BIOLOOGIA-GEOGRAAFIA TEADUSKOND
MOLEKULAAR- JA RAKUBIOLOOGIA INSTITUUT
BIOINFORMAATIKA ÕPPETOOL

Kadiliina Kuusik

Automaatne PCR-i praimerite disain kõikidele inimese genoomi SNP-dele

Bakalaureusetöö

Juhendaja: Maido Remm, Prof., PhD

TARTU 2004

Lühendid ja mõisted

ADR	*adverse drug response*
amplikon	ühe PCR–i produktina amplifitseeritav DNA järjestus
bp	aluspaar, nukleotiidialuste paar
CGI	programmeerimiskeel dünaamilise veebilehe disainimiseks, (*Common Gateway Interface)*
cSNP	kodeeriva ala ühenukleotiidne polümorfism
flanking piirkond	amplikoni mõlemalt poolt külgnev järjestus praimerite seostumiseks
Gb	gigabait, 10^9 bp
geenikiip	= microarray = DNA chip
GST	*genespecific sequence tag*
e-PCR	praimerite unikaalsuse kontroll genoomsel järjestusel
kb	kilobait, 10^3 bp
Mb	megabait, 10^6 bp
LD	tasakaalustamata aheldus *(linkage disequilibrium)*
PCR	polümeraasi ahelreaktsioon *(Polymerase Chain Reaction)*
RFLP	*restriction fragment length polymorphism*
sensitiivsus	suurus, mis kajastab algoritmi võimet leida "õigeid" positiivseid (sn, *sensitivity*)
SNP	ühenukleotiidne polümorfism (*single nucleotide poymorfism)*
STR	*short tandem repeats*
target	märklaud järjestus e uuritav järjestus
Tm	sulamistemperatuur, temperatuur, mille juures pooled DNA ahelad on üheahelalised ja pooled kaheahelalised
VNTR	*variable number of tandem repeats*

Sissejuhatus

Seoses inimese genoomi täieliku sekveneerimisega on hakatud paremini mõistma selles esinevaid varieeruvusi**.** Kõige suurema osa genoomis esinevatest polümorfismidest moodustavad SNP-d (*single nucleotide polymorfisms*). SNP–d on ühealuspaarilised positsioonid genoomses DNA–s, kus mõnedes populatsioonides teatud normaalsetel indiviididel esinevad erinevad järjestuse alternatiivid (alleelid). SNP-de keskmiseks esinemissageduseks genoomis peetakse keskmiselt üks nukleotiid 300 aluspaari kohta. Praeguseks on Inimese Genoomi Projekti raames kaardistatud ligi 6 miljonit SNP-d. SNP markeritel on mitmeid eeliseid laiaulatuslike genoomiuuringute teostamiseks. Seoses SNP-de kasutamisega inimese genoomi geneetilise aheldatuse uuringutel on teoksil suuremahuline HapMap Projekt, mis püüab koostada võimalikult tihedat inimese genoomi haplotüübi kaarti. Samuti on teoksil erinevad epidemioloogilised uuringud SNP-de baasil.

Genotüpiseerimise esimeseks sammuks on tavaliselt uuritava järjestuse PCR–i amplifikatsioon. PCR ehk polümeraasi ahelreaktsioon on tänapäeva molekulaarbioloogide üks põhilisi meetodeid, leides rakendamist nii vundamentaal-, kui kliiniliste uuringute juures. PCR-i läbiviimiseks on meil vaja teada vastavat DNA järjestust, mille alusel disainitakse praimeripaar. Automaatseks praimeri disainiks on välja töötatud mitmeid algoritme ja programme, kuid siiani ei ole bioinformaatika jõudnud selles vallas täielikult rahuldava tulemuseni.

Käesoleva töö eesmärk on luua automaatne vahend SNP-de amplifitseerimiseks vajalike PCR-i praimerite disainiks.

I KIRJANDUSE ÜLEVAADE

1. Polümorfsed markerid inimese genoomis ja SNP-d

Inimese genoomis leidub hulgaliselt varieeruvaid järjestuse motiive, mida kasutatakse polümorfsete markeritena.

Esimesed kasutusele võetud geneetilised markerid olid restriktsioonifragmentide pikkuse polümorfismid (RFLP). Need on ühealuspaarilised varieeruvused (asendused, insertsioonid ja deletsioonid), mille esinemine restriktsioonisaidis muudab ära erinevate restriktaaside äratundmiskohad (Campbell *et al*., 2000). Kui DNA-s esineb polümorfism, on restriktaasi äratundmiskohad DNA-s inaktiveerunud, või vastupidi- on tekkinud uus lõikamiskoht, mida varem ei esinenud.

Minisatelliitjärjestused e VNTR-d (*variable number of tandem repeats*) koosnevad 6-64 bp pikkustest tandemitena korduvatest aladest, on kõrge alleelide arvuga ning paiknevad genoomis valdavalt telomeerides. Nende alade erinevate jaotuste põhjal saab koostada indiviidi genoomile unikaalse DNA „sõrmejälje" (Campbell *et al.,*2000). VNTR-de põhiliseks puuduseks on kordusühikute suurest pikkusest tingitud analüüsitulemuste keerukas interpreteerimine.

Laialt kasutatud markerid on mikrosatelliitjärjestused e STR-d (*short tandem repeat*), mis koosnevad tandemitena korduvatest 2-6 bp motiividest. Kõige levinumad dinukleotiidsed tandeemsed kordused on AC ja AT, vastavalt 50% ja 35%. Kõige levinumad trinukleotiidsed kordused on AAT ja AAC, vastavalt 33% ja 21% (Lander *et al*., 2001). STR markerite eeliseks on kõrge heterosügootsus, sage esinemine genoomis ning kuna nad on ka lühikesed, siis leiavad nad laialdast kasutamist genoomi kaardistamisel, evolutsioonilistes uurimustöödes, kohtumeditsiinis ja isaduse tuvastamisel (Strachan *et al*., 1999). Geenikaardistamisel on mikrosatelliitide puhul puuduseks kõrgest mutatsioonikiirusest (10^{-3} asendust nukleotiidi kohta põlvkonnas) põhjustatud ebastabiilsus

(Schafer, Hawkins, 1998). Alleelipikkuste täpne määramine on töömahukas protsess, mis teeb keerukaks nende kasutamise suuremahulistes genotüpiseerimise projektides ning raskendab täieliku automatiseeritud analüüsi väljatöötamist.

Viimase aja tööd põhinevad aga enamasti ühenukleotiidsetel polümofismidel ehk SNP-del (*single nucleotide polymorfism*). 90% DNA polümorfismidest moodustavad SNP–d (Collins *et al*., 1998). SNP–d on ühealuspaarilised positsioonid genoomses DNA–s, kus mõnedes populatsioonides teatud normaalsetel indiviididel esinevad erinevad järjestuse alternatiivid (alleelid). Definitsiooni järgi nimetatakse polümorfismiks sellist DNA järjestuse muutust, kus harvemini esineva alleeli sagedus populatsioonis on vähemalt 1%. Ühenukleotiidilised insertsioonid ja deletsioonid ei ole SNP–d (Brookes, 1999; Campbell *et al*., 2000).

SNP–d võivad olla kahe-, kolme-, või nelja-alleelsed polümorfismid. Kuna kahte viimast varianti peaaegu ei esine, siis kutsutakse tihti SNP–si bialleelseteks markeriteks. Bialleelsed SNP-d koosnevad neljast tüübist: ühest transitsioonist C↔T (G↔A) ja kolmest transversioonist C↔A (G↔T), C↔G (G↔C),T↔A (A↔T) (Brookes, 1999).

Nelja põhilise SNP tüübi sagedus ei ole inimese genoomis võrdne. 2/3 SNP–dest moodustavad C↔T (G↔A) variatsioonid. See võib olla seotud 5-metüültsütosiini deaminatsiooni reaktsioonidega, mis leiavad sagedasti aset CpG dinukleotiidides (Holliday, Grigg, 1993). Seega asuvad nad just geenisisestes alades, mistõttu seostatakse neid tihti haiguste tekkega (Brookes, 1999).

Põhjuseid, miks SNP-de populaarsus geneetiliste markerite seas järjest suureneb, on mitmeid. Esiteks paiknevad SNP-d genoomis teiste markeritega võrreldes tunduvamalt tihedamalt. Samuti asuvad nad genoomis nii valku kodeerivates, regulatoorsetes kui ka mittekodeerivates alades. SNP-de mutatsioonikiirus on madal (ligikaudu 10^{-8} asendust nukleotiidi kohta põlvkonnas) (Li *et al*., 1996), seega on ka nende põlvnemine stabiilsem. Peamiseks SNP-de eeliseks loetakse tänu nende bialleelsele loomusele analüüsi lihtsust ja unifitseeritavust, mis võimaldab neid analüüsida korraga tuhandeid või isegi sadu tuhandeid markereid.

1.1. SNP-de arv, tihedus ja genoomi katvus

Inimese genoom sisaldab 3,2 miljardit nukleotiidi (Kruglyak, 2001). Kaks inimese genoomi erinevad üksteisest keskmiselt iga 1000 bp-i tagant (Li *et al*., 1991; Reich *et al*., 2003). On palju uuringuid, kus inimese populatsiooni SNP-de keskmiseks tiheduseks arvatakse 300 bp-d. Järelikult leidub inimese genoomis ligikaudu 10 miljonit SNP-d (Kruglyak *et al*., 2001; Reich *et al*., 2003). Eesmärk omaette ei ole kõigi 10 miljoni SNP kaardistamine igas inimeses. Oluline on, et kaardistatud saavad genoomi haplotüübi blokkide markerid (lähema ülevaate teemast annab järgmine peatükk 1.2.).

Kogu genoomis võib SNP-de tihedus erineda kuni 100 korda. On leitud, et mõnes genoomi piirkonnas on see alla 0,1% (Nachman *et al*., 1998). Mitte-kodeerivates HLA regioonides jääb SNP-de tihedus 5% ja 10% vahele (Guillaudeux *et al*., 1998; Horton *et al*., 1998). Täpselt ei ole teada, millistes genoomi piirkondades oleks vaja tihedamat kaardistamist, seetõttu peetakse praegu parimaks lahenduseks identifitseerida võimalikult palju SNP-id (Tsui *et al*., 2003).

Tabel 1. SNP-de vaheliste kauguste analüüs *dbSNP* (2003) andmete põhjal. 24,6% intervallidest on suuremad kui 10 kb ja 54,9% on üle 3 kb pikad (Tsui *et al*., 2003).

SNP-de vaheliste kauguste (intervallide) jaotus *dbSNP* andmete põhjal, 2003

Intervalli suurus (kb)	intervallide arv	genoomist kui suure osa hõlmavad (%)
< 0.1	540 376	0.76
0.1 kuni 1	1 059 491	13.99
1 kuni 3	399 275	22.67
3 kuni 5	105 666	13.32
5 kuni 10	77 348	17.52
10 kuni 500	**38 497**	**24.22**
Kokku	2 220 653	92.32

Inimese SNP kaardi polümorfismide vahelisi kaugusi uurides on leitud, et suur hulk SNP-sid asuvad üksteisest kaugemal kui 3 kb (Tsui *et al*., 2003). Tabelist 1 on näha, et inimese genoomi kaardi koostamisel 300 000-1 miljoni SNP-ga jääb 24,6-54,9% inimese genoomist katmata.

Levinumateks uurimisobjektideks on kodeerivates alades leiduvad SNP-d e. cSNP-d. Kodeerivates eksonites on nukleotiidne diversiteet umbes neli korda madalam (eksonid

rohkem konseveerunud). cSNP-de keskmiseks esinemissageduseks on üks SNP 1,08 kb kohta. cSNP-dest pooled põhjustavad genoomis aminohappelisi e võimalikke funktsionaalseid muudatusi (Li, Sadler, 1991; Nickerson *et al*., 1998).

1.2. HapMap Projekt

Haplotüübi moodustavad vähemalt kaks ühel kromosoomi ahelal asuvat markeri varianti. Kromosoomil füüsiliselt järjestikku asuvad markerid moodustavad haplobloki, kui markerite vahel on tugev LD (*linkage disequilibrium*). LD on kahe (või enama) genoomse lookuse alleelide mittejuhuslik, reeglipärane assotsieerumine. Mida lähemal asub geneetiline marker haigusgeenile, seda väiksem on rekombinatsiooni tõenäosus nende kahe vahel ja suurem tõenäosus tasakaalustamata ahelduse tekkeks.

