Tampereen yliopisto

Reko Salonen

YKSITTÄISTEN SOLUJEN RNA-SEKVENSOINTIDATAN KLUSTEROINTI

Lääketieteen ja terveysteknologian tiedekunta
Kandidaatin tutkielma
Huhtikuu 2021

TIIVISTELMÄ

Reko Salonen: Yksittäisten solujen RNA-sekvensointidatan klusterointi
Kandidaatin tutkielma
Tampereen yliopisto
Bioteknologian tutkinto-ohjelma
Huhtikuu 2021

Yksittäisten solujen RNA-sekvensointidata (scRNA-sekvensointidata) on niukan RNA-lähtömateriaalin takia paljon häiriösignaaleja ja puuttuvaa tietoa sisältävä aineisto. Tämän tuoreen teknologian pohjalta voidaan kuitenkin tarkastella ihmisen elimistön kudoksia ennennäkemättömällä tarkkuudella. Ainutlaatuisen datatyypin hyödyntämiseksi on tärkeää kehittää tietokoneella tapahtuvaa analyysityönkuvaa. Siinä keskiössä on aineiston klusterointi, sillä siihen pohjautuu suuri osa jatkoanalyyseista, joiden perusteella tehdään lopullisia päätelmiä kudoksen toiminnasta. Klusterointi on hyvin tunnettu koneoppimisen tieteenalan menetelmä, jonka tarkoituksena on ryhmitellä samanlaiset aineiston havainnot yhteen. Yksittäisten solujen RNA-sekvensointidatan kohdalla tämä tarkoittaa transkriptomiltaan samanlaisten solujen, tyypillisesti samojen solutyyppien, ryhmittelemistä yhteen.

Tässä työssä tarkoituksena oli tutustua korkeaulotteisen eturauhasen scRNA-sekvensointidatan klusterointiin ja saada se onnistumaan niin, että klusterit edustaisivat kudoksen solubiologiaa. Tavoitteena oli siis saada klusteroinnilla muodostumaan solurykelmät, jotka vastaisivat eturauhasen solutyyppejä, eivätkä esimerkiksi teknistä tai biologista häiriösignaalia. Korkeaulotteisen datan käsittelyssä haasteena on solujen välisten etäisyyksien merkityksen pieneneminen, mikä hankaloittaa solujen eroavaisuuksien ja yhtäläisyyksien havaitsemista. Lisäksi korkeaulotteisen aineiston käyttäminen sellaisenaan vaatisi tietokoneelta paljon muistia ja laskentatehoa, ja joka tapauksessa analyysivaiheet kestäisivät huomattavan kauan. Tähän haasteeseen tässä työssä vastattiin ulotteisuuden pienentämisellä eli dimensioreduktiolla.

Ulotteisuutta aineistossa pienennettiin paljon varioivien geenien valinnalla, mikä myös vahvistaa biologisesti merkittävää signaalia. Keskeisimpänä menetelmänä dimensioreduktiossa käytettiin pääkomponenttianalyysia (PCA), joka pyrkii puristamaan alkuperäisessä aineistossa olevan informaation pienempään määrään muuttujia, joita kutsutaan pääkomponenteiksi. Näistä valittiin jatkoon informatiivisimmat, joiden avulla suoritettiin graafipohjainen klusterointi. Graafipohjainen klusterointi toteutettiin luomalla ensin jaetun lähimmän naapurin graafi eli verkkorakenne, joka sitten eroteltiin samankaltaisten solujen rykelmiksi Louvainin algoritmilla. Lopuksi klusteroinnin onnistumista tarkasteltiin kaksiulotteisesti epälineaarisen UMAP-dimensioreduktion avulla.

Käytetyllä työnkuvalla saatiin aikaiseksi onnistunut klusterointi, joka erotteli eturauhasen erilaiset solutyypit toisistaan. Tämä pystyttiin havaitsemaan geenimarkkereiden ilmenemisen visualisoinnilla kaksiulotteisessa UMAP-koordinaatistossa. Klustereiden vastaavuus tarkasteltavan kudoksen solubiologiaan onkin tärkeä varmistaa jälkikäteen. Koko klusterointiprosessin havaittiin olevan aikaa vievää ja osin subjektiivista. Tämä johtuu muun muassa yksikäsitteisen ratkaisun puuttumisesta dimensioreduktiolla saatujen pääkomponenttien ja klusterointiresoluutioarvon valinnassa. Tulevaisuudessa saadaan toivottavasti luotua yhtenäistettyä työnkuvaa erilaisilla scRNA-sekvensointiprotokollilla tuotetuille erikokoisille aineistoille, mikä nopeuttaisi analyysia ja tekisi tuloksista vertailukelpoisempia.

Avainsanat: klusterointi, dimensioreduktio, scRNA

ALKUSANAT

Tämä luonnontieteiden kandidaatin tutkielma on tehty Tampereen yliopiston Lääketieteen ja terveysteknologian tiedekunnassa. Tutkielman aihe pohjautuu Laskennallisen biologian tutkimusryhmässä vuoden 2020 lopussa tekemääni työhön.

Tahdon kiittää Laskennallisen biologian tutkimusryhmän johtajaa ja työni ohjaajaa Matti Nykteriä mahdollisuudesta työskennellä mielenkiintoisen aiheen parissa. Häneltä olen saanut aina pyytäessäni apua ja ohjausta matkan varrella. Kiitän myös läheisiäni ja ystäviäni kommenteista ja hetkistä kanssanne muuten niin opiskeluntäyteisessä arjessani.

Tampereella, 24.04.2021

Reko Salonen

SISÄLLYSLUETTELO

1 JOHDANTO

Molekyylibiologian keskusdogmi on DNA:n geenien sisältämän informaation siirtäminen RNA-molekyyleihin, joista osa edelleen translaatiossa käännetään proteiineiksi. Solun RNA-koostumus eli transkriptomi määrittää perustavanlaatuisesti solun identiteettiä, minkä takia sen tutkiminen on välttämätöntä solun normaalin ja häiriintyneen toiminnan ymmärtämiseksi. Perinteisillä RNA-sekvensointiteknologioilla (bulk RNA-sekvensointi) voidaan selvittää, millaisia RNA-molekyylejä tietty kudos ilmentää ja missä määrin. Ongelmana on niiden rajoittuminen kudostasolle, minkä takia saadut tulokset kertovat ainoastaan usean solun keskivertoisen geenieksprression, jolloin yksittäisten solujen väliset biologisesti merkittävät erot jäävät havaitsematta (Picelli 2017; Hwang ym. 2018).

