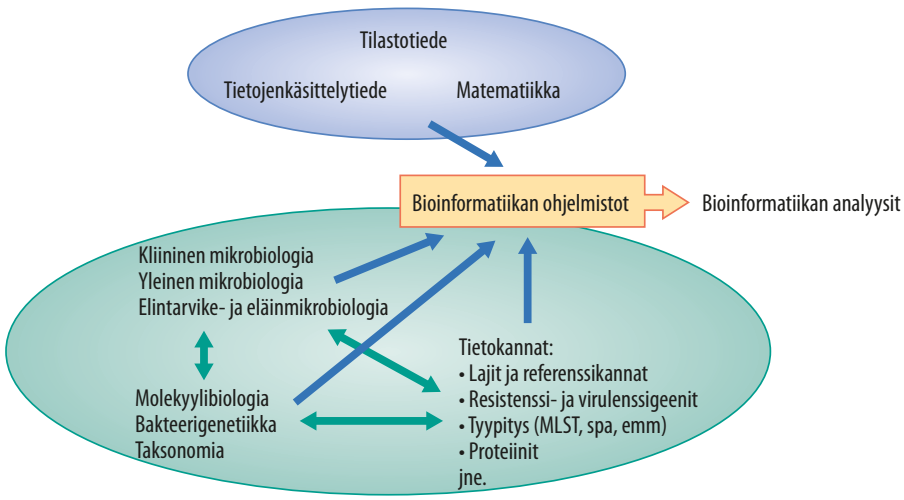
VERTAISARVIOITU
KOLLEGIALT GRANSKAD
PEER-REVIEWED
www.tsv.fi/tunnus

Teemu Kallonen ja Antti Hakanen

Bioinformatiikka bakteriologiassa – lupauksia ja haasteita

Bakteriologiassa bioinformatiikalla pyritään vastamaan usein hyvin tavanomaisiin kysymyksiin, kuten mikä bakteerilaji on kyseessä, onko tutkittu kanta sukua tunnetuille taudinaiheuttajaklooneille tai onko kannalla resistenssi- ja virulenssitekijöitä. Lisäksi se mahdollistaa vastaamisen monimutkaisempiin kysymyksiin, kuten mikä bakteerin geneettinen muutos on assosioitunut potilaan tautitilaan. Bioinformatiikkaa tarvitaan käytettäessä moderneja sekvensointitekniikoita, jotka tuottavat merkittävästi enemmän dataa kuin yleisesti käytössä olleet tekniikat ja siksi vaativat analyysien tekemiseen yleensä tehokkaita tietokoneita ja kehittyneitä analyysiohjelmia. Näin myös saadaan tarkkoja tuloksia tutkittavista kannoista. Keskitymme tässä katsauksessa bakteerien bioinformaattisiin analyyseihin, mutta osittain samoja työkaluja voidaan käyttää myös muiden domeenien kuten eukaryoottien ja arkkien sekä virusten perimän tutkimiseen.

Bioinformatiikka on kasvava tieteenala lääke- ja biotieteissä (1). Yleisesti sillä tarkoitetaan biologisten ja tässä artikkelissa bakteriologisten aineistojen käsittelyä, organisointia ja analysointia tietojenkäsittelytieteen, matematiikan ja tilastotieteen menetelmin (**KUVA 1**). Termi ei ole tarkkaan määritelty, ja sen käyttö on laajentunut vasta vuosituhannen vaihteessa. Rajaamme tässä artikkelissa käsiteltävän kokonaisuuden pääasiassa bakteerien kokogenomisekvenssien (WGS) analysointiin bioinformaattisin menetelmin. Mikrobistojen analysointia käsittelemme vain lyhyesti. Analyysimenetelmiä on paljon, ne ovat usein samankaltaisia mutta eivät välttämättä silti sovi kaikille mikrobeille ja kaikkiin tilanteisiin. On hyvä tietää käyttämiensä menetelmien erityispiirteet, jotta niitä voi soveltaa oikeisiin kohteisiin. Bakteriologisen bioinformatiikan tärkeimpiä kliinisiä sovellusalueita tällä hetkellä on esitetty **TAULUKOSSA 1**.



KUVA 1. Bakteriologisen bioinformatiikan analyyseihin vaikuttavia tekijöitä

Duodecim 2024;140:165–72

TAULUKKO 1. Bioinformatiikan tärkeimmät kliiniset sovellusalueet bakteriologiassa

Sovellus	Mahdollinen muutos kliinisessä työssä tai muu hyöty
Epidemiaselvitykset	Epidemioiden rajaaminen, potilaiden hoitaminen
Yksittäisen kannan tyypitys sekä resistenssi- ja virulenssitekijöiden havaitseminen	Oikean mikrobilääkkeen valinta, vaikeiden infektioiden hoito, potilaan eristys
Vakavien, vaikeasti tutkittavien infektioiden diagnostiikka (esim. aivoabsessit)	Mikrobilääkehoidon aloittaminen ja kohdentaminen
Tutkimuksellisten bakteerikantakokoelmien analysointi	Tutkimukselliset mahdollisuudet
Mikrobistojen analyysit (lajit ja niiden osuudet, metabolinen potentiaali, jne.)	Ei-tarttuvien tautien diagnostiikka ja uudenlaiset hoidot

