TARTU ÜLIKOOL
LOODUS- JA TEHNOLOOGIATEADUSKOND
MOLEKULAAR- JA RAKUBIOLOOGIA INSTITUUT
BIOINFORMAATIKA ÕPPETOOL

Andres Veidenberg

DNA koopiaarvu määramine ja koopiaarvu muutuste pärandumise uurimine genotüpiseerimiskiipide andmete põhjal

Magistritöö

juhendaja: Priit Palta, M.Sc.

Tartu 2009

Sisukord

Kasutatud lühendid 4

Sissejuhatus 5

1 Kirjanduse ülevaade 6

1.1 DNA koopiaarvu varieeruvus inimese genoomis 6

1.2 DNA mikrokiipide kasutamine CNV-de määramisel 6

1.2.1 Mikrokiibil põhinev võrdlev genoomne hübridisatsioon 7

1.2.2 Genotüpiseerimiskiibid 9

1.3 CNV-de määramine genotüpiseerimisandmetest 11

2 Töö eesmärk 14

3 Materjal ja metoodika 15

3.1 Algandmed 15

3.2 Kasutatud programmid ja nende tööpõhimõte 16

3.2.1 Arvutuslik meetod CNV-de määramiseks 16

3.2.2 Arvutusliku meetodi valideerimine 18

3.2.3 CNV-de pärandumise analüüs 18

4 Tulemused 19

4.1 CNV-de määramise meetodi väljatöötamine 19

4.2 Väljatöötatud metoodikaga leitud CNV-de valideerimine 20

4.3 CNV-de päranduvuse uurimine 24

4.3.1 CNV-d HapMap triodes 24

4.3.2 CNV-d Eesti VAM uuringu perekondades 27

Arutelu ja järeldused 29

Kokkuvõte 32

Summary (kokkuvõte inglise keeles) 33

Tänuavaldused 34

Kasutatud kirjandus 35

Lisad 39

Kasutatud lühendid

aCGH	Mikrokiibipõhine genoomne hübridisatsioon (*array-based Comparative Genomic Hybridisation*)
BAC	Bakteriaalne tehiskromosoom (*Bacterial Artificial Chromosome*)
BAF	B-alleeli sagedus (*B-Allele Frequency*)
CNV	Koopiaarvu variant/varieeruvus (*Copy-Number Variant/Variation*)
DASH	Dünaamiline alleelispetsiifiline hübridisatsioon (*Dynamic Allele-Specific Hybridisation*)
HMM	Peidetud Markovi mudel (*Hidden Markov Model*)
kb	10^3 aluspaari (*kilo base*)
Mb	10^6 aluspaari (*Mega base*)
LOH	Heterosügootsuse kadu (*Loss Of Heterozygosity*)
LRR	Proovi summaarne signaaliintentsiivsus (*Log R Ratio*)
MLAP	Multipleks ligeerimissõltuvate proovide aplifikatsioon (*Multiplex Ligation depednant Allelic Probe hybridisation*)
MAPH	Multipleks amplifitseeritavate proovide hübridisatsioon (*Multiplex Aplifable Probe Hybridisation*)
qPCR	Kvantitatiivne polümeraasi ahelreaktsioon (*quantitative Polymerase Chain Reaction*)
RT-qPCR	Reaalaja kvantitatiivne polümeraasi ahelreaktsioon (*Real-Time quantitative Polymerase Chain Reaction*)
ROMA	*Represantional Oligonucleotide Microarray Analysis*
SNP	Ühenukleotiidne polümorfism (*Single Nucleotide Polymorphism*)

Sissejuhatus

Kuigi tsütogeneetikas on kromosomaalsel tasemel DNA lõikude koopiaarvu muutusi indiviidide vahel uuritud juba pikka aega, on submikroskoopsel tasemel selline varieeruvus leidnud suuremat tähelepanu alles viimastel aastatel. Koopiaarvult varieeruvad DNA piirkonnad katavad inimgenoomist suurema ala kui ükski teine polümorfism, mõjutades lisaks normaalsele geneetilisele varieeruvusele ka paljude geneetiliste haiguste ja kasvajate arengut.

DNA koopiaarvu variantide ehk *CNV*-de leidmiseks on levinud mikrokiibipõhised meetodid. Kuigi mikrokiibid võimaldavad kättesaadava hinnaga teha ülegenoomseid analüüse, on meetodi nõrgaks kohaks keerukas andmeanalüüs. Mikrokiibilt loetud pidevat tüüpi arvulised signaaliintensiivsused tuleb jaotada kindla koopiaarvu ja otspunktidega genoomseteks piirkondadeks.

Uueks väljakutseks *CNV*-de uurimisel on nende alleelse päranduvuse analüüsimine, mis eeldab *CNV*-de koopiaarvu leidmist diploidse genoomi kummagi alleeli jaoks eraldi. Selle probleemi üheks võimalikuks lahenduseks on genotüüpide kui lisainformatsiooni kasutamine genotüpiseerimiskiipidelt.

Antud magistritöö tutvustab *CNV*-de määramist genotüpiseerimiskiipide abil ja kirjeldab täpsemalt *in silico* metoodikat genotüpiseerimisandmete põhjal *CNV*-de leidmiseks ja *CNV*-de pärandumise uurimiseks.

1 Kirjanduse ülevaade

1.1 DNA koopiaarvu varieeruvus inimese genoomis

DNA koopiaarvu varieeruvus (*CNV*) on üks olulisemaid varieeruvuse tüüpe inimese genoomis (Redon *et al.*, 2006). Kuigi DNA lõikude koopiate muutusi indiviidide vahel on täheldatud tsütogeneetika algusaegadest peale, seostati seda pigem haigustega kui üldise varieeruvusega. Viimastel aastatel on aga mikrokiibi tehnoloogiate areng võimaldanud uurida koopiaarvu varieeruvusi submikroskoopsel skaalal (u. 1 *kb* - 5 *Mb*) ning alates esimestest submikroskoopsete *CNV*-de kaardistamistest 2004. aastal (Iafrate *et al.*, 2004; Sebat *et al.*, 2004) on saanud selgeks, et DNA koopiaarvu varieeruvused on palju laiema levikuga kui seni arvatud. *CNV*-d mõjutavad (nii kattuvate kui lähedalasuvate) geenide ekspressiooni, põhjustavad paljusid haigusi ja osalevad vähi arengus (Sebat *et al.*, 2007; Cook & Scherer 2008; Henrichsen *et al.*, 2009).

Redon *et al*. leidsid 2006. aasta uuringus, et *CNV*-sid leidub üle terve inimese genoomi, hõlmates rohkem aluspaare (12% genoomist) kui ühenukleotiidsed polümorfismid. Seega on koopiaarvu variatsioonid väga oluline variatsiooni tüüp inimese genoomis, ja selle paremaks mõistmiseks tuleb läbi viia detailseid kaardistamisi. *CNV*-de kaardistamiseks on välja töötatud terve rida meetodeid, millest üheks enam levinud vahendiks on mikrokiibitehnoloogiad (Carter *et al.*, 2007).

1.2 DNA mikrokiipide kasutamine *CNV*-de määramisel

CNV-de leidmiseks kasutatakse laialdaselt mikrokiipidel põhinevaid meetodeid, kuigi see ei ole kaugeltki ainus võimalus. *CNV*-de määramiseks on kasutusel ka kvantitatiivne polümerisatsioonireaktsioon (*qPCR*), multipleks amplifitseeritavate proovide hübridisatsioon (*MAPH*), multipleks ligeerimissõltuvate proovide amplifikatsioon (*MLPA*) ja dünaamiline alleelispetsiifiline hübridisatsioon (*DASH*), kuigi need on piiratud analüüsiulatuse ja -mahuga (Carter *et al.*, 2007). *CNV*-de määramisel täpsusega hiilgavad sekveneerimispõhised meetodid on aga väga kallid, kuigi uue põlvkonna tehnoloogiad (454, SOLiD) annavad lootust hinna odavnemiseks tulevikus. Levinuim meetod *CNV*-de määramiseks on siiski DNA mikrokiibid, millega saab ülegenoomselt analüüsida paljusid lookuseid korraga. Kasutatavamad DNA

mikrokiipide tehnoloogiad põhinevad võrdleval genoomsel hübridisatsioonil või *SNP*-de genotüpiseerimisel.

1.2.1 Mikrokiibil põhinev võrdlev genoomne hübridisatsioon

Üheks DNA koopiaarvu määramisel kasutatavaks tehnikaks on võrdlev genoomne hübridisatsioon. See 1992. aastal tutvustatud meetod kasutab *FISH* (*fluoresence in situ hybridisation*) analüüsi normaalsetel metastaasi kromosoomidel, et tuvastada DNA koopiaarvu erinevusi test- ja referentsgenoomi vahel (Kallioniemi *et al.*, 1992). Referentsgenoomi suhtes DNA regioonide kadu või juurde lisandumist loetakse välja fluorokroomidega märgitud proovide hübridiseerumisel eralduva fluoresentssignaalide suhtest.

Kuigi võrdlev genoomne hübridisatsioon (*CGH*) oli pöördelise tähtsusega tsütogeneetikas, eriti vähiuuringutes, on meetodi oluliseks puuduseks suhteliselt madal lahutusvõime (olenevalt koopiaarvust 2-10 *Mb*) (Kallioniemi *et al.*, 1996; Carter *et al.*, 2007). Ühe lahendusena kirjeldas Solinas-Toldo koos kolleegidega 1997. aastal *CGH* modifikatsiooni, kus metastaasi kromosoomide asemel kasutati tahkele kandjale kinnitatud genoomsete kloonide maatriksit. Kuna tehiskromosoomides paljundatud järjestuste (näiteks *BAC* kloonid) asukoht genoomis on teada, (suur kaardistatud kloonide allikas on Inimese Genoomi Projekt), saab kiibilt loetud järjestikuste kloonide fluoresentssignaalide suhtarvudest kokku panna uuritava regiooni koopiaarvu profiili.

ArrayCGH nime saanud tehnoloogia baseerub võrdleva genoomse hübridisatsiooni tehnikal ja selle tööpõhimõte on kokkuvõtvalt järgmine: testitav DNA ja referents DNA (mille suhtes test DNA-d võrreldakse) märgitakse kumbki erineva fluorokroomiga. Märgistatud DNA hübridiseeritakse mikrokiibil olevatele proovidele koos kordusjärjestusi supresseeriva *Cot1* DNA-ga. Seejärel pestakse kiibilt halvasti või mitteseondunud DNA. Lõpuks skaneeritakse kiibid fluorokroome ergastava laservalgusega ja saadud fluoressentssignaalide intensiivsused peegeldavad proovidega hübridiseerunud test- ja referents DNA hulka.

Kiibipõhise genoomse hübridisatisooni lahutusvõime sõltub kiibile kinnitatud järjestuste pikkusest ning tihedusest (omavahelisest kaugusest genoomis) (Oostlander *et al.*, 2004). *CNV* uuringute laiemale levikule aluse pannud tööd kasutasid proovidena *BAC* kloone (Iafrate *et al.*, 2004) ja *in situ* sünteesitud oligonukleotiide (Sebat *et al.*, 2004). Mõlemal lähenemisel on omad eelised ja puudused, lisaks teoreetilisele fragmentide pikkusele ja paiknemisele genoomis ka signaalikvaliteedi ja töömahukuse osas.

Inimese genoomile kaardistatud *BAC* kloonid, mis panid aluse esimestele *arrayCGH* töödele (Solinas-Toldo *et al.*, 1997, Pinkel *et al.*, 1998) sisaldavad inserte vahemikus 20-100 *kb* ja seavad seega piirangu seda tüüpi mikrokiipide lahutusvõimele. Genoomsete kloonide kasutamisel on hübridisatsioonisignaalid küll hea signaali-müra suhtega, kuid ka lühemate kloonide puhul (fosmiidid, kosmiidid) jääb detekteerimislaeks kuni 15 *kb* suurused *CNV*-d (Carter *et al.*, 2007). Lisaks ei ole genoomseid kloone võimalik selekteerida nendes sisalduvate järjestuste alusel, mistõttu võib kiibile kinnitatud proovide hulgas olla ka kordusjärjestusi ja muid analüüsi keerukust tõstvaid DNA elemente (uuritava järjestusega sarnased pseudogeenid jms.) (Mantripragada *et al.*, 2004).