Yksittäisten solujen RNA-sekvensointi (single-cell RNA sequencing, scRNA-sekvensointi) mahdollistaa solun transkriptomin mittaamisen parhaimmalla mahdollisella tarkkuudella. Tämän avulla voidaan esimerkiksi löytää uusia solutyyppejä kudoksesta, selvittää geenisäätelyverkostoja ja tutkia hyvin heterogeenisten syöpäkasvainten syntyä, kehitystä ja vastetta lääkehoidolle (Hwang ym. 2018; Papalexi ja Satija 2018).

Yksittäisten solujen RNA-sekvensoinnista saatu data on korkeaulotteista sisältäen suuren määrä soluja ja ekspressioarvoja jopa 25 000 geenille (Luecken ja Theis 2019). Korkeaulotteisen datan analysointi on haastavaa, minkä takia ulotteisuutta pyritään pienentämään dimensioreduktiomenetelmillä. Dimensioreduktion avulla vaadittu laskentateho pienenee, kun voidaan verrata vain muutamia dimensioita monien tuhansien geenien sijaan. Aineistossa olevaa häiriösignaalia saadaan myös vähennettyä. Lisäksi solujen samankaltaisuuden ja erilaisuuden vertaaminen helpottuu, kun niin kutsuttua dimensiokirousilmiötä pienennetään (Andrews ja Hemberg 2018; Kiselev ym. 2019; Amezquita ym. 2020, luku 9). Ulotteisuuden pienentäminen on mahdollista, sillä scRNA-sekvensointidata on pohjimmiltaan pieniulotteista (Heimberg ym. 2016). Tämä on seurausta geenisäätelystä, jossa tietty biologinen prosessi, kuten ligandin sitoutuminen reseptoriin, muuttaa usean geenin ilmenemistä samanaikaisesti (Heimberg ym. 2016; Amezquita ym. 2020, luku 9). Dimensioreduktiota hyödynnetään toiseenkin päämäärään, datan visualisointiin, sillä ihmisen on helpompi hahmottaa aineistoa kaksi- tai kolmiulotteisesti kuvattuna.

Klusteroinnilla tarkoitetaan samankaltaisten näytteiden ryhmittelemistä yhteen. Yksittäisten solujen RNA-sekvensointidatan kohdalla tämä merkitsee RNA-ekspressioprofiililtaan samanlaisten solujen tunnistamista, mikä tähtää eri solutyyppien havaitsemisen. Erilaisia klusterointimenetelmiä on useita – muun muassa k-means ja hierarkkinen klusterointi – mutta scRNA-sekvensointidatalle tyypillisesti käytetään graafipohjaista klusterointia, joka soveltuu hyvin isoille dataseteille eikä tee oletuksia klustereiden muodosta tai niiden sisältämien solujen lukumäärästä (Andrews ja Hemberg

2018; Kiselev ym. 2019). Graafi- eli verkkopohjaisessa klusteroinnissa muodostetaan ensin verkkorakenne, joka koostuu solmuista (soluista) ja niiden välisistä viivoista. Viivoille voidaan edelleen antaa painoarvolukuja, jotka kuvaavat niiden yhdistämien solujen samankaltaisuutta. Graafin muodostuksen jälkeen siitä pyritään erottelemaan ryppäitä eli klustereita, jotka sisältävät keskenään samankaltaisia soluja. Klusteroinnin onnistuminen on keskeistä, jotta voidaan löytää datan sisältämät todelliset solutyypit. Niiden tunnistamiseen perustuu suuri osa jatkoanalyyseista, joissa pyritään selvittämään solutyyppien normaalia ja häiriintynyttä toimintaa ja vuorovaikutusta (Kiselev ym. 2019). Haasteena klusteroinnissa on varmistaa, että solut eivät ryhmity teknisten tai biologisten häiriösignaalien, kuten erävaikutuksen (batch effect) tai solusyklin, mukaan (Kiselev ym. 2019). Lisäksi sopivan klusteroinnin tunnistamista vaikeuttaa se, että yleensä ei ole tarkkaa etukäteistietoa kudoksen sisältämistä solutyypeistä ja näiden alatyypeistä ja lukumääristä (Zappia ja Oshlack 2018).

Tutkimuksen tavoitteena oli klusteroida korkeaulotteinen eturauhasen scRNA-sekvensoinnista saatu datasetti niin, että muodostuneet klusterit vastaisivat kudoksen solubiologiaa. Päämäärän saavuttamiseksi datan ulotteisuutta pienennettiin pyrkien samalla kuitenkin säilyttämään keskeinen informaatio. Tämän jälkeen muokattu data ryhmiteltiin klustereiksi, jotka vastaisivat solutyyppejä eivätkä olisi teknisten tai biologisten häiriösignaalien aiheuttamia.

2 MATERIAALIT JA MENETELMÄT

2.1 Aineisto ja ohjelmointipaketit

Tässä tutkimuksessa hyödynnettiin Karthausin ym. 2020 Science-lehdessä julkaistun artikkelin aineistoa. Se koostuu 8 eri ihmiseltä otettujen eturauhaskudosnäytteiden scRNA-sekvensointituloksista kattaen yhteensä 120 300 solua. Yksittäisten solujen RNA-sekvensointiin käytettiin 10X Chromium -pohjaista tekniikkaa, jolla saadaan sekvensoitua pieni osuus lähetti-RNA:n 3'-pään emäsjärjestyksestä. Karthaus ym. muodostivat sekvensointidatan pohjalta geenieksпressiomatriisin, josta havaitaan, kuinka monta RNA-molekyyliä kustakin geenistä ilmennetään kussakin solussa. Edelleen he esikäsittelivät dataa suodattamalla pois siitä elinkelvottomat solut ja solut, jotka todennäköisesti olivat peräisin sekvensointiteknologian aiheuttamista virheistä. Lopuksi ekspressiomatriisin arvot oli normalisoitu transkriptia per 10 000 (TP10K) -muotoon ja $\log_2(x+1)$-muunnettu.

Edellä mainittua dataa analysoitiin tässä tutkimuksessa käyttäen R-ohjelmointikielen Seurat-pakettia ja tämän versiota 3.1 (Butler ym. 2018; Stuart ym. 2019). Klusteroinnin resoluutioparametrin

valitsemisessa hyödynnettiin R-ohjelmointikielen clustree-pakettia versiolla 0.4.3 (Zappia ja Oshlack 2018).

