Kuidas kaitsta geneetilisi andmeid?

Viimasel ajal on jõutud juhtudeni, kus inimesi on geeniandmete põhjal tuvastatud ka täielikku anonüümsust eeldavatel puhkudel. Ühe võimaluse, kuidas koguda ja töödelda andmeid inimeste privaatsust ohtu seadmata, on välja pakkunud Eesti teadlased.

Inimese, nagu ka loomade, lindude, taimede ja mikroobide pärilikkuse infot kannab DNA. Igal inimesel on unikaalne DNA järjestus, mis põhineb juhuslikul kombinatsioonil tema isa ja ema DNA-st, millele lisanduvad juhuslikult tekkinud täiesti uued mutatsioonid. Inimese DNA-s on ligikaudu 3200 miljonit nukleotiidi kahes kromosoomis ehk kaks korda 3,2 miljardit aluspaari. Laste ja vanemate ning õdede-vendade DNA on seega väga pikkades lõikudes ühesugune.

Põhilised erinevuste kohad on inimkonna DNA-s juba kinnistunud ja teada, nagu näiteks sagedamini esinevad ühe nukleotiidi variatsioonid (SNP, *single-nucleotide polymorphism*) või mõne piirkonna koopiaarvu muutused. Just nende positsioonide kaudu otsitakse geenivariantide seoseid haigusi tekitavate mutatsioonidega. Samas annavad need kombinatsioonid väga täpselt infot iga inimese pärinemise kohta.

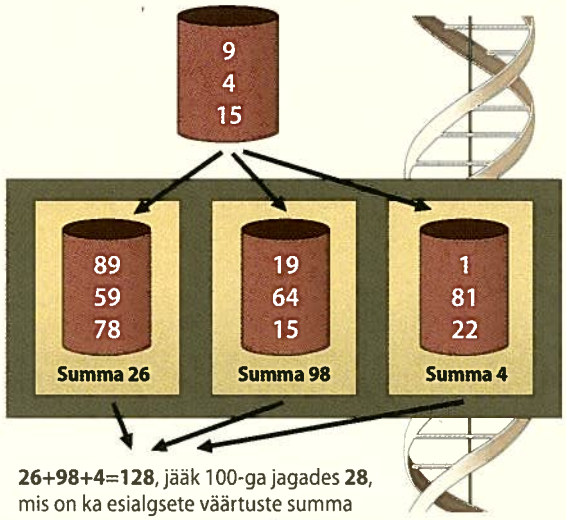
Siit tulenebki geneetika andmete suurim probleem. Esiteks on selge, et ilma arvutiteta ei saa analüüsida väga paljude inimeste omavahelisi miljoneid erinevusi, et otsida haigusi soodustavaid geenivariante. Samas on iga inimese DNA küll unikaalne, kuid siiski sarnane oma vanemate ja lähisugulaste DNA-ga. Piisavalt paljude DNA-s varieeruvate positsioonide kohta info teadmine võimaldab tuvastada isiku identiteeti. Seda kasutatakse ühelt poolt kriminalistikas. Kuid teisalt on jõutud olukorrani, kus inimesi on tuvastatud ka täielikku anonüümsust eeldavatel puhkudel. Näiteks muutub võimalikuks anonüümse spermadoonorist isa päritolu tuvastamine. Hiljuti avaldati teadusajakirjas Science artikkel, kus avalike sugupuu (näiteks *geni.com*) ja avalike, inimeste endi jagatud geneetilise info andmebaaside teavet kombineerides tuvastati anonüümsetest, teadusuuringuteks mõeldud DNA andmetest paljude isikute nimed. Lihtsustatult öeldes – meie DNA sisaldab infot meie nime kohta.

Selline olukord on äärmiselt keeruline. Ühelt poolt on hädavajalik analüüsida paljude inimeste DNA-d ja nende seoseid terviseandmetega (haigusi, ravimite kõrvalmõjusid, töökeskkonna ja tervise seoseid jne). Sest ainult nii saab leida seoseid inimese DNA ja tema tuleviku tervise ning ravi vajavatele isikutele personaalselt sobivate ravimeetodite vahel. Teiselt poolt seatakse geneetilise info kogumisega ja avalikustamisega aga ohtu isikute privaatsus. Isikuid on võimalik tuvastada isegi siis, kui tema enda DNA ei ole avalikus andmebaasis, kuid näiteks mõne tema lähisugulase oma on. Ka kriminalistikas piisaks, kui kuriteopaiga DNA-le leitakse piisavalt sarnane vaste inimese enda jagatud avalikest andmetest.

