

Matti Kankainen ja Soili Kytölä

Laajojen kliinisten sekvensointitutkimuksien bioinformatiikka nyt ja tulevaisuudessa

DNA-sekvensointia hyödyntävät laajat geenitutkimukset ovat yleistyneet terveydenhuollossa viime vuosina. Erityisen tärkeään asemaan nämä pienien emäsmuutoksien määrittämisessä erinomaisesti pärjäävät ratkaisut ovat nousseet syöpätautien ja perinnöllisten harvinaissairauksien diagnostiikassa ja hoidon ohjaamisessa. Viime vuosina tutkimuksia on yhä enemmän käytetty myös laajempien rakennepoikkeamien tunnistamisessa. Sekvensoinnissa tuotettujen valtavien tietoaisteistojen tulkinta ja analyysi on kuitenkin monimutkaista, ja tuotetun raakatiedon käsittely hoidollisesti merkittävässä ajassa vaatii tehokasta ja erikoistunutta tietotekniikkaa ja bioinformatiikkaa. Käytettävien tiedontallennus-, laskenta- ja ohjelmistoratkaisujen tulisi myös olla kustannusvaikuttavia ja vikasietoisia sekä geenitiedon arkaluonteisuuden vuoksi mahdollisimman tietoturvallisia kansallisten ja kansainvälisten lakien mukaisesti.

Sekvensointitekniologioiden nopean kehittyamisen myötä perimän luenta on arkipäiväistynyt. Nykymenetelmien avulla henkilön koko perimän tai perimän osan emäsjärjestys voidaan selvittää nopeasti ja vähäisin kustannuksin (1,2). Erityisesti tästä on hyötynyt syöpätautien ja perinnöllisten harvinaissairauksien diagnostiikka. Perimän sekvensoinnissa syntyvä raakatieto on yleensä kuitenkin ihmislukukelvotonta, ja sen käyttökelpoisuus riippuu pitkälti analyysissa käytettävistä algoritmeista. Esimerkiksi tuhansien referenssistä poikkeavien mutaatioiden havaitseminen miljoonista referenssinmukaisista emäshavainnoista on mahdotonta ilman tehokasta bioinformatiikkaa. Laadullisesti huonot tai tehtävään muuten epäsoyvät algoritmit voivat taas sivuuttaa sairauteen liittyviä mutaatioita. Koneoppimistoratkaisujen voidaan olettaa helpottavan tulevaisuudessa sekvensointitutkimuksien nykyistä ongelmakohtaa eli sairauteen tai sen hoitoon liittyvien muutosten erottelua harmittomista (3).

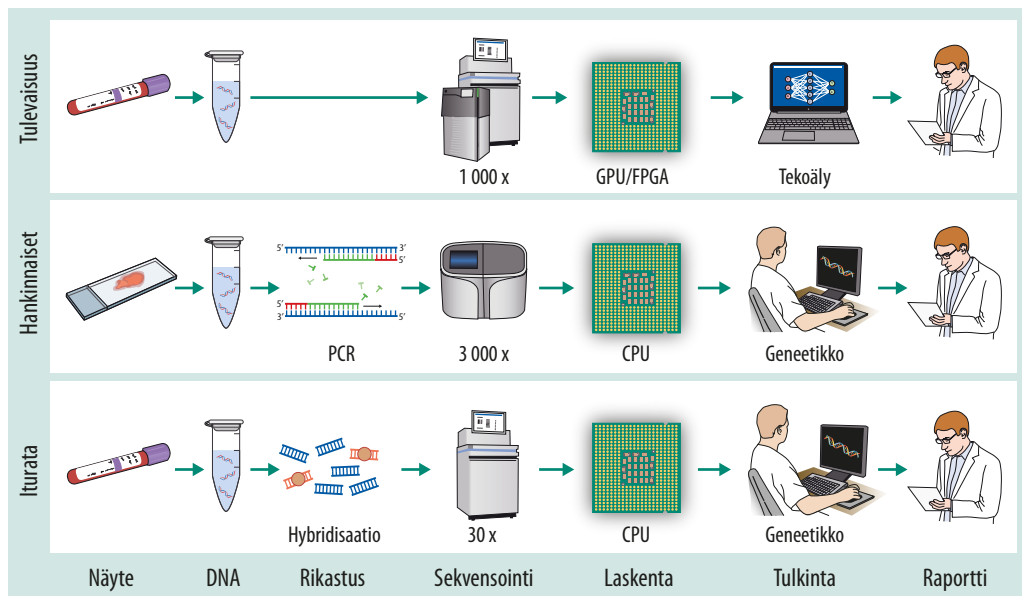
tegiaan ja -teknologiaan perustuvia laajoja DNA-sekvensointitutkimuksia (KUVA 1). Nämä tutkimukset kuuluvat suurimmilta osin testauslaboratorioiden pätevyysalueille ja tehdään usein SFS-EN ISO/IEC 15189:2013 (tai uudemman) standardin vaatimuksien mukaisesti. Pääasiallisesti tutkimuksien teossa käytetään Illuminan ja Ion Torrentin lyhyitä lukuja tuottavia sekvensointitekniologioita. Luonteeltaan tutkimukset ovat lähes poikkeuksetta kohdennettuja. Tutkimuksien kohdennuksissa on kuitenkin eroja, ja tarjolla on niin rajallisia määriä kliinisesti merkittäviä alueita käsittäviä paneelisekvensointitutkimuksia kuin myös lähes kaikki genomien koodaavat alueet sisältäviä eksomitutkimuksia. Esimerkiksi HUS-diagnostiikkakeskuksessa molempia käytetään laajalti. Myös rakennepoikkeavuuksien sisällyttämisissä tutkimuksiin on eroja. Tulevaisuutta silmällä pitäen testauslaboratorioissa on aloitettu kartoittamaan koko genomien sekvensoinnin etuja ja haittoja.