2.2 Datan esikäsittely

Datan esikäsittely on tärkeätä, jotta todellinen biologisesti merkittävä informaatio saataisiin aineistosta esiin. Solujen heterogeenisyyden kannalta merkityksettömän tiedon poistamiseksi ja teknisten häiriösignaaleiden vähentämiseksi ekspressiomatriisista karsittiin pois geenit, joita ilmennettiin alle 3 solussa, minkä jälkeen datamatriisi sisälsi 27 220 geeniä. Tämän jälkeen valittiin 2000 paljon varioivaa geeniä, joita käytettiin dimensioreduktiossa ja klusteroinnissa. Tällä piirrevalinnalla (feature selection) pyritään vahvistamaan aineiston merkityksellistä biologista signaalia ja pienentämään moniulotteisen datan ulottuvuuksien määrää (Luecken ja Theis 2019; Amezquita ym. 2020, luku 8). Lopuksi kullekin geenille tehtiin ekspressioarvojen skaalaus niin, että keskiarvoksi saatiin 0 ja varianssiksi 1, mikä antaa jokaiselle geenille yhtäläisen painoarvon jatkoanalyyseissä (Luecken ja Theis 2019).

2.3 Lineaarinen dimensioreduktio pääkomponenttianalyysilla

Moniulotteisen datan ulotteisuuden pienentämiseen käytettiin pääkomponenttianalyysia (principal component analysis, PCA). Pääkomponenttianalyysi pyrkii säilyttämään mahdollisimman paljon merkityksellistä informaatiota etsimällä datasta pääkomponentit, jotka säilyttävät järjestyksessään maksimaalisen määrän aineiston variaatiosta (Jolliffe ja Cadima 2016; Amezquita ym. 2020, luku 9). Matemaattisesti tämä toteutetaan selvittämällä geenieksressiomatriisista vektorit, joista ensimmäinen säilyttää eniten variaatiota, minkä jälkeen selvitetään tälle ortogonaalinen vektori, joka tallettaa mahdollisimman paljon jäljellä olevaa variaatiota jne. (Amezquita ym. 2020, luku 9).

Pääkomponenttien soveltamisessa scRNA-sekvensointidataan on taustaoletuksena, että biologiset prosessit säätelevät samanaikaisesti useiden geenien toimintaa (Heimberg ym. 2016; Amezquita ym. 2020, luku 9). Tämän takia ensimmäiset pääkomponentit tallentavat biologista informaatiota todennäköisesti paremmin kuin myöhemmät pääkomponentit, sillä variaatiota saadaan talletettua enemmän tarkastelemalla useiden geenien toisistaan riippuvaa käytöstä (Amezquita ym. 2020, luku 9). Se, kuinka monta ensimmäistä pääkomponenttia kannattaa valita jatkoon, ei ole yksiselitteistä. Tässä tutkimuksessa hyödynnettiin niin sanottua kyynärpääkuvaajaa (elbow plot), joka havainnollistaa kunkin pääkomponentin selittämän varianssin määrää datasta. Kuvaajasta voidaan arvioida taitekohta, ”kyynärpää”, minkä jälkeen pääkomponentit eivät enää sisällä merkittävästi tilastollista informaatiota (Amezquita ym. 2020, luku 9).

2.4 Epälineaarinen dimensioreduktio UMAP:illa datan visualisoimiseksi

Datan visualisoimiseksi käytettiin UMAP-tekniikkaa (Uniform Manifold Approximation and Projection) (McInnes ym. 2018). UMAP pyrkii muodostamaan moniulotteisesta datasta pienempiulotteisen esityksen, jossa moniulotteisessa avaruudessa lähellä toisiaan olevat pisteet olisivat lähellä myös kaksiulotteisessa esityksessä (Amezquita ym. 2020, luku 9). Teoreettiselta taustaltaan UMAP pohjautuu matemaattisten monistojen tarkastelemiseen ja se käyttää epälineaarista datan transformaatiota toisin kuin lineaarinen PCA (McInnes ym. 2018; Amezquita ym. 2020, luku 9). Verrattuna t-SNE:hen, joka on toinen suosittu scRNA-sekvensointidatan visualisointiin käytetty dimensioreduktiotekniikka, UMAP säilyttää datan globaalin rakenteen eli klustereiden eroavaisuudet paremmin kuin alustamatonta upotusta käyttävä t-SNE. Lisäksi UMAP on huomattavasti nopeampi, minkä takia se on nousemassa käytetyimmäksi menetelmäksi (Becht ym. 2019; Amezquita ym. 2020, luku 9; Kobak ja Linderman 2021). UMAP-algoritmi suoritettiin datalle Seuratin RunUMAP-funktiolla käyttäen oletusparametreja ja 30 ensimmäistä pääkomponenttia.

2.5 Klusterointi

Solujen graafipohjainen klusterointi toteutettiin Seuratin FindNeighbors- ja FindClusters-funktioilla. Ensin FindNeighbors-funktiota käytettiin jaetun lähimmän naapurin (shared nearest neighbor, SNN) graafin muodostamiseen, minkä jälkeen FindClusters-funktiolla eroteltiin graafin solut omiksi ryhmikseen eli klustereikseen.

Graafi muodostetaan aluksi k-lähimmän naapurin (k-nearest neighbor, KNN) menetelmällä. Siinä jokaiselle solmulle eli tässä tapauksessa solulle etsitään aiemmin lasketussa PCA-avaruudessa k-lähintä solua Euklidisen etäisyyden perusteella. Tämän jälkeen KNN-graafista muokataan SNN-graafi hiomalla solujen välisten viivojen painoarvoa Jaccard indeksien avulla. (https://satijalab.org/seurat/articles/pbmc3k_tutorial.html#cluster-the-cells-1, 22.4.2021) Siinä idea on antaa paljon samoja naapureita jakavien solujen viivoille isompi painoarvo kuin vähemmän yhteisiä naapureita jakaville soluille (Levine ym. 2015). FindNeighbors-funktio ajettiin 30 ensimmäisellä pääkomponentilla ja oletusparametrilla k.param = 20.

SNN-graafin jakaminen klustereihin Seuratilla pohjautuu niin sanottuun yhteisöjen havaitsemiseen (community detection). Yhteisöt ovat graafin tihentymiä, joissa solmut ovat tiheästi toisiinsa linkitettyjä verrattuna muiden yhteisöjen solmuihin (Blondel ym. 2008; Luecken ja Theis 2019). Käytetyin algoritmi yksittäisten solujen RNA-sekvensointidatan yhteisöjen havaitsemiseen on Louvain, joka optimoi yhteisön tiiviyttä mallintavaa modulaarisuusarvoa (Luecken ja Theis 2019). FindClusters-funktio suoritettiin muuten oletusparametreilla, mutta resoluutioarvoksi päätettiin clustree-paketin perusteella 0.3.

