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

Arno Pilvar

Y-kromosomaalse haplogrupi O2a alamharude esinemine India austroaasia keeli kõnelevatel rahvastel

Bakalaureusetöö

Juhendajad *MSc* Monika Karmin
PhD Gyaneshwer Chaubey
PhD Ene Metspalu

Tartu 2015

SISUKORD

KASUTATUD LÜHENDID

AAKR – austroaasia keeli kõnelevad rahvad

ap – aluspaar

CNV – c*opy number variation*, koopiaarvu varieeruvus

GRCh37 – Genome Reference Consortium human build 37, inimese referentsgenoomi 37. kokkupanek

GRCh38.p3 – Genome Reference Consortium human build 38 patch 3, inimese referents genoomi 38. kokkupanek, kolmas parandus.

hg - haplogrupp

HGP – The Human Genome Project, Inimese Genoomi Projekt

kb – *kilobase*, tuhat aluspaari

Mb – *megabase*, miljon aluspaari

MHC – Major Histocompability Complex, peamine koesobivuskompleks

mtDNA – mitokondriaalne DNA

PCR – polymerase chain reaction, polümeraasi ahelreaktsioon

RFLP – *restriction fragment length polymorphism*, restriktsioonifragmentide pikkuspolümorfism

SNP – *single nucleotide polymorphism*, ühenukelotiidne polümorfism

TAT – tuhat aastat tagasi

YCC – Y-Chromosome Consortium, Y-kromosoomi konsortsium

Y-STR – *Y-short tandem repea*ts, Y-kromosoomi lühikesed tandeemsed kordused

SISSEJUHATUS

Populatsioonigeneetikas aastakümneid kasutusel olnud Y-kromosoom on aidanud uurida rahvaste rändeid ning meessoopõhiseid demograafilisi protsesse. Y-kromosoom on laialdaselt kasutuses olnud seetõttu, et suurem osa temast on haploidne ja jääb rekombinatsioonist praktiliselt puutumata. Sellest tulenevalt kanduvad muutumatult edasi eelmiste inimpõlvede jooksul Y-kromosoomi kogunenud mutatsioonid. Nende mutatsioonide põhjal koostatava fülogeneetilise puu abil võime tuletada populatsioonide lahkenemise aega ning võimalikke rändeid. See annab väärtusliku infot vanemate lahknemiste kohta, millest puuduvad kirjalikud ajalooallikad ja arheoloogilisi leide on väga vähe.

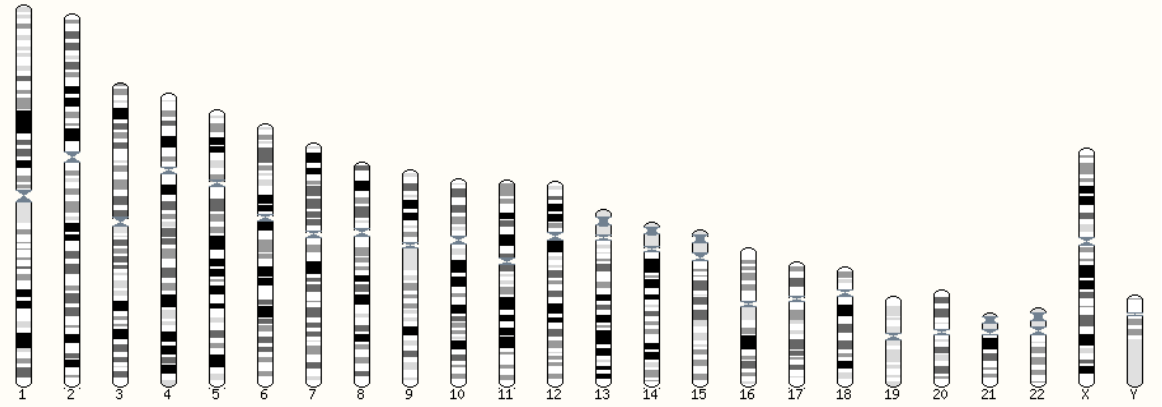
India ühtedeks vanimateks populatsioonideks peetakse austroaasia keeli kõnelevaid rahvaid (AAKR). Nende pärinemise kohta on püstitatud kaks hüpoteesi. Esimese järgi jäid nad India piirkonnas paikseks pärast väljarännet Aafrikast. Teise järgi läksid nad Indiast läbi Ida- ja Kagu-Aasiasse ning osa populatsioonist migreerus hiljem Kagu-Aasiast Indiasse tagasi. AAKRid omavad erinevalt ülejäänud India populatsioonidest kõrget Y-kromosomaalse haplogrupi O2a sagedust. Selle poolest sarnanevad nad rohkem Ida- ja Kagu-Aasia populatsioonidele, kui oma naabritele Indias.

Käesoleva töö eesmärgiks on täpsustada Y-kromosoomi haplogrupi O2a-M95 alamklaadide levikut India AAKRidel kasutades uusi haplogrupisiseseid markereid, mis on hiljuti saadud Y-kromosoomi täisjärjestuste põhjal (Karmin *et al*. 2015). Uurides haplogrupi O2a alamharusid, millest sai alguse populatsiooni kasvuga haplogrupisisene mitmekesistumine, defineerida asutajaliine ja nende andmete valguses uuesti hinnata AAKRi päritolu hüpoteese.

1. KIRJANDUSE ÜLEVAADE

1.1 Inimese genoom

Inimese genoom koosneb raku tuumas asuvast pärilikkusainest ehk DNAst ja raku energiat tootva organelli mitokondri genoomist. Tuumas olev DNA on pakitud 23 kromosoomipaari, millest 22 on autosomaalsed ja 1 paar sugukromosoome (X ja Y) (joonis 1). Mitokondriaalne DNA (mtDNA) on oluliselt lühem kui kromosoomid (Jobling *et al*., 2014. lk. 28-31).



Joonis 1. Inimese pärilikkuseaine jagunemine kromosoomidesse. Joonisel on 22 autosomaalset kromosoomi järjestatud suuruse järgi ning seejärel on toodud sugukromosoomid (X ja Y) (kohandatud Ensembl: *Whole Genome*, järgi, http://www.ensembl.org/Homo_sapiens/Location/Genome).

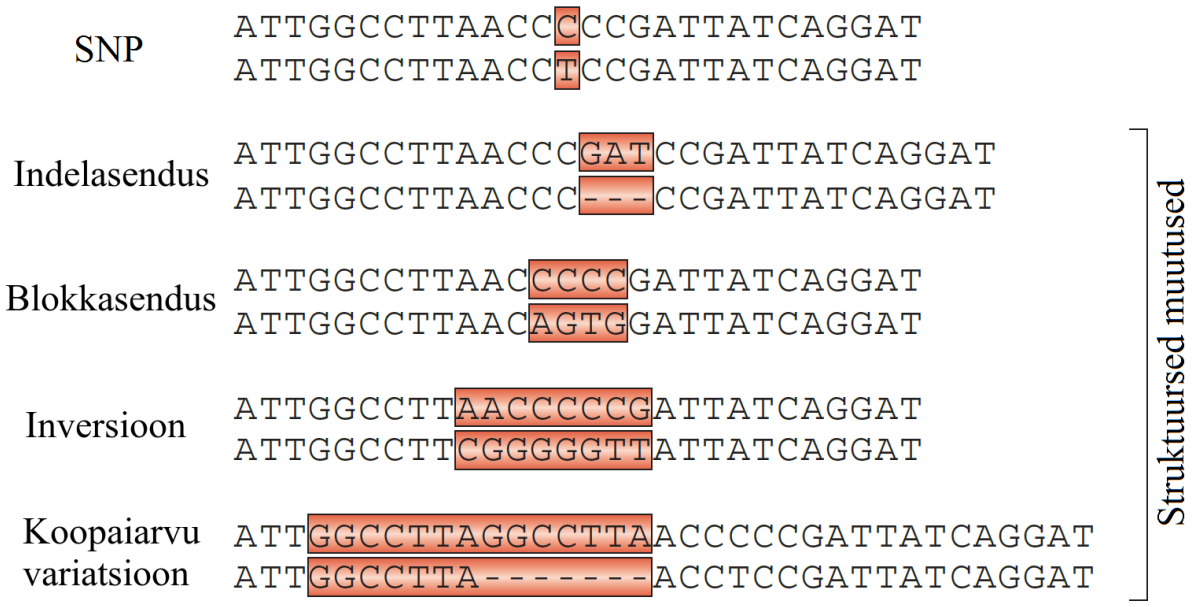
Autosomaalsed kromosoomid ehk autosoomid on nimetatud numbritega pikkuse järgi kahanevas järjekorras. Erandina on 20. pikem kui 19. kromosoom. Nende pikkused on vahemikus ~250Mb kuni ~50Mb (Jobling *et al.* 2014. lk. 28-31).

Sugukromosoomid X ja Y määravad ära isiku soo XY süsteemi järgi. Naistel on sugukromosoomide paariks XX ja meestel XY. X-kromosoom on ~155 Mb pikk, jäädes sellega 7. ja 8. kromosoomi vahele. Y-kromosoom on suuremas osas haploidne ning kromosoomide seas üks väiksematest, vaid ~59 Mb pikk ja jääb sellega 20. ja 19. kromosoomi vahele (Jobling *et al.* 2014. lk. 28-31).

MtDNA on 16,6 kb pikk rõngasmolekul, mis ei rekombineeru ja muteerub polümeraasi veatuvastamise mehhanismi puudumise tõttu ligi 20 korda kiiremini kui ülejäänud DNA (Douglas *et al.*, 1999).

1.1.1 Genoomi varieeruvuse tüübid

Valdava osa geneetilisest varieeruvusest moodustavad ühenukleotiidsed polümorfismid ehk SNPd. Neid raporteeriri 1092-s indiviidis kogu genoomi peale umbes 38 miljonit (The 1000 Genomes Project Consortium, 2012). SNPd jagunevad transitsioonideks (C ↔ T, A ↔ G) ja transversioonideks (C/T ↔ A/G) (Zhongming ja Boerwinkle, 2002). Transitsioone esineb ligikaudu kaks korda sagedamini (Levy *et al.*, 2007). Struktuursed genoomimuutused hõlmavad endas nii ülekromosoomseid kui ainult ühenukleotiidseid muutusi (joonis 2), kuid praktikas peetakse alampiiriks 1 kb (Jobling *et al.* 2014. lk 28-31). Struktuurse muutuse alla kuuluvad indelasendus, blokiasendus, inversioon ja CNV (koopiaarvu varieeruvus) (Frazer *et al.*, 2009).



Joonis 2. Genoomi varieeruvuse tüübid (kohandatud Frazer *et al.*, 2009 järgi).

Indelid on järjestuses toimunud muutused nukleotiidi lisandumise või kaotsiminekuga. Lühendit "indel" kasutatakse juhul, kui pole esivanemat, kellega võrrelda, ja me ei saa olla kindlad, kas toimus nukleotiidi lisandumine või kadumine (Scherer *et al.*, 2007). Üldiselt on indelid mõne aluspaari pikkused, aga võivad ulatuda ka üle 80kb (Levy *et al.*, 2007). Blokiasendus on olukord, kus rida kõrvutiasetsevaid nukleotiide erineb kahe genoomi vahel (Frazer *et al.*, 2009). Inversioon on nukleotiidide ümberpööratud järjestus genoomis. Koopiaarvu variatsioon tekib siis, kui identsed või väga sarnased järjestused korduvad osades kromosoomides, kuid mitte kõikides (Frazer *et al.*, 2009).

1.1.2 Referentsgenoom ja genoomi varieeruvus populatsioonis

1.1.2.1 Referentsgenoom

Inimese genoomi täielikuks sekveneerimiseks algatati 1990. aastal 3 miljardi dollari suurune projekt *The Human Genome Project*. Selle abil kirjeldati 2001. aastal 1,42 millionit SNPd (Sachidanandam, 2001). *The Human Genome Project* saavutas oma eesmärgi 2 aastat enne tähtaega 14.aprillil 2003.

Esimene referentsgenoom on koostatud 13 anonüümse vabatahtliku DNAst, kes on pärit USA New Yorki osariigis olevast Buffalo linnast ning see anti välja 2001. aastal. Algselt sisaldas see ~150 000 lünka (Anon, 2010). Tänaseks viimases 19. versioonis (GRCh38) on lünkade arv oluliselt vähenenud. Genome Reference Comsortiumi andmeil neid esineb neid veel alla tuhade.

Suurema osa inimgenoomi puhul on võimalik sekveneeritud lugemeid paigutada referentsgenoomile, aga on genoomipiirkondi, kus on liiga palju muutlikkust usaldusväärse tulemuse saamiseks. Üheks selliseks raskeks piirkonnaks on näiteks MHCd kodeeriv ala (*The MHC Sequencing Consortium*, 1999).

