TARTU ÜLIKOOL
BIOLOOGIA-GEOGRAAFIATEADUSKOND
MOLEKULAAR- JA RAKUBIOLOOGIA INSTITUUT
BIOINFORMAATIKA ÕPPETOOL

Priit Palta

Hübridiseerimisproovide disaini metoodika geeni koopiaarvu määramiseks

Bakalaureusetöö

Juhendaja: Maido Remm, prof., PhD

TARTU
2005

Sisukord

Lühendid ja mõisted

BAC	bakteri kunstlik kromosoom (*bacterial artificial chromosome*)
bp	aluspaar (*base pair*)
cDNA	mRNA-lt pöördtranskriptaasi abil sünteesitud mRNA-ga komplementaarne DNA
CGH	võrdlev genoomne *in situ* hübridisatsioon
DNA	desoksüribonukleiinhape
FISH	fluorestsents *in situ* hübridisatsioon
FR	heleduste intensiivsuste suhe
GB	gigabait, 10^9 baiti (*gigabyte*)
kb	10^3 aluspaari (*kilo base*)
MAPH	multipleks amplifitseeritavate proovide hübridisatsioon
MB	megabait, 10^6 baiti (*megabyte*)
Mb	10^6 aluspaari (*mega base*)
MFISH	mitmevärvi fluorestsents *in situ* hübridisatsioon
mRNA	matriits- ehk informatsiooniline RNA (*messenger RNA*)
märklaud	käesolevas töös mikrokiibi pinnale seotud DNA järjestus (*target*)
PAC	P1 kunstlik kromosoom (*P1 artificial chromosome*)
parser	programm, millega muudetakse ühe programmi väljund teisele programmile sobivaks sisendiks
PCR	polümeraasi ahelreaktsioon (*polymerase chain reaction*)
proov	hübridiseerimisproov, käesolevas töös genoomsele DNA-le või mikrokiibile hübridiseeritav järjestus
RNA	ribonukleiinhape
SKY	spektraalne karüotüpiseerimine
speisser	järjestuse pikendamiseks kasutatav molekul (*spacer*)
spot	mikrokiibile prinditud märklaud-järjestus
stacking	aluste kuhjumine, aheldumine (*base stacking*)

Sissejuhatus

Geneetilised haigused võivad olla põhjustatud mitmetest erinevatest muutustest genoomis. Sellisteks muutusteks võivad olla punktmutatsioonid või suuremad ümberkorraldused geneetilises materjalis. Viimasel aastatel on järjest rohkem leitud, et muutused DNA koopiaarvus ja järjestuste duplitseerumine/deleteerumine on tõenäoliselt paljude geneetiliste haiguste põhjustajaks. On näidatud, et geenidoosi ehk geeni koopiaarvu muutused on iseloomulikud kasvajarakkudele ja otseses seoses arenguhäirete ja vaimse alaarengu esinemisega (Albertson & Pinkel, 2003; Pollack *et al.*, 2002). Tekkinud muutused võivad olla väga erineva suurusega, alates tervete kromosoomide kordistumisest/kaotsiminekust kuni geenide ja üksikute eksonite kordistumiseni/kaotsiminekuni.

Et eristada patoloogilisi muutusi inimeste vahelisest varieeruvusest, on tähtis aru saada sellest, kui suur on inimeste vaheline „normaalne" varieeruvus genoomides ehk genoomi plastilisus. Seetõttu on oluline töötada välja uusi ja arendada edasi olemasolevaid meetodeid, millega detekteerida erinevate lookuste koopiaarvu inimese genoomis.

I KIRJANDUSE ÜLEVAADE

1. Näiteid aberratsioonide tuvastamise meetoditest

Mitmed erinevad molekulaarse tsütogeneetika meetodid nagu kromosoomide vöödistamine (*Q-banding*), radioaktiivne *in situ* hübridisatsioon (*radioactive in situ hybridization*), fluorestsents *in situ* hübridisatsioon (*fluorescence in situ hybridization, FISH*), spektraalne karüotüpiseerimine (*spectral karyotyping, SKY*) ja mitmevärvi fluorestsents *in situ* hübridisatsioon (*multiplex fluorescence in situ hybridization, MFISH*) võimaldavad uurida erineva ulatusega genoomsete aberratsioonide, sealhulgas geeni koopiaarvu muutusi. Enamasti on nende meetoditega võimalik määrata siiski vaid suuremaid ja teadaolevaid muutusi genoomis. Vaja oleks aga meetodeid, millega saaks usaldusväärselt leida ja kaardistada ka väiksemaid ning uusi, seni teadmata aberratsioone uuritavas genoomis.

Viimastel aastatel on arendatud uusi, mikrokiipidel (*microarray, array*) põhinevad meetodeid, mis ühendavad traditsioonilisi tsütogeneetilisi ja moodsaid mikrokiipidel põhinevad võimalusi. Sellised uued meetodid võimaldavad muutunud genoomseid piirkondi kaardistada suurema täpsusega ja oluliselt kiiremini, mis omakorda lihtsustab aberratsioonide uurimist ja nendega seotud haiguste/häirete diagnostikat, aidates meil paremini mõista selliste muutuste ja nendega tihtipeale kaasnevate võimalike haiguste olemust.

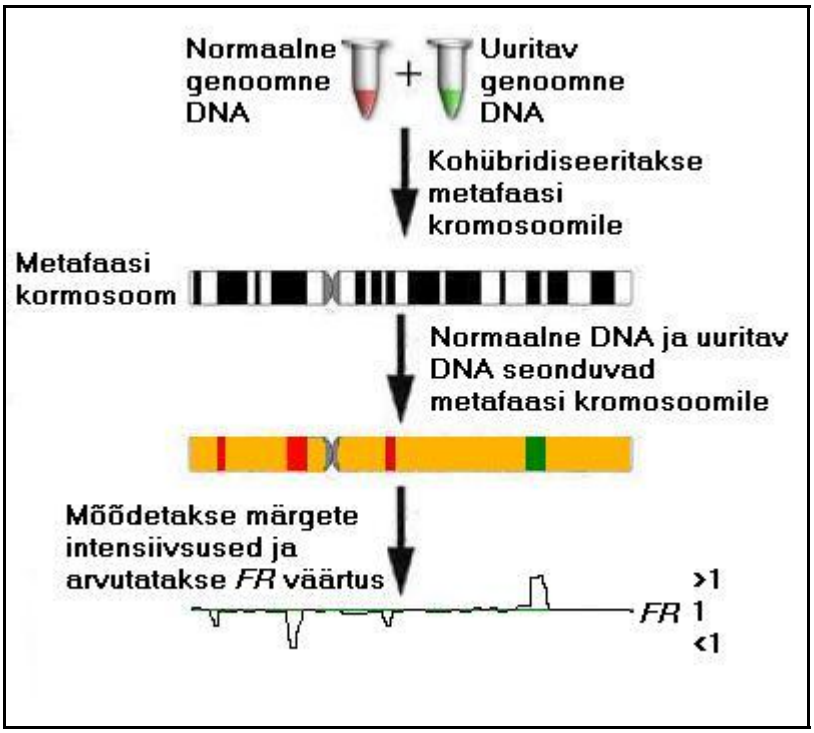
1.1 Traditsiooniline võrdlev genoomne *in situ* hübridisatsioon

Võrdlev genoomne *in situ* hübridisatsioon ehk *CGH* (*comparative genomic in situ hybridization*) on fluorestsents *in situ* hübridisatsioonist (*fluorestsents in situ hybridization, FISH*) (Langer-Safer *et al.*, 1982) välja arenenud meetod, mis võimaldab ühe eksperimendi käigus üle kogu genoomi avastada seal toimunud DNA koopiaarvu

muutusi (Kallioniemi *et al.*, 1992). Traditsiooniline *CGH*-i meetod (*chromosome CGH*) põhineb DNA pöördvärvimise ideel, mille puhul erinevate fluorofooridega märgistatud uuritav DNA (testi-DNA) ja tsütogeneetiliselt kontrollitud normaalne DNA (kontroll-DNA) hübridiseeritakse normaalsetele metafaasi kromosoomidele, mis on seotud tahkele kandjale (Kallioniemi *et al.*, 1992; Kallioniemi *et al.*, 1994a; Kallioniemi *et al.*, 1994b). Sealjuures lisatakse ka liigispetsiifiline märgistamata konkurent-DNA (inimese *Cot-1* DNA), et vähendada märgitud proovide mittespetsiifilist seondumist kordusjärjestustele. Hübridisatsioonil seonduvad nii testi-DNA kui ka kontroll-DNA oma märklaudjärjestusele metafaasi kromosoomil. Spetsiifiliselt seondunud testi-DNA ja kontroll-DNA järjestused tehakse detekteeritavaks, märkides neid näiteks erinevate fluorofooridega. Detekteerimine toimub kohe pärast hübridiseerimist fluorestsentsmikroskoobi või spetsiaalse *CCD* (*charge-coupled device*) kaamera ja vastava digitaalset andmetöötlust võimaldava tarkvara abil. Mõõdetakse erinevate fluorofooride intensiivsus (*intensity*) ja arvutatakse iga kromosoomisegmendi jaoks märgete heleduste intensiivsuste suhe ehk *FR* väärtus (*fluorescence ratio*). Piirkonnad, kus on toimunud muutused DNA-s (deletsioonid, duplikatsioonid), on nähtavad kui kahe fluorofoori muutunud intensiivsuse suhtega alad märklaud-kromosoomidel. Saadud *FR* väärtus näitab kromosoomide või nende osade esindatust testi genoomis kontroll-genoomi suhtes ehk lahtimõtestatult: kui palju kordi esines mingi kindel piirkond uuritava raku genoomis rohkem/vähem kordi, kui normaalses genoomis (Kallioniemi *et al.*, 1994a; Kallioniemi *et al.*, 1994b). Kui uuritavas ja kontrolli DNA-s on kindlas lookuses järjestusi võrdselt, on tegemist normaalse koopiaarvuga ja vastava piirkonna *FR* väärtus on 1. Kui *FR* väärtus mingis piirkonnas on suurem või väiksem 1-st, on selles piirkonnas uuritavas DNA-s suure tõenäosusega vastavalt duplikatsioon või deletsioon.

CGH-i eelisteks teiste *in situ* hübridisatsiooni meetodite ees on suur kiirus ja see, et kromosoomide ja nende alade lisandumist ja/või kadumist uuritavas genoomis saab hinnata ilma uuritavast DNA-st kromosoomipreparaati tegemata. See on aga oluline näiteks kasvajarakkude uurimisel, kuna tihtipeale on kromosoomipreparaatide tegemine kasvajarakkude komplekssetest genoomidest raskendatud, kui mitte võimatu. Meetodi puuduseks on aga võrdlemisi limiteeritud lahutusvõime (resolutsioon), mis kromosoomide kõrge kondenseerituse tõttu ei ületa kadude uurimise puhul 10 Mb ja juurdetulekute puhul 2 Mb. Puuduseks on ka see, et kogu protsess on ainult osaliselt automatiseeritud ja lõpuks peab ikkagi kogenud tsütogeneetik leidma muutunud

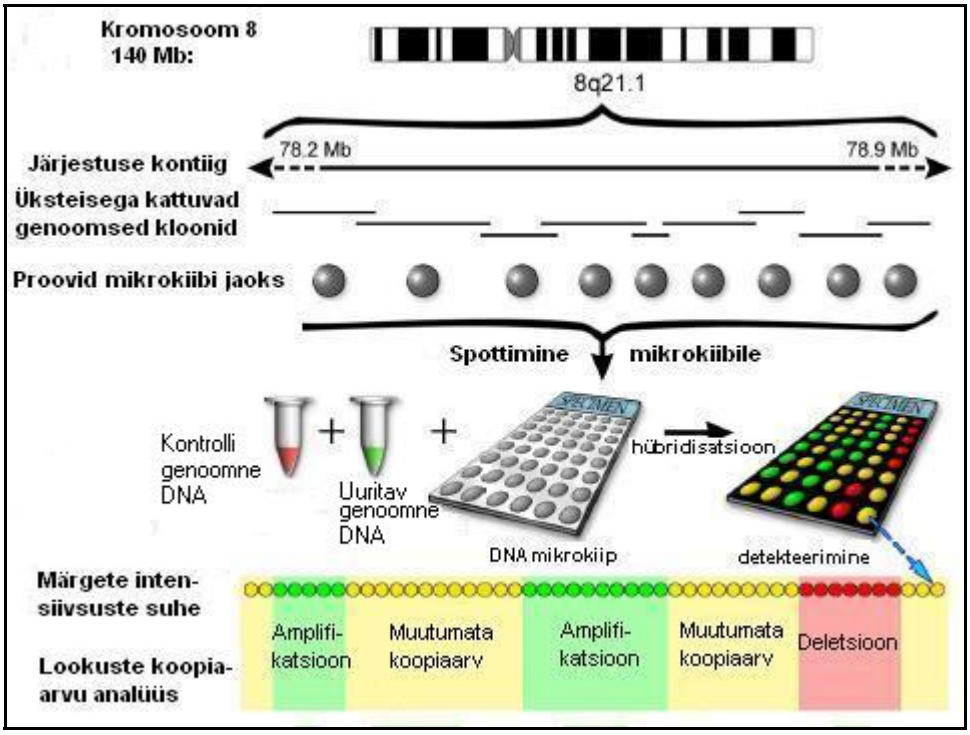
regioonid uuritavas genoomis (Beheshti *et al.*, 2002). Miinuseks võib pidada sedagi, et võrdleva genoomse *in situ* hübridisatsiooniga saab hinnata vaid kromosoomimaterjali lisandumist ja kadumist, samas kui ümberpaigutused translokatsioonide ja inversioonide näol jäävad avastamata (Forozan *et al.*, 1997; Kallioniemi *et al.*, 1992; Kallioniemi *et al.*, 1994b). Traditsioonilise võrdleva genoomse *in situ* hübridisatsiooni põhietapid on kujutatud joonisel 1.



Joonis 1. Traditsiooniline võrdlev genoomne *in situ* hübridisatsioon. Normaalne ja uuritav DNA märgistatakse erinevalt ja kohübridiseeritakse tahkele kandjale seotud metafaasi kromosoomidele. Hübridisatsioonil seonduvad nii uuritav testi-DNA kui ka normaalne kontroll-DNA oma märklaudjärjestusele metafaasi kromosoomil. Seejärel mõõdetakse erinevate fluorofooride intensiivsus ja arvutatakse iga kromosoomisegmendi jaoks märgete intensiivsuste suhe ehk *FR* väärtus. Uuritava genoomse DNA piirkonnas, millele vastav *FR* väärtus on oluliselt erinev 1-st, on suure tõenäosusega tegemist deletsiooni või duplikatsiooniga.

1.2 Mikrokiipidel põhinev võrdlev genoomne *in situ* hübridisatsioon

Tänaseks on lisaks traditsioonilisele *CGH*-le lisandunud ka mikrokiipidel põhinev võrdlev genoomne *in situ* hübridisatsioon (*array CGH, matrix CGH*). Mikrokiipidel *CGH*-i puhul kasutatakse metafaasi kromosoomide asemel hübridisatsiooni märklaudadena (*target sequence*) mikrokiibile seotud järjestusi. Uuritavat regiooni esindavate järjestuste olemasolu korral mikrokiibil on selle meetodiga võimalik tunduvalt tõsta süsteemi lahutusvõimet, et kindlaks määrata muutunud koopiaarvuga regioone (Solinas-Toldo *et al.*, 1997). Hübridisatsiooni märklaudadena kasutatakse mikrokiibil enamasti genoomseid järjestusi, cDNA järjestusi või ka oligonukleotiide, mis on kindlas järjekorras seotud mikrokiipidele (Mantripragada *et al.*, 2004). Mikrokiipidel läbiviidava *CGH*-i puhul toimub kõigepealt uuritava DNA (testi-DNA) ja normaalse DNA (kontroll-DNA) erinev märkimine (enamasti fluorestseeruvate värvidega). Seejärel proovid kohübridiseeritakse mikrokiibil olevatele järjestustele, lisades mittespetsiifiliste seondumiste ärahoidmiseks konkurent-DNA-d. Pärast hübridisatsiooniprotsessi kiibid pestakse alaneva soola kontsentratsiooniga lahustega, et eemaldada mitteseondunud ja mittespetsiifiliselt seondunud järjestused. Seejärel kiipidel olevad erivärvilised signaalid detekteeritakse spetsiaalse seadmega, mis skaneerides laseritega loeb iga spoti signaali intensiivsuse mõlema märke fluorestseerumise lainepikkusel. Saadud signaalid salvestatakse ja analüüsitakse spetsiaalse tarkvara abil. Iga mikrokiibil olnud järjestuse jaoks arvutatakse testi-DNA ja kontroll-DNA märgete heleduste suhe, mis näitab, kas vastav regioon on uuritavas genoomis võrreldes normaalse genoomiga üle- või alaesindatud ehk teisisõnu duplitseerunud või deleteerunud (Solinas-Toldo *et al.*, 1997). Mikrokiipidel läbiviidava *in situ* hübridisatsiooni põhietapid on kujutatud joonisel 2:



Joonis 2. Genoomsete järjestuste kloonidega mikrokiibil läbiviidav võrdlev genoomne *in situ* hübridisatsioon. Huvipakkuv regioon paljundatakse üksteisega osaliselt kattuvate genoomsete kloonidena ja spotitakse mikrokiibile. Normaalne (kontroll-DNA) ja uuritav DNA märgistatakse erinevalt ja kohübridiseeritakse mikrokiibil olevate järjestustega. Kiibilt saadavad signaalid detekteeritakse ning analüüsitakse. Regioonid, kus uuritavas DNA-s on toimunud muutused, paistavad välja kui muutunud märgete intensiivsuste suhtega spotid mikrokiibil.

Genoomseid järjestusi kandvate mikrokiipide märklaud-järjestused (*target sequence*) pärinevad tavaliselt genoomsetest *BAC*, *PAC* või kosmiidsetest kloonidest, mis on mitu suurusjärku väiksemad, kui tavalises lähenemises kasutatud terved kromosoomid (Pinkel *et al.*, 1998; Solinas-Toldo *et al.*, 1997). Märklaud-järjestuse lühenemine toob aga kaasa lahutusvõime tõusu detekteerimisel, tagades muutunud koopiaarvuga lookuste täpsema lokaliseerimise. Samuti on võimalik huvipakkuvaid regioone esindavaid järjestusi piisava ja ühtlase katvusega mikrokiibile panna, mis parandab tulemuste täpsust. On näidatud, et niimoodi on võimalik avastada 100 kb ja isegi 50 kb suurusi muutusi uuritavas genoomis (Albertson & Pinkel, 2003; Wessendorf *et al.*, 2001). Genoomsete järjestustega mikrokiipidel läbiviidava võrdleva genoomse *in situ* hübridisatsiooni eelisteks võrreldes traditsioonilise lähenemisega on suurem

resolutsioon ja täielik automatiseeritus, mis teeb võimalikuks paljude paralleelsete testide läbiviimise (Albertson, 2003; Buckley *et al.*, 2002; Pinkel *et al.*, 1998).

