RNA:n sekvensointi ilman referenssigenomia

Satu Salekari

Kandidaatintutkielma
HELSINGIN YLIOPISTO
Tietojenkäsittelytieteen laitos

Helsinki, 8. toukokuuta 2017

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta — Fakultet — Faculty	Laitos — Institution — Department
Matemaattis-luonnontieteellinen	Tietojenkäsittelytieteen laitos

Tekijä — Författare — Author
Satu Salekari

Työn nimi — Arbetets titel — Title

RNA:n sekvensointi ilman referenssigenomia

Oppiaine — Läroämne — Subject
Tietojenkäsittelytiede

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidoantal — Number of pages
Kandidaatintutkielma	8. toukokuuta 2017	18

Tiivistelmä — Referat — Abstract

RNA-Seq-menetelmässä solun transkriptomi eristetään ja sekvensoidaan. Menetelmä on bioinformatiikan perustyökalu ja tärkeässä roolissa esimerkiksi geenien ekspression tutkimuksessa.

Transkriptomin muodostavat RNA-molekyylit pilkkoutuvat ennen sekvensointia, ja fragmentit täytyy koostaa uudestaan sekvensoinnin jälkeen. Prosessia kutsutaan transkriptomin rekonstruktioksi. Se helpottuu, jos tutkitun lajin genomi on saatavilla, mutta aina näin ei ole. Tämä tutkielma on katsaus transkriptomin rekonstruktioon *de novo* eli ilman referenssigenomia.

De novo-rekonstruktio on mielenkiintoinen algoritminen haaste. Vakiintunut ratkaisu on hyödyntää de Bruijnin verkkoja. Ne ovat laskennallisesti tehokas ratkaisu ongelmaan, jossa rajatusta aakkostosta halutaan koostaa syklinen koostesana, jossa esiintyy jokainen mahdollinen annetun vakion k mittainen sana täsmälleen yhden kerran. Tämän ansiosta ne ovat yleishyödyllinen bioinformatiikan työväline, jolla on sovelluksia myös RNA-Seq:n ulkopuolella.

Transkriptomin rekonstruktion lisäksi RNA-Seq:ssä pyritään kvantifikaatioon eli transkriptien pitoisuuden määritykseen. Luettujen sekvenssien lukumäärää käytetään siinä korvikemuuttujana transkriptin pitoisuudesta. Sekvenssit, joita ei voida liittää yksiselitteisesti yhteen transkriptiin, tuottavat kvantifikaatiotuloksiin epävarmuutta. Jotta tulokset ovat vertailtavissa, ne täytyy normalisoida. Erilaisia normalisointiyksikköjä ovat muun muassa RPKM, FPKM ja TPM.

RNA-Seq-menetelmä on vielä varsin nuori ja kehittyy jatkuvasti. Sekvensointiteknologioiden kolmas sukupolvi eliminoinee monia menetelmään liittyviä epävarmuustekijöitä. RNA-Seq:lle voikin ennustaa varsin valoisaa tulevaisuutta sen siirtyessä hiljalleen teini-iästä kohti täyttä aikuisuutta.

ACM Computing Classification System (CCS):

General and reference → Document types → Surveys and overviews
Applied computing → Transcriptomics
Mathematics of computing → Graphs and surfaces

Avainsanat — Nyckelord — Keywords
RNA, RNA-seq, sekvensointi, geenien ekspressio, de Bruijnin verkot

Säilytyspaikka — Förvaringsställe — Where deposited

Muita tietoja — Övriga uppgifter — Additional information

Sisältö

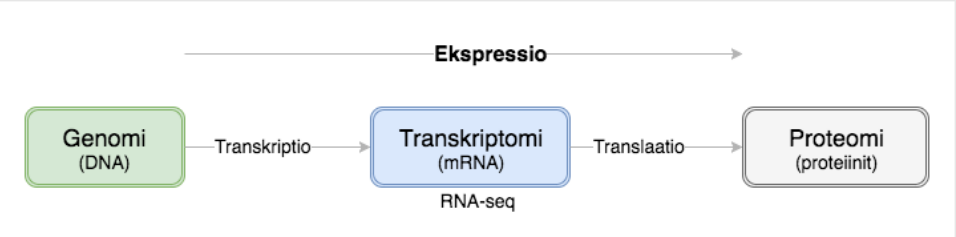
1 Johdanto

RNA-sekvensointi eli RNA-Seq on menetelmä, joka on merkittävässä osassa geenin ilmenemisen eli ekspression tutkimuksessa. Kun geeni ilmenee, siitä tuotetaan ensin RNA-molekyyli eli transkripti, joka sisältää kaiken tarvittavan tiedon proteiinin rakentamiseksi. Prosessia kutsutaan transkriptioksi. RNA-Seq-menetelmässä solun tai solujoukon sisältämät RNA-molekyylit eristetään ja sekvensoidaan.

Ekspressiotutkimus on ensiarvoisen tärkeää solun toiminnan ymmärtämiseksi, sillä solun toiminnanohjaus tapahtuu proteiinien välityksillä. Proteiinit toimivat rakenneosina (esimerkiksi kollageeni) ja ohjaavat entsymaattisesti solun kemiallisia reaktioita. Ne säätelevät myös toisten proteiinien ekspressiota eli toimivat säätelyproteiineina (regulatory proteins). Esimerkiksi solun jakautuminen ja erilaistuminen ovat proteiinien ohjaamia prosesseja, ja niinpä esimerkiksi jatkuvasti jakautuvien syöpäsolujen tai erilaistuneiden kudossolujen ymmärtämiseksi on välttämätöntä tuntea niissä vaikuttava proteiinipopulaatio eli proteomi.

Tutkielmassa esittelen ensin menetelmän biologisen perustan. Tämä on tarpeellista, koska RNA-Seq:ssä (ja bioinformatiikassa yleisemmin) laboratoriomenetelmien rajoitukset vaikuttavat tiedonkäsittelyyn ratkaisevasti. Lisäksi tulokset ovat merkityksellisiä vain elävän solun kontekstissa – datan kerääminen ei ole itseisarvo. Siksi käsittelen tutkielmassa myös esimerkiksi sitä kysymystä, kuinka tarkasti geenien ekspressio voidaan RNA-Seq:n avulla ennustaa.

Kuva 1 on korkean tason esitys proteiinisynteesistä. Nuoli kuvastaa prosessia, ja sen vasemmalla puolella on mallina toimiva molekyyli ja oikealla puolella lopputuote.



Kuva 1: Datan kulku solussa

Kolmas kappale esittelee tarkemmin menetelmän keskeisen ongelman, transkriptomin rekonstruktion. Aihe on rajattu siten, että tutkielma syventyy tarkemmin transkriptomin rekonstruktioon *de novo* eli ilman referenssigenomia. Tämän myötä esittelen tarkemmin de Bruijn-verkot, jotka ovat tärkeä työkalu *de novo* rekonstruktiossa. De Bruijin-verkkojen käytön havainnollistamiseksi olen valinnut Trinity-menetelmän.

Tutkielmassa esitellään myös RNA-Seq-datan kvantifikaatio, joka vastaa kysymykseen *kuinka paljon* kutakin transkriptia on. Tämä luonnollisesti on mahdollista vasta, kun ensin on selvitetty mitä transkripteja tuotetaan. Siksi kvantifikaatio on usein RNA-Seq-menetelmän viimeinen vaihe ja päämäärä.