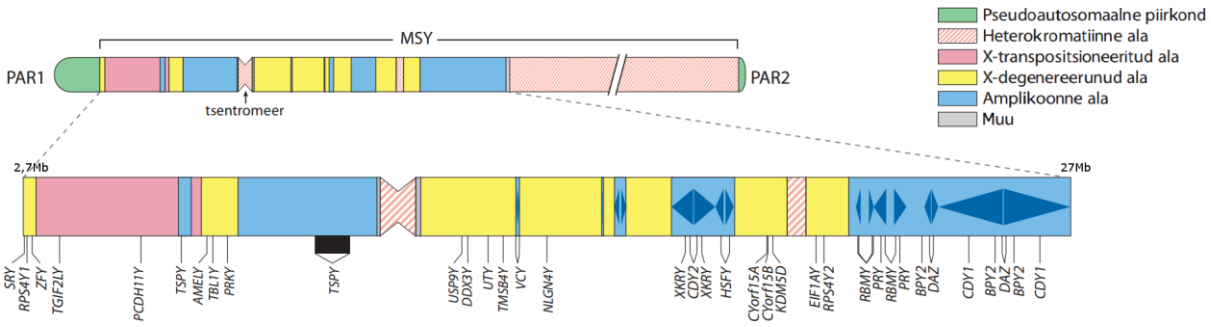
1.1.2.2 Genoomi varieeruvuse kaardistamise projektid

On tehtud mitmeid inimgenoomi varieeruvuse kaardistamise projekte, mis kirjeldavad varieeruvuse levikut populatsioonides. Üks neist, rahvusvaheline HapMap-projekt, on koostöö erinevate akadeemiliste keskuste, mittetulunduslike biomeditsiini uurimisgruppide ja erafirmade vahel, mis on pärit Kanadast, Hiinast, Jaapanist, UKst ja USAst. Selle eesmärgiks on kindlaks teha ühiseid jooni inimese genoomi variatsioonis ning avalikustada see info kõigile tasuta. Esimese faasi ülesandeks oli genotüpiseerida vähemalt üks levinud SNP (sagedus >1%) iga 5 kb tagant (International HapMap Consortium, 2003). I faasis (2005) kirjeldati ~1,3 miljonit, II faasis (2007) üle 3,1 miljoni ja III faasis (2009) 1,6 miljonit SNPd (International HapMap Consortium, 2003; 2007; 2010).

Esialgsed HapMap projekti andmed mängisid keskset rolli GWASi (*Genome Wide Assosiation Studi*es, ülegenoomsed assotsiatsiooniuuringud) väljatöötamisel. Järgmiseks sammuks kogu genoomi uurimisel oli 1000 Genoomi Projekt, mille eesmärgiks oli leida, genotüpiseerida ja pakkuda täpset haplotüüpide informatsiooni erinevate polümorfismide kohta mitmetes inimpopulatsioonides (The 1000 Genomes Project Consortium, 2012). Need kõigile vabalt kätte saadavad andmed on olnud oluliseks tõukeks nii medisiinigenoomikale kui populatsioonigeneetikale.

1.2 Y-kromosoom

Erinevalt teistest kromosoomidest jääb Y suuremalt jaolt kõrvale meiootilisest rekombinatsioonist. Siiski on kaks pseudoautosomaalset piirkonda kromosoomi otstes, PAR1 ja PAR2, mis rekombineeruvad X kromosoomiga (joonis 3). PAR2 asetseb X ja Y-kromosoomi pika õla otsas, on 329 kb pikk (GRCh38.p3), sisaldab nelja geeni ning rekombinatsioon selles piirkonnas on harvem. PAR2 tekkis evolutsioonilisest seisukohast suhteliselt hiljuti ja on sellisel kujul inimesele ainuomane. PAR1 aga asetseb X ja Y-kromosoomi lühikese õla tipus, on oluliselt suurem, 2,7 Mb pikk (GRCh38.p3), ning mahutab endas 24 geeni (Mangs ja Morris 2007). Nende suurus on kokku ~3 Mb, mis on ligikaudu 5% Y-kromosoomi pikkusest. Leiti ka PAR3 (3,5 Mb), kuid võrreldes teiste pseudoautosomaalsete piirkondadega on selle esinemise sagedus üldpopulatsioonis ainult 2% (Veerappa *et al.* 2013). Ülejäänud piirkonda kutsutakse Y-kromosoomi mitterekombineeruvaks osaks (NRPY või NRY) või meesspetsiifiliseks Y-kromosoomi alaks (MSY). Selles eristuvad X-transpositsioneeritud piirkonnad, mis tekkisid 3-4 miljonit aastat tagasi transpositsiooni käigus X-kromosoomilt pärit järjestuse liitmisel Y-kromosoomile (Page, 1984; Skaletsky *et al.*, 2003); X-degenereerunud ala, mis on X-kromosoomi deletsioonidest räsitud analoog (Skaletsky *et al.*, 2003); amplikoonne ala, mis sisaldab palju kordusjärjestusi ja geene ning on funktsioonilt spetsialiseerunud (joonis 3) (Bhowmick, *et al.*, 2007).



Joonis 3. Y-kromosoomi struktuuri skeem. Ülemisel paneelil on kujutatud kogu Y-kromosoom. Värvidega on eristatud pseudoautosomaalsed alad PAR1 ja PAR2 rohelisega, heterokromatiinne ala roosa-valge triibulise või heleroosana, X-transpositsioneeritud ala tumeroosaga, X-degenereerunud ala kollasega ja amplikoonne ala sinisega. Tumesinised kolmnurgad illustreerivad amplikoni orientatsiooni. MSY – meesspetsiifiline Y-kromosoomi ala (kohandatud Hughes ja Rozen, 2012 järgi).

Juba üle poole sajandi on teada, et Y-kromosoom määrab imetajatel isassugu, kuid pikemat aega teadmised Y-kromosoomi kohta sellega piirdusidki. Y-kromosoomi peeti pikka aega geneetiliseks tühermaaks (Charlesworth ja Charlesworth 2000). Osalt seetõttu, et võrreldes X-kromosoomiga on hulgaliselt deletsioone ja geenifunktsioonide kadumist ning teisalt seetõttu, et klassikaliste geneetiliste meetodite (sugupuu ja aheldatusemeetodid) abil ei

leitud seost Y-liiteliste geenide ja fenotüüpide vahel (Hughes ja Rozen 2012). Praeguseks on uurimsmeetodite täiustumise tulemusel teadmised Y-kromosoomi ja selles leiduvate erinevte strukruursete elementide kohta oluluselt täienenud (Skaletsky et al. 2003; Hughes ja Rozen 2012)

Y-kromosoom on tundlikum asutajaefekti ja geenitriivi suhtes kui X- ja autosomaalsed kromosoomid, sest ühe mees-naine paari kohta on neli autosomaalse DNA koopiat, kolm X-kromosoomi koopiat aga ainult üks Y-kromosoom. Seega on Y-kromosoomi efektiivne populatsiooni suurus veerand autosoomi ning kolmandik X kromosoomi omast (Jobling ja Tyler-Smith, 2003).

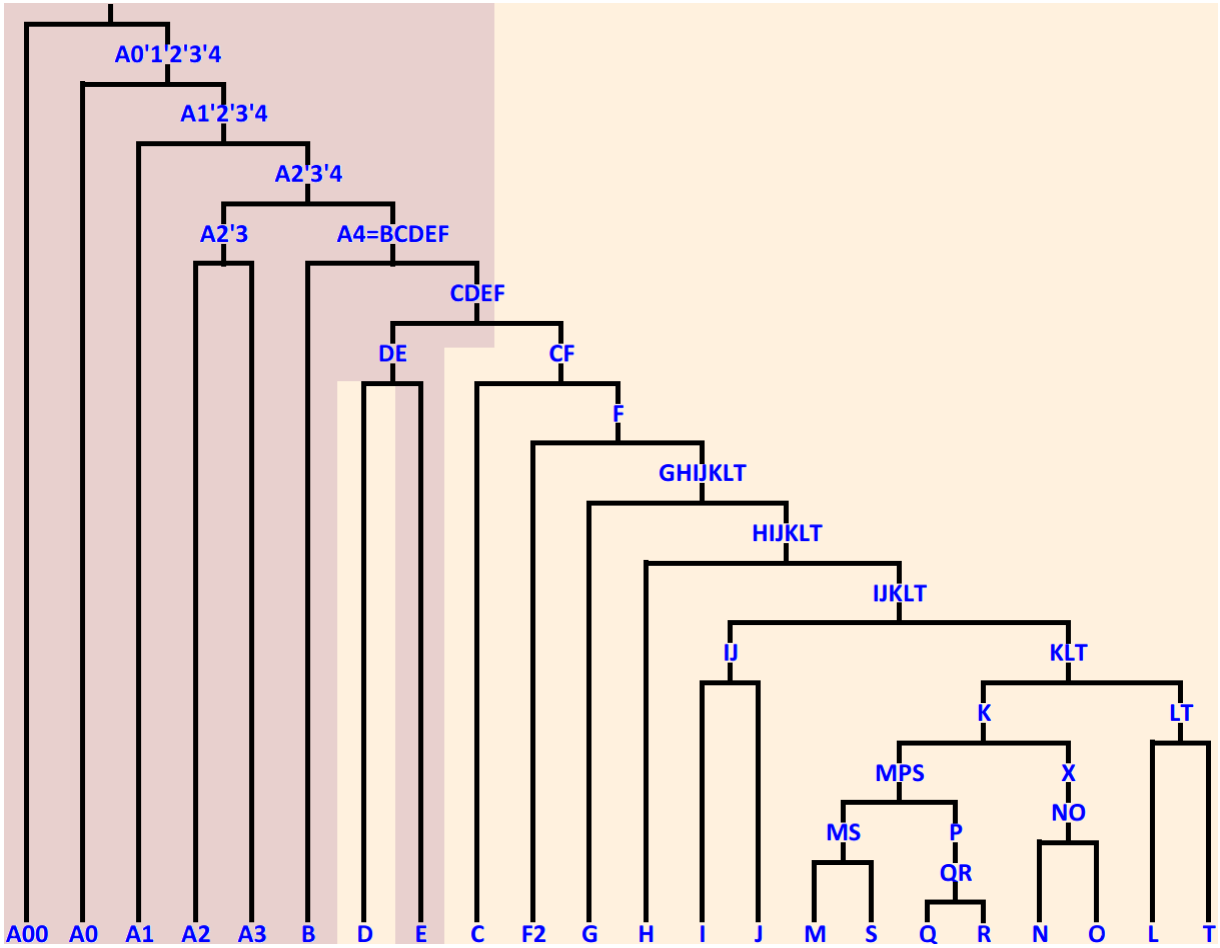
Kuigi Y-kromosoom on väike osa genoomist, isaliini pidi klonaalselt päranduva sugukromosoomi bialleelsed markerid (SNPd) laialdaselt kasutusel antropoloogias ja populatsioonigeneetikas (Jobling ja Tyler-Smith, 2003). Nende põhjal on võimalik arvutada populatsioonide lahknemise aega eeldusel, et Y-kromosoomi molekulaarne kell on ajas konstantne (Wei *et al.*, 2013). Y-kromosoomi mutatsioonikiiruste uurimustöödes on saadud tulemusi alates $0{,}6*10^{-9}$ (Mendez *et al.*, 2013) kuni $1*10^{-9}$ (Xue *et al.*, 2009) asenduseni nukleotiidi kohta aastas. Kõige uuem pakutud kiirus, $0{,}74*10^{-9}$ asendust nukleotiidi kohta aastas, on kalibreeritud kahe iidse DNA järjestuse järgi (Karmin *et al*., 2015).

Y-kromosoomi varieeruvus sisaldab peale bialleelsetele markerite veel ka mini- ja mikrosatelliite (STRe), millel põhinevad testid on kasutusel isiku- ja isaduse tuvastamisel ning genealoogias (Diegoli, 2015).