Kõige parema lahutusvõimega mikrokiibipõhist genoomset hübridisatsiooni võimaldab sünteesitud oligonukleotiidproovide kasutamine. Kuna sünteesitavatele järjestustele ei ole genoomsete kloonidega sarnaseid piiranguid, on näiteks ühenukleotiidse nihkega genoomses ülekattes proovide korral võimalik saavutada aluspaari tasemel lahutusvõime (Gribble *et al.*, 2007), kuigi sel moel kogu genoomi katmine (u. 2 miljoni prooviga) osutuks väga kulukas. Kuna erinevalt *BAC* proovidest on oligonukleotiidproovid unikaalse järjestusega, ei ole hübridisatsioonisegusse vaja lisada kordusjärjestusi blokeerivat DNA-d. Meetodi puuduseks on madal signaali-müra suhe, mis põhjustab suurt varieeruvust signaaliintensiivsustes (võrreldes *BAC* kloonidega võib esineda kuni 5-kordne erinevus). Probleemi lahendamiseks on proovitud vähendada inimese genoomi komplekssust, lõigates hübridiseeritava DNA restriktaasidega väiksemateks fragmentideks. *ROMA* nime saanud meetod kasutab uuritava DNA fragmentidest edasisel hübridiseerimisel vaid alla 1.2 *kb* pikkusi järjestusi, mille tulemusel väheneb koos analüüsitava DNA kompleksusega paraku ka analüüsikiibi esindatus (kattes vaid ~2,5% genoomist) (Lucito *et al.*, 2003).

Sarnaselt võrdleva genoomse hübridisatsiooniga on kiibile viidud ka multipleks amplitseeritavate proovide hübridisatsioonitehnika (*MAPH*) (Armour *et al.*, 2000). Mikrokiibipõhine *MAPH* e. *ArrayMAPH* (Patsalis *et al.*, 2007) on *arrayCGH*-st küll tundlikum, kuid ka töömahukam, kuna nõuab kahte hübridisatsiooniprotsessi ja uuritava DNA eelnevat immobiliseerimist nailonfiltrile.

Lisaks kahele eelkirjeldatud mikrokiibipõhisele *CNV*-de analüüsimeetodile (*arrayCGH* ja *arrayMAPH*) on viimasel ajal järjest rohkem kasutust leidnud genotüpiseerimiskiibid, mis lisaks ühenukleotiidsetele polümorfismidele (*SNP*-dele) inimese genoomis võimaldavad hinnata ka DNA koopiaarvu uuritava DNA genotüpiseeritavates *SNP* lookustes.

1.2.2 Genotüpiseerimiskiibid

Kuigi genotüpiseerimiskiibid arendati algselt välja ainult ühenukleotiidsete polümorfismide (*SNP*-de) detekteerimiseks, on viimastel aastatel sama tehnoloogiat edukalt kasutatud ka *CNV*-de määramisel. Genotüpiseerimiskiipide tööpõhimõte on sarnane *arrayCGH*-le, kus uuritav DNA hübridiseeritakse lookusspetsiifiliselt disainitud oligonukleotiidproovidele ning seondunud DNA hulgale vastavad fluoressentssignaalide intensiivsused on aluseks *CNV*-de koopiaarvu määramisel. Lisaks oligonukleotiidproovide heale lahutusvõimele on genotüpiseerimiskiipide suureks eeliseks *SNP* alleelide tuvastamine paralleelselt *CNV*-de määramisega, seega on uuritavad *CNV*-d paremini kirjeldatud. Lisaks sellele saab genotüpiseerimiskiipidega tuvastada olulisi koopiaarvu neutraalseid piirkondi (näiteks *LOH*-regioonid) (Bignell *et al.*, 2004; Peiffer *et al.*, 2006).

Genotüpiseerimiskiibid on laialt levinud töövahendid geneetilistes uuringutes ja seda paljuski tänu kommertsiaalsetele kiibitootjatele nagu Affymetrix ja Illumina. Selle tulemusena on genotüpiseerimiskiibid ja nende kasutamine hästi standardiseeritud ja stabiilse kvaliteediga, mis aitab kaasa analüüsitulemuste varieeruvuse ja reprodutseeritavuse paranemisele (Carter *et al.*, 2007). Affymetrixi GeneChip® perekonna genotüpiseerimiskiibid (näiteks 10K, 100K, 500K) kasutavad iga uuritava *SNP* määramisel komplekti alleelispetsiifilisi 26-meerseid oligonukleotiidproove. Kuna mõlemaid alleele esindavad oligonukleotiidjärjestused on varieeruva hübridiseerumisefektiivsusega, saab iga *SNP* jaoks proovide fluoressentssignaali intensiivsusi vastava tarkvara abil võrreldes arvutada nii genotüübi- kui koopiaarvu andmed (Affymetrix, Inc., 2005; Macconaill *et al.*, 2007).

Teiseks levinud genotüpiseerimisplatvormiks on Illumina BeadArray® tehnoloogial põhinevad mikrokiibid. Erinevalt Affymetrixi kiipidest on oligonukleotiidproovid kinnitatud indekseeritud mikrokerakestele, mis on juhuslikus järjekorras paigutatud klaaskandjale. *SNP* alleelide määramisel kasutatakse alleelispetsiifilist märklaudjärjestuste praimerekstensiooni, mis koos lookusspetsiifilise hübridisatsiooniga annab hea signaali-müra suhte. Lisaks on BeadArray® kiipide eeliseks kogu uuritava DNA kasutamine analüüsiprotsessis, samas kui Affymetrixi tehnoloogia puhul selekteeritakse kiipidele hübridiseerimiseks kuni 200-aluspaarilised DNA fragmendid (nagu *ROMA* puhul, komplekssuse vähendamiseks).

Illumina genotüpiseerimiskiibi tööpõhimõte on ülevaatlikult järgnev: kõigepealt uuritav DNA amplifitseeritakse ja fragmenteeritakse ning hübridiseeritakse *SNP* lookustele vastavate oligonukleotiidsete proovidega kiibile. Pärast hübridisatsiooni pikendatakse DNA-ga seondunud oligonukleotiidid kiibipõhise ensümaatilise praimerekstensiooni abil hübridiseerunud DNA *SNP*

positsioonile komplementaarse nukleotiidi võrra. Praimerekstensioonis kasutatavad dideoksünukleotiidid on immunokeemiliselt märgistatud alleelispetsiifiliste fluorokroomidega ja mida rohkem on DNA-ga hübridiseerunud oligonukleotiidproove, seda suurem arv järjestusi praimerekstensiooni käigus märgitakse. Viimase sammuna pestakse kiibilt halvasti või mitteseondunud järjestused ja fluoressentssignaalide intensiivsused skaneeritakse. Tulemuseks on pildifailid (kummagi fluorokroomi jaoks eraldi), millest on vastava tarkvara abil võimalik välja lugeda alleelispetsiifilised signaaliintensiivsused (Gunderson *et al*., 2005; Steemers *et al*., 2006).

Koos *arrayCGH*-ga on genotüpiseerimiskiibid kõige levinumad mikrokiibitehnoloogiad *CNV*-de analüüsiks. Erinevalt *arrayCGH*-st ei kohübridiseerita genotüpiseerimiskiipidele kahte DNA-d vaid iga uuritav DNA analüüsitakse eraldi kiibiga. *CNV*-de määramiseks võrreldakse uuritava DNA signaaliintensiivsusi ühe või mitme referents DNA genotüpiseerimisel saadud signaaliintensiivsustega. See võimaldab lisaks *CNV* määramisele vahet teha kas koopiaarvu muutus pärineb test- või referents DNA-st. Kuigi paljud genotüpiseerimiskiibid katavad suhteliselt hea lahutusvõimega kogu uuritava genoomi, ei ole *SNP* markerid jaotunud üle genoomi ühtlaselt, kuna teatud regioonidesse (kordusjärjestustega alad) on keeruline polümorfseid markereid disainida. Seepärast on nii Illumina kui Affymetrix oma *CNV*-uuringutele suunatud genotüpiseerimiskiipidele (Illumina Human 1M, Affymetrix 6.0) lisanud peale *SNP*-proovide ka mittepolümorfseid markereid, et katta olulisemad *CNV*-de piirkonnad (mida leidub rohkelt just kordusjärjestustega piirkondades) (Macconaill *et al*., 2007; Carter *et al*., 2007).

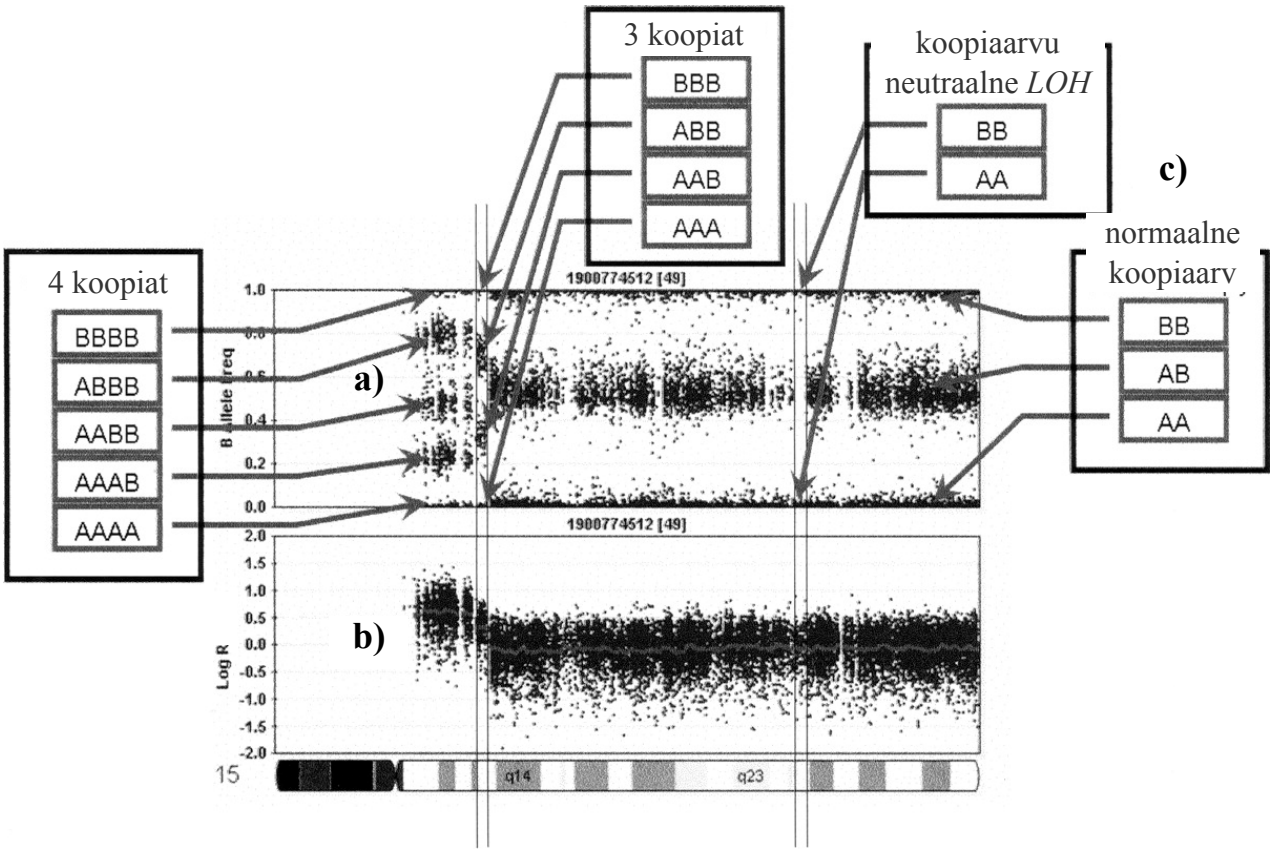
Üheks suuremaks probleemiks genotüpiseerimiskiipide analüüsil on signaaliintensiivsuste varieeruvus. Muutlikkust põhjustab sageli *PCR*-põhine DNA amplifitseerimine, mida Affymetrixi kiipidel kasutatakse muuhulgas genoomse komplekssuse vähendamiseks (väikeste fragmentide analüüs). Kuigi Illumina tehnoloogia ei kasuta DNA amplifitseerimiseks *PCR*-i, on varieeruvus siiski probleemiks. Seetõttu ei piisa usaldusväärseks *CNV* analüüsiks ainult ühe andmepunkti (*SNP*) hübridiseerimise andmetest vaid korraga tuleb arvesse võtta mitme järjestikuse markeri signaaliintensiivsused. Näiteks Bignell *et al.* kasutasid oma analüüsides kolme järjestikuse *SNP* andmeid (Bignell *et al.*, 2004), samas Peiffer ja kolleegid pidasid optimaalseks kümne markeri signaaliintensiivsuste keskmist või mediaani (Peiffer *et al.*, 2006). Mitme markeri andmete summeerimisel tuleb samas arvestada analüüsi lahutusvõime vähenemisega.

