TARTU ÜLIKOOL
MATEMAATIKA-INFORMAATIKATEADUSKOND
Matemaatilise statistika instituut

Tatjana Iljašenko

Geneetiliste markerite imputeerimine

Magistritöö

Juhendaja: Märt Möls

TARTU 2013

Sisukord

Sissejuhatus

Inimeste ülegenoomsed uuringud omavad suurt tähtsust tänapäevases teaduses, aidates mõista, kuidas geneetiline informatsioon ja selle muutused mõjutavad üksikisikuid, nende arengut, vananemist, heaolu ja haigusi, aga ka inimeste käitumist ja psühholoogiat seoses muutustega, mis on tingitud neid ümbritsevast keskkonnast ja inimeste elustiilist.

On teada, et kahe erineva inimindiviidi DNA on täiesti identne enam, kui 99% ulatuses [1]. Seega, on inimindiviidi geneetiline unikaalsus tingitud vähem kui 1% DNA järjestuse varieeruvusest ning just taolised "varieeruvad" DNA osad, mida on kombeks nimetada geneetilisteks markeriteks, pakuvad huvi geneetilises analüüsis.

Eksisteerivad erinevad meetodid geneetiliste andmete kogumiseks ehk genotüpiseerimiseks ning huvialuste DNA piirkondade detekteerimiseks ehk määramiseks, kuid enamus nendest suudavad määrata vaid osa huvipakkuvate markerite väärtuse.

Sel juhul kasutatakse erinevaid imputeerimismeetodeid, mis võimaldavad genotüpiseerimata jäänud huvipakkuvate geneetiliste markerite ennustamist, seega ka uuringu võimsuse tõstmist läbi analüüsitavate DNA piirkondade arvu suurendamise.

Üheks levinuks imputeerimisinstrumendiks on spetsiaalne tarkvara/programm IMPUTE2 [2], mille arvutusalgoritm põhineb varjatud Markovi ahelate algoritmil. Antud tarkvara võimaldab kasutada ja kombineerida referentspaneelidena (vt. 1.2 Referentspaneelid. lk. 5) erinevaid haplotüüpide referentse, näiteks 1000 Genome Project, HapMap2 ja HapMap3 [2].

Käesoleva töö põhieesmärgiks on kontrollida IMPUTE2 programmi abil imputeeritud geneetiliste markerite kvaliteeti ja programmi poolt väljastatavate kvaliteedihinnangute kvaliteeti.

Eesmärgi saavutamiseks teostatakse kolm erinevat imputeerimisprotsessi, milledest esimene viiakse läbi nö ideaaltingimustes. Võetakse juhuslik valim 1000 Genoomi Projekti haplotüüpide seast. Valimisse sattunud haplotüüpidest eemaldatakse osa geneetiliste markerite väärtustest. Seejärel imputeeritakse puuduole-

vate markerite väärtused, kasutades referentspaneelina esialgse koguandmestiku ilma valimisse sattunud haplotüüpideta.

Teise imputeerimise eesmärk seisneb selles, et eurooplaste referenshaplotüüpe kasutades (see ongi 1000 Genoomi Projekti raames kogutud andmed), imputeerida genotüpiseeritud eestlaste andmetes puuduolevate markerite väärtusi. Selleks kasutatakse samuti 1000 Genoomi Projekti raames kogutud andmeid referentspaneelina, kuid seekord jäetakse välja teatud hulk geneetilisi markereid, et viia referentspaneelina kasutatava andmestiku vastavusse eestlaste genotüüpiseeritud andmetega markerite nimekirja suhtes. Valim moodustatakse seekord eestlaste genotüpiseeritud andmetest, korjates sealt välja hulk teatud markereid. Seejärel imputeeritakse ettevalmistatud eurooplaste haplotüüpide andmestiku abil väljakorjatud markerite väärtusi tagasi.

Kolmanda imputeerimise ülesandeks on ennustada eestlaste geeniandmeid, kasutades referentspaneelina eestlaste genotüpiseeritud andmeid, mis koosnevad 49 indiviidi sekveneeritud andmetest ehk kindlaks määratud DNA molekulide aminohapete ja nukleotiitide järjestusest [3]. Valimi moodustamisel valitakse juhuslikult 15 indiviidi andmed 49 indiviidi andmete seast, kust jäetakse välja hulk teatud markereid. Referentspaneeli jääb 34 indiviidi. Eestlaste andmed on saadud Eesti Geenivaramust.

Töö esimeses ja teises peatükkides antakse detailne ülevaade imputeerimisprotsessist ja sellega seotud mõistetest ning programmis IMPUTE2 kasutatud meetodist. Kolmandas peatükis analüüsitakse imputeerimistulemusi ja nende kvaliteeti. Neljas peatükis tutvustatakse imputeerimise kvaliteedihinnangu analüüsil kasutatud meetodeid ning analüüsitakse programmi IMPUTE2 poolt raporteeritute kvaliteedihinnangute usaldusväärsust.

1 Andmetest

1.1 SNP, genotüüp, haplotüüp

Inimindiviidide geneetiline varieeruvus on tingitud vähem kui 1 % DNA järjestuse varieeruvusest. Neid varieeruvaid DNA piirkondi, mis pakuvad suurt huvi geneetilises analüüsis, nimetatakse $geneetilisteks\ polümorfismideks$ ehk erinevate indiviidide geenide ja geenidevaheliste alade teatud järjestuste erinevusteks [4].

Polümorfismide roll geneetilises analüüsis on kõigepealt seotud geneetiliste andmete kogumisega ehk inimindiviidide genotüpiseerimisega.

Eristatakse mitu erinevat polümorfismide klassi, milliste seas on kõige tavalisem ja sagedasem (ligikaudu 90 % inimese genoomi variatsioonidest) - ühe nukleotiidi (A, T, C või G) muutumine genoomis, mida nimetatakse SNP-ks (ingl. $Single\ Nucleotide\ Polymorphism$) [1].

SNP-i võimalikeks variantideks on $alleelid$ [5] ning enamusel SNP-idel on ainult kaks alleeli, millepärast nimetatakse neid vahel ka binaarseteks markeriteks. Kõikide alleelide sagedused annavad kokku 100%. Alleeli mõistet kasutades saab defineerida ka genotüübi ning haplotüübi.

Definitsioon 1.
Segu mõlema kromosoomi alleelidest nimetatakse genotüübiks. SNP-ide puhul on tavaliselt 3 võimalikku genotüübi: 11, 12, 22 (kui tähistame ühe alleeli 1-ga ja teise - 2-ga) [5].

Definitsioon 2.
Haplotüübiks nimetatakse ühel kromosoomil järjestikku esinevad alleelid [5].

Võrreldes teiste klasside polümorfismidega (ehk teiste geneetiliste markeritega), paiknevad SNP-d genoomis suhteliselt tihedalt (üks SNP 100 kuni 300 aluspaari DNA kohta,kusjuures inimese genoom koosneb hinnanguliselt umbes 3 miljardist nukleotiidipaarist [1]). Samuti asuvad nad genoomis nii valku kodeerivatel, regulatoorsetel kui ka teadaolevat funktsiooni mitteomavatel aladel ehk erinevates

huvipakkuvates genoomi piirkondades.

Lisaks sellele on SNP-e võimalik suhteliselt odavalt ja täpselt detekteerida (määrata/avastada) geenikiibi tehnoloogiat kasutades ning nende mutatsioonikiirus on suhteliselt madal, mis teeb SNP-e stabiilseks uurimismaterjaliks [4].

Kõike ülalmainitud arvestades, pole üllatav, et enamikes geneetilistes assotsiatsiooniuuringutes kasutatakse just SNP-e inimeste geneetilise muutlikuse kirjaldamiseks.

1.2 Referentspaneelid

Eksisteerivad erinevad meetodid geneetiliste markerite imputeerimiseks. Enamus neist meetoditest kasutab imputeerimiseks referentspaneeli abi. Sellised meetodid eeldavad, et referentsandmestik ja uuritav valim on pärit samast populatsioonist.
Referentspaneel kujutab endast tihedat hulka haplotüüpiseeritud SNP-e. Tänapäeval on enamkasutatavateks referentspaneelideks 1000 Genoomi Projekti raames koostatud andmestikud: HapMap2, HapMap3, 1000 Genomes Pilot, 1000 Genomes Phase I (interim ja Integrated versioonid) referentshaplotüüpe [2]. Nimetatud paneelid erinevad üksteisest ehk sisaldavad erinevaid SNP-e ja baseeruvad erinevatel inimpopulatsioonidel. Näiteks, HapMap3 sisaldab vähem SNP-e, kui HapMap2, kuid HapMap3 sobib väidetavalt paremini haruldaste SNP-ide imputeerimiseks, kui HapMap2 [6].

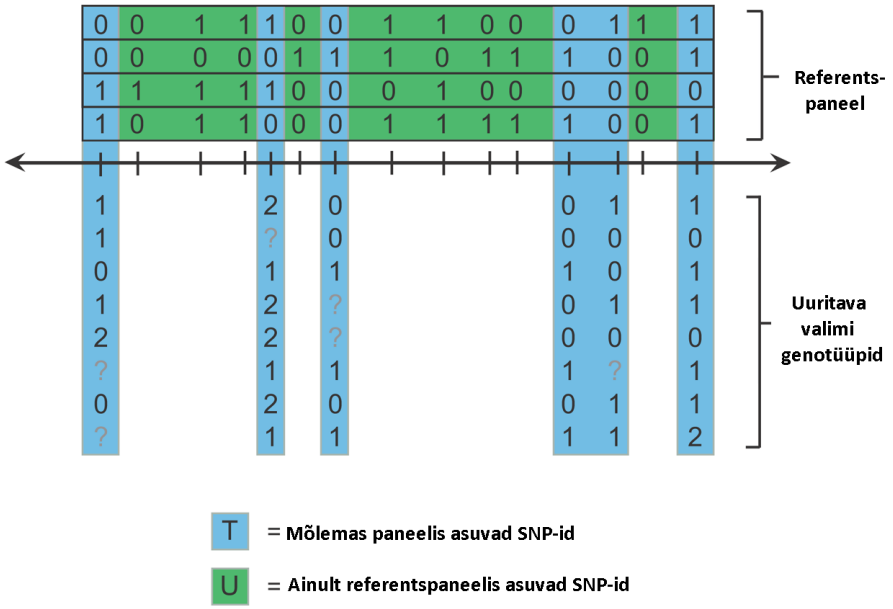
Mõnedes imputeerimisstsenaariumides kasutatakse kombineeritud referentspaneele, näiteks HapMap 3 + 1,000 Genoomi pilootuuringu haplotüüpe.
Kõiki nimetatud referentspaneele perioodiliselt uuendatakse ning täiendatakse.
Käesolevas töös on kasutatud referentshaplotüüpidena “1000 Genomes Phase I Integrated” referentshaplotüüpe ning Eesti Geenivaramu poolt saadud eestlaste sekveneeritud andmed.

2 Imputeerimine

2.1 Imputeerimise idee

Nagu mainiti üleval, jääb hulk SNP-e genotüüpiseerimise käigus identifitseerimata, kuid antud probleemi üritatakse lahendada imputeerimise abil.
Geneetiline imputeerimine on protsess, mille käigus ennustatakse genotüpiseerimata jäänud SNP-d. Antud töös analüüsitakse geneetiliste markerite imputeerimisprogrammi IMPUTE2 abil. Antud meetodi lihtsama stsenaariumi põhiideed saab esitada alljärgneva skeemi abil [7].

Joonis 1: Ühe referentspaneeliga imputeerimisskeem.



Joonisel 1 referentshaplotüübid esitatud 0 ja 1 sisalduva hulgana, kus 0-ga ja 1 -ga tähistatud SNP-i alternatiivsed alleelid. Uuritava valimi genotüübid on tähistatud numbritega: 0 ja 2 - homosügootsed SNP-id (koosnevad kahest ühesugustest alleelidest, ehk omavad kuju 11 või 22 (vt. 1.1 SNP, genotüüp, haplotüüp. Definitsioon 1., lk. 4) ja 1 - heterosügootsed SNP-id (need, mis koosnevad erinevatest alleelidest, ehk omavad kuju 12 (vt. 1.1 SNP, genotüüp, haplotüüp. Definitsioon 1., lk. 4). Puuduvad andmed tähistatud ?-ga.

Referentspaneeli ridadeks on haplotüübid ja veergudeks - SNP-id. Uuritava valimi paneeli veergudes on SNP-id ja ridades - genotüübid.

Nagu skeemilt on näha, sisaldab uuritava valimi paneel vähem geneetilisi markereid (ainult hulga T SNP-id), samal ajal referentspaneelil on nii uuritava valimi SNP-id, kui ka täiendav hulk SNP-e (hulk U). Imputeerimise eesmärk - hinnata hulga U SNP-ide genotüüpe uuritavas valimis.

Suurem osa imputeerimismeetodeid eeldab kõigepealt uuritava valimi SNP-ide haplotüpiseerimist, kasutades teadaolevaid genotüüpe. Saadud haplotüüpe võrreldakse seejärel referentshaplotüüpidega. Eeldatakse, et kui haplotüüpide mustrid täielikult või peaaegu langevad kokku hulgas T, langevad nad ka kokku hulgas U. Selles seisnebki imputeerimise põhiidee.

Siinjuures peab märkima, et olulisim nendest kahest etapist (siin mõeldakse haplotüpiseerimise ja imputeerimise etapid) on uuritava valemi teadaolevate SNP-ide haplotüpiseerimine, mis teostatakse varjatud Markovi mudelite abil. Imputeerimisetapp teostatakse suhteliselt kiiresti, arvestades, et esimesel etapil hinnatud haplotüüpid on õiged.

Ülaltoodust järelduvad mõned tähtsad märkused imputeerimistäpsuse kohta:

1. Imputeerimise täpsus suures osas sõltub uuritava valimi haplotüpiseerimisest.

2. Uuritava valimi puuduvate andmete arvutamine - kallis ja keeruline protsess, mis lisab imputeerimisele ebatäpsusi.

3. Sageli kasutatakse haplotüpiseerimisel ainult referentspaneelis sisalduvat informatsiooni, mis tähendab, et haplotüpiseerimise täpsus ei sõltu uuritava valimi mahust.

2.2 Meetodi kirjeldus

Antud alampeatükk baseerub J. Marchini, B. Howie, S. Myers, G. McVean, P. Donnelly artiklil [8].

Et vaadata imputeerimismehhanismi detailsemalt, eeldame, et meil on L dialleelseid SNP-e, kus alternatiivsed alleelid kodeeritud 0 ja 1 ning võimalikute teadaolevate genotüüpide variandid - 0, 1 ja 2, kusjuures 0-ga ja 2-ga tähistatud homosügootsed SNP-id ja 1-ga - heterosügootsed SNP-id. Olgu meil N referentshaplotüüpi ja K indiviidi uuritavas valimis. Tähistame haplotüüpide hulka H-ga ning genotüüpide hulka G-ga, nii et $H_n = (H_{n1}, H_{n2}, ..., H_{nL})$, $n = 1, ..., N$ ja $G_k = (G_{k1}, G_{k2}, ..., G_{kL})$, $k = 1, ..., K$.

Nagu üleval mainiti, imputeerimisprotsessi põhiline ülesanne seisneb uuritava valimi puuduvate SNP-ide genotüüpide ennustamises, kusjuures olulisemaks etapiks on valimi teadaolevate SNP-ide haplotüpiseerimine. Antud ülesande lahendamiseks kasutatakse uuritava valimi iga indiviidi genotüübi G_k jaoks varjatud Markovi mudeli:

$$P(G_k \mid H) = \sum_{Z_i^{(1)}, Z_i^{(2)}} P(G_k \mid Z_i^{(1)}, Z_i^{(2)}, H) P(Z_i^{(1)}, Z_i^{(2)} \mid H). \tag{1}$$

kus $Z_i^{(1)} = \{Z_{i1}^{(1)}, Z_{i2}^{(1)}, ..., Z_{iL}^{(1)}\}$ ja $Z_i^{(2)} = \{Z_{i1}^{(2)}, Z_{i2}^{(2)}, ..., Z_{iL}^{(2)}\}$ on kaks pikkusega L varjatud seisundite jadat/ahelat ning $Z_{il}^{(j)} \in \{1, ..., N\}$, $j = 1, 2$, $i = 1, ...M$, $l = 1, ..., L$, kusjuures $M = (N^2)^L$ tähistab kõigi võimalike varjatud seisundite jadade arvu.
Nõnda, lookuses l asuva markeri korral $1 \leq l \leq L$ võib mõelda ahela varjatud seisundi all hulga H haplotüübipaari $(Z_{il}^{(1)}, Z_{il}^{(2)})$ saamist antud lookuses, mille abil moodustatakse l-nda SNP-i k-s genotüüpi, $1 \leq k \leq K$.
Varjatud seisundite eeljaotust, mis kirjeldab seisundite muutumist mööda ahelat, esitab valemi (1) osa: $P(Z_i^{(1)}, Z_i^{(2)} \mid H)$:

$$P(Z_i^{(1)}, Z_i^{(2)} \mid H) = P(Z_{i1}^{(1)}, Z_{i1}^{(2)} \mid H) \prod_{l=1}^{L-1} P(\{Z_{il}^{(1)}, Z_{il}^{(2)}\} \to \{Z_{i(l+1)}^{(1)}, Z_{i(l+1)}^{(2)}\} \mid H), \tag{2}$$

kus seisundite algjaotuseks on $P(Z_{i1}^{(1)}, Z_{i1}^{(2)} \mid H) = \frac{1}{N^2}$ ning üleminekumaatriksi A elemendiks on $P(\{Z_{il}^{(1)}, Z_{il}^{(2)}\} \to \{Z_{i(l+1)}^{(1)}, Z_{i(l+1)}^{(2)}\} \mid H)$. Sel juhul võtab

üleminekumaatrks A kuju:

$$A = \begin{cases} (e^{\frac{-p_l}{N}} + \frac{(1-e^{\frac{-p_l}{N}})}{N})^2, Z_{il}^{(1)} = Z_{i(l+1)}^{(1)}, Z_{il}^{(2)} = Z_{i(l+1)}^{(2)} \\ (e^{\frac{-p_l}{N}} + \frac{(1-e^{\frac{-p_l}{N}})}{N})(\frac{1-e^{\frac{-p_l}{N}}}{N}), Z_{il}^{(1)} = Z_{i(l+1)}^{(1)}, Z_{il}^{(2)} \neq Z_{i(l+1)}^{(2)}, \\ Z_{il}^{(1)} \neq Z_{i(l+1)}^{(1)}, Z_{il}^{(2)} = Z_{i(l+1)}^{(2)} \\ (\frac{1-e^{\frac{-p_l}{N}}}{N})^2, Z_{il}^{(1)} \neq Z_{i(l+1)}^{(1)}, Z_{il}^{(2)} \neq Z_{i(l+1)}^{(2)} \end{cases}$$

kus $p_l = 4N_e r_l$, kus omakorda r_l on l ja $l+1$ SNP-i geneetiline kaugus (mõõdetuna sentiMorganites) ühe generatsiooni kohta (andmed geneetilise kauguse kohta saadakse koos referentspaneelidega), $N_e = 11,418$ [9].

Valemi (1) osa $P(G_k \mid Z_i^{(1)}, Z_i^{(2)}, H)$ modeleerib, kui hästi uuritava valimi genotüübid langevad kokku moodustatud haplotüübidega, samal ajal imiteerides mutatsioonide effekti eeldusel, et mõõtmisvead või mutatsioonid toimuvad sõltumatult. Ühe alleelide paari muteerimistõenäosus on $\lambda = \frac{\theta}{2(\theta+N)}$, kus $\theta = (\sum_{n=1}^{N-1} \frac{1}{n})^{-1}$ [10].
Seega:

$$P(G_k \mid Z_i^{(1)}, Z_i^{(2)}, H) = \prod_{l=1}^{L} P(G_{kl} \mid Z_{il}^{(1)}, Z_{il}^{(2)}, H) = \prod_{l=1}^{L} P((H_{Z_{il^{(1)}}l} + H_{Z_{il^{(2)}}l}) \rightarrow G_{kl}), \tag{3}$$

kus $P((H_{Z_{il^{(1)}}l} + H_{Z_{il^{(2)}}l}) \rightarrow G_{kl})$ on tõenäosus, et positsioonis l haplotüüpe $H_{Z_{il^{(1)}}l}$ ja $H_{Z_{il^{(2)}}l}$ omaval k. indiviidil nähakse positsioonis l genotüüpi G_{kl}, ning arvutatakse järgmise tabeli abil:

Tabel 1: Muteerimistõenäosus

	$G_{kl} = 0$	$G_{kl} = 1$	$G_{kl} = 2$
$(H_{Z_{il^{(1)}}l} + H_{Z_{il^{(2)}}l}) = 0$	$(1-\lambda)^2$	$2\lambda(1-\lambda)$	λ^2
$(H_{Z_{il^{(1)}}l} + H_{Z_{il^{(2)}}l}) = 1$	$\lambda(1-\lambda)$	$\lambda^2 + (1-\lambda)^2$	$\lambda(1-\lambda)$
$(H_{Z_{il^{(1)}}l} + H_{Z_{il^{(2)}}l}) = 2$	λ^2	$2\lambda(1-\lambda)$	$(1-\lambda)^2$

Mudeli seisundi $(Z_i^{(1)}, Z_i^{(2)})$ tõenäosust etteantud genotüübi G_k tingimusel

(see ongi nn haplotüüpiseerimine) saame tuginedes tõenäosuste korrutamise reeglile ja kasutades valemiga (1) kirjeldatud tingliku tõenäosuse:

$$P(Z_i^{(1)}, Z_i^{(2)}, H \mid G_k) = \frac{P(G_k \mid Z_i^{(1)}, Z_i^{(2)}, H) P(Z_i^{(1)}, Z_i^{(2)} \mid H)}{P(G_k \mid H)}. \qquad (4)$$

Vaatame antud mudeli rakendamist näidisandmetel.