Kui *Human Genome Project* (HGP) (*The International Human Genome Sequencing Consortium*, 2001) on sekveneerinud 99,9% inimese genoomist, mille järgi oleme me kõik „sarnased", siis *HapMap Project*-i eesmärgiks on üles leida ja kaardistada indiviidide vahelised erinevused genoomis. HapMap Projekt (www.hapmap.org/) koostab inimese genoomi haplotüübi kaardi. Haplotüübi kaart e "HapMap" on vahend leidmaks kindlate haigustega seotud geene ja geneetilisi variatsioone. Geneetiliste analüüside lihtsustamiseks on eesmärgiks seatud selliste SNP-de kaardistamine, mis on ühele haplotüübile unikaalsed. Selliseid SNP-sid kutsutakse *tag*SNP-deks (*The International HapMap Consortium*, 2003).

Inimese genoomi haplotüübi arhitektuur varieerub suuresti. Küsimusele, kui pikk on see ala genoomis, kus endiselt on säilinud esialgne SNP-de muster, ei ole praegu teadlaste ringkondades ühest vastust. Haplotüübi blokk võib olla <1 kb kuni >100 kb (Daly *et al*., 2001; Reich *et al*., 2001; Gabriel *et al*., 2002).

Uuringud on näidanud, et praegune inimese genoomi haplotüübi kaart ei ole piisava tihedusega toetamaks HapMap Projekti. Umbes 50% genoomist moodustavad SNP-de vahelised intervallid, mis on suuremad kui 3 kb. 25%-l seni kaardistatud SNP-dest asuvad üksteisest kaugemal kui 10 kb (vaata tabel 1) (Tsui *et al*., 2003). Inimese genoomi haplotüübi arhitektuuri defineerimiseks on vajalik võimalikult paljude SNP-de kaardistamine.

1.3. Andmebaasid

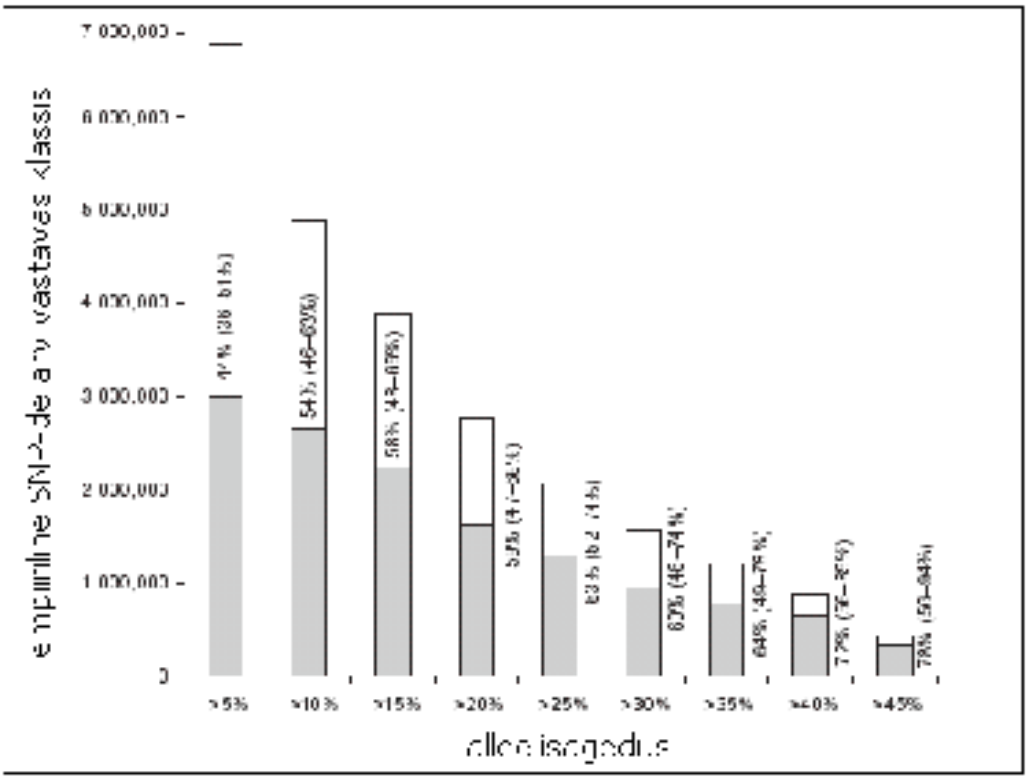
Tähtsaimat rolli SNP–de avastamises ja SNP–de andmebaasidesse kogumises omab the National Genome Research Institute (http://www.nhgri.nih.gov/), kellega teevad koostööd Genset, akadeemilised asutused, Genbank, The SNP Consortium (TSC) ja paljud teised. Kõige suurem SNP–de andmebaas on dbSNP (http://www.ncbi.nlm.nih.gov/SNP/). dbSNP hõlmab endas detailset informatsiooni seni avastatud, erinevate liikide genoomsete ja cSNP–de kohta, samuti on kirjas kasutatud meetodid, külgnevate järjestuste PCR–i tingimused jms.

Samuti avastatakse mõningal määral uusi SNP-sid erinevate kommertsiaalsete toodete välja töötamisel. Sellised on näiteks chr21 and Whole Genome Scanning Project (http://www.perlegen.com/), Affymetrix 120K chip Project (http://www.affymetrix.com/) ja ABI 150K set (http://press.appliedbiosystems.com/corpcomn/applerapress.nsf/ABIDisplayPress/), kus uute SNP-de detekteerimine põhineb teatavate genoomi piirkondade korduval sekveneerimisel erinevates indiviidides.

Kommertsiaalsetest andmebaasidest on suurim Celera andmebaas (www.celera.com/; Venter *et al*., 2001). Celera ja Applera on sidusettevõtted. ABI 150K on tehtud Celera andmeid kasutades.

On läbi viidud mitmeid uuringuid, mis püüavad hinnata hetkel olemasolevate SNP andmebaaside terviklikkust. Hinnang antakse peamiselt juhusliku valimjärjetuse resekveneerimise tulemuste alusel (Reich *et al*., 2003). Joonis 1.

Joonis 1. Ennustatav osa inimese genoomis olevatest, avalikes andmebaasides kättesaadavatest SNP-dest alleelisageduse funktsioonina. Uuringus resekveneeriti 173 kb Lääne-Aafrika ja Euroopa indiviidide 150-st kromosoomist. Halliga on märgitud osa, mis näitab seni identifitseeritud ja andmebaasides olevaid SNP-sid võrreldes kogu oletatava inimese genoomis leiduva SNP hulgaga (Reich *et al*., 2003).



1.4. SNP-de kasutamisest

Viimastel aastatel nähakse suurt potentsiaali SNP-de rakendamisel komplekshaiguste kaardistamisel, kasutades kogu genoomi hõlmavat assotsiatsioonianalüüsi (Collins *et al.,* 1997; Brookes, 1999; Kruglyak, 1999; Goldstein *et al*., 2003; Cardon, Palmer, 2003).

Põhiline SNP-de kasutusala on assotsitsiooniuuringud, kus uuritakse SNP-de kui geneetiliste markerite seotust haigusgeenidega nii tervetel kui haigutunnustega indiviididel (Kruglyak, 1999).

SNP-de kaardid kiirendavad haigusgeenide leidmist ning annavad sellega aluse märklaud-ravimite väljatöötamisele farmakogenoomikas (Rothberg, 2001).

Igal aastal sureb Ameerikas 100 000 inimest ADR-i (*adverse drug response*) tõttu. Uut suunda farmakogenoomikas nimetatakse personaliseeritud meditsiiniks (*personalized medicines*), mille põhimõte on, et iga inimene saaks haiguse korral õige ravimi (Roses, 2000).

Samuti leiavad SNP-d kasutust kohtumeditsiinilistes ja isikutuvastamise analüüsides ning molekulaarevolutsioonilistes uurimustöödes. Mitmed teadlased on seisukohal, et mitte-kodeerivates alades olevad SNP-d on parimateks markeriteks molekulaarevolutsioonilistes ja populatsioonigeneetilistes uurimustes, võimaldades meil lahendada ja mõista kaasaegsete inimpopulatsioonide päritolu ja levikut (Hacia *et al.,* 1999; Jorde *et al.*, 2000).

2. PCR-i praimerite disain

Annoteeritud genoomide järjestuste kättesaadavus tekitab üha rohkem vajadust eksperimentaalsete lähenemiste järele, mis tegelevad suure hulga geenide analüüsiga. Tänapäeval on vähe selliseid DNA-uuringuid, kus ei kasutataks polümeraasset ahelreaktsiooni ehk PCR-analüüsi. Olulisel kohal PCR-i õnnestumises on sobivalt valitud praimeripaar ehk kaks praimerit, mis kinnituvad meid huvitava DNA piirkonna 5` ja 3` otstele ning millest algab DNA süntees. PCR-i praimerite disainimise käigus leitakse kaks praimerit, mis seostuvad genoomis unikaalsesse kohta, andes sobiva pikkusega produkti. PCR õnnestumine on otseses sõltuvuses PCR-i kvaliteeti mõjutavatest parameetritest, milleks on praimerite sulamistemperatuur, PCR-i produktide pikkus, praimerite GC-sisaldus ning praimerite omavahelised interaktsioonid ja seostumiskohad genoomis. Seega, et praimerid ei seostuks genoomsel DNA-l juhuslikku kohta, iseenda või mõne teise praimeri külge, peavad nad olema väga petsiifilised. On oluline kontrollida PCR-i praimereid genoomi vastu, et vältida PCR-i kvaliteedi langust ja lisaproduktide teket. Eelnevaid ja muid probleeme, mis puudutavad kvaliteetsete PCR-i praimerite automaatset disaini, püüavad lahendada spetsiaalsed programmid ja algoritmid.

2.1. PCR-i praimerite kasutamine genotüpiseerimisel

Laiaulatuslik genotüpiseerimine nõuab korraga tuhandete kuni miljonite SNP markerite analüüsi (Syvänen, 2001).

SNP-de analüüsimiseks on tänapäeval võimalik kasutada mitmeid erinevaid tehnoloogiaid (Syvänen, 2001). Enamik kasutusel olevaid genotüpiseerimismeetodeid nõuavad kas enne või pärast genotüpiseerimist iga märklaud-järjestuse amplifitseerimist PCR-i abil, milleks on vaja SNP-d sisaldava *target*-spetsiifilisi praimereid.

Hilisemad tööd püüavad välja arendada tehnoloogiaid, kus ei ole vaja iga SNP jaoks unikaalset praimeripaari. Selliseks on WGA meetod (*whole genome amplifying*), kus ühe restriktsiooniensüümiga lõigatakse kogu genoomne järjestus tükkideks, ja seejärel kasutatakse selle üles amplifitseerimiseks ainult ühte universaalset praimerit (Matsuzaki *et al*, 2004). Meetodi puuduseks on, et selle abil saab genotüpiseerida väga väikese hulga genoomist, kuna restriktsioonifragmendid on väga erineva pikkusega, ja nende hulgast peab välja sorteerima ainult sobivad. Palju leidub väga suuri ja väga väikeseid fragmente, mida edasistesse genotüpiseerimisetappidesse ei kaasata.

PCR-i reaktsioon peab tagama nõutud sensitiivsuse/spetsiifilisuse eristamaks genoomides hetero- ja homosügootseid genotüüpe (Kirk *et al*., 2002).