Tutkielman viimeisessä kappaleessa tehdään yhteenveto RNA-Seq:n nykytilasta. Lisäksi luodaan katse sen jännittävältä näyttävään tulevaisuuteen, joka voi jo lähivuosina tuoda menetelmään aivan uudenlaista tarkkuutta ja helppoutta.

2 Tausta

2.1 Geeniekspressio

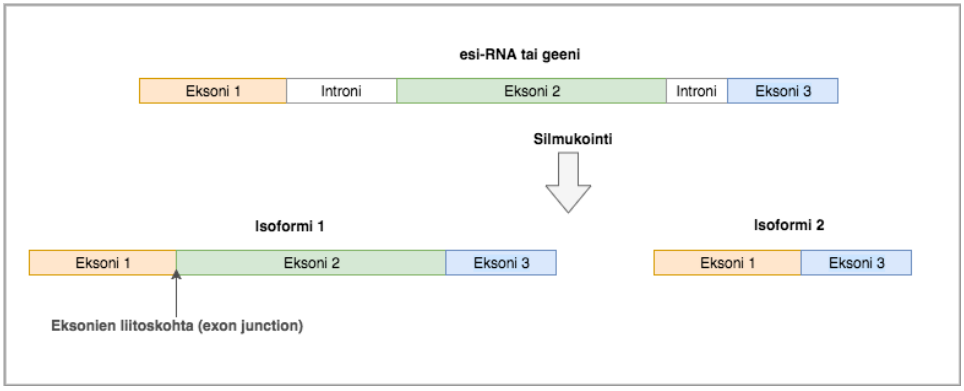
Solun koko perimä eli genomi sisältää sekä geenejä että geenityhjää ainesta. Geeni on genomi-informaation perusyksikkö, joka kykenee itsenäiseen ekspressioon. Samasta geenistä voi syntyä yksi tai useita vaihtoehtoisia proteiineja silmukoinniksi kutsutussa prosessissa, joka esitellään seuraavassa kappaleessa tarkemmin. Geeni sisältää usein myös säätelyalueita, joihin säätelyproteiinit kiinnittyvät joko lisäten tai hilliten kyseisen geenin ekspressiota.

Transkriptiossa tuotetaan geenin koodaaman informaation mukaisesti RNA-transkripti. Solun mRNA-populaatiota kutsutaan transkriptomiksi. Kaikista transkripteista ei kuitenkaan välttämättä tuoteta proteiinia, vaan myös RNA-molekyyli voi olla biologisesti aktiivinen lopputuote solussa. Yhteisnimitys tälle RNA-tyypille on koodaamaton RNA (non-coding RNA eli ncRNA). Vaikka tutkielma jatkossa keskittyy RNA-Seq:n rooliin geenieksprression tutkimuksessa, on hyvä todeta, että ncRNA:lla on tärkeitä esimerkiksi säätelyyn liittyviä tehtäviä solussa ja että RNA-Seq on tärkeä menetelmä sen tutkimuksessa. Tyypillisesti RNA-molekyyli on kuitenkin proteiinisynteesin välituote, jolloin sitä kutsutaan lähetti-RNA:ksi (messenger-RNA eli mRNA). Tällöin transkription jälkeen tapahtuu translaatio, jossa soluelin nimeltä ribosomi rakentaa proteiinin käyttäen mRNA:ta mallina. Myös itse ribosomi koostuu pitkälti proteiineista, mikä on eräs esimerkki proteomin tärkeästä roolista solussa.

2.2 Silmukointi

Eukaryoottigeeneissä on myös proteiinia koodaamattomia intronialueita (introns), jotka poistetaan lopullisesta transkriptista ennen translaatiota. Tätä prosessia kutsutaan silmukoinniksi. Silmukoinnin jälkeen jäljelle jääviä osia kutsutaan eksoneiksi (exons). Intronit muodostavat usein pääosan geeniä, jolloin eksoneiden osuus jää suhteessa pieneksi.

Silmukointi voi usein tapahtua monella eri tavalla samasta molekyylistä, mikä on esitetty kuvassa 2. Näin samasta geenilokuksesta voi syntyä vaihtoehtoisia transkripteja, joita kutsutaan isoformeiksi. Vaihtoehtoista silmukointia säätelee joukko proteiineja, joita kutsutaan silmukointitekijöiksi (splicing factor).



Kuva 2: Esimerkki vaihtoehtoisesta silmukoinnista (ei mittakaavassa)

2.3 Uuden sukupolven sekvensointi

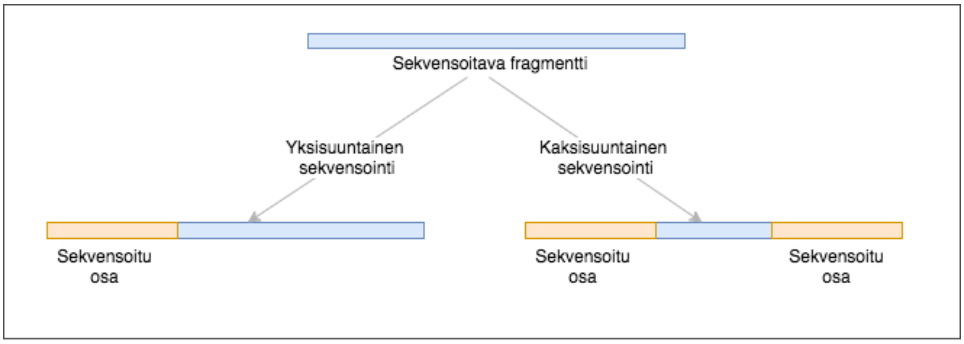
Ensimmäiset DNA-sekvensointimenetelmät kehitettiin jo 1970-luvulla, mutta viimeisen vuosikymmenen aikana ne ovat tehostuneet huomattavasti, mikä on vähentänyt sekvensoinnin kustannuksia ja lisännyt räjähdysmäisesti sekvensoidun datan määrää. Näihin kertaluokkaa tehokkaampiin menetelmiin viitataan uuden sukupolven sekvensoinnilla (next generation sequencing eli NGS). Uuden sukupolven menetelmät ovat toteutukseltaan monimuotoisia mutta niille kaikille on yhteistä samanaikaisuuden hyödyntäminen, minkä vuoksi niitä kutsutaan usein myös laajamittaiseksi rinnakkaissekvensoinniksi (massively parallel sequencing).

Jotta sekvensointi olisi luotettavaa, näytteen sekvenssejä on monistettava. Mitä suurempi sekvensoitujen fragmenttien määrä eli sekvensointisyvyys, sitä luotettavammin transkription laatu ja määrä voidaan määrittää. Toisaalta syvä sekvensointi on kallista ja saattaa lisätä kohinaa testituloksissa

erityisesti vähiten kiinnostavien transkriptien osalta[3]. Siksi sopivan sekvensointisyvyyden valinta on tasapainottelua näiden tekijöiden välillä ja vaihtelee koejärjestelyjen välillä.