Näide 1. *Olgu H koosneb kahest haplotüübist ehk $N = 2$ ja uuritavas valimis G on 1 indiviid.*

	$SNP1$	$SNP2$	$SNP3$
$hapl.1$	0	1	1
$hapl.2$	1	1	0

	$SNP1$	$SNP2$	$SNP3$
$indiv.1$	1	2	1

Oletame, et vaadeldavate SNP-ide geneetilised kaugused (cM) on vastavalt $15.0,\ 15.05$ *ja* 15.1. *Sellest lähtuvalt saame arvutada parameetri p_l:*
$p_1 = 4 \cdot 11.418 \cdot (15.05 - 15.0)$
$p_2 = 4 \cdot 11.418 \cdot (15.1 - 15.05)$

Varjatud seisundite ruum koosneb kõikidest võimalikest haplotüübipaaridest antud lookuses, ehk antud näites $N = 2$ haplotüübi korral $N^2 = 4$ seisundist, milliseid tähistame: $a = 11,\ b = 12,\ c = 21,\ d = 22$. Sel juhul kõiki varjatud seisundite jadade ruum $\Omega = \{aaa, aab, aac, aba, abb, abc, ..., ccc\}$ sisaldab 64 elemendi (4^3). Valemi (2) alusel on seisundite algjaotuseks $\pi_0 = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$ ning üleminekumaatriks võtab üldkuju:

	a	b	c	d
a	$(e^{\frac{-p_l}{N}} + \frac{(1-e^{\frac{-p_l}{N}})}{N})^2$	$(e^{\frac{-p_l}{N}} + \frac{(1-e^{\frac{-p_l}{N}})}{N})(\frac{1-e^{\frac{-p_l}{N}}}{N})$	$(e^{\frac{-p_l}{N}} + \frac{(1-e^{\frac{-p_l}{N}})}{N})(\frac{1-e^{\frac{-p_l}{N}}}{N})$	$(\frac{1-e^{\frac{-p_l}{N}}}{N})^2$
b	$(e^{\frac{-p_l}{N}} + \frac{(1-e^{\frac{-p_l}{N}})}{N})(\frac{1-e^{\frac{-p_l}{N}}}{N})$	$(e^{\frac{-p_l}{N}} + \frac{(1-e^{\frac{-p_l}{N}})}{N})^2$	$(\frac{1-e^{\frac{-p_l}{N}}}{N})^2$	$(e^{\frac{-p_l}{N}} + \frac{(1-e^{\frac{-p_l}{N}})}{N})(\frac{1-e^{\frac{-p_l}{N}}}{N})$
c	$(e^{\frac{-p_l}{N}} + \frac{(1-e^{\frac{-p_l}{N}})}{N})(\frac{1-e^{\frac{-p_l}{N}}}{N})$	$(\frac{1-e^{\frac{-p_l}{N}}}{N})^2$	$(e^{\frac{-p_l}{N}} + \frac{(1-e^{\frac{-p_l}{N}})}{N})^2$	$(e^{\frac{-p_l}{N}} + \frac{(1-e^{\frac{-p_l}{N}})}{N})(\frac{1-e^{\frac{-p_l}{N}}}{N})$
d	$(\frac{1-e^{\frac{-p_l}{N}}}{N})^2$	$(e^{\frac{-p_l}{N}} + \frac{(1-e^{\frac{-p_l}{N}})}{N})(\frac{1-e^{\frac{-p_l}{N}}}{N})$	$(e^{\frac{-p_l}{N}} + \frac{(1-e^{\frac{-p_l}{N}})}{N})(\frac{1-e^{\frac{-p_l}{N}}}{N})$	$(e^{\frac{-p_l}{N}} + \frac{(1-e^{\frac{-p_l}{N}})}{N})^2$

Kasutades arvutatud parameetri p_l, hindame üleminekumaatriksi, mis antud näites ei muutu, sest $r_1 = r_2 = 0.05$, järelikult ka $p_1 = p_2 = 2.2836$:

	a	b	c	d
a	0.4351011	0.2245208	0.2245208	0.1158572
b	0.2245208	0.4351011	0.1158572	0.2245208
c	0.2245208	0.1158572	0.4351011	0.2245208
d	0.1158572	0.2245208	0.2245208	0.4351011

Esitame arvutustulemused joonistel 2 ja 3, kus igale võimalikule ahela seisundile vastab:

1. *Teises veerus asuv etteantud genotüübi tõenäosus $P(G_k \mid Z_i^{(1)}, Z_i^{(2)}, H)$ mudeli seisundite $(Z_i^{(1)}, Z_i^{(2)})$ tingimusel, mida arvutatud valemi (3) põhjal,*

2. *Kolmandas veerus asuv mudeli seisundite tõenäosus $P(Z_i^{(1)}, Z_i^{(2)} \mid H)$, arvutatud valemi (2) põhjal,*

3. *Neljandas veerus asuv tõenäosus $P(Z_i^{(1)}, Z_i^{(2)}, H \mid G_k)$ haplotüübi hindamiseks etteantud genotüübi G_k tingimusel, arvutatud valemi (4) põhjal, kus valemis (4) kasutatud k-nda indiviidi genotüübi G_k tõenäosust $P(G_k \mid H) = 0.1739673$, arvutame valemi (1) põhjal, summeerides läbikorrutatud*

vektorid, mis on saadud valemite (2) ja (3) kasutamise tulemustena (teine ja kolmas veerud):

Joonis 2: Tõenäosuste $P(Z_i^{(1)}, Z_i^{(2)}, H \mid G_k)$ vektori arvutamise tulemused. Osa 1.

	Seisund	Valem (3)	Valem (2)	Valem (4)
1	aaa	0.05358368	0.047328241	0.014577572
2	aab	0.13931756	0.024422316	0.019558026
3	aac	0.13931756	0.024422316	0.019558026
4	aad	0.05358368	0.012602402	0.003881666
5	aba	0.05358368	0.012602402	0.003881666
6	abb	0.13931756	0.024422316	0.019558026
7	abc	0.13931756	0.006503091	0.005207844
8	abd	0.05358368	0.012602402	0.003881666
9	aca	0.05358368	0.012602402	0.003881666
10	acb	0.13931756	0.006503091	0.005207844
11	acc	0.13931756	0.024422316	0.019558026
12	acd	0.05358368	0.012602402	0.003881666
13	ada	0.05358368	0.003355724	0.001033597
14	adb	0.13931756	0.006503091	0.005207844
15	adc	0.13931756	0.006503091	0.005207844
16	add	0.05358368	0.012602402	0.003881666
17	baa	0.13931756	0.024422316	0.019558026
18	bab	0.36222565	0.012602402	0.026240062
19	bac	0.36222565	0.012602402	0.026240062
20	bad	0.13931756	0.006503091	0.005207844
21	bba	0.13931756	0.024422316	0.019558026
22	bbb	0.36222565	0.047328241	0.098544387
23	bbc	0.36222565	0.012602402	0.026240062
24	bbd	0.13931756	0.024422316	0.019558026
25	bca	0.13931756	0.006503091	0.005207844
26	bcb	0.36222565	0.003355724	0.006987114
27	bcc	0.36222565	0.012602402	0.026240062
28	bcd	0.13931756	0.006503091	0.005207844
29	bda	0.13931756	0.006503091	0.005207844
30	bdb	0.36222565	0.012602402	0.026240062
31	bdc	0.36222565	0.012602402	0.026240062
32	bdd	0.13931756	0.024422316	0.019558026
33	caa	0.13931756	0.024422316	0.019558026

Joonis 3: Tõenäosuste $P(Z_i^{(1)}, Z_i^{(2)}, H \mid G_k)$ vektori arvutamise tulemused. Osa 2.

	Seisund	Valem (3)	Valem (2)	Valem (4)
34	cab	0.36222565	0.012602402	0.026240062
35	cac	0.36222565	0.012602402	0.026240062
36	cad	0.13931756	0.006503091	0.005207844
37	cba	0.13931756	0.006503091	0.005207844
38	cbb	0.36222565	0.012602402	0.026240062
39	cbc	0.36222565	0.003355724	0.006987114
40	cbd	0.13931756	0.006503091	0.005207844
41	cca	0.13931756	0.024422316	0.019558026
42	ccb	0.36222565	0.012602402	0.026240062
43	ccc	0.36222565	0.047328241	0.098544387
44	ccd	0.13931756	0.024422316	0.019558026
45	cda	0.13931756	0.006503091	0.005207844
46	cdb	0.36222565	0.012602402	0.026240062
47	cdc	0.36222565	0.012602402	0.026240062
48	cdd	0.13931756	0.024422316	0.019558026
49	daa	0.05358368	0.012602402	0.003881666
50	dab	0.13931756	0.006503091	0.005207844
51	dac	0.13931756	0.006503091	0.005207844
52	dad	0.05358368	0.003355724	0.001033597
53	dba	0.05358368	0.012602402	0.003881666
54	dbb	0.13931756	0.024422316	0.019558026
55	dbc	0.13931756	0.006503091	0.005207844
56	dbd	0.05358368	0.012602402	0.003881666
57	dca	0.05358368	0.012602402	0.003881666
58	dcb	0.13931756	0.006503091	0.005207844
59	dcc	0.13931756	0.024422316	0.019558026
60	dcd	0.05358368	0.012602402	0.003881666
61	dda	0.05358368	0.012602402	0.003881666
62	ddb	0.13931756	0.024422316	0.019558026
63	ddc	0.13931756	0.024422316	0.019558026
64	ddd	0.05358368	0.047328241	0.014577572

Antud vektori maksimaalne väärtus 0.09854439 *esineb antud vektoris 2 korda:* 22*-s reas ja* 43*-s veerus ning vastab kombinatsioonidele* bbb *ja* ccc *vastavalt, millised peale ümberkodeerimist võtavad kuju:*

	$SNP1$	$SNP2$	$SNP3$
$Z_i^{(1)}$	0	1	1
$Z_i^{(2)}$	1	1	0

	$SNP1$	$SNP2$	$SNP3$
$Z_i^{(1)}$	1	1	0
$Z_i^{(2)}$	0	1	1

vastavalt, ning mõlemad osutuvad korrektseks lahendiks.

2.3 Meetodi kasutamine töös

Käesoleva töö raames imputeerimist viidi läbi kolmel erineval tingimusel, kusjuures imputeeriti iga kord 20-nda kromosoomi osa, mis moodustab 20-nda kromosoomi pikkusest ligikaudu $\frac{3}{4}$ genoomi füüsilise positsiooni mõttes. Imputeerimiseks vajalikud andmed muudeti ja valmistati ette iga imputeerimise jaoks eraldi vastavalt vajadusele ja eesmärgile (vt. Sissejuhatus, lk.3). Nii andmete muutmise kui ka imputeerimise skeeme on esitatud allpool.

Esimest imputeerimist teostati nö ideaaltingimustes, kus referentspaneelina kasutati 1000 Genomes Phase I Integrated referenshaplotüüpe ning uuritav valim moodustati 100-st juhuslikult valitud haplotüübist referenshaplotüübide hulgast. Esimese imputeerimise skeemi saab näha joonisel 4.

Teist imputeerimisprotsessi teostati kasutades referentspaneelina samast 1000 Genomes Phase I Integrated andmestikust saadud haplotüüpe, kuid uuritav valim moodustati eestlaste sekveneeritud andmetest (vt. Sissejuhatus, lk. 3). Antud imputeerimisstsenaariumi kasutamise eesmärgiks oli kontrollida eestlaste genotüübi ennustamise kvaliteeti eurooplaste referentshaplotüüpe kasutades. Teise imputeerimise skeemi saab näha joonisel 5.

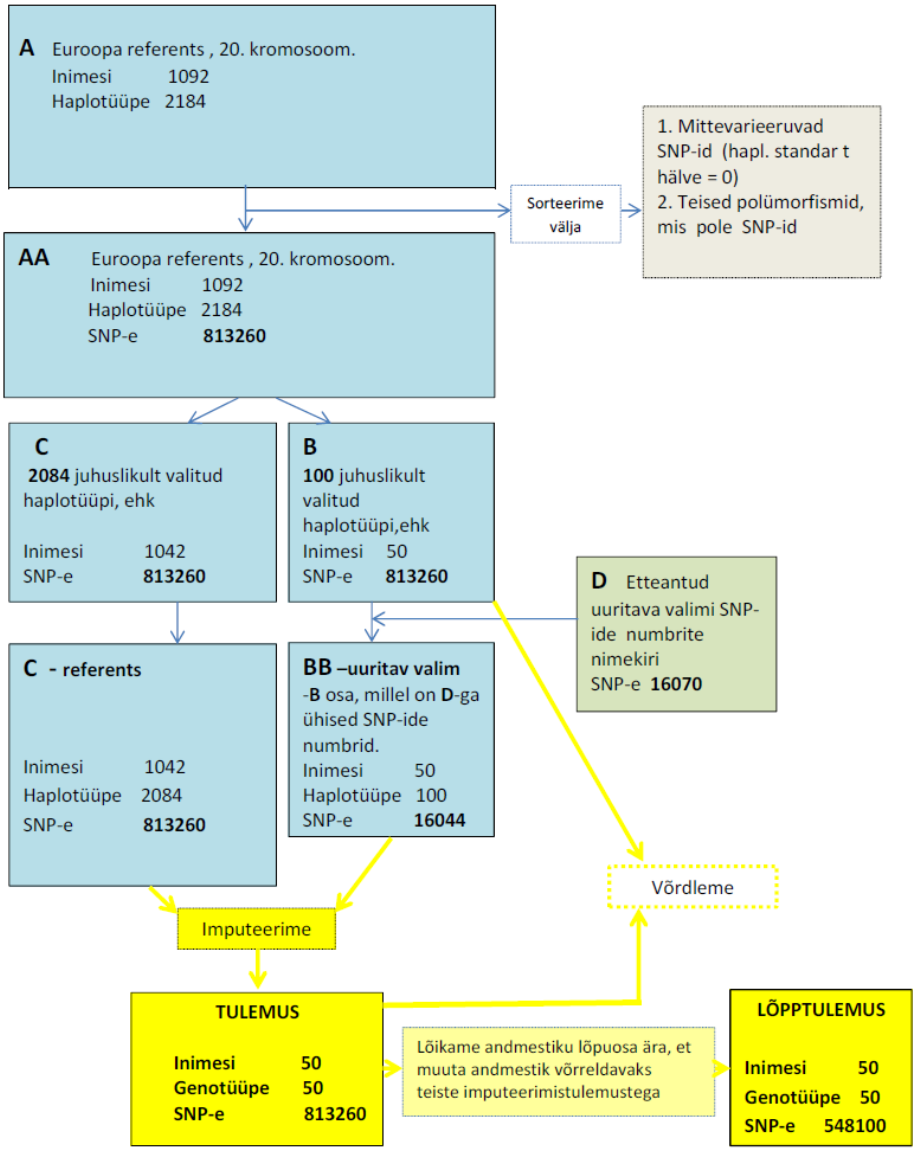
Mainime, et teise imputeerimise jaoks andmete ettevalmistamise käigus korjati välja SNP-id genotüübidega AT ja CG (vt. 2. imputeerimisskeemi). Selle põh-

juseks on asjaolu, et genotüübide AT ja CG pööratud genotüübid omavad kuju TA ja GC ning kohati võivad tingida andmete sekveneerimise käigus segadust. Et minimiseerida riski kaasata analüüsi valed andmeid, otsustati nimetatud SNP-id analüüsist välja jätta.

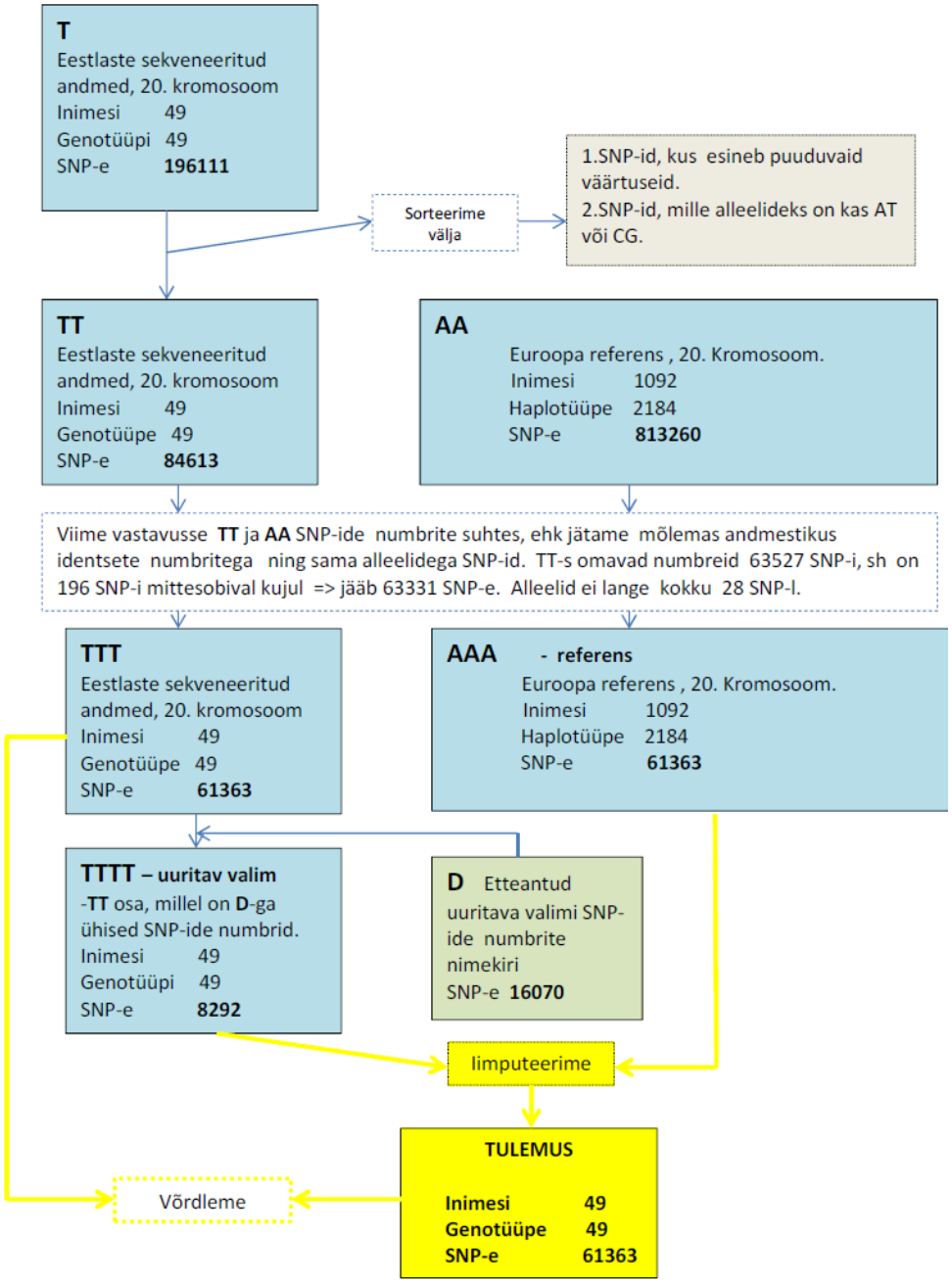
Kolmanda imputeerimise jaoks kasutati nii referentspaaneelina, kui ka valimi moodustamiseks eestlaste sekveneeritud andmeid (vt. Sissejuhatus, lk. 3), mis võimaldas kontrollida eestlaste genotüübi ennustamise kvaliteeti eestlaste referentsgenotüüpe kasutades. Kolmanda imputeerimise skeem on esitatud joonisel 6.

On oluline, et kuigi esimesel ja kolmandal imputeerimisel nii uuritava valimi, kui ka referentspaneeli andmed pärinevad samast populatsioonist, on erinevuseks see, et esimesel imputeerimisel kasutatakse haplotüpiseeritud andmeid, kolmandal aga genotüpiseeritud. Kolmanda imputeerimise genotüüpiseeritud andmed haplotüpiseeriti, kasutades programmi IMPUTE2.

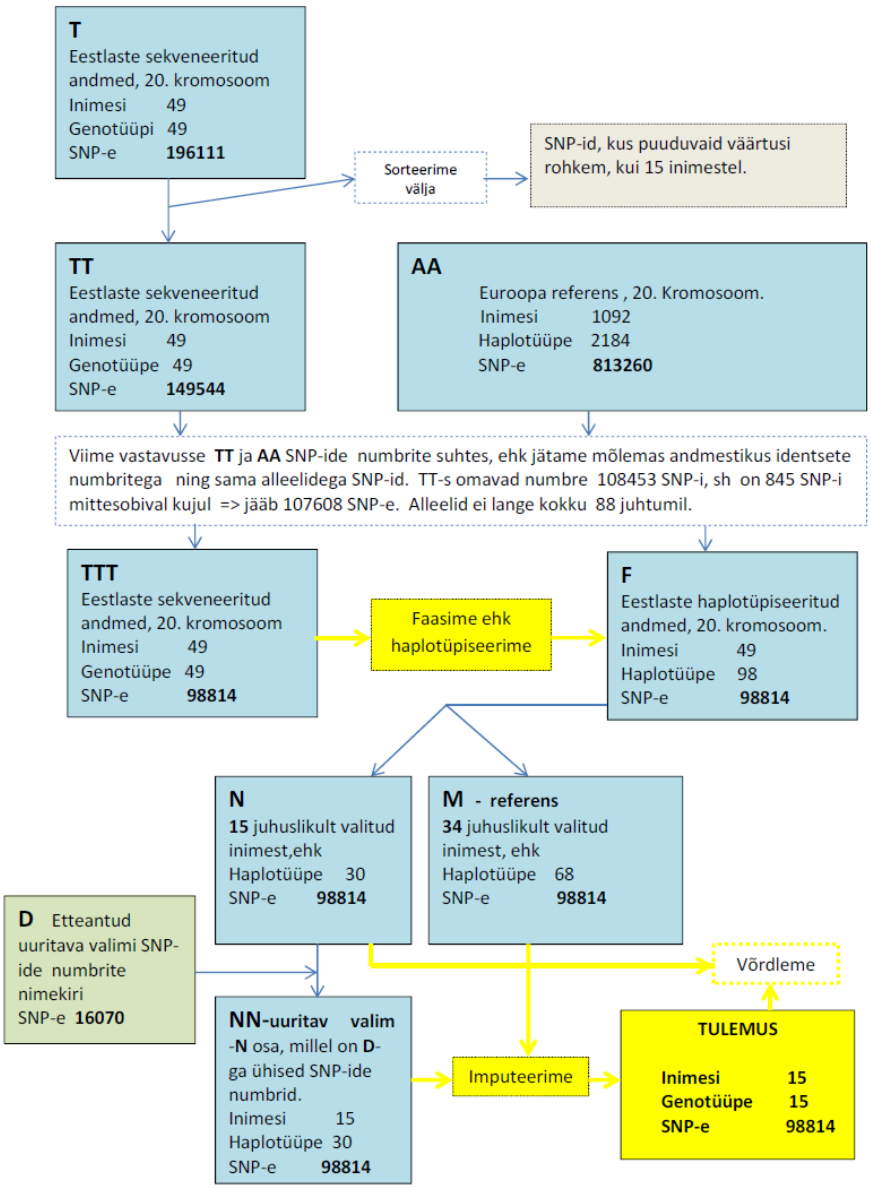
Joonis 4: Eurooplaste genotüübi imputeerimine eurooplaste referentshaplotüüpe kasutades.



Joonis 5: Eestlaste genotüübi imputeerimine eurooplaste referentshaplotüüpe kasutades.



Joonis 6: Eestlaste genotüübi imputeerimine eestlaste referentshaplotüüpe kasutades.



