6 Kokkuvõte

Käesoleva magistritöö eesmärk oli leida võimalusi, kuidas anda TÜ Geenivaramu doonoritele tagasisidet nende päritolu kohta nii rahvuse tasandil kui ka Eesti-siseselt. Saadud tulemused näitasid, et välja pakutud võimalused päritolu ennustamiseks annavad enamasti hästi interpreteeritavaid ja loogilisi tulemusi, kuid kindlasti saaks prognoose veelgi enam täpsustada. Töö käigus lähtuti analüüside tegemisel peakomponentidest ning üks suuremaid küsimusi töö käigus oli sobiva peakomponentide arvu valik korrektseima ennustuse saamiseks.

Rahvuse klassifitseerimisel ning vastavate klassitõenäosuste leidmisel saavutas täpseima tulemuse lineaarne diskriminantanalüüs. Leidmaks korrektseks tõenäosuse arvutamiseks sobivat peakomponentide arvu, teostati simulatsioonikatse, milles võrreldi arvatavasti teadaoleva päritoluga genotüübi tõenäosuseid vastavate tõenäosuste prognoosidega sellele genotüübile. Paraku simulatsioonikatse vastus ei olnud ühene ning lõplikku vastust peakomponentide arvu valikuks ei selgunud: põhjendatuks saab lugeda nii väikse kui ka suure peakomponentide arvu kasutamist ning teema vajab täpsemat uurimist. Samas annavad mõlemad lähenemised loogilisi vastuseid ning saadud tulemust halvaks pidada ei saa.

Eesti-sisesel prognoosimisel kontrolliti esialgu võimalust klassifitseerida maakondade alusel. Sealjuures saavutas täpseima tulemuse lineaarne diskriminantanalüüs, kuid mitmete Sise-Eesti maakondade puhul oli klassifitseerimistäpsus madal. Seetõttu otsustati alternatiivina leida K-keskmiste klasterdamise abil uued klassid, mis võiksid moodustada loogilisemaid ja täpsemaid kogumeid, milledesse inimesi liigitada. Leiti seos klasterdamisel kasutatavate peakomponentide arvu ja klastrite arvu vahel ning klastrite arvudena pakuti *gap*-statistiku alusel välja 3, 7, 12 ja 18, mis kirjeldavad erineva detailsusega tekkivaid kogumeid. Leitud klastrite alusel testisikute klassifitseerimine andis häid tulemusi ning enamasti klassifitseeriti testinimesed samasse klastrisse või naaberklastrisse. Klassifitseerimismeetoditena kasutati 3, 7, 12 puhul tugivektormasinaid ja 18 klastri puhul lineaarset diskriminantanalüüsi.

Viimaseks anti näide tagasisidest kahe inimese põhjal, kelle päritolu on suuresti teada. Nende puhul oli näha, et tõenäosusliku hinnangu andmine rahvusele võib olla vägagi varieeruv sõltuvalt valitud peakomponentide arvust ning õigustatud on anda kaks hinnangut. Teisalt olid Eesti-sisesed prognoosid pigem täpsed. Nii maakonna alusel kui ka klastrite alusel klassifitseerides jõuti enamasti ootuspäraste tulemusteni.

Suurim küsimus edasise uurimise osas on ilmselt rahvusprognooside korrigeerimine. Üks osa sellest on kindlasti peakomponentide edasine uurimine ning võimaluste leidmine sobiva peakomponentide arvu valikuks. Teisalt on võimalik, et see pole piisav, sest ka käesolevas töös on näidatud, et sobivat kompromissi peakomponentide arvu valikuks ei pruugigi leiduda. Hoopis tulemuslikum võib olla referentsvalimite korrigeerimine. Ka töö käigus selgus, et vene referentspopulatsiooni on vaja kindlasti täiendada ning pole täielikult selge, kui hästi kajastab töös kasutatud vene referentsvalim varieeruvust venelaste seas. Lisaks tuleks kindlasti kaaluda Geenivaramu olemasolevate doonorite baasil ka valgevenelaste ning ukrainlaste referentsgruppide moodustamist ning ühtlasi võib saada Geenivaramu andmetest täiendust lätlaste ja soomlaste referentspopulatsioonidele.

Tasub mõelda ka selle peale, et rahvuse määramise referentsandmestikust eemaldada Põhja-Soome vaatlused, mis on ülejäänud referentsidega võrreldes

selgelt erandlikud. Selle grupi eemaldamisega tekib võimalus, et peakomponentanalüüs suudaks eristada paremini ülejäänud eurooplaste vahel eksisteerivat varieeruvust ning seeläbi jõuda klassifitseerimisel paremate tulemusteni. Teatav analoogia on olemas Eesti-sisese päritolu määramisega, kus peakomponendid on leitud vaid Geenivaramu andmestest ning nende peakomponentide kasutamine annab Eesti-siseselt sisukaid tulemusi.

Seega, edasise päritolu uurimise seisukohalt on oluline tagada, et referentsvalimid oleksid piisavalt suured ning samaaegselt esinduslikud. Eriti oluline on see rahvuse määramise kontekstis, kuid näiteks referentsvalimi suurendamine omab kindlasti positiivset mõju ka Eesti-sisese päritolu uurimise kontekstis, kus oleks seeläbi võimalik klastreid paremini defineerida. Aastal 2018 viiakse läbi 100 000 täiendava geenidoonori andmete kogumine Geenivaramusse. Kindlasti aitab ka see andmemahu suurenemine kaasa paremate referentside välja töötamisele.

Käesolevas magistritöös anti ülevaade meetoditest ning võimalustest, mille abil anda tagasisidet päritolu kohta. Tulemused näitavad, et töös pakutud võimaluste abil saab esialgsel kujul anda TÜ Geenivaramu doonoritele tagasisidet. Töös välja pakutud ideid kasutades saab meetodeid kindlasti veel parandada, et lõppkokkuvõttes tagada geenidoonoritele võimalikult täpne tagasiside.