Potilasnäytteistä nopea DNA-analyysi tekoälyn avulla

Ihmisen perimä sisältää miljoonia geneettisiä muunnelmia, variantteja, jotka tekevät jokaisesta yksilöstä ainutlaatuisen. Jotkut variantit vaikuttavat silmien väriin tai verityyppiin, toiset perinnöllisiin sairauksiin. DNA-sekvenssistä voi löytyä myös patogeeninen sekvenssivariantti, joka aiheuttaa geenin toiminnassa erilaisia häiriötä. Häiriöt ilmenevät perinnöllisinä sairauksina. Suomalainen Blueprint Genetics luokittelee potilasnäytteistä perimästä havaittuja geneettisiä variaatioita ja analysoi niiden yhteyden potilaista kuvattuihin oireisiin.

Blueprint Genetics aloitti toimintansa keskittymällä sydän- ja verisuonitautien diagnostiikkaan. Nyt yritys pystyy analysoimaan saamistaan potilasnäytteistä valtaosan perinnöllisistä sairauksista. Ihmisistä tunnetaan yli 6000 yhden geenin virheestä syntynyttä sairautta. Keskimäärin yksi kahdesta sadasta perii geenivirheen vanhemmiltaan. On myös paljon monitekijäisiä sairauksia, joissa useamman geenivariaation yhdistelmä aiheuttaa sairauden tai nostaa sairastumisriskiä. Sellaisia ovat esimerkiksi Alzheimer, diabetes, nivelreuma tai syöpäsairaudet.

Blueprint Geneticsin datatieteen johtaja ja Itä-Suomen yliopiston tutkija **Jussi Paananen** on taustaltaan tietojenkäsittelytieteilijä, joka on erikoistunut data-analytiikkaan. Paananen kiinnostui jo varhain biolääketieteestä, koska siinä hyödynnetään teknologioita, jotka tuottavat paljon dataa. Viime vuosina häntä on kiinnostanut koneoppiminen ja tekoäly, jotka ovat tulossa bioinformatiikan tutkimusmenetelmiksi kasvavan laskentatehon myötä.

”Minua kiinnostaa, miten tekoälyllä voidaan auttaa geneetikkoja päätöksenteossa ja isojen datamäärien käsittelyssä.”

Tekoäly auttaa varianttien tunnistamisessa

Tekoälyn tutkimus on kovassa kasvussa ja menetelmät muuttuvat. Koneoppimisessa tietokone oppii itsenäisesti päätymään tiettyyn lopputulokseen. Koneoppimisen algoritmit löytävät sellaisia säännönmukaisuuksia isoista aineistojoukoista, joita ihminen ei havaitse. Koneoppimisessa hyödynnetään neuroverkkotutkimusta, jossa Suomessa on pitkä perinne. Neuroverkko oppii muuttujien epälineaariset riippuvuussuhteet suoraan havaintoaineistosta. Se osaa esimerkiksi luokitella eläinaiheisista kuvista korvat.

Kun kromosomista häviää jokin osa puhutaan häviämästä eli deleetiosta. Tällöin kromosomi usein katkeaa kahdesta eri kohdasta, jolloin irronnut pala häviää. Tämän seurauksena myös osa geeneistä häviää, mikä aiheuttaa kehityshäiriöitä. Näkymä IGV (Integrative Genomics Viewer) -ohjelmasta, jossa geneetikon tarkasteltavana on RPGR-geenin deleetio ORF15-alueella. ORF15 on yksi RPGR-geenin osa. Se on käytännössä yksi eksoni, joka ohjaa RGPR-geenin proteiintuotantoa. Mutaatiot RPGR-geenissä aiheuttavat kaksi kolmasosaa X-kromosomaalisista verkkokalvonrappeumatapauksista. Kuvassa näkyvät väripalkit ovat potilasnäytteestä sekvensoituja nukleotidisekvenssejä. Väri ilmaisee, kummasta suunnasta DNA-molekyyliä on luettu. Potilasnäytteestä luettujen sekvenssien keskellä näkyy kahden nukleotidin pituinen deleetio.

”Kaikkein parhaimpia neuroverkot ovat juuri luokitteluongelmien ratkaisuissa”, sanoo Paananen.

