BIOLOOGIA-GEOGRAAFIATEADUSKOND
MOLEKULAAR- JA RAKUBIOLOOGIA INSTITUUT
BIOINFORMAATIKA ÕPPETOOL

Priit Palta

Hübridiseerimisproovide automaatne disain inimese genoomsetele järjestustele

Keskastme lõputöö

Juhendaja: Maido Remm, prof., PhD

TARTU
2004

Sisukord

Sisukord 2
Lühendid ja mõisted. 3
Sissejuhatus. 4
I KIRJANDUSE ÜLEVAADE 5
1. Mikrokiibid. 5
1.1 Mikrokiipide lühitutvustus. 5
1.2 Mikrokiipide rakendusalad. 5
2. Ülevaade erinevatest aberratsioonide tuvastamise meetoditest. 6
2.1 *In situ* hübridisatsioon. 7
2.1.1 Radioaktiivne *in situ* hübridisatsioon. 8
2.1.2 Mitteradioaktiivne *in situ* hübridisatsioon. 8
2.1.3 Mitme värvi fluorestsents *in situ* hübridisatsioon. 9
2.2 Võrdlev genoomne *in situ* hübridisatsioon. 10
2.2.1 Traditsiooniline võrdlev genoomne *in situ* hübridisatsioon. 10
2.2.2 Mikrokiipidel põhinev võrdlev genoomne *in situ* hübridisatsioon. 12
2.3 Multipleks amplifitseeritavate proovide hübridisatsioon. 15
3. Hübridiseerimisproovide disain. 16
3.1 Hübridiseerimisproovide olulised parameetrid. 17
3.1.1 Hübridiseerimisproovide pikkus. 17
3.1.2 Hübridiseerimisproovide GC nukleotiidide sisaldus. 17
3.1.3 Hübridiseerimisproovide sekundaarstruktuur. 18
3.1.4 Hübridiseerimisproovide sulamistemperatuur. 18
3.2 Hübridiseerimisproovide seostumiskohtade ja unikaalsuse kontroll. 19
3.3 Hübridiseerimisproovide disaini programmid. 21
4. PCR-i praimerite disain hübridiseerimisproovidele. 21
4.1 PCR-i praimerite olulised parameetrid. 22
4.1.1 Praimerite sulamistemperatuur. 22
4.1.2 PRC-i produkti pikkus. 22
4.1.3 Praimerite GC nukleotiidide sisaldus. 23
4.1.4 Praimerite pikkus. 23
4.2 Praimerite seondumiskohtade kontroll. 23
4.3 Praimerite disaini programmid. 24
II PRAKTILINE TÖÖ 26
Töö eesmärgid. 26
Kasutatud meetodid. 27
1. Andmete päritolu ja struktuur. 27
2. Kasutatud programmide kirjeldus. 27
3. Tööks vajalik mälumaht ja aeg. 29
Tulemused. 32
1. Veebi kasutajaliides. 32
2. Algoritmi kirjeldus. 33
3. Tulemuste kirjeldus. 35
Arutelu. 40
Kokkuvõte. 42
Summary. 43
Kasutatud kirjandus. 44

Lühendid ja mõisted.

BAC	bakteri kunstlik kromosoom
BIO	biotiin
cDNA	mRNA-lt pöördtranskriptaasi abil sünteesitud DNA
Cy-värvid	tsüaniinvärvid
DIG	digoksigeniin
DNA	desoksüribonukleiinhape
FITC	fluorestseiin
Gb	gigabait
in silico	arvutis
in situ	algse materjaliga
in vitro	katseklaasis
kb	*kilobase*, 10^3 aluspaari
Mb	megabait
mRNA	*messenger*-RNA
P1	bakteriofaag P1
PAC	P1 kunstlik kromosoom
PCR	polümeraasi ahelreaktsioon
RNA	ribonukleiinhape
TRITC	rodamiin
YAC	pärmi kunstlik kromosoom

Sissejuhatus.

Alates inimese kogu genoomi esmase kavandi väljatulemisest aastal 2000 on inimese genoomi uuringud aasta-aastalt kasvanud. Järjestuste uurimisega ja analüüsimisega on võimalus tegeleda peaaegu kõigil, kuna andmed on vabalt kättesaadavad. Inimese genoomi järjestuse teadmine võimaldab uurijatel lisaks kõigele muule otsida ka geneetiliste haiguste põhjuseid DNA tasandil, aidates paremini aru saada inimese kui terviku puhul toimivatest põhjus-tagajärg sündmustest. Teades aga vähimaidki pisiasju inimesse puhul, oskame paremini näha ennast tervikuna ja vastupidi. Meil on võimalus otsida ja leida sarnasusi ning erinevusi enda ja teiste liikide vahel, mis saavad alguse madalaimast võimalikust korrastatult edasikanduvast süsteemist – pärilikkuseaine DNA järjestusest.

Uurides DNA järjestusi, vajame me meetodeid, mis aitaksid meil ennast paremini organiseerida inimese 3,2 miljardi aluspaari seas. Sellise andmemahuga töötamiseks vajame me automaatseid tööriistasid, mis teostaksid just meile vajaliku informatsiooni töötlust ja analüüsi. Samuti vajame me tööriistasid, mis aitaksid meile teadaolevat genoomi infot üle kanda ja rakendada praktilistes lahendustes. Et sellised lahendused toimiksid reaalses elus võimalikult hästi, on vajalik nende katsetamine, kontrollimine ja optimeerimine *in silico*.

I KIRJANDUSE ÜLEVAADE

1. Mikrokiibid.

Tänapäeval on mikrokiibid (või ka DNA/RNA kiibid, biokiibid, geenikiibid) laialdaselt kasutusel kogu geenitehnoloogiaga aga eriti biotehnoloogiaga seotud aladel. Mikrokiibid teevad võimalikuks suure hulga katsete tegemise paralleelselt ja ka üheaegselt. Pidades silmas üha kasvavat eksperimentide arvu ja mahtu, on mikrokiibid samm tuleviku ja genoomi-uuringute ratsionaliseerimise suunas.

1.1 Mikrokiipide lühitutvustus.

Mikrokiip on korrastatud struktuuriga tahke kandja, millele planaarsele pinnale on kinnitatud mikroskoopilised molekulid (DNA, RNA, valgud), milledele seonduvad neile komplementaarsed molekulid. Vastavalt definitsioonile on mikrokiip ükskõik milline analüütiline seade, mis on: a) korrastatud, b) mikroskoopiline, c) planaarne ja d) spetsiifiline (Schena, 2003).
Mikrokiipide tehnoloogia põhineb biomolekulide hübridiseerimisel, mille puhul kahe molekuli komplementaarsus viib nende seostumisele. Üks molekulidest on tavaliselt seotud tahkele kandjale ehk kiibile (Gabig *et al*., 2001).

1.2 Mikrokiipide rakendusalad.

Kõige esimesed mikrokiipidega tehtud eksperimendid viidi läbi cDNA mikrokiipidega juba 1990-ndate alguses (Schena *et al*., 1995) ja pärast seda on mikrokiipide tehnoloogia ja meetodid arenenud järjest kasvava kiirusega.

Tänapäeval kasutatakse mikrokiipe kõige rohkem geeniekspressiooni uurimiseks, mutatsioonide ja polümorfismide leidmiseks, nende kaardistamiseks, primaarstruktuuri määramiseks ja ka liikide eristamiseks ja määramiseks. Protsentuaalselt on kõigist mikrokiipidega seotud publikatsioonidest ülekaalukalt

esikohal geeniekspressiooni uuringud (81,5 %), järgnevad genotüpiseerimine (12 %), koe analüüsid (4,6 %) ja valgu- uuringud (1 %) (Schena *et al*., 2003).

Vastavalt selle järgi, millist biomolekuli mikrokiibid kannavad, jaotatakse nad mitmeks erinevaks tüübiks. Kõige enam kasutatakse cDNA-d kandvaid mikrokiipe (65 % kõikidest vastavatest publikatsioonidest), oligonukleotiidide mikrokiipe (26 %), koe-mikrokiipe (6,6 %), valguproovidega mikrokiipe (1,4 %) ja teisi erinevaid proove kandvaid mikrokiipe (1 %) (Schena *et al*., 2003).

Erinevaid molekule kandvate mikrokiipidega saab küll tihtipeale teha sarnaseid analüüse, kuid nende tootmise meetodite vahel on fundamentaalsed erinevused. cDNA kiipide tootmisel immobiliseeritakse robotite poolt tahkele kandjale suhteliselt pikad DNA molekulid, enamasti pikemad kui 100 nukleotiidi. Proovide cDNA-d paljundatakse enamasti PCR-ga. Seda tüüpi kiibid on kasutusel ulatuslikes skriinimis- ja ekspressioonianalüüsides. Oligonukleotiidide kiipidel olevad järjestused sünteesitakse kas *in situ* valguse poolt kontrollitud keemiliste reaktsioonidega või eraldi oligote sünteesi teel, millele järgneb nende kinnitamine klaasile või muule kandjale. Koe mikrokiipe, mille puhul märgitud proovid (enamasti riboproovid) hübridiseeritakse koelõikudes oleva RNA-ga, kasutatakse mRNA ekspressiooni uurimiseks. Valgu mikrokiipidega uuritakse kompleksseid süsteeme; näiteks on nii võimalik uurida valkude, DNA ja RNA seondumissaite.

Olenemata sellest, milliste proovidega on tegemist, on lõpp-tulemusena spetsiifilised molekulid kiibil ja eksponeeritud lahuses olevatele märgitud molekulidele. Väga täpselt määratletud soolasisaldusega lahuses ja kindla temperatuuri juures hübridiseerub kiibil olev molekul ainult temale komplementaarse molekuliga, mitte aga ükskõik millisele teise molekuliga, mis on lahuses. Valesti paardunud molekulide vahel tekkinud vesiniksidemed on lühikese elueaga, samas kui õigesti paardunud molekulide vahel jäävad sidemed stabiilseks, andes determineeritava signaali (Miller *et al*., 2001).

2. Ülevaade erinevatest aberratsioonide tuvastamise meetoditest.

Kuigi väga paljusid geneetilisi haiguseid põhjustavad punktmutatsioonid, on viimasel ajal järjest rohkem leitud, et muutused DNA koopiaarvus ja järjestuste inserteerumine/deleteerumine on tõenäoliselt paljude geneetiliste haiguste põhjustajaks.

Sellised muutused võivad olla väga erineva suurusega, alates tervete kromosoomide ja kromosoomialade kuni geenide ja üksikute eksonite kordistumisest/kaotsiminekust. On näidatud, et geenidoosi ja koopiaarvu muutused on iseloomulikud kasvajarakkudele ja otseses seoses arenguhälvete ja/või vaimse alaarengu esinemisega (Pollack *et al*., 2002; Albertson *et al*., 2003a). Seetõttu on väga oluline välja töötada ja arendada meetodeid selliste muutuste detekteerimiseks.

On olemas mitmeid erinevaid klassikalisi meetodeid, mis võimaldavad uurida erineva pikkusega koopiaarvu muutusi. Kuid enamasti on nendega võimalik määrata vaid suuremaid ja seni teadaolevaid ja põhjalikult uuritud muutusi genoomis. Viimastel aastatel on aga rakendatud ka uusi, mikrokiipidel põhinevad meetodeid, mis oluliselt kiirendavad ja lihtsustavad geenidoosi ja kromosoomiaberratsioonide uurimist ja diagnostikat ning aitavad meil seega ka paremini mõista selliste muutuste ja nendega kaasnevate võimalike haiguste olemust.

2.1 *In situ* hübridisatsioon.

In situ hübridisatsioon e. ISH (*in situ hybridization*) on suhteliselt vana, kuid viimasel aastakümnel väga kiiresti arenenud tehnoloogia. Meetod võimaldab tänapäeval metafaasi kromosoomides kaardistada juba 1-2 kb pikkusi DNA järjestusi, interfaasi kromatiinis aga isegi mõnesaja aluspaari pikkusi RNA või DNA järjestusi. Hübridiseerimise eesmärgiks on proovi (*probe*) teadaolevat järjestust kasutades tuvastada homoloogilisi järjestusi märklaud-nukleiinhappes (*target*). Tüüpiliselt kasutatakse *in-situ* hübridiseerimisel kloneeritud ja denatureeritud DNA fragmente, mis märgistatakse radioaktiivse, fluorestseeruva või antigeense märkega ja hübridiseeritakse vastaval alusel olevale fikseeritud metafaasi kromosoomile. Edasi saab autoradiograafiliselt detekteerida kohad, kuhu *in-situ* hübridiseerimise käigus on seostunud radioaktiivselt märgitud proovid. Fluorestsentsiga või ensüüm-vahendatud detekteerimisega leitakse kohad, kuhu on hübridiseerunud mitteradioaktiivselt, fluorokroomidega või antikehadega märgitud proovid.