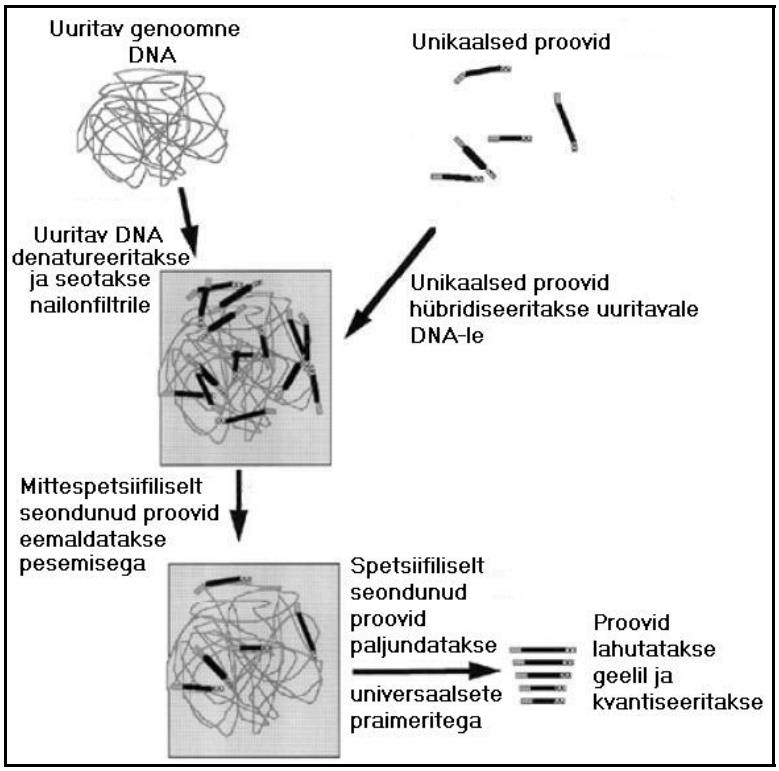
CGH-i mikrokiibid, millel kasutatakse sihtmärgina cDNA järjestusi, võeti kasutusele 1999 aastal (Pollack *et al.*, 1999). Kuigi cDNA mikrokiipe kasutatakse enamasti geenikespressiooniuuringuteks, on näidatud, et neid saab edukalt kasutada ka huvipakkuva genoomse lookuse koopiaarvu määramiseks. cDNA mikrokiipide abil on võimalik avastada isegi heterosügootset deletsiooni ja ühekordse täpsusega ka geeni amplifitseerumist. Liites need mõlemad cDNA kiipide funktsioonid, saadakse tundlikud süsteemid, millega on paralleelselt võimalik uurida geenikespressiooni taset ja geeni koopiaarvu muutusi ning nendevahelisi seoseid kasvajarakkudes (Heidenblad *et al.*, 2004; Heidenblad *et al.*, 2005; Pollack *et al.*, 1999).

cDNA mikrokiipide eeliseks teiste *CGH*-i meetodite ees on automatiseeritus ja suur lahutusvõime aberratsiooniga piirkonna leidmisel ning ka see, et need on kommertsiaalselt kõige paremini kättesaadavad. Puudusteks on aga küllaltki suur valepositiivsete (15 %) ja valenegatiivsete (15 %) tulemuste arv, sõltumine genoomi järjestuse kvaliteedist (Beheshti *et al.*, 2002) ja see, et nendega on võimalik uurida vaid teadaolevate geenide lookuseid.

Oligonukleotiidsete proovidega *CGH* mikrokiipide (*oligo-array CGH*) puhul pannakse mikrokiibile suhteliselt lühikesed (tavaliselt 50-70 nukleotiidi) sünteesitud järjestused, millele kohübridiseeritakse testi- ja kontroll-DNA (Mantripragada *et al.*, 2003). Suurim erinevus võrreldes genoomseid kloone ja cDNA-d kandvate platvormidega on see, et oligonukleotiidsete järjestustega mikrokiibil läbiviidava võrdleva genoomse *in situ* hübridisatsioonil proovide kohübridiseerimisel enamasti ei lisata konkurent-DNA-d. Seda sellepärast, et oligonukleotiidsed järjestused on disainitud ja sünteesitud enamasti unikaalsetena, st. nii, et igale proovile vastaks normaalses genoomis üksainus kindel järjestus, mille koopiaarvu uuritavas genoomis kontroll-genoomi suhtes tuvastatakse. Meetodi eeliseks teiste kiibipõhiste võrdleva genoomse *in situ* hübridisatsiooni meetodite ees on suurem resolutsioon ja väiksem taustamüra (Lucito *et al.*, 2003; Mantripragada *et al.*, 2004).

1.3 Multipleks amplifitseeritavate proovide hübridisatsioon

Multipleks amplifitseeritavate proovide hübridisatsioon ehk *MAPH* (*multiplex amplifiable probe hybridization*) on meetod, millega saab uurida spetsiifiliselt valitud ja huvipakkuvaid lookuseid seal toimunud võimalike koopiaarvu muutuste tuvastamiseks (Armour *et al.*, 2000; Hollox *et al.*, 2002). Meetod võimaldab tuvastada huvipakkuvate lookuste koopiaarvu uuritavas genoomis, seda isegi 100 bp lahutusvõimega (Armour *et al.*, 2000). *MAPH*-i puhul fikseeritakse uuritav (patsiendi) denatureeritud genoomne DNA nailonmembraanile ja hübridiseeritakse spetsiifiliste proovidega, mida on hübridisatsioonilahuses suures ülehulgas (Hollox *et al.*, 2002). Hübridisatsioonisegusse lisatakse ka konkurent-DNA (inimese *Cot-1* DNA), et blokeerida hübridiseerimisproovide mittespetsiifilist seondumist uuritava genoomse DNA kordusjärjestustele. Hübridiseerimisproovid on disainitud nii, et normaalse genoomse DNA puhul seondub iga proov genoomis vaid ühte teadaolevasse kohta. Proovid on tavaliselt saadud märklaud-järjestuse kloneerimisega plasmiidsesse vektorisse ja nende paljundamisega sealt universaalsete praimeritega, mille tulemusena on kõikide proovide otstes universaalsele praimeripaarile vastavad järjestused. Et proove hiljem detekteerida, on üks universaalsetest praimeritest tavaliselt märgitud kas radioaktiivse isotoobiga või fluorofooriga. Pärast hübridiseerimist nailonmembraan pestakse korduvalt alaneva soola kontsentratsiooniga lahustega, et eemaldada sealt mitteseondunud ja mittespetsiifiliselt seondunud proovid. Membraanil oleva DNA-ga jäävad seotuks ainult need proovid, mis on seondunud spetsiifiliselt. Lühiajalise kuumutamisega vabastatakse seondunud proovid lahusesse ja paljundatakse universaalsete praimeritega (millest üks on märgitud kas radioaktiivse või fluorestseeruva märkega) *PCR*-i meetodiga. Amplifitseerimine toimub ainult *PCR* kvantitatiivse faasi (5-20 tsüklit) jooksul. Saadud produktid lahutatakse üksteisest ja proovide signaalitugevused määratakse geelelektroforeesil või kapillaarsekvenaatorit kasutades. Võrreldes uuritavatele aladele vastavate proovide signaalitugevusi kontrollproovide omadega, leitakse suhteline amplifikatsiooniprodukti hulk ehk DNA koopiaarv analüüsitavas lookuses (Armour *et al.*, 2000; Hollox *et al.*, 2002). *MAPH* meetodi põhietapid on toodud joonisel 3.



Joonis 3. Traditsioonilise multipleks amplifitseeritavate proovide hübridisatsiooni põhietapid. Uuritav DNA denatureeritakse ja seotakse nailonfiltrile, kus see hübridiseeritakse proovidega, mis seonduvad kindlasse lookusesse genoomis. Oma märklaud-järjestusele spetsiifiliselt seondunud proovid amplifitseeritakse unikaalsete praimeritega. Seejärel paljundatud proovid lahutatakse geel-elektroforeesil ja kvantiseeritakse. Proovi hulk võrreldes teadaolevate normaalsete lookuste proovide hulgaga näitab, kas proovidele vastavad lookused uuritavas DNA-s oli üle- või alaesindatud (Armour *et al.*, 2000).

Meetodi suurimaks puuduseks on see, et proovide pikkused peavad olema teineteisest piisavalt erinevad, et proove geelil identifitseerida. See seab omakorda piirangud korraga analüüsitavate lookuste arvule (Sellner & Taylor, 2004). Enamikus 2005 a. mai seisuga publitseeritud töödes on korraga analüüsitud 30-50 lookust (Armour *et al.*, 2000; Hollox *et al.*, 2002). Meetodi eelisteks teiste koopiaarvu määramise meetodite ees on kiirus ja odavus. Lisaks ka see, et meetod ei vaja eriaparatuuri olemasolu ja on hea spetsiifilisuse ja sensitiivsusega.

Lisaks traditsioonilisele *MAPH*-le on arendamisel ka mikrokiipidel põhinev multipleks amplifitseeritavate proovide hübridisatsioon (*array MAPH*). Kiipidel põhineva *MAPH*-i puhul detekteeritakse amplifitseeritud proovid kvantitatiivsete signaalide suhte alusel mikrokiip-platvormil (Hollox *et al.*, 2002). Kuni kindlaid lookusi

esindavate unikaalsete proovide amplifitseerimiseni toimub kõik sarnaselt tavalisele *MAPH* lähenemisele. Seejärel paljundatud proovid puhastatakse ja märgitakse fluorestseeruva märkega *nick* translatsioonil. Seejärel hübridiseeritakse proovid kiibile, mille kindlatesse positsioonidesse on suures ülehulgas seotud *MAPH* proovidele identsed järjestused. Nii seondub iga hübridiseerimisproov mikrokiibil justkui iseendaga. Kiibilt skaneerimisel saadavad fluorestseeruvad signaalid peegeldavad iga positsiooni ja seega ka uuritava DNA-ga seondunud proovi kvantitatiivset hulka. Võrreldes igale uuritavale lookusele vastavaid fluorestsentsintensiivsuste suhteid patsiendi ja kontrollindiviidi DNA-s, leitakse analüüsitavate järjestuste koopiaarvud uuritavas DNA-s. (Patsalis *et al*., käsikiri ettevalmistamisel).

2. Hübridiseerimisproovide disain

Nagu elusloodus, on DNA ahelate komplementaarsusel tähtis roll ka biotehnoloogias. Suur osa molekulaarbioloogia ja biotehnoloogia meetoditest (polümeraasi ahelreaktsioon, sekveneerimine Sangeri meetodiga, nukleiinhapete hübridiseerimine, mikrokiibid) põhinevad nukleiinhapete ahelate komplementaarsuse printsiibil. Kuna ühe- ja kaheahelaliste DNA molekulide kasutamine on molekulaarbioloogias ja eriti biotehnoloogias väga oluline, on ka tähtis aru saada DNA struktuuri mõjutavatest faktoritest.

2.1 DNA interaktsioonide biokeemia

Mõned faktorid on peaaegu alati DNA kaksikheeliksit stabiliseerivad ja mõned peaaegu alati DNA kaksikheeliksit destabiliseerivad. Stabiliseerivate jõudude hulka kuuluvad Watson-Crick'i vesiniksidemed ja lämmastikaluste *stacking*. Korrektse kaksikheeliksi moodustumise seisukohast on oluline ka aluste õige ruumiline paardumine. Viimase eelduseks on lämmastikaluste omavaheline paardumine korrektse geomeetriaga ehk tähtsat rolli mängivad ka aluste paardumise ruumilised efektid, millest räägitakse edaspidi. Destabiliseerivaks jõuks on negatiivsete fosfaatrühmade tõukumine üksteisest, mis toimub nii ühe ahela sisemiselt ja kaheahelalises DNA-s ka

kahe ahela vahel ning mis lahuses vähemalt osaliselt neutraliseeritakse katioonide poolt (Dimitrov & Zuker, 2004; Kool, 2001).

2.1.1 DNA ahelate seondumise ruumilised efektid

Korrektse kaheahelalise heeliksi moodustumise seisukohalt on kõige olulisem aluspaaride geomeetria. G-C ja A-T paarid on oma keemiliste struktuuride ja paardumise geomeetria poolest isomorfsed: kaugused glükosiidsete lämmastike vahel on mõlemas paaris võrdsed. DNA, mis koosneb täielikult Watson-Crick'i aluspaaridest, moodustab paardunud ahelate kogu ulatuses regulaarse struktuuriga spiraali, sõltumata sellest, mis järjekorras nukleotiidid ahelates on. Kuna ainult korrektselt paardunud Watson-Crick'i aluspaarid säilitavad geomeetrilise struktuuri, mis on iseloomulik kaheahelalisele DNA-le, on õigete aluste omavaheline seondumine eelistatud. See asjaolu võimaldab hoolimata nukleotiidide järjestusest ahelas säilitada kaksikheeliksi korrektset suhkur-fosfaatselgroogu. Kahe komplementaarse ahela aluspaarid hakkavad kooperatiivselt seonduma, kui ahelate vahel on tekkinud 3-aluspaariline nukleatsiooni tsenter. Ahelad keerduvad teineteise ümber ja astmeliselt, nn „tõmbluku" (*zipper*) sarnaselt, moodustub kaksikheeliks (Saenger, 1984). Arvatakse, et kuni 20 nukleotiidi pikkuste ahelate seondumine toimub kahe-astmeliselt (kokku-lahku), kuid pikemate ahelate seondumine on mitme-astmeline, sisaldades ka vahepealsete (osaliselt koos-osaliselt lahus) vormide moodustumist (Dimitrov & Zuker, 2004; SantaLucia *et al.*, 1996). Pikkades ahelates arvatakse esmalt seonduvat GC-rikkad regioonid ja seejärel ülejäänud järjestused (*random sequence*) (Saenger, 1984).

Kui omavahel seonduvad mittekanoonilised nukleotiidid (*mismatch'*id - ei teki Watson-Crick'i aluspaarid), vähendavad need DNA kaksikheeliksi lokaalse ja summaarse seondumise tugevust. Seondumise tugevust vähendab tõenäoliselt see, et vale aluspaari puhul on häiritud nii vale aluspaari enda kui ka kõrvalasuvate aluspaaride standardsed geomeetriad kaksikheeliksis. On leitud, et tõenäoliselt ei tule *mismatch'*idest tulenev väiksem summaarne seondumisenergia mittestandardsetest vesiniksidemetest või vesiniksidemete puudumisest, sest paljud või isegi enamus valedest aluspaaridest ehk *mismatch'*idest on omavahel seotud vesiniksidemetega (Kool, 2001; SantaLucia & Hicks, 2004). On pakutud, seondumisenergia vähenemine tuleb asjaolust, et *mismatch'*itud nukleotiidide lämmastikaluste *stacking* oma

naaberalustega on häiritud ja erinev tavapärasest (Kool, 2001). Samuti on näidatud, et kaksikheeliksi otsas olev *mismatch* on kaksikheeliksit stabiliseeriv (-1.23…-0.21 kcal/mol) (SantaLucia & Hicks, 2004), kuigi väiksema energiaga, kui õige paardumise puhul (SantaLucia, 1998; SantaLucia & Hicks, 2004).

2.1.2 Vesiniksidemed

Vee keskkonnas moodustavad üheahelaliste nukleiinhapete lämmastikalustega vesiniksidemeid vee molekulid. Sellised sidemed vee molekulidega on väga lühiajalised ja vahetuvad väga kiiresti. Seetõttu on lahustunud DNA seisukohalt kahe DNA ahela vahelised vesiniksidemed soodustatumad, kuna kahe DNA ahela seondumisel tekkivad aatom-aatom interaktsioonid on optimeeritumad ja vesiniksidemed korrastatumad kui ühe ahela ja lahusti molekulide vaheliste interaktsioonide korral (Lane & Jenkins, 2000).

Kahe DNA ahela vahel olevad vesiniksidemed moodustuvad vastakuti sobivalt paiknevate nukleotiidide aluste vesiniksidemete doonorite ja aktseptorite vahel. Vesiniksidemete doonorid on lämmastikaluste amino- (NH_2) ja iminorühmadega (NH) kovalentselt seotud vesinikud. Vesiniksidemete aktseptorid on aluste hapniku aatomid ja aromaatses ringis olevad lämmastiku aatomid. Aktseptori kahe elektroniga orbitaalid ja doonori peaaegu vabad orbitaalid kattuvad, moodustades nukleotiidide lämmastikaluste vahele vesiniksidemed. Guaniini ja tsütosiini vahel moodustub 3 ja adeniini ja tümiini vahele 2 vesiniksidet.

2.1.3 Lämmastikaluste *stacking*

Lisaks vesiniksidemetele stabiliseerib nukleiinhapetes ahelaid ka lämmastikaluste *stacking*. Lämmastikaluste *stacking'*u all mõistetakse seda, et kaks aromaatset molekuli on geomeetriliselt paigutunud nii, et üks on teise kohal ja nad on seotud mittekovalentsete jõududega, mis otseselt soodustavad sellist seondumise geomeetriat energeetiliselt. *Stacking* on kompleksne mittekovalentne interaktsioon, sõltudes erinevatest jõududest, mis tõenäoliselt kõik mängivad tähtsat osa. On pakutud,

et *stacking* aluste vahel tekib kolme faktori mõjul: hüdrofoobne efekt, elektrostaatilised jõud ja Van der Waals'i jõud (Kool, 2001; Lane & Jenkins, 2000).

DNA teadaolevatest erinevatest struktuuridest enamuse puhul on kõrvutiolevate nukleotiidide lämmastikalused paralleelselt üksteisega ja osaliselt ka kohakuti. DNA B-vormis ei ole lämmastikalused täpselt üksteise kohal, vaid üksteise suhtes natuke nihkes. Selline aluste üksteise suhtes nihkes olemine võib olla eelistatud DNA suhkur-fosfaatselgroo sobiva konformatsiooni tõttu ja ka aluste endi elektrostaatiliste efektide tõttu. DNA B-vormi puhul on kahe aromaatse tasapinna vaheline kaugus ca 3,4 Ångstromi, mis vastab DNA B-vormi tõusule ühe aluspaari kohta. *Stacking'*u uurimisel on leitud, et lämmastikaluste *stacking* DNA-s mõjub stabiliseerivalt kaksikheeliksile ja võib isegi olla dominantne kaksikheeliksit kooshoidev jõud (Bommarito *et al.*, 2000; Kool, 2001; Lane & Jenkins, 2000). *Stacking'*u stabiliseerivat mõju näitab juba seegi, et ühe üleulatuva paardumata nukleotiidi olemasolu kaksikahela otsas stabiliseerib kaheahelalist DNA-d vabaenergiaga (ΔG) kuni -0.98 kcal/mol (Bommarito *et al.*, 2000; SantaLucia & Hicks, 2004).

*Stacking'*u uurimisel on näidatud, et üheahelalised nukleiinhappe molekulid ei moodusta lahuses korrapäratuid struktuure (*random coil*), vaid pigem korrapäraseid, osaliselt *stacking'*uga seotud spiraalseid struktuure. Sellistes struktuurides on *stacking'*u määr on oluliselt sõltuv nii temperatuurist kui ka ahela järjestusest ja lahusest, milles DNA on lahustatud (Dimitrov & Zuker, 2004; Gellman *et al.*, 1996; Lane & Jenkins, 2000). On näidatud, et *stacking'*u määr on suurem polaarsetes lahustes nagu vesi ja väiksem orgaanilistes lahustes nagu metanool ja kloroform (Norberg & Nilsson, 1998). Samuti on näidatud seda, et polüpuriinsetes oligonukleotiidides on rohkem aluseid omavahel *stacking'*us kui polüpürimidiinsetes oligonukleotiidides (Asensio *et al.*, 1998), kuigi ka viimastes on täheldatud märkimisväärset aluste *stacking'*ut. Üheahelalises DNA-s toimub *stacking'*uga seotud struktuuride pidev muutumine – *stacking* tekib ja kaob erinevate naabrite vahel. See toob kaasa selliste korrastatud struktuuride pikkuse pideva muutumise. Üheahelalise DNA sellisele agregeerumisele aitab kaasa see, et suhkur-fosfaatselgroog seob omavahel kõrvutiolevaid alused üksteisest õigele kaugusele, vähendades sellega ebasoodsat entroopiat, mis takistaks kõrvutiolevate aluste omavahelist interakteerumist DNA-s.

Uurijad on näidanud, et erinevate lämmastikaluste *stacking* oma naaberalusega on erineva tugevusega. Nii on näiteks leitud, et puriinide omavaheline *stacking* on tugevam kui pürimidiinidel, seda arvatavasti sellepärast, et neil on suurem väline pind ja

rohkem polaarseid rühmi (Bommarito *et al.*, 2000). Üldiselt võib üheahelalise nukleiinhappe dinukleotiidide (kõrvutipaiknevate nukleotiidide) *stacking'*u tugevuse järjestada järgnevalt: RR > YR > RY > YY, kus R tähistab puriini ja Y pürimidiini. Eranditeks on GC dinukleotiidi tugevaim *stacking* ja GG dinukleotiidi küllaltki nõrk seondumine (Kool, 2001). Huvitav on asjaolu, et DNA-s on *stacking* tugevam 3' suunas, st. aluse seondumine endast 3' suunas olevale alusele on tugevam, kui selle seondumine endast 5' suunas olevale alusele (Bommarito *et al.*, 2000; Kool, 2001). RNA-s, vastupidiselt DNA-le, on näidatud, et *stacking* on tugevam 5' suunas (Freier *et al.*, 1985; Kool, 2001).

2.2 Koopiaarvu määramisel kasutatavad hübridiseerimisproovid

Koopiaarvu määramisel kasutatakse hübridiseerimisproovidena mitmeid erinevalt saadud järjestusi, mis võivad olla nii sünteetilised kui *in vitro* paljundatud. Viimastel aastatel on koopiaarvu määramisel hübridiseerimisproovidena kasutatud *PCR*-i produkte (Buckley *et al.*, 2002) ja sünteetilisi oligonukleotiide (Kane *et al.*, 2000; Reymond *et al.*, 2004).

