TARTU ÜLIKOOL
LOODUS- JA TEHNOLOOGIATEADUSKOND
MOLEKULAAR- JA RAKUBIOLOOGIA INSTITUUT
EVOLUTSIOONILISE BIOLOOGIA ÕPPETOOL

Kai Tätte

Austroaasia keeli kõnelevate India rahvaste geneetiline päritolu

Magistritöö

Juhendaja: PhD Mait Metspalu

TARTU 2015

SISUKORD

KASUTATUD LÜHENDID

ANI – ingl *Ancestral North Indians*, iidsed põhja-indialased

ASI – ingl *Ancestral South Indians,* iidsed lõuna-indialased

IBD – ingl *identity by descent*, sama järjestusega DNA lõik kahel või rohkemal inimesel, mis on päritud ühiselt eellaselt ilma et see oleks rekombinatsiooni käigus katkenud

LD – ingl *linkage disequilibrium*, aheldustasakaalutus on alleelidevahelise sõltumatuse puudumine

NRY – ingl *non-recombining region of the Y-chromosome,* 95% Y-kromosoomist, mis meioosi käigus ei rekombineeru X-kromosoomiga

SNP – ingl *single nucleotide polymorphism*, üksiknukleotiidne polümorfism

STR – ingl *short tandem repeat*, lühikesed kordusjärjestused DNA-s

KASUTATUD MÕISTED

autohtoonne – millegi tänapäevane esinemine samas paigas, kus ta kunagi moodustus

autosoom – mittesugukromosoom, esineb rakus kahes korduses kõigil isenditel sõltumata soost

eksogaamia – tava, mis keelab sama sugulusrühma liikmete omavahelise abielu

endogaamia – sotsiaalne norm, mille järgi abielud sõlmitakse vaid sama sotsiaalse rühma (nt kasti) liikmete vahel

faasitud andmed – andmed, mille puhul on kindlaks tehtud, millised alleelid on ühelt vanemalt pärandunud ja millised teiselt

geenitriiv – alleelide sageduste juhuslikud muutused populatsiooni järjestikustes põlvkondades juhuvaliku tõttu

geenivool ehk geenisiire – geneetilise materjali vahetus populatsioonide või populatsiooni allosade vahel isendite migratsiooni ja ristumise teel

haplotüüp – alleelide kombinatsioon, mis pärandub üheskoos järglasele

haplogrupp – sarnaste haplotüüpide, mis omavad ühist esivanemat, kogum või ka alamhaplogruppide, mis kuuluvad ühte klaadi, kogum

patrilokaalsus – tava, mille puhul naine kolib abielludes mehe juurde elama

varieeruvate saitide kallutatus – ingl SNP *ascertainment bias*, genotüpiseerimisplatvormile valitud varieeruvad saidid ei ole juhuvalim kõigist varieeruvatest saitidest, vaid mingi populatsiooni seas levinud varieeruvatest saitidest

SISSEJUHATUS

Lõuna-Aasia on koduks rohkem kui miljardile inimesele, kes kuuluvad tuhandetesse kultuuri, keele, eluviisi ja ka rahva väljanägemise poolest eristuvatesse gruppidesse. Sellise mitmekesisuse tõttu on Lõuna-Aasia huvitavaks piirkonnaks populatsioonigeneetiliste uuringute läbiviimiseks. Mitmete mitokondriaalsete uuringute põhjal on leitud, et Lõuna-Aasia oli esimene piirkond, kus pärast Aafrikast välja rändamist toimus kiire populatsiooni kasv ja mitmekesistumine (Atkinson *et al.*, 2008; Basu *et al.*, 2003; Kivisild *et al.*, 2003; Macaulay *et al.*, 2005; Metspalu *et al.*, 2004; Palanichamy *et al.*, 2004). Lõuna-Aasia rahvaste uurimise muudab oluliseks ning huvitavaks ka fakt, et piirkond on geneetilise mitmekesisuse poolest Aafrika järel teisel kohal (Xing *et al.*, 2010).

Lõuna-Aasiast suurema osa moodustab India. Vaatamata suurele mitmekesisusele viitab peaaegu kõigi India rahvaste geneetiline profiil kahe esivanemliku populatsiooni segunemisele (Metspalu *et al.*, 2011; Reich *et al.*, 2009). Üheks erandiks on Indias elavad munda keelte kõnelejad. Tegu on peamiselt Kagu-Aasias levinud austroaasia keelkonna ühe Indias leiduva haruga. Antropoloogide hulgas on olnud erinevaid hüpoteese mundade päritolu kohta alates sellest, et nad on India esmaasustajad kuni selleni, et nad on ühed viimastest sisserändajatest (Basu *et al.*, 2003; Majumder, 2001; Tamang & Thangaraj, 2012).

Antud magistritöö praktilises osas püütakse selgeks teha, kes mundad geneetilises mõttes on. Täpsemalt otsitakse vastust küsimustele, millised Lõuna- ja Kagu-Aasia rahvad on mundadele sarnaseimad ning millal toimus populatsioonide segunemine, mille tulemusena tekkis mundade populatsioon.

1. KIRJANDUSE ÜLEVAADE

1.1. Ülevaade Lõuna-Aasia rahvastikust

Lõuna-Aasia on geograafiline mõiste, mis viitab Aasia maailmajao lõunaosale. Läänest ümbritseb Lõuna-Aasiat Iraani platoo, põhjast ja idast Himaalaja mäestik ning lõunast India ookean. Riikidest katavad suurema osa Lõuna-Aasiast India, Pakistan ja Bangladesh. Koosseisu kuuluvad ka Himaalaja mägedes asuvad Nepal ja Bhutan ning India ookeani saareriigid Sri Lanka ja Maldiivid.

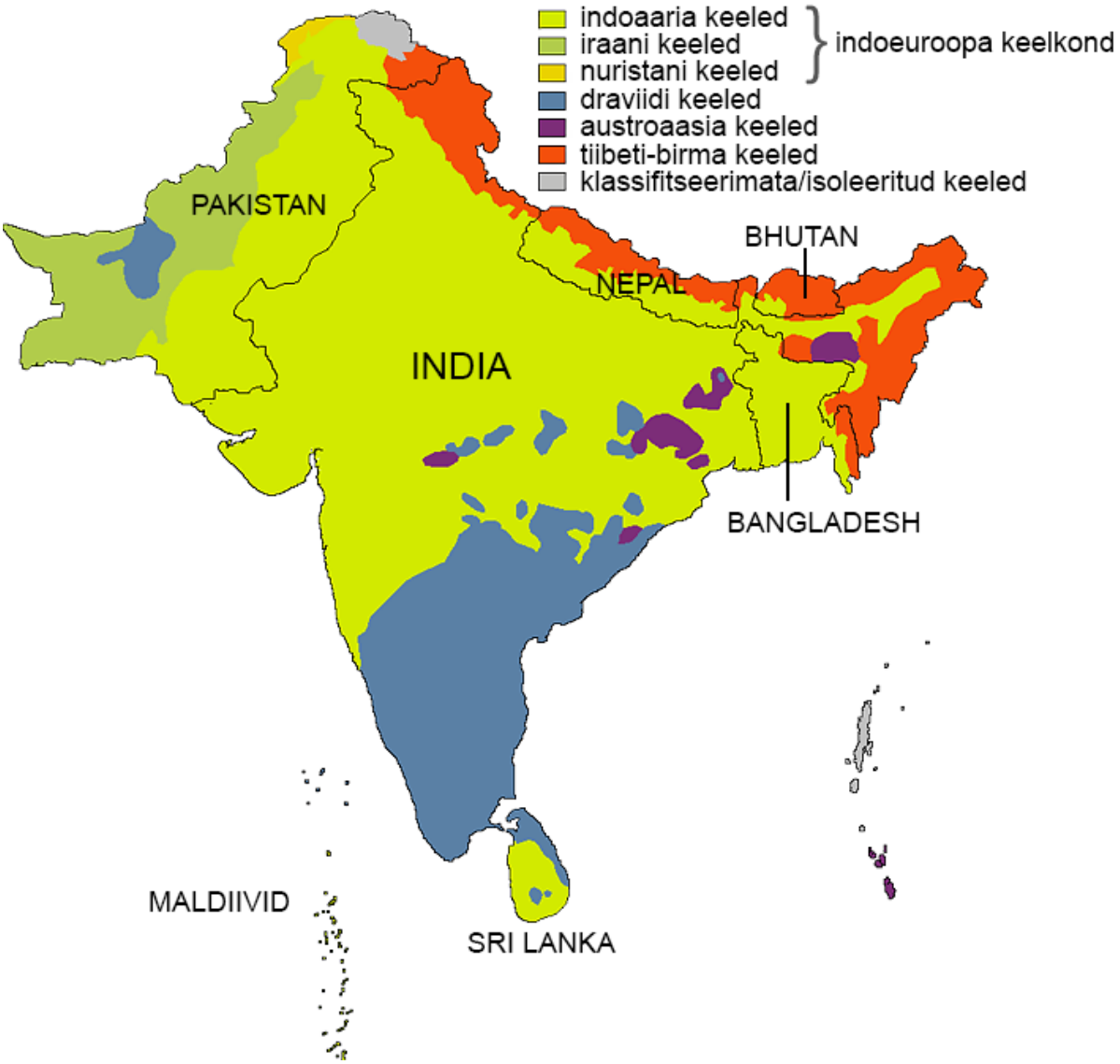
Lõuna-Aasias elab üle viiendiku maailma rahvastikust (esa.un.org/unpd/wpp/unpp/panel_population.htm). Vaatamata tihedale asustusele on sealsed rahvad kultuuriliselt ja keeleliselt väga mitmekesised. Indias kõneldavad keeled jagunevad nelja keelkonda, kuid keeli on seal sadu ja murdeid tuhandeid (www.ethnologue.com). Ka religioosne varieeruvus on Lõuna-Aasias suur. Näiteks Indias on kõige levinumad religioonid hinduism, islam, kristlus, sikhism, budism ja džainism (www.censusindia.gov.in/Census_Data_2001/India_at_glance/religion.aspx), kuid nagu igast keelest räägitakse erinevaid dialekte, siis ka religioossete rituaalide ja tavade järgimine on varieeruv iga religiooni siseselt (Chaubey, 2010). Selline suur ning mitmetasandiline varieeruvus on põhjustatud paljude faktorite koosmõjust. Esiteks piiravad populatsioonide segunemist geograafilised barjäärid nagu kõrbed ja mäestikud (Field *et al.*, 2007). Teiseks on juba ainuüksi viimastest aastatuhandetest ehk kirjapandud ajaloo algusest teada mitmeid sissetunge lähemate ja kaugemate naabrite poolt (Bamshad *et al.*, 2001; Diamond & Bellwood, 2003). Sellised invasioonid on mõjutanud erinevate populatsioonide keeli ja kultuure, aga ka geneetilist materjali, ja seeläbi mitmekesistanud kogu Lõuna-Aasia rahvastikku. Veel üks põhjus, miks me siiani Lõuna-Aasias nii palju geneetiliselt erinevaid populatsioone näeme (Watkins *et al.*, 2008), seisneb sealsete rahvaste ranges reeglistikus abikaasa valikul, mis tekitab sotsiaalseid barjääre endogaamia näol. Endogaamia on sotsiaalne norm, mille järgi abielusid sõlmitakse vaid sama sotsiaalse rühma liikmete vahel. Selline käitumine on võimaldanud ka tihedalt asustatud piirkondades erinevatel inimgruppidel säilitada oma kultuuri, keele ja geneetilised eripärad. Lõuna-Aasia riikides on sellisteks sotsiaalseteks rühmadeks kastid ja hõimud (Chaubey, 2010). Igas riigis on kastisüsteem natuke erinev. Indias on 4635 selgelt eristatavat populatsiooni, millest 532 on hõimud, kusjuures 72 hõimu puhul on tegu küttide-korilastega (www.censusindia.gov.in/). Peamisi kaste, millega ajalooliselt on määratud inimese tegevusvaldkond, on neli. Iga kasti sees on veel palju endogaamseid alamkaste ning nende sees omakorda eksogaamseid üksusi, mida nimetatakse gotrateks. See tähendab, et abikaasa valikul tuleb lisaks muudele reeglitele arvestada, et kaasa kuuluks

samasse alamkasti, kuid erinevasse gotrasse. Hõimude seas pole endogaamia nii range (Chaubey, 2010).

Keeruliseks muudab Lõuna-Aasia demograafilise ajaloo uurimise see, et ühine või suguluses olev keel ei tähenda ilmtingimata Lõuna-Aasia populatsioonide puhul suuremat geneetilist sarnasust või ühist geneetilist päritolu võrreldes erinevaid keeli rääkivate rahvastega (Chaubey *et al.*, 2008b; Sharma *et al.*, 2012). Samuti ei tähenda erinevate populatsioonide geneetilist sarnasust sama kultuurikombestiku või religiooni järgimine (Kivisild *et al.*, 2003; Metspalu *et al.*, 2004). Geenide, keele ja kultuuri vahelist korrelatsiooni muudab nõrgemaks näiteks erinevate sisserändajate erinev mõju. Mõned toovad ainult kultuuri ja keele (Eaaswarkhanth *et al.*, 2009; Eaaswarkhanth *et al.*, 2010), teised võivad aga muuta ainult genofondi (Narang *et al.*, 2011; Shah *et al.*, 2011). Mõned populatsioonid eristuvad teistest peamiselt vaid isaliini pidi kanduvate markerite poolest, mis viitab soospetsiifilistele rännetele (Chaubey *et al.*, 2011; McElreavey & Quintana-Murci, 2005). Geneetilist pilti on segasemaks muutnud ka see, et ajalooliselt on populatsioonid korduvalt lõhestunud mitmeks populatsiooniks (Basu *et al.*, 2003). Kui aga üks neist populatsioonidest on esialgu väga väike, siis on pudelikaelaefektil ja geenitriivil tugev mõju geenialleelide sagedustele.

1.1.1. Lõuna-Aasia keeled

Lõuna-Aasia keeled kuuluvad indoeuroopa, draviidi, austroaasia ja tiibeti-birma keelkondadesse (joonis 1), millest esimene on kõige levinum nii Lõuna-Aasias kui ka Maal üldiselt. Draviidi keeli kõnelevatest rahvastest elab enamik India lõunaosas, kus arheobotaaniku Dorian Fulleri (2003) mudeli kohaselt see keelkond ka tekkis. Austroaasia keeli kõnelevad hõimud elavad India kesk- ja idaosas väikeste gruppidena üksteisest teiste keelkondade keeli kõnelevate rahvaste poolt eraldatuna. Peamine austroaasia keelte kõnelejaskond elab Kagu-Aasias. Ida-Aasias levinud tiibeti-birma keelkonna serv ulatub Tiibeti ja Myanmari poolt Lõuna-Aasiasse. Indias on indoeuroopa ja suures osas ka draviidi keeli kõnelevate rahvaste seas levinud kastisüsteem, ülejäänud keelkondade rahvad elavad peamiselt hõimudena. Endogaamse eluviisi tõttu on erinevad sotsiaalsed grupid suutnud säilitada oma keele ning seetõttu elavad Indias tänapäeval 447 erineva keele kõnelejad (www.ethnologue.com/country/IN). Sealhulgas esineb ka isoleeritud keeli, mida ei suudeta klassifitseerida, näiteks nihali keel Lääne-Indias ja mõned Andamani saarte keeled (Moseley, 2008).



Joonis 1. Lõuna-Aasias levinud keelegrupid (Allikas: BishkekRocks, Wikimedia Commons [CC-BY-SA-3.0], modifitseeritud)

Indias on avastatud mõned juhud, kus on esinenud keelevahetus (Chaubey *et al.*, 2008b; Sharma *et al.*, 2012). See on nähtus, mille puhul populatsioon läheb üle uuele keelele, ilma et see nende geneetilist profiili märkimisväärselt mõjutaks. Põhjuseks võib olla näiteks soov suhelda paremal järjel oleva naaberpopulatsiooniga. Üks keelevahetuse läbinud hõim on India idaosas elavad musharid, kelle Y-kromosoomi ja mitokondriaalse DNA haplogruppide esinemissagedused on sarnased läheduses asuvate austroaasia (munda) keelte kõnelejate haplogruppidega, kuid kes ise räägivad indoeuroopa keelkonda kuuluvat keelt (Chaubey *et al.*, 2008b). Keelevahetus on toimunud ka Kesk-Indias asuvas hõimus bharia, mille liikmed praeguseks räägivad draviidi keelkonda kuuluvat keelt. Võrdlusel lähedal asuvate indoeuroopa ja austroaasia keelte kõnelejatega selgus, et nii uniparentaalsete kui ka biparentaalsete geneetiliste markerite poolest on bharia hõimu geneetiline profiil väga sarnane austroaasia keelte kõnelejatega (Sharma *et al.*, 2012).

1.2. Lõuna-Aasia paleoantropoloogiline ja arheoloogiline ajalugu

Rohkelt arheoloogilisi tõendusmaterjale viitab sellele, et varased hominiinid (ld *Homininae)* elasid Lõuna-Aasias juba enne anatoomiliselt moodsat inimest (Paddayya *et al.*, 2002; Patnaik *et al.*, 2009). Lõuna-Aasia ainus hominiini fossiil, mis ei kuulu anatoomiliselt moodsale inimesele, leiti Kesk-Indiast Narmada jõe orust 1982. aastal (Sonakia & Kennedy, 1985). Tegu on koljulaega (ld *calvaria*), mis arvatakse olevat vähemalt 236 tuhat aastat vana (Cameron *et al.*, 2004). Erinevad teadlased on klassifitseerinud Narmada pealuud nii Aasia püstiseks inimeseks (ld *Homo erectus*) (Sonakia & Kennedy, 1985), kui ka Heidelbergi inimese (ld *H. heidelbergensis*) ja neandertallase (ld *H. neanderthalensis*) vahepealseks liiniks (Cameron *et al.*, 2004). Athreya (2007) leidis, et kui vaadata kolju morfoloogiat, on tegu püstise inimesega, kuid aju suuruse poolest meenutab kolju pigem Heidelbergi inimest. Arvestades, et taksonite vahelised seosed on pidevad, mitte diskreetsed, soovitas Athreya Narmada hominiini kutsuda lihtsalt Kesk-Pleistotseeni inimeseks. Inimeste demograafilisele ajaloole mõeldes tuleb silmas pidada ~74 tuhat aastat tagasi toimunud Toba vulkaanipurset Sumatra saarel, mis kattis 15 cm paksuse tuhakihiga kogu Lõuna-Aasia (Oppenheimer, 2002). Ambrose (1998) on välja käinud teooria, et vulkaanipurske põhjustatud 5-6 aastane vulkaaniline talv ning 1000-aastane jahe periood tekitasid tollastes *Homo* populatsioonides tugeva pudelikaela, mis pani aluse erinevate tänapäevaste populatsioonide tekkele.

Mitokondriaalse DNA andmete põhjal arvatakse, et anatoomiliselt moodne inimene asustas Lõuna-Aasia kuni 65 tuhat aastat tagasi, vahetult pärast Aafrikast välja rändamist (Atkinson *et al.*, 2008; Macaulay *et al.*, 2005; Mellars *et al.*, 2013), kuid inimese säilmeid sellest ajast leitud ei ole. Pärast viimase jääaja maksimumi on veetase tõusnud ja arvatavasti jääb kunagine inimeste rändetee Aafrikast Lõuna-Aasiasse praegu vee alla (Field *et al.*, 2007). On ka võimalik, et paljud luud pole säilinud ebasobiliku kliima tõttu. Küll aga on Lõuna-Aasiast leitud palju kiviaja tööriistu (Misra, 2001). Lõuna-Aasia varaseimad kivist tööriistad on leitud Pakistani põhjaosast ning paleomagnetismi abil on nende vanuseks hinnatud ~2 miljonit aastat (Dennell *et al.*, 1988). Erinevate kiviaegade tööriistu on leitud kõikjalt Lõuna-Aasiast. Vanema paleoliitikumi Acheuli kultuur sai alguse juba 2–0,7 miljonit aastat tagasi. Alates 150 tuhat aastat tagasi tekkis ka keskmise paleoliitikumi kultuur, mille tööriistad olid juba arenenumad, kuid Acheuli kultuur ei kadunud sel hetkel, vaid kestis paralleelselt edasi. Keskmise paleoliitikumi kultuur on Lõuna-Aasias edasi arenenud noorema paleoliitikumi kultuuriks, mis tööriistade poolest erineb Lääne-Euraasia nooremast paleoliitikumist (James & Petraglia, 2005; Misra, 2001). Noorema paleoliitikumi tööriistad hakkasid peamiselt ilmuma 30 tuhat aastat tagasi (Misra, 2001), kuid vanimate selle kultuuri tööriistade, mis on leitud Pakistainst,

vanuseks on termoluminestsentsi abil hinnatud 45 tuhat aastat (Dennell *et al.*, 1992). Lisaks on leitud mitmeid loomi ja inimesi kujutavaid koopamaalinguid noorema paleoliitikumi ja mesoliitikumi-aegsete inimeste elukohtadest (Misra, 2001).

Vanimad nüüdisinimese jäänused Lõuna-Aasias on leitud kahest Sri Lanka koopast ning nende vanusteks on hinnatud 31 ja 28,5 tuhat aastat (James & Petraglia, 2005; Kennedy & Deraniyagala, 1989). Vaatamata sellele, et 35–28 tuhat aastat vanu inimese säilmeid pole Lõuna-Aasiast rohkem leitud, arvatakse mitokondri haplogruppide mitmekesistumise põhjal, et inimpopulatsioon kasvas seal sel ajal kiiresti. Sellega on kooskõlas ka tõsiasi, et 35–30 tuhat aastat tagasi hakkas levima suure leiukohtade arvuga uus mikroliitiderohke kultuur (Petraglia *et al.*, 2009).