Bakteriologisen bioinformatiikan erityispiirteitä

Laboratoriodata-analytiikassa ja varsinkin kliinisessä mikrobiologiassa on aina syytä ottaa huomioon sekä näytemateriaalin laatu että käytettyjen testausmenetelmien ja analyysityökalujen mahdollisuudet, rajoitukset ja virhelähteet. TAULUKOSSA 2 on esitetty mikrobiologiseen bioinformatiikkaan liittyviä virhelähteitä.

Kuten yleensäkin mikrobiologian analyyseissa näytteen laatu on ensiarvoisen tärkeää myös bakteriologisen bioinformatiikan analyysien kannalta. Huonolaatuisten ja vääränlaisten, kuten kontaminoituneiden näytteiden analysointi on hankalaa ja aikaa vievää, ellei jopa mahdotonta. Esimerkiksi vaikka bakteerin puhdasviljelmässä pitäisi olla vain yhtä bakteerikantaa, siinä voi olla laatupoikkeamana useita saman lajin kantoja, joita ei silmämääräisesti pysty erottamaan toisistaan tai useita toisistaan erottuvia kantoja tai lajeja selvän kontaminaation merkkinä. Bakteerien alkuperäiset määräsuhteet mikrobistonäytteissä voivat muuttua kuljetuksessa tai säilytyksessä. Pahimmillaan perimäaines voi hajota ja testitulos siten vääristyä.

Kliinisessä mikrobiologiassa pitää varmistaa tulosten luotettavuus ja toistettavuus myös uusien ohjelmaversioiden kanssa. Myös analyysiparametrien määrittely on oleellista, koska niitä muokkaamalla tulokset voivat muuttua huomattavasti. Esimerkki tällaisesta parametrista on mikrobilääkeresistenssigeenien samankaltaisuus ja se, kuinka suuri osa geenistä pitää löytää. Jos asetukset ovat liian tiukat, tutkittavassa bakteerikannassa olevat resistenssigeenit voivat jäädä virheellisesti löytymättä. Erityisesti kliinisessä laboratoriossa parametrit pitää testata ja niiden toimintaa seurata, laatu varmistaa esimerkiksi standardoiduilla referenssimateriaaleilla, kuten yleensäkin kliinisen laboratorion validaatioprosesseissa. Toimivat parametrit on hyvä kirjata ja lukita tulosten laadun ja toistettavuuden varmistamiseksi.

Bakteerigenomit ovat huomattavasti pienempiä kuin ihmisgenomi. Näitä voidaan sekvensoida samassa ajossa kymmenistä satoihin. Periytyminen on yksinkertaisempaa, rekombinaatio vähäisempää eikä tekijäinvaihduntaa tapahdu samalla mekanismilla kuin ihmisellä. Erityispiirteenä on taudinaiheuttajien luonne: Tutkittavat mikrobit voivat olla alkuperäisessä näytteessä valtagenomina tai hyvin pahasti normaalimikrobiston peitossa. Myös humaanigenomi on näissä analyyseissa ”kontaminatti”. Se voidaan poistaa joko DNA-eristysvaiheessa tai myöhemmin bioinformatiikan keinoin vertaamalla saatuja sekvenssejä ihmisen referenssigenomiin.

Sekvensointiin liittyviä taustatekijöitä

Sekvensointiteknologian ymmärtäminen on oleellista, jotta voidaan valita oikeat bioinformatiikan työkalut. Illumina-yhtiön teknologia on tällä hetkellä yleisimmin käytössä oleva niin sanottu uuden sukupolven sekvensointimenetelmä (NGS, next generation sequencing) maailmalla, mutta myös tavanomaista Sanger-sekvensointia käytetään edelleen. Kaikkien NGS-menetelmien ongelma on ollut sekvensaattorin tuottamien yksittäisten sekvenssilukujen eli luettujen DNA-sekvenssien lyhyt lukupituus. Tämä aiheuttaa ongelmia esimerkiksi genomien kokoamisessa, koska readit eivät yllä

genomeissa olevien toistojaksojen yli, jolloin yhtenäistä genomia ei saada kootuksi. Varsinkin plasmidien kokoaminen lyhyistä readeista on haasteellista.

Uusimmat pitkän lukupituuden (long-read) sekvensointimenetelmät ratkaisevat lyhyistä readeista aiheutuvan ongelman. Tärkeimmät toimijat tällä alueella ovat Oxford Nanopore (ONP) ja Pacific Biosciences (PacBio). Long-read-teknologioilla voidaan tehdä myös mikrobistoanalyyseja koko *16S rRNA* -geeni sekvensoimalla. Lyhyillä readeilla tätä ei voida tehdä, vaan sekvensointi pitää kohdentaa lyhyemmille variaabeleille alueille.