PCR-i produktid, mida kasutatakse hübridiseerimisproovidena, on enamasti 100-1000 bp pikad ja paljundatud tavaliselt kas genoomsete järjestuste kloonidelt või otse genoomselt DNA-lt. Hübridiseerimisproovidena kasutatavate *PCR*-i produktide eeliseks teiste proovide ees on see, et nad on pikemad, annavad intensiivsema signaali ja neil on rohkem seondumiskohti (Stillman & Tonkinson, 2001).

Oligonukleotiidsed proovid, mida kasutatakse hübridiseerimisproovidena on enamasti 25-75 nukleotiidi pikad ja saadud keemilise sünteesi tulemusel. Nende suureks eeliseks on see, et erinevalt *PCR*-i produktidest saab neid disainida väga spetsiifiliselt, st. praktiliselt unikaalsetena, mis võimaldab viia proovide rist-hübridiseerumise (*cross-hybridization*) miinimumini (Kane *et al.*, 2000; Zhou, 2003).

2.3 Teadaolevad proovide hübridiseerumist mõjutavad tegurid

Hübridiseerimisproovide disain ja spetsiifilisus on kriitilise tähtsusega nii traditsiooniliste kui ka kiipidel läbiviidavate hübridiseerimismeetodite kvaliteetseks teostamiseks.

Erinevad hübridiseerimismeetodid esitavad hübridiseerimisproovidele väga erinevaid nõudeid, mis on seotud konkreetse meetodi eripära ja tingimustega, kuid põhilised parameetrid on sarnased. Kõige tähtsamaks proovi omaduseks võib pidada tema unikaalsust uuritavas DNA-s. Tähtsateks proovide omadusteks on ka proovide pikkus, GC nukleotiidide sisaldus, sekundaarstruktuuride moodustamine ja sulamistemperatuur (Benita *et al.*, 2003).

On näidatud, et kui proov sisaldab regiooni, mis on 50 nukleotiidi ulatuses üle 75% identne mõne teise prooviga või regiooniga uuritavas genoomis, võib see viia proovi rist-hübridiseerumisele ehk seondumisele valel järjestusel. Kui proov on marginaalselt (50-75%) identne mõnele teisele samas katses kasutatavale järjestusele, ei tohi see proov sisaldada üle 15 nukleotiidi pikkusi mõne teise järjestusega 100% identseid alasi (Kane *et al.*, 2000). Sarnaseid tulemusi on saadud ka pikemate järjestustega tehtud katsetes (Hughes *et al.*, 2001).

Hübridiseerimisproovide pikkus on erinevate meetodite vahel kõige enam varieeruv parameeter, igale meetodile on omane optimaalne proovi pikkus. Mikrokiipidel läbiviidavatel katsetel peaksid ilma speisseriteta (*spacer*) mikrokiibile seotud proovid olema vähemalt 60 nukleotiidi pikad, kuna lühemad järjestused ei ole kahemõõtmelisele kiibile seotuna piisavalt ligipääsetavad ega taga piisavat hübridiseerumise efektiivsust (Hughes *et al.*, 2001). On näidatud, et pikemad proovid (üle 100 nt) annavad hübridiseerimisel intensiivsema signaali (Relogio *et al.*, 2002) ning rohkemate proovide ja sihtmärk-järjestuste seondumise, kuna neil on rohkem seondumiskohti (Stillman & Tonkinson, 2001).

Vastavalt kasutatavale tehnoloogiale valitakse tavaliselt ka proovide saamise ja paljundamise meetod, mis seab omakorda piirangud proovide pikkusele. Sünteetilised hübridiseerimisproovid on tavaliselt 25-70 nukleotiidi pikad (Chen *et al.*, 2002; Kane *et al.*, 2000), samas kui *PCR*-i produktid, mida kasutatakse hübridiseerimisproovidena, on 100-3000 bp pikad (Tõnisson *et al*., 2000). Ühe katse raames kasutatakse tavaliselt

küllaltki kindlas pikkuse vahemikus olevaid proove, et tagada kõikide proovide ühetaoline käitumine katses.

Hübridiseerimisproovide GC nukleotiidide sisaldus näitab G (guaniini) ja C (tsütosiini) nukleotiidide osakaalu järjestuses. On näidatud, et paljude lühikeste GC rikaste regioonide sisaldumine proovis mõjutab negatiivselt proovi- ja märklaud-järjestuse hübridiseerumist, soodustades mittespetsiifilist rist-hübridiseerumist. See tuleneb arvatavasti DNA seondumise termodünaamikast - GC-rikkad regioonid seonduvad kõigepealt (Rouillard *et al.*, 2003). Kui kasutatav hübridiseerimise meetod sisaldab ka proovide paljundamist *PCR*-i meetodil (*MAPH*, *array-MAPH*), on oluline meeles pidada ka seda, et proovid ei tohiks sisaldada väga kõrge või väga madala GC sisaldusega regioone, kuna on näidatud, et sellised piirkonnad on raskelt amplifitseeritavad (Benita *et al.*, 2003).

Hübridiseerimisproovide disainimisel on oluline jälgida, et proovid ei annaks stabiilseid sekundaarstruktuure ei iseendaga ega ka teiste proovidega. Sekundaarstruktuuride tekke võimalikkus proovis võib viia selle seondumisele iseendaga (moodustub nn. *stem-loop* struktuur) või teiste hübridiseerimisproovidega (*slipped* struktuur). Kuigi iseendaga seondumine on enamasti nõrgem kui seondumine oma sihtmärk-järjestusega, võib proovide sekundaarstruktuuride tekkimine viia ebakorrektsete tulemusteni katses.

Sulamistemperatuur (*melting temperature,* T_m) on temperatuur, mille juures pooled DNA molekulidest lahuses on denatureerunud (ahelad on teineteisest lahus) ja pooled mitte (ahelad on kaksikheeliksina). Sulamistemperatuur on oluline, kuna vastavalt sellele arvutatakse välja proovi ja tema sihtmärk-järjestuse seondumise energia teisendatuna temperatuuriks (*annealing temperature*). Tegelikust erineva seondumistemperatuuri kasutamine võib kaasa tuua proovide spetsiifilisuse vähenemise. Sulamistemperatuur arvutatakse omavahel paardunud nukleiinhapetele välja vastavalt kindlatele valemitele. Kõige lihtsam valem sulamistemperatuuri arvutamiseks on toodud järgnevana (Suggs *et al*., 1981):

$$\mathbf{T_m = 4 * (G + C) + 2 * (A + T),} \qquad \mathbf{(1)}$$

kus G ja C on guaniini ja tsütosiini ning A ja T adeniini ja tümiini nukleotiidide arv.
Täpsemad valemid hübridiseerimisproovide sulamistemperatuuri väljaarvutamiseks arvestavad ka mono- ja divalentsete soolade kontsentratsiooni hübridisatsioonilahuses,

kuna on näidatud, et nii mono- kui divalentsed katioonid mõjutavad hübridiseerimisproovide sulamistemperatuuri, stabiliseerides lahuses DNA kaksikheeliksit (Ahsen *et al.,* 2001). Üheks selliseks valemiks on erinevate autorite poolt välja pakutud Marmur-Schildkraud-Doty valem, mis arvestab monovalentsete soolade sisaldust lahuses ja sobib ka pikemate proovide sulamistemperatuuri arvutamiseks (Marmur & Doty, 1962; Schildkraut, 1965):

$$\mathbf{T_m = 81.5 + 16.6 * \log_{10}([Na^+]) + (0.41 * GC\%) - (b / l),} \quad \mathbf{(2)}$$

kus $\log_{10}([Na^+])$ on kümnendlogaritm monovalentsete soolade kontsentratsioonist, GC% on G ja C nukleotiidide osakaal proovis, b on konstant (erinevad autorid on kasutanud b-na väärtusi vahemikus 500-750) ja l on proovi pikkus.

Kõige täpsemaks sulamistemperatuuri arvutamise meetodiks on termodünaamilisel *Nearest-Neighbour* mudelil põhinev valem (Borer *et al.*, 1974; SantaLucia, 1998):

$$\mathbf{T_m = \Delta H / (\Delta S + R * \ln C_T/4),} \quad \mathbf{(3)}$$

kus ΔH on entalpia muut proovi ja märklaud-järjestuse seondumisprotsessis, ΔS on entroopia muut samas protsessis, R on universaalse gaasi konstant (R = 8,31 J / K * mol = 1,987 cal / K * mol) ja C_T on hübridiseerimisproovide kontsentratsioon lahuses. Mõned autorid soovitavad *Nearest-Neighbour* mudelit kasutada vaid kuni 50-meersete proovide sulamistemperatuuri arvutamiseks, kuna pikemate proovide puhul on sulamistemperatuurid mõnevõrra väiksemad *Nearest-Neighbour* mudeli poolt eeldatavast (Haas *et al.*, 2003; Rouillard *et al.*, 2003). Erinevad töögrupid on kõigile kümnele Watson-Crick'i paardumisega dinukleotiidile välja arvutanud *Nearest-Neighbour* mudelile vastava entalpia muudu (ΔH), entroopia muudu (ΔS) ja seondumisenergia (ΔG), mida saab kasutada proovide täpsete sulamistemperatuuride väljaarvutamiseks, kasutades valemit nr. 3 (SantaLucia, 1998; SantaLucia & Hicks, 2004; Sugimoto *et al.*, 1996). Üheskoos hübridiseeritavad proovid peaksid olema sarnase sulamistemperatuuriga, et tagada nende ühtlane käitumine katsetes (Reymond *et al.*, 2004).

3. Suured genoomid ja kordusjärjestuste vältimine

On teada, et suure osa eukarüootsetest genoomidest moodustavad kordusjärjestused. Erinevate imetajate genoomides moodustavad kordusjärjestused isegi kuni pool kogu järjestustest (Makalowski, 2000). Erandiks ei ole siinkohal ka inimene, kelle genoomist 49.9% moodustavad kordusjärjestused (Makalowski, 2001). Seetõttu on erinevate proovide disainimisel tähtis silmas pidada, et disainitavad proovid ei satuks kordusjärjestustele, mis toob kaasa proovide spetsiifilisuse kadumise ja seetõttu näiteks *PCR*-il valede regioonide paljundamise või hübridiseerimisel valede järjestuste seondumise sihtmärk-järjestusele (*target*).

3.1 Inimese kordusjärjestuste iseloomustus

Ligi poole inimese genoomse DNA järjestusest hõlmavad erineva pikkusega ja korduste arvuga kordusjärjestused, mis enamasti paiknevad geenidevahelistes regioonides (Makalowski, 2000). Kordusjärjestused jaotatakse üldiselt kaheks: tandeemseteks kordusteks (*tandem repeats*) ja hajutatud korduselementideks (*interspersed repetitive elements*).

Tandeemsed kordusjärjestused esinevad genoomis suurte korduvate blokkidena ja jaotatakse enamasti kolme klassi. Kõige lühemad, mikrosatelliidid (*short tandem repeat, STR*) koosnevad lühikestest (1-4 bp) ja lihtsatest kordustest (kõige levinum on CA dinukleotiidide kordus), mis esinevad kuni mõnesaja bp pikkuste kordustena (Csink & Henikoff, 1998; Makalowski, 2000; Makalowski, 2001). Minisatelliidid (*variable number of tandem repeats, VNTR*) on tandeemselt korratud 6-64 bp pikkuste järjestuste (kõige levinumad mustrid on TTAGGG ja GGGCAGGANG) 1-15 kb suurused blokid, mis paiknevad kõikide kromosoomide telomeerides ja nende lähedal (Makalowski, 2001). Kõige pikemad, makrosatelliidid on 5-200 bp pikkused järjestused, mis moodustavad kuni mõnesaja kb pikkuseid tandeemseid korduseid ja esinevad kromosoomides tsentromeeride lähedastes regioonides (Csink & Henikoff, 1998; Makalowski, 2001).

Hajutatud korduselemendid paiknevad üksikult ja hajutatult üle kogu inimese genoomi. Ka hajutatud korduselemendid jaotatakse omakorda klassideks, mis on ära toodud tabelis 1.

Tabel 1. Hajusate korduselementide klassid inimese genoomis. Inimesel esinevate hajusate kordusjärjestuste klassid ja erinevate hajuskorduste esinemiste arvud ning osakaalud kogu inimese genoomi järjestusest (Lander *et al.*, 2001).

Klass	Koopiate arv (x1000)	Kogupikkus genoomis (bp)	Osa genoomist (%)
SINE elemendid	1558	359.6	13.14
LINE elemendid	868	558.8	20.42
LTR elemendid	443	227	8.29
DNA transposoonid	294	77.6	2.84
Klassifitseerimata elemendid	3	3.8	0.14

Nagu tabelist 1 näha, on kõige levinumad hajusad kordusjärjestused inimese genoomis SINE (*short interspersed repetitive elements*) ja LINE (*long interspersed repetitive elements*) elemendid, mis kokku võtavad enda alla kolmandiku inimese genoomist. Kõige levinum SINE element on primaatide spetsiifiline Alu element (pikkus ~280 bp), mida inimesel arvatakse olevat 500-900 tuhat koopiat, moodustab inimese kogu genoomist 10.6% (Lander *et al.*, 2001; Makalowski, 2001). Kõige levinum LINE element on L1 (*LINE1*), mille pikkuseks on ligikaudu 6.1 kb ja mis võtab kogu inimese genoomist enda alla koguni 16.89% (Lander *et al.*, 2001).

3.2 Kordusjärjestuste maskeerimine

Et kiirendada suurte genoomide puhul proovide disainimiseks ja unikaalsuse kontrollimiseks kasutatavate programmide tööd, kasutatakse järjestuste maskeerimist kordusjärjestuste suhtes. Proovide disainimine maskeeritud järjestustelt vähendab proovide sarnasuse tõenäosust kordusjärjestuste suhtes ja seega suurendab nende unikaalsust. Levinumad järjestuste maskeerimise programmid on DUST ja RepeatMasker. Programm DUST (*ftp://ftp.ncbi.nih.gov/pub/tatusov/dust*) maskeerib etteantud DNA järjestuses madala keerukusega kordusjärjestused: polü-N traktid, di-, tri- ja tetranukleotiidsed kordused. RepeatMasker (*http://www.repeatmasker.org*) on programm, mis maskeerib etteantud järjestustes spetsiaalse andmebaasi (RepBase, *http://www.girinst.org*) abil kõik inimese genoomis teadaolevad kordusjärjestused (Bedell *et al.*, 2000).

3.3 Programmid unikaalsuse tuvastamiseks

Et saada kvaliteetseid hübridiseerimisproove, on enne praktilisi katseid vaja *in silico* leida kõikide proovide võimalikud seondumiskohad uuritavas genoomis. See aitab tagada proovide unikaalsuse ja vältida nende rist-hübridiseerumist. Samas on vaja kontrollida, et ühes katses koos kasutatavad proovid ei seonduks üksteisega.

Selleks, et tagada hübridiseerimisproovide unikaalsus genoomis, on arendatud mitmeid erinevaid programme, mis *in silico* otsivad etteantud proovide seondumiskohti genoomis. Tuntumateks sellisteks programmideks on SSAHA, BLAST2 ja MegaBLAST.

SSAHA (*Sequence Search and Alignment by Hashing Algorithm*) programmi algoritm kasutab metoodikat, mille puhul andmebaas (näiteks inimese genoom kromosoomide kaupa), millest hakatakse kasutajat huvitavat järjestust (*query sequence*) otsima, kõigepealt indekseeritakse ja kirjutatakse ühte suurde paisktabelisse (*hash*). See on aeglane, kuid see-eest ühekordne protsess. Kui programm hakkab kasutaja poolt etteantud järjestusi võrdlema andmebaasiga, loetakse enne valmistehtud paisktabel mällu ja järjestuse otsimine ise toimub juba kiiresti. SSAHA otsib etteantud järjestuse ja indekseeritud paisktabeli vahel kasutaja poolt kindlaksmääratud sõnapikkusega identseid (100% kokkulangevaid) järjestusi. Programmi töö tulemusena leitakse järjestused, mis on andmebaasi suhtes unikaalsed (Ning *et al.*, 2001).

BLAST2 (*Basic Local Alignment Search Tool, version 2*) ehk *Gapped BLAST* programm kasutab sarnasuse leidmiseks päring-järjestuse ja andmebaasi vahel heuristilisi algoritme. Kui kasutaja poolt etteantud järjestus ja mõni andmebaasi järjestus mingis osas ühtivad, arvutatakse sellele sarnasele osale skoor (*hits*) ja kui see ületab kasutaja poolt määratud piiri, tulemus väljastatakse. BLAST2 suureks eeliseks on asjaolu, et see oskab otsida ka omavahel nihkes (*gap*) olevaid sarnaseid järjestusi. Seda tehes võetakse sarnasuse otsimisel arvesse ka väiksemaid erinevusi järjestustes (näiteks aja jooksul toimunud mutatsioone), mis arvestades järjestuste bioloogilist päritolu on ainuõige (Altschul *et al.*, 1997).

MegaBLAST on programm, mis võrdleb päring-järjestuse ja andmebaasi järjestuste sarnasuse leidmiseks nende erinevust üksteisest. Programm kasutab ahnet (*greedy*) algoritmi, mis võimaldab kahe järjestuse võrdlemisel kõrvale jätta ebatõenäolised muutused (*gap'*id) nendes järjestustes. Seetõttu sobib MegaBLAST rohkem selliste järjestuste otsimiseks, millel on küllaltki suur oodatav sarnasus

andmebaasi järjestustega. Ahne algoritm võimaldab aga oluliselt vähendada programmi ajalist keerukust. Kasutades sobivaid järjestusi, võib päring-järjestusele sarnased regioonid andmebaasis leida kuni 10 korda kiiremini kui kõige lihtsamat identsuse otsingut - dünaamilist programmeerimist (*dynamic programming*) kasutades (Zhang *et al.*, 2000). Seega sobib MegaBLAST hästi ka pikkade päring-järjestuste võrdlemiseks andmebaasiga.

II PRAKTILINE TÖÖ

Töö eesmärgid

Üha enam tegeleb funktsionaalne genoomika inimese genoomsetes järjestustes toimunud muutuste (aberratsioonide) ja geneetiliste haiguste omavaheliste seoste uurimise ja diagnostikaga. Seetõttu on vajalik ka spetsiifiliste bioinformaatiliste lahenduste väljatöötamine ja *in silico* rakendamine.

Käesoleva töö eesmärgiks oli välja töötada metoodika hübridiseerimisproovide automaatseks disainimiseks ja veebipõhise kasutajaliidese loomine hübridiseerimisproovide disainimiseks inimese genoomsetele järjestustele. Lisaks oli käesoleva töö eesmärgiks uurida, millised hübridiseerimisproovide omadused võiksid mõjutada pikkade (200-600 bp) hübridiseerimisproovide signaali intensiivsust ja signaali intensiivsuse varieeruvust mikrokiipidel põhinevates katsetes.

Lahendusena töötati välja unikaalsel metoodikal põhinev algoritm ja programmid hübridiseerimisproovide automaatseks disainimiseks. Programmide lihtsamaks kasutamiseks loodi veebipõhine kasutajaliides MAPHDesigner, millega saab automaatselt disainida *PCR*-i produktidena paljundatavaid hübridiseerimisproove mikrokiipidel läbiviidavate *MAPH* (*multiplex amplifiable probe hybridization*) ja *CGH* (*comparative genomic hybridization*) meetodite jaoks.

1. Kasutatud meetodid

1.1 Lähteandmete päritolu ja struktuur

Töös kasutatud nukleotiidsed järjestused ja nende andmed on saadud ENSEMBL-i avalikust andmebaasist (*ensembldb.ensembl.org*), kust kohalikku serverisse on alla laetud inimese genoomi järjestuse versioon nr. 35. Järjestused on salvestatud kohalikku MySQL andmebaasi *core 35*. Järjestuste annotatsiooniandmed on saadud ENSEMBL-i, NCBI ja VEGA annoteerimiskeskustest ning samuti salvestatud kohalikku MySQL-i andmebaasi *human 35* (Triinu Kõressaar, TÜ MRI, bioinformaatika õppetool). Järjestuse andmetest on kasutatud inimese genoomi nukleotiidseid järjestusi, mis on saadud *core* 35 andmebaasist kujul *kromosoom, alguspositsioon, järjestus.* Annotatsiooniandmetest on kasutatud geenide süstemaatilisi ja sünonüümseid nimesid ning geenide algus- ja lõpp-koordinaate, mis on kättesaadavad *human 35* andmebaasist kujul *geeni nimi, alguspositsioon, lõpp-positsioon.*

Töös kasutatud mikrokiibi signaalide toorandmed (*raw data*) pärinevad inimese X kromosoomi spetsiifiliselt mikrokiibilt, mis on saadud TÜ Molekulaar- ja Rakubioloogia Instituudi biotehnoloogia õppetoolist, dr. Ants Kure töögrupilt.