Teadusuuringuteks jääb justkui üle kaks varianti – lubada geneetilisi ja terviseandmete uuringuid läbi viia üksnes väga väikesel hoolikalt valitud usaldusväärsel teadlaskonnal, piirates oluliselt ligipääsu andmetele. Või pakkuda pikas perspektiivis välja viise, kuidas koguda ja analüüsida andmeid nii, et isegi mitte teadlased „ei näe" andmeid, küll aga saavad läbi viia vajalikke analüüse. See viimane tundub hetkel veidi vastuolulise nõudena. Kuid ometi on just Eesti noored teadlased jõudnud uue teetähiseni ja näidanud, kuidas see oleks võimalik.

Meie töörühm koosseisus Liina Kamm, Dan Bogdanov, Sven Laur ja Jaak Vilo pakkus välja, kuidas saaks kasutada turvaliste ühisarvutuste platvormi, et just selliseid analüüse läbi viia. Tarkvara Tehnoloogia Arenduskeskuse (STACC.ee), Cybernetica AS-i ja Tartu Ülikooli koostöös valmis uurimistöö, kuidas viia läbi geneetilisi assotsiatsiooniuuringuid, kasutades Sharemind-platvormi. Sellisel juhul töötavad omavahel koos kolm osapoolt, näiteks riiklik andmeturbe agentuur, patsientide usalduskogu ja tervishoiuasutus. Igaüks saab andmete kohta täiesti juhuslikuna näiva versiooni. Isegi juhul, kui kaks osapoolt „kaotavad" oma andmed või need varastatakse, siis nende põhjal ei saa taastada esialgseid päris väärtusi. Seega andmeturve on tagatud niikaua, kui vähemalt üks osapool on oma andmeid suutnud piisavalt kaitsta. Andmete analüüsiks kasutatakse aga krüptograafilisi protokolle, nii et arvutusi teostatakse iga osapoole juhuslikena näivate väärtuste peal, kuid õige lõpptulemus saadakse kolme osapoole vahetulemuste kombineerimisel.

Joonisel on toodud selle printsiibi väga väike näide. Samamoodi saab esitada kõikide geenivariantide väärtusi. Ühe nukleotiidi variatsioonide puhul alleelide väärtust, kas vastav nukleotiid on 2, 1 või 0 koopias, ehk kahel, ühel või mitte kummalgi kromosoomil. Selleks, et lugeda kokku, kui mitmel isikul, kellel on mingi diagnoos, esineb vastav geenivariant, saab kirjutada Sharemind'i jaoks vastava programmi, mis isikute algseid andmeid kusagil, isegi mitte programmi sees, välja ei arvuta. Praktiliste eksperimentidega näitasidki Liina Kamm ja Dan Bogdanov, et olulisemad geenivariantide tuvastamise algoritmid on võimalik realiseerida nii, et need töötavad vaid mõned korrad aeglasemalt kui senised, avatud andmete peal töötavad programmid. Igal juhul on tegu ka praktikas kasutatava lahendusega.



Sharemind'i põhiidee (jääk 100). **Näide väärtuste (9, 4, 15) jaotamisest kolmele sõltumatule andmebaasile ja summeerimisest. Kõik andmebaasid sisaldavad juhuslikke väärtusi, nii et kolme arvu summa on esialgne väärtus (kasutades 100-ga jagamise jääki). Näiteks 89+19+1=109, jääk 100-ga jagades 9. Summeerimiseks ei pea algseid väärtusi kusagil kasutama ja ükski kolmest osapoolest nende kohta midagi „ei lekita".**

• **Liina Kamm, Dan Bogdanov, Sven Laur, Jaak Vilo**

LOE VEEL

- Alison Motluk. Anonymous sperm donor traced on Internet. – New Scientist, November 03, 2005.
- Melissa Gymrek *et al.* Identifying Personal Genomes by Surname Inference. – Science, January 18, 2013.
- Liina Kamm, Dan Bogdanov, Sven Laur, Jaak Vilo. A new way to protect privacy in large-scale genome-wide association studies. – Bioinformatics. First published online: February 14, 2013.