Laajat kliiniset sekvensointitutkimukset

Suomalaisissa kliinisen genetiikan laboratorioissa on tarjolla useaan sekvensointistra-

Laajojen kliinisten sekvensointitutkimuksien bioinformatiikka

Laajoihin geenitutkimuksiin liittyvien bioinformatiivisten analyysien tarkoituksena on ja-



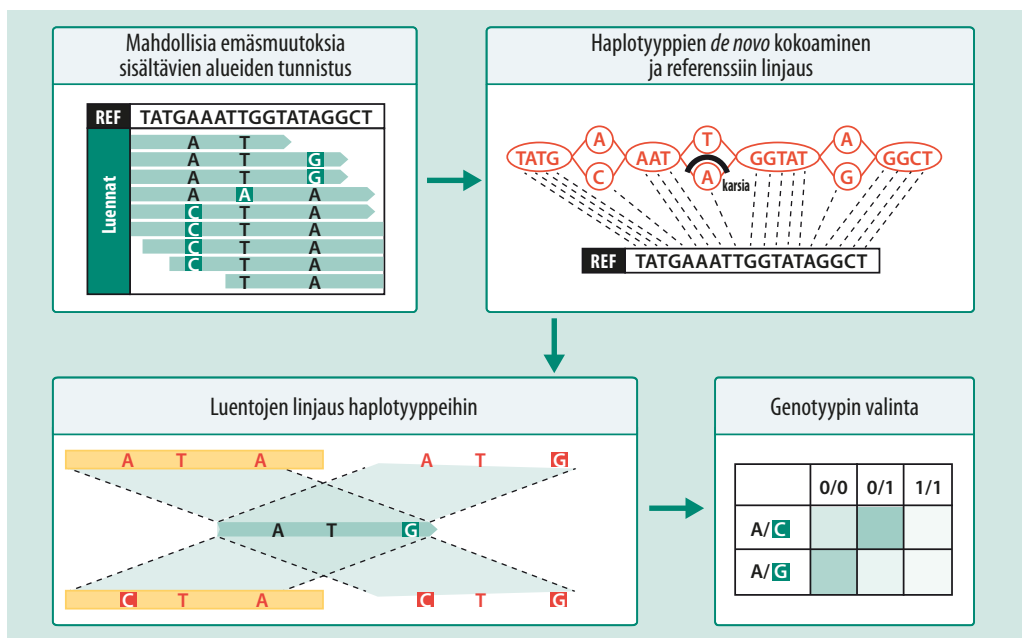
KUVA 1. Laajojen kliinisten DNA-sekvensointitutkimuksien polku näytteestä diagnoosiin. Tutkimuksien edellytyksenä on sairautta edustava ja tarpeeksi hyvälaatuista DNA:ta sisältävä näyte. Näytekesittelyssä näytteen DNA eristetään ja DNA:han liitetään sekvensointi-, näytetunniste- tai PCR-tunnisteadapterit. Kohdealueet rikastetaan tarvittaessa PCR-monistuksella tai hybridisaatioalavalmilla. Ennen rinnakkaissekvensointia useampi näytekirjasto yhdistetään instrumentin sekvensointikapasiteetin jakamiseksi. Kasvainnäytteiden sekvensoinnissa käytetään usein suurta lukupeittoa (> 3 000 x) kun taas perinnöllisten näytteiden kohdalla käytetään maltillisempaa (~ 30 x) lukupeittoa. Sekvensointi tehdään nykyisin yleensä lyhyitä sekvenssilukuja tuottavilla sekvensointitekniikoilla. Tulevaisuudessa pidempiä sekvenssilukuja tuottavien teknologioiden käyttö sekvensoinnissa todennäköisesti yleistyy. Bioinformatiivisessa analyysissä emäs-

kutsusignaalit muutetaan näytekohtaisiksi sekvenssiedostoiksi, luvuista poistetaan huonolaatuiset- sekä adapteriemäkset, luvut linjataan referenssigenomiin, PCR-monistetut luvut tunnistetaan linjaustietojen tai PCR-tunnisteadapterien avulla, emäs-kohtaiset laatuarvot arvioidaan uudelleen ja emäsmuutokset tunnistetaan. Analyysissä voidaan käyttää tavanomaisia CPU- tai kiihdytettyjä GPU/FPGA-suorittimia. Tulkinassa havainnot luokitellaan tietokoneavusteisesti. Monet työhön soveltuvat ohjelmistot esittäytävät arvioitavia luokittelukohtia tekemiensä tietokantahakujen perusteella. Vaikka arvioitavista 28 kohdasta useimmat ovat esitettävissä, luokittelu voi olla työlästä. Tulevaisuudessa syväoppimismenetelmät voivat tehostaa luokittelutyötä. Patogeeniset, todennäköisesti patogeeniset ja merkitykseltään epäselvät muutokset raportoidaan, kun taas todennäköisesti harmittomiksi ja harmittomiksi tulkittuja muutoksia ei raportoida.

lostaa sekvensoinnissa tuotettu massiivinen ja ihmislukukelvoton tieto hyödyllisempään muotoon ja auttaa tulkitsijaa ymmärtämään sitä. Yleensä työvaihe alkaa emäskutsusignaalin muuntamisella sekvenssiluvuiksi ja sisältää eri näytekirjastoihin kuuluvien sekvenssilukujen erottelun toisistaan. Työvaihe tehdään tavallisesti laitevalmistajien ohjelmistoilla, ja sen lopputuloksena on näytekohtaiset sekvenssiluvut sisältävä tiedosto. Tallennustilan säästämiseksi tulostiedostot yleensä pakataan. Työvaiheessa yleisesti käytetyt pakkausalgoritmit kuitenkin pakkaavat sekvenssitetietoa varsin epäoptimaalisesti, ja tarvittaessa tiedostoja voidaan huomattavasti tiivistää lisää esimerkiksi SPRING ja

Original Read Archive (ORA) pakkausmenetelmien avulla (4).

Seuraavaksi tehtävässä linjausanalyysissä selvitetään alkuperäisten tai adapterisekvensseistä ja muista virheistä siivottujen sekvenssilukujen sijainnit referenssigenomissa. Yleensä linjaukset tehdään ultranopeiden Burrows–Wheeler-muunnosta hyödyntävien linjausalgoritmien avulla, vaikkakin myös muunlaisia ratkaisuja on olemassa (5). Nykyalgoritmien linjaustulokset ovat yleensä laadukkaita. Ongelmat liittyvät lähinnä moniselitteisesti ja hankalille genomialueille linjautuviin sekä pitkiä poistuma- ja lisäysmuutoksia (indel) sisältäviin sekvenssilukuihin. Kliinisissä suosituksissa linjauksissa suositel-



KUVA 2. Pienien emäsmuutoksien tunnistus sekvensointitiedosta graafiteorian avulla. Emäsmuutoksien tunnistuksessa potentiaaliset emäsmuutosalueet tunnistetaan linjaustiedon perusteella, potentiaalisten muutosalueille sekä niitä reunustavilla genomialueille linjattujen sekvenssilukujen tukemat haplotyytit rakennetaan De Bruijn kaltaisten graafien avulla ja havaitut haplotyytit linjataan referenssigenomiin aina optimaalisen linjauksen löytävällä Smith–Waterman-algoritmillä. Viimeisessä vaiheessa muutosalueen

sekvenssiluvut linjataan haplotyyppisiin ja haplotyyppien laadut arvioidaan niitä tukevien sekvenssilukujen perusteella. Monissa ohjelmistoissa emäsmuutoshavaintojen herkkyyttä voidaan parantaa tekemällä analyysi yhtäaikaaisesti verrokkipotilaiden tai sukulaisnäytteiden kanssa. Menetelmässä lainataan tietoa näytteiden välillä ja hyväksytään heikkolaatuiset emäsmuutokset, jotka on luotettavasti tunnistettu toisissa näytteissä. Kuvan esimerkissä kuvattu GATK-kutsuohjelmiston tapa tunnistaa emäsmuutoksia (9).

laan käytettävän uusimpia referenssigenomeita, joissa on muun muassa paikattu aukkoja ja mahdollista analysoida poikkeaviakin haplotyyppisiä (6). Esimerkiksi uusin T2T-referenssigenomi tarkoittaa emäsmuutoksien tunnistusta lähes puolella aiemmasta (7). Laboratorion lausuntokäytäntöiden päivityshankaluuksien vuoksi kliinisessä työssä kuitenkin usein turvaututaan vanhempiin referenssigenomeihin.