Inimeste elustiili tõi suure muutuse see, kui õpiti taimi ja loomi kodustama. Paikne eluviis, tihedam asustus, süsivesikuterikkam toit – selline suur elukorralduse muutus käivitas valiku surve osadele geenidele ning mõned selle tulemusena populatsioonis sagedaseks muutunud geenivariandid on praeguseks ka avastatud. Näiteks seoses teraviljade levikuga on suurenenud amülaasi geeni koopiate arv (Coyne & Hoekstra, 2007; Novembre *et al.*, 2007; Perry *et al.*, 2007) ja karjapidamisega tegelevatel rahvastel on muutunud laktaasi aktiivsuse regulatsioon (Beja-Pereira *et al.*, 2003; Romero *et al.*, 2011; Tishkoff *et al.*, 2007). Kus toidu kasvatamise oskused aga täpselt tekkisid ja millal nad Lõuna-Aasiasse jõudsid? Arheoloogiliste uuringute põhjal on leitud, et paljud kohalikud taimeliigid nagu hirss, hobukaun ja munguba, on Lõuna-Aasias koha peal kultuursordiks aretatud ning seda peamiselt viies erinevas India piirkonnas (Fuller, 2003). Diamond ja Bellwood (2003) leiavad, et tuntud põllukultuurid on sisserändajate poolt Lõuna-Aasiasse toodud kahelt poolt: riisikasvatus ning seapidamine Ida-Aasiast, nisu ja odra kasvatamise oskus ning kodustatud veised, lambad ja kitsed Lähis-Idast. Lähis-Ida põllumajandusoskuste komplekt jõudis Lõuna-Aasia loode osasse 7–9 tuhat aastat tagasi, kuid Kesk-Indiasse levis alles 4 tuhat aastat tagasi (Fuller, 2003). Kuigi kodustatud veised on Lõuna-Aasiasse sisse toodud, siis mtDNA põhjal on leitud, et veised on kaks korda kodustatud ning jagunevad erinevateks alamliikideks, millest üks (tuntud kui seebu) pärineb ilmselt Lõuna-Indiast (Loftus *et al.*, 1994). Riisikasvatuse levikut Lõuna-Aasias on seostatud austroaasia keelte kõnelejate sisserändega (Diamond & Bellwood, 2003). Kui varem arvati, et riisi India alamliik *Oryza sativa indica* on Ida-Aasia alamliigist *Oryza sativa japonica* eraldiseisvalt aretatud (Fuller, 2003), siis tänu mahukale riisi genoomi uuringule on praeguseks teada, et Lõuna-Hiinas aretatud sort (*japonica*) jõudis koos rännetega Lõuna-Aasiasse, kus teda kohaliku metsiku riisiga ristati ning selle tulemusena saadi alamliik *indica* (Huang *et al.*, 2012). See leid on paremini kooskõlas ka Diamondi ja Bellwoodi (2003) tulemustega.

1.3. Geneetilised uuringud

1.3.1. Erinevad genoomipiirkonnad populatsioonigeneetikas

Inimese genoom koosneb tuumagenoomist ja mitokondriaalsest genoomist. Nendel genoomidel on erinev edasikandumise mehhanism ja mutatsioonikiirus, mistõttu on nad heaks tööriistaks fülogeneesi eri tasandite uurimisel. Inimese populatsioonigeneetika uurimisele on palju juurde andnud ka tuumagenoomis leiduva Y-kromosoomi meesliini-spetsiifiline edasikandumine.

1.3.1.1. Uniparentaalselt päranduvate lookuste analüüsid

Rakkude tsütosoolis paiknevad mitokondrid on peamiselt tuntud kui organellid, mis toodavad rakkude elutegevuseks vajalikku energiat ATP näol. 1963. aastal avastati, et mitokondrites leidub ka DNA-d (Nass & Nass, 1963). Inimese mitokondriaalne DNA (mtDNA) on kaheahelaline rõngasmolekul, mis koosneb 16568-st aluspaarist (Anderson *et al.*, 1981; Andrews *et al.*, 1999) ja 37-st geenist (Bandelt *et al.*, 2006). Mitokondriaalne DNA pärandub vaid emalt lastele, sest ainult munaraku mitokondrid kanduvad järgmisesse põlve. Rakutuumas paiknev Y-kromosoom on samuti uniparentaalne marker, sest pärandub vaid isalt poegadele. Siiski, Y-kromosoomi otsad on homoloogsed X-kromosoomi otstele ning nende alade vahel võib meioosi käigus toimuda rekombinatsioon (Skaletsky *et al.*, 2003). Seetõttu räägitakse meesliini uuringute puhul umbes 60 Mb pikkusest mitte-rekombineeruvast Y-kromosoomi osast (NRY – *non-recombining region of the Y-chromosome*), mis moodustab umbes 95% tervest Y-kromosoomist. Uniparentaalse DNA abil saab uurida soo-spetsiifilist geenisiiret, mis autosoomsete markerite puhul pole võimalik. Rekombinatsiooni puudumine muudab uniparentaalsed markerid unikaalseteks tööriistadeks populatsioonigeneetikas, sest saame kindlad olla, et iga erinevus DNA-s on põhjustatud mutatsioonist, mitte rekombinatsioonist.

Viimase kalibreerimise kohaselt, mis teostati kümne kuni 40 000 aasta vanuse inimese luude abil, on mtDNA keskmine mutatsioonikiirus $2{,}67\times10^{-8}$ mutatsiooni nukleotiidi kohta aastas (Fu *et al.*, 2013). See on umbes 50 korda suurem, kui tuumagenoomil (vt ptk 1.3.1.2.). Tegelikult on mtDNA erinevate osade mutatsioonikiirus erinev – nimelt leidub mtDNA molekulil kolm mittekodeerivat regiooni HVS-I, HVS-II ja HVS-III (*hypervariable segment* – hüpervarieeruv ala), mille mutatsioonikiirus on tunduvalt suurem, näiteks HVS-I puhul $1{,}64\times10^{-7}$ mutatsiooni nukleotiidi kohta aastas (Soares *et al.*, 2009). Lisaks on leitud, et mtDNA mutatsioonikiirus ei ole ajas ühtlane – hiljutisemas minevikus on mtDNA „kell“ olnud kiirem, sest natuke kahjulikke mutatsioone ei ole jõutud veel kõrvaldada (Loogväli *et al.*, 2009). Leiti, et see seaduspära kehtib nii šimpansitel kui ka neandertallastel, aga eriti tuleb see esile inimeste puhul, sest pärast Aafrikast välja rändamist on populatsioonisuurus kõvasti kasvanud.

Ka NRY-kromosoomi mutatsioonikiirus on suurem autosoomsetest kromosoomidest, kuid mitte märkimisväärselt. Y-kromosoomi puhul tuleneb mutatsioonikiiruse tõus sellest, et see kromosoom kandub edasi vaid meesliinipidi, läbides selleks alati spermatogeneesi, mille käigus toimub aga rohkem rakujagunemisi kui ovogeneesi käigus, ja seega on rohkem võimalusi mutatsioonide tekkeks (Jobling & Tyler-Smith, 2003). Arheoloogiliste leidude põhjal on Y-kromosoomi mutatsioonikiiruseks saadud erinevates hiljutistes uuringutes 0,74–0,76×10^{-9} mutatsiooni aluspaari kohta aastas (Fu *et al.*, 2014; Karmin *et al.*, 2015). Kolossaalses Islandi suguvõsade uuringus saadi 274 meesliini uurimisel mutatsioonikiiruseks 0,87×10^{-9} mutatsiooni aluspaari kohta aastas (Helgason *et al.*, 2015). Nii nagu mtDNA puhul, on ka Y-kromosoomi puhul näha, et hiljutine „kell“ on kiirem, kuna natuke kahjulikud mutatsioonid pole veel puhastava valiku poolt kõrvaldatud.

Y-kromosoomi erinevad osad muteeruvad erineva kiirusega. Nimelt lühikeste kordusjärjestuste (STR – *short tandem repetas*) keskmine mutatsioonikiirus on mitu korda suurem kui punktmutatsioonidel (Xue *et al.*, 2009). Sellised erinevused mtDNA ning NRY mutatsioonikiiruste seas on teadlaste jaoks mugavad, sest võimaldavad uurida erineva ajalise sügavusega sündmusi (de Knijff, 2000).

Kuna nii mtDNA kui ka Y-kromosoom on haploidsed, siis on nende efektiivne populatsioonisuurus (N_e) neli korda väiksem autosoomide N_e-st (Jobling & Tyler-Smith, 2003). Seetõttu on uniparentaalsed markerid tugevamalt mõjutatud geenitriivi ja rajajaefekti poolt (Thomson *et al.*, 2000).

1.3.1.2. Ülegenoomsed analüüsid

Inimese haploidses tuumagenoomis on >3,2 miljardit aluspaari. Lisaks sellele toimub tuumagenoomis (v.a NRY) homoloogiline rekombinatsioon, mis võimaldab paljudel erinevatel lookustel eraldiseisvalt edasi kanduda vähendades sellega juhuslikkuse faktorit. Seega sisaldavad autosoomid tunduvalt rohkem informatsiooni inimkonna ajaloo kohta kui mtDNA ja NRY. Inimese tuumagenoomi mutatsioonikiiruseks on inimese ja inimahvide genoome võrreldes saadud 10^{-9} mutatsiooni aluspaari kohta aastas (Takahata & Satta, 1997), kuid viimastel aastatel võimalikuks saanud resekveneerimise tõttu saab nüüd mutatsioonikiirust leida ka otseselt. Saadud kiirused on umbes 2 korda väiksemad, jäädes 0,5×10^{-9} mutatsiooni juurde aluspaari kohta aastas (Scally & Durbin, 2012).

Kogu genoomi uurimise muudab keeruliseks homoloogilise rekombinatsiooni esinemine, kuid samas annab see omadus genoomile ka informatiivsust juurde. Rekombinatsioon ei toimu võrdse tõenäosusega kõigis genoomi piirkondades (Jeffreys *et al.*, 2001). Seetõttu päranduvad

mõned genoomi piirkonnad suurema tõenäosusega koos edasi, teised piirkonnad aga lahutatakse tihti rekombinatsiooni käigus. Kohti, kus rekombinatsioon toimub sagedamini, nimetatakse rekombinatsiooni *hotspot*ideks. Lisaks rekombinatsiooni *hotspot*idele põhjustab genoomipiirkondade vahelist sõltuvust ka looduslik valik ning füüsiline lähestikku paiknemine DNA-l. Sellist korrelatsiooni erinevate DNA segmentide vahel nimetatakse aheldus-tasakaalutuseks (LD – *linkage disequilibrium*) ning koos edasikanduvaid DNA segmente nimetatakse haplotüübiks. Aheldustasakaalutuse arvesse võtmine on oluline enamiku ülegenoomsete uuringute puhul. Haplotüüpide põhjal populatsioonide struktuuri uurimisega on tegelenud HapMap'i projekt (hapmap.ncbi.nlm.nih.gov).

Ülegenoomset geneetilist informatsiooni saadakse peamiselt kahel moel. Genotüpiseerimine võimaldab teada saada, millised nukleotiidid paiknevad konkreetsel indiviidil sadades tuhandetes saitides, mille polümorfsus inimeste seas on juba varem teada. Genotüpiseerimise suureks murekohaks ülemaailmsete populatsioonigeneetiliste uuringute puhul on uuringus kasutatavate varieeruvate saitide kallutatus (ingl SNP *ascertainment bias*) (Jobling & Tyler-Smith, 2003). Nimelt kui genotüpiseerimisplatvorme luuakse, ei võeta arvesse kõigi populatsioonide seas levinud varieeruvaid saite ja nii võib juhtuda, et eurooplaste varieeruvuse põhjal loodud kiibi abil leitakse laiahaardelisemas uuringus, et eurooplaste geneetiline mitmekesisus on tunduvalt suurem kui näiteks aafriklastel, kuigi see tegelikult nii ei ole. Sellest probleemist on vaba teine kasutusel olev meetod, genoomide resekveneerimine, mis on kallim, kuid täpsem, sest võimaldab leida uusi varieeruvaid saite ning avastada ka genoomseid ümberkorraldusi, mitte ainult ühenukleotiidilisi erinevusi (Colonna *et al.*, 2011). Resekveneerimise abil inimeste geneetilise varieeruvuse kirjeldamisega tegeleb 1000 Genoomi Projekt (1000genomes.org).

Ülegenoomsete andmete analüüsiks kasutatavaid meetodeid kirjeldatakse täpsemalt antud magistritöö peatükkides 2.2.2.–2.2.5.

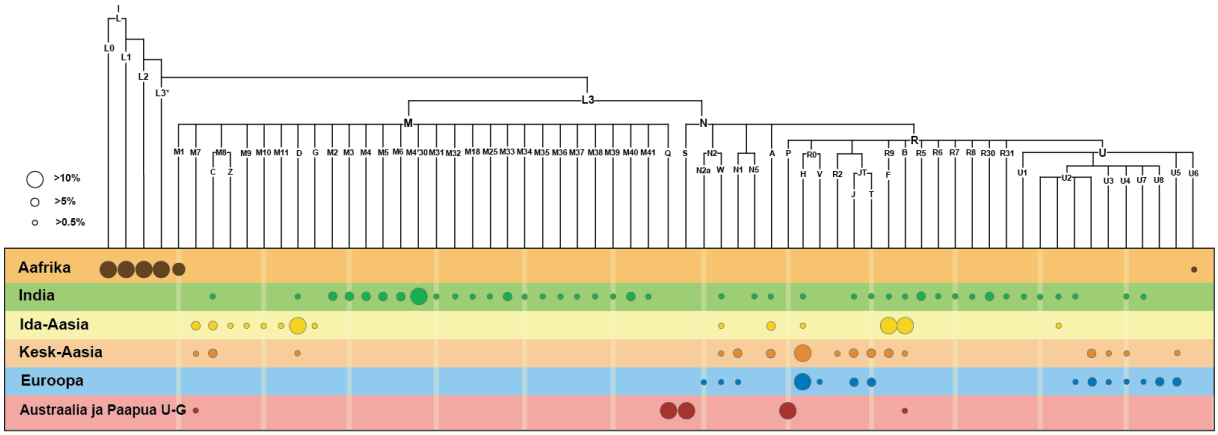
1.3.2. Geneetilised uuringud Lõuna-Aasias

Kui uuringud inimeste geneetilise mitmekesisuse osas alles algasid, olid Lõuna-Aasia rahvad paljudes projektides alaesindatud (Genomes Project *et al.*, 2010; Gibbs *et al.*, 2003; Hinds *et al.*, 2005; International HapMap *et al.*, 2007; Li *et al.*, 2008). Mõistes sealse geneetilise mitmekesisuse ulatust, on nüüdseks läbi viidud juba mitmeid uuringuid, mis hõlmavad rohkelt Lõuna-Aasia populatsioone (Kivisild *et al.*, 2003; Metspalu *et al.*, 2004; Metspalu *et al.*, 2011; Reich *et al.*, 2009; Xing *et al.*, 2010). Palju uuringuid on tehtud uniparentaalselt päranduvate mtDNA ja Y-kromosoomi põhjal (Atkinson *et al.*, 2008; Basu *et al.*, 2003; Chaubey *et al.*, 2008a; Kivisild *et al.*, 2003; Metspalu *et al.*, 2004; Sahoo *et al.*, 2006; Sengupta *et al.*, 2006),

kuid üha rohkem artikleid avaldatakse ka sadade tuhandete ülegenoomsete markerite põhjal tehtud uuringutest (Metspalu *et al.*, 2011; Moorjani *et al.*, 2013; Reich *et al.*, 2009). Resekveneerimisandmetele tuginedes näitasid Xing *et al.* (2010), et Lõuna-Aasia on geneetiliselt mitmekesisuselt teisel kohal pärast Aafrikat.

1.3.2.1. Mitokondriaalse DNA uuringute tulemused Lõuna-Aasias

Mitokondriaalse DNA haplogruppide uuringud on andnud esmase ülevaate inimpopulatsioonide päritolust ja omavahelistestest seostest. Haplogrupp on monofüleetiline rühm ehk klaad, mistõttu ühe fülogeneetilise puu peal on palju haplogruppe erinevatel hierarhia tasemetel. Populatsioonidel, mis on hiljuti lahknenud või mille vahel on toimunud tugev geenivool, on haplogruppide sagedused sarnasemad. Selleks, et mõista, kuhu paigutub Lõuna-Aasia selles haplogruppide rägastikus, on vaja omada taustainfot haplogruppide jaotumise kohta Maal (joonis 2). On leitud, et kogu Aafrika-väline mitokondri haplogruppide mitmekesisus on tekkinud vaid kahest asutaja-haplogrupist M ja N, mis on Aafrika haplogrupi L3 tütarhaplogruppideks (Van Oven & Kayser, 2009). Sageli tuuakse eraldi välja ka haplogrupi N tütarhaplogrupp R, sest see on väljaspool Aafrikat saavutanud suure mitmekesisuse. Euroopas on levinud vaid haplogrupi N tütarhaplogrupid, mujal maailmas leidub mõlema haplogrupi tütarhaplogruppe (Chaubey *et al.*, 2007; Torroni *et al.*, 2006). Nii on ka Lõuna-Aasias palju erinevaid haplogruppe nii M kui ka N liinist, kusjuures erinevate juurest hargnevate haplogruppide arv on seal suurem kui mujal Euraasias või Okeaanias, mis tähendab seda, et rohkem mitmekesisust on säilunud Lõuna-Aasia asustamisperioodist (Chaubey *et al.*, 2007).



Joonis 2. Mitokondri haplogruppide fülogeneetilised seosed ning levik Aafrikas, Indias, Ida-Aasias, Kesk-Aasias, Euroopas, Austraalias ja Paapua Uus-Guineas. Ringi suurus näitab, kui levinud on vastav haplogrupp antud piirkonnas. (Allikas: Chaubey *et al.* (2007), kohandatud)

Haplogrupp L3 tekkis Ida-Aafrikas 60–70 tuhat aastat tagasi (Soares *et al.*, 2011). Oli klimaatiliselt soodne aeg, inimpopulatsioon kasvas kiirelt ning mitokondri genoom mitmekesistus. Seetõttu lahknesid algsest haplogrupist L3 makrohaplogrupid N ja M juba õige

pea pärast L3 teket. Soares *et al.* (2009) on hinnanud, et Lõuna-Aasias leiduva haplogrupi N vanus jääb samasse vahemikku, kuhu L3 tekegi. See tähendab, et kõigi praeguste Lõuna-Aasia N haplogruppi kuuluvate alamhaplogruppide ühine esiema elas 60–70 tuhat aastat tagasi. Keskmiselt 60 protsendi indialaste mitokondri genoom kuulub umbes sama vanasse makrohaplogruppi M, mille osakaal on kõrgeim India lõuna- ja idaosas ning hõimude seas (Metspalu *et al.*, 2004). Ka Siberis, Hiina põhjaosas ja Jaapanis on haplogrupp M haplogrupist N veidi rohkem levinud, kuid Kagu-Aasias on N veidi sagedasem ja Edela-Aasias puudub M praktiliselt üldse (Metspalu *et al.*, 2006). Lääne- ja Ida-Euraasia haplogrupi N (sealhulgas ka R) tütarhaplogrupid ei kattu üksteisega, aga Kesk-Aasias seevastu leidub mõlema diferentseerunud populatsiooni N haplogruppe umbes võrdse osakaaluga (Comas *et al.*, 2004). Ameerika põliselanike seas on levinud väike osa Ida-Aasia haplogruppidest (Torroni *et al.*, 1993). Austraalias ja Uus-Guineas on levinud haplogruppide R ja M autohtoonsed (st kujunenud samas kohas, kus praegu levinud on) liinid (Hudjashov *et al.*, 2007).

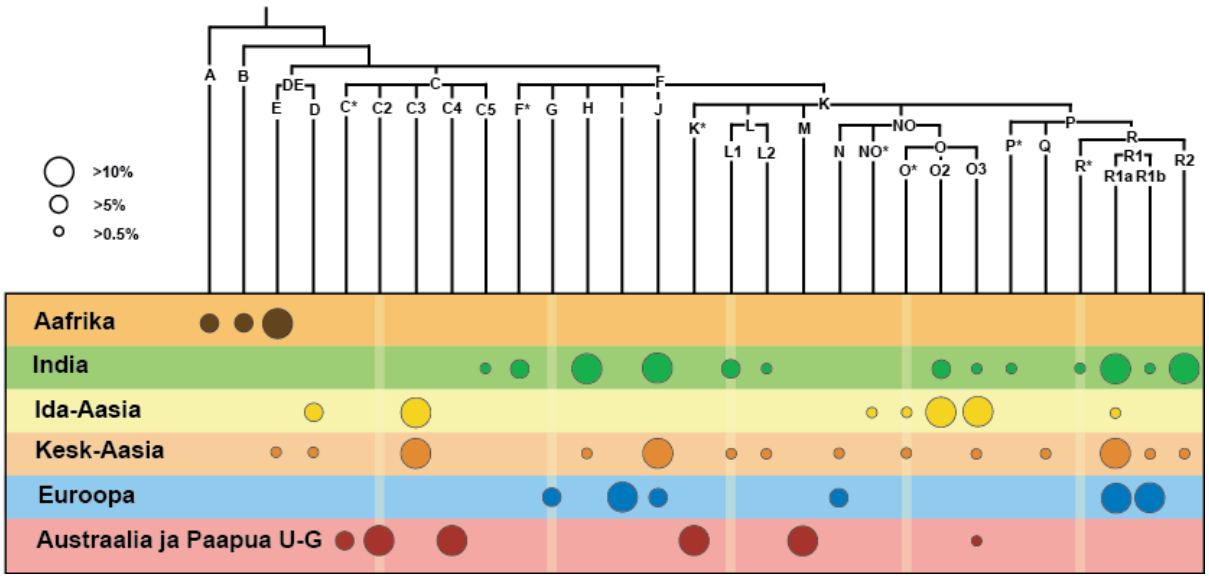
Peaaegu kõik Lõuna-Aasias levinud haplogrupid on samuti autohtoonsed (Kivisild *et al.*, 2003). Seega on nad otseselt algse makrohaplogrupi järglased ning pole üldjuhul mujale maailma levinud ega muid teid pidi Lõuna-Aasiasse jõudnud. India haplogruppide iidset päritolu kinnitab ka see, et erinevate haplogruppide jaotus populatsioonis pole seotud keelega ega kastidesse või hõimudesse kuulumisega. Kuigi triivi tõttu haplogruppide sagedused pisut erinevad populatsioonide vahel, on siiski samad haplogrupid esindatud üle kogu India (Kivisild *et al.*, 2003). Järelikult olid erinevad haplogrupid Indias juba laialdaselt levinud enne, kui tekkis kastisüsteem, mis inimeste ristumist piirama hakkas. Lõuna-Aasias leidub ka mõningaid Ida-Aasiale spetsiifilisi haplogruppe, kuid vaid väikses koguses ja peamiselt India kirdeosas. Lääne-Euraasia haplogruppidest leidub Loode-Indias mõningaid vanu haplogrupi R liine, mis ei ole hiljutise migratsiooni teel sinna sattunud (Chaubey *et al.*, 2007). See, et nii Lääne-Euraasias, Ida-Aasias, Lõuna-Aasias, Kagu-Aasias kui ka Okeaanias ja Austraalias esineb palju autohtoonseid haplogruppe nii N, R kui ka M liinist, tähendab, et need kolm liini tekkisid peaaegu ühel ajal vahetult enne seda, kui tänapäevane inimene üle kogu Maa levis. Samuti saab järeldada, et Euraasia ja Austraalia asustamine oli kiire protsess ning geneetiline mitmekesistumine toimus hiljem koha peal, mitte ei ole tegu järkjärgulise protsessiga, mille puhul ühe populatsiooni haplogruppide mitmekesisus moodustab alamhulga teise populatsiooni mitmekesisusest (Macaulay *et al.*, 2005).

Populatsiooni struktuur ja põlvnemine pole ainsad demograafilised omadused, mida mitokondri genoomi abil uurida saab. Kasutades informatsiooni mitokondri genoomi mitmekesisuse kohta, uurisid Atkinson *et al.* (2008) populatsiooni suhtelist suurust ja suuruse muutusi erinevates Maa

piirkondades erinevatel aegadel. *Bayesian skyline plot*i põhjal leiti, et esimene inimpopulatsiooni kiire kasv väljaspool Aafrikat toimus Lõuna-Aasias. Selle uuringu tulemuste põhjal viiekordistus Lõuna-Aasia populatsioon lühikese ajavahemiku jooksul umbes 52 tuhat aastat tagasi. Vahemikus 45–20 tuhat aastat tagasi elas Lõuna-Aasias üle 50 protsendi kogu Maa rahvastikust ning protsentuaalsesse haripunkti jõudis Lõuna-Aasia 38 tuhat aastat tagasi, kui seal elas vähemalt 60% kõigist inimestest.