1.2.1 Y-kromosoomi fülogenees

Y-kromosoomi fülogeneesipuu klaadide eristamiseks kasutatakse mitterekombineeruvas alas olevate SNPde haplotüüpe. Nende põhjal on võimalik määrata indiviidi kuuluvust uuritavasse haplogruppi. Nomenklatuuri ühtlustamiseks avaldas YCC (Y Chromosome Consortium) 2002. aastal põhimõtted ja reeglistiku, millest lähtuda Y-kromosoomi markerite ja haplogruppide nimetamisel (YCC, 2002). Suuremad haplogrupid on tähistatud üksiku tähega alustades Ast. Haplogrupid jagunevad veel üks kuni mitu korda alamhaplogruppideks, mille tagajärel moodustub puu (Jobling ja Tyler-Smith, 2003). Algselt oli defineeritud 18 haplogruppi (A-st R-ni), mis jagunesid kokku 153 haruks. Selleks kasutati 243 markerit (Jobling ja Tyler-Smith, 2003). 2008. aastal tuli markereid juurde, kokku 599, lahutuvus suurenes (20 suuremat ja 311 alamharu) ja osad harud tõsteti ümber, aga suures plaanis jäi puu kuju samaks (Karafet *et al.*, 2008). Haplogrupid on tähistati tähtedega A-st kuni T-ni.

Hea ülevaate praegustest suurematest haplogruppidest annab 2014. aastal avaldatud minimalistlik Y-kromosoomi fülogeneesipuu (joonis 4) (van Oven, 2014).

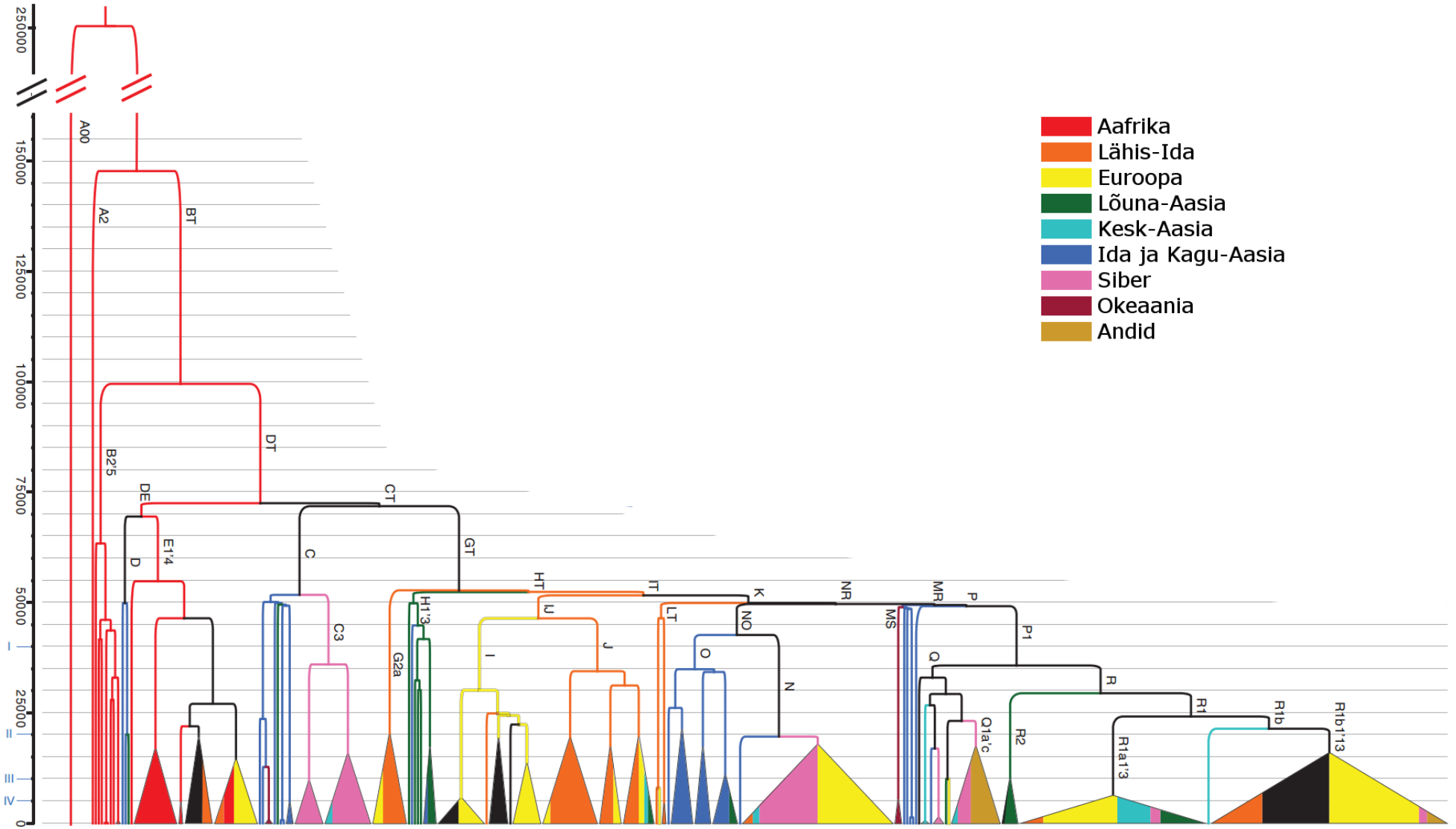


Joonis 4. Lihtsustatud Y-kromosoomi fülogeneesipuu. Puu kõige vanemad lahknemised – vaid Aafrikas levinud haplogrupid on tähistatud roosa taustaga. Väljaspool Aafrikat levinud haplogrupid on tähistatud kollasega ning moodustavad esimeste alamhulga. (kohandatud van Oven, 2014; Jobling *et al*., 2014. lk. 357 järgi).

Ajakohase Y-DNA markerite info kättesaadavana hoidmiseks uuendab vabatahtlikest koosnev mittetulundusühing ISOGG (International Society of Genetic Genealogy) regulaarselt oma veebilehte.

Aja jooksul on juurde lisandunud aina rohkem Y-kromosoomi markereid, mis lubab koostada detailsemaid fülogeneesipuid. Teise põlvkonna sekveneerimise abil saadud Y-kromosoomi täisjärjestused on markerite hulka paarkümmend korda suurendanud(Hallast *et al.*, 2014; Karmin *et al.*, 2015). See ei ole Y-kromosoomi fülogeneesipuu üldstruktuuri muutnud, aga see-eest on muutunud harude pikkused (vanused üksteise suhtes) ning juurde on tekkinud hulgaliselt tipuharusid. See informatsioon võimaldab nüüd palju täpsemini iseloomustada demograafilisi protsesse nagu rände, asutajaefekt, populatsiooni kasv (Jobling *et al*., 2014. lk. 174).

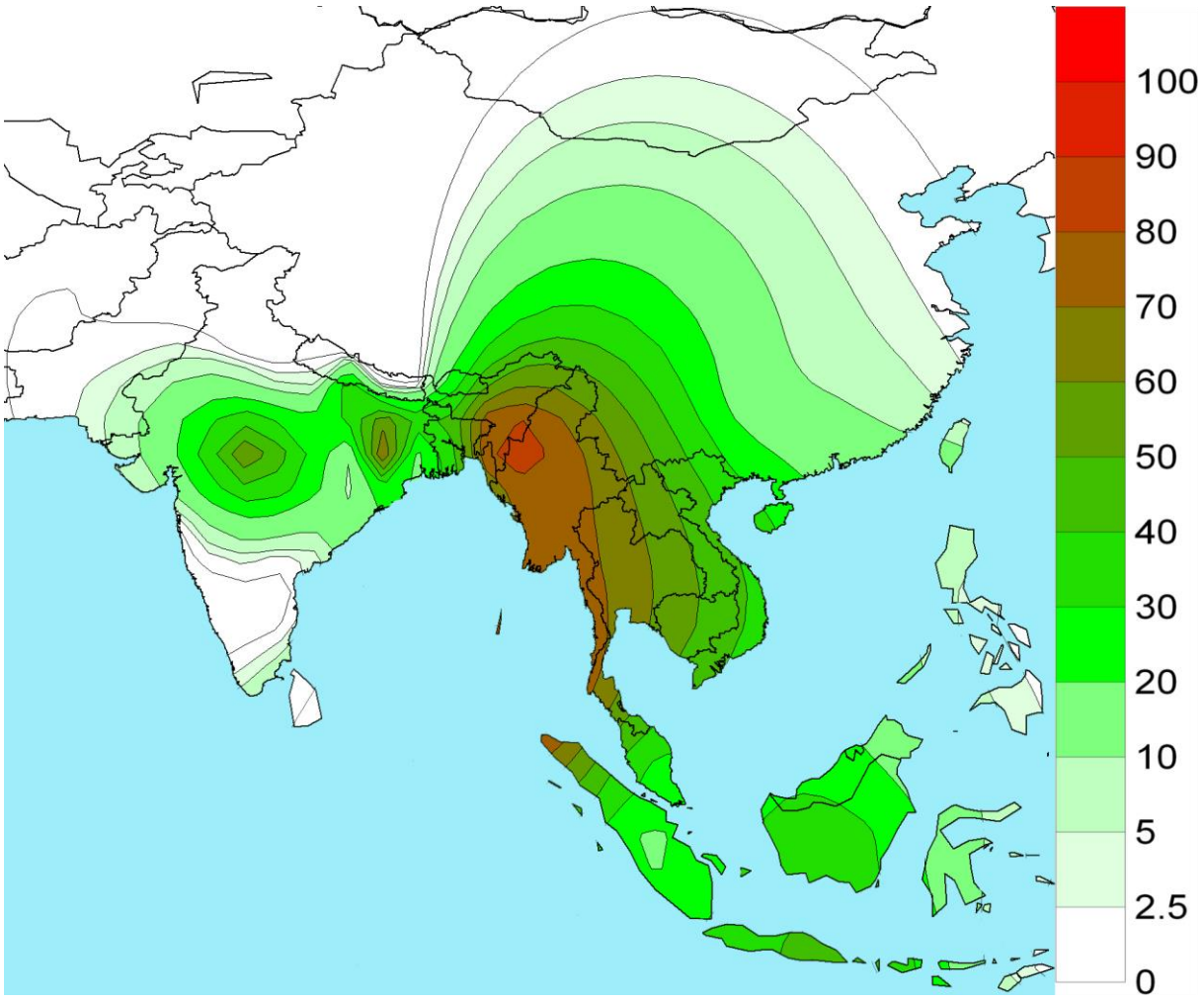
Hetkel kõige uuem globaalse valimiga Y-kromosoomi fülogeneetiline puu on avaldatud märtsis 2015 (joonis 5) (Karmin *et al.*, 2015).



Joonis 5. Y-kromosoomi täisjärjestustest saadud markeritel põhinev fülogeneesipuu koos levikuga maailmas (kohandatud Karmin *et al.*, 2015 järgi)**.**

1.2.1.1 Haplogrupp O levik

Haplogrupp O on kõige laiemalt levinud Ida-ja Kagu-Aasias ning vähemal määral Lõuna-Aasias. (Karmin *et al.*, 2015) Ida-Aasias on selle keskmiseks sageduseks 63,75% (Zhong *et al.*, 2011). Seda leidub mõõduka või madala sagedusga ka Kesk-Aasias ja Okeaanias (Hammer *et al.*, 2005). ISOGGi andmetes (jaanuar 2015 seisuga) on haplogrupile O omased 81 markerit (http://www.isogg.org/tree/ISOGG_HapgrpO.html). Haplogrupp O on alamhaplogruppide poolest väga mitmekesine, viimaste andmete põhjal on sellel 92 alamhaplogruppi (Karmin *et al.*, 2015). Antud töös uuritav haplogrupp O2a-M95 esineb kõige kõrgema sagedusega just AAKR populatsioonides (joonis 6) (Basu *et al.*, 2003; Kumar *et al.*, 2007). Täisjärjestuste põhjal saadud uued markerid kirjeldavad võrreldes varasemaga palju detailsemalt haplogrupi O2a-M95 sisestruktuuri, kuid leitud alamharude laevikut AAKR populatsionides pole kirjeldatud.



