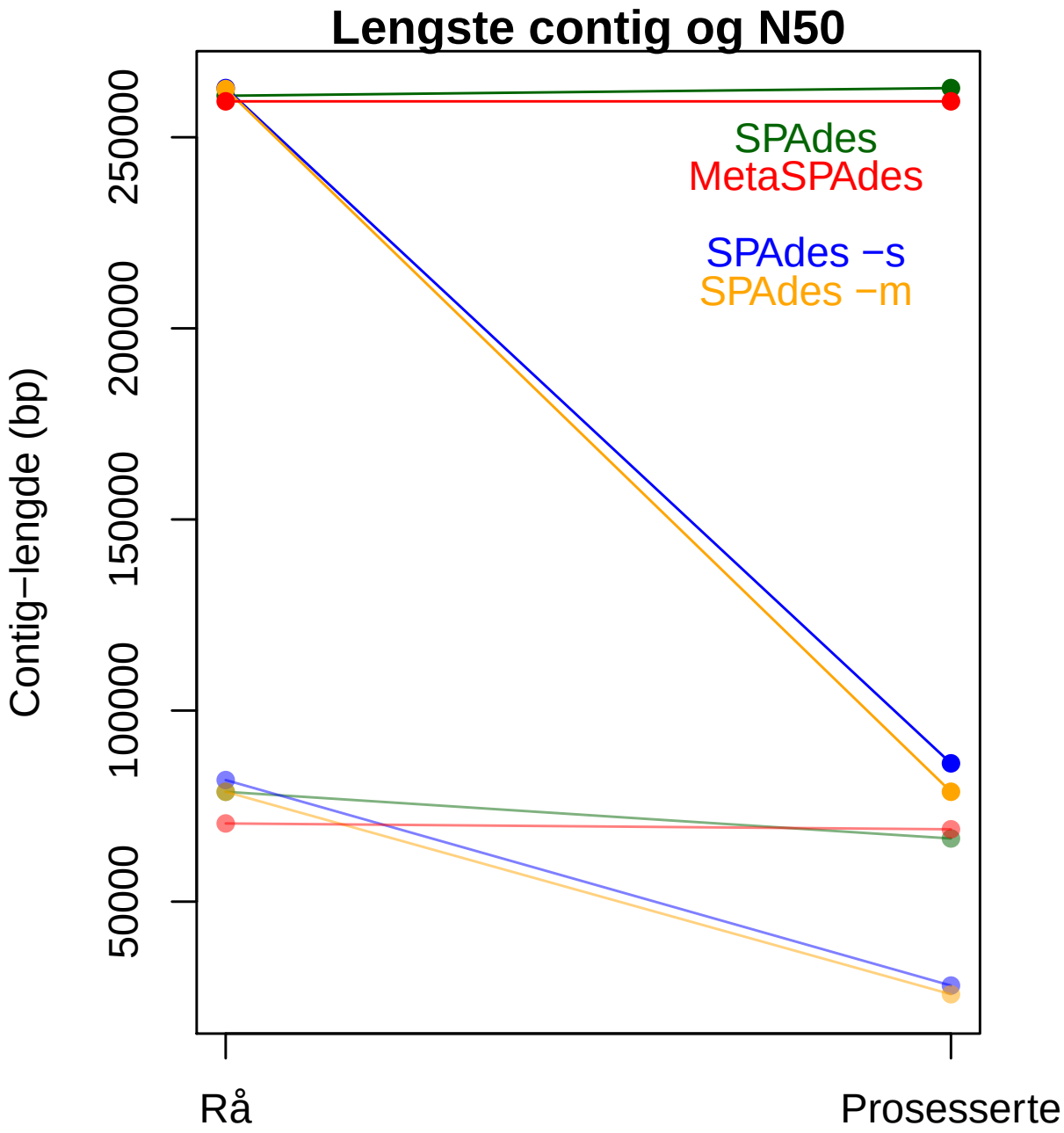


Figur 12: Antall Illumina-reads før prosessering, etter mapping mot menneske-genomet og etter prosessering. Røde punkter er antall R1- og R2-reads, blå punkter er antall forlengede reads kombinert av FLASH. Det ene grønne punktet viser antall parvise-reads som mistet partneren sin under prosessering.

Figur 14 på neste side viser at lengste contig og N50 stort sett sank fra rå til prosesserte reads. Lengden på lengste contig faller betydelig for assemblyene av de forlengede readene, men forblir på samme verdi med MetaSPAdes og blir faktisk lengre med SPAdes på kun R1 og R2. N50 sank veldig for assembly med forlengede reads, men ikke like mye for assemblering av kun R1 og R2 med SPAdes og MetaSPAdes. Antall contiger lengre enn 500 bp ble undersøkt for å finne ut om det fantes mange contiger med samme lengde. Høyeste forekomsten av contiger av samme lengde av de som var lengre enn 500 bp var 2.



Figur 13: Antall contiger fra alle assemblyene av reelle Illumina-reads. Grønne punkter er fra SPAdes på R1-og R2-readene. Blå og gule punkter er SPAdes på R1 og R2 med forlengede reads i tillegg, gitt på to ulike måter (enkelt-read-opsjonen eller ”merged”-read-opsjonen). De røde punktene er MetaSPAdes på R1 og R2.



Figur 14: Øverste fire linjene er lengde på lengste contig, mens nederste fire linjene er N50 for assemblyene av de reelle Illumina-readene før de ble prosessert (rå) og etter. Verdiene er basert på vedlegg tabell 12 på side 58 og tabell 13 på side 58.

3.3.2 Hybridassemblering av reelle Illumina- og Nanopore-reads

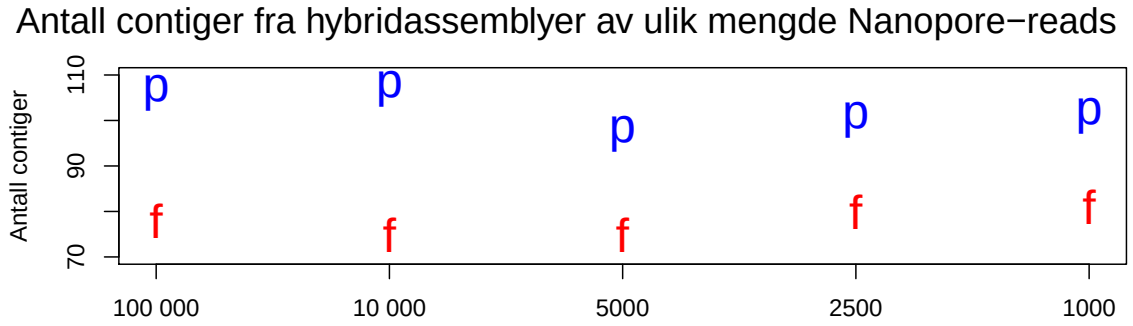
Det totale antallet Nanopore-reads er 1 350 595 og av disse ble det valgt ut de 100 000, 10 000, 5000, 2500 og 1000 lengste til assemblering. Hybridassemblyer ble gjort i tre deler, kalt A, B og C. I del A ble alle utvalgene av de lengste Nanopore-readene assemblert med både rå og prosesserte Illumina-reads. I del B assemblerte HybridSPAdes med MetaSPAdes de 100 000, 10 000 og 5 000 lengste Nanopore-readene med både rå og prosesserte Illumina-reads. I del C ble de 5 000 lengste Nanopore-readene trimmet og korrigert med Canu, og assemblert med flere assemblere sammen med både rå og prosesserte Illumina-reads.

Resultater felles for alle delene er at på det meste hadde to contiger over 500 bp samme lengde i hybridassemblyene. Det ble gjort en opptelling av antall contiger med samme lengde i alle hybridassemblyene, men det er kun én contig av hver lengde i alle assemblyene. Et utvalg av de korte contigene fra alle assemblyene ble alignet med shufflon-sekvensen som ble brukt ved simulering, og enkelte contiger inneholdt deler av shufflonet, men ingen av contigene besto av hele denne shufflon-sekvensen. Alle plottene i alle delene viser at ingen contiger hadde samme lengde som shufflonet som ble brukt til simulering. Ingen contiger hadde samme lengde som plasmidet på 91109 bp eller kromosomet på rundt 5 000 000 bp heller.

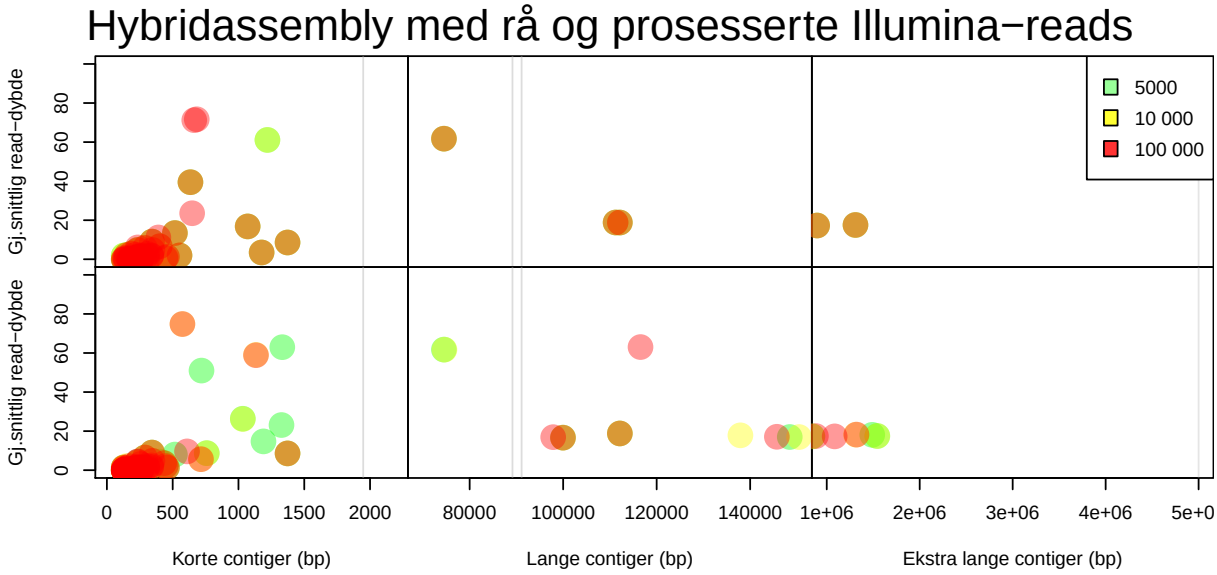
Del A 1000 - 100 000 lengste Nanopore-readene

Figur 15 viser at antall contiger er færrest i assembly med de 5000 og 10 000 lengste Nanopore-readene og rå Illumina-reads. Hybridassembly med de prosesserte Illumina-readene ga færrest contiger sammen med de 5000 lengste Nanopore-readene.