Peale uuritavate lookuste amplifikatsiooni järgneb kas DNA järjestusvariantide kindlakstegemine lühikeste alleelspetsiifiliste hübridisatsiooniproovidega või ensümaatiliselt. Peamisteks ensüümideks, mida SNP-de tuvastamiseks kasutatakse, on DNA polümeraas või DNA ligaas, harvem kasutatakse restriktaase. Üks vanimaid genotüpiseerimismeetodeid on hübridisatsioon aleelspetsiifiliste oligonukleotiididega (ASO) (Sachidanandam *et al*., 2001; Venter *et al*, 2001). Tuntumaid hübridiseerumisel põhinevaid genotüpiseerimisplatforme pakuvad firmad Affymetrix (http://www.affymetrix.com/technology/index.affx) ja vähem tuntud Febit (http://www.febit.de/geniom/technology.htm). Spetsiifilisuse lisamiseks tekkisid genotüpiseerimistehnoloogiad, kus lisaks hübridisatsioonile rakendati ensümaatilisi reaktsioone. Ensüümideks võivad olla näiteks DNA polümeraas (Syvänen *et al*., 1990) või DNA ligaas (Landegren *et al*., 1998). DNA polümeraasi kasutavat ektensioonimeetodit (= *minisequencing*) rakendavad genotüpiseerimisel järgmised firmad: Orchid (http://www.orchid.com/technology/index.asp) , Sequenom (http://www.sequenom.com/Assets/pdfs/posters/abraunposter_8x11.pdf), nii nagu ka Eesti biotehnoloogia firma Asper Biotech (http://www.asperbio.com/technology.htm). DNA

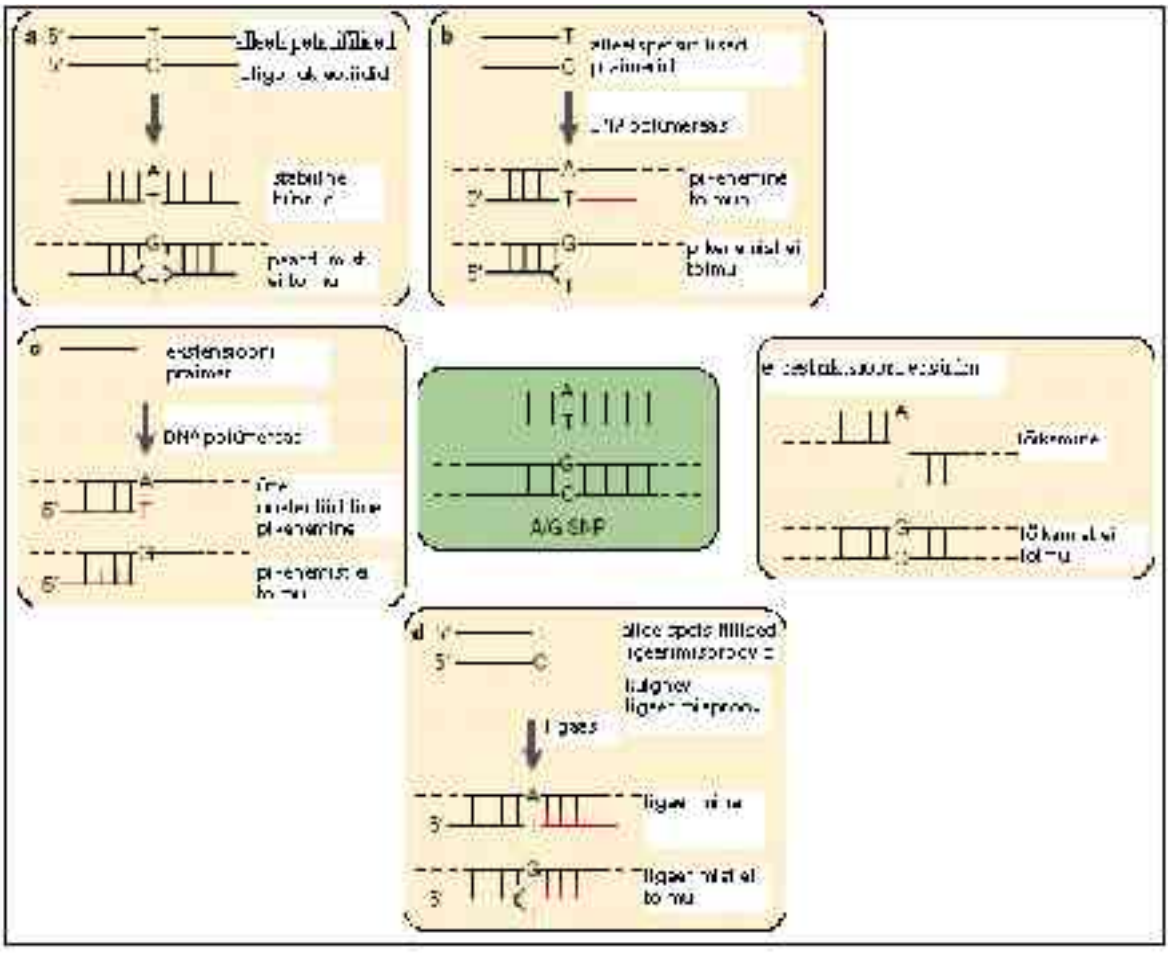
ligaasi kasutatavat meetodit nimetatakse oligo-ligeerimiseks, ja seda kasutab peamiselt Illumina (http://www.illumina.com/tech_overview.htm).

Igal neist meetoditest on spetsiifilised nõudmised oligote/praimerite disaini seisukohalt. Hübridisatsioonil alleelspetsiifiliste oligonukleotiididega (ASO-dega) kasutatakse kahte alleelspetsiifililist oligot, millede keskkohtades asuvad vastava uuritava järjestuse amplikoni SNP alleelidele komplementaarsed nukleotiidid. *Target*-iga moodustavad stabiilse hübriidi sellised oligod, mis sisaldavad endas keskelt amplikoni SNP alleeli. Sellist tehnoloogilist lähenemist kasutavad Affymetrix ja Febit platvormid. (Vaata joonis 2, paneel a.) Hübridisatsiooni alla kuulub veel genotüpiseerimine alleel-spetsiifiliste praimeritega, mille järjestus on komplementaarne *target*-i SNP-d ühele poole külgeneva alaga ja mis seostub sinna ainult siis, kui tema 3`otsas on nukleotiid, mis on komplementaarne *targeti*-i SNP alleelile. Pärast seostumist viib amplikoni ekstensiooni läbi DNA polümeraas. Sellist protsessi kasutavad meetodid on Taqman (Heid *et al*., 1996) ja *Molecular Beacon* (Marras *et al*., 1999; Tyagi *et al*., 1998). (Vaata joonis 2, paneel b.)

Ekstensioonimeetodite puhul kasutatakse ühte või kahte praimerit, mis on komplementaarsed *target*-i SNP ühele poole külgnevale järjestusele. Hübriidi SNP alleel määratakse DNA polümeraasi poolt, mis lisab ühe komplementaarse nukleotiidi SNP kohal vastavalt sellele, millist alleeli SNP sisaldab. Sellist genotüpiseerimistehnoloogiat kasutavad ettevõtted Orchid, Sequenom, Asper. (Vaata joonis 2, paneel c.)

Oligo-ligeerimisel kasutatakse kahte alleelspetsiifilist ligeerimisoligot ning amplikoniga komplementaarset ligeerimisproovi. Hübriidi moodustab see oligo, mille 3`ots on komplementaarne uuritava *target*-i SNP alleeliga. Oligo-ligeerimist kasutab Illumina. (Vaata joonis 2, paneel d.)

RFLP (*Restriction Fragment Length Polymorfism*) puhul kasutatakse restriktaasi, mis valitakse nii, et ta lõikaks DNA-d ainult SNP ühe alleeli väärtuse korral, teise alleeliga SNP koha pealt ei lõika. SNP väärtuse saab teada selle järgi, kas PCR produkt läks katki või mitte. Meetodit on raske automatiseerida, sest raske on valida sobivaid ensüüme ning neid omavahel kokku sobitada. (Vaata joonis 2, paneel e.)



Joonis 2. Erinevate genotüpiseerimistehnoloogiate biokeemiliste reaktsioonide põhimõtted (Syvänen, 2001). Joonisel on näidatud A/G transitsiooni A alleeli detektsioon (G alleel detekteerimine toimub analoogselt). a – hübridisatsioon alleelspetsiifiliste oligonukleotiididega (ASO-dega), b – alleelspetsiifiline praimeri ekstensioon, c - *minisequencing* e ühe nukleotiidiline praimeri ekstensioon, d – oligo-ligeerimine, e – alleelspetsiifiline restriktsioon.

2.2. PCR-i kvaliteeti mõjutavad parameetrid

2.2.1. Tm

Üks kõige olulisemaid PCR praimerite disaini juures arvestatavaid parameetreid on sulamistemperatuur (Tm). Sulamistemperatuuri ligikaudseks arvutamiseks kasutatakse valemit (Kämpke *et al*., 2001):

$$Tm = 4\ ºC\ *(G + C) + 2\ ºC\ *(A + T),$$

kus on näha, et sulamistemperatuur sõltub ainult nelja erineva nukleotiidi esinemissagedusest järjestusel. Vastavalt valemile, lisab iga A/T paar juurde 2 ºC, G/C paar aga 4 ºC. Seega, mida GC rikkama sisaldusega järjestus on, seda rohkem vesiniksidemeid on vaja lõhkuda ja seda kõrgem on sulamistemperatuur. Praimeri Tm-i täpsemaks arvutamise algoritmiks praimer-*target* interaktsiooni korral on *Nearest-Neighbor* e Lähima-Naabri meetod (Borer *et al*., 1974).

$$T_m = \frac{\Delta H}{\Delta S - R\ln(c/4)} + 16.6 \log \frac{[K^+]}{1 + 0.7[K^+]} - 273.15$$

kus H- dupleksi entalpia (kirjeldab protsessi soojussisaldust), S- dupleksi entroopia (kirjeldab süsteemi korrapäratust), R- universaalne gaasikonstant (R=8.31J/mol), c- oligonukleotiidide totaalne molaarne konsentratsioon (Rychlik *et al*., 1990). Breslauer koos kaasautoritega on välja arvutanud kõigi võimalike nukleotiidipaaride seostumise kombinatsioonidel tekkivad energiaväärtused (Breslauer *et al*., 1986). Ideaalseks praimeri pikkuseks loetakse vahemikku [20, 25] bp-d. Optimaalne sulamistemperatuur langeb vahemikku [50, 70] ºC .

2.2.2. Produkti pikkus

Erinevate PCR meetoditega saab üles amplifitseerida erineva pikkusega DNA fragmente. Produkti pikkus on tavaliselt vahemikus 100 kuni 3000 bp. Kuigi eksperimentaalselt on tõestatud, et lühemate PCR produktide korral on PCR-i tulemused paremad ehk saadavad produktid on spetsiifilisemad (identsus lähima paraloogse järjestusega väiksem) (Thareau *et al*., 2003), on viimasel ajal arendatud välja mitmeid PCR-i meetodeid just pikkade, kuni 6kb pikkuste DNA fragmentide amplifitseerimiseks

(Kämpke *et al*., 2001). Paljud hilisemad uuringud on jõudnud järeldusele, et PCR-i ebaõnnestumise põhjuseks on pigem ebakvaliteetne DNA uuritav järjestus, kui valesti valitud praimerid ja et PCR-i praimerite automaatsel disainil pööratake rohkem tähelepanu PCR-i praimerite valikut optimiseerivatele parameetritele (Benita *et al*., 2003). Produkti iseloomustavateks parameetriteks on tavaliselt ainult produkti pikkus ja produkti Tm (Rychlik *et al*., 1993). Samas on teada, et raske on amplifitseerida väga kõrge või madala GC sisaldusega produkte (Baskaran *et al.*, 1996).

2.2.3. Praimerite GC sisaldus

Väiksema GC sisaldusega praimereid tuleks pikendada vastavalt nii palju, et vajalik sulamistemperatuur oleks üle 50°C. Praimeri sptsiifiliseks seostumiseks DNA järjestusega on oluline tema 3`otsa GC nukleotiidide sisaldus. 3`otsa nukleotiidse sisalduse optimaalsuse printsiip ei ole veel täielikult välja töötatud, sest mõned teooriad väidavad, et õige oleks kasutada GC rikast 3`otsa, mõned mitte. On läbi viidud uuringuid, mis on näidanud, et väga kõrge GC sisaldus võib põhjustada genoomis vale seostumise (Li *et al*., 1997). On väidetud, et praimeri 3`otsas peaks selle stabiliseerimiseks olema G- või C-nukleotiid (Henegariu *et al*.,1997).

2.2.4. Praimerite omavahelised interaktsioonid ja seostumiskohad genoomis

Unikaalse PCR-i produkti saamiseks tuleb disainitud praimereid testida, sest peale selle, et nad seostuvad uuritava genoomse DNA-ga, võivad praimerid hübridiseeruda iseendaga või vastaspraimeriga, samuti on võimalik praimeri alternatiivne seostumine genoomsel DNA-l. Rohkema kui ühe produkti amplifitseerimine võib anda vale-positiivseid signaale genotüpiseerimisel. Viimase põhjustajateks on põhiliselt praimerid, mis sisaldavad DNA kordusjärjestuste motiive. Seega, enne praimerite disaini on oluline kordusjärjestuste maskeerimine (vt. peatükk 2.3.2).

Praimerite seondumiskohtade arvu ja asukoha leidmiseks saab kasutada erinevaid programme, millest tuntumad on BLAST (*Basic Local Alignment Search Tool*) (Altschul *et al*., 1997), MEGABLAST (Zhang *et al*., 2000), SSAHA (*Sequence Search and Alignment by Hashing Algorithm)* (Ning *et al*., 2001) ja GTESTER GenomeMasker-i paketist (Reppo *et al*., unpublished).