Pienien emäsmuutoksien tunnistuksessa pyritään löytämään tutkittavan näytteen ja referenssigenomin väliset pistemutaatiot sekä pienehköt (noin parikymmentä emästä pitkät) indel-muutokset. Koska virheiden seurauksena potilas voi saada väärän diagnoosin tai hoidon, vain tarkimpia kutsuohjelmia tulisi hyödyntää. Nykytietämyksen valossa monet parhaista kutsuohjelmista ovat kaupallisia ja hyödyntävät graafiteoriaa mutaatioiden tunnistamisessa (KUVA 2) (8). Lisäksi monien ohjelmistojen

avulla on mahdollista käsitellä useampia näytteitä yhtäaikaisesti (9). Tämä analyysintapa tarkoittaa tuloksia huonosti katetuilla alueilla ja auttaa todentamaan periytymättömien de novo-mutaatioiden puuttumisen sekvenssoiduista perheenjäsenistä. Koska liian pieni lukupeitto voi estää mutaation havaitsemisen, laboratorioden tulisi määrittää tutkimuskohtaiset lukupeittokriteerit, joiden avulla ne arvioivat saavuttavansa tutkimuksellensa asettamansa suorituskykyvaatimukset. Ihannetapauksissa lukupeittokriteerit huomioivat kohdealueiden ominaispiirteet ja olisivat kohdealuekohtaisia.

Hankitut ja perinnölliset emäsmuutokset tunnistetaan pääpiirteittäin samoin. Koska hankitut emäsmuutokset voivat esiintyä millä tahansa alleelifrekvenssillä 0 ja 100 %:n välillä, somaattisten muutoshavaintojen seassa voi olla runsaastikin sekvensointi- tai muista virhelähteistä kumpuvia virheitä. Virheiden määrää voi-

daan kuitenkin merkittävästi vähentää hyödyn-
tämällä analyysissa potilaan normaalinäytettä
ja raportoimalla kasvainnäytteille tunnusomai-
set muutokset (10). Monissa kutsuohjelmissa
virheellisiä havaintoja voidaan myös suodattaa
syöpään liittymättömistä normaalinäytteistä
kerättyjen mutaatiohavaintojen avulla. Vaik-
ka suodatus vähentää virheitä, potilaan nor-
maalinäytteen käyttö mahdollistaa tarkemmat
analyysit, minkä vuoksi niiden käyttöä tulisi
näytteenottohankaluuksista huolimatta suosia.
Esimerkiksi HUS-diagnostiikkakeskuksessa
tehdyissä validointikokeissa potilaskohtaisten
normaalinäytteiden avulla voitiin poistaa kol-
mannes eksomitutkimuksessa havaituista ami-
nohappoa vaihtavista emäsmuutoksista. Myös
lukupeitto vaikuttaa suorituskykyyn (11).

Kromosomimuutoksien bioinformatiikka

Pienien emäsmuutoksien lisäksi ihmisgenomi
sisältää tuhansia suuremman kokoluokan ra-
kennepoikkeamia (12). Näiden tunnistamiseen
sekvensointitiedosta on kehitetty useita lasken-
nallisia lähestymistapoja (KUVA 3). Lähesty-
mistavoista erityisesti epätavallisesti linjattuja
parittaisia sekvenssilukupareja ja epäjatkuvasti
linjattuja sekvenssilukuja käyttävät kutsuohjel-
mat tunnistavat laadukkaasti muutoksia koko
genomin sekvensoinnilla tuotetuista sekvens-
sitietoaaineistoista (13). Kohdennetuissa tutki-
muksissa muutoksien tunnistus taas rajautuu
lähinnä lukupeittovertailuun ja kopiolumuu-
toksien havainnointiin. Kopiolukumuutoksien
tunnistuksella voidaan kliinistä vaikuttavuutta
kuitenkin parantaa jopa 5 % (14). Tulokset tu-
lisi tosin virhelöydöksen suuren määrän vuoksi
varmistaa muilla menetelmillä. Kohdennetuis-
sa tutkimuksissa muiden lähestymistapojen
kompastuskivenä on sensitiivisyys, koska muu-
tosten katkeamiskohdat sijaitsevat harvoin
koodaavilla alueilla, jonka vuoksi negatiivinen
tulos ei sulje pois geneettistä taustaa. Kasvain-
näytteisiin liittyvät rakennemuutokset ennuste-
taan pitkälti kuten ituradalliset. Eroavuudet liit-
tyvät tarpeeseen huomioida kasvaimen osuus
näytteestä ja mahdollinen alikloonaalisuus sekä
tunnistaa heterotsygootian menetykset. Lisäksi

kasvaimille tyypillisten kromosomi- ja käsivar-
sikokoisten poikkeamien tunnistus voi vaatia
erillisten kutsuohjelmien käyttöä.