Sekvensointilaitteet pystyvät lukemaan rajallisen mittaisia sekvenssejä, joten kirjasto täytyy useimmiten pilkkoa sopivaan mittaan. Lyhyen lukupituuden teknologiat pystyvät lukemaan 35–700 emäsparin pituisia sekvenssejä [6]. Pilkkomisella on perustavanlaatuinen vaikutus sekvenssidatan analyysiin. Viime vuosina maksimilukupituus on kuitenkin kasvanut pitkälukuisten teknologioiden kehittymisen myötä. Esimerkiksi Pacific Biosciences BioSMRT-sekvensointilaite pystyy lukemaan useimmat transkriptit täysmittaisina [3]. Pitkän lukupituuden teknologioiden leviämistä hidastavat kuitenkin niiden kalleus ja hitaus [6]. Koska ne eivät vielä ole laajamittaisessa käytössä, tässä tutkielmassa keskitytään fragmentoituneiden lyhyen lukupituuden sekvensointitulosten analyysiin.

Fragmentista voidaan sekvensoida molemmat päät (kaksisuuntainen sekvensointi) tai vain toinen niistä (yksisuuntainen sekvensointi), mikä on esitetty kuvassa 3. Yksisuuntainen sekvensointi on kaksisuuntaista halvempaa [3]. Sekvenssikirjastot jakautuvat tämän perusteella kahteen ryhmään: parittomiin (single-ended) ja parillisiin (pair-ended). Parillisissa kirjastoissa jokaisen sekvenssin yhteyteen talletetaan tieto vastinparista sekä siitä, kuinka pitkä parin välinen etäisyys on. Aina kaksisuuntainen sekvensointi ei kuitenkaan onnistu, ja tämän vuoksi parillisissakin kirjastoissa osa sekvensseistä on parittomia.



Kuva 3: Sekvensoinnin kaksi tapaa (ei mittakaavassa)

2.4 RNA-Seq-menetelmä

Laajamittainen RNA-Seq tuli mahdolliseksi vasta uuden sukupolven menetelmien kehittymisen myötä, sillä vanhan polven menetelmille urakka on kallis. Ennen RNA-Seq:n kehittymistä transkriptomin tutkimuksessa käytettiinkin

lähinnä räätälöityihin DNA-siruihin (microarray) perustuviin menetelmiä. Ne edellyttävät ennakkotietoa siitä, mitä etsitään, eikä niiden avulla saada RNA-Seq:n kaltaista transkriptomin kokonaiskuvaa. RNA-Seq:n avulla onkin mahdollista löytää aiemmin tuntemattomia geenejä tai isoformeja [13].

RNA-Seq:ssä solusta eristetään ensin laboratoriossa transkriptomi, josta tuotetaan käänteiskomplementti-DNA (cDNA). cDNA:lla tarkoitetaan RNA:sta valmistettua DNA:ta, joka valmistetaan tietyistä viruksista saatavan käänteiskopioijaentsyymin avulla. Näin saadaan DNA-kirjasto, joka informaatioltaan vastaa transkriptomia. Tämä sekvensoidaan käyttäen uuden sukupolven menetelmiä.

RNA-Seq-tulosten tulkinta on haastavaa monesta syystä. Ensinnäkin transkriptien pitoisuudet vaihtelevat huomattavasti, ja heikosti ekspressoidun geenin RNA-tuotetta saattaa edustaa vain muutama molekyyli. Toiseksi sekvensoidut fragmentit voivat olla sekä käsittelemätöntä intronit sisältävää esi-RNA:ta tai jo käsiteltyä pelkästään eksoneista koostuvaa mRNA:ta [5]. Kolmanneksi sekvensoidut fragmentit ovat lyhyitä: ne pilkkoutuvat näytteessä ennen sekvensointia, joten kukin fragmentti on osa suurempaa kokonaisuutta. Geenillä voi myös olla useita isoformeja, joilla on niin paljon yhteisiä osia, että voi olla mahdotonta erottaa, mistä isoformista luettu sekvenssi on peräisin. Neljänneksi eri geenit saattavat olla toisteisia [7]: esimerkiksi geenien kahdentumat eli paralogit tuottavat samankaltaisia transkripteja. Viidenneksi sekvensointi on virhealtista [5].

Oikea koesuunnittelu auttaa saamaan luotettavan lopputuloksen kustannustehokkaalla tavalla. Heikosti ekspressoitujen geenien tutkimus tarkentuu syvällä sekvensoinnilla [3]. Kirjaston parillisuus ja sekvenssien lukupituuden kasvattaminen auttaa, jos päämääränä on *de novo*-rekonstruktio tai isoformien erottelu [3]. Toisaalta jos tutkimuskohteena olevan organismin transkriptomi on jo valmiiksi kattavasti kartoitettu, saadaan halvemmalla parittomalla kirjastolla usein jo riittävän hyvä tulos [3].

Esimerkiksi: jos tutkitun organismin genomi ei ole kovin toisteinen, kuten lituruohon(*Arabidopsis thaliana*) tapauksessa, riittää 50 miljoonaa 50 emäsparin mittaista paritonta sekvenssiä kattamaan ekspressoidut geenit lähes täydellisesti [13]. Jos tutkitaan erityisesti vaihtoehtoista silmukointia, vaaditaan 50-100 emäsparin pituisia parillisia sekvenssejä [13]. Ja jos tutkittava genomi on erittäin toisteinen, kuten riisin (*Oryza sativa*), pitää sekvenssien olla parillisia ja 100-150 emäsparin pituisia[13].

2.5 Proteiini- ja mRNA-pitoisuuksien välinen korrelaatio

Koska mRNA-transkriptin valmistus on ensimmäinen välivaihe proteiinin tuotannossa, voisi olla houkuttelevaa ajatella, että sen pitoisuudesta voisi suoraan ennustaa sen koodaaman proteiinin pitoisuuden. Näin ei kuitenkaan ole. Aiheesta kirjoittavat muun muassa Vogel ja Marcotte Nature-lehdessä [14].

Vogelin ja Marcotten mukaan suoraan – esimerkiksi massaspektrometrilla – mitattuja proteiinien pitoisuuksia verrattaessa vastaavan mRNA-transkriptin pitoisuuteen näiden välillä on keskimäärin korrelaatio muttei kovin vahva sellainen. Tämä pätee sekä eukaryooteilla että prokaryooteilla. On arvioitu, että keskimäärin noin 40 prosenttia proteiinien pitoisuuksien vaihtelusta selittyy eroilla mRNA-pitoisuuksissa (Pearsonin korrelaatiokerroin noin 0,4) [4]. Toisaalta mRNA-pitoisuuden on havaittu olevan hyvä indikaattori vastaavan proteiinin olemassaololle solussa[14]. Jotta jäljelle jäävä 60% vaihtelusta saa selityksen, selittäviä tekijöitä voidaan etsiä sekä transkription jälkeisestä säätelystä että mittausmenetelmien epätäydellisyydesta.

Proteiinin pitoisuuteen vaikuttavat monet muut tekijät mRNA-pitoisuuden lisäksi: esimerkiksi translaation ja hajotusnopeuden säätely. Proteiinien elinikä vaihtelee suuresti: toiset ovat harvoin tuotettuja mutta stabiileja, jolloin niiden pitoisuus voi silti olla suuri. Toisten proteiinin pitoisuuden on havaittu olevan lähempänä ns. dynaamista tasapainoa, jolloin ne ovat jatkuvan translaation ja hajotuksen kohteena. Vogelin ja Marcotten mukaan erot kertovat usein erilaisesta funktiosta: ylläpitoproteiinit (housekeeping proteins), jotka vastaavat solun perustoiminnoista, ovat usein pitkäikäisempiä. Sitä vastoin monet säätelyyn osallistuvat geenit, joiden tehtävä on reagoida muutoksiin, hajotetaan nopeasti. Nämä tekijät on otettava huomioon arvioitaessa RNA-Seq-menetelmässä saatujen tulosten biologista merkitystä.