1.3 *CNV*-de määramine genotüpiseerimisandmetest

CNV-de määramine genotüpiseerimisandmetest ei ole lihtne ülesanne, kuna lisaks varieeruvusele tuleb arvestada ka analüüsiandmete tüübiga - kui *CNV* koopianumber on diskreetne täisarvuline väärtus (0 ja 1 - deletsioon; 3, 4, jne. - duplikatsioon), siis hübridiseerumistugevust näitavad signaaliintensiivsused on pidevad arvtunnused. Kiibipõhistest hübridiseerimisandmetest *CNV*-sid määraval arvutitarkvaral tuleb signaaliintensiivsuste jada piisavalt suure statistilise usaldusväärsusega jagada kindlateks *CNV* ja normaalse koopiaarvuga regioonideks. Selle ülesande jaoks on arendatud terve rida meetodeid, alates lihtsatest (ühel või mitmetel) läviväärtustel põhinevatest lähenemistest (Vermeesch *et al*., 2005; Fiegler *et al*., 2006) kuni komplekssete statistiliste modelleerimisteni. Nii analüüsimeetodi vead kui andmete varieeruvus viib valepositiivsete (*CNV* detekteerimine normaalse koopiaarvuga regioonis) ja valenegatiivsete (leidmata jäänud *CNV*-d) tulemusteni. Paljud avastatud *CNV*-d pannakse kirja üldkasutatavatesse andmebaasidesse (näiteks *Database of Genomic Variants*: [http://projects.tcag.ca/variation]), mille andmekogusid kasutatakse järgnevates uurimustöödes. Seetõttu on oluline, et *CNV*-de määramise metoodikad vähendaksid valepositiivsete tõenäosuse miinimumini, samas ohverdamata detekteerimisvõimet (leidmata või valetulemuseks loetud *CNV*-d).

Lihtsamate *CNV*-de määramise meetodite hulka kuuluvad segmentatsioonialgoritmid, nagu SW-ARRAY (Price *et al.*, 2005) ja CBS (Olshen *et al.*, 2004) eeldavad minimaalset andmete struktuuri ja valepositiivsete tulemuste vältimiseks seatakse signaaliintensiivsuste segmenteerimisel suhteliselt range lävi, mille tõttu jääb suur osa *CNV*-sid leidmata. Samas nii *arrayCGH* kui genotüpiseerimiskiipide andmete puhul on *CNV*-de määramiseks välja töötatud ka mitmeid signaaliintensiivsuste struktuuri modelleerivaid statistilisi mudeleid ja vastavaid algoritme (näiteks GADA (Pique-Regi *et al.*, 2008) ja ITALICS (Rigaill *et al.*, 2008)), millest kõige enam on levinud peidetud Markovi mudelil põhinevad meetodid. Peidetud Markovi mudel (*HMM*) on statistiline meetod, mis modelleerib andmeid lähtuvalt Markovi protsessist, kus iga kindla väärtuse esinemise tõenäosus teatud ajahetkel sõltub ainult eelmistel ajahetkedel esinenud väärtustest. *HMM* eeldab, et iga vaadeldud andmepunkti väärtuse jaotus sõltub vaatlemata (peidetud) väärtusest, mis on modelleeritud eelkirjeldatud Markovi protsessiga. Kuna *CNV* määramisel kasutatakse tihti mitme järjestikuse markeri andmeid, on *HMM* sobiv lähedalasuvate markerite koopiaarvu sõltuvusstruktuuride modelleerimiseks (Wang *et al*., 2008; Zöllner *et al*., 2009).

Kui mitmed varasemad peidetud Markovi mudelil põhinevad algoritmid nagu dChip (Zhao *et al.*, 2004) ja CNAG (Nannya *et al.*, 2005) kasutasid *CNV*-de määramisel kiibiproovide summaarseid signaaliintensiivsusi, siis keerulisemad algoritmid lisavad mudelisse alleelispetsiifilised signaaliintensiivsused ning võimalusel ka muud saadaolevat infot. Colella *et al.* poolt 2007. a. avaldatud algoritm QuantiSNP kasutab *SNP* markerite summaarseid signaaliintensiivsusi koos *SNP* alleelide osaintensiivsustega ja võimaldab statistilise võimsuse suurendamiseks korraga analüüsida mitme mikrokiibi andmeid. QuantiSNP-s rakendatakse *HMM*-i edasiarendust, milles mudeli parameetreid hinnatakse *Objective Bayesi* paradigma abil (lisanduvate andmete valguses hinnatakse eelnevalt määratud parameetrid ümber).



Joonis 1. *Illustratsioon B-alleeli sageduste (a) ja Log R väärtuste (b) kohta uuritava indiviidi 15q kromosoomiõlas.* Normaalses kromosoomi regioonis on kolm B-alleeli sageduste klastrit, millele vastab kolm erinevat genotüüpi: AA, AB ja BB (c). *LOH* regioonil on normaalse regiooniga sama Log R väärtus (koopiaarv), kuid puudub AB genotüüp. Kõrgema koopiaarvuga regiooni näitavad nii suurenenud B-alleeli klastrite arv kui kasvav signaaliintensiivsus (LogR). Erinevad B-alleeli sageduste ja Log R väärtuste kombinatsioonid võimaldavad seega määrata nii koopiaarvu kui genotüüpi (Wang *et al*., 2007 järgi).

PennCNV (Wang *et al*., 2007) algoritm on sarnane QuantiSNP-ga, kuid koopiaarvu muutuste modelleerimisel on võimalik arvesse võtta ka perekonnaandmed (uuritavate indiviidide pärilikkussuhted). Täiendandmete kasutamine *HMM*-is suurendab nii PennCNV kui QuantiSNP puhul *CNV*-de määramise täpsust. Võrdlusena - QuantiSNP suudab korrektselt kaardistada kaks korda rohkem *CNV* regioonide piire kui Illumina enda tarkvara LOH+ (Colella *et al*., 2007).

Nii QuantiSNP kui PennCNV algoritmi puhul on *HMM*-i Markovi protsessis peidetud väärtusteks DNA koopiaarv igas uuritud *SNP* lookuses. Modelleermise aluseks on genotüpiseerimiskiipide andmed, mis koosnevad iga *SNP* puhul kahest komponendist - esiteks LogR väärtusest, mis näitab *SNP* proovi summaarset fluoressentssignaali intensiivsust ja teiseks B-alleeli sagedusest (*BAF*), mis näitab ühe *SNP* alleeli signaaliintensiivsuse osakaalu teise alleeli suhtes. PennCNV välja töötanud töörühm (Wang *et al.*) leidis, et LogR eksponentsiaalväärtus kasvab koos koopiaarvuga enam-vähem lineaarselt. Seetõttu on genotüpiseerimisandmetest lihtsam leida deletsiooni kui duplikatsiooni ning maksimaalseks määratavaks koopiaarvuks on 4 (sellest suuremad koopiaarvud ei ole 4-st eristatavad). Igale *SNP*-le vastava lookuse võimalikud koopiaarvud (0 kuni 4) ning genotüübid (ühele koopiaarvule võib vastata mitu erinevat genotüüpi - vt. joonis nr. 1) moodustavad *HMM*-is peidetud väärtuste kogu. Arvestades vaadeldavat väärtust (summaarset ja alleelispetsiifilist signaaliintensiivsust) ja erinevaid mudeli parameetreid (ja ka täiendandmeid, näiteks perekonnaandmeid), leiab *HMM* iga *SNP* puhul (peidetud väärtuste hulgast) vastava genoomse lookuse kõige tõenäolisema koopiaarvu ja sellele vastava genotüübi (Collela *et al.*, 2007; Wang *et al.*, 2007; Zöllner *et al*., 2009).

Genotüpiseerimiskiipide suurimaks eeliseks *CNV* uuringutes on hübridisatsioonisignaalide ja alleelse koosseisu paralleelne analüüs. Kuigi *HMM*-põhised algoritmid nagu QuantiSNP ja PennCNV määravad hea statistilise usaldusväärsusega nii koopiaarvu kui vastava genotüübi andmed, on tegemist diploidse raku summaarse genotüübiga (näiteks ABBB nelja koopia puhul). Kuna reaalselt on *CNV* jaotunud homoloogiliste kromosoomide vahel, on *CNV*-de bioloogilise tähtsuse (mõju fenotüübile) paremaks mõistmiseks oluline teada kromosoomispetsiifilist koopiaarvu. Teades lisaks uuritava indiviidi *CNV* andmetele ka genoomi alleelset koosseisu ja perekonnainfot, peaks olema võimalik arendada algoritm, mis teatud statistilise tõepäraga võimaldaks tuletada kromosoomispetsiifilisi genotüüpe (näiteks ABB ja B, summaarselt neli koopiat) (Wang *et al.*, 2008; Yau & Holmes, 2008).

2 Töö eesmärk

Käesoleva magistritöö praktilise osa eesmärgiks oli:

1. Välja töötada ja valideerida arvutuslik metoodika genotüpiseerimiskiipide andmetest *CNV*-de määramiseks ning *CNV-de* pärandumismustrite ja nende sageduste leidmiseks.
2. Analüüsida ja võrrelda *CNV*-de pärandumist kahe valimi perekonnaandmete põhjal, kasutades antud töö raames arendatud arvutuslikke meetodeid.

3 Materjal ja metoodika

3.1 Algandmed

CNV-de määramisel oli algandmeteks HapMap projekti raames analüüsitud kuuekümne isa-ema-laps trio genotüpiseerimisandmed. Kokku 180-st HapMap indiviidist koosnevas grupis on võrdselt esindatud kaks populatsiooni: pooled indiviidid pärinevad Nigeeria populatsioonist Aafrikas (YRI), ülejäänud on Euroopa juurtega Põhja-Ameerika asurkonnast (CEU). Affymetrix 500K EA *SNP* mikrokiibiga analüüsitud DNA-de genotüpiseerimisandmed laeti tekstifailidena alla HapMap konsortsiumi kodulehelt (http://www.hapmap.org).

Lisaks HapMap andmetele kasutasime Eesti vaimse arengu mahajäämuse uuringu kohordi genotüpiseerimisandmeid, mida kogub ja analüüsib prof. Ants Kure juhitud töögrupp (biotehnoloogia õppetool, TÜ MRI). VAM kohordi DNA-d on genotüpiseeritud Illumina HumanCNV370 Duo *SNP* mikrokiipidega.

Lähteandmetena saadi mõlemast allikast (HapMap'i koduleht ja prof. Ants Kure töörühm) mikrokiibi katsete toorandmed tekstifailide kujul. Failid sisaldasid infot analüüsitud SNP markerite fluoressetnssignaali intensiivsuse, genoomse asukoha ja detekteeritud alleelide kohta. Lisaks oli iga uuritud indiviidi kohta teada tema sugu ja pärilikkussuhted teiste indiviididega.

In silico leitud *CNV*-de valideerimisel kasutati *RT-qPCR* meetodil määratud *CNV*-de andmeid, mis saadi prof. Ants Kure töörühmalt.