1.3.2.2. Y-kromosoomi uuringute tulemused Lõuna-Aasias

Vaid meesliini pidi edasi kanduva Y-kromosoomi haplogruppide esinemismustrid Maal on sarnased mitokondri genoomile, kuid leidub ka mõningaid väikeseid erinevusi, mis võivad olla tingitud patrilokaalsusest või soospetsiifilistest migratsioonidest (Seielstad *et al.*, 1998). Kuna Y-kromosoomi efektiivne populatsiooni suurus on väike, siis on haplogruppide levikut suuresti määranud pigem geenitriiv kui geenivool. Karmima kliimaga aladel, kus läbi ajaloo on esinenud vaid väikesed inimpopulatsioonid, on geenitriivil eriti suur mõju olnud ja inimgrupid on diferentseerunud (Karafet *et al.*, 2002; Zerjal *et al.*, 2002). Piirkondades, kus populatsiooni suurus on pidevalt suurem olnud, on geenitriivil nõrgem mõju olnud ning haplogruppide jaotus on populatsioonide vahel sarnasem.



Joonis 3. Y-kromosoomi haplogruppide fülogeneetilised seosed ning levik Aafrikas, Indias, Ida-Aasias, Kesk-Aasias, Euroopas, Austraalias ja Paapua Uus-Guineas. Ringi suurus näitab, kui levinud on vastav haplogrupp antud piirkonnas. (Allikas: Chaubey *et al.* (2007), kohandatud)

Y-kromosoomi haplogruppide esinemismustrid (joonis 3) on peamiselt väljakujunenud paleoliitikumis, kui populatsiooni suurused olid väiksemad ja toimus rohkem rändeid (Jobling & Tyler-Smith, 2003). Kõigi tänapäeva meeste hiljutisim ühine eellane elas viimase kalibreerimise kohaselt umbes 254 tuhat aastat tagasi (Karmin *et al.*, 2015) Lääne-Aafrikas

(Cruciani *et al.*, 2011). Haplogruppe A ja B, mis on kõige varem ülejäänutest lahknenud, leidub peamiselt vaid Aafrika lõuna-, ida- ja lääneosas ning sedagi enamasti madala sagedusega (Cruciani *et al.*, 2002; Ellis & Hammer, 2002; Underhill *et al.*, 2000). Kõige rohkem esineb neid haplogruppe khoisanide ja etiooplaste seas (Semino *et al.*, 2002). Kagu-Aasias, Austraalias ja Okeaanias on levinud haplogrupid C, K, O ja M, millest C ja K liine esineb ka mujal Aasias (Chaubey *et al.*, 2007; Underhill & Kivisild, 2007). Euraasia peamisteks Y-kromosoomi haplogruppideks on I, J, N ja R. Sellise selge haplogruppide kaheks jagunemise taga arvatakse olevat kaks eraldiseisvat rännet Aafrikast: esimene viis mööda lõuna-rannikut Austraaliani ning teise rände käigus asustati Euroopa ja ülejäänud Aasia (Jobling & Tyler-Smith, 2003). Ameerika põliselanike seas on levinuimaks C ja Q haplogrupi liinid, mis sarnanevad Edela-Siberi haplogruppidele (Zegura *et al.*, 2004). Enamik Lõuna-Aasia Y-kromosoomi haplogruppe on evolutsioneerunud asutajahaplogruppidest C, F ja K (Sengupta *et al.*, 2006; Underhill & Kivisild, 2007). Sarnaselt mitokondri haplogruppidele on paljud kõige tavalisemad Lõuna-Aasia Y-kromosoomi haplogrupid (C5, F*, H, R2, L1) autohtoonsed ning mujal maailmas neid eriti ei leidu (Sahoo *et al.*, 2006; Sengupta *et al.*, 2006). Ülejäänud haplogruppidest on Lõuna-Aasias levinud näiteks R1a, mis on samuti väga sage Ida-Euroopas (Underhill *et al.*, 2015). R1a mitmekesisus ja sagedus viitavad haplotüübi Euroopa päritolule, kuid haplogrupi Indias esinev mitmekesisus annab aimu sellest, et R1a ei ole sinna hiljuti saabunud, vaid vähemalt 6000 aastat tagasi (Underhill *et al.*, 2010). Seda haplogruppi on seostatud Kaukasuse piirkonnast pärit indoeurooplaste sissetungiga Indiasse, sest seda esineb peamiselt vaid kastipopulatsioonides ning arvatakse, et just indoeurooplaste sissetung tõi Indiasse kastisüsteemi ja indoeuroopa keeled (Chaubey *et al*., 2007 ja viited selle sees). Mitmed teadlased siiski ei usu seda teooriat, kuna see eeldab liigset lihtsustamist (Kivisild *et al.*, 2003; Sahoo *et al.*, 2006; Sengupta *et al.*, 2006).

Euroopa vanimat Y-kromosoomi haplogruppi BR* (Rosser *et al.*, 2000) leidub ka Indias, kuid selle levimine pole kooskõlas indoeurooplaste sissetungiga, sest selle osakaal on hõimudes keskmiselt suurem kui kastipopulatsioonides (Basu *et al.*, 2003). H1 alamhaplogrupp H1a1a, mis iseloomustab roma mustlasi, on pärit Loode-Indiast (Rai *et al.*, 2012). Haplogrupp O2a on eriti levinud India kirdeosas tiibeti-birma ja austroaasia keelte kõnelejate seas ning pärineb Kagu-Aasiast (Chaubey *et al.*, 2011; Cordaux *et al.*, 2004). Nende samade India hõimude seas leidub palju ka vana haplogruppi K*, mis on ühtlasi levinud hiinlaste seas (Basu *et al.*, 2003; Su *et al.*, 2000).

1.3.2.3. Kogu genoomi uuringute senised tulemused Lõuna-Aasias

Rekombineeruvate genoomi alade uurimine on suurema hoo sisse saanud pärast kommertsiaalsete mikrokiipide kasutuselevõttu, kuid ka enne seda üritati autosoomsete lookuste põhjal populatsioonigeneetilisi järeldusi teha. Näiteks uurisid Basu *et al.* (2003) lisaks uniparentaalsetele markeritele ka 25 autosoomset lookust, mille põhjal leiti nelja peamise India keelkonna rahvaste võrdluses, et indoeuroopa ja draviidi keeli kõnelevad rahvad on sarnaseimad. Kivisild *et al.* (2003) võtsid oma uuringusse ühe 21. kromosoomi lookuse, millel asub kaheksa varieeruvat saiti. Nad leidsid, et India kasti- ja hõimupopulatsioonide vahel puudub neis saitides statistiliselt oluline erinevus, aga võrreldes kõigi teiste vaatluse all olnud piirkondadega – näiteks Ida-Aasia, Euroopa, Pakistani ja Austraaliaga – on India rahvad selgelt erinevad. Rosenberg *et al.* (2006) uurisid 1200 ülegenoomset varieeruvat saiti 432 indialasel, kes kuuluvad 15 erinevasse populatsiooni. Nende tulemuste kohaselt olid uurimisalused geneetiliselt üllatavalt homogeensed ning ei klasterdunud gruppidesse. Uuringu tulemusi tuleb siiski kriitiliselt võtta, sest valim koosnes vaid Ameerika Ühendriikidesse immigreerunutest. Järgmises suuremas uuringus võeti valimisse 55 populatsiooni üle kogu India ning uuriti neid 405 SNP-i põhjal (Brahmachari *et al.*, 2008). Vastupidiselt eelmainitud uuringule leiti selles töös, et India populatsioonid klasterduvad geograafia ja keelte põhjal gruppidesse. Siiski pole see SNP-ide arv piisav, et väga põhjalikku geneetilist profiili Indiast kuvada.

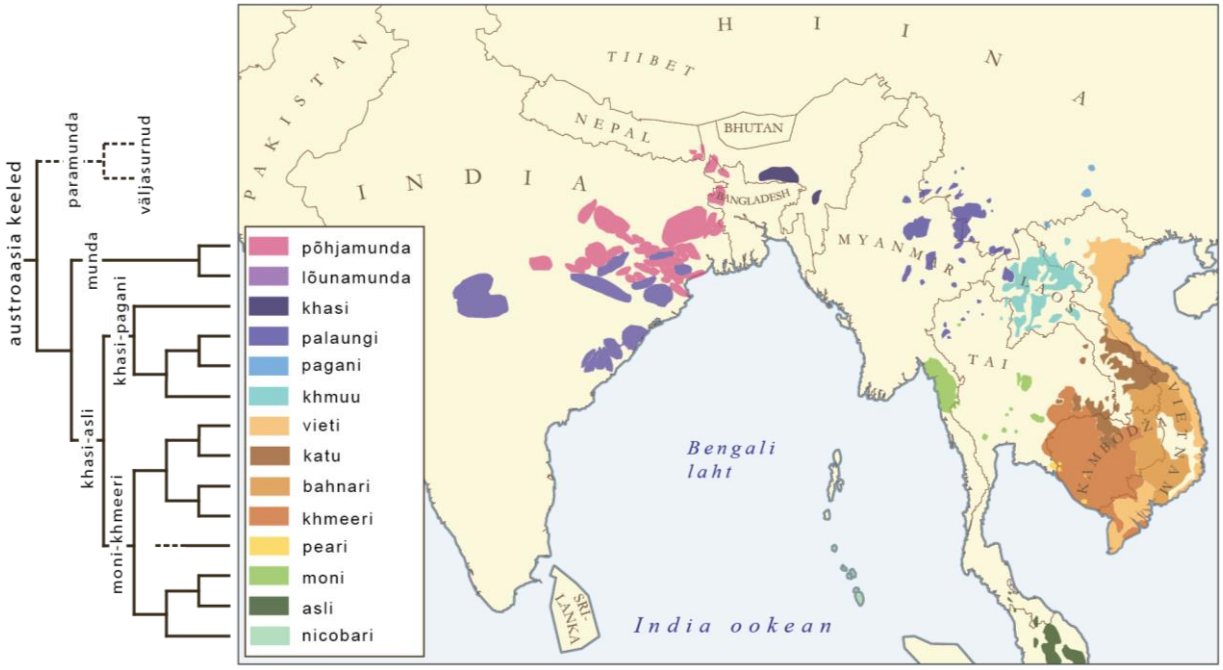
Ülegenoomsete markerite põhjal on Lõuna-Aasias ka geenivoolu uuritud. On teada, et islam levis Indiasse juba 7. sajandil läbi Araabia kaupmeeste. Moslemitel on olnud suur roll India kultuuris ja majanduse arengus ning nad on erinevatel ajaperioodidel isegi piirkonna valitsejaks olnud (Esposito, 1999). Selle põhjal võiks arvata, et India moslemid on geneetilist mitmekesisust Indias suurendanud. Uurides 13 autosoomset STR markerit, selgus siiski, et pigem on islam olnud kultuuriline nähtus Indias ning vaid osade sealsete moslemite genoomist võib leida nõrku signaale Lähis-Ida päritolu kohta (Eaaswarkhanth *et al.*, 2009). Üks selline Lõuna-Aasia populatsioon, kus geenivoolu viimase 200 aasta jooksul on esinenud, on Indias ja Pakistanis elavad siddid. Nende lähedast sugulust Aafrika bantu keele kõnelejatega näitas mtDNA, Y-kromosoomi ja 250 tuhande autosoomse SNP-i uuring (Shah *et al.*, 2011). Ka ajaloost on teada, et siddid sattusid Lõuna-Aasiasse, kui portugallased 300–500 aastat tagasi Lõuna-Aasiasse orjasid tõid (Bhattacharya, 1970).

Esimene mahukas ülegenoomsetel SNP-del põhinev Lõuna-Aasia rahvaste uuring avaldati 2009. aastal. Reich *et al.* (2009) uurisid 560 tuhandet SNP-i 25 populatsioonil erinevatest sotsiaalsetest ja keelelistest gruppidest üle kogu India ja Andamani saarte. Leiti, et tänane geneetilise varieeruvuse muster Lõuna-Aasias on tekkinud kahe varasemalt eksisteerinud

populatsiooni, mida autorid nimetasid ANI (*Ancestral North Indians* – iidsed põhja-indialased) ja ASI (*Ancestral South Indians* – iidsed lõuna-indialased), ristumisel. ANI komponent on lahknenud eurooplaste eellasest, ASI komponent on levinud vaid Indias. Kuigi mõlemad komponendid on peaaegu kõigil India rahvastel olemas, esineb komponentide osakaaludel genoomis selge gradient – nimelt Põhja-India populatsioonidel on suurem osakaal ANI komponendil (77%), mis aga lõuna poole liikudes sujuvalt langeb kuni 39%-ni genoomist. India austroaasia keelte kõnelejatel (Chaubey *et al.*, 2011) ning Andamani saarte rahvastel puudub ANI komponent täielikult (Reich *et al.*, 2009). Samade tulemusteni indialaste kahe peamise geneetilise komponendi koha pealt jõudsid ka Metspalu *et al.* (2011). Haplotüüpide mitmekesisuse analüüsi põhjal leidsid nad, et ANI komponent on Indias olnud juba umbes 12500 aastat lükates ümber teooria, et see komponent jõudis Indiasse 3500 aastat tagasi toimunud indoeurooplaste sisserändega. Lisaks toodi välja, et Pakistan jääb geneetilises mõttes Euroopa ja India lõuna-osa keskele, mistõttu on varasemalt Lõuna-Aasia geneetiline mitmekesisus ja autohtoonsus jäänud tihti tähelepanuta, sest mitmetes uuringutes, nt Li *et al.* (2008), on pakistanlasi kasutatud kui Lõuna-Aasia esindajaid. Moorjani *et al.* (2013) hindasid aheldustasakaalutuse põhjal, et ANI-ASI segunemine toimus 1900–4200 aastat tagasi. Nad järeldasid, et tihe üle-Indialine segunemine rauges sel ajal endogaamse eluviisi levimise tõttu. Esimene resekveneerimisel põhinev uuring Indias (Xing *et al.*, 2010) leidis 92 indiviidil kõigest 100 kb DNA sekveneerimisel 137 uut varieeruvat saiti asetades India geneetiliselt mitmekesisuselt Aafrika järel teisele kohale.

1.4 Mundad – kas India põliselanikud või hoopis uusimad sisserändajad?

Mundadeks kutsutakse mitmeid India kesk- ja idaosa hõimurahvaid, kes räägivad munda keeli. Munda keeled on üks austroaasia keelkonda kuuluvatest keelerühmadest. Munda keeled jaotatakse omakorda põhjamunda keelteks ja lõunamunda keelteks (www.britannica.com/EBchecked/topic/397435/Munda-languages). Teine austroaasia keelte haru munda keelte kõrval on khasi-asli, mille keeli räägitakse peamiselt Kagu-Aasias (joonis 4). Mundade arvukus oli 20. sajandi lõpul umbes 9 miljonit. Piirkond, kus mundad elavad, on mägine ja metsaderohke ning suurematest India keskustest eraldatud (www.britannica.com/EBchecked/topic/397427/Munda).



Joonis 4. Austroaasia keelte puu ja levik kaardil (Allikas: Chaubey *et al.* (2011), kohandatud)

Mundade päritolu on olnud põletav uurimisküsimus. Ühe spekulatsiooni kohaselt on austroaasia keelte kõnelejad, sealhulgas ka India mundad, pärit Kagu-Aasiast, teisalt on pakutud välja, et austroaasia keeled said alguse Lõuna-Aasias ning liikusid sealt edasi Kagu-Aasiasse (Fuller, 2007). Hüpoteesi, et austroaasia keeli kõnelevad rahvad on India esmaasustajad, toetavad nii Majumder (2001) kui ka Basu *et al.* (2003) leid, et austroaasia keeli kõnelevatel hõimudel esineb suurim mitokondri HVS-1 varieeruvus erinevate India sotsiaalsete ja keeleliste gruppide seas. Selle varieeruvuse põhjal hinnati, et austroaasia keelte kõnelejate seas toimus kiire rahvastiku kasv umbes 15 000 aastat varem kui teiste populatsioonigruppide seas. Lisaks leidsid Basu *et al.* (2003), et India austroaasia keele kõnelejate seas puudub noor mitokondri haplogrupp M4, mis on mujal Indias väga tavaline (15%). Viimane väide on siiski valimi omapäradest tingituna seatud kahtluse alla (Metspalu *et al.*, 2004). Mundade

mitokondriaalsed liinid on täiesti erinevad Kagu-Aasia austroaasia ehk khasi-asli keelte kõnelejate liinidest, sarnanedes hoopis India indoeuroopa ja draviidi keelte kõnelejate liinidele (Thangaraj *et al.*, 2005). Vastupidiselt mitokondri haplogruppidele on nii mundade kui ka Kagu-Aasia austroaasia keelte kõnelejate seas levinud ühise Y-kromosoomi haplogrupi O2a liinid. (Chaubey *et al.*, 2011; Sahoo *et al.*, 2006; Sengupta *et al.*, 2006). Kusjuures, mundade O2a noort vanust (<10 000 aastat) on tõlgendatud kui märki hiljutisest Kagu-Aasia päritolust (Chaubey *et al.*, 2011; Sahoo *et al.*, 2006). Mundade Kagu-Aasia päritolule viitab ka O2a haplogrupi suurem mitmekesisus Kagu-Aasias (Chaubey *et al.*, 2011; Tamang & Thangaraj, 2012).

Chaubey *et al.* (2011) leidsid autosoomseid mikrokiibi andmeid kasutades, et peakomponentanalüüsi põhjal sarnanevad mundad enim India draviidi keeli kõnelevate hõimudega, kuid on nihutatud veidi Kagu-Aasia populatsioonide suunas. Ka populatsioonide struktuuri näitav ADMIXTURE analüüs (vt ptk 2.2.2.) kinnitas, et ligi 75% mundade geneetilisest varieeruvusest kirjeldab Indias levinud komponent, mida vähesel määral esineb ka khasi-asli keelte kõnelejatel. Samas selgus, et mundadel puudub täielikult üks kahest põhilisest komponendist, mis kõigil teistel indialastel esineb. Ülejäänud 25% mundade geneetilisest varieeruvusest kirjeldab Ida- ja Kagu-Aasia esivanemlik komponent, mis India draviidi ja indoeuroopa keelte kõnelejatel puudub. Erinevate analüüside kokkuvõtteks leidsid Chaubey *et al.* (2011), et mundad on Indiasse saabunud Kagu-Aasiast, kus nad on kohalike India populatsioonidega soo-spetsiifiliselt segunenud.

2. PRAKTILINE OSA

2.1 Töö eesmärgid

Varasematest uuringutest on teada, et India munda keelte kõnelejad (mundad) on isaliini pidi geneetiliselt väga sarnased Kagu-Aasia austroaasia keelte kõnelejatega, kuid Y-kromosoomi haplogruppide mitmekesisus on Kagu-Aasias suurem (Chaubey *et al.*, 2011; Sahoo *et al.*, 2006; Sengupta *et al.*, 2006). See viitab mundade Y-kromosoomide Kagu-Aasia päritolule. Emaliinidelt sarnanevad munda keele kõnelejad geograafiliselt lähedal asuvate India rahvastega (Basu *et al.*, 2003; Metspalu *et al.*, 2004). Korduvalt on väidetud, et just mundad on anatoomiliselt moodsast inimesest rääkides India subkontinendi esmaasukad (Basu *et al.*, 2003; Majumder, 2001). Kuivõrd keelepuude rekonstrueeritavad ajalised sügavused ei ulatu ligilähedalegi Euraasia asustamise aegadesse, ei ole sellised väited adekvaatsed. Samuti on mitokondriaalse DNA varieeruvust uurides näidatud, et India mundade hulgas ei ole vanad haplogrupid (nt M2) üleesindatud (Metspalu *et al.*, 2004) nagu eelnevalt väideti (Basu *et al.*, 2003). Ülegenoomsed analüüsid on näidanud, et mundad on tekkinud indialaste ja Kagu-Aasia päritolu populatsioonide segunemisel (Chaubey *et al.*, 2011). Lahtiseks on jäänud nii selle segunemise aeg kui ka küsimus sellest, millised tänapäeva Kagu-Aasia populatsioonid on lähimad populatsioonile, mis Indias kohalikega segunedes mundadele aluse pani.

Käesoleva uuringu eesmärgiks on kasutada varasemast rohkem võrdluspopulatsioone Kagu-Aasiast, et täpsemalt hinnata India munda keele kõnelejate päritolu. Lisaks uuritakse selles töös esmakordselt lähtepopulatsioonide segunemisaega, et saada teada India mundade vanus. Selleks kasutatakse suure hulga autosoomsete SNP-ide andmeid.

Töö eesmärk on vastata küsimustele:

- Millised Lõuna- ja Kagu-Aasia populatsioonid on sarnaseimad India mundadele?
- Millal toimus populatsioonide segunemine, mille tulemusena tekkis mundade populatsioon?

2.2. Materjal ja metoodika

2.2.1. Andmed

Kokku on uuringusse võetud 841 indiviidi 67 populatsioonist. Kasutatud on erinevates artiklites (Chaubey *et al.*, 2011; Li *et al.*, 2008; Metspalu *et al.*, 2011; Migliano *et al.*, 2013; Rasmussen *et al.*, 2011; Yunusbayev *et al.*, 2012) varem avaldatud andmeid ning seega on DNA proovid juba varem genotüpiseeritud Illumina 610K, 650K, 660K ja 710K SNP mikrokiipidega. Lisaks on kasutatud varem avaldamata Eesti Biokeskuse ja Cambridge'i uurimisgrupi andmeid. Enamik indiviide on Indiast ja mujalt Aasiast, kuid võrdluseks on kaasatud ka Aafrika, Euroopa, Okeaania, Lähis-Ida ja Kaukaasia populatsioonid. Kokkuvõtet valimist võib näha tabelis 1. Aasia populatsioonide asukohad on esitatud kaardil (joonis 5). Täpsem ülevaade valimist (populatsioon, riik, keel, allikas) on ära toodud lisas 1.