”Kuva-analytiikassa kuvia tunnistetaan tai kuvista tunnistetaan osia ja niitä luokitellaan. Kone pystyy tunnistamaan esineitä ja asioita: tässä on ihminen, tässä auto, tässä syöpäkasvain. Se, mitä me teemme, on DNA-varianttien luokittelu. Yritämme löytää potilasnäytteistä, mitkä DNA-variantit aiheuttavat sairauksia ja mitkä geneettiset variaatiot ovat osa normaalia perimäämme.”

Geenimuunnos tunnistetaan eri lähteitä seulomalla

Blueprint Geneticsin asiakkaina ovat potilaita hoitavat lääkärit. Lääkärit haluavat selvittää, johtuvatko heidän potilaidensa sairaudet perinnöllistä tekijöistä vai eivät. Lääkärit eri puolilta maailmaa lähettävät Blueprint Geneticsille potilaidensa veritai sylkinäytteen, josta eristetty DNA sekvensoidaan. Sekvensointi tuottaa valtavan määrän dataa, joista poimitaan kiinnostavat variantit. Käytännössä se tarkoittaa, että potilaan geenimuunnoksia verrataan keskimääräiseen ihmisen referenssi-DNA:han.

Blueprint Geneticsin palveluksessa on huippuammattilaisia, geneetikkoja ja lääkäreitä, jotka luokittelevat variantteja. He käyvät läpi datamassaa, jota on jo käsitelty ja pilkottu pienempiin osiin. Asiantuntijat käyvät käytännössä läpi olemassa olevaa tieteellistä kirjallisuutta ja tietokantoja.

”Yritämme selvittää, mitkä näistä varianteista selittäisivät sairauden tai sen oireet”.

Koska vastaavaa informaatiota on kerätty ympäri maailmaa, usein tieteellisistä artikkeleista ja tietokannoista löytyy yksittäinen DNA-variantti, joka selittää sairauden.

”Teemme aineistosta kliinisen lausunnon joka lähetetään asiakaslääkärille. Lääkäri käyttää lausuntoa apuna diagnoosissa ja hoidon suunnittelussa.”

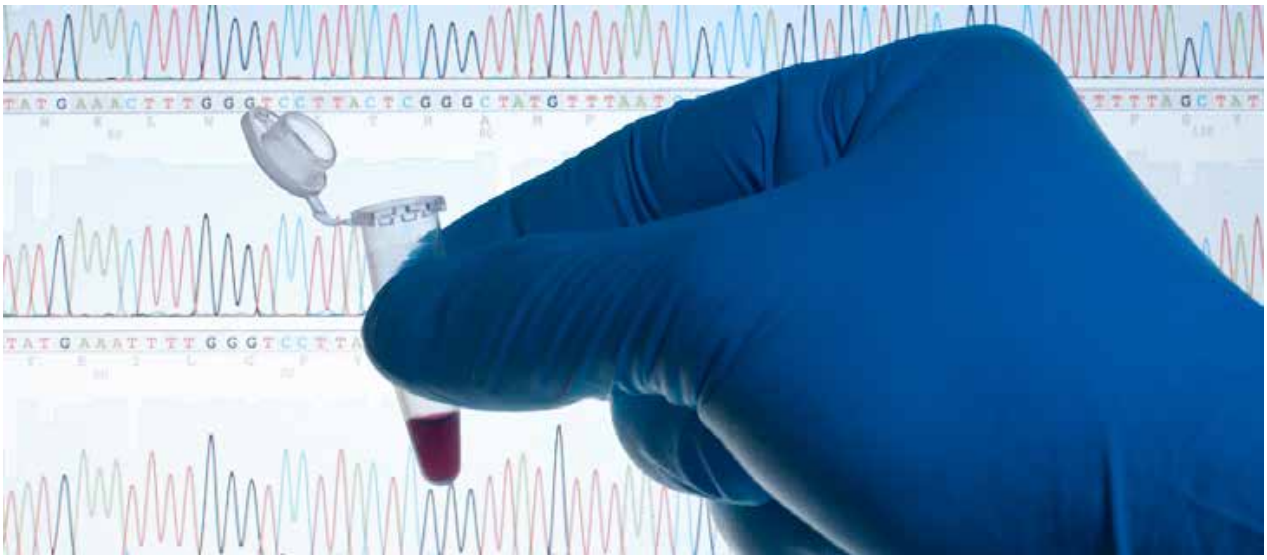
Blueprint Genetics hyödyntää erilaisia datalähteitä. Mahdollisuuksien mukaan data-aineiston analysointi automatisoidaan. Ohjelmistot analysoivat dataa ja tekevät monimutkaista datankäsittelyä. Ala on jatkuvassa kehityksessä. Ohjelmistoja päivitetään useita kertoja vuodessa, datamäärät ja laskentatehot kasvavat. Menetelmät kehittyvät ja muuttuvat nopeasti.