2.3. Praimerite unikaalsuse tagamise meetodid

2.3.1. Homoloogia otsinguteks mõeldud programmid

BLAST (http://www.ncbi.nih.gov/blast/; Altschul *et al*., 1997) on mõeldud sarnasuse otsinguteks erinevatest järjestuse andmebaasidest. BLAST-i üheks tugevaks küljeks on mitmekülgsus, kuna tal on palju erinevaid alamprogramme erinevat tüüpi otsingute jaoks. Pikkade järjestuste uurimisel kasutatakse MEGABLAST-i (Zhang *et al*., 2000).

SSAHA (http://www.sanger.ac.uk/Software/analyses/; Ning *et al*., 2001) on BLAST tüüpi programmidest kiireim, paisates järjestusinformatsiooni paisktabeli (*hash)* kujulisse andmestruktuuri, kust edasi toimub efektiivne sarnasuste otsing. Suur kiirus aga saavutatakse mälumahu arvelt (programm vajab vähemalt 1GB muutmälu). Kõige sobivam PCR-i praimerite testimiseks on spetsiaalselt selleks otstarbeks kirjutatud programm GTESTER, mis loendab genoomis kõik vastava praimeripaari esinemiskohad ja testib mitu PCR produkti nad tekitavad.

2.3.2. Korduste leidmiseks ja maskeerimiseks mõeldud programmid

Imetajate genoomist moodustavad kordusjärjestused väga suure osa, jäädes enamus ektrageensesse genoomi osasse. Samas ei saa öelda, et sellised piirkonnad on ebainformatiivsed, kuna nende abil uuritakse evolutsiooni mehhanisme ja kiirust. Samuti võivad ka rohkete kordusjärjestustega alad sisaldada endas geene, geenide jäänukeid, SNP-sid jne. Praimerite disainil on aga kasulik korduseid sisaldavad alad ära maskeerida, kuna see vähendab võimalust praimerite seostumist alternatiivsetesse kohtadesse genoomis.

DUST on kordusjärjestuste maskeerimise programm, mille abil saame järjestusest välja filtreerida bioloogiliselt ebaolulised kohad (ftp://ftp.ncbi.nih.gov/pub/tatusov/dust/version1/). DUST maskeerib DNA järjestuses vähese keerukusega alad (*low complexity DNA*), st alad, mille nukleotiidne koostis ei ole piisavalt mitmekesine (lühikesed ühe-, kahe-, ja kolmenukleotiidsed kordused). Algoritm on vaikimisi kasutatav BLASTN programmi poolt teostavates DNA homoloogiaotsingutes. Programmi autoriteks on Roman Tatusov ja David Lipman.

RepeatMasker on programm, mis kontrollib FASTA formaadis DNA järjestust teadaolevate kordusjärjestuste andmebaasi vastu, tagastades maskeeritud järjestuse, mis on valmis erinevateks andmebaasi otsinguteks ja vastava maskeeritud regiooni annoteerimiseks (http://repeatmasker.genome.washington.edu/RM/RepeatMasker). Programmi väljundiks on sisendjärjestuse detailne korduste annotatsioon ja järjestus, kus kõik annoteeritud kordused on maskeeritud ehk asendatud N-idega. Üheks võimalikuks väljundfailiks on otsinguks kasutatud järjestuse joondus kokkulangevate kordusjärjestustega andmebaasis. Kasutajal on võimalus valida, milliste korduste vastu tahab ta maskeerimist sooritada. Otsingu sooritamiseks on kolm võimalust. Vaikimisi maskeeritakse nii hajuskordusjärjestused kui ka lihtsad tandeemsed kordused, polypuriini, AT- ja/või GC rikkad piirkonnad ja mikrosatelliidid. "*Mask only simple repeats"* optsiooni korral saadakse tulemus ainult polümorfsete lihtsate korduste kohta. Vastupidiselt, "*do not mask simple repeats"* valiku korral maskeeritakse ainult hajuskordusjärjestused. Järjestuste võrdlemine viiakse läbi programmi *cross-match* abil, mis on Smith-Waterman-Gotoh algoritmi implementatsioon.

Programmi kiirus ja sisendjärjestuse pikkus on lineaarses seoses, ehk mida lühem järjestus, seda lühem tööaeg. Programmi tundlikkust on võimalik suurendada või vähendada kiiruse arvelt. Tundlikum otsing võtab kolm korda kauem aega, aga maskeeritakse 0-5% rohkem kordusi kui vaikimisi. Kiire otsing on küll 3-5 korda kiirem, kuid võib märkamata jätta 5-10% vaikimisi maskeeritud kordustest. Programmi autoriks on Phil Green.

GenomeMasker on programmipakett, mis on mõeldud korduste leidmiseks ja vältimiseks PCR-i praimerite disainil. GenomeMasker-i programmipaketti kuuluvad kaks korduvate järjestuste maskeerimise ja nende leidmise programmi GMASKER ja

GTESTER (http://www.ismb02.org/posters/poster/Remm_1.pdf; Reppo *et al*, unpublished). GMASKER maskeerib FASTA formaadis järjestusest üleesindatud sõnad vältimaks praimerite disainil praimeri sattumist kordusjärjestusi sisaldavatele piirkondadele. Programm ei kasuta järjestuse joondamise algoritme ja seega on tema töökiirus küllalt kõrge. Selle asemel kasutab ta programmi BLACKLISTER poolt koostatud üleesindatud sõnade nimekirja (*black list*). Sõna pikkust on võimalik varieerida vahemikus [8, 16]. Kasutades sellist nimekirja, maskeerib programm GMASKER kõik korduvad piirkonnad. GMASKER-i maskeerimisprotsessi põhimõte on sama, mis RepeatMasker-il, kuid esimese puhul võib tuua välja mõningaid eeliseid. Esiteks, otsing teostatakse fikseeritud sõnapikkuste vastu, mis on täpsem, kui otsingu teostamine levinumate kordusjärjestuste vastu. Teiseks, maskeerib GMASKER ainult üleesindatud sõnade 3`otsa, mis võimaldab praimerite disaini ka siis, kui praimeri 5`osa asub osaliselt kordusjärjestuses. Kogu sõna maskeerimisel võib juhtuda, et maskeeritakse liiga suured alad, mis võivad siiski sisaldada meile huvipakkuvat informatsiooni. Kuna maskeeritakse ainult korduses olevate sõnade 3`ots, on GMASKER-i maskeerimine asümmeetriline ja maskeeritud nukleotiidid võivad kummalgi DNA ahelal olla erinevad.

Programmi GTESTER abiga saab ka leida PCR-i produktide arvu genoomis.

2.4. Erinevad PCR-i praimerite disaini programmid

Suuremamahuliste projektide jaoks on mõeldud PRIMER3 (Rozen *et al*.,1998), PRIDE (Haas *et al*., 1998), GST-PRIME (Varotto *et al*., 2001), PRIMO (Li *et al*., 1997), PRIMER MASTER (Proutski *et al*., 1996) jt. Väiksemate praimerikoguste disainimiseks on programmid - OLIGO (Rychlik *et al*., 1989), PRIME, PRIMERSELECT jt. Paljud programmid on mõeldud praimerite disainimiseks ainult bakteriaalsetes genoomides, mistõttu ei saa neid kasutada geenide puhul, mis sisaldavad introneid. Üheks selliseks on näiteks PRIMEARRAY (Raddatz *et al*., 2001). Teine programmide klassifitseerimise võimalus on nende automatiseeritus ehk kui palju peab praimeri disaini juures vahele sekkuma kasutaja. Heaks näiteks on siinkohal PRIDE, kus ei eemaldata ühtegi praimeri kandidaati enne, kui on välja arvutatud kõik parameetrid kõikidele praimeritele (Haas *et*

al., 1998). GST-PRIME-i teeb eelistatuks see, et disainimist saab alustada juba valgu järjestuse tasemel. PRIMERSELECT pakub palju võimalusi nii praimerite disainimiseks kui kasutaja enda olemasolevate praimerite analüüsimiseks. Vajaduse korral saab kasutaja PRIMERSELECT-i poolt vaikimisi valitud termodünaamilistele parameetritele valida erinevate alternatiivide vahel. Tulemuste analüüsimise hõlpsustamiseks on illustratsioonid, graafikud, statistilised aruanded. OLIGO on samuti laialt kasutatav kommertsiaalne programm, mis on väga mitmekülgse kasutajaliidesega ja mida on täiustatud pidevalt juba aastast 1989.

Esimeseks sammuks iseloomustamaks geeni funktsionaalsust on tema ekspressiooni mustri identifitseerimine. Siin on olulised programmid, mis praimerite disainil võtavad arvesse geenistruktuuri, kuna meil on vaja geeni mRNA-le spetsiifilisi praimereid (Freeman, 2000). Koos mikrokiipide arenguga on viimasel ajal loodud üha rohkem programme, mis püüavad leida geenist temale kõige geenispetsiifilisema osa ehk GST (*gene specific tag*) ning seejärel disainivad sellele piirkonnale vastavad praimerid. Selliste unikaalsete piirkondade kasutamine praimeri disainil asendab edukalt cDNA-de kasutamise, aidates samamoodi vältida rist-hübridisatsiooni sündmusi DNA mikrokiipidel (Hijum *et al*., 2003; Thareau *et al*., 2003). Üheks selliseks on Unifrag ja GenomePrimer-i pakett, mis on rakendatav ainult bakteriaalsetele genoomidele (Hijum *et al*., 2003). SPADS (*The Specific Primer and Amplicon Design Software*; http://genoplante-info.infobiogen.fr/spads) on rakendatav nii bakteriaalsete kui eukarüootsete genoomide puhul ja mille peamiseks kriteeriumiks geenispetsiifiliste praimerite valikul on amplikoni spetsiifilisus (Thareau *et al*., 2003). Üheks samalaadseks on ka kommertsiaalne GenomePRIDE, mis võimaldab nii pikkade oligomeeride (40-70 bp) kui geenispetsiifiliste PCR-i praimerite automaatset disaini (Haas *et al*., 2003).

Osasid praimeri disaini tarbeks loodud programme saab kasutada väga spetsiifilistel eesmärkidel. Näiteks on loodud programm, mis disainib spetsiaalselt minisekveneerimiseks mõeldud parimereid (Kaderali *et al*., 2003).

II PRAKTILINE OSA

Eesmärk

Käesoleva bakalaureusetöö eesmärgiks oli luua terviklik lahendus kõikide inimese SNP-de üles amplifitseerimiseks vajalike PCR-i praimerite automaatseks disainiks. Töö sisaldas olemasolevate praimerite disainimise ja testimise programmide kombineerimist nii, et praimerid leitakse võimalikult suurele osale seni identifitseeritud SNP-dele. Iga bioinformaatiline programm ja selle tulemused ei oma lõplikku funktsionaalsust, kui neid ei saaks hõlpsalt ja kiirelt kasutada. Selle tarbeks on viidud tulemused andmebaasi, et luua kasutajasõbralik veebilahendus. Minu poolt tehtud töö on mooduliks/osaks veebi portaalist nimega *PrimerParadise*, kus lisaks SNP-dele, saab praimereid leida ka eksonitele ja

hübridisatsiooni oligotele. Kaugemaks eesmärgiks on programmi tulemuste õigsuse kontrollimine praktikas ja selle pidev täiustamine (näiteks võimalikult paljude annoteeritud organismide kaasamine portaali).

Meetodid

1. Andmete päritolu ja struktuur

Andmed on saadud kohalikust bioinformaatika õppetooli serveri mysql andmebaasi *human_34* tabelist *dbsnp_119*, mis on annotatsioonikeskuse NCBI kodulehel (http://www.ncbi.nih.gov/) asuva tabeli dbsnp versioon 119-e kohandatud versioon (ftp://ftp.ncbi.nih.gov/snp/human/).

2. Kasutatud riistvara

Praimeridisaini programmide kirjutamiseks ja käivitamiseks kasutasime riistvarana arvutit MicroLink Novator 5000HG. Antud arvuti on kaheprotsessoriline, mõlemate protsessorite kiirus on 2.6 GHz. Arvuti maksimaalne põhimälu ehk RAM on 6 GB, kõvaketta maht on 5 * 150 GB.

3. Praimeri disainiks vajalike skriptide kirjeldus, tööpõhimõte

Esimeseks etapiks SNP-dele PCR-i praimerite disainil on SNP kohta käivate andmete allalaadimine. Seda viis läbi programm nimega *SNP_tombamine.pl*, mille töö resultaadiks oli tekstifail nimega *snp.txt*, mis koosnes järgmistest SNP kohta käivatest andmetest: SNP ID; kromosoomi number; positsioon; ja signaal.