3 Imputeerimistulemuste analüüs

3.1 Analüüsiks kasutatud andmed

Imputeerimistulemuste kontrollimise põhiidee seisnes tegelike ja prognoositud genotüüpide võrdlemises. Võrdlemise käigus arvutati erinevaid näitajaid ja teststatistikuid, mis koondati tabelitesse. Tulemuste analüüsi teostati kolme tabeli põhjal (iga imputeerimise kohta üks tabel). Kõik autori poolt koostatud tabelid sisaldasid järgmiseid tunnuseid:

1. SNP-i füüsiline positsioon genoomis, pidev tunnus.
2. Imputeeritava SNP-i minimaalne kaugus teadaolevast SNP-ist, ehk kui kaugel asub imputeeritud SNP lähimast teadaolevast markerist füüsilise positsiooni mõttes (andmed geneetiliste markerite füüsilise positsiooni kohta saadakse koos referentspaneelidega), pidev tunnus.
3. Minoorse alleeli sagedus,lühendatuna MAF (ingl. $Minor\ allele\ frequency$), ehk harvemesineva alleeli esinemissagedus populatsioonis, kust on pärit imputeeritavad andmed, pidev tunnus.
4. SNP-i alleelid, nominaalne tunnus.
5. Valesti arvutatud genotüüpide arv, ehk kui paljude indiviidide jaoks oli genotüüp valesti ennustatud, diskreetne tunnus.
6. Valesti arvutatud genotüüpide osa protsentides (sama, mis eelmine tunnus, kuid arvutatud protsentides), pidev tunnus.
7. Hoshmer-Lemeshew teststatistiku väärtus, kus antud test osutus võimalikuks, vastasel juhul arvutati χ^2- statistiku väärtust, pidev tunnus.
8. Punktis 7 nimetatud teststatistiku olulisuse tõenäosus, pidev tunnus.
9. AUC statistiku väärtus, pidev tunnus.

Esimest viit tunnust kasutati imputeerimise kvaliteedi hindamisel ja kolm viimast tunnust kasutati imputeerimise kvaliteedihinnangu hindamisel.

3.2 Imputeerimise kvaliteet

Antud töö osas kirjeldame SNP-ide imputeerimise täpsust ja imputeerimise täpsuse sõltuvust SNP-i iseloomustavatest tunnustest, kasutades logistilise regressiooni. Uurime kui palju valesti imputeeritud genotüüpe on vaadeldavas SNP-is, kusjuures esitame SNP-i imputeerimisvea tegemise tõenäosuse prognoosi protsentides (mugavuse mõttes).

Mudelite kuju töös ei esitata, kuna kasutatud mudelite polünoomide järgud on kõrged, ning mudelite kasutamise põhiliseks eesmärgiks on tunnustevaheliste seoste visualiseerimine.

Imputeerimiskvaliteedi üldise hinnangu iseloomustavad näitajad (ümardatuna) on:

1. Täiesti korrektselt imputeeritud SNP-de osakaal on esimese imputeerimise korral 79%, teise imputeerimise korral 72% ja kolmanda imputeerimise korral 77%.

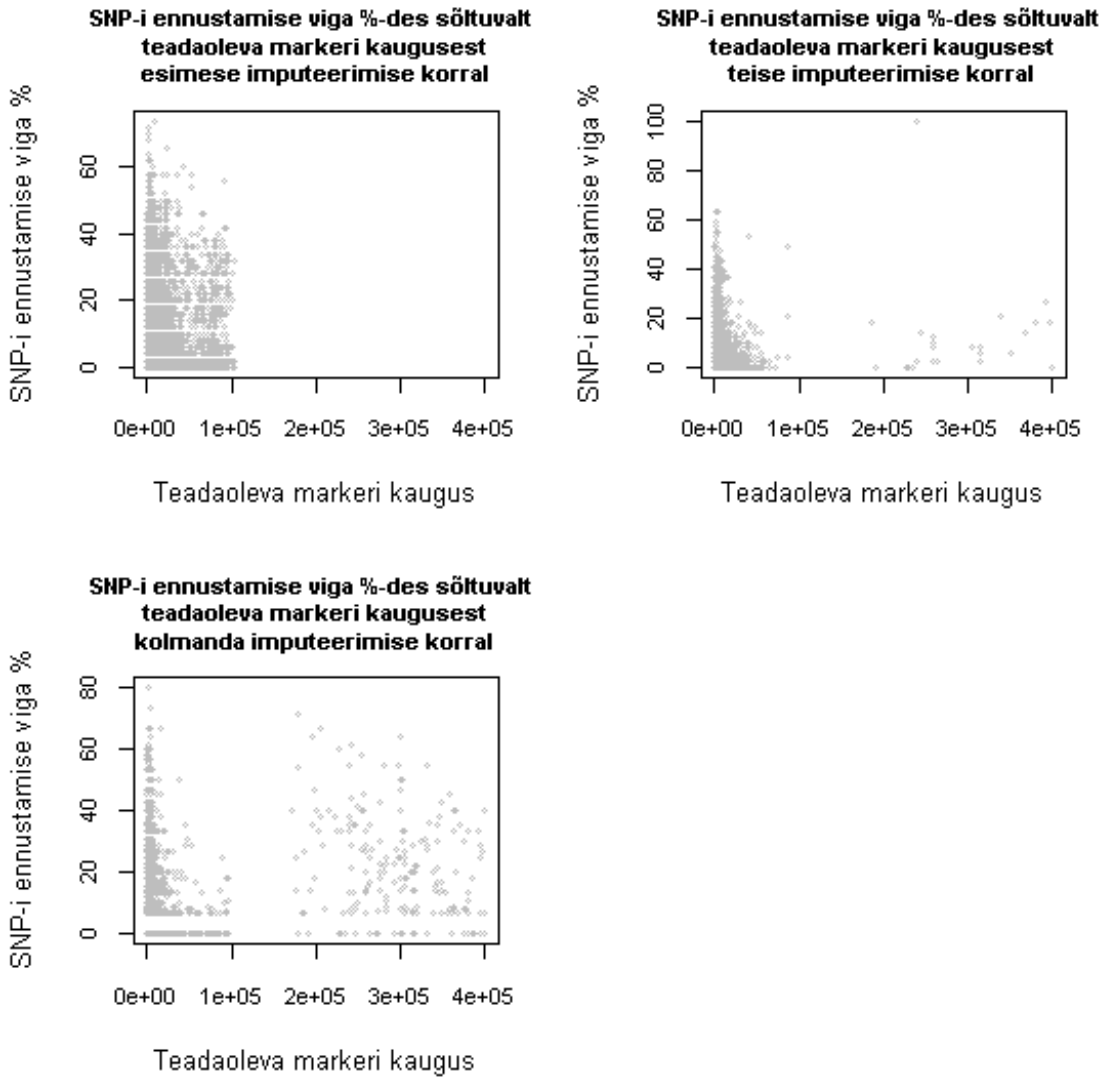
2. SNP-de osakaal, kus imputeerimisviga suurem 0% ja ei ületa 20% on esimese imputeerimise korral 20%, teise imputeerimise korral 27% ja kolmanda imputeerimise korral 22%.

3. SNP-de osakaal, kus imputeerimisviga suurem 20% ja ei ületa 50% on kõikide imputeerimise korral ligikaudu 1%.

4. Täiesti ebakorrektselt imputeeritud SNP-de osakaal on kõikide imputeerimise korral ligikaudu 0%.

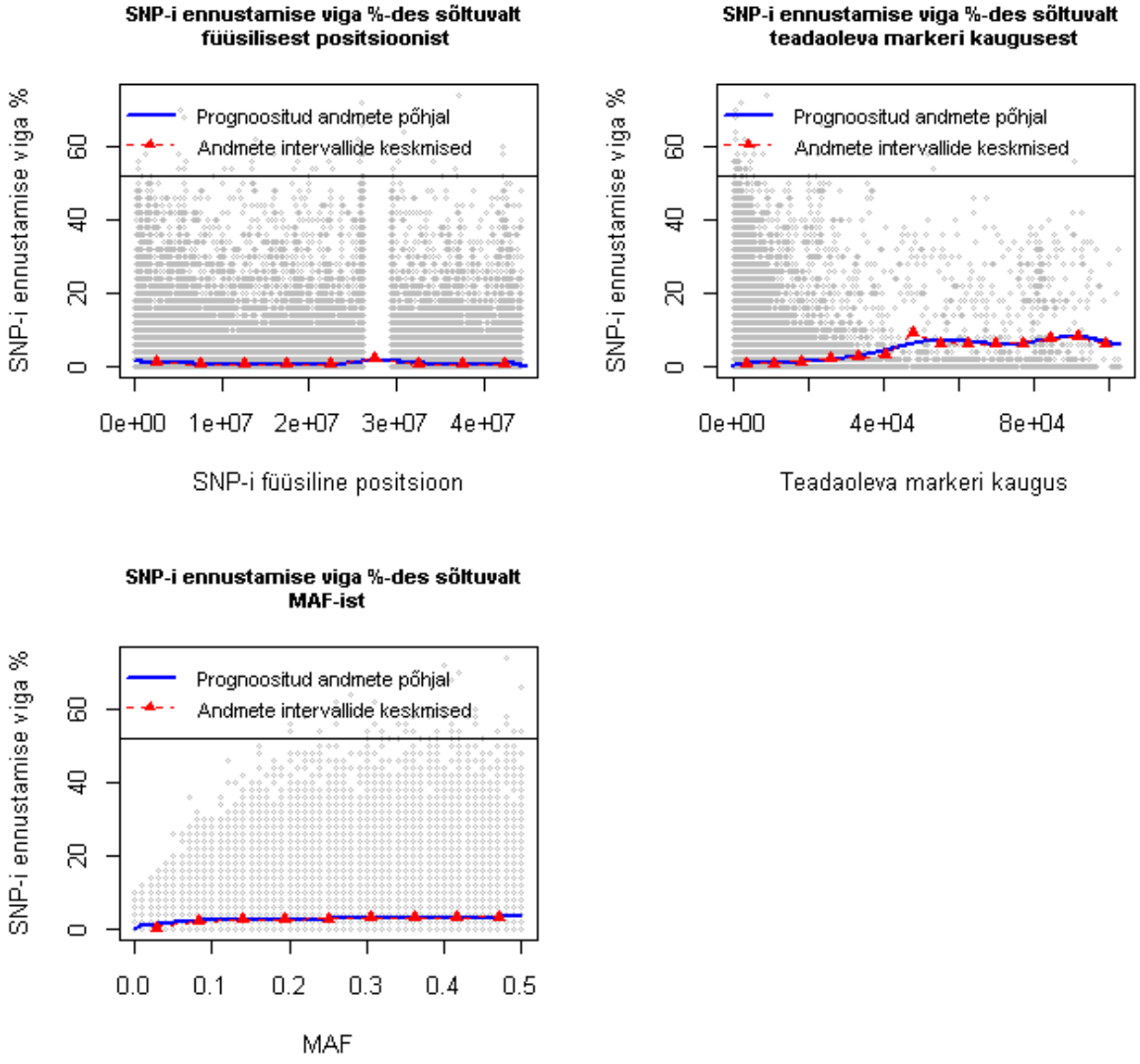
Märkame, et teisel ja kolmandal imputeerimisel esines üksikuid SNP-e, mille kaugus lähimast teadaolevast markerist oli suur (vt. joonis 7). Hilisematel teadaoleva markeri kauguse mõju kirjeldavatel joonistel jätame ekstreemselt kaugel paiknevad markerid jooniselt välja ja piirdume kirjeldamisel väiksemate (sagedamini esinevate) kaugustega.

Joonis 7: Imputeerimisvea sõltuvus teadaoleva markeri ja imputeeritava SNP-i vahekaugusest.

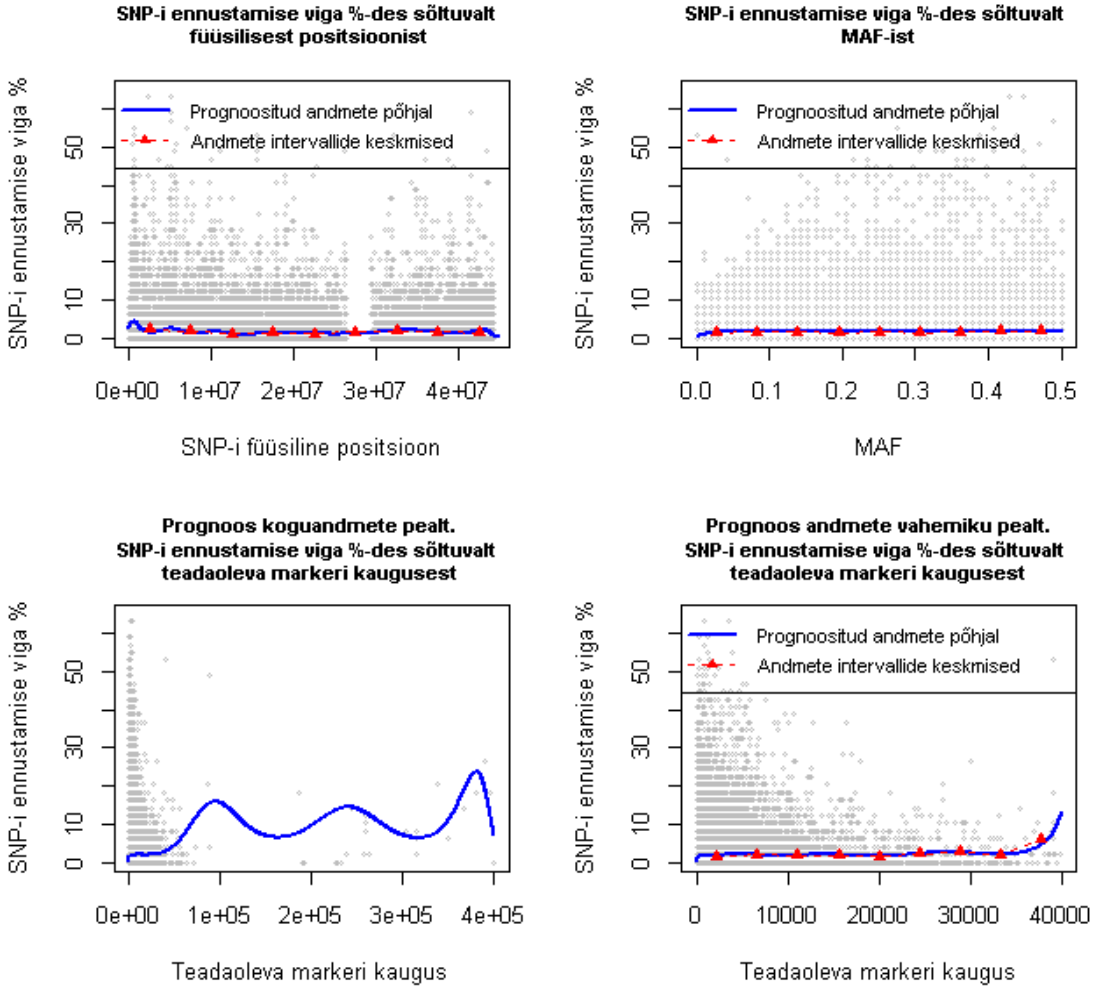


Joonistel 8,9,10 füüsilise positsiooni mõju kirjeldavatel graafikutel on selgelt näha positsiooni 30000000 aluspaari piirkonnas “tühja koridori”, ehk piirkonda, kus mõõtmised puuduvad. Selle põhjuseks on asjaolu, et antud piirkond vastab kromosoomi tsentromeeri [11] asukohale. Tsentromeeris paikneda võivad geneetilisi markereid pole aga tänapäevaste tehnoloogiate abil võimalik uurida.

Joonis 8: 1. Imputeerimine. Imputeerimisvea sõltuvus uuritavatest pidevatest tunnustest.

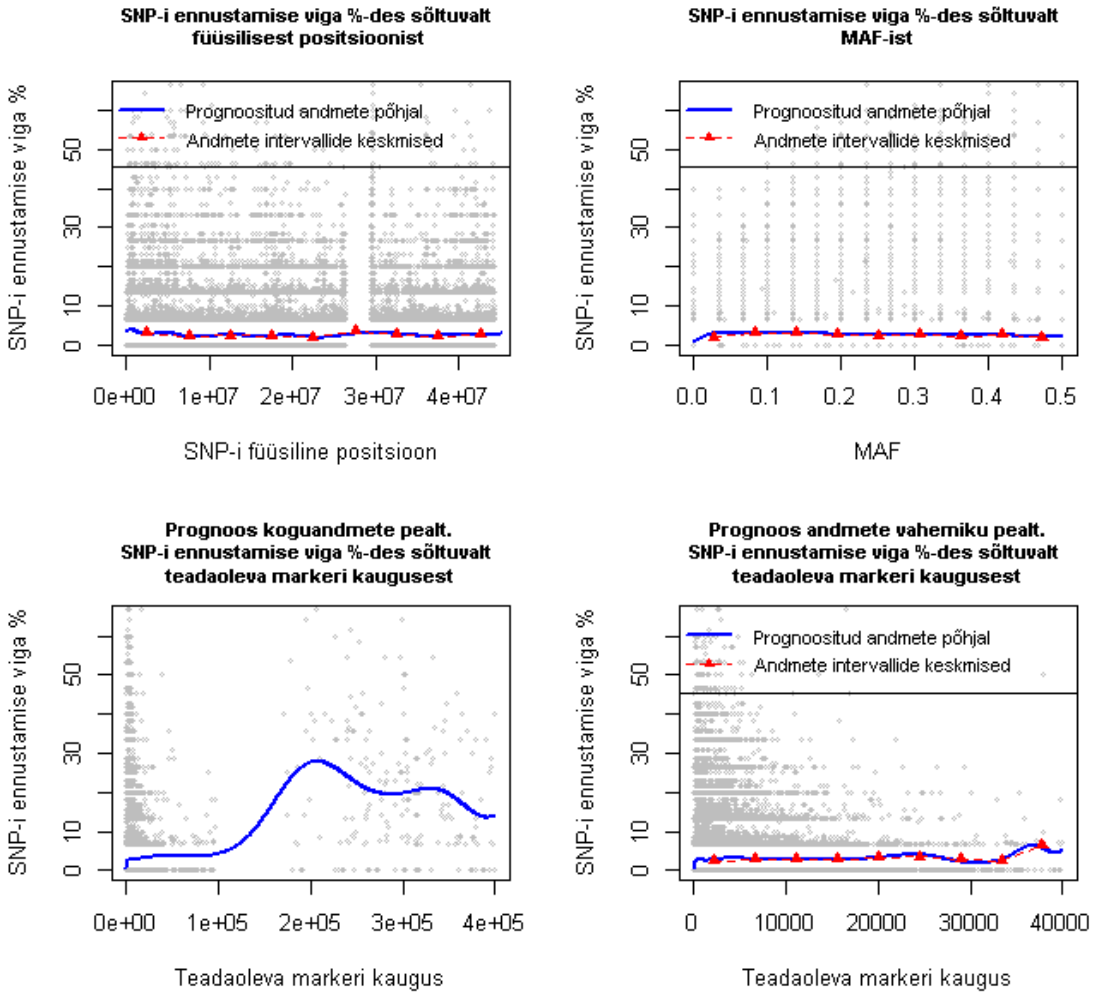


Joonis 9: 2. Imputeerimine. Imputeerimisvea sõltuvus uuritavatest pidevatest tunnustest.



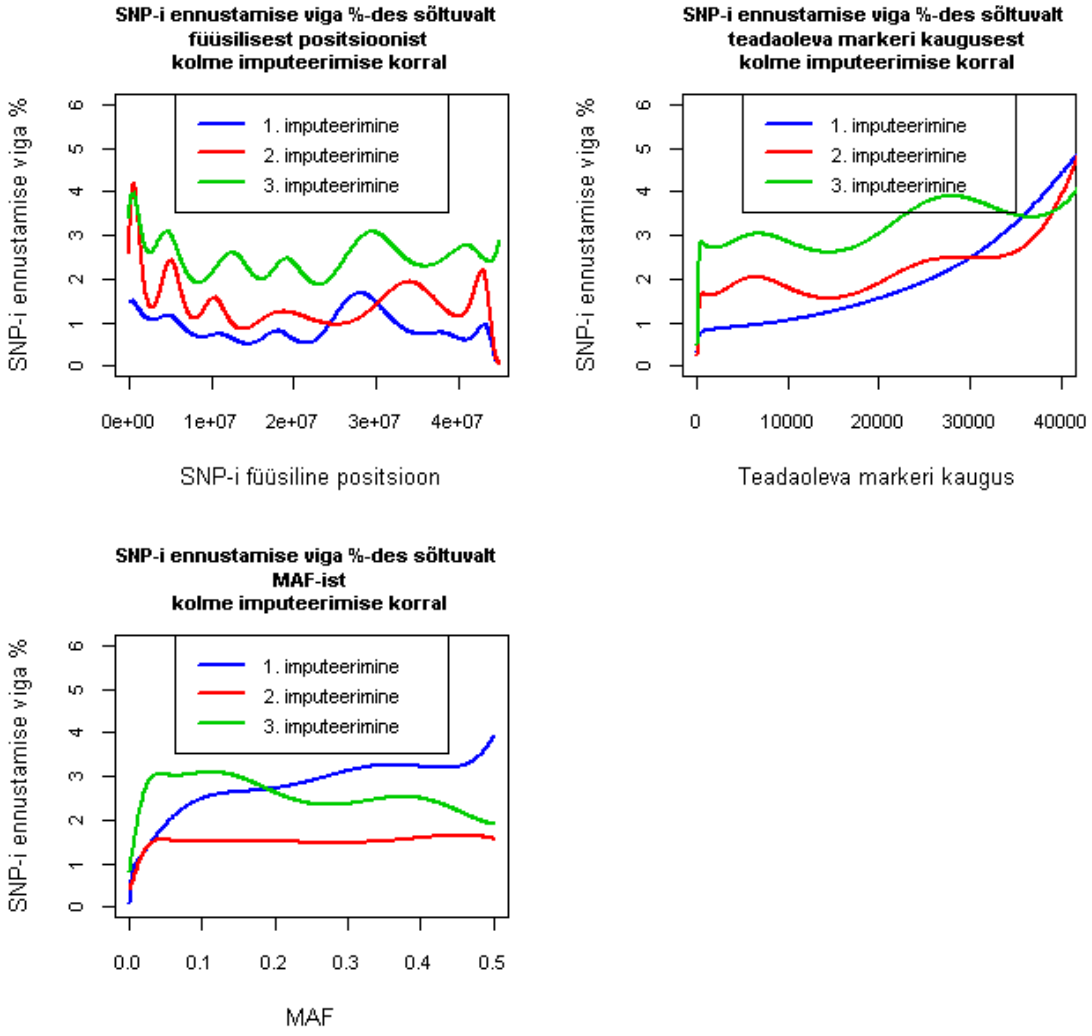
Märkame ka, et mudelite prognoosid sobivad ilusti andmetega, sest seletavate tunnuste väärtuste intervallidele vastavad imputeerimisvea keskmised järgivad logistilise regressioonimudeli abi leitud prognoosikõverat (vt joonised 8,9,10).

Joonis 10: 3. Imputeerimine. Imputeerimisvea sõltuvus uuritavatest pidevatest tunnustest.