DNA:n eristys ja sekvensointi olisi hyvä pyrkiä standardoimaan. Tästä riippumatta sekvenssidatan analyysi on tärkeää aloittaa tarkoilla sekvenssien laatua mittaavilla analyyseilla. Sekvensaattorilta saatujen readien laatua, kuten Q-arvoja, on syytä tarkastella mutta tehdä myös muita menetelmään soveltuvia analyyseja, kuten kokogenomisekvensoitujen bakteerikantojen tapauksessa tarkastella koonnin (assembly) eli genomin kokoa ja yhtenäisten koottujen DNA-pätkien eli kontigien määrää sekä mahdollisesti myös heterotsygoottisten SNP:iden (single nucleotide polymorphism) määrää. Laaduntarkistukseen voidaan käyttää myös erilaisia ohjelmia kuten fastQC, multiQC tai Quast (2–4).

Bakteerien kokogenomianalyysien aloitus

Bioinformatiikan menetelmiä käyttäviä työkaluja on käytetty DNA-sekvenssien analysoinnissa jo kauan ennen NGS-menetelmien käyttöönottoa. Sekvensointimenetelmien kehittyessä ja sekvenssien määrän lisääntyessä ohjelmat ovat kuitenkin kehittyneet vastaamaan tarpeita, joihin tavanomaiset Sanger-sekvensoinnille suunnitellut ohjelmat eivät voi vastata. **KUVASSA 2** on kuvattu kaksi hieman toisistaan poikkeavaa mutta toisiaan täydentävää analyysilinjaa kokogenomisekvensoiduille kannoille. **TAULUKOSSA 3** on esimerkkejä analyysiohjelmista.

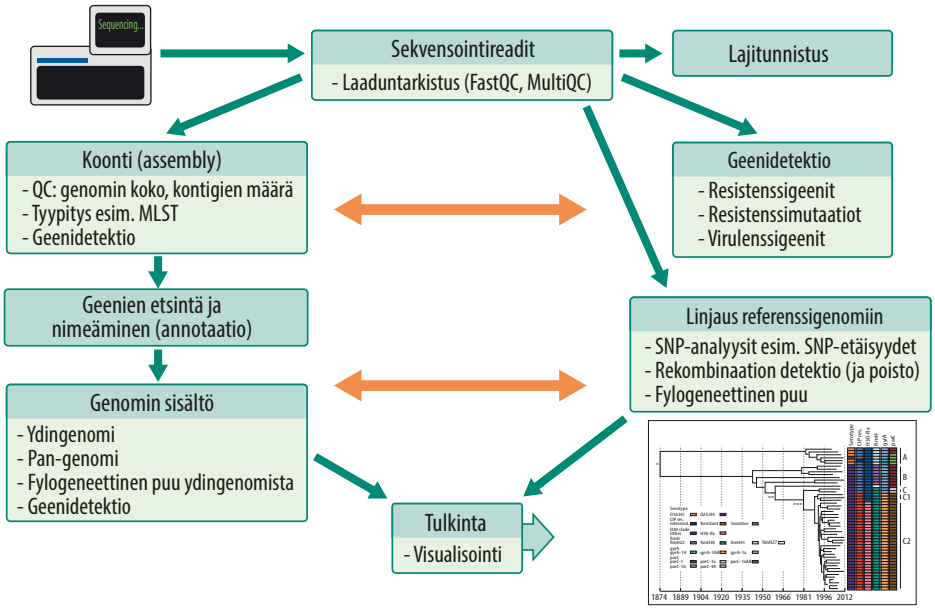
Nykyisillä bioinformatiikan menetelmillä voidaan tehdä monia erilaisia analyyseja yksittäisen kannan tyypittämiseksi. Ensimmäiseksi on kuitenkin hyvä varmistaa, minkä lajin sekvenssiä tutkitaan. Tämä voidaan selvittää esimerkiksi Kraken-ohjelmalla, joka antaa jokaiselle näytteen readille taksonomisen osoitteen (5). Tällä voidaan varmistaa myös, onko näytteessä kontaminaatioita toisesta lajista, joka voisi hankaloittaa tulosten tulkintaa, sekä varmistaa mahdollinen alustava lajintunnistus.