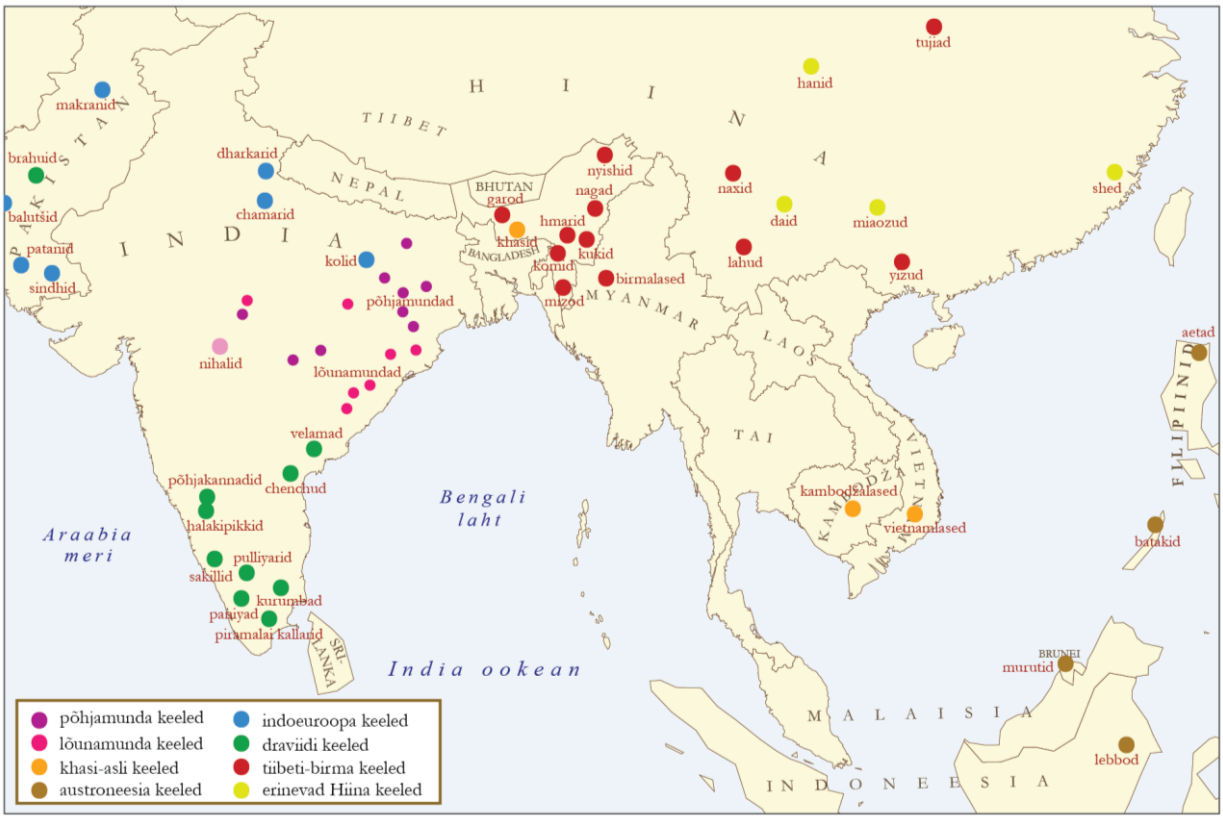
Tabel 1. Kokkuvõte valimist

Piirkond	**Indiviidide arv**	**Populatsioonide arv**
Aafrika	61	3
Kaukaasia	53	3
Lähis-Ida	90	2
Euroopa	58	3
Ida-Aasia	166	12
Lõuna-Aasia	260	35
sh põhjamundad	*9*	*4*
lõunamundad	*11*	*5*
Kagu-Aasia	126	7
Okeaania	27	2
Kokku	**841**	**67**

2.2.1.1. Andmete filtreerimine

Selleks, et kätte saada erinevate kiipide peal kattuvad autosomaalsed SNP-id ning tagada andmestiku kvaliteet, on andmed filtreeritud kasutades tarkvara PLINK v1.07 (Purcell *et al.*, 2007). Andmestikust eemaldati indiviidid, kelle genotüpiseerimisedukus oli alla 97 protsendi. Lisaks eemaldati SNP-id, mille genotüpiseerimisedukus oli alla 97% ja ka need SNP-id, mille puhul haruldasema alleeli sagedus oli alla 1%. Pärast neid etappe jäi alles 305 782 SNP-i. Kuna osade analüüside (peakomponentanalüüs, ADMIXTURE) eelduseks on andmete sõltumatus, siis on nende jaoks eemaldatud SNP-id, mis on tugevas aheldustasakaalutuses (korrelatsioon r^2

>0,4), liigutades 200 SNP-i pikkust akent 25 SNP-i võrra edasi. Pärast sellist filtreerimist jäi andmestikku 841 indiviidi (kõik) ja 186 242 SNP-i.



Joonis 5. Valimis olevate Aasia populatsioonide ligikaudne asukoht ja kõneldav keel

2.2.2. Meetodid andmetest ülevaate saamiseks: peakomponentanalüüs ja ADMIXTURE

Selleks, et saada esmast ülevaadet sellest, millised populatsioonid on geneetilises mõttes omavahel sarnased, kasutati peakomponentanalüüsi. Kui tegelikult on andmestik maksimaalselt nii mitme mõõtmeline, kui palju on seal varieeruvaid saite, siis moodustades algsetet dimensioonidest lineaarkombinatsioone, võimaldab peakomponentanalüüs visualiseerida vaid kõige informatiivsemaid mõõtmeid ehk kõige rohkem varieeruvaid dimensioone korraga. Peakomponentanalüüs viidi läbi kasutades tarkvarapaketti EIGENSOFT 5.0.2 (Patterson *et al.*, 2006). Kuigi peakomponentanalüüsi väljund on indiviidi, mitte populatsiooni kohta, siis kasutades keskmistamist on võimalik visualiseerida ka populatsioonide kaugusi üksteisest.

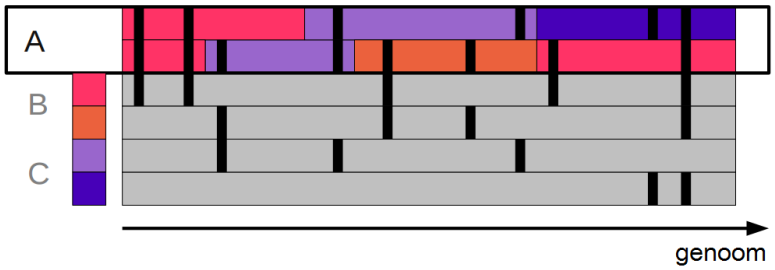
Teine meetod, mida andmetest ülevaate saamiseks kasutati, võimaldab aimu saada populatsiooni struktuurist, väljendades iga indiviidi geneetilist profiili komplektina konstrueeritud hüpoteetiliste eellaspopulatsioonide osakaaludest. Tegemist ei ole spetsiifiliselt mingite eellaspopulatsioonide osakaalude leidmisega vaid olemasoleva geneetilise struktuuri kirjeldamisega. Meetodi põhiolemus seisneb analüüstavate genoomide geneetilise

mitmekesisuse esitamises eelpool kirjeldatud komponentidena nii, et maksimeeritakse Hardy-Weinbergi tasakaal ja minimeeritakse LD-d. Meetod ei näita otseselt, kas populatsioon on tekkinud mitme populatsiooni segunemise käigus või on ise olnud allikaks teiste populatsioonide tekkele. Kui ühe populatsiooni indiviidide eellaskomponentide proportsioonid väga palju kõiguvad, siis see viitab hiljutisele segunemisele.

Populatsioonide struktuuride leidmiseks kasutati antud töös programmi ADMIXTURE (Alexander *et al.*, 2009). Kuna täpne eellaspopulatsioonide arv (K) pole teada, siis jooksutati programmi eeldades erinevaid K väärtusi alates K=2 kuni K=19. Iga K puhul korrati analüüsi 100 korda, et tõepära skooride (*LogLikelihood score*, LLs) hajuvuse põhjal oleks võimalik hinnata, milline K väärtus on parim. Parimaks loetakse sellise eellaspopulatsioonide arvuga mudelit, mille puhul LLs on võimalikult kõrge, kuid skooride hajuvus on võimalikult väike. Antud analüüsis oli selliseks K väärtuseks 9 (Lisa 3b). Nii EIGENSOFT kui ka ADMIXTURE tarkvara jooksutamiseks ja tulemuste visualiseerimiseks statistikapaketis R on kasutatud dr Mait Metspalu kirjutatud koodi.

2.2.3. Meetodid geneetiliselt sarnaste populatsioonide leidmiseks: fineSTRUCTURE ja RefinedIBD

Sarnaselt peakomponentanalüüsile ja ADMIXTURE analüüsile annab fineSTRUCTURE meetod (Lawson *et al.*, 2012) aimu sellest, millised populatsioonid on üksteisele sarnasemad tänu ühistele esivanematele ning võimaldab näha andmestiku struktuuri erinevatel tasanditel, alates maailmajagude tasandist kuni perekondade eristumiseni välja. fineSTRUCTURE on eelnevalt kirjeldatud meetoditest täpsem, sest kasutab eelteadmisi selle kohta, millised varieeruvad saidid on aheldatud (joonis 6), mitte ei vaata kõiki saite sõltumatutena. Selleks, et kasutada seda informatsiooni, vajab fineSTRUCTURE faasitud andmeid. Antud magistritöö jaoks on andmed faasitud programmiga Beagle 3.3.2 (Browning & Browning, 2007) kasutades dr Gyaneshwer Chaubey kirjutatud skripti. fineSTRUCTURE tarkvara kasutati antud töös, et leida, millistest Kagu- ja Lõuna-Aasia populatsioonide genoomidest saab kõige paremini



Joonis 6. Esimese sammuna kromosoomid „värvitakse“ vastavalt sarnasusele teiste indiviididega. Antud näites on indiviidi A mõlemad haploidsed kromosoomid värvitud indiviidide B ja C kromosoomide põhjal. (Allikas: www.paintmychromosomes.com/)

mundade genoomi kokku panna. fineSTRUCTURE eeliseks eelmises peatükis kirjeldatud meetodite ees on ka tulemuste parem tõlgendatavus. Tulemused on visualiseeritud kasutades kasutajaliidest fineSTRUCTURE GUI.

Selleks, et leida, millistel populatsioonidel on mundadega ühised esivanemad, on antud töös kasutatud ühiselt esivanemalt päritud DNA lõikude (IBD – *identity by descent*) detekteerimisel põhinevat meetodit Refined IBD (Browning & Browning, 2013). Täpsemalt võimaldab antud meetod leida, kui palju ühise päritoluga DNA lõike jagab uuritav populatsioon teiste populatsioonidega, kusjuures tulemusi on võimalik vaadata DNA lõikude pikkuse lõikes. Pikkade DNA lõikude jagamine viitab hiljutisele ühisele esivanemale või segunemisele ning lühemad jagatud lõigud viitavad vanemale geneetilisele seosele populatsioonide vahel. Evolutsiooniliste sündmuste aegu selle meetodiga siiski määrata ei saa. Refined IBD on võimas meetod, kuna ei nõua aheldunud saitide eemaldamist, vaid hoopis kasutab LD-s sisalduvat informatsiooni ära. Täpsuse suurendamiseks kasutab Refined IBD tõenäosuslikku lähenemist vastupidiselt näiteks samade autorite varasemale meetodile fastIBD (Browning & Browning, 2011). Programmi kasutamine nõudis eelnevat varieeruvate saitide geneetiliste kauguste (sentimorganites) lisamist andmestikku. Lisaks tuli andmed viia VCF formaati, mida tehti PLINK/SEQ paketi abil. Analüüsi jooksutamisel oli parameetri *ibdtrim* väärtuseks 20. Alles jäeti kõik IBD lõigud, mille LOD väärtus (tõepäral põhinev hinnang) oli üle kolme. Andmetöötluseks ja programmi jooksutamiseks on kasutatud dr Alena Kushniarevichi ja dr Märt Mölsi kirjutatud koode. Tulemused on visualiseeritud kasutades R tarkvara.

2.2.4. Meetod geenivoolu tuvastamiseks: TreeMix

Klassikalise dihhotoomse puu kasutamine liigisiseste populatsioonide põlvnemise uurimiseks pole geenivoolu tõttu mõistlik. Tegelikkusele vastab pigem võrgustik, kus lisaks populatsioonide lahknemisele on ära toodud ka populatsioonide segunemised. Selleks, et suure hulga geneetiliste andmete abil leida parim selline võrgustik, lõid Pickrell ja Pritchard (2012) tarkvarapaketi TreeMix. Programm leiab esmalt suurima tõepära meetodil andmetega hästi klappiva puu. Seejärel vaadatakse jääkide kovariatsiooni maatriksilt, milline populatsioon kõige halvemini puu topoloogiasse sobitub ning lisatakse tema ja mõne teise praeguse või oletatava eellaspopulatsiooni vahele joon, mis tähistab migratsiooni ja parandab andmete sobivust puu topoloogiaga. Pärast seda leitakse järgmine populatsioon, mille geneetilised andmed kõige halvemini puu topoloogiaga sobivad ja korratakse protsessi. Seda tehakse seni, kuni saavutatakse soovitud migratsioonide arvuga puu. Kui puule palju migratsioone lisada, muutub see raskesti jälgitavaks.

Antud magistritöös kasutatakse TreeMix tarkvara, et tuvastada, millised populatsioonid on segunenud mundadega. Programmi jooksutati andmestikul, millest tugevas aheldatuse tasakaalutuses olevad saidid olid eelnevalt kärbitud. Programmi autorid on soovitanud isegi pärast sellist töötlust eeldada lähestikku asuvate saitide aheldatust ja seda programmi jooksutamisel vastava parameetriga veelgi vähendada. Seetõttu on antud juhul LD akna suuruseks võetud 280 polümorfset saiti. Esmalt jooksutati programmi kaasates kõik populatsioonid (va kaks Aafrika populatsiooni) kuni kümne migratsiooni lisamiseni puule. Kuna programm ei leidnud, et nende kümne esimese lisatava migratsiooni seas oleks oluline lisada mundadega seotud migratsioon, siis eemaldati mõned populatsioonid valimist (Euroopa, Kaukaasia, Lähis-Ida), mis mundade kontekstis pole olulised ning jooksutati programmi uuesti, seekord kuni 15 migratsiooni lisamiseni.

2.2.5. Meetod segunemisaja määramiseks: ALDER

Selleks, et määrata, millal toimus populatsioonide segunemine, mis kulmineerus mundade tekkega, on antud magistritöös kasutatud tarkvara ALDER (Loh *et al.*, 2013). Tegu on Moorjani *et al.* (2011) ja Patterson *et al.* (2012) meetodi ROLLOFF edasiarendusega, mis kasutab informatsiooni lookuste omavahelise seotuse, täpsemalt segunemisest põhjustatud LD kohta. Segunenud populatsiooni kromosoomid koosnevad pikkadest DNA plokkidest, mis on päritud ühelt või teiselt algselt populatsioonilt. Rekombinatsiooni käigus need plokid katkevad ning muutuvad lühemaks jättes sellega maha märgi segunemisest möödunud ajast. Kui ALDERi algoritmile anda ette uuritav (segunenud) populatsioon ja üle kahe referentspopulatsiooni, siis on töö käik järgmine. Esmalt eemaldatakse analüüsist referentspopulatsioonid, millel esinevad pikad LD plokid, mis on tugevalt korreleeritud testpopulatsiooni omadega – ilmselt on tegu hiljutise populatsioonide lahknemise, mitte kunagise segunemisega. Teise sammuna leitakse kõigi alles jäänud populatsioonide ja testpopulatsiooni vahel kaalutud LD väärtused ning nende põhjal LD lagunemiskõverad. Kui kõverat pole võimalik leida, siis referentspopulatsioon eemaldatakse edasisest analüüsist. Sellisel juhul on suure tõenäosusega tegu geneetiliselt kauge populatsiooniga. Kolmanda sammuna moodustatakse alles jäänud referentspopulatsioonidest kahekaupa kõikvõimalikud paarid ning leitakse kahe referentspopulatsiooni põhjal arvutatud kaalutud LD väärtused ning nende põhjal LD lagunemiskõverad. Saadud lagunemiskõveraid võrreldakse varem leitud üksikpopulatsiooni lagunemiskõveratega ning väga erinevate kõveratega populatsioonipaaridele juhitakse tähelepanu. Need populatsioonid, vähemalt koos, ei sobi referentspopulatsioonideks. Kui aga mõlema paaris oleva referentspopulatsiooni individuaalse kõvera amplituud erineb vähem kui 25 protsenti, loetakse populatsioonide LD ajalugu testpopulatsiooni suhtes piisavalt sarnaseks (st mõlema populatsiooni LD

lagunemiskõver viitab segunemisele ligikaudu samas ajavahemikus), et kaaluda neid populatsioone kui uuritava populatsiooni tekke põhjustanud segunemispopulatsioone.

ALDERiga samal ajal tuli välja ka teine meetod segunemisaja määramiseks, StepPCO (Pugach *et al.*, 2011), kuid Yunusbayev *et al.* (2015) uuringus selgus, et hiljutise segunemise puhul annab see meetod kallutatud tulemusi ning ka vanemate segunemisaegade puhul ei ole StepPCO nii täpne kui ALDER, mistõttu otsustati selles töös kasutada just viimast meetodit.

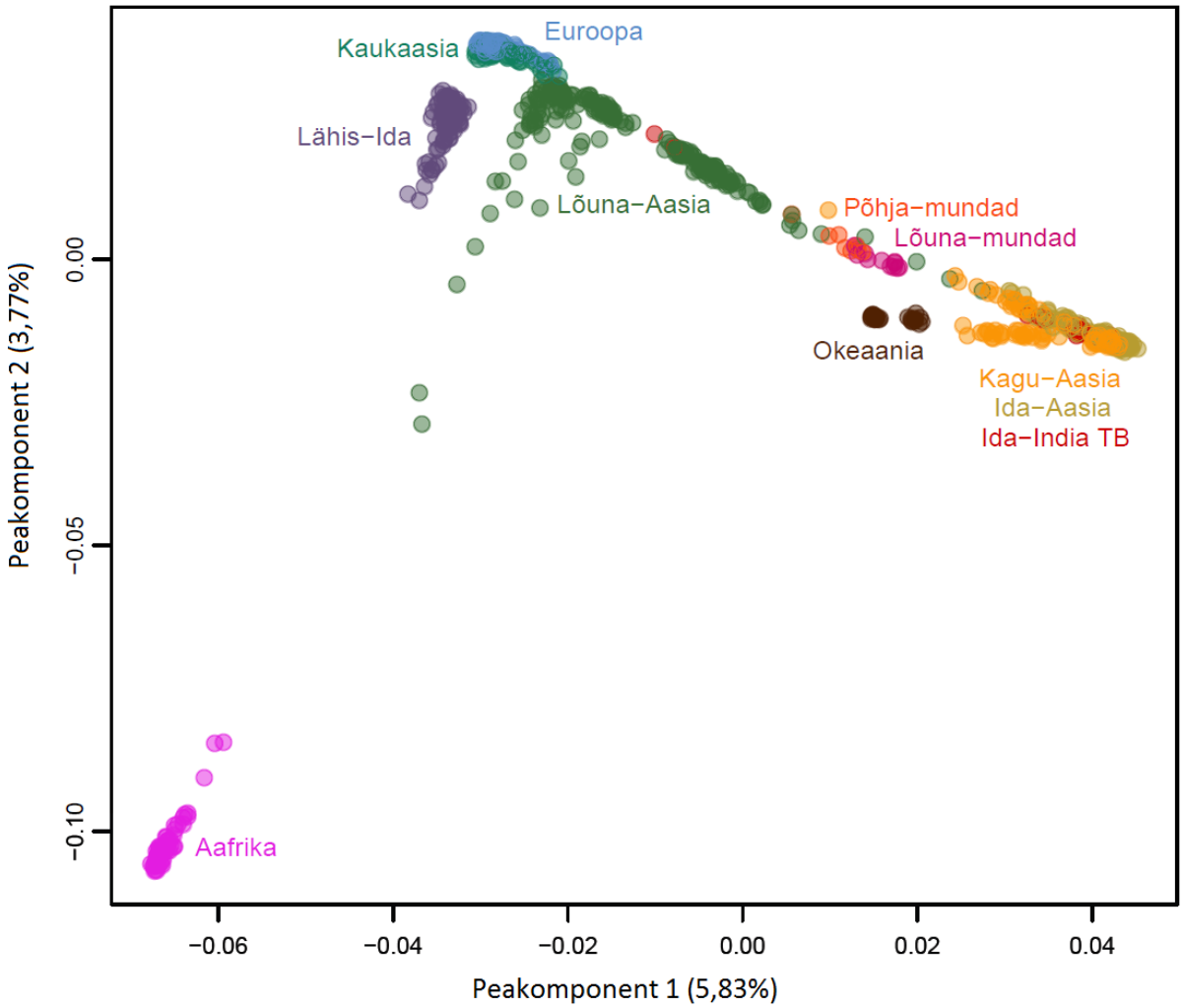
Andmete sobivale kujule viimine ja ALDERi jooksutamine on läbi viidud Anne-Mai Ilumäe koostatud skriptide abil.

Käesoleva töö analüüsid on läbi viidud Tartu Ülikooli teadusarvutuste keskuses (High Performance Computing Center of University of Tartu) ja Eesti Biokeskuse arvutusklastris (EBC Core Computing Unit).