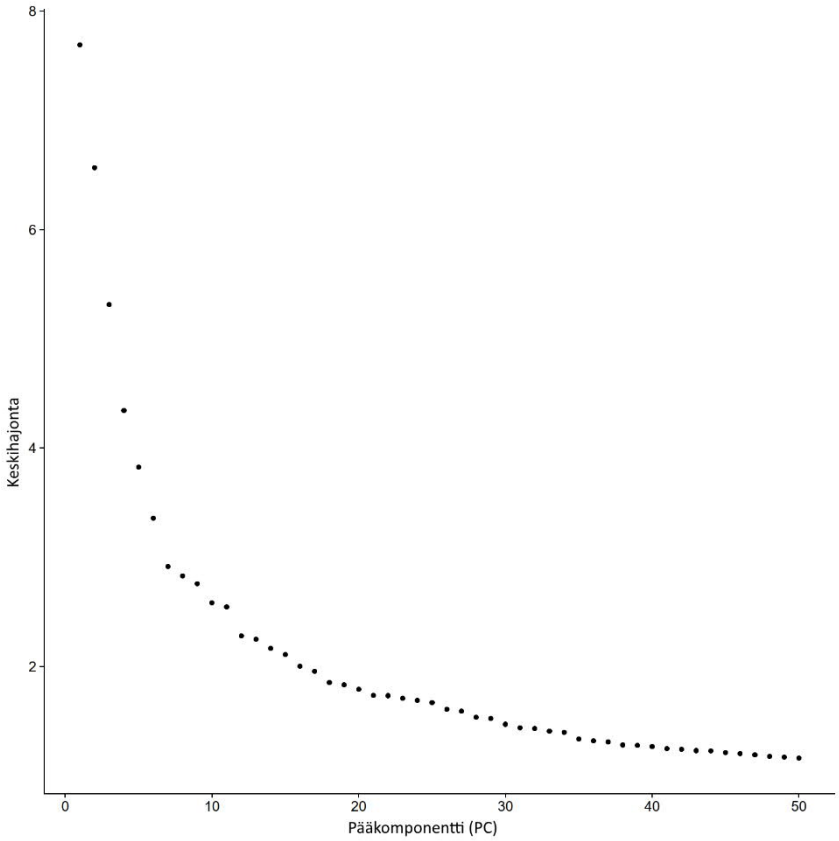
2.6 Klusterointiresoluution valinta clustree:lla

Klusterointia varten päätetään joku arvo niin sanotulle resoluutioparametrille. Mitä suurempi lukuarvo annetaan, sitä enemmän klustereita muodostuu. Yksittäisten solujen RNA-sekvensoinnissa haasteena on se, että ei tiedetä etukäteen, montako erilaista solutyyppiä datassa on, joten oikeaa klusterointiresoluutiotakaan ei tiedetä. Voidaan kuitenkin haarukoida sopiva resoluutioväli ja havaita resoluutiot, jotka todennäköisesti eivät ole oikeita. Tämä tehtiin clustree-ohjelmointipaketin avulla.

Clustree vertailee klusterointeja useilla eri resoluutioilla ja muodostaa klusterointipuun. Sen avulla voidaan visuaalisesti arvioida kyseiselle aineistolle järkevät resoluutioparametrin arvot. Puurakenteessa ylimpänä ovat pienimmällä resoluutiolla saadut klusterit ja seuraavalla tasolla seuraavaksi pienimmällä resoluutiolla muodostuneet klusterit jne. Suunnatut viivat (nuolet) vierekkäisten resoluutiotasojen klustereiden välille muodostetaan vertaamalla, kuinka monta yhteistä solua löytyy kunkin kahden klusterin väliltä. Tämä luku jaetaan vielä suuremman resoluutiotason klusterin sisältämien solujen kokonaislukumäärällä, jotta eri kokoisten klusterien väliset viivat olisivat vertailukelpoisia keskenään. Viivan läpinäkyvyys ilmoittaa saadun suhdeluvun arvoa kuvaten pienet suhdeluvut läpinäkyvimmiksi kuin suuret. Näin muodostetusta klusterointipuusta voidaan arvioida resoluutioarvot, joilla klusterointi on vakaata ja todennäköisesti ilmentää datassa esiintyviä ryhmiä. Epävakaaseen klusterointiin viittaavat hyvin läpinäkyvät viivat, uusien klustereiden muodostuminen useista aikaisemmista klustereista ja klusterit, jotka muodostuvat, mutta myöhemmin katoavat resoluution kasvaessa. (Zappia ja Oshlack 2018)

3 TULOKSET JA TULOSTEN TARKASTELU

Pääkomponenttianalyysin jälkeen klusterointia varten määritettiin käytettävien pääkomponenttien lukumäärä, jolla mahdollisimman paljon datan sisältämästä biologisesta informaatiosta saataisiin talteen samalla kun vähennetään sotkevaa häiriösignaalia. Tätä tarkoitusta varten muodostettiin kyynärpääkuvaaja (kuva 1), josta arvioitiin biologisesti merkittävää informaatiota tallettavien pääkomponenttien lukumäärä.