2.1.1 Radioaktiivne *in situ* hübridisatsioon.

Radioaktiivne *in* situ hübridisatsioon võeti esmakordselt kasutusele 1969-ndal aastal (Pardue *et al*., 1969). Radioaktiivse *in situ* hübridisatsiooni korral märgistatakse uuritav proov radioaktiivse märkega, denatureeritakse ja hübridiseeritakse denatureeritud metafaasi kromosoomidele. Seejärel kromosoomipreparaati pestakse, et eemaldada nõrgalt seostunud või üldse mitte hübridiseerunud proovid. Pärast pesemist kaetakse preparaat fotoemulsiooniga ja inkubeeritakse pimedas kuni paar nädalat. Proovidega integreeritud radioaktiivne märgis indutseerib hõbedagraanulite ergastumise fotoemulsioonis, mida saab kvantitatiivselt detekteerida. Emulsiooniga detekteerimise asemel võib kasutada ka röntgenfilmi autoradiograafiat, mis lühendab inkubeerimise aja mõnele päevale. On olemas ka kommertsiaalselt pakutavad tehnoloogiad (näiteks detekteerimine „PhosphoImager" aparatuuri abil), mis on veelgi kiiremad. Radioaktiivse *in situ* hübridisatsiooni eelisteks on suur tundlikkus, proove on suhteliselt lihtne märgistada ja hiljem detekteerida. Meetodi puudusteks on selle ohtlikkus (töötamine radioaktiivsete isotoopidega) ja ajanõudlikkus.

2.1.2 Mitteradioaktiivne *in situ* hübridisatsioon.

Mitteradioaktiivne *in situ* hübridisatsioon ehk NISH (*non-radioactive in situ hybridization*) on olnud kasutusel 80-ndate aastate algusest, kui radioaktiivse märgise asemel hakati proove märgistama biotiiniga ja töötati välja uued, tundlikumad meetodid märgiste detekteerimiseks (Pinkel *et al*., 1986). Mitteradioaktiivse *in situ* hübridisatsiooni puhul märgitakse proovid kas fluorokroomiga või retseptormolekuliga. Võimalik on ka erinevate märgiste omavaheline kombineerimine. Hübridiseerimine ja pesemine toimub üldjuhul samamoodi, kui radioaktiivse *in situ* hübridiseerimise puhul, detekteerimine on aga fundamentaalselt erinev. Kui proovid on märgitud vahetult nukleotiidile seotud fluorokroomiga (Cy-värvid, fluorestseiin, rodamiin), toimub detekteerimine kohe pärast hübridiseerimist fluorestsentsmikroskoobi või spetsiaalse CCD (*charge-coupled device*) kaamera ja vastava digitaalset andmetöötlust võimaldava tarkvara abil. Kui proovid on seotud retseptormolekulidega (biotiin, digoksigeniin), tuleb neid enne detekteerimist töödelda vastava retseptormolekuli spetsiifilise

antikehaga või retseptormolekulile afiinse ühendiga (näiteks biotiini puhul avidiini või streptavidiiniga), mis on konjugeeritud fluorokroomiga. Seejärel on võimalik hübridiseerunud proove detekteerida CCD kaamera ja vastava tarkvara abil. Kui kasutatakse kahte eelpool mainitud meetodit, on tegemist fluorestsents *in situ* hübridisatsiooni ehk FISH-iga (*fluorescence in situ hybridization*). Juhul kui retseptormolekuli vastane antikeha või retseptori molekulile afiinne ühend ei ole seotud fluorokroomiga, vaid mõne ensüümiga (näiteks aluselise fosfataasi või peroksüdaasiga), lisatakse proovidele nende detekteerimiseks veel spetsiaalset substraati, mis ensümaatilise reaktsiooni käigus lokaalselt värvub. Seda spetsiifilist meetodit nimetatakse ensüüm-vahendatud *in situ* hübridisatsiooniks ehk EISH-iks (*enzyme-linked in situ hybridization*). Ensüüm-vahendatud *in situ* hübridisatsiooni puhul võib ensüümi (aluselise fosfataasi) substraadina kasutada ka spetsiaalset ainet – luminooli, mille defosforüleerimisel tekib tasapisi lagunev ja valgust eraldav produkt. Saadav signaal detekteeritakse visualiseerimisega fotopaberile või röntgenfilmile.
Mitteradioaktiivse *in situ* hübridisatsiooni eelisteks on kiire visualiseerimine ja suhteliselt kõrge eraldatavus ning võrreldes radioaktiivse *in situ* hübridisatsiooniga suurem ohutus. Puuduseks on aga väiksem tundlikkus.

2.1.3 Mitme värvi fluorestsents *in situ* hübridisatsioon.

1980-ndate lõpus modifitseeriti traditsioonilist *in situ* hübridisatsiooni nii, et ühes eksperimendis sai samaaegselt kasutada mitmeid erinevalt märgistatud DNA proove ning mitmete erinevate fluorokroomidega märgistatud detekteerimissüsteeme (Nederlof *et al*., 1989). Meetodit hakati nimetama mitme värvi fluorestsents *in situ* hübridisatsiooniks (*multicolor fluorescence in situ hybridization*). Meetodi põhierinevuseks fluorestsents *in situ* hübridisatsioonist on see, et korraga kasutatakse kombineeritult mitmeid erinevaid fluorokroome. Seetõttu on detekteerimiseks peale fluorestsentsmikroskoobi ja CCD kaamera vajalikud ka spetsiaalsed filtrid ja digitaaltöötluse programmid. Mitme värvi fluorestsents *in situ* hübridisatsiooni edasiarendusteks võib pidada MFISH-i (*multiplex fluorescence in situ hybridization*) ja spektraalse karüotüpiseerimise ehk SKY (*spectral karyotyping*) meetodeid. Neid kasutatakse peamiselt kromosoomide ja nende osade diferentsiaalseks värvimiseks ja

nad erinevad mitme värvi fluorestsents *in situ* hübridisatsioonist ja teineteisest minimaalselt modifitseeritud detekteerimise meetodite poolest.

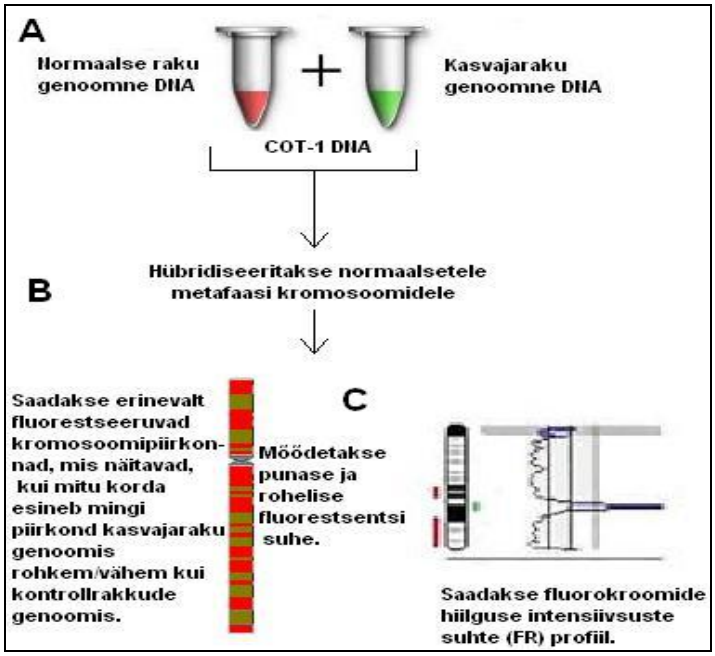
2.2 Võrdlev genoomne *in situ* hübridisatsioon.

Võrdlev genoomne *in situ* hübridisatsioon ehk CGH (*comparative genomic in situ hybridization*) on molekulaarne tsütogeneetikal põhinev meetod, mis võimaldab kogu kasvajaraku genoomis avastada seal toimunud aberratsioone: geeni koopiaarvu muutusi, kromosoomialade juurdetulekuid ja kadusid. Meetod võeti kasutusele 1990-ndate alguses (Kallioniemi *et al.*, 1992) ja põhineb kromosoomide pöördvärvimise ideel, mille puhul erinevate fluorokroomidega märgistatud testi-DNA ja kontroll-DNA hübridiseeritakse korraga normaalsetele märklaud-kromosoomidele. Piirkonnad, kus on toimunud muutused DNA-s (näiteks deletsioonid, duplikatsioonid, amplifikatsioonid) on nähtavad kui kahe fluorokroomi muutunud intensiivsuse suhtega alad märklaud-kromosoomidel. Võrdleva genoomse *in situ* hübridisatsiooni eeliseks teiste koopiaarvu tuvastavate meetodite ees on see, et ta võimaldab korraga uurida kogu genoomis toimunud muutusi ja on palju kiirem. Tänapäeval on lisaks traditsioonilisele võrdlevale genoomsele *in situ* hübridisatsioonile lisandunud ka mikrokiipidel põhinev võrdlev genoomne *in situ* hübridisatsioon, mis ühendab traditsioonilised tsütogeneetilised ja moodsad mikrokiipidel põhinevad meetodid.

2.2.1 Traditsiooniline võrdlev genoomne *in situ* hübridisatsioon.

Traditsiooniline võrdlev genoomne *in situ* hübridisatsioon (*chromosome* CGH) põhineb meetodil, mille puhul kasvajarakkudest saadud ja märgistatud (näiteks biotinileeritud) kogu genoomne DNA (testi-DNA) segatakse normaalsetest rakkudest saadud ja teisiti märgistatud (näiteks digoksigeniiniga) genoomse DNA-ga (kontroll-DNA) vahekorras üks-ühele. Mõned autorid on näidanud, et parem on siiski kasutada otseselt fluorokroomidega märgitud DNA-d, mis tagab tulemuste parema kvaliteedi (Karhu *et al.*, 1997). Saadud segu hübridiseeritakse normaalsetele metafaasi märgistamata kromosoomidele, kusjuures segule lisatakse ka liigispetsiifiline

märgistamata konkurent-DNA (*competitor DNA*), et vähendada märgitud proovide mittespetsiifilist seondumist. Sellist lähenemist nimetatakse kromosomaalseks *in situ* supressiooniks ehk CISS-ks (*chromosomal in situ supression*) ja see aitab tunduvalt vähendada proovide mittespetsiifilist seondumist näiteks kordusjärjestustele. Hübridiseerunud proovid detekteeritakse erinevate fluorokroomidega seotud märgetele afiinsete molekulidega (näitele vastavalt FITC-avidiiniga ja TRITC-antidigoksigeniiniga). Seejärel mõõdetakse fluorokroomide hiilgus ja arvutatakse iga kromosoomisegmendi jaoks rohelise ja punase (FITC/TRITC) hiilguse intensiivsuste suhe ehk FR (*fluorescence ratio*). Saadud FR väärtus näitab kromosoomide või nende osade esindatust testi genoomis kontroll genoomi suhtes ehk lahtimõtestatult; kui palju kordi esineb mingi kindel piirkond kasvajarakkudes rohkem/vähem, kui normaalsetes rakkudes. Kromosoomide ja/või nende osade kadumist või lisandumist saab määrata ka vaid märgitud testi-DNA CISS-hübridiseerimisel normaalsetele kromosoomidele. Teisiti märgitud kontroll-DNA lisamine aga kindlustab selle, et tasakaalustamata (monosoomiad, trisoomiad) ja tasakaalustatud (disoomiad) kromosoomide kohta saadakse selgelt eristuvad FR-väärtused.



Joonis 1. Traditsiooniline võrdlev genoomne hübridisatsioon. Kõigepealt eraldatakse normaalse ja kasvajaraku DNA-d (A), mis kohübridiseeritakse metafaasi kromosoomidele (B) ja seejärel detekteeritakse (C).

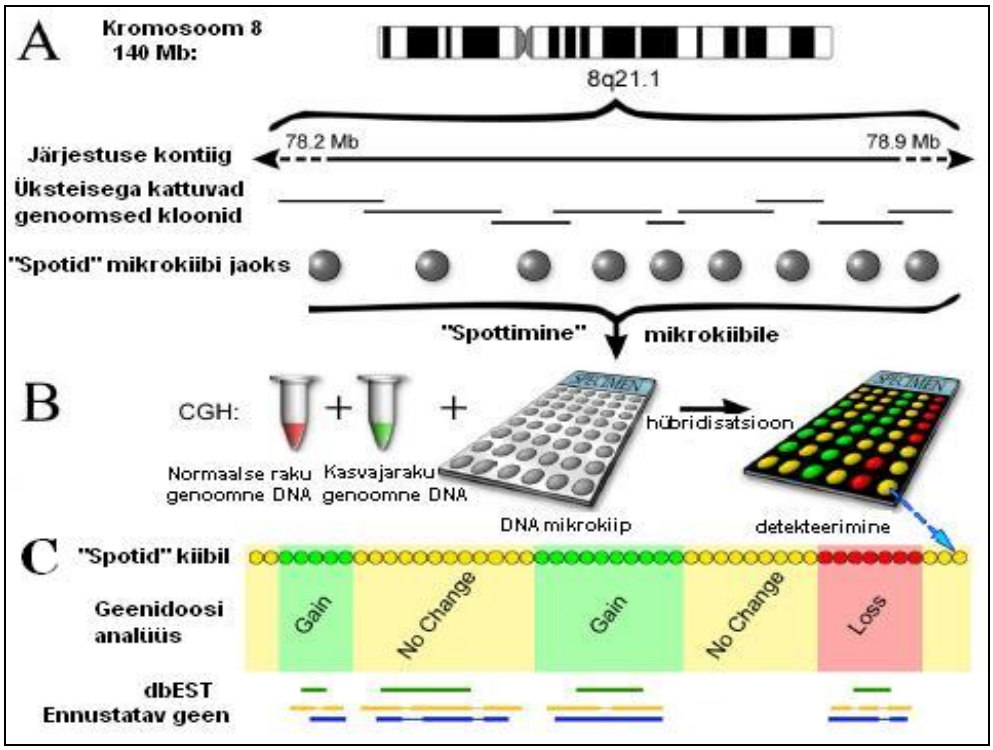
Võrdleva genoomse *in situ* hübridisatsiooni eelisteks teiste *in situ* hübridisatsiooni meetodite ees on suurem kiirus ja see, et kromosoomide ja nende alade lisandumist ja/või kadumist saab testi-genoomis hinnata ilma kromosoomipreparaati tegemata. See ongi kasvajarakkude uurimise puhul oluline, kuna tihtipeale on kromosoomipreparaadi tegemine kasvajarakkude segipaisatud genoomidest raskendatud, kui mitte võimatu. Puuduseks on aga võrdlemisi limiteeritud resolutsioon (võime positsioneerida muutusi), mis kromosoomide kõrge kondenseerituse tõttu ei ületa kadude uurimise puhul 10 Mb ja juurdetulekute puhul 2 Mb (Beheshti *et al*., 2002), andes küll kindlad koordinaadid edasisteks uuringuteks, kuid mitte piisava täpsusega. Puuduseks on ka see, et kogu protsess on ainult osaliselt automatiseeritud ja lõpuks peab ikkagi kogenud tsütogeneetik leidma muutunud regioonid kromosoomides. Lisaks veel see, et võrdleva genoomse *in situ* hübridisatsiooniga saab hinnata vaid kromosoomimaterjali lisandumist ja kadumist. Ümberpaigutused translokatsioonide ja inversioonide näol jäävad aga avastamata (Kallioniemi *et al*., 1992, 1994; Forozan *et al*., 1997).

