TARTU ÜLIKOOL
BIOLOOGIA-GEOGRAAFIA TEADUSKOND
MOLEKULAAR- JA RAKUBIOLOOGIA INSTITUUT
BIOINFORMAATIKA ÕPPETOOL

Mikk Eelmets

GEENIDE JA EKSONITE PIIRIDE JA HAPLOTÜÜBI BLOKI PIIRIDE KORRELATSIOON

Bakalaurusetöö

Juhendaja prof. Maido Remm, PhD

Tartu 2006

SISUKORD

KASUTATUD LÜHENDID

cDNA - komplementaarne *(complementary)* DNA

EST - ekspresseeruva järjestuse tunnus *(expressed sequence tag)*

HMM - varjatud Markovi mudel *(Hidden Markov Model)*

htSNP - haplotüüpi määrav SNP *(hapotype tagSNP)*

kb - tuhat aluspaari *(kilobase)*

LD - aheldatuse mittetasakaalustatus *(linkage disequilibrium)*

Mb - miljon aluspaari

miRNA - mikro-RNA *(microRNA)*

mRNA - informatsiooni (*messenger-)* RNA

nt - nukleotiid

PCR - polümeraasi ahelreaktsioon *(polymerase chain reaction)*

pre-mRNA - eel-mRNA

rRNA - ribosomaalne RNA *(ribosomal RNA)*

snoRNA - väikese tuumakese RNA *(small nucleolar RNA)*

SNP - üksiknukleotiidpolümorfism *(single-nucleotide polymophsim)*

snRNA - väike tuuma RNA *(small nuclear RNA)*

tagSNP - esindav SNP

SISSEJUHATUS

Vaatamata genotüpiseerimise tehnoloogia ja analüüsimeetodite arengule pole endiselt komplekstunnuste geenide määramisel suurt edu saavutatud. Kitsaskohtadeks on genotüpiseerimise kõrge hind ja polümorfismide rohkus inimese genoomis. Edu tõotavad tuua meetodid, mis võtavad arvesse informatsiooni haplotüübi struktuuri kohta erinevates genoomi piirkondades. Selle tarvis on käivitatud laiaulatuslik HapMap projekt, mille eesmärgiks on inimpopulatsiooni haplotüübi struktuuri kindlaks tegemine. Uuringud on näidanud, et kromosoomidel olevad polümorfismid võivad olla korreleerunud, moodustades diskreetseid bloki laadseid struktuure, mida kirjanduses on hakatud kutsuma haplotüübi blokkideks. Käesolevas töös uuritakse haplotüübi blokkide ja geenide vahelisi seoseid. Kirjandusülevaates tuuakse ülevaade aheldatuse mittetasakaalustatuse hindamisest, haplotüübi blokkide arvutusmeetoditest ja geenide annotatsiooni probleemidest. Praktilises osas uuritakse geeni ja eksonite ning haplotüübi blokkide vahelisi seosed. Töö valmis Tartu Ülikooli Molekulaar– ja rakubioloogia instituudis Bioinformaatika õppetoolis.

I KIRJANDUSE ÜLEVAADE

1. HAPLOTÜÜBI BLOKID

1.1 Haplotüübid

Haplotüüp on lühend terminist haploidne genotüüp ning tähistab individuaalset kromosoomi. Haplotüüp võib viidata ühele lookusele või ka tervele genoomile. Ülegenoomne haplotüüp on pool diploidsest genoomist, omades igas lookuses ainult ühte alleeli. Kitsamas tähenduses on haplotüüp ühel kromatiidil asetsevate polümorfismide kogum.

Diploidse organismi genotüpiseerimise tulemusel ei saa otseselt määrata haplotüüpe. Haplotüübi faasi määramiseks on mitu võimalust. Üheks võimaluseks on kasutada laboratoorseid tehnikaid nagu alleelspetsiifiline *long-range* PCR või diploid – haploid konversioon, mille tulemusena eraldatakse kromosoomid tehniliselt (Zhang *et al.* 2004). Laboratoorsed meetodid ei sobi laiaulatuslikeks genoomi uuringuteks, kuna on tehniliselt nõudlikud ning kallid. Teiseks võimaluseks on kasutada algoritme, mis ennustavad haplotüüpe genotüpiseerimise andmetelt. Haplotüüpide määramise algoritmid saab jagada kombinatoorikal või statistikal põhinevateks. Kombinatoorikal põhinevad meetodid määravad genotüübid haplotüüpidesse, mille põhjal hinnatakse haplotüüpide sagedused. Statistikal põhinevad meetodid hindavad esimesena haplotüüpide sagedused ning seejärel määratakse iga indiviidi genotüübile tema haplotüübid (Zhang *et al.* 2004).

1.2 Aheldatuse mittetasakaalustatus

Aheldatuse mittetasakaalustatus on populatsiooni tasemel erinevates positsioonides olevate alleelide seotus, mis on tekkinud mitmete geneetiliste faktorite ja populatsioonis toimunud demograafiliste sündmuste läbi (Li *et al.* 2003). Aheldatuse mittetasakaalustatuse asemel võib kasutada terminit gameetiline mittetasakaalustatus,

mis võtab vaatluse alla ühel kromatiidil olevad lookused. Samuti võib kasutada sünonüümselt termineid haplotüüp ja gameet (Hedrick 1987). Tähtsaimaks aheldatuse mittetasakaalustatust mõjutavaks geneetiliseks faktoriks on rekombinatsioon, mis vähendab alleelide vahelist seotust. Kahe lähestikku asetseva lookuse vahel toimuva rekombinatsiooni tõenäosus ühes meioosis on äärmiselt väike, kuid kui vaadelda kõiki populatsioonis toimunud rekombinatsioone läbi mitme generatsiooni, siis omab olulist mõju LD mustri kujunemisel (Li *et al.* 2003).

Gameetilise mittetasakaalustatuse ulatust saab hinnata mitmete suuruste abil. Lihtsaim LD mõõt väljendub järgnevalt:

$$D_{ij} = x_{ij} - p_i q_j \quad (1)$$

kus x_{ij} on gameedi $A_i A_j$ sagedus populatsioonis, p_i ja q_j on alleelide A_i ja Bj sagedused lookustes *A* ja *B* ning $p_i q_j$ on gameedi $A_i B_j$ teoreetiline sagedus, kui lookused *A* ja *B* pole statistiliselt assotsieerunud (Hedrick 1987).

D' on normaliseeritud *D* :

$$D'_{ij} = \frac{D_{ij}}{D_{\max}} \quad (2)$$

kus $D_{\max}$ avaldub järgnevalt:

$$D_{\max} = \min\left\lfloor p_i q_j \, , (1-p_i)(1-q_j) \right\rfloor \text{kui } D_{ij} < 0 \text{ või}$$
$$D_{\max} = \min\left\lfloor p_i (1-q_j) \, , (1-p_i) q_j \right\rfloor \text{kui } D_{ij} > 0$$

Gameetilist mittetasakaalustatust kahe lookuse kõikide alleelide vahel saab hinnata mitmel viisil. Üks võimalus väljendub valemiga:

$$D^2 = \sum_{i=1}^{k} \sum_{j=1}^{l} D_{ij}^2 \quad (3)$$

kus *k* ja *l* on alleelide arv vastavalt lookustes *A* ja *B*. Kuna D_{ij} väärtused sõltuvad oluliselt alleelisagedustest, siis D^2 samuti tugevalt seotud alleelsagedustega.

Vähendamaks seotust alleelsagedustega on loodud mitmeid normaliseeritud mittetasakaalustuse mõõte (Hedrick 1987).

Üks selline mõõt on standardiseeritud üksiklookuse heterosügootususega. Hardy-Weinbergi homosügootsus lookuses A, kus on k alleeli on

$$F_A = \sum_{i=1}^{k} p_i^2 \quad (4a)$$

ning lookuses B, kus on l alleeli on

$$F_B = \sum_{j=1}^{l} q_j^2 \quad (4b).$$

Hardy-Weinbergi heterosügootsus lookuses A on $H_A = 1 - F_A$ ja lookuses B $H_B = 1 - F_B$. Heterosügootsuse abil standardiseeritud kahelookusese vahelise assotsiatsiooni mõõt on $D*$, mida tähistatakse ka R. Heterosügootusega standardiseeritud aheldatuse mittetasakaalustuse mõõt väljendub:

$$D* = \frac{D^2}{H_A H_B} \quad (5a).$$

Juhul kui mõlemas lookuses on kaks alleeli, siis valem (5a) avaldub kujul

$$D* = \frac{D^2}{4p_1(1-p_1)q_1(1-q_1)} \quad (5b)$$

ja $D^2 = 4D_{ij}^2$. Sellisel juhul on $D*$ võrdne korrelatsiooni koefitsiendi ruuduga (Hedrick 1987). Kahe lookuse vaheline korrelatsiooni koefitsient avaldub valemiga:

$$\varphi = \frac{D}{\sqrt{p_1 p_2 (1-p_1)(1-p_2)}} \quad (6),$$

kus p_1 ja p_2 alleelide sagedused kahes lookuses ning omab väärtuseid vahemikus nullist üheni (Franklin *et al.* 1970). Korrelatsiooni koefitsiendi ruutu tähistatakse kirjanduses ka r^2-ga.

Praktikas kaks enim kasutatavat LD mõõtu on D' ja r^2. Mõlema väärtuste skaala on nullist üheni, kus null viitab aheldatuse puudumisele ning üks aheldatusele. Mõõtude väärtuste interpreteerimine on pisut erinev. D' väärtus üks tähendab, et väikesema sagedusega alleel esineb ühes haplotüübis erandidult ainult ühe alleeliga teisest lookusest, kui r^2 omab väärtus üks kui mõlemas lookuses on alleelide sagedused

identsed ja esimese lookuse alleel esineb ainult ühe alleeliga teisest lookusest (Zondervan *et al.* 2004).

1.3 Haplotüübi blokid

Genoomis olev LD ei ole homogeenne. Kõrge LD-ga piirkonnad vahelduvad madala LD-ga piirkondadega. Regioonides, milles on kõrge LD, on madal haplotüüpide mitmekesisus. Nendes piirkondades on mõned erinevad haplotüübid, mis esinevad enamikes populatsioonis olevates kromosoomides (Cardon *et al.* 2003).