Mittaustuloksia arvioitaessa on otettava huomioon proteiinien ja RNA-molekyylien erilainen elinikä ja tuotantovauhti. RNA-molekyylin puoliintumisaika on keskimäärin 2.6 - 7 tuntia, ja proteiinimolekyylin vastaava on 46 tuntia [14]. Lisäksi proteiinien keskimääräinen tuotantovauhti on moninkertainen mRNA-molekyyleihin verrattuna. Eräässä yksisolututkimuksessa (single-cell study), jossa verrattiin *E. coli* -bakteerien mRNA- ja proteiinipitoisuuksia, ei havaittu mitään korrelaatiota näiden välillä [11]. Vogel ja Marcotte esittävät, että tällainen tutkimustulos saattaa selittyä juuri edellä mainituilla tekijöillä, sillä erityisesti yhden solun mittakaavassa mRNA- ja proteiinipitoisuuksien aaltoliike tapahtuu eri tahdissa. Solupopulaatiota tutkittaessa ilmiö tasoittuu, ja korrelaatio tulee jälleen näkyviin.

3 Transkriptomin rekonstruktio

Transkriptomin rekonstruktiolla tarkoitetaan transkriptoitujen geenien ja niiden RNA-isoformien kartoitusta. Kuten edellä on todettu, sekvensointivaiheessa näyte fragmentoituu, joten koko loppuanalyysin onnistuminen riippuu siitä, kuinka luotettavasti alkuperäiset transkriptit saadaan koostettua fragmenteista. Transkriptomin rekonstruktio voidaan tehdä joko referenssigenomiin nojautuen (genome-guided reconstruction) tai referenssigenomista riippumattomasti (genome-independent reconstruction tai *de novo* reconstruction). Jos referenssigenomi on olemassa, saadaan siihen nojaavaa menetelmää käyttämällä tarkempi tutkimustulos[5].

Kun referenssigenomi on saatavilla, ensimmäinen askel on tehdä syötteenä oleville fragmenteille kohdistus (read alignment) eli fragmenttien alkuperän etsintä referenssigenomista. Tämän jälkeen pyritään löytämään isoformi, josta kukin löydetty sekvenssivastaavuus on peräisin. Tässä käytettyjä menetelmiä ovat muun muassa Cufflinks[12] ja Scripture[8]. Niiden läheisempi tarkastelu on tämän tutkielman aihealueen ulkopuolella.

Kaikille lajeille, esimerkiksi monille mikrobeille, ei ole olemassa referenssigenomia. Tässä tapauksessa ensimmäinen työvaihe on yhdistää sekvenssifragmenttien joukko suuremmiksi kokonaisuuksiksi, joita kutsutaan jatkumoiksi (contig) [3]. Yleinen jatkumoiden muodostamiseksi käytetty menetelmä on rakentaa fragmenteista de Bruijnin verkko. De Bruijnin verkkoja käyttäviä menetelmiä ovat muun muassa Trinity[7] ja transABySS[10]. De Bruijnin verkot ja Trinity esitellään tutkielmassa tarkemmin.

Joskus valinta referenssiin nojautuvan tai siitä riippumattoman kokoamisen välillä ei ole yksinkertainen. Esimerkiksi syöpäsolujen tapauksessa solun genomi on saattanut mutatoitua niin laajamittaisesti, että vertailu referenssigenomiin ei ole enää mielekäs, vaikka se olisikin olemassa [9]. Referenssigenomin puuttuminen kuormittaa kuitenkin laskennallisesti [5].

4 De Bruijnin verkot

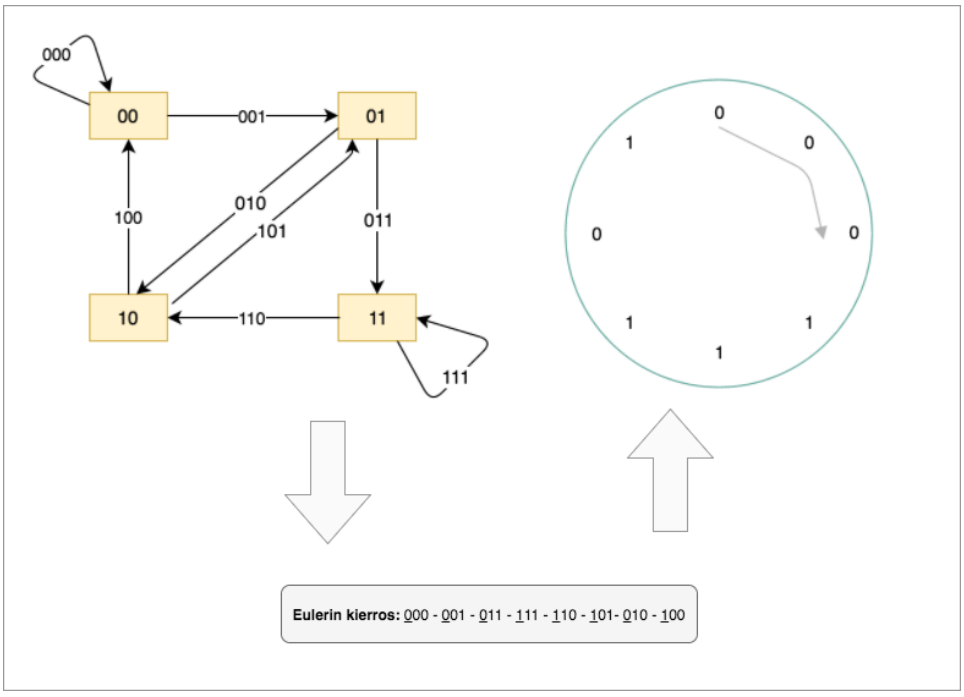
4.1 Teoreettinen tausta

De Bruijnin verkko on saanut nimensä hollantilaisen matemaatikon Nicolaas de Bruijnin mukaan. Muun muassa kombinatoriikkaa tutkinut de Bruijn halusi tietää, voidaanko rajatusta aakkostosta koostaa syklinen koostemerkkijono (superstring), jossa esiintyy jokainen annetun vakion k mittainen sana (k-meeri) täsmälleen yhden kerran [1].

De Bruijn ratkaisi ongelman hyödyntämällä verkkoteoriaa, jonka perusti Leonhard Euler 1700-luvulla. Euler oli sveitsiläinen matemaatikko, joka kiinnostui tunnetusta Königsbergin siltoihin liittyvästä ongelmasta. Siinä kysymyksenä on, kuinka löytää polku, joka kiertää kaupungin jokaisen sillan kautta kerran, ja onko tällainen polku ylipäätään olemassa. Eulerin oivallus oli kuvata jokainen erillinen maa-alue verkon solmuna ja jokainen silta näiden välisenä kaarena[2].