Pärast SNP andmete kättesaamist käivitasime programmi nimega *primers.pl*, mis hakkas SNP-dele praimereid disainima. *Primers.pl*-s on kasutatud kahte andmebaasi: kohalikku ning Sangeri Instituudi poolt pakutavat (kaka.sanger.ac.uk). Kohaliku

andmebaasi nimi on *human_34*, millest käesoleva töö eesmärgi täitmiseks kasutame tabelit *dbsnp_119*. Viimase alusel saime esimese etapi tulemusena tekstifaili *snp.txt*. Sangeri Instituudi poolt pakutav andmebaas on inimese genoomi 34 versiooni *core_34* andmebaas, kus asub inimese genoomi nukleotiidne järjestus.

Järgmisena võtsime failist *snp.txt* SNP ID ja liitsime tema ette ja taha 100 nukleotiidi. Selle tulemusena saime 201 nukleotiidi pikkuse PCR-i abil ülesamplifitseeritava produkti ehk amplikoni. Et tagada igale amplikonile unikaalne praimeripaari disain, siis maskeerisime kõik amplikonid kordusjärjestuste vastu. Kõigepealt kasutasime DUST programmi, mis maskeerib kõik lihtsad ja lühikesed kordused. Teisena maskeerisime programmi GMASKER abiga, mis aitas meil vabaneda pikematest ja keerulisematest kordusjärjestustest. GMASKER kasutab oma töö siseselt kahte faili: *blacklist_10*-t ja *blacklist_30*-t. Esimeses on üle 10 korra, teises üle 30 korra genoomis esinevad 16 nukleotiidi pikkused sõnad (kirjed).

Peale korduste maskeerimist amplikonis algas praimerite disain. Selleks kasutasime programmi PRIMER3 (meie poolt kohandatud programmi nimi *primers.pl*), mille eesmärgiks oli igale amplikonile disainida üks praimeripaar. *Primers.pl* sisendiks on sellised PCR-i praimerite parameetrite lubatud väärtused nagu on näha tabelis 2.

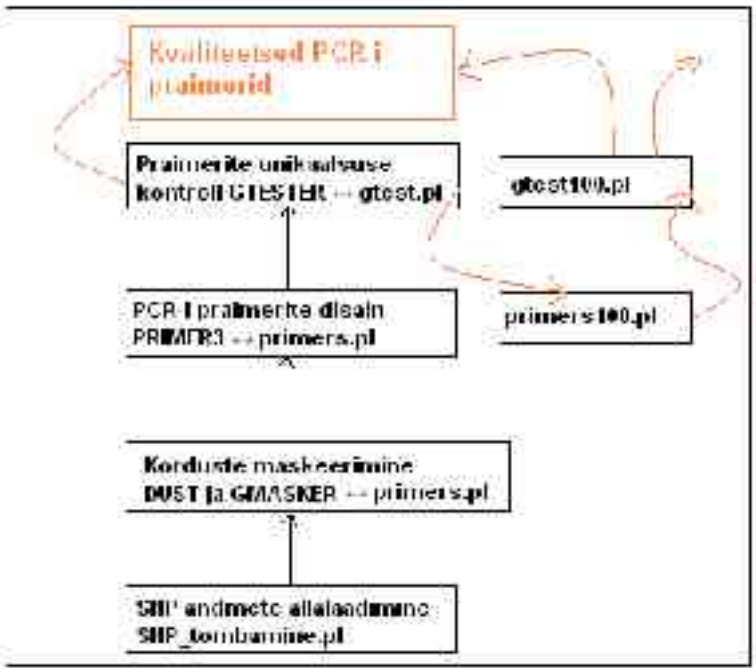
Tabel 2. Primer3-e sisendformaadi parameetrite lubatud väärtused.

PRIMER_PRODUCT_OPT_SIZE=600
PRIMER_OPT_SIZE=21
PRIMER_MIN_SIZE=16
PRIMER_MAX_SIZE=30
PRIMER_OPT_TM=62
PRIMER_MIN_TM=59
PRIMER_MAX_TM=65
PRIMER_MAX_DIFF_TM=4
PRIMER_OPT_GC_PERCENT=35
PRIMER_MIN_GC=20
PRIMER_MAX_GC=80
PRIMER_SALT_CONC=20
PRIMER_MAX_POLY_X=4
PRIMER_NUM_RETURN=1

Paraku ei suutnud PRIMER3 ühe sellise tsükli tulemusena kõigile amplikonidele häid praimereid leida, seega korrati protsessi tsükliliselt. Iga uue tsükli käikuminekuga pikendati külgnevate piirkondade (*flanking* piirkondade) pikkusi 100 nukleotiidi võrra ehk siis maksimaalselt võis amplikon olla 2401 nukleotiidi pikk. Pärast 5. tsüklit toimus kordusjärjestuste maskeerimine vähem range kordusjärjestuste nimekirjaga (*blacklist*_30-ga). Sobiv praimeripaar püüti leida kõikidele amplikonidele. Selliseid tsükleid korrati kuni 12 korda. Kui ka sel juhul ei suutnud programm praimeripaari leida, siis neid amplikone enam edasistesse etappidesse ei kaasatud. Meie poolt disainitud head praimerid läksid kataloogi nimega *good/*.

Järgmiseks etapiks oli programmi GTESTER kasutamine. GTESTER kontrollis kõikide praimerite esinemist genoomis. Programmi sisendiks oli PRIMER3 poolt genereeritud väljundkataloog nimega *good/*. Kui praimer esineb genoomis rohkem kui üks kord, siis tema ID jäetakse meelde ja vastav praimeripaar kustutatakse mälust. Vastava SNP ID, mille praimeriga oli tegu, panime kataloogi nimega *bad_primers_id/*. Genoomis üks kord esinevad praimeripaarid läksid kataloogi nimega *final_primers/*.

Eelmises etapis välja läinud praimeritega SNP-dele üritasime disainida uued praimerid põhimõttel, et kohandatud programm PRIMER3 disainib nendele amplikonidele ühe asemel maksimaalselt 100 praimeripaari. Seejärel kohandasime ka GTESTER programmi sel viisil, et kontrollitakse kõigi 100 praimeripaari esinemist genoomis. Kui leitakse unikaalne praimeripaar, siis toimitakse sarnaselt eelnevatele etappidele.



Joonis 3. PCR-i praimerite disaini etapid ja meie poolt originaalprogrammide (suured tähed) alusel loodud analoogprogrammid (väikesed tähed, laiendiga .pl).

Andmebaasi tabeli koostasime kataloogi *final_primers/* failide ja tekstifaili *snp.txt* alusel.

4. Kasutatud programmide tööaeg ja mälukasutus

Tabel 3. Programmide tööaeg sekundites.

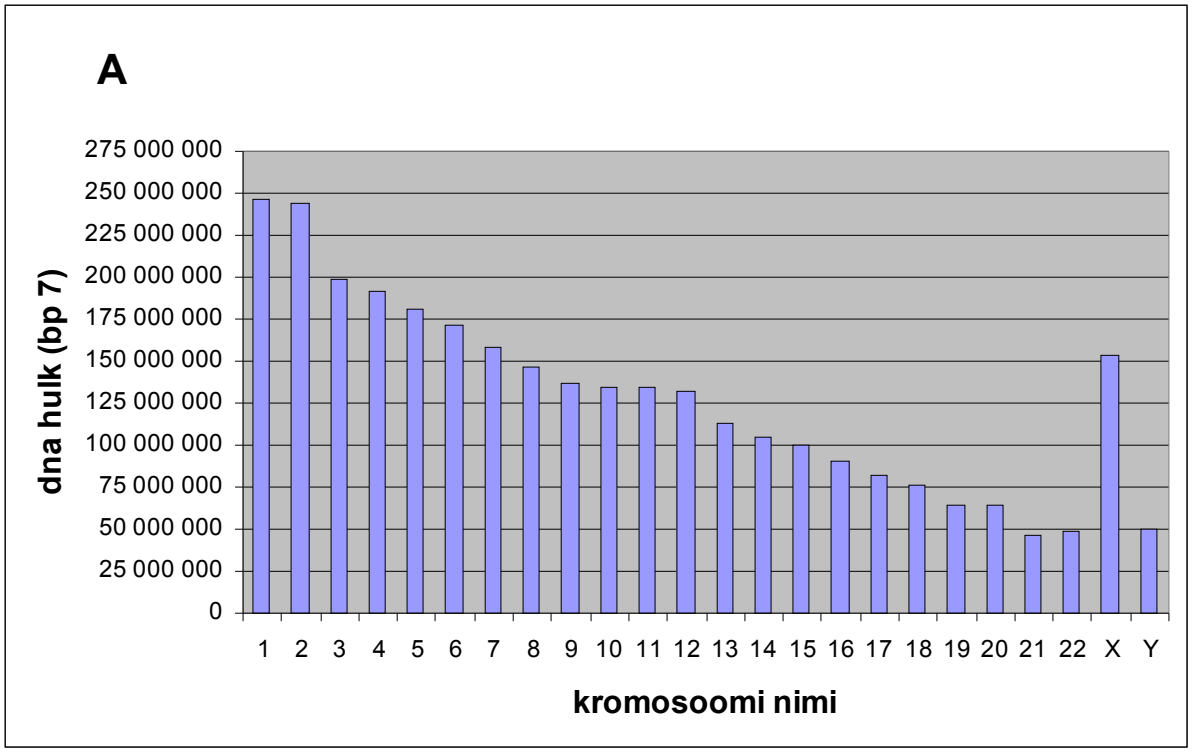
jrk.nr.	protsessi nimetus	programmi nimi	tööaeg (sek)
1.	SNP andmete allalaadimine	*SNP_tombamine.pl*	2940 (6 h;13 min)
2.	igale amplikonile disainitakse 1 praimeripaar	*primers.pl*	45960 (5 ööpäeva;5 h;16 min)
3.	praimerite unikaalsuse kontroll	*gtest.pl*	4320 (9 h;18 min)
4.	gtestiga välja läinud amplikonidele disainitakse		
	maksimaalselt 100 praimeripaari	*primers100.pl*	17760 (2 ööpäeva;8 min)
5.	praimerite unikaalsuse kontroll	*gtest100.pl*	65160 (7 ööpäeva;12 h;6 min)
	Kõikide inimese genoomi SNP-de PCR-i praimerite disainiks kulunud aeg kokku		136 140 (15 ööpäeva;18 h;1 min)

Kõige rohkem mälumahtu vajasid programmid, mis kontrollisid konstrueeritud praimerite seondumiskohtade unikaalsust genoomis. Vajalik mälumaht küündis 350–450 Mb-ni ja viimast peamiselt tänu programmile GTESTER, mis jätab meelde praimerite järjestused ja nende esinemised genoomis ning samuti indeksid, mis on loodud kogu inimese genoomist programmi GTESTER algoritmi realiseerimiseks.

Tulemused

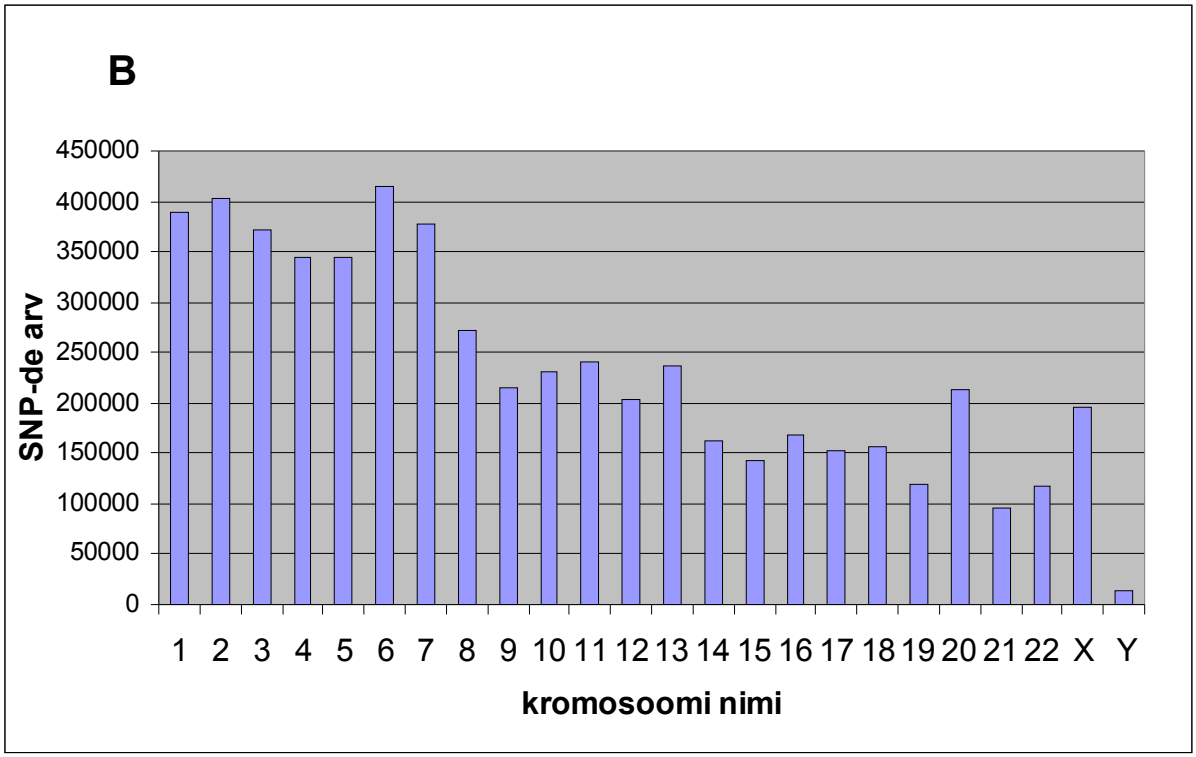
1. Tulemuste kirjeldus

Inimese tuuma DNA hulk on määratud kromosoomide arvu ja suuruse poolt. Inimese genoomi moodustab 22 kromosoomi ning X ja Y kromosoom. Joonis A näitab DNA hulga jaotumist inimese genoomi kromosoomide vahel. Andmed on saadud ENSEMBL kodulehelt (www.ensembl.org/). Kõige hilisema versiooni 21.34d.1 (viimati muudetud 7. mai 2004.a.) järgi koosneb inimese genoom 3 223 443 491 bp-st.