1.2 Kasutatud riistvara

Hübridiseerimisproovide omaduste arvutamisel kasutati serverit MicroLink Novator 5000HG. Server on kaheprotsessoriline (2 x 2.66 GHz Intel Xeon), 6 GB (*gigabait*) operatiivmäluga ja viie Maxtor Atlas RAID 0 10K Ultra320 SCSI (147 GB) kõvakettaga. Serveri operatsioonisüsteemiks on Mandrakelinux 9.2. Lisaks kasutati parameetrite arvutamisel Eesti Biokeskuse arvutiklastrit (16 arvutit), kus olevatel arvutitel on Mandrakelinux 9.2 operatsioonisüsteem, 2.8 GHz Intel Pentium 4 HT protsessor ja 1 GB operatiivmälu. Proovide automaatseks disainimiseks loodud programmid ja veebi-kasutajaliides MAPHDesigner töötavad Mandrakelinux 10.1 operatsioonisüsteemiga kaheprotsessorilises (2 x 800Mhz Intel Pentium III) ja 2 GB operatiivmäluga serveris.

1.3 Kasutatud programmid ja nende käivitamise põhimõte

Proovide omaduste arvutamiseks kasutati selleks käesoleva töö raames programmeerimiskeeles Perl (*Practical Extraction and Reporting Language,* http://www.perl.org) kirjutatud programme. Programm *probe_param.pl* sai sisendiks hübridiseerimisproovide nukleotiidsed järjestused ja arvutas igale järjestusele selle pikkuse, GC nukleotiidide sisalduse (GC%) ja sulamistemperatuuri kahe erineva valemi järgi. Lisaks kasutati proovide omaduste arvutamiseks programmeerimiskeeles Python (http://www.python.org) kirjutatud programmi *get_auc.py* ja teiste autorite poolt C keeles kirjutatud programme *fastagrep* (Lauris Kaplinski, TÜ MRI, bioinformaatika õppetool), SSAHA (Ning *et al.*, 2001) ja MegaBLAST (Zhang *et al.*, 2000).

Programmiga *get_auc.py* kutsuti välja programmi *AUC.py* (Benita *et al.*, 2003) alamprogramm, mis arvutas igale proovile vajalikud parameetrid. Programmiga *mmgc.pl* arvutati iga proovi jaoks proovi lokaalne (30 nukleotiidi pikkuses aknas oleva järjestuse) minimaalne ja maksimaalne GC nukleotiidide sisaldus, liikudes aknaga 1 nukleotiidse sammuga üle proovi järjestuse.

Programmi DUST modifitseeritud versioonidega *dust_lower* ja *dust_score* arvutati igale proovile DUST'i skoor, mis näitab, kui palju on proovi järjestuses madala keerukusega kordusjärjestusi.

Programmiga *jupita.pl* tehti igale proovile etteantud pikkusega (15, 20, 25, 30, 35, 40, 45 ja 50 nukleotiidi) alamstringid (järjestuste osad), liikudes proovi järjestusel edasi ühe nukleotiidi kaupa. Igale proovile vastavaid alamstringe saadi seega iga erineva pikkuse kohta: proovi_pikkus - alamstrigi_pikkus + 1. Näiteks 424 bp pikkusele proovile tehti 424 - 15 + 1 = 410 15-ne nukleotiidi pikkust alamstringi.

Programmiga SSAHA kontrolliti, millistesse regioonidesse tehtud alamstringid seonduvad oma järjestuse kogu ulatuses. Programm SSAHA käivitati võtmetega *-gf fasta* (täpsustab päring-järjestuste formaadi), *-sf hash* (täpsustab, millise andmebaasi formaadiga päring-järjestusi võrreldakse), *-pf* (väljastab parserile sobiva formaadi) ja *-mp N* (minimaalne identsete aluspaaride arv päringu ja andmebaasi vahel). Programmiga MegaBLAST kontrolliti, millistesse regioonidesse alamstringid seonduvad vähemalt 75%-lise identsusega. Seda tehti samuti otsides proovide seondumisi inimese genoomis kromosoomide kaupa. Programm MegaBLAST käivitati võtmetega *-F F* (lülitab välja lihtsaid kordusi maskeeriva programmi DUST filtri), *-D 2*

(väljastab parserile sobiva formaadi), *-W 12* (täpsustab kasutatava sõnapikkuse), *-e N* (täpsustab oodatava seondumiste arvu).

Alamstringide seondumised loeti kokku programmiga *loe_seond.pl*. Igale proovile leiti kõige rohkem kordi seondunud alamstringid ja nende seondumiste arvud, mis on lisas 2 esitatud kujul S_N (SSAHA-ga leitud 100%-ed seondumised) ja MB_N (MegaBLAST-ga leitud seondumised 75% ulatuses), kus N on alamstringide pikkus.

Lisaks identsusel põhinevatele seondumiskohtadele otsiti ka proovide alamstringide termodünaamilisi seondumisi. Termodünaamilise tugevuse alusel otsimise teeb keeruliseks asjaolu, et sama tugevusega seonduvate alamstringide pikkused võivad olla isegi kuni 3 korda erinevad. Programmid *fgrep.pl* ja *deltaG.pl* leidsid proovidest kõik alamstringid, mis seondusid ilma *mismatch'*ideta etteantud *ΔG* tugevusega. Seejuures kutsuti programmi *fgrep.pl* poolt järjestuse seondumistugevuse (*ΔG*) arvutamiseks välja programm *fastagrep*, mis käivitati võtmega *-dglist G*. Selle tulemusena printis programm *fastagrep* välja etteantud järjestuse kõik alamjärjestused (-stringid), mille perfektse seondumise vabaenergia (*ΔG*) on väikseim võimalik, kuid suurem kui väärtus G. Programmiga *shortest_slice.pl* leiti proovide alamstringidele paikapandud seondumistugevusele vastavad lühimad alamstringid. Sama seondumistugevusega alamstringid jaotati pikkuse järgi gruppidesse ja programmiga SSAHA leiti nende seondumiskohad inimese genoomis. Seejärel loeti programmiga *loe_seond.pl* kokku kõikide alamstringide seondumiskohad inimese genoomis. Kuna lühemad alamstringid andsid inimese genoomis väga palju seondumisi (728703), toimus seondumiste kokkulugemine kromosoomide kaupa. Programmiga *max_seond.pl* leiti iga proovi jaoks jällegi kõige rohkem seondumiskohti andnud alamstring, mille seondumiste arv on lisas 2 antud kujul DG_N, kus N on alamstringi seondumisenergia. Programmiga SSAHA otsiti proovide seondumiskohti inimese genoomis. Kuna terve inimese genoom oli korraga operatiivmällu lugemiseks liiga suur, otsiti proovide seondumiskohti kromosoomide kaupa.

Mikrokiipide andmete normaliseerimiseks kasutati kahte programmi. Mediaani järgi signaalide normaliseerimine tehti tabelarvutusprogrammiga Microsoft Office Excel 2003 (Microsoft Corporation, Redmond, WA, USA). Kolmandat järku polünoomiga normaliseerimine tehti kasutades statistikapaketti SAS 8.2 (SAS Institute Inc., Cary, NC, USA) Normaliseeritud signaalide teisendamiseks ja analüüsimiseks kasutati Perl-is kirjutatud programme *mean.pl, kokku.pl ja sg.pl.*

Signaali tugevuse ja varieeruvuse ning proovide omaduste vaheliste seoste otsimisel ja analüüsimisel kasutati statistikapaketti SAS 8.2 ja tabelarvutusprogrammi Microsoft Office Excel 2003.

2. Tulemused

2.1 Hübridiseerimisproovide omaduste arvutamine

Et uurida, millised hübridiseerimisproovi omadused võiksid mõjutada signaalide intensiivsust mikrokiibil ja signaali intensiivsuse varieeruvust korduvkatsetel (erinevatel mikrokiipidel), arvutati igale väljatöötatud metoodikat järgides disainitud inimese X kromosoomi spetsiifilisel mikrokiibil olnud hübridiseerimisproovile (499) seda iseloomustavad omadused ehk parameetrid. Arvestati teiste autorite poolt leitud hübridisatsiooni mõjutavaid parameetreid ja arvutati ka proovide parameetreid, mis võiksid mõjutada proovide seondumist oma sihtmärk-järjestustega. Kokku arvutati igale proovile 33 parameetrit, mis võib tinglikult jagada kaheks: proovide järjestuse omadustele vastavad parameetrid ja proovide seondumistele vastavad parameetrid. Näide proovidele arvutatud parameetrite failist on toodud lisas 2.

Igale proovile arvutati tema pikkus ja GC nukleotiidide sisaldus (GC%). Proovi pikkus võib mõjutada proovide hübridiseerumist, kuna erinevate pikkustega proovide migreerumisvõime mikrokiibil võib olla erinev ja kuna pikematel proovidel on rohkem seondumiskohti, kuhu detekteeritav sihtmärk-järjestus (*target*) võib seonduda. Proovide GC sisaldus on oluline, sest arvatakse, et GC% rikkad regioonid proovides rist-hübridiseeruvad sagedamini, kui madala GC sisaldusega piirkonnad. Leiti ka proovide sulamistemperatuurid, mis on olulised kuna erinevate sulamistemperatuuridega proovid võivad seonduda ühel kindlal temperatuuril läbiviidavas katses erinevalt. Kvaliteetsete hübridiseerimisproovide disainimiseks, mida saab edukalt kasutada ühes katses (ühel kiibil) on vaja leida optimaalne sulamistemperatuuride vahemik. Proovide sulamistemperatuur arvutati kahe erineva valemi järgi. Esimene, T_{m1} on proovi sulamistemperatuur, mis on arvutatud monovalentse soola kontsentratsiooni arvestava Marmur-Schildkraut-Doty valemi järgi (Marmur & Doty, 1962; Schildkraut, 1965):

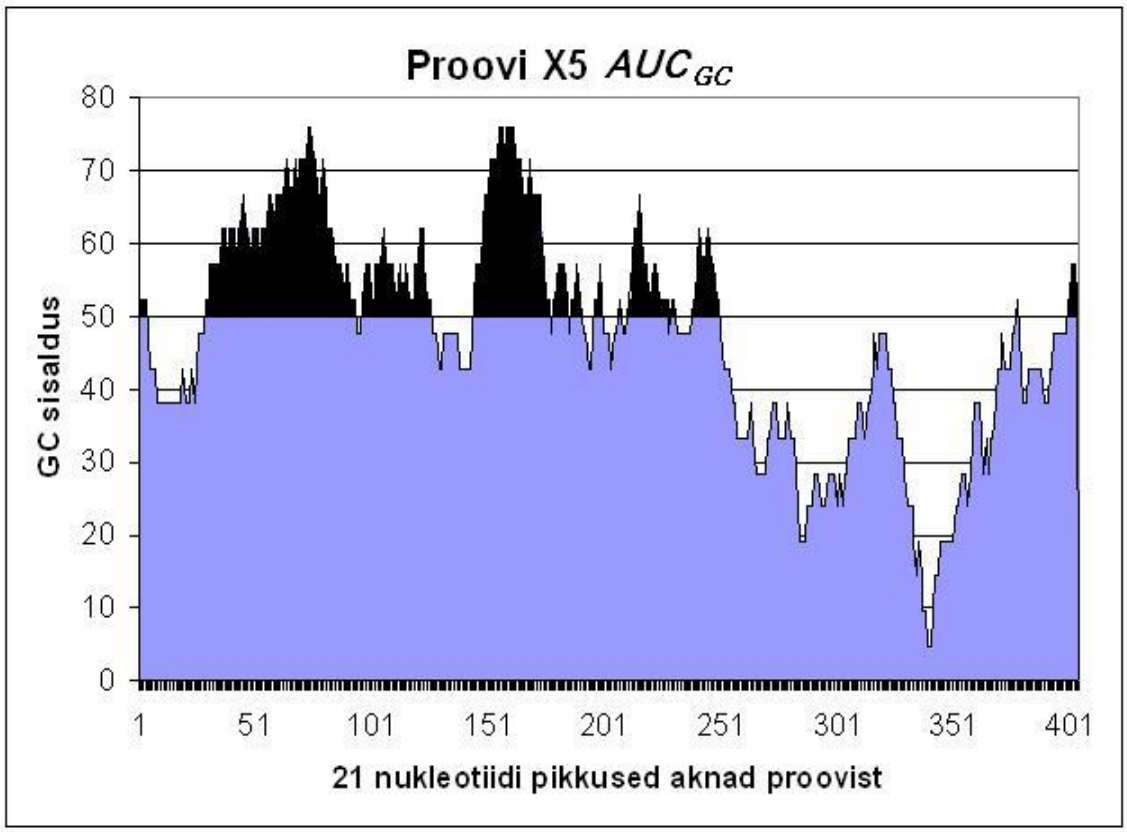
$$T_{m1} = 81.5 + 16.6 * \log_{10}([Na^+]) + (0.41 * GC\%) - (600 / l),$$

kus $\log_{10}([Na^+])$ on kümnendlogaritm monovalentsete soolade kontsentratsioonist, GC% on G ja C nukleotiidide osakaal proovis ja *l* on proovi pikkus. Teine, T_{m2} on proovi sulamistemperatuur, mis on arvutatud lihtsama valemi järgi:

$$T_{m2} = 64.9 + 41 * (nG + nC - 16.4) / (nA + nT + nG + nC),$$

kus n on vastavate nukleotiidide arv järjestuses.

Proovidele arvutati ka nende GC nukleotiidide sisaldusest sõltuvad omadused: AUC_{GC}, $ratio_{GC}$, AUC_{Tm} ja $ratio_{Tm}$. AUC_{GC} (*Area under the GC curve*) on parameeter, mis väljendab näitlikult seda, kui suur osa proovi järjestusest ja millisel määral jääb ülespoole kindlaksmääratud GC% ülempiiri. Käesolevas töös valiti GC sisalduse ülempiiriks 50%. Parameetri AUC_{GC} sisu väljendab joonis 4, kus on näitena toodud ühe analüüsitud proovi (X5) AUC_{GC}.



Joonis 4. Parameetri AUC_{GC} näitlik sisu. Graafikule on kantud libiseva aknaga üle proovi liikudes igas aknas saadud GC sisaldus (sinine ja must ala). Akendest, kus GC% on üle määratud piiri (50%), on arvutatud kokku parameeter AUC_{GC}, mis näitlikult moodustub mustaks värvitud alade summaarsest pindalast.

AUC_{Tm} (*Area under the* T_m *curve*) on sarnane parameeter AUC_{GC}-le, näidates seda, kui suur osa proovi järjestusest ja kui palju on suurem kindlaksmääratud T_m-i ülempiirist (käesolevas töös valiti T_m-i ülempiiriks 70^o C). $Ratio_{GC}$ näitab seda, kui suur osa proovi järjestusest on üle määratud GC sisalduse piiri (käesolevas töös 50%) ja $ratio_{Tm}$ sedasama arvutatuna T_m-i kohta. $Ratio_{GC}$ ja $ratio_{Tm}$ arvutatakse vastavalt valemile:

***ratio* = *N* / proovi pikkus,**

kus *N* on akende arv, milles GC% või T_m oli suurem määratud ülempiirist (50% ja 70^o C vastavalt). Järjestuse aknad saadi liikudes 21 nukleotiidi pikkuse raamiga ja 1 nukleotiidse sammuga üle kogu proovi järjestuse ja arvutades igal sammul raami sees oleva 21 nukleotiidi pikkuse järjestuse GC% ja T_m. T_m on siinkohal arvutatud identselt T_{m1}-le. Parameetrite *ratioGC* ja *ratioTm* paremaks väljendamiseks korrutati nende väärtused 100-ga. Igale proovile arvutati ka tema lokaalne minimaalne (min_{GC}) ja maksimaalne GC (max_{GC}) nukleotiidide sisaldus, mis näitavad proovi sees kindla pikkusega aknas (30 nt) olevat minimaalset ja maksimaalset GC sisaldust.

Parameetrid AUC_{GC}, $ratio_{GC}$, AUC_{Tm}, $ratio_{Tm}$, min_{GC} ja max_{GC} on arvutatud, kuna teised uurijad on näidatud, et need näitajad on kõige paremad *PCR*-i edukuse ennustamisel (Benita *et al.*, 2003). Kuna töös kasutatud mikrokiipide andmed on pärit mikrokiipidel põhineva multipleks amplifitseeritavate proovide hübridisatsioonilt, mille protokoll sisaldab proovide paljundamist *PCR*-i meetodil, on oluline arvestada ka nende parameetritega.

Igale proovile arvutati ka DUST'i skoor, mis näitab, kui palju on proovi järjestuses selliseid madala keerukusega kordusjärjestusi nagu polü-N trakte, di-, tri- ja tetranukleotiidseid korduseid. Kordusjärjestuste sisaldus on oluline, kuna tihtipeale toimub proovide sekundaarstruktuuride moodustumine ja ebaspetsiifiline seondumine just proovides olevate kordusjärjestuste vahendusel.

Hübridiseerimisproovide signaali tugevust mõjutab kindlasti ka proovide seondumiskohtade arv genoomis, sest proovid, mis seonduvad uuritavale DNA-le ebaspetsiifiliselt, annavad vale tugevusega signaale mikrokiibil. Et mikrokiibil olnud hübridiseerimisproovid on täispikkuses unikaalsed, arvutati iga proovi erineva pikkusega alamstringidele ehk järjestuse alamosadele (*substring, subsequence*) nende seondumiste arv inimese genoomis. Seondumiskohtade leidmisel kasutati kolme

erinevat meetodit. Esimesel juhul leiti proovide alamstringide seondumiskohad inimese genoomis alamstringide täies ulatuses (100% identsed piirkonnad). Teisel juhul leiti seondumiskohad, kuhu proovide alamstringid seondusid vähemalt 75% ulatuses. Kolmandal juhul leiti proovide kõikide alamstringide seondumiskohad, kuhu need seonduvad kindla termodünaamilise tugevusega (ΔG).

2.2 Mikrokiibi signaalide normaliseerimine

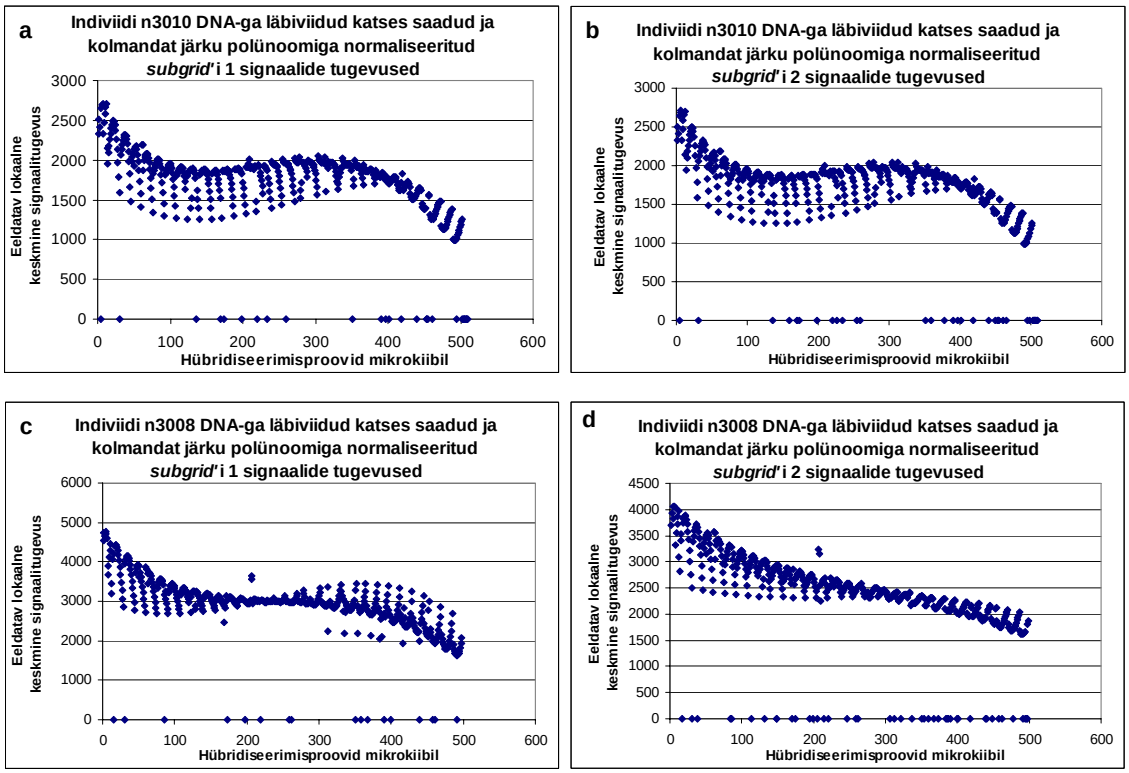
Et analüüsida mikrokiipidelt saadud andmeid, on vajalik esmalt mikrokiipide andmed normaliseerida, kuna toorandmed sisaldavad tihti vigu. Normaliseerimisel on kõige tähtsam leida parim meetod erinevate kiipide (korduvkatsete) omavahel võrreldavaks muutmisel, kuna korduvkatsete signaalitugevused võivad üksteisest erineda mitmeid kordi. Tähtis on ka kiipide nii-öelda füüsilise pinna erinevustest tulenevate nihete ja vigade kõrvaldamine, seda eriti enamtundlike meetodite juures.