Diskreetseid kromosoomi piirkondi, mis on tugevas aheldatuse mittetasakalus ning milles on madal haplotüübi mitmekesisus, nimetatakse haplotüübi blokkideks. Eeldatakse, et kõik blokki kuuluvad polümorfismide paarid on tugevas aheldatuse mittetasakaalus ning polümorfismid, mis ei kuulu blokki, ei ole nendega mittetasakaalus. Hüpoteesi kohaselt on vastavates kromosoomi regioonides rekombinatsiooni tõenäosus madal ja regioone ümbritsevad piirkonnad, kus on rekombinatsiooni tõenäosus kõrge. Mitmed uuringud on leidnud selliseid piirkondi kromosoomides, kus esinevad haplotüübi blokid (Cardon *et al.* 2003).

Kuna haplotüübi blokid on tihedalt seotud aheldatuse mittetasakaalustatusega, siis on neid mõjutavad faktorid samad. LD taseme tõus viib haplotüübi blokkide tekkimisele ning langus blokkide lagunemisele. Haplotüübi blokid võivad tekkida erinevate mehhanismide abil. Üheks haplotüübi blokkide tekke põhjuseks on heterogeenne rekombinatsioon, mille tulemusena tekivad genoomis kõrgema LD-ga alad. Samuti on populatsiooni ajalugu oluline LD ning haplotüübi blokkide mõjutaja. Populatsioonis toimuvatest protsessidest tõstavad LD taset looduslik valik ja „pudelikaela" mehhanism, mille käigus tekivad ka haplotüübi blokid. Erinevate alleelsagedustega populatsioonide segunemisel võib tõusta LD tase populatsioonis. Viimane teeb esivanematelt päritud konserveerunud segmentide jälgimise eriti raskeks. Lisaks geneetilistele faktoritele mõjutab ka eksperimendi planeerimine leitud halotüübi blokkide asukohti ja struktuure. Olulisteks mõjutajateks on uuringuks valitud markerite tihedus erinevates genoomi piirkondades ning nende alleelsagedused populatsioonis (Phillips *et al.* 2003).

1.3.1 Haplotüübi blokkide arvutamise meetodid

Haplotüübi blokkide kirjeldamiseks on arendatud mitmeid meetodeid. Eristada võib kahte lähenemisviisi. Esimene käsitleb haplotüübi blokke kui lihtsalt vähese haplotüübi varieeruvusega piirkondi seostamata neid otseselt rekombinatsiooniga (Wall *et al.* 2003). Teine seostab haplotüübi blokid rekombinatsiooniga ning blokkide arvutamise aluseks kasutatakse lookuste vahelist LD-d (Wall *et al.* 2003).

1.3.1.1 Madalal haplotüüpide mitmekesisusel põhinevad meetodid

Meetodi eesmärgiks on leida esindavate SNP-ide kogum, mille abil on võimalik kirjeldada kõik sagedasemad haplotüübid. Selle saavutamiseks võetakse arvesse kõik blokid, mis sisaldavad ühte või enam järjestikust markerit ning mis vastavad ette antud kriteeriumitele. Bloki loomise kriteeriumiks võib olla kitsendus, et 80% kõikidest blokki moodustavatest haplotüüpidest peavad esinema populatsioonis sagedamini kui 5% (Patil *et al.* 2001). Kõikide võimalike blokkide hulgast valitakse lõplikud blokid, mis kirjeldavad kõiki peamisi haplotüüpe vähemalt etteantud kitsenduste piires. Lõpliku bloki hulga valimiseks võib kasutada ahnet algoritmi. Ahne algoritm valib võimalike blokkide hulgast pikima bloki ning valitud blokiga kattuvad blokid eemaldatakse. Protseduuri korratakse seni, kuni ei leidu ühtegi blokki, mis teisega kattuks ja kogu uuritav piirkond on blokkide ning tagSNP-dega kaetud (Patil *et al.* 2001). Efektiivsem meetod valib dünaamilise programmeerimise algoritmi abil võimalike kandidaatblokkide hulgast blokke, mis kirjeldavad minimaalse arvu tagSNP-de abil kogu uuritava piirkonna (Zhang *et al.* 2002).

Teiseks võimaluseks madala mitmekesisusega piirkondade leidmiseks on kasutada heterosügootsuse mõõtu. Selleks luuakse kõik viiest järjestikusest SNP-ist koosnevad kogumid - aknad. Arvestatakse tegelikku ja teoreetilist heterosügootsust ning igale viiest markerist koosnevale aknale määratakse skoor $S_5 = HET_{tegelik} / HET_{teoreetiline}$, kus *HET* tähistab heterosügootsust. Mida väiksem on skoor, seda madalam on mitmekesisus. Madala skooriga aknaid pikendatakse või lühendatakse markerite lisamise või eemaldamisega ning leitakse pikimad markerite järjestused, mille skoor on

võimalikult madal. Saadud akende piirid on haplotüübi bloki piirideks (Daly *et al.* 2001).

1.3.1.2 LD-l põhinevad meetodid

LD-l põhineva haplotüübi bloki moodustavad järjestikku asetsevad lookused, mille vahel pole võimalik tuvastada populatsiooni ajaloos toimunud rekombinatsioone või nende vahel on rekombinatsioonide sagedus väga madal (Gabriel *et al.* 2002). Iga lookuse paari vahel leitakse populatsiooni LD usaldusintervallid, mille abil vähendatakse populatsiooni valimist tulenevaid kõrvalekaldeid ja genotüpiseerimise iseärasusest tekkinud ebamäärasust. Lookuste vahelised D' väärtused jaotatakse kolme kategooriasse: tugev LD (|D'| läheneb ühele, mis viitab lookuste vahel toimunud rekombinatsiooni puudumise või harvale esinemisele populatsiooni ajaloos), nõrk LD (|D'| on oluliselt madalam ühest, millest järeldub, et lookuste vahel on toimunud populatsiooni ajaloos rekombinatsioon) ning vahepealne/määramatu LD. Kolmandasse kategooriasse kuuluvad lookuste paarid, mille LD väärtus on nulli ja ühe vahel või mille usaldusintervallide piirid on ebamääraselt laiad (Gabriel *et al.* 2002). Kaks või rohkem lookust moodustavad bloki, kui äärmiste lookuste vahel on tugev LD ja bloki moodustavate lookuste tugevas LD-s olevate paaride osakaal on nõrgas LD-s olevatest paaridest 19 korda kõrgem (Wall *et al.* 2003).

1.3.1.3 LD-l ja madalal haplotüüpide mitmekesisusel põhinevad meetodid

Madala mitmekesisusega piirkondadeks loetakse need, milles viie haplotüübi sageduste summa vaadeldavas populatsioonis on vähemalt 75% ning aheldatuse mittetasakaalustatus iga regioonis oleva markeri ja regiooni ümbritseva markeri vahel oleks kõrgem kui 0,75 (Dawson *et al.* 2002). Vastavate markerite kogumite (haplotüübi võrkude) leidmist, mis vastaksid nendele tingimustele, alustatakse ühest algusmarkerist, millele lisatakse markereid juurde. Haplotüübi võrgustikus ei pea olema järjestikused markerid, kuid võib seada piiranguid välja jäävate markerite arvu suhtes. Näiteks

lõpetada uute markerite otsimine, kui võrgustikku pole sobinud kuus järjestikust markerit (Dawson *et al.* 2002).

1.3.1.4 Nelja gameedi reeglil põhinev meetod

Mitmed haplotüübi blokkide definitsioonid ja nende leidmise algoritmid vajavad eelnevalt määratud läviväärtusi. Kui aluseks on võetud markerite vaheline LD, siis blokki kuuluvate markerite vaheline keskmine D' väärtus peab ületama mingi etteantud väärtuse. Kui blokkide moodustamisel jälgitakse, et blokk sisaldaks minimaalse arvu markereid, mille abil saab kirjeldada enamuse sagedastest haplotüüpidest, on samuti vaja eelnevalt määrata väärtused, mille abil otsustada markerite blokki kuuluvus. Erinevad blokkide definitsioonid, blokkidesse määramise algoritmid ja algoritmide erinevad parameetrid annavad tulemuseks erinevad haplotüübi blokkide mustrid (Wang *et al.* 2002). Üheks lihtsaks ja alternatiivseks võimaluseks on kasutada nelja gameedi reeglit, mille puhul ei ole läviväärtust ning mille abil saab määrata, kas lookuste vahel on toimunud rekombinatsioon.

Nelja gameedi reegli puhul eeldatakse, et uuritavates lookustes pole toimunud tagasi ja/või korduvat mutatsiooni. Kahe lookuse A ja B puhul, kus mõlemal on kaks alleeli, vastavalt A_1 ja A_2 ning B_1 ja B_2, saab olla ilma nende vahel toimunud rekombinatsioonita kolm haplotüüpi (n. A_1B_1, A_1B_2 ja A_2B_1). Juhul kui leidub neli haplotüüpi (n. A_1B_1, A_1B_2, A_2B_1 ja A_2B_2) ning lookuses pole toimunud korduvat mutatsiooni, siis saab seda põhjustada ainult lookuste vahel toimunud rekombinatsioon (Wang *et al.* 2002).

Nelja gameedi testiks nimetatakse lookuste vahelist nelja gameedi reegli kontrollimist. Haplotüübi blokkide leidmiseks m markerist koosnevast kogumist viiakse iga markeri paari vahel läbi nelja gameedi test, et leida lookused, mille vahel on toimunud rekombinatsioon. Haplotüübi bloki moodustavad järjestikkused markerid, mille vahel pole toimunud rekombinatsiooni. Blokkide moodustamist alustatakse uuritava regiooni alusest nelja gameedi testi tulemuste põhjal järjestikkuste markerite lisamisega. Blokki lisatakse järjest markereid, kuni haplotüüpide arv ei ületa kolme. Kui jõutakse markerini, mis moodustab neli haplotüüpi kõigi talle eelnevate markeritega, siis loetakse see marker uue bloki algusmarkeriks (Wang *et al.* 2002).

1.3.2 Haplotüüpe määravad markerid

Kui on määratud populatsioonis enam levinud haplotüübid, saab kindlaks teha, millised markeritest on taandatavad ning millised olulised assotsiatsiooni uuringuteks. Informatiivsed markerid on htSNP-d, mille abil saab määrate haplotüüpe aheldatuse mittetasakaalus olevates piirkondades (Johnson *et al.* 2001). Eelnevalt määratud haplotüübi struktuuri teadmine ning haplotüüpe määravate markerite valimine vähendaks genotüpiseerimise kulutusi uuringutes, kus võrreldakse geneetilisi variatsioone fenotüübiliste tunnustega. Tähtsam on asjaolu, et sellise lähenemise puhul osalevad uuringus kõik vairatsioonid, mis jäävad haplotüüpi määrava markeri kirjeldatavasse piirkonda (Johnson *et al.* 2001).