2.3. Tulemused ja arutelu

2.3.1. Ülevaade populatsioonide geneetilisest struktuurist

Andmestikust esmase ülevaate saamiseks viidi läbi peakomponentanalüüs. Esimesed kaks komponenti kirjeldavad vastavalt 5,83% ja 3,77% summaarsest geneetilisest varieeruvusest. Mõlema komponendi puhul eristuvad teistest populatsioonidest kõige rohkem Aafrika populatsioonid: mandenkad, jorubad ja bantud. Esimesel põhikomponendil esineb lääne-ida suunaline gradient, st telje ühes otsas on Aafrika, Lähis-Ida, Kaukaasia ja Euroopa populatsioonid ning teises otsas Ida- ja Kagu-Aasia populatsioonid. Mundad ei kattu sel gradiendil otseselt ühegi teise populatsiooniga, nad jäävad ühelt poolt Lõuna-Aasia populatsioonide ning teiselt poolt Melaneesia ja Kagu-Aasia populatsioonide vahele. Põhjamundad on lõunamundadest sarnasemad teistele Lõuna-Aasia populatsioonidele, st paiknevad lääne pool sel gradiendil (joonis 7). Lõuna-Aasia populatsioonidest on mundadele sarnaseimad Lõuna-India draviidi keelt kõnelev paniya populatsioon ja lingvistiline isolaat



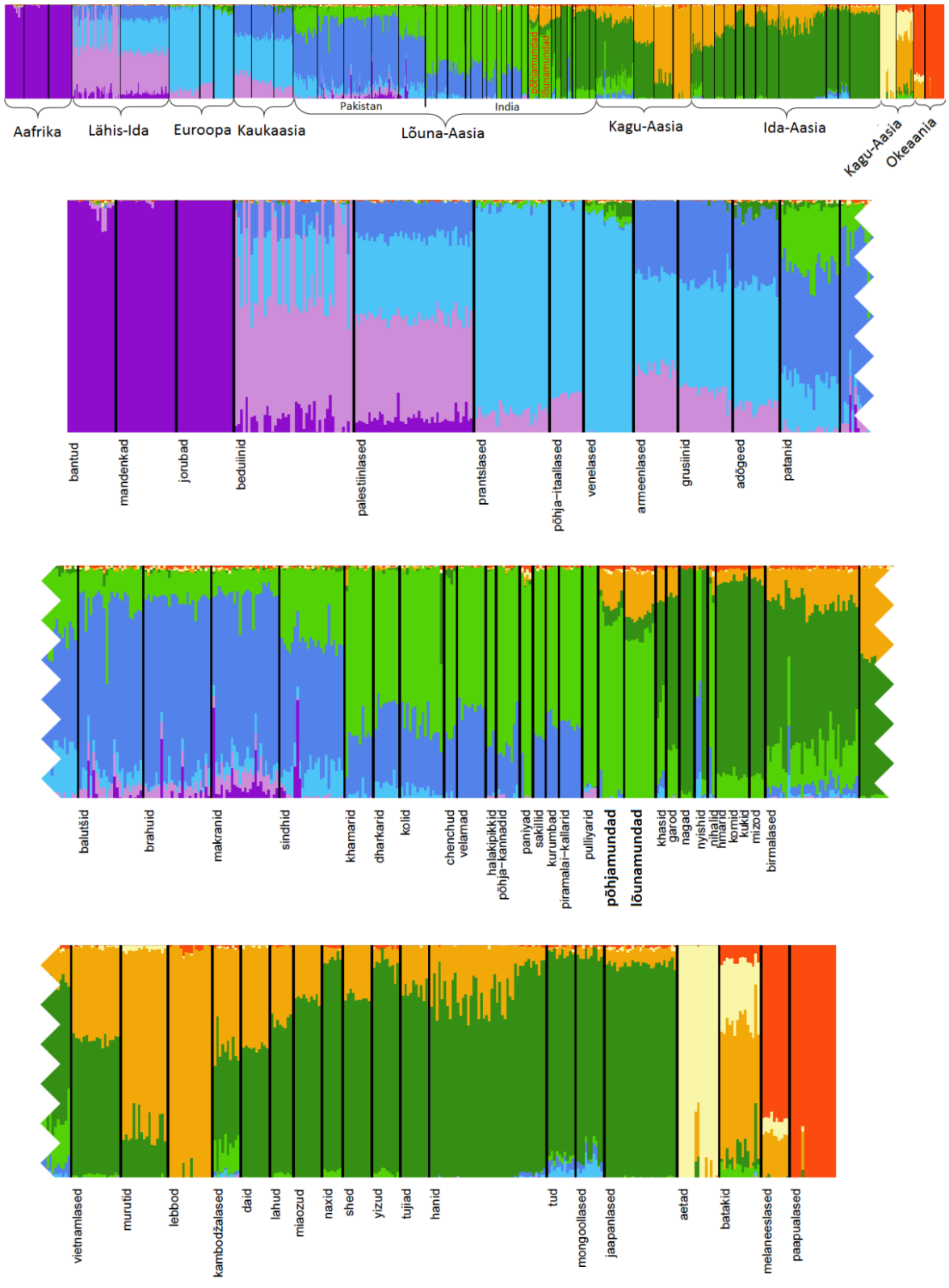
Joonis 7. Peakomponentanalüüsi kaks kõige rohkem varieeruvust kirjeldavat peakomponenti. Erinevat värvi ringidega on tähistatud erinevatest piirkondadest pärit indiviidid. Populatsiooninimedega joonis on toodud lisas 2.

Kesk-Indiast nihali. Segadust võib tekitada tiibeti-birma keelt kõnelev nyishi populatsioon, mille mediaanväärtus asetseb mundade juures (lisa 2). Tegelikult ükski indiviid siiski joonisel mundade juures ei asu. Nimelt paigutuvad osad nyishi indiviidid teiste Ida-India tiibeti-birma populatsioonide juurde, aga mõned sarnanevad esimese põhikomponendi poolest hoopis dharkari ja teiste indoeuroopa või draviidi keeli kõnelevate Lõuna-Aasia populatsioonidega ning selle lahknevuse tõttu satub nyishi mediaanväärtus mundade vahele. Järgnevad kaheksa enim varieeruvust kirjeldavat peakomponenti mundade kohta uudset informatsiooni ei anna.

Selleks, et saada ülevaade valimis olevate populatsioonide struktuurist, viidi läbi ADMIXTURE analüüs, mille tulemused on toodud joonisel 8. Sobivaimaks eellaspopulatsioonide arvuks (K) osutus 9, tulemused K=2 kuni K=19 on esitatud lisas 3a.

K=9 juures iseloomustab mundasid peamiselt 3 eellaskomponenti: heleroheline, oranž ja tumeroheline. Heleroheline komponent on iseloomulik Lõuna-Aasia populatsioonidele – suure osakaaluga esineb seda India rahvastel (v.a Ida-India tiibeti-birma keelte kõnelejad), väiksema osakaaluga Pakistani populatsioonidel. Märkimisväärne on see, et mundadel puudub sinine komponent, mis teistel Lõuna-Aasia populatsioonidel on olemas. Tumeroheline ja oranž komponent on levinud Ida- ja Kagu-Aasia rahvaste seas. Oranžil komponendil on veidi suurem osakaal Kagu-Aasia populatsioonide geneetilises profiilis ning tumerohelisel komponendil Ida-Aasia populatsioonide omas, kuid peaaegu kõigis Ida- ja Kagu-Aasia populatsioonides on mõlemad komponendid olemas. Seega on tõenäoline, et mundad said need komponendid korraga mõnelt Ida- või Kagu-Aasia populatsioonilt. Mundadele sarnaseim tumerohelise ja oranži komponendi suhe esineb Borneo saarel elavatel muruttidel. Tõenäolisem on siiski, et mundad said need komponendid mõnelt mandripopulatsioonilt, kellest sarnaseima komponentide suhtega on tänapäeval vietnamlased, kambodžalased ja daid. ADMIXTURE tulemuste põhjal võib spekuleerida, et varem kattis Kagu-Aasiat oranž komponent, kuid Ida-Aasias levinud tumeroheline komponent on aja jooksul Kagu-Aasias oma proportsioone suurendanud. Mundad võisid Indiasse migreeruda enne tumerohelise komponendi suuremat levikut Kagu-Aasias.

Eeldades, et mundade populatsioon on tekkinud India ja Ida-/Kagu-Aasia populatsiooni segunemisel, annavad ADMIXTURE tulemused erinevate komponentide osakaalu kaudu aimu segunemisproportsioonidest. Kuna ei ole teada, milline K väärtus selleks konkreetseks uurimisküsimuseks parim on, ei saa neid proportsioone väga tõsiselt võtta (eellaskomponentide arvu muutusega hinnatakse praegused komponendid samuti ümber ning seetõttu komponentide proportsioonid erinevad natuke erinevate K väärtuste juures). Kuna on täheldatud, et kaugete populatsioonide ristumisel on esimese põlvkonna järglastel ADMIXTURE tulemustes selgelt



Joonis 8. Ülemine riba kujutab ADMIXTURE analüüsi tulemusi 9 eellaskomponendi puhul (K=9). Alumised kolm riba on sama joonise piirkonnad suurendatuna. Eellaspopulatsioonidest pärit komponendid on tähistatud erinevate värvidega. Iga vertikaalne joon tähistab üht indiviidi, mustad jooned eraldavad populatsioone. Ülemisel joonisel on ära toodud piirkonnad, kus populatsioonid asuvad ning alumistel joonistel on näha populatsioonide nimed.

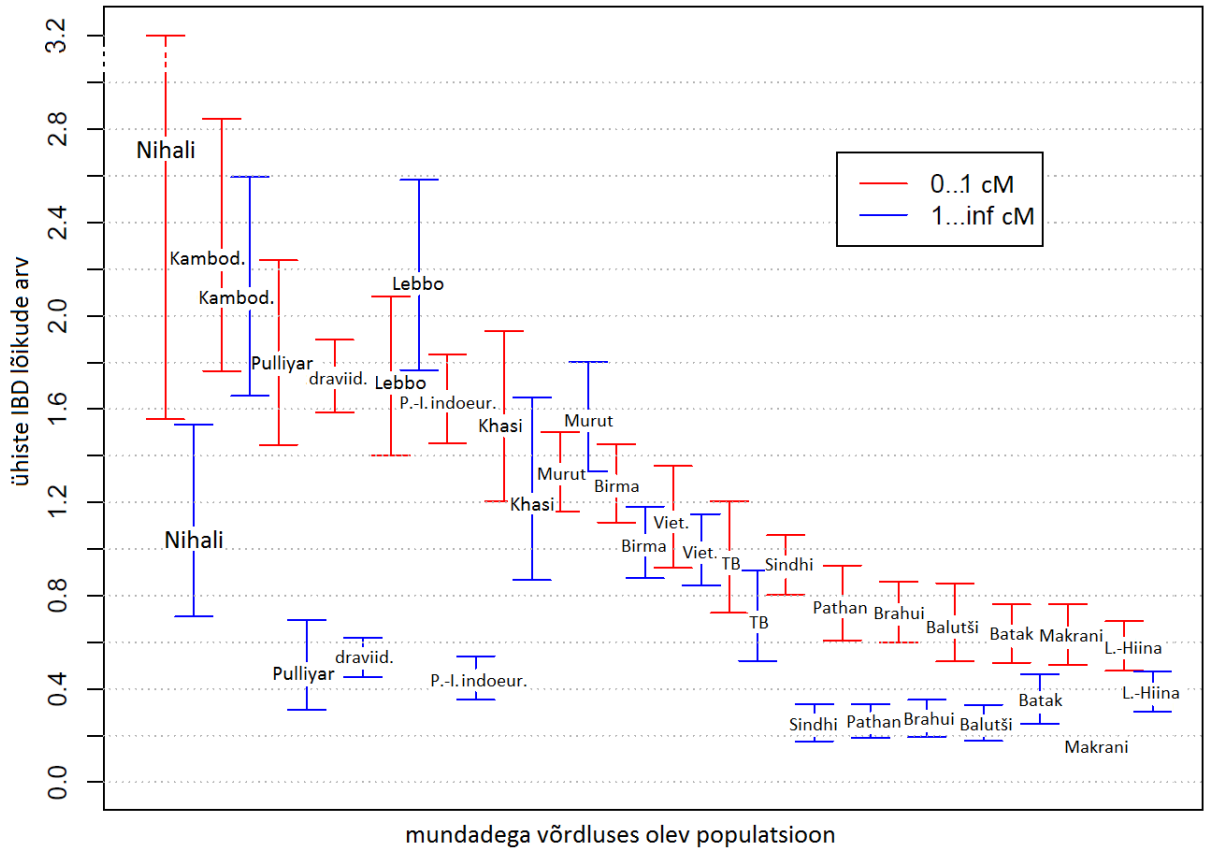
50% komponente ühelt eellaspopulatsioonilt ning 50% komponente teiselt, siis saab neid proportsioone käsitleda siiski informatiivsetena. Valimis olevatel põhjamundadel varieerub Lõuna-Aasia komponendi (heleroheline) osakaal 73 ja 80 protsendi vahel, Ida- ja Kagu-Aasia komponentide (tumeroheline ja oranž) osakaal on 11–24%. Lõunamundadel on Lõuna-Aasia komponenti 67–75% ning Ida- ja Kagu-Aasias levinud komponente 20–32%. Mõlematel munda populatsioonidel esineb 0–2% paapualaste eellaskomponenti (punane).

ADMIXTURE tulemustest leiti, et muidu homogeenses kolide populatsioonis on üks indiviid segunenud mõne idapoolse populatsiooniga ning kannab 40% ulatuses tumerohelist komponenti. Seetõttu on see indiviid järgmistest analüüsidest välja jäetud.

2.3.2. Geneetiliselt sarnased populatsioonid

Täpsema ülevaate saamiseks sellest, milliste populatsioonide genoomidest saaks kokku panna mundade genoomi, on kasutatud fineSTRUCTURE tarkvara. Tulemused on esitatud suure maatriksina, kus iga ruuduke on värvitud vastavalt kindlale eeskirjale (ingl *heatmap*). *Heatmap*i vasakul servas on esitatud iga indiviid kui DNA juppide doonor, üleval servas aga kui vastuvõtja. Kui vaadata näiteks mundade horisontaalset triipu, siis näeme sellel erineva värviga alasid, mis on värvitud vastavalt sellele, mitme DNA jupi doonoriks mundad üleval real olevatele vastuvõtjatele sobiksid. Värvide tähendust saab kindlaks teha paremal servas olevalt skaalalt. Mida sinisem/tumedam on piirkond, seda rohkem esineb maatriksis ristuvate populatsioonide vahel jagatud DNA juppe. Üleval ja ka vasakul on ära toodud sarnasusdendrogramm, mis sorteerib sarnased indiviidid gruppidesse. Vastavalt sellele on *heatmap*il ühisest grupist pärit indiviidide ridade ja ühisest grupist pärit indiviidide veergude kohtumisala värv keskmistatud. Joonise suurte mõõtmete tõttu on see kättesaadav vaid internetist aadressilt kodu.ut.ee/~kairemm/Joonis_fS.png. Joonisel on erinevatest piirkondadest pärit indiviidid tähistatud erinevate värvidega. Seda, millisesse populatsiooni täpsemalt mingi konkreetne indiviid kuulub, saab vaadata tabelist kodu.ut.ee/~kairemm/Tabel_fS.xlsx. Vaadates joonisel mundade vertikaalset riba, on näha, et teistest punasema ala moodustavad ehk rohkemate DNA juppide doonoriks sobivad India indoeuroopa ning draviidi keeli kõnelevad populatsioonid. Oranže ruute mundadega moodustavad ehk keskmiselt head doonorid oleks birmalased ning khasid ja garod Ida-Indiast, samuti Pakistani indoeuroopa keeli kõnelevad populatsioonid. Vaadates horisontaalset mundade riba näeme muidu sama pilti, kuid rohkem oranži Ida- ja Kagu-Aasia populatsioonide juures. See tähendab, et mundad sobivad DNA juppide doonoriks paremini kui vastuvõtjaks nende populatsioonide puhul. Ülejäänud populatsioonide puhul sobivad mundad doonoriks ja vastuvõtjaks sama hästi.

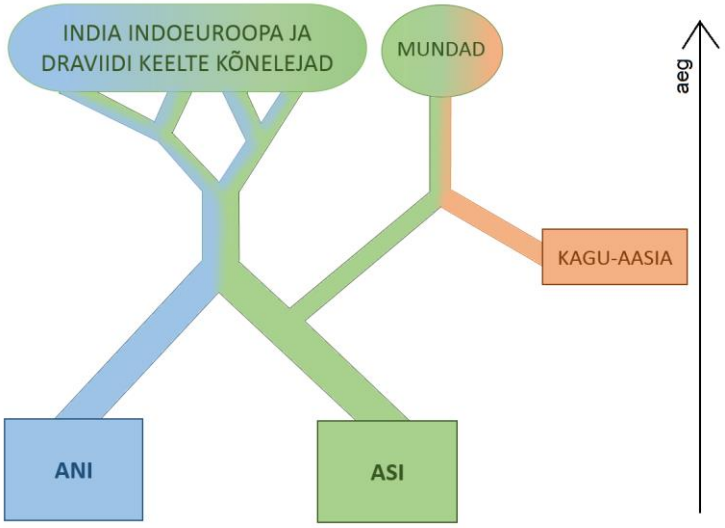
Selleks, et leida populatsioone, mille indiviidide genoomid sisaldavad sama päritoluga DNA lõike kui mundade genoom, viidi läbi IBD analüüs nimega Refined IBD. Suurema valimi saamiseks liideti lõuna- ja põhjamundad kokku üheks populatsiooniks. Sama tehti ka teiste populatsioonidega, kes elavad lähestikku, räägivad sarnast keelt ning ADMIXTURE tulemuste põhjal üksteisest märkimisväärselt ei erine. Mundadega jagatud IBD lõikude arvu mediaanväärtus on iga populatsiooni(grupi) puhul välja toodud kahes klassis: kuni 1 cM pikkused DNA lõigud ja pikemad lõigud. Tulemustest (joonis 9) selgub, et enim lühikesi IBD lõike jagavad mundadega kesk-India Madhya Pradeshi osariigis elavad nihalid. Kuna valimis oli vaid kaks nihalit, on jagatud IBD lõikude mediaanväärtusel selle populatsiooni puhul väga suured 95%-lised usaldusintervallid. Nihalisid ei saanud valimi suurendamiseks mõne teise populatsiooniga liita, sest tegu on lingvistilise isolaadiga. Nii lühikeste kui ka pikkade jagatud IBD lõikude arvu poolest paistavad silma kambodžalased ja lebbod. Borneo saarel elavad hõimurahvad lebbod ja murutid on ainsad populatsioonid, kes jagavad mundadega rohkem pikki IBD lõike kui lühikesi. India draviidi (sh pulliyar) ja indoeuroopa keeli kõnelevate rahvastega on mundadel vähe ühiseid pikki IBD lõike, mis viitab sellele, et hiljuti neil kokkupuudet ei ole olnud. Lühikesi mundadega jagatud IBD lõike on neil Indiat katvatel



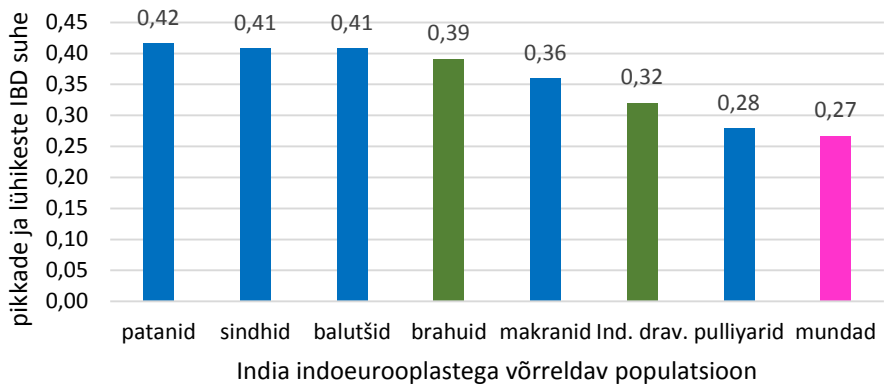
Joonis 9. Refined IBD analüüsi tulemused. Mundadega jagatud pikkade ja lühikeste IBD lõikude mediaanväärtus ja usaldusintervallid populatsiooniti. draviid. – draviidi keeli kõnelevad India populatsioonid; P.-I. indoeur. – Põhja-India indoeuroopa keelte kõnelejad; TB – Ida-India tiibeti-birma keelte kõnelejad

populatsioonidel aga palju, millest võib järeldada ammust ühist päritolu. Ida-Indias elavad tiibeti-birma keelte kõnelejad jagavad mundadega pikki ühist päritolu DNA lõike rohkem kui teised India rahvad, kuid lühikesi lõike jällegi vähem. Refined IBD tulemusi vaadati ka 2 cM ja pikemate IBD lõikude läbilõikes, kuid leiti, et nii pikki lõike ei jaga mundadega peaaegu ükski populatsioon. Siiski kambodžalastel, lebbodel ja muruttidel leidus keskmiselt 0,2 nii pikka IBD lõiku, mis viitab väga hiljutisele kokkupuutele nende populatsioonide vahel. Populatsioonid, kellel detekteeritavad mundadega ühist päritolu DNA lõigud peaaegu puudusid, on jooniselt välja jäetud. Sinna kuulusid näiteks Euroopa, Kaukaasia, Lähis-Ida, Aafrika ja Okeaania populatsioonid, aga ka mitmed Hiina populatsioonid, jaapanlased ja põlised Filipiinde elanikud aetad.

Selle analüüsi tulemused viitavad sellele, et mundad on India draviidi ja indoeuroopa keelte kõnelejatest lahknenud varem kui toimus segunemine Kagu-Aasia rahvastega, sest mundad jagavad viimastega pikemaid IBD lõike. ALDERi analüüs näitas, et mundade India komponendi segunemine Kagu-Aasia komponendiga lõppes 2250-3700 aastat tagasi (vt ptk 2.3.4.). ADMIXTURE analüüsist on teada, et mundadel puudub üks geneetiline komponent, mis teistel indialastel esineb (ptk 2.3.1.) Lisaks on teada, et 1900–4200 aastat tagasi toimus Indias ANI-ASI eellaspopulatsioonide segunemine (ptk 1.3.2.3.). Kõik see viitab ajaloolisele stsenaariumile (joonis 10), milles mundad lahknesid India põliselanikest enne ANI ja ASI komponentide segunemist ning nende genoomi Indiast pärit osa (~75%) koosneb tänapäevani vaid põliste lõuna-indialaste (ASI) genoomist. Selle mudeli kohaselt, ei ole mundad peale Kagu-Aasiast pärit populatsiooniga segunemist märkimisväärselt segunenud India teiste populatsioonidega. Selle aspekti kontrollimiseks võrreldi India indoeuroopa ja draviidi keeli kõnelevate rahvaste omavaheliste jagatud IBD lõikude hulka. Selgus, et India draviidi keelte kõnelejad ja indoeuroopa keelte kõnelejad jagavad omavahel rohkem pikki ühise päritoluga DNA lõike kui mundad kummagi populatsioonide grupiga ning see kehtib ka siis, kui tulemused standardiseerida lühikeste IBD



Joonis 10. India draviidi, indoeuroopa ja munda keelte kõnelejate põlvnemise mudel



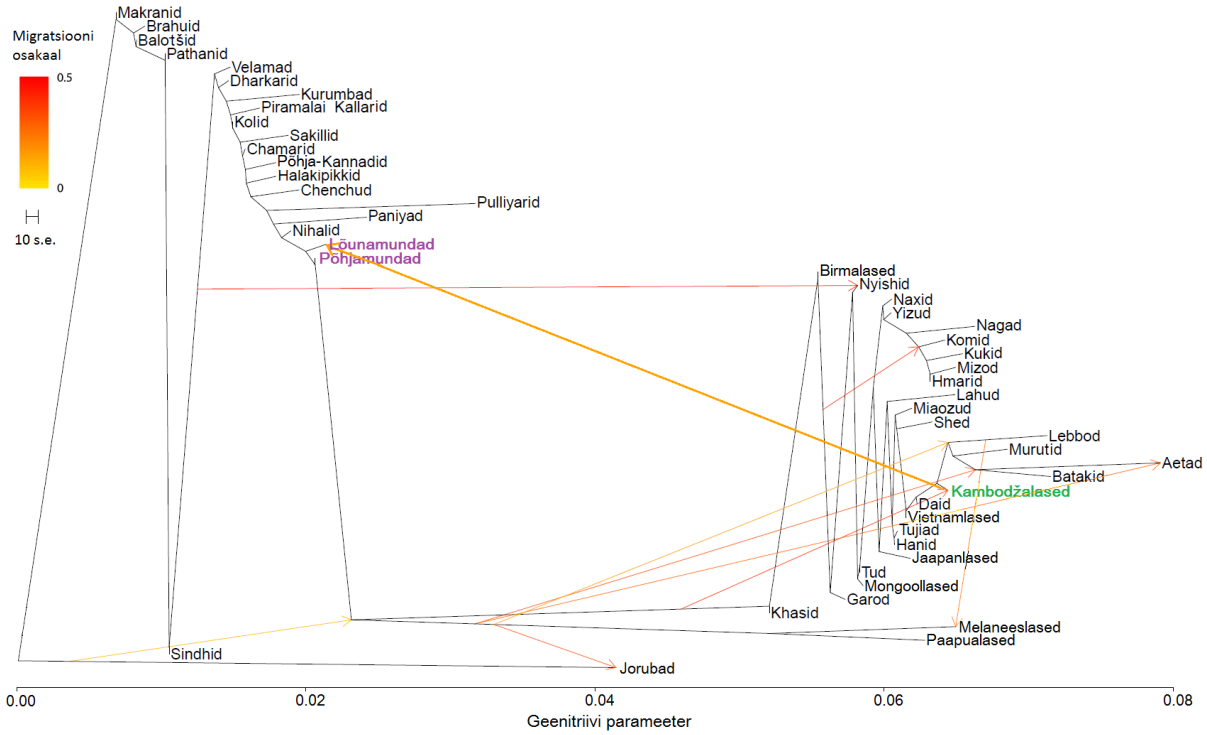
Joonis 11. India indoeuroopa keelte kõnelejate ja teiste India ning Pakistani populatsioonide vahelised IBD lõigud. Võrreldavuse eesmärgil on tulemused esitatud pikkade ja lühikeste lõikude pikkuste suhtena. Draviidi keelte kõnelejad on tähistatud rohelisega ning indoeuroopa keelte kõnelejad on tähistatud sinisega.

lõikude kaudu (joonis 11), võttes arvesse mundade segunemist ja sellest tulenevat väiksemat India komponendi osakaalu genoomis. Kusjuures India draviidi keeli kõnelevad populatsioonid ja indoeuroopa keeli kõnelevad populatsioonid asuvad keskmiselt geograafiliselt üksteisest kaugemal kui kumbki populatsioonide grupp mundadest. See leid on kooskõlas ülalkirjeldatud mudeliga, kuid ei välista ka mõningaid teisi võimalusi ning vajab edasist uurimist.