Võrdlemaks imputeerimise kvaliteeti kolmel erineval imputeerimisel esitame imputeerimiskvaliteeti kirjeldavad kõverad samadel graafikutel (vt. joonis 11). Võrreldes omavahel kolme imputeerimiskvaliteeti kirjeldavaid seoseid, saab öelda, et sobitatud mudelid (mis sobivad andmetega hästi) käituvad väga sarnaselt erinevate imputeerimiste korral, välja arvutatud mudel, mis kirjeldab imputeerimistäpsuse sõltuvust MAF-ist, millele anname seletuse hiljem (vt. joonis 12, tabel 2, seletus lk. 26-27).

Joonis 11: Imputeerimise vea sõltuvus pidevatest tunnustest.



Uurides füüsilise positsiooni mõju imputeerimistäpsusele, saab märgata, et tsentromeeri piirkonna läheduses imputeerimisviga kergelt suureneb, samuti ka uuritud kromosoomi alguses (vt. joonis 11). Üldiselt aga ei mõjuta füüsiline asukoht genoomis märkimisväärselt imputeerimistulemust.

Imputeeritava SNP-i ja temaga lähima teadaoleva markeri vahelise kaugu-se mõju imputeerimistäpsusele avaldub kasvavalt kõikide imputeerimiste korral, kuid teise ja kolmanda imputeerimistäpsuse prognoosid on suurema võnkumisega, kui esimese oma, mis võib olla tingitud tunnuse “teadaoleva markeri kaugus”

Tabel 2: Keskmine kaugus lähimast markerist

	1.imputeerimine	2.imputeerimine	3.imputeerimine
Väike MAF-[0,0.001]	2798	2440	2830
Suur MAF-[0.4,0.5]	2362	2255	2213

ebaühtlasemast jaotusest teise ja kolmanda imputeerimise korral. (joonis 11).

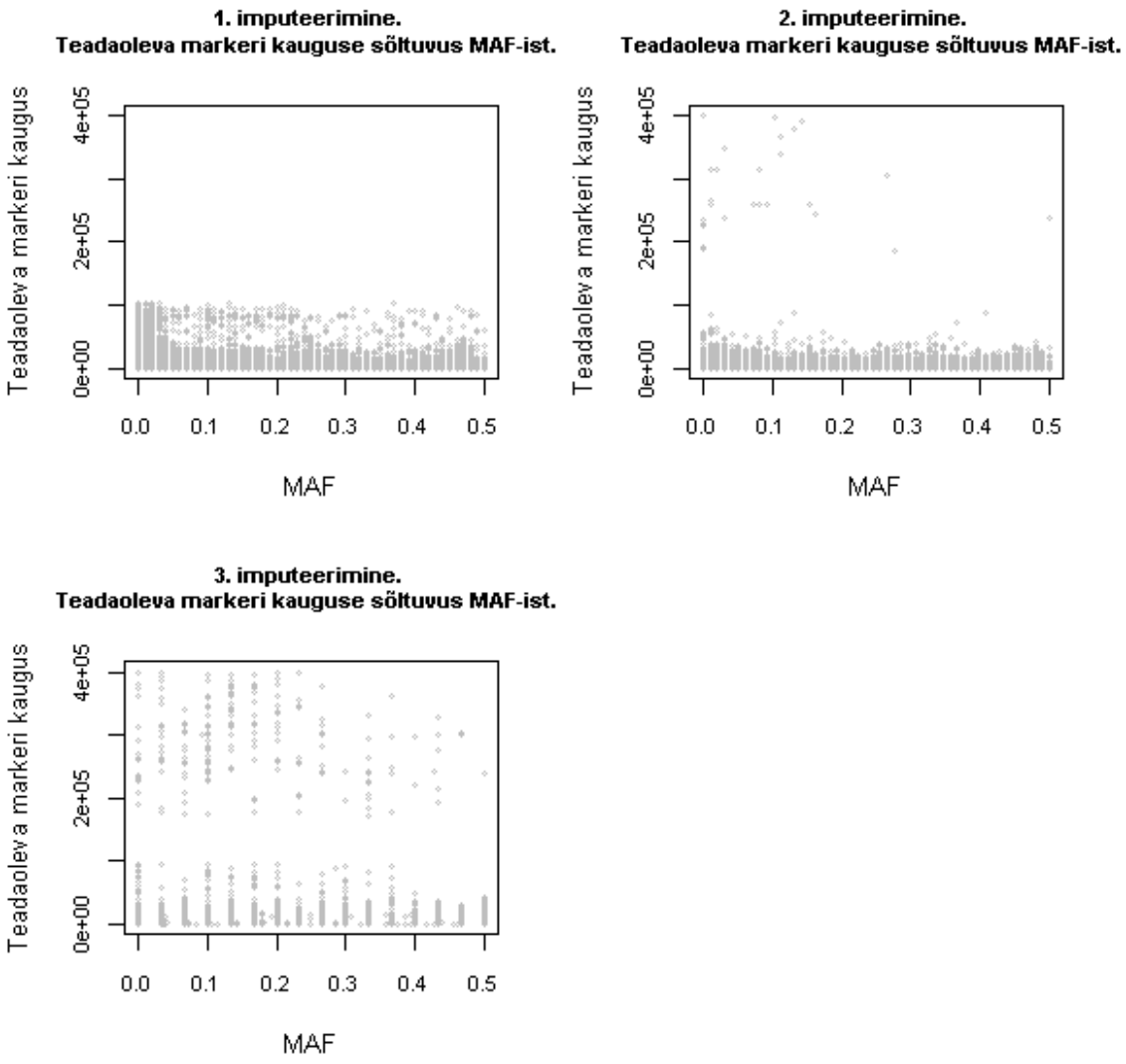
On märgata, et MAF-i mõju imputeerimisvea tõenäosuse prognoosile erineb kolme imputeerimise korral: esimese imputeerimise korral on see kasvava mõjuga, kolmandal imputeerimisel on MAF-ist tingitud imputeerimisviga alguses oluliselt kõrgem kui teisel ja esimesel imputeerimisel.

Lisaks sellele, MAF-i kasvades kolmanda imputeerimise vea tõenäosus hakkab aeglaselt kahanema. Samal ajal teise imputeerimise korral on imputeeerimisvea tõenäosuse prognoos kõige madalam ja MAF-ist peaaegu ei sõltu (joonis 11).

Taoline mudelite käitumise erinevus imputeerimiste korral on tingitud sellest, et kolmandal imputeerimisel suur osa vähevarieeruvatest SNP-idest (s.t. väiksema MAF-i väärtustega) asetsevad kaugel uuritava valimi teadaolevast markerist, mis omakorda mõjutab imputeerimistäpsust suurendades imputeerimisviga.

Samal ajal teisel imputeerimisel asuvad SNP-id teadaoleva markeri läheduses peaaegu kõikide MAF-i väärtuste korral, välja arvatud üksikud vaatlused (vt. tabel 2, hajuvusdiagramm joonisel 12).

Joonis 12: Teadaoleva markeri kauguse sõltuvus MAF-ist.

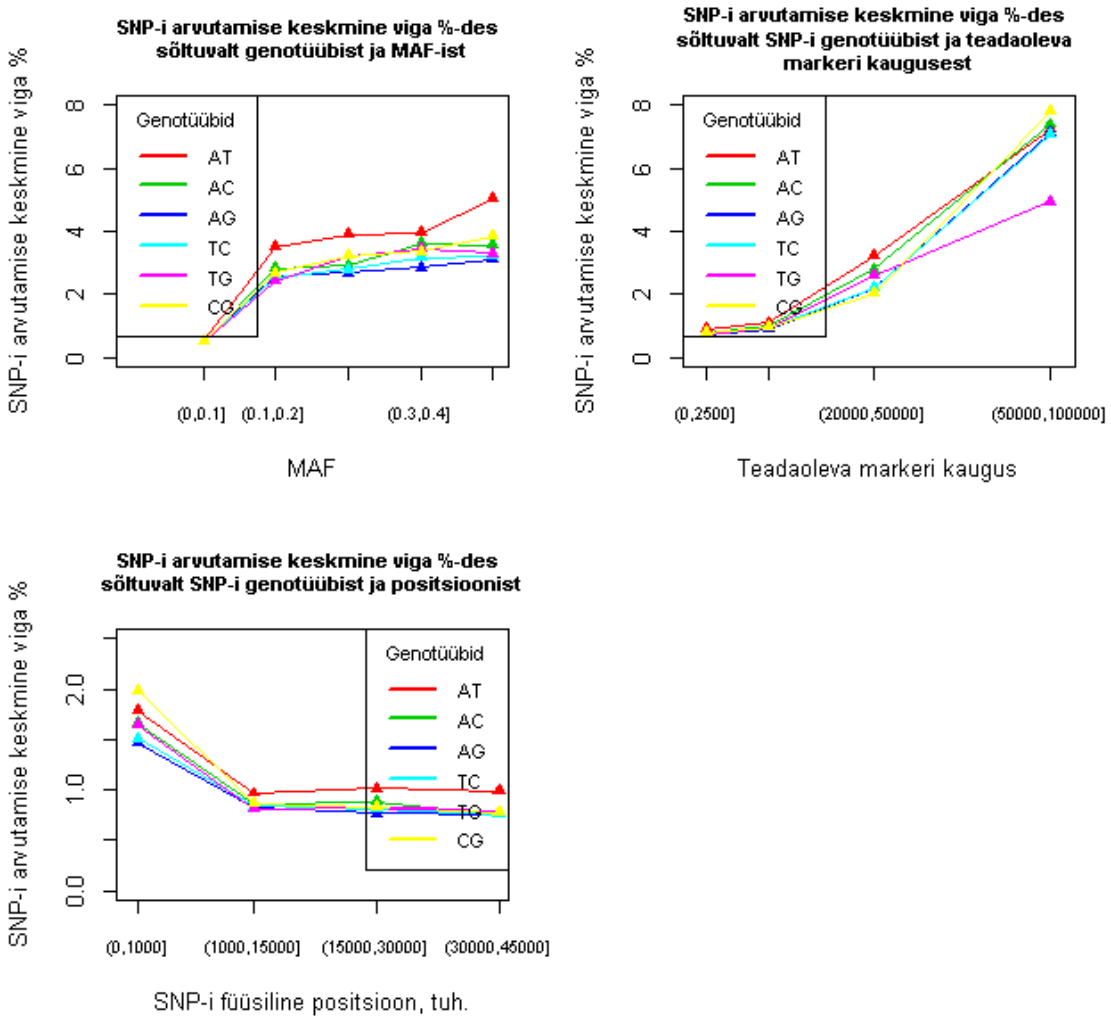


Üldiselt, joonise 11 põhjal saab järeldada, et imputeerimiskvaliteet kõigi kolme imputeerimise puhul on väga hea ning esimese ja teise imputeerimise korral peaaegu ei erine. Tuletame ka meelde, et nimetatud juhtudel kasutati eurooplaste referentspaneeli, kuhu eestlaste andmeid pole kaasatud, kuid esimene kord imputeeriti sama referentspaneeli haplotüüpidelt juhuslikult moodustatud valimisse, teine kord aga valim oli juhuslikult moodustatud eestlaste sekveneeritud andmetest. Seega, polnud referentspaneel ja valim moodustatud samast populatsioonist.

Kolmanda imputeerimise kvaliteet on testest mitteoluliselt madalam - keskmiselt ligikaudu 1-2 % võrra madalam (vt joonis 11). Mainime ka, et kolmanda imputeerimise referentspaneelina kasutati eestlaste geeniandmeid, kust olid eemaldatud nii SNP-id genotüübidega AT ja CG, kui ka puuduvaid väärtuseid sisaldavad SNP-id, kusjuures referentspaneelis oli kõigest 68 haplotüübi ehk 34 indiviidi.

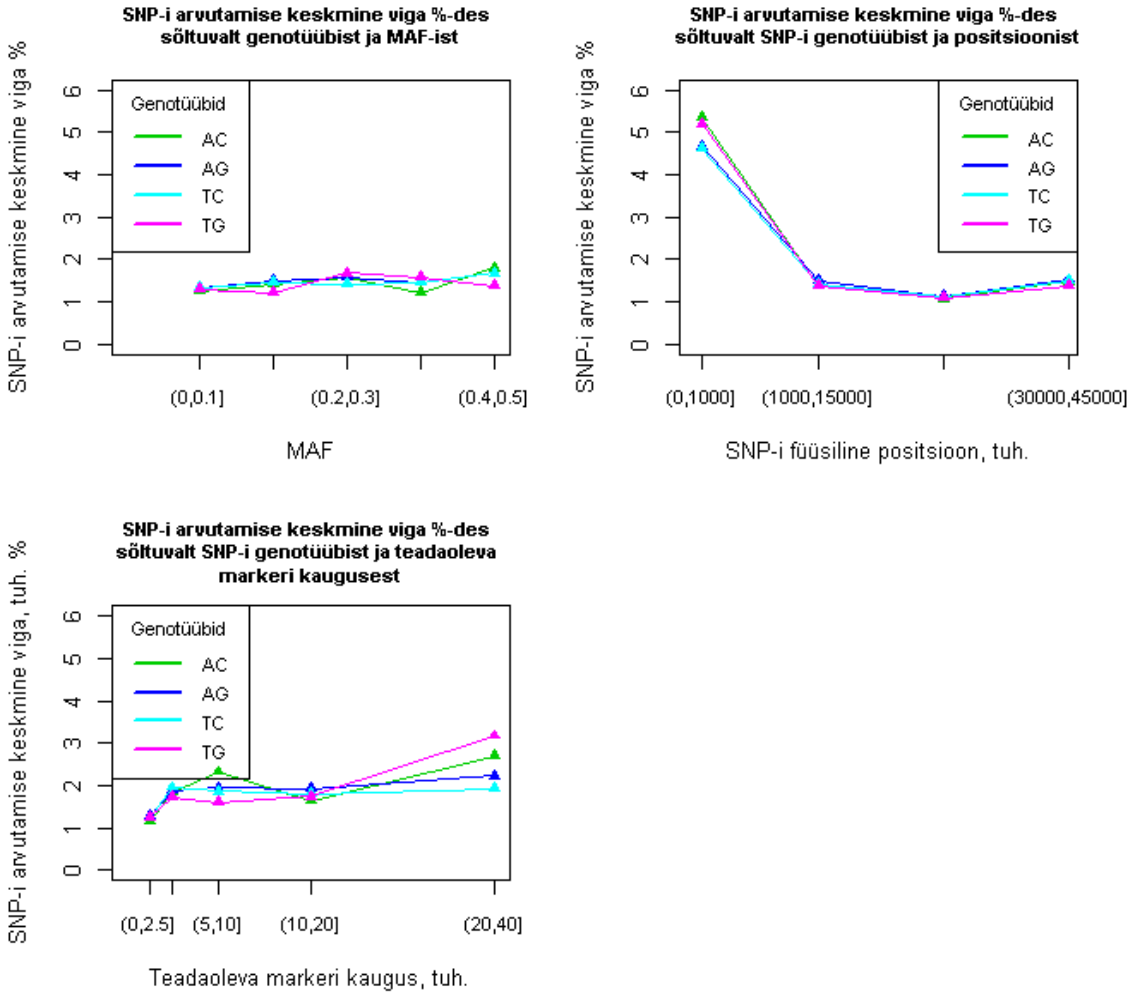
Referentspaneeli väiksuse tõttu võib haplotüpiseerimine olla raskendatud ja sellest tulenevalt võib kannatada ka imputeerimiskvaliteet. Seega, ülalmainitud arvestades, võib teise imputeerimise parema kvaliteedi (võrreldes kolmanda imputeerimisega) põhjuseks olla referentspaneeli nõudlikum ja rangem töötlus.
Lõpuks uurime ka SNP-i alleelide mõju imputeerimistäpsusele (joonis 13,14,15).

Joonis 13: SNP-i imputeerimisvea sõltuvus pidevatest tunnustest ja MAF-ist 1. imputeerimise korral.



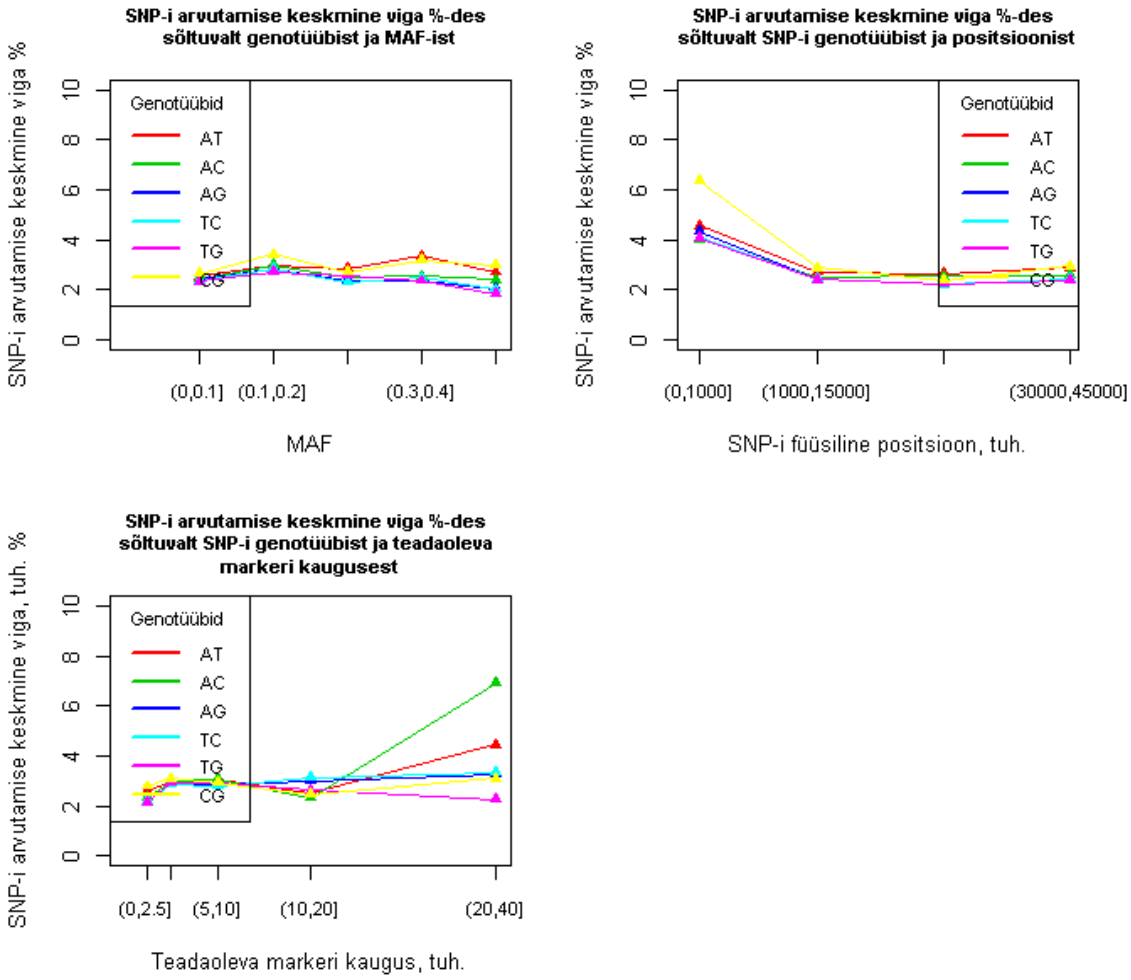
Nagu eespool mainitud (vt 2.3 Meetodi kasutamine töös), teostati teist imputeerimist ilma SNP-ta , mille genotüüp on AT või CG (vt joonis 14). Antud juhul ei esine silmapaistvat süstemaatilist erinevust imputeerimiskvaliteedi vahel genotüüpide lõikes.

Joonis 14: SNP-i imputeerimisvea sõltuvus pidevatest tunnustest ja MAF-ist 2. imputeerimise korral.



Esimese ja kolmanda imputeerimise puhul (vt. joonised 13, 15) hakkab silma alleelidega AT ja CG SNP-ide imputeerimisel tehtavate vigade erinev käitumine teistest SNP-idest, ehk nimetatud alleelidega SNP-id imputeritakse suurema veaga iga füüsilise positsiooni- ja MAF-i väärtuste korral nii esimesel , kui ka kolmandal imputeerimisel.

Joonis 15: SNP-i imputeerimisvea sõltuvus pidevatest tunnustest ja MAF-ist 3. imputeerimise korral.



Graafikutel, mis kirjeldavad imputeerimisvea prognoosi sõltuvust valimi teadaoleva markeri kaugusest genotüüpide lõikes (joonised 13-15) saab märgata, et SNP-id alleelidega AT ja AC suurendavad imputeerimisvea tõenäosust.

Taoline mõju AT ja CG poolt on mõnel määral ootuspärane (vt. 2.3 Meetodi kasutamine töös) ning kinnitab otsust eemaldada vastavad SNP-id analüüsist.

Üllatav on ka alleelidega AC SNP-ide imputeerimisvea seos teadaoleva markeri kaugusest. Kui aga kontrollime keskmise vea usalduspiire huvialuses piirkonnas ning leiame, et usaldusintervallid kattuvad, järelikult ei saa väita, et üldkogumi

tasemel antud genotüübiga SNP-id imputeeritakse suurema veaga.

Antud peatüki põhjal saab järeldada, et IMPUTE2 tarkvara kasutades saab eestlaste genotüüpe imputeerida sama edukalt nii eurooplaste haplotüüpide abil kui ka eestlaste sekveneeritud andmete abil.

4 Imputeerimise kvaliteedihinnangu analüüs.

Vigaselt imputeeritud SNP-id ei pruugi tekitada probleeme, kui me teame, et imputeerimistulemus antud SNP-i puhul pole usaldusväärne.

Käesolevas peatükis üritame anda hinnangut programmi IMPUTE2 poolt imputeerimistulemuste kvaliteedile antud hinnangule. Nagu näidatakse alapunktis 2.2 “Meetodi kirjeldus” arvutab programm IMPUTE2 varjatud Markovi mudelit kasutades genotüübi saamise tõenäosust antud lookuses . Tulemuseks imputeeritakse genotüüp, mille saamise tõenäosus on antud lookuses maksimaalne. Kuna bialleelse markeri korral valitakse kolme erineva variandi vahel ehk 11, 12 (mis on samaväärne variandiga 21) ja 22 (vt. 1.1 SNP, genotüüp, haplotüüp, Definitsioon 1, lk. 4), võib parimaks hinnatud genotüübi tõenäosuseks olla kas 0.333333339 või 0.9999999. Ja seda nii õige imputeerimistulemuse korral, kui ka vale imputeerimisetulemuse korral.

Uurime, millesel määral vastavad programmi IMPUTE2 poolt pakutud tõenäosused korrektselt (või ka valesti) imputeeritud genotüüpide tegelikkele tõenäosustele.