Informatiivseid SNP-e võib defineerida mitmeti ning nende leidmiseks on välja töötatud mitmeid algoritme. Näiteks tarkvara programmi HapBlock (http://www.cmb.usc.edu/msms/HapBlock/) on implementeeritud viis erinevat informatiivsete SNP-ide valiku algoritmi (Zhang *et al.* 2005). Üheks võimaluseks on määrata SNP-id, mis on assotsieerunud teiste genoomis olevate SNP-idega. Teine lähenemine püüab SNP-ide abil kirjeldada eelnevalt määratud haplotüübi blokke. Kirjanduses nimetatakse informatiivseid SNP-e tagSNP-ideks või htSNP-ideks.

LD mõõdu r^2 abil saab määrata nii tag- kui ka htSNP-e. TagSNP-ide puhul kasutatakse ahnet algoritmi, mis esimese sammuna leiab SNP-ide kogumid, mille omavaheline r^2 väärtus ületab etteantud nivoo. Ühte gruppi kuuluvad SNP-id ei pea asuma kromosoomis järjest, vaid võivad paikneda vaheliti. Grupi moodustavatest SNP-idest valitakse üks, mille keskmine r^2 väärtus teiste grupi SNP-ide suhtes on kõige kõrgem ning see SNP valitakse grupi tagSNP-iks. Polümorfismide grupeerimist jätkatakse kuni kõik polümorfismid kuuluvad mingisse gruppi. Grupi võib moodustada ka üksik SNP juhul kui ta ei ole LD-s ühegi teise SNP-iga (Carlson *et al.* 2004). Haplotüübi tagSNP-ide määramiseks peab eelnevalt määrama haplotüübi blokid. Parima htSNP-ide kogumi valikul võetakse arvesse SNP-i haplotüüpide eristamise võime, optimaalne r^2 väärtus teiste blokki moodustavate SNP-ide vahel ning SNP-i asukoht haplotüübi blokis (Zhang *et al.* 2003).

2. GEENIDE ANNOTATSIOON

Mitmete organismide genoomid on sekveneeritud. Genoomi järjestuste hüppelise kasvuga on oluline töötada välja meetodid, mis suudaks nendest järjestustest leida geene ning muud olulist informatsiooni. Automaatne geenide ennustamine on keeruline protsess, mille tarvis töötatakse välja üha keerulisemaid algoritme. Hinnanguliselt on inimese genoomis 30000 geeni, millest 70% on automaatselt ennustatavad (Ashurst *et al.* 2003).

Enamus inimese geene on annoteeritud cDNA andmete põhjal. Oluliselt on edendanud geenide annoteerimist suuremahulised cDNA sekveneerimise programmid nagu Mammalian Gene Collection (MGC) (Strausberg *et al.* 1999) Ameerika Ühendriikides ning RIKEN (http://genome.rtc.riken.go.jp/) Jaapanis. Ligi pooled RefSeq andmebaasis annoteeritud geenidest pärinevad sellistest laiaulatuslikest cDNA sekveneerimise projektidest. Samas arendatakse ka järjest paremaid arvutuslikke meetodeid geenide ennustamiseks.

2.1 Geenide ennustamine

Eukarüootsete organismide valke kodeerivad geenid sisaldavad introneid ja eksoneid. Eksonid jagunevad nelja klassi: 5' eksonid, sisemised eksonid, 3' eksonid ning intronite vabad eksonid. Neli eksonite klassi jagunevad veel omakorda kaheteistkümneks alaklassiks, millel on erinevad statistilised omadused. Iga klassi eksonite ennustamine on pisut erinev ning vajab erinevat lähenemist (Zhang 1998).

2.1.1 Sisemiste eksonite ennustamine

Sisemiste eksonite ennustamiseks kasutati varasemates algoritmides pre-mRNA intronites sisalduvaid konsensus järjestusi, mis moodustavad splaisosoomi RNA elementidega aluspaaride vahelisi sidemeid ning aitavad läbi viia splaisingut. Doonor saidi ehk 5' splaisingsaidi konsensusjärjestust kirjeldab AG|**GU**RAGU ning aktseptor

saidi ehk 3’ splaisingsaidi konsensusjärjestust kirjeldab YYYYYYYYYYNC**AG**|G või vähem konserveerunud järjestus CUR**AY** (Senapathy *et al.* 1990). Pärmis ja teistes madalamates eukarüootides on lühikesed intronid, mille otstega interakteeruvad splaisingu faktorid ning seovad need. Sellist introni põhist geeniennustust on kasutatud mõningates algoritmides. Paraku ei ole aga selline lähenemine piisav, kuna lisaks konsensusjärjestustele leidub teisi peidetud splaisingu saite.

Selgroogsete genoomis on sisemised eksonid väikesed (keskmiselt 140 aluspaari) ja intronid suured. 1990. aastal välja pakutud ja siiani ümberlükkamata eksoni definitsiooni mudeli kohaselt seondub eksoniga enne, kui intronid molekulaarselt ära tuntakse ja välja lõigatakse, valguline faktor, mis moodustab üle eksoni silla teiste pre-mRNA-le seostunud valkudega (Robberson *et al.* 1990). Vastavate faktorite seostumine pre-mRNA-le on tingitud nukelotiidjärjestusest, mille abil on püütakse geene ennustada. Järjetuse tunnused jaotatakse kaheks: signaalideks ja sisuks. Signaalid on lühikesed cis-elemendid või splaisingsaidid ning sisu on ulatuslikumad funktsionaalsed regioonid nagu eksonid või intronid. Tunnusele määratakse hinnangu funktsioon, mille abil saab tunnust iseloomustada. Parimaks hinnangu funktsiooniks on tõenäosus, et järjestus s sisaldab tunnust a - P(a/s). Vastavalt Bayesi võrrandile P(a/s)=P(s/a)P(a)/P(s), kus P(s/a) on tõenäosus, et s sisaldab a-d. Edasiseks moodustatakse proovid, milles on eelnevalt teada tunnus a ning hinnatakse järjestust s. Mitmed hinnangud koondatakse ühte skoori, mis kirjeldab tervet objekti. Geeni ennustamisel leitakse sellised struktuurid, millel on kõrgeim skoor. Varieerida saab uuritavaid tunnuseid, skoori funktsiooni ning integratsiooni meetodeid. Kui eksoni ennustamise probleem on defineeritud kui statistiline mustri otsimise probleem, võib rakendada mitmeid statistilisi või masinõppe vahendeid vastavate mustrite leidmiseks (Zhang *et al.* 2002).

2.1.2 3’ eksonite ennustamine

Geenide 3’ otsa ennustamine on lihtsam kui geeni 5’ otsa ennustamine. Selle põhjuseks on asjaolu, et GenBank andmebaaside on paljude mRNAde ja EST järjestuste 5’ otsad kärbitud. Eksoni definitsiooni mudelit saab kasutada 3’-eksonite puhul asendades mudelis 5’-spalising saidi polü(A)-saidiga. Kuna selgoogsete 3’-eksonid on sisemistest eskonitest pikemad, ulatudes kuni mõne tuhande aluspaarini, siis tuleb arvestades 3’-

eksonite pikkusjaotust. 3'-eksonite splaisingu käigus moodustavad molekulaarse silla splaisingu faktor U2AF65 ja polü(A)-saiti ära tundva polü(A)-polümeraasi karbonüülterminuse domään (Zhang *et al.* 2002).

EST 3' järjestuste võrdlemisel genoomi järjestusega on leitud mitmeid polü(A) saite. Mitmed järjestuse tunnused nagu polü(A) signaal AAUAAA ja G+U rikas sait on kirjeldatud kuue organismi (pagaripärm, riis, müürlook, äädikakärbes, hiir ja inimene) genoomis ning nende abil on ennustatud polü(A) saite (Tabaska *et al.* 1999). Usaldusväärsemaid geeni 3' otste järjestusi saab määrata mRNA võrdlemisel genoomi järjestustega.

2.1.2 5' eksonite ennustamine

Geeni 5' otsa ennustamine on kõige raskem ülesanne geeni ennustamisel. Selle põhjuseks on promootorite ja transkriptsiooni algussaidi järjestuse määramise keerulisus. Enamus GenBank-is olevatest mRNA-delt saadud cDNA järjestusi on kärbitud 5' otsast, sest pöördtranskriptaas ei suuda cDNA ahelat alati lõpuni sünteesida (Zhang *et al.* 2002).

Promootorite aktivatsioon ja transkriptsiooni initsiatsioon on keeruline protsess. Promootorite ümbruses olev kromatiin viiakse hüperatsetüleeritud ja lõdvestatud olekusse, järgmisena seostub tuum-promootorile initsiatsiooni eelne kompleks. Transkriptsiooni initsiatsiooni reguleerivad peamiselt transkriptsiooni faktorid, mis seostuvad promootori lähedal asuvale regioonile ning esimese introni piirkonda. Mitmed programmid püüavad ennustada promootoreid ja translatsiooni algussaite. Paraku pole nende ennustused veel rahuldaval tasemel.

Sarnaselt 3' eksonite ennustamisega ei suuda enamus programme ennustada tegelikke 5' eksoneid, vaid 5' kodeerivat järjestust. 5' kodeeriva järjesutse ennustamiseks kasutatakse sisemiste eksonite ennustamisele sarnast lähenemist asendades ostitava 3' splasingu signaali initsiatsiooni signaaliga, võttes arvesse translatsiooni initsiatsiooni konsensusjärjestusi ja arvestades 5' eksonite pikkuste jaotust (Zhang *et al.* 2002).

2.1.3 Introniteta geenide määramine

Introneid mittesisaldavate geenide ennustamise teeb keeruliseks pseudogeenide olemasolu. Heksameerne kodeerimismõõt, perioodiline või entroopial põhinev kodeerimismõõt töötavad tõhusalt pikkade kodeerivate järjestuste ennustamisel, kuid need ei suuda eristada introniteta geeni ja pikka sisemist eksonit. Kuna paljud pseudogeenid on tekkinud splaisitud geenidest, siis on väga raske eristada neid introniteta kodeerivatest järjestustest kui pole teada millest geenist pseudogeen on tekkinud või pseudogeeni pole akumuleerunud nonsens mutatsioonid (Zhang *et al.* 2002).