2.2.2 Mikrokiipidel põhinev võrdlev genoomne *in situ* hübridisatsioon.

Mikrokiipidel põhineva võrdleva genoomse *in situ* hübridisatsiooni (*array CGH, matrix CGH*) puhul kasutatakse metafaasi kromosoomide asemel hübridisatsiooni märklaudadena (*target*) mikrokiibile seotud järjestusi. Õige regiooni piisava esindatuse korral mikrokiibil on selle meetodiga võimalik tunduvalt tõsta süsteemi eraldusvõimet, positsioneerimaks muutunud regioone. Hübridisatsiooni märklaudadena kasutatakse enamasti genoomsete järjestuste kloone, cDNA järjestusi või ka oligonukleotiide, mis on korrastatult ja kindlas järjekorras seotud mikrokiipidele. Nagu ka traditsioonilise võrdleva genoomse *in situ* hübridisatsiooni puhul, toimub kõigepealt proovide (testi-DNA ja kontroll-DNA) erinev märkimine (enamasti Cy-värvidega). Seejärel proovid kohübridiseeritakse mikrokiibil olevatele järjestustele, lisades mittespetsiifiliste seondumiste ärahoidmiseks konkurent-DNA-d. Kiibid pestakse, et eemaldada mitteseondunud ja mittespetsiifiliselt seondunud nukleiinhappejärjestused. Kiipidel olevad signaalid detekteeritakse ja analüüsitakse (Solinas-Toldo *et al*., 1997).

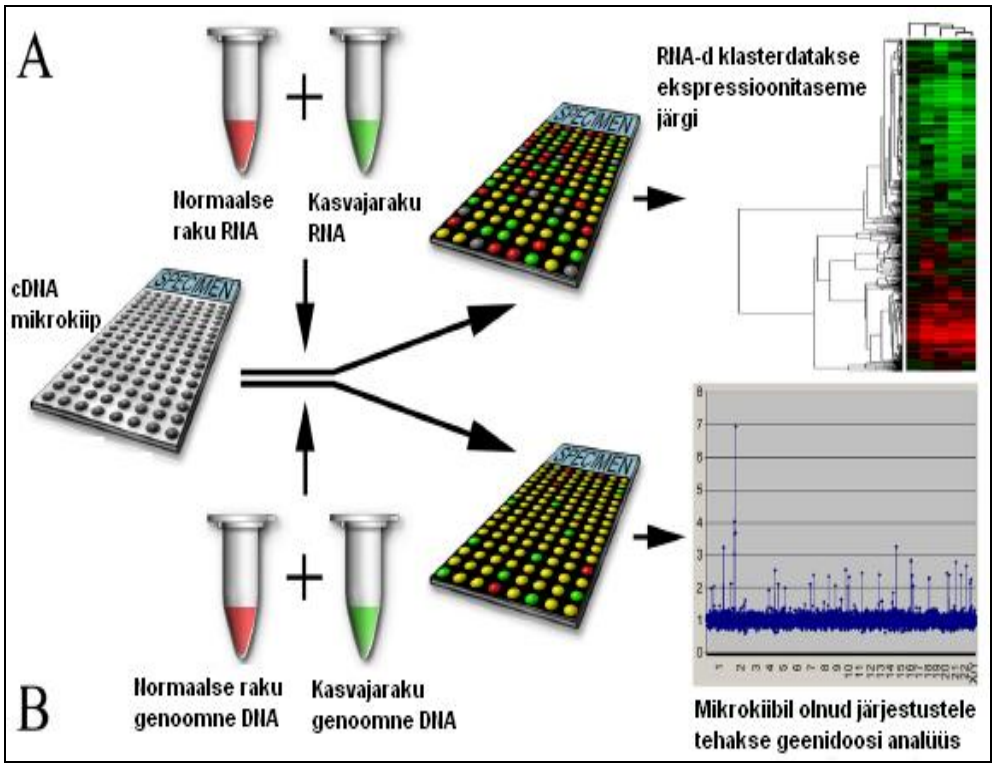
Genoomseid järjestusi kandvate mikrokiipide märklaud-järjestused pärinevad tavaliselt genoomsetest kloonidest, mida paljundatakse spetsiaalsetes vektorites.

Sellised vektorid on YAC-id (sisestatava DNA suurus 0,2-2 Mb), BAC-id (kuni 300 kb), P1-d (~ 70-100 kb), PAC-id (~ 130-150 kb) ja kosmiidid (~ 30-45 kb), mis on mitu suurusjärku väiksemad, kui traditsioonilises lähenemises kasutatud terved kromosoomid. Märklaud-järjestuse lühenemine toob kaasa resolutsiooni tõusu detekteerimisel, tagades muutunud koopiaarvuga geenide täpsema lokaliseerimise. Samuti on võimalik huvipakkuvaid regioone piisava ja ühtlase katvusega mikrokiibile panna, mis parandab tulemuste täpsust. On näidatud, et on võimalik avastada 100 kb (Wessendorf *et al*., 2001) ja isegi 50kb (Albertson, 2003a) suurusi aberratsioone uuritavas genoomis. 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 (Pinkel *et al*., 1998; Buckley *et al*., 2002; Albertson *et al*., 2003b).



Joonis 2. Genoomsete järjestustega mikrokiibil läbiviidav võrdlev genoomne *in situ* hübridisatsioon. Huvipakkuv regioon paljundatakse üksteisega osaliselt kattuvate genoomsete kloonidena (A) ja spotitakse mikrokiibile. Normaalse ja kasvajaraku genoomid märgistatakse erinevalt ja hübridiseeritakse mikrokiibil olevatega järjestustega (B) ning seejärel detekteeritakse ja analüüsitakse (C).

CGH mikrokiibid, millel kasutatakse sihtmärgina cDNA järjestusi, võeti kasutusele 1999 aastal (Pollack *et al*., 1999). Kuigi cDNA mikrokiipe kasutatakse enamasti ekspressiooniuuringuteks, näidati, et neid saab edukalt kasutada ka geenidoosi määramiseks genoomis. cDNA mikrokiipide abil on võimalik avastada ühe geeni kadu ja ühekordse täpsusega ka geeni amplifitseerumist. Liites mõlemad cDNA kiipide funktsioonid, saadakse tundlikud süsteemid, millega on paralleelselt võimalik uurida geeniekspressiooni ja geeni koopiaarvu muutusi ühes koes või kasvajarakkudes (Zhou *et al*., 2004; Mantripragada *et al*., 2004).



Joonis 3. cDNA mikrokiipidel läbiviidavad analüüsid: geeniekspressiooni (A) ja geenidoosi analüüs võrdleva genoomse *in situ* hübridisatsiooni meetodil (B).

cDNA mikrokiipide eeliseks võrdleva genoomse *in situ* hübridisatsiooni läbiviimisel on automatiseeritus ja suur resolutsioon muutunud piirkonna täpsustamisel. Puudusteks on aga küllaltki suur valepositiivsete (15 %) ja valenegatiivsete (15 %) tulemuste arv (Beheshti *et al*., 2002) ja sõltumine järjestuste andmebaaside annoteerituse kvaliteedist.

Oligonukleotiididel baseeruvate mikrokiipide (*oligo-array CGH*) puhul on kiibil suhteliselt lühikesed (~ 50-70 nukleotiidi) järjestused, millele hübridiseeritakse testi- ja

kontroll-DNA. Suurem erinevus võrreldes genoomseid järjestusi ja cDNA-d kandvate platvormidega on, et oligonukleotiidsete järjestustega mikrokiibil läbiviidava võrdleva genoomse *in situ* hübridisatsioonil proovide kohübridiseerimisel enamasti ei lisata konkurent-DNA-d, kuna oligonukleotiidsed järjestused on sünteesitud või paljundatud unikaalsetena. Meetodi eeliseks on aga suurem resolutsioon ja väiksem taustamüra (Lucito *et al*., 2003).

2.3 Multipleks amplifitseeritavate proovide hübridisatsioon.

Multipleks amplifitseeritavate proovide hübridisatsioon ehk MAPH (*multiplex amplifiable probe hybridization*) on meetod, mis võimaldab terves genoomis spetsiifiliste geneetiliste lookuste uurimist seal toimunud võimalike aberratsioonide tuvastamiseks (Armour *et al*., 2000). Meetod võimaldab täpselt tuvastada vastava lookuse (geeni) koopiaarvu genoomis. Multipleks amplifitseeritavate proovide hübridisatsioonil fikseeritakse uuritav denatureeritud genoomne DNA nailonmembraanile ja hübridiseeritakse spetsiifiliste proovidega. Hübridisatsioonisegusse lisatakse ka konkurent-DNA (näiteks Cot-1). Proovid on disainitud nii, et igale proovile vastab nailonmembraanil olevas järjestuses üks kindel komplementaarne märklaud-järjestus, mida soovitakse detekteerida. Proovid on tavaliselt saadud märklaud-järjestuse kloneerimisega plasmiidsesse vektorisse ja nende paljundamisega sealt universaalsete praimeritega, mille tulemusena on kõikide proovide otstes (*flanking*) identsed järjestused (Sellner *et al*., 2004). Pärast hübridiseerimist nailonmembraan pestakse, et eemaldada sealt mitteseondunud proovid. Membraanile jäävad proovid, mis on seondunud genoomsele DNA-le spetsiifiliselt. Lühiajalise kuumutamisega vabastatakse seondunud proovid lahusesse ja paljundatakse (universaalsete praimeritega) PCR-i meetodiga. Saadud produktid lahutatakse üksteisest polüakrüülamiid-geelil. Iga saadud bändi kvantitatiivse hinnangu alusel saab hinnata genoomile seondunud identsete proovide arvu ja seega ka vastava geeni koopiaarvu genoomis/lookuses. Et proove polüakrüülamiid-geelil identifitseerida ja detekteerida, on üks universaalsetest praimeritest tavaliselt märgistatud kas radioaktiivse või fluorestseeruva märgisega (Hollox *et al*., 2002). Meetodi suurimaks puuduseks on see,

et proovide pikkused peavad olema teineteisest piisavalt erinevad, et proove geelil identifitseerida.

Traditsioonilise MAPH-i kõrvale on tulemas ka kiipidel põhinev multipleks amplifitseeritavate proovide hübridisatsioon, mis on praegu arendatav mitme erineva uurimisgrupi poolt. Kiipidel põhineva MAPH-i puhul detekteeritakse amplifitseeritud proovid mitte enam geelil, vaid kvantitatiivse signaali alusel mikrokiibil, mis saadakse paljundatud proovide tagasi hübridiseerimisega iseendile (Hollox *et al*., 2002). Proovide hübridiseerimine uuritavale DNA-le toimub sarnaselt klassikalisele lähenemisele: proovid kohübridiseeritakse nailonile seotud DNA-ga, lisades konkurent-DNA-d. Mitteseondunud proovid eemaldatakse pesemisega. Seondunud proovide paljundamine toimub samamoodi kui ennegi: proovid amplifitseeritakse unikaalsete praimeritega, millest üks on märgitud fluorestseeruva märkega. Needsamad proovid, mis kohübridiseeriti uuritava DNA-ga, on seotud ka kindlas järjekorras mikrokiibile. Uuritava DNA-ga seondunud ja amplifitseeritud proovid hübridiseeritakse kiibil olevate järjestustega, mis ei ole märgitud. Iga proov seondub justkui iseendaga. Kiibilt saadavad fluorestseeruvad signaalid peegeldavad iga positsiooni ja seega ka uuritava DNA-ga seondunud proovi kvantitatiivset hulka. Arvestades eeldust, et iga proovi järjestus on genoomis esindatud vaid üks kord, saab niimoodi määrata muutunud intensiivsusega proovide koopiaarvu genoomis (Kurg *et al*., unpublished).