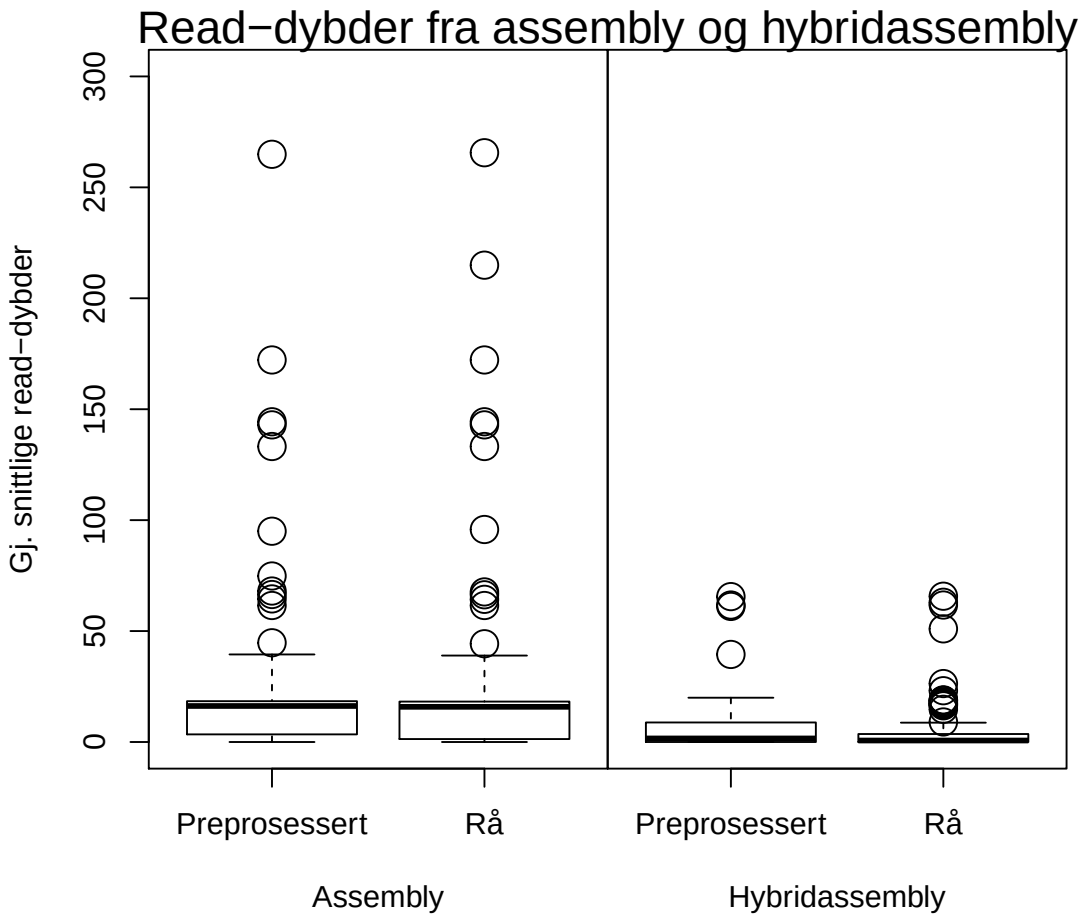
Figur 17 på side 37 viser at gjennomsnittlig read-dybde er høyere for assembly av kun Illumina-reads enn hybridassembly av de 5000 lengste rå Nanopore-readene.



Figur 15: Antall contiger fra hybridassembly med henholdsvis 100 000, 10 000, 5000, 2500 og 1000 av de lengste Nanopore-readene. P betyr prosesserte og f betyr før preprosessering. Plottet er basert på verdier fra tabell 14 på side 59 og tabell 15 på side 59.



Figur 16: Hvert punkt er en contig plottet med gjennomsnittlig read-dybde mot contig-lengde. Øverste rekke med plott er for rå og nederste rekke for prosesserte Illumina-reads brukt i hybridassembly med forskjellig antall av de lengste Nanopore-redaene indikert med hver sin farge. Gjennomsnittlig read-dybde er antall Illumina-reads som i gjennomsnitt dekker én posisjon på hver contig. Fullstendig oversikt over verdiene som er brukt til å lage plottet finnes i tilleggsdata tabell 16 på side 60, tabell 17 på side 61 og tabell 18 på side 62. Det er kun disse vinduene som vises av resultatene, fordi det blir for uoversiktelig å vise alle contigene.

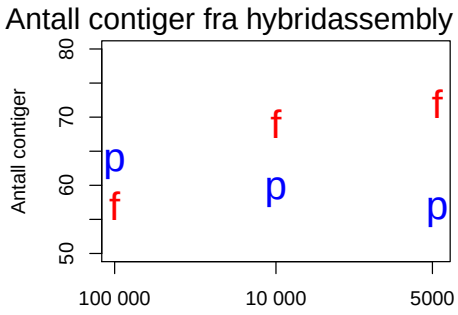


Figur 17: Gjennomsnittlige read-dybder fra både assembly av kun Illumina-reads og hybridassembly med de 5000 lengste rå Nanopore-readene. Boksplottene viser en størst spredning i gjennomsnittlig read-dybde for assembly av kun Illumina-reads.

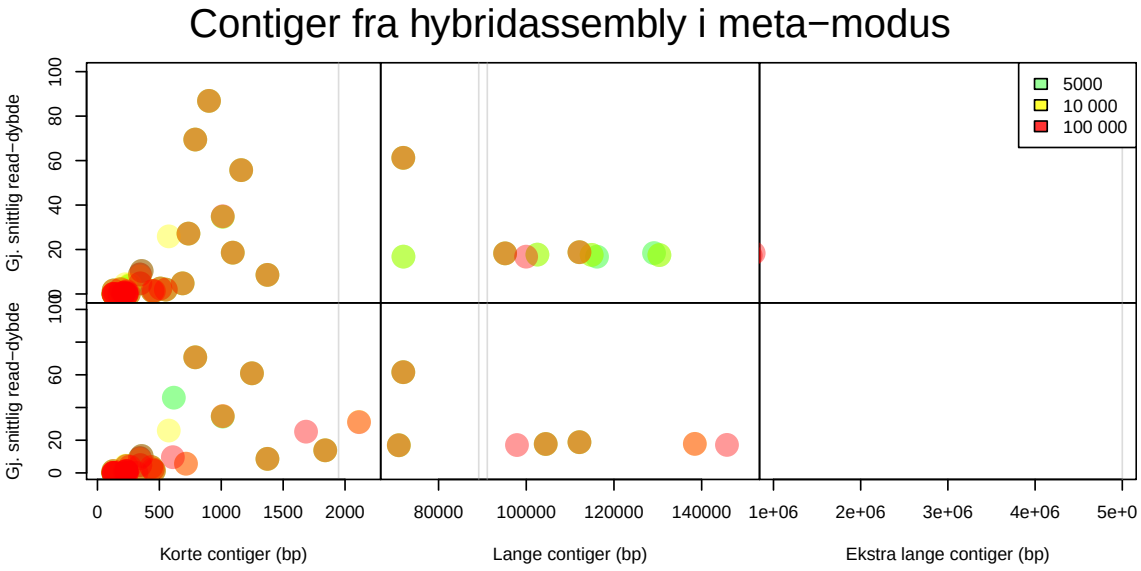
Del B MetaSPAdes samtidig som HybridSPAdes på 5000 - 100 000 Nanopore-reads

Assemblyer av de 100 000, 10 000 og 5000 lengste Nanopore-readene er også gjort med HybridSPAdes samtidig som MetaSPAdes. Resultatene fra disse assemblyene vises i figur 18 og figur 19. Utfyllende verdier for assembly kan finnes i tilleggsdata; tabell 16 på side 60, tabell 17 på side 61 og tabell 18 på side 62.

Figur 18 sammenlignet med figur 15 på side 35 viser at MetaSPAdes med HybridSPAdes gir stort sett færre contiger enn med kun HybridSPAdes. Unntaket er rå Illumina-read og 5000 lengste Nanopore-readene med Meta/HybridSPAdes som gir over 70 contiger samme som med assembly med kun HybridSPAdes. Ulikt resultatene i figur 15 var det assembly med prosesserte Illumina-reads og mengdene 10 000 og 5000 lengste Nanopore-reads som ga færrest contiger.



Figur 18: Resultater fra hybridassembly i meta-modus av rå (f) og prosesserte (p) Illumina-reads med ulik mengde av de lengste Nanopore-readene. Utfyllende verdier for assembly kan finnes i tilleggsdata; tabell 16 på side 60, tabell 17 på side 61 og tabell 18 på side 62.

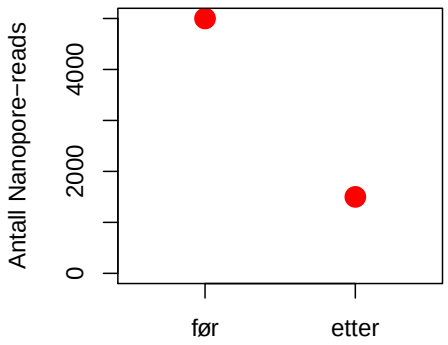


Figur 19: Hvert punkt er en contig med gjennomsnittlig read-dybde plottet mot contig-lengde. Øverste rekke med plott er for rå og nederste rekke for prosesserte Illumina-reads brukt i hybridassembly med forskjellig antall av de lengste Nanopore-redaene indikert med hver sin farge. Gjennomsnittlig read-dybde er antall Illumina-reads som i gjennomsnitt dekker én posisjon på hver contig. Fullstendig oversikt over verdiene som er brukt til å lage plottet finnes i tilleggsdata tabell 16 på side 60, tabell 17 på side 61 og tabell 18 på side 62.

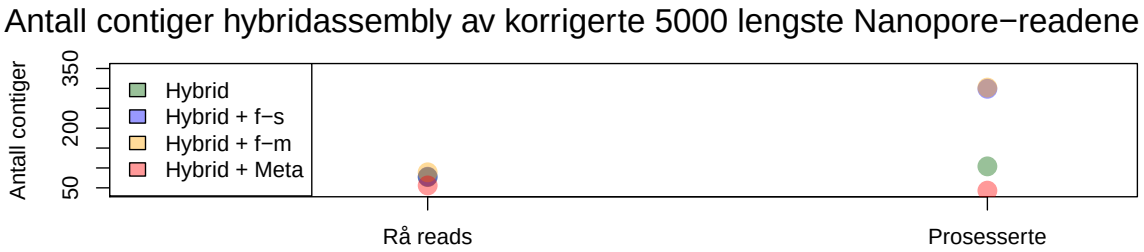
Del C Korrigerte 5000 lengste Nanopore-reads

De 5000 lengste Nanopore-readene ble korrigert og trimmet med Canu, og forskjell i antall reads før og etter vises i figur 20. Resultatene fra assemblering av de korrigerte og trimmede Nanopore-readene vises figur 22 og figur 23 på neste side.