De Bruijn muodosti aakkostosta verkon, jonka solmuja ovat mahdolliset $(k-1)$-pituiset merkkijonot eli $(k-1)$-meerit. Solmusta A suuntautuu kaari solmuun B, jos A:n merkkijono *ensimmäistä merkkiä* lukuunottamatta on sama kuin B:n merkkijono *viimeistä merkkiä* lukuunottamatta. Kaaren merkkijono saadaan lisäämällä sen lähtösolmun merkkijonoon sen maalisolmun merkkijonon viimeinen merkki. Verkon kaarien joukko on samalla mahdollisten k-meerien joukko[2]. De Bruijnin etsimä koostemerkkijono saadaan muodostamalla verkosta Eulerin kierros[2]: polku joka kulkee jokaisen kaaren kautta *täsmälleen kerran* ja palaa alkusolmuun.

4.2 Esimerkki de Bruijnin verkosta



Kuva 4: De Bruijnin verkko kun k=3 ja aakkosto on {0, 1}

Kuvassa 4 on esitetty De Bruijnin verkon muodostus kun k on kolme ja aakkosto koostuu merkeistä 0 ja 1. Koska aakkosto koostuu kahdesta merkistä, mahdollisten k-meerien lukumäärä on $2^k = 2^3 = 8$. Verkon solmut muodostavien $(k-1)$-meerien lukumäärä $2^{k-1} = 4$, ja itse joukko on {00, 01, 10, 11}. 01 suuntautuu kaari solmuun 10, koska 01-solmun viimeinen merkki ja 10-solmun ensimmäinen merkki on sama eli 1. Tällöin kaaren merkkijonoksi tulee näiden yhdistelmä 010. Kuvasta havaitaan myös, että verkon kaarien lukumäärä on sama kuin k-meerien lukumäärä (kahdeksan) ja että kukin mahdollinen k-meeri on esiintyy kaaren merkkijonona täsmälleen kerran.

Kuvassa 4 esitetään yksi verkkoon muodostuva Eulerin kierros alkaen vasemman ylänurkan 00-solmusta. Koostemerkkijono saadaan Eulerin kierroksesta valitsemalla jokaisen kaaren ensimmäinen merkki (kuvassa alleviivattu). Näin syntyvä koostemerkkijono on 00011101. Tämän syklistä esitystä tarkastelemalla voidaan varmistua kunkin k-meerin esiintyvän siinä kerran.

Euler todisti, että suuntautuneesta verkosta löytyy Eulerin kierros, jos se on tasapainotettu eli jokaisesta solmusta johtaa ulos yhtä monta kaarta kuin sisään[2]. Tämä pätee kuvan 4 verkolle, jonka Eulerin kierros esitettiin edellä. Eräs verkon ominaisuus on, että solmun lähtevien ja tulevien määrä kertoo sen, kuinka monta kertaa solmun merkkijono esiintyy koostemerkkijonossa[2].

4.3 Soveltaminen sekvenssidataan

De Bruijnin verkkojen rooli sekvenssidatan analyysissa on muodostaa fragmenteista suurempia kokonaisuuksia eli jatkumoita. Kuten edellä todetaan, de Bruijnin alkuperäinen ongelmanasettelu oli muodostaa koostemerkkijono, jossa jokainen k:n mittainen osamerkkijono esiintyy kerran. Tämän myötä on helppo nähdä, kuinka de Bruijnin verkkoja voidaan hyödyntää jatkumoiden muodostamisessa.

Aakkosto sisältää merkit A, C, G ja T. Syötteenä olevat eripituiset DNA-fragmentit pilkotaan k:n mittaisiksi paloiksi. Verkon solmut muodostuvat siis *kaikkien löydettyjen sekvenssien* pohjalta (toisin kuin esitetyssä esimerkissä 4, jossa solmut muodostettiin kaikkien mahdollisten kombinaatioiden pohjalta).

De Bruijnin verkkoja sovellettaessa sekvenssidatan konstruktioon prosessi asettaa tiettyjä haasteita. Sekvensoinnissa tapahtuu virheitä (1), sekvensoidussa syötteessä on aukkoja (2), sama k:n mittainen sekvenssi esiintyy rekonstruoitavassa transkriptissa useammin kuin kerran (3), ja koostemerkkijono on lineaarinen eikä syklinen (4) [2]. Erityisesti lineaarisuusongelma voidaan ratkaista siten, että De Bruijnin verkosta etsitään Eulerin kierroksen sijaan Eulerin polku, jossa lähtösolmu voi olla eri kuin loppusolmu[2].

De Bruijnin verkkoja käyttävät menetelmät tasapainoilevat haasteiden välillä esimerkiksi säätämällä k:n arvoa. Mitä matalampi k:n arvo on, sitä monimutkaisempi verkosta tulee, sillä saman solmun kautta kuljetaan useammin. Optimaalinen k:n arvo riippuu sekvensointisyvyydestä: kun sekvensointisyvyys on matala, matalat k:n arvot ovat edullisia, sillä näin erilaisten k-meerien määrä lisääntyy [5]. Tällä tavoin monet sellaisetkin k-meerit, jotka ovat jääneet sekvensoimatta, tulevat edustetuksi verkossa. Toisaalta kun sekvensointisyvyys on korkea ja k:n arvo on matala, sekvensointivirheet vaikuttavat verkon muodostukseen liikaa [5]. Sekvensointisyvyyden ollessa korkea kannattaa siten k:n arvoakin kasvattaa.

De Bruijnin verkkoa voidaan käyttää sekä genomin että transkriptomin konstruktioon. Genomidatan konstruktiossa syntyy vain joitakin suuria verkkokokonaisuuksia, joista kukin edustaa yhtä kromosomia. Yksi geenilokus kytkeytyy vain yhteen jatkumoon. Monien organismien DNA-sekvenssi sisältää geenityhjiä toistojaksoja, joiden tarkka rekonstruktio voi olla haastavaa fragmenttien samanlaisuuden vuoksi.

Transkriptomidata on taas luonteeltaan sirpaloitunutta, jolloin oletettu lopputulos sisältää monia verkkoja, joista kukin edustaa yhtä geeniä [7]. Transkription määrästä riippuen samasta lokuksesta voi syntyä lukemattomia transkripteja. Toisaalta transktiptomidata on genomista dataa tiiviimpää eikä sisällä toistojaksoja [9].

De Bruijnin verkko ei ole ainoa tapa yhdistää fragmentoitunut DNA-data, mutta se on laskennallisesti tehokas verrattuna vaihtoehtoisiin menetelmiin. Esimerkiksi ihmisen genomin kartoitusprojektissa vuonna 2001 jatkumoiden muodostamisessa käytettiin Hamiltonin menetelmää. Se eroaa jo esitellystä menetelmästä muun muassa seuraavalla tavalla: Eulerin kierros käy kerran jokaisen *kaaren* kautta mutta Hamiltonin kierros käy kerran jokaisen *solmun* kautta. Tämä kuuluu ns. NP-täydellinen ongelmien joukkoon, joille ei ole löydettävissä todennettua tehokasta ratkaisua ainakaan tällä hetkellä. Verkon esittämistavalla on siten merkittävä vaikutus laskennan kuormittavuuteen.