3.2 Kasutatud programmid ja nende tööpõhimõte

3.2.1 Arvutuslik meetod *CNV*-de määramiseks

CNV-de määramisel oli esimeseks sammuks lähteandmete analüüs (genotüpiseerimine ja esmane *CNV*-de detekteerimine), milleks kasutati teiste autorite poolt välja töötatud programme QuantiSNP (Colella *et al.*, 2007) ja PennCNV (Wang *et al.*, 2007), mis on ainsad Illumina genotüpiseerimisandmete analüüsiks arendatud HMM-põhised algoritmid (kuigi saab kasutada ka Affymtrixi kiipide andmeid). Mõlemad programmid kasutavad iga *SNP* markeri koopiaarvu ja genotüübi arvutamiseks lähteandmetest kahte väärtust: vastava proovi normaliseeritud summaarset signaaliintensiivsust (*log R Ratio, LRR*) ja minoorse alleeli normaliseeritud signaaliintensiivsuse osakaalu teise alleeli suhtes (*B Allele Frequency, BAF*). Nii *LRR* kui *BAF* on iseloomulikud Illumina genotüpiseerimisandmetele, mis arvutatakse mikrokiibikatse skanneerimispildist Illumina BeadStudio tarkvara abil (Illumina, San Diego, USA). Affymetrixi mikrokiipide genotüpiseerimisandmed sisaldavad iga proovi kohta normaliseerimata summaarset signaaliintensiivsust ja standardset kahealleelset genotüüpi (näit. T/G). Et Affymetrix'i andmeid saaks kasutada QuantiSNP ja PennCNV sisendina, konverteeriti Affymetrix'i-spetsiifilised signaaliintensiivsused *LRR* ja *BAF* väärtusteks, kasutades Affymetrix Powertools nimelist tarkvara (Affymetrix, Santa Clara, USA). QuantiSNP ja PennCNV poolt tehtud arvutuste tulemused kirjutas kumbki programm oma väljundfaili, mis sisaldas iga määratud *CNV* regiooni kohta tulpadesse jaotatult järgmisi andmeid: analüüsitud indiviidi ID (unikaalne tähemärkide kombinatsioon), *CNV* regiooni genoomne asukoht (kujul kromosoom, alguspositsioon, lõpupositsioon), CNV regiooni pikkus (aluspaarides), DNA koopiaarv, *CNV* hinnangu statistiline tõepära (Log Bayes'i faktor, näitab programmi kindlust antud *CNV* määramisel).

Antud töös *CNV*-de määramise metoodika tarbeks välja töötatud programmid on kirjutatud programmeerimiskeeles Perl (*Practical exrtraction and reporting language*, http://www.perl.org). Programm *find_bulletproof_CNVs.pl* kasutas sisendandmetena QuantiSNP ja PennCNV poolt genereeritud tekstifaile ja ühendas kahe programmi poolt leitud koopiaarvu piirkonnad ühtseks nimekirjaks, jättes väljundfaili alles vaid need *CNV*-d, mida detekteerisid nii QuantiSNP kui PennCNV. Programm *find_bulletproof_CNVs.pl* käivitati UNIX operatsioonisüsteemi käsurealt järgmiselt (sisendparameetrid nurksulgudes):

```
find_bulletproof_CNVs.pl <PennCNV väljundfail> <QuantiSNP
    väljundfail> <populatsioon> <praakide nimekiri>
```

Vastavalt käsurea süntaksile kasutas antud programm lisaks QuantiSNP ja PennCNV väljundfailide nimedele veel kahte lisaparameetrit: uuritava populatsiooni nime ja praagitavate indiviidide nimekirja failinime. Käivitatud programmi töö käik oli järgnev: kõigepealt avas programm sisendparameetrina saadud PennCNV faili, mida hakati rea kaupa lugema. Iga rea puhul loeti esimesest tulbast uuritava indiviidi ID ja võrreldi seda praagitavate indiviidide nimekirjaga (sisaldas halvakvaliteediliste katsetulemuste tõttu analüüsist välja jäetavaid indiviide). Juhul, kui antud indiviidi ID-d ei olnud praagitavate nimekirjas, võrreldi järgmise sammuna käesolevat *CNV* regiooni (PennCNV failist) sama indiviidi *CNV* regioonidega QuantiSNP failis. Kui kahest erinevast failist pärinevad (ehk erinevate programmidega määratud) *CNV* regioonid kattusid, määrati kattumisalale vastav regioon uueks *CNV* regiooniks ning vastavad algus- ja lõpupositsioonid kirjutati koos ülejäänud andmetega (indiviidi ID, *CNV* koopiaarv, Log Bayes'i faktor) väljundfaili. Lisaks pandi väljundfaili kirja ka uuritava indiviidi sugu, mis loeti uuritava populatsiooni nimele vastavast tekstifailist. Kirjeldatud tööprotsessi tulemusel olid programmi *find_bulletproof_CNVs.pl* väljundfailis kirjas *CNV*-d mis on saadud QuantiSNP ja PennCNV poolt määratud *CNV*-de kattuvusalast (indiviidide kaupa). Seega on uueks CNV-ks määratud vaid need regioonid, mille CNV-ks määramisega nõustuvad mõlemad programmid (nii QuantiSNP kui PennCNV). Sarnaselt *find_bulletproof_CNVs.pl* sisendfailidele on ka programmi väljundfailis iga *CNV* piirkonna andmed kirjutatud eraldi reale.

Määratud CNV-de nimekirja lõplikule kujule viimiseks järjestati *CNV*-de nimekiri indiviidide kaupa kõigepealt kromosoomi ja seejärel *CNV* alguspositsiooni järgi ning viimase sammuna praagiti järjestatud *CNV*-de nimekirjast välja halva kvaliteediga *CNV*-d. Halva kvaliteediga *CNV*-deks loeti sellised regioonid, mille pikkus oli alla tuhande aluspaari ja mille Log Bayes'i faktor (LBF, programmi QuantiSNP poolt arvutatud CNV usaldusväärtuse skoor) oli alla viie. CNV-de filtreerimiseks kirjutati programm *filter_CNVs.pl*, mis käivitati UNIX'i käsurealt koos filtreerimise parameetritega järgneval kujul :

```
filter_CNVs.pl <find_bulletproof_CNVs.pl väljundfail>
      <minimaalne CNV pikkus> <minimaalne LBF>
```

Pärast filtreerimist on lõpptulemuseks tekstifail, mis sisaldab genoomse positsiooni järgi sorteeritud *CNV*-sid.

3.2.2 Arvutusliku meetodi valideerimine

Eelnimetatud programmide abil määratud *CNV*-de valideerimiseks kirjutati programm *confirm_CNVs.pl*, mis võrdles *in silico* määratud *CNV* piirkondi *RT-qPCR* katsete tulemustega. Programm kirjutas väljundfaili genoomi positsioonid, kus *RT-qPCR* katsete praimerid kattusid *in silico* leitud *CNV*-dega. Lisaks analüüsis *confirm_CNVs.pl* kahe erineva meetodiga leitud *CNV*-de võrdluse tulemusi erinevate kirjeldavate statistikute abil (kokkulangevate regioonide pikkuste keskväärtus, mediaan, standardhälve jne.). Programm *confirm_CNVs.pl* sai sisendandmetena arvutuslikult saadud *CNV*-de nimekirja ning *RT-qPCR*-i katsete tulemused:

```
confirm_CNVs.pl <in silico CNV-de nimekiri> <RT-qPCR tulemused>
```

Andmete võrdluse tulemused kirjutas programm tekstifaili, mille alusel sai hinnata käesolevas töös kirjeldatud *in silico CNV*-de määramise meetodi efektiivsust.

3.2.3 *CNV*-de pärandumise analüüs

Pärnduvusanalüüsides kasutati algandmetena eelkirjeldatud ühendmeetodiga leitud *CNV*-sid. Edasiseks analüüsiks klasterdas vastav programm (*find_familyCNVs.cpp*) iga perekonna *CNV*-d ühendi leidmise teel (ühe perekonna indiviide kattuvad *CNV* lookused liideti üheks pikemaks regiooniks, et võrdsustada perekonna liikmete *CNV* piirid (vt. joonis 4)). Kõigi perekondade *CNV*-d (nii ühendmeetodiga leitud *CNV*-d kui ka üksikute programmidega leitud *CNV*-d) koondati ühte faili, mida analüüsiti päranduvusmustrite leidmise programmiga *CNV_heritage.pl:*

```
CNV_heritage.pl <perekondlike CNV-de tabel>
```

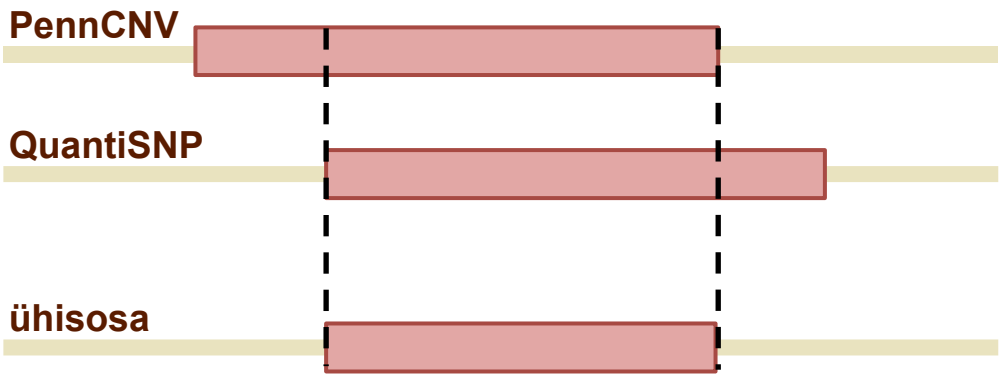
Programm otsis sisendfailist üles kõik CNV-de päranduvussündmused (lookused, kus vähemalt ühel trio liikmel on *CNV*), kusjuures arvestati vaid selliseid lookuseid, mille koopiaarv oli kinnitatud mõlema *CNV* leidmise programmiga (nii *CNV* kui normaalse koopiaarvuga regioonide puhul). Seejärel jagas programm lookused päranduvuse järgi tüüpidesse ning arvutas vastavad esinemissagedused. Tulemuseks on analüüsitud perekonna *CNV*-de päranduvust iseloomustav fail (näitena vt. lisa 1).

4 Tulemused

Antud magistritöö raames töötati välja arvutuslik meetod genotüpiseerimisandmetest *CNV*-de määramiseks. Meetod võimaldab nii Affymetrixi kui Illumina genotüpiseerimiskiipide andmete analüüsimist ja valepositiivsete tulemuste vähendamiseks kasutatakse kahe erineva *HMM*-põhise *CNV*-de määramise programmi arvutusi.

4.1 CNV-de määramise meetodi väljatöötamine

In silico CNV-de määramisel on tulemuste hulgas alati teatav hulk valepositiivseid tulemusi, mis vähendab leitud *CNV*-de usaldusväärsust (Perry *et al.*, 2008). Et vähendada valepositiivsete tulemuste hulka, arendati välja algoritm, mis rakendas samadele andmetele kahte erinevat *CNV*-de leidmise programmi (QuantiSNP ja PennCNV). Mõlema programmi tulemustest sõeluti seejärel välja ebakvaliteetsed (lühikesed ja väikese usaldusväärusega) *CNV*-d ning saadud lookused pandi kokku ühisosa (∩) leidmise kaudu, mida illustreerib joonis nr. 2. Välja töötatud algoritm leidis kõik kahe programmi kattuvad *CNV*-d, mille ühisosa piirid määrasid uue *CNV*. Mittekattuvad *CNV*-d jäeti edasisest analüüsist välja. Seega kasutab algoritm konservatiivset lähenemist, kus *CNV*-ks loetakse vaid need lookused, mille koopiaarvu muutust kinnitavad mõlemad kaasatud programmid.



Joonis 2. *Kahe programmi poolt arvutuslikul teel leitud CNV-de ühendamine.* Kahe programmi poolt leitud osaliselt kattuvate *CNV*-de ühine osa määrab uue *CNV* piirid. *CNV*-d on märgitud punaste ristkülikutena ja normaalse koopiaarvuga genoomne piirkond heleda triibuna. *CNV*-de ühisosa on piiritletud punktiirjoonega.