2.1.4 Eksonite kokkupanek ja üksiku transkripti ennustamine

Geeni ennustamismudel annab paremaid tulemusi kui sellesse on kaasatud erinevaid parameetreid. Splaingsaidi signaalid koos kodeerimist iseloomustavate parameetritega võimaldavad paremini ennustada splaisingsaite. Eksoneid saab paremini ennustada kui võtta arvesse eksonite paigutust ühes transkriptis. DNA järjestus geenis pole juhuslik ning kui kaasata erinevad parameetrid ühte mudelisse, siis võimaldab see täpsemalt ennustada geene.

Esimest ja viimast eksonit on kõige keerulisem ennustada, seetõttu paljud programmid ennustavad kodeerivat järjestust defineerides 5’ eksoni ATG-GT, sisemise eksoni AG-GT, 3’ eksoni AG-STOP ning introniteta kodeeriva järjesutse ATG-STOP. Vähe on programme, mis püüavad leida tervet mRNA-d mitte kodeerivaid järjestusi. Mitte kodeerivat 5’otsa defineeritakse TSS-ATG ning 3’ otsa STOP-polü(A) (Zhang 2002).

Erinevate fragmentide geeniks kokku liitmiseks kasutatakse dünaamilist programmeerimist või HMM’i. Leides kõik võimalikud geenifragmendid ning määrates neile skoori, saab dünaamilise programmeerimise abil leida parima skooriga osade kombinatsiooni. Eelnevalt tuleb fragmentide skoorid optimiseerida, et need oleks omavahel võrreldavad (Zhang *et al.* 2002). HMM puhul jagatakse DNA järjestus eraldi seisvateks fragmentideks või olekuteks. Tõenäosust, et alus s on olekus q tähistame P(s/q). Ülemineku tõenäosus, et olek q-le järgneb olek q’ tähistame T(q/q’). Olekute

järjestust tähistatakse Φ -ga. Kui tõenäosused P ja T olekute paigutusele $\{q_i : i = 1,2,...,N\}$ on teada, siis ühine tõenäosus olekute järjestusele on $P(\Phi, S) = P(s_1|q_1)T(q_1|q_2)P(s_2|q_2)...T(q_{N-1}|q_N)P(s_N|q_N)P_0(q_N)$. Viterbi algoritmi abil saab leida kõige tõenäolisema Φ, mis viitab optimaalseimale transkriptile (Zhang *et al.* 2002).

HMM eelis on tema paindlikus, kuna mudelisse on võimalik lihtsalt liita erinevaid olekuid nagu intergeensed piirkonnad, promootorid, mitte transleeritavad piirkonnad, polü(A) ja raami- või ahelda sõltuvad eksonid ja intronid. Samuti võimaldab HMM lisada olekuid ja üleminekuid, mis võimaldavad mitmekordseid geene, osalisi geene ning mõlemal ahelal asuvaid geene üheskoos ennustada.

2.1.5 Sarnasuse skooride kombineerimine

Andmebaasidest otsimise- ja joondamise programmide nagu BLASTX (Gish *et al.* 1993) ja Sim4 (Florea *et al.* 1998) kasutamine geenide ennustamiseks on olnud populaarne, kuna varem teada olevate valgu või cDNA/EST järjestuste võrdlemine võib palju parandada geenide ennustamise täpsust. Traditsiooniliselt viidi sarnasuste otsing läbi eraldi geeni ennustusest ning tulemusi kombineeriti käsitsi. Praegu leidub mitmeid programme, mis teevad seda automaatselt.

Eksonite piiride identifitseerimine on efektiivsem kui geenile leidub lähedane homoloog, mille võrdluse andmeid saab siduda muude parameetritega. Ensembli andmebaasi poolt kasutatav automaatne geeni annotatsiooni algoritm seob geeni ennustuse HMMi valgu profiili HMMiga (Pfam) ning viib läbi üheaegselt geeniennustust ning joondamist (Birney *et al.* 2000). Homoloogide kasutamine geeni ennustuseks eeldab loomulikult vastavate homoloogide leidumist genoomis, on arvtuslikult raske ning vajab eelnevalt vastava homoloogi piirkonda leidmist. Üldiselt sarnasuste otsing geeni ennustamisel parandab algoritmide täpsust (Zhang *et al.* 2002).

2.1.6 Võrdleva genoomika meetodid

Lähedases suguluses olevate genoomide olemasolu võimaldab läbi viia genoomide võrdlust. Kui kaks genoomi on hiljuti lahknenud, siis on paljud genoomis sisalduvad geenid, geenide arv, geenide asetus ning isegi geenide struktuurid kõrgelt konserveerunud. Uusi geene on võimalik leida otseselt genoomide võrdlemisel. Lähedaste liikide genoomide võrdlemisel on võimalik tuvastada konserveerunud regulatoorseid piirkondi .(Pennacchio *et al.* 2001) Seetõttu omab võrdlev genoomika tähtsat rolli geenide ennustamisel. Selle tarvis on loodud mitmeid programme ning mitmeid erinevaid strateegiaid.

Geeni ennustamisel on abiks genoomide joondamine. Genoomide võrdlemiseks on loodud visualiseerimise programmid, mis väljastavad kahe või enama genoomi joondatud järjestused. Joondamise parandamiseks võtab näiteks programm WABA (Kent *et al.* 2000) arvesse kodeerivate eksonite kolmanda aluspaari võbelust.

Mitmed programmid kasutavad üldistatud paari HMMi, mille abil on võimalik ennustada ortoloogseid aluspaare vastavalt kaksik-HMMile. Erinevalt tavalisest HMMist loob kaksik-HMM olekute paari, millest üks ühest teine teisest organismist.

Inimese genoomi võrdlemisel hiire genoomiga on selgunud, et ligi pooled konserveerunud alad ei asu kodeerivates piirkondades ning need põhjustavad segadust algoritmi töös. Probleemi lahendamiseks on mudelisse lisatud konserveerunud mittekodeeriv olek. Paraku ei osata seda olekut piisavalt täpselt defineerida ning pole piisavalt andmeid, millel vastavaid programme treenida (Zhang *et al.* 2002).

II PRAKTILINE OSA
TÖÖ EESMÄRGID

Käesoleva töö eesmärgiks on geenide ja eksonite koordinaatide ja haplotüübi bloki koordinaatide võrdlemisel selgitada, kas geenid ja eksonid asuvad eelistatud haplotüübi blokkide sees, blokkidest väljas või juhuslikult.

1. MEETODID

1.1 Katseandmete kirjeldus

Töös on kasutatud HapMap projekti genotüpiseerimise andmeid. Vaatluse all on kahe erinevat populatsiooni: Nigeeria päritolu populatsiooni Yoruba ja Lääne- ja Põhja Euroopa päritolu CEPH populatsioon. Mõlemad andmed sisaldavad 30 ema-isa-laps kolmikut, kokku 90 indiviidi (The International HapMap Consortium 2003).
HapMap andmebaasist pärinevad andmed:
a) http://www.hapmap.org/genotypes/latest_ncbi_build35/non-redundant/:
genotypes_chr1_CEU.b35.txt.gz (281773 markerit),
genotypes_chr1_YRI.b35.txt.gz (279637 markerit),
genotypes_chr2_CEU.b35.txt.gz (307541 markerit),
genotypes_chr2_YRI.b35.txt.gz (298646 markerit),
genotypes_chr3_CEU.b35.txt.gz (240283 markerit),
genotypes_chr3_YRI.b35.txt.gz (232846 markerit),
genotypes_chr4_CEU.b35.txt.gz (228604 markerit),
genotypes_chr4_YRI.b35.txt.gz (220582 markerit),
genotypes_chr5_CEU.b35.txt.gz (234026 markerit),
genotypes_chr5_YRI.b35.txt.gz (226865 markerit),
genotypes_chr6_CEU.b35.txt.gz (255354 markerit),
genotypes_chr6_YRI.b35.txt.gz (252062 markerit),
genotypes_chr7_CEU.b35.txt.gz (194065 markerit),
genotypes_chr7_YRI.b35.txt.gz (188175 markerit),

genotypes_chr8_CEU.b35.txt.gz (202830 markerit),
genotypes_chr8_YRI.b35.txt.gz (200915 markerit),
genotypes_chr9_CEU.b35.txt.gz (170858 markerit),
genotypes_chr9_YRI.b35.txt.gz (168844 markerit),
genotypes_chr10_CEU.b35.txt.gz (197408 markerit),
genotypes_chr10_YRI.b35.txt.gz (195705 markerit),
genotypes_chr11_CEU.b35.txt.gz (191161 markerit),
genotypes_chr11_YRI.b35.txt.gz (184716 markerit),
genotypes_chr12_CEU.b35.txt.gz (180613 markerit),
genotypes_chr12_YRI.b35.txt.gz (178091 markerit),
genotypes_chr13_CEU.b35.txt.gz (148170 markerit),
genotypes_chr13_YRI.b35.txt.gz (147631 markerit),
genotypes_chr14_CEU.b35.txt.gz (116348 markerit),
genotypes_chr14_YRI.b35.txt.gz (112921 markerit),
genotypes_chr15_CEU.b35.txt.gz (100889 markerit),
genotypes_chr15_YRI.b35.txt.gz (97703 markerit),
genotypes_chr16_CEU.b35.txt.gz (102501 markerit),
genotypes_chr16_YRI.b35.txt.gz (98868 markerit),
genotypes_chr17_CEU.b35.txt.gz (83500 markerit),
genotypes_chr17_YRI.b35.txt.gz (80793 markerit),
genotypes_chr18_CEU.b35.txt.gz (112827 markerit),
genotypes_chr18_YRI.b35.txt.gz (110834 markerit),
genotypes_chr19_CEU.b35.txt.gz (51694 markerit),
genotypes_chr19_YRI.b35.txt.gz (49941 markerit),
genotypes_chr20_CEU.b35.txt.gz (114478 markerit),
genotypes_chr20_YRI.b35.txt.gz (111464 markerit),
genotypes_chr21_CEU.b35.txt.gz (46515 markerit),
genotypes_chr21_YRI.b35.txt.gz (46680 markerit),
genotypes_chr22_CEU.b35.txt.gz (51499 markerit),
genotypes_chr22_YRI.b35.txt.gz (52310 markerit).
b) http://www.hapmap.org/genotypes/2005-06_16c_phaseI/full/non-redundant/:
genotypes_chr22_CEU.txt.gz (19120 markerit),

genotypes_chr22_YRI.txt.gz (19854 markerit).
c) http://www.hapmap.org/genotypes/2005-09_phaseII_6chrs/non-redundant/:
genotypes_chr11_CEU.txt.gz (190330 markerit),
genotypes_chr11_YRI.txt.gz (185258 markerit),
genotypes_chr14_CEU.txt.gz (116616 markerit),
genotypes_chr14_YRI.txt.gz (113177 markerit),
genotypes_chr15_CEU.txt.gz (100498 markerit),
genotypes_chr15_YRI.txt.gz (97993 markerit),
genotypes_chr21_CEU.txt.gz (46689 markerit),
genotypes_chr21_YRI.txt.gz (46843 markerit).