4.1 Kasutatud metoodika.

Antud tööosas anname ülevaate sellest, kuidas on võimalik kontrollida mingi kindla meetodiga arvutatud positiivse katsetulemuse tõenäosuse õigsust, ehk uurida, kas kontrollitava meetodi poolt arvutatud tõenäosused vastavad vaadeldava sündmuse toimumise tegelikule tõenäosusele.

Tavaliselt, kui prognoositud tõenäosuste seas esinevad kordused ehk korduvad väärtused, siis räägime $rühmitatud$ andmetest. Sel juhul, korjame kokku vaatlused, mille arvutatud tõenäosused langevad kokku ehk korduvad ning saame sagedustabeli, (vt. tabel 3):

Tabel 3: Vaatluste sagedus prognoositud tõenäosuse suhtes.

Positiivse katsetulemuse arvutatud tõenäosus $\hat{\pi}_i$:	$\hat{\pi}_1$	$\hat{\pi}_2$	...	$\hat{\pi}_k$
Vaatluste arv $i-$*ndas rühmas* n_i:	n_1	n_2	...	n_k
Positiivsete katsetulemuste arv $i-$*ndas rühmas* y_i:	y_1	y_2	...	y_k
Negatiivsete katsetulemuse arvutatud tõenäosus $1-\hat{\pi}_i$:	$1-\hat{\pi}_1$	$1-\hat{\pi}_2$	...	$1-\hat{\pi}_k$
Negatiivsete katsetulemuste arv $i-$*ndas rühmas* n_i-y_i:	n_1-y_1	n_2-y_2	...	n_k-y_k
Positiivsete katsetulemuste prognoositud arv $i-$*ndas rühmas* $\hat{\pi}_i n_i$:	$\hat{\pi}_1 n_1$	$\hat{\pi}_2 n_2$	...	$\hat{\pi}_k n_k$
Negatiivsete katsetulemuste prognoositud arv $i-$*ndas rühmas* $(1-\hat{\pi}_i)n_i$:	$(1-\hat{\pi}_1)n_1$	$(1-\hat{\pi}_2)n_2$	...	$(1-\hat{\pi}_k)n_k$

k - unikaalsete arvutatud tõenäosuse $\hat{\pi}_i$ arv.

n - valimi kogusuurus, $n = \sum_{i=1}^{k} n_i$.

Ülaltoodud sagedustabeli põhjal ning arvestades, et huvialune tunnus on binaarne (y_i võimalikud väärtused on "oige" ja "vale"), on mõistlik kontrolli teostada $\mathcal{X}^2$- testi abil. Sel juhul võrdleb Pearsoni $\mathcal{X}^2$-statistik tegelikke ning prognoositud (arvutatud tõenäosuste põhjal) vaatluste arve ja kontrollitakse hüpoteese:

H_0: P(positiivne katsetulemus | $\hat{\pi}_i$=x) = x ehk arvutatud tõenäosus on õige,
H_1: P(positiivne katsetulemus | $\hat{\pi}_i = x) \neq x$ ehk arvutatud tõenäosus ei ole õige.

Eeldades, et vaatlused on sõltumatud ja uuritava meetodi poolt arvutatud tõenäosus on õige (ehk kehtib H_0), on uuritav tunnus y_i antud n_i korral binoomjaotusega ehk $y_i \sim B(n_i, \pi_i)$, $i = 1, ..., k$. Antud juhul $\mathcal{X}^2$ statistik avaldub kujul:

$$
\begin{aligned}
\mathcal{X}^2 &= \sum_{i=1}^{k} \Big[\frac{(y_i - n_i\hat{\pi}_i)^2}{n_i\hat{\pi}_i} + \frac{((n_i - y_i) - n_i(1-\hat{\pi}_i))^2}{n_i(1-\hat{\pi}_i)}\Big] \\
&= \sum_{i=1}^{k} \Big[\frac{(y_i - n_i\hat{\pi}_i)^2(1-\hat{\pi}_i) + (n_i\hat{\pi}_i - y_i)^2\hat{\pi}_i}{n_i\hat{\pi}_i(1-\hat{\pi}_i)}\Big] \\
&= \sum_{i=1}^{k} \Big[\frac{(y_i - n_i\hat{\pi}_i)^2(1- \cancel{\hat{\pi}_i} + \cancel{\hat{\pi}_i})}{n_i\hat{\pi}_i(1-\hat{\pi}_i)}\Big] \\
&= \sum_{i=1}^{k} \frac{(y_i - n_i\hat{\pi}_i)^2}{n_i\hat{\pi}_i(1-\hat{\pi}_i)}
\end{aligned}
$$

On teada, et nullhüpoteesi kehtides, on $\mathcal{X}^2$ statistiku väärtus ligikaudu $\mathcal{X}^2$ jaotusega, vabadusastmete arvuga $df = rühmade\ arv - valimi\ põhjal\ hinnatud\ teoreetiliste\ parameetrite\ arv$ $(meie\ konkreetsel\ juhul,\ see\ on\ rühmade\ koguarv - hinnatud\ erinevate\ n_i - de\ arv,\ ehk\ df = 2k - k)$ eeldusel, et nullhüpoteesile vastavad tõenäosused (sagedused) on piisavalt suured (suurem ühest ja vähemalt 75% nendest suurem viiest) [12], ehk :

$$
\mathcal{X}^2 = \sum_{i=1}^{k} \frac{(y_i - \hat{\pi}_i)^2}{\hat{\pi}_i(1-\hat{\pi}_i)} \xrightarrow{D,H_0} \mathcal{X}^2_{K-r},
$$

r - valimi põhjal hinnatud teoreetiliste parameetrite arv,
k - erinevate rühmade arv,
K - grupide koguarv (igas rühmas on 2 gruppi - positiivsete ja negatiivsete katsetulemustega), $K = 2k$, kusjuures k on ligikaudselt võrdne või võrdne valimi mahuga n.

Selge on see, et binaarse tunnuse korral, kui rühmade arv k on ligikaudselt võrdne või võrdne valimi mahuga n ülalmainitud eeldus pole täidetud, järelikult asümptootika ei kehti.
Sel juhul on uuritava tunnuse y_i jaotuseks Bernoulli jaotus ehk $y_i \sim B(1, \pi_i)$, $i = 1, ..., k$ ja rühmade suurused $n_1 = ... = n_k = 1$ ning räägitakse, et tegu on $rühmitamata$ andmetega. Seega, ülalmainitu põhjal, on sel juhul:

$$\mathcal{X}^2 = \sum_{i=1}^{k} \frac{(y_i - \hat{\pi}_i)^2}{\hat{\pi}_i(1-\hat{\pi}_i)} \overset{D}{\longrightarrow} \mathcal{X}^2_{K-r},$$

Rühmitamata andmete korral püstitatud hüpoteesi kontrollimiseks võib kasutada Hoshmer-Lemeshew testi.

Järgmine tekstilõik baseerub K. J. Archer and S. Lemeshow artiklil [13, lk.99]. Hoshmer-Lemeshew testi idee seisneb andmete grupeerimises kas lähtuvalt prognoositud sündmuse (ehk prognoositud positiivse vaatluse) tõenäosusest või fikseeritud etteantud lõikepunktide järgi. Siinjuures nõutakse, et subjektide arv gruppides oleks võrdne, ning gruppide arv oleks kahest suurem (soovitavalt 10).

Etteantud töös on andmed grupeeritud vastavalt hinnatud tõenäosuste kas 0,1-kvantiilidele, 0,2-kvantiilidele või kui see ei osutu võimalikuks (arvutatud tõenäosuste seas on palju kordusi), siis vastavalt hinnatud tõenäosuste 0,33-kvantiilidele.

Gruppides võrreldatakse vaadeldud sagedusi prognoositud (H_0 kehtides) sagedustega Pearsoni $\mathcal{X}^2$ statistikuga.

Näiteks, olgu meil jagatud prognoositud tõenäosused G gruppideks, kus iga grupi suurus on ligikaudselt $\frac{n}{G}$. Tähistame g-ndas grupis ($g = 1, ..., G$):

positiivsete katsetulemuste arvu

$$o_{1g} = \sum_{i=1}^{n_g} y_i,$$

negatiivsete katsetulemuste arvu

$$o_{0g} = \sum_{i=1}^{n_g} (1 - y_i),$$

prognoositud positiivsete katsetulemuste arvu

$$e_{1g} = \sum_{i=1}^{n_g} \hat{\pi}_i,$$

prognoositud negatiivsete katsetulemuste arvu

$$e_{0g} = \sum_{i=1}^{n_g} (1 - \hat{\pi}_i).$$

Saadud tulemused esitame sagedustabelina (vt. tabel 4):

Tabel 4: Vaadeldud ning prognoositud vaatluste sagedused gruppides

$Gruppid$	1	2	...	G
$Positiivsete\ katsetulemuste\ arv$ $g-ndas\ gruppis\ o_{1g}$:	o_{11}	o_{12}	...	o_{1G}
$Negatiivsete\ katsetulemuste\ arv$ $g-ndas\ gruppis\ o_{0g}$:	o_{01}	o_{02}	...	o_{0G}
$Positiivsete\ katsetulemuste\ prognoositud\ arv$ $g-ndas\ gruppis\ e_{1g}$:	e_{11}	e_{12}	...	e_{1G}
$Negatiivsete\ katsetulemuste\ prognoositud\ arv$ $g-ndas\ gruppis\ e_{0g}$:	e_{01}	e_{02}	...	e_{0G}

kus $o_{1g} + o_{0g} = e_{1g} + e_{0g} = \frac{n}{G}$ ja $o_1 + o_0 = e_1 + e_0 = n$.

Siis Hoshmer-Lemeshew teststatistik avaldub kujul:

$$\hat{C}_i^* = \sum_{k=0}^{1} \sum_{g=1}^{G} \frac{(o_{kg} - e_{kg})^2}{e_{kg}} \sim \mathcal{X}^2_{2G-G},$$

tingimusel, et arvutatud tõenäosuste unikaalsete väärtuste arv on ligikaudne võrdne või võrdne valimi mahuga.

Veel üheks laialdaselt kasutatavaks lähenemiseks hinnata binaarse tunnuse prognoosimiseks kasutatava meetodi täpsust on arvutada meetodi $tundlikkus$ ja $spetsiifilisus$. Nende näitajate/statistikute defineerimiseks vajame mõnede mõistete lahtiseletamist.

Esmalt määrame positiivse katsetulemuse tõenäosusele piiri, mida tähistame c. Anname ette järgmise otsustamisreegli: kui arvutatud tõenäosus on suurem c väärtusest, siis prognoosime positiivset katsetulemust, c - st väiksema või võrdse tõenäosuse korral prognoosime negatiivset katsetulemust, ehk:

$$\text{kui } \hat{\pi}_i > c, \hat{y}_i = 1, i = 1, ..., k$$
$$\text{kui } \hat{\pi}_i \leq c, \hat{y}_i = 0, i = 1, ..., k$$

Kasutades ülalkirjeldatud otsustamisreeglit, esitame tulemused klassifitseerimistabelina (vt tabel 5):

Tabel 5: Vaadeldud ning prognoositud vaatluste klassifitseerimistabel

$Katsetulemuse\ tegelik\ väärtus\ y_i$	$y_i = 0$	$y_i = 1$
$Otsustamisreegli\ \hat{\pi}_i \leq c\ põhjal\ prognoositud$ $katsetulemuse\ väärtus\ \hat{y}_i = 0$	TN	FN
$Otsustamisreegli\ \hat{\pi}_i > c\ põhjal\ prognoositud$ $katsetulemuse\ väärtus\ \hat{y}_i = 1$	FP	TP

TN - tõeselt negatiivsete katsetulemuste arv, ehk nende sündmuste arv, mis prognoosi kohaselt ei tohi toimuda ja ei toimu ka tegelikkuses (ingl. $true\ negative$),
FN -valenegatiivsete katsetulemuste arv, ehk eksklikult negatiivseteks prognoositud katsetulemuste arv (ingl. $false\ negative$),
TP -tõeselt positiivsete katsetulemuste arv (ingl. $true\ positive$),
FP -valepositiivsete katsetulemuste arv, ehk eksklikult positiivseteks prognoositud katsetulemuste arv (ingl. $false\ positive$).

Tundlikkus (ingl. $sensitivity$) näitab, kui suure osa tegelikult positiivsete katsetulemuste arvust ennustab meie poolt kontrollitav meetod õigesti:

$$Tundlikkus = TP/(TP + FN).$$

Spetsiifilisus (ingl. $specificity$) näitab, kui suure osa tegelikult negatiivsete katsetulemuste arvust ennustab meie poolt kontrollitav meetod õigesti:

$$Spetsiifilisus = TN/(TN + FP).$$

Tundlikkuse ja spetsiifilisuse karakteristikuid kasutatakse nn $ROC - analüüsis$ (ingl. $receiver\ operating\ characteristic\ analysis$), mille osaks on nn $ROC - kõver$ (ingl. $receiver\ operating\ characteristic\ curve$). ROC-kõvera punktide koordinaatideks on tundlikkuse ja spetsiifilisuse väärtused (tavaliselt tundlikkuse

väärtused y-teljel ning (1 - spetsiifilus) ehk $valepositiivsete\ katsetulemuste\ määra$ väärtused x-teljel) etteantud c korral.

Siinjuures c väärtusteks võime valida arvutatud tõenäosuse väärtuseid, ehk arvutada ülalkirjeldatud klassifitseerimistabeli ning ka tundlikkuse ja spetsiifilisuse iga prognoositud tõenäosuse väärtuse korral.

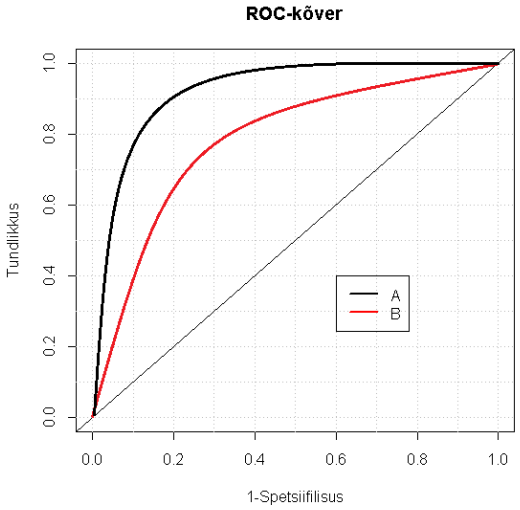
ROC-kõvera abil arvutatakse palju erinevaid karakteristikuid, milliseid kasutatakse uuritava meetodi analüüsimisel. Üheks oluliseks näitajaks on $ROC - kõvera\ alune\ pindala$, (ingl. $area\ under\ the\ curve$, AUC). Antud näitaja/statistik kasutatakse prognoosi täpsuse kirjeldamiseks ning seda võib mitmel moel interpreteerida.

Enam kasutatavaks interpretatsiooniks on:

AUC näitab tõenäosust, et juhuslikult valitud positiivse katsetulemusega vaatluse $y_i = 1$ arvutatud tõenäosus $\hat{\pi}_i$ on suurem, kui juhuslikult valitud negatiivse katsetulemusega vaatluse $y_j = 0$ arvutatud tõenäosus $\hat{\pi}_j$.

Üldiselt, mida suurem on AUC väärtus, seda paremini kontrollitava meetodi poolt arvutatud tõenäosus vastab vaadeldava sündmuse toimumise tegelikule tõenäosusele. Ideaalset olukorda kirjeldav AUC väärtus on 1.

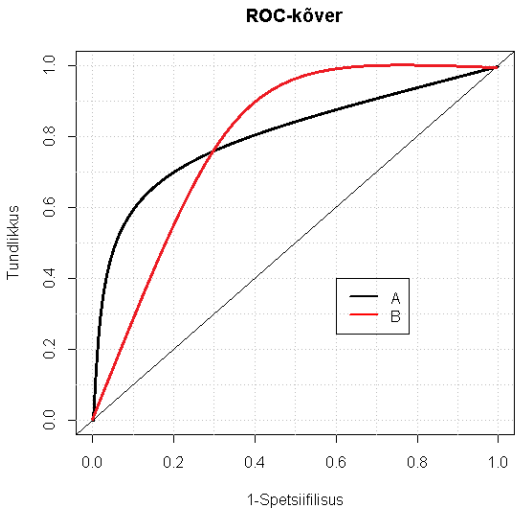
Joonis 16: Kahe erineva uuritava meetodi $ROC - kõverad$.



Jooniselt 16 on näha, et uuritava meetodi A korral on AUC väärtus tunduvalt suurem, kui uuritava meetodi B korral, mille põhjal saab järeldada, et meetodi A abil saab paremini ennustada huvipakkuva sündmuse toimumist.

Kuid alati pole AUC väärtus piisav otsustamaks kumb meetod on parem, sest esinevad olukorrad, kus erinevalt ennustavate meetodite korral saame tulemuseks võrdseid AUC väärtusi. Sellist olukorra kirjeldab joonis 17.

Joonis 17: Kahe erineva uuritava meetodi $ROC - kõverad$.



Joonise 17 korral otsuse langetamisel meetodi A või B kasuks, peab arvesse võtma asjaolu, et mõlemad meetodid prognoosivad keskmiselt võrdse täpsusega uuritava sündmuse toimumise tõenäosust, kuid meetod A tagab kõrgema tundlikkuse, kui spetsiifilisuse näitaja on vahemikkus (0,7,1), meetodi B korral on tundlikkuse näitaja kõrgem, kui spetsiifilisus on väiksem, kui 0,7.

Taolises olukorras tuleb lähtuda uuringu kontekstist, see tähendab valida sobivaim meetod sõltuvalt meetodi rakendamise eesmärgist. Kui meetodi rakendamisel üritatakse vältida valepositiivseid tulemusi, siis eelistavamaks osutub meetod A. Kui aga rohkem ebameeldivamaks loetakse valenegatiivne tulemus, siis tuleb otsustada meetodi B kasuks.

4.2 Imputeerimiskvaliteedi hinnangu hinnang

Antud alamosas kasutame arvutatud AUC- ja Hoshmer-Lemeshew teststatistikuid (või kui meil on palju korduvaid väärtusi, siis $\mathcal{X}^2$-statistiku) programmi IMPUTE2 poolt raporteeritud imputeerimiskvaliteedi hinnangute paikapidavuse kontrollimiseks.

Korrektselt imputeeritud SNP-ide jaoks ei saa teha Hoshmer-Lemeshew testi ega ka arvutada AUC väärtust. Selliste SNP-ide korral leiame hinnatud tõenäosuste (et imputeerimine toimus korrektselt) keskmise.

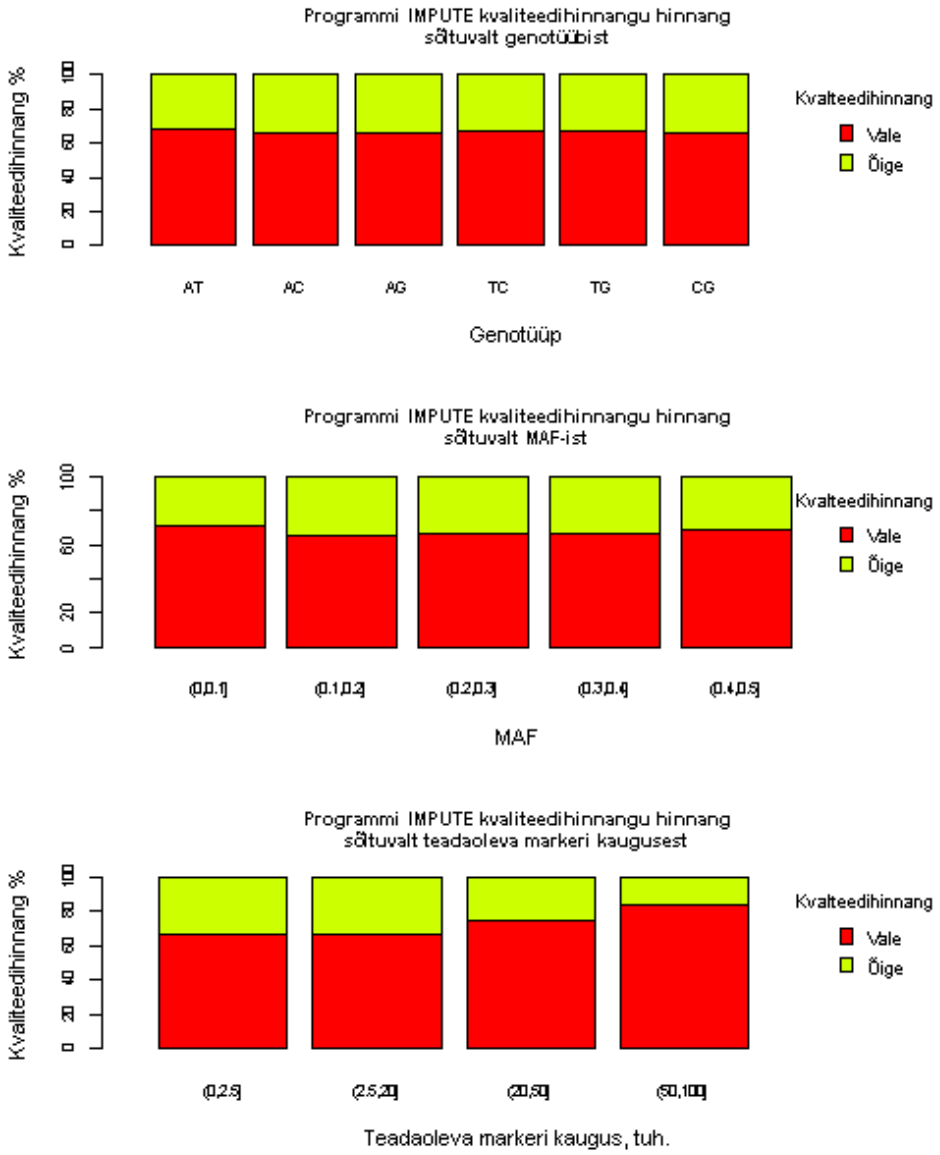
Jagame imputeerimistulemuste andmestiku täiesti korrektseks ja vigu sisaldavaks osadeks (vt tabel 5).