3. Hübridiseerimisproovide disain.

Hübridiseerimisproovide disain ja spetsiifilisus on kriitilise tähtsusega nii traditsiooniliste kui kiipidel läbiviidavate hübridiseerimismeetodite kvaliteetseks realiseerimiseks. Hübridiseerimisproovidena kasutatakse kõige enam PCR-i produkte (tavaliselt üle 100 nukleotiidi pikad) ja sünteetilisi oligonukleotiide (25-70 nukleotiidi pikad) (Reymond *et al*., 2003). Proovid peavad olema kõrge tundlikkuse ja spetsiifilisusega.

3.1 Hübridiseerimisproovide olulised parameetrid.

Erinevad tänapäeval kasutatavad molekulaarbioloogia meetodid rakendavad hübridiseerimisproovidele väga erinevaid nõudeid, kuid põhilised muutuvad parameetrid on sarnased. Tähtsamateks parameetriteks on: proovide pikkus, GC nukleotiidide sisaldus, sekundaarstruktuuride moodustamine ja sulamistemperatuur (Benita *et al*., 2003).

3.1.1 Hübridiseerimisproovide pikkus.

Hübridiseerimisproovide pikkus on erinevate meetodite vahel kõige enam varieeruv parameeter, igale meetodile on omane optimaalne proovi pikkus. Mikrokiipidel läbiviidavatel katsetel peaksid ilma speisseriteta (*spacer*) kiibile seotud proovid olema vähemalt 60 nukleotiidi pikad, kuna lühemad järjestused ei taga piisavat hübridiseerumise efektiivsust (Hughes *et al*., 2001). Pikemad proovid seonduvad reeglina efektiivsemalt, kui lühikesed (Relógio *et al*., 2002).

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 (Kane *et al*., 2000; Chen *et al*., 2002), samas kui PCR-i produktid, mida kasutatakse hübridiseerimisproovidena, on 100-3000 nukleotiidi pikad (Tõnisson *et al*., 2000; Schena *et al*., 2003). Ühe katse raames kasutatakse tavaliselt küllaltki kindlas pikkuse vahemikus olevaid proove, et tagada kõikide proovide ühtlane käitumine.

3.1.2 Hübridiseerimisproovide GC nukleotiidide sisaldus.

Hübridiseerimisproovide GC nukleotiidide sisaldus näitab G (guaniini) ja C (tsütosiini) nukleotiidide osakaalu järjestuses. GC sisalduse teeb oluliseks asjaolu, et G ja C nukleotiidid seonduvad hübridiseeruvates järjestustes omavahel tugevamini kui A ja T nukleotiidid. G ja C nukleotiidid moodustavad paardumisel kolm vesiniksidet, samas kui A ja T nukleotiidi paardumisel moodustub kaks vesiniksidet. On näidatud, et

paljude lühikeste GC rikaste regioonide sisaldumine proovis mõjutab negatiivselt proovi- ja märklaud-järjestuse hübridiseerumist, soodustades mittespetsiifilist rist-hübridiseerumist (*cross-hybridization*) (Rouillard *et al.*, 2003). Suurema GC nukleotiidide sisaldusega kaasneb ka järjestuse kõrgem sulamistemperatuur.

3.1.3 Hübridiseerimisproovide sekundaarstruktuur.

Väga oluline on hübridiseerimisproovide disainimisel jälgida, et proovid ei annaks stabiilseid sekundaarstruktuure. Sekundaarstruktuuride tekke võimalikkus proovis võib viia selle seondumisele iseendaga (moodustuvad nn. *stem-loop* ja *hairpin* struktuurid), mitte aga oma märklaud-järjestusega.

3.1.4 Hübridiseerimisproovide sulamistemperatuur.

Sulamistemperatuur (T_m) on temperatuur, mille juures pool DNA dupleksist on denatureerunud (ahelad on üksteisest lahus) ja pool mitte (ahelad on dupleksina). Sulamistemperatuur arvutatakse omavahel paardunud nukleiinhapetele välja vastavalt kindlatele valemitele. Kõige lihtsam valem lühikeste järjestuste (16-28 nukleotiidi) sulamistemperatuuri arvutamiseks on toodud järgnevana (Suggs *et al.*, 1981):

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

kus G ja C on guanosiini ja tsütosiini ning A ja T adeniini ja tümidiini nukleotiidide arv. Pikemate hübridiseerimisproovide sulamistemperatuuri väljaarvutamiseks kasutatav meetod põhineb termodünaamilisel *Nearest-Neighbor* mudelil (Borer *et al.*, 1974; SantaLucia, Jr., 1998):

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

kus ΔH on entalpia muut protsessis, ΔS on entroopia muut, R on universaalse gaasi konstant (R = 8,31 J / K * mol = 1,987 cal / K * mol) ja C_T on oligonukleotiidide

kontsentratsioon lahuses. Mõned autorid soovitavad siiski *Nearest-Neighbor* mudelit kasutada vaid kuni 50-meersete proovide sulamistemperatuuri arvutamiseks, viidates faktile, pikemate proovide puhul on sulamistemperatuurid mõnevõrra väiksemad mudeli poolt eeldatavast (Rouillard *et al*., 2003; Haas *et al*., 2003). Tegelikust kõrgema sulamistemperatuuri rakendamine võib kaasa tuua proovide spetsiifilisuse vähenemise. On välja arvutatud erinevate autorite poolt välja pakutud dinukleotiidide seondumiste energiate ühtlustatud keskmised, mis on siinkohal ära toodud tabelina:

Tabel 1. Erinevate autorite poolt publitseeritud *Nearest-Neighbor* seondumiste vabaenergiate võrdlus ja nende ühtlustatud (*unified*) tulemus (SantaLucia, Jr., 1998).

	Parameter, kcal/mol							
Sequence	Gotoh (1981)	Vologodskii (1984)	Breslauer (1986)	Blake (1996)	Benight (1992)	SantaLucia (1996)	Sugimoto (1996)	Unified
AA/TT	−0.43	−0.89	(−1.66)	−0.67	−0.93	−1.02	−1.20	−1.00
AT/TA	−0.27	−0.81	−1.19	−0.62	−0.83	−0.73	−0.90	−0.88
TA/AT	−0.22	−0.76	−0.76	−0.70	−0.70	−0.60	−0.90	−0.58
CA/GT	−0.97	−1.37	−1.80	−1.19	−1.26	−1.38	−1.70	−1.45
GT/CA	−0.98	−1.35	−1.13	−1.28	−1.52	−1.43	−1.50	−1.44
CT/GA	−0.83	−1.16	−1.35	−1.17	−1.03	−1.16	−1.50	−1.28
GA/CT	−0.93	−1.25	−1.41	−1.12	−1.56	−1.46	−1.50	−1.30
CG/GC	−1.70	−1.99	(−3.28)	−1.87	(−1.65)	−2.09	(−2.80)	−2.17
GC/CG	−1.64	−1.96	(−2.82)	−1.85	−2.44	−2.28	−2.30	−2.24
GG/CC	−1.22	−1.64	(−2.75)	−1.55	−1.67	−1.77	−2.10	−1.84
Average	−0.92	−1.32	−1.82	−1.20	−1.36	−1.39	−1.64	−1.42
Init. w/term. G·C*	NA	NA	(+2.60)	NA	NA	0.91	(+1.70)	0.98
Init. w/term. A·T*	NA	NA	(+2.60)	NA	NA	1.11†	(+1.70)	1.03
Sodium concentration, M	0.0195	0.195	1.0	0.075	0.115	1.0	1.0	1.0
Rank of stacking matrix	8	8	11	8	9	10	11	12

3.2 Hübridiseerimisproovide seostumiskohtade ja unikaalsuse kontroll.

Et saada kvaliteetseid hübridiseerimisproove, on enne praktilisi katseid vaja *in silico* leida kõikide proovide võimalikud seostumiskohad uuritavas genoomis, tagamaks proovide unikaalsust ja vältimaks nende rist-hübridiseerumist. Samas on vaja kontrollida, et ühes katses koos kasutatavad proovid ei seonduks üksteisega.

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 sarnane (50-75 %) 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). Sarnased tulemusi on saadud ka pikemate järjestustega tehtud katsetes (Hughes *et al*., 2001).

Selleks, et leida ja kontrollida hübridiseerimisproovide seondumiskohti genoomis, on arendatud mitmeid erinevaid programme, mis *in silico* otsivad etteantud proovide seondumiskohti genoomis. Tuntumateks sellisteks programmideks on SSAHA (*Sequence Search and Alignment by Hashing Algorithm*), BLAST2 (*Basic Local Alignment Search Tool, ver 2*) ja MegaBLAST.

SSAHA programmi algoritm kasutab metoodikat, mille puhul andmebaas (näiteks inimese esimene kromosoom), millest hakatakse järjestust otsima, kõigepealt indekseeritakse ja kirjutatakse ühte suurde paisktabelisse (*hash*). See on aeglane, kuid see-eest ühekordne protsess. Kui programm hakkab etteantud järjestust (*query sequence*) otsima, 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 protsenti kokkulangevaid) elemente. Programmi töö tulemusena saadakse unikaalsed ehk andmebaasis vaid üks kord esinevad järjestused (Ning *et al*., 2001).

BLAST2 ehk *Gapped BLAST* programm kasutab sarnasuse leidmiseks päring-järjestuse ja andmebaasi vahel heuristilisi algoritme. Kõikidele sõnadele pikkusega *T* arvutatakse skoor, kasutades selleks spetsiaalseid tabeleid. Kui kasutaja poolt etteantud järjestus ja mõni andmebaasi järjestus mingis osas ühtivad, arvutatakse nende sarnasele osale skoor ja kui see ületab kasutaja poolt määratud piiri, väljastatakse tulemus. BLAST2 suureks eeliseks on asjaolu, et ta võimaldab otsida ka *gap*-idega. 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. Sellega vähendatakse oluliselt

programmi ajalist keerukust. Kasutades sobivaid järjestusi, võib päring-järjestusele sarnased regioonid andmebaasis leida kuni 10 korda kiiremini kui dünaamilist programmeerimist (*dynamic programming*) kasutades (Zhang *et al*., 2000). Seetõttu sobib MegaBLAST hästi ka pikkade päring-järjestuste võrdlemiseks andmebaasiga.

Kõik eelmainitud programmid kasutavad sarnasuse otsimisel evolutsioonilist kaugust kahe järjestuse vahel. Uuemad programmid kasutavad aga proovide sarnasuse leidmisel termodünaamilist lähenemist, kalkuleerides järjestuste seondumise tugevust ja leides kõik piirkonnad, kuhu proovid võivad seonduda piisavalt tugevasti. Ebasoovitavaid seondumisi andvad proovid saab seejärel eemaldada.

3.3 Hübridiseerimisproovide disaini programmid.

Suurem enamus hübridiseerimisproovide disaini programme on loodud suhteliselt lühikeste (20-70) oligonukleotiidide disainimiseks. Kvaliteetsed programmid pikkade hübridiseerimisproovide automaatseks disainimiseks aga puuduvad.

Tabel 2. Erinevad programmid hübridiseerimisel kasutatavate oligote disainimiseks.

Programm	Kättesaadavus
PROBEmer	http://probemer.cs.loyola.edu/
PRIMROSE	http://www.cf.ac.uk/biosi/research/biosoft/Primrose/index.html
OligoArray 2.0	http://berry.engin.umich.edu/oligoarray2/
GenomePRIDE	http://pride.molgen.mpg.de/genomepride.html
OLIGO6	http://oligo.net/
OligoWiz	http://www.cbs.dtu.dk/services/OligoWiz/
ProbeSel	Kirja teel küsitav : schliep@molgen.mpg.de
ProbeSelect	Kirja teel küsitav : lif@ural.wustl.edu

4. PCR-i praimerite disain hübridiseerimisproovidele.

Väga paljud tänapäeval kasutatavad genoomide uurimiseks mõeldud funktsionaalse genoomika meetodid baseeruvad genoomist PCR-iga amplifitseeritavatel järjestustel – PCR-i produktidel ja nende hübridiseerimisel. PCR-i produkte kasutatakse genotüpiseerimisel, mutatsioonide uurimisel, jne. PCR-iga paljundatud proove

kasutavad ka sellised võrdleva genoomika meetodid nagu *array CGH* (*comparative genomic in situ hybridization*) ja MAPH (*multiplex amplifiable probe hybridization*). Et need praktilised meetodid toimiksid veatult, on enne PCR-i rakendamist vajalik spetsiifiliste PCR-i praimerite ja produktide disainimine *in silico*. Et saada kvaliteetseid PCR-i produkte, mida saab kasutada hübridiseerimisproovidena, on vajalik märklaud-järjestuse amplifitseerimine kahe PCR-i praimeriga.

4.1 PCR-i praimerite olulised parameetrid.

Polümeraasi ahelreaktsiooni õnnestumiseks peavad kasutatavad praimerid vastama mitmele erinevale kvaliteedinõudele. Praimerite disainimisel on kõige tähtsam jälgida praimerite sulamistemperatuuri ja produkti pikkust, aga ka praimerite GC nukleotiidide sisaldust ja pikkust (Benita *et al*., 2003).

4.1.1 Praimerite sulamistemperatuur.

Praimerite sulamistemperatuuri arvutatakse sarnaselt hübridisatsiooniproovide sulamistemperatuurile. Tähtis on, et ühe kindla regiooni ülesamplifitseerimiseks kasutatavad praimerid (*sense*- ja *antisense* praimerid) oleksid ühesuguse sulamistemperatuuriga. Näiteks inimese genoomse DNA fragmendi amplifitseerimisel kasutatavad 18-24 nukleotiidi pikkused praimerid võiksid olla sulamistemperatuuriga 55-58° C (Henegariu *et al*., 1997).