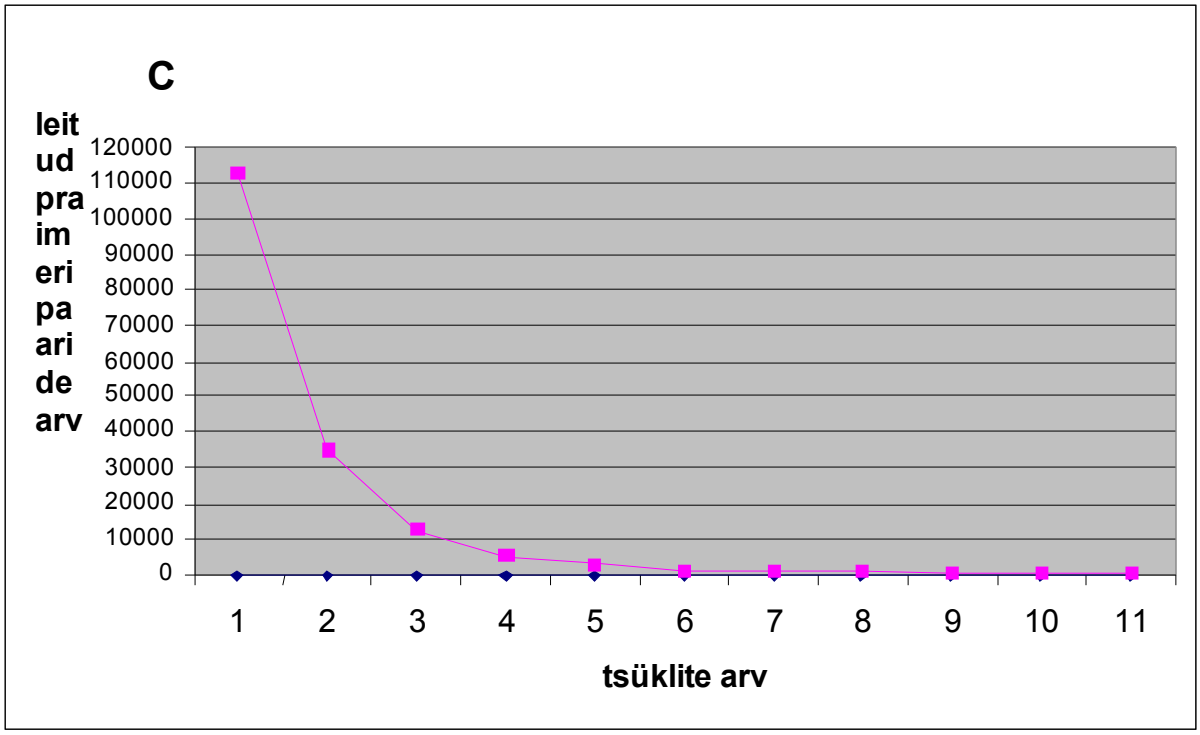
Joonis A. Inimese genoomi kromosoomide suurused.

SNP-d asuvad ühtlaselt üle kogu genoomi. Joonisel B on näha seni identifitseeritud SNP-de arv erinevatel kromosoomidel. Võrreldes jooniseid A ja B näeme, et SNP-de arvu ja kromosoomi suuruste vahel ei ole otsest korrelatsiooni.



Joonis B. Seni identifitseeritud SNP-d kromosoomiti.

Praimeri disaini programm koosnes seesmiselt 12 tsüklist. Kui esimese tsükli jooksul programm praimereid ei leidnud, liideti amplikoni mõlemale poole 100 bp-d ja prooviti uuesti leida unikaalset praimeripaari. Maksimaalseks tsüklite arvuks oli 12. Analüüsides erinevate tsüklite jooksul disainitud praimeripaaride arvu selgus, et esimese kuue tsükliga suudeti praimerid disainida keskmiselt 99%-ile amplikonidest/SNP-dest. Joonis C näitab juhuslikult valitud kromosoomi peal jooksutatud tsüklite tulemusi.



Joonis C. Erinevate tsüklite jooksul leitud praimeripaaride arv. Joonisel on näidatud juhuslikult valitud kromosoom, kuna kõikidel kromosoomidel oli tulemus analoogne.

Tabel 4. Tsüklite analüüsimine. Eraldi on välja toodud SNP-de hulk, mis leiti vastava tsükli jooksul ja SNP-de hulk, millele leiti praimeripaar vastava tsükli ja temale eelnevate tsüklite jooksul. Tabelis on juhuslikult valitud kromosoomid, kuna tsükleid analüüsides tuli välja, et tulemus on analoogne üle kogu kromosoomi komplekti. Pärast kuuendat tsüklit leidus keskmiselt 1% kogu vastava kromosoomi SNP-dest, millele praimeripaari ei olnud veel leitud. Pärast 12. tsüklit leidus selliseid SNP-sid keskmiselt 0,1%.

Kromosoom 1			
Tsüklid		%SNP-sid mis leiti vastava tsükli jooksul	**SNP-de % millele leiti praimeripaar vastava tsükli ja temale eelnevate tsüklite jooksul**
1	389742	68,97246177	68,97246
2	113005	19,99844267	88,97090444
3	34990	6,192164143	95,16306858
4	13009	2,302196723	97,4652653
5	5488	0,971208826	98,43647413
6	3228	0,571257669	99,0077318
7	1340	0,237139181	99,24487098
8	1077	0,190596193	99,43546717
9	914	0,161750158	99,59721733

10	827	0,146353808	99,74357114
11	755	0,133612001	99,87718314
12	694	0,12281686	99,9
SNP-d kokku	565069		
Kromosoom 16			
1	168826	74,02473834	74,02473834
2	39488	17,3142103	91,33894864
3	11109	4,870937049	96,20988569
4	4014	1,76000912	97,96989481
5	1697	0,744079591	98,7139744
6	1010	0,44285232	99,15682672
7	436	0,191171892	99,34799861
8	358	0,156971416	99,50497003
9	316	0,138555775	99,6435258
10	282	0,123647875	99,76717368
11	271	0,118824731	99,88599841
12	260	0,114001587	99,9
SNP-d kokku	228067		

Pärast praimerite disainimist leidis aset nende unikaalsuse kontrollimine. Juhul, kui programm *gtest.pl* leidis, et praimeripaar omab genoomis enam kui ühte seostumiskohta, suunati see ebakvaliteetseid praimeripaare sisaldavasse faili, millega tegeles edasi programm *primers100.pl*. Tabel 5 näitab, kui suur hulk praimeritest läks välja pärast nende unikaalsuse kontrolli genoomis ehk kui paljudele SNP-dele suudeti unikaalsed praimerid disainida esimese (*gtest.pl*-ga) ja teise korraga (*gtest100.pl*-ga).

Tabel 5. Kvaliteetset praimeripaari omavate SNP-de hulk peale esimest (*gtest.pl*) ja teist (*gtest100.pl*) unikaalsuse kontrolli. Esimese tsükliga leiti 93,7%-ile SNP-dest unikaalne praimeripaar. Teise tsükliga tuli juurde 2,6% SNP-d, millele suudeti praimeripaar leida.

SNP-de arv kokku	gt est.pl	gtest100.pl
5 584 078	5 232 616	5 375 805
100%	93.7%	96.3%

Tabel 6. Palju leidus SNP-sid iga kromosoomi kohta, millele ei suudetud praimereid disainida. Tabelist näeme, et iga kromosoomi kohta leidus keskmiselt 1% SNP-sid, mille amplifitseerimiseks vajalikke praimereid ei suudetud disainida.

Kromosoomi nimi	%SNP-sid, millele praimereid ei leitud
1	1,040388
2	1,036383
3	1,028816
4	1,021306
5	1,025153
6	1,031355
7	1,030621
8	1,025791
9	1,103356
10	1,031042
11	1,023007
12	1,030154
13	1,019471
14	1,03348
15	1,034395
16	1,042297
17	1,049122
18	1,029476
18	1,042943
20	1,091888
21	1,035739
22	1,04999
X	1,062196

Tabel 7. SNP-de lõplik hulk, millele suudeti praimerid disainida. Programmide töö tulemusena leiti unikaalne praimeripaar 96,15%-le SNP-dest.

kõik SNP-d kokku	5 584 078	100%
SNP-d,millele suudeti praimerid disainida	5 369 100	96.15%
SNP-d, millele ei suudetud praimereid disainida	214 987	3.85%

2. Andmebaasi loomine ja kasutajaliides

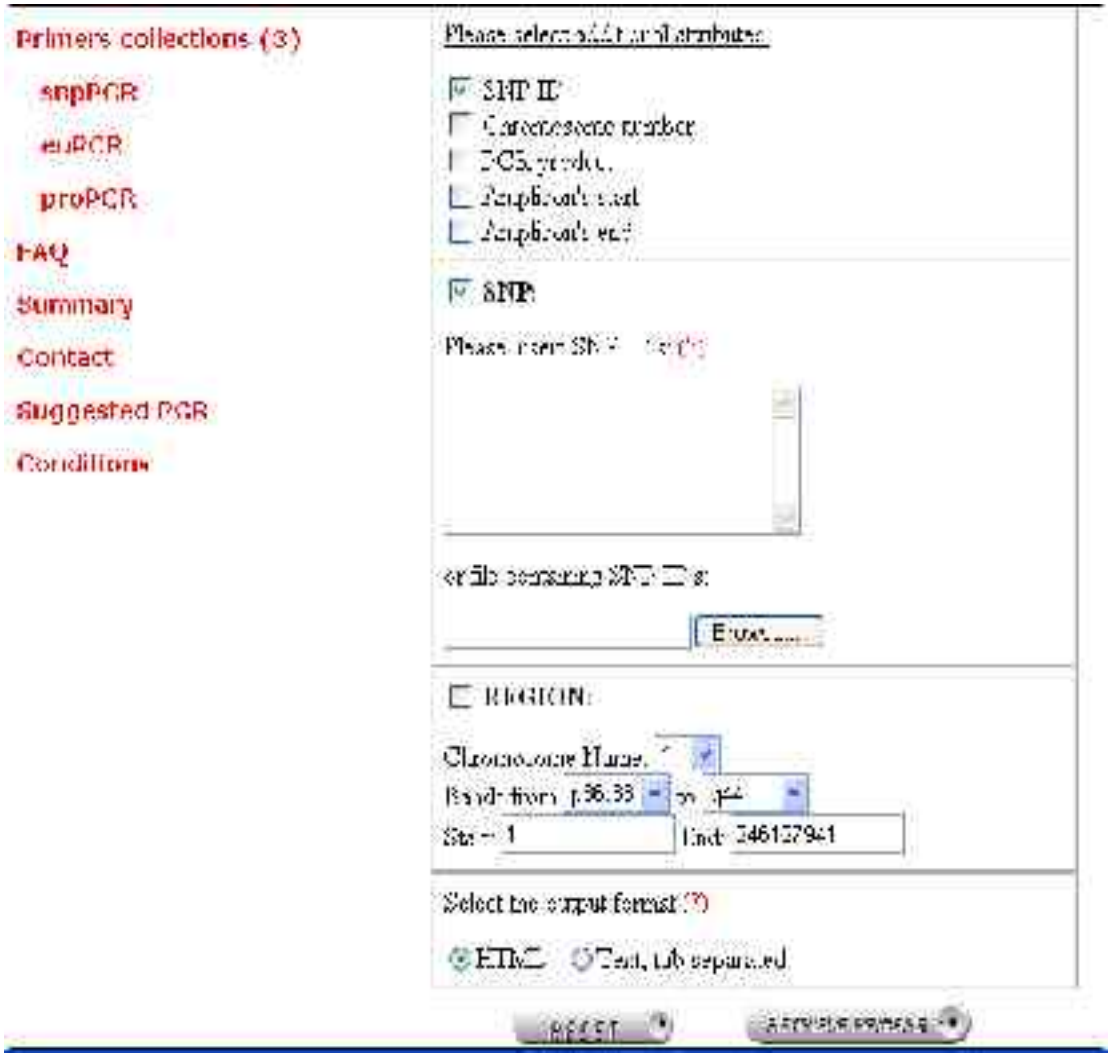
Kuna iga bioinformaatilise analüüsi ja töö kasulikkus sõltub sellest, kui hea on ligipääsetavus töötulemustele, siis kasutades MySQL-i võimalusi, lõime konstrueeritud praimerite ja vastavate produktide jaoks andmebaasi.

```
mysql> describe human_119;
+---------+------------------+------+-----+---------+-------+
| Field   | Type             | Null | Key | Default | Extra |
+---------+------------------+------+-----+---------+-------+
| id      | varchar(40)      |      | PRI |         |       |
| chr     | varchar(15)      |      | MUL |         |       |
| aid     | int(10) unsigned | YES  |     | NULL    |       |
| a_start | int(10) unsigned | YES  |     | NULL    |       |
| a_end   | int(10) unsigned | YES  |     | NULL    |       |
| primer1 | varchar(40)      |      |     |         |       |
| primer2 | varchar(40)      |      |     |         |       |
| product | text             |      |     |         |       |
+---------+------------------+------+-----+---------+-------+
8 rows in set (0.01 sec)
```