Lisaks on töös võrdlusena kasutatud ahne algoritmi (vt. punkt 1.2.1.2) abil arvutatud 21. kromosoomi haplotüübi blokke, dünaamilise programmeerimise algoritmi (vt. punkt 1.3.1.1) abil arvutatud 21. kromosoomi blokke ja 19. kromosoomi haplotüübi blokke.
19. kromosoomi haplotüübi blokid on arvutatud 10 CEPH perekonna (Dausset *et al.* 1990) kromosoomide genotüpiseerimise andmetelt (Phillips *et al.* 2003). Iga sugupuu sisaldab nelja vanavanemat, kahte vanemat ja kahte järglast (kokku 80 kromosoomi).
21. kromosoomi haplotüübid pärinevad 21. kromosoomi haplotüübi blokkide andmebaasist (http://genome.perlegen.com/haplotype/). Haplotüübi andmed sisaldavad 20 haploidset kromosoomi, mis esindavad Aafrika, Aasia ja Kaukaasia kromosoome. Igal kromosoomil on analüüsitud 24047 SNP-i, keskmiselt üks SNP iga 1300 aluspaari kohta (Patil *et al.* 2001).
Töös kasutatud geenide ja eksonite koordinaadid pärinevad Ensembl andmebaasist http://oct2004.archive.ensembl.org/Multi/martview, mis on vastavuses *Homo sapiens* genoomi verisooniga NCBI 34, ja http://feb2006.archive.ensembl.org/Multi/martview, mis on vastavuses *Homo sapiens* genoomi verisooniga NCBI 35. Uuringusse on kaasatud kõik Endsembli andmebaasi geenid ja eksonid. Nende hulgas on pseudogeenid, RNA geenid ja valke kodeerivaid geenid.
HapMap andmed a on vastavuses genoomi verisooniga NCBI 35, HapMap andmed b ja c genoomi verisooniga NCBI 34. Dünaamilise programmeerimise ja ahne algoritmi abil arvutatud blokid viisime vastavusse genoomi versiooniga NCBI 34. Sellest tulenevalt võrdlesime edasises töös HapMap andmete a põhjal arvutatud haplotüübi blokke

geenidega ja eksonitega, mis pärinevad andmebaasist http://feb2006.archive.ensembl.org/Multi/martview, kõiki ülejäänud haplotüübi blokke geenidega ja eksonitega, mis pärinevad andmebaasist http://feb2006.archive.ensembl.org/Multi/martview.

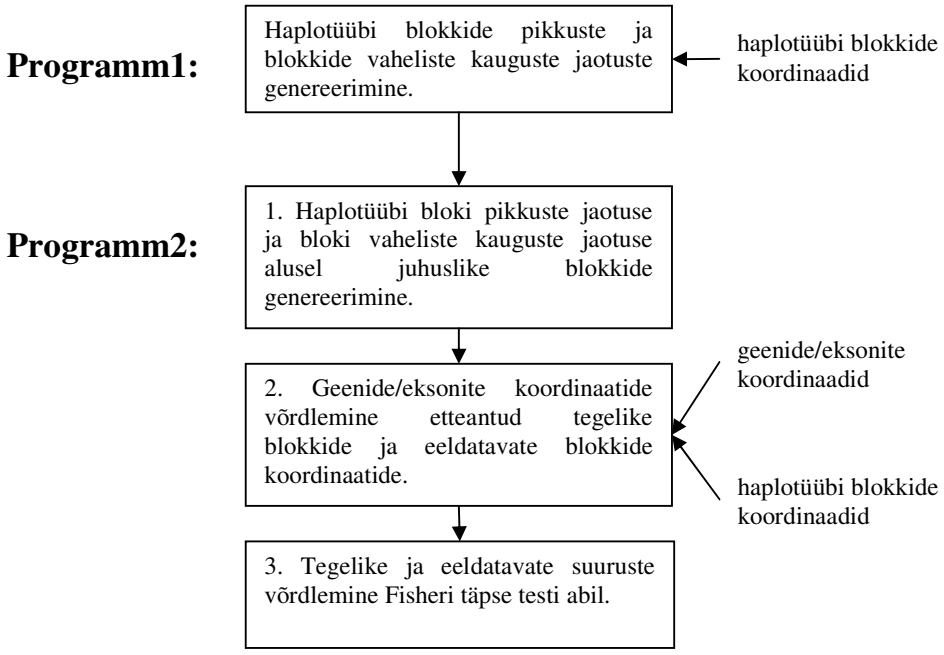
1.2 Kasutatud programmide kirjeldus

21. kromosoomi haplotüübi blokkide koordinaatide viimiseks geeni piiridega ühtsesse koordinaatsüsteemi kasutasime programmi GenomeTester 1.1, mis leiab etteantud nukleotiidse järjestuse koordinaadid genoomis. Andmebaasis (http://genome.perlegen.com/haplotype/) on haplotüübi blokkide koordinaadid esitatud kontiigide NT_002836, NT_00135, NT_003545 ja NT_002835 positsioonidena. Genoomi koordinaatide leidmiseks võrdlesime GenomeTester-i abil SNP-i ümbritsevat kontiigi järjestust genoomi järjestustega.

Hapmap projekti genotüpiseerimise andmetest haplotüübi blokkide arvutamiseks kasutasime programmi Haploview 3.2 (http://www.broad.mit.edu/mpg/haploview/download.php). Blokke arvutasime LD-l põhineval meetodil (vt. punkt 1.3.1.2) ja nelja gameedi reeglil põhineval meetodil (vt. punkt 1.3.1.4.) kasutades Haploview vaikeparameetreid.

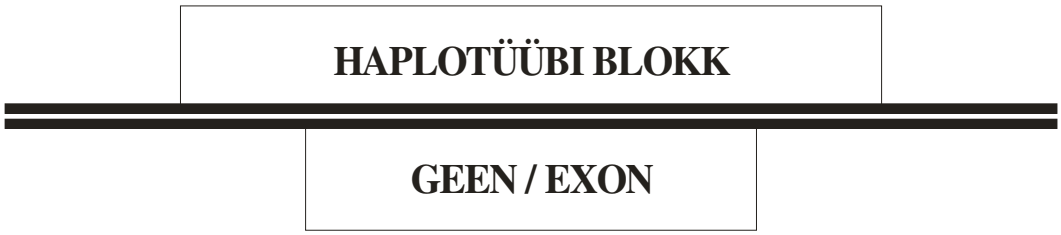
Töö käigus valmisid programmid Programm1 ja Programm2. Programm1 sisendandmeteks on haplotüübi blokkide alg- ja lõppkoordinaadid (joonis 1). Haplotüübi bloki koordinaatidena käsitletakse blokki kuuluva esimese ja viimase SNP-i koordinaate. Programm1 väljastab haplotüübi blokkide pikkuste ja haplotüübi blokkide vahelise kauguse jaotuse tuues välja vastava pikkusega blokkide arvu ning kumulatiivväärtuse kõigi blokkide hulgas.

Programm2 genereerib juhuslikke haplotüübi blokke ja võrdleb haplotüübi blokkide geenide/eksonite koordinaatide. Programm2 sisendandmeteks on haplotüübi blokkide algus- ja lõppkoordinaadid, geenide või eksonite algus- ja lõppkoordinaadid, Programm1 poolt väljastatud haplotüübi blokkide pikkuste jaotus ja bloki vaheliste kauguste jaotus.



Joonis 1. Programm1 ja Programm2 töö etapid.

Programm2 töö jaotub kolme etappi (joonis 1). Esimeses etapis genereeritakse juhuslikud haplotüübi blokid, mille pikkuste jaotus vastab etteantud tegelike blokkide pikkuste ja blokkide vaheliste kauguste jaotusele. Teises etapis võrreldakse geeni algus ja lõpp koordinaatide tegelike haplotüübi blokkide ja juhuslike haplotüübi blokkide koordinaatidega. Programm2 loendab juhte, kus geeni algus- ja lõppkoordinaat paiknevad sama haplotüübi bloki sees (joonis 2). Kõiki ülejäänud paiknemise juhte käsitletakse kui loendatu vastandit.



Joonis 2. Programm 2 poolt loendatud geeni koordinaatide asetus haplotüübi bloki piiride suhtes. Must kahekordne joon tähistab kromosoomi. Ristküliku küljed viitavad algus- ja lõppkoordinaatidele.

Programm2 kasutab juhuslike blokkide tulemuste ja tegelike blokkide tulemuste võrdlemiseks Fisheri täpset testi. Fisheri täpse testi abil saab määrata iga erineva kaks korda kaks sagedustabeli (tabel 1) saamise tõenäosuse vastavate ridade ja tulpade summade korral. Iga konkreetse tabeli seisu saamise tõenäosus on väljendatav hüpergeomeetrilise jaotusega ning arvutakse valemiga

$$p = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{n!a!b!c!d!}.$$

Andmete võrdlemisel viisime läbi kahepoolse hüpoteesi kontrolli. Statistiliseks oluliseks lugesime p-väärtust, mis on väiksem kui 0,001.

Tabel 1. Andmete esitamine kaks korda kaks sagedustabelis.