5 Trinity

Trinity[7] on De Bruijnin verkkoihin nojaava RNA-Seq-datan *de novo*-konstruktioon erikoistunut menetelmä. Se jakautuu kolmeen vaiheeseen, mistä johtuu nimi Trinity (kolminaisuus). Vaiheet ovat Inchworm, Chrysalis ja Butterfly, ja niiden välinen työnjako on karkeasti seuraava: Inchwormissa tapahtuu alustavien jatkumoiden muodostus, joiden pohjalta Chrysalis-vaiheessa rakennetaan De Bruijnin verkot. Butterfly-vaiheessa verkkojen polut analysoidaan [7]. Seuraavassa kukin vaihe esitellään tarkemmin.

Trinityn käyttämä oletusarvo k:lle on 25. Analyysin onnistumisen kannalta on tärkeää, että arvo on paljon lyhyempi kuin sekvenssin lukupituus [7].

5.1 Inchworm

Inchworm ottaa syötteenä koko sekvenssijoukon. Alustavat jatkumot muodostetaan seuraavalla tavalla[7]:

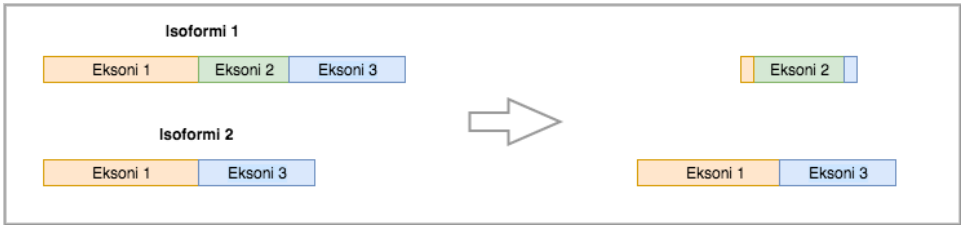
1. **K-meerilistan luonti** Syötteenä olevat fragmentit pilkotaan k-mittaisiksi fragmenteiksi, joiden esiintymisfrekvenssit lasketaan.

2. **Siemenen valinta** K-meerilistasta valitaan useimmin esiintyvä k-meeri siemeneksi.

3. **Laajennus** Siementä laajennetaan molemmista päistä etsimällä fragmentteja, joilla on $(k-1)$-pituinen sekvenssivastaavuus sen kanssa. Jos tällainen fragmentti löytyy, se yhdistetään jatkumoksi siemenen kanssa. Samalla jatkumoon liitetyt fragmentit poistetaan k-meerilistasta.

4. **Toisto** Prosessin vaiheet 2 ja 3 toistetaan kunnes k-meerilista on tyhjä.

Todettakoon, että lopputuloksena saatava jatkumoiden joukko ei vielä tässä vaiheessa ole käyttökelpoinen esitys eri isoformeista ja voi antaa vääristyneen kuvan transkriptomista [7]. Esimerkki tästä on esitetty kuvassa 5. Oletetaan että näytteessä on todellisuudessa kahta isoformia 1 ja 2 (kuvan vasen puoli) ja että isoformia 2 on runsaasti ja isoformia 1 vähemmän. Tällöin Inchworm saattaa rakentaa oikein vain isoformin 2, ja muodostaa virheellisesti isoformin 1 yksilöllisestä osasta jatkumon (kuvan oikea puoli).



Kuva 5: Esimerkki Inchwormin virheellisistä jatkumoista (ei mittakaavassa)

5.2 Chrysalis

Chrysalis-vaiheessa Inchwormin tuottamista jatkumoista rakennetaan de Bruijnin verkko. Siinä tapahtuu kaksi asiaa [7]:

1. **Jatkumoiden ryhmittely** Jatkumot sijoitetaan samaan ryhmään, jos:

    1. niiden välillä on $(k-1)$–pituinen sekvenssivastaavuus ja

2. on löydettävissä riittävästi yhdistäviä fragmentteja, jotka ulottuvat $(k-1)/2$-merkkiä liitoskohdan molemmin puolin.

Jatkokäsittelyn kannalta on oleellista, että ryhmät on muodostettu siten, että *kaikki samaan ryhmään kuuluvat jatkumot ovat todennäköisesti peräisin saman geenin eri isoformeista tai geenien paralogeista.* Ryhmittelystä eteenpäin Trinityn laskenta on myös helposti rinnakkaistettavissa eri säikeisiin[9].

2. De Bruijnin verkkojen rakentaminen Kustakin ryhmästä rakennetaan oma De Bruijnin verkko. Kaaren pituus on k ja solmun pituus on $(k-1)$ kuten esimerkissä 4. Jokainen verkon kaari painotetaan alkuperäisen fragmenttijoukon perusteella.

5.3 Butterfly

Butterfly on Trinityn kolmas ja viimeinen vaihe, jossa jokainen Chrysaliksen rakentama verkko redusoidaan joukoksi transkripteja. Ensimmäisenä tapahtuu verkon yksinkertaistaminen, jossa verkon solmuja sulautetaan yhteen pidempien kokonaisuuksien muodostamiseksi (1) sekä sellaisia kaaria karsitaan (2), joiden Trinity arvelee syntyneen sekvensointivirheiden seurauksena [7].

Viimeisenä vaiheena Butterflyssa tapahtuu polkujen pisteytys (plausible path scoring)[7], jossa lasketaan lopulliset polut eli transktiptit. Butterly hyödyntää polkujen todennäköisyyksien laskennassa alkuperäistä sekvenssidataa, jossa fragmentti on usean k:n pituinen [7]. Tämä perustuu siihen, että de Bruijnin verkon kaaren pituus on k, jolloin yksi fragmentti vastaa useaa kaarta eli verkon osapolkua.

6 Transkriptomin kvantifikaatio

Rekonstruktion jälkeinen vaihe on transkriptomin kvantitatiivinen määritys, jossa voidaan tutkia joko yksittäisen transkriptin pitoisuutta (transkriptitason tutkimus) tai kaikkien geenin tuottamien transkriptien pitoisuutta (geenitason tutkimus). Pitoisuuksia ei voida suoraan mitata, joten löydettyjen sekvenssien lukumäärää käytetään korvikemuuttujana pitoisuudesta[4]. Tämä edellyttää, että mitä enemmän transkriptia näytteessä on, sitä todennäköisemmin siitä syntynyt fragmentti päätyy sekvensoitavaksi. Tämän oletuksen paikkansapitävyys on vahvistettu kokeellisesti.

RNA-Seq-tuloksissa osa vaihtelusta on aina koeteknistä: esimerkiksi sekvensointisyvyys tai jopa sekvensoinnin aikana vallinnut lämpötila tuottavat tuloksiin vaihtelua, joka ei kerro mitään tutkimushypoteesista. Biologinen

taas on juuri sitä vaihtelua, jota tutkimuksella pyritäänkin mittaamaan (eli vaihtelua organismien tai muiden verrokeiden välillä). Esimerkiksi kahden eri kudoksen välillä oleva ero tietyn geenin ekspressiossa on biologista vaihtelua. Koe on mielekäs vain, jos se mittaa ensisijaisesti biologista eikä koeteknistä vaihtelua.

6.1 Monikytkeytyvät sekvenssit

Määrällisen analyysin ensimmäinen askel on käydä läpi rekonstruoitu transkriptomi ja laskea niiden alkuperäisestä raakadatasta löytyvien sekvenssien lukumäärä, jotka ovat peräisin kustakin geenistä tai transkriptista. Jatkossa tähän lukuun viitataan tutkielmassa lähdesekvenssimäärällä (LSM).