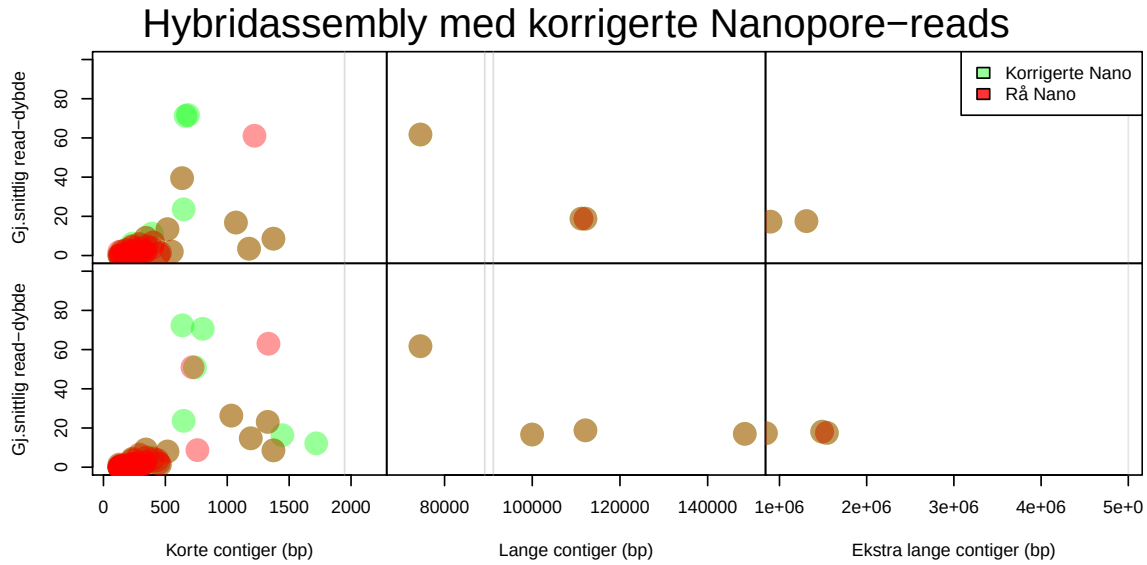
Figur 21 viser at korrigerte lengste 5000 Nanopore-reads og korrigerte Illumina-reads ga færrest contiger med HybridSPAdes samtidig med MetaSPAdes. Korrigerte Nanopore-reads ga ikke bedre assemblyer fra andre assemblere, enn Hybrid/MetaSPAdes.



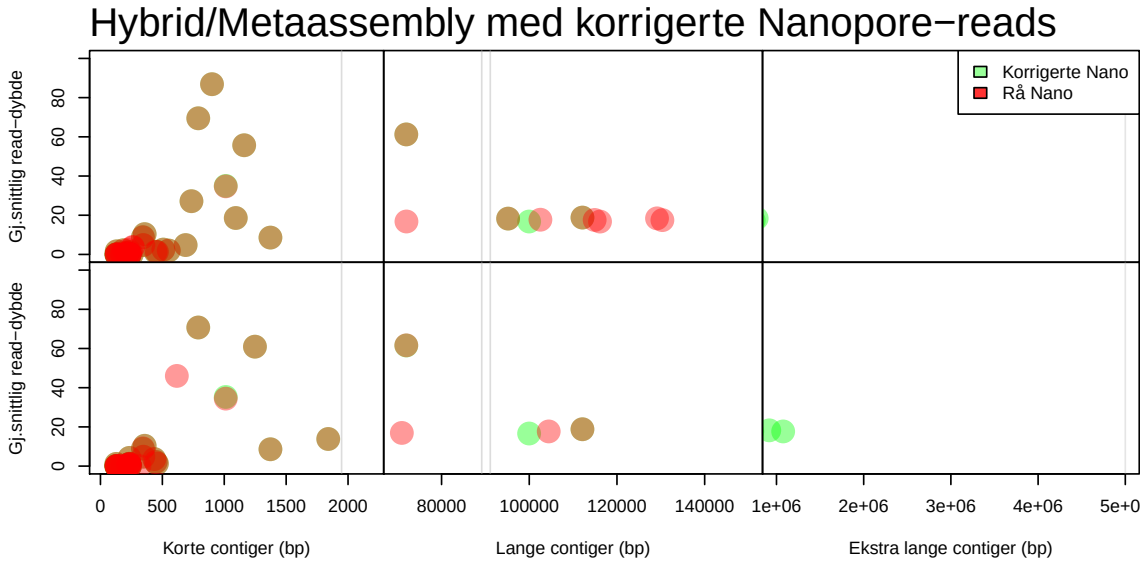
Figur 20: Antall Nanopore-reads før og etter prosessering.



Figur 21: Antall contiger for rå og prosesserte Illumina-reads i hybridassemblyer med korrigerte 5000 lengste Nanopore-readene.



Figur 22: Hvert punkt er en contig plottet med gjennomsnittlig read-dybde mot contig-lengde. Øverste rekke med plott er for rå og nederste rekke for prosesserte Illumina-reads brukt i hybridassembly med korrigert eller rå Nanopore-reads. Rå Nanopore-reads er de 5000 lengste, og de korrigerte er de samme men korrigert med Canu. Gjennomsnittlig read-dybde er antall Illumina-reads som i gjennomsnitt dekker én posisjon på hver contig. Fullstendig oversikt over verdiene som er brukt til å lage plottet finnes i tilleggsdata tabell 19 på side 63 og tabell 20 på side 63.



Figur 23: Hvert punkt er en contig plottet med gjennomsnittlig read-dybde mot contig-lengde. Dette er resultater fra HybridSPAdes og MetaSPAdes kjørt samtidig. Øverste rekke med plott er for rå og nederste rekke for prosesserte Illumina-reads brukt i hybrid/metaassembly med korrigert eller rå Nanopore-reads. Rå Nanopore-reads er de 5000 lengste, og de korrigerte er de samme men korrigert med Canu. Gjennomsnittlig read-dybde er antall Illumina-reads som i gjennomsnitt dekker én posisjon på hver contig. Fullstendig oversikt over verdiene som er brukt til å lage plottet finnes i tilleggsdata tabell 19 på side 63 ogtabell 20 på side 63.

4 Diskusjon

4.1 Assemblering av plasmidene med repetert sekvens

4.1.1 Assemblering av Illumina-reads

Illumina MiSeq v3-reads som er simulert fra et plasmid med repetert område vil være utfordrende å assemblere dersom det repeterte området er lengre enn fragmentlengde.

Ufullstendige assemblyer fra plasmider med RO lengre enn 701 bp og gjennomsnittlig fragmentlengde for de simulerte readene på 750 bp gir grunn til å bekrefte at RO lengre enn fragmentlengde gir utfordringer for assemblering. Fragmentlengde er lengden på fragmentet som et R1- og R2-par er sekvensert fra. Under simulering av Illumina-readene ble fragmentlengden satt til 750 bp. Ifølge Salzberg *et. al* [37] vil RO lengre enn fragmentlengden gi hull i assembly og dette stemmer med resultatene i denne oppgaven. Med read-lengde 250 bp gir RO på 701 bp og kortere gir én contig, mens RO på 801 bp og lengre gir flere contiger. Med en fragmentlengde på 750 bp, som er midt mellom 701 bp og 801 bp, er det grunn til å tro at fragmentlengden er avgjørende for assembly av RO. SPAdes utnytter da at hvert par av de parvise readene har en gitt avstand mellom hverandre og at enkelte par-reads overlapper hverandre delvis. Ifølge forfatterne av FLASH-programvaren er det nettopp dette FLASH skal hjelpe assembleren med å utnytte [43], men dersom assembleren gjør dette selv vil ikke forlengede reads fra FLASH tilføre ny informasjon til assembleren.

Assembly blir ikke bedre med forlengede reads, fordi de ikke gir ny informasjon til assembleren. Forfatterne i [43] skriver at forlengede reads fra FLASH bidro til bedre assembly, med lengre N50. Resultatene i tabell 5 på side 24 viser at N50 blir høyere for *K801*, men antall contiger øker sammenlignet med tabell 4 på side 24. Resultatene fra *K901* er veldig like som uten forlengede reads. I begge tilfellene kan det sies at assembly ikke blir forbedret med forlengede reads. Grunnen til det kan være at FLASH kombinerer reads som SPAdes hadde kommet til å gjøre uansett, sånn som tidligere beskrevet i forrige avsnitt. De forlengede readene blir gitt i tillegg til de parvise, men de forlengede readene vil ikke gi ny informasjon til assembleren. Dersom assembly gir flere contiger eller lavere identitet kan grunnen være at FLASH gjør feil ved skjøting av readene. Dette er mulig, fordi FLASH har en feilrate på litt under 1 % [43].

Grunnen til at Magoč hevder at skjøting av overlappende reads før assemblering har effekt kan være fordi FLASH ble utviklet før assemblere begynte å utnytte parvise reads. Ifølge Bankevich *et. al* [32] manglet det i 2012 assemblere som utnyttet parvise reads riktig. FLASH-programvaren ble utviklet i 2011 [43]. Derfor er det mulig at FLASH på det tidspunktet hadde en positiv effekt sammen med en assembler som ikke utnyttet de parvise readene riktig. På tross av resultatene nevnt i forrige avsnitt og at SPAdes utnytter parvise reads har dagens versjon av SPAdes en egen opsjon for input av slike ferdige skjøtede parvise reads [63]. To senere utviklede programmer [67] [68] som også skjøter sammen parvise reads står foreslått til å generere forlengede reads som skal gis inn med denne opsjonen [63].

Standardavviket til fragmentlengden kan være grunnen til at feil-reads fra plasmid *K701* gir assembly med lav identitet og kort total assembly-lengde. ART simulerte både feilfrie reads og reads med normal feilrate for MiSeq v3-reads, her kalt feil-reads. Sammenlignet med plasmidene med kortere RO som også ga én contig fikk *K701* lavere identitet og kortere total assembly-lengde. Readene ble simulert med gjennomsnittlig fragmentlengde 750 bp og standardavvik

100 bp, og derfor vil fragmentlengden i enkelte simuleringer være en del lavere enn 701 bp. I disse simuleringene kan RO på 701 bp bidra til assembleren gjør enkelte feil under assemblering som ikke gjøres med lengre fragmentlengder. Gjennomsnittlig read-dybde er derimot lik som de plasmidene med kortere RO, som betyr at alle plasmidene med kortere RO enn 750 bp har like mange Illumina-reads som dekker deres ene contig.

Høy read-dybde og lavere identitet på assembly som gir flere contiger kan være en konsekvens av RO. Både tabell 3 og tabell 4 på side 24 viser at når RO blir 801 bp eller lengre gir assemblering flere contiger, og read-dybdene varierer mellom contigene. For eksempel er den gjennomsnittlige største read-dybden til plasmid med RO på 901 bp omtrent 10 ganger så høy som de fleste andre read-dybdene. I tillegg viser det seg at de contigene med 10 ganger så høy gjennomsnittlig read-dybde har lengde 901 bp. I følge Phillippy *et. al* blir read-dybden henholdsvis høyere og lavere, dersom identiske reads blir tolka som at de kommer fra færre eller flere områder enn de gjør [39]. Dette kan bety at alle de 10 kopiene av RO kollapset til samme contig og derfor mapper readene fra alle kopienes områder til denne contigen og ga 10 ganger så høy read-dybde. Når kopiene av RO kollapser til én contig vil identiteten også synke, fordi assembly mangler disse kopiene.

Mengden feil i Illumina feil-readene kan være grunnen til lavere read-dybde og identitet i tabell 4 sammenlignet med tabell 3. Feil-readene ble simulert med samme feilrate som reads fra MiSeq v3 [51]. De fleste av read-dybdene i tabell 3 er oppimot like høye som den dybden readene ble simulert med, mens read-dybdene i tabell 4 ligger rundt 82. Alle Illumina-readene ble simulert med dybde 100x, og den lave read-dybden på 82 kan komme av at mengden feil i readene som skaper variasjon mellom sekvensen som reads ble simulert fra og assembly. Dette vises også i lavere identiteter i tabell 4 enn i tabell 3.