4.2 Väljatöötatud metoodikaga leitud *CNV*-de valideerimine

Et hinnata väljatöötatud meetodi efektiivsust *CNV*-de määramisel, võrreldi arvutuslikul teel leitud *CNV*-sid sama andmestiku peal läbi viidud reaalaja kvantitatiivse *PCR*-i katsete tulemustega.

Analüüside algandmed pärinevad Eesti VAM kohordist, kellest 46 indiviidi puhul leiti *RT-qPCR* meetodil kokku 75 lookuse koopiaarv (1-3 lookust indiviidi kohta). *RT-qPCR* katsetes kasutati iga lookuse kohta 1 kuni 10 praimerit, mille alguspositsioonide järgi määrati lookuste pikkuseks 1 bp kuni 6,2 Mb. Suurem osa lookustest olid koopiaarvu muutuseta regioonid (55 lookust 75-st) ning ülejäänud 20 *CNV* hulgas oli rohkem deletsioone (13 deletsiooni, 7 duplikatsiooni).

Reaalaja *qPCR* analüüsides kasutatud 46 indiviidile leiti *CNV*-d ka käesolevas töös välja töötatud arvutusliku meetodiga. Kuna antud meetod põhineb kahe erineva *CNV*-de leidmise programmi (QuantiSNP ja PennCNV) arvutustulemuste kombineerimisel, leiti *CNV*-d ka kummagi programmiga eraldi ning kõiki kolme arvutatud *CNV*-de kogu võrreldi *RT-qPCR* referentsanalüüsi tulemustega. Võrdluse eesmärgiks oli valideerida väljatöötatud arvutusliku meetodiga leitud *CNV*-sid ja hinnata meetodi efektiivsust *CNV*-de leidmisel võrreldes QuantiSNP ja PennCNV-ga. Selleks hinnati kolme arvutusliku meetodi täpsust *CNV*-de leidmisel võrreldes *RT-qPCR*-iga - kui palju leidsid arvutuslikud meetodid valepositiivseid tulemusi (*RT-qPCR*-iga leitud koopiaarvu neutraalne regioon hinnati *CNV*-ks), valenegatiivseid tulemusi (*CNV* regioon jäi leidmata) ja kui palju erinevad valideeritud *CNV*-de asukoht (CNV otspunktide positsioonid) ja koopiaarv *RT-qPCR*-iga leitud *CNV*-dest.

Kolme arvutusliku meetodi võrdluse tulemused on kokkuvõtvalt toodud tabelis nr.1. Kui valepositiivsete tulemuste osakaal oli kõigi kolme arvutusliku meetodi puhul sama siis valenegatiivseid tulemusi leidus ühendmeetodi puhul teistest meetoditest rohkem. Seega on väljatöötatud meetod suhteliselt konservatiivsem - koopiaarvu neutraalseks hinnati piirkond, mida kumbki programm eraldi (ja *RT-qPCR*) hindas *CNV*-ks. Õigesti määratud *CNV*-de hulgast (programmide puhul kõik kõik 20 RT-qPCR-iga kinnitatud CNV-d) leidus mõlema programmi puhul üks *CNV*, mille koopiaarv ei langenud referentsanalüüsi andmetega kokku. Kuna väljatöötatud arvutusmeetodi puhul ei olnud eraldi protseduuri kahe programmi poolt leitud koopiaarvu kombineerimiseks (konsensuse leidmiseks) siis koopiaarvu leidmise täpsust ühendmeetodi puhul ei hinnatud.

Tabel 1. *Arvutuslikult leitud CNV-de valideerimise tulemused.* Tabelis on toodud kolme arvutusliku meetodi poolt leitud *CNV*-de erinevus *RT-qPCR* meetodil leitud *CNV*-dest (kasutades samu algandmeid). Valepositiivsete tulemuste protsent on osakaal kõigist referentsanalüüsi neutraalsetest lookustest. Valenegatiivsete ja valesti hinnatud koopiaarvuga *CNV*-de protsentarv on osakaal kõigist valideeritud *CNV* regioonidest. Arvutuslikult leitud *CNV*-de algus- ja lõpp-positsioonide keskmised erinevused referents-*CNV*-dest on antud tuhandetes aluspaarides.

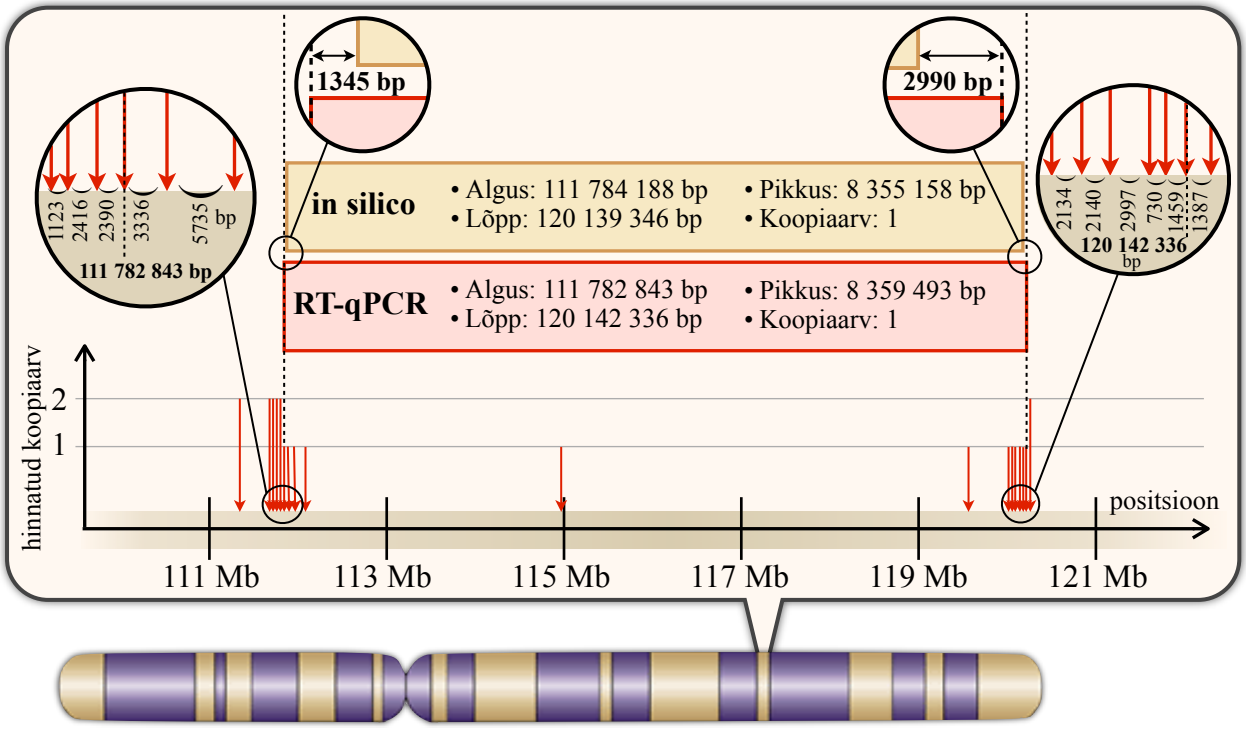
	QuantiSNP	PennCNV	PennCNV ∩ QuantiSNP
valepositiivsed	5% (3/55)	5% (3/55)	5% (3/55)
valenegatiivsed	0	0	4% (1/20)
CNV koopiaarvu hinnanguviga	4% (1/20)	4% (1/20)	-
CNV alguspos. väiksem	866 kb	560 kb	568 kb
CNV lõpp-pos. suurem	255 kb	114 kb	137 kb
CNV alguspos. kaugus	886 kb	572 kb	591 kb
CNV lõpp-pos. kaugus	257 kb	189 kb	141 kb

Kõik arvutuslikud meetodid hindasid *CNV* pikkuse enamasti suuremaks kui *RT-qPCR*. Võrreldes *RT-qPCR*-iga leitud *CNV* piiridega hindas QuantiSNP *CNV* alguspositsiooni keskmiselt väiksemaks ja lõpp-positsiooni suuremaks kui PennCNV. Ühendmeetodi tulemused olid kahe programmi tulemustega võrreldes vahepealsed. Erinevused *CNV*-de algus- ja lõpppositsioonide hindamisel näitavad ka arvutuslikult leitud *CNV*-de asukoha nihutatuse tendentsi referentsanalüüsi *CNV*-de suhtes - kõigi kolme arvutusliku metoodika puhul olid *CNV*-d küll pikemad kui *RT-qPCR*-iga leitud *CNV*-d kuid mitte võrdselt mõlemast *CNV* otspunktist vaid keskmiselt enam nihutatud referents-*CNV* suhtes genoomse 5' otsa poole.

CNV otspunktide hindamise täpsust näitab arvutatud *CNV*-de otspunktide keskmine kaugus vastavate referents-*CNV*-de algus- ja lõpp-positsioonidest. *CNV* piiride täpsuse võrdlemine demonstreerib väljatöötatud konservatiivse arvutusmeetodi eelist - kui ühendmeetodiga arvutatud *CNV*-de alguspositsiooni keskmine kaugus on veidi suurem kui PennCNV tulemus, siis *CNV* lõpp-positsiooni hindamisel on kahe programmi arvutusi ühendav meetod mõlemast programmist täpsem. Arvutuslike meetodite täpsus *CNV* piiride hindamisel sõltub ka *CNV* suurusest - pikemate arvutatud *CNV*-de otspunktid on referents-*CNV* piiridele lähemal (*CNV* kogupikkuse suhtes) kui lühikeste *CNV*-de puhul. *CNV* otspunktide hinnangu täpsusest pikemate *CNV*-de puhul annab aimu joonis nr. 4, kus on võrdlevalt kujutatud *in silico* (kahe programmi ühendmeetodiga) ja katseliselt (*RT-qPCR* meetodiga) leitud *CNV* lookust ühe VAM kohordi indiviidi seitsmenda kromosoomi pikas õlas. Jooniselt kujutatud valideeritud *CNV* pikkusega võrreldes (8,4 Mb) on arvutuslikult leitud *CNV* piirid suhteliselt lähedal *RT-qPCR*-iga leitud *CNV*-le (alla 3 kb), samas on *CNV* lõpp-positsiooni hinnanguviga umbes kaks korda suurem kui algpositsioonide erinevus. Joonisele on kantud ka *RT-qPCR*-i analüüsis kasutatud praimerite positsioonid ja vastavad koopiaarvud. Et *RT-qPCR* meetodiga oleksid *CNV*-de piirid võimalikult täpselt määratud, on suurem osa praimereid paigutatud tihedalt analüüsitava *CNV* oletatavate piiride lähedusse.

Lisaks meetodite efektiivsuse võrdlemisele vaadati detailsemalt arvutuslikult leitud ja valideeritud *CNV* regioone, et leida kui palju erinevad üksteisest kolme arvutusliku meetodiga leitud *CNV*-d. Kahekümnest *RT-qPCR*-iga kinnitatud *CNV*-st pooltel juhtudel katsid mõlema kasutatud programmi poolt arvutatud *CNV*-d samu *RT-qPCR* analüüsi praimereid, mistõttu oli programmide tulemused identsed ühisosa rakendava arvutusmeetodiga. Ülejäänud kümne *CNV* puhul kattusid kahe programmi poolt leitud *CNV* regioonid vaid osaliselt ja vastavalt algoritmile kasutas kolmas e. väljatöötatud arvutuslik meetod CNV määramisel kattuvate regioonide ühisosa. Ühes valideeritud *CNV* lookuses ei olnud PennCNV ja QuantiSNP poolt arvutatud *CNV*-de vahel ülekattuvust ja seetõttu oli selles lookuses ühendmeetodil valenegatiivne tulemus.

Kolme arvutusliku meetodi valideerimis- ja võrdlusanalüüsi tulemused näitavad, et väljatöötatud kahe programmi ühisosal põhinev arvutuslik meetod täidab oma peamist eesmärki, olles *CNV*-de ennustamisel pigem konservatiivne. Teoreetiliselt vähendab see valepositiivsete tulemuste hulka. Kuigi see meetodi eelis on saavutatud võimaliku valenegatiivsete tulemuste hulga suurenemise arvelt, näidati ühisosa meetodi paremat täpsust *CNV* piiride hindamisel.