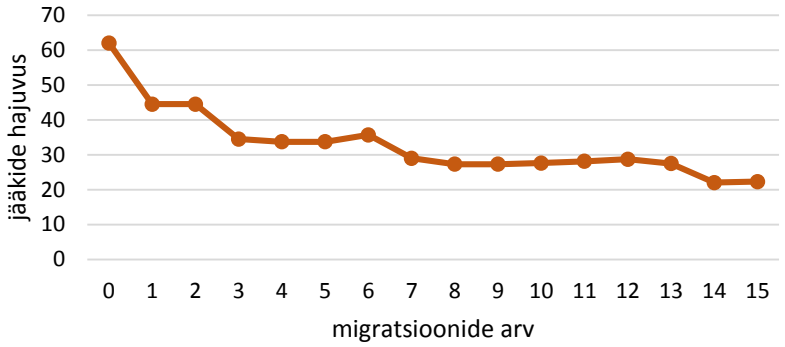
Antud magistritöös näidati esmakordselt mundadega jagatud IBD pikkusdiferentseeritust. Tulevikus on plaanis kasutada sarnasel põhimõttel töötavat meetodit DiCal-IBD, millega saab lisaks ajas toimunud populatsioonide suurusmuutusi leida ning populatsioonigeneetiliste sündmuste aegu määrata (Tataru *et al.*, 2014).

2.3.3. Ajaloos toimunud migratsioonid

Küsimusele, millised migratsioonid ja segunemised on toimunud valimis olevate populatsioonide ja/või nende eellaspopulatsioonide vahel, otsiti vastust TreeMixi abil. Esimest mundadega seotud migratsiooni oli näha 10. migratsiooni lisamisel puule. See migratsioon näitab geenivoolu kamboodžalastelt lõunamundadele (joonis 12). Siinkohal tuleb märkida, et programmi autorid on öelnud, et kõige rohkem vigu teeb programm migratsiooni suuna määramisel (Pickrell & Pritchard, 2012). Põhjamundasid see migratsioon üllatuslikult ei puuduta. Järgnevate nelja migratsiooni lisamine puule mundasid ei puuduta. Alles 15. migratsiooni lisamisel ilmneb uus mundadega seotud migratsioon. See näitab geenivoolu lõunamundadest Kagu-Aasia saarte põliselanike aetade, batakkide, lebbode ja muruttide ühise eellase juurde. Põhjamundad jäävad TreeMixi tulemuste põhjal ilma igasugustest viidetest migratsioonile. Üks põhjus võib olla selles, et mundad paiknevad puul niigi Ida-Aasia ja Okeaaniaga ühes klaadis, mis vähendab vajadust lisada migratsioonijooni nende piirkondade ja mundade vahele.



Joonis 12. TreeMixi poolt tuletatud kümne migratsioonisündmusega puu. Migratsiooninoole värv viitab migratsiooni osakaalule segunenud populatsioonis. Kui see on suurem kui 0,5, siis tehakse puu ümber ilma migratsiooni lisamata. Puu harude pikkused (horisontaalsel teljel) on proportsionaalsed harus toimunud geneenitriiviga. Skaala näitab vaatluste kovariatsiooni-maatriksist leitud kümnekordset keskmist standardviga.

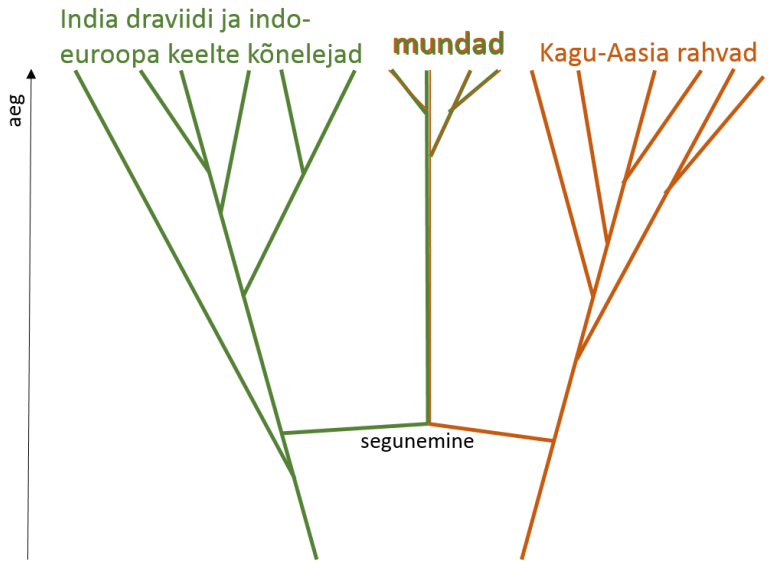


Joonis 13. Jääkide hajuvus muutub hüppeliselt 1., 3., 7. ja 14. migratsiooni-noole lisamisel puule. Need migratsioonid on puu topoloogia parandamiseks olulisemad.

Jääkide maatriksi hajuvusest (joonis 13) on näha, et kumbki lõunamundadega seotud migratsioon (10., 15.) pole väga oluline, sest ei muuda puu topoloogiat nii palju paremaks, et keskmine jääkide väärtus eriti muutuks.

2.3.4. Segunenud populatsioonid ja segunemisajad

Ükski seni läbiviidud analüüs pole olnud sobilik mundade tekkega päädinud populatsioonide segunemisaja leidmiseks. Sobilikke eellaspopulatsioone otsiti ning segunemisaega hinnati antud töös programmi ALDER abil. Algsest valimist on välja jäetud Kaukaasia ja Lähis-Ida populatsioonid ning itaallased, venelased, bantud ja mandenkad. Kas valimi väiksuse või muude eripärade tõttu ei leitud põhjamundadele vähemalt kaht sobivat referentspopulatsiooni. Analüüsi tulemusel leiti, et vaid vietnamlased oleks sobilikuks referentspopulatsiooniks, kuid kuna teine sobilik referentspopulatsioon puudus, jäi segunemisaeg leidmata. Lõunamundade puhul leiti 25 potentsiaalset referentspopulatsiooni ehk sellist populatsiooni, mis testpopulatsiooniga moodustavad LD lagunemiskõvera. Need populatsioonid on daid, hanid, lahud, miaozud, naxid, shed, tud, tujiad, yizud, komid, kukid, brahuid, balutšid, nagad, makranid, halakipikkid, sakillid, patanid, sindhid, velamad, dharkarid, aetad, batakid, lebbod ja kambodžalased. Kõiki neid populatsioone prooviti kahekaupa sobitada lõunamundade eellaspopulatsioonideks, kuid osad paarid ei andnud üldse ühist LD lagunemiskõverat, osade puhul olid üksikkõverad liialt erinevad. Lõpuks jäi alles 19 võimalikku referentspopulatsioonide paari. Need paarid ning neile vastavad segunemisajad on esitatud tabelis 2. Nagu tabelist näha, on igas paaris üks populatsioon Indiast/Pakistanist ja teine Ida/Kagu-Aasiast. Populatsioonide segunemisajaks on hinnatud 2250–3700 aastat tagasi, mida tuleks võtta kui segunemise lõppaega. Segunemisaegade seas esineb seaduspära – kui referentspopulatsiooniks on halakipikkid, siis segunemisaja hinnang on varasem. Halakipikkid on ka geograafiliselt kaugem populatsioon paiknedes Karnataka osariigis.



Joonis 14. Joonis näitlikustamaks seda, miks ei saa kaht konkreetset tänapäevast populatsiooni lugeda mundade eellasteks. Tegu on illustratiivse joonisega, mis ei peegelda täpselt ajalugu ega tulemusi.

Miks ei suutnud ALDER üht referentspopulatsioonide paari välja valida? Kuna segunemine toimus 2–4 tuhat aastat tagasi, siis algsetes segunenud populatsioonides on pärast seda

toimunud palju muutusi läbi populatsioonide lahknemiste ja geenitriivi ning mitmed tänapäevased populatsioonid võivad olla kunagise mundade segunemises osalenud populatsiooni järglased (joonis 14). Seetõttu pole imekspandav, et ALDER valis välja mitu sobivat referentspopulatsioonide paari ja seega ei saagi vaid kaht konkreetset tänapäevast populatsiooni mundade eellasteks lugeda.

Tabel 2. ALDERi poolt leitud populatsioonipaarid, mis sobivad kõige paremini mundade eellaspopulatsioonideks. p-väärtus referentspopulatsioonide segunemisele ja selle tulemusena mundade tekkele on leitud z-skooride põhjal, mis omakorda on leitud LD kõvera amplituudi ja LD lagunemismäära standardvea põhjal, mis on leitud jackknife meetodil. Segunemisaeg generatsioonides on leitud kahe ref. populatsiooni LD kõvera amplituudi põhjal, vastav usaldusvahemik aga üksikute ref. populatsioonide LD kõvera amplituudide erinevuste põhjal. Segunemisaeg on aastatesse teisendatud eeldades 25-aastast põlvkonna pikkust. Populatsiooninimed on värvitud vastavalt kõneldavale keelele: sinine – indoeuroopa keeled, roheline – draviidi keeled, oranž – austroneesia keeled, punane – tiibeti-birma keeled, violetne – muud Hiinas kõneldavad keeled, tumekollane – austroaasia keeled.

ref. pop. 1	ref. pop. 2	p-väärtus	segunemisaeg			
			gen. tagasi	+/-	aastat tagasi	+/-
dharkarid	aetad	$6{,}7\times10^{-6}$	90,06	15,74	2252	394
patanid	nagad	$3{,}2\times10^{-6}$	93,83	13,95	2346	349
sindhid	nagad	0,0014	95,23	15,52	2381	388
dharkarid	tud	$4{,}3\times10^{-5}$	95,69	15,46	2392	387
velamad	nagad	0,0013	97,76	17,55	2444	439
bolutšid	nagad	0,00048	98,05	14,86	2451	372
velamad	tud	0,017	103,00	24,27	2575	607
dharkarid	nagad	0,00048	108,02	17,86	2701	447
velamad	lahud	0,029	111,77	27,7	2794	693
velamad	daid	0,00063	115,19	23,72	2880	593
velamad	komid	0,0077	115,69	21,97	2892	549
velamad	shed	$5{,}2\times10^{-6}$	117,64	17,95	2941	449
halakipikkid	murutid	0,008	124,81	22,82	3120	571
halakipikkid	miaozud	0,023	125,95	23,88	3149	597
halakipikkid	daid	0,0003	127,69	20,03	3192	501
halakipikkid	naxid	0,022	131,89	23,27	3297	582
halakipikkid	lahud	0,011	132,87	29,18	3322	730
halakipikkid	tujiad	0,00062	133,66	21,01	3342	525
halakipikkid	kambodž.	0,015	147,89	28,37	3697	709

Antud töös hinnati esmakordselt mundade tekkega päädinud segunemise aega ülegenoomsete andmete põhjal. Varasemad arheoloogilised hinnangud mundade vanusele on olnud vanemad: Diamond ja Bellwood (2003) on leidnud, et mundad saabusid Indiasse umbes 5000 aastat tagasi tuues kaasa riisikasvatuse kunsti. Fulleri arheolingvistiliste uuringute põhjal puutusid mundad viimati kokku Kagu-Aasia austroaasia keelte kõnelejatega vähem kui 7000 aastat tagasi (Fuller, 2003, 2007). Y-kromosoomi uuringud pole praeguste tulemustega võrreldavad, sest annavad hinnangu mundadel levinud O2a haplogrupi lahknemisajale ülejäänud Kagu-Aasia haplogruppidest, mitte otsesele segunemisajale. Olgu siiski öeldud, et Y-kromosoomi haplogrupi lahknemine toimus hinnanguliselt 15 000 aastat tagasi, mis on seega ülemiseks piiriks mundade tekkele (Chaubey *et al.*, 2011).

KOKKUVÕTE

Käesoleva magistritöö eesmärgiks oli uurida Indias elavate austroaasia keeli kõnelevate hõimude, mundade, geneetilist ajalugu. Täpsemalt oli töö eesmärgiks esmakordselt kasutada ülegenoomseid andmeid, et leida mundadele sarnaseimad Aasia populatsioonid ning teada saada, millal toimus mundade tekkeni viinud segunemine India ja Kagu Aasia populatsioonide vahel.

Varasemast on teada, et mundade geneetilisest varieeruvusest umbes kolmveerand on jagatud teiste tänapäevaste indialastega ning ligi veerand pärineb Ida-/Kagu-Aasia populatsioonidelt (Chaubey *et al.*, 2011). Tegu pole aga klassikalise populatsioonide segunemisega, sest mundade genoomi indialastelt pärit osa ei ole sama struktuuriga nagu teistel India populatsioonidel ning ka Kagu-Aasia komponendi proportsioonid ei klapi ühegi maismaa Kagu-Aasia populatsiooniga. Seetõttu ei sobi ükski tänapäevane Aasia populatsioon otseselt mundade tekkega päädinud segunemise doonorpopulatsiooniks.

Antud magistritöö käigus selgus, et kuigi mundad jagavad indialastega palju ühise päritoluga DNA lõike, siis need lõigud on peamiselt lühikesed, mis viitab varasele lahknemisele ülejäänud indialastest, mitte hiljutisele kokkupuutele. Seevastu kambodžalaste ja Borneo saare hõimudega jagavad mundad palju nii pikki kui ka lühikesi ühise päritoluga DNA lõike, mis viitab hiljutisemale kokkupuutele.

Lisaks leiti antud töös, et mundade eellaspopulatsioonide segunemine võis lõppeda 2250–3700 aastat tagasi, mis on varasemate hinnangutega võrreldes noorem vanus. Samas pole varem ülegenoomsete andmete põhjal mundade segunemisaega määratud ning otseselt võrreldavaid uuringuid pole tehtud. Selles vahemikus (Moorjani *et al.*, 2013) toimus Indias ka ANI-ASI segunemine (vt ptk 1.3.2.3.). Arvestades, et mundade genoomis puudub ülejäänud indialaste seas levinud geneetiline komponent ning et mundade viimane kokkupuude teiste indialastega toimus varem kui segunemine Kagu-Aasia populatsiooniga, võib oletada, et mundad lahknesid põlistest India rahvastest enne ANI ja ASI komponentide segunemist Indias ning seega tänapäevaste indialaste teket. Selle mudeli kontrollimiseks võrreldi ülejäänud India populatsioonide omavahelist ühise päritoluga DNA lõikude jagamist. Selgus, et India draviidi keelte kõnelejad ja indoeuroopa keelte kõnelejad jagavad omavahel rohkem pikki ühise päritoluga DNA lõike kui mundad kummagi populatsioonide grupiga. Kusjuures need populatsioonide grupid asuvad geograafiliselt üksteisest keskmiselt kaugemal kui mundadest. See leid toetab ülalkirjeldatud stsenaariumit.

Genetic Ancestry of Indian Austroasiatic Speakers

Kai Tätte

SUMMARY

South Asia serves as a home for more than billion people who belong to thousands of diverse population groups with different culture, language, lifestyle and appearance. It has also been shown that South Asia is genetically the most diverse region after Africa (Xing *et al.*, 2010). Mitochondrial DNA analyses have shown that South Asia was the first region to have a fast population growth and diversification after exodus from Africa (Atkinson *et al.*, 2008; Basu *et al.*, 2003; Kivisild *et al.*, 2003). For these reasons, South Asia is an area of great interest for studies of population genetics.

The purpose of this master's thesis is to investigate the ancestry of Indian Austroasiatic speaking tribes – the Mundas. To be more precise, the goal was to use genome-wide data to detect the most similar populations to the Mundas and also to determine the admixture time that led to the genesis of the Mundas as we know them today.

It is known that about three quarters of the Mundas' genome consists of Indian heritage and about one quarter of East/Southeast Asian heritage (Chaubey *et al.*, 2011). This is not a typical recent admixture case as the Indian part of the Mundas' genome is missing a component that exists in all the other Indians. Moreover, the Southeast Asian part of the Mundas' genome does not tie in well with other Southeast Asian populations, too. Therefore, none of the modern populations in Asia work well as a donor population for the Mundas' admixture.

In current thesis, it was found that although the Mundas share a lot of DNA segments of identity by decent (IBD) with other Indians, these segments are very short. What this means is that the Mundas diverged from other Indians long time ago. On the other hand, IBD segments shared with Cambodians and Borneo tribes were long which means relatively recent contact.

Another novelty of this thesis is a finding that the admixture of the Mundas' ancestors ended 2250–3700 years ago. This date is more recent than previous estimates but as other studies have been based on archeology or uniparental markers, they are not comparable. During the same era (Moorjani *et al.*, 2013), Ancestral North Indians (ANI) admixed with Ancestral South Indians (ASI) giving Indians (except Mundas) their distinctive genetic profile (Metspalu *et al.*, 2011; Reich *et al.*, 2009). Considering that the Mundas lack one genetic component in their genome that all the other Indians have and also considering that the Mundas' last contact with

other Indians was before admixture with Southeast Asians, we can assume that the Mundas diverged from native Indians before ANI component admixed with ASI component in India. To examine this model, another IBD analysis was run to check IBD sharing between North-Indian Indo-European speaking populations and South-Indian Dravidian speaking populations. The results show that these two population groups share more long IBD segments than they share with the Mundas meaning they have had more recent contact than with the Mundas and therefore reassuring the abovementioned scenario.

TÄNUAVALDUSED

Tänan Alena Kushniarevichi, Gyaneshwer Chaubeyd ja Anne-Mai Ilumäed abi eest erinevate analüüside läbiviimisel, aga ka teisi kolleege evolutsioonilise bioloogia õppetoolist. Eriti tänulik olen oma juhendajale Mait Metspalule, kes oli ööpäevaringselt valmis mu küsimustele vastama ning andis palju kasulikku tagasisidet. Samuti olen tänulik oma abikaasale Kunterile hea nõu eest sõnastus- ja vormistusküsimuste puhul.

KASUTATUD KIRJANDUSE LOETELU

Alexander, D. H., Novembre, J. & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Res, 19*(9), 1655-1664.

Ambrose, S. H. (1998). Late Pleistocene human population bottlenecks, volcanic winter, and differentiation of modern humans. *J Hum Evol, 34*(6), 623-651.

Anderson, S., Bankier, A. T., Barrell, B. G., De Bruijn, M., Coulson, A. R., Drouin, J., Eperon, I., Nierlich, D., Roe, B. A. & Sanger, F. (1981). Sequence and organization of the human mitochondrial genome.

Andrews, R. M., Kubacka, I., Chinnery, P. F., Lightowlers, R. N., Turnbull, D. M. & Howell, N. (1999). Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nature genetics, 23*(2), 147-147.

Athreya, S. (2007). Was Homo heidelbergensis in South Asia? A test using the Narmada fossil from central India *The evolution and history of human populations in South Asia* (pp. 137-170): Springer.

Atkinson, Q. D., Gray, R. D. & Drummond, A. J. (2008). mtDNA variation predicts population size in humans and reveals a major Southern Asian chapter in human prehistory. *Mol Biol Evol, 25*(2), 468-474.

Bamshad, M., Kivisild, T., Watkins, W. S., Dixon, M. E., Ricker, C. E., Rao, B. B., Naidu, J. M., Prasad, B. R., Reddy, P. G. & Rasanayagam, A. (2001). Genetic evidence on the origins of Indian caste populations. *Genome Res, 11*(6), 994-1004.

Bandelt, H.-J., Richards, M. & Macaulay, V. (2006). *Human mitochondrial DNA and the evolution of Homo sapiens* (Vol. 18): Springer Science & Business Media.

Basu, A., Mukherjee, N., Roy, S., Sengupta, S., Banerjee, S., Chakraborty, M., Dey, B., Roy, M., Roy, B. & Bhattacharyya, N. P. (2003). Ethnic India: a genomic view, with special reference to peopling and structure. *Genome Res, 13*(10), 2277-2290.

Beja-Pereira, A., Luikart, G., England, P. R., Bradley, D. G., Jann, O. C., Bertorelle, G., Chamberlain, A. T., Nunes, T. P., Metodiev, S., Ferrand, N. & Erhardt, G. (2003). Gene-culture coevolution between cattle milk protein genes and human lactase genes. *Nature genetics, 35*(4), 311-313.

Bhattacharya, D. (1970). Indians of African origin. *Cahiers d'études africaines*, 579-582.

Brahmachari, S. K., Majumder, P. P., Mukerji, M., Habib, S., Dash, D., Ray, K. & Bahl, S. (2008). Genetic landscape of the people of India: a canvas for disease gene exploration. *Journal of genetics, 87*(1), 3-20.

Browning, B. L. & Browning, S. R. (2011). A fast, powerful method for detecting identity by descent. *Am J Hum Genet, 88*(2), 173-182.

Browning, B. L. & Browning, S. R. (2013). Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics, 194*(2), 459-471.

Browning, S. R. & Browning, B. L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet, 81*(5), 1084-1097.

Cameron, D., Patnaik, R. & Sahni, A. (2004). The phylogenetic significance of the Middle Pleistocene Narmada hominin cranium from central India. *International Journal of Osteoarchaeology, 14*(6), 419-447.

Chaubey, G. (2010). *The demographic history of India: A perspective based on genetic evidence.*

Chaubey, G., Karmin, M., Metspalu, E., Metspalu, M., Selvi-Rani, D., Singh, V. K., Parik, J., Solnik, A., Naidu, B. P., Kumar, A. *et al.* (2008a). Phylogeography of mtDNA haplogroup R7 in the Indian peninsula. *BMC Evol Biol, 8*, 227.