4.1.2 PRC-i produkti pikkus.

Kuigi viimastel aastatel on saanud tänu PCR-i meetodite arenemisele võimalikuks kuni 10 kb pikkuste alade amplifitseerimine (Zakour *et al*., 2004), on näidatud, et edukamalt amplifitseeruvad lühikesed PCR-i produktid. Kui produkti pikkus on üle 400 nukleotiidi, langeb PCR-i kvalite*et al*la 80 protsendi (R. Andreson,

2002). Kui produkti pikkus on 10 protsendi võrra suurem või väiksem kui soovitud, siis võib produkte pidada kvaliteetseteks (Benita *et al*., 2003).

4.1.3 Praimerite GC nukleotiidide sisaldus.

Enamasti peaks praimerite GC nukleotiidide sisaldus olema valitud selliselt, et see on sama või natuke kõrgem ülesamplifitseeritava DNA regiooni GC nukleotiidide sisaldusest (Benita *et al*., 2003). Inimese puhul on näidatud, et optimaalsete praimerite GC sisaldus on ca 50 protsenti (Haas *et al*., 1998). Praimerite GC sisaldus sõltub ka sulamistemperatuurist; madalama sulamistemperatuuriga praimerid disainitakse väiksema ja kõrgema sulamistemperatuuriga praimerid suurema GC nukleotiidide sisaldusega.

4.1.4 Praimerite pikkus.

PCR-i praimerite pikkus sõltub oluliselt sulamistemperatuurist ja GC nukleotiidide sisaldusest. Praimerite pikkus võib olla 15-30 nukleotiidi, optimaalseks peetakse 18-26 nukleotiidi pikkuseid praimereid. Kuna erinevate organismide genoomide GC nukleotiidide sisaldused varieeruvad, erinevad ka praimerite keskmised pikkused organismide vahel (Haas *et al*., 1998, 2003).

4.2 Praimerite seondumiskohtade kontroll.

Et saada kvaliteetseid PCR-i praimereid ja seega ka produkte, peab *in silico* kontrollima, et disainitud praimerid ei seonduks genoomis valedele järjestustele ega annaks mittespetsiifilisi produkte. Samas peab jälgima, et praimerid ei seonduks iseendaga ega teiste praimeritega (Matveeva *et al*., 2003).

Praimerite seondumiskohtade *in silico* leidmiseks genoomis saab rakendada samasid programme, mida rakendatakse hübridiseerimisproovide unikaalsuse ja seondumiskohtade leidmiseks: SSAHA, MegaBLAST, BLAST2. On olemas ka spetsiaalselt praimerite või teiste lühikeste järjestuste seondumiskohtade leidmiseks

kirjutatud programme. Üheks selliseks programmiks on GenomeMasker-i paketist programm GenomeTester (Reppo *et al.*, unpublished). GenomeTester-i tööpõhimõte on sarnane programmi SSAHA omale; enne järjestuste otsimist indekseeritakse genoomsed järjestused ja kirjutatakse paisktabelitesse (*hash*). Järjestuste otsimisel loetakse paisktabel mällu ja identsete piirkondade leidmine toimub juba kiiresti. Eeliseks SSAHA ees on see, et indekseerimisel kasutatakse fikseeritud sõnapikkust (vaikimisi 16). Seetõttu on GenomeTester teistest sarnastest programmidest ka kiirem. Teiseks GenomeTester-i eeliseks on see, et väheste päring-järjestuste puhul ei pea programm indekseeritud järjestusi mälusse lugema. Programmi töö tulemusena leitakse praimerite seondumiskohad genoomis ja kui palju ja millistest regioonidest produkte nende praimeritega amplifitseeritakse. Üleliigseid seondumisi ja produkte andvad praimerid eemaldatakse.

On näidatud, et kui praimerite paari seostumiste summa kogu genoomis on suurem kui 10, langeb PCR-i kvalite*et al*la 80 protsendi. Samavõrra alaneb PCR-i kvaliteet juhul, kui praimerite paari ennustatavaid produkte on rohkem kui 2 (R. Andreson, 2002).

4.3 Praimerite disaini programmid.

Palju programmid, mis sobivad lühikeste hübridiseerimisproovide disainimiseks, sobivad ka praimerite disainimiseks. Disainimisel on oluline jälgida väga paljude erinevate parameetrite sobivust, head praimerite disainimise programmid võimaldavad neid parameetreid valida ja muuta.

Praimerite disainimise programmid võib üldjoontes jaotada kaheks: programmid, mis on mõeldud suuremahuliste projektide tarvis praimerite disainimiseks ja programmid, mis on väiksema mahuga projektidel vajaminevate praimerite disainimiseks. Esimesed on enamasti rohkem automatiseeritud (suurte andmehulkade kiirem läbitöötamine) kui väiksemahuliste projektide jaoks mõeldud programmid. Viimaste heaks omaduseks on aga see, et kasutajal on kogu aeg ülevaade (tihti ka visualiseeritud) disainimise protsessist. Tabelis 3 on ära toodud enamlevinud programmid praimerite disainiks.

Tabel 3. Erinevad programmid praimerite disainimiseks.

Programm	Kättesaadavus
Primer3	http://frodo.wi.mit.edu/primer3/primer3_code.html
	http://frodo.wi.mit.edu/cgi-bin/primer3/primer3.cgi
Web Primer	http://genome-www2.stanford.edu/cgi-bin/SGD/web-primer
MEDUSA	http://www.cgr.ki.se/cgr/MEDUSA/
OLIGO6	http://www.oligo.net/
PrimerDesign	http://www.chemie.uni-marburg.de/~becker/pd.exe
Primer Premier 5	http://www.premierbiosoft.com/primerdesign/index.html
PrimerQuest	http://biotools.idtdna.com/primerquest/
PrimerSelect	http://www.dnastar.com/web/r11.php

II PRAKTILINE TÖÖ

Töö eesmärgid.

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

Käesoleva töö eesmärgiks oli luua terviklik lahendus hübridiseerimisproovide automaatseks disainimiseks inimese genoomsetele järjestustele, veebipõhise kasutajaliidese loomine hübridiseerimisproovide disaini programmide kasutamiseks ja näidisproovide tegemine praimerite andmebaasi PrimerStudio jaoks.

Lahendusena on välja töötatud unikaalsel metoodikal põhinev programmide pakett ja veebiportaal nimega MAPHDesigner. MAPHDesigner-iga saab automaatselt disainida mikrokiipidel läbiviidavate MAPH-i (*multiplex amplifiable probe hybridization*) ja CGH (*comparative genomic hybridization*) hübridiseerimisproove.

Kasutatud meetodid.

1. Andmete päritolu ja struktuur.

Töös kasutatud algandmed ja järjestused on saadud EnsEMBL-i andmebaasist. EnsEMBL-ist on alla laetud inimese genoomi andmebaasi (*Human 34*) versioon 18.34.1 (avalikustatud 04.11.2003) ja sama versiooni inimese genoomsete järjestuste kontiigide andmebaas (*Core 34*). Mõlemad andmebaasid on assambleeritud NCBI inimese genoomi versiooni nr 34 järgi. Tabelis 4 on näha Human 34 andmebaasis olevad, käesolevas töös kasutatud tabelid ja nende väljad.

Tabel 4. Andmebaasi *Human 34* struktuur, töös kasutatud andmebaasi tabelid ja väljad.

Human 34 andmebaas:	
Tabel:	**Väljad:**
karyotype	chromosome_id, chr_start, chr_end, band, stain
ens_gene	gene_id, chr, start, end, strand, external_id, mim_id, hugo_id, locuslink_id, refseq_id
vega_gene	gene_id, chr, start, end, strand, mim_id, hugo_id, locuslink_id, refseq_id
ncbi_gene	id, gene_id, chr, start, end, strand, locus_id, mim_id, hugo_id, locuslink_id,
	refseq_id, external_id

Core 34 andmebaasist saadi järjestuste info kujul: *kromosoom, alguspositsioon, järjestus*. Selleks kasutati spetsiaalset EnsEMBL-i moodulit (*SliceAdaptor*), mis teeb andmebaasist päringu ja tagastab tulemused (Stajich *et al*., 2002).

2. Kasutatud programmide kirjeldus.

MAPHDesigner-i poolt hübridiseerimisproovide disainiks kasutatavad programmid võib tinglikult jagada kaheks. Esimese grupi moodustavad programmid, mis teostavad töö sisulise poole; järjestuste päringu andmebaasist, proovide järjestuse maskeerimise, proovide ja praimerite disainimise, unikaalsuse kontrolli, jne. Teise grupi moodustavad veebi kasutajaliidese genereerimiseks ja töötamiseks vajalikud programmid.

Töö sisulist poolt teostavate programmide kirjeldus on toodud tabelis 5.

Tabel 5. Hübridiseerimisproovide disaini teostavad programmid.

Programm	Kirjeldus
1. proc*	Tõmbab andmebaasist DNA järjestused, käivitab kindlas järjekorras ülejäänud programmid.
2. dust_lower (modifitseeritud DUST)	Maskeerib madala keerukusega piirkonnad proovides (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	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, mis seonduvad valedes kohtades ja annavad sekundaarseid produkte.
9. prod2fas	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*	Kirjutab liiga kõrge DUST-i skooriga proovide ID-d vastavasse faili.
12. exclude_from_fasta	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	Parsib MegaBLAST-i tulemused tabuleeritud formaati.
16. parsers_parser* 17. wrong_chr_out* 18. wrong_doubled_out* 19. one_per_start* 20. extract_from_file	Eemaldavad proovide hulgast need, mis annavad seondumisi ka mittesoovitavates regioonides.
21. extract*	Kirjutab tulemused ümber fasta formaadis faili.
22. first_name	Kirjutab "heade" proovide ID-d selleks ettenähtud faili.
23. first* 24. extrtract_lines_from_file*	Eraldavad erinevatest failidest "heade" proovide vajalikud andmed ja kirjutavad need faili.
25. one_per_start_pos*	Eemaldab redundantsed proovid.

Käesoleva töö raames loodud programmid (*) on kirjutatud Perl-is (http://www.perl.com/) ja shelli-skriptis. On kasutatud ka teiste autorite programme. Enamus programmidest on 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, seda vahepealt kõvakettale kirjutamata. Praimerite ja

hübridiseerimisproovide disaini teostavate programmide väljundiks on fail *results.txt*. *Results.txt* sisaldab: disainitud proovi järjekorranumbrit; ID-d; kromosoomi, kuhu proov seondub; proovi alguspositsiooni kromosoomil; sense ja antisense praimereid proovi amplifitseerimiseks ja proovi järjestust.

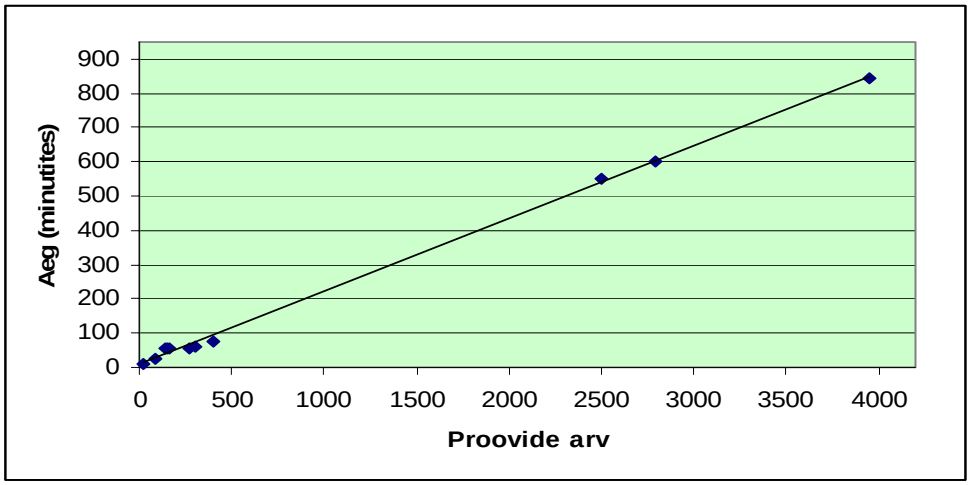
Veebiliidese genereerimisel osalevad mitu erinevat programmi. MAPHDesigner-i esmase portaali paneb kahest html-failist (index.p1.html ja index.p2.html) kokku programm index.pl. Edasised leheküljed pannakse kokku failides Index.pm ja Bioprimer.pm (R. Andreson) olevate alamprogrammidega.

3. Tööks vajalik mälumaht ja aeg.

MAPHDesigner-i töös on kõige ressursinõudlikumad programmid, mis kontrollivad praimerite ja proovide unikaalsust genoomis: GenomeTester, SSAHA ja MegaBLAST. Mälunõudlikkus on kõige suurem programmil SSAHA, mis otsides proove inimese esimese kromosoomi pealt, 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 nõudlikkus ei ületa 150 Mb piiri.