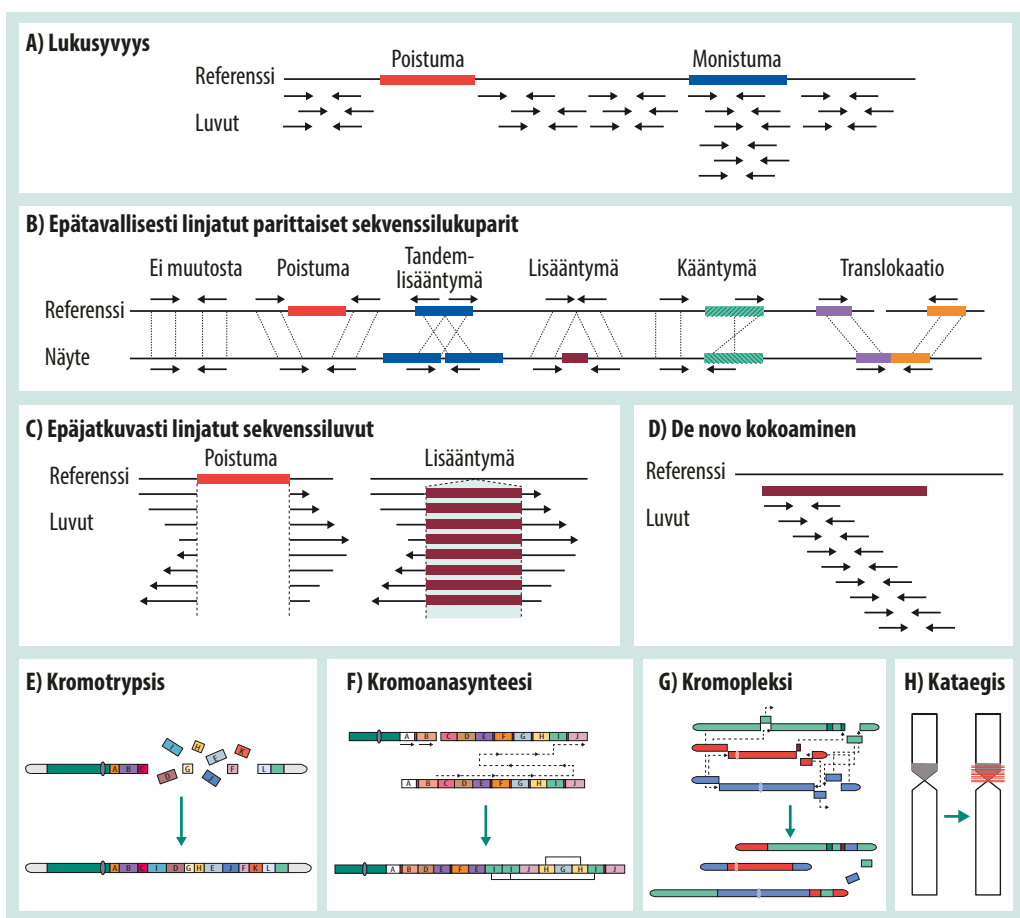
Kompleksisten kromosomipoikkeamien bioinformatiikka

Kromotripsis, kromoanasynteesi ja kromo-
pleksis ovat yksittäisiä katastrofitapahtumia
(KUVA 3), joissa muutama (kromotripsis ja kro-
moanasynteesi) tai useampi (kromopleksis)
kromosomi hajoaa osiksi ja osista muodostuu
kaoottisessa järjestyksessä olevia uusia kro-
mosomeja (15). Niitä voi esiintyä ituradassa,
mutta huomattavasti yleisempiä ne ovat syöpä-
kasvaimissa, vaikkakin näiden kompleksisten
kromosomipoikkeamien roolia syövän kehityk-
sessä ei täysin ymmärretä (16,17). Menetelmäl-
lisesti koko genomin sekvensointi on yliver-
tainen verrattuna kohdennettuihin tutkimuksiin
sensitiivisyyden vuoksi.

Mutaatiohavaintojen tulkinta

Emäsmuutoksien tulkinnassa harmittomat
emäsmuutokset erotellaan tutkittavaan sairau-
teen tai sen hoitoon liittyvistä keräämällä ja
arvioimalla niihin, sairauteen ja potilaaseen liit-
tyvää tietoa. Tehtävään on useita ohjelmistoja
ja toimintamalleja, ja se on tutkimuksen työ-
läimpiä vaiheita. Erityisesti eksomitutkimuksi-
en, joiden lopputulemana voi olla kymmeniä-
tuhansia havaintoja, tulkinta voi olla työlästä.
Näiden ja muiden tutkimuksien työkuorman
kohtuullistamiseksi tulkintaa voidaan rajata, ja
usein rajataankin, ennalta valittuihin tautigee-
neihin tai oirekuvaan sopiviin geeneihin. Koska
aihetta on käsitelty kattavasti aiemmin, käsitte-
lemme tulkintatyötä seuraavassa yleisluonteis-
emmin (18,19).

Useimmat kliiniset testauslaboratoriot käsit-
televät havaintoja kansainvälisten suositusten
mukaisesti (20–22). Näissä suosituksissa emäs-
poikkeamien patogeenisuuden arviointi perus-
tuu monenlaisiin tietoihin. Esimerkiksi itu-
ratamuutoksien kohdalla luokittelutulokseen
vaikuttaa muutoksen vaikutus (hiljainen, ami-
nohappoa vaihtava ja proteiiniirakennetta rikko-



KUVA 3. Laskennalliset lähestymistavat rakennepoikkeamien etsintään. **A)** Lukupeittovertailussa poistumat ja monistumat tunnustetaan esimällä genomialueiden tai näytteiden välisiä lukupeittopoitkeamia. Koska saman kopiolumun genomialueet voivat kohdennetussa sekvensoinnissa tulla sekvensoiduksi erilaisilla lukupeitoilla, lukupeittolukuja usein korjataan normalisointi- tai monimuuttujamenetelmillä. Analyysin päätteeksi peräkkäiset ja saman lukupeiton alueet yhdistetään yhtenäisiksi ennustuksiksi. Jos analyysissä käsitellään useampia näytteitä, niiden näytekäsittelyt ja näytetyypit olisi hyvä olla yhteneväiset. **B)** Epätavallisesti linjattuja parittaisia sekvenssilukupareja hyödyntävässä analyysissä rakennepoikkeamat tunnustetaan vertaamalla lukuparien etäisyyksiä ja linjaussuuntia oletettuihin. **C)** Epäjatkuvasti linjattuja sekvenssilukuja hyödyntävässä lähestymistavassa etsitään lukuja, jotka ovat linjautuneet epäjatkuvasti

useampaan kohtaan referenssigenomissa. **D)** De novo kokoamisessa sekvenssilukujen tukemat haplotyyppit rakennetaan De Bruijn kaltaisten graafien avulla. Toisin kuin pienien emäsmuutoksien kohdalla syötteenä voi tosin olla kaikki luvut. **E)** Kromotrypsiksessä yksi tai muutama kromosomi hajoaa osiksi yksittäisessä katastrofitapahtumassa ja osista muodostuu yhtä monta kaoottisessa järjestyksessä olevaa uutta kromosomia. **F)** Kromoanasynteessä yhden tai muutaman kromosomin replikaatio epäonnistuu ja synnyttää kromosomaalisia monistumia sekä muita kromosomivirheitä. **G)** Kromopleksissä useampi kromosomi hajoaa osiksi yksittäisissä katastrofitapahtumissa ja osista muodostuu useampi kaoottisessa järjestyksessä oleva kromosomi. **H)** Kataegis kuvaa paikallista hypermutaatiotapahtumaa, missä pienelle DNA-alueelle kohdentuu monia pistemutaatioita.

va emäsmuutos), laskennallisten ennusteohjelmien arviot haitallisuudesta, esiintyvyys väestössä, aiemmat luokittelu- ja patogeenisuusarviot, kirjallisuustieto sekä periytymismalli. Näiden ja muiden olennaisten kuvailutietojen

perusteella genetiikan asiantuntijat luokittelevat havainnot ennalta sovittujen sääntöjen mukaan patogeenisuusluokkiin. Vaikka iturata-tulkintaprosessissa arvioitavista 28 kohdasta 18 on kerättävissä laskennallisesti, luokitusprosessi