Käesolevas töös kasutati mikrokiipide toorandmete normaliseerimiseks kahte erinevat meetodit: normaliseerimist signaalide mediaani järgi ja normaliseerimist kolmandat järku polünoomiga (kuuppolünoomiga). Viimane on täiesti uus lähenemine mikrokiibi toorandmete normaliseerimisel, mis võimaldab leida mikrokiibi füüsilise pinna mõjudest ja/või hübridisatsiooni eripärast tulenevaid erinevusi ja elimineerida signaali väärtuste juhuslikku, normaaljaotusele alluvat kõikumist.

Mediaani järgi normaliseerimisel signaalitugevused kõigepealt logaritmitakse. Seejärel leitakse igale normaliseeritavale mikrokiibile (ja *subgrid'*ile) selle signaalitugevuste mediaan med_i. Iga kiibi i signaalid jagatakse selle kiibi mediaani väärtuse med_i-ga läbi ja saadakse iga signaali nihe mediaanist. Mediaani järgi normaliseerimise miinuseks on see, et kui mõni kiibi või *subgrid'*i osa on andnud kas palju tugevamaid või nõrgemaid signaale, mõjutab see mediaani ja normaliseerimisel viib ka teised signaalitugevused nihkesse.

Kolmandat järku polünoomiga normaliseerimine teostati (Tõnu Möls, TÜ MRI, bioinformaatika õppetool) kasutades statistikapaketti SAS 8.2. Kiibi signaalide normaliseerimisel kolmandat järku polünoomiga paigutatakse piltlikult iga kiibi (*subgrid'*i) logaritmitud signaaliväärtuste peale kolmemõõtmeline pind, mis asetub nii, et kõikide signaalide kaugus sellest oleks minimaalne, kuid samas säiliks polünoomi

ühtlane pind. Joonisel 5 on kujutatud kahe erineva mikrokiibi mõlema *subgrid'*i kuuppolünoomi pinnad.



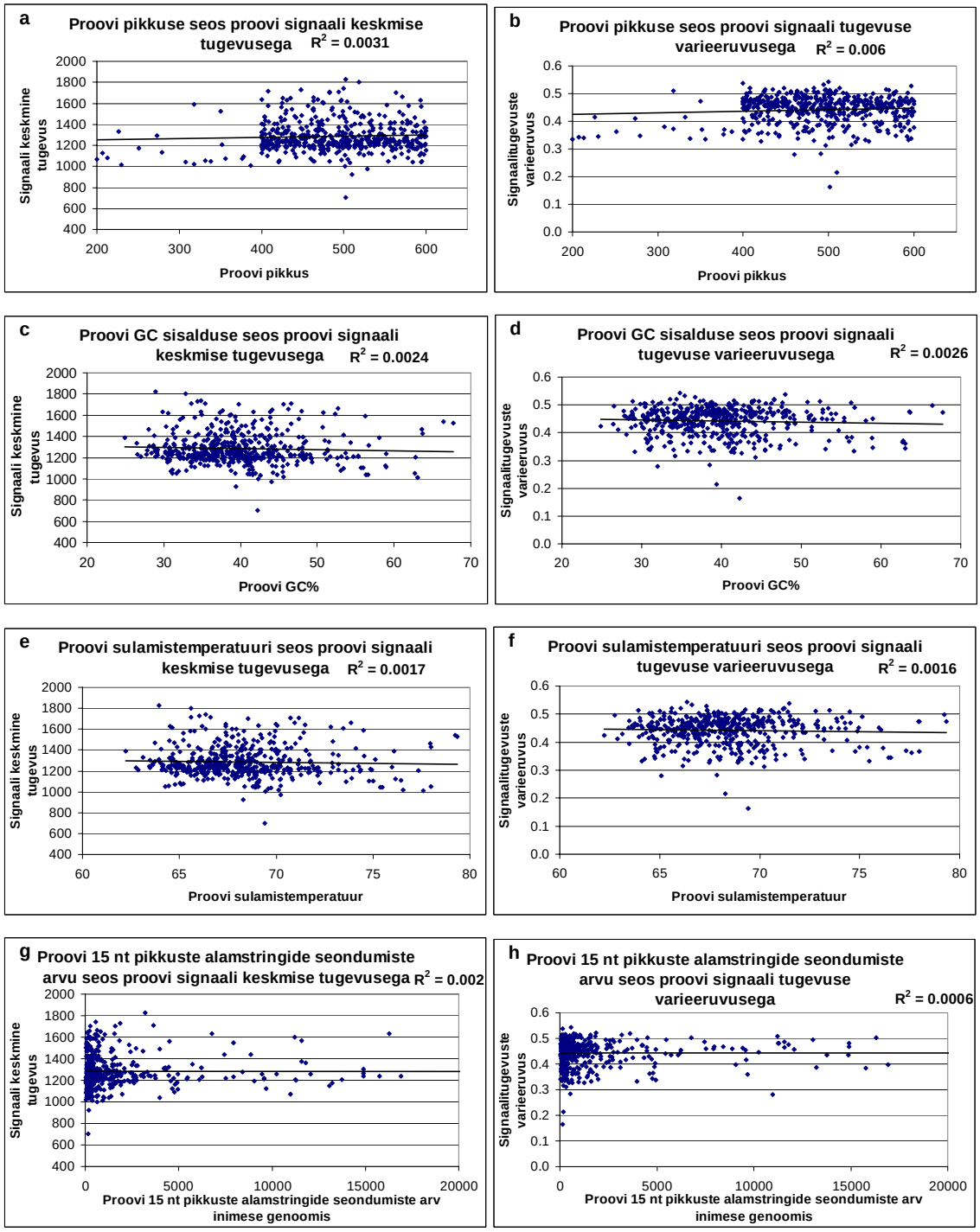
Joonis 5. Kahe erineva mikrokiibi mõlema *subgrid'*i eeldatavad lokaalsed keskmised signaalitugevused. Joonistel 5a ja 5b on indiviidi n3010 DNA-ga läbiviidud katse kolmandat järku polünoomiga normaliseeritud signaalide tugevused ühe mikrokiibi kahel erineval *subgrid'*il. Joonistel 5c ja 5d on indiviidi n3008 DNA-ga läbiviidud katse (kolmandat järku polünoomiga normaliseeritud) eeldatavad signaalide keskmised tugevused ühe mikrokiibi kahel erineval *subgrid'*il.

Signaaliväärtused, mis jäävad mikrokiibi logaritmitud signaalidele sobitatud polünoomi pinnast kõrgemale (signaal on tugevam) või madalamale (signaal on nõrgem) korrigeeritakse vastavalt. Sellega saadakse normaliseeritud signaaliväärtused, mis ideaaljuhul (kui ei esine mikrokiibi füüsilisest pinnast või hübridisatsiooni eripärast tulenevaid mõjusid) asetseks tasapinnale, st. graafikul oleks näha vaid üks sirgjoon.

2.3 Seosed proovide omaduste ja signaali tugevuse ja varieeruvuse vahel

Et optimeerida ja parandada hübridiseerimisproovide disaini metoodikat, uuriti käesolevas töös, millised proovide omadused võiksid mõjutada hübridiseerimisproovide signaali tugevust ja varieeruvust mikrokiipidel. Selleks analüüsiti 39 *subgrid'*i kolmandat järku polünoomiga normaliseeritud signaale ja proovide omadustele vastavaid parameetreid. Kõikide proovide omaduste ja seondumiste parameetrid pandi vastavusse nende signaalitugevustega 39-lt *subgrid'*ilt.

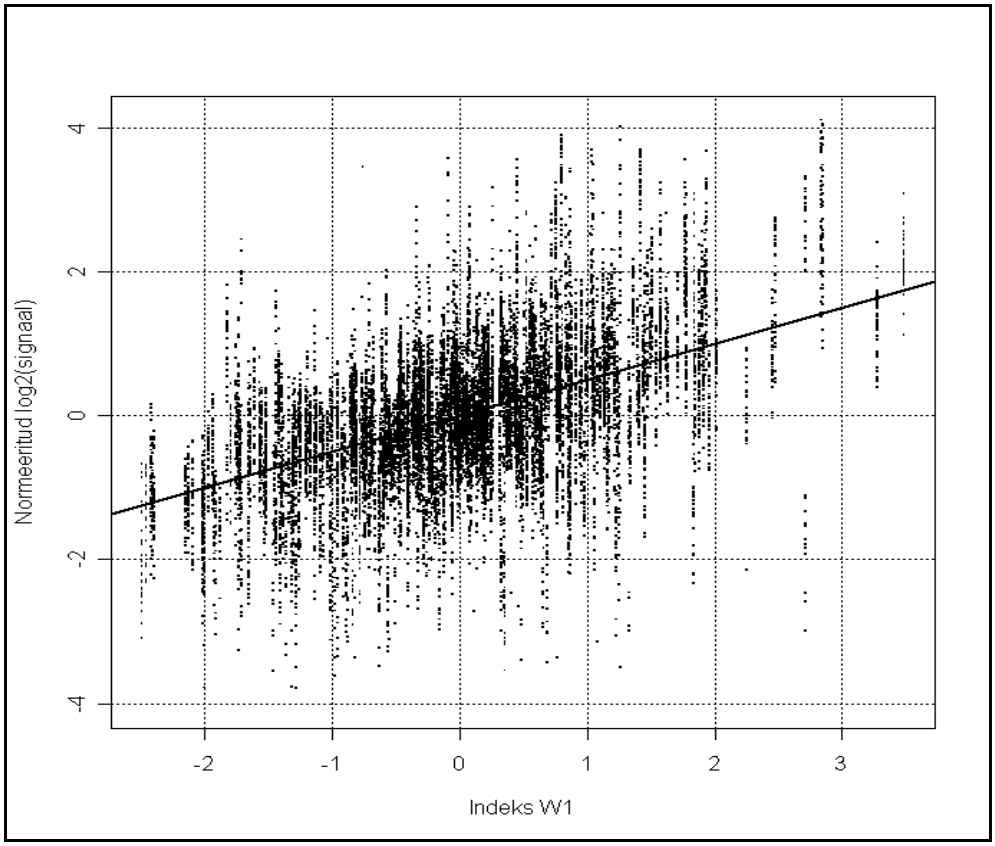
Proovide omadustele ja seondumistele vastavate parameetrite seoseid proovide signaali tugevuse ja varieeruvusega otsiti tabelarvutusprogrammi Microsoft Office Excel abil. Ühtegi proovi signaali tugevust või varieeruvust oluliselt mõjutavat proovi omadusele või seondumise arvule vastavat parameetrit aga ei õnnestunud leida, kuna kiipidevaheline varieeruvus oli liiga suur. Joonistel 6a - 6h on toodud proovide pikkuse, GC sisalduse, sulamistemperatuuri ja 15 nukleotiidi pikkuste alamstringide seondumiste arvu seosed proovide signaali tugevuse ja varieeruvusega.



Joonis 6. Proovide parameetrite ja signaali tugevuse ning varieeruvuse vahelised seosed. Graafikutel on toodud proovide pikkuse (6a, 6b), GC sisalduse (6c, 6d), sulamistemperatuuri (6e, 6f) ja 15 nukleotiidi pikkuste alamstringide seondumiste arvu (6g, 6h) seosed proovide signaali tugevuse ja varieeruvusega.

Analüüsid tehti ka kasutades SAS 8.2 statistikapaketti. Proovide parameetrite ja signaali tugevuse ja varieeruvuse vaheliste seoste leidmiseks kasutati omakorrelatsiooni (*canonical correlations*) analüüsi. Selle tegemiseks teisendati väga suures skaalas

varieeruvad (mittepidevad) proovide parameetrid (seondumiste parameetrid S_15…MB_45) pidevateks. Omakorrelatsiooni analüüsil leiti kõikide faktorite lineaarkombinatsioonist koosnev proovide parameetrite indeks W1, mis kõige paremini korreleerus logaritmiliselt teisendatud ja normaliseeritud signaalidega. Analüüsist selgus, et proovide parameetrite ja signaali varieeruvuse vaheline seos on olemas, kuid ühegi parameetri mõju eraldi proovi signaali tugevusele ega varieeruvusele ei õnnestunud tuvastada, kuna erinevate mikrokiipide ja *subgrid'*ide vahelised varieeruvused olid liiga suured. Joonisel 7 on toodud proovide parameetrite indeksi W1 ja kolmandat järku polünoomiga normaliseeritud ning logaritmitud signaalide omavaheline sõltuvus.

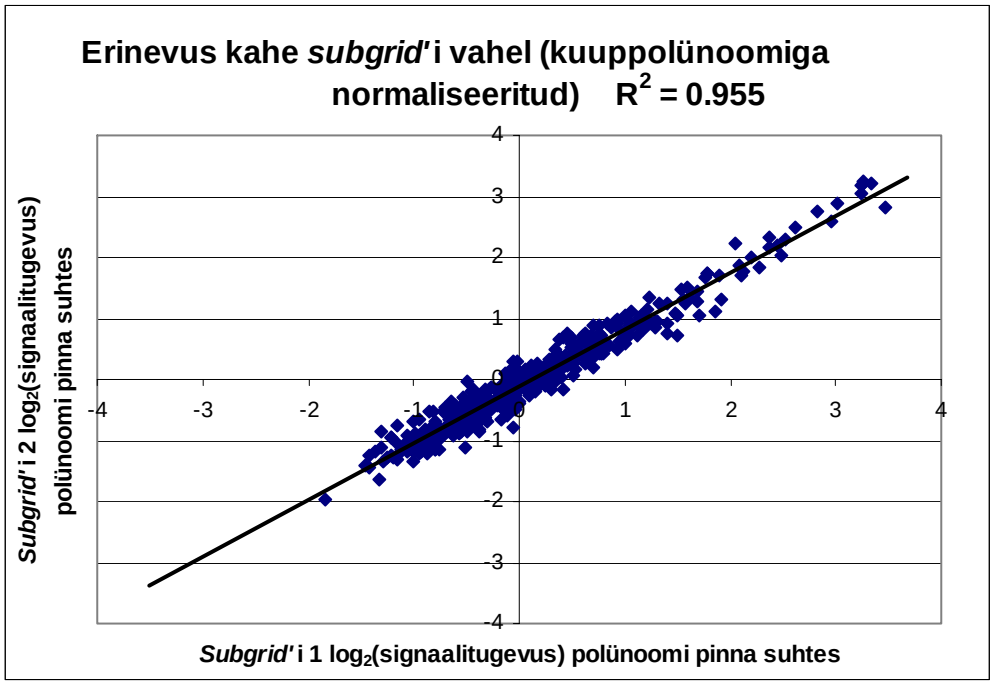


Joonis 7. Proovide parameetrite indeksi W1 ja kolmandat järku polünoomiga normaliseeritud ning logaritmitud signaalide omavaheline sõltuvus. Graafiku y-teljele on kantud kõikide analüüsis olnud *subgrid'*idelt (39) saadud kõikidele proovidele vastavad logaritmitud ja kuuppolünoomiga normaliseeritud signaalitugevused (kokku 17849). Graafiku x-teljel on hübridiseerimisproovide parameetrite indeks W1, mis kõige paremini korreleerus mikrokiipidelt saadud proovide signaalidega.

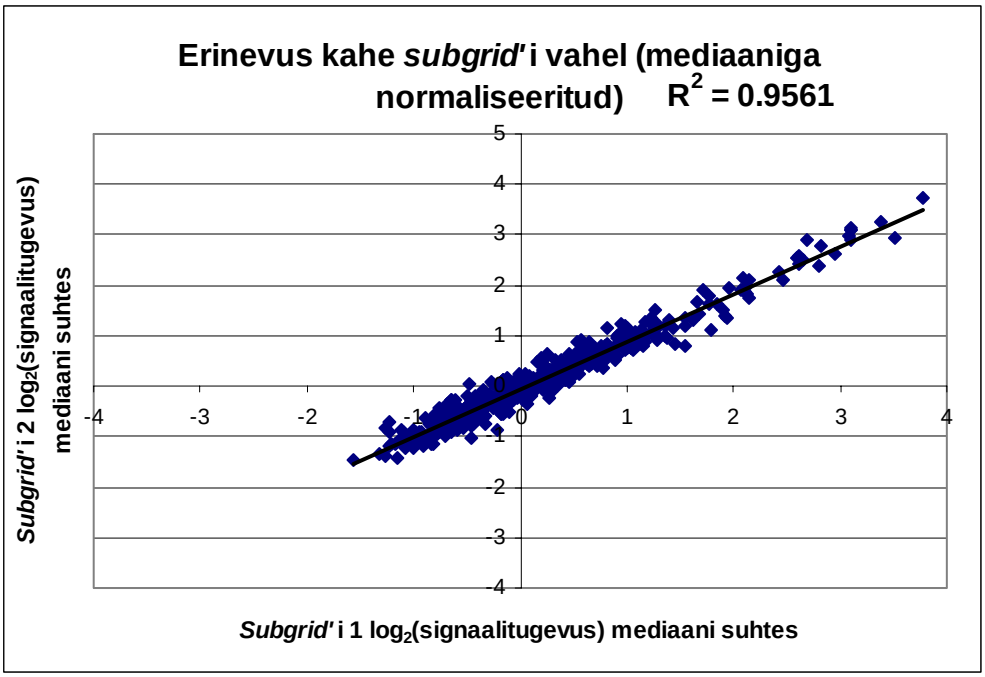
2.4 Mikrokiibi signaalide varieeruvuse allikate selgitamine

Et täpsemalt uurida, millest tuleb suur varieeruvus mikrokiipide signaalides, analüüsiti ühe kontroll-indiviidi DNA-ga tehtud katseid. Ainult ühe kontroll-indiviidi DNA-ga tehtud mikrokiipe kasutati sellepärast, et välistada erinevate indiviidide normaalsest DNA koopiaarvu varieeruvusest tulenevaid erinevusi mikrokiibi signaalides. Kasutati nii mediaani järgi kui ka kolmandat järku polünoomi abil normaliseeritud andmeid. Võrdluseks kasutati ka toorandmeid.

Analüüsis võrreldi ühe kontroll-indiviidi DNA-ga tehtud ja ühel mikrokiibil olnud *subgrid'*e, mis andsid omavahel väga hea korrelatsiooni (R^2 mõnikord isegi 1), seda nii mediaani järgi normaliseeritud signaalitugevustega kui ka kuuppolünoomi abil normaliseeritud signaalitugevustega. Sellest võib järeldada, et ühe kindla proovi signaali varieeruvus erinevatel mikrokiipidel ei tulene ühe katse (ühe mikrokiibi ehk kahe *subgrid'*i) varieeruvusest, sest ühel mikrokiibil olevad signaalid annavad reeglina väga hea omavahelise korrelatsiooni. Ühe mikrokiibi kahe *subgrid'*i vaheline signaalitugevuste korrelatsioon on toodud joonistel 8 ja 9.