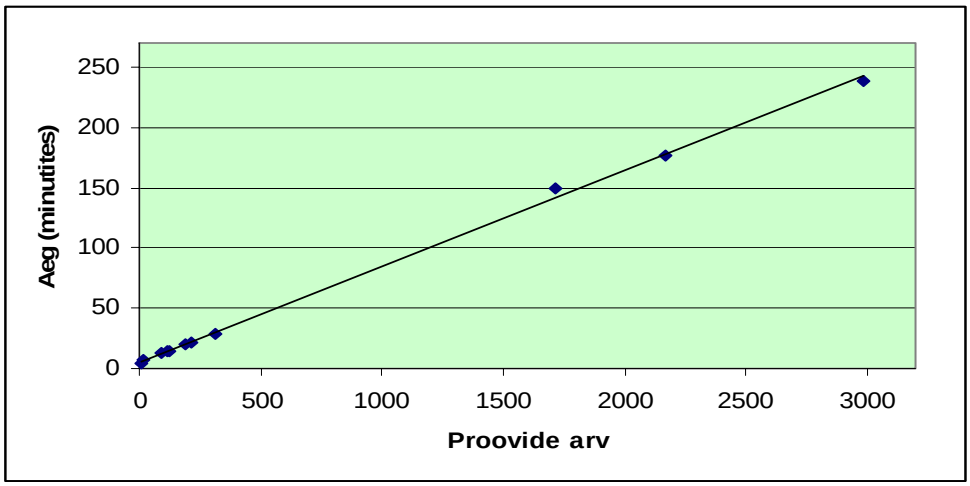
Nagu mälukasutuse puhul, on ka ajalises mõttes kõige keerukamad programmid MegaBLAST ja SSAHA. Ülejäänud programmide tööajad on suhteliselt lühikesed. Teoreetiliselt on nii SSAHA kui ka MegaBLAST-i ajaline keerukus võrdelises seoses päringute arvuga; päringute arvu suurenemisel kasvab lineaarselt ka 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 4 on ära toodud programmi SSAHA praktiliselt mõõdetud ajaline kulu sõltuvalt kontrollitavate proovide arvust.



Joonis 4. Programmi SSAHA ajaline kulu sõltuvalt proovide arvust. Proovide arvu suurenemisel kasvab lineaarselt ka programmi SSAHA tööaeg.

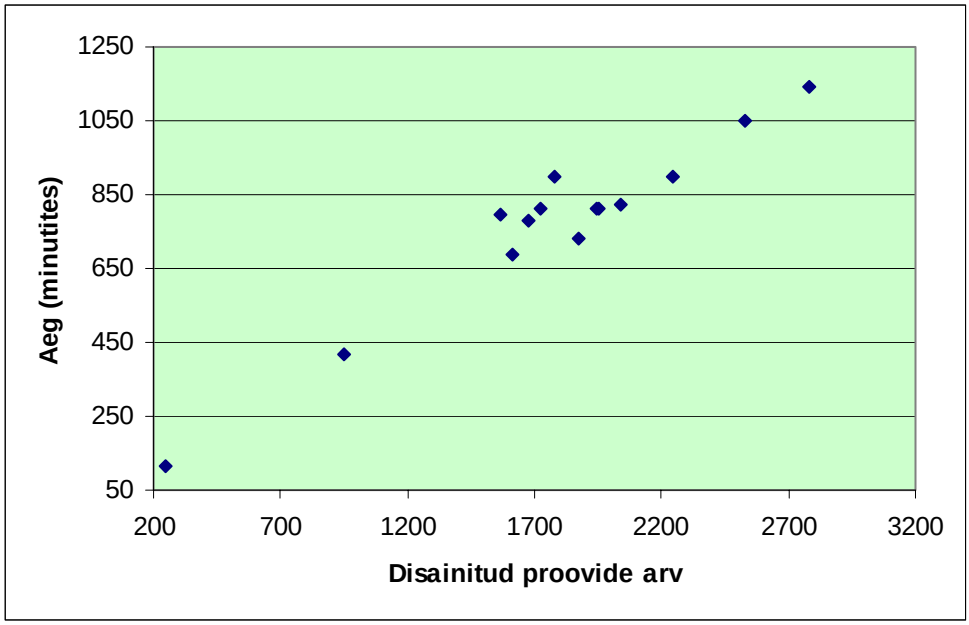
Siinjuures on oluline märkida, et proovide arvu all on silmas peetud programmide SSAHA ja MegaBLAST poolt kontrollitud proovide arvu, mitte positiivset tulemust. Joonisel 5 on programmi MegaBLAST ajaline kulu sõltuvalt kontrollitavate proovide arvust.



Joonis 5. Programmi MegaBLAST ajaline kulu sõltuvalt proovide arvust. Proovide arvu suurenemisel kasvab lineaarselt ka programmi MegaBLAST tööaeg.

Nagu joonistelt 4 ja 5 näha, on programmide SSAHA ja MegaBLAST praktiliselt saadud tööajad lineaarses seoses päring-järjestuste arvuga, mis on kooskõlas ka teoreetilise mudeliga.

Kogu MAPHDesigner-i tööaja moodustab kõikide programmide summaarne töötamise aeg. Hübridiseerimisproovide disaini protsessi ajalist kulu kirjeldab joonis 6.



Joonis 6. Hübridiseerimisproovide disainimise ajaline kulu sõltuvalt disainitud proovide arvust. Protsessi ajaline kulu on lineaarses sõltuvuses disainitud proovide arvust.

Protsessi ajaline kulu on lineaarses sõltuvuses disainitud proovide arvust ja 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 ajana (indeksite kettalt mällu lugemine, jne).

Käesoleval juhul töötas MAPHDesigner-i programmipakett Linux-i operatsioonisüsteemiga kaheprotsessorilises (2 * Pentium III @800Mhz) serveris, millel on 1 Gb operatiivmälu. Ühe proovi disainimise ajaks (K) tuli keskmiselt 0,4 minutit ehk 24 sekundit.

Tulemused.

1. Veebi kasutajaliides.

Loodud programmide paketi lihtsamaks kasutamiseks valmis veebipõhine kasutajaliides ehk portaal MAPHDesigner (http://kobra.ebc.ee/MAPH), mis võimaldab kasutajal disainida hübridiseerimisproove inimese genoomsetele järjestustele.

Kasutajal on võimalik portaalis disainida proove kindlatele regioonidele ja/või geenidele. Saab valida terveid kromosoome ja/või nende piirkondi (vaikimisi on valitud regiooniks terve esimene kromosoom). Valikute hulgas olevate piirkondade määramisel pannakse nende piirkondade algus- ja lõpp-koordinaadid (positsioonid kromosoomil) automaatselt paika. Samas on kasutajal võimalik ka ise määrata soovitavad positsioonid. 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. Valikud on: EnsEMBL, Vega ja RefSeq. Erinevate projektide käigus on geene erinevalt annoteerinud ning seetõttu erinevad ka nende koordinaadid andmebaasides. Vega andmebaasis on hetkel saadaval 6, 13, 14, 20 ja 22 kromosoomis olevate geenide koordinaadid, EnsEMBL-i ja RefSeq-i andmebaasides on esindatud kõik inimese kromosoomid. Vaikimisi on geenikoordinaatide andmebaasiks EnsEMBL.

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

Valida saab ka disainitavate proovide pikkuse. On võimalik määrata minimaalne proovi pikkus (vaikimisi 200 nukleotiidi), optimaalne proovi pikkus (vaikimisi 400) ja maksimaalne proovi pikkus (vaikimisi 600).

Lisaks saab kasutaja paika panna ka hübridiseerimisproovide minimaalse (vaikimisi 70 C^o) ja maksimaalse (vaikimisi 80 C^o) sulamistemperatuuri.

Kui kasutaja soovib proovide disaini lõppemisest teada saada e-maili teel, on tal 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, kust on võimalik nupu „Check Results“ vajutamisega kontrollida, kas proovide disain on lõpetatud. Kui hübridiseerimisproovide disain on lõpetatud, suunatakse kasutaja tulemuste lehele, kus on link failile *results.txt* ja kust on võimalik alla laadida tulemuste faili ka pakitud (.zip) kujul.

2. Algoritmi kirjeldus.

Kasutaja poolt käsurealt või veebi kasutajaliidese kaudu sisestatud andmed on argumendiks programmile proc.pl. Kui kasutaja on defineerinud kindla regiooni inimese genoomist (näiteks esimene kromosoom), jaotatakse see regioon (*span*) soovitud arvu (*num*) proovide alusel väiksemateks lõikudeks (*subdiv*) ja programm proc.pl teeb päringu Core 34 andmebaasist, kust iga lõigu (*subdiv*) keskelt võetakse 3000 nukleotiidi pikkune järjestus. Kui kasutaja on defineerinud kindla geeni, teeb proc.pl kõigepealt päringu Human 34 andmebaasist, kust saadakse vastava geeni algus- ja lõpp-koordinaadid, mille alusel toimub soovitud arvu järjestuste päring Core 34 andmebaasist. Saadud järjestused kirjutatakse koos genereeritud ID-ga (kujul *kromosoom:järjestuse alguspositsioon*) faili *regions.fas*. Kui andmebaasist saadud järjestus sisaldab üle ühe protsendi N-tähti (teadmata nukleotiide), ei kirjutata järjestust andmefaili.

Edasi toimub järjestustes olevate korduste maskeerimine programmidega dust_lower ja gmasker, maskeeritud järjestused kirjutatakse faili *regions.gm*. Programm fasta2primer3.pl kirjutab maskeeritud järjestused Primer3-ele sobivas formaadis faili *primers.txt*. Gm_primer3 üritab igas 3000 nukleotiidi pikkuses regioonis leida sobivad praimerid, mis vastaksid soovitud tingimustele ja kontrollib, et nende praimeritega tekkiv produkt vastaks proovile seatud nõuetele. Programm primer3_to_table.pl kirjutab Primer3-e tulemused (ID, praimerid, produktid) sobivas formaadis faili *primers2.txt*.

Programm gtester kontrollib seejärel genoomitestiga, kuhu saadud praimerid inimese genoomis seonduvad ja kas tekib ka sekundaarseid produkte. Praimerid, mis seonduvad ka mittespetsiifiliselt ja annavad sekundaarseid produkte, eemaldatakse

programmiga gtester_filter.pl. Sobivate praimeritega *in silico* saadavad produktid ehk proovid kirjutatakse koos ID-ga 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*.

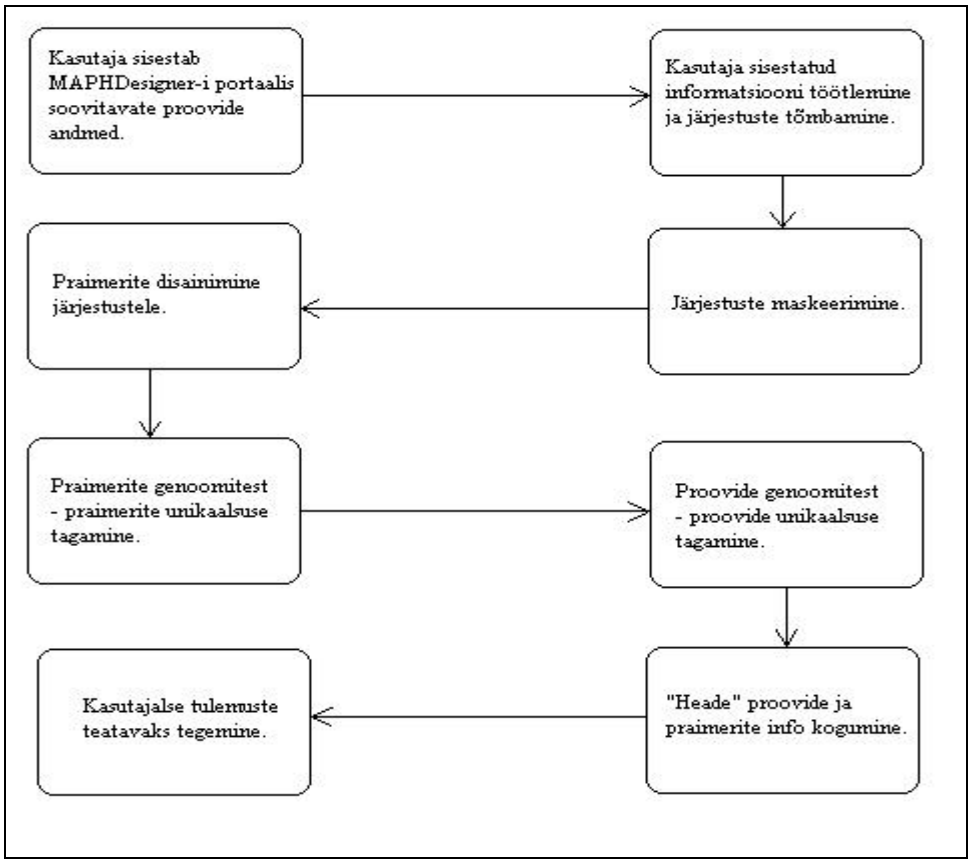
Proovide unikaalsuse kontroll tehakse nii genoomi kui proovide iseendi vastu. Kõigepealt kontrollitakse proove iseendi ja üksteise vastu. Seda tehakse programmiga MegaBLAST. Proovid, mis võivad anda omavahel seostumisi, eemaldatakse. Võimalikeks seostumisteks on loetud juhud, mille puhul kaks proovi on vähemalt 50 nukleotiidi pikkusel lõigul enam kui 75 protsenti sarnased. Unikaalsuse kontroll inimese genoomi vastu tehakse programmidega SSAHA ja MegaBLAST. Kontrollitakse failis *prod.fas* olevate järjestuste võimalikke sekundaarseid seostumiskohtasid inimese genoomis.

MegaBLAST-i tulemused parsitakse 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 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.