Paikallinen	IaaS (Infrastruktuuri palveluna)	PaaS (Palvelualusta palveluna)	SaaS (Ohjelmisto palveluna)	
Sovellukset	Sovellukset	Sovellukset	Sovellukset	
Data	Data	Data	Data	
Suoritus aika	Suoritus aika	Suoritus aika	Suoritus aika	
Väliohjelmisto	Väliohjelmisto	Väliohjelmisto	Väliohjelmisto	
Käyttöjärjestelmä	Käyttöjärjestelmä	Käyttöjärjestelmä	Käyttöjärjestelmä	
Virtualisointi	Virtualisointi	Virtualisointi	Virtualisointi	
Palvelimet	Palvelimet	Palvelimet	Palvelimet	
Tallennus alusta	Tallennus alusta	Tallennus alusta	Tallennus alusta	
Verkko	Verkko	Verkko	Verkko	

Laboratorion vastuulla

Toimittajan vastuulla

KUVA 4. Laboratorion omissa tiloissa ja pilvikonesalissa sijaitsevien laskentapalvelimien ja analyysiohjelmistojen ylläpitovastuiden tyypillinen jako. Laboratorion omissa tiloissa olevat palvelimet ovat tyypillisesti laboratorion itsensä tai ulkopuolisen toimittajan ylläpitovastuulla. Vaihtoehdon etuna ovat lyhyet vasteajat ja nopea tiedonsiirto. Ratkaisu ei kuitenkaan skaalaudu käytön mukaan, palvelin- ja ohjelmistoylläpitovastuut voivat olla epäselvät ja testi- ja varaympäristöjen toteutus voi olla puutteellinen. Tallennusratkaisu ei myöskään skaalaudu ja tallenteiden varmennus voi olla hankalasti toteutettavissa. Pilvipalvelujen ylläpito-

vastuu riippuu tuotteesta ja palveluntarjoajan ylläpitovastuu voi rajoittua vain verkko-, palvelin- tai tallennus alustoihin (IaaS palvelu), ylettyä myös käyttöjärjestelmään ja väliohjelmistoihin (PaaS) tai kattaa kaiken varsinaiseen analysointiohjelmistoon asti (SaaS). Palvelujen etuna on skaalautuvuus, palvelin- ja ohjelmistoylläpitovastuiden selkeys, laitekannan ajantasaisuus ja mahdollisuus maantieteelliseen tai konesali kohtaiseen hajautukseen. Haittana on tiedonsiirron hitaus. Pilvialustojen epäoptimaalinen varaus voi myös tuottaa yllättäviä kustannuksia.

on työläs (23). Tulevaisuudessa koneoppimis-perusteiset tulkintatyökalut todennäköisesti tehostavat työtä (3).

Yksittäisten kasvainmutaatioiden raportointiin on myös olemassa kansainvälisiä suosituksia (24–26). Useimmissa näistä mutaatiot luokitellaan moniportaisesti huomioiden diagnostinen, ennusteellinen tai hoidollinen merkitys. Tämän lisäksi kasvainanalyysien yhteydessä raportoidaan usein mutaatiohavaintoihin liittyviä tunnuslukuja. Näistä keskeisimpiä ovat mutaatioiden määrää tutkimus aluetta kohden kuvaava mutaatiokuorma, mikä ennustaa immuunoterapian hyötyä, sekä mikrosatelliittitoistojaksojen stabiiliutta kuvaavat indeksit (27,28). Rakennemuutoksien jatkoanalyysillä voidaan taas arvioida homologisen rekombinaatioreitin (HR) häiriintymistä ja PARP-estäjien hyötyä hoidossa (29). Koska kasvainten mutageeniset

prosessit synnyttävät niihin tunnusomaisia mutaatioita, kasvainmutaatioiden avulla voidaan tunnistaa kasvaimeen vaikuttaneita mutageenisia prosesseja (30). Näitä profilointianalyysijä tehdään kuitenkin harvalti vielä diagnostiikkas- sa. Muiden mainittujen tunnuslukujen raportointi on taas perustoiminto useimmissa kaupallisissa tulkintaohjelmistoissa. Vaihtoehtoisesti tunnusluvut ovat laskettavissa tehtävään suunnitelluilla erillisohjelmistoilla (28,29,31).

Potilaan taudinkuvaan sopivien geenien lisäksi laajoissa sekvensointitutkimuksissa voidaan havaita muihin sairauksiin liittyviä periytyviä geenivirheitä. American College of Medical Genetics and Genomics (ACMG) nykysuosituksessa, jota laajalti noudatetaan myös Suomessa, nämä niin sanotut sivulöydökset tulisi raportoida tutkittavan niin toivoessa vähintään 78 geeniin liittyen (32,33).

Alati kasvava emäsmuutostieto voi muuttaa aiemmin tehtyjä luokituksia ja mahdollistaa ajan saatossa toisenlaisen diagnoosin ja hoitosuosituksen. Tämän vuoksi genetiikan alan useat kansainväliset kattojärjestöt kannustavat analysoimaan uudelleen sekvensointitietoa (34–36). Myös Suomessa useimmat testauslaboratoriot suorittavat tarvittaessa analysointia uudelleen, kuten esimerkiksi HUS-diagnostiikkakeskus.

Geenitietojärjestelmät

Geenidiagnostiikassa käytettävät analyysiohjelmistot ovat yleensä kaupallisia, vaikka lainsäädäntö ei estä omien tai tutkimuksellisten ohjelmistojen käyttöä omavalmisteina, jos potilashoidon tarpeita ei muuten voida täyttää. Koska ohjelmistojen suorittamat analyysit vaativat laskentatehoa, muistia ja tallennuskapasiteettia, analyysit suoritetaan useimmiten laboratorion tiloissa, terveydenhuoltoyksikön konesaleissa tai pilvessä sijaitsevilla teholaskentapalveluissa. Vaihtoehtoja on omat vahvuutensa ja heikkoutensa, ja useampi voi olla varteenotettava vaihtoehto. Niillä on myös erilaiset palvelumallit (KUVA 4). Mikäli tutkimukset on vastattava nopeasti, GPU- tai FPGA-prosessorit ovat huomionarvoinen vaihtoehto. Näitä resursseja on tosin vähäisesti saatavilla, ja ne ovat kalliita. Kaikki algoritmit eivät myöskään osaa hyödyntää näitä erikoisprosessoreja.