Assemblyer av plasmid med lengde på RO kortere enn 801 bp gir ikke nødvendigvis én contig. Det ene resultatet fra 10 simuleringer fra plasmid *K401* ga 2 contiger. Den lengste contigen var like lang som de contigene i datasettene som ga én contig. Den korteste contigen var ca. 450 bp med en read-dybde på 0. En read-dybde på 0 vil si at assembleren har konstruert en sekvens fra *k*-merer av readene som ikke har samme sekvens som readene.

4.1.2 Hybridassemblering av Illumina- og Nanopore-reads

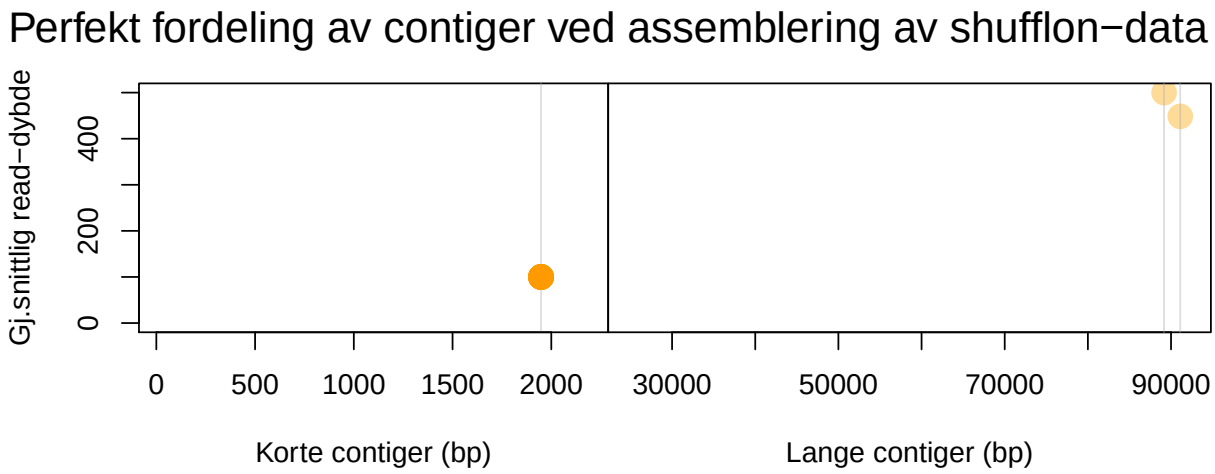
Resultatene fra simuleringen viser at repetert område som ikke løses med vanlig assembly kan løses med hybridassembly dersom de lange readene har nok dybde.

Hybridassembly med dybde av Nanopore-reads på 15x løste opp i assembly som var usammenhengende grunnet for langt RO. Nanopore-readene som ble brukt i hybridassembly var simulert med dybde 5x, 10x, 15x og økende, og tabell 6 på side 26 viser at på dybde 15x gir hybridassembly én contig. Sammenlignet med resultatene i tabell 4 på side 24 kan det se ut som at hybridassembly løste opp de repeterte områdene. Grunnen til dette kan være som beskrevet i [40] at de lange readene hjelper SPAdes med å skjøte sammen de delene av assembly med kun korte reads som ikke henger sammen.

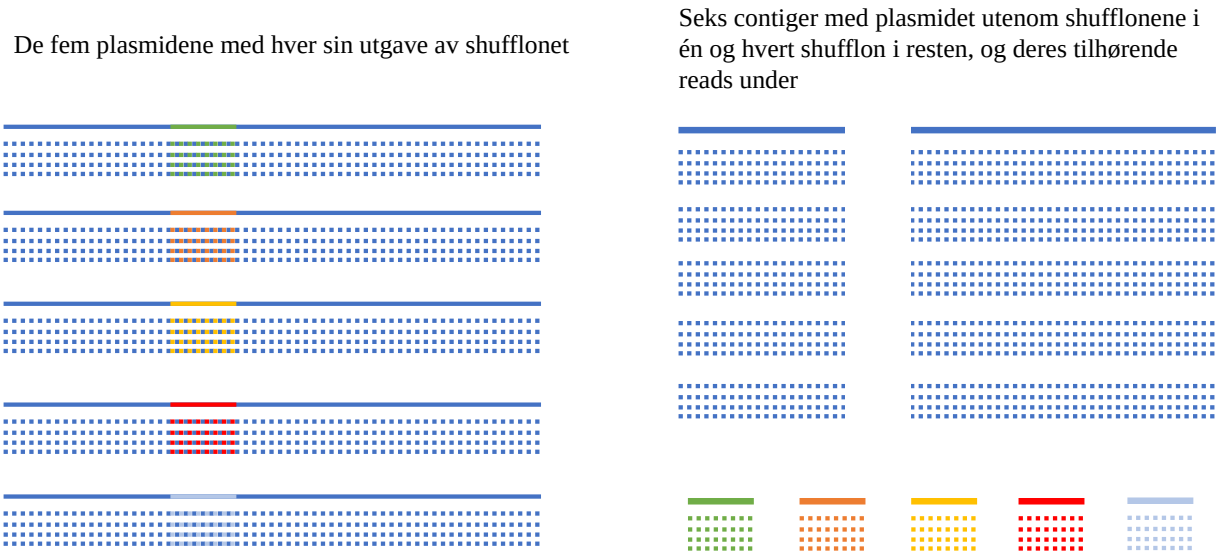
Samme som med vanlig assembly ga hybridassembly av forlengede reads ikke forbedret assembly. Resultatene i tabell 7 på side 27 viser faktisk at assembly fikk lavere identitet og flere contiger med de forlengede readene. Som diskutert i avsnitt 4.1.1 på side 41 skulle assembly bli forbedret ifølge forfatterne av FLASH [43], men resultatene i denne oppgaven viser at de forlengede readene tilfører ingen ny informasjon til assembleren. Resultatene i tabell 7 på side 27 viser at assembly inneholdt mer feil når de forlengede readene ble brukt, og dette kan være grunnet feilraten til FLASH på mindre enn 1 % [43].

4.2 Assemblering av shufflon-variantene

En vellykket assemblering av shufflon-plasmidene kan være slik som vist i figur 24 og figur 25. Hver av de korte contigene er én variant av de fem shufflonene og alle de lange contigene er plasmidet med eller uten shufflon.



Figur 24: Denne figuren viser hvordan denne typen plott skulle sett ut dersom assemblering av shufflon-variantene hadde blitt slik som ønsket. Det som er utydelig med denne figuren er at det er mange contiger korte contiger med samme lengde. Det er én lang contig som er enten 89161 bp eller 91109 bp lang og resten av contigene er like lange som shufflonet og derfor samles i samme punkt.



Figur 25: Figuren visualiserer assemblering av shufflonene på en forenklet måte. Til venstre vises data før assemblering, med sekvensene av de 5 plasmidene med hver sin utgave av shufflonet indikert med egen farge, med tilhørende reads under. Til høyre vises assembly og de readene som utgjorde den delen av assembly (contigen) under. De lengste contigene øverst til høyre, kan egentlig sies å være én contig med et hull der shufflonet skulle vært. Men det er viktig å huske på at dette er sirkulært DNA slik at det likevel kan være én sammenhengende contig.

4.2.1 Assemblering av Illumina-reads

Assemblering av simulerte Illumina-reads fra shufflon-varianter ga ikke sammenhengende sekvenser av alle shufflonene.

Ingen av assemblerne klarte å assemblere plasmidene slik at sekvensene av alle shufflonene ble sammenhengende slik som vist i figur 24 på side 44 og figur 25 på side 44. Figur 10 på side 30 viser at ingen korte contiger er like lange som shufflonet. Det samme resultatet fikk også Brouwer *et. al* [8], der de også assemblerte korte reads, deriblant Illumina-reads fra IncI1-plasmider som inneholdt ulike utgaver av shufflonet. Forfatterne beskriver at assembly av Illumina-reads ga contiger med bare deler av shufflonet. Det viser våre resultater også, de fleste korte contigene består av deler av shufflonet.

Likheten mellom plasmidene med ulike shuffloner og et metagenom kan være årsaken til at MetaSPAdes kommer godt ut blant assemblerne. Et metagenom er samling av flere genomer, der enkelte kan være nært beslektet med for det meste identisk genom. Sekvensene av de fem plasmidene med shufflon-varianter er også for det meste identiske utenom shufflonene. MetaSPAdes ga høyest identitet, korte contiger med lengde mest lik shufflon-lengden og mest lik den forventede gjennomsnittlige read-dybden av shufflonene. De fleste korte contigene besto av deler av shufflonet, men én contig per simulering ble litt lengre enn shufflonet og besto av hel shufflon-sekvens. Men med én contig med én av de fem shufflon-variantene mangler fortsatt de fire andre shufflon-variantene. Når det gjelder de lange contigene kan de to tilsammen utgjøre hele lengden av plasmid-sekvensen, som nevnt i figur 24 på side 44. Resultatet fra sammenligning av contigene med shufflon-sekvensen viser at de lange contigene ikke inneholder shufflon, og det viser at MetaSPAdes klarer å skille mellom plasmid og shufflon bedre enn kun SPAdes.

Forlengede reads forbedret ikke assembly, fordi SPAdes allerede utnytter de overlappende R1- og R2-readene. Forskjellene mellom resultatene fra assemblering av kun Illumina-reads og assemblering av forlengede reads av Illumina-reads er liten og resultatene viser også at det er samme hvordan de forlengede readene blir gitt til SPAdes. Grunnen til at assembly ikke blir bedre kan være at de forlengede readene ikke bidrar med ny informasjon til assembleren slik som nevnt for resultatene i del 1 i avsnitt 4.1.1 på side 41.

Utifra det som er beskrevet i figur 25 på side 44 skulle $5 \cdot 100x$ reads i gjennomsnitt dekke hver posisjon i contigen(e) som er felles for alle de 5 plasmidene. Dette er fordi sekvenseringsdybden av Illumina-readene ble satt til 100x under simulering. Resultatene viser at den gjennomsnittlige største gjennomsnittlige read-dybden er rundt 400. Grunnen til dette kan være at readene som simuleres og assembleres inneholder feil, sånn som med de resultatene fra del 1 i avsnitt 3.1 på side 23

4.2.2 Hybridassemblering av Illumina- og Nanopore-reads

Hybridassembly med Nanopore- og Illumina-reads løser ikke utfordringen med assemblering av shufflon-varianter.

Ingen hybridassemblyer ga alle de fullstendige sekvensene av shufflon-variantene. Hybridassembly ga lange contiger med samme lengde som plasmidet og korte contiger kortere enn shufflonet bestående kun av deler av shufflonene. Tilsvarende assemblyer ga assemblering av kun Illumina-reads beskrevet i avsnitt 4.2.1 på side 45. Brouwer *et al.* anbefalte å forsøke med lange reads for å løse opp i assembleringsproblemet [8], men våre resultater viser at HybridSPAdes med lange Nanopore-reads ikke løste det.