TAULUKKO 2. Mikrobiologiseen bioinformatiikkaan liittyviä virhelähteitä

Mikrobiologinen bioinformatiikka	Virhelähteet
Preanalytiikka	Näytteen otto, kuljetus ja säilytys
	Mahdollisen puhdasviljelmän laatu
Sekvensointi	DNA-eristys ja puhtaus
	Sekvenssin laatu
Bioinformatiikka	Laji- ja referenssikantatietokannat
	Resistenssi- ja virulenssitietokannat
	Kaikki muut tietokannat
	Virheet ohjelmissa
	Väärät asetukset ohjelmissa
	Tulosten väärä tulkinta

Analyysi voi lähteä liikkeelle joko suoraan sekvensointilaitteen readeista tai se voidaan aloittaa tekemällä niin sanottu raakaversio (draft) genomista suorittamalla de novo -koonti (**KUVA 2**). Eri menetelmät aloittavat analyysin erilaisista tiedostoista, joten niiden toiminta poikkeaa jo tästä syytä toisistaan. Yksittäiselle kannalle voidaan tehdä esimerkiksi MLST-tyypitys (multilocus sequence typing) joko tavanomaista seitsemän geenin kaavaa käyttäen tai niin sanottuun ydingenomiin (core-genome) perustuvilla cgMLST-tietokannoilla. Nämä analyysit voivat perustua esimerkiksi sekvenssireadien linjaamiseen (mapping) tai BLAST:iin (basic local alignment search tool). Perinteisen MLST-analyysin voi tehdä esimerkiksi SRST2- tai mlst_check -ohjelmilla (6,7). Yksittäisten kantojen serotyyppi voidaan myös ennustaa sekvenssistä (8).

Joissain tapauksissa uusien kohdegeenien havaitseminen voi olla nopeaa ja helppoa, mutta tämä edellyttää, että vanhoista analyyseista on säilytetty kootut genomit. Kun Kiinasta raportoitiin vuonna 2016 uusi plasmidivälitteinen kolistiiniresistenssiä välittävä geeni (9), jotkut



KUVA 2. Analyysiesimerkkejä. Molemmat tavat alkavat sekvensaattorilta saatavista readeista, jotka siis ovat lyhyitä geenisekvenssejä. Nämä ovat yleensä FASTQ-tiedostomuodossa. Yleensä on hyvä tarkastella sekvensoinnin onnistumista esimerkiksi riittävän saannon ja sekvenssien laadun pohjalta. Sen jälkeen voidaan suoraan varmistaa lajintunnistus tai alkaa tunnistaa geenejä, esimerkiksi resistenssi- tai virulenssigeenejä. Suoraan sekvenssiluvuista voidaan tehdä myös joitakin tyypityksiä kuten MLST. Readeista voidaan myös tehdä niin sanottu koonti (assembly), jossa readeista muodostetaan mahdollisimman täydellinen genomi eli niin sanottu draft-genomi. Tästä yleensä myös tunnistetaan ja nimetään automaattisesti geenit (annotaatio). Annotoidusta genomista voidaan tehdä genomin sisältöön liittyviä analyyseja, kuten ydin- ja pangenomianalyyseja. Ydingenomilinjauksesta voidaan tehdä myös fylogeneettinen puu, johon taas voidaan lisätä esimerkiksi tyypitystuloksia. Tarkin mahdollinen fylogeneettinen puu saadaan kuitenkin linjaamalla sekvensointireadit hyvään referenssigenomiin, joka on mahdollisimman läheistä sukua analysoiduille bakteerikannoille. Saadusta sekvenssistä tulee kuitenkin poistaa liikkuvat geneettiset elementit ja rekombinaatio, koska nämä eivät noudata normaalia mendelististä periytymistä. Saadusta linjauksesta voidaan sitten luoda fylogeneettinen puu tai tarvittaessa esimerkiksi ajallisesti linjattuja sukupuita.

maat pystyivät tarkistamaan aiemmin sekvensoitujen genomien kokoelmista muutamassa tunnissa, oliko tätä geeniä kantavia bakteereita jo sekvensoitu jossain muussa kontekstissa. Tässä tapauksessa nopea analyysi antoi tiedon näiden maiden tilanteesta tämän potentiaalisesti tärkeän geenin suhteen. Vastaavan analyysin voi tehdä esimerkiksi käyttämällä julkaistuja primer-sekvenssejä ja in silico -PCR:ää (polymerace chain reaction, polymeraasiketjureaktio) tallennettujen genomien kanssa.

Bakteerien aiheuttamien epidemioiden analyysit

Epidemioita on tutkittu vertailemalla mahdollisten epideemisten kantojen lajintunnistus- ja tyypitystuloksia herkkyysmääritystulosten kanssa. WGS antaa mahdollisuuden vertailla kantoja paljon tarkemmin. Sillä voidaan periaatteessa havaita yhden SNP:n erot genomeissa. Tämä lisää huomattavasti epidemiaselvitysten erottelukykyä ja luotettavuutta. Esimerkiksi saman sekvenssityypin (ST) kannoista osa saattaa kuulua epideemiseen klusteriin, mutta toiset eivät. Usein vain SNP- tai cgMLST-analyysi voi erotella suurella varmuudella epideemiset kannat siihen kuulumattomista kannoista. Kokogenomeista voidaan analysoida myös geenisisältöä, joka voi esimerkiksi erotella muuten läheistä sukua olevat kannat toisistaan.