Chaubey, G., Metspalu, M., Choi, Y., Magi, R., Romero, I. G., Soares, P., van Oven, M., Behar, D. M., Rootsi, S., Hudjashov, G. *et al.* (2011). Population genetic structure in Indian Austroasiatic speakers: the role of landscape barriers and sex-specific admixture. *Mol Biol Evol, 28*(2), 1013-1024.

Chaubey, G., Metspalu, M., Karmin, M., Thangaraj, K., Rootsi, S., Parik, J., Solnik, A., Rani, D. S., Singh, V. K. & Naidu, B. P. (2008b). Language shift by indigenous population: a model genetic study in South Asia. *International Journal of Human Genetics, 8*(1/2), 41.

Chaubey, G., Metspalu, M., Kivisild, T. & Villems, R. (2007). Peopling of South Asia: investigating the caste-tribe continuum in India. *Bioessays, 29*(1), 91-100.

Colonna, V., Pagani, L., Xue, Y. & Tyler-Smith, C. (2011). A world in a grain of sand: human history from genetic data. *Genome Biol, 12*, 234.

Comas, D., Plaza, S., Wells, R. S., Yuldaseva, N., Lao, O., Calafell, F. & Bertranpetit, J. (2004). Admixture, migrations, and dispersals in Central Asia: evidence from maternal DNA lineages. *European Journal of Human Genetics, 12*(6), 495-504.

Cordaux, R., Weiss, G., Saha, N. & Stoneking, M. (2004). The northeast Indian passageway: a barrier or corridor for human migrations? *Mol Biol Evol, 21*(8), 1525-1533.

Coyne, J. A. & Hoekstra, H. E. (2007). Evolution of protein expression: new genes for a new diet. *Current Biology, 17*(23), R1014-1016.

Cruciani, F., Santolamazza, P., Shen, P., Macaulay, V., Moral, P., Olckers, A., Modiano, D., Holmes, S., Destro-Bisol, G., Coia, V. *et al.* (2002). A back migration from Asia to sub-Saharan Africa is supported by high-resolution analysis of human Y-chromosome haplotypes. *Am J Hum Genet, 70*(5), 1197-1214.

Cruciani, F., Trombetta, B., Massaia, A., Destro-Bisol, G., Sellitto, D. & Scozzari, R. (2011). A revised root for the human Y chromosomal phylogenetic tree: the origin of patrilineal diversity in Africa. *Am J Hum Genet, 88*(6), 814-818.

de Knijff, P. (2000). Messages through bottlenecks: on the combined use of slow and fast evolving polymorphic markers on the human Y chromosome. *Am J Hum Genet, 67*(5), 1055-1061.

Dennell, R., Rendell, H. & Hailwood, E. (1988). Late Pliocene artefacts from northern Pakistan. *Current Anthropology*, 495-498.

Dennell, R. W., Rendell, H. M., Halim, M. & Moth, E. (1992). A 45,000-year-old open-air Paleolithic site at Riwat, northern Pakistan. *Journal of Field Archaeology*, 17-33.

Diamond, J. & Bellwood, P. (2003). Farmers and their languages: the first expansions. *Science, 300*(5619), 597-603.

Eaaswarkhanth, M., Dubey, B., Meganathan, P. R., Ravesh, Z., Khan, F. A., Singh, L., Thangaraj, K. & Haque, I. (2009). Diverse genetic origin of Indian Muslims: evidence from autosomal STR loci. *J Hum Genet, 54*(6), 340-348.

Eaaswarkhanth, M., Haque, I., Ravesh, Z., Romero, I. G., Meganathan, P. R., Dubey, B., Khan, F. A., Chaubey, G., Kivisild, T., Tyler-Smith, C., Singh, L. & Thangaraj, K. (2010). Traces of sub-Saharan and Middle Eastern lineages in Indian Muslim populations. *Eur J Hum Genet, 18*(3), 354-363.

Ellis, N. & Hammer, M. F. (2002). A nomenclature system for the tree of human Y-chromosomal binary haplogroups. *Genome Res, 12*(2), 339-348.

Esposito, J. L. (1999). *The Oxford History of Islam*: Oxford University Press.

Field, J. S., Petraglia, M. D. & Lahr, M. M. (2007). The southern dispersal hypothesis and the South Asian archaeological record: Examination of dispersal routes through GIS analysis. *Journal of Anthropological Archaeology, 26*(1), 88-108.

Fu, Q., Li, H., Moorjani, P., Jay, F., Slepchenko, S. M., Bondarev, A. A., Johnson, P. L., Aximu-Petri, A., Prüfer, K. & de Filippo, C. (2014). Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature, 514*(7523), 445-449.

Fu, Q., Mittnik, A., Johnson, P. L., Bos, K., Lari, M., Bollongino, R., Sun, C., Giemsch, L., Schmitz, R., Burger, J. *et al.* (2013). A revised timescale for human evolution based on ancient mitochondrial genomes. *Current Biology, 23*(7), 553-559.

Fuller, D. Q. (2003). An agricultural perspective on Dravidian historical linguistics: archaeological crop packages, livestock and Dravidian crop vocabulary.

Fuller, D. Q. (2007). Non-human genetics, agricultural origins and historical linguistics in South Asia *The evolution and history of human populations in South Asia* (pp. 393-443): Springer.

Genomes Project, C., Abecasis, G. R., Altshuler, D., Auton, A., Brooks, L. D., Durbin, R. M., Gibbs, R. A., Hurles, M. E. & McVean, G. A. (2010). A map of human genome variation from population-scale sequencing. *Nature, 467*(7319), 1061-1073.

Gibbs, R. A., Belmont, J. W., Hardenbol, P., Willis, T. D., Yu, F., Yang, H., Ch'ang, L.-Y., Huang, W., Liu, B. & Shen, Y. (2003). The international HapMap project. *Nature, 426*(6968), 789-796.

Helgason, A., Einarsson, A. W., Guðmundsdóttir, V. B., Sigurðsson, Á., Gunnarsdóttir, E. D., Jagadeesan, A., Ebenesersdóttir, S. S., Kong, A. & Stefánsson, K. (2015). The Y-chromosome point mutation rate in humans. *Nature genetics, 47*(5), 453-457.

Hinds, D. A., Stuve, L. L., Nilsen, G. B., Halperin, E., Eskin, E., Ballinger, D. G., Frazer, K. A. & Cox, D. R. (2005). Whole-genome patterns of common DNA variation in three human populations. *Science, 307*(5712), 1072-1079.

Huang, X., Kurata, N., Wei, X., Wang, Z.-X., Wang, A., Zhao, Q., Zhao, Y., Liu, K., Lu, H. & Li, W. (2012). A map of rice genome variation reveals the origin of cultivated rice. *Nature, 490*(7421), 497-501.

Hudjashov, G., Kivisild, T., Underhill, P. A., Endicott, P., Sanchez, J. J., Lin, A. A., Shen, P., Oefner, P., Renfrew, C. & Villems, R. (2007). Revealing the prehistoric settlement of Australia by Y chromosome and mtDNA analysis. *Proceedings of the National Academy of Sciences, 104*(21), 8726-8730.

International HapMap, C., Frazer, K. A., Ballinger, D. G., Cox, D. R., Hinds, D. A., Stuve, L. L., Gibbs, R. A., Belmont, J. W., Boudreau, A., Hardenbol, P. *et al.* (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature, 449*(7164), 851-861.

James, H. A. & Petraglia, M. (2005). Modern Human Origins and the Evolution of Behavior in the Later Pleistocene Record of South Asia1. *Current Anthropology, 46*(S5), S3-S27.

Jeffreys, A. J., Kauppi, L. & Neumann, R. (2001). Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nature genetics, 29*(2), 217-222.

Jobling, M. A. & Tyler-Smith, C. (2003). The human Y chromosome: an evolutionary marker comes of age. *Nat Rev Genet, 4*(8), 598-612.

Karafet, T. M., Osipova, L. P., Gubina, M. A., Posukh, O. L., Zegura, S. L. & Hammer, M. F. (2002). High levels of Y-chromosome differentiation among native Siberian populations and the genetic signature of a boreal hunter-gatherer way of life. *Human biology, 74*(6), 761-789.

Karmin, M., Saag, L., Vicente, M., Sayres, M. A. W., Järve, M., Talas, U. G., Rootsi, S., Ilumäe, A.-M., Mägi, R. & Mitt, M. (2015). A recent bottleneck of Y chromosome diversity coincides with a global change in culture. *Genome Res*.

Kennedy, K. A. & Deraniyagala, S. U. (1989). Fossil remains of 28,000-year-old hominids from Sri Lanka. *Current Anthropology*, 394-399.

Kivisild, T., Rootsi, S., Metspalu, M., Mastana, S., Kaldma, K., Parik, J., Metspalu, E., Adojaan, M., Tolk, H.-V. & Stepanov, V. (2003). The genetic heritage of the earliest settlers persists both in Indian tribal and caste populations. *Am J Hum Genet, 72*(2), 313-332.

Lawson, D. J., Hellenthal, G., Myers, S. & Falush, D. (2012). Inference of population structure using dense haplotype data. *PLoS genetics, 8*(1), e1002453.

Li, J. Z., Absher, D. M., Tang, H., Southwick, A. M., Casto, A. M., Ramachandran, S., Cann, H. M., Barsh, G. S., Feldman, M. & Cavalli-Sforza, L. L. (2008). Worldwide human relationships inferred from genome-wide patterns of variation. *Science, 319*(5866), 1100-1104.

Loftus, R. T., MacHugh, D. E., Bradley, D. G., Sharp, P. M. & Cunningham, P. (1994). Evidence for two independent domestications of cattle. *Proceedings of the National Academy of Sciences, 91*(7), 2757-2761.

Loh, P.-R., Lipson, M., Patterson, N., Moorjani, P., Pickrell, J. K., Reich, D. & Berger, B. (2013). Inferring admixture histories of human populations using linkage disequilibrium. *Genetics, 193*(4), 1233-1254.

Loogväli, E.-L., Kivisild, T., Margus, T. & Villems, R. (2009). Explaining the imperfection of the molecular clock of hominid mitochondria. *PLoS One, 4*(12), e8260.

Macaulay, V., Hill, C., Achilli, A., Rengo, C., Clarke, D., Meehan, W., Blackburn, J., Semino, O., Scozzari, R. & Cruciani, F. (2005). Single, rapid coastal settlement of Asia revealed by analysis of complete mitochondrial genomes. *Science, 308*(5724), 1034-1036.

Majumder, P. P. (2001). Ethnic populations of India as seen from an evolutionary perspective. *Journal of Biosciences, 26*(4), 533-545.

McElreavey, K. & Quintana-Murci, L. (2005). A population genetics perspective of the Indus Valley through uniparentally-inherited markers. *Annals of human biology, 32*(2), 154-162.

Mellars, P., Gori, K. C., Carr, M., Soares, P. A. & Richards, M. B. (2013). Genetic and archaeological perspectives on the initial modern human colonization of southern Asia. *Proceedings of the National Academy of Sciences, 110*(26), 10699-10704.

Metspalu, M., Kivisild, T., Bandelt, H.-J., Richards, M. & Villems, R. (2006). The pioneer settlement of modern humans in Asia *Human Mitochondrial DNA and the Evolution of Homo sapiens* (pp. 181-199): Springer.

Metspalu, M., Kivisild, T., Metspalu, E., Parik, J., Hudjashov, G., Kaldma, K., Serk, P., Karmin, M., Behar, D. M., Gilbert, M. T. *et al.* (2004). Most of the extant mtDNA boundaries in south and southwest Asia were likely shaped during the initial settlement of Eurasia by anatomically modern humans. *BMC Genet, 5*, 26.

Metspalu, M., Romero, I. G., Yunusbayev, B., Chaubey, G., Mallick, C. B., Hudjashov, G., Nelis, M., Magi, R., Metspalu, E., Remm, M. *et al.* (2011). Shared and unique components of human population structure and genome-wide signals of positive selection in South Asia. *Am J Hum Genet, 89*(6), 731-744.

Migliano, A. B., Romero, I. G., Metspalu, M., Leavesley, M., Pagani, L., Antao, T., Huang, D.-W., Sherman, B. T., Siddle, K. & Scholes, C. (2013). Evolution of the pygmy phenotype: Evidence of positive selection from genome-wide scans in African, Asian, and Melanesian pygmies. *Human biology, 85*(1), 251-284.

Misra, V. (2001). Prehistoric human colonization of India. *Journal of Biosciences, 26*(4), 491-531.

Moorjani, P., Patterson, N., Hirschhorn, J. N., Keinan, A., Hao, L., Atzmon, G., Burns, E., Ostrer, H., Price, A. L. & Reich, D. (2011). The history of African gene flow into Southern Europeans, Levantines, and Jews. *PLoS genetics, 7*(4), e1001373.

Moorjani, P., Thangaraj, K., Patterson, N., Lipson, M., Loh, P. R., Govindaraj, P., Berger, B., Reich, D. & Singh, L. (2013). Genetic evidence for recent population mixture in India. *Am J Hum Genet, 93*(3), 422-438.

Moseley, C. (2008). *Encyclopedia of the world's endangered languages*: Routledge.

Narang, A., Jha, P., Rawat, V., Mukhopadhyay, A., Dash, D., Indian Genome Variation, C., Basu, A. & Mukerji, M. (2011). Recent admixture in an Indian population of African ancestry. *Am J Hum Genet, 89*(1), 111-120.

Nass, M. M. & Nass, S. (1963). Intramitochondrial fibers with DNA characteristics I. Fixation and electron staining reactions. *The Journal of cell biology, 19*(3), 593-611.

Novembre, J., Pritchard, J. K. & Coop, G. (2007). Adaptive drool in the gene pool. *Nature genetics, 39*(10), 1188-1190.

Oppenheimer, C. (2002). Limited global change due to the largest known Quaternary eruption, Toba≈ 74kyr BP? *Quaternary Science Reviews, 21*(14), 1593-1609.

Paddayya, K., Blackwell, B., Jhaldiyal, R., Petraglia, M., Fevrier, S., Chaderton, D., Blickstein, J. & Skinner, A. (2002). Recent findings on the Acheulian of the Hunsgi and Baichbal valleys, Karnataka, with special reference to the Isampur excavation and its dating. *Current Science, 83*(5), 641-647.

Palanichamy, M. G., Sun, C., Agrawal, S., Bandelt, H. J., Kong, Q. P., Khan, F., Wang, C. Y., Chaudhuri, T. K., Palla, V. & Zhang, Y. P. (2004). Phylogeny of mitochondrial DNA macrohaplogroup N in India, based on complete sequencing: implications for the peopling of South Asia. *Am J Hum Genet, 75*(6), 966-978.

Patnaik, R., Chauhan, P. R., Rao, M. R., Blackwell, B. A., Skinner, A. R., Sahni, A., Chauhan, M. S. & Khan, H. S. (2009). New geochronological, paleoclimatological, and archaeological data from the Narmada Valley hominin locality, central India. *J Hum Evol, 56*(2), 114-133.

Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., Webster, T. & Reich, D. (2012). Ancient admixture in human history. *Genetics, 192*(3), 1065-1093.

Patterson, N., Price, A. L. & Reich, D. (2006). Population structure and eigenanalysis. *PLoS genetics, 2*(12), e190.

Perry, G. H., Dominy, N. J., Claw, K. G., Lee, A. S., Fiegler, H., Redon, R., Werner, J., Villanea, F. A., Mountain, J. L. & Misra, R. (2007). Diet and the evolution of human amylase gene copy number variation. *Nature genetics, 39*(10), 1256-1260.

Petraglia, M., Clarkson, C., Boivin, N., Haslam, M., Korisettar, R., Chaubey, G., Ditchfield, P., Fuller, D., James, H., Jones, S. *et al.* (2009). Population increase and environmental deterioration correspond with microlithic innovations in South Asia ca. 35,000 years ago. *Proc Natl Acad Sci U S A, 106*(30), 12261-12266.

Pickrell, J. K. & Pritchard, J. K. (2012). Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS genetics, 8*(11), e1002967.

Pugach, I., Matveyev, R., Wollstein, A., Kayser, M. & Stoneking, M. (2011). Dating the age of admixture via wavelet transform analysis of genome-wide data. *Genome Biol, 12*(2), R19.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., De Bakker, P. I. & Daly, M. J. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet, 81*(3), 559-575.

Rai, N., Chaubey, G., Tamang, R., Pathak, A. K., Singh, V. K., Karmin, M., Singh, M., Rani, D. S., Anugula, S. & Yadav, B. K. (2012). The phylogeography of Y-chromosome haplogroup h1a1a-m82 reveals the likely Indian origin of the European Romani populations. *PLoS One, 7*(11), e48477.

Rasmussen, M., Guo, X., Wang, Y., Lohmueller, K. E., Rasmussen, S., Albrechtsen, A., Skotte, L., Lindgreen, S., Metspalu, M., Jombart, T. *et al.* (2011). An Aboriginal Australian genome reveals separate human dispersals into Asia. *Science, 334*(6052), 94-98.

Reich, D., Thangaraj, K., Patterson, N., Price, A. L. & Singh, L. (2009). Reconstructing Indian population history. *Nature, 461*(7263), 489-494.

Romero, I. G., Mallick, C. B., Liebert, A., Crivellaro, F., Chaubey, G., Itan, Y., Metspalu, M., Eaaswarkhanth, M., Pitchappan, R. & Villems, R. (2011). Herders of Indian and European cattle share their predominant allele for lactase persistence. *Mol Biol Evol*, msr190.

Rosenberg, N. A., Mahajan, S., Gonzalez-Quevedo, C., Blum, M. G., Nino-Rosales, L., Ninis, V., Das, P., Hegde, M., Molinari, L. & Zapata, G. (2006). Low levels of genetic divergence across geographically and linguistically diverse populations from India. *PLoS genetics*, *2*(12), e215.

Rosser, Z. H., Zerjal, T., Hurles, M. E., Adojaan, M., Alavantic, D., Amorim, A., Amos, W., Armenteros, M., Arroyo, E. & Barbujani, G. (2000). Y-chromosomal diversity in Europe is clinal and influenced primarily by geography, rather than by language. *Am J Hum Genet, 67*(6), 1526-1543.

Sahoo, S., Singh, A., Himabindu, G., Banerjee, J., Sitalaximi, T., Gaikwad, S., Trivedi, R., Endicott, P., Kivisild, T., Metspalu, M., Villems, R. & Kashyap, V. K. (2006). A prehistory of Indian Y chromosomes: evaluating demic diffusion scenarios. *Proc Natl Acad Sci U S A, 103*(4), 843-848.

Scally, A. & Durbin, R. (2012). Revising the human mutation rate: implications for understanding human evolution. *Nature Reviews Genetics, 13*(10), 745-753.

Seielstad, M. T., Minch, E. & Cavalli-Sforza, L. L. (1998). Genetic evidence for a higher female migration rate in humans. *Nature genetics, 20*(3), 278-280.

Semino, O., Santachiara-Benerecetti, A. S., Falaschi, F., Cavalli-Sforza, L. L. & Underhill, P. A. (2002). Ethiopians and Khoisan share the deepest clades of the human Y-chromosome phylogeny. *Am J Hum Genet, 70*(1), 265-268.

Sengupta, S., Zhivotovsky, L. A., King, R., Mehdi, S. Q., Edmonds, C. A., Chow, C. E., Lin, A. A., Mitra, M., Sil, S. K., Ramesh, A. *et al.* (2006). Polarity and temporality of high-resolution y-chromosome distributions in India identify both indigenous and exogenous expansions and reveal minor genetic influence of Central Asian pastoralists. *Am J Hum Genet, 78*(2), 202-221.

Shah, A. M., Tamang, R., Moorjani, P., Rani, D. S., Govindaraj, P., Kulkarni, G., Bhattacharya, T., Mustak, M. S., Bhaskar, L. V., Reddy, A. G. *et al.* (2011). Indian Siddis: African descendants with Indian admixture. *Am J Hum Genet, 89*(1), 154-161.

Sharma, G., Tamang, R., Chaudhary, R., Singh, V. K., Shah, A. M., Anugula, S., Rani, D. S., Reddy, A. G., Eaaswarkhanth, M. & Chaubey, G. (2012). Genetic affinities of the central Indian tribal populations. *PLoS One, 7*(2), e32546.

Skaletsky, H., Kuroda-Kawaguchi, T., Minx, P. J., Cordum, H. S., Hillier, L., Brown, L. G., Repping, S., Pyntikova, T., Ali, J. & Bieri, T. (2003). The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature, 423*(6942), 825-837.

Soares, P., Alshamali, F., Pereira, J. B., Fernandes, V., Silva, N. M., Afonso, C., Costa, M. D., Musilová, E., Macaulay, V. & Richards, M. B. (2011). The expansion of mtDNA haplogroup L3 within and out of Africa. *Mol Biol Evol*, msr245.

Soares, P., Ermini, L., Thomson, N., Mormina, M., Rito, T., Rohl, A., Salas, A., Oppenheimer, S., Macaulay, V. & Richards, M. B. (2009). Correcting for purifying selection: an improved human mitochondrial molecular clock. *Am J Hum Genet, 84*(6), 740-759.

Sonakia, A. & Kennedy, K. A. (1985). Skull cap of an early man from the Narmada valley alluvium (Pleistocene) of central India. *American Anthropologist, 87*(3), 612-616.

Su, B., Xiao, C., Deka, R., Seielstad, M. T., Kangwanpong, D., Xiao, J., Lu, D., Underhill, P., Cavalli-Sforza, L. & Chakraborty, R. (2000). Y chromosome haplotypes reveal prehistorical migrations to the Himalayas. *Human genetics, 107*(6), 582-590.

Zegura, S. L., Karafet, T. M., Zhivotovsky, L. A. & Hammer, M. F. (2004). High-resolution SNPs and microsatellite haplotypes point to a single, recent entry of Native American Y chromosomes into the Americas. *Mol Biol Evol, 21*(1), 164-175.

Zerjal, T., Wells, R. S., Yuldasheva, N., Ruzibakiev, R. & Tyler-Smith, C. (2002). A genetic landscape reshaped by recent events: Y-chromosomal insights into central Asia. *Am J Hum Genet, 71*(3), 466-482.

Takahata, N. & Satta, Y. (1997). Evolution of the primate lineage leading to modern humans: phylogenetic and demographic inferences from DNA sequences. *Proceedings of the National Academy of Sciences, 94*(9), 4811-4815.

Tamang, R. & Thangaraj, K. (2012). Genomic view on the peopling of India. *Investigative genetics, 3*(1), 20.

Tataru, P., Nirody, J. A. & Song, Y. S. (2014). diCal-IBD: demography-aware inference of identity-by-descent tracts in unrelated individuals. *Bioinformatics, 30*(23), 3430-3431.

Thangaraj, K., Sridhar, V., Kivisild, T., Reddy, A. G., Chaubey, G., Singh, V. K., Kaur, S., Agarawal, P., Rai, A. & Gupta, J. (2005). Different population histories of the Mundari-and Mon-Khmer-speaking Austro-Asiatic tribes inferred from the mtDNA 9-bp deletion/insertion polymorphism in Indian populations. *Human genetics, 116*(6), 507-517.

Thomson, R., Pritchard, J. K., Shen, P., Oefner, P. J. & Feldman, M. W. (2000). Recent common ancestry of human Y chromosomes: evidence from DNA sequence data. *Proceedings of the National Academy of Sciences, 97*(13), 7360-7365.