For få av de korte contigene fra HybridSPAdes med MetaSPAdes inneholder en hel sekvens av et shufflon til å kunne kalle assembleringen vellykket. En vellykket assemblering ville i dette tilfellet ha gitt én contig med hele shufflon-sekvensen for hver variant av shufflonet. Da ville rekkefølgen på delene i alle shufflonene bli karakterisert. Resultatene viser at enkelte av de korte contigene var litt lengre enn shufflonet og inneholdt hele shufflon-sekvensen. Men det var kun én eller to slike contiger per simulering som betyr at ikke alle shufflon-variantene ble funnet. HybridSPAdes med MetaSPAdes var den eneste assembleren som ga contiger med hele sekvensen av shufflonet. Ifølge manualen til SPades [63] er kjøring av HybridSPAdes samtidig som MetaSPAdes eksperimentelt og noe forfatterne ikke kan garantere at gjennomføres optimalt. Men resultatene i denne oppgaven viser at MetaSPAdes er den mest lovende assembleren av simulerte reads fra plasmidene med shufflon-varianter.

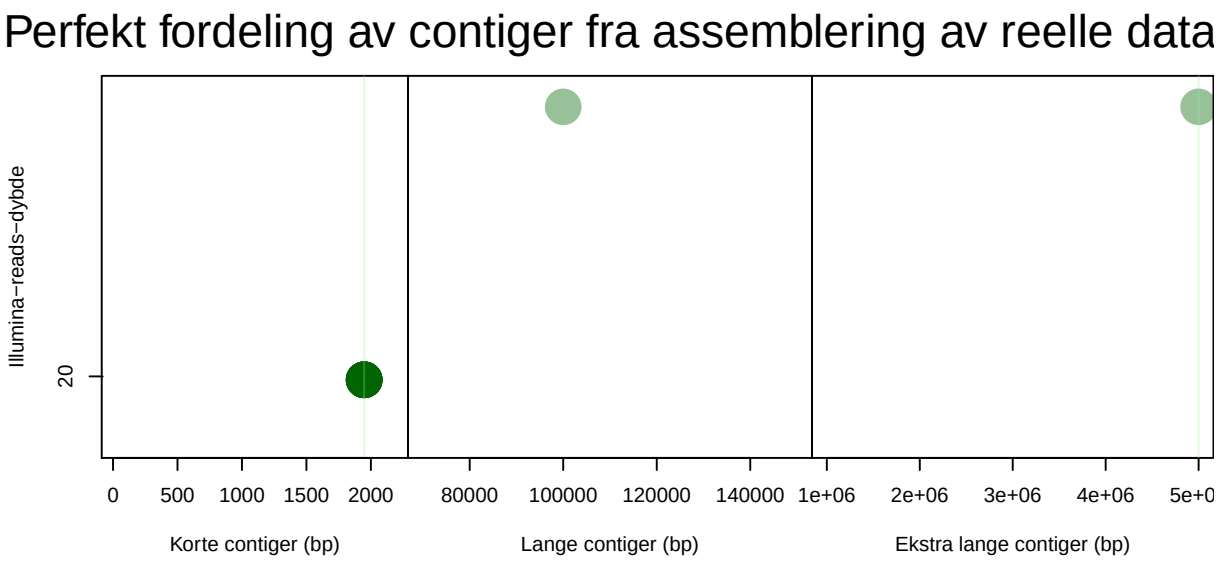
HybridSPAdes med MetaSPAdes skiller bedre mellom shufflon og plasmid enn kun HybridSPAdes. De lange contigene fra HybridSPAdes med MetaSPAdes inneholdt ikke shufflon-sekvens. Dette viser at MetaSPAdes skiller på plasmid og shufflon, noe som er forventet av MetaSPAdes, fordi den jobber for å finne identiske sekvenser for alle genomene som assembleres [53].

MetaSPAdes sammen med HybridSPAdes er eksperimentelt [63], og kan være grunnen til at assemblyer fra disse er vanskelig å forklare. De identiske resultatene fra MetaSPAdes med ulike mengde Nanopore-reads beskrevet i figurteksten i tabell 11 på side 31 viser at assemblering av MetaSPAdes med HybridSPAdes er uavhengig av dybde på Nanopore-readene, hvertfall mellom 5x og 200x. En mulig forklaring på dette kan være at MetaSPAdes med HybridSPAdes bruker enkelte Nanopore-reads, men ikke alle. Det er også verdt å nevne at MetaSPAdes med HybridSPAdes gir enkelte kortere contiger og større spredning i contig-lengde enn MetaSPAdes med kun Illumina-reads. Lengre contiger er intuitivt fordi HybridSPAdes skjøter contiger, men de kortere contigene er vanskelig å forklare. Som forfatterne av SPADes skriver er HybridSPAdes med MetaSPAdes eksperimentelt [63], og resultatene kan være vanskelig å forklare av den grunn.

Forlengede reads forbedret ikke hybridassembly, fordi de forlengede readene ikke bidrar med ny informasjon til assembleren og fordi FLASH-programvaren kan gjøre feil. Forlengede reads av R1- og R2- reads i tillegg til R1- og R2-readene og Nanopore-reads ble gitt til HybridSPAdes. Resultatene viser at assembly med forlengede reads kan inneholde mer feil enn assembly uten. Årsakene til ingen forbedring eller mer feil i assembly kan være at de forlengede readene ikke gir ny informasjon til assembleren (som tidligere nevnt i avsnitt 4.1.1 på side 41) og at FLASH-programvaren gjør feil [43].

4.3 Assemblering av de reelle dataene

De reelle dataene er reads fra sekvensering av en prøve fra kultur av *[E. coli*. Disse *E. coliene* antas å ha kromosomalt DNA med lengde 5 000 000 bp og ett eller to IncI1-plasmider. Lengden på IncI1-plasmider varierer mellom 91 000 bp til 120 000 [69]. IncI1-plasmidene antas å ha shufflon-område med innhold som vil variere i rekkefølge fra plasmid til plasmid. Kromosomet inneholder 16S-genet, som er et repetert område. Siden alle de sekvenserte *E.coliene* kommer fra samme kultur vil alle genomene utenom shufflonene være identiske. Derfor kan det tenkes at et optimalt assembly er én contig med kromosomet, én contig med plasmidet uten shufflonet og så mange contiger som det er ulike shuffloner med de i. Dersom shufflonet er av samme type som beskrevet i IncI-alpha-plasmidet R64 i 1986 [7] vil det ha lengde på litt under 2000 bp. Lengden på IncI-alpha-plasmidet R64 er 120 000 bp [69].



Figur 26: Denne figuren viser hvordan denne typen plott skulle sett ut dersom resultatene av assemblering av de reelle dataene hadde blitt slik vi ønsket. De tre punktene skal indikere contiger med henholdsvis shufflonet, plasmidet og kromosomet. Det som er utydelig med denne figuren er at det er mange contiger, altså punkter oppå hverandre i punktet som har samme contig-lengde som shufflonet. Read-dybden på plasmidet og kromosomet er ikke spesifisert fordi dette er utfordrende å forutsi. Men de bør ha den samme read-dybden, eller plasmidet bør ha dobbelt så stor read-dybde dersom det forekommer i kopitall 2.

4.3.1 Assemblering av reelle Illumina-reads

Assemblering av kun reelle Illumina-reads gir ikke de fullstendige sekvensene til shufflonene. Resultatene viser at det ikke er flere enn 2 contiger av samme lengde av de contigene over 500 bp. Dersom assembly hadde gitt contiger med fullstendige sekvenser av shufflonene ville det vært én contig per shufflon-variant og disse contigene ville hatt samme lengde. De vil ha samme lengde fordi de delene shufflonet består av er høyt konservert selvom de blir flyttet på.

For mye trimming og filtrering med Trimmomatic kan være årsaken til at assembly med prosesserte reads gir flere contiger og lavere n50 enn med rå reads. SPAdes korrigerer readene under assemblering, og sammen med Trimmomatic kan det hende at det ble for streng prosessering av readene. På denne måten kan reads med viktig informasjon for assembly forsvinne, og konsekvensene kan være et usammenhengende assembly med flere contiger og lavere N50.

At ingen reads mapper til menneske-genomet betyr at Illumina-readene ikke er kontaminert av menneske-DNA. Ifølge Knut Rudi er de sekvenserte *E. coliene* transkonjuganter av de *E. coliene* som ble funnet i tarmen til tvillingene som er beskrevet i studien til Mari Hagbø [3]. Dette betyr at det ikke er de faktiske prøvene som ble sekvensert, men rendyrka kultur av de. Dette kan være grunnen til at de ikke er kontaminert av menneske-DNA, selvom det alltid kan være en sjanse for kontaminering.

Forskjell mellom assembly av kun prosesserte R1- og R2-reads og assembly med forlengede reads i tillegg viser at de forlengede readene har annen informasjon enn R1 og R2 etter at readene ble prosessert. Før prosessering ga assemblering veldig like resultater for alle assemblerne, men etter prosessering ble assembly med de forlengede readene mindre sammenhengende. Med mindre sammenhengende menes tre ganger så mange contiger, lavere n50 og kortere lengste contig. Lavere N50 tyder på at assembly har mange korte contiger, og det er ikke nødvendigvis negativt i denne sammenhengen, fordi vi ønsker mange korte contiger av samme lengde som består av de fullstendige shufflon-sekvensene. Men etter at de korte contigene over 500 bp ble telt opp, var det maksimalt 2 contiger med samme lengde. At assembly med forlengede reads blir svært annerledes assembly av kun R1 og R2 stemmer ikke med de tidligere resultatene fra de simulerte assemblyene, der de forlengede readene ikke bidro med ny informasjon til assembleren. Dersom FLASH har gjort mange feil under skjøting av readene, kan dette være en forklaring. Men ifølge forfatterne har FLASH en feilrate på mindre enn 1% [43], så dette er sannsynligvis ikke forklaringen.

4.3.2 Hybridassemblering av de reelle Illumina- og Nanopore-readene

Fordi alle assemblyene hadde maksimalt 2 contiger over 500 bp med samme lengde betyr dette at ingen av assemblyene gir contiger bestående kun av fullstendige sekvenser av alle shufflonene. Som beskrevet i figur 26 på side 47 hadde ønsket resultat vært mange korte contiger av samme lengde som besto av hele sekvensen til alle varianter av shufflonet i de sekvenserte *E. coliene*. Alle hybridassemblyene beskrevet i avsnitt 3.3.2 på side 35 ga korte contiger med varierende read-dybde og contig-lengde. Disse resultatene ligner på assembly beskrevet i [8] der de korte contigene besto av deler av shufflonet. Dette gjelder også for våre resultater, der enkelte contiger består av shufflon-sekvens, men det er også flere korte contiger som ikke gjør det. Selvom nesten ingen contiger har eksakt samme lengde, kan det hende at enkelte contiger likevel inneholder en eller flere hele av shufflon-sekvensene.

At ”MismatchCorrector” deaktiveres ved kjøring av MetaSPAdes kan være grunnen til at resultatene fra Meta/HybridSPAdes er bedre for korrigerte Illumina-reads. Hybridassembly med de prosesserte Illumina-readene gir færrest contiger med HybridSPAdes og MetaSPAdes samtidig. Derimot gir HybridSPAdes (og SPAdes) færrest contiger fra assemblering av rå Illumina-reads. Grunnen til dette kan være at MetaSPAdes ikke kjører modulen ”MismatchCorrector”, mens SPAdes og HybridSPAdes gjør det. Når HybridSPAdes kjøres samtidig som MetaSPAdes skrus ”MismatchCorrector” av og assembly blir ikke korrigert for feil, og da kan det være en fordel at Illumina-readene ble prosessert av Trimmomatic før assemblering.