Geenitiedon käsittelyssä käytettävien tietojärjestelmien lainsäädäntö

Useat lait, asetukset, ohjeet ja määräykset säätelevät sekvensointitutkimuksissa tuotetun arkaluonteisen geenitiedon hallintaa ja käsittelyä sekä näiden tietojen käsittelyyn tarkoitettujen tietojärjestelmien toteutusta. Näistä erityismaininnan arvoinen on asiakastietolaki (784/2021). Laki velvoittaa suunnittelemaan ja valmistamaan henkilötietojen käsittelyssä käytettävät tietojärjestelmät tietoturva-, tietosuojaa ja yhteentoimivuutta koskevien lakien, säännösten ja määrittelyjen mukaisesti. Laki

myös velvoittaa terveyspalvelujen tuottajat laatimaan itselleen tietoturvasuunnitelman ja määrittelee miten, milloin ja kuka tietojärjestelmiä ylläpitää ja käyttää. Geenitiedon arkistointia ohjaava potilasasiakirja-asetus (94/2022) määrittää lopullisten laboratoriotuloksien säilytysajaksi vähintään 12 vuotta ja perinnöllisyyslääketieteellisten yksiköiden tutkimuksissa muodostuvat asiakirjat pysyvässä säilytettäväksi. Henkilötietojen käsittelyä säätelevä EU:n yleinen tietosuojalaki (EU 2016/679) velvoittaa arvioimaan henkilötiedon käsittelytoimien tarpeellisuutta ja oikeasuhteisuutta sekä henkilötiedon käsittelystä aiheutuvia riskejä ja niiden hallintakeinoja. Lisäksi tietojärjestelmien ja niiden valmistajien on täytettävä EU:n IVD-asetuksen (EU 2017/746) ehdot, mikäli valmistaja on tarkoittanut niitä käytettäväksi in vitro -diagnostiikkasäätölaiteiksi. Kaupallisten laitteiden lisäksi IVD-asetus säätelee osaltaan myös terveydenhuoltoyksiköiden omaa laitevalmistusta ja velvoittaa laatimaan omavalmisteille käyttötarkoitusta, suunnittelu- ja valmistusprosessia sekä suorituskäytännöt koskevat asiakirja-aineistot.

Akkreditointialueella käytettävät geenitietojärjestelmät

Laajojen geenitutkimuksien teossa käytettävien tietojärjestelmien on täytettävä laboratorion noudattamien standardien vaatimukset, jos niitä käytetään akkreditoitujen tutkimuksien osana. Tällöin niihin kohdistuu samat vaatimukset kuin muihinkin akkreditointialueen menettelyihin, ja ne on hankittava, dokumentoitava, otettava käyttöön ja ylläpidettävä standardin ja sisäisten menettelytapojen mukaisesti. Ne on myös verifioitava tai validoitava käyttötarkoitukseensa. Suorituskäytännöt voisi esimerkiksi olla tunnistaa verinäytteiden mutaatiot ja indelit 99,5 %:n herkkyydellä ja 99,5 %:n positiivisella ennustearvolla. Lisäksi standardi edellyttää määrittämään hallintavastuut, käytänteet häiriötilanteissa ja menettelyt tallenteiden hallintaan. Standardit myös asettavat vaatimuksia henkilökunnan perehdyttämiseen, pätevyyden valvontaan ja ammattitaidon ylläpitoon.

Useat standardit asettavat omia erityisvaati-

Laaja geenitutkimus tai sekvensointitutkimus: useamman kohdealueen emäsmuutokset tunnistava rinnakkaissekvensointi-instrumenteilla tehtävä DNA-tutkimus.

Pistemuutos (eng. single nucleotide variant): perimäaineen yksittäisen emäksen vaihtuminen toiseksi emäkseksi.

Insertio-deleetiomuutos (eng. insertion-deletion variant): perimäaineen poistuma tai lisäys.

Pienet emäsmuutokset (eng. short variant): Emäsmuutosryhmä, joka käsittää yksittäisten emästen pistemuutokset sekä yleisimmin noin parikymmentä emästä pitkät indel-muutokset. Kategoriaan voidaan laskea kuuluvaksi myös pidempiä indel-muutoksia algoritmin mukaan.

De novo -mutaatio (eng. de novo mutation): potilaan ituradan genomissa oleva mutaatio, jota ei ole potilaan vanhempien ituradan genomeissa.

Referenssigenomi (eng. reference genome): useiden eri yksilöiden DNA-sekvensseistä koostettu kuvaus lajin genomista. Uusimmat ihmisreferenssigenomit kuvaavat kunkin alueen yleisintä vaihtoehtoa.

Linjausalgoritmi (eng. read alignment tool): sekvenssilukuja referenssigenomiin vertaileva algoritmi. Algoritmin avulla selvitetään sekvenssilukujen sijainnit referenssigenomissa. Sekvenssiluvut, joiden lähtökohtaa ei voida selvittää yksiselitteisesti saavat huonomman laatuarvon kuin sekvenssiluvut, joiden lähtökohta on yksiselitteinen.

Kutsuohjelma (eng. variant caller): sekvensointitiedon emäshavainnoiksi muuttava algoritmi. Algoritmin avulla voidaan mm. tunnistaa, miltä osin potilas eroaa referenssigenomista.

Emäshavainto (eng. base-call): Emäsmuutoksien kutsussa käytettävän kutsuohjelman ennustus tietyn genomikohdan sisältämistä emäksistä. Tuloksena voi olla referenssin mukainen havainto tai referenssistä poikkeava mutaatiohavainto.

Mutaatiokuorma (eng. tumor mutation burden): hankinnaisten mutaatiohavaintojen lukumäärä per sekvensointitutkimuksen kattaman genomialueen koko kertaa miljoona. Tuloksena voi olla esimerkiksi luku 11 mutaatiota per Mb.

Haplotyyppi (eng.haplotype): isän tai äidin puoleisen kromosomin tietyn (yleensä lyhyehkön) genomialueen sisältämien alleelien yhdistelmä. Alueen alleelit periytyvät usein yhdessä

Sekvenssiluku (eng. read): kuvaus yksittäisen DNA- tai RNA-molekyylin emäsjärjestyksestä. Molekyylin emäsjärjestys voidaan sekvensaattorista riippuen selvittää molekyylin yhdestä tai molemmista päistä. Kun sekvensointi on tehty molemmista päistä, molekyylistä saadut kaksi lukua muodostavat yhdessä parittaisen sekvenssilukuparin (eng. paired-end read).

Rakennepoikkeama (eng. structural variant): perimäaineuksessa oleva laajahko poikkeama. Rakennepoikkeamien koolle ei ole olemassa tarkkaa kokorajaa ja joissain tapauksissa ryhmään on sisällytetty kaikki yli 50 emästä käsittävät poikkeamat. Yleisimmin rakennepoikkeamilla kuitenkin viitataan yli tuhat emästä käsittäviin poikkeamiin. Useimmiten ryhmään on sisällytetty niin poistumat, lisäykset, translokaatiot, kääntymät ja monistumat kuin myös laajat ja monimutkaiset kromosomitason muutokset, mutta tähänkin sääntöön on poikkeuksia.