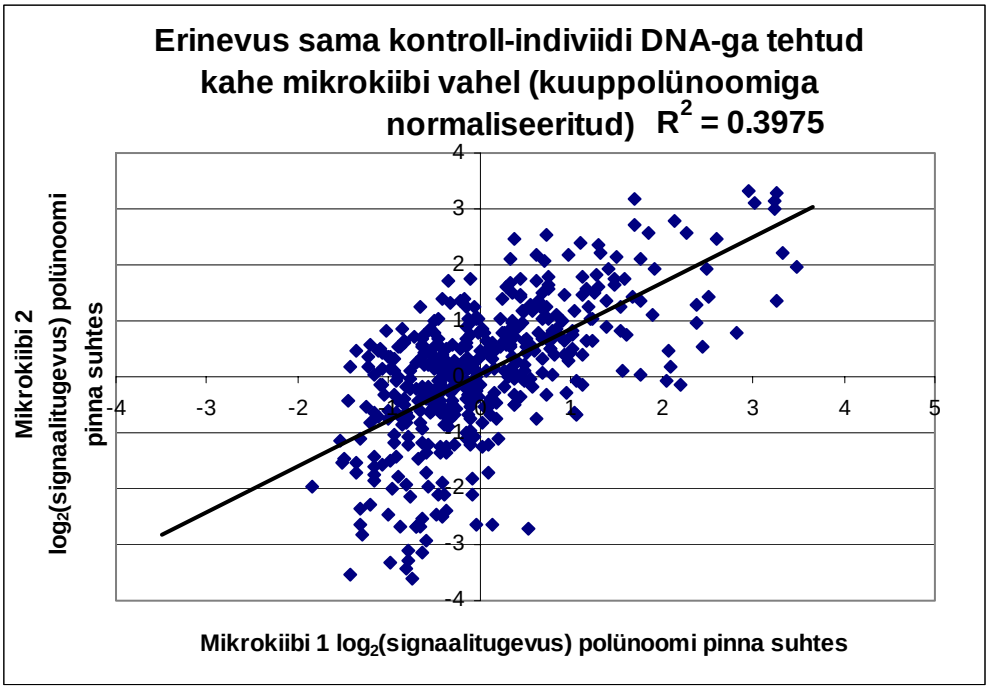


Joonis 8. Ühe mikrokiibi kahe *subgrid'*i signaali tugevuste omavaheline seos. *Subgrid'*ide signaalid on normaliseeritud kolmandat järku polünoomiga. Graafikule kantud signaaliväärtused on arvutatud polünoomi pinna suhtes.

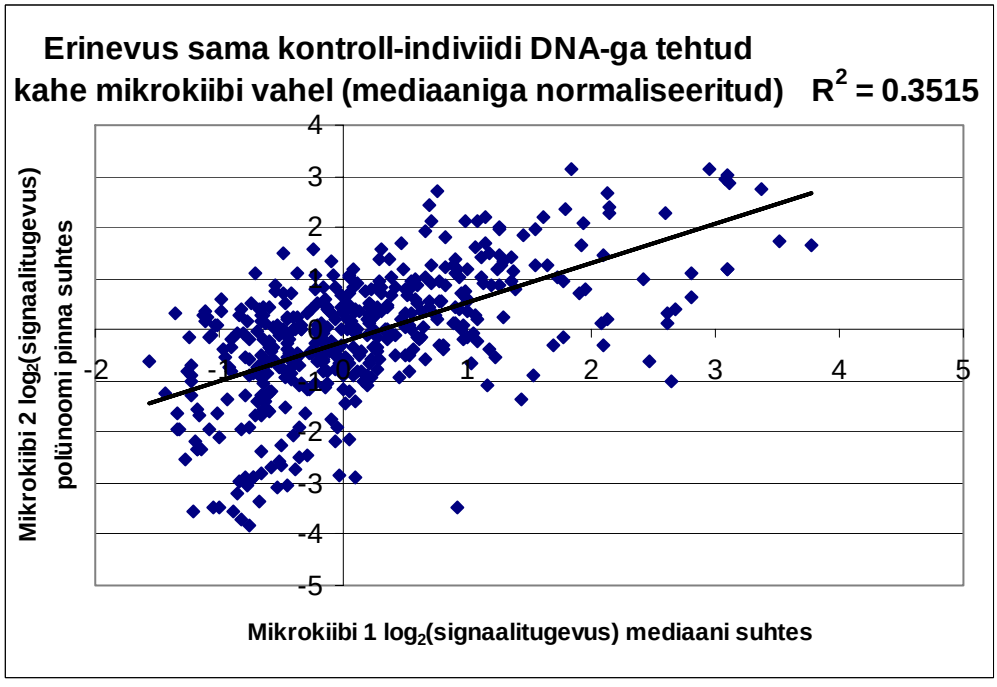


Joonis 9. Ühe mikrokiibi kahe *subgrid'*i signaali tugevuste omavaheline seos. *Subgrid'*ide signaalid on normaliseeritud mediaani (leitud mõlemale *subgrid'*ile eraldi) järgi. Graafikule kantud signaaliväärtused on arvutatud mediaani suhtes.

Edasi võrreldi eri aegadel ühe kontroll-indiviidi DNA-ga tehtud eksperimentidest pärinevaid *subgrid'*e. Joonistel 10 ja 11 on toodud ühe kontroll-indiviidi DNA-ga erinevatel aegadel tehtud katsetevaheline signaalitugevuste korrelatsioon.



Joonis 10. Kahe mikrokiibi signaali tugevuste omavaheline seos. Mikrokiipide signaalid on normaliseeritud kolmandat järku polünoomiga. Graafikule kantud signaaliväärtused on arvutatud polünoomi pinna suhtes.



Joonis 11. Kahe mikrokiibi signaali tugevuste omavaheline seos. Mikrokiipide signaalid on normaliseeritud mediaani (leitud mõlemale mikrokiibile eraldi) järgi. Graafikule kantud signaaliväärtused on arvutatud mediaani suhtes.

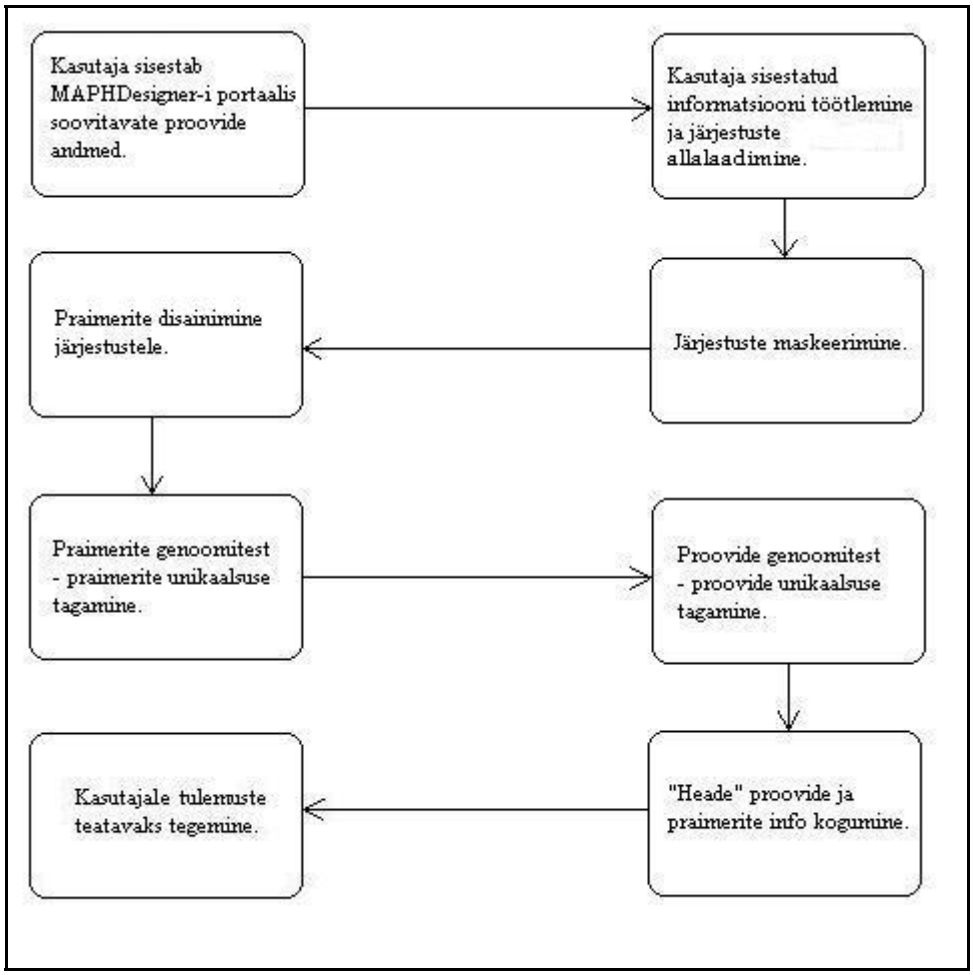
Tehtud analüüs näitas, et ühe kontroll-indiviidi DNA-ga erinevatel aegadel tehtud mikrokiipide vaheline signaalitugevuste korrelatsioon on palju väiksem, kui ühel mikrokiibil tehtud katsete vaheline korrelatsioon (joonised 10 ja 11). Kolmanda järgu polünoomiga signaalide normaliseerimisel saadi erinevatelt mikrokiipidelt signaalide vahel parem korrelatsioon kui mediaani järgi normaliseerimisel, mis kinnitab hüpoteesi, et kuuppolünoomi meetodiga normaliseerides on võimalik paremini vabaneda kiipidel esinevast signaalide juhuslikust kõikumisest.

2.5 Hübridiseerimisproovide automaatse disaini metoodika ja algoritm

Hübridiseerimisproovide automaatseks disainiks kasutatav algoritm võimaldab suhteliselt kiiresti leida proovideks sobivad järjestused, mida spetsiaalselt selleks disainitud praimerite abil on võimalik genoomselt DNA-lt amplifitseerida. Kuna pikkade hübridiseerimisproovide disainimine on aeganõudev protsess, on proovide disaini metoodika üles ehitatud selliselt, et hübridiseerimisproovideks mittesobivad järjestused kõrvaldatakse võimalikult kiiresti, läbides järgnevaid etappe vaid nende proovidega, mis on eelmise etapi „kvaliteedikontrolli" edukalt läbinud.

Enamus kasutatud programmidest on töötamisel rakendatud toru (*pipe*) põhimõttel; eelmise programmi väljund on järgneva programmi sisendiks (STDOUT > STDIN). See võimaldab mingil määral optimeerida programmide tööd, kuna väljundinfo suunatakse otse mälust järgmisele programmile. Nii jääb ära failide kõvakettale kirjutamine ja sealt lugemine, mis võrreldes andmete operatiivmälust lugemise kiirusega on aeglane protsess. Kõik algoritmis töötavad programmid koos lühikirjeldustega on toodud lisas 1.

Hübridiseerimisproovide automaatse disaini lihtsustatud algoritm on toodud joonisel 12.



Joonis 12. Hübridiseerimisproovide automaatse disaini algoritmi lihtsustatud skeem. Skeemil on toodud hübridiseerimisproovide disaini tähtsamad etapid ja nende läbiviimise järjekord.

2.5.1 Algoritmi detailne kirjeldus

Kasutaja poolt käsurealt või veebi kasutajaliidese kaudu sisestatud andmed on sisendiks programmile *proc.pl*. Kui kasutaja on defineerinud kindla regiooni inimese genoomist (näiteks mõni kindel kromosoom), jaotatakse see regioon (*span*) soovitud arvu (*num*) proovide alusel väiksemateks lõikudeks (*subdiv*) ja programm *proc.pl* teeb päringu *core 35* andmebaasi, kust iga lõigu (*subdiv*) keskelt võetakse 3000 nukleotiidi pikkune järjestus. Kui kasutaja on defineerinud kindla geeni, teeb *proc.pl* kõigepealt päringu *human 35* andmebaasi, kust saadakse vastava geeni algus- ja lõpp-koordinaadid, mille alusel toimub soovitud arvu järjestuste päring *core 35* andmebaasist. Saadud järjestused kirjutatakse koos genereeritud ID-ga (kujul

kromosoom:järjestuse alguspositsioon) faili *regions.fas*. Kui andmebaasist *core 35* saadud järjestus sisaldab üle ühe protsendi teadmata nukleotiide, ei kirjutata järjestust andmefaili.

Edasi toimub järjestustes olevate korduste maskeerimine programmidega *dust_lower* ja GenomeMasker. Sarnaselt teistele programmidele ei kasuta programm GenomeMasker teadaolevate kordusjärjestuste andmebaasi, vaid maskeerib etteantud järjestuses kõik regioonid, mis esinevad järjestuses (inimese genoom) rohkem kui X korda, kusjuures X on kasutaja poolt muudetav parameeter. Üleesindatud korduste otsimine järjestustest on aeganõudev, kuid selle eest ühekordne tegevus, mille alusel tehakse üleesindatud järjestuste nimekiri (*black list*). Lisaks kiiremale maskeerimisele on GenomeMasker'i eeliseks see, et ta ei maskeeri kogu kordusjärjestuse motiivi etteantud järjestuses, vaid ainult selle kõige 3' otsa poolsema nukleotiidi, mistõttu praimerit ei ole võimalik disainida täpselt kordusjärjestuse peale. Samas aga säilib võimalus disainida praimer kordusjärjestuse 5' otsa, mis koos kõrvalasuva mitte-korduva regiooniga võib anda piisavalt unikaalse piirkonna, et see sobiks praimeriks. See võimaldab disainida praimereid ka kordusjärjestuste rikastesse regioonidesse (Andreson *et al.,* käsikiri ettevalmistamisel). GenomeMasker'i poolt maskeeritud järjestused kirjutatakse faili *regions.gm*.

Programm *fasta2primer3.pl* kirjutab maskeeritud järjestused praimerite disainimise programmile Primer3 sobivas formaadis faili *primers.txt*. Primer3'e modifitseeritud versioon *gm_primer3* üritab igas sisendiks olevas 3000 nukleotiidi pikkuses maskeeritud järjestuses leida sobivad praimerid, mis vastaksid soovitud tingimustele ja kontrollib, et disainitud praimeripaariga amplifitseeritav produkt vastaks hübridiseerimisproovile seatud nõuetele. Programm *primer3_to_table.pl* kirjutab *gm_primer3*-e tulemused sobivas formaadis (*ID, sense-praimer, antisense-praimer, produkt*) faili *primers2.txt*.

Programm GenomeTester kontrollib seejärel genoomitestiga, kuhu saadud praimerid inimese genoomis seonduvad. Sarnaselt programmile SSAHA kasutab GenomeTester järjestuste otsimisel paisktabelisse paigutatud ja eelindekseeritud andmebaasi. Praimerite seondumiskohtade otsimisel loetakse paisktabel mällu ja identsete piirkondade leidmine toimub juba kiiresti. GenomeTester'i töö tulemusena leitakse praimerite seondumiskohad genoomis ning kui palju ja millistest regioonidest nende abil produkte amplifitseeritakse (Andreson *et al.,* käsikiri ettevalmistamisel). Praimerid, mis seonduvad mittespetsiifiliselt ja annavad sekundaarseid produkte,

eemaldatakse programmiga *gtester_filter.pl*. Sobivate praimeritega *in silico* saadavad produktid ehk hübridiseerimisproovid kirjutatakse faili *prod.fas*.

Seejärel arvutab programm *dust_score* vastavalt failis *prod.fas* olevate proovide maskeeritud tähtedele igale järjestusele vastava skoori. Järjestused, millel on liiga palju korduvaid elemente (skoor üle 50), kõrvaldatakse programmidega *cut_off.pl* ja *exclude_from_fasta.pl* failist *prod.fas*.

Proovide unikaalsuse kontroll tehakse nii proovide kui ka genoomsete järjestuste vastu. Kuna pikkade proovide unikaalsuse kontroll kogu inimese genoomi vastu on aeganõudev protsess, tehakse see kolmes etapis, mis aitab kiirendada proovide disainimist, sest igas järgnevas etapis töötatakse edasi vaid proovidega, mis on edukalt läbinud eelneva etapi. Palju aega nõudvate etappide läbiviimine minimaalse arvu negatiivsete proovidega võimaldab oluliselt kiirendada kogu disainimise protsessi.

Esmalt kontrollitakse proove üksteise vastu programmiga MegaBLAST. Proovid, mis seonduvad 50 nukleotiidi pikkusel lõigul üksteisega rohkem kui 75% ulatuses, eemaldatakse.

Seejärel kontrollitakse programmiga SSAHA, kuhu disainitavad proovid inimese genoomis seonduvad. Proovid, mis seonduvad enam kui 30 nukleotiidi ulatuses mitteoodatud regioonidega, eemaldatakse.

Proovide unikaalsuse kontrolli kõige aeganõudvam osa, genoomsete seondumiskohtade kontroll programmiga MegaBLAST viiakse läbi allesjäänud proovidega. Proovid, mis on vähemalt 50 nukleotiidi pikkusel lõigul enam kui 75 protsenti identsed mitteoodatud genoomsete regioonidega, eemaldatakse järgmises etapis.

MegaBLAST'i tulemused parsitakse ehk formaaditakse ümber programmiga *megablast_parser* tabuleeritud formaati ja nende tulemuste järgi eemaldatakse programmidega *parsers_parser.pl*, *wrong_chr_out.pl*, *wrong_doubled_out.pl*, *one_per_start* ja *extract_from_file.pl* inimese genoomis mittespetsiifilisi seondumisi andvad proovid.

Viimases etapis võetakse erinevatest failidest „heade" proovide vajalikud andmed ja kirjutatakse need koos praimerite ja proovide järjestustega lõpp-faili *results.txt*. Selles etapis osalevad programmid *extract.sh*, *first_name.pl*, *first.pl*, *extract_lines_from_file.pl* ja *one_per_start_pos.pl*. Fail r*esults.txt* sisaldab: disainitud proovi järjekorranumbrit; ID-d; kromosoomi, kuhu proov seondub; proovi

alguspositsiooni kromosoomil; proovi amplifitseerimiseks vajalike praimerite järjestusi ja proovi järjestust.

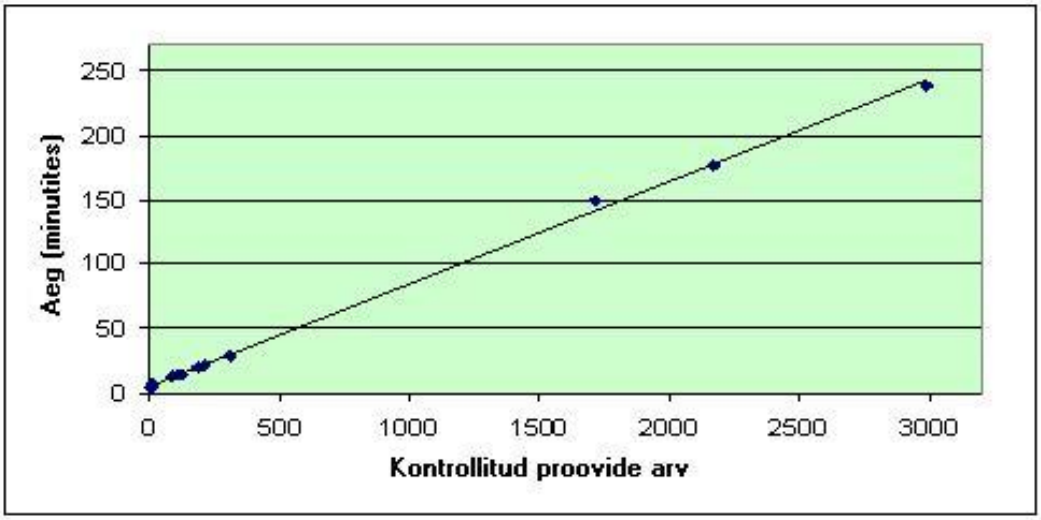
Et kasutajal oleks võimalik mugavamalt tulemusi alla laadida, pakitakse koopia failist *results.txt* faili *results.zip*. Kui kasutaja sisestas töö alguses oma e-maili aadressi, saadab programm *proc.pl* disainimisprotsessi lõpus kasutajale vastavasisulise e-maili. Kirjas on kasutaja päringu ID (*request ID*, unikaalne igal disainimisel) ja disainimise protsessi alguse aeg. Lisaks sisaldab kiri ka hüperlinke tulemuste failile ja tulemuste pakitud failile.

2.5.2 Algoritmi tööaeg ja mälukasutus

Hübridiseerimisproovide automaatse disaini metoodikas kasutatud programmidest on kõige ressursinõudlikumad programmid, mis kontrollivad praimerite ja proovide seondumiskohti genoomis: GenomeTester, SSAHA ja MegaBLAST. Tööks vajalik mälumaht on kõige suurem programmil SSAHA, mis otsides proove inimese suurematelt kromosoomidelt, hõivab kuni 1 GB mälu. MegaBLAST'i sama näitaja on 500 MB. GenomeTester vajab praimerite seondumiskohtade otsimisel samuti kuni 500 MB mälu. Vähem mälu kasutavate programmide (näiteks GenomeMasker ja *gm_primer3*) nõudlikkus ei ületa 150 MB piiri.

Nagu mälukasutuse puhul, on ka ajalises mõttes kõige keerukamad programmid MegaBLAST ja SSAHA. Ülejäänud programmide tööajad on suhteliselt lühikesed. Teoreetiliselt on nii SSAHA kui ka MegaBLAST'i ajaline keerukus võrdelises seoses päringute arvuga; päringute arvu suurenemisel kasvab lineaarselt ka programmide töötamise aeg (Ning *et al.*, 2001; Zhang *et al.*, 2000). Programmil SSAHA lisandub proovide arvust sõltuvale tööajale veel indeksite kõvakettalt mälusse lugemise aeg.