Tabel 6: Imputeerimistulemuste esitamine kahe osadena

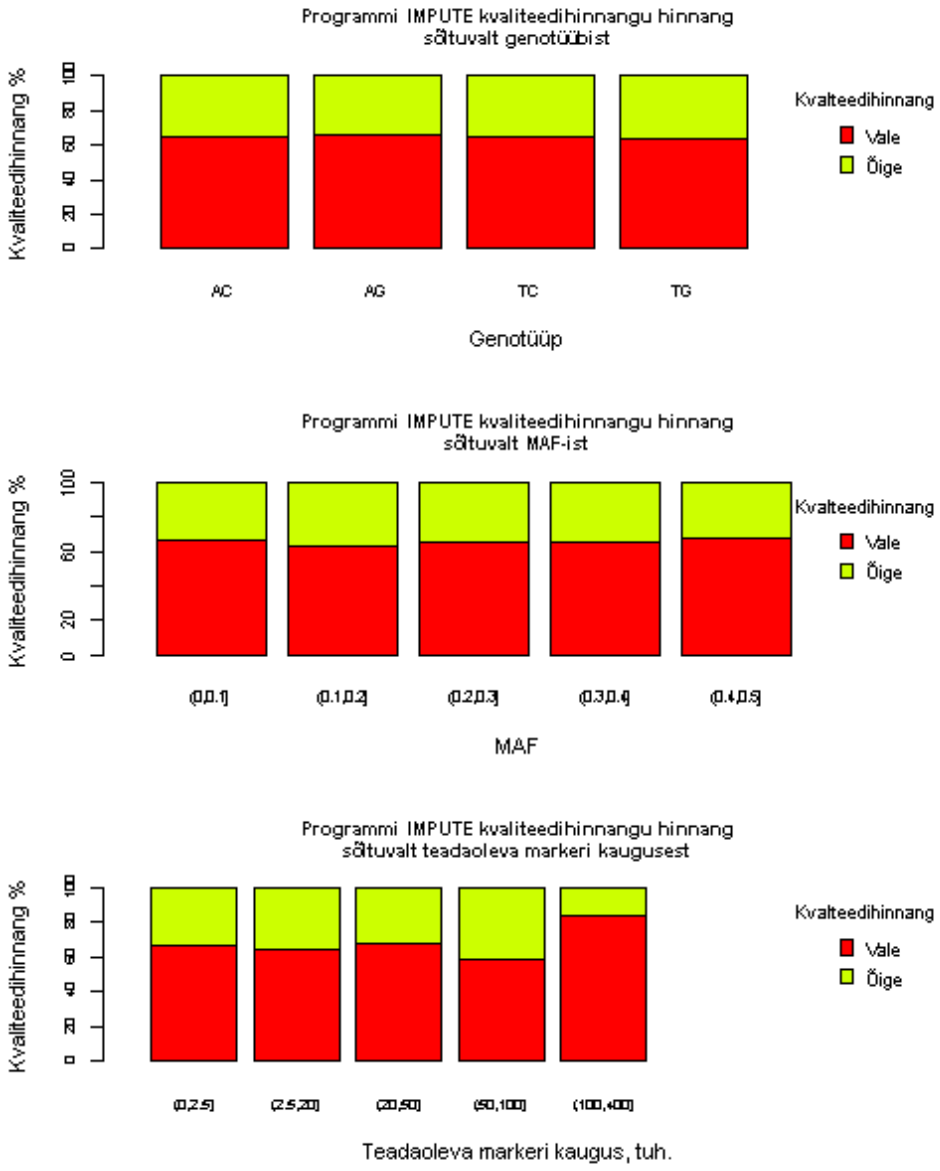
	Täiesti korrektselt imputeeritud SNP-id	%	Vigadega imputeeritud SNP-id	%
$1.imputeerimine$	433316	79	114784	21
$2.imputeerimine$	44293	72	17070	28
$3.imputeerimine$	75840	77	22974	23

Osaliselt vigaselt imputeeritud SNP-ide korral kontrollime kasutades Hoshmer-Lemeshew (või $\mathcal{X}^2$) testi, kas IMPUTE2 poolt arvutatud tõenäosused on õiged (H_0) või valed (H_1). Siinjuures märkame, et kõigi täiesti korrektselt arvutatud SNP-ide jaoks saame $\mathcal{X}^2$ testi olulisuse tõenäosuse põhjal jääda alati H_0 juurde. Nende SNP-ide jaoks, mis on imputeeritud vigadega, esitame Hoshmer-Lemeshew testi tulemused joonistel 18-20.

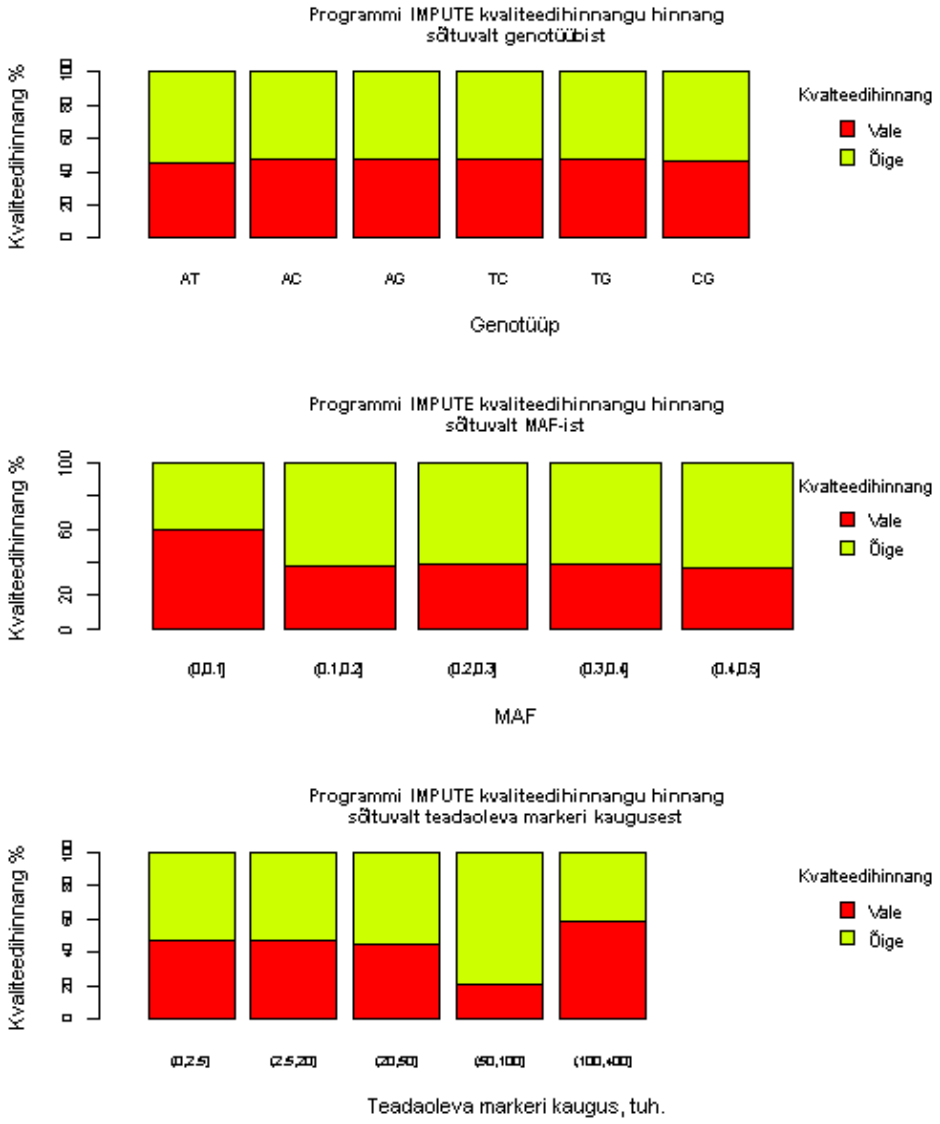
Joonis 18: 1. imputeerimine. Hoshmer-Lemeshew testi tulemus.



Joonis 19: 2. imputeerimine. Hoshmer-Lemeshew testi tulemus.



Joonis 20: 3. imputeerimine. Hoshmer-Lemeshew testi tulemus.



Jooniste 18-20 põhjal saab järeldada, et kvaliteedihinnangute kvaliteet on esimese ja teise imputeerimise korral märkimisväärselt ei erine. Tuletame ka meelde, et joonisel 11 esitatud imputeerimistulemuste prognoosid käituvad samamoodi, ehk ka sarnanevad suurel määral esimese ja teise imputeerimise korral.

On märgata, et nii esimesel, kui ka teisel imputeerimisel väheneb H_0 osakaal (ehk õigeks loetud tõenäosuste osakaal Hoshmer-Lemeshew testi põhjal) imputeeritava SNP-i ja temast lähima teadaoleva markeri vahelise kauguse kasvades (joonised 18-20).

Kolmanda imputeerimise tulemusena korrektselt arvutatud tõenäosuste osakaal on märgatavalt suurem, kui esimese ja teise imputeerimise puhul (vt. joonis 11).

AUC statistikut kasutame kontrollimaks, kas edukalt imputeeritud genotüüpidele antakse paremaid kvaliteedihinnanguid, kui vigaselt imputeeritud genotüüpidele. Uurime ka AUC statistiku sõltuvust SNP-i iseloomustavatest pidevatest tunnustest ja SNP-i alleelidest.

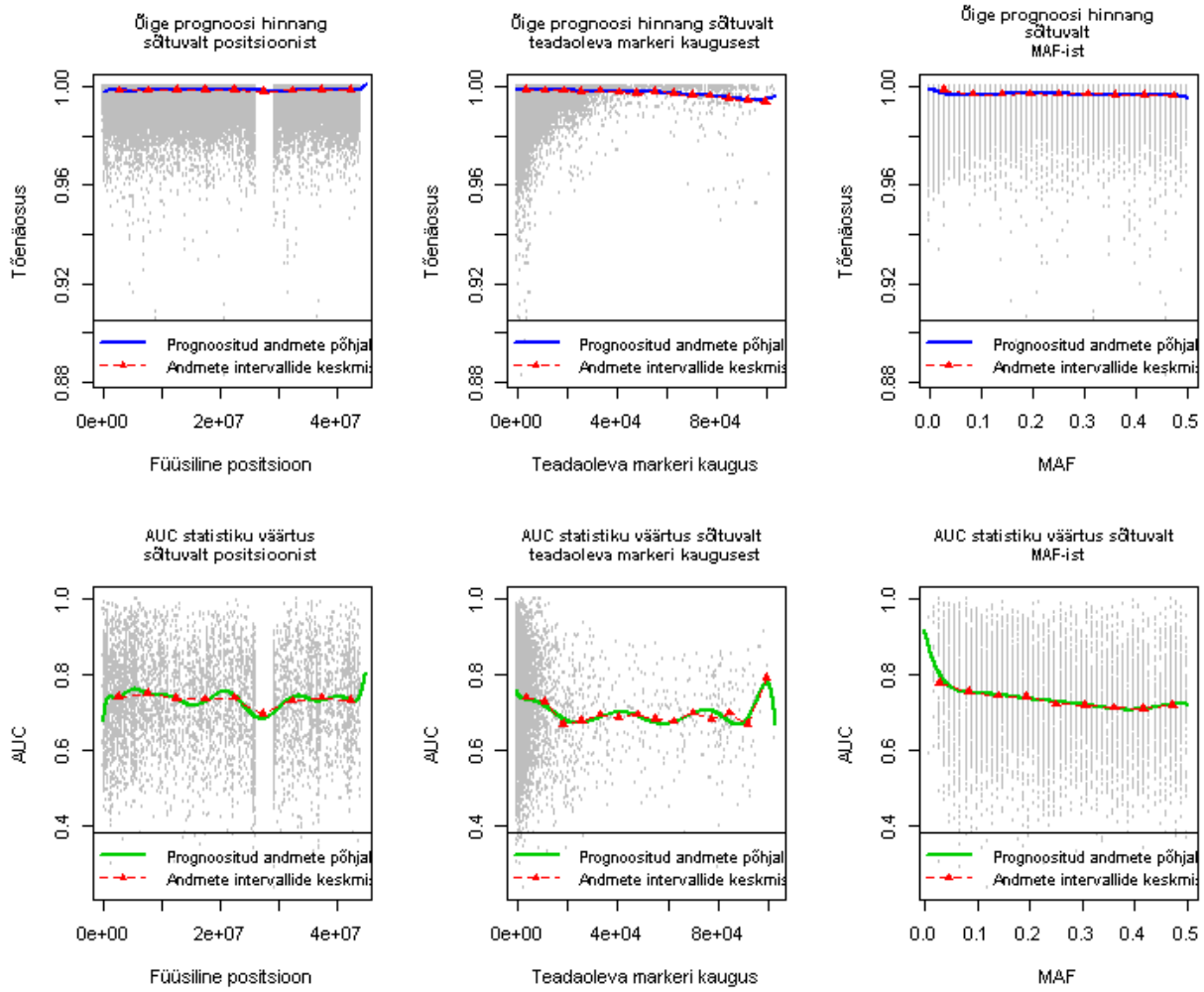
Siinjuures märkame, et ilmselt ei ole mõtet arvutada AUC väärtust nende SNP-ide jaoks, kus imputeerimine on alati olnud edukas või alati osutunud valeks, mis omakorda tähendab, et tundlikkuse ja spetsiifilisuse näitajaid on raske hinnata.

Tõepoolest, tinglikku tõenäosuse
$P(Väidatavalt\ valesti\ arvutatud\ haplotüüp \mid Tegelikult\ valesti\ arvutatud\ haplotüüp)$
väärtus tuleb arvatavasti suhteliselt ebatäpne, kui $Tegelikult\ valesti\ arvutatud\ haplotüüpide$ arv on väike.

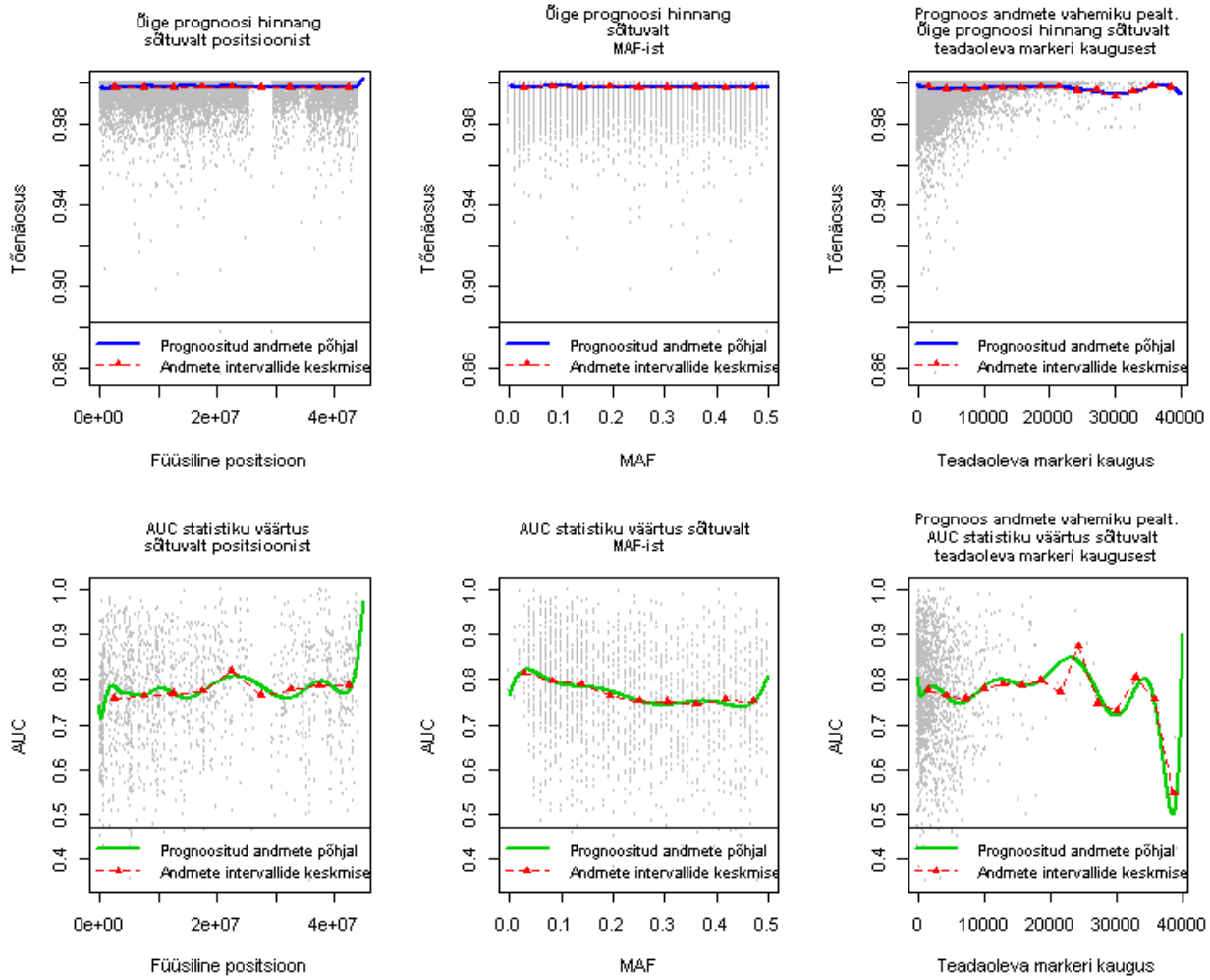
Seega, leiame AUC väärtuse vaid nende SNP-ide jaoks, kus on selgelt eristuvad valed ja õiged imputeerimistulemused (näiteks, kus nii korrektselt, kui ka valesti imputeeritud genotüüpe olivähemalt 10% kõigist genotüüpidest).

Joonistel 21-23 on esitatud nii täiesti korrektselt arvutatud SNP-ide jaoks IMPUTE poolt hinnatud tõenäosuste keskmine kui ka vigadega imputeeritud SNP-ide jaoks arvutatud AUC väärtused. Korrektselt imputeeritud SNP-ide puhul on IMPUTE2 poolt raporteeritud korrektse imputeerimise tõenäosuste keskmine ootuspäraselt peaaegu 1.

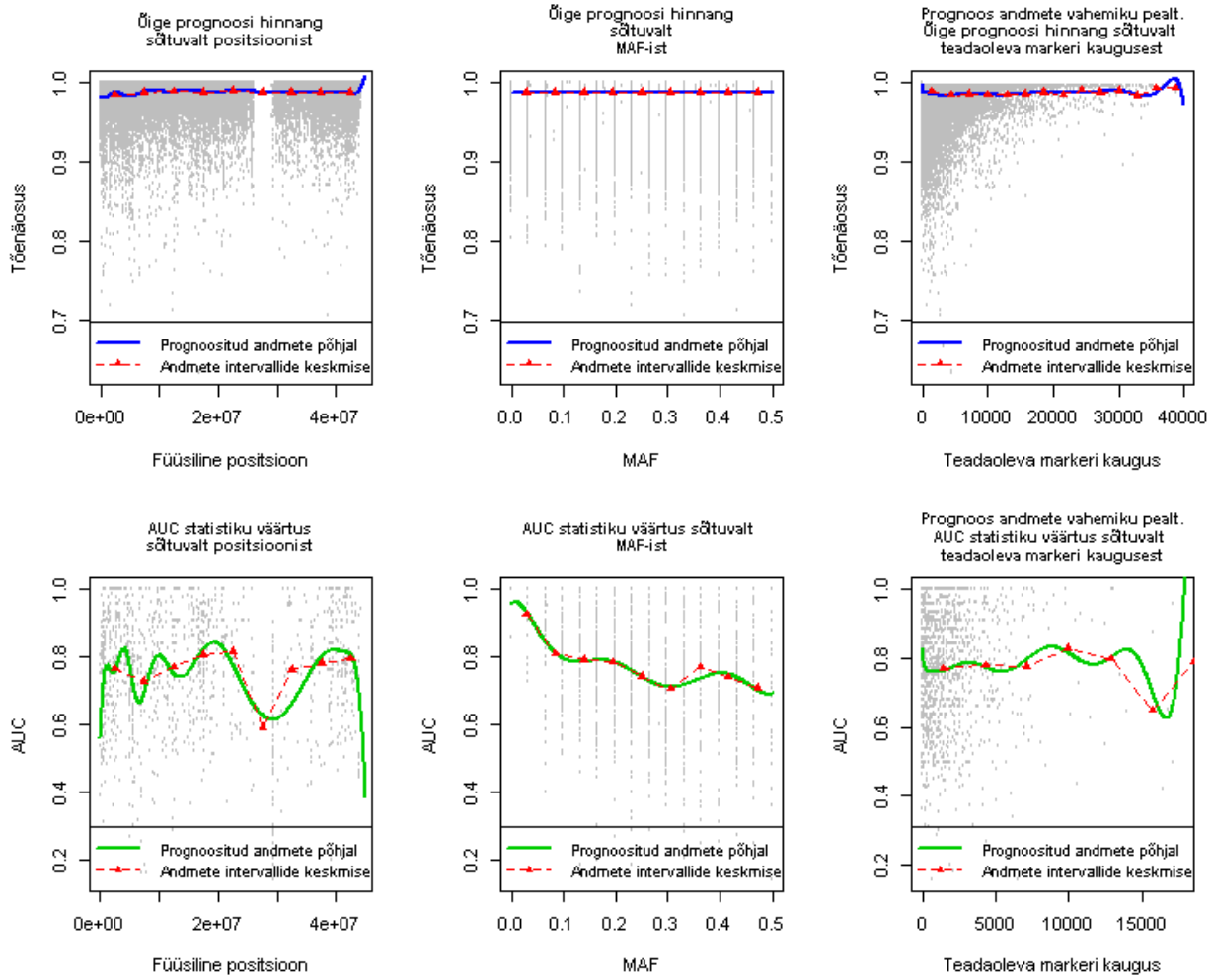
Joonis 21: 1. Imputeerimine. Imputeerimiskvaliteedi hinnangute kvaliteedi sõltuvus erinevatest SNP-i iseloomustavatest pidevatest tunnustest.



Joonis 22: 2. Imputeerimine. Imputeerimiskvaliteedi hinnangute kvaliteedi sõltuvus erinevatest SNP-i iseloomustavatest pidevatest tunnustest.



Joonis 23: 3. Imputeerimine. Imputeerimiskvaliteedi hinnangute kvaliteedi sõltuvus erinevatest SNP-i iseloomustavatest pidevatest tunnustest.