Koska geenin isoformit jakavat keskenään samoja eksoneita, ja monilla geeneillä on genomissa paralogeja, osaa sekvensseistä on mahdoton kytkeä vain yhteen transkriptiin. Näitä kutsutaan monikytkeytyviksi sekvensseiksi (multireads). Ilmiötä nimitetään alkuperäepävarmuudeksi (read assignment uncertainty) [5]. Monikytkeytyvät sekvenssit voivat johtaa systemaattiseen virheeseen kvantifikaatiossa sen perusteella, kuinka suuri osa transkriptista on yksilöllistä suhteessa muihin transkripteihin[13].

Eräs strategia on yksinkertaisesti jättää laskematta useampaan transkriptiin mahdollisesti liittyvät sekvenssit. Tämä kuitenkin johtaa vähän yksilöllisiä osia sisältävien transkriptien pitoisuuden aliarviointiin [13]. Toisaalta monikytkeytyvien sekvenssien jakaminen tasaisesti mahdollisten alkuperätranskriptien voi myös johtaa vääristyneeseen pitoisuusarviooon [13]. Ongelman ratkaisemiseksi on kehitetty herkempiä tilastollisia malleja, jotka pyrkivät minimoimaan kvantifikaatiotarkkuuden menetyksen [13].

6.2 Normalisointi

RNA-Seq-menetelmän kvantifikaatio on luonteeltaan pikemmin vertailevaa kuin absoluuttista. Kysymyksenasettelu on siten, että onko näytteiden välillä eroa (esimerkiksi syöpäsolu ja verrokki). Jotta vertailu on mahdollista, tulokset näytteiden välillä täytyy normalisoida.

Tyypillisessä sekvensointikokeessa on kaksi merkittävää normalisoitavaa tekijää. Ensimmäinen on sekvensointisyvyys [5]. Jos esimerkiksi näytteen A:n sekvensointisyvyys on puolet näytteen B sekvensointisyvyydestä, näytteen B jokainen transkripti - vaikka niiden ekspressio olisi yhtä suuri - tuottaa kaksinkertaisen määrän laskennallisia sekvenssejä A:han nähden. Toinen normalisoitava tekijä on transkriptin pituus, mikä seuraa näytteen fragmentoitumisesta. Pilkkoutuessaan pitkä transkripti tuottaa enemmän fragmentteja, jolloin se olisi analyysissa yliedustettu ilman normalisointia [5].

Differentiaaliekspressiossa tutkitaan *saman geenin tai transkriptin* ekspressiota eri näytteissä kuten esimerkiksi syöpäsolussa ja verrokissa. Koska vertailu kohdistuu yhteen geeniin, tulokset on siten normalisoitava vain sekvensointisyvyyden osalta muttei transkriptin pituuden osalta [3].

Seuraavaksi esittelen eri normalisointiyksikköjä taulukossa 1 esitetyn esimerkkiaineiston avulla. Aineisto sisältää kahden näytteen transkriptikohtaiset lähdesekvenssimäärät (sarake LSM). Kokonaissekvenssimäärällä (KSM) tarkoitetaan *yhden näytteen* sekvenssien yhteismäärää.

Luvut on valittu havainnollistamaan laskentaa, eivätkä ne välttämättä vastaa kovin hyvin tyypillistä RNA-Seq-dataa. Esimerkiksi jokaisen transkriptin pituudeksi on valittu kaksi kiloemästä laskennan helpottamiseksi. Lisäksi oikeassa kokeessa transkripteja löytyisi enemmän.

	Näyte 1 LSM	Näyte 2 LSM
Transkripti A (2Kb)	60	180
Transkripti B (2Kb)	20	60
Transkripti C (2Kb)	20	160
KSM	100	400

Taulukko 1: Esimerkkiaineisto (luvut miljoonaa)

Esimerkkiaineistosta voi tehdä huomion, että transkriptien A ja B välinen suhde näytteiden välillä on sama. Toisin sanoen näytteestä 2 löytyy kolminkertainen määrä kumpaakin transkriptia suhteessa näytteeseen 1. Tällainen löydös selittyy erolla näytteiden sekvensointisyvyydessä eikä edellytä eroa transkriptiotasoissa. Sitä vastoin tämä ei päde transkriptiin C, joka selvästi esiintyy runsaammin näytteessä 2 verrattuna näytteeseen 1.

6.2.1 RPKM ja FPKM

RPKM (**reads** per kilobase million) ja FPKM (**fragments** per kilobase million) normalisoivat sekä transkriptin pituuden että sekvensointisyvyyden. RPKM lasketaan seuraavalla kaavalla [3]. Kaavassa TP tarkoittaa transkriptin pituutta.

$$RPKM = \frac{LSM}{KSM(milj.sekvenssiä) * TP(kiloemästä)}$$

	Näyte 1		Näyte 2	
Transkr.	LSM	RPKM	LSM	RPKM
A	60M	60M/(2*100) = 300K	180M	180M/(2*400) = 225K
B	20M	20M/(2*100) = 100K	60M	60M/(2*400) = 75K
C	20M	20M/(2*100) = 100K	160M	160M/(2*400) = 200K
KSM	100M		400M	

Taulukko 2: RPKM-arvot (M=miljoona, K=tuhat)

Taulukon 2 tiivistämiseksi miljoona on lyhennetty M-kirjaimella ja tuhat K-kirjaimella. Kunkin transkriptin pituus on kaksi kiloemästä. RPKM-sarakkeessa jakaja on siten kaksi kertaa vastaava kokonaissekvenssimäärä(KSM) miljoonissa.

FPKM eroaa RPKM-arvosta siten, että se laskee fragmentteja eikä yksittäisiä sekvenssejä. Parillisissa kirjastoissa yksi fragmentti luetaan molemmista päistä, jolloin yksi fragmentti tuottaa kaksi sekvenssiä. FPKM-arvossa samasta fragmentista syntynyt sekvenssipari lasketaan vain kerran. Siksi parittomien sekvenssikirjastojen tapauksessa raportoidaan yleensä RPKM-arvo ja parillisten tapauksessa FPKM-arvo.

6.2.2 TPM

Kuten RPKM, myös TPM (**transcripts** per million) normalisoi sekvensointisyvyyden ja transkriptin pituuden. TPM:n laskemisessa käytetään RPK-arvoa (reads per kilobase million), joka on LSM jaettuna transkriptin pituudella. Toisin sanoen se vastaa RPKM-arvoa ilman kokonaissekvenssimäärän normalisointia.

TPM lasketaan transkriptille t seuraavalla tavalla:

1. RPK-arvon laskeminen transkriptille.
2. Skaalauskertoimen laskeminen. Skaalauskerroin lasketaan summaamalla kaikkien saman näytteen transkriptien RPK-arvot ja jakamalla tulos miljoonalla. Transkripteja on yhteensä n kappaletta.
3. TPM saadaan jakamalla transkriptin RPK-arvo skaalauskertoimella.

$$TPM_t = \frac{RPK_t}{\sum_{i=1}^{n} RPK_i * (1/miljoona)}$$

	Näyte 1			Näyte 2		
Trans.	LSM	RPK	TPM	LSM	RPK	TPM
A	60M	60M/2=30M	30M/50=600K	180M	180M/2=90M	90M/200=450K
B	20M	20M/2=10M	10M/50=200K	60M	60M/2=30M	30M/200=150K
C	20M	20M/2=10M	10M/50=200K	160M	160M/2=80M	80M/200=400K
Yht.	100M	50M	1M	400M	200M	1M

Taulukko 3: RPK- ja TPM-arvot (M=miljoona, K=tuhat)

Taulukossa 3 esitetään jälleen esimerkkiaineistosta lasketut RPK- ja TPM-arvot. Skaalauskerroin ensimmäiselle näytteelle on kaikkien RPK-arvojen summa eli 50 miljoonaa jaettuna miljoonalla, minkä tuloksena saadaan 50. Toiselle näytteelle se on vastaavasti 200.