”Meillä on omaa ohjelmistotuotantoa, joka yhdistää eri datalähteitä ja helpottaa kirjallisuushakuja. Lopullisen tulkinnan tekee kuitenkin aina geneetikko. ”

Potilasdatan analysointi ja tulkitseminen on vaativaa työtä, koska siihen liittyy paljon lainsäädäntöä ja säätelyä. Blueprint Genetics tarjoaa lääkäreille käsiteltyä tietoa, mutta lääkärit tekevät aina varsinaisen päätöksen.

Blueprint Genetics on myös kiinnostunut myös julkisen- ja yksityissektorin välisestä yhteistyöstä.

”Geneettisen tiedon hyödyntämisessä on kyse koko ihmiskuntaa koskevasta valtavasta haasteesta. Ratkaisu vaatii yhteistoimintaa niin yrityksiltä, akateemisilta tutkimusryhmiltä kuin julkisrahoitteisilta järjestöiltä. Blueprint Genetics pyrkii osallistumaan avoimen tieteen ratkaisujen kehittämiseen ja etsii jatkuvasti uusia yhteistyötahoja.”



Blueprint Genetics ottaa vastaan veri- tai sylkinäytteen, josta saadusta DNA:sta etsitään mahdollisen taudin aiheuttama geenimuunnos. Analyysiin menee noin kolme viikkoa.

Geenimuunnosten luettelotietokannat tärkeitä

Alun perin Blueprint Genetics keskittyi potilaan oireiden perusteella tiettyihin kiinnostaviin geeneihin eli geenipaneeleihin. Paneelissa on tyypillisesti noin sata kappaletta tiettyyn tautiin liittyviä tunnettuja geenejä. Geneetikoista koostuva tiimi käy paneelin avulla tutkitut noin 2000 varianttia läpi. Nyt yritys on siirtynyt eksomisekvensointiin eli se sekvensoi kaikki proteiineja koodaavat geenit, joita meidän perimässämme on noin 21 000.

Ihmisen eksomi on se osa DNA:sta, jonka avulla tuotetaan kaikki ihmisen proteiinit. Sitä geenin osaa, joka koodaa ja suoraan ohjaa proteiinien tuotantoa kutsutaan eksoniksi. Kaikkia ihmisen eksoneita perimässämme kutsutaan kokonaisuudessa eksomiksi. Ihmisen eksomi noin 1,5% koko genomista.

”Kun analyysimme kohdistui geenipaneeleihin, saimme esimerkiksi 2000 varianttia, jota geneetikkojen tiimi kävi läpi. Nyt variantteja voi tulla 200 000. Kun olemme menossa koko genomin sekvensoitiin, variantteja saadaan 5 miljoonaa. Tätä datamäärää ei voida käsipelillä käydä läpi.”

Potilasnäytteistä kerätyn datan tulkinnassa ulkopuoliset tietokannat ovat tärkeitä. Genomin variaatioita on luetteloitu erilaisiin kansainvälisiin tietokantoihin, näistä tärkeimmät sijaitsevat eurooppalaisessa EMBL-EBI:ssä sekä yhdysvaltalaisessa NCBI (The National Center for Biotechnology Information) organisaatioissa. Lisäksi ELIXIR koordinoi Euroopassa julkisen biolääketieteen infrastruktuuria mahdollistaen geneettisten variaatioiden louhimisen näistä kansainvälisistä tietokannoista.