	Tegelikud haplotüübi blokid	**Juhuslikud haplotüübi blokid**	**Kokku**
Geen või ekson haplotüübi bloki sees	a	b	a + b
Geen või ekson haplotüübi blokist väljas	c	d	c + d
Kokku	a + c	b + d	n = a + b + c +d

2. TULEMUSED

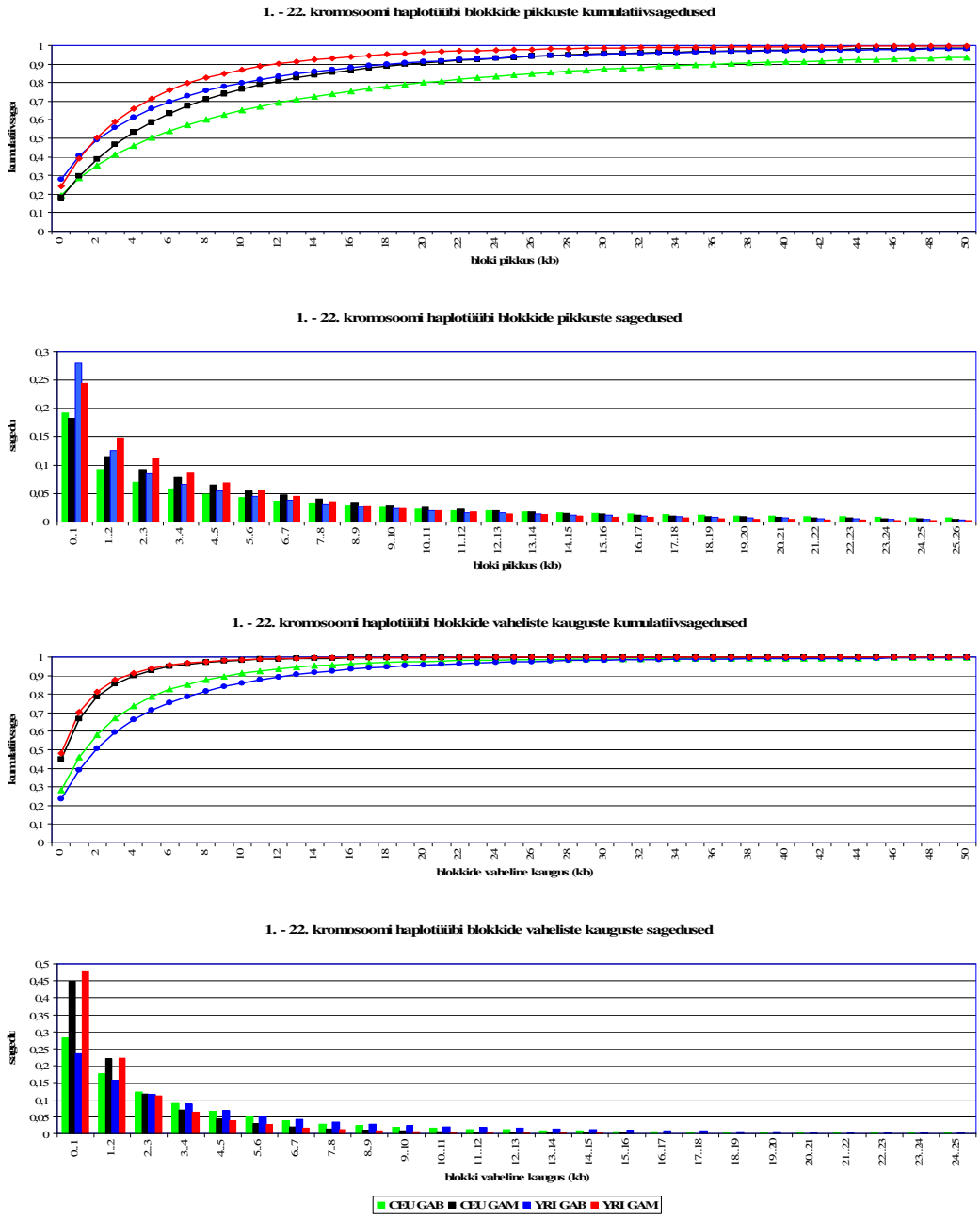
2.1 Haplotüübi blokkide kirjeldus

Lähema vaatluse all on 1. kuni 22. kromosoomi haplotüübi blokid, mis on arvutatud kahe erineva algoritmiga HapMap a andmetelt (vt. punkt II 1.1). Haplotüübi blokke iseloomutavad nende pikkused, omavaheline kaugus ja kogu arv. Joonisel 3 on välja toodud 22 kromosoomi haplotüübi blokkide ja nende vaheliste kauguste kumulatiivsed ja üldised sagedused.

Erinevate blokiarvutusmeetodite poolt leitud blokkide pikkuste jaotused on erinevad. LD-l põhinev meetod (vt. punkt 1.3.1.2) tuvastab nelja gameedi reegli meetodist (vt. punkt 1.3.1.4.) pikemaid blokke. Lühikeste, kuni üks kb, blokkide osakaal on LD-l põhinevatel blokkidel väiksem kui nelja gameedi reegli blokkidel (tabel 2). Erinevalt jaotuvad ka blokkide vahelised kaugused. LD-l põhinevate blokkide vahelised kaugused on pikemad kui nelja gameedi reegli meetodil arvutatud blokkide vahelised kaugused. Blokkide pikkus ja blokivaheliste kauguste jaotus pole erinevates populatsioonides sama. Töös on vaatluse all kaks populatsiooni: Nigeeria päritolu populatsioonis Yoruba ja Lääne- ja Põhja Euroopa päritolu CEPH. Nigeeria populatsioonis on lühemate blokkide osakaal suurem kui Euroopa päritolu populatsioonis. Töös kasutatud bloki arvutamise algoritmid leidsid Yoruba populatsioonis rohkem haplotüübi blokke kui CEPH populatsioonis (tabel 2).

Tabel 2. Erinevate populatsioonide genotüpiseerimise andmetelt kahe algoritmidega arvutad haplotüübi blokkide ja blokkide vaheliste kauguste erinevused. Blokkide arv on saadud summeerides 1. kuni 22. kromosoomi haplotüübi blokid. CEU viitab CEPH populatsioonile, YRI Yoruba populatsioonile. GAB tähendab LD-l põhinevat blokkide arvutamise meetodit (vt. punkt 1.3.1.2) ning GAM nelja gameedi reegli meetodit (vt. punkt 1.3.1.4.).

	CEU		**YRI**	
	GAB	**GAM**	**GAB**	**GAM**
blokkide arv	132662	249308	196857	361512
kuni 1 kb pikkuste blokkide osakaal	19,2%	18,2%	28,0%	24,4%
95 % blokkidest jääb alla	60 kb	30 kb	29 kb	18 kb
kuni 1 kb pikkuste vahede osakaal	28,2%	44,9%	23,6%	48,1%
95% vahedest jääb alla	14 kb	7 kb	19 kb	6 kb



Joonis 3. Haplotüübi blokkide ja nende vaheliste kauguste jaotused. CEU viitab CEPH populatsioonile, YRI Yoruba populatsioonile. GAB tähendab LD-l põhinevat blokkide arvutamise meetodit (vt. punkt 1.3.1.2) ning GAM nelja gameedi reegli meetodit (vt. punkt 1.3.1.4.). Sagedused on saadud vastava pikkusega blokkide või vahede arvu jagamisel kõikide blokkide või vahede arvuga. Kumulatiivsagedus saadakse vastava pikkusega bloki või vahe sageduse ja kõikide temast lühemate blokkide või vahede sageduste summaga.

2.2 Geenide ja eksonite kirjeldus

Töös on vaatluse all 1. kuni 22. kromosoomi geenid ja eksonid. Kromosoomid erinevad geenide ja eksonite arvu, tiheduse ning katvuse poolest. Kromosoome iseloomustavad arvulised väärtused on välja toodud tabelis 3 ja tabelis 4.

Tabel 3. Käesolevas töös uuritud kromosoomide pikkused, geenide ja SNP-ide arvud. (http://feb2006.archive.ensembl.org/Homo_sapiens/index.html). Teadaolevaid valke kodeerivad geenid on geenid, mis on kaardistatud ning mille valguline produkt on teada ning on kantud avalikesse andmebaasidesse. Ennustatud valke kodeerivad geenid on geenid, mida ei ole võimalik täieliku kindlusega kromosoomile kaardistada.

Kromosoom	Pikkus (Mb)	Teadaolevaid valke kodeerivad geenid	Ennustatud valke kodeerivad geenid	Pseudogeenid	miRNA	rRNA	snRNA	snoRNA	muud RNA-d	SNPs
1	245,5	2062	140	152	40	52	187	88	92	799984
2	243,0	1295	164	144	14	24	118	45	73	765356
3	199,5	1076	66	139	19	22	93	37	67	640775
4	191,4	762	90	110	9	13	86	19	58	659079
5	180,9	867	86	117	14	22	75	22	70	584533
6	171,0	1070	69	137	8	17	85	30	57	663380
7	158,6	957	133	111	24	14	66	35	63	551352
8	146,3	710	72	66	11	14	63	23	39	490444
9	138,4	802	91	79	32	10	44	23	48	509422
10	135,4	777	79	62	4	17	68	9	45	518949
11	134,5	1294	89	104	21	19	56	63	48	506846
12	132,4	1022	63	93	14	15	81	26	66	471782
13	114,1	334	53	58	16	9	30	15	34	364926
14	106,4	651	60	85	32	14	46	59	39	298796
15	100,3	610	89	75	11	6	48	145	35	283822
16	88,8	864	83	48	5	13	42	20	31	323251
17	78,8	1129	92	63	26	10	48	45	52	258192
18	76,1	278	35	40	8	5	43	16	21	268402

Tabel 3. (järg). Käesolevas töös uuritud kromosoomide pikkused, geenide ja SNP-ide arvud. (http://feb2006.archive.ensembl.org/Homo_sapiens/index.html). Teadaolevaid valke kodeerivad geenid on geenid, mis on kaardistatud ning mille valguline produkt on teada ning on kantud avalikesse andmebaasidesse. Ennustatud valke kodeerivad geenid on geenid, mida ei ole võimalik täieliku kindlusega kromosoomile kaardistada.