Sairaaloissa leviäviä plasmidiepidemiota on myös voitu osoittaa sekvensoimalla. Tässä tapauksessa normaalit SNP- tai cgMLST-ana-

TAULUKKO 3. Analyysimenetelmiä (5–7,13–16,22–35)

	Sekvensointireadit	Koonti
Koonti (assembly)	Shovil, Unicycler, SPAdes, Velvet (optimizer)	
Geenien etsintä (annotaatio)	Prokka	
Lajinmääritys	Mash, Kraken	
Tyypitys MLST, serotyypitys	MLST (SRST2 ja vastaavat)	Mlst_check (MLST by BLAST)
Resistenssi SNP ja geenit Valitse tietokanta	ARIBA, SRST2, Resfinder ja PointFinder	Resfinder ja PointFinder, AMRFinderPlus, Abricate
Virulenssi	ARIBA, SRST2	
Fylogenia	Linjaus referenssiin: BWA (bamtools) SNP:t ja variantit: bcftools Rekombinaation poisto: Gubbins Puun rakennus: RAxML, IqTree	Roary: ydingenomi Snp-sites RAxML, FastTree, IqTree

lyysit eivät toimi, sillä niissä ei yleensä analysoida liikkuvaa geneettistä materiaalia, kuten plasmideja. Plasmidiepidemioissa samalla tai toisistaan eroavilla bakteerilajeilla on havaittu sama tai hyvin samankaltainen plasmidi tai resistenssigeenikasetti. Näissä tapauksissa analyysin voi aloittaa resistenssigeenien detektiolla sekä plasmidireplikonin tunnistuksella ja esimerkiksi sekvensoida tämän jälkeen potentiaalisia kantoja long-read-menetelmillä kokonaisten plasmidisekvenssien saamiseksi. Näitä voidaan käyttää myöhemmin referenssiplasmideina uusien kantojen vertailussa, jolloin voidaan saada selville, onko näissä kannoissa samaa tai hyvin samankaltaista plasmidia.

Bakteerikantojen sukulaisuuden analysointi

Kuvaamme kaksi hieman toisistaan poikkeavaa tapaa kokogenomisekvensoitujen bakteerikantojen sukulaisuussuhteiden eli fylogenian analysoimiseksi. Tuloksena syntyviin fylogeneettisiin puihin on hyvä liittää muita tyypitystuloksia ja kliinisiä tietoja. Näillä tiedoilla voidaan arvioida paitsi kantojen sukulaisuutta myös selvittää, ovatko kannat osa epidemiaa. Tulosten visualisointiin on saatavilla monia ilmaisia ohjelmia kuten esimerkiksi iTOL Interactive tree of life, Phandango ja Microreact (10–12).

Linjaus referenssigenomiin vaatii referenssigenomisekvenssin. Yleensä referenssiksi on hyvä valita tutkittavan bakteerikannan mahdollisimman läheinen sukulainen. Eri lajeille on valikoitunut referenssikantoja erilaisin perustein ja joillekin on myös omia referenssikantoja tunnetuille sekvenssityypeille. Sekvenssien linjaamisen jälkeen tehdään yleensä fylogeneettinen puu, jollakin siihen tarkoitetulla ohjelmalla. Puun saa rakennettua nopeasti esimerkiksi FastTree2:lla, mutta yleensä luotettava puu tehdään jotain ML-menetelmää (maximum likelihood) käyttävällä ohjelmalla, kuten RaxML:lla tai IqTree:lla (13–15). Luotettava fylogenia vaatii myös, että normaalin periytymisen rikkovat alueet poistetaan tai peitetään (masking) sekvenssistä ennen puun luontia. Tämä tarkoittaa bakteerien tapauksessa liikkuvia geenielementtejä kuten faageja, erilaisia patogeenisuus- ja resistenssisaarekkeita ja rekombinaatioalueita. Tähän voi käyttää esimerkiksi Gubbins-ohjelmaa (16). Huolellisesti luodusta linjauksesta on mahdollista tehdä myös ajallisesti linjattuja sukupuita esimerkiksi TempEst- tai BEAST2-ohjelmalla (17,18). Näitä analyyseja tarkastelemalla voidaan havaita esimerkiksi erilaisten patogeenisten bakteerilinjojen eriytymisen ajankohta.