Joonis 3. *Arvutuslikul teel leitud CNV võrdlus katseliselt valideeritud CNV-ga.* Joonisel on toodud näide valideeritud *CNV*-st ühe VAM kohordi indiviidi seitsmenda kromosoomi pikas õlas. Genoomne skaala on antud miljonites aluspaarides, ülejäänud positsioonid ja vahemaad aluspaarides. Punaste noolte asukoht märgib *RT-qPCR* katsetes kasutatud praimerite asukohta genoomsel skaalal, noole pikkus näitab antud praimeri hinnatud koopiaarvu (1 või 2). Ringidega on esile toodud detailsem vaade *CNV* piiridest. Punktiirjoon märgib valideeritud *CNV* otspunkte tähistavate praimerite asukohta. Ringi sees on noolte all toodud praimeritevahelised kaugused, *CNV* otspunkti tähistava praimeri positsioon on jämedas kirjas.

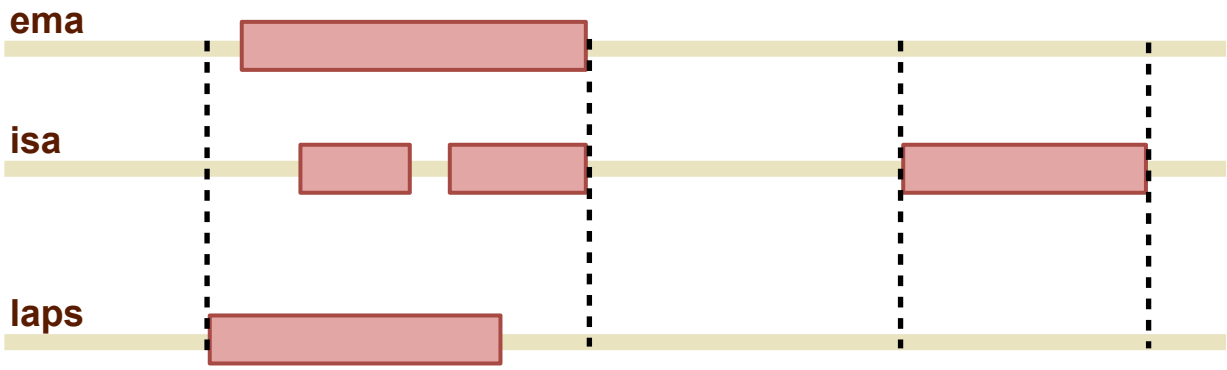
4.3 *CNV*-de päranduvuse uurimine

Kahe programmi tulemusi ühendav algoritm võimaldas *CNV* piirkondi määrata täpsemalt kui üksikprogrammid, andes kvaliteetsemat algmaterjali edasisteks uuringuteks. Seetõttu olid sel meetodil leitud *CNV*-d aluseks järgnevatele analüüsidele, et uurida koopiaarvult varieeruvate lookuste päranduvust. Eelkõige pöörasime tähelepanu *de-novo CNV*-sid iseloomustavatele päranduvusskeemidele ja selle varieerumisele Hapmapi ja Eesti vaimse arengu mahajäämuse uuringu kohordi indiviidide näitel.

4.3.1 *CNV*-d HapMap triodes

Antud töö üks eesmärkidest oli analüüsida *CNV*-de pärandumist ühe põlvkonna piires, et vaadata detailsemalt erinevat tüüpi *CNV*-de ülekannet vanematelt lapsele ning seda eriti *de-novo CNV*-de puhul. Kuna *de-novo CNV*-d tunduvad olevat lapse genoomi tekkinud ilma mendeliaalse pärandumiseta (antud lookuses ei ole kummagil vanemal koopiaarvu muutust) võib selliste *CNV*-de perekonnapõhine uurimine aidata paremini mõista uute *CNV*-de tekkimise seaduspärasid.

Päranduvusanalüüsides alustati *CNV*-de määramisest eelpool kirjeldatud kahe programmi tulemusi ühendava algoritmiga. Et edasises analüüsis oleksid kõigi perekonnaliikmete *CNV* lookused omavahel võrreldavad, klasterdasime perekonnasiseselt *CNV* piirid ühendi leidmise teel. Kui ühisosa algoritm viskab kahe programmi andmeid ühendades ära kõik mittekattuva siis ühendi meetodil liidetakse kõik ühe perekonna indiviide kattuvad *CNV* lookused üheks pikemaks regiooniks (vt. joonis 4). Tulemuseks on diskreetsed lookused, mis on ühe perekonna piires kõigil indiviididel sama pikkuse ja asukohaga (osaliselt kattuvad regioonid loetakse üheks lookuseks). Sel viisil on kõik *CNV* lookused eraldatud isa + ema → laps päranduvussündmusteks, mida saab jälgida ja analüüsida.



Joonis 4. *Ühe perekonna liikmete CNV lookuste ühendamine.* Osaliselt kattuvad *CNV* piirkonnad liidetakse üheks suuremaks regiooniks (võetakse ühend), mis seab uued CNV piirid kõigil pereliikmeil võrdseks. Eraldiseisvaid *CNV*-sid ei ignoreerita vaid need moodustavad omaette lookuse. *CNV*-d on märgitud punaste ristkülikutena ja normaalse koopiaarvuga genoomne piirkond heledama triibuna. *CNV*-de ühend on piiritletud punktiirjoonega.

Päranduvuse uurimist alustati laialdaselt referentsandmetena kasutatavast (Wang *et al.*, 2009; McCarroll *et al.*, 2008) HapMap populatsiooni indiviididest. Selleks arvutati genotüpiseerimisandmetest kuuekümne HapMap ema-isa-laps trio jaoks *CNV* piirkonnad, kasutades programmitulemuste ühisosa algoritmi. Seejärel klasterdati eelkirjeldatud ühendimeetodil iga perekonna (trio) *CNV*-d diskreetseteks lookusteks. Lõpuks pandi kõik andmed *CNV*-d ühte tabelisse ja loeti kokku kõik vanemad-laps *CNV*-de pärandumised ning arvutati kõigi erinevate *CNV* pärandumistüüpide leidumissagedused. Siinjuures on oluline märkida, et *CNV* olemasolu arvestati vaid siis, kui see oli leitud käesolevas töös väljatöötatud ühendmeetodiga (*CNV* leiti indiviidil samas regioonis nii QuantiSNP-ga kui ka PennCNV-ga) ja *CNV* puudumist arvestati vaid siis, kui vastavas regioonis polnud leitud koopiaarvu muutust kummagi programmiga. Juhud kus mõnel perekonnaliikmel oli uuritavas regioonis leitud CNV ainult ühe programmiga (QuantiSNP-ga või PennCNV-ga) jäeti pärandumise analüüsist kõrvale.

HapMap populatsiooni triode *CNV*-de pärandumist ja sagedusi kajastavad tulemused on kokkuvõtvalt toodud tabelis nr. 2. Tabelis on kirjas erinevad *CNV*-de päranduvuse võimalused ja nende esinemissagedused, kusjuures number 1 iga päranduvustüübi juures tähistab *CNV* olemasolu vastaval trio liikmel ning 0 selle puudumist. Seega märgib päranduvustüüp “0 0 1” (00→1 ehk lookus, kus isal on 0, emal 0 ja lapsel 1 *CNV*) *de-novo CNV*-d, mille esinemissagedus HapMap triode puhul on 7% kõigist analüüsitud pärandumissündmustest. Lisaks on iga pärandumistüübi juures jagatud lapse *CNV*-d vastavalt koopiaarvule kas deletsiooniks või duplikatsiooniks. Kõikide

HapMap triode päranduvussündmuste puhul, kus lapsel esineb antud lookuses *CNV*, on erinevus deletsioonide ja duplikatsioonide esinemissageduste vahel ligilähedaselt kahekordne. Üldiselt esineb deletsioone rohkem kui duplikatsioone (vastavalt 62% ja 38% laste *CNV*-dest), välja arvatud juhul kus ühes lookuses on kõigil trio liikmetel koopiaarvu muutus. Levinuim *CNV* pärandumismuster HapMap triode puhul on 0/1→0 (ühel vanematest on *CNV*, mis ei pärandu edasi lapsele; sagedus 56% kõigist juhtudest) ja kõige harvem esineb sündmus, kus ühes lookuses on mõlemal vanemal *CNV*, samas kui lapsel seal koopiaarvu muutust ei ole (muster 11→0; 1% kõigist pärandumissündmustest).

Tabel 2. *CNV pärandumistüüpide jaotus HapMap triodes.* Ridades on toodud *CNV*-de päranduvuse võimalused koos nende esinemissagedusega HapMap triode populatsioonis. Number 1 tähistab *CNV* olemasolu vastaval trio liikmel ning 0 selle puudumist, 0/1 tähistab juhtu kus *CNV* esineb ühel kahest vanemast. Iga pärandumistüübi juures on lastel esinenud *CNV*-d jaotatud duplikatsioonideks ja deletsioonideks. Numbrid sulgudes näitavad vastava lookuse absoluutset esinemist, protsendid aga suhtarvu, mis duplikatsioonide ja deletsioonide puhul on esinemissagedus laste *CNV*-de koguarvu suhtes, "kokku" tulbas aga kõikide pärandumissünduste hulga suhtes.

isa	ema	laps	deletsioone	duplikatsioone	kokku
0	0	1	12% (80)	5% (34)	7% (114)
0/1		0	-	-	56% (861)
0/1		1	47% (314)	28% (188)	32% (502)
1	1	0	-	-	1% (21)
1	1	1	3% (17)	5% (36)	4% (53)
Kokku			1	0	100% (1551)

4.3.2 *CNV*-d Eesti VAM uuringu perekondades

Vaimse arengu mahajäämust (VAM) on seostatud mitmete geneetiliste muutustega inimgenoomis (Chelly & Mandel, 2001). Et võrrelda *CNV*-de pärandumist ja *de-novo CNV*-de esinemissagedust normaalse kontroll-populatsiooni (HapMap) ja VAM indiviidide vahel, viisime läbi *CNV*-de päranduvusanalüüsi Eesti VAM kohordi perekondade andmetes.

VAM-iga indiviidide *CNV*-de päranduvuse uurimiseks viisime läbi samad arvutused, mis HapMap triode puhul, kasutades algandmetena Eesti VAM kohordi indiviidide genotüpiseerimistulemusi ja perekonnaandmeid. Kokku analüüsiti 79 VAM kohordi indiviidi 18 perekonnast, mis olid jaotatud 43 isa-ema-laps trioks. Analüüside tulemused on kokkuvõtvalt toodud tabelis nr. 3.

Kui kõrvutada VAM kohordi analüüsitulemusi HapMap triode andmetega, hakkab silma rida erinevusi. Kui HapMap järglaspõlvkonna *CNV*-de seas oli deletsioonide ülekaal duplikatsioonide suhtes veidi vähem kui kahekordne siis VAM-ide puhul on vahe enam kui neljakordne (suhe 4,5:1). Deletsioonide osakaal on eriti suur 11→1 tüüpi päranduvuste puhul ehk kui mõlemal vanemal on antud lookuses *CNV*, siis pea alati pärandub edasi just deletsioon, kuigi selliste *CNV*-dega vanematel on duplikatsioone pea kaks korda rohkem kui deletsioone. Samuti on VAM perekonna lastel *de-novo CNV*-de esinemissagedus poolteist korda (11/7 = 1,57 korda) suurem kui HapMap triode puhul ning vahe tuleneb jällegi deletsioonide arvelt.