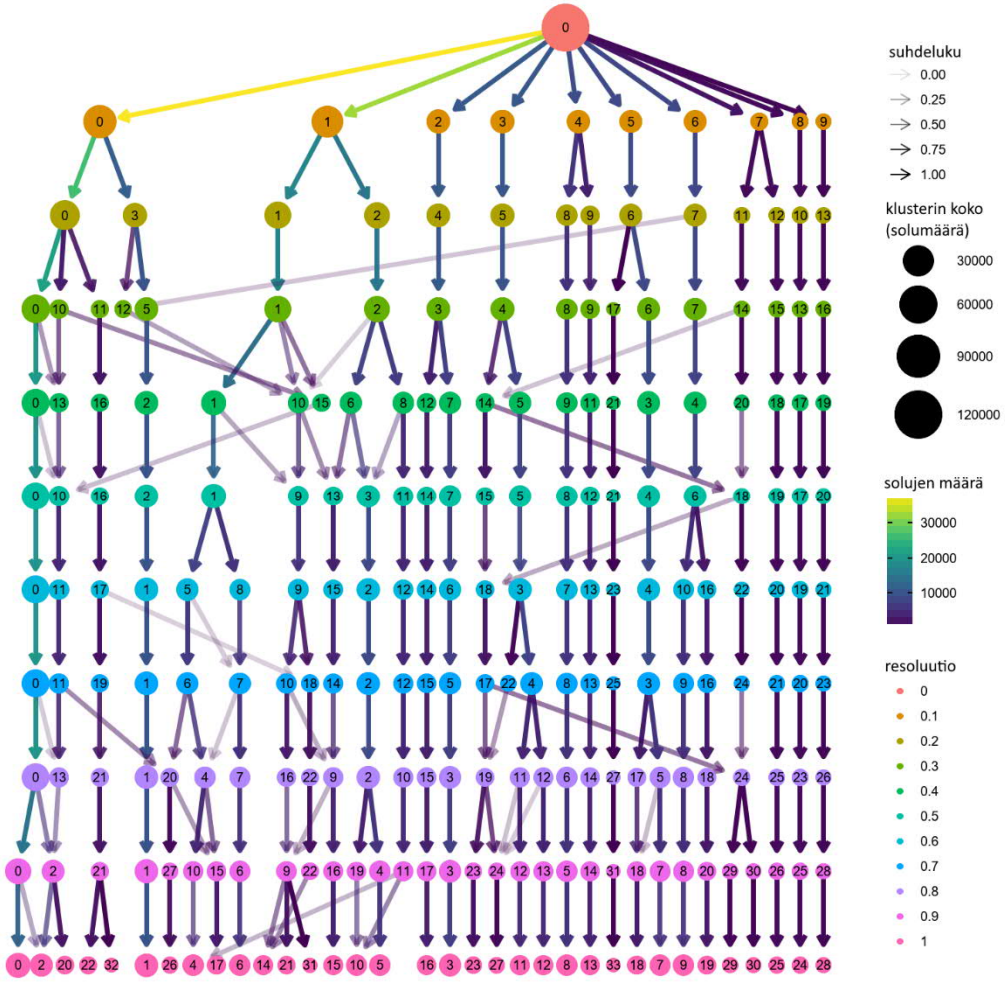


Kuva 1. Pääkomponenttien datasta selittämä variaatio (keskihajonta).

Saadussa kuvaajassa ei ole kovin selkeää taitekohtaa, minkä perusteella valita käytettävien pääkomponenttien lukumäärä. Voidaan havaita, että pääkomponentit 1–6 tallettavat suuren määrän aineiston sisältämästä variaatiosta, minkä jälkeen käyrässä havaitaan pieni loiventuma pääkomponenttien 7–17 alueella. Taitekohdan voisi arvioida alkavan pääkomponentin 18 kohdalta. Menetelmän heuristisen luonteen ja selkeän taitekohdan puuttumisen takia valittiin 30 ensimmäistä pääkomponenttia, millä pyrittiin takaamaan biologisesti merkittävän signaalin saaminen jatkoanalyyseihin. Korkeamman määrän valintaa puoltaa se, että liian pieni pääkomponenttien lukumäärä sulkee pois tärkeää informaatiota ja lisäksi tulokset eivät juurikaan eronneet esimerkiksi 20 pääkomponentilla tehtyihin klusterointeihin. Vastaava suositus ja havainto esitetään myös Seurat-paketin opetusohjelmassa (https://satijalab.org/seurat/articles/pbmc3k_tutorial.html#determine-the-dimensionality-of-the-dataset-1, 22.4.2021).

Klusterointiresoluution valitsemiseksi muodostettiin klusterointipuu, joka havainnollistaa klustereiden vakautta ja solujen siirtymistä eri resoluutiotasojen eli resoluutioiden välillä (kuva 2). Puurakenteen oikeassa laidassa on nähtävissä kolme pientä, mutta hyvin vakaata klusteria, jotka pysyvät muuttumattomina resoluutiosta 0.2 eteenpäin. Toinen, lähes yhtä vakaat klusterit omaava kohta, on resoluution 0.1 klusterista 2 alkava reitti, joka jakaantuu resoluution 0.4 kohdalla kahdeksi klusteriksi. Muut klusterit ilmentävät vaihtelevaa vakautta resoluutiosta riippuen. Kuitenkin klusterointipuussa on havaittavissa kaksi suhteellisen vakaata aluetta, joissa useat eri klusterit vaikuttavat

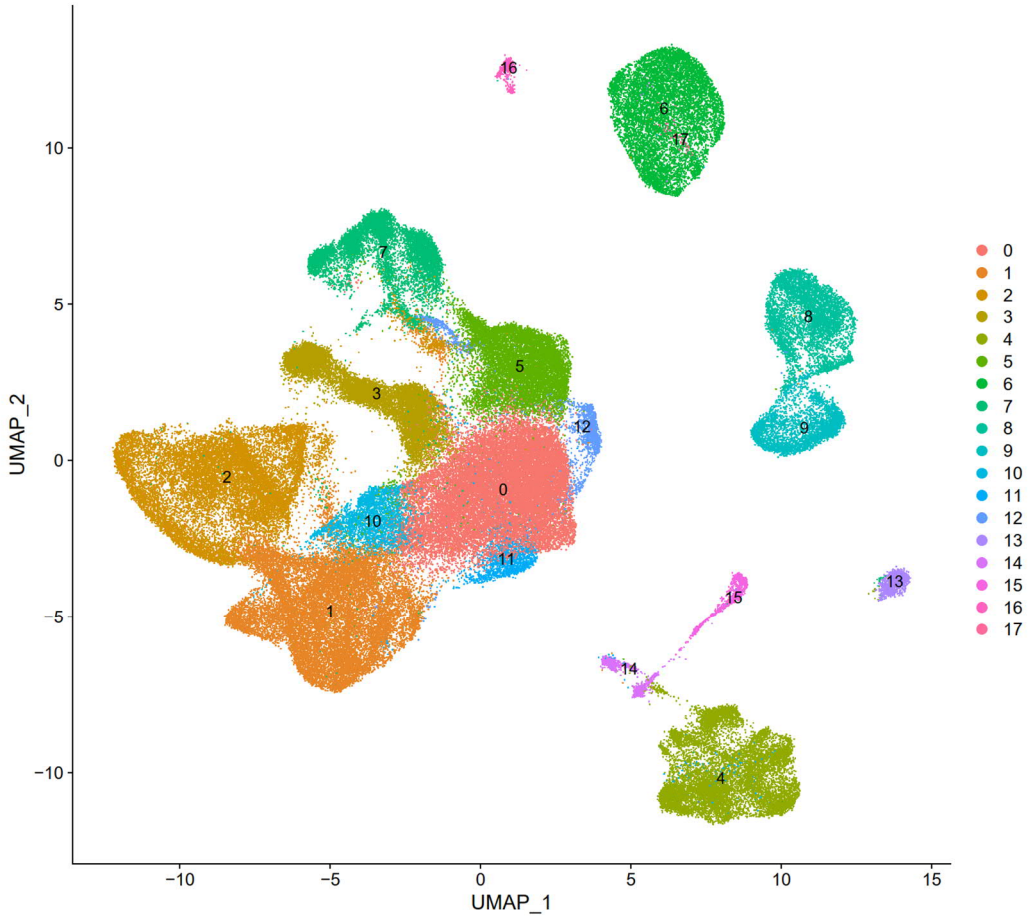
olevan stabiileja: resoluutiovälit 0.1–0.3 ja 0.5–0.6. Klusterointiresoluutioksi valittiin 0.3, sillä välittömästi sen jälkeen ilmenee paljon alhaisen suhdeluvun viivoja, mikä merkitsee näiden viivojen alkuklustereista tulevan loppuklustereihin vain pienen määrän soluja suhteutettuna loppuklusterien kokonaissolumäärään. Lisäksi resoluutiossa 0.4 havaitaan klusteri 15, joka ilmenee vain tässä resoluutiossa. Vastaavasti epävakaa klusteri 12 havaitaan tosin myös resoluutiossa 0.3, joka muuten on paljon stabiilimpi klustereiltaan.