Joonis 4. Andmebaasis *snpPCR* sisaluva praimereid mahutava tabeli *human_119* struktuur.

Disainitud praimerite jaoks lõime andmebaasi *snpPCR*, mis sisaldab endas genereeritud veebiliidese töö jaoks vajalikke tabeleid ning praimereid sisaldavat tabelit. Tabeli nimi on *human_119*, mis väljendab organismi ja kasutatud *dbsnp* versiooni. Tabeli struktuur on toodud joonisel 4. Tabeli esimese veeru andmed on saadud NCBI-st. Väli *chr* kujutab endast kromosoomi nime, *aid* on amplikoni id, *a_start* ja *a_end* on amplikoni algus- ja lõpu koordinaadid, *primer1* ja *primer2* meie poolt leitud praimerite ning *product* on vastava praimeripaariga amplifitseerimistulemusena saadava produkti järjestus.

Kasutajaliidese loomisel lähtusime eelkõige sellest, et on vajalik kasutaja poolt sisestatud andmete dünaamiline ja automaatne liikumine arvutiprogrammile ja programmi töö tulemusena genereeritud andmete visualiseerimine kasutajale. Programmide kodeerimisel oleme kasutanud programmeerimiskeelt Perl. Dünaamilise veebilehe ehitamise vahendiks valisime CGI (*Common Gateway Interface*) tänu tema heale ühilduvusele Perl-iga. Samuti on CGI võimas vahend andmebaaside haldamiskeele MySQL-iga suhtlemiseks ning HTML formaadis teksti genereerimiseks.



Joonis 5. Graafiline kasutajaliides (http://kobra.ebc.ee/exonPCR/primerPCR/). Kõigepealt saab kasutaja valida genoomi, seejärel millise annotatsioonikeskuse andmeid ta tahab kasutada ning viimaks saab ta piirata sisendit vastavalt sellele, kas tal on olemas SNP ID(-d), fail, mis seda/neid sisaldab, või soovib ta väljundina saada kõigi teatud piirkonnas sisalduvate SNP-de amplifitseerimiseks vajalikke praimereid. Väljundiks saab olla teksti- või HTML formaat.

Arutelu

Antud töö tulemus on osaks veebiportaalist nimega *PrimerParadise*, kus lisaks SNP-dele saab praimereid disainida veel eksonitele ja hübridisatsioonioligotele. Portaal on terviklik lahendus kasutajatele, kes tahavad PCR-i praimereid leida efektiivselt ja mugavalt.

Üle inimese genoomi suutsime leida peaaegu kõikide seni identifitseeritud SNP-de amplifitseerimiseks vajalikud praimerid. Nimelt, programmide töö tulemusena leiti unikaalne praimeripaar 96,15%-le SNP-dest. Kui kokku on andmebaasis 5 584 078, siis ilma praimeriteta jäid 214 987 SNP-d. Programmide töökindlust näitab seegi, et kromosoomiti eraldi leidus ühtlaselt kesmiselt 1% SNP-sid, millele ei suudetud üldse või unikaalset praimeripaari disainida.

Praimeri disaini läbi viiv programm koosnes 12-st tsüklist. Kui tsükliga praimereid ei leitud, pikendati külgnevat ala ja nii kuni 12 korda. 12 on maksimaalse tulemi saamiseks optimaalne tsüklite arv, sest tsükleid analüüsides selgus, et peale 12. ringi enam praimereid juurde ei leita. Kui kasutaja tahab tulemusi saada väga kiirelt, siis võib ta programmi koodi siseselt tsüklite arvu vähendada vastavalt tabelis 4 toodud tulemustele. Kõikide kromosoomide puhul oli tulemus analoogne, st esimese kuue tsükliga leiti praimeripaar keskmiselt 99%-ile SNP-dest. 12. tsükliga suurenes see arv 99,9%-ile.

Edasi toimus leitud praimerite unikaalsuse kontrollimine. Esimese e-PCR-ga jäi järele 93,7% SNP-sid, mille praimerid tunnistati kvaliteetseteks. Väljaläinud amplikonidele disainiti uued praimerid, mis läksid samuti e-PCR-i. Sel moel tuli juurde 2,6% SNP-d, millele suudeti unikaalne praimeripaar disainida. Siit ka lõpptulemus: 96,15%.

Et tulemused oleksid veelgi paremad, peame me tehtud tööga viima läbi täiendavaid analüüse, manipuleerides erinevate võimalustega. Üheks võimaluseks on kohe praimerite disainimise algetapis muuta GM_PRIMER3 sisendfailis PCR-i parameetrite lubatud varieeruvuse vahemikke. Näiteks vähendada GC-sisaldust, mis on praegu 20-80% ning piirata produktide pikkusi. On läbi viidud mitmeid uuringuid, et lühemad amplikonid annavad paremaid tulemusi (näiteks on PCR-i produktid unikaalsemad) (Thareau *et al*., 2003). Praktikas tõestamist leidnud, PCR-i kvaliteediga mitte otseselt seotud olevaid väärtuseid ei ole mõtet muuta. Näiteks, ei ole praimeri pikkusel otseses korrelatsioonis PCR-i väljatulemise kvaliteediga (Haas *et al*., 1998),

küll aga ei ole tark vähendada minimaalset praimeri pikkust, kuna väikesed praimerid on suurema tõenäosusega mitte-spetsiifilised (Andreson, 2002). Samuti on uuringutest leidnud tõestamist see, et AT-rikka piirkonna puhul oleks otstarbekas suurendada maksimaalset praimeri suurust ja vähendada minimaalset sulamistemperatuuri (Rychlik, 1993).

Analüüsides SNP-de, millele praimereid ei leitud, asukohti genoomis selgus, et sellised SNP-d asusid amplikonides, mis sisaldasid palju korduseid või ei olnud nende ümber piisavalt külgnevat ala (sekveneerimata või väga katkendlikult sekveneeritud), mille peale saaks praimerid disainida. Sellised amplikonid tuleks võtta uuesti vaatluse alla ja lõdvendada erinevad parameetreid nii kaua, kuni siiski praimerid leitakse. Siis on jälle küsimus selles, et kui hea kvaliteediga ja kui spetsiifilised leitud praimerid on. Tulevikus võiks välja töötada programmi täienduse, kus disainitud praimerid omistavad endale skoori, mis näitab nende kvaliteeti. Viimane on eraldiseisev ja keeruline küsimus, kuna välja tuleb töötada algoritm, mis, kuidas ja mille alusel vastavat skoori arvutatakse.

Programmide töö tulemust võib lugeda heaks ja seega oleks esimene asi mitte raisata aega töö meetodite muutmise, vaid praimeripaaride testimise peale. Tulemuste usaldusväärsuse suurendamiseks on kavas viia läbi eksperimentaalsed katsed.

Kahtlemata on väga oluline andmebaasi täiendamine. Kui praegu on andmebaasis ainult inimese genoomi SNP-de praimerid, siis varsti on kasutaja valikus rohkem organisme. Samuti on praegu saadaval ainult NCBI-st saadud lähteandmed. Lisada võiks veel mõned annotatsioonikeskused. Graafilise kasutajaliidese peaks kasutajale tegema veelgi mugavamaks ning paremini hoomatavaks (kujundus, navigeerimine). Väljundfailis võiks kasutaja saada soovi korral rohkem lisainformatsiooni. Näiteks võiks vastava programmi töö tulemusena väljundis olla erinevad graafikud, joonised ning tabelid.

Kokkuvõte

Geneetiliste variatsioonide ja bioloogilise funktsiooni vaheliste seoste uurimine on tänapäeval muutunud üheks juhtivamaks suunaks bioloogias, evolutsioonis ja patofüsioloogias. Et mõista täielikult inimese genoomi geneetiliste variatsioonide ja fenotüübiliste muutustevahelisi seoseid, vajame me SNP-de analüüsimiseks efektiivseid, odavaid ning sensitiivseid genotüpiseerimistehnikaid. Enamik genotüpiseerimistehnikaid sisaldab endas esmalt PCR-i abil uuritava, SNP-d sisaldava järjestuse üles amplifitseerimist. Seega, on PCR-i jaoks vajaliku unikaalse praimeripaari valik väga oluline, sest genotüpiseerimistulemused on otseselt sõltuvad PCR-i amplifikatsiooni produktist. Meie töö eesmärgiks oli luua terviklik lahendus kõikide inimese SNP-de üles amplifitseerimiseks vajalike PCR-i praimerite automaatseks disainiks.

Praimeri disain algas SNP-d sisaldavate amplikonide moodustamisega. Seejärel maskeerisime selles sisalduvad kordusjärjestused, et vältida praimerite mittespetsiifilisi seondumisi. Kui programm praimereid ei leidnud, siis pikendasime amplikoni selle mõlemalt poolt maksimaalselt 12 korda. Eukarüootseid praimereid testisime e-PCR-ga, et oleksime kindlad, et PCR-i tulemusena ei genereerita alternatiiveid produkte ja et praimerid ei omaks genoomis sekundaarseid seostumisi. Amplikonid, mille praimerid praagiti e-PCR-ga välja, läksid uuele ringile, mis erines esimesest ringist selle poolest, et amplikonidele disainiti kuni 100 praimeripaari, mille hulgast püüti leida temale spetsiifiline.

Käesolev töö moodustab osa kompaktsest praimerite veebiportaalist *PrimerParadise*, kus lisaks SNP-dele saab praimereid leida ka eksonitele ja hübridisatsioonioligotele. Andmebaasis on 96,15% (5 369 100) seni identifitseeritud, NCBI dbSNP andmebaasist saadud SNP-d, millele PCR-i praimeripaar leiti. Portaali praimereid saab kasutada kogu genoomi hõlmavates projektides (resekveneerimine, mutatsiooni analüüs, SNP-de detekteerimine jne).

Summary

Understanding the relationship between genetic variation and biological function on a genomic scale is expected to provide fundamental new insights into the biology, evolution and pathophysiology. The hope that single nucleotide polymorfisms (SNPs) will allow genes that underlie complex disease to be identified, together with progress in identifying large sets of SNPs, are the driving forces behind intense efforts to establish the effective, cheap and sensitive technology for large-scale analysis of SNPs. Most genotyping methods include the PCR amplification of target sequence. The selection of PCR primers is very important because the quality of genotyping results directly depends on the quality of PCR amplification products. The purpose of this work was to create the solution of automated design of PCR primers to amplify all SNPs in the human genome.