Færre contiger for hybridassembly kan bety at repeterte områder er løst opp. Som for de simulerte dataene med repetert område ble antall contiger lavere ved hybridassembly enn for assembly. Grunnen til dette var at de repeterte områdene kunne kobles til de delene av sekvensen de tilhørte istedenfor å havne i for eksempel én contig. I tillegg viser figur 17 på side 37 at read-dybden blir mye lavere for hybridassembly, noe som også bekrefter at kopiene av repeterte områder ikke havner i samme contig.

Spredning i contig-lengder i andre vindu i alle plott gjør det utfordrende å finne den reelle lengden på plasmidet. Det er som tidligere nevnt at lengden på plasmidet ikke er eksakt definert, og det kan derfor være kortere eller lengre enn 100 000 bp. Det er en rekke av contiger som har ulik lengde, men med samme read-dybde og disse er litt lengre enn plasmidet i simuleringen (indikert med vertikale linjer). Denne read-dybden er på ca. 20 som kan være samme som den antatte gjennomsnittlige dybden Illumina-readene ble sekvensert ved. Samme read-dybde som sekvenseringsdybden betyr at ingen av de samme readene mapper til contigene. Derfor kan ingen av disse contigene inneholde plasmidet, dersom plasmidet er identisk for alle *E. coliene*.

Utifra våre resultater er det usikkert hvorfor flere av de lengste Nanopore-readene ikke nødvendigvis gir bedre assembly. Resultatene for hybridassembly viser at de korrigerte Illumina-readene gir flere contiger enn med de rå. Samtidig viser resultatene at hybridassembly med flere av de lengste Nanopore-readene ikke forbedres. En mulig årsak kan være at Nanopore-readene inneholder så mye feil at flere Nanopore-reads ødelegger for assembly, men resultatene i figur 21 på side 39 viser at med korrigering av de 5000 lengste Nanopore-readene forbedres ikke hybridassembly. Grunnen til dette kan være at korrigeringen av Nanopore-reads ikke forbedrer Nanopore-readene, eller så er det ikke feilene i Nanopore-readene som er årsak til at fler Nanopore-reads gir mindre korrekt hybridassembly.

Årsaken til forskjell mellom assembly fra simulerte reads og reelle reads kan være at ART ikke simulerer realistisk nok. Resultatene viser en tydelig forskjell på resultatene fra simulerte data og de reelle. Grunnen til dette kan være at de simulerte dataene ikke gjenspeiler kompleksiteten som de reelle dataene har. Med kompleksiteten menes det at når readene sekvenseres er sekvenseringsprosessen påvirket av innholdet i DNA og på grunn av dette er ikke sekvenseringsdybden kontinuerlig for hele genomet [70]. ART beskrives ikke som et simuleringsverktøy som tar høyde for disse systematiske skjevhetene ved sekvensering [71]. Derfor kan det hende at de simulerte readene ikke gjenspeiler kompleksiteten i sekvenserte reads.

5 Videre arbeid

ART tar ikke høyde for ujevn fordeling av sekvenseringsdybde, men det er andre simuleringsverktøy som tar det [71] og som mulig kunne gitt mer realistisk simulering av reads. Når det gjelder de reelle readene er sekvenseringsdybden av Illumina-readene svært lav (litt under 20x), og en økning i sekvenseringsdybden av Illumina-readene kunne kanskje forbedret assembly.

FLASH ga ikke forbedrede assemblyer, men det kan være interessant å se hvilken effekt de to programmene [67] [68] som SPAdes har foreslått til å generere forlengede reads [63] har på assembly.

Da Komano *et. al* først fant shufflonet i R64-plasmidet [72] brukte de ulike typer restriksjonsenzymer og gel-elektroforese [73]. Det kjent at shuffloner sitter mellom *pilV*-genet og *rci*-genet. Dersom det er ønskelig å kun undersøke sekvensene av shufflonene kan en bruke primere som binder spesifikt til *pilV* eller *rci* for å få fragmenter av shufflonene [8]. En annen måte er å bruke Illumina-reads til å korrigere lange reads [16]. De lange readene strekker over shufflonene og ved å korrigere disse kan shufflon-sekvensene detekteres ved å sortere for de som inneholder de svært konserverte delene av shufflonet eller genene *pilV*/*rci*. For å finne shufflon-variantene trengs altså en tilpasset løsning og ikke de-novo-assemblering.

6 Konklusjon

HybridSPAdes med Illumina- og Nanopore-reads egner seg ikke for å finne de fullstendige sekvensene av alle shufflon-variantene i en kultur av *E. coli*. Dette betyr at det ikke vil være mulig å finne ut hvilke forbindelser *E. coliene* får en stabil binding til. HybridSPAdes egner seg derimot bra til å løse opp i assemblyer som er ufullstendige på grunn av repetert område.

Repetert område gir ikke problemer for de-novo-assemblering av parvise-reads, dersom området er kortere enn fragmentlengde. Årsaken er at SPAdes-assembleren utnytter at hver par av de parvise readene har en gitt avstand mellom hverandre og at enkelte av dem overlapper delvis. Med dette i tankene er det forståelig at forlengede reads satt sammen av FLASH-programvaren ikke tilfører ny informasjon til assembleren. Dersom det repeterte området er lengre enn fragmentlengde gir assembly flere enn én contig med varierende gjennomsnittlige read-dybder. FLASH kan gjøre feil ved skjøting av de delvis overlappende parvise readene [43], og dette kan gjøre at assembly inneholder mer feil enn med kun parvise reads. Hybridassemblering med Nanopore-reads med dybde 15x løste det repeterte området og ga én contig per assembly.

Illumina- og Nanopore-reads fra plasmider med ulike versjoner av shufflonet kan ikke assembleres de-novo med SPAdes, HybridSPAdes eller MetaSPAdes slik at variantene av shufflonene kommer fram. Resultatet fra de-novo-assemblering var én contig bestående hovedsakelig av plasmid-sekvens uten shufflonet, og korte contiger bestående av deler av og ikke fullstendige sekvenser av shufflonet. MetaSPAdes skilte seg ut fra de andre assemblerne fordi den klarte å skille tydelig mellom plasmid og shufflon, og klarte å gjenskape enkelte, men ikke alle, de fullstendige shufflon-sekvensene. Grunnen til at MetaSPAdes skilte seg ut fra de andre assemblerne kan være fordi plasmidene med ulike utgaver av shufflonet ligner på metagenom, som MetaSPAdes er programmert til å assemblere. MetaSPAdes sammen med HybridSPAdes er ifølge forfatterne av SPAdes eksperimentelt [63], men disse assembleringsalgoritmene samtidig ga på det meste to contiger med hele sekvenser av shufflonet.

De-novo-assemblering med SPAdes eller HybridSPAdes på reelle Illumina-og Nanopore-reads gir heller ikke fullstendige sekvenser av alle shufflon-variantene. Resultatene var mye vanskeligere å lese enn for assemblering av de simulerte readene. Grunnen til dette er mest sannsynligvis at read-simulatoren ikke simulerer sekvenseringsprosessen realistisk nok [71].

SPAdes-assembleren korrigerer Illumina-reads slik at det ikke er nødvendig å bruke annen programvare til å gjøre dette før assemblering. Dersom for eksempel Trimmomatic brukes til trimming og filtrering kan prosesseringen bli for streng og reads med viktig informasjon kan bli kastet slik at assembly blir satt sammen feil og gir flere contiger. Et unntak kan være ved kjøring av MetaSPAdes, fordi da blir ikke assembly korrigert for feil.

Til videre arbeid anbefales det å bruke en annen metode for å finne shufflon-sekvensene. Shufflon-variantene er tidligere blitt detektert ved å korrigere lange reads med korte og lete etter shufflon-sekvensene i de lange, korrigerte readene.

Bibliografi

[1] Ahn, S. «Introduction to bioinformatics: sequencing technology». I: *Asia Pacific Allergy* årg. 1, nr. 2 (2011), s. 93–97.

[2] Luo, J. mfl. «EPGA: de novo assembly using the distributions of reads and insert size». I: *Bioinformatics* årg. 31, nr. 6 (2014), s. 825–833.

[3] Hagbø, M. E. S. «Characterization of conjugative plasmids in the gut microbiota from a preterm twin pair». I: (2017).

[4] Szmolka, A. mfl. «Conjugative IncF and IncI1 plasmids with tet (A) and class 1 integron conferring multidrug resistance in F18+ porcine enterotoxigenic E. coli». I: *Acta Veterinaria Hungarica* årg. 63, nr. 4 (2015), s. 425–443.

[5] Wong, M. H.-y. mfl. «IncI1 plasmids carrying various blaCTX-M genes contribute to ceftriaxone resistance in Salmonella enterica serovar Enteritidis in China». I: *Antimicrobial agents and chemotherapy* årg. 60, nr. 2 (2016), s. 982–989.

[6] Mazel, D. og Davies, J. «Antibiotic resistance in microbes». I: *Cellular and Molecular Life Sciences CMLS* årg. 56, nr. 9-10 (1999), s. 742–754.

[7] Komano, T., Kubo, A. og Nisioka, T. «Shufflon: multi-inversion of four contiguous DNA segments of plasmid R64 creates seven different open reading frames». I: *Nucleic acids research* årg. 15, nr. 3 (1987), s. 1165–1172.

[8] Brouwer, M. S. mfl. «IncI shufflons: assembly issues in the next-generation sequencing era». I: *Plasmid* årg. 80 (2015), s. 111–117.

[9] Carattoli, A. mfl. «Contemporary IncI1 plasmids involved in the transmission and spread of antimicrobial resistance in Enterobacteriaceae». I: *Plasmid* (2018).

[10] Kim, J. S. mfl. «Complete nucleotide sequence of the IncI1 plasmid pSH4469 encoding CTX-M-15 extended-spectrum beta-lactamase in a clinical isolate of Shigella sonnei from an outbreak in the Republic of Korea». I: *International journal of antimicrobial agents* årg. 44, nr. 6 (2014), s. 533–537.

[11] UiO, D. m.-n. f. *Pilus*. URL: https://www.mn.uio.no/ibv/tjenester/kunnskap/plantefys/leksikon/p/pilus.html (sjekket 21.02.2019).

[12] Bradley, D. E. «Characteristics and function of thick and thin conjugative pili determined by transfer-derepressed plasmids of incompatibility groups I1, I2, I5, B, K and Z». I: *Microbiology* årg. 130, nr. 6 (1984), s. 1489–1502.

[13] Komano, T. mfl. «Transfer region of IncI1 plasmid R64 and role of shufflon in R64 transfer». I: *Journal of bacteriology* årg. 172, nr. 5 (1990), s. 2230–2235.

[14] Horiuchi, T. og Komano, T. «Mutational analysis of plasmid R64 thin pilus prepilin: the entire prepilin sequence is required for processing by type IV prepilin peptidase». I: *Journal of bacteriology* årg. 180, nr. 17 (1998), s. 4613–4620.

[15] Ishiwa, A. og Komano, T. «PilV adhesins of plasmid R64 thin pili specifically bind to the lipopolysaccharides of recipient cells». I: *Journal of molecular biology* årg. 343, nr. 3 (2004), s. 615–625.

[16] Sekizuka, T. mfl. «Elucidation of quantitative structural diversity of remarkable rearrangement regions, shufflons, in IncI2 plasmids». I: *Scientific reports* årg. 7, nr. 1 (2017), s. 928.