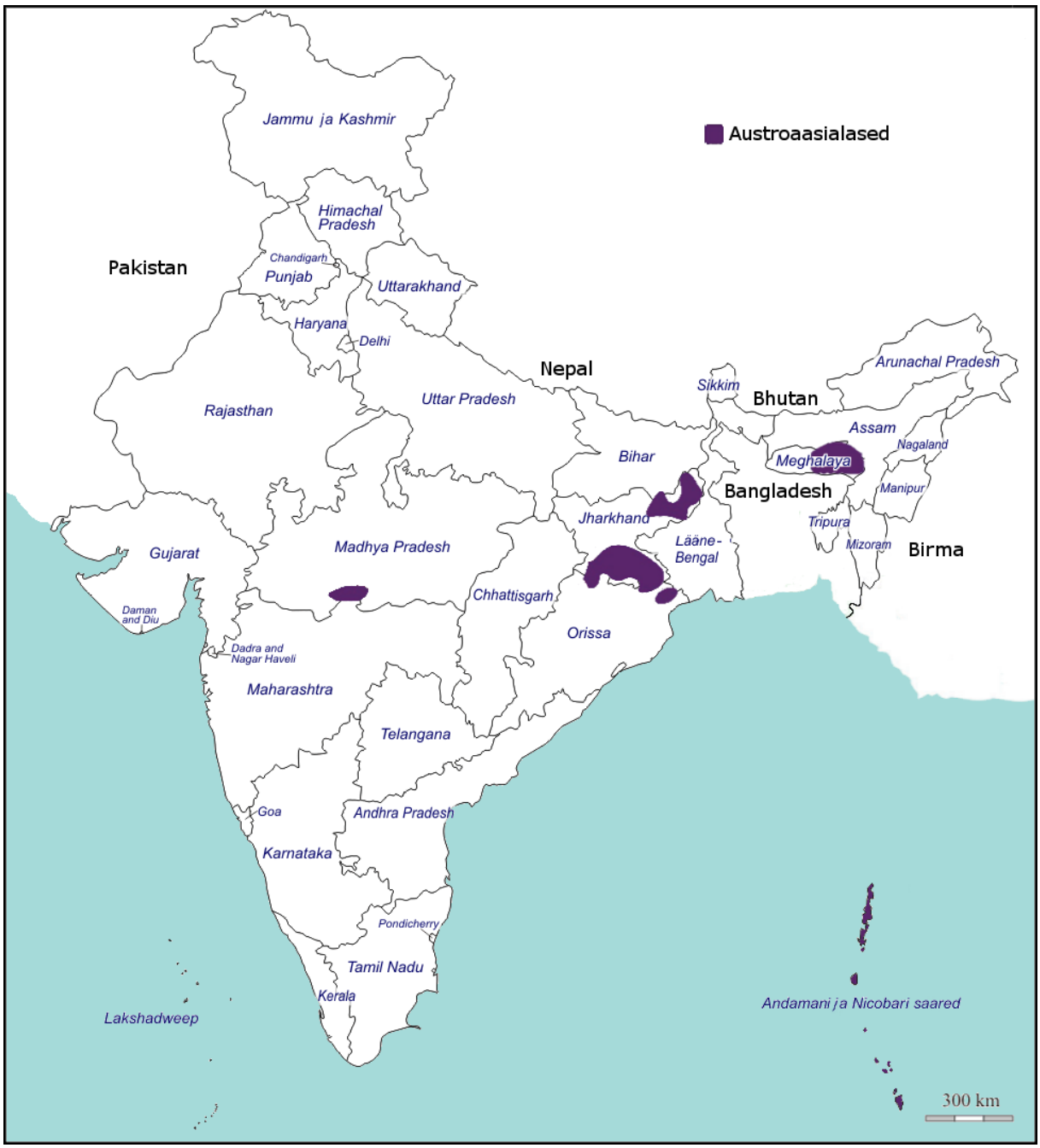
Joonis 6. Y-kromosoomi haplogrupp O2a sagedusi kujutav fülogeograafiline skeem Lõuna- ja Kagu-Aasiast (kohandatud Chaubey *et al.*, 2011 järgi)**.**

1.3 Austroaasia keelkond ja selle levik Indias

Austroaasia keelkonda kuulub ligikaudu 169 keelt ning see on 102,5 miljoni rääkijaga suuruselt maailmas üheksas (Ethnologue: Languages of the World, https://www.ethnologue.com/statistics/family). Austroaasia keelte rääkijad on eranditult hõimurahvad (Basu *et al*., 2003). Suuremalt jaolt paiknevad nad hajutatult Kagu-Aasias. Vähemal määral on AAKRe Kesk-, Ida- ja Kirde-Indias (joonis 7) (Ayub ja Tyler-Smith, 2009). Oletatavalt on üle 70 miljoni AAKRist Vietnamlased, 10 miljonit räägivad khmeri keelt, 5 miljonit santhali keelt ning üljäänud jagunevad 150 keele vahel paarisaja kuni mõnekümne tuhande inimesega (van Driem, 2001).

Tänapäeva India elanikkond koosneb paljudest erinevatest populatsioonidest, mis kogurahvaarvuna moodustavad 1267,4 miljonit inimest. Ametliku staatusega on Indias 22 keelt, kuid India elanikud räägivad väga paljusid keeli, mis valdavalt jagunevad hõimulistel rahvastel austroaasia, draviidi ja tiibeti-birma keelkonnaks ja mittehõimulistel rahvastel indo-euroopa ja draviidi keelkonnaks ((United Nations Population Division, 1. juuni 2014).

Indias levivad austroaasia keeled jagunevad kahte suuremasse keelkonna harusse, need on munda haru (Kesk-, Ida- ja Kirde-Indias) ning khasi-asliani haru (peamiselt Meghalaya osariik ja Nicobari saared). Enamus Kagu-Aasia AAKRe on mõne khasi-asliani haru rääkijad (Diffloth, 2001). Selline kahte suuremasse rühma jaotamine (munda ja teised austroaasia keeled) on valdavalt levinud (Anderson 2006), kuid välja on käidud ka võimalus keeli jagada kolme rühma: munda, nikobari ja mon-khmeri keeled (van Driem, 2001).



Joonis 7. Austroaasia keelte levik Indias tähistatud lilla värviga (kohandatud Ayub ja Tyler-Smith, 2009 järgi).

1.3.1 India AAKRide päritolu lingvistika ja riisi fülogeneetika põhjal

Indias elavate austroaasia keeli kõnelevate populatsioonide päritolu kohta on kaks põhilist hüpoteesi, mis üksteisele vastandlikult paigutavad nende algkodu kas Kagu-Aasiasse või Indiasse (Van Driem 2001, Fuller 2007, Kumar *et al.* 2007, Chaubey *et al.* 2011). Ühe teooria järgi pärinevad nad Kagu-Aasias riisikavatajatest. Traditsiooniline riisi kodustamise hüpotees väidab, et riisi kasutuselevõtt algas Hiinas umbes 6000 aastat tagasi. Higham-Bellwoodi mudel arvestab India munda ja India khasi-asliani keelt rääkivate korilaste ning

Kagu-Aasia khasi-asliani keelt rääkivate korilaste sõnavara sarnasust riisi kasvatamisega seotud sõnade puhul (Higham 2003; Bellwood 2005). Kagu-Aasia khasi-asliani populatsioon võis riisi kasvatamise hiinlastelt üle võtta ja hiljem neoliitikumis, mis oli 4-6,5 tuhat aastat tagasi (TAT) Indiasse migreeruda (Higham, 2003; Bellwood, 2005). Keeleteadlaste seas on endiselt levinud arvamus, et varastel AAKRidel on seos riisi kasvatamisega (Van Driem, 2012).

Teine hüpotees pakub vastupidiselt esimesele, et AAKRide levik algas juba enne neoliitikumi Lõuna-Aasiast (Fuller, 2007). Gerard Diffloth väidab, et AAKRide floora ja fauna sõnavara põhjal võib nende päritolu pigem eeldada troopilisest (Indiast) kui parasvöötme piirkonnast (Hiinast) (Diffloth, 2005).

Varaste austroaasia keelte kõnelejate seostamine riisikasvatamisega teeb keerukakas ka see, et riisi päritolu pole ka veel üheselt mõistetav. Esimese hüpoteesi järgi arenes riisi liigist *Oryza rufipogon* kodustamise tagajärel välja *Oryza indica* (pikateraline riis, mitte kleepuv riis) ning hiljem lahknes sellest *Oryza japonica* (lühiteraline, kleepuv riis) (Lu *et al.*, 2002).

Teise hüpoteesi järgi toimus diferentseerumine erinevate ökoloogiliste ja geograafiliste piirkondadega kohanemise tõttu. Sellele järgnes ühekordne *O.sativa* kodustamine liigist *O.rufipogon* (Oka ja Morishima, 1982).

Kolmanda hüpoteesi järgi kodustati *O.sativa* vähemalt kaks korda diferentseerunud *O.rufipogon*ist (Londo *et al*., 2006). Dorian Q.Fuller leidis geneetilisi tõendeid *Oryza indica* ja *Oryza japonica* iseseisva kodustamise kohta, mis toetab võimalust, et austroaasia keelkonna populatsioonid on India päritolu (Fuller, 2007).

Ükski eelnevalt mainitud riisi päritolu teooriatest ei suuda aga seletada ülegenoomset variatsiooni ei *O.rufipogon* ega *O.sativa* puhul ning pakuti välja „kombineeritud mudel“ (Kovach *et al.*, 2007). Selle järgi kodustati *O.japonica* ja *O.indica* diferentseerunud *O.rufipogonist* ning alguses toimus kodustatud liikide hübridisatsioon, mis seletaks ühiseid kodustamise tulemusel tekkinud alleele kõigis tänapäevastes sortides (Kovach *et al.*, 2007). Fuller pakkus, et *japonica* riis võis proto-*indica* riisiga kokku saada pigem kauplemise kui järk-järgulise põllukasvatajate rände läbi (Fuller, 2011). Seega riisi päritolu kohta käiva info seostamine austroaasia keeli kõnelevate inimeste võimaliku päritoluga annab vastuolulisis tulemusi.

1.3.2 India AAKRide päritolu populatsioonide geneetilise info põhjal

Erinevate meetodite (mtDNA, Y-kromosoomi ja autosoomi) põhjal tehtud uuringutega on saadud erinevaid tulemusi. Austroaasia keelte kõnelejate emaliinide uuringud on näidanud, et Indias elavate munda keelt kõnelevate rahvaste mtDNA haplogrupp on sarnane kohalike draviidi ja indo-euroopa keeli rääkivate populatsioonidega (Basu *et al*., 2003; Chaubey *et al*., 2007; Thangaraj *et al*., 2009). Samas näitab mtDNA selget erinevust India munda keele ja Kagu-Aasia khasi-asliani keele rääkijate vahel, mis viitab soopõhisele migratsioonile (Chaubey *et al.,* 2011).

AAKRide seas esineb Y-kromosoomi haplogrupp O-M95 kõrge sagedusega, keskmiselt üle 55% (mitmes populatsioonis üle 85%) (Kumar *et al.*, 2007; Chaubey *et al.*, 2011). Selle lahknemisajaks on pakutud 65 (25– 132) TATi, mis pooldab AAKRide India päritolu (Kumar *et al.*, 2007). Kaks tööd on saanud aga oluliselt hilisema lahknemise: 8 800 (3 900 – 23 200) (Kayser *et al.,* 2003) ja 20 000 (17 300 – 22 700) (Chaubey *et al.,* 2011). Viimased toetavad austroaasia keelte kõnelejate päritolu Kagu-Aasiast.

Autosomaalsete markeritega tehtud uuringud näitavad kahesuunalist geenivoolu üle Bengali lahe austroaasia keelte rääkijate ja tiibeti-birma keelte rääkijate vahel. India munda keele rääkijate genoomis on näha Kagu-Aasia geneetilist komponenti (umbes veerand), mis viitab hiljutisele migratsioonile millele järgnes põhjalik segunemine kohalike India populatsioonidega (Chaubey *et al*., 2011).

Hiljutised molekulaarsed uuringud toetavad India austroaasia keelte (munda keelkond) kõnelejate Kagu-Aasia päritolu (Chaubey et al 2011, Metspalu et al 2011). Varasemad uuringud on näidanud ka Y-kromosoomi haplogrupi O2a-M95 seotust selle rändega (Sahoo et al 2006, Sengupta et al 2006, aga ka Chaubey et al 2011). Samas, selle rände aeg vajab täpsemaid hinnanguid

2. EKSPERIMENTAALOSA

2.1 Töö eesmärgid

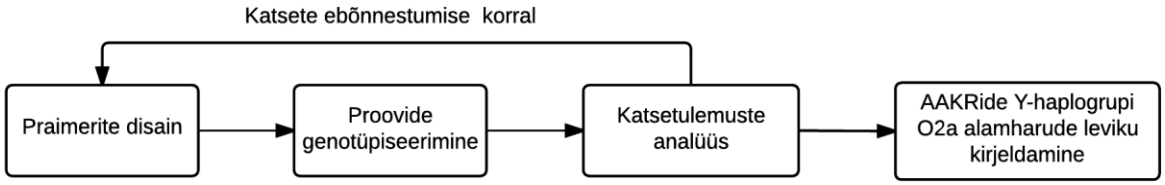
Hiljutine meie uurimisgrupi töö andis olulise panuse tänapäeva inimeste isaliinide kõrgresolutsiooniga fülogeneesile (Karmin 2015). See uuring tõi esile ohtralt uusi O2a-M95st fülogeneetiliselt allavoolu olevaid markereid ning olulisi tõendeid selle haplogrupi neoliitilise päritolu kohta.