Primer design started with generating SNPs containing amplicons. To prevent unspecific primer binding we masked all repeated regions in amplicon. If the program could not find primers we added sequence to both sides of amplicon maximally twelve times. The primers for eukaryotic genomes have been tested with e-PCR to make sure that no alternative products will be generated and that excessively binding eukaryotic primers are excluded. With the cycle described before we tried to designed maximally 100 new primerpairs to the amplicons with primers e-PCR said unqualified. Next e-PCR was used again to reject primerpairs with secondary binding sites in genome.

Present work is part of the compact web portal named *PrimerParadise* (http://kobra.ebc.ee/exonPCR/primerPCR) where you can retrieve primers for SNPs, exons and hybridization oligos. Our database contains of 96,15% (5 369 100) of identified SNPs uploaded from the NCBI dbSNP database to whom PCR primerpair with quality was automatly designed. Primers can be used for genome-wide projects (resequencing, mutation analysis, SNP detection etc).

Kasutatud kirjandus

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ, 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997 Sep 1;25(17):3389-402.

Andreson R, 2002 Erinevate *in silico* meetodite võrdlus PCR praimerite kvaliteedi parandamiseks. Magistritöö.

Baskaran N, Kandpal RP, Bhargava AK, Glynn MW, Bale A, Weissman SM, 1996. Uniform amplification of a mixture of deoxyribonucleic acids with varying GC content. Genome Res. 1996 Jul;6(7):633-8.

Benita Y, Oosting RS, Lok MC, Wise MJ, Humphery-Smith I, 2003. Regionalized GC content of template DNA as a predictor of PCR success. Nucleic Acids Res. 2003 Aug 15;31(16):e99.

Borer PN, Dengler B, Tinoco I.Jr., 1974. Stability of Ribonucleic acid Double-stranded helices. J.Mol. Biol. 86:843.

Breslauer JK, Dengler R, Blocker H, Marky L, 1986. Predicting DNA duplex stability from the base sequence. Proc.Natl.Acad.Sci. 83:3746-3750.

Brookes Anthony J, 1999. The essence of SNPs. Elsevier Science B.V. Gene 234:177-186

Cardon R, Palmer J, 2003. Population stratification and spurious allelic association. Lancet 361:598-604.

Collins FS, Brooks LD, Chakravarti A, 1998. A DNA polymorphism discovery resource for research on human genetic variation. Genome Res. 8:1229–1231.

Collins FS, Guyer MS, Charkravarti A, 1997 Variations on a theme: cataloging human DNA sequence variation. Science 278:1580-1.

Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES., 2001. High resolution haplotype structure in the human genome. Nature Genet.,29:229-332.

Goldstein DB, 2001. Island of linkage disequilibrium. Nature Genet.,29:109-111.

Freeman T, 2000. High troughput gene expression screening:its emerging role in grug discovery. Med.Res.Rev., 20:197-202.

Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D., 2002. The structure of haplotype blocks in the human genome. Science. 2002 Jun 21;296(5576):2225-9.

Goldstein DB, Tate SK, Sisodiya SM, 2003. Pharmacogenetics goes genomic. Nat Rev Genet. 2003 Dec;4(12):937-47.

Guillaudeux T, Janer M, Wong GK, Spies T, Geraghty DE, 1998. The complete genomic sequence of 424 015 bp at the centromeric end of the HLA class I region: gene content and polymor- are therefore essential, and certainly an increased empha- phism. Proc. Natl. Acad. Sci. USA 95:9494–9499.

Haas S, Vingron M, Poustka A, Wiemann S, 1998. Primer design for large scale sequencing. Nucleic Acid Res 26(12):3006-3012.

Haas SA, Hild M, Wright AP, Hain T, Talibi D, Vingron M, 2003. Genome-scale design of PCR primers and long oligomers for DNA microarrays. Nucleic Acids research. 19:5576-5581.

Hacia JG, Fan JB, Ryder O, Jin L, Edgemon K, Ghandour G, Mayer RA, Sun B, Hsie L, Robbins CM, Brody LC, Wang D, Lander ES, Lipshutz R, Fodor SP, Collins FS, 1999. Determination of ancestral alleles for human single-nucleotide polymorphisms using high-density oligonucleotide arrays. Nature Genet. 22**:**164–167.

Heid CA, Stevens J, Livak KJ, Williams PM, 1996. Real time quantitative PCR. Genome Res. 1996 Oct;6(10):986-94.

Henegariu O, Heerema NA, Dlouhy SR, Vance GH, Vogt PH, 1997. Multiplex PCR: Critical Parameters and Step-by-Step Protocol. BioTechniques 23:504-511.

Holliday R, Grigg GW, 1993. DNA methylation and mutation. Mutat. Res. 285:61–67.

Hijum SA, de Jong A, Buist G, Kok J, Kuipers OP, 2003. Unifrag and GenomePrimer: selection of primers for genome-wide production of unique amplicons. Bioinformatics, 19:1580-1582.

Horton R, Niblett D, Milne S, Palmer S, Tubby B, Trowsdale J, Beck S., 1998. Large-scale sequence comparisons reveal unusually high levels of variation in the HLA-DQB1 locus in the class II region of the human MHC. J. Mol. Biol. 282:71–97.

Jorde LB, Watkins WS, Bamshad MJ, Dixon ME, Ricker CE, Seielstad MT, Batzer MA, 2000. The distribution of human genetic diversity: a comparison of mitochondrial, autosomal, and Y-chromosome data. Am. J. Hum. Genet. 66:979-988.

Kaderali L, Deshpande A, Nolan JP, White PS, 2003. Primer-design for multiplexed genotyping. Nuc.Acids Res. 6:1796-1802.

Kirk BW, Feinsod M, Favis R, Kliman RM, Barany F, 2002. Single nucleotide polymorfism seeking long term association with complex disease. Nucleic Acids Res. 2002 Aug 1;30(15):3295-311.

Kruglyak L, Nickerson DA, 2001. Variation is the spice of life. Nature 27:234-235.

Kämpke T, Kieninger M, Mecklenburg M, 2001. Efficient primer design algorithms. Bioinformatics 17(3):214-25.

Landegren U, Nilsson M, Kwok PY, 1998. Reading bits of genetic information: methods for single-nucleotide polymorphism analysis.Genome Res. 8:769–776.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E., 2001 Initial sequencing and analysis of the human genome. Nature 409:860-921.

Li P, Kupfer KC, Davies CJ, Burbee D, Evans GA, Garner HR, 1997. PRIMO: A Primer Design Program That Applies Base Quality Statistics for Automated Large-Scale DNA Sequencing. Genomics 40(3):476-85.

Li W, Sadler LA, 1991. Low nucleotide diversity in man. Genetics 129:513–523.

Marras SA, Kramer FR, Tyagi S, 1999. Multiplex detection of single-nucleotide variations using molecular beacons. Genet Anal. 1999 Feb;14(5-6):151-6.

Matsuzaki H, Loi H, Dong S, Tsai YY, Fang J, Law J, Di X, Liu WM, Yang G, Liu G, Huang J, Kennedy GC, Ryder TB, Marcus GA, Walsh PS, Shriver MD, Puck JM, Jones KW, Mei R, 2004. Parallel Genotyping of Over 10,000 SNPs Using a One-Primer Assay on a High-Density Oligonucleotide Array. Genome Res. 2004 Mar;14(3): 414-25.

Nachman MW, Bauer VL, Crowell SL, Aquadro CF, 1998. DNA variability and recombination rates at X-linked loci in humans. Genetics 150:1133–1141.

Nickerson DA, Taylor SL, Weiss KM, Clark AG, Hutchinson RG, Stengard J, Salomaa V, Vartiainen E, Boerwinkle E, Sing CF, 1998. DNA sequence diversity in a 9.7 kb region of the human lipoprotein lipase gene. Nature Genet. 19:233–240.

Ning Z, Cox AJ, Mullikin JC, 2001. SSAHA: a fast search method for large DNA databases. Genome Res. 2001 Oct;11(10):1725-9.

Proutski V, Holmes EC, 1996. Primer Master: a new program for te design and analyses of PCR primers. Comput. Appl. Biosci., 12:253-255.

Raddatz G, Dehio M, Meyer TF, Dehio C, 2001. PrimeArray: genome-scale primer design for DNA microarray construction. Bioinformatics. Jan;17(1):98-9.

Reich E, Gabriel S, Altschuler D, 2003. Quality and completeness of SNP databases. Nature Genet. Online publication.

Roses Allen D, 2000. Pharmacogenetics and the practice of medicine. Nature Vol 405:857-865.

Rozen S, Skaletsky H, 2000. Primer3 on the WWW for general users and for biologist programmers. Methods Mol Biol 132:365-86.

Rothberg BE, 2001. The use of animal models in expression pharmacogenomic analyses. Pharmacogenomics J. 2001;1(1):48-58.

Rychlik W, Spencer WJ, Rhoads RE, 1990. Optimization of the annealing temperature for DNA amplification in vitro. Nucleic Acid Res 18(21):6409-6412.

Rychlik W, 1993. Selection of primers for polymerase chain reaction. In White, B.A. Humana Press, Totowa, NJ, Vol 15: pp.31-40.

Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, Sherry S, Mullikin JC, Mortimore BJ, Willey DL, Hunt SE, Cole CG, Coggill PC, Rice CM, Ning Z, Rogers J, Bentley DR, Kwok PY, Mardis ER, Yeh RT, Schultz B, Cook L, Davenport R, Dante M, Fulton L, Hillier L, Waterston RH, McPherson JD, Gilman B, Schaffner S, Van Etten WJ, Reich D, Higgins J, Daly MJ, Blumenstiel B, Baldwin J, Stange-Thomann N, Zody MC, Linton L, Lander ES, Altshuler D; International SNP Map Working Group, 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorfisms. Nature. 2001 Feb 15;409(6822):928-33.

Schafer AJ, Hawkins JR, 1998. DNA variation and the future of human genetics. Nat Biotechnol. 1998 Jan;16(1):33-9.

Syvanen AC, Aalto-Setala K, Harju L, Kontula K, Soderlund H, 1990. A primer-guided nucleotide incorporation assay in the genotyping of apolipoprotein E. Genomics. 1990 Dec;8(4):684-92.

Taillon-Miller P, Gu Z, Li Q, Hillier L, Kwok PY, 1998. Overlapping genomic sequences: A treasure trove of single-nucleotide polymorphisms. Genome Res. 8:748–754.

Tatusova TA, Madden TL , 1999**.** BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. FEMS Microbio Lett 174(2):247-50.

Thareau V, Dehais P, Serizet C, Hilson P, Rouze P, Aubourg S, 2003. Automatic design of gene-specific sequence tags for genome-wide functional studies. Bioinformatics. 2003 Nov 22;19(17):2191-8.

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

The International SNP Map Working Group, 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorfisms. Nature, 409:928-933.

Tsui C, Coleman LE, Griffith JL, Bennett EA, Goodson SG, Scott JD, Pittard WS, Devine SE., 2003. Single nucleotide poymorfisms (SNPs) that map to gaps in the human SNP map. Nucleic Acids Research, 16:4910-4916.

Tyagi S, Bratu DP,Kramer FR, 1998. Multicolor molecular beacons for allele discrimination. Nat. Biotechnol., 16:49-53.

Varotto C, Richly E, Salamini F, Leister D, 2001**.** GST-PRIME: a genome wide primer design software for the generation of gene sequence tags. Nucleic Acid Res 29(21):4373-4377.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, Di Francesco V, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue B, Sun J, Wang Z, Wang A, Wang X, Wang J, Wei M, Wides R, Xiao C, Yan C, Yao A, Ye J, Zhan M, Zhang W, Zhang H, Zhao Q, Zheng L, Zhong F, Zhong W, Zhu S, Zhao S, Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An H, Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, Busam D, Carver A, Center A, Cheng ML, Curry L, Danaher S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Ferriera S, Garg N, Gluecksmann A, Hart B, Haynes J, Haynes C, Heiner C, Hladun S, Hostin D, Houck J, Howland T, Ibegwam C, Johnson J, Kalush F, Kline L, Koduru S, Love A, Mann F, May D, McCawley S, McIntosh T, McMullen I, Moy M, Moy L, Murphy B, Nelson K, Pfannkoch C, Pratts E, Puri V, Qureshi H, Reardon M, Rodriguez R, Rogers YH, Romblad D, Ruhfel B, Scott R, Sitter C, Smallwood M, Stewart E, Strong R, Suh E, Thomas R, Tint NN, Tse S, Vech C, Wang G, 2001. The sequence of the human genome. Science 291:1304-1351.

Wang DG, Fan JB, Siao CJ, Berno A, Young P, *et al.*, 1998. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. Science 280:1077–1082.

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