Muus osas on kahe võrdlusaluse populatsiooni *CNV*-de pärandumine sarnane - nii HapMap kui VAM perekodnade puhul on levinuim *CNV*-de pärandumisskeem 0/1→0 (ühel vanematest on *CNV*, mis edasi ei pärandu), kuigi VAM kohordis on vahe populaarsuselt järgmise pärandumismustriga (0/1→1) tunduvalt suurem kui HapMap indiviididel (vastavalt HapMap 56% ja 32% VAM 64% ja 17% vastu). Samuti on mõlemas populatsioonis kõige haruldasem pärandumismuster 11→0 (mõlemal vanemal *CNV*, kuid kumbki edasi ei pärandu), kuigi VAM indiviididel esineb sellist pärandumist veidi rohkem (VAM 3% HapMap 1% vastu).

Tabel 3. *CNV pärandumistüüpide jaotus Eesti VAM perekondades.* Ridades on toodud *CNV*-de päranduvuse võimalused koos nende esinemissagedusega HapMap triode populatsioonis. Number 1 tähistab *CNV* olemasolu vastaval trio liikmel ning 0 selle puudumist, 0/1 tähistab juhtu kus *CNV* esineb ühel kahest vanemast. Iga pärandumistüübi juures on lastel esinenud *CNV*-d jaotatud duplikatsioonideks ja deletsioonideks. Numbrid sulgudes näitavad vastava lookuse absoluutset esinemist, protsendid aga suhtarvu, mis duplikatsioonide ja deletsioonide puhul on esinemissagedus laste *CNV*-de koguarvu suhtes, "kokku" tulbas aga kõikide pärandumissünduste suhtes.

isa	ema	laps	deletsioone	duplikatsioone	kokku
0	0	1	27% (74)	5% (14)	11% (88)
0 / 1		0	-	-	64% (521)
0 / 1		1	40% (107)	13% (35)	17% (142)
1	1	0	-	-	3% (23)
1	1	1	15% (40)	~0% (1)	5% (41)
Kokku			1	0	100% (914)

Arutelu ja järeldused

Antud töö käigu arendati välja arvutuslik meetod genotüpiseerimisandmetest *CNV*-de leidmiseks. Selle meetodi valideerimiseks ja efektiivsuse hindamiseks võrreldi *in silico* leitud *CNV*-sid katseliselt määratud *CNV*-dega, võttes referentsiks *RT-qPCR* meetodil leitud *CNV*-d. Kuigi arvutuslike meetodite valideerimisel eeldatakse, et kvantitatiivse *PCR*-i katsetulemused on arvutatud *CNV*-dest usaldusväärsemad (MacConaill *et al.*, 2007), ei pruugi see alati nii olla. Ka *RT-qPCR* katsetulemuste analüüsimiseks kasutatakse arvuteid ja spetsiaaltarkvara, mis võivad anda valepositiivseid ja valenegatiivseid tulemusi sarnaselt *in silico* meetodites kasutatud progammidele. Seepärast võib kahelda antud töö valideerimise osas ühe katseliselt leitud *CNV* koopiaarvus, mis erines nii QuantiSNP kui PennCNV poolt hinnatud koopiaarvust ning samamoodi tasub üle vaadata lookused, kuhu mõlemad programmid leidsid suure tõepäraga *CNV* kuid mida ei kinnitanud kvantitatiivse *PCR*-i katsed.

Väljatöötatud *in silico* meetodi puuduseks võib pidada seda, et algoritmis ei ole rakendatud viisi, kuidas ühendada kahe programmi poolt *CNV* lookusele arvutatud koopiaarvu hinnangud. Kuna meie analüüsides ei olnud PennCNV ja QuantiSNP koopiaarvu hinnangud piisavalt erinevad, et valideerimsel oleks referentsandmetega olulisi vasturääkivusi tekkinud, siis määras algoritm kahe programmi *CNV* lookusele lihtsalt PennCNV vastava koopiaarvu hinnangu. Samas on alati võimalus, et kaks programmi leiavad samale *CNV*-le oluliselt erinevaid koopiaarve (näiteks kui PennCNV hindab *CNV*-d deletsiooniks ja QuantiSNP duplikatsiooniks) ja seni kuni puudub alus ühe programmi hinnangut teise omale eelistada või mehhanism kompromisshinnangu arvutamiseks, ei ole koopiaarvu hindamise osas antud ühendamismeetodil eelist üksikprogrammi ees.

Kahe programmi tulemusi ühendaval meetodil on nõrgaks kohaks ka komplekssete *CNV*-de määramine, kus ühe *CNV* lookuse sees on mitu erineva koopiaarvuga lõiku. Probleemi illustreeriva näite võib tuua antud töö valideerimisanalüüsist, kus RT-qPCR katsetes oli üks osa katkematu *CNV* lookuse praimeritest suurema hinnatud koopiaarvuga kui sama *CNV* ülejäänud praimerid, samas kui arvutuslik meetod leidis samale regioonile vaid ühe koopiaarvu. Kuid nagu eelpool mainitud, ei ole alati kindel, kumb meetod (arvutuslik või qPCR) on *CNV* hindamisel teinud vea. Samas sõltub algoritmide efektiivsus *CNV*-de leidmisel palju algandmete täpsusest. Seega tuleks edaspidi komplekssete *CNV*-de uurimisel kasutada suurema markerite arvuga genotüpiseerimiskiipe kui antud töös.

Käesolevas töös kasutati arendatud *CNV*-de leidmise meetodit koopiaarvu muutuste päranduvuse uurimisel. Töös analüüsiti *CNV*-de päranduvuse erinevusi Eesti VAM kohordi

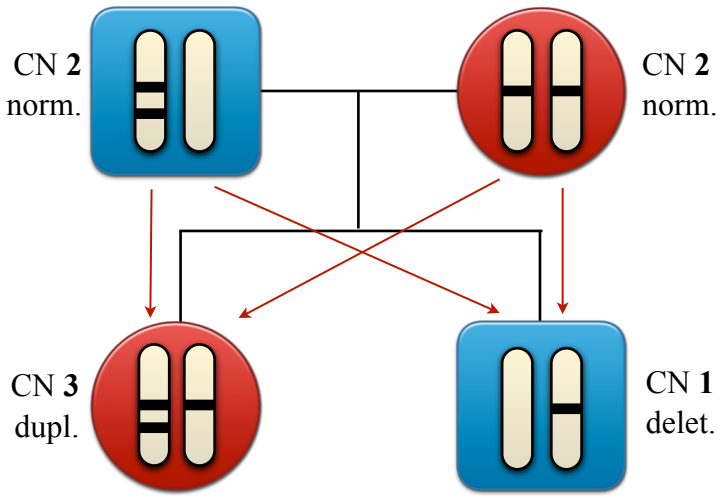
perekondade ja HapMap triode vahel. Deletsioonide suurema suhtarvu põhjust VAM kohordi järglaspõlvkonna puhul on ilma lisakatseid tegemata keeruline analüüsida, kuid edaspidistes VAM uuringutes tasub kandidaatgeenide otsimisel tähelepanu pöörata deletsioonidega rikastatud genoomipiirkondadele, kuna analüüsitud VAM kohordi järglaspõlvkond paistis silma just deletsioonide suhteliselt suure arvu poolest.

Varasemates uuringutes on täheldatud geneetiliste aberratsioonide seost vanusega (*Croen et al.*, 2001) - indiviidi vanuse kasvades suureneb genoomis struktuursete muutuste hulk (k.a. *CNV*-d), mis mõjutab ka järglaspõlvkonna genoomi. Antud töös ei arvestatud päranduvuse analüüsimisel vanemate vanust, mis võis olla üks teguritest VAM kohordi suhteliselt suure deletsioonide ja *de-novo CNV*-de hulga leidmisel. Kuigi HapMap triode puhul meil vanuse andmed puuduvad, on see teada VAM kohordi indiviididel. Hetkel on analüüsimisel valitud perekonnad Eesti üldpopulatsioonist (Eesti Geenivaramust), mida saab edukalt kasutada päranduvusuuringutes sarnaselt HapMap triodele referentsandmeteks. Kuna Eesti üldpopulatsiooni indiviidide kohta on teada ka nende vanus, saab võrdleval analüüsil VAM kohordiga arvesse võtta vanemate vanuse mõju *CNV*-de päranduvusele (ja *de-novo CNV*-de tekkimise sagedusele), mis annab võrdlustulemustele tunduvalt suurema kaalu.

Enamikes koopiaarvu analüüsides ei eristata diploidse genoomi haploidseid koopiaid, vaid koopiaarv määratakse summaarselt mõlema alleeli kohta. Seega tähistab somaatiliste kromosoomide puhul koopiaarv 2 normaalset (referents) koopiaarvu, kuna eeldatakse, et kumbki sama lookuse koopia on eraldi kromosoomis (üks koopia pärandunud emalt, teine isalt). Summaarne koopiaarv võib aga olla eksitav. Näiteks tähistab *de-novo CNV* lookust, kus vanematel on normaalne koopiaarv (somaatiliste kromosoomide puhul 2), kuid lapsel esineb *CNV*. Kuigi selline järglaspõlvkonna *CNV* tundub olevat *de-novo* (mitte pärandunud vanematelt vaid tekkinud iseseisvalt), võib alleelset koopiaarvu arvesse võttes selgitada selle *CNV* teket vanemate alleelide teatud kombinatsioonidega (näiteks kui vanematel esineb kahe või null koopiaga alleele). Seega võivad ka osad antud töös leitud *de-novo CNV*-d olla hoopis mendeliaalse päranduvuse tulemus. Sellise päranduvusmehhanismi kohta on toodud näide joonisel 5, kus järglase *CNV* tekke põhjuseks on ühe vanema antud lookuse mõlema koopia koondumine ühte alleeli.

Genotüpiseerimiskiipide kasutamise eeliseks CNV-de määramisel on genotüpiseerimis-analüüsist saadud lisainfo markerite alleelse koostise kohta. Antud töös analüüsitud indiviidide puhul (HapMap triod ja VAM kohort) on teada nii *CNV*-de summaarne koopiaarv, genotüpiseerimisandmed kui ka perekonnainfo, mida ühendades on iga *CNV* jaoks (piisava hulga heterosügootsete markerite olemasolul) võimalik arvutada alleelne koopiaarv. Suuremate

koopiaarvude puhul on summaarse koopiaarvu alleelideks lahutamine keerukas, sest antud lookuse koopiate jaotumiseks isa-ema-laps trio alleelide vahel on eksponentsiaalselt rohkem võialusi, millest kõige tõenäolisema kombinatsiooni leidmine ei ole alati võimalik. Sellest hoolimata on alleelse koopiaarvu arvutamise algoritm hetkel arenduses ning esialgsed tulemused näitavad, et olenevalt päranduvusmustrist ja koopiaarvust on alleelne koopiaarv osade *CNV*-de jaoks arvutatav.



Joonis 5. *CNV-de alleelne päranduvus.* Näiliselt *de-novo CNV-d* lastel võivad tegelikult olla mendeliaalselt pärandunud alleelid, mis annavad normaasest erineva koopiaarvu alleelide kombinatoorika tõttu. Mustad triibud tähistavad antud lookuse alleelset koopiaarvu, summaarne koopiaarv (CN) on toodud indiviidi sümboli (punane ring/sinine nelinurk) kõrval. Nooled näitavad pärandunud *CNV* alleelide jaotumist lastel.

DNA koopiaarvu määramine ja koopiaarvu muutuste pärandumise uurimine genotüpiseerimiskiipide andmete põhjal

Andres Veidenberg

Kokkuvõte

Koopiaarvult varieeruvad DNA piirkonnad katavad inimgenoomist suurema ala kui ükski teine teada olev polümorfism, mõjutades lisaks normaalsele geneetilisele varieeruvusele ka paljude geneetiliste haiguste ja kasvajate arengut. DNA koopiaarvu variantide (*CNV*-de) leidmisel on üheks sagedasti kasutatavaks meetodiks genotüpiseerimiskiipide rakendamine.

Antud magistritöö andis ülevaate *CNV*-de uurimisest ja võrdles erinevaid mikrokiibipõhiseid meetodeid *CNV*-de leidmiseks, pöörates rohkem tähelepanu genotüpiseerimiskiipide kasutamisele ja sealt *in silico* meetodil *CNV*-de määramisele.