[17] Komano, T., Kim, S.-R. og Yoshida, T. «Mating variation by DNA inversions of shufflon in plasmid R64». I: *Advances in biophysics* årg. 31 (1995), s. 181–193.

[18] Cock, P. J. mfl. «The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants». I: *Nucleic acids research* årg. 38, nr. 6 (2009), s. 1767–1771.

[19] Caporaso, J. G. mfl. «Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms». I: *The ISME journal* årg. 6, nr. 8 (2012), s. 1621.

[20] Fadrosh, D. W. mfl. «An improved dual-indexing approach for multiplexed 16S rRNA gene sequencing on the Illumina MiSeq platform». I: *Microbiome* årg. 2, nr. 1 (2014), s. 6.

[21] Illumina. *What is NGS Library Preparation?* URL: https://www.illumina.com/techniques/sequencing/ngs-library-prep.html (sjekket 05.04.2019).

[22] Illumina. *Explore Illumina sequencing technology*. URL: https://www.illumina.com/science/technology/next-generation-sequencing/sequencing-technology.html (sjekket 05.04.2019).

[23] Bentley, D. R. mfl. «Accurate whole human genome sequencing using reversible terminator chemistry». I: *Nature* årg. 456, nr. 7218 (2008), s. 53.

[24] Fu, S., Wang, A. og Au, K. F. «A comparative evaluation of hybrid error correction methods for error-prone long reads». I: *Genome Biology* årg. 20, nr. 1 (feb. 2019), s. 26.

[25] Ashton, P. M. mfl. «MinION nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island». I: *Nature biotechnology* årg. 33, nr. 3 (2015), s. 296.

[26] Lu, H., Giordano, F. og Ning, Z. «Oxford pore MinION sequencing and genome assembly». I: *Genomics, proteomics & bioinformatics* årg. 14, nr. 5 (2016), s. 265–279.

[27] Laehnemann, D., Borkhardt, A. og McHardy, A. C. «Denoising DNA deep sequencing data—high-throughput sequencing errors and their correction». I: *Briefings in bioinformatics* årg. 17, nr. 1 (2015), s. 154–179.

[28] Jain, M. mfl. «The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community». I: *Genome biology* årg. 17, nr. 1 (2016), s. 239.

[29] Dekker, C. «Solid-state nanopores». I: *Nature nanotechnology* årg. 2, nr. 4 (2007), s. 209.

[30] Lander, E. S. og Waterman, M. S. «Genomic mapping by fingerprinting random clones: a mathematical analysis». I: *Genomics* årg. 2, nr. 3 (1988), s. 231–239.

[31] Ayling, M., Clark, M. D. og Leggett, R. M. «New approaches for metagenome assembly with short reads». I: *Briefings in bioinformatics* (2019).

[32] Bankevich, A. mfl. «SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing». I: *Journal of computational biology* årg. 19, nr. 5 (2012), s. 455–477.

[33] Alkan, C., Sajjadian, S. og Eichler, E. E. «Limitations of next-generation genome sequence assembly». I: *Nature methods* årg. 8, nr. 1 (2011), s. 61.

[34] Case, R. J. mfl. «Use of 16S rRNA and rpoB genes as molecular markers for microbial ecology studies». I: *Appl. Environ. Microbiol.* Årg. 73, nr. 1 (2007), s. 278–288.

[35] Goldstein, S. mfl. «Evaluation of strategies for the assembly of diverse bacterial genomes using MinION long-read sequencing». I: *BMC genomics* årg. 20, nr. 1 (2019), s. 23.

[36] Koren, S. og Phillippy, A. M. «One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly». I: *Current opinion in microbiology* årg. 23 (2015), s. 110–120.

[37] Salzberg, S. L. mfl. «GAGE: A critical evaluation of genome assemblies and assembly algorithms». I: *Genome research* årg. 22, nr. 3 (2012), s. 557–567.

[38] Goodwin, S., McPherson, J. D. og McCombie, W. R. «Coming of age: ten years of next-generation sequencing technologies». I: *Nature Reviews Genetics* årg. 17, nr. 6 (2016), s. 333.

[39] Phillippy, A. M., Schatz, M. C. og Pop, M. «Genome assembly forensics: finding the elusive mis-assembly». I: *Genome biology* årg. 9, nr. 3 (2008), R55.

[40] Antipov, D. mfl. «hybridSPAdes: an algorithm for hybrid assembly of short and long reads». I: *Bioinformatics* årg. 32, nr. 7 (2015), s. 1009–1015.

[41] Venter, J. C. mfl. «Environmental genome shotgun sequencing of the Sargasso Sea». I: *science* årg. 304, nr. 5667 (2004), s. 66–74.

[42] Antipov, D. mfl. «plasmidSPAdes: assembling plasmids from whole genome sequencing data». I: *bioRxiv* (2016), s. 048942.

[43] Magoč, T. og Salzberg, S. L. «FLASH: fast length adjustment of short reads to improve genome assemblies». I: *Bioinformatics* årg. 27, nr. 21 (2011), s. 2957–2963.

[44] Goodwin, S. mfl. «Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome». I: *Genome research* årg. 25, nr. 11 (2015), s. 1750–1756.

[45] Magi, A. mfl. «Nanopore sequencing data analysis: state of the art, applications and challenges». I: *Briefings in bioinformatics* årg. 19, nr. 6 (2017), s. 1256–1272.

[46] Chin, C.-S. mfl. «Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data». I: *Nature methods* årg. 10, nr. 6 (2013), s. 563.

[47] Berlin, K. mfl. «Assembling large genomes with single-molecule sequencing and locality-sensitive hashing». I: *Nature biotechnology* årg. 33, nr. 6 (2015), s. 623.

[48] Deshpande, V. mfl. «Cerulean: A hybrid assembly using high throughput short and long reads». I: *International workshop on algorithms in bioinformatics*. Springer. 2013, s. 349–363.

[49] Kim, J. S. mfl. *Shigella sonnei strain SS084469 plasmid pSH4469, complete sequence*. URL: https://www.ncbi.nlm.nih.gov/nuccore/696158084 (sjekket 06.04.2019).

[50] Alonso, C. A. mfl. *Escherichia coli plasmid pCAZ590, complete sequence 95782*. URL: https://www.ncbi.nlm.nih.gov/nuccore/LT669764.1?report=genbank&from=99769&to=101716 (sjekket 13.03.2019).

[51] Huang, W. mfl. «ART: a next-generation sequencing read simulator». I: *Bioinformatics* årg. 28, nr. 4 (2012). 10.1093/bioinformatics/btr708, s. 593–594.

[52] Li, Y. mfl. «DeepSimulator: a deep simulator for Nanopore sequencing». I: *Bioinformatics* årg. 34, nr. 17 (). 10.1093/bioinformatics/bty223, s. 2899–2908.

[53] Nurk, S. mfl. «metaSPAdes: a new versatile metagenomic assembler». I: *Genome research* årg. 27, nr. 5 (2017), s. 824–834.

[54] Anthony M. Bolger Bjoern Usadel, M. L. «Trimmomatic: a flexible trimmer for Illumina sequence data». I: *Bioinformatics* årg. 30, nr. 15 (apr. 2014), s. 2114–2120. eprint: http://oup.prod.sis.lan/bioinformatics/article-pdf/30/15/2114/17143152/btu170.pdf.

[55] Koren, S. mfl. «Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation». I: *Genome research* årg. 27, nr. 5 (2017), s. 722–736.

[56] Langmead, B. og Salzberg, S. L. «Fast gapped-read alignment with Bowtie 2». I: *Nature methods* årg. 9, nr. 4 (2012), s. 357.

[57] Subgroup, 1. G. P. D. P. mfl. «The Sequence Alignment/Map format and SAMtools». I: *Bioinformatics* årg. 25, nr. 16 (jun. 2009), s. 2078–2079. eprint: http://oup.prod.sis.lan/bioinformatics/article-pdf/25/16/2078/531810/btp352.pdf.

[58] Gurevich, A. mfl. «QUAST: quality assessment tool for genome assemblies». I: *Bioinformatics* årg. 29, nr. 8 (2013), s. 1072–1075.

[59] Desai, A. mfl. «Identification of optimum sequencing depth especially for de novo genome assembly of small genomes using next generation sequencing data». I: *PloS one* årg. 8, nr. 4 (2013), e60204.

[60] Yu Li Renmin Han, C. B. og Li, M. *DeepSimulator manual*. URL: https://github.com/lykaust15/DeepSimulator (sjekket 12.02.2019).

[61] Anthony M. Bolger Bjoern Usadel, M. L. *Manual for Trimmomatic: A flexible trimmer for Illumina Sequence Data. Bioinformatics, btu170.* 2014. URL: http://www.usadellab.org/cms/?page=trimmomatic (sjekket 02.04.2019).

[62] Adam Phillippy Sergey Koren, B. W. *Canu quick start*. URL: https://canu.readthedocs.io/en/latest/quick-start.html.

[63] Bankevich, A. mfl. *SPAdes 3.13.0 Manual*. URL: http://cab.spbu.ru/files/release3.13.0/manual.html (sjekket 19.02.2019).

[64] Nikolenko, S. I., Korobeynikov, A. I. og Alekseyev, M. A. «BayesHammer: Bayesian clustering for error correction in single-cell sequencing». I: *BMC genomics*. Bd. 14. BioMed Central. 2013, S7.

[65] Li, H. og Durbin, R. «Fast and accurate short read alignment with Burrows–Wheeler transform». I: *bioinformatics* årg. 25, nr. 14 (2009), s. 1754–1760.

[66] Lab, B. *Metabat Bickbucket-page*. URL: https://bitbucket.org/berkeleylab/metabat/src/master/ (sjekket 13.02.2019).

[67] Bushnell, B., Rood, J. og Singer, E. «BBMerge–accurate paired shotgun read merging via overlap». I: *PLoS One* årg. 12, nr. 10 (2017), e0185056.

[68] Qiuming Yao, C. P. *STORM (A Scalable Tool for Merging Paired-End Reads with Variable Insert Sizes)*. URL: https://bitbucket.org/yaoornl/align_test/src/master/.

[69] Johnson, T. J. mfl. «Comparative genomics and phylogeny of the IncI1 plasmids: a common plasmid type among porcine enterotoxigenic Escherichia coli». I: *Plasmid* årg. 66, nr. 3 (2011), s. 144–151.

[70] Sims, D. mfl. «Sequencing depth and coverage: key considerations in genomic analyses». I: *Nature Reviews Genetics* årg. 15, nr. 2 (2014), s. 121.

[71] Escalona, M., Rocha, S. og Posada, D. «A comparison of tools for the simulation of genomic next-generation sequencing data». I: *Nature Reviews Genetics* årg. 17, nr. 8 (2016), s. 459.

[72] Komano, T. mfl. «Highly mobile DNA segment of IncI alpha plasmid R64: a clustered inversion region.» I: *Journal of bacteriology* årg. 165, nr. 1 (1986), s. 94–100.

[73] Furuichi, T., Komano, T. og Nisioka, T. «Physical and genetic analyses of the Inc-I alpha plasmid R64.» I: *Journal of bacteriology* årg. 158, nr. 3 (1984), s. 997–1004.