Kuva 2. Klusterointipuu klusteroinnin vakauden arvioimiseksi eri klusterointiresoluutioilla. Klusterointipuusta nähdään klusterien määrä eri resoluutioilla ja erityisesti solujen siirtyminen klusterista toiseen resoluution muuttuessa.

Klusterointiresoluution valintaan vaikutti myös oletettu solutyyppien lukumäärä eturauhasessa. Eturauhasen tiedetään sisältävän immuunisoluja, kudoksen toimintaa tukevia stroomasoluja sekä kolmea erilaista pääepiteelisolutyyppiä (Henry ym. 2018). Erilaisia alatyyppejä löytyy vielä jokaisesta joukosta, minkä takia erilaisia solupopulaatioita on suuremmissa eturauhasen scRNA-sekvensointidatoissa usein havaittu 16 – 20 (Karthaus ym. 2020; Dong ym. 2020; Chen ym. 2021). Tämän perusteella järkevät resoluutioarvot voidaan rajata välille 0.3–0.6, sillä niissä klustereiden määrä vaihtelee välillä 18–24 eikä siten poikkea paljon aiempien tutkimusten tuloksista.

Klusterointitulosten kaksiulotteinen esitys UMAP-dimensioreduktiotekniikalla on nähtävissä kuvassa 3. Siinä vasemmalla erottuu selkeästi monen klusterin muodostama isompi rykelmä. Koska UMAP pyrkii säilyttämään myös datan globaalin rakenteen, voidaan tämän isomman rykelmän klusterien sanoa olevan geeniekspressioltaan lähempänä toisiaan verrattuna kauempana oleviin erillisiin klustereihin. Solutyypeille ominaisten geenien ilmentämisen perusteella näiden lähellä toisiaan olevien klusterien voidaan todeta olevan eturauhasen epiteelisoluja (liite A).



Kuva 3. Klusterien visualisointi UMAP-dimensioreduktiotekniikalla.

Kuvasta havaitaan myös Louvainin algoritmin erottelemien klusterien ja UMAP-dimensioreduktiotekniikalla saadun klustereiden kaksiulotteisen visualisoinnin ero. Kaksiulotteisessa UMAP-visualisoinnissa Louvainin algoritmilla havaitun tietyn klusterin kaikki solut eivät välttämättä ole täysin yhtenäisesti toistensa lähellä. Tästä esimerkkinä voidaan havaita pienen joukon klusterin 12 soluista olevan erillään klustereiden 7 ja 5 välissä. Tätä epäyhtenäisyyttä selittänee se, että kaksiulotteinen visualisointi pakottaa solut yhteen tasoon, kun taas itse klusterointi on tehty 30-ulotteisessa PCA-avaruudessa. Yksi mahdollinen syy on myös itse Louvainin algoritmi. Suuresta suosiostaan huolimatta sen on toisinaan todettu tuottavan yhteisöjä, jotka ovat sisäiseltä rakenteeltaan heikosti kytköksissä (Traag ym. 2019). Toinen kaksiulotteisessa visualisoinnissa havaittava epäyhtenäisyys on pienen määrän tietystä klusterista olevien solujen ryhmittyminen toiseen klusteriin.

Klusterin 13 vasemmalla reunalla voidaan esimerkiksi nähdä täplä vihertäviä soluja ja klusterin 4 sisällä sinertäviä soluja. Edellä mainitun moniulotteisen datan kahteen ulottuvuuteen puristamisen lisäksi selittävänä tekijänä voi myös olla jokin tekninen häiriösignaali, esimerkiksi erävaikutus, minkä seurauksena saman potilaan solut ryhmittyisivät yhteen vain kahta ulottuvuutta käytettäessä. Eräs mahdollinen kiinnostava selittäjä on eturauhassyöpäsolujen kyky muokata kasvainympäristöään (tumour microenvironment, TME) (Dong ym. 2020). Dong ym. havaitsivat tutkimuksessaan muun muassa T-soluille poikkeavaa KLK3-geenin – tunnetaan myös prostataspresifisenä antigeeninä, PSA) – ekspressiota, minkä katsottiin johtuvan kasvainsolujen ekstrasellulaaristen vesikkelien välittämästä viestinnästä. Tutkimuksessa huomattiin muitakin eturauhaselle tyypillisten geenien aktivaatiota immuunisoluissa. Tällainen vuorovaikutus, joka muokkaa toisten solujen transkriptio-ohjelmia, voi johtaa eri solutyyppien lähentymiseen transkriptomiltaan. Näin ollen klusteroinnin UMAP-visualisoinnissa voitaneenkin havaita Louvainin algoritmin useampiulotteisessa avaruudessa erottelemia eri solutyyppejä kaksiulotteisen koordinaatiston samassa paikassa. Kokonaisuudessaan klusteroinnin voi kuitenkin todeta olevan onnistunut, sillä se pääasiallisesti erottaa eturauhasen eri solutyypit omiksi ryhmikseen (liite A).

4 YHTEENVETO

Korkeaulotteisen yksittäisten solujen RNA-sekvensointidatan klusterointi vaatii onnistuakseen ulotteisuuden pienentämistä sekä biologisesti merkittävän signaalin vahvistamista ja teknisten häiriösignaalien karsimista. Datan dimensioreduktio tehdään tyypillisesti pääkomponenttianalyysilla, mutta se, montako pääkomponenttia eli ulottuvuutta kannattaa valita jatkoanalyyseihin ei ole yksiselitteistä. Kysymys on tasapainoilusta sen suhteen, paljonko biologista informaatiota halutaan säilyttää, koska sen lisääntyessä myös teknisen häiriösignaalin määrä lisääntyy. Suurille dataseteille nopea vaihtoehto on tarkastella pääkomponenttien aineistosta selittämän variaation – tilastollisen informaation – määrää ja jättää pois ne komponentit, joiden selityskyky on hyvin pieni. Vastaava osin subjektiivinen ja vailla yhtä tarkkaa oikeaa vastausta oleva vaihe on sopivan klusterointiresoluution valinta. Tämän parametrin sopivan arvon päättäminen riippuu tarkkuudesta, jolla kudoksen solutyyppejä halutaan tarkastella. Pienemmät resoluutioarvot riittävät erottelemaan isomman tason solutyypit, kun taas suurempia resoluutioita tarvitaan harvinaisten solupopulaatioiden tai saman solutyypin toiminnaltaan hieman poikkeavien ryhmien erottamiseen. Sopivan resoluution haarukoinnissa hyväksi avuksi havaittiin clustree-paketti, jolla voidaan visualisoida klustereiden vakautta eri resoluutioarvoilla.