Kromosoom	Pikkus (Mb)	Teadaolevaid valke kodeerivad geenid	Ennustatud valke kodeerivad geenid	Pseudogeenid	miRNA	rRNA	snRNA	snoRNA	muud RNA-d	SNPs
19	63,8	1365	83	45	24	6	16	17	18	208635
20	62,4	583	36	40	7	8	33	23	34	290207
21	46,9	247	20	31	8	3	10	5	6	149536
22	49,6	483	59	32	10	2	20	12	20	186404

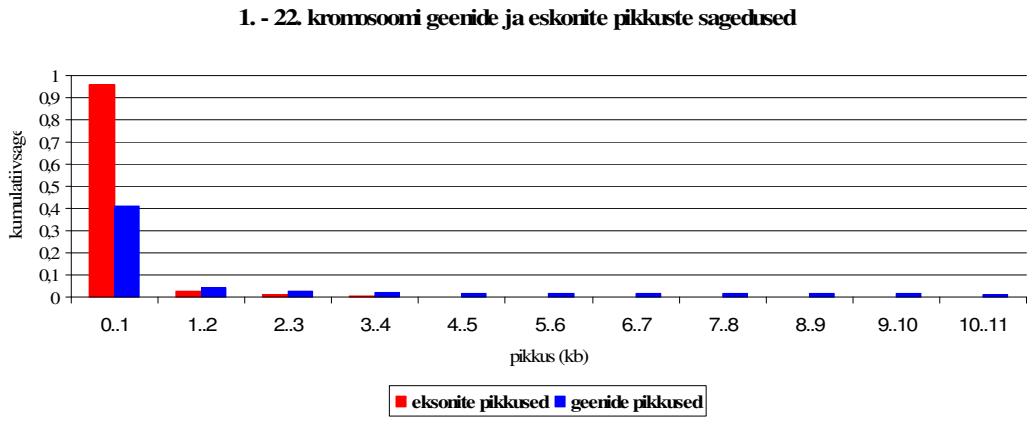
Geenide ja eksonite arv ja nende suhteline tihedus miljoni aluspaari kohta varieerub kromosoomide vahel. Geenide arv ja geenide tihedus pole otseselt seotud kromosoomi pikkusega. Kõige rohkem geene ja eksoneid leidub 1. ja kõige vähem 21. kromosoomil (tabel 3). Kuigi 1. kromosoom on kõige pikem ning 21. kromosoom kõige lühem, pole 1. kromosoomi geenide tihedus kõige suurem ning 21. kromosoomi geenide tihedus kõige väiksem. Kõrge suhtelise tihedusega on 19., 17. ja 22. kromosoom ning madala tihedusega 13., 18. ja 4. kromosoomil (tabel 4). Geenide järjestused moodustavad kõige suurema osa kromosoomi järjestusest 17., 22. ja 19. kromosoomil ning kõige väiksema osa 4., 5. ja 8. kromosoomil (tabel 4). Eksonid moodustavad kromosoomi üldjärjestusest kõige suurema osa 19. 17. ja 22. kromosoomil ning kõige väiksema osa 13., 4., ja 18. kromosoomil (tabel 4).

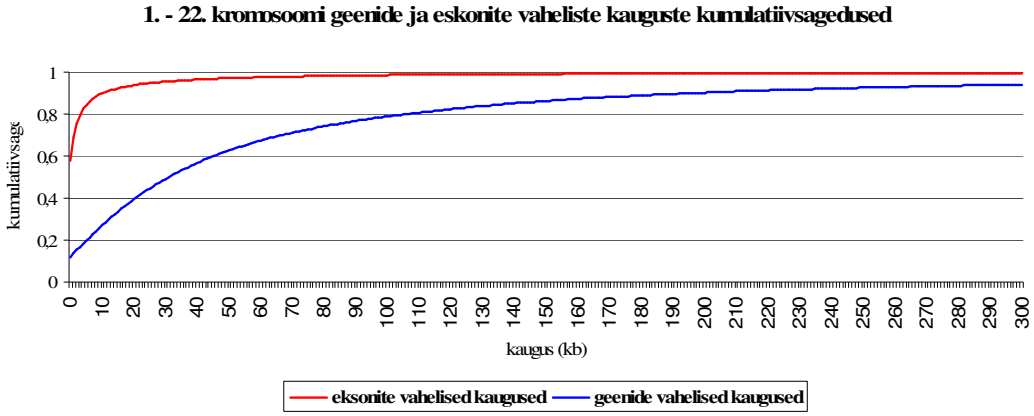
Tabel 4. Geenide ja eksonite katvus ja tihedus miljoni aluspaari kohta. Katvuse saamiseks jagasime kromosoomi aluspaaride arvu geene või eksoneid moodustavate aluspaaride arvuga. Geenide ja eksonite arvu miljoni aluspaari kohta saime kromosoomi kromosoomil olevate geenide või eksonite arvu jagamisel kromosoomi pikkusega.

kromosoom	Geenid		Eksonid	
	katvus	Geene 1 Mb kohta	katvus	Eksoneid 1 Mb kohta
1	0,40	11,46	0,04	146,24
2	0,37	7,72	0,03	106,16
3	0,39	7,61	0,02	102,27
4	0,32	5,99	0,02	66,48
5	0,32	7,04	0,02	80,35
6	0,37	8,61	0,03	105,92
7	0,44	8,85	0,03	107,98
8	0,32	6,82	0,02	77,33
9	0,33	8,16	0,03	107,15
10	0,44	7,84	0,03	122,21
11	0,40	12,59	0,04	140,61
12	0,40	10,42	0,03	138,19
13	0,34	4,81	0,02	54,93
14	0,41	9,27	0,03	110,89
15	0,45	10,16	0,03	120,19
16	0,40	12,45	0,04	162,96
17	0,54	18,59	0,06	255,98
18	0,33	5,86	0,02	70,37
19	0,50	24,67	0,07	301,13
20	0,39	12,24	0,04	166,67
21	0,33	7,04	0,03	85,46
22	0,52	12,86	0,06	160,08

Üle 95% eksonitest on alla tuhande aluspaari pikad, samal ajal moodustavad alla tuhande aluspaari pikkused geenid 41,2% kõikidest geenidest. 95% kõikidest geenidest jäävad alla 150 kb. Geenide vahelised kaugused varieeruvad palju. Kõige sagedasem geenide vaheline kaugus jääb alla 1 kb, kuid see moodusta vaid 11,8% kõikidest kaugustest. Eksonid asuvad kromosoomis kontsentreeritumalt. 95% eksonitest pole

üksteisest kaugemal kui 26 kb. Geenide ja eksonite pikkuste ja nende vaheliste kauguste jaotused on välja toodud joonisel 4.





Joonis 4. Geenide ja eksonite pikkuste ning geenide vaheliste ja eksonite vaheliste kauguste jaotused. Sagedused on saadud vastava pikkusega geenide või eksonite arvu jagamisel kõikide geenide või eksonite arvuga. Kumulatiivsagedus saadakse vastava pikkusega geeni või eksoni sageduse ja kõikide temast lühemate geenide või eksonite sageduste summaga.

2.3 Haplotüübi blokkide, geenide ja eksonite koordinaatide võrdlus

Töös võrdlesime haplotüübi blokkide koordinaate geenide ja eksonite koordinaatidega ning juhuslikult kromosoomile asetatud haplotüübi blokkide koordinaate geenide ja eksonite koordinaatidega. Saadud tegelike ja juhuslike blokkide tulemusi võrdlesime

Fisheri täpse testi abil. Statistiliselt oluliseks lugesime tulemust, mille p-väärtus oli väiksem kui 0,001. HapMap andmebaasist a (vt. punkt II 1.1) pärinevate genotüpiseerimise andmete põhjal arvutatud haplotüübi blokkidega läbi viidud katse tulemused on tabelis 5.

Tabel 5. Tegelike haplotüübi blokkide ja geenide/eksonite koordinaatide ning juhuslike haplotüübi blokkide ja geenide/eksonite koordinaatide võrdlustulemuste suhe. Suhte väärtus üks viitab, et geene ja eksoneid asub tegelikes ja eeldatavates haplotüübi blokkides võrdselt. Ühest väiksema väärtuse korral on geene ja eksoneid tegelikes haplotüübi blokkides vähem kui eeldatavates blokkides. Ühest suurema väärtuse korral on geene ja eksoneid tegelikes haplotüübi blokkides rohkem kui eeldatavates blokkides. CEU tähistab CEPH populatsiooni, YRI Yoruba populatsiooni. GAB tähistab LD-l põhinevat blokkide arvutamise meetodit ning GAM nelja gameedi reegli meetodit. Sinise ja punase kirjaga on välja toodud statistiliselt olulised erinevused. Statistiliselt oluliseks loeme p-väärtusi, mis on väiksemad kui 0,001. Sinine tähistab olukorda, kus geenide või eksonite arv tegelikes haplotüübi blokkides on suurem kui eeldatavates blokkides. Punane tähistab olukorda, kus geene või eksoneid on tegelikes haplotüübi blokkides vähem kui eeldatavates blokkides.

kromosoom	Eksonid				Geenid			
	CEU		YRI		CEU		YRI	
	GAB	GAM	GAB	GAM	GAB	GAM	GAB	GAM
1	1,00	1,01	1,01	**1,02**	0,98	1,01	**0,91**	1,01
2	**1,02**	**1,01**	1,00	**1,03**	1,00	1,07	1,00	1,04
3	**0,97**	1,01	1,02	**1,02**	**1,12**	1,04	1,09	1,08
4	1,00	1,01	1,02	**1,03**	0,99	0,96	0,96	0,98
5	0,99	1,01	0,99	1,00	1,06	1,03	1,00	1,01
6	1,00	**1,02**	**1,06**	**1,05**	1,00	1,04	1,02	1,02
7	**0,92**	1,01	**0,93**	**1,02**	0,98	1,02	1,02	1,01
8	**0,96**	0,98	**0,93**	0,99	0,99	0,98	0,97	0,97
9	1,00	0,99	1,00	1,02	1,05	0,96	0,92	0,96
10	**0,96**	1,01	1,00	**1,05**	1,01	0,99	1,10	1,08
11	**0,97**	0,99	**0,96**	1,01	1,02	1,00	0,97	1,04
12	0,99	**1,02**	0,98	1,01	1,03	1,04	0,94	1,01

Tabel 5. (jätk). Tegelike haplotüübi blokkide ja geenide/eksonite koordinaatide ning juhuslike haplotüübi blokkide ja geenide/eksonite koordinaatide võrdlustulemuste suhe.

kromosoom	Eksonid				Geenid			
	CEU		YRI		CEU		YRI	
	GAB	GAM	GAB	GAM	GAB	GAM	GAB	GAM
13	**1,03**	**1,03**	**1,04**	**1,05**	1,10	1,01	0,96	1,01
14	**0,97**	1,01	0,97	1,01	0,97	0,96	0,91	1,02
15	**1,10**	**1,05**	**1,05**	**1,04**	1,03	1,03	1,02	0,98
16	0,98	1,00	1,00	1,00	1,02	1,06	1,10	1,07
17	0,99	**1,02**	0,99	1,02	1,00	0,98	0,96	0,96
18	1,01	1,02	1,03	**1,04**	1,06	1,02	0,97	0,95
19	**0,95**	0,99	**0,91**	**0,97**	**0,83**	0,92	**0,78**	0,89
20	1,00	0,99	**0,95**	0,99	1,05	1,02	1,00	1,04
21	**0,96**	1,01	0,95	1,00	1,05	0,97	1,05	1,08
22	**1,07**	**1,05**	1,04	1,02	0,99	0,96	1,04	1,02
Summaarne	**0,991**	**1,010**	**0,993**	**1,018**	1,009	1,009	0,980	1,011

Katse tulemused, mis on läbi viidud HapMap b ja c andmetelt (vt. punkt II 1.1) leitud haplotüübi blokkidega ja ahne ja dünaamilise programmeerimise algoritmi abil arvutatud haplotüübi blokkidega, toetasid tabelis 5 olevaid tulemusi.