Käesoleva magistritöö tulemusena arendati välja arvutuslik meetod Illumina ja Affymetrixi genotüpiseerimiskiipide andmetest *CNV*-de leidmiseks. Väljatöötatud meetodiga leitud *CNV*-sid valideeriti reaalaja kvantitatiivse *PCR*-i andmete abil. Lisaks analüüsiti väljatöötatud metoodikaga leitud *CNV*-de päranduvust HapMap populatsiooni triode ja Eesti VAM uuringu kohordi perekonnaandmete põhjal.

CNV-de päranduvusanalüüsides leidsime deletsioonide ja *de-novo CNV*-de suuremat esinemist Eesti VAM kohordi järglaspõlvkonnas võrreldes vastavate referentsandmetega HapMap populatsioonist.

DNA Copy Number Assessment and Heritage Analysis of Copy Number Variations from Genotyping Data

Andres Veidenberg

Summary (kokkuvõte inglise keeles)

Copy number variations (CNVs) cover more genetic content in human genome than any other known type of polymorphisms, having its effect on various genetic traits and tumorigenesis as well as shaping general genetic variation. For CNV discovery, genotyping microarrays are often used.

In the current thesis an overview was given to introduce different methods in CNV studies and discovery. Particular emphasis was put on describing genotyping microarrays and *in silico* methods that are used to discover CNVs from genotyping data.

As a practical outcome of this thesis, *in silico* method was developed for finding CNVs by using genotyping data from Illumina or Affymetrix genotyping microarrays. Some CNVs found with this method were validated by real-time quantitative PCR analysis. In addition, CNVs were found and their heritage was comparatively analyzed for HapMap trios and for families from Estonian Mental Retardation (MR) study.

The results of our CNV heritage analysis show that the offspring in the MR cohort has comparatively more deletions and *de-novo* CNVs than that of HapMap families.

Tänuavaldused

Autor tänab Prof. Ants Kurge ja tema töörühma, kellega on olnud rõõm koos töötada kogu magistriõpingute vältel. Suurimad tänusõnad kuuluvad erakordselt abivalmis ja professionaalsetele kolleegidele bioinformaatika õppetoolist, kes on prof. Maido Remm visionäärsel juhtmisel mind toetanud ja õpetanud alates minu esimestest bioinformaatiku päevadest.

Eriti tahan tänada oma juhendajat teadur Priit Paltat, kes on oma kannatliku ja süsteemse lähenemisega olnud minu teejuhiks ja valgustajaks teadusmaailmas.

Lõpetuseks tahan tänada oma sõpru, perekonda ja kallist Liinat, kelle toetus on olnud lihtsalt asendamatu.

Kasutatud kirjandus

Affymetrix, Inc. (2005) Affymetrix: GeneChip Human Mapping 500 K Array Set Data Sheet. [http://www.affymetrix.com].

Armour, J.A., Sismani, C., Patsalis, P.C., Cross, G. (2000) Measurement of locus copy number by hybridisation with amplifiable probes. Nucleic Acids Res., **28**, 605 - 609.

Carter, N.P. (2007) Methods and strategies for analyzing copy number variation using DNA microarrays. Nat. Genet., **39(7 Suppl)**, S16 - S21.

Chelly, J. & Mandel, J.L. (2001) Monogenic causes of X-linked mental retardation. Nat. Rev. Genet., **2**, 669-680.

Colella, S., Yau, C., Taylor, J.M., Mirza, G., Butler, H., *et al.* (2007) QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. Nucleic Acids Res., **35**, 2013-2015.

Cook, E.H. & Scherer, S. (2008) Copy-number variations associated with neuropsychiatric conditions. Nature, **455**, 919 - 923.

Croen, L.A., Grether, J.K. & Selvin, S. (2001) The epidemiology of mental retardation of unknown cause. Pediatrics, **107**, E86.

Fiegler, H., Redon, R., Andrews, D., Scott, C., Andrews, R., *et al.* (2006) Accurate and reliable high-throughput detection of copy number variation in the human genome. Genome Res., **16**, 1566 - 1574.

Gribble, S.M., Kalaitzopoulos, D., Burford, D.C., Prigmore, E., Selzer, R.R., *et al.* (2007) Ultra-high resolution array painting facilitates breakpoint sequencing. J. Med. Genet., **44**, 51 - 58.

Gunderson, K.L., Steemers, F.J., Lee, G., Mendoza, L.G, Chee, M.S. (2005) A genome-wide scalable SNP genotyping assay using microarray technology. Nat. Genet., **37**, 549 - 554.

Henrichsen, C.N., Vinckenbosch, N., Zöllner, S., Chaignat, E., Pradervand, S., *et al.* (2009) Segmental copy number variation shapes tissue transcriptomes. Nat. Genet., **41**, 424 - 429.

Iafrate, A.J., Feuk, L., Rivera, M.N., Listewnik, M.L., Donahoe, P.K., Qi, Y., *et al.* (2004) Detection of large-scale variation in the human genome. Nat. Genet., **36**, 949 - 951.

Kallioniemi, A., Kallioniemi, O. P., Sudar, D., Rutovitz, D., Gray, J. W., *et al.* (1992) Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. Science, **258**, 818 - 21.

Kallioniemi, A., Visakorpi, T., Karhu, R., Pinkel, D., Kallioniemi, O.P. (1996) Gene copy number analysis by fluorescence in situ hybridization and comparative genomic hybridization. Methods, **9**, 113 - 121.

Lucito, R., Healy, J., Alexander, J., Reiner, A., Esposito, D., Chi, M., *et al.* (2003) Representational oligonucleotide microarray analysis: a high-resolution method to detect genome copy number variation. Genome Res., **13**, 2291 - 2305.

Macconaill, L.E., Aldred, M.A., Lu, X., Laframboise, T. (2007) Toward accurate highthroughput SNP genotyping in the presence of inherited copy number variation. BMC Genomics, **8**, 211.

Mantripragada, K.K., Buckley, P.G., de Stahl, T.D., Dumanski, J.P. (2004) Genomic microarrays in the spotlight. Trends Genet., **20**, 87 - 94.

Nannya, Y., Sanada, M., Nakazaki, K., Hosoya, N., Wang, L., *et al.* (2005) A robust algorithm for copy number detection using high-density oligonucleotide single nucleotide polymorphism genotyping arrays. Cancer Res., **65**, 6071 - 6079.

Olshen, A.B., Venkatraman, E.S., Lucito, R., Wigler, M. (2004) Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics, **5**, 557 - 572.

Oostlander, A.E., Meijer, G.A., Ylstra, B. (2004) Microarray-based comparative genomic hybridization and its applications in human genetics. Clin. Genet., **66**, 488 - 495.

Perry, G.H., Ben-Dor, A., Tsalenko, A., Sampas, N., *et al.* (2008) The Fine-Scale and Complex Architecture of Human Copy-Number Variation. Am. J. Hum. Genet., **82**, 685 - 695.

Pinkel, D., Segraves, R., Sudar, D., Clark, S., Poole, I., Kowbel, D., *et al.* (1998) High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. Nat. Genet., **20**, 207 - 211.

Pique-Regi, R., Monso-Varona, J., Ortega, A., Seeger, R.C., Triche, T.J., *et al.* (2008) Sparse representation and Bayesian detection of genome copy number alterations from microarray data. Bioinformatics, **24**, 309 - 318.

Price, T.S., Regan, R., Mott, R., Hedman, A., Honey, B., *et al.* (2005) SW-ARRAY: a dynamic programming solution for the identification of copy-number changes in genomic DNA using array comparative genome hybridization data. Nucleic Acids Res., **33**, 3455 - 3464.

Redon, R., Ishikawa, S., Fitch, K.R., Feuk, L., Perry, G.H., Andrews, T.D., *et al.* (2006) Global variation in copy number in the human genome. Nature, **444**, 444 - 454.

Rigaill, G., Hupe, P., Almeida, A., La Rosa, P., Meyniel, J.P., *et al.* (2008) ITALICS: an algorithm for normalization and DNA copy number calling for Affymetrix SNP arrays. Bioinformatics, **24**, 768 - 774.

Sebat, J., Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P., *et al.* (2004) Large-scale copy number polymorphism in the human genome. Science, **305**, 525 - 528.

Sebat, J., Lakshmi, B., Malhotra, D., Troge, J., Lese-Martin, C., Walsh, T., *et al.* (2007) Strong association of de novo copy number mutations with autism. Science, **316**, 445 - 449.

Solinas-Toldo, S., Lampel, S., Stilgenbauer, S., Nickolenko, J., Benner, A., *et al.* (1997) Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances. Genes Chromosom. Cancer, **20**, 399 - 407.

Steemers, F.J., Chang, W., Lee, G., Barker, D.L., Shen, R., Gunderson, K. L. (2007) Whole-genome genotyping with the single-base extension assay. Nat. Methods, **3**, 31 - 33.

Vermeesch, J.R., Melotte, C., Froyen, G., *et al.* (2005) Molecular karyotyping: array CGH quality criteria for constitutional genetic diagnosis. J. Histochem. Cytochem., **53**, 413 - 422.

Wang, K., Li, M., Hadley, D., Liu, R., Glessner, J., *et al.* (2007) PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genomse Res., **17**, 1665 - 1674.

Wang, K., Chen, Z., Tadesse, M.G., Glessner, J., *et al*. (2008) Modeling genetic inheritance of copy number variations. Nucleic Acids Res., **36**, e138.

Yau, C., Holmes, C.C. (2008) CNV discovery using SNP genotyping arrays. Cytogenet. Genome Res., **123**, 307-312.

Zhao, X., Li, C., Paez, J.G., Chin, K., Jänne, P.A., Chen, T.H., *et al.* (2004) An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism arrays. Cancer Res., **64**, 3060 - 3071.

Zöllner, S., Su, G., Stewart, W.C., Chen, Y., McInnis, *et al*. (2009) Bayesian EM algorithm for scoring polymorphic deletions from SNP data and application to a common CNV on 8q24. Genet. Epidem., **33**, 357 - 368.

Lisad

Lisa 1. Osaline väljavõte skripti *CNV_heritage.pl* väljundfailist HapMap andmete analüüsil.

```
STATISTICS:
==================================================
Families: 60 Trios: 60 Individuals: 180 Cases: 1551 (plus unvalidated: 802)

HERITAGE    RATIO
mom 0 dad 0 => daughter 1     2.515% (39/1551)
mom 0 dad 0 => son 1    4.836% (75/1551)
mom 0 dad 1 => daughter 0     8.575% (133/1551)
mom 0 dad 1 => daughter 1     5.674% (88/1551)
mom 0 dad 1 => son 0    19.536% (303/1551)
mom 0 dad 1 => son 1    12.057% (187/1551)
mom 1 dad 0 => daughter 0     11.283% (175/1551)
mom 1 dad 0 => daughter 1     6.125% (95/1551)
mom 1 dad 0 => son 0    16.119% (250/1551)
mom 1 dad 0 => son 1    8.511% (132/1551)
mom 1 dad 1 => daughter 0     0.709% (11/1551)
mom 1 dad 1 => daughter 1     1.483% (23/1551)
mom 1 dad 1 => son 0    0.645% (10/1551)
mom 1 dad 1 => son 1    1.934% (30/1551)
...where X chr:
mom 0 dad 1 => daughter 0     0.387% (6/1551)
mom 0 dad 1 => son 0    0.129% (2/1551)
mom 0 dad 1 => son 1    0.967% (15/1551)
mom 1 dad 0 => daughter 0     0.129% (2/1551)
mom 1 dad 0 => daughter 1     0.258% (4/1551)
mom 1 dad 0 => son 0    0.193% (3/1551)
mom 1 dad 0 => son 1    0.129% (2/1551)
mom 1 dad 1 => daughter 1     0.064% (1/1551)

Summary:
heritage              ratio   (cases)     del/dup_ratio (cases)   parents
parents 01 => child 0 55.513% (861/1551)                          299/137
parents 01 => child 1 32.366% (502/1551) 20.24/12.12  (314/188)  172/103
parents 0 => child 1  7.350%  (114/1551) 5.158/2.192  (80/34)    0/0
parents 1 => child 0  1.354%  (21/1551)                           22/20
parents 1 => child 1  3.417%  (53/1551)  1.09/2.32     (17/36)   37/69
```