Tishkoff, S. A., Reed, F. A., Ranciaro, A., Voight, B. F., Babbitt, C. C., Silverman, J. S., Powell, K., Mortensen, H. M., Hirbo, J. B., Osman, M. *et al.* (2007). Convergent adaptation of human lactase persistence in Africa and Europe. *Nature genetics, 39*(1), 31-40.

Torroni, A., Achilli, A., Macaulay, V., Richards, M. & Bandelt, H.-J. (2006). Harvesting the fruit of the human mtDNA tree. *TRENDS in Genetics, 22*(6), 339-345.

Torroni, A., Schurr, T. G., Cabell, M. F., Brown, M. D., Neel, J. V., Larsen, M., Smith, D. G., Vullo, C. M. & Wallace, D. C. (1993). Asian affinities and continental radiation of the four founding Native American mtDNAs. *Am J Hum Genet, 53*(3), 563.

Underhill, P. A. & Kivisild, T. (2007). Use of y chromosome and mitochondrial DNA population structure in tracing human migrations. *Annu Rev Genet, 41*, 539-564.

Underhill, P. A., Myres, N. M., Rootsi, S., Metspalu, M., Zhivotovsky, L. A., King, R. J., Lin, A. A., Chow, C.-E. T., Semino, O. & Battaglia, V. (2010). Separating the post-Glacial coancestry of European and Asian Y chromosomes within haplogroup R1a. *European Journal of Human Genetics, 18*(4), 479-484.

Underhill, P. A., Poznik, G. D., Rootsi, S., Jarve, M., Lin, A. A., Wang, J., Passarelli, B., Kanbar, J., Myres, N. M., King, R. J. *et al.* (2015). The phylogenetic and geographic structure of Y-chromosome haplogroup R1a. *Eur J Hum Genet, 23*(1), 124-131.

Underhill, P. A., Shen, P., Lin, A. A., Jin, L., Passarino, G., Yang, W. H., Kauffman, E., Bonné-Tamir, B., Bertranpetit, J. & Francalacci, P. (2000). Y chromosome sequence variation and the history of human populations. *Nature genetics, 26*(3), 358-361.

Van Oven, M. & Kayser, M. (2009). Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Human mutation, 30*(2), E386-E394.

Watkins, W. S., Thara, R., Mowry, B. J., Zhang, Y., Witherspoon, D. J., Tolpinrud, W., Bamshad, M. J., Tirupati, S., Padmavati, R., Smith, H., Nancarrow, D., Filippich, C. & Jorde, L. B. (2008). Genetic variation in South Indian castes: evidence from Y-chromosome, mitochondrial, and autosomal polymorphisms. *BMC Genet, 9*, 86.

Xing, J., Watkins, W. S., Hu, Y., Huff, C. D., Sabo, A., Muzny, D. M., Bamshad, M. J., Gibbs, R. A., Jorde, L. B. & Yu, F. (2010). Genetic diversity in India and the inference of Eurasian population expansion. *Genome Biol, 11*(11), R113.

Xue, Y., Wang, Q., Long, Q., Ng, B. L., Swerdlow, H., Burton, J., Skuce, C., Taylor, R., Abdellah, Z. & Zhao, Y. (2009). Human Y chromosome base-substitution mutation rate measured by direct sequencing in a deep-rooting pedigree. *Current Biology, 19*(17), 1453-1457.

Yunusbayev, B., Metspalu, M., Järve, M., Kutuev, I., Rootsi, S., Metspalu, E., Behar, D. M., Varendi, K., Sahakyan, H. & Khusainova, R. (2012). The Caucasus as an asymmetric semipermeable barrier to ancient human migrations. *Mol Biol Evol, 29*(1), 359-365.

Yunusbayev, B., Metspalu, M., Metspalu, E., Valeev, A., Litvinov, S., Valiev, R., Akhmetova, V., Balanovska, E., Balanovsky, O., Turdikulova, S. *et al.* (2015). The Genetic Legacy of the Expansion of Turkic-Speaking Nomads across Eurasia. *PLoS genetics*, *11*(4), e1005068.

KASUTATUD VEEBIAADRESSID

World Population Prospects, http://esa.un.org/unpd/wpp/unpp/panel_population.htm (25.01.2015)

Ethnologue: Languages of the World, http://www.ethnologue.com (27.01.2015)

Census of India 2011, http://www.censusindia.gov.in/Census_Data_2001/India_at_glance/religion.aspx (27.01.2015)

Encyclopedia Britannica, www.britannica.com/EBchecked/topic/397427/Munda (11.04.15)

Encyclopedia Britannica, http://www.britannica.com/EBchecked/topic/397435/Munda-languages (11.04.15)

LISAD

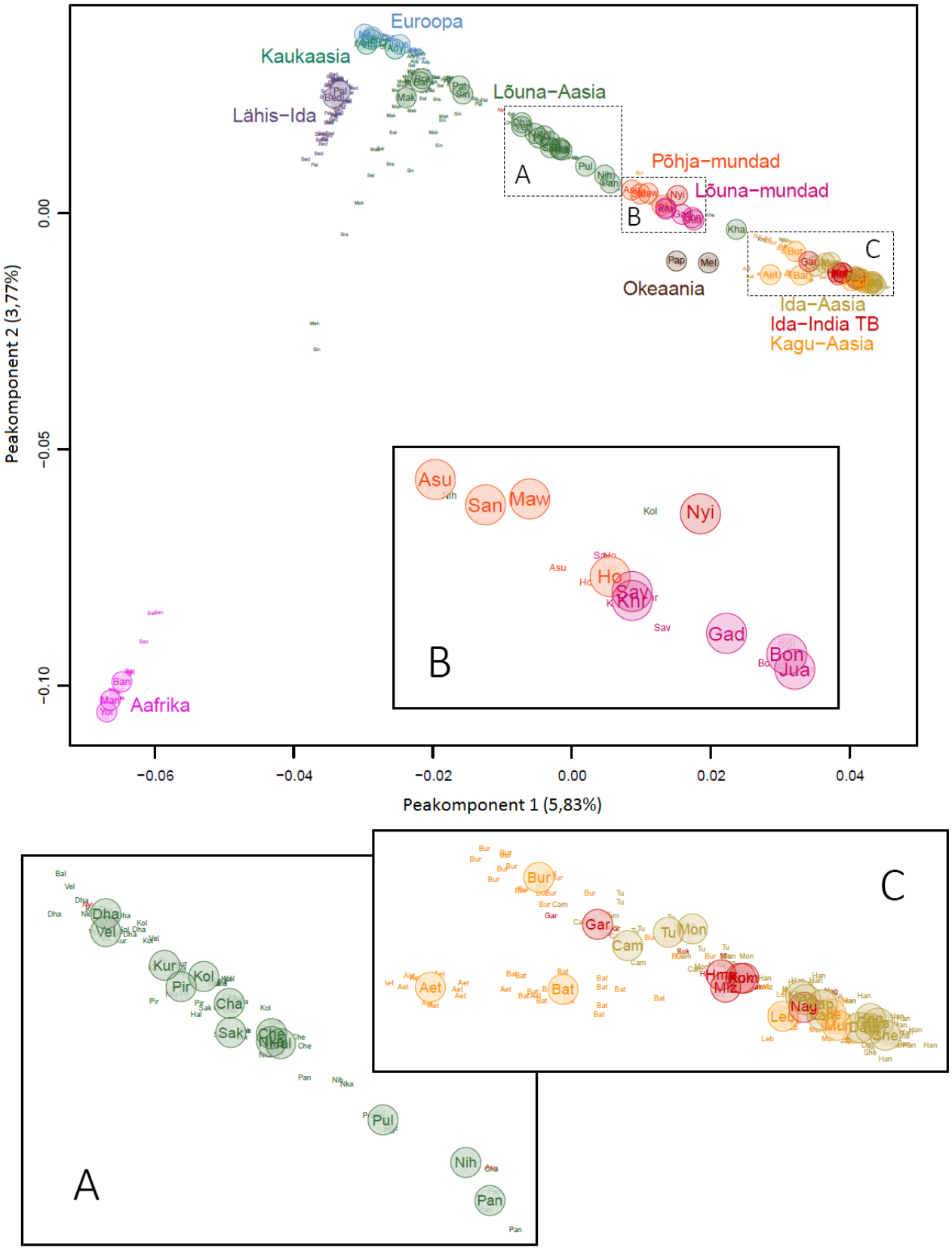
Lisa 1

Ülevaade valimist. [&] – põhjamunda, * – lõunamunda, N – arv valimis

Populatsioon	N	Riik (piirkond)	Keel	Allikas
Aafrika (N = 61)				
bantud	18	Lõuna-Aafrika, Keenia	nigeri-kongo	Li *et al* 2008
mandenkad	22	Senegal	nigeri-kongo	Li *et al* 2008
jorubad	21	Nigeeria	nigeri-kongo	Li *et al* 2008
Kaukaasia (N = 53)				
adõgeed	17	Venemaa (Põhja-Kaukaasia)	kaukaasia	Li *et al* 2008
armeenlased	16	Armeenia	indoeuroopa	Yunusbayev *et al* 2012
grusiinid	20	Gruusia	kaukaasia	Behar *et al* 2010
Ida-Aasia (N = 166)				
kambodžalased	10	Kambodža	austroaasia	Li *et al* 2008
daid	10	Hiina	tai-kadai	Li *et al* 2008
hanid	44	Hiina	hiina-tiibeti	Li *et al* 2008
jaapanlased	27	Jaapan	jaapani	Li *et al* 2008
lahud	8	Hiina	tiibeti-birma	Li *et al* 2008
miaozud	10	Hiina	miao-jao	Li *et al* 2008
mongoollased	10	Hiina	altai	Li *et al* 2008
naxid	7	Hiina	tiibeti-birma	Li *et al* 2008
shed	10	Hiina	miao-jao	Li *et al* 2008
tud	10	Hiina	altai	Li *et al* 2008
tujiad	10	Hiina	tiibeti-birma	Li *et al* 2008
yizud	10	Hiina	tiibeti-birma	Li *et al* 2008
Euroopa (N = 58)				
prantslased	28	Prantsusmaa	indoeuroopa	Li *et al* 2008
põhja-itaallased	12	Itaalia	indoeuroopa	Li *et al* 2008
venelased	18	Venemaa	indoeuroopa	Yunusbayev *et al* 2013
Lähis-Ida (N = 90)				
beduiinid	45	Iisrael	afroaasia	Li *et al* 2008
palestiinlased	45	Iisrael	afroaasia	Li *et al* 2008
Okeaania (N = 27)				
melaneeslased	10	Paapua Uus-Guinea	paapua	Li *et al* 2008
paapuad	17	Paapua Uus-Guinea	paapua	Li *et al* 2008
Lõuna-Aasia (N = 260)				
asurid[&]	2	India	austroaasia (munda)	Chaubey 2010
hod[&]	5	India	austroaasia (munda)	Chaubey 2010
mawasid[&]	1	India	austroaasia (munda)	Metspalu *et al* 2011
santhalid[&]	1	India	austroaasia (munda)	Chaubey 2010
bondad*	4	India	austroaasia (munda)	Chaubey 2010
gadabad*	1	India	austroaasia (munda)	Chaubey 2010
juangid*	2	India	austroaasia (munda)	Chaubey 2010
khariad*	2	India	austroaasia (munda)	Chaubey 2010
savarad*	2	India	austroaasia (munda)	Chaubey 2010
chamarid	10	India	indoeuroopa	Metspalu *et al* 2011
chenchud	4	India	draviidi	Metspalu *et al* 2011
dharkarid	9	India	indoeuroopa	Metspalu *et al* 2011
garod	4	India	tiibeti-birma	Chaubey 2010
halakipikkid	3	India	draviidi	Metspalu *et al* 2011
hmarid	4	India	tiibeti-birma	EBK (avaldamata)
khasid	3	India	austroaasia (khasi-asli)	Chaubey 2010

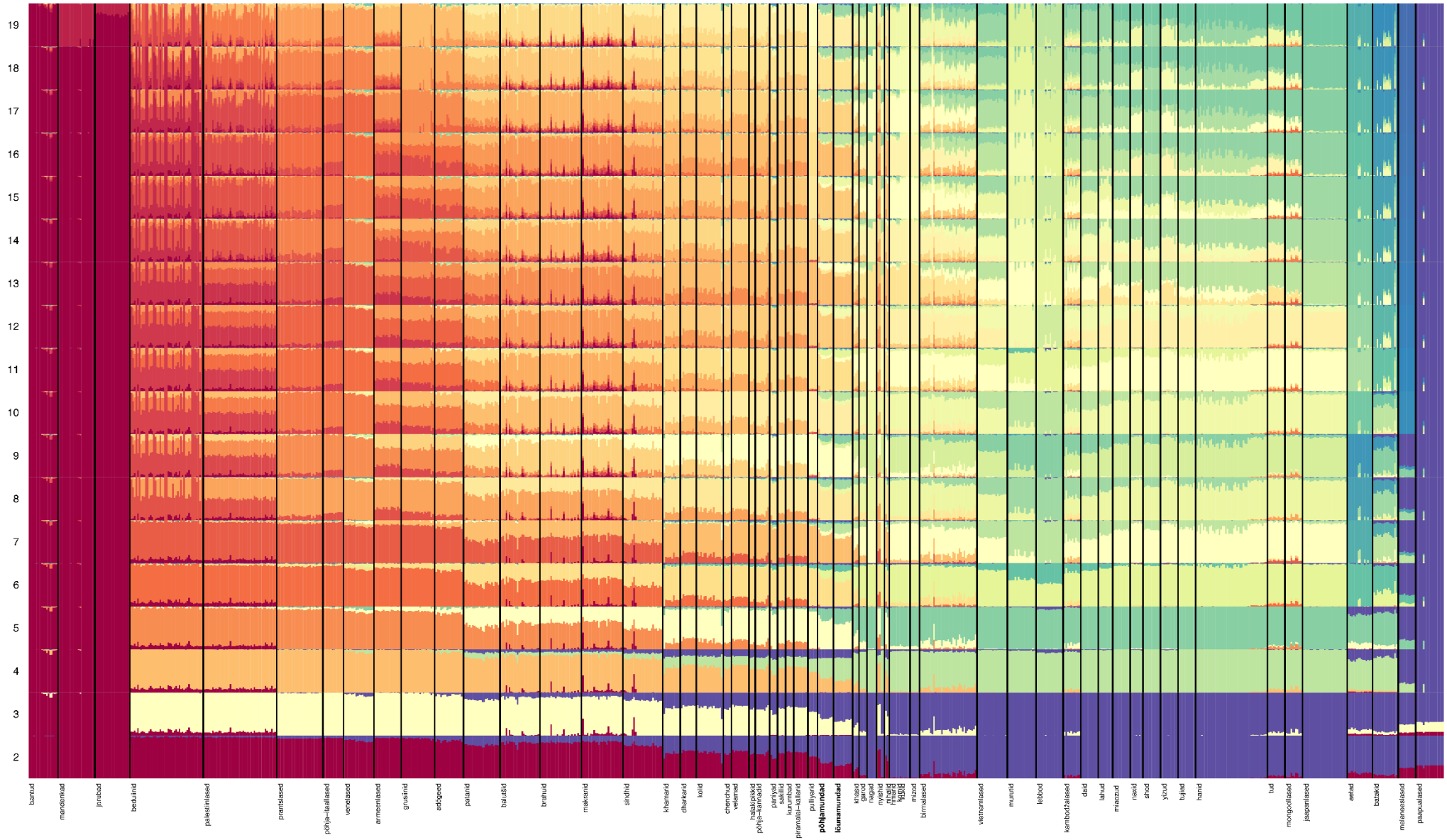
kolid	16	India	indoeuroopa	Metspalu *et al* 2011
komid	2	India	tiibeti-birma	EBK (avaldamata)
kukid	6	India	tiibeti-birma	EBK (avaldamata)
kurumbad	4	India	draviidi	Metspalu *et al* 2011
mizod	5	India	tiibeti-birma	EBK (avaldamata)
nagad	5	India	tiibeti-birma	Metspalu *et al* 2011
nihalid	2	India	lingvistiline isolaat	Metspalu *et al* 2011
põhja-kannadid	8	India	draviidi	Behar *et al* 2010
nyishid	4	India	tiibeti-birma	EBK (avaldamata)
paniyad	4	India	draviidi	Behar *et al* 2010
piramalai-kallarid	8	India	draviidi	Metspalu *et al* 2011
pulliyarid	5	India	draviidi	Metspalu *et al* 2011
sakillid	4	India	draviidi	Behar *et al* 2010
velamad	10	India	draviidi	Metspalu *et al* 2011
balutšid	24	Pakistan	indoeuroopa	Li *et al* 2008
brahuid	25	Pakistan	draviidi	Li *et al* 2008
makranid	25	Pakistan	indoeuroopa	Li *et al* 2008
patanid	22	Pakistan	indoeuroopa	Li *et al* 2008
sindid	24	Pakistan	indoeuroopa	Li *et al* 2008
Kagu-Aasia (N = 126)				
aetad	15	Filipiinid	austroneesia	Rasmussen 2011
batakid	15	Filipiinid	austroneesia	Migliano 2013
birmalased	35	Myanmar	tiibeti-birma	Cambridge (avaldamata)
lebbod	16	Indoneesia (Borneo)	austroneesia	Cambridge (avaldamata)
murutid	17	Brunei	austroneesia	Cambridge (avaldamata)
vietnamlased	18	Vietnam	austroaasia (khasi-asli)	Cambridge (avaldamata)
kambodžalased	10	Kambodža	austroaasia (khasi-asli)	Li *et al* 2008

Lisa 2

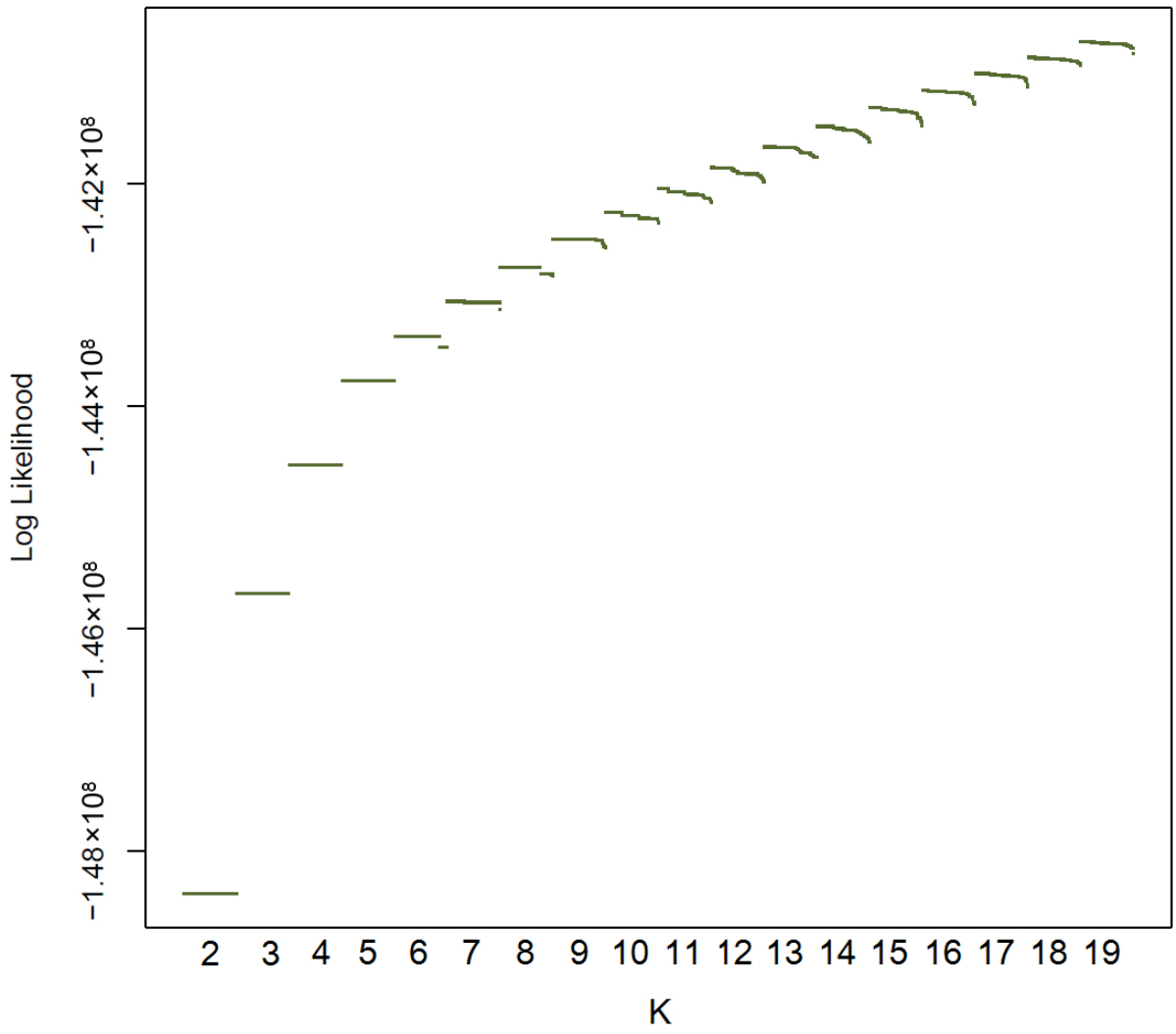


Kaks kõige rohkem varieeruvust kirjeldavat põhikomponenti. Väikesed kolmetähelised populatsiooninimede lühendid tähistavad indiviide, ringides olevad lühendid populatsioonide mediaanväärtusi. Kolm piirkonda joonisel on suurendatud (A, B, C). Populatsiooninimede lühendid koosnevad kolmest populatsiooninime esitähest. Populatsioonide nimekirja võib leida lisast 1.

Lisa 3



a) ADMIXTURE analüüsi tulemused K=2 kuni K=19 juures



b) ADMIXTURE programmi 100-kordsel jooksutamisel saadud mudelite tõepära skooride väärtused iga K puhul. Jooniselt on näha, et kõrgeimad stabiilsed LL väärtused on K=9 juures.

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina **Kai Tätte** (sünnikuupäev: 3. aprill 1990)

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose

Austroaasia keeli kõnelevate India rahvaste geneetiline päritolu,

mille juhendaja on **Mait Metspalu**,

1.1. reprodutseerimiseks säilitamise ja üldsusele kättesaadavaks tegemise eesmärgil, sealhulgas digitaalarhiivi DSpace-is lisamise eesmärgil kuni autoriõiguse kehtivuse tähtaja lõppemiseni;

1.2. üldsusele kättesaadavaks tegemiseks Tartu Ülikooli veebikeskkonna kaudu, sealhulgas digitaalarhiivi DSpace-i kaudu kuni autoriõiguse kehtivuse tähtaja lõppemiseni.

2. olen teadlik, et punktis 1 nimetatud õigused jäävad alles ka autorile.

3. kinnitan, et lihtlitsentsi andmisega ei rikuta teiste isikute intellektuaalomandi ega isikuandmete kaitse seadusest tulenevaid õigusi.

Tartus, 26.05.2015