Itse klusterointi tapahtuu automaattisesti edellä mainittujen parametrien valitsemisen jälkeen. Näin saatujen klusterien ei voi kuitenkaan olettaa edustavan totuutta kudoksen solubiologiasta, minkä takia klusteroinnin laadun tarkastaminen on tärkeää. Graafinjakoalgoritmin muodostamia ryhmiä tulisikin tutkia sen suhteen, ovatko ne muodostuneet erävaikutuksen, jonkin muun teknisen häiriösignaalin tai epätoivotun biologisen variaation seurauksena. Tätä varten on hyödyllistä käyttää epälineaarista dimensioreduktiotekniikkaa, jonka avulla klustereita voidaan piirtää kaksiulotteiselle kuvaajalle, jolla voidaan visualisoida tietyn piirteen ilmenemistä eri ryhmissä.

Tulevaisuuden haasteena ja tavoitteena yksittäisten solujen RNA-sekvensointidatan klusteroinnissa näkisin selkeiden yhtenäisten toimintatapojen kehittämisen. Johtuen datan monimuotoisuudesta suositellut klusterointikäytännöt voisi määritellä erikseen esimerkiksi pienille vs. isoille solumäärille sekä kokopitkiä transkripteja vs. vain osan transkriptista lukeville sekvensointiprotokollille. Yhtenäiset ja selkeät ohjeet helpottaisivat tutkimusalan ymmärrettävyyttä, nopeuttaisivat data-analyysiä ja tekisivät tutkimustulokset vertailukelpoisemmiksi keskenään.

5 LÄHTEET

Amezquita R, Lun A, Hicks S, ym. Orchestrating Single-Cell Analysis with Bioconductor 2020; http://bioconductor.org/books/release/OSCA/index.html (e-kirja, versio 1.0.6).

Andrews TS, Hemberg M. Identifying cell populations with scRNASeq. Mol Aspects Med 2018;59:114-22.

Becht E, McInnes L, Healy J ym. Dimensionality reduction for visualizing single-cell data using UMAP. Nat Biotechnol 2019;37:38-44.

Blondel VD, Guillaume J, Lambiotte R ym. Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment 2008;2008:P10008.

Butler A, Hoffman P, Smibert P ym. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol 2018;36:411-20.

Chen S, Zhu G, Yang Y ym. Single-cell analysis reveals transcriptomic remodellings in distinct cell types that contribute to human prostate cancer progression. Nat Cell Biol 2021;23:87-98.

Dong B, Miao J, Wang Y ym. Single-cell analysis supports a luminal-neuroendocrine transdifferentiation in human prostate cancer. Communications Biology 2020;3:778.

Heimberg G, Bhatnagar R, El-Samad H ym. Low Dimensionality in Gene Expression Data Enables the Accurate Extraction of Transcriptional Programs from Shallow Sequencing. Cell systems 2016;2:239-50.

Henry GH, Malewska A, Joseph DB ym. A Cellular Anatomy of the Normal Adult Human Prostate and Prostatic Urethra. Cell reports 2018;25:3530,3542.e5.

Hwang B, Lee JH, Bang D. Single-cell RNA sequencing technologies and bioinformatics pipelines. Exp Mol Med 2018;50:96.

Jolliffe IT, Cadima J. Principal component analysis: a review and recent developments. Philosophical transactions.Series A, Mathematical, physical, and engineering sciences 2016;374:20150202.

Karthaus WR, Hofree M, Choi D ym. Regenerative potential of prostate luminal cells revealed by single-cell analysis. Science 2020;368:497.

Kiselev VY, Andrews TS, Hemberg M. Challenges in unsupervised clustering of single-cell RNA-seq data. Nature Reviews Genetics 2019;20:273-82.

Kobak D, Linderman GC. Initialization is critical for preserving global data structure in both t-SNE and UMAP. Nat Biotechnol 2021;39:156-7.

Levine JH, Simonds EF, Bendall SC ym. Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. Cell 2015;162:184-97.

Luecken MD, Theis FJ. Current best practices in single-cell RNA-seq analysis: a tutorial. Molecular systems biology 2019;15:e8746.

McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction 2018.

Papalexi E, Satija R. Single-cell RNA sequencing to explore immune cell heterogeneity. Nature Reviews Immunology 2018;18:35-45.

Picelli S. Single-cell RNA-sequencing: The future of genome biology is now. RNA biology 2017;14:637-50.

Regev A, Teichmann SA, Lander ES ym. The Human Cell Atlas. eLife 2017;6:e27041.

Stuart T, Butler A, Hoffman P ym. Comprehensive Integration of Single-Cell Data. Cell 2019;177:1888,1902.e21.

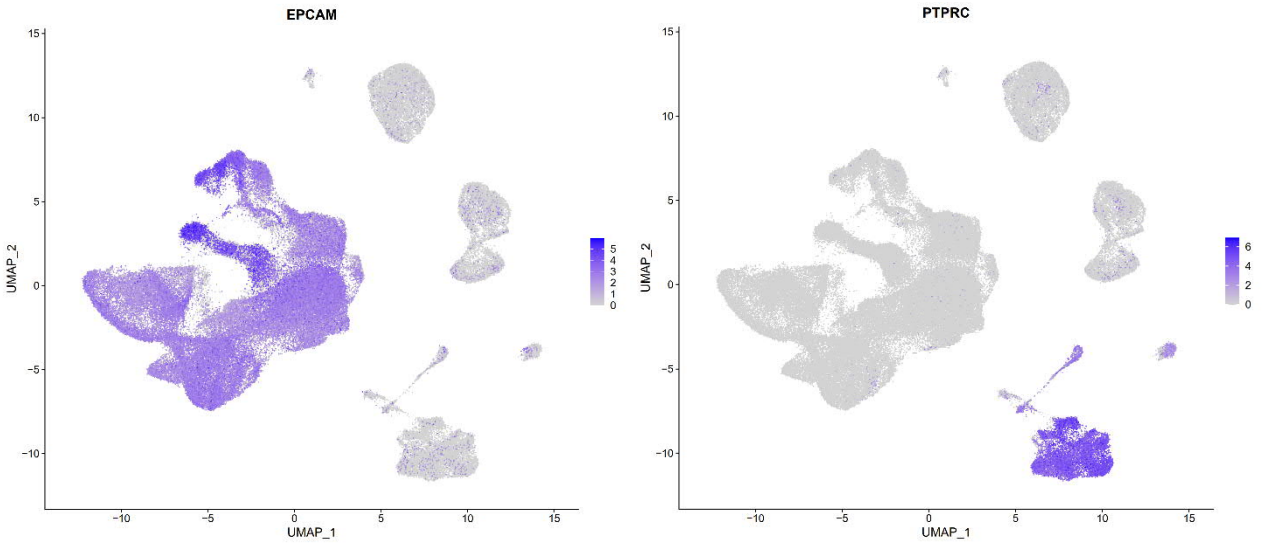
Traag VA, Waltman L, van Eck NJ. From Louvain to Leiden: guaranteeing well-connected communities. Scientific Reports 2019;9:5233.