India austroaasia keelte kõnelejate rännetega seotud alamklaadide ja ajaskaala ledmiseks disainisime uuetele markeritele praimerid ning genotüpiseerisime need India M95 markeri järgi O2a haplogruppi kuuluvate munda ning indoeuroopa keelte kõnelejate seas.
Töö alaeesmärgid on selgitada:

1) miinimum arv asutajaliine, mis on seotud haplogrupi O2a Lõuna-Aasiasse toomisega
2) haplogrupis O2a-M95st allavoolu olevate asutajaliinide ajaline ja ruumiline jaotus

2.2 Materjal ja metoodika

Allpool olev skeem kirjeldab antud töö praktilise poole ülesehitust (joonis 8).



Joonis 8. Üldine tööskeem. Praimerite disain sisaldab endas ka nende valideerimist ning katset naise DNAga, et välistada seondumine X-kromosoomile.

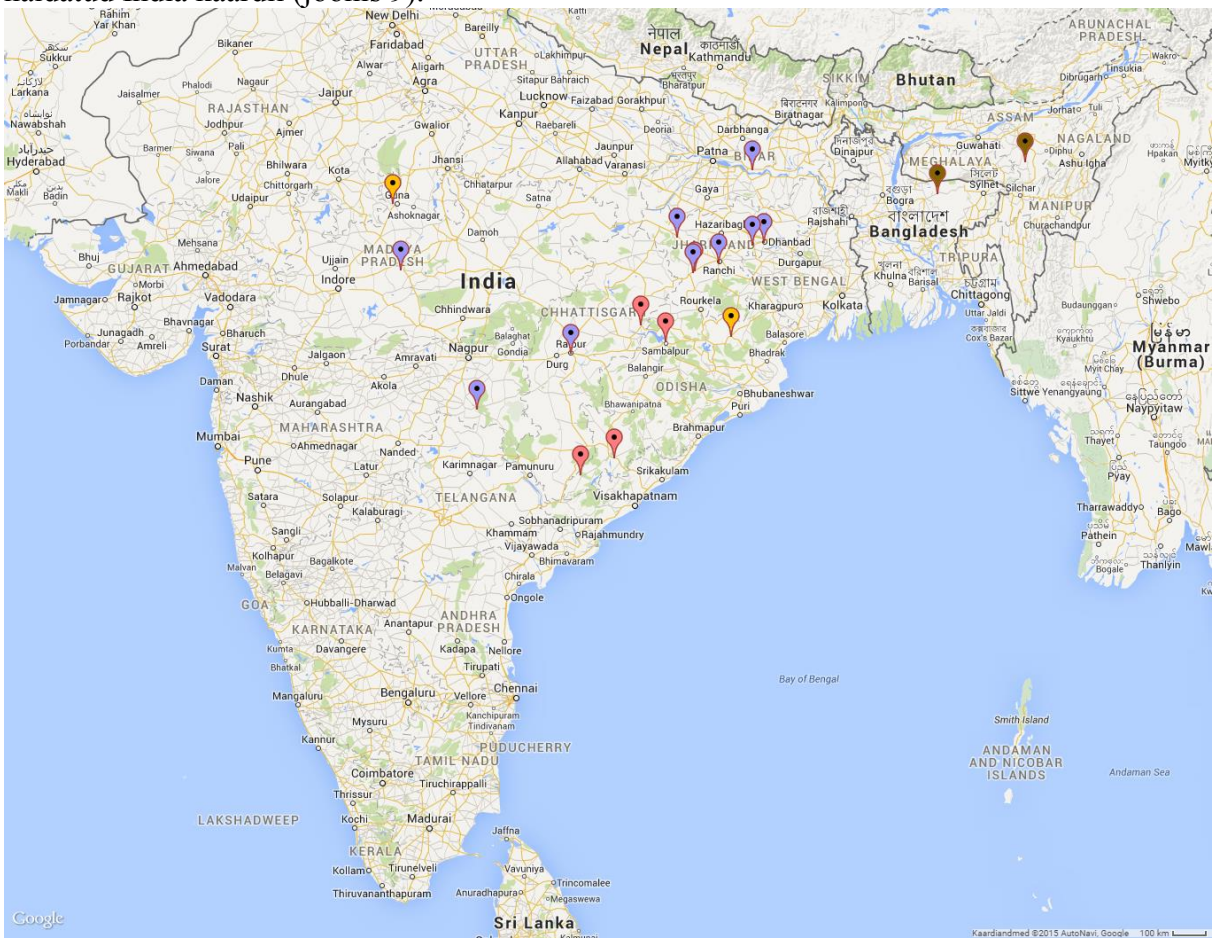
2.2.1 Valim

Valim koosneb 173st 1970. aastal ja perioodil 2002-2005 kogutud vabatahtlikest (tabel 1), kes andsid kirjaliku nõusoleku oma vereproovi kasutamiseks uuringutes. Kirjaoskamatutele seletati suuliselt protseduuri eesmärk ja võeti allkirja asemel pöidla sõrmejälg. DNA eraldati verest Sambrooki protokolli järgi (Sambrook *et al*., 1987). Uuringugruppi võeti ka täna mitmekeelse, kuid valdavalt indo-euroopa keelt kõneleva Baiga rahva esindajaid (23). Pärimuse järgi rääkisid nende esivanemad austroaasia keelkonna keelt (Grierson, 1904). Samas on neid kirjeldatud ka draviidi keelkonna rahvana (Russel ja Hiralal, 1916; Menz ja Hansda 2009). Valim on eelnevalt genotüpiseeritud Y-kromosoomi haplogruppi O2a-M95.

Tabel 1. Valimis olevad AAKRide populatsioonid.

Populatsioon	Indiviide	Keel	Keeleperekond	Keelkond	Osariik
Asur	29	asuri	põhja-munda	austroaasia	Jharkhand
Baiga*	23	chhattisgarhi	indo-iraani	indo-euroopa	Madhya Pradesh, Orissa
Birhor	24	birhor	põhja-munda	austroaasia	Maharastra, Chhattisgarh
Bonda	23	bonda	lõuna-munda	austroaasia	Orissa
Gadaba	6	bodo gadaba	lõuna-munda	austroaasia	Orissa
Ho	9	ho	põhja-munda	austroaasia	Bihar
Juang	16	juang	lõuna-munda	austroaasia	Orissa
Kharia	3	kharia	lõuna-munda	austroaasia	Chhattisgarh
Khasi	8	khasi	khasi-aslian	austroaasia	Meghalaya
Mahali	5	mahali	põhja-munda	austroaasia	Jharkhand
Mawasi	22	korku	põhja-munda	austroaasia	Jharkhand, Madhya Pradesh
Santhal	5	santali	põhja-munda	austroaasia	Jharkhand

Proove on varem kasutatud töödes Chaubey *et al.* 2008 ja 2011. Proovide kogumise kohad on näidatud India kaardil (joonis 9).



Joonis 9. Vereproovide võtmise kohad. Tähistused: sinine – põhja-munda keelte kõnelejad, punane – lõuna-munda keelte kõnelejad , kollane – baiga rahvas, pruun – khasi rahvas.

2.2.2 RFLP disain

RFLP analüüs on antud töös eelistatud meetod, sest see on sekveneerimisega võrreldes kiirem ja odavam. Lisaks on võimalik RFLP analüüsi teostada olukorras, kus PCR annab korraga erineva pikkusega produkte. Sellises olukorras sekveneerimiseks peaks kasutama *nested* PCRi, mis on kulukam ja annab suurem ajakulu.

Uuritavaid haplogruppe iseloomustavate SNP-de asukoht võeti uute markerite tabelist (Karmin *et al.*, 2015). Positsioonid määrati referentsgenoom GRCh37(veebruar 2014) järgi. Uuritav positsioon võeti koos külgnetavate aladega välja Ensembl-i andmebaasis asuvast referentsgenoomist GRCh37 (http://grch37.ensembl.org/Homo_sapiens/Info/Index). Sobivate restriktaaside olemasolu kontrolliti veebilehe *Insilicase RFLP enzyme picker* abil (http://www.insilicase.co.uk/Web/RFLP.aspx). Programm väljastab nimekirja sobivatest ensüümidest. Juhul kui programm ei tagastastanud ühtegi vastet, siis valiti uus positsioon. Seda korrati seni kuni leidus positsioon, mida on võimalik lõigata. Lõikekohtade arvu kontrolliti programmiga NEBcutter 2 (http://nc2.neb.com/NEBcutter2/index.php). Üle ühe lõikega ensüümid jäeti kõrvale ning kui ei jäänud alles ühtegi sobivat kandidaati, siis alustati algusest uue positsiooniga. Järgmise sammuna disainiti praimerid.

2.2.3 PCR praimerite disain

Esimesel viiel markeril olid praimerid juba disainitud Gyaneshwer Chaubey poolt. Järgnevad markerid (tabel 2) valiti töö käigus vastavalt tulemustele.

Praimerite genereerimiseks kasutati NCBI Primer Blast programmi (http://www.ncbi.nlm.nih.gov/tools/primer-blast/). Praimerite sobivus kontrolliti üle programmiga Primer 3 (http://primer3.ut.ee/) (Koressaar *et al.*, 2007; Untergrasser *et al.*, 2012).

Tabel 2. Genotüpiseerimisel kasutatud Y-kromosoomi markerid.

Marker	Positsioon	Muutus	Ensüüm	Lõikekoht (↓)	PCR produkti pikkus (ap)	RFLP tulemus		Puhver
B418	Y: 22578932	T → C	*BclI*	5′..T↓GATCA..3′	352	+	352	Tango
						-	286 + 66	
B419	Y: 7750676	A → G	*Bsh1236I*	5′..CG↓CG..3′	262	+	146 + 116	Tango
						-	262	
B422	Y: 2803226	C → T	-	-	233	+		-
						-		
B424	Y: 18249536	T → C	*TaaI*	5′..ACN↓GT..3′	299	+	121 + 178	Tango
						-	299	
B426	Y: 15097700	A → G	-	-	215	+		-
						-		
B426_eq	Y: 16794293	T → C	*MspI*	5′..C↓CGG..3′	309	+	144 + 165	Tango
						-	309	
M1284	Y: 6838496	C → G	*BseLI*	5′..CCNNNNN↓NNGG..3′	273	+	125 + 148	Tango
						-	273	

RFLP tulemuses + tähistab uuritavas positsioonis SNP olemasolu ja - SNP puudumist. Nende järel on vastav fragmendi pikkus geelil. Sobiva restriktaasi puudumisel positsioon sekveneeriti. Valgel taustal on Gyaneshwer Chaubey disainitud ning sinisel taustal minu poolt hiljem juurde disainitud markerid.

2.2.4 PCR ehk polümeraasi ahelreaktsioon

Genoomne DNA oli kontsentratsioonil 5-25ng/µl. PCR reaktsioonisegu maht oli 11,2 µl, millest 1 µl oli genoomne DNA. PCR reaktsioonisegu koostis:

7,1 µl milliQ deioniseeritud vesi

2,5 µl 5x Solis BioDyne FIREPol® Master Mix Ready to Load (12,5mM $MgCl_2$, reaktsioonipuhver B, dNTPd, kollane ja sinine värv.)

0,3 µl *forward* praimer (0,27 pmol/µl)

0,3 µl *reverse* praimer (0,27 pmol/µl)

1,0 µl uuritav genoomne DNA

Ilma DNA-ta kontrolli segusse jäeti DNA lisamata. Kasutusel olnud praimerid on lisas (lisa 1).

PCR reaktsioon toimus Biometra UNO II masinas. PCRi programm:

DNA esmane denaturatsioon	94 °C	3s	
DNA denaturatsioon	94 °C	20s	37 tsüklit
Praimerite seondumine DNA-ga	53-58 °C	20s	
Ekstensiooni faas	72 °C	35s	
Ekstensiooni lõpetamine	72 °C	3s	

2.2.4.1 Geelelektroforees

PCR-i õnnestumise selgitamiseks ning RFLP reaktsiooni järgselt fragmentide pikkuste kontrolliks valmistati 2%-ne agaroosgeel. Selleks kasutati 2 g agaroosi ja 100 ml 0,5-kordset TBE puhvrit, mille koostises oli 45 mM Tris-boraat ja 0,05 M EDTA-Na_2 ning pH oli 8,3. Tris-boraat hoiab DNA deprotoneeritud kujul ja vees lahustuvana ning EDTA-Na_2 seob kahevalentseid katioone, mis on kofaktoriteks nukleaasidele. DNA UV-kiirguses nähtavaks muutmiseks lisati jahtunud geelilahusele etiidiumbromiidi lõppkontsentratsiooniga 0,5 µl/ml. Geelelektroforeesi pikkusmarkeriks kasutati Fermentas O'RangeRuler™ 20bp DNA Ladder Ready to Load. Geelipilt jäädvustati UVI Pro Gold masinaga kasutades 260nm kiirgust. Pildid jäädvustati tarkvaraga UVI Pro v12.5.