7 Vedlegg

Tabeller fra assemblering av reelle Illumina-reads

Assembly av rå, reelle Illumina-reads

Tabell 12: Resultater fra ulike assembleringer av rå Illumina-reads. S betyr SPAdes, f-s betyr forlengede reads gitt med enkelt-reads-opsjon, f-m betyr forlengede reads gitt med merged-reads-opsjon og M betyr MetaSPAdes. Read-dybde er gjennomsnittlig antall Illumina-reads som dekker én posisjon i assembly. Gjennomsnittlig read-dybde er et gjennomsnitt av gjennomsnittlige read-dybder for alle contigene. Dersom alle contigene sorteres etter lengde og legges etter hverandre på én rekke vil lengden av den korteste contigen i midten av denne rekka være N50.

	S	S + f-s	S + f-m	M
Antall contiger	235	214	229	239
Minste read-dybde	0	0	0	0
Største read-dybde	265	264	264	277
Gjennomsnittlig read-dybde	17	19	29	18
Median read-dybde	17	17	17	17
Total assembly-lengde (bp)	4642353	4634518	4634921	4639956
Lengste contig (bp)	260858	262873	262607	259399
Korteste contig (bp)	128	128	128	128
Gjennomsnittlig contig-lengde (bp)	19755	21657	20240	19414
Median contig-lengde (bp)	1742	2709	1906	2599
N50 (bp)	78748	81815	78748	70462

Assembly av prosesserte, reelle Illumina-reads

Tabell 13: Resultater fra assembly av prosesserte Illumina-reads. S betyr SPAdes, f-s betyr forlengede reads gitt med enkelt-reads-opsjon, f-m betyr forlengede reads gitt med merged-reads-opsjon og M betyr MetaSPAdes.

	S	S + f-s	S + f-m	M
Antall contiger	269	662	682	232
Minste read-dybde	0	0	0	0
Største read-dybde	266	219	226	277
Gjennomsnittlig read-dybde	16	11	13	17
Median read-dybde	16	7	8	17
Total assembly-lengde (bp)	4648538	4713933	4716198	4635936
Lengste contig (bp)	262873	86190	78753	259399
Korteste contig (bp)	128	128	128	128
Gjennomsnittlig contig-lengde (bp)	17281	7121	6915	19982
Median contig-lengde (bp)	1327	447	518	3104
N50 (bp)	66528	28047	25770	68935

Tabeller fra hybridassemblering av reelle reads

Hybridassembly med SPAdes av alle mengdene Nanopore-reads

Tabell 14: Resultater fra hybridassemblering av rå Illumina-reads og synkende antall Nanopore-reads. Første kolonna viser resultater fra hybridassemblering der de 100 000 lengste Nanopore-readene er brukt, og resten av kolonnene viser henholdsvis 5000, 2500 og 1000 lengste Nanopore-reads.

	100 000	1 000	5000	2500	1000
Antall contiger	77	74	75	80	81
Minste read-dybde	0	0	0	0	0
Største read-dybde	72	66	66	65	63
Gjennomsnittlig read-dybde	8	7	7	7	6
Median read-dybde	1	2	2	1	1
Total assembly-lengde (bp)	4710429	4709265	4709419	4704915	4706027
Lengste contig (bp)	1309158	1309148	1309159	850601	850595
Korteste contig (bp)	128	128	128	128	128
Gjennomsnittlig contig-lengde (bp)	61174	63639	62792	58811	58099
Median contig-lengde (bp)	260	258	260	261	260
N50 (bp)	676972	676972	676973	631201	631203

Tabell 15: Resultater fra assemblering av korrigerte Illumina-reads og ulikt antall Nanopore-reads.

	100 000	1 000	5000	2500	1000
Antall contiger	105	106	97	100	101
Minste read-dybde	0	0	0	0	0
Største read-dybde	75	75	66	75	63
Gjennomsnittlig read-dybde	5	5	6	6	4
Median read-dybde	1	1	1	1	1
Total assembly-lengde (bp)	4708851	4710712	4713351	4711666	4709839
Lengste contig (bp)	1319798	1543903	1543918	1593489	1593473
Korteste contig (bp)	128	128	128	128	128
Gjennomsnittlig contig-lengde (bp)	44846	44441	48591	47117	46632
Median contig-lengde (bp)	241	240	251	251	251
N50 (bp)	1085300	1314917	1489475	1085356	1085354

Hybridassemblyer med 5000 lengste Nanopore-readene

Tabell 16: Resultater fra ulike typer hybridassemblyering av uprosesserte og prosesserte Illumina-reads. Hybrid er hybridassemblering av Illumina- og Nanopore-reads. Hybrid + f-m er Hybridassemblering av Illumina-, forlengede Illumina- gitt med opsjonen -m og Nanopore-reads. Hybrid + Meta er det samme som Hybrid, men i metagenom-modus (HybridSPAdes og MetaSPAdes).

5000 lengste Nanopore-readene			
Rå Illumina-reads	H	H + f-m	H + M
Antall contiger	75	76	72
Minste read-dybde	0	0	0
Største read-dybde	66	75	87
Gjennomsnittlig read-dybde	7	8	16
Median read-dybde	2	1	17
Total assembly-lengde (bp)	4709419	4702997	4716958
Lengste contig (bp)	1309159	1075876	629222
Korteste contig (bp)	128	128	128
Gjennomsnittlig contig-lengde (bp)	62792	61882	65513
Median contig-lengde (bp)	260	255	1267
N50 (bp)	676973	784256	268024
Prosesserte Illumina-reads	H	H + f-m	H + M
Antall contiger	97	293	56
Minste read-dybde	0	0	0
Største read-dybde	66	76	71
Gjennomsnittlig read-dybde	6	3	16
Median read-dybde	1	0	16
Total assembly-lengde (bp)	4713351	4752030	4693245
Lengste contig (bp)	1543918	1458735	549120
Korteste contig (bp)	128	128	128
Gjennomsnittlig contig-lengde (bp)	48591	16219	83808
Median contig-lengde (bp)	251	251	1311
N50 (bp)	1489475	1075656	332610

Hybridassemblyer med 10 000 lengste Nanopore-readene

Tabell 17: Resultater fra ulike typer hybridassemblyering av uprosesserte og prosesserte Illumina-reads. Hybrid er hybridassemblering av Illumina- og Nanopore-reads. Hybrid + f-m er Hybridassemblering av Illumina-, forlengede Illumina- gitt med opsjonen -m og Nanopore-reads. Hybrid + Meta er det samme som Hybrid, men i metagenom-modus (HybridSPAdes og MetaSPAdes).

10 000 lengste Nanopore-readene			
Rå Illumina-reads	H	H f-m	H + M
Antall contiger	74	75	68
Minste read-dybde	0	0	0
Største read-dybde	66	75	87
Gjennomsnittlig read-dybde	7	8	16
Median read-dybde	1	1	17
Total assembly-lengde (bp)	4709265	4702843	4718443
Lengste contig (bp)	1309148	1075865	629217
Korteste contig (bp)	128	128	128
Gjennomsnittlig contig-lengde (bp)	63639	62705	69389
Median contig-lengde (bp)	258	255	1267
N50 (bp)	676972	784249	272872
Prosesserte Illumina-reads	H	H f-m	H + M
Antall contiger	106	292	58
Minste read-dybde	0	0	0
Største read-dybde	75	81	71
Gjennomsnittlig read-dybde	5	3	15
Median read-dybde	1	0	16
Total assembly-lengde (bp)	4710712	4749653	4692933
Lengste contig (bp)	1543903	1465914	549115
Korteste contig (bp)	128	128	128
Gjennomsnittlig contig-lengde (bp)	44441	16266	80913
Median contig-lengde (bp)	240	251	1311
N50 (bp)	1314917	619067	361267

Hybridassemblyer med 100 000 lengste Nanopore-readene

Tabell 18: Resultater fra ulike typer hybridassemblyering av uprosesserte og prosesserte Illumina-reads. Hybrid er hybridassemblering av Illumina- og Nanopore-reads. Hybrid + f-m er Hybridassemblering av Illumina-, forlengede Illumina- gitt med opsjonen -m og Nanopore-reads. Hybrid + Meta er det samme som Hybrid, men i metagenom-modus (HybridSPAdes og MetaSPAdes).

100 000 lengste Nanopore-readene			
Rå Illumina-reads	H	H f-m	H + M
Antall contiger	77	78	56
Minste read-dybde	0	0	0
Største read-dybde	72	75	87
Gjennomsnittlig read-dybde	8	8	16
Median read-dybde	1	1	13
Total assembly-lengde (bp)	4710429	4704014	4720904
Lengste contig (bp)	1309158	1075878	772842
Korteste contig (bp)	128	128	128
Gjennomsnittlig contig-lengde (bp)	61174	60308	84302
Median contig-lengde (bp)	260	255	763
N50 (bp)	676972	784257	504342
Prosesserte Illumina-reads	H	H f-m	H + M
Antall contiger	105	292	62
Minste read-dybde	0	0	0
Største read-dybde	75	70	71
Gjennomsnittlig read-dybde	5	2	15
Median read-dybde	1	0	16
Total assembly-lengde (bp)	4708851	4750777	4693207
Lengste contig (bp)	1319798	1096668	549081
Korteste contig (bp)	128	128	128
Gjennomsnittlig contig-lengde (bp)	44846	16270	75697
Median contig-lengde (bp)	241	251	1763
N50 (bp)	1085300	619028	324173

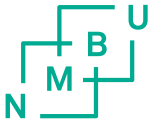
Hybridassembly med korrigerte 5000 lengste Nanopore-readene

Tabell 19: Resultater fra ulike typer assemblering av ikke korrigerte Illumina-reads og korrigerte 5000 Nanopore-reads.

	Hyb	Hyb + f-m	Hyb + f-s	Hyb + Meta
Antall contiger	77	78	89	56
Minste read-dybde	0	0	0	0
Største read-dybde	72	75	75	87
Gjennomsnittlig read-dybde	8	8	9	16
Median read-dybde	1	1	1	13
Total assembly-lengde (bp)	4710429	4704014	4697814	4720904
Lengste contig (bp)	1309158	1075878	1185209	772842
Korteste contig (bp)	128	128	128	128
Gjennomsnittlig contig-lengde (bp)	61174	60308	52784	84302
Median contig-lengde (bp)	260	255	255	763
N50 (bp)	676972	784257	672430	504342

Tabell 20: Resultater fra ulike typer assemblering av korrigerte Illumina-reads og korrigerte 5000 Nanopore-reads.

	Hyb	Hyb + f-m	Hyb + f-s	Hyb + Meta
Antall contiger	104	299	308	43
Minste read-dybde	0	0	0	0
Største read-dybde	72	69	73	71
Gjennomsnittlig read-dybde	7	2	3	15
Median read-dybde	1	0	0	9
Total assembly-lengde (bp)	4710696	4751592	4747197	4694977
Lengste contig (bp)	1543914	1465920	1226386	1076645
Korteste contig (bp)	128	128	128	128
Gjennomsnittlig contig-lengde (bp)	45295	15892	15413	109185
Median contig-lengde (bp)	254	251	251	446
N50 (bp)	1489475	1075655	657275	722375



Norges miljø- og biovitenskapelige universitet
Noregs miljø- og biovitskapelege universitet
Norwegian University of Life Sciences

Postboks 5003
NO-1432 Ås
Norway