ARUTELU

Blokkide arvutamiseks vajalikud genotüpiseerimise andmed pärinevad HapMap projektist. Uuritud on kahte populatsiooni: Nigeeria päritolu populatsiooni Yoruba ja Lääne- ja Põhja Euroopa päritolu populatsioon CEPH. Mõlemast populatsioonist on vaatluse all 22 kromosoomi. Lisaks on töös kasutatud 21. kromosoomi haplotüübi blokke, mis pärinevad 21. kromosoomi haplotüübi blokkide andmebaasist (http://genome.perlegen.com/haplotype/). HapMap andmete põhjal arvutasime haplotüübi blokid kahel meetodil: LD-l põhinev meetodil (vt. punkt 1.3.1.2) ja nelja gameedi reegli meetodil (vt. punkt 1.3.1.4.). Haplotüübi blokkide pikkuste jaotusest viitas, et Euroopa päritolu haplotüübi blokid olid pikemad Aafrika päritolu blokkidest, mis on kooskõlas eelnevate uuringutega (Gabriel *et al.* 2002). Lisaks on töös kasutatud dünaamilise programmeerimise algoritmiga(vt. punkt 1.3.1.1) arvutatud 19. ja 21. kromosoomi blokke ning ahne algoritmiga (vt. punkt 1.2.1.2) 21. kromosoomi blokke.

LD-l põhineval meetodil, nelja gameedi reegli meetodil ning madalal haplotüüpide mitmekesisusel põhinevate algoritmidega leitud haplotüübi blokkide koordinaatide ja geeni koordinaatide võrdlemisel ei saa väita, et haplotüübi blokkide ja geenide asukoht genoomis oleksid omavahel seotud. Geenide koordinaadid asuvad võrdselt juhuslikult genereeritud blokkide koordinaatide ja tegelike blokkide koordinaatide vahel. Erandiks on 19. ja 1. kromosoom, mille puhul geenid paiknesid harvemini blokkide sees ja 3. kromosoomi, mille puhul geenid paiknesid sagedamini haplotüübi blokkide sees. Edasises uuringus võiks vaadelda kitsamaid kromosoomi piirkondi, et selgitada 1., 3. ja 19. kromosoomi ja erinevust ülejäänud kromosoomidest.

Eksonite ja LD-l põhineva Gabrieli meetodiga leidud haplotüübi blokkide võrdlemisel olid eksonid harvemini blokkide sees kui võrdluses eeldatavate blokkiga. Nelja gameedi reegli meetodi blokkide ja eksonite võrdlusel olid eksonid eelistatult blokkide sees. Kuigi tulemused olid mõlema meetodi puhul statistiliselt olulised, on raske teha antud tulemustest bioloogilist sisu omavaid järeldusi, sest ühe meetodi puhul on eksonid eelistatult blokkide sees, teise puhul blokkidest väljas või blokkide piiriga ülekattes. Samuti on suhteline erinevus tegelike ja juhuslike tulemuste vahel madal, jäädes alla 2 protsendi. Statistiline olulisus tekib sellise väikse erinevuse puhul tänu suurele eksonite koguarvule.

Edasistesse uuringutes võiks uurida kindlat tüübi geene ning kindlate omadustega geene ja eksoneid. Samuti võiks lisada uuringusse erinevatest andmebaasidest usaldusväärsemad käsitsi annoteeritud geenid.

KOKKUVÕTE

Käesoleva töö eesmärgiks oli geeni ja haplotüübi bloki koordinaatide ja eksonite ja haplotüübi bloki koordinaatide võrdlemisel selgitada, kas geenid ja eksonid asuvad genoomis eelistatult haplotüübi blokkide sees, neist väljas või juhuslikult. Töös võrdlesime erinevate populatsioonide ning erinevate bloki arvutamise algoritmi poolt leitud haplotüübi blokkide koordinaate geenide ja eksonite koordinaatidega. Antud lähteandmetele tuginedes ei saa järeldada haplotüübi blokkide ja geenide koordinaatide vahelist korrelatsiooni. Eksonite ja haplotüübi blokkide koordinaatide võrdlusel leidsime, et erinevate algoritmide abil arvutatud blokkide puhul paiknevad eksonid eelistatult kas blokkide sees või blokkidest väljas. Kuigi tulemused olid statistiliselt olulised, on raske antud tulemustest bioloogilisi järeldusi teha, sest suhteline erinevus tegelike ja eeldatavate tulemuste vahel oli madal, jäädes alla 2 protsendi.

SUMMARY

The aim of current paper was to examine if genes and exons reside preferably inside haplotype blocks, outside haplotype blocks or randomly in genome. We calculated haplotype blocks using different population data and different haplotype block partitioning algorithms. The comparison of gene coordinates and haplotype block coordinates showed no preference of gene residing. The comparison of exon and haplotype block coordinates showed eristic results. Using different haplotype partitioning algorithms the exons reside preferably inside or outside haplotype blocks. Although the differences were statistically significance, it is hard to make biological conclusions, because the difference between observed and expected results were less than 2%.

VIITED

Ashurst, J. L. and J. E. Collins (2003). "Gene annotation: prediction and testing." Annu Rev Genomics Hum Genet **4**: 69-88.

Birney, E. and R. Durbin (2000). "Using GeneWise in the Drosophila annotation experiment." Genome Res **10**(4): 547-8.

Cardon, L. R. and G. R. Abecasis (2003). "Using haplotype blocks to map human complex trait loci." Trends Genet **19**(3): 135-40.

Carlson, C. S., M. A. Eberle, et al. (2004). "Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium." Am J Hum Genet **74**(1): 106-20.

Daly, M. J., J. D. Rioux, et al. (2001). "High-resolution haplotype structure in the human genome." Nat Genet **29**(2): 229-32.

Dausset, J., H. Cann, et al. (1990). "Centre d'etude du polymorphisme humain (CEPH): collaborative genetic mapping of the human genome." Genomics **6**(3): 575-7.

Dawson, E., G. R. Abecasis, et al. (2002). "A first-generation linkage disequilibrium map of human chromosome 22." Nature **418**(6897): 544-8.

Florea, L., G. Hartzell, et al. (1998). "A computer program for aligning a cDNA sequence with a genomic DNA sequence." Genome Res **8**(9): 967-74.

Franklin, I. and R. C. Lewontin (1970). "Is the gene the unit of selection?" Genetics **65**(4): 707-34.

Gabriel, S. B., S. F. Schaffner, et al. (2002). "The structure of haplotype blocks in the human genome." Science **296**(5576): 2225-9.

Gish, W. and D. J. States (1993). "Identification of protein coding regions by database similarity search." Nat Genet **3**(3): 266-72.

Hedrick, P. W. (1987). "Gametic disequilibrium measures: proceed with caution." Genetics **117**(2): 331-41.

Johnson, G. C., L. Esposito, et al. (2001). "Haplotype tagging for the identification of common disease genes." Nat Genet **29**(2): 233-7.

Kent, W. J. and A. M. Zahler (2000). "Conservation, regulation, synteny, and introns in a large-scale C. briggsae-C. elegans genomic alignment." Genome Res **10**(8): 1115-25.

Li, N. and M. Stephens (2003). "Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data." Genetics **165**(4): 2213-33.

Patil, N., A. J. Berno, et al. (2001). "Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21." Science **294**(5547): 1719-23.

Pennacchio, L. A. and E. M. Rubin (2001). "Genomic strategies to identify mammalian regulatory sequences." Nat Rev Genet **2**(2): 100-9.

Phillips, M. S., R. Lawrence, et al. (2003). "Chromosome-wide distribution of haplotype blocks and the role of recombination hot spots." Nat Genet **33**(3): 382-7.

Robberson, B. L., G. J. Cote, et al. (1990). "Exon definition may facilitate splice site selection in RNAs with multiple exons." Mol Cell Biol **10**(1): 84-94.

Senapathy, P., M. B. Shapiro, et al. (1990). "Splice junctions, branch point sites, and exons: sequence statistics, identification, and applications to genome project." Methods Enzymol **183**: 252-78.

Strausberg, R. L., E. A. Feingold, et al. (1999). "The mammalian gene collection." Science **286**(5439): 455-7.

Zhang, K., M. Deng, et al. (2002). "A dynamic programming algorithm for haplotype block partitioning." Proc Natl Acad Sci U S A **99**(11): 7335-9.

Zhang, K. and L. Jin (2003). "HaploBlockFinder: haplotype block analyses." Bioinformatics **19**(10): 1300-1.

Zhang, K., Z. Qin, et al. (2005). "HapBlock: haplotype block partitioning and tag SNP selection software using a set of dynamic programming algorithms." Bioinformatics **21**(1): 131-4.

Zhang, K., Z. S. Qin, et al. (2004). "Haplotype block partitioning and tag SNP selection using genotype data and their applications to association studies." Genome Res **14**(5): 908-16.

Zhang, M. Q. (1998). "Statistical features of human exons and their flanking regions." Hum Mol Genet **7**(5): 919-32.

Zhang, M. Q. (2002). "Computational prediction of eukaryotic protein-coding genes." Nat Rev Genet **3**(9): 698-709.

Zondervan, K. T. and L. R. Cardon (2004). "The complex interplay among factors that influence allelic association." Nat Rev Genet **5**(2): 89-100.

Tabaska, J. E. and M. Q. Zhang (1999). "Detection of polyadenylation signals in human DNA sequences." Gene **231**(1-2): 77-86.

The International HapMap Consortium (2003). "The International HapMap Project." Nature **426**(6968): 789-96.

Wall, J. D. and J. K. Pritchard (2003). "Haplotype blocks and linkage disequilibrium in the human genome." Nat Rev Genet **4**(8): 587-97.

Wang, N., J. M. Akey, et al. (2002). "Distribution of recombination crossovers and the origin of haplotype blocks: the interplay of population history, recombination, and mutation." Am J Hum Genet **71**(5): 1227-34.