Pangenomianalyyseilla voidaan tarkastella bakteerikantojen lähisukulaisuutta myös genomin geenisisällön kautta. Pangenomi tarkoittaa tässä tapauksessa kaikkia kokoelman, esimer-

Ydinasiat

- Bioinformatiikkaa tarvitaan sekä yksittäisten bakteerien että kokonaisten mikrobistojen sekvenssianalyyseihin, mikä avaa uutta aikakautta sekä infektioiden että mikrobistosta ja sen häiriöistä johtuvien terveysvaikutusten tutkimuksessa.
- Infektioepidemiologia on voimakkaimmin bakteriologisesta bioinformatiikasta hyötyviä aloja.
- Pitää tuntea käyttämiensä bioinformatiikkaohjelmien ja tietokantojen vaatimukset, mahdollisuudet, rajoitteet ja virhelähteet.
- Kaupallisia ohjelmia on usein helpompi käyttää kuin avoimen lähdekoodin ohjelmia, mutta niillä tehtävien analyysien yksityiskohdat eivät välttämättä ole käyttäjän saatavilla.

kiksi bakteerilajin, geenejä. Lähisukulaisilla on todennäköisesti samankaltaisempi genomin sisältö kuin kaukaisemmilla sukulaisilla. Tätä eroa ei kuitenkaan voida nähdä, jos tarkastellaan vain ydingenomia eli geenejä, jotka ovat kaikilla saman lajin edustajilla. Accessory-geenit eli geenit joita ei ole lajin kaikilla kannoilla, voivat kuitenkin paljastaa sukulaisuussuhteita, joita ei havaita ydingenomissa. Näitä voivat olla esimerkiksi plasmidit tai muut geenit, jotka eivät ole kaikille lajin kannoille välttämättömiä. Tätä voidaan tutkija esimerkiksi Roary- tai kehittyneemmällä Panaroo-ohjelmalla (19,20). Myös geenien säätelyalueita voidaan tutkia vastaavasti Piggy-ohjelmalla (21). Nämä menetelmät vaativat yleensä de novo -koonnin ja annotaation eli geenien havaitsemisen ja nimeämisen genomista esimerkiksi Prokka-ohjelmalla (22). Roary-ohjelmalla voidaan luoda myös ydingenomilinjaus, jota voidaan puolestaan käyttää luomaan fylogeneettinen puu, johon voidaan liittää muita pan-genomianalyysilla ja muilla tyypitysmenetelmillä saatuja tuloksia esimerkiksi epidemioiden selvittämiseksi. Fylogeneettisen puun luominen ydingenomin kautta voi olla nopeampaa kuin readien linjaaminen referenssigenomiin, mutta se ei luo yhtä luotettavaa fylogoniaa.

Mikrobistonäytteet

Mikrobistonäytteistä sekvensoidaan esimerkiksi kaikki näytteestä eristetty DNA (shotgun sekvensointi) tai monistetaan bakteereille ominaista DNA-sekvenssiä eli *16S rRNA* -geeniä, joka sekvensoidaan. Näytteen kaikkien organismien DNA:n analyysia kutsutaan metagenomiikaksi. Shotgun-menetelmillä voidaan saada selville näytteen bakteerien, alkueläinten, sienten ja virusten monimuotoisuus ja niiden määrälliset osuudet. Lisäksi voidaan tutkia näytteen ja sen sisältämien bakteerien metabolista potentiaalia, sekä esimerkiksi mikrobilääkeresistenssi- ja virulenssigeenejä.

Niin sanotulla 16S-sekvensoinnilla saadaan myös selville bakteerien monimuotoisuus ja määräsuhteet mutta ei tietoa muista bakteerien geeneistä. Sitä voidaan tehdä sekvensoimalla eri variaabeleita alueita kyseisestä geenistä, mutta eri alueet antavat usein huomattavastikin erilaisen tuloksen näytteen lajikirjosta. Menetelmä myös perustuu geenimonistukseen, joka voi aiheuttaa haasteita, koska se monistaa kaiken bakteeri-DNA:n näytteestä, tuli se sitten näytteestä itsestään tai kontaminaationa reagenssseista tai ympäristöstä. Shotgun-sekvensointi onkin korvaamassa 16S-sekvensointia mikrobistotutkimuksissa.

Analyysien haasteet

Monet analyysit, kuten resistenssi- ja virulenssigeenien detektio, perustuvat tietokantoihin. Yksittäisen bakteerikannan bioinformatiikka-analysoinnissa merkittävin virhelähde voikin olla vertailutietokanta. Osa niistä on yksittäisten tutkimusryhmien ja osa yliopistojen tai kansallisten instituuttien, kuten yhdysvaltalaisen NCBI:n ylläpitämiä. Vaikka monet näistä kehittyvät jatkuvasti, vanhojen tietokantojen ylläpito saattaa loppua yllättäen, eikä niiden laatu välttämättä ole parhaalla mahdollisella tasolla ylläpitäjän taloudellisen tilanteesta riippuen. Joskus onkin hyödyllistä harkita oman tietokannan rakentamista. Tällöin on tarkkaan

tiedossa, millä kriteereillä ja mihin tarkoitukseen se on rakennettu. Lisäksi voidaan olla varmoja, että kaikki tietokannan sekvenssit on tarkistettu ja nimetty oikein. Toisaalta myös oman tietokannan tekeminen ja ylläpitäminen vaatii resursseja kuten aikaa ja rahaa, joita ei usein ole riittävästi saataville.