2.2.5 RFLP ehk restriktsioonifragmentide pikkuspolümorfism

PCR produktile lisati RFLP reaktsioonisegu, mis sisaldas järgmisi reagente:

3 µl milliQ deioniseeritud vett

1,7 µl puhvrit 10x Tango

0,05 µl restriktsiooni ensüümi (0,5 U)

Tuubid pandi inkubaatorisse kolmeks tunniks kuni üleöö ensüümile vastava optimaalse temperatuuri juurde.

Bsh1236I , MspI	37 °C
BclI, BseLI	55 °C
TaaI	65 °C

2.2.6 PCRi produkti puhastamine sekveneerimiseks

PCRi produktile lisati 1µl 1:1 segu lõppkontsentratsiooniga 1U/µl eksonukleaasist *ExoI* (Thermo Scientific), mis lagundab üheahelalise DNA (praimerid) ja SAPst (Thermo Scientific) ehk kreveti aluselisest fosfataasist, mis defosforüleerib vabad nukleotiidid (dNTP). Reaktsioon viidi läbi Biometra UNO II masinas, seekord tingimustel 37 °C 20min nukleaasi ja fosfataasi tööks ning 80 °C 15min nende inaktivatsiooniks.

2.2.7 Sekveneerimine

Markereid mille jaoks ei leidunud sobivaid restriktaase sekveneeriti kasutades BigDye® Terminator v3.1 Cycle Sequencing Kit-i, mis põhineb Sangeri ensümaatilisel didesoksüterminatsioonil (Sanger *et al*., 1977). Selle käigus liidetakse praimeri pikendamise faasis suvalises kohas pikendatava ahela lõppu ddNTP (didesoksünukleotiidtrifosfaat), mis peatab edasise reaktsiooni antud ahela jaoks. Iga ddNTP on seotud (vastavalt nukleotiidile) ühega neljast fluorestseeuvast ühendist. Sekveneerimisproduktid puhastati ja anti edasi Eesti Biokeskuse tuumiklaborile. Saadud järjestused analüüsti programmi BioEdit v7.2.5 abil.

2.2.7.1 Sekveneerimise reaktsioon

Reaktsioonisegu, kogumahuga 10 µl, koosnes järgnevatest reagentidest:

6,1 µl milliQ deioniseeritud vesi

2 µl 5x BigDye® Terminator v3.1 Buffer

0,75 µl BigDye® Terminator v3.1 reagent premix

0,16 µl *forward* või *reverse* praimer (10 pmol/ µl)

1 või 2 µl PCRi produkt (vastavalt PCR produkti intensiivsusele geelil)

Praimerite järjestused on toodud tabelis lisa 1. Reaktsiooni läbiviimiseks kasutati Biometra UNO II masinat. Programm oli järgmine:

DNA denaturatsioon	95 °C	15 sek	30 tsüklit
Praimerite seondumine	52-56 °C	10 sek	
Ekstensioon	60 °C	1 min	

2.2.7.2 DNA sadestamine

Sekveneerimisreaktsioonile lisati 2 µl 1:1 ammoniumatsetaadi ja punase dekstraani segu (lõppkontsentratsioonidega 1,5 M NH4Ac-d ja 20 mg/ml dekstraani). Segul lasti seista 5 minutit. Seejärel lisati 30 µl -20°C 96% etanooli ning segu suspendeeriti. Proovid asetati 10 minutiks –20°C juurde sadenema ning seejärel tsentrifuugiti Hettich Zentrifugen MIKRO 22 tsentrifuugis 10 minutit RCF = 15 871g juures. Eemaldati supernatant ning pesuks lisati 200 µl –20°C 70% etanooli, tsentrifuugiti 5 minutit samal kiirusel. Eemaldati supernatant ning pesti veelkord 200 µl –20°C 70% etanooliga ning tsentrifuugiti 5 minutit samal kiirusel. Eemaldati supernatant ja proovid asetati termostaati 37°C juurde 10 minutiks kuivama. Sademele lisati 10 µl 70% formamiidi ja lasti toatemperatuuril 10 minutit seista.

2.3 Tulemused ja arutelu

Antud töö katsete järjekord ja uudikalleeliga(*derived*) proovide arv on esitatud tabelis (tabel 3) ning detailsed tulemused on kujutatud kokkuvõtvalt koos fülogeneesipuuga (joonis 11). Kõigi RFLP analüüsi tulemuse kontrollimiseks sekveneeriti kaks proovi, välja arvatud B418, mille PCR andis mitu produkti.

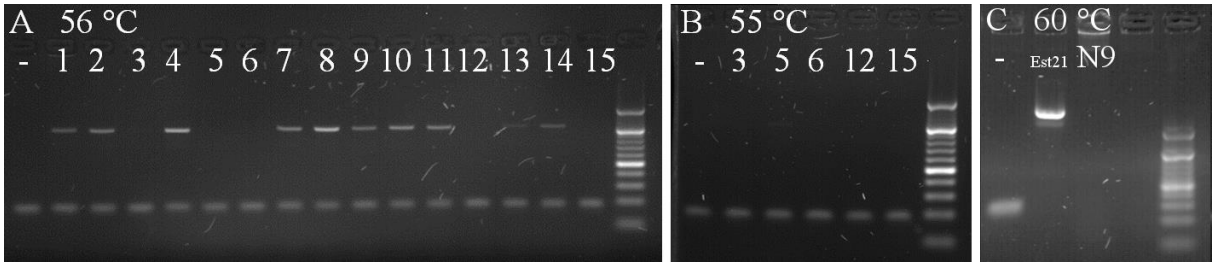
Tabel 3. Katsed ja tulemused.

Marker	Meetod	Uudik.al
B418	RFLP	61
B424	RFLP	0
B426	Sekveneerimine	17*
B426_eq	RFLP	9*
M1284	RFLP	16*
B422	Sekveneerimine	*

Markerid järjestatud uurimise järjekorras. Uudik.al - uudikalleeliga proovide arv. * numbri lõpus tähistab osade katsete ebaõnnestumist. B422 ebõnnestus täielikult.

Tööd alustati RFLP analüüsiga markerist B418. Kolmandikel proovidel oli antud markeri uudikalleel. Järgmiseks võeti marker B424, sest Y-kromosoomi täisjärjestuste põhjal kuulus üks AAKR esindaja Y-kromosoomi sellesse haplogruppi (Ho, Bihari osariigist, joonis 11) ja eeldati, et see võib olla AAKR populatsioonidele iseloomulik alamklaad. Tulemus oli vastupidine – kõik proovid omasid eellasalleeli (*ancestral*). Kahe proovi sekveneerimine kinnitas saadud tulemusi.

Edasi sekveneeriti marker B426, millele ei leidunud sobivat restriktaasi, aga kahjuks ligi kolmandik PCRe ei andnud produkte. Ebaõnnestunud katseid korrati seondumistemperatuuril 55 °C. Kuna algse katse (A joonis 10) ilma DNAta kontrollproovis oli näha soovimatut produkti, siis saastunud reagentide välistamiseks võeti korduskatsete jaoks uus milliQ, uus PCR master mix ning tehti uued praimeri lahjendused (joonis 10).

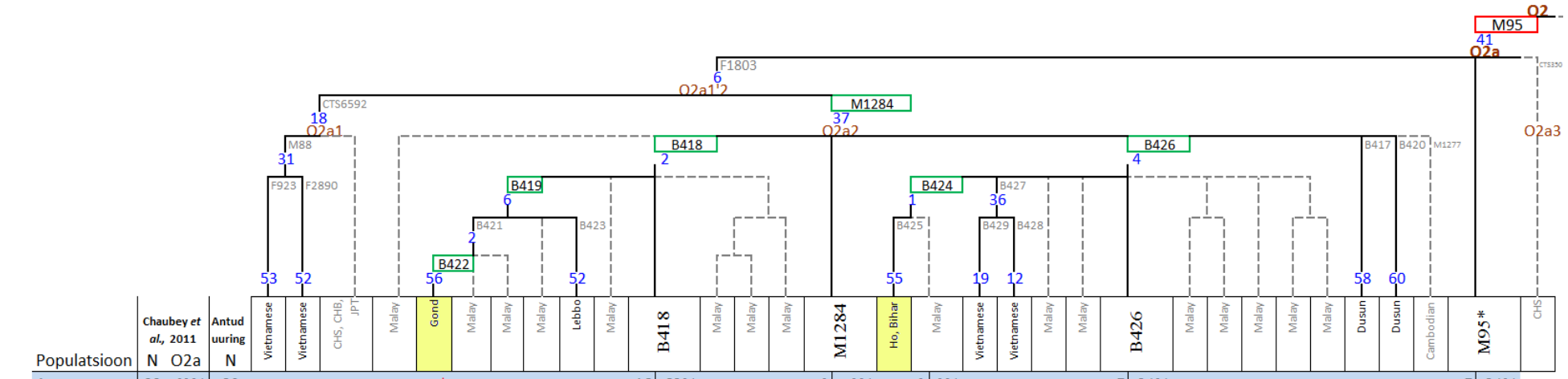


Joonis 10. Näiteid tehtud katsete geelipiltidest. Miinus tähistab kontrollproovi (ilma DNAta). A) B426 PCR seondumistemperatuuriga 56 °C. Oodatud produkti pikkus on 215 ap.. B)B426 ebaõnnestunud katsed tehtud uuesti 55 °C juures. C) M1284 kontrollreaktsioon koos naise DNAga (N9) 60 °C juures. Kasutatud pikkusmarker on vahedega 20 ap ning 100, 200, 300 ap on eredamad.

Kuna markerit B426 ei saanud kõikides proovides testida, disainiti uued praimerid ühele (kolmest) sama klaadi defineerivast alternatiivsetest positsioonidest. Uued praimerid ebaõnnestusid sarnase sagedusega. Kuigi enamus ebaõnnestumisi ei kattunud, jäid 21 proovi markeri suhtes testimata.

Järgnev marker M1284 asetseb fülogeneesipuul B426st ja B418st ühe astme võrra juurele lähemal. Selle põhjal lahenes 21st B426 suhtes analüüsimata jäänud proovist 10, sest need omasid markeri M1284 eellasalleeli ning jäid seega paragruppi M95*. Katsed markeriga M1284 ei läinud ka ilma probleemideta, 22 proovi puhul soovitud produkt ei amplifitserunud. Kokku jäi parahaplogrupi M95* alla 70 proovi, millest 29 ebaõnnestunud katsete tõttu.