Zappia L, Oshlack A. Clustering trees: a visualization for evaluating clusterings at multiple resolutions. Gigascience 2018;7:giy083.

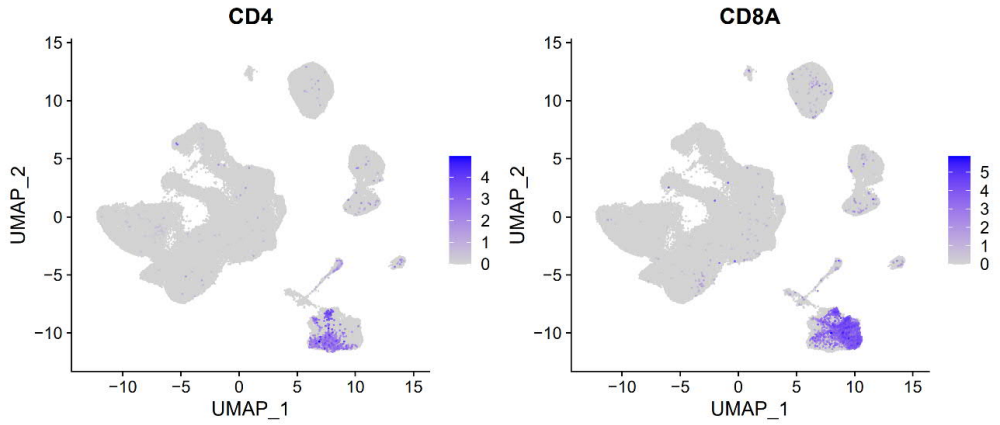
LIITE A: KLUSTEREIDEN ANNOTOINTI

Klustereiden annotoinnilla tarkoitetaan muodostuneiden soluryhmien identiteetin määrittämistä. Tyypillisesti halutaan tietää, mitä solutyyppiä klusterin solut edustavat. Tällä hetkellä tämä vaihe on scRNA-sekvensointianalyysin työläimpiä vaiheita. Tämä johtuu pitkälti siitä, että aiemman tiedon pohjalta kerättyjä ja muodostettuja soluontologioita ja soluatlaksia ollaan vasta kehittämässä (Regev ym. 2017; Kiselev ym. 2019; Luecken ja Theis 2019). Lisähaastetta annotointiin ja vertailutietokantojen luomiseen tuo solutyypin käsitteen epämääräisyys (Regev ym. 2017; Luecken ja Theis 2019). T-solu voi olla riittävä määritelmä, jos kiinnostuksen kohteena tarkasteltavassa kudoksessa ovat muut kuin immuunisolut, mutta juuri tätä joukkoa tutkivalle jaottelu CD4-positiivisiin auttaja-T-soluihin ja CD8-positiivisiin tappaja-T-soluihin ei välttämättä riitä.

Tämänhetkiset klustereiden annotointimenetelmät voidaan karkeasti jaotella kirjallisuudesta koottujen kullekin solutyypille ominaisten geenien ekspression visualisointiin sekä klustereiden geeniekspression vertaamiseen valmiiksi annotoituihin vertailudatasetteihin. Ensimmäinen keino perustuu tiettyä kudosta tutkivan tiedeyhteisön käsitykseen tärkeimmistä solutyypeistä ja niiden ilmentämistä geeneistä, niin sanotuista kanonisista markkereista, joita sitten visualisoidaan datan klustereissa (Kiselev ym. 2019). Näin voidaan tehdä isomman tason erottelua sen selvittämiseksi, mitkä klusterit vastaavat kudoksen epiteelisoluja, immuunisoluja tai strooman soluja (kuva 4) sekä hyvinkin yksityiskohtaista solutyyppimääritystä (kuva 5). Lähestymistavan onnistuminen riippuu pitkälti siitä, yksilöivätkö käytetyt markkerigeenit tarkasti halutun solutyypin vai voivatko muutkin solut ilmentää niitä jossain määrin.



Kuva 4. Korkean tason solutyyppien annotointi kanonisilla markkereilla. EPCAM (epithelial cell adhesion molecule) paljastaa epiteelisolujen sijainnin ja PTPRC eli CD45 on immuunisoluille yksilöllinen markkeri.



Kuva 5. T-solujen tunnistaminen koreseptoreita CD4 ja CD8 koodaavien geenien ilmentämisen avulla.

Jälkimmäinen vertailutietokantojen käyttöön perustuva tapa voidaan toteuttaa kahdella eri tavalla. Ensinnäkin solujen ekspressioprofiilia voidaan suoraan verrata valmiiksi annotoitujen solujen ekspressioprofiileihin. Tässä niin kutsutussa projisoinnissa kullekin tarkasteltavan aineiston solulle etsitään vertailudatasta ekspressioltaan mahdollisimman vastaava solu, jonka annotointi sitten siirretään aiemmin tuntemattomalle solulle. Tämä ei vaadi manuaalista solutyyppimarkkereiden vertailua, minkä takia se on huomattavan nopea keino. Toisena vaihtoehtona on differentiaalisella geenieksprессioanalyysilla selvittää klustereista geenit, joita niissä ilmennetään huomattavasti muita klustereita enemmän tai vähemmän, ja vertailla näin saatua geenijoukkoa tietokannan solutyyppien vastaavasti selvitettyihin geenijoukkoihin. Tämä vaatii lisäanalyysivaihetta, mutta on kuitenkin automatisoitu toimintatapa, joka helpottaa klustereiden solujen annotointia. (Luecken ja Theis 2019)