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 disainimise protsessi lõpus kasutajale vastavasisulise e-maili. Kirjas on kasutaja päringu ID (*request ID*, unikaalne igal disainimisel) ja disainimise protsessi alguse aeg. Lisaks on kirjas ka hüperlingid tulemuste failile ja tulemuste pakitud failile.

Joonisel 7 on toodud MAPHDesigner-i algoritm lihtsustatud kujul.



Joonis 7. MAPHDesigner-i algoritmi lihtsustatud skeem. Disainimine algab kasutaja poolt sisestatud järjestuste allalaadimisega andmebaasist ja lõpeb kasutajale tulemuste teatavaks tegemisega.

3. Tulemuste kirjeldus.

Hübridiseerimisproovide disaini metoodika väljatöötamisel arvestati MAPH-i meetodi eripäradega. Pandi paika erinevad kriitilised parameetrid, millele praimerid ja proovid peavad vastama. Proovide pikkuseks määrati 200 – 600 nukleotiidi (Hollox *et al*., 2002). Proovide ja praimerite GC nukleotiidide sisalduseks pandi 20 – 50 protsenti, mis määrati praktiliselt saadud tulemuste (Patsalis *et al*., unpublished) analüüsil ja mida toetavad ka kirjanduse andmed (Benita *et al*., 2003). Unikaalsuse tagamisel arvestati seda, et kuigi kõik PCR-i produktid (proovid) amplifitseeritakse eraldi, on nad koos hübridiseerimislahuses. Seetõttu peavad proovid olema unikaalsed genoomis ja nad ei tohi anda seostumisi omavahel. Kuna proovide kontrollimine on kogu disainimise protsessi juures kõige aeganõudvam, tuli see korraldada võimalikult aega säästvalt.

Seetõttu toimus proovide kontrollimine järjekorras: 1. proovid proovide vastu programmiga MegaBLAST (võimaldas kiiresti elimineerida omavahel seonduvad proovid), 2. proovid inimese genoomi vastu programmiga SSAHA (võimaldas suhteliselt kiiresti eemaldada proovid, mis sisaldasid ulatuslikke komplementaarseid alasid genoomiga) ja 3. proovid inimese genoomi vastu programmiga MegaBLAST (kõige aeganõudvam kontrollimise osa, mis viidi läbi vaid allesjäänud proovidega). Kõige aeganõudvama kontrollimise osa läbiviimine minimaalse arvu negatiivsete proovidega võimaldas oluliselt kiirendada kogu protsessi.

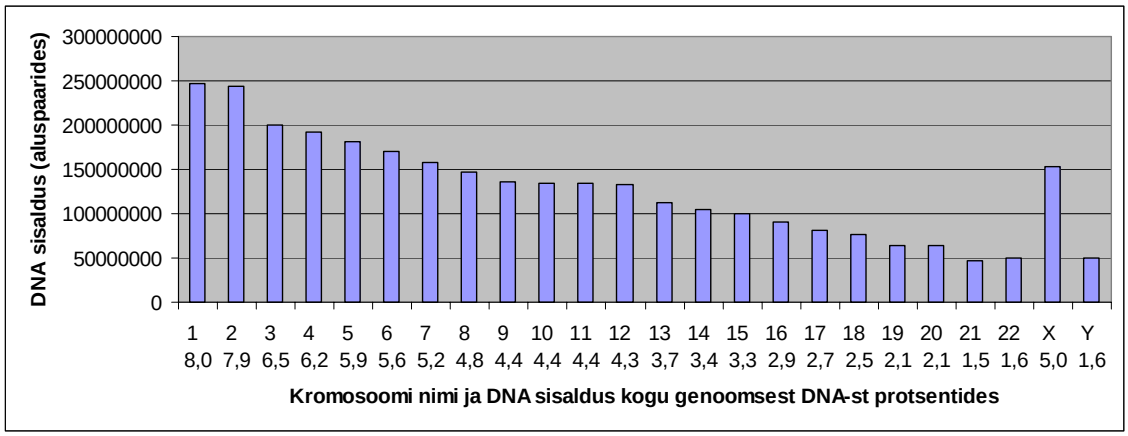
Metoodika väljatöötamisel kasutati katseliselt saadud PCR-i praimerite kvaliteedi andmeid (Patsalis *et al.*, unpublished), mis sisaldasid informatsiooni MAPHDesigner-ile sarnase metoodikaga disainitud (M. Remm) praimerite ja nende produktide (proovide) kohta. Nende proovide hulgas oli kaks eraldi (kuid ühe algoritmiga) disainitud gruppi: proovid, mis olid disainitud inimese X kromosoomile (grupp 1) ja proovid, mis olid disainitud autosomaalsetele kromosoomidele (grupp 2). Grupi 1 PCR-i õnnestumise protsent oli 96,6 (96,6 protsenti produktidest õnnestus amplifitseerida neile disainitud praimeritega) ja grupi 2 PCR-i õnnestumise protsent oli 93,8. Ebaõnnestunud praimerite hulka loeti praimerid, mis: a) ei andnud produkti, b) andsid mitu produkti, c) andsid vale pikkusega produkte. Grupi 1 ja grupi 2 praimerite ning proovide *in silico* võrdlemisel selleks kirjutatud programmiga MA (*MAPHAnalyser*) saadi tulemused, mis on kokkuvõtvalt esitatud tabelis 6.

Tabel 6. Kahe erineva PCR-i kvaliteediga proovide ja praimerite grupi võrdlus. Võrreldud on praimerite ja proovide tähtsamaid parameetreid.

Grupp 1 (kromosoom X)		
Proovide andmed:	Keskmine pikkus:	**481**
	Max pikkus:	589
	Min pikkus:	386
	Keskmine GC sisaldus:	**38,6**
Praimerite andmed:	Keskmine pikkus:	**23,88**
	Max pikkus:	26
	Min pikkus:	22
	Keskmine GC sisaldus:	**48,36**
Grupp 2 (autosoomid)		
Proovide andmed:	Keskmine pikkus:	**395**
	Max pikkus:	600
	Min pikkus:	200
	Keskmine GC sisaldus:	**50,1**
Praimerite andmed:	Keskmine pikkus:	**21,25**
	Max pikkus:	26
	Min pikkus:	18
	Keskmine GC sisaldus:	**60,57**

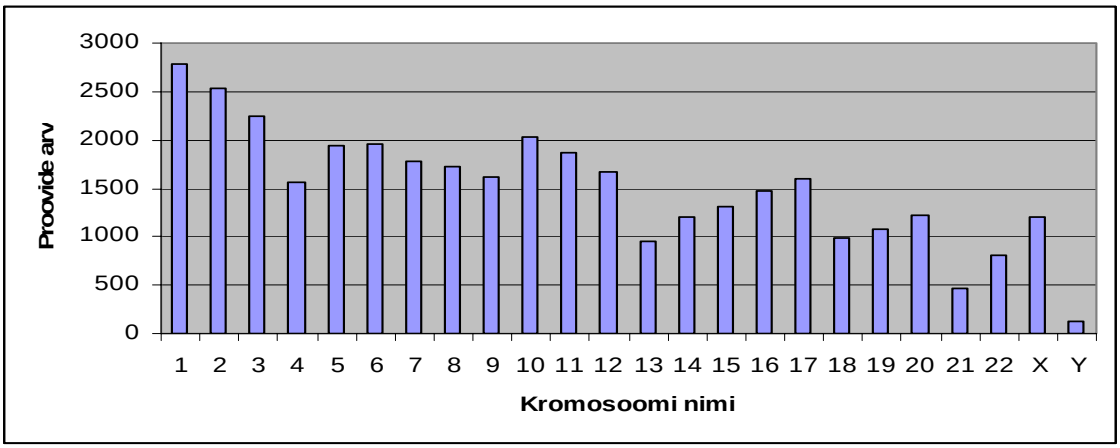
Võrdlusest tuli välja, et madalama PCR-i kvaliteediga praimerite ja proovide (grupp 2) keskmine GC nukleotiidide sisaldus on tunduvalt kõrgem (ca 12 % võrra) kui PCR-il paremini õnnestunud praimeritel ja proovidel (grupp1). Ka olid grupi 2 praimerid tunduvalt lühemad, kui grupi 1 praimerid. Lühemate ja GC-rikkamate praimerite kasutamine võib viia nende seostumisele valele genoomsele järjestusele ja valede produktide tekkele (Haas *et al*., 2003). On näidatud, et pikemate praimerite kasutamine tõstab PCR-i kvaliteeti (R. Andreson, 2002). Samas ei olnud märkimisväärseid GC nukleotiidide sisalduse ega ka pikkuse erinevusi ühes grupis olnud õnnestunud ja mitteõnnestunud proovide ning praimerite vahel. Nendest tulemustest sõltuvalt otsustati praimerite pikkuseks määrata 22 – 26, optimaalse pikkusega 24 nukleotiidi.

Hübridiseerimisproovide disainimisel oli eesmärgiks teha PrimerStudio andmebaasi jaoks 30 000 praimerite paari ja proovi üle kogu genoomi. Proovide disainimisel arvestati inimese genoomse DNA ebaühtlast jaotumist kromosoomide vahel. Erinevate kromosoomide DNA sisaldust kujutab joonis 8.



Joonis 8. Inimese genoomse DNA jaotus erinevate kromosoomide vahel. Kromosoomide DNA sisaldus aluspaarides ja protsentuaalselt kogu genoomsest DNA-st.

Proovide disain inimese genoomsetele järjestustele teostati kromosoomide kaupa ja automaatse disainimise tulemusena saadi 36 172 proovi üle kogu genoomi. Proovide jaotumist kromosoomide vahel kujutab joonis 9.



Joonis 9. MAPHDesigner-iga disainitud hübridiseerimisproovide arv erinevatel kromosoomidel.

Tabelis 7 on ära toodud soovitud proovide arv (vastavalt kromosoomide suurusele) ja tulemused kromosoomide kaupa.

Tabel 7. MAPHDesigner-iga teostatud proovide disainimise tulemused. Disainida proovitud ja edukalt disainitud proovide arv erinevate kromosoomide kohta.

Kromosoom	Soovitud proovide arv	Proovitud teha	Saadud proovide arv	Mitu protsenti saadi soovitud proovide arvust	Mitu protsenti saadi teha proovitud proovide arvust
1	2405	9999	2781	115,6	27,8
2	2381	9999	2530	106,3	25,3
3	1948	9740	2243	115,1	23,0
4	1874	9368	1567	83,6	16,7
5	1769	8845	1944	109,9	22,0
6	1670	8351	1956	117,1	23,4
7	1549	7746	1776	114,6	22,9
8	1430	7148	1726	120,7	24,1
9	1333	6663	1611	120,9	24,2
10	1320	6598	2037	154,4	30,9
11	1314	6571	1873	142,5	28,5
12	1291	6453	1676	129,9	26,0
13	1105	5523	952	86,2	17,2
14	1029	5145	1197	116,3	23,3
15	980	4898	1316	134,3	26,9
16	880	4399	1466	166,6	33,3
17	800	4000	1603	200,4	40,1
18	744	3719	995	133,8	26,8
19	624	3118	1083	173,7	34,7
20	623	3114	1227	197	39,4
21	459	2295	470	102,4	20,5
22	483	2413	813	168,4	33,7
X	1502	7509	1211	80,6	16,1
Y	491	2457	119	24,2	4,8
Summa/keskmine	**30000**	**146070**	**36172**	**125,6**	**25,5**

Disainida üritatud proovide arvust saadi keskmiselt 25,5 protsenti. See tähendab, et ca 75 protsenti proovidest langeb disainimise käigus välja, kuna ei vasta kvaliteedinõuetele. Kõige vähem proove saadi Y kromosoomist. See on seletatav sellega, et inimese Y kromosoomis on väga palju kordusjärjestusi, mida aga proovide disainimisel välditi. Disainitud praimerite ja proovide olulisemad parameetrid on toodud tabelis 8. Siinkohal peab arvestama, et kõik parameetrid on oma keskmise (optimaalse) väärtuse suhtes normaaljaotusega.

Tabel 8. MAPHDesigner-iga disainitud praimerite ja proovide tähtsamad parameetrid.

Proovide andmed:	Keskmine pikkus:	**390.97**
	Max pikkus:	600
	Min pikkus:	200
	Keskmine GC sisaldus:	**46,39**
Praimerite andmed:	Keskmine pikkus:	**24,06**
	Max pikkus:	26
	Min pikkus:	22
	Keskmine GC sisaldus:	**49,2**

Proovide disainimiseks kulunud arvuti-aeg oli ligikaudu 240 tundi ehk ca 10 ööpäeva. Kõik disainitud proovid ja nende amplifitseerimiseks vajalikud praimerid on kättesaadaval andmebaasi PrimerStudio alajaotuses *maphPCR* (T. Kõressaar, http://bioinfo.ebc.ee/PrimerStudio/).

Arutelu.

Saadud tulemuste järgi võib arvata, et disainitud hübridiseerimisproovid on kõrge kvaliteediga ja sobivad hästi nii kiibil läbiviidava CGH (*array-CGH*) kui ka multipleks paljundatavate proovide hübridisatsiooniks. Muidugi oleks vaja proovide kvaliteedi tõestamiseks teha praktilisi katseid, sest kuigi *in silico* on proovid sobivad, ei tea me täpselt, kuidas nad käituvad *in vitro*. Pisut ettearvamatult võivad proovid käituda juba sellepärast, et siiamaani pole päris täpselt teada, kuidas toimub pikkade järjestuste omavaheline hübridiseerumine ja kui pikad võivad olla komplementaarsed alad, et proovid ei rist-hübridiseeruks. Levinud arvamus on, et kõigepealt seonduvad omavahel komplementaarselt kõrgema GC nukleotiidide sisaldusega alad ja alles seejärel ülejäänud proovi alad.