Varianttitietokannat tarjoavat hyödyllisiä luetteloita, joilla voidaan löytää korrelaatioita geenimuunnosten ja fenotyyppidatan välillä. EMBL-EBI luokittelee, tallentaa ja jakaa tietoa geenimuunnoksista. Tärkeimpiä tietokantoja ovat European Genome-phenome Archive (EGA), johon säilötään biolääketieteen tutkimuksen aineistot potilaista, European Variation Archive (EVA), joka sisältää geneettiset variaatiot, Ensembl tarjoaa näille variaatioille tulkinnan, gnomAD palvelu väestötasoiselle varianttien esiintymistiedolle sekä kliinisesti merkittävien varianttien varasto ClinVar. Lääkäri tarvitsee siis usein tiedon useammasta palvelusta, jotta oikea tulkinta genomivariaatiosta voidaan tuottaa potilasta varten. Tästä syystä eurooppalaiset ja amerikkalaiset palvelut vaihtavat säännöllisesti tietoa viimeisimmistä tutkimustuloksista, jotta palvelut tarjoaisivat aina viimeisimmän tiedon perimästämme tutkimuksen ja lääketieteen käyttöön.

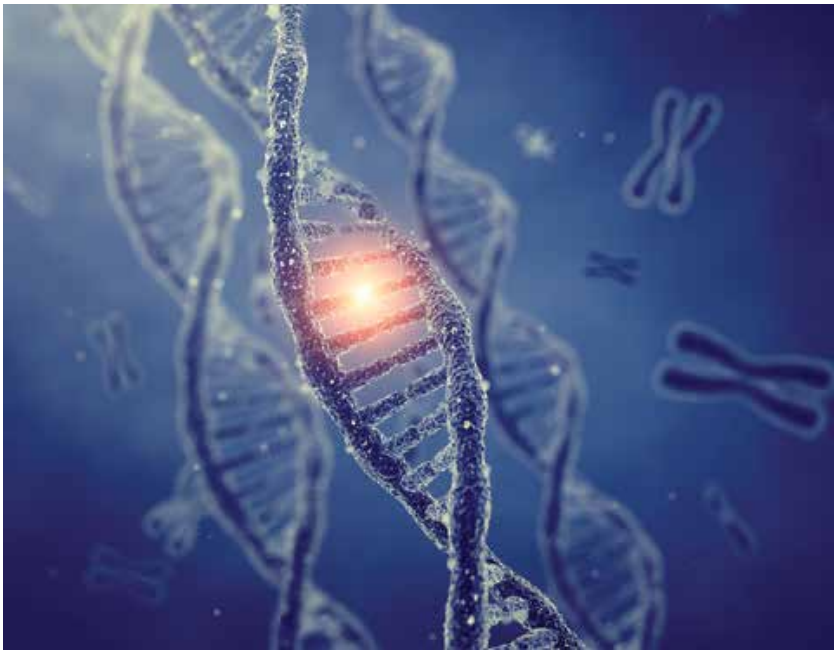
”Geenivarianttitietokannat ovat tärkeitä, koska sieltä löytyy tietoa varianttien yleisyydestä terveessä ihmisessä. Tätä tietoa voidaan käyttää hyödyksi esimerkiksi silloin, kun tiedetään, että tietty periytyvä harvinainen sairaus on vain yhdellä prosentilla ihmisistä. Kun nähdään, että siellä on variantti, joka on viidellä prosentilla ihmisistä, voidaan todeta, että tämä ei voi olla se tautia aiheuttava variantti. Voidaan siis suodattaa pois isoja yleisiä DNA-variantteja, jotka eivät voi liittyä tähän sairauteen.”

ELIXIRin tarjoamat julkisen sektorin datapalvelut ovat tärkeitä.

”Me hyödynnämme omia paikallisia kopioita eri datalähteistä. Fyysinen etäisyys ja tietoliikenneyhteydet vaativat, että lähteet ovat samassa paikassa. Julkisilla palveluilta toivoisin lisää toimenpiteitä tietokantojen versioimiseen liittyen. Vanhoja versioita ei pitäisi hävittää pois. Eri versioihin pitäisi tarjota pitkäaikaissäilytystä.”

Metadatan standardointi haasteellista

Iso haaste niin julkisissa tutkimusorganisaatioissa kuin yksityissektorillakin on



tulkintaan käytetyn datan standardointi. Datan merkintätavat voivat vaihdella suuresti. Blueprint Geneticsin iso haaste on ns. fenotyyppidata.