Bioinformatiikassakin sattuu inhimillisiä virheitä, ja ohjelman tai tietokannan uusi versio voi olla viallinen ja antaa virheellisiä tuloksia. Varsinkin kliinisessä käytössä olevien ohjelmien ja tietokantojen uudet versiot täytyy testata normaaliin tapaan laatujärjestelmien mukaisesti ennen käyttöönottoa.

Muita tietoteknisiä haasteita voivat tuottaa esimerkiksi ohjelmien käyttämät lukuisat eri tiedostomuodot, laitevaatimukset sekä henkilökunnan kokemattomuus Linux- tai Unix-komentorivin käytöstä. Komentorivi on yleensä käytössä varsinkin avoimen lähdekoodin ohjelmissa, mutta kaupalliset ohjelmat käyttävät yleensä graafista käyttöliittymää ja ovat siksi usein helpompia käyttää. Toisaalta kaupalliset ohjelmat eivät aina ole avoimia käytettyjen analyysimenetelmien yksityiskohdista.

Yksittäisten kantojen analyysit eivät ole teknisesti vaativia, ja niitä voi suorittaa ainakin yleensä normaaleilla pöytätietokoneilla. Suurten kokoelmien analyysit taas vaativat huomattavaa panostusta laskentakapasiteettiin ja tallennustilaan. Tätä on kuitenkin saatavilla ainakin yliopistoille ja tutkimuslaitoksille CSC:ltä, Tieteen tietotekniikan keskus Oy:ltä. Tietoteknisiä haasteita voi helpottaa se, että monia työkaluja voi käyttää ilmaiseksi verkossa. Näiden käytettävyyttä voi hankaloittaa se, että analyyseja voidaan ajaa usein vain yhdelle tai muutamalle kannalle kerrallaan. Tämä kuitenkin mahdollistaa esimerkiksi lajinmäärityksen sekä resistenssigeenien analyysin ja usein myös muita tyypitys- ja analyysivaihtoehtoja. Esimerkiksi seuraavilla verkkosivuilla voi tehdä ilmaiseksi analyyseja sekvensoiduille kannoille: www.genomicepidemiology.org/services/, www.pathogen.watch/.

Lopuksi

Bioinformatiikkaa käytetään tällä hetkellä bakteriologisessa tutkimuksessa hyvin monenlaisen sekvensointidatan analytiikassa. Kliinisessä mikrobiologiassa eli mikrobiologisessa potilasdiagnostiikassa kaikkein lähimpänä rutiinikäyttöä ovat resistenttien bakteerien epidemiaselvitykset esimerkiksi *Staphylococcus aureus* -bakteerille ja enterobakteereille sekä mykobakteerien herkkyysmääritykset. Mikrobistoja analysoivassa metagenomiikassakin voitaisiin bioinformatiikan osalta tarjota diagnostisia tutkimuksia, mutta ongelmana ovat sekä hinta, diagnostiikan tekniset haasteet että kaikkein eniten löydösten kliinisen merkityksen ymmärtämisen keskeneräisyys. Tässä on paljon työsarkaa monien eri alojen asiantuntijoiden yhteistyölle.

Yhdistettynä muiden omiikoiden, kuten metabolomiikan ja proteomiikan analytiikkaan sekä koneoppimiseen, bioinformatiikka tulee jo lähitulevaisuudessa myös bakteriologiassa avaamassa kokonaan uutta aikakautta. Yksi voimakkaimmin kehittyvä ala on infektioepidemiologia. Lisäksi mikrobistoista ja niiden häiriöistä johtuvien terveysvaikutusten tutkimus ja seuraavassa vaiheessa myös diagnostiikka tuovat uutta näkökulmaa myös monien ei-tarttuvien tautien diagnostiikkaan ja hoitoon. ■

TEEMU KALLONEN, FT, erikoistutkija

ANTTI HAKANEN LT, dosentti, kliinisen mikrobiologian ylilääkäri, tulosryhmäjohtaja
Kliininen mikrobiologia, Turun kliininen mikrobistopankki
Tyks Laboratoriot, Tyks
Mikrobikeskus TY, Varha

TEEMAN TOIMITTAJAT
Sampsa Hautaniemi ja Tuomas Mirtti

SIDONNAISUUDET
Teemu Kallonen: Ei sidonnaisuuksia
Antti Hakanen: Koulutus-, konsultointi- asiantuntijatoiminta (Labquality, BioCodex, MSD ja Pfiser), luottamustoimet (Kliiniset mikrobiologit ry, Puheenjohtaja, Infektiopäivien järjestelytoimikunta/hallitus, Puheenjohtaja
Suomen biopankkien osuuskunta FINBB, Hallituksen jäsen)
Terveydenhuollon ohjaukseen pyrkivät hankkeet (FiRe, Suomalainen mikrobilääkeresistenssin tutkimusryhmä (Finnish Study Group for Antimicrobial Resistance)
Koordinaattori 2007-2014, Puheenjohtaja 2014-2023, Hallituksen jäsen 2023-)