Epätavallisesti linjattu parittainen sekvenssilukupari (eng. discordantly mapped read-pair): parittainen sekvenssilukupari, jonka sekvenssiluvut linjautuvat oletettua lähemmäksi tai kauemmaksi toisistaan tai epätavallisiin suuntiin referenssigenomissa.

Epäjatkuvasti linjattu sekvenssiluku (eng. split read): sekvenssiluku, jonka linjaus hajautuu eri genomialueille. Esimerkiksi ensimmäiset 40 emästä kattavat referenssikromosomissa alueen 500–540 ja viimeiset 60 emästä alueen 2500–2560.

Henkilötieto: Tieto, mikä liittyy tunnistettuun tai tunnistettavissa olevaan henkilöön (EU 2016/679). Geenitiedot, jotka sisältävät potilaan yksilöiviä mutaatioita tai mutaatioyhdistelmiä, ovat itsestään henkilötietoja. Geneettiset tiedot ovat myös ns. erityisiin henkilötietoryhmiin kuuluvaa henkilötietoa kun niistä selviää yksilöllistä tietoa kyseisen potilaan fysiologiasta tai terveydentilasta

muksia tietojärjestelmille. Standardi SFS-EN ISO/IEC 15189:2013 edellyttää ylläpitämään ohjelmistot ja laitteet valmistajan ohjeistuksen mukaisesti ja varmistamaan muutoksien vaikutukset suorituskykyyn ennen käyttöönottoa. Geenitiedon käsittelyyn tarkoitettujen ohjelmistojen kohdalla tulisi esimerkiksi määrittää, kuka seuraa päivitystiedotteita, miten ja missä päivityksien toimivuus tarkastetaan sekä ketkä vastaavat päivityksien asennuksesta. Standardi

myös velvoittaa toteuttamaan tietojärjestelmät voimassa olevien tietoturva- ja tietosuojalakien mukaisesti sekä varmistamaan toiminnan jatkuvuus häiriötilanteissa.

Kliinisen ja tutkimuksellisen toiminnan eroavaisuudet

Kliiniset ja tutkimustyöhön liittyvät sekvensointitutkimukset eroavat toisistaan tietyiltä

osin. Tieteelliset tutkimukset pyrkivät usein tuottamaan ennennäkemätöntä tietoa tutkittavasta ilmiöstä, mikä monesti onnistuu parhaiten uusimmilla menetelmillä, joita ei ole laajalti käyttöönotettu diagnostiikassa. Samasta syystä tutkimusasetelmissa tosin voidaan myös päätyä käyttämään menetelmiä, joiden soveltuvuutta tutkimuskohteeseen ei ole varmistettu. Eroa on myös mahdollisuuksista käyttää tutkimustyössä erikoisjärjestelyjä näyttemateriaalin keruussa. Tutkimushankkeissa toisin kuin diagnostiikassa voidaan esimerkiksi suunnitella kerättävän useita tuoreinäyteitä kasvaimesta ja viereisestä normaalikudoksesta. Tutkimustyössä usein myös käsitellään potilasjoukkoja, miksi systemaattiset ja satunnaiset tekijät voidaan erottaa tilastotieteen keinoin tulkintatyötä tehostaen. Tutkimustyössä usein myös sallitaan pidempi aikajänne, jolloin tutkimusnäytteet voidaan sekvensoida isoissa erissä ja bioinformatiiviset analyysit räätälöidä tavoitteita vastaavaksi.

Lopuksi

Laajat sekvensointitutkimukset tuottavat valtavia määriä geenitietoa. Tämä geenitieto on ollut avainasemassa sairauksien taustalla olevien geenivirheiden selvittämisessä ja luonut perusteet kliinisille geenitutkimuksille. Kasvaneiden tietomäärien myötä geenitutkimuksista on yhä enemmän myös tullut riippuvaisia niissä käytettävistä bioinformatiivisista ratkaisuista. Koska tulevaisuudessa DNA-sekvensointitutkimukset todennäköisesti tuottavat nykyistä runsaammin geenitietoa, riippuvuus bioinformatiikasta oletettavasti vahvistuu.

Genomitutkimukset ja alati kasvavat tietomäärät asettavat uusia suorituskykyvaatimuksia tulevaisuuden geenitietojärjestelmille. Oletettavasti tulevaisuudessa bioinformatiiviset analyysit tehdään yhä enemmän kiihdytetyillä (grafikka) prosessoreilla, jotka suoriutuvat minuuteissa analyyseista, joiden tekoon muuten kuluisi päiviä. Tiedon säilytyksessä todennäköisesti nojataan sairaalatietotaltisiin tai genomikeskukseen, mikä voi edistää myös geenitiedon tutkimuksellista käyttöä. Aiempaa laajemmin ja syvemmin luettujen näytteiden myötä yhä useammalle potilaalle voidaan myös

Ydinasiat

- » Bioinformatiikka muuntaa laajojen sekvensointitutkimuksien tuottaman tiedon hyödylliseen ja ihmisluettavaan muotoon.
- » Siinä missä pienet emäsmuutokset tunnistetaan laadukkaasti, rakennepoikkeamien tunnistamisessa on vielä työnsarkaa.
- » Kiihdytettyjen prosessorien avulla genominlaajuisetkin tietoaineistot voidaan prosessoida tunneissa päivien sijaan.
- » Tekoälypohjaisten emäsmuutosluokittelutyökalujen esiinmarssin myötä tulkintatyön tehokkuus voi kasvaa merkittävästi.

löytää hoidollisesti merkittävä tulos. Tällöin järjestelmällä on tosin pystyttävä analysoimaan laadukkaasti niin pieniä emäsmuutoksia kuin rakennepoikkeavuuksia.

Geenitiedon käsittelyn eräs suurimmista pullonkauloista on havaittujen muutoksien luokittelu sairauteen liittyviin ja harmittomiin. Ajan saatossa ongelma todennäköisesti ratkeaa tekoälypohjaisten luokittelutyökalujen avulla. Esimakua tulevasta tarjoavat ensimmäiset markkinoille jo tulleet tekoälypohjaiset ratkaisut, jotka muutamassa minuutissa luokittelevat genominlaajuiset ituradan mutaatiot jopa 93 %:n herkkyydellä (37). Kehitystyön edessä ohjelmiston herkkyys luultavasti parantuu ja vapauttaa näin arvokkaita resursseja tärkeämpään tehtävään. ■

MATTI KANKAINEN, FT

Genetiikan laboratorio, HUS Diagnostiikkakeskus
Hematologinen tutkimusyksikkö, Helsingin yliopisto

SOILI KYTÖLÄ, FT, Dos.