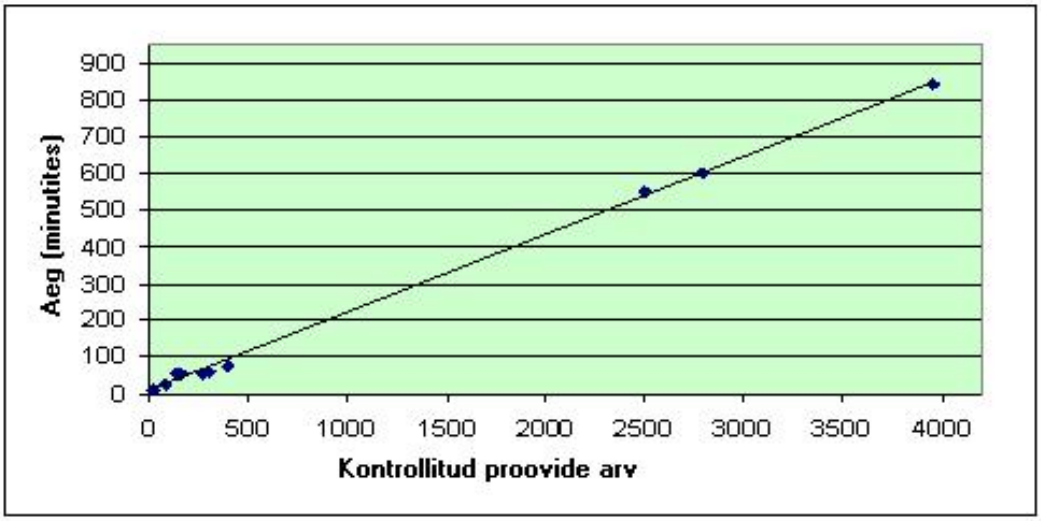
Joonisel 13 on ära toodud programmi SSAHA praktiliselt mõõdetud ajaline kulu sõltuvalt kontrollitavate proovide arvust.



Joonis 13. Programmi SSAHA tööaeg sõltuvalt kontrollitud proovide arvust.

Siinjuures on oluline märkida, et proovide arvu all on silmas peetud programmide SSAHA ja MegaBLAST poolt kontrollitud proovide arvu, mitte positiivseid tulemusi.

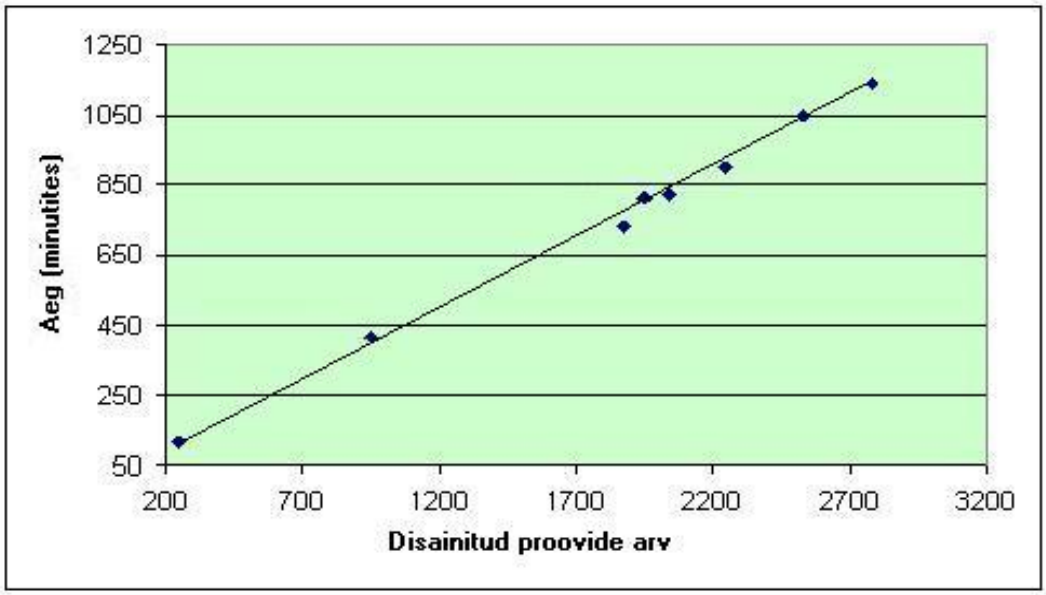
Joonisel 14 on programmi MegaBLAST ajaline kulu sõltuvalt kontrollitavate proovide arvust.



Joonis 14. Programmi MegaBLAST tööaeg sõltuvalt kontrollitud proovide arvust.

Nagu joonistel 13 ja 14 näidatud, on programmide SSAHA ja MegaBLAST praktiliselt saadud tööajad lineaarses seoses päring-järjestuste arvuga, mis on kooskõlas programmide teoreetilise tööajaga (Ning *et al.*, 2001; Zhang *et al.*, 2000).

Kogu algoritmi tööaja moodustab kõikide programmide summaarne töötamise aeg. Hübridiseerimisproovide disaini protsessi ajalist kulu kirjeldab joonis 15.



Joonis 15. Hübridiseerimisproovide disainimise tööaeg sõltuvalt disainitud proovide arvust.

Nagu joonisel 15 näidatud, on protsessi ajaline kulu lineaarses sõltuvuses disainitud proovide arvust. Selle keskmist suuremate päringute puhul väljendab valem:

$$T = 30 + K * \text{proovide arv},$$

kus T on aeg ja K on ühe proovi disainimiseks keskmiselt kuluv aeg minutites. 30 minutit lisandub programmide „tühjalt“ töötamise (indeksite kettalt mällu lugemine, jne) ajana.

Antud juhul töötab hübridiseerimisproovide automaatse disainimise algoritm Mandrekelinux 10.1 operatsioonisüsteemiga kaheprotsessorilises (2 x 800Mhz Intel Pentium III) serveris, millel on 2 GB operatiivmälu. Ühe proovi disainimise ajaks (K) tuli keskmiselt 0,4 minutit ehk 24 sekundit.

2.6 Veebi kasutajaliides

Loodud programmide paketi lihtsamaks kasutamiseks on tehtud veebipõhine kasutajaliides MAPHDesigner (*http://bioinfo.ebc.ee/MAPH*), mis võimaldab kasutajal disainida inimese genoomsetele järjestustele vastavaid hübridiseerimisproove.

Kasutajal on võimalik MAPHDesigner'iga disainida proove kindlatele genoomi regioonidele (kromosoomidele ja nende piirkondadele) ja/või geenidele. Vaikimisi on valikuks inimese 1. kromosoom). Valikute hulgas olevate kromosoomivöötide määramisel pannakse nende alade algus- ja lõpp-koordinaadid (positsioonid kromosoomil) automaatselt paika. Samas on kasutajal võimalik ka ise määrata regioonid, kuhu proove soovitakse disainida. Geenispetsiifiliste proovide disainimisel peab kasutaja vastavasse kohta sisestama geenide nimed, millele soovitakse proove disainida. Seejärel on võimalus määrata, millisest andmebaasist soovib kasutaja sisestatud geenide koordinaadid saada. Võimalik on valida ENSEMBL-i, Vega ja RefSeq'i vahel. Erinevate projektide käigus on geene erinevalt annoteeritud ning seetõttu erinevad andmebaasides ka nende koordinaadid. Andmebaasis Vega on 2005 a. mai seisuga saadaval 6., 7., 13., 14., 20. ja 22. kromosoomil paiknevate geenide koordinaadid. Andmebaasides ENSEMBL ja RefSeq on esindatud kõik inimese kromosoomid. Vaikimisi on geenikoordinaatide andmebaasiks ENSEMBL.

Ka soovitud disainitavate proovide arvu määrab MAPHDesigner'is kasutaja. Kui kasutaja on valinud mitu erinevat genoomset regiooni ja/või geeni, üritab programm teha soovitud arvu proove igale valitud regioonile. Vaikimisi on soovitavate proovide arvuks 1.

Samuti saab valida disainitavate proovide pikkuse. Võimalik on määrata minimaalne proovi pikkus (vaikimisi 400 bp), optimaalne proovi pikkus (vaikimisi 500 bp) ja maksimaalne proovi pikkus (vaikimisi 600 bp).

Lisaks saab kasutaja määrata hübridiseerimisproovide minimaalse (vaikimisi 60^{o} C) ja maksimaalse (vaikimisi 70^{o} C) sulamistemperatuuri.

Kui kasutaja soovib proovide disaini lõppemisest teada saada e-maili teel, on tal selleks võimalik sisestada oma e-maili aadress.

Hübridiseerimisproovide disain algab, kui kasutaja vajutab nupule „*Select Probes*". Kasutaja poolt sisestatud andmed saadetakse läbi *CGI* (*Common Gateway Interface*) programmile *proc.pl*, mis käivitab proovide disaini. Kasutaja suunatakse edasi ootelehele, kus on võimalik nupu „*Check Results*" vajutamisega kontrollida, kas proovide disain on lõpetatud. Kui hübridiseerimisproovide disain on lõpetatud, suunatakse kasutaja tulemuste lehele, kus on otsetee failile *results.txt* ja kust on võimalik alla laadida tulemuste faili ka pakitud (*.zip)* kujul.

ARUTELU

Käesolevas töös tehtud analüüside põhjal võib eeldada, et välja töötatud metoodika sobib automaatseks hübridiseerimisproovide disainimiseks suurtele, komplekssetele genoomidele. Realiseeritud algoritmid võimaldavad disainida hübridiseerimisproove ka suuremahuliste projektide jaoks, kuna teoreetiliselt on proovide disainimisel ainsateks limiteerivateks faktoriteks aeg ja kasutatavate järjestuste kvaliteet.

Muutes erinevate programmide töötamise järjekorda, on välja töötatud algoritmi tõenäoliselt võimalik ka edasi arendada ning optimeerida. Võimalik, et see aitab algoritmi veelgi kiiremaks muuta. Programmide töötamise järjekord peaks siiski alluma loogikale, mille puhul hübridiseerimisproovid ja nende amplifitseerimiseks vajalikud praimerid, mis ei vasta kvaliteedinõuetele, eemaldatakse disainimisprotsessi alguses ja kõige aeganõudvamad protsessid (unikaalsuse testimine) viiakse läbi minimaalse arvu mittesobivate proovidega.

Samuti oleks võimalik disainimisprotsessi oluliselt kiirendada, kui jagada üksteisest vähesõltuvad, kuid ajaliselt keerukad protsessid (proovide unikaalsuse kontrollimine programmidega SSAHA ja MegaBLAST) arvutamiseks erinevatele arvutitele või arvutiklastrile.

Tuleb märkida, et proovide disainimiseks väljatöötatud algoritm ja selle kasutamiseks tehtud veebiliides on unikaalsed, sest algoritmi arendamise ajal olemas olnud programmidega saab disainida suhteliselt lühikesi, oligonukleotiidseid proove. Samas on aga vajadus ka pikkade hübridiseerimisproovide järele. Viimased on eelistatud mitmete DNA koopiaarvu määramisel kasutatavate meetodite (kiipidel läbiviidavad *CGH* ja *MAPH*) puhul, kuna annavad intensiivsema signaali ja tagavad seega madalatasemeliste koopiaarvu muutuste täpsema detekteerimise ja kaardistamise.

Kasutaja jaoks saab proovide disainimise mugavamaks muuta, arvestades kui suur protsent soovitud proovide arvust õnnestub keskmiselt disainida. Sel juhul oleks võimalik kohe alguses võtta kadude võrra rohkem proove. Üheks lahenduseks oleks korrutada kasutaja poolt sisestatud proovide hulk disainimise alguses läbi kindla kordajaga nii, et lõpptulemuses oleks piisav arv proove. Käesolevas töös saadud

tulemuste põhjal võib oletada, et piisava arvu proovide saamiseks võiks kasutaja poolt sisestatud proovide arvu korrutada 4 või 5-ga, kuna keskmiselt langeb disainimise käigus välja 75 protsenti kasutaja poolt soovitud proovide arvust.

Hübridiseerimisproovide disainimisel proovidele seatud kvaliteedinõudeid (pikkus, GC sisaldus, sulamistemperatuur, unikaalsus) võib pidada piisavateks. Selle kasuks räägib ka asjaolu, et väljatöötatud metoodika alusel disainitud hübridiseerimisproovid, mida kasutati eksperimentaalselt inimese X kromosoomi spetsiifilisel mikrokiibil, andsid erinevatel kiipidel ühetaolisi signaale. Viimast võib järeldada ühe mikrokiibi *subgrid'*ide omavahelisest väga heast korrelatsioonist ja proovide omaduste, nende signaalitugevuse ning varieeruvuse vahelisest analüüsist, kus ei eristunud ühtegi proovi omadusele vastavat parameetrit, mis oleks oluliselt mõjutanud proovi signaalide tugevust ja varieeruvust korduvkatsetes. Võimalik, et proovi signaalitugevust ja signaali varieeruvust oluliselt mõjutavaid parameetreid oleks õnnestunud leida juhul, kui analüüsis oleks olnud ka juhuslikult, nõutud kvaliteedimäära rakendamata disainitud proove. Siiski pole välistatud, et proovide signaalitugevust ja selle varieeruvust mõjutavaid faktorid õnnestub määrata, kui viia korduvkatsete vahelised kõikumised nii katseliselt kui normaliseerimisega miinimumini.

Mikrokiipide signaalide normaliseerimise analüüsil leiti, et tulemused erinevate normaliseerimismeetoditega võivad mitte kokku langeda. Normaliseerimiseks tuleb leida sobivaim meetod või mitme meetodi kombinatsioon. Üheks selliseks võimaluseks oleks koos kasutada kolmandat järku polünoomiga normaliseerimist ja mediaani järgi normaliseerimist. Kolmandat järku polünoomiga on võimalik efektiivselt likvideerida signaalide juhuslikud kõikumised ühel mikrokiibil. Normaliseerides seejärel kõiki kasutatavaid mikrokiipe mediaani järgi, on võimalik muuta erinevad korduvkatsed omavahel paremini võrreldavaks.

Normaliseerimine on tähtis proovide omaduste ja signaalitugevuse ning varieeruvuse vaheliste seoste leidmiseks. Kui on teada, millised faktorid (proovi omadused) mõjutavad signaalide varieeruvust, on võimalik neid teadmisi edaspidi rakendada väljatöötatud algoritmi täiustamiseks ja seega paremate hübridiseerimisproovide disainimiseks. Lisaks sellele on andmete normaliseerimine väga tähtis mikrokiipide rakenduslikus vallas, olukorras kus kontroll-indiviidide mikrokiipidelt saadud signaalide alusel tuleb hinnata muutusi uuritava patsiendi lookuste koopiaarvus, nagu seda tehakse mikrokiipidel läbiviidava *MAPH* meetodi puhul. Kiipidevahelise varieeruvuse vähendamine aitab parandada kontroll-indiviidide

signaalide usaldusintervalli. See teeb lihtsamaks ja täpsemaks patsientide lookuste koopiaarvu määramise, mis on eelduseks, et antud metoodikat oleks võimalik rakendada DNA koopiaarvu muutustega seotud häirete uurimisel ja vastavate patsientide skriinimisel.

KOKKUVÕTE

Käesolevas töös anti ülevaade erinevatest aberratsioonide tuvastamise meetoditest. Erilist tähelepanu pöörati mikrokiipidel põhinevatele meetoditele (mikrokiipidel põhinev *CGH* ja *MAPH*). Mitmete autorite poolt on välja toodud erinevaid proovide hübridiseerumist mõjutavaid tegureid ning kuna üheks selliseks on hübridiseerimisproovide unikaalsus, räägiti enamlevinud unikaalsuse tuvastamise meetoditest.

Töö raames arendati välja metoodika ja algoritmid mikrokiipidel põhinevate *CGH* ja *MAPH* meetoditel kasutatavate hübridiseerimisproovide automaatseks disainimiseks. Loodud algoritmi ja programmide mugavamaks kasutamiseks tehti veebipõhine kasutajaliides MAPHDesigner (*http://bioinfo.ebc.ee/MAPH*), mille kaudu saab kasutaja disainida oma nõudmistele vastavaid hübridiseerimisproove.

Hübridiseerimisproovide disaini optimeerimiseks analüüsiti käesoleva töö raames väljatöötatud metoodikaga disainitud proove. Neid kasutati eksperimentaalselt *MAPH*-i meetodil inimese X kromosoomi spetsiifilisel mikrokiibil. Selleks arvutati igale mikrokiibil olnud proovile parameetrid, mis iseloomustasid selle omadusi ja seondumiste arvu inimese genoomis. Kasutatud mikrokiipide andmed normaliseeriti kolmandat järku polünoomi abil. Leiti, et hübridiseerimisproove iseloomustavate parameetrite koosmõju proovide signaali tugevusele ja varieeruvusele on olemas, kuid üksikuid tegureid ei õnnestunud leida, kuna kiipidevaheline varieeruvus oli liiga suur.

Kiipidevahelise suure varieeruvuse põhjuste leidmiseks võrreldi kahe erineva meetodiga (kolmandat järku polünoomiga ja mediaani järgi) normaliseeritud mikrokiipide andmeid. Selgus, et ühe kindla proovi signaali varieeruvus ei tule ühe katse (ühe mikrokiibi) sisesest varieeruvusest, vaid erinevate korduskatsete vahelisest suurest varieeruvusest. Selle tõestuseks on ka asjaolu, et ühe mikrokiibi erinevatelt *subgrid*'idelt loetud signaalid andsid omavahel väga hea (R^2 0.95…1) korrelatsiooni. Kasutatud normaliseerimismeetoditest saadi paremad tulemused kolmandat järku polünoomi puhul, mille abil oli võimalik vabaneda kiipidelt loetud signaalide juhuslikust kõikumisest.

SUMMARY

In the past few years it has been found that changes in DNA copy number (duplications/deletions) are most likely the cause of many genetic diseases, including mental retardation. To distinguish pathological aberrations from normal variability between individuals, it is important to develop new and improve existing methodologies for measuring locus copy number alterations in human genome. Two of such methodologies are microarray MAPH (Multiplex Amplifiable Probe Hybridization) and CGH (Comparative Genomic *in situ* Hybridization).

The purpose of this project was to develop a unique methodology and programs for designing PCR-amplifiable hybridization probes (200-600 bp) that can be used for microarray MAPH and CGH. We have also developed a web interface (*http://bioinfo.ebc.ee/MAPH*) for our programs.

To optimize our methodology, we analyzed 499 probes that were made with our programs and used on human X chromosome-specific microarray, using MAPH methodology. Therefore we calculated 33 parameters for all 499 probes used. Calculated parameters characterized probe sequence properties and their binding sites in human genome. Data from microarray assays was normalized using cubic polynomial. As the result of these analyses, it was found, that probe parameters have influence to the mean and the variations in signal intensity. However, we found no specific parameters influencing the signal intensity due to the substantial variability between microarrays used.

To find the cause of variability between the microarrays, normalized data from different microarrays was compared. Normalization was done using two separate methods: cubic polynomial and normalization by median signal intensity. If was found that the variability in signals of one specific probe on different microarrays is not caused by the inner variability of a single test (one microarray). It is rather caused by large variability between different microarrays. The fact that the signals read from different subgrids of the same microarray gave good correlation proves it. We also found, that normalization done with cubic polynomial method gave the best results of the methods used, because of its higher efficiency in getting rid of accidental fluctuation of the signals read from the microarrays.

KASUTATUD KIRJANDUS

Albertson, D. G. (2003). Profiling breast cancer by array CGH. *Breast Cancer Res Treat* 78, 289-298.

Albertson, D. G. & Pinkel, D. (2003). Genomic microarrays in human genetic disease and cancer. *Hum Mol Genet* 12 Spec No 2, R145-152.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25, 3389-3402.

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.

Asensio, J. L., Lane, A. N., Dhesi, J., Bergqvist, S. & Brown, T. (1998). The contribution of cytosine protonation to the stability of parallel DNA triple helices. *J Mol Biol* 275, 811-822.

Bedell, J. A., Korf, I. & Gish, W. (2000). MaskerAid: a performance enhancement to RepeatMasker. *Bioinformatics* 16, 1040-1041.

Beheshti, B., Park, P. C., Braude, I. & Squire, J. A. (2002). Microarray CGH. *Methods Mol Biol* 204, 191-207.

Benita, Y., Oosting, R. S., Lok, M. C., Wise, M. J. & Humphery-Smith, I. (2003). Regionalized GC content of template DNA as a predictor of PCR success. *Nucleic Acids Res* 31, e99.