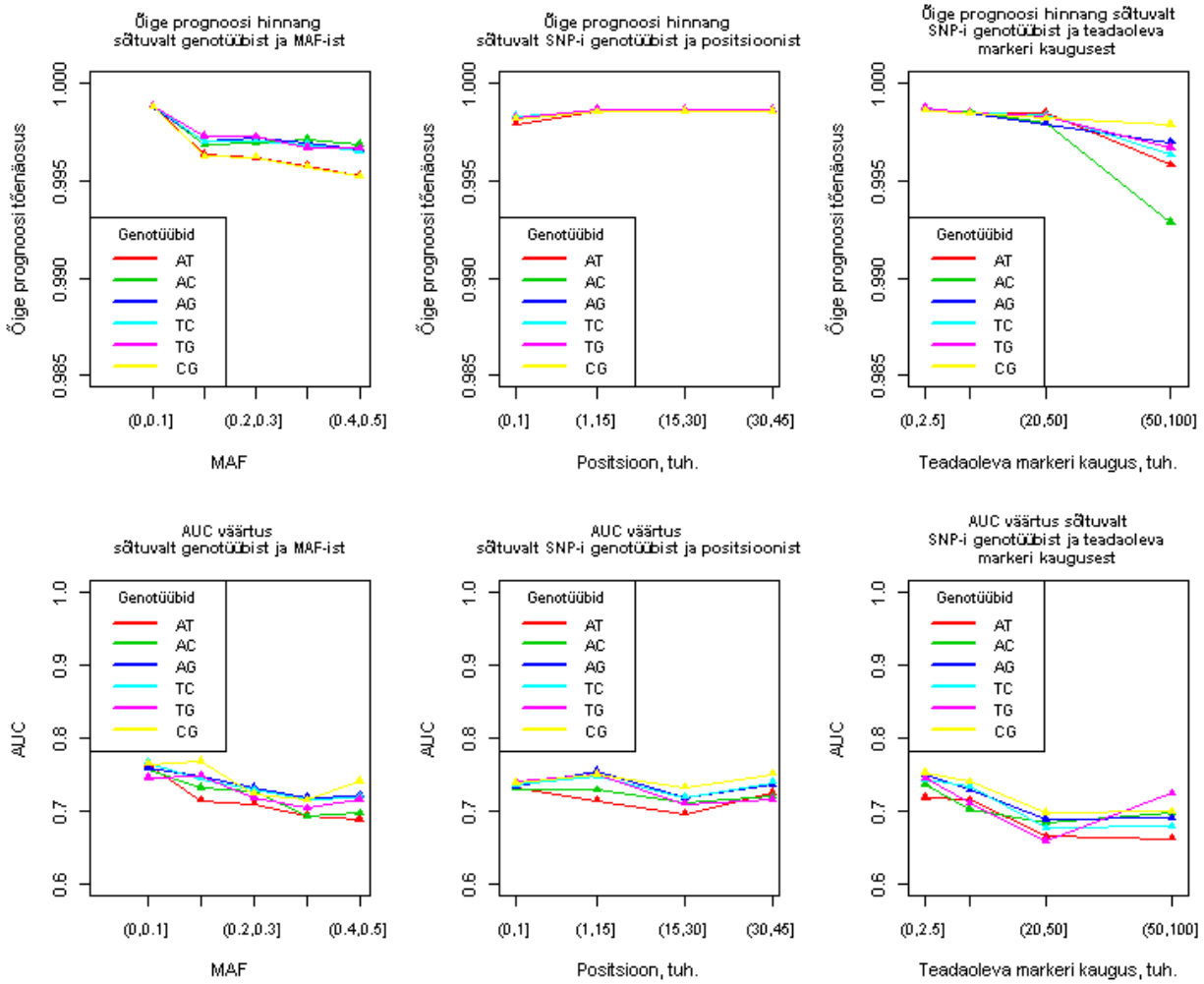
AUC väärtuste prognoosid ehk imputeerimiskvaliteedi hinnangu prognoosid (vt. joonised 21-23, rohelised kõverad) käituvad väga sarnaselt imputeerimise kvaliteedi prognoosidega (vt. joonised 8-11): selgelt eristub AUC väärtuste vähenemine, ehk kvaliteedihinnangute kvaliteedi langus tsentromeeri piirkonnas ja kromosoomi alguses.

Kõigi imputeerimiste korral on märgata AUC väärtuste kahanemist MAF-i kasvades (joonised 21-23).

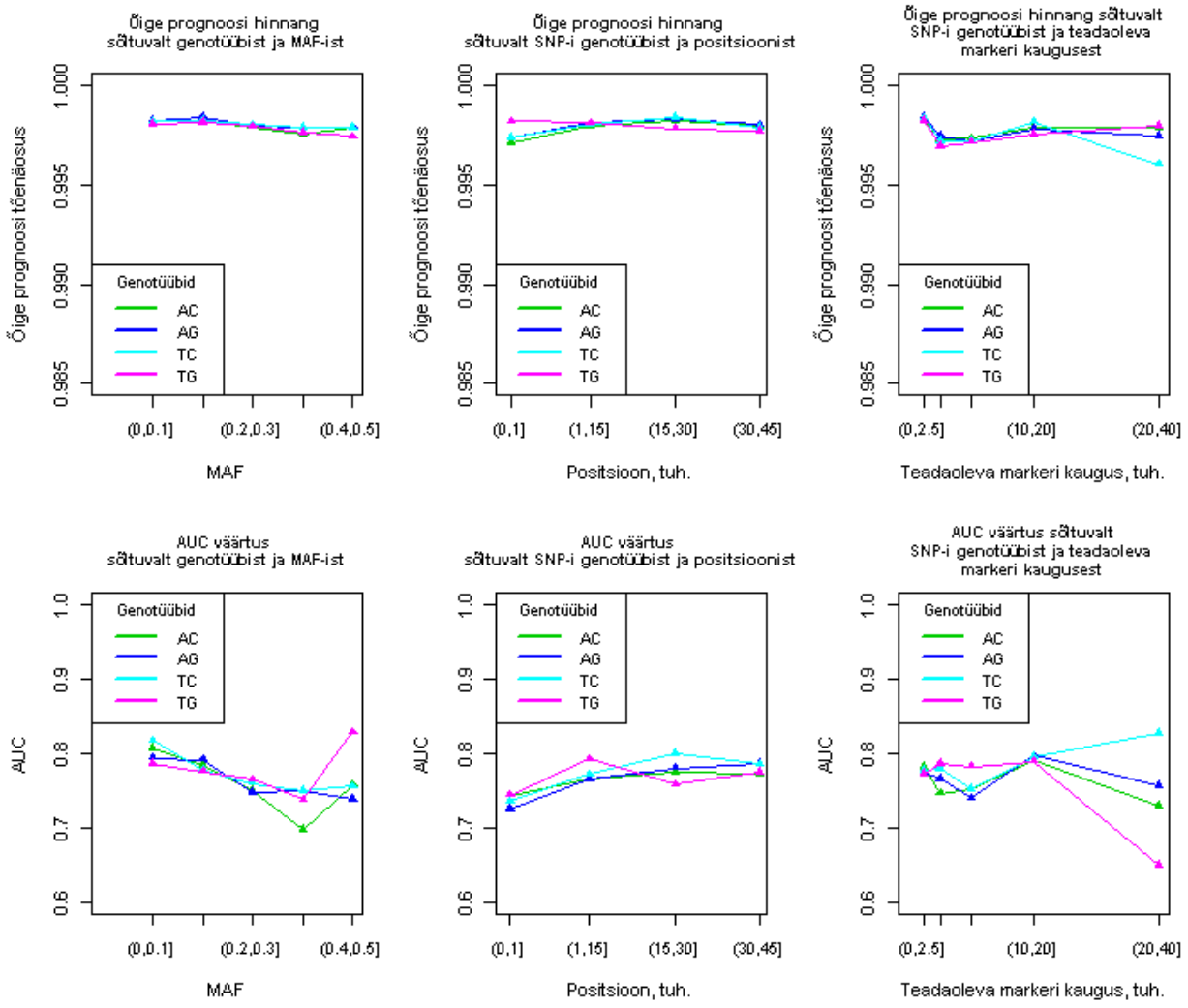
Esimese imputeerimise korral lähima teadaoleva markeri kauguse mõju imputeerimise kvaliteedihinnangu kvaliteedile (ehk AUC prognoosile) on kahaneva iseloomuga prognoosikõvera algosas, kus vaatluste arv on piisavalt suur, samal ajal teisel imputeerimisel kirjeldatud mõju pigem puudub ning kolmandal ta on vastusuunaline (joonised 21-23, rohelised jooned). Taoline imputeerimiskvaliteedi hinnangu kvaliteedi käitumine täielikult vastab imputeerimisvea prognoosikõverate käitumisele (vt. 3.2 Imputeerimise kvaliteet, joonis 11), mille tõlgendus on toodud lk 24.

Viimasena on esitatud imputeerimiskvaliteedi hinnangute kvaliteedi sõltuvus pidevatest tunnustest genotüüpide lõikes (joonised 24-26). Siinjuures märkame, et teise ja kolmanda imputeerimise korral vaatleme tunnuse “teadaoleva markeri kaugus” väärtusi, mis ei ületa 40000 aluspaari, sest antud tunnuse jaotuse põhiline mass asetseb just antud vahemikus (esimesel imputeerimisel veelgi kitsamas vahemikus, vt. joonis 7).

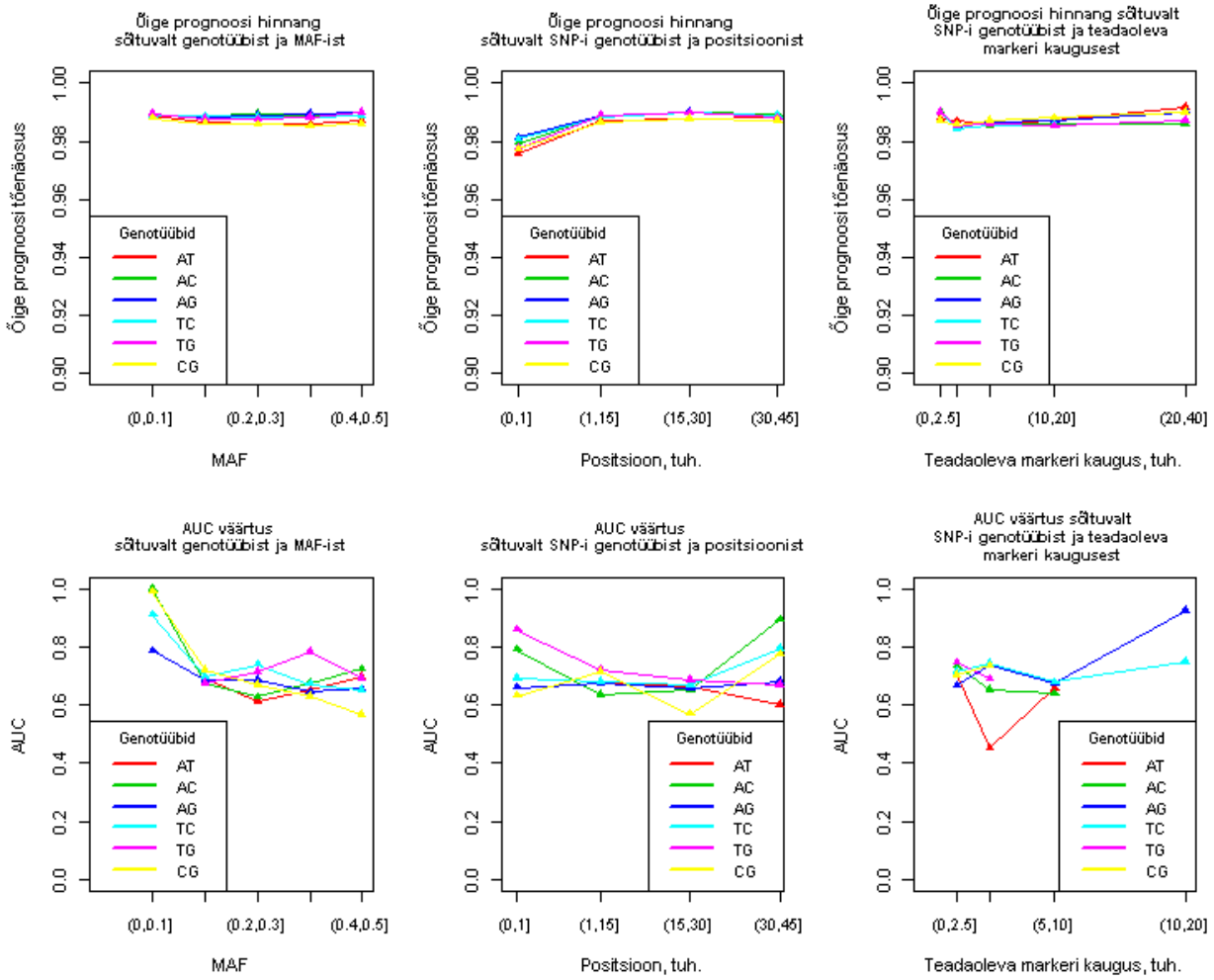
Joonis 24: 1. Imputeerimine. Imputeerimiskvaliteedi hinnangute kvaliteedi sõltuvus SNP-i iseloomustavatest pidevatest tunnustest ja MAF-ist.



Joonis 25: 2. Imputeerimine. Imputeerimiskvaliteedi hinnangute kvaliteedi sõltuvus SNP-i iseloomustavatest pidevatest tunnustest ja MAF-ist.



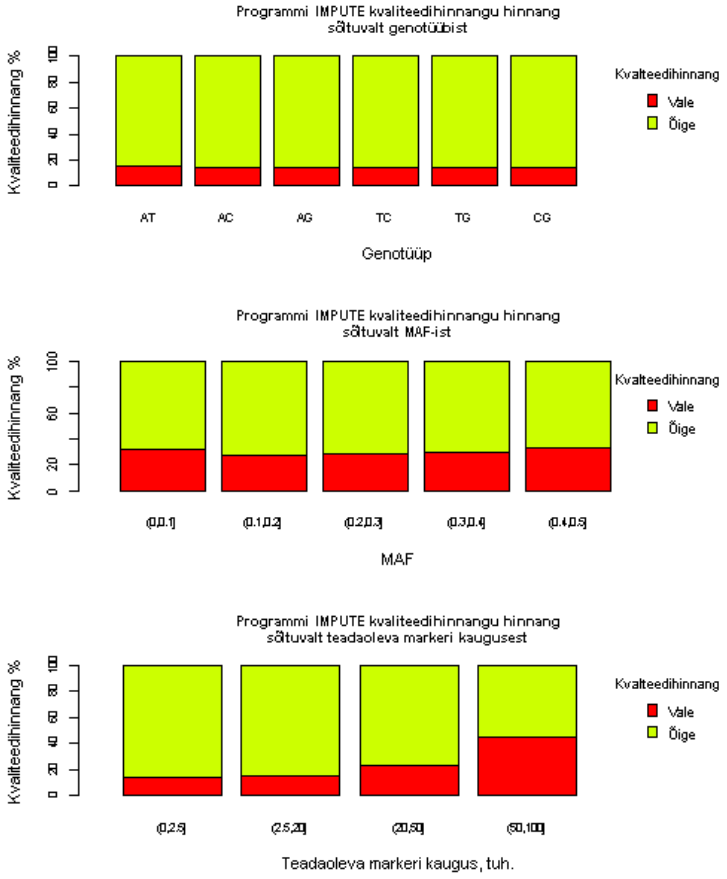
Joonis 26: 3. Imputeerimine. Imputeerimiskvaliteedi hinnangute kvaliteedi sõltuvus SNP-i iseloomustavatest pidevatest tunnustest ja MAF-ist.



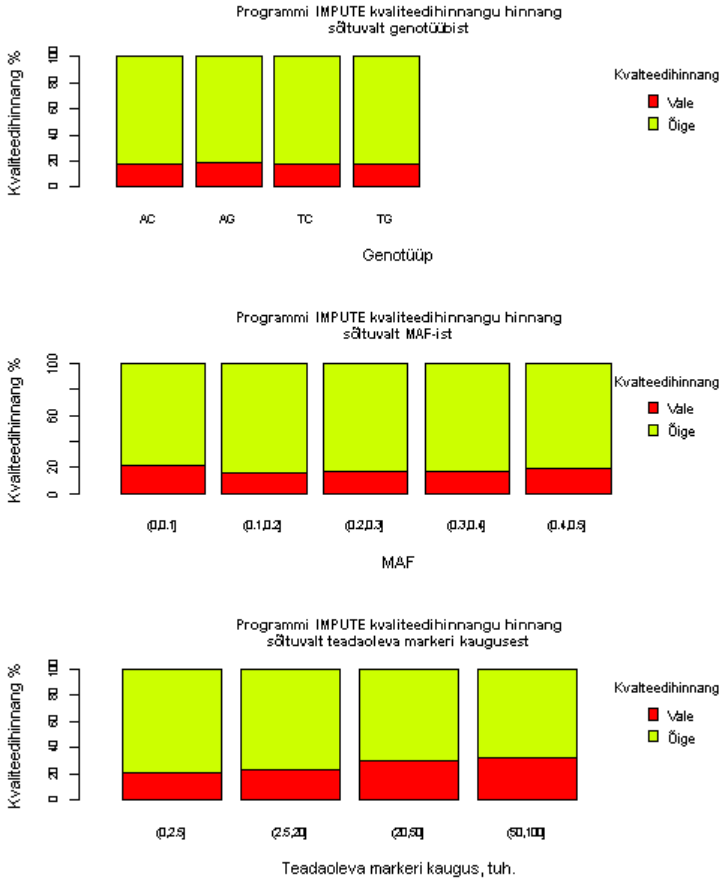
Peamine, mis hakkab silma graafikutel 22-24 on täielik vastavus graafikutele, mis kirjeldavad imputeerimise kvaliteedi sõltuvust nimetatud tunnustest (joonised 13-15). Antud asjaolu näitab, et imputeerimiskvaliteedi langedes langeb ka kvaliteedihinnangute kvaliteet. Kui antud SNP-i ei õnnestu kuigi hästi imputeerida, siis ei tasu uskuda ka IMPUTE2 raporteerimist, et antud indiviidi genotüübis on ta kindel, kuid järgmise indiviidi genotüübi ei tea, sest tegelikkus võib olla vastupidi).

Lisaks sellele tuletame meelde, et enamus SNP-idest on täiesti korrektselt imputeeritud (vt. tabel 5) ning nende SNP-ide jaoks arvutatud tõenäosuste hinnang on õige. Vigadega imputeeritud SNP-ide imputeerimiskvaliteedi korral raporteerib ka IMPUTE2 korrektse imputeerimise tõenäosuse olevat kõigil indiviididel ligikaudu 1. Seega vigadelt imputeeritud SNP-ide imputeerimiskvaliteedi hinnangu kvaliteet ei ole kõrge (vt. joonised 18-20), kuid vaadates imputeerimiskvaliteedi hinnangut kõigi SNP-ide pealt (vt. joonised 27-29), jõuame järelduseni, et programm IMPUTE2 annab üsnagi täpseid hinnanguid imputeerimiskvaliteedile.

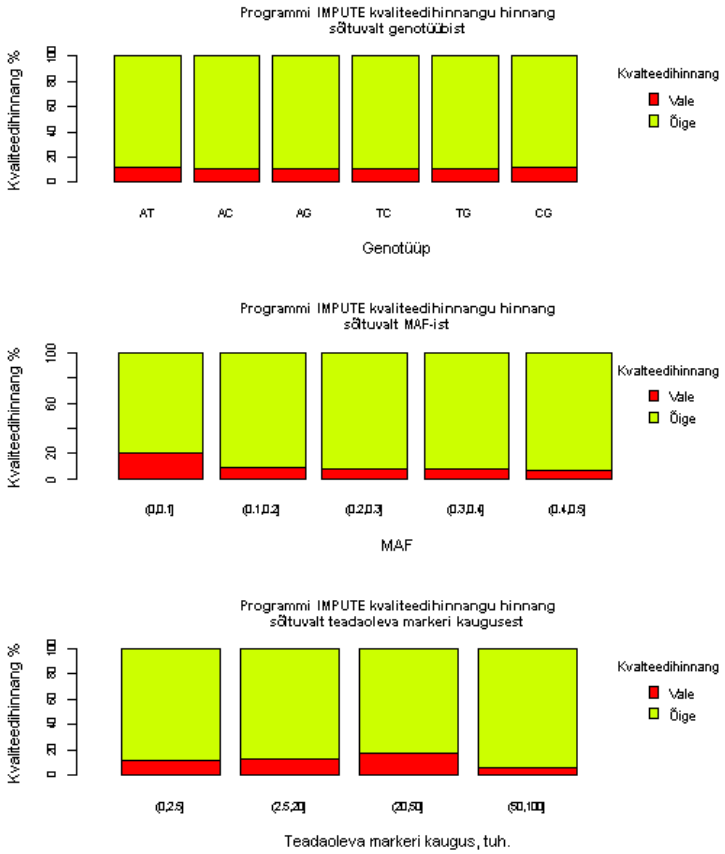
Joonis 27: 1. imputeerimise Hoshmer-Lemeshew testi tulemus koguandmete pealt.



Joonis 28: 2. imputeerimise Hoshmer-Lemeshew testi tulemus koguandmete pealt.



Joonis 29: 3. imputeerimise Hoshmer-Lemeshew testi tulemus koguandmete pealt.



5 Kokkuvõte

Tänapäevases geeniteaduses kasutatakse inimeste genoomi imputeerimiseks erinevaid meetodeid ja programme, selliseid nagu IMPUTE2, MACH, fastPHASE või BIMBAM, EMINIM, BEAGLE. Üheks levinumaks nendest võib pidada programmi IMPUTE2, mille viimane versioon ühendab endas palju võimalikke imputeerimisega seotud protseduure ning imputeerimisstsenaariume.

Antud töö põhiliseks eesmärgiks on uurida programmi IMPUTE2 abil teostatava imputeerimisprotsessi iseloomu, kontrollida programmi abil imputeeritud geneetiliste markerite kvaliteeti ning hinnata programmi poolt väljastatavate kvaliteedihinnangute kvaliteeti.

Imputeerimisprotsessi põhiline ülesanne seisneb määramata jäänud geneetiliste markerite (enamasti SNP-ide) ennustamises, kusjuures olulisemaks etapiks on teadaolevate SNP-ide haplotüpiseerimine. Antud ülesande lahendamiseks kasutab programm IMPUTE2 varjatud Markovi mudeli, mida rakendatakse uuritava valimi iga indiviidi haplotüübi määramiseks, ning määratud haplotüüpide põhjal puuduolevate SNP-ide genotüüpide imputeerimiseks. Töös püstitaud eesmärgi saavutamiseks uuriti, kuidas kasutatakse programmis mainitud varjatud Markovi mudelit ning teostati imputeerimisprotsessi kolmel erineval tingimusel:

1. Esimest imputeerimist teostati nö ideaaltingimustes, kus referentspaneelina kasutati 1000 Genomes Phase I Integrated referenshaplotüüpe ning uuritav valim moodustati 100-st juhuslikult valitud haplotüübist referenshaplotüübide hulgast.

2. Teist imputeerimisprotsessi teostati kasutades sama referentspaneeli, mis esimesel imputeerimisel ehk 1000 Genomes Phase I Integrated referenshaplotüüpe, kuid uuritav valim moodustati eestlaste sekveneeritud andmetest. Eesmärgiks oli kontrollida eestlaste genotüübi ennustamise kvaliteeti eurooplaste referentshaplotüüpe kasutades. Taolise imputeerimisstsenaariumi, kus valimi ja referentspaneeli andmed pärinevad erinevatest populatsioonidest, kasutatakse laialdaselt.

3. Kolmanda imputeerimise jaoks kasutati nii referentspaaneelina, kui ka valimi moodustamiseks eestlaste sekveneeritud andmeid, mis võimaldas kont-

rollida eestlaste genotüüpide ennustamise kvaliteeti eestlaste referentshaplotüüpe kasutades (referentspaneeli saamiseks haplotüpiseeriti eestlaste genotüübid programmi IMPUTE2 abil).

Imputeerimistulemuste põhjal arvutati imputeerimise kvaliteeti kirjaldavaid näitajaid, näiteks leiti valesti imputeeritud genotüüpide osakaal iga imputeeritava SNP-i jaoks.