KIRJALLISUUTTA

1. Christensen H, toim. Introduction to bioinformatics in microbiology. Learning materials in biosciences. New York: Springer 2018.
2. Andrews S. FastQC: a quality control tool for high throughput sequence data. Babraham Institute 2010 www.bioinformatics.babraham.ac.uk/projects/fastqc.
3. Ewels P, Magnusson M, Lundin S, ym. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics 2016;32:3047–8.
4. Gurevich A, Saveliev V, Vyahhi N, ym. QUAST: quality assessment tool for genome assemblies, Bioinformatics 2013;29:1072–5.
5. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol 2014;15:R46.
6. Inouye M, Dashnow H, Raven LA, ym. SRST2: rapid genomic surveillance for public health and hospital microbiology labs. Genome Med 2014;6:90.
7. Page AJ, Taylor B, Keane JA. Multilocus sequence typing by blast from de novo assemblies against PubMLST. J. Open Source Softw 2016;1:118.
8. Epping L, van Tonder AJ, Gladstone RA, ym. SeroBA: rapid high-throughput serotyping of Streptococcus pneumoniae from whole genome sequence data. Microb Genom 2018;4:e000186.
9. Liu YY, Wang Y, Walsh TR, ym. Emergence of plasmid-mediated colistin resistance mechanism MCR-1 in animals and human beings in China: a microbiological and molecular biological study. Lancet Infect Dis 2016;16:161–8.
10. Letunic I, Bork P. Interactive tree of life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Res 2021;49:W293–6.
11. Hadfield J, Croucher NJ, Goater RJ, ym. Phandango: an interactive viewer for bacterial population genomics. Bioinformatics 2018;34:292–3.
12. Argimón S, Abudahab K, Goater RJE, ym. Microreact: visualizing and sharing data for genomic epidemiology and phylogeography. Microb Genom 2016;2:e000093.
13. Price MN, Dehal PS, Arkin AP. FastTree 2 – approximately maximum-likelihood trees for large alignments. PLoS ONE 2010;5:e9490.
14. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies, Bioinformatics 2014;30:1312–3.
15. Nguyen LT, Schmidt HA, von Haeseler A, ym. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol 2015;32:268–74.
16. Croucher NJ, Page AJ, Connor TR, ym. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. Nucleic Acids Res 2015;43:e15.
17. Rambaut A, Lam TT, Max Carvalho L, ym. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). Virus Evol 2016;2:vew007.
18. Bouckaert R, Vaughan TG, Barido-Sottani J, ym. BEAST 2.5: an advanced software platform for Bayesian evolutionary analysis. PLoS Comput Biol 2019;15:e1006650.
19. Page AJ, Cummins CA, Hunt M, ym. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics 2015;31:3691–3.
20. Tonkin-Hill G, MacAlasdair N, Ruis C, ym. Producing polished prokaryotic pangenomes with the Panaroo pipeline. Genome Biol 2020;2:180.
21. Thorpe HA, Bayliss SC, Sheppard SK, ym. Piggy: a rapid, large-scale pan-genome analysis tool for intergenic regions in bacteria. Gigascience 2018;7:1–11.
22. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics 2014;30:2068–9.
23. Seeman T. Shovil 2020. https://github.com/tseemann/shovill.
24. Wick RR, Judd LM, Gorrie CL, ym. Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. PLoS Comput Biol 2017;13:e1005595.
25. Bankevich A, Nurk S, Antipov D, ym. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 2012;19:455–77.
26. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 2008;18:821–9.
27. Ondov BD, Treangen TJ, Melsted P, ym. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol 2016;17:132.
28. Hunt M, Mather AE, Sánchez-Busó L, ym. ARIBA: rapid antimicrobial resistance genotyping directly from sequencing reads. Microb Genom 2017;3:e000131.
29. Bortolaia V, Kaas RS, Ruppe E, ym. ResFinder 4.0 for predictions of phenotypes from genotypes. J Antimicrob Chemother 2020;75:3491–500.
30. Zankari E, Allesøe R, Joensen KG, ym. PointFinder: a novel web tool for WGS-based detection of antimicrobial resistance associated with chromosomal point mutations in bacterial pathogens. J Antimicrob Chemother 2017;72:2764–8.
31. Feldgarden M, Brover V, Gonzalez-Escalona N, ym. AMRFinderPlus and the Reference Gene Catalog facilitate examination of the genomic links among antimicrobial resistance, stress response, and virulence. Sci Rep 2021;11:12728.
32. Seemann T, Abricate 2020. https://github.com/tseemann/abricate.
33. Li H, Handsaker B, Wysoker A, ym. The sequence alignment/map format and SAMtools. Bioinformatics 2009;25:2078–9.
34. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009;25:1754–60.
35. Page AJ, Taylor B, Delaney AJ, ym. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. Microb Genom 2016;2:e000056.