”Se on tavallaan metadata itsessään eli potilasnäytteen mukana tuleva informaatio: oireet, diagnoosi ja muut taustatiedot. Voi olla, että näytteen mukana saadaan paljon metadataa tai sitten sitä ei saada ollenkaan.”

Fenotyyppidatan standardoinnissa on sama ongelma kuin terveydenhuollon potilasdatassa, jossa haasteena ovat erilaiset merkintätavat.

”Meille tulee eri maista erilaisilla käytänteillä varustettua tietoa. Taustatieto vaihtelee.”

Blueprint Geneticsin tapaisten firmojen on Jussi Paanasen mielestä hankalaa hyödyntää julkisrahoitteisten ja tutkimuskeskeisten organisaatioiden tuottamaa ja hallinnoimaa dataa.

”Tutkimusorganisaatiot ja yhteiset infrastruktuurit ovat kiinnostuneita isoista väestökohorteista, jolloin kyse on valtavista datamäärästä, joita koetetaan harmonisoida. Me käsittelemme tietoa eri tavoin kuin kohorteissa, jossa vaikkapa kootaan kymmenien tuhansien samalla alueella asuvien ihmisten tietoa. Meillä on kyse kuitenkin aina yksilöistä.”

Blueprint Genetics pyrkii käyttämään kansainvälisesti yhdenmukaista luokittelua, terminologiaa ja standardeja toiminnassaan.

”Tuotamme itse DNA-datan ja voimme päättää missä muodossa ja missä standardeissa se on. Me kuitenkin hyödynnämme muiden tekemää ohjeistusta kun tulkitsemme tuloksia.”

Muutama vuosi sitten tuli ensimmäinen yritelmä tällaisesta standardista. Yhdysvaltalainen American College of Medical Genetics and Genomics (ACMG) on laatinut ohjeistuksen, miten sekvenssivariantteja voisi luokitella. ACMG on ehdottanut seuraavanlaista yhteistä terminologiaa yksittäisen geenin aiheuttamille sairauksille: patogeeninen, todennäköisesti patogeeninen (likely pathogenic), epävarma merkitys (uncertain significance), todennäköisesti hyvälaatuinen (likely benign) ja hyvälaatuinen (benign).

”Meillä on ACMG:n luokittelusta oma muokattu versio.”

Blueprint Geneticsin tapaisten yritysten haasteena on tiedon hyödynnettävyys. Tietoa on paljon referoiduissa julkaisuissa ja tavoitteena on kehittää hyviä tekstilouhintatyökaluja, jolloin artikkelien seulominen voitaisiin automatisoida.

”Pitäisi saada keskitetty pääsy kaikkiin julkaisuihin. Nyt on pitkään neuvoteltu akateemisten kustantajien kanssa lisenssimaksuista, jotka ovat korkeat.”

Ari Turunen

LISÄTIETOA:

https://blueprintgenetics.com
https://www.elixir-europe.org/platforms/data/core-data-resources
https://www.ebi.ac.uk/ena/
https://www.ebi.ac.uk/eva/
https://www.ensembl.org/
https://www.ncbi.nlm.nih.gov/clinvar/
http://gnomad.broadinstitute.org
https://www.ebi.ac.uk/dgva

CSC – Tieteen tietotekniikan keskus Oy

on valtion omistama, opetus- ja kulttuuriministeriön hallinnoima, voittoa tavoittelematon osakeyhtiö. CSC ylläpitää ja kehittää valtion omistamaa keskitettyä tietotekniikkainfrastruktuuria.

http://www.csc.fi
https://research.csc.fi/cloud-computing

ELIXIR

rakentaa infrastruktuurin bioalan tutkimuksen tueksi. Se yhdistää 21 Euroopan maan ja Euroopan molekyylibiologian laboratorion EMBL:n johtavat organisaatiot yhteiseksi biologisen informaation infrastruktuuriksi. Sen Suomen keskus on CSC – Tieteen tietotekniikan keskus Oy.

http://www.elixir-finland.org
http://www.elixir-europe.org

SUOMEN ELIXIR
Puh. +358 9 457 2821 ● e-mail: servicedesk@csc.fi
www.elixir-europe.org/about-us/who-we-are/nodes/finland

www.elixir-finland.org

ELIXIR PÄÄMAJA
EMBL-European Bioinformatics Institute
www.elixir-europe.org