Bommarito, S., Peyret, N. & SantaLucia, J., Jr. (2000). Thermodynamic parameters for DNA sequences with dangling ends. *Nucleic Acids Res* 28, 1929-1934.

Borer, P. N., Dengler, B., Tinoco, I., Jr. & Uhlenbeck, O. C. (1974). Stability of ribonucleic acid double-stranded helices. *J Mol Biol* 86, 843-853.

Buckley, P. G., Mantripragada, K. K., Benetkiewicz, M. & other authors (2002). A full-coverage, high-resolution human chromosome 22 genomic microarray for clinical and research applications. *Hum Mol Genet* 11, 3221-3229.

Chen, Y., Kamat, V., Dougherty, E. R., Bittner, M. L., Meltzer, P. S. & Trent, J. M. (2002). Ratio statistics of gene expression levels and applications to microarray data analysis. *Bioinformatics* 18, 1207-1215.

Csink, A. K. & Henikoff, S. (1998). Something from nothing: the evolution and utility of satellite repeats. *Trends Genet* 14, 200-204.

Dimitrov, R. A. & Zuker, M. (2004). Prediction of hybridization and melting for double-stranded nucleic acids. *Biophys J* 87, 215-226.

Forozan, F., Karhu, R., Kononen, J., Kallioniemi, A. & Kallioniemi, O. P. (1997). Genome screening by comparative genomic hybridization. *Trends Genet* 13, 405-409.

Freier, S. M., Alkema, D., Sinclair, A., Neilson, T. & Turner, D. H. (1985). Contributions of dangling end stacking and terminal base-pair formation to the stabilities of XGGCCp, XCCGGp, XGGCCYp, and XCCGGYp helixes. *Biochemistry* 24, 4533-4539.

Gellman, S. H., Haque, T. S. & Newcomb, L. F. (1996). New evidence that the hydrophobic effect and dispersion are not major driving forces for nucleotide base stacking. *Biophys J* 71, 3523-3526.

Haas, S. A., Hild, M., Wright, A. P., Hain, T., Talibi, D. & Vingron, M. (2003). Genome-scale design of PCR primers and long oligomers for DNA microarrays. *Nucleic Acids Res* 31, 5576-5581.

Heidenblad, M., Schoenmakers, E. F., Jonson, T., Gorunova, L., Veltman, J. A., van Kessel, A. G. & Hoglund, M. (2004). Genome-wide array-based comparative genomic hybridization reveals multiple amplification targets and novel homozygous deletions in pancreatic carcinoma cell lines. *Cancer Res* 64, 3052-3059.

Heidenblad, M., Lindgren, D., Veltman, J. A., Jonson, T., Mahlamaki, E. H., Gorunova, L., van Kessel, A. G., Schoenmakers, E. F. & Hoglund, M. (2005). Microarray analyses reveal strong influence of DNA copy number alterations on the transcriptional patterns in pancreatic cancer: implications for the interpretation of genomic amplifications. *Oncogene* 24, 1794-1801.

Hollox, E. J., Akrami, S. M. & Armour, J. A. (2002). DNA copy number analysis by MAPH: molecular diagnostic applications. *Expert Rev Mol Diagn* 2, 370-378.

Hughes, T. R., Mao, M., Jones, A. R. & other authors (2001). Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nat Biotechnol* 19, 342-347.

Kallioniemi, A., Kallioniemi, O. P., Sudar, D., Rutovitz, D., Gray, J. W., Waldman, F. & Pinkel, D. (1992). Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science* 258, 818-821.

Kallioniemi, A., Kallioniemi, O. P., Piper, J. & other authors (1994a). Detection and mapping of amplified DNA sequences in breast cancer by comparative genomic hybridization. *Proc Natl Acad Sci U S A* 91, 2156-2160.

Kallioniemi, O. P., Kallioniemi, A., Piper, J., Isola, J., Waldman, F. M., Gray, J. W. & Pinkel, D. (1994b). Optimizing comparative genomic hybridization for analysis of DNA sequence copy number changes in solid tumors. *Genes Chromosomes Cancer* 10, 231-243.

Kane, M. D., Jatkoe, T. A., Stumpf, C. R., Lu, J., Thomas, J. D. & Madore, S. J. (2000). Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays. *Nucleic Acids Res* 28, 4552-4557.

Kool, E. T. (2001). Hydrogen bonding, base stacking, and steric effects in dna replication. *Annu Rev Biophys Biomol Struct* 30, 1-22.

Lander, E. S., Linton, L. M., Birren, B. & other authors (2001). Initial sequencing and analysis of the human genome. *Nature* 409, 860-921.

Lane, A. N. & Jenkins, T. C. (2000). Thermodynamics of nucleic acids and their interactions with ligands. *Q Rev Biophys* 33, 255-306.

Langer-Safer, P. R., Levine, M. & Ward, D. C. (1982). Immunological method for mapping genes on Drosophila polytene chromosomes. *Proc Natl Acad Sci U S A* 79, 4381-4385.

Lucito, R., Healy, J., Alexander, J. & other authors (2003). Representational oligonucleotide microarray analysis: a high-resolution method to detect genome copy number variation. *Genome Res* 13, 2291-2305.

Makalowski, W. (2000). Genomic scrap yard: how genomes utilize all that junk. *Gene* 259, 61-67.

Makalowski, W. (2001). The human genome structure and organization. *Acta Biochim Pol* 48, 587-598.

Mantripragada, K. K., Buckley, P. G., Jarbo, C., Menzel, U. & Dumanski, J. P. (2003). Development of NF2 gene specific, strictly sequence defined diagnostic microarray for deletion detection. *J Mol Med* 81, 443-451.

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

Marmur, J. & Doty, P. (1962). Determination of the base composition of deoxyribonucleic acid from its thermal denaturation temperature. *J Mol Biol* 5, 109-118.

Ning, Z., Cox, A. J. & Mullikin, J. C. (2001). SSAHA: a fast search method for large DNA databases. *Genome Res* 11, 1725-1729.

Norberg, J. & Nilsson, L. (1998). Solvent influence on base stacking. *Biophys J* 74, 394-402.

Pinkel, D., Segraves, R., Sudar, D. & other authors (1998). High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat Genet* 20, 207-211.

Pollack, J. R., Perou, C. M., Alizadeh, A. A., Eisen, M. B., Pergamenschikov, A., Williams, C. F., Jeffrey, S. S., Botstein, D. & Brown, P. O. (1999). Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nat Genet* 23, 41-46.

Pollack, J. R., Sorlie, T., Perou, C. M. & other authors (2002). Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proc Natl Acad Sci U S A* 99, 12963-12968.

Relogio, A., Schwager, C., Richter, A., Ansorge, W. & Valcarcel, J. (2002). Optimization of oligonucleotide-based DNA microarrays. *Nucleic Acids Res* 30, e51.

Reymond, N., Charles, H., Duret, L., Calevro, F., Beslon, G. & Fayard, J. M. (2004). ROSO: optimizing oligonucleotide probes for microarrays. *Bioinformatics* 20, 271-273.

Rouillard, J. M., Zuker, M. & Gulari, E. (2003). OligoArray 2.0: design of oligonucleotide probes for DNA microarrays using a thermodynamic approach. *Nucleic Acids Res* 31, 3057-3062.

Saenger,W. Principles of Nucleic Acid Structure. *Springer Verlag*, New York, 1984.

SantaLucia, J., Jr., Allawi, H. T. & Seneviratne, P. A. (1996). Improved nearest-neighbor parameters for predicting DNA duplex stability. *Biochemistry* 35, 3555-3562.

SantaLucia, J., Jr. (1998). A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc Natl Acad Sci U S A* 95, 1460-1465.

SantaLucia, J., Jr. & Hicks, D. (2004). The thermodynamics of DNA structural motifs. *Annu Rev Biophys Biomol Struct* 33, 415-440.

Schildkraut, C. (1965). Dependence of the melting temperature of DNA on salt concentration. *Biopolymers* 3, 195-208.

Sellner, L. N. & Taylor, G. R. (2004). MLPA and MAPH: new techniques for detection of gene deletions. *Hum Mutat* 23, 413-419.

Solinas-Toldo, S., Lampel, S., Stilgenbauer, S., Nickolenko, J., Benner, A., Dohner, H., Cremer, T. & Lichter, P. (1997). Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances. *Genes Chromosomes Cancer* 20, 399-407.

Stillman, B. A. & Tonkinson, J. L. (2001). Expression microarray hybridization kinetics depend on length of the immobilized DNA but are independent of immobilization substrate. *Anal Biochem* 295, 149-157.

Sugimoto, N., Nakano, S., Yoneyama, M. & Honda, K. (1996). Improved thermodynamic parameters and helix initiation factor to predict stability of DNA duplexes. *Nucleic Acids Res* 24, 4501-4505.

Zhang, Z., Schwartz, S., Wagner, L. & Miller, W. (2000). A greedy algorithm for aligning DNA sequences. *J Comput Biol* 7, 203-214.

Zhou, J. (2003). Microarrays for bacterial detection and microbial community analysis. *Curr Opin Microbiol* 6, 288-294.

Wessendorf, S., Lichter, P., Schwanen, C., Fritz, B., Baudis, M., Walenta, K., Kloess, M., Dohner, H. & Bentz, M. (2001). Potential of chromosomal and matrix-based comparative genomic hybridization for molecular diagnostics in lymphomas. *Ann Hematol* 80 Suppl 3, B35-37.

LISAD

Lisa 1. Hübridiseerimisproovide disaini teostavad programmid.

Programm	Programmi töö kirjeldus
1. proc.pl	Laeb andmebaasist alla DNA nukleotiidsed järjestused, käivitab kindlas järjekorras ülejäänud programmid.
2. dust_lower (modifitseeritud DUST)	Maskeerib madala keerukusega piirkonnad proovi järjestustes (polü-N traktid, di-, tri- ja tetranukleotiidsed kordused).
3. gmasker (GenomeMasker)	Maskeerib nn "musta nimekirja" alusel järjestustes olevad üle-esindatud kuni 16 nukleotiidi pikkused sõnad.
4. fasta2primer3.pl	Kirjutab fasta failis olevad järjestused Primer3-e sisendformaati.
5. gm_primer3 (modifitseeritud Primer3)	Üritab igale järjestusele disainida etteantud parameetritele vastava *sense* ja *antisense* praimeri.
6. primer3_to_table	Kirjutab Primer3 tulemused ümber tabuleeritud formaadis faili.
7. gtester (GenomeTester)	Kontrollib praimerite seondumiskohti genoomis ja sekundaarsete produktide tekke võimalust.
8. gtester_filter	Eemaldab praimerid, millel on sekundaarseid seondumiskohti ja mis annavad sekundaarseid produkte.
9. prod2fas.pl	Kirjutab gtester_filter-i tulemused ümber fasta formaadis faili.
10. dust_score (modifitseeritud DUST)	Vastavalt sellele, kui palju on proovis maskeeritud nukleotiide, arvutab igale proovile kordusjärjestuste skoori.
11. cut_off.pl	Kirjutab liiga kõrge DUST-i skooriga proovide ID-d selleks ettenähtud faili.
12. exclude_from_fasta.pl	Eemaldab ühes failis olevate proovide ID-de alusel vastavad proovid teisest failist.
13. SSAHA	Otsib etteantud sõnapikkusega identseid alasid proovide ja genoomi vahel.
14. MegaBLAST	Otsib sarnaseid alasid proovide ja genoomi järjestuste vahel.
15. megablast_parser_b	Formaadib MegaBLAST-i tulemused tabuleeritud formaati.
16. parsers_parser.pl 17. wrong_chr_out.pl 18. wrong_doubled_out.pl 19. one_per_start.pl 20. extract_from_file.pl	Eemaldavad proovide hulgast need, mis annavad seondumisi ka mitteoodatud regioonides.
21. extract.sh	Kirjutab tulemused ümber fasta formaadis faili.
22. first_name.pl	Kirjutab "heade" proovide ID-d selleks ettenähtud faili.
23. first.pl 24. extrtract_lines_from_file.pl	Eraldavad erinevatest failidest "heade" proovide vajalikud andmed ja kirjutavad need tulemusfaili.
25. one_per_start_pos.pl	Eemaldab redundantsed proovid.

Tabelis on toodud hübridiseerimisproovide disainimiseks kasutatavad programmid ja nende lühikirjeldused.

Lisa 2. Näide proovide omadustele ja seondumistele vastavate parameetrite failist.

ID	ahel	pos	pikkus	GC%	Tm1	Tm2	AUC_GC	ratio_GC	AUC_Tm	ratio_Tm	min_GC	max_GC	dust_score	S_15	S_20	S_25	S_30	S_35
X5	1	2771276	424	47.64%	71.42	82.85	1752.38	39.11	298.17	14.60	13.33	73.33	26	1824	6	1	1	1
X7	-1	3365285	426	44.13%	69.98	81.42	883.33	28.33	29.59	2.71	16.67	63.33	26	807	139	2	2	2
X11	-1	2956134	468	41.67%	69.10	80.55	1154.76	26.12	79.76	4.46	13.33	70.00	34	16291	433	7	2	1
X17	1	3841429	532	42.48%	69.59	81.05	590.48	16.41	0.00	0.00	16.67	60.00	16	451	7	1	1	1
X18	1	3940527	483	46.58%	71.15	82.61	597.62	23.11	0.00	0.00	30.00	60.00	12	3633	51	10	3	1
X21	-1	4289524	448	34.38%	66.05	77.49	0.00	0.00	0.00	0.00	16.67	50.00	31	1863	23	2	2	2
X25	-1	4415278	502	28.88%	63.94	75.40	188.10	5.60	0.00	0.00	6.67	60.00	19	3188	18	1	1	1
X34	-1	5952843	518	32.82%	65.59	77.06	169.05	3.82	0.00	0.00	10.00	63.33	22	21505	36	2	2	2
X35	-1	5853123	483	45.55%	70.73	82.18	780.95	26.78	0.00	0.00	20.00	63.33	14	1593	236	49	16	1
X37	-1	5580295	408	33.58%	65.59	77.02	414.29	7.73	71.36	5.41	6.67	66.67	45	119384	87	3	3	2
X39	-1	5379623	494	35.43%	66.61	78.06	0.00	0.00	0.00	0.00	16.67	53.33	37	22476	11302	796	4	1
X42	1	5059745	500	34.80%	66.37	77.82	0.00	0.00	0.00	0.00	16.67	53.33	12	566	69	18	3	2
X49	1	7103806	474	52.32%	73.48	84.93	2047.62	51.54	36.01	3.30	33.33	70.00	10	169	82	39	4	1
X52	-1	7337285	431	34.80%	66.17	77.61	0.00	0.00	0.00	0.00	16.67	53.33	16	587	4	2	2	2
X53	1	7499669	405	38.27%	67.51	78.93	52.38	3.64	0.00	0.00	16.67	53.33	16	360	2	1	1	1
X55	1	7700591	440	43.41%	69.73	81.17	885.71	30.48	0.00	0.00	20.00	63.33	12	320	15	4	3	3
X58	1	8000729	442	31.22%	64.74	76.18	0.00	0.00	0.00	0.00	13.33	46.67	13	1390	3	2	1	1
X60	1	8145274	486	32.10%	65.22	76.68	73.81	2.36	0.00	0.00	13.33	56.67	36	4507	18	1	1	1
X63	1	8443337	408	34.07%	65.79	77.22	142.86	6.19	0.00	0.00	13.33	56.67	12	933	5	1	1	1
X65	1	8643678	425	38.59%	67.71	79.14	497.62	15.56	0.00	0.00	10.00	60.00	22	11218	366	15	3	2

ID	S_40	S_45	S_50	MB_20	MB_25	MB_30	MB_35	MB_40	MB_45	MB_50	DG_17.8	DG_24.1	DG_30.3	DG_36.7	DG_43	DG_49.3	DG_55.5	DG_61.8
X5	1	1	1	174	61	3	2	1	1	1	600	5	1	1	1	1	1	1
X7	2	1	1	1088	1407	37	4	2	2	2	775	4	2	2	2	2	1	1
X11	1	1	1	16300	5401	1792	114	113	83	46	6065	59	8	3	1	1	1	1
X17	1	1	1	287	21	3	2	1	1	1	3245	53	1	1	1	1	1	1
X18	1	1	1	232	1024	1266	1175	697	1185	277	2194	64	12	3	1	1	1	1
X21	2	2	1	419	37	2	2	1	2	2	55	2	2	2	2	2	1	1
X25	1	1	1	1413	44	10	2	1	1	1	65	3	1	1	1	1	1	1

X34	1	1	1	519	646	13	6	4	3	2	286	3	2	2	1	1	1	1
X35	1	1	1	2324	879	1138	1060	1316	1084	931	2741	307	52	19	1	1	1	1
X37	2	2	2	10006	496	20	6	4	4	4	643	3	3	2	2	2	2	2
X39	1	1	1	26800	24782	3806	291	20	2	1	13458	3527	38	1	1	1	1	1
X42	1	1	1	443	641	755	603	625	721	445	701	47	18	3	1	1	1	1
X49	1	1	1	142	249	141	88	72	70	38	495	87	24	9	2	1	1	1
X52	1	1	1	382	15	3	2	1	2	2	22	2	2	2	2	1	1	1
X53	1	1	1	111	25	4	2	1	1	1	600	1	1	1	1	1	1	1
X55	1	1	1	88	119	93	175	104	46	15	424	13	3	3	3	1	1	1
X55	1	1	1	293	36	7	4	1	1	1	141	2	1	1	1	1	1	1
X60	1	1	1	524	56	6	2	2	1	1	480	2	1	1	1	1	1	1
X63	1	1	1	219	53	8	2	1	1	1	816	2	1	1	1	1	1	1
X65	1	1	1	2434	4790	4564	4828	756	791	685	17461	279	16	3	1	1	1	1

Veergudes on toodud erinevatele proovidele vastavad omadused ja seondumiste arvud. Veergudes on toodud proovide ***ID***-d, kummalt **ahel**alt genoomis nad on disainitud, 5' otsa esimese nukleotiidi **pos**itsioon inimese X kromosoomil, proovide **pikkus** ja **GC** nukleotiidide sisaldus. T_{m1} on monovalentse soola kontsentratsiooni arvestav sulamistemperatuur, T_{m2} on lihtsama valemi järgi arvutatud sulamistemperatuur. ***AUC_GC*** on parameeter, mis väljendab seda, kui suur osa proovi järjestusest ja millisel määral jääb ülespoole kindlaksmääratud GC% ülempiiri (50%). ***ratio_GC*** on akende (21 nt) arv, mis on üle GC% ülempiiri (50%) jagatuna kõikide akende arvuga (proovi pikkus - akna pikkus + 1) korrutatud 100-ga. ***AUC_Tm*** väljendab seda, kui suur osa proovi järjestusest ja millisel määral jääb ülespoole kindlaksmääratud sulamistemperatuuri ülempiiri (70^{o} C). ***ratio_Tm*** on akende (21 nt) arv, mis on üle T_m ülempiiri (70^{o} C) jagatuna kõikide akende arvuga (proovi pikkus - akna pikkus + 1) korrutatud 100-ga; T_m on arvutatud samamoodi kui T_{m1}. ***min_GC*** on proovi sees kindla pikkusega aknas (30 nt) esinev minimaalne GC%. ***max_GC*** on proovi sees kindla pikkusega aknas (30 nt) esinev maksimaalne GC%. ***dust_score*** on vastavalt proovi sees olevatele di-, tri- ja tetranukleotiidsetele kordustele ja polü-N traktidele arvutatud skoor; mida rohkem madala keerukusega kordusi proovis, seda suurem skoor. ***S_N*** on programmiga SSAHA leitud seondumiste arv. Leitud on iga proovi jaoks *N* nukleotiidi pikkune alamstring, mis annab kõige rohkem seondumisi inimese genoomis, parameetrina on kirjas selle alamstringi seondumiste arv inimese genoomis. Seondumiskohtade otsimisel nõuti 100% identsust. ***MB_N*** on programmiga MegaBLAST leitud seondumiste arv. Leitud on iga proovi jaoks *N* nukleotiidi pikkune alamstring, mis annab kõige rohkem seondumisi inimese genoomis, parameetrina on kirjas selle alamstringi seondumiste arv inimese genoomis. Seondumiskohtade otsimisel nõuti 75% identsust. ***DG_N*** on programmiga SSAHA leitud seondumiste arv seondumistugevuse ΔG_{37} järgi. Leitud on iga proovi jaoks -*N* (kcal/mol) suuruse seondumistugevusega alamstring, mis annab kõige rohkem seondumisi inimese genoomis, parameetrina on kirjas selle alamstringi seondumiste arv inimese genoomis.