Kokkuvõtteks kolmveerand proovidest jagunesid markeri B418 ja M95* vahel ära. Ülejäänud olid haplogrupis B426 või M1284. Põhja-munda populatsioonidel oli keskelt läbi kõrgeim haplogrupi B418 sagedus (48%)

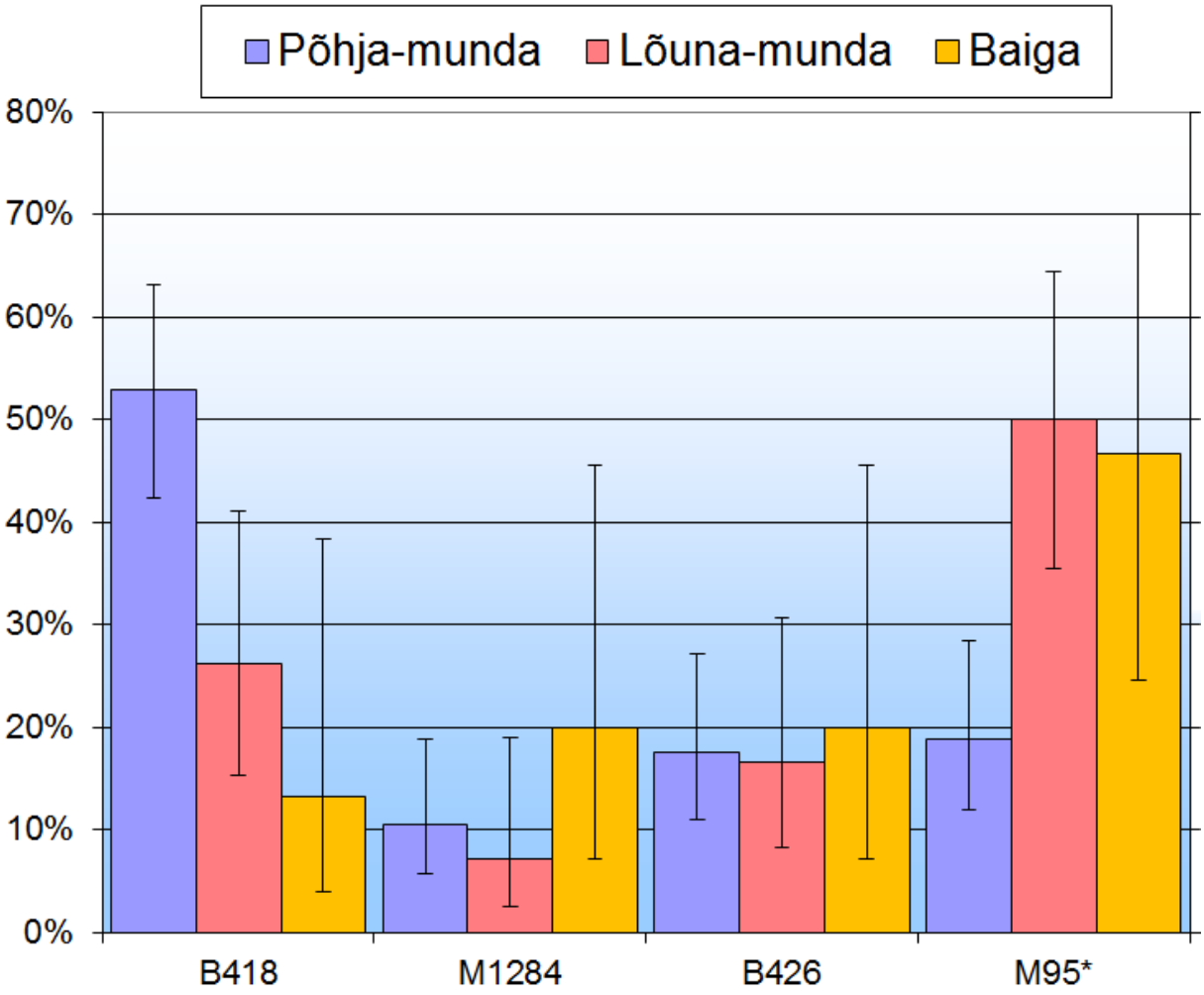


	Populatsioon	Chaubey *et al.*, 2011 N	Chaubey *et al.*, 2011 O2a	Antud uuring N	Vietnamese	Vietnamese	CHS, CHB, JPT	Malay	Gond	Malay	Malay	Malay	Lebbo	Malay	B418	Malay	Malay	Malay	M1284	Ho, Bihar	Malay	Vietnamese	Vietnamese	Malay	Malay	B426	Malay	Malay	Malay	Malay	Malay	Dusun	Dusun	Cambodian	M95*	CHS
Põhja-munda	Asur	88	60%	29					*		-			15	52%		0	0%		0	0%			7	24%									7	24%	
	Birhor	62	95%	24					*		-			6	25%		4	17%		0	0%			4	17%									10	42%	
	Ho	45	84%	9					*		-			4	44%		2	22%		0	0%			1	11%									2	22%	
	Mahali	32	19%	5					*		-			1	20%		0	0%		0	0%			0	0%									4	80%	
	Mawasi	39	72%	22					*		-			15	68%		3	14%		0	0%			2	9%									2	9%	
	Santhal	20	35%	5					*		-			4	80%		0	0%		0	0%			1	20%									0	0%	
				94										**45**	**48%**		**9**	**10%**						**15**	**16%**									**25**	**27%**	
Lõuna-munda	Bonda	42	95%	23					*		-			4	17%		0	0%		0	0%			2	9%									17	74%	
	Gadaba	27	89%	6					*		-			1	17%		0	0%		0	0%			1	17%									4	67%	
	Juang	54	89%	16					*		-			6	38%		3	19%		0	0%			3	19%									4	25%	
	Kharia	37	49%	3					*		-			0	0%		0	0%		0	0%			1	33%									2	67%	
				48										**11**	**23%**		**3**	**6%**						**7**	**15%**									**27**	**56%**	
	Baiga	65	71%	**23**					*		-			**2**	**9%**		**3**	**13%**		**0**	**0%**			**3**	**13%**									**15**	**65%**	
	Khasi	21	48%	**8**					*		-			**3**	**38%**		**1**	**13%**		**0**	**0%**			**1**	**13%**									**3**	**38%**	
	Kokku			**173**										**61**	**35%**		**16**	**9%**						**26**	**15%**									**70**	**40%**	

Joonis 11. Tulemused esitatud haplogrupi O2a-M95(punasega märgitud) alamharude fülogeneetilise puu all (kohandatud Karmin *et al*. 2015 järgi). Alamharude tipus on Karmin *et al*., 2015 töös genotüpiseeritud proovid. Kollase taustaga märgitud proovid on Indiast pärit. Tärniga haru M95* tähistab proove, mis potentsiaalselt kuuluvad mõnda alamharusse. Genotüpiseeritud markerid on märgitud rohelisega. Sinised numbrid tähistavad SNPde arvu, mis defineerivad vastavat haru. Miinusega on märgitud katsed milleni ei ole veel jõutud. Punane tärn tähistab katsete ebaõnnestumist.

M95* haplogruppi jäid need proovid,millel kõik testitud markerid olid eellasallelid ja/või PCR ebaõnnestus. Proovid, millel katse ei õnnestunud, tehti uus PCR reaktsioon samade praimeritega, aga madalama seondumistemperatuuriga. Osadel juhtudel see aitas, kuid enamasti tekkisid juurde ainult ebaspetsiifilised produktid. Ebaõnnestunud katsete tõttu M95* alla liigitatud proove oli põhja-munda proovides kokku 7 (7,4%), lõuna-munda proovides 6 (12,5%), Baiga proovides 8 (34,8%). 8 khasi prooviga probleeme ei esinenud.

Määratud haplogruppide sagedused on esitatud karp-vurrud diagrammina joonisel 12. Sellest analüüsist on välja jäetud ainukesed khasi-aslani keele kõneleja khasid, kunda valimi suurus on liiga väike. Samuti ei kaasatud ebõnnestunud katsete tõttu parahaplogruppi M95* jäänud proove.



Joonis 12. Uuritud markerite sagedused. Põhja- ja lõuna-munda populatsioonid on eraldi kokku grupeeritud, sest neid loetakse erinevateks keeleharudeks. Baiga rahvas on allikate vasturääkivuste tõttu eraldi. Khasi-asliani keeli esindas ainult 8 proovi, mida on liiga vähe täpse alleelisageduse saamiseks.

Antud töös tuvastati kõrge sagedusega (üle 50%) esinev haplogrupp B418 põhja-munda rahvastes. Teistes populatsioonides esinesid uuritud alamliinid madalama keskmise sagedusega ning sageduste statistilised usalduspiirid väga suured. Tulemuste analüüsil kahepoolse t-testiga selgus, et põhja- ja lõuna-munda keeleperekonna populatsioonid olid omavahel kahe markeri, B418($p<0,0001$) ja M95*($p<0,0001$), sageduste poolest oluliselt erinevad. Teiste markerite poolt defineeritud alamklaadide sageduste erinevused polnud statistiliselt olulised.

M95* sagedus näitab proovide hulk, millel on alamklaad veel genotüpiseerimata. Antud tulemustes paistab võimalik, et lõuna-munda keelte kõnelejate Y-kromosoomi haplogrupi O2a sisestruktuur on mitmekesisem kui põhja-munda keelte rääkijatel. Lõuna-munda populatsioonide suuremale geneetilisele mitmekesisusele räägivad kaasa ka meie uurimisgrupi veel avaldamata mtDNA andmed, mis näitavad mtDNA haplogruppide suuremat varieeruvust kui põhja-munda rahvastel. Proovide vähesuse tõttu olid baiga ja khasi standardhälbed suured ning nende kohta ei saa järeldusi teha.

Tulevikus tasub kindlasti edasi uurida proove, mis jäid parahaplogruppi M95*. Esialgu tuleks disainida praimerid markerile F1803, et teha kindlaks kas M95* haplogruppi langenud proovid on üldse O2a1’2 haplogrupis. Seejärel tuleks sekveneerida kogu MSY piirkond mõnel indiviidil, kelle Y-DNA haplogruppi ei õnnestu RFLP analüüsidega tuvastada.Võimalik, et saadud andmete põhjal saab defineerida juurde uue haplogrupp O2 alamharu. Kõikide proovide eduka genotüpiseerimisega saaks ka piisavalt andmeid, et anda objektiivsem hinnang Indias elavate AAKRide päritolu hüpoteeside kohta.

KOKKUVÕTE

Antud töös uuriti Y-kromosoomi haplogrupp O2a-M95 alamklaadide levikut India austroaasia keelte rääkijate seas, et harude leviku põhjal anda hinnang India austroaasia keelte rääkijate päritolu kohta käivatele hüpoteesidele. Selleks kasutati hiljuti Y-kromosoomi täisjärjestuste uuringus avaldatud markereid, mis annab võimaluse eristada haplogrupi O2a-M95 alamharusid kõrgema lahutusvõimega.

Töös genotüpiseeriti neli uut markerit M95st allavoolu 94 põhja-munda, 48 lõuna-munda, 23 baiga ning 8 khasi keele kõnelejate proovides. Antud proovidel oli varasemalt määratud markeri M95 uudikalleel. Töö käigus õnnestus alamhaplogrupp ära määrata 60% proovidest. Baiga ja khasi populatsioonide proove on liiga vähe, mille tõttu on haplogrupi sageduste standardhälve liiga suur usaldusväärsete järelduste tegemiseks.

Tulemustest leiti, et põhja-munda keelte kõnelejatel esineb kõrge sagedusega B418 alamliini., mis on statistiliselt oluliselt erinev lõuna-munda keelte kõnelejate markeri B418 sagedusest.

Parahaplogrupp M95* sagedusega oli olukord vastupidine, mis võib viidata lõuna-munda keelte rääkijate suuremale geneetilisele mitmekesisusele võrreldes põhja-munda keelte kõnelejatega. Seda väidet toeatavad ka uurimisrgrupi avaldamata mtDNA tulemused.

India AAKRide pärituolu hüpoteeside hindamiseks peaks tulevikus markerist M95 allavoolu haplogrupeerimata proovid genotüpiseerima, et saada piisavalt infot Y-kromosoomi haplogrupi O2a-M95 asutajaliinide leviku kohta neis populatsioonides. Selleks tuleks tuleks potentsiaalsete uute alamliinide leidmiseks mõned seni M95* paragruppi jäänud Y-kromosoomid täielikult sekveneerida ja saadud andmete põhjal genotüpiseerida ülejäänud proovid.

Counting the most recent O2a-M95 male founders of Indian Austroasiatics

Arno Pilvar

Summary

Y-chromosome has been a good tool in population genetics for decades, helping us study migrations and sex specific demographic processes. Y-chromosome is especially interesting because it is haploid and it is mostly left untouched by recombination. For that reason the accumulated mutations can be carried on to the next generation unchanged. Phylogenetic tree of Y-chromosome is based on these mutations and with it we can predict coalesence times between populations and possible migrations.

This study takes a look at Y-chromosomal haplogroup O2a-M95 among Indian Austroasiatic populations. Recent studies on long stretch (8-10MB) of Y-chromosome have increased the number of biallelic markers up to ten-fold. This gives us an opportunity to segregate the downstream branches of the haplogroup O2a-M95 with higher resolution. This would not only allow us to identify the minumum number of founders involved in bringing the language as well as genes from Southeast Asia, but also to define a precise timeline for this migration.

We have genotyped 94 North Munda, 48 South Munda, 8 Khasi-Khumaic speakers and 23 Baiga people, derived to O2a-M95 marker, for four newly defined markers. We have used RFLP method and validated the results by sequencing 2 samples per marker with Sanger sequencing method.

By genotyping four novel markers we were able to assign 60% of Indian Austroasiatic samples in to three subclades. We also observed a significant difference of marker B418 for North vs South Munda groups. B418 was higher in North Munda. The opposite was true for M95*, which may suggest that South Munda populations have greater genetic diversity than North Munda populations. Our workgroups unpublished mtDNA results also support this conclusion. In the timeline perspective it is noted that the sequencing based analysis has estimated the time of arrival after 5KYA which is 2-3 fold lower than STR estimates. Our next focus will be on revealing the phylogenetic status of M95 samples by seqencing a few of them followed by genotyping them in the pool.