Taulukosta 3 ilmenee myös eräs TPM-arvon hyödyllinen ominaisuus, joka on etu RPKM-arvoon nähden. Eri näytteiden TPM-arvot nimittäin summautuvat lukuun kaikki samaan lukuun (esimerkkiaineistossa tämä on yksi miljoona). Tämän ansiosta eri näytteiden välillä on helppo vertailla eri transkriptien suhteellisia osuuksia.

7 Yhteenveto

RNA-Seq-menetelmässä solun transkriptomi sekvensoidaan, ja sekvenssidata analysoidaan. RNA-Seq:n haasteet liittyvät siten sekä laboratoriomenetelmiin että tiedonkäsittelyyn ja algoritmeihin. Vaikka menetelmä on vielä varsin nuori, se on jo ehtinyt levitä laajalle ja tullut osaksi bioinformatiikan perustyökalupakkia. RNA-Seq:n ympärille on kehittynyt monipuolinen apumenetelmien ekosysteemi, joka auttaa ratkomaan niin transkriptomin rekonstruktioon kuin kvantitatiiviseen määritykseen liittyviä ongelmia.

De Bruijn-verkot on tärkeä *de novo*-rekonstruktiossa käytetty työkalu. De Bruijnin verkkoa voidaan käyttää sekä genomin että transkriptomin rekonstruktiossa, sillä kummankin sekvensointi edellyttää lähdemateriaalin pilkkomista. Genomidatan kokoamisessa syntyy vain joitakin suuria verkkokokonaisuuksia, joista kukin edustaa yhtä kromosomia. Transkriptomidata

on taas luonteeltaan sirpaloitunutta, jolloin oletettu lopputulos sisältää monia verkkoja, joista kukin edustaa yhden geenin transkriptiotuotteita eli eri isoformeja.

Tällä hetkellä sekvensointiteknologia kehittyy huimaa vauhtia, mikä eliminoinee tulevaisuudessa monia tiedonkäsittelyn ongelmia. Vielä kehitteillä oleva yksittäissekvensointi (single molecule sequencing) pystyy lukemaan transkriptit ilman pilkkomista ja monistamista [13]. Tällöin sekä transkriptomin rekonstruktioon, monikytkeytyviin sekvensseihin että normalisointiin liittyvät ongelmat jäänevät historiaan. Tällä on niin syvällinen vaikutus bioinformatiikan koko kenttään, että yksittäissekvensoinnista puhutaan jo sekvensoinnin kolmantena sukupolvena[13].

Kolmannen sukupolven sekvensointi sekä vähentää ratkaisevasti laskennan tarvetta että parantaa tulosten tarkkuutta [13]. Sen myötä transkriptomi voidaan määrittää ennennäkemättömän tarkasti ja helposti, mikä voi tulevaisuudessa johtaa biologisen tutkimuksen uusille ja jännittäville poluille.

Lähteet

[1] Bruijn, N. G. de: *A Combinatorial Problem.* Koninklijke Nederlandsche Akademie Van Wetenschappen, 49(6):758–764, kesäkuu 1946.

[2] Compeau, P.E.C., Pevzner, P.A. ja Tesler, G.: *How to apply de Bruijn graphs to genome assembly.* Nature Biotechnology, 29(11):987–991, 2011.

[3] Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M.W., Gaffney, D.J., Elo, L.L., Zhang, X. ja Mortazavi, A.: *A survey of best practices for RNA-seq data analysis.* Genome Biology, 17(1), 2016.

[4] De Sousa Abreu, R., Penalva, L.O., Marcotte, E.M. ja Vogel, C.: *Global signatures of protein and mRNA expression levels.* Molecular BioSystems, 5(12):1512–1526, 2009.

[5] Garber, M., Grabherr, M.G., Guttman, M. ja Trapnell, C.: *Computational methods for transcriptome annotation and quantification using RNA-seq.* Nature Methods, 8(6):469–477, 2011.

[6] Goodwin, Sara, McPherson, John D. ja McCombie, W. Richard: *Coming of age: ten years of next-generation sequencing technologies.* Nat Rev Genet, 17(6):333–351, kesäkuu 2016.

[7] Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z.,

Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., Di Palma, F., Birren, B.W., Nusbaum, C., Lindblad-Toh, K., Friedman, N. ja Regev, A.: *Full-length transcriptome assembly from RNA-Seq data without a reference genome.* Nature Biotechnology, 29(7):644–652, 2011.

[8] Guttman, M., Garber, M., Levin, J.Z., Donaghey, J., Robinson, J., Adiconis, X., Fan, L., Koziol, M.J., Gnirke, A., Nusbaum, C., Rinn, J.L., Lander, E.S. ja Regev, A.: *Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs.* Nature Biotechnology, 28(5):503–510, 2010.

[9] Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., Couger, M.B., Eccles, D., Li, B., Lieber, M., Macmanes, M.D., Ott, M., Orvis, J., Pochet, N., Strozzi, F., Weeks, N., Westerman, R., William, T., Dewey, C.N., Henschel, R., Leduc, R.D., Friedman, N. ja Regev, A.: *De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis.* Nature Protocols, 8(8):1494–1512, 2013.

[10] Robertson, Gordon, Schein, Jacqueline, Chiu, Readman, Corbett, Richard, Field, Matthew, Jackman, Shaun D, Mungall, Karen, Lee, Sam, Okada, Hisanaga Mark, Qian, Jenny Q, Griffith, Malachi, Raymond, Anthony, Thiessen, Nina, Cezard, Timothee, Butterfield, Yaron S, Newsome, Richard, Chan, Simon K, She, Rong, Varhol, Richard, Kamoh, Baljit, Prabhu, Anna Liisa, Tam, Angela, Zhao, YongJun, Moore, Richard A, Hirst, Martin, Marra, Marco A, Jones, Steven J M, Hoodless, Pamela A ja Birol, Inanc: *De novo assembly and analysis of RNA-seq data.* Nat Meth, 7(11):909–912, marraskuu 2010.

[11] Taniguchi, Y.: *Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells (Science (2010) (533)).* Science, 334(6055):453, 2011.

[12] Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., Van Baren, M.J., Salzberg, S.L., Wold, B.J. ja Pachter, L.: *Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation.* Nature Biotechnology, 28(5):511–515, 2010.

[13] Verk, Marcel C. Van, Hickman, Richard, Pieterse, Corné M.J. ja Wees, Saskia C.M. Van: *RNA-Seq: revelation of the messengers.* Trends in Plant Science, 18(4):175 – 179, 2013, ISSN 1360-1385.

[14] Vogel, C. ja Marcotte, E.M.: *Insights into the regulation of protein abundance from proteomic and transcriptomic analyses.* Nature Reviews Genetics, 13(4):227–232, 2012.