Genetiikan laboratorio, HUS Diagnostiikkakeskus

TEEMAN TOIMITTAJAT

Sampsa Hautaniemi ja Tuomas Mirtti

SIDONNAISUUDET

Matti Kankainen: Terveystieteiden tutkimuskeskukseen pyrkivät hankkeet (STM, yksilöllistetyn terveydenhuollon ja tutkimuksen kokonaisarkkitehtuurialaryhmän jäsen)

Soili Kytölä: Luentopalkkio/asiantuntijapalkkio (Novartis, Amgen, Astra Zeneca), korvaukset koulutus- ja kongressikuluista (Amgen, Incyte, Astra Zeneca)

KIRJALLISUUTTA

- Hou YC, Neidich JA, Duncavage EJ, ym. Clinical whole-genome sequencing in cancer diagnosis. *Hum Mutat* 2022;43:1519–30.
- Hu T, Chitnis N, Monos D, ym. Next-generation sequencing technologies: an overview. *Hum Immunol* 2021;82:801–11.
- McInnes G, Sharo AG, Koleske ML, ym. Opportunities and challenges for the computational interpretation of rare variation in clinically important genes. *Am J Hum Genet* 2021;108:535–48.
- Chandak S, Tatwawadi K, Ochoa I, ym. SPRING: a next-generation compressor for FASTQ data. *Bioinformatics* 2019;35:2674–6.
- Trapnell C, Salzberg SL. How to map billions of short reads onto genomes. *Nat Biotechnol* 2009;27:455–7.
- Souche E, Beltran S, Brosens E, ym. Recommendations for whole genome sequencing in diagnostics for rare diseases. *Eur J Hum Genet* 2022;30:1017–21.
- Aganezov S, Yan SM, Soto DC, ym. A complete reference genome improves analysis of human genetic variation. *Science* 2022;376:eabl3533.
- Olson ND, Wagner J, McDaniel J, ym. PrecisionFDA Truth Challenge V2: calling variants from short and long reads in difficult-to-map regions. *Cell Genom* 2022;2:100129.
- DePristo MA, Banks E, Poplin R, ym. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 2011;43:491–8.
- Mandelker D, Ceyhan-Birsoy O. Evolving significance of tumor-normal sequencing in cancer care. *Trends Cancer* 2020;6:31–9.
- Chen Z, Yuan Y, Chen X, ym. Systematic comparison of somatic variant calling performance among different sequencing depth and mutation frequency. *Sci Rep* 2020;10:3501.
- Collins RL, Brand H, Karczewski KJ, ym. A structural variation reference for medical and population genetics. *Nature* 2020;581:444–51.
- Kosugi S, Momozawa Y, Liu X, ym. Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biol* 2019;20:117.
- Koboldt DC. Best practices for variant calling in clinical sequencing. *Genome Med* 2020;12:91.
- Pellestor F, Gatinois V. Chromoanagenesis: a piece of the macroevolution scenario. *Mol Cytogenet* 2020;13:3.
- Bertelsen B, Nazaryan-Petersen L, Sun W, ym. A germline chromothripsis event stably segregating in 11 individuals through three generations. *Genet Med* 2016;18:494–500.
- Cortes-Ciriano I, Lee JJ, Xi R, ym. Comprehensive analysis of chromothripsis in 2,658 human cancers using whole-genome sequencing. *Nat Genet* 2020;52:331–41.
- Saarela J, Kettunen K. Kenelle eksomisekvensointi? *Duodecim* 2017;133:481–8.
- Paakkola T, Keskinen S, Pohjola P, ym. Laajojen geenitutkimusten sivi- ja sattumalöydökset. *Duodecim* 2022;138:2111–7.
- Richards S, Aziz N, Bale S, ym. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med* 2015;17:405–24.
- Houge G, Laner A, Cirak S, ym. Stepwise ABC system for classification of any type of genetic variant. *Eur J Hum Genet* 2022;30:150–9.
- Masson E, Zou WB, Genin E, ym. Expanding ACMG variant classification guidelines into a general framework. *Hum Genomics* 2022;16:31.
- Li Q, Wang K. InterVar: clinical interpretation of genetic variants by the 2015 ACMG-AMP guidelines. *Am J Hum Genet* 2017;100:267–80.
- Li MM, Datto M, Duncavage EJ, ym. Standards and guidelines for the interpretation and reporting of sequence variants in cancer: a joint consensus recommendation of the Association for Molecular Pathology, American Society of Clinical Oncology, and College of American Pathologists. *J Mol Diagn* 2017;19:4–23.
- Horak P, Griffith M, Danos AM, ym. Standards for the classification of pathogenicity of somatic variants in cancer (oncogenicity): joint recommendations of Clinical Genome Resource (ClinGen), Cancer Genomics Consortium (CGC), and Variant Interpretation for Cancer Consortium (VICC). *Genet Med* 2022;24:1991.
- Mateo J, Chakravarty D, Dienstmann R, ym. A framework to rank genomic alterations as targets for cancer precision medicine: the ESMO Scale for Clinical Actionability of molecular Targets (ESCAT). *Ann Oncol* 2018;29:1895–902.
- Chan TA, Yarchoan M, Jaffee E, ym. Development of tumor mutation burden as an immunotherapy biomarker: utility for the oncology clinic. *Ann Oncol* 2019;30:44–56.
- Bonneville R, Krook MA, Chen HZ, ym. Detection of microsatellite instability biomarkers via next-generation sequencing. *Methods Mol Biol* 2020;2055:119–32.
- Wagener-Rydzek S, Merkelbach-Bruse S, Siemanowski J. Biomarkers for homologous recombination deficiency in cancer. *J Pers Med* 2021;11:612.
- Alexandrov LB, Kim J, Haradhvala NJ, ym. The repertoire of mutational signatures in human cancer. *Nature* 2020;578:94–101.
- Islam SMA, Diaz-Gay M, Wu Y, ym. Uncovering novel mutational signatures by de novo extraction with SigProfilerExtractor. *Cell Genom* 2022;2:none.
- Miller DT, Lee K, Abul-Husn NS, ym. ACMG SF v3.1 list for reporting of secondary findings in clinical exome and genome sequencing: a policy statement of the American College of Medical Genetics and Genomics (ACMG). *Genet Med* 2022;24:1407–14.
- Green RC, Berg JS, Grody WW, ym. ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. *Genet Med* 2013;15:565–74.
- Carrieri D, Howard HC, Benjamin C, ym. Recontacting patients in clinical genetics services: recommendations of the European Society of Human Genetics. *Eur J Hum Genet* 2019;27:169–82.
- David KL, Best RG, Brenman LM, ym. Patient re-contact after revision of genomic test results: points to consider-a statement of the American College of Medical Genetics and Genomics (ACMG). *Genet Med* 2019;21:769–71.
- Loong L, Garrett A, Allen S, ym. Reclassification of clinically-detected sequence variants: framework for genetic clinicians and clinical scientists by CanVIG-UK (Cancer Variant Interpretation Group UK). *Genet Med* 2022;24:1867–77.
- Nostos Genomics. A clinical validation study of AION on the Genomics England 100,000 Genomes project. Nostos Genomics GmbH 2022.