Saadud imputeerimistulemuste analüüsi käigus uuriti imputeerimiskvaliteedi sõltuvust väljaarvutatud imputeeritava SNP-i ja temast lähima teadaoleva markeri vahelisest kaugusest, minoorse alleeli sagedusest ja alleelidest logistilise ja lineaarse regressiooni abil, lisades prognoosikõveratele SNP-i kirjeldavate näitajate keskmiseid väärtusi.

Imputeerimisekvaliteedi programmi IMPUTE2 poolt arvutatava hinnangu kontrollimiseks arvutati Hoshmer-Lemeshew ja AUC teststatistikuid.

Nii imputeerimiskvaliteet, kui ka imputeerimishinnangu kvaliteet sõltuvad SNP-i kirjeldavatest tunnustest ühtemoodi:

1. Füüsiline positsioon genoomis ei mõjuta märkimisväärselt tulemust iga imputeerimise korral (kuid imputeerimise kvaliteet ja kvaliteedihinnangute kvaliteet on madalam tsentromeeride lähistel ja kromosoomi algusosas).

2. MAF-i mõju on erinev erinevatel imputeerimisel:
   esimesel imputeerimisel MAF-i kasv langetab nii imputeerimiskvaliteeti, kui ka imputeerimishinnangu kvaliteeti; teisel imputeerimisel nii imputeerimiskvaliteet, kui ka imputeerimishinnangu kvaliteet MAF-ist peaaegu ei sõltu; kolmandal imputeerimisel nii imputeerimiskvaliteet, kui ka imputeerimishinnangu kvaliteet tõusevad MAF-i kasvades, mis seletatakse sellega, et suure varieeruvusega SNP-id asetsevad suhteliselt lähedal uuritava valimi teadaolevast markerist.

3. Kaugus lähimast teadaolevast markerist mõjtaub sarnaselt imputeerimiskvalitedi iga imputeerimise korral - tema väärtuste kasvades, langeb imputeeri-

miskvaliteet ning ka kvaliteedihinnangu kvaliteet.

Analüüsides nimetatud kvaliteedinäitajaid ning kvaliteedihinnangut iseloomustavaid teststatistikuid, leiti, et:

1. IMPUTE2 tarkvara kasutades saab eestlaste genotüüpe imputeerida sama edukalt nii eurooplaste haplotüübide abil kui ka eestlaste sekveneeritud andmete abil.

2. Imputeerimisekvaliteet suurel määral vastab programmi IMPUTE2 poolt antud imputeerimisekvaliteedi hinnangule.

IMPUTATION OF GENETIC MARKERS

Master's Thesis

Tatjana Iljashenko

Summary

There are quite a few of different methods that have been proposed for human genotype imputation in genetic research to date. Some of these statistical methods for imputing genotypes (for example IMPUTE2, MACH, fastPHASE or BIMBAM, BEAGLE) are widely used in the analysis of genome-wide association studies. But one of the most common methods is provided by program for phasing observed genotypes and imputing missing genotypes, named IMPUTE2.

The latest versions of IMPUTE2 supply a flexible modelling framework that increases accuracy and combines information across multiple reference panels while remaining computationally feasible.

The general goal of the present thesis is to provide an overview of process of predicting (or imputing) genotypes that are not directly assayed in a sample of individuals and to discuss and illustrate the factors that affect the accuracy of genotype imputation.

First of all, we consider the method, used by IMPUTE2, that is based on an Hidden Markov Models of each individual′s vector of genotypes, conditional on reference haplotypes and set of parameters. Additionally, we explore in detail using of Hidden Markov Models for genotype imputation by example.

The main focus of present study is carrying out of genotype imputation in three distinct situations based on the conditions:

1. In the first imputation we assume "ideal" condition and use reference dataset of the 1000 Genomes Project as a reference panel and collect a study sample from the same reference dataset by random (it means, that our study sample and reference haplotypes are collected from the same population).

2. For the second imputation we combine a reference panel by using the 1000 Genomes Project, but the study sample is composed from an Estonian popu-

lation by random (the study sample is genotyped at a set of specified SNPs, which is received from Estonian Genome Bank).

3. For the third imputation we use the same study sample as for the second imputation, but the reference panel consist of individuals genotyped at a set of SNPs , which come from Estonian Genome Bank.

Finally we estimate the quality of imputation by calculation and analysing values of minor allele frequency, smallest interval between imputing SNPs and known genetic locus, percentage of wrongly imputed haplotypes et cetera. After that we model the dependence of wrongly imputed haplotypes with minor allelle frequency, smallest interval between imputing SNPs and known genetic locus and physical positsion by using logistic regression.
Additionally, we estimate through calculating and analysing Hosmer-Lemeshow statistics and AUC values the estimating probabilities of imputed genotypes by using linear regression.

To illustrate the estimation of quality of estimation of imputations we show the results of Hosmer-Lemeshow test graficly.

The main conclusions are:

1. In our experiment the imputation of genotype of Estonian study sample is as success as the imputation of genotype of European study sample using the European reference panel.

2. The IMPUTE2 estimates the quality of estimation of imputation well, it means, that when the results of imputation are wrong, IMPUTE2 reports the small probability to see these genotypes. And vice versa.

Viited

[1] Human Genome Project Information. SNP Fact Sheet. [WWW]
$http://www.ornl.gov/sci/techresources/Human\_Genome/faq/snps.shtml$.

[2] IMPUTE2 koduleht. [WWW]
$https://mathgen.stats.ox.ac.uk/impute/impute\_v2.html\#reference$.

[3] Vikipeedia. [WWW]
$http://et.wikipedia.org/wiki/Sekveneerimine$.

[4] Kurg, A. $Geenivaramu: mis, kuidas ja milleks$? [WWW]
$http://www.loodusajakiri.ee/eesti\_loodus/EL/vanaweb/0008/geen.html$.

[5] Remm, M. $Bioinformaatika\ ja\ genotüpiseerimine$ (ainekursuse konspekt). [WWW]
$http://www.cs.ut.ee/\ varmo/tday-arula/remm-slides.pdf$.

[6] Marchini, J., Howie, B. $Genotype\ imputation\ for\ genome-wide\ association\ studies$. 2010. Macmillan Publishers Limited. [WWW]
$http://home.uchicago.edu/\ bhowie/papers/\ marchini\_howie\_nat\_rev\_genet\_2010pdf$.

[7] Howie, B., Donnelly, P., Marchini, J. $A\ Flexible\ and\ Accurate\ Genotype\ Imputation\ Method\ for\ the\ Next\ Generation\ of\ Genome-Wide\ Association\ Studies$. [WWW]
$http://www.plosgenetics.org/article/info\%3Adoi\%2F10.1371\%2Fjournal.pgen.1000529$.

[8] Marchini, J., Howie, B., Myers, S., McVean, G., Donnelly P.
$A\ new\ multipoint\ method\ for\ genome-wide\ association\ studies\ via\ imputation\ of genotypes: Supplementary\ Methods$. [WWW]
$http://www.nature.com/ng/journal/v39/n7/extref/ng2088-S4.pdf$.

[9] The International HapMap Consortium. The Phase II Hapmap. 2007.

[10] Li, N., Stephens, M. $Modelling\ linkage\ disequilibrium,\ and\ identifying\ recombination\ hotspots\ using\ snp\ data$. 2003. Genetics, 165:2213Ű2233.

[11] $Kromatiinjakromosoomid$. [WWW] $http://cellbio.ebc.ee/rakubio/kromatii.html$.

[12] $Chi-square:\ Testing\ for\ goodness\ of\ fit$. lk 7. [WWW] $http://physics.ucsc.edu/\ drip/133/ch4.pdf$

[13] Archer K. J., Lemeshow, S. $Goodness-of-fit\ test\ fo\ a\ logistic\ regression\ model\ fitted\ using\ survey\ sample\ data$. 2006. The Stata Journal, Number 1, lk. 99.

Lisad

Joonis 30: Magistritöös kasutatavate koodide fragmendid. Osa 1.

```
#Imputeeritava SNP-i ja temast lähima teadaoleva markeri vahelise kauguse arvutamise kood.
olgu refpos  on referentsandmestiku markerite f"u"usilise positsioonide vektor
olgu valimpos on uuritava valimi SNP-ide f"u"usilise positsioonide vektor

valimpos=c(valimpos,0)
minvahe=c(0)        # see on otsitav vektor, deklareerime
i=1
for (j in 1:length(refpos)){
if (abs(valimpos[i]-refpos[j]) <= abs(valimpos[i+1]-refpos[j])){
minvahe[j]=abs(refpos[j]-valimpos[i])}
if (abs(valimpos[i]-refpos[j])>abs(valimpos[i+1]-refpos[j])){
minvahe[j]=abs(refpos[j]-valimpos[i+1])
i=i+1}
}

Programmilõik, mis arvutab m~oned anal"u"usis kasutatavad n"aitajad:


tulem=matrix(rep(0,12),4,3)# valmistab ette maatriksi, mille dimensiooniks tuleb

a0a0_kokku=0 # $a_0a_0$ genot"u"upide arv SNP-is
a0a1_kokku=0 # $a_0a_0$ genot"u"upide arv SNP-is
a1a1_kokku=0 # $a_0a_0$ genot"u"upide arv SNP-is
a0a0_0=0    # nende genot"u"upide arv SNP-is, mille kohta ei saa otsustada, millise genot"u"upiga ta on

m=0
i=1
hii=rep(NA,dim(vastus)[1])# $vastus$ on imputeerimisvastuste maatriks
MAF=c(0)              # imputeerimistulemuste MAF
teg_MAF=c(0)          # referentspaneeli MAF
Vale_gen_arv=c(0)       # valesti arvutatud genot"u"upide arv SNP-is
Kadu=c(0)             # nende SNP-ide vektor, mille kohta IMPUTE

while (i<dim(vastus)[1]+1) {
qhap=as.vector(v[i+m,])   # vas on referentspaneeli tabel (haplot"u"ubide kujul, 1 indiviidi kohta 2 numbri)
imp=as.vector(vas[i,])    # v on imputeerimisvastuste tabel (genot"u"ubi kujul, 1 indiviidi kohta 3 numbri)

Q=matrix(qhap,ncol=2,byrow=TRUE)
Imp=matrix(imp,ncol=3,byrow=TRUE)

a0a0=Imp[(Q[,1]==0 & Q[,2]==0),]# eraldi tegelik a0a0
a1a1=Imp[(Q[,1]==1 & Q[,2]==1),]# eraldi tegelik a0a1
a0a1=Imp[(Q[,1]==1 & Q[,2]==0|Q[,1]==0 & Q[,2]==1),]# eraldi tegelik a1a1


a0a0_dim=length(a0a0)/3
a0a1_dim=length(a0a1)/3
a1a1_dim=length(a1a1)/3


a0a0_a0a0=0
a0a0_a0a1=0
a0a0_a1a1=0
```

Joonis 31: Magistritöös kasutatavate koodide fragmendid. Osa 2.

```
a0a1_a0a0=0
a0a1_a0a1=0
a0a1_a1a1=0

a1a1_a0a0=0
a1a1_a0a1=0
a1a1_a1a1=0
#a0a0_kadu=0

# vaatame tegelikku a0a0 jaotust - palju seal arvutatud a0a0,a0a1,a1a1
if (a0a0_dim==1){
        if (min(a0a0)!= max(a0a0)){
                if (which.max(a0a0)==1){a0a0_a0a0=1}
                if (which.max(a0a0)==2){a0a0_a0a1=1}
                if (which.max(a0a0)==3){a0a0_a1a1=1}}}

if (a0a0_dim !=1){
        a0a0_a0a0=length(a0a0[a0a0[,1]>a0a0[,2]& a0a0[,1]>a0a0[,3],1])
        a0a0_a0a1=length(a0a0[a0a0[,2]>a0a0[,1]& a0a0[,2]>a0a0[,3],2])
        a0a0_a1a1=length(a0a0[a0a0[,3]>a0a0[,1]& a0a0[,3]>a0a0[,2],3])}
        #a0a0_kadu=length(a0a0[a0a0[,3]==a0a0[,2] & a0a0[,2]==a0a0[,1] |a0a0[,3]==a0a0[,2] &
a0a0[,3]>a0a0[,1]|a0a0[,3]==a0a0[,1] & a0a0[,3]>a0a0[,2]|a0a0[,1]==a0a0[,2] & a0a0[,2]>a0a0[,3],3])}


# vaatame tegelikku a0a1 jaotust - palju seal arvutatud a0a0,a0a1,a1a1
if (a0a1_dim==1){
        if (min(a0a1)!= max(a0a1)){
                if (which.max(a0a1)==1){a0a1_a0a0=1}
                if (which.max(a0a1)==2){a0a1_a0a1=1}
                if (which.max(a0a1)==3){a0a1_a1a1=1}}}
if (a0a1_dim!=1)
{a0a1_a0a0=length(a0a1[a0a1[,1]>a0a1[,2]& a0a1[,1]>a0a1[,3],1])
        a0a1_a0a1=length(a0a1[a0a1[,2]>a0a1[,1]& a0a1[,2]>a0a1[,3],2])
        a0a1_a1a1=length(a0a1[a0a1[,3]>a0a1[,1]& a0a1[,3]>a0a1[,2],3])}


# vaatame tegelikku a1a1 jaotust - palju seal arvutatud a0a0,a0a1,a1a1
if (a1a1_dim==1){
        if (min(a1a1)!= max(a1a1)){
                if (which.max(a1a1)==1){a1a1_a0a0=1}
                if (which.max(a1a1)==2){a1a1_a0a1=1}
                if (which.max(a1a1)==3){a1a1_a1a1=1}}
if (a1a1_dim!=1)
        {a1a1_a0a0=length(a1a1[a1a1[,1]>a1a1[,2]& a1a1[,1]>a1a1[,3],1])
        a1a1_a0a1=length(a1a1[a1a1[,2]>a1a1[,1]& a1a1[,2]>a1a1[,3],2])
        a1a1_a1a1=length(a1a1[a1a1[,3]>a1a1[,1]& a1a1[,3]>a1a1[,2],3])}


# t"aidame maatriksi
t=cbind(c(a0a0_a0a0,a0a0_a0a1,a0a0_a1a1,a0a0_dim-
sum(a0a0_a0a0,a0a0_a0a1,a0a0_a1a1)),c(a0a1_a0a0,a0a1_a0a1,a0a1_a1a1,a0a1_dim-
```

Joonis 32: Magistritöös kasutatavate koodide fragmendid. Osa 3.

```
sum(a0a1_a0a0,a0a1_a0a1,a0a1_a1a1)),c(a1a1_a0a0,a1a1_a0a1,a1a1_a1a1,a1a1_dim-
sum(a1a1_a0a0,a1a1_a0a1,a1a1_a1a1)))

# akumuleerib iga SNP-i tulemused
tulem=tulem+cbind(c(a0a0_a0a0,a0a0_a0a1,a0a0_a1a1,a0a0_dim-
sum(a0a0_a0a0,a0a0_a0a1,a0a0_a1a1)),c(a0a1_a0a0,a0a1_a0a1,a0a1_a1a1,a0a1_dim-
sum(a0a1_a0a0,a0a1_a0a1,a0a1_a1a1)),c(a1a1_a0a0,a1a1_a0a1,a1a1_a1a1,a1a1_dim-
sum(a1a1_a0a0,a1a1_a0a1,a1a1_a1a1)))

tegelik_jaotus=c(a0a0_dim,a0a1_dim,a1a1_dim)
arvutatud_jaotus=c(sum(c(a0a0_a0a0,a0a1_a0a0,a1a1_a0a0)),sum(c(a0a0_a0a1,a0a1_a0a1,a1a1_a0a1)),sum(c(a
0a0_a1a1,a0a1_a1a1,a1a1_a1a1)))

h=0
for(j in 1:3){
        if (tegelik_jaotus[j]!=0){
                z=(arvutatud_jaotus[j]-tegelik_jaotus[j])**2/tegelik_jaotus[j]}
        else {
                z=0}
h=h+z}

hii[i]=h # arvutab hii-ruut statistiku iga SNP-i jaoks #töös ei kasuta

#  imputeeritud SNP-i MAF

if (sum(t[1,])>sum(t[3,]))
        {mafsagedus=(2*sum(t[3,])+sum(t[2,]))/(2*sum(t))}
if (sum(t[1,])<=sum(t[3,]))
        {mafsagedus=(2*sum(t[1,])+sum(t[2,]))/(2*sum(t))}


# Referents SNP-i MAF

if (sum(t[,1])>sum(t[,3]))
        {teg_mafsagedus=(2*sum(t[,3])+sum(t[,2]))/(2*sum(t))}
if (sum(t[,1])<=sum(t[,3]))
        {teg_mafsagedus=(2*sum(t[,1])+sum(t[,2]))/(2*sum(t))}


vale=a0a0_a0a1+a0a0_a1a1+a0a1_a1a1+a0a1_a0a0+a1a1_a0a0+a1a1_a0a1

kadu=a0a0_dim-sum(a0a0_a0a0,a0a0_a0a1,a0a0_a1a1)+a0a1_dim-
sum(a0a1_a0a0,a0a1_a0a1,a0a1_a1a1)+a1a1_dim-sum(a1a1_a0a0,a1a1_a0a1,a1a1_a1a1)

MAF[i]=mafsagedus
teg_MAF[i]=teg_mafsagedus
Vale_gen_arv[i]=vale
Kadu[i]=kadu
i=i+1}


AUC v"a"artuse arvutamine, testi l"abiviimine
```

Joonis 33: Magistritöös kasutatavate koodide fragmendid. Osa 4.

```
# funktsioonid

hosmerlem = function(y, yhat, g) {
cutyhat = cut(yhat,breaks = quantile(yhat, probs=seq(0,1, 1/g)), include.lowest=TRUE)
  obs = xtabs(cbind(1 - y, y) ~ cutyhat)
  expect = xtabs(cbind(1 - yhat, yhat) ~ cutyhat)
ind=rep(NA,2*g)
for (k in 1:(2*g)){# nüüd korjame välja neid read, kus mitte expect, mitte observed ei võrdu nulliga
ind[k]=!(expect[k]==0 & obs[k]==0)}
expect=expect[ind]
obs=obs[ind]
  chisq = sum((obs - expect)^2/expect)
  P = 1 - pchisq(chisq, g - 2)
  return(as.vector(c(chisq,P)))
}


hii_ruut = function(y, yhat) {# see on praeguseks viimane ja õige variant
  g=length(table(yhat))
  obs1 = as.vector(by(y,yhat,sum))
  obs0 = as.vector(by(1-y,yhat,sum))
  exp1=table(yhat)*as.numeric(names(table(yhat)))
  p0=(1-as.numeric(names(table(yhat))))
  exp0=table(yhat)*p0
  ind=rep(NA,g)
if (g > 1){
for (k in 1:g){# nüüd korjame välja neid read, kus mitte expect, mitte observed ei võrdu nulliga
            ind[k]=(exp1[k]+obs1[k]!=0 & exp0[k]+obs0[k]!=0)}
            exp1=exp1[ind]
            obs1=obs1[ind]
            exp0=exp0[ind]
            obs0=obs0[ind]
            chisq=sum(((obs1-exp1)^2/exp1)+((obs0-exp0)^2/exp0))}# obs1=y1, exp1=n_1*Pi_1, p0=Pi_0
if (g==1){
            if(exp1[1]+obs1[1]!=0 & exp0[1]+obs0[1]!=0){chisq=((obs1-exp1)^2/exp1)+((obs0-exp0)^2/exp0)}
            if(exp1[1]+obs1[1]==0 | exp0[1]+obs0[1]==0){
            chisq=0}}
  P = 1 - pchisq(chisq, g)
  return(c(chisq,P))
}

testimine=function(y,yhat){
if (length(unique(quantile(yhat, probs = seq(0, 1, 1/3))))>=4){hosmerlem(y,yhat,g=3)}
if (length(unique(quantile(yhat, probs = seq(0, 1, 1/5))))>=6){hosmerlem(y,yhat,g=5)}
if (length(unique(quantile(yhat, probs = seq(0, 1, 1/10))))>=11){hosmerlem(y,yhat,g=10)}
#if (length(unique(quantile(yhat, probs = seq(0, 1, 1/2))))<3){hii_ruut(y,yhat)}
else {hii_ruut(y,yhat)}
}

# AUC arvutamiseks

install.packages("MKmisc")
library(MKmisc)
```

Joonis 34: Magistritöös kasutatavate koodide fragmendid. Osa 5.

```
m=0
AUC=c(0)
test_kokku=c(0,0)
i=1
while (i<dim(vastus)[1]+1) {
qhap=as.vector(v[i+m,])
imp=as.vector(vas[i,])
Q=matrix(qhap,ncol=2,byrow=TRUE)
Imp=matrix(imp,ncol=3,byrow=TRUE)

otsus=c(0)# observed
maximum=c(0)# expected

# võrdleme tulemusi
for (j in 1:dim(Q)[1]){
if (min(Imp[j,])!= max(Imp[j,]) | Imp[j,1]!=Imp[j,2]| Imp[j,2]!=Imp[j,3] | Imp[j,1]!=Imp[j,3]){
        if (sum(Q[j,])==1 & which.max(Imp[j,])==2 |
        sum(Q[j,])==2 & which.max(Imp[j,])==3 |
        sum(Q[j,])==0 & which.max(Imp[j,])==1){otsus[j]=1}
        else {otsus[j]=0}

        maximum[j]=max(Imp[j,])}}

# mittevarieeruvate SNP-ide jaoks AUC asemel arvutame keskmist tõenäosust
if (mean(otsus)!=1  &  mean(otsus)!=0){
        AUC[i]=AUC(maximum, group = 1-otsus, switchAUC = FALSE)}
if(mean(otsus)==1 | mean(otsus)==0){
        AUC[i]= -mean(maximum)}#

test_kokku=rbind(test_kokku,testimine(otsus,maximum))# väljastab HM testi(või hii-ruut) test statistiku ja
#p-value

i=i+1}
```

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina, Tatjana Iljashenko 22.02.1973a.

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose "Geneetiliste markerite imputeerimine", mille juhendaja on Märt Möls.

1.1. reprodutseerimiseks säilitamise ja üldsusele kättesaadavaks tegemise eesmärgil, sealhulgas digitaalarhiivi DSpace-is lisamise eesmärgil kuni autoriõiguse kehtivuse tähtaja lõppemiseni;
1.2. üldsusele kättesaadavaks tegemiseks Tartu Ülikooli veebikeskkonna kaudu, sealhulgas digitaalarhiivi DSpace´i kaudu kuni autoriõiguse kehtivuse tähtaja lõppemiseni.

2. olen teadlik, et punktis 1 nimetatud õigused jäävad alles ka autorile.

3. kinnitan, et lihtlitsentsi andmisega ei rikuta teiste isikute intellektuaalomandi ega isikuandmete kaitse seadusest tulenevaid õigusi.

Tartus , **20.05.2013.**