Disainimise algoritmi täiustamiseks võiks proovida erinevate programmide töötamise järjekorra varieerimist. Võimalik, et algoritmi saab niimoodi veelgi kiiremaks teha. Üldine programmide töötamise järjekord peaks alluma loogikale, mille puhul kohe alguses eemaldatakse proovid ja praimerid, mis ei sobi ja kõige aeganõudvamad protsessid (genoomitest) viiakse läbi minimaalse arvu negatiivsete proovidega.

Samuti aitaks proovide kiiremale disainimisele kaasa, kui jagada üksteisest mittesõltuvad protsessid arvutamiseks erinevatele arvutitele või panna kogu protsess tööle arvutiklastris.

Kasutaja jaoks saaks MAPHDesigner-it mugavamaks teha, kui arvestada, kui suur protsent proove õnnestub keskmiselt disainida sõltuvalt soovitud proovide arvust. Sel juhul oleks võimalik alguses võtta kadude võrra rohkem proove; see tähendab korrutada kasutaja poolt sisestatud proovide hulk läbi kindla kordajaga, et lõpptulemuses oleks piisav proovide arv esindatud. 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 soovitud proovide arvust.

MAPHDesigner on üks esimesi, kui mitte esimene omalaadne programmide pakett, millega saab automaatselt disainida pikki hübridiseerimisproove inimese genoomsetele järjestustele. Seetõttu on MAPHDesigner-i meetodi ja programmide

täiustamine edasises töös veel võimalik, arvestades tulevikus praktiliselt saadavate PCR-i ja hübridisatsiooni tulemusi.

Kokkuvõte.

Käesolevas töös tehti ülevaade mikrokiipidel põhinevatest aberratsioonide tuvastamise meetoditest, samuti anti ülevaade mikrokiipidel kasutatavate proovide ja nende amplifitseerimiseks vajalike praimerite disainimisest ja nii proovide kui ka praimerite kvaliteedi kontrollimisest.

Praktilises töös alustati ülesande püstituse ja eesmärkide seadmisega. Järgnes hübridiseerimisproovide disainimiseks vajaliku metoodika järk-järguline väljatöötamine ja arendamine ning tööks vajalike programmide kirjutamine. Meetodi arendamisel ja optimeerimisel arvestati nii kirjanduse andmetega kui ka praktiliste tulemuste analüüsil saadud andmetega.

Praimerite ja proovide disain algas inimese genoomi järjestuse ja muu vajaliku informatsiooni allalaadimisega EnsEMBL-i andmebaasist ja töötlemisega programmide poolt käsitletavale kujule. Järgnes kromosoomide kaupa toimunud praimerite ja proovide disain. Esimeseks sammuks selle juures oli sobivate järjestuste allalaadimine kohalikust andmebaasist, järjestuste maskeerimine, praimerite disainimine ja nende seondumiskohtade kontroll. Seejärel toimus PCR-i produktide ehk hübridiseerimisproovide kvaliteedi ja unikaalsuse kontroll kahekordse genoomitestiga ning andmete kokkupanek tulemus-faili. Disainitud hübridiseerimisproovide saamiseks vajalikud praimerid pandi praimerite andmebaasi PrimerStudio.

Kogu protsessi lihtsamaks käsitlemiseks loodi veebipõhine kasutajaliides MAPHDesigner, mis võimaldab kasutajal mugavalt disainida hübridiseerimisproove inimese genoomsetele järjestustele.

Summary

In this paper, we made overview about different methods for detecting aberrations in human genomic DNA. We also made overview and had a discussion about designing probes that can be hybridized on microchips and about primers for amplifying these probes.

In practical work, we started with defining our objectives. This was followed by development of the method and writing programs for automatic probe design. In the method elaboration, we considered with literature facts and the data, which we got from analysis of practically tested materials.

We first downloaded human genome sequences from EnsEMBL database and decoded them to be appropriate for our programs. Then probes were designed separately for every chromosome. First, we downloaded specific sequences from local database and masked them for repetitive sequences. Then we designed primers for each required sequence and tested those primers for cross-hybridization in the human genome. We also checked witch products (probes) would amplify with those primers. Designed probes were tested twice for cross-hybridization in the genome with programs SSAHA and MegaBLAST. The final step in probe design process was assembling all information about „good" probes and primers. All probes designed were added in the database called PrimerStudio.

To simplify the designing process for user, we made web interface called MAPHDesigner, witch can be easily utilized for designing long hybridization probes from human genomic DNA.

Kasutatud kirjandus.

Albertson D. G., Pinkel D. Genomic microarrays in human genetic disease and cancer. Human Molecular Genetics, 12:R145-R152, 2003a

Albertson D. G. Profiling breast cancer by array CGH. Breast Cancer Research and Treatment, 78:289-298, 2003b

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

Andreson R. Erinevate *in silico* meetodite võrdlus PCR praimerite kvaliteedi parandamiseks. M.Sc. thesis, 2002

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

Beheshti B., Park P. C., Brande I., Squire J. A. Microarray CGH. Methods Mol Biol, 204:191-207, 2002

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

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

Buckley P. G., Mantripragada K.K., Benetkiewicz M., Tapia-Paez I., de Stahl D. T., Rosenquist M., Ali H., Jarbo C., de Bustos C., Hirvelä C., Wilen B. S., Fransson I., Thyr C., Johnsson B-I., Bruder C. E. G., Menzel U., Hergersberg M., Mandahl N., Blennow E., Wedell A., Beare D. M., Collins J. E., Dunham I., Albertson D., Pinkel D., Bastian B. C., Faruqi A. F., Lasken R. S., Ichimura K., Collins V. P., Dumanski J. P. A full-coverage, high-resolution human chromosome 22 genomic microarray for clinical and research applications. Human Molecular Genetics, 11:3221-3229, 2002

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

Forozan F., Karhu R., Kononen J., Kallioniemi A., Kallioniemi O-P. Genome screening by comparative hybridization. TIG, 13:405-409, 1997

Gabig M., Wegrzyn G. An introduction to DNA chips: principles, technology, applications and analysis. Acta Biochimica Polonica, 48:615-622, 2001

Haas S., Vingron M., Poustka A., Wiemann S. Primer design for large scale sequencing. Nucleic Acids Research, 26:3006-3012, 1998

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

Henegariu O., Heerema N. A., Dlouhy S. R., Vance G. H., Vogt P. H. Multiplex PCR: critical parameters and step-by-step protocol. Biotechniques, 23:504-511, 1997

Hollox E. J., Akrami S. M., Armour J. A. L. DNA copy number analysis by MAPH: molecular diagnostic applications. Expert Rev Mol Diagn, 2:370-278, 2002

Hughes T. R., Mao M., Jones A. R., Burchard J., Marton M. J., Shannon K. W., Lefkowitz S. M., Ziman M., Schelter J. M., Meyer M. R., Kobayashi S., Davis C., Dai H., He Y. D., Stephaniants S. B., Cavet G., Walker W. L., West A., Coffey E., Shoemaker D. D., Stoughton R., Blanchard A. P., Friend S. H., Linsley P. S. Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. Nature Biotechnology, 19:342-347, 2001

Kallioniemi A., Kallioniemi O-P., Sudar D., Rutovitz D., Gray J. W., Waldman F., Pinkel D. Comparative Genomic Hybridization for Molecular Cytogenetic Analysis of Solid Tumors. Science, 258:818-821, 1992

Kallioniemi O-P., Kallioniemi A., Piper J., Isola J., Waldman F., Gray J. W., Pinkel D. Optimizing Comparative Genomic Hybridization for Analysis of DNA Sequence Copy Number in Solid Tumors. Genes, Chromosomes & Cancer, 10:231-243, 1994

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

Karhu R., Kahkonen M., Kuukasjarvi T., Pennanen S., Tirkkonen M., Kallioniemi O-P. Quality control of CGH: impact of metaphase chromosomes and the dynamic range of hybridization. Cytometry, 28:198-205, 1997

Lucito R., Healy J., AlexanderJ., Reiner A., Esposito D., Chi M., Rodgers L., Brady A., Sebat J., Troge J., West J. A., Rostan S., Nguyen K. C. Q., Powers S., Ye K. Q., Olshen A., Venkatraman E., Norton L., Wigler M. Representational Oligonucleotide Microarray Analysis: A High-Resolution Method to Detect Genome Copy Number Variation. Genome Research, 13:2291-2305, 2003

Mantripragada K. K., Buckley P. G., Diaz de Stahl T., Dumanski J. P. Genomic microarrays in the spotlight. Trends in Genetics, 20:87-94, 2004

Matveeva O. V., Shabalina S. A., Nemtsov V. A., Tsodikov A. D., Gesteland R. , Atkins J. F. Thermodynamic calculations and statistical correlations for oligo-probes design. Nucleic Acids Research, 31:4211-4217, 2003

Miller O. J., Therman E. Human Chromosomes. Springer-Verlag Inc., 2001

Nederlof P. M., Robinson D., Abuknesha R., Wiegant J., Hopman A. H., Tanke H. J., Raap A. K. Three-color fluorescence in situ hybridization for the simultaneous detection of multiple nucleic acid sequences. Cytometry, 10:20-27, 1989

Ning Z., Cox A. J., Mullikin J. C. SSAHA: A Fast Search Method for Large DNA Databases. Genome Research, 11:1725-1729, 2001

Pardue M. L., Gall J. G. Molecular hybridization of radioactive DNA to the DNA of cytological preparations. Proc Natl Acad Sci U S A, 64:600-604, 1969

Pinkel D., Straume T., Gray J. W. Cytogenetic analysis using quantitative, high-sensitivity, fluorescence hybridization. Proc Natl Acad Sci U S A, 83:2934-2938, 1986

Pinkel D., Segraves R., Sudar D., Clark S., Poole I., Kowbel D., Collins C., Kuo W. L., Chen C., Zhai Y., Dairkee S. H., Ljung B. M., Gray J. W., Albertson D. G. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. Nature Genetics, 20:207-211, 1998

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. Genome-wide analysis of DNA copy-number changes using cDNA microarrays. Nature Genetics, 23:41-46, 1999

Pollack J. R., Sorlie T., Perou C. M., Rees C. A., Jeffrey S. S., Lonning P. E., Tibshirani R., Botstein D., Borresen-Dale A-L., Brown P. O. Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. PNAS, 99:12963-12968, 2002

Relógio A., Schwager C., Richter A., Ansorge W., Valcárcel J. Optimization of oligonucleotide-based DNA microarrays. Nucleic Acids Research, 30:e51, 2002

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

Rouillard J-M., Zuker M., Gulari E. OligoArray2: design of oligonucleotide probes for DNA microarrays using a thermodynamic approach. Nucleic Acids Research, 31:3057-3062, 2003

SantaLucia J. Jr. A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. Biochemistry, 95:1460-1465, 1998

Schena M., Shalon D., Davis R.W., Brown P.O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science, 20:368-371, 1995

Schena M. Microarray Analysis. John Wiley & Sons, Inc., 2003

Sellner L. N., Taylor G. R. MLPA and MAPH: New Techniques for Detection of Gene Deletions. Human Mutations, 23:413-419, 2004

Solinas-Toldo S., Lampel S., Stilgenbauer S., Nickolenko J., Benner A., Döhner H., Cremer T., Lichter P. Matrix-based Comparative Genomic Hybridization: Biochips to Screen for Genomic Imbalances. Genes, Chromosomes & Cancer, 20:399-407, 1997

Stajich J. E., Block D., Boulez K., Brenner S. E., Chervitz S. A., Dagdigian C., Fuellen G., Gilbert J. G. R., Korf I., Lapp H., Lehväslaiho H., Matsalla C., Mungall C. J., Osborne B. I., Pocock M. R., Schattner P., Senger M., Stein L. D., Stupka E., Wilkinson M. D., Birney E. The Bioperl Toolkit: Perl Modules for the Life Sciences. Genome Research, 12:1611-1618, 2002

Suggs S. V., Hireose T., Kawashima M. J. Using Purified Genes. Symposium on Developmental Biology, 23, 1981

Tõnisson N., Kurg A., Lõhmussaar E., Metspalu A. Arrayed Primer Extension on the DNA Chip: Method and Applications. Microarray Biochip Technology, 2000

Wessendorf S., Lichter P., Schwänen C., Fritz B., Baudis M., Walenta K., Kloess M., Döhner H., Bentz M. Potential of chromosomal and matrix-based comparative genomic hybridization for molecular diagnostics in lymphomas. Ann Hematol, 80:B35-B37, 2001

Zakour B. N., Gautier M., Andonov R., Lavenier D., Cochet M. F., Veber P., Sorokin A., Le Loir Y. GenoFrag: software to design primers optimized for whole genome scanning by long-range PCR amplification. Nucleic Acids Research, 32:17-24, 2004.

Zhang Z., Schwartz S., Wagner L., Miller W. A Greedy Algorithm for Aligning DNA Sequences. Journal of Computational Biology, 7:203-214, 2000

Zhou X., Jordan R. C. K., Mok S., Birrer M. J., Wong D. T. W. DNA copy number abnormality of oral squamous cell carcinoma detected with cDNA array-based comparative genomic hybridization. Cancer, Genetics & Cytogenetics, 151:90-92, 2004