KASUTATUD KIRJANDUS

Anon 2010. E pluribus unum. Nature Methods. p. 331

Ayub Q, Tyler-Smith C. (2009). Genetic variation in South Asia: assessing the influences of geography, language and ethnicity for understanding history and disease risk. Brief Funct Genomic Proteomic. 8(5):395-404

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

Bellwood PS, editor. 2005. First farmers. London: Wiley-Blackwell.

Bhowmick BK, Satta Y, Takahata N. (2007). The origin and evolution of human ampliconic gene families and ampliconic structure. Genome Res. 17:441–50

Charlesworth B. ja Charlesworth D. (2000) The degeneration of Y chromosomes. Philos Trans R Soc Lond B Biol Sci. 355(1403): 1563–1572.

Chaubey G *et al.*,(2008) Phylogeography of mtDNA haplogroup R7 in the Indian peninsula. BMC Evol Biol. 4;8:227.

Chaubey 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):1013-24

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.
chromosomes based on full sequence data, p. 357 *Human Evolutionary Genetics, second edition*. Garland Science, Taylor & Francis Group, LLC

Diegoli, (2015). Forensic typing of short tandem repeat markers on the X and Y chromosomes.

Diffloth G. 2001. Tentative calibration of time depths in Austroasiatic branches. Paper presented at the Colloque Perspective sur la Phylogenie des Langues d'Asoe Orientales at Perigueux, 30 August 2001.

Diffloth G. 2005. The contribution of linguistic palaeontology and Austro-Asiatic. in Laurent Sagart, Roger Blench and Alicia Sanchez-Mazas, eds. The Peopling of East Asia: Putting Together Archaeology, Linguistics and Genetics. 77–80. London: Routledge Curzon

Douglas C. Wallace, Michael D. Brown, Marie T. Lott, (1999). Mitochondrial DNA variation in human evolution and disease. Gene 238(1): 211–230
Forensic Sci Int Genet. pii: S1872-4973(15)00064-2

Frazer, K. A., Murray, S. S., Schork, N. J., Topol, E. J. (2009). Human genetic variation and ts contribution to complex traits. Nature Reviews Genetics, 10(4), 241-251.

Fuller D. 2007. Non-Human Genetics, Agricultural Origins and Historical Linguistics in South Asia. In: Petraglia M, Allchin B, editors. Vertebrate Paleobiology and Paleoanthropology. The Netherlands: Springer. p. 393–443.

Fuller D. 2011. Rice and the Austroasiatic and Hmong-Mien homelands. In *Dynamics of Human Diversity: The Case of Mainland Southeast Asia*m Nick J. Enfiled (ed.), 361-389. Canberra: Pacific Linguistics.

Grierson GA, 1904. Linguistic survey of India, Volume 6. lk 241

Hallast, *et al.* (2014). The Y-Chromosome Tree Burrsts into Leaf: 13,000 High-Confidence SNPs Covering the Majority of Known Clades. Mol. Biol. Evol. 32(3):661–673

Hammer MF, Karafet TM, Park H, Omoto K, Harihara S, Stoneking M, Horai S. (2005). Dual origins of the Japanese: common ground for hunter-gatherer and farmer Y chromosomes. J Hum Genet.51(1):47-58.

Higham C. 2003. Languages and farming dispersals: austroasiatic languages and rice cultivation. In: Bellwood P, Renfrew C, editors. Examining the farming/language dispersal hypothesis. Cambridge: The McDonald Institute for Archaeological Research. p. 223–233.

Hughes JF, Rozen S, (2012). Genomics and Genetics of Human and Primate Y Chromosomes. Annu. Rev. Genomics Hum. Genet. 13:3.1–3.26

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

Jobling, Hollox, Hurles, Kivisild, Tyler-Smith, 2014. Human Chromosomes and Human Karyotype, p. 28-31; Making inferences form diversity, p. 174; Phylogeny of 29 Y

Karafet TM., Mendez FL., Meilerman MB., Underhill PA., Zegura SL., Hammer MF. (2008). New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. Genome Res. 18(5): 830–838.

Karmin, M *et al*. (2015) A recent bottleneck of Y chromosome diversity consides with a global change in culture. Genome Res. 25(4):459-66

Kayser M, Brauer S, Weiss G, Schiefenhövel W, Underhill P, Shen P, Oefner P, Tommaseo-Ponzetta M, Stoneking M. (2003). Reduced Y-Chromosome, but Not Mitochondrial DNA, Diversity in Human Populations from West New Guinea. Am J Hum Genet. 72(2): 281–302.

Koressaar T, Remm M (2007) Enhancements and modifications of primer design program Primer3. Bioinformatics 23(10):1289-91

Kovach MJ, Sweeney MT, McCouch SR. (2007) New insights into the history of rice domestication. Trends Genet. 23(11):578-87.

Kruglyak, L., Nickerson, D. 2001. Variation is the spice of life. Nature Genetics, 27(3), 234.

Kumar V, Reddy AN, Babu JP, Rao TN, Langstieh BT, Thangaraj K, Reddy AG, Singh L, Reddy BM. (2007) Y-chromosome evidence suggests a common paternal heritage of Austro-Asiatic populations. BMC Evol Biol. 28;7:47.

Levy S, *et al*., (2007). The Diploid Genome Sequence of an Individual Human. PLoS Biology (5), 254.
Londo, J.P. *et al*. (2006) Phylogeography of Asian wild rice, Oryza rufipogon reveals multiple independent domestications of cultivated rice, Oryza sativa. Proc. Natl. Acad. Sci. U. S. A. 103, 9578–9583

Lu BR, Zheng KL, Qian HR, Zhuang JY. (2002) Genetic differentiation of wild relatives of rice as assessed by RFLP analysis. Theor Appl Genet. 106(1):101-6.

Lu, B.R. *et al*. (2002) Genetic differentiation of wild relatives of rice as assessed by RFLP analysis. Theor. Appl. Genet. 106, 101–106

Mangs HA, Morris BJ. (2007). The human pseudoautosomal region (PAR): origin, function and future. Curr. Genomics 8:129–36

Mendez F. L., Krahn T., Schrack B., Krahn A. M., Veeramah K. R., Woerner A. E., Fomine F. L., Bradman N., Thomas M. G., Karafet T. M., Hammer M. F. (2013). An African American paternal lineage adds an extremely ancient root to the human Y chromosome phylogenetic tree. *Am J Hum Genet* 92: 454-459.

Menz, Hansda, 2009. Baiga, p. 17 *Encyclopedia of Scheduled Tribes of Jharkhand*. Kalpaz Publications Delhi

Oka, H.I. ja Morishima, H. (1982) Phylogenetic differentiation of cultivated rice, potentiality of wild progenitors to evolve the *indica* and *japonica* types of rice cultivars. Euphytica 31, 41–50

Page DC, Harper ME, Love J, Botstein D. (1984). Occurrence of a transposition from the X-chromosome long arm to the Y-chromosome short arm during human evolution. Nature 311:119–23.

Russell RV, 1916. The Tribes and Castes of the Central Provinces of India, Volume 2. lk 77
Sachidanandam R, 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature. 2001 Feb 15;409(6822):928-33..

Sambrook *et al*. 1987, *Molecular Cloning Manual Vol 2*, pt. 9.16-9.19. Cold Spring Harbor Laboratory Press,U.S.

Scherer, (2007). Challenges and standards in integrating surveys of structural variation. Nat Genet. 39(7 Suppl):S7-15.

Skaletsky H, Kuroda-Kawaguchi T, Minx PJ, Cordum HS, Hillier L, et al. (2003). The malespecific region of the human Y chromosome is a mosaic of discrete sequence classes. Nature 423:825–37

Zhong H, Shi H, Qi XB, Duan ZY, Tan PP, Jin L, Su B, Ma RZ. (2011) Extended Y chromosome investigation suggests postglacial migrations of modern humans into East Asia via the northern route. Mol Biol Evol. 28(1):717-27

Zhongming, Z., Boerwinkle, E. (2002). Neighboring-Nucleotide Effects on Single Nucleotide Polymorphisms: A Study of 2.6 Million Polymorphisms Across the Human Genome. Genome Research (12), 1679-1686.

Thangaraj K, *et al*., 2009. Deep rooting in-situ expansion of mtDNA Haplogroup R8 in South Asia. PloS One. 4:e6545.

The 1000 Genomes Project Consortium. (2012) An integrated map of genetic variation from 1,092 human genomes, , Nature 491, 56–65 (01. november 2012) doi:10.1038/nature11632

The International HapMap 3 Consortium (2010) Integrating common and rare genetic variation in diverse human populations. Nature 467: 52-58

The International HapMap Consortium (2003) The International HapMap Project. Nature 426: 789-796

The International HapMap Consortium (2007) A second generation human haplotype map over 3.1 million SNPs. Nature 449: 851-861

The MHC sequencing consortium (1999). Complete sequence and gene map of a human major histocompatibility complex. Nature, 401(6756), 921.
Untergrasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, Rozen SG (2012) Primer3 - new capabilities and interfaces. Nucleic Acids Research 40(15):e115

van Driem G. (2012) The ethnolinguistic identity of the domesticators of Asian rice, Comptes Rendus Palevol, 11 (2): 117-132.

van Driem G. 2001. Languages of the Himalayas: An Ethnolinguistic Handbook of the Greater Himalayan Region, containing an Introduction to the Symbiotic Theory of Language (2 vols.). Leiden: Brill. [xxvi + 1375 = 1401 pp.]

van Oven M, Van Geystelen A, Kayser M, Decorte R, Larmuseau MH. (2014) Seeing the wood for the trees: a minimal reference phylogeny for the human Y chromosome. Hum Mutat. 35(2):187-191.

Veerappa, A. *et al.* (2013) Copy number variation-based polymorphism in a new pseudoautosomal region 3 (PAR3) of a human X-chromosome-transposed region (XTR) in the Y chromosome. Funct. Integr. Genomics 3, 285-293

Wei, W *et al*. (2013) A calibrated human Y-chromosomal phylogeny based on resequencing. Genome Res. 23(2):388-395.

Xue Y., Wang Q., Long Q., Ng B. L., Swerdlow H., Burton J., Skuce C., Taylor R., Abdellah Z., Zhao Y., Asan, MacArthur D. G., Quail M. A., Carter N. P., Yang H., Tyler-Smith C. (2009). Human Y chromosome base-substitution mutation rate measured by direct sequencing in a deep-rooting pedigree. *Current Biology* 19: 1453-1457.

KASUTATUD VEEBIAADRESSID

http://genome.wellcome.ac.uk/doc_WTD020876.html

http://web.ornl.gov/sci/techresources/Human_Genome/project/index.shtml

http://www.ensembl.org/Homo_sapiens/Location/Genome?r=1

http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/

http://www.biokeemiaselts.ee/?mid=9&lang=et

http://lcweb2.loc.gov/frd/cs/profiles/India.pdf

https://www.ethnologue.com/statistics/family

http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human/

http://www.isogg.org/

http://www.phylotree.org/Y/

http://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.29

LISAD

Lisa 1. Kasutatud praimerite nimekiri. Roosal taustal on Gyaneshwer Chaubey poolt ning sinisel minu poolt disainitud praimerid. Praimerid telliti firmast DNA Technology A/S, mis asub Taanis.

Praimer	Järjestus	Pikkus	GC%	T_m (°C)
B418F	GGGGGTGAGAAGGTTTGTTA	20	50	57,3
B418R	CAGGGAGAAGGAACACTTTCAA	22	45,5	58,4
B419F	GCCCAGTGATATGACACAAT	20	45	55,3
B419R	GCAGATCACAATGTGAGGAG	20	50	57,3
B422F	AGGAACGTTGTGACGGAAAC	20	50	57,3
B422R	GTTCGGAGCTGACAAAAAGC	20	50	57,3
B424F	ACGTGCAATCTCACAGGTTT	20	45	55,3
B424R	ATGTCCCAGATTGCTGAGAA	20	45	55,3
B426F	GCTCCCAAACATTTCTGCTA	20	45	55,3
B426R	TACCCAGGGTTCAAGATAGC	20	50	57,3
B426_eqF	TTCATTGTGAGAAAGGGCCTC	21	47,6	53,7
B426_eqR	ATCATCCAACTGCTTCCCAACT	22	45,5	55,2
M1284F	TTTCACCCAGGCTGCAGT	18	55,6	55,6
M1284R	AGTACTTAGGCCAGGCACAG	20	55	55,0

LIHTLITSENTS

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina, Arno Pilvar

(sünnikuupäev: 04.01.1992)

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

„Y-kromosomaalse haplogrupi O2a alamharude esinemine India austroaasia keeli kõnelevatel rahvastel“,

mille juhendajad on *M. Sc.* Monika Karmin, *Ph.D* Gyaneshwer Chaubey ja *Ph. D.* Ene 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 alates 1.06.2018 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