TARTU ÜLIKOOL
LOODUS- JA TÄPPISTEADUSTE VALDKOND
MOLEKULAAR- JA RAKUBIOLOOGIA INSTITUUT
BIOINFORMAATIKA ÕPPETOOL

Madis Sarapuu

Inimese *Per3* geeni tandeemse korduse koopiaarvu määramine teise põlvkonna sekveneerimisandmetest.

Bakalaureusetöö
Geenitehnoloogia
(12 EAP)

Juhendaja Tarmo Puurand MSc
Kaasjuhendaja Maris Teder-Laving MSc

Tartu 2016

Infoleht

Inimese *Per3* geeni tandeemse korduse koopiaarvu määramine teise põlvkonna sekveneerimisandmetest.

Varieeruva koopiaarvuga tandeemsed kordused on üle inimese genoomis laialt levinud. VNTRi (Variable Number of Tandem Repeats) perekonna moodustavad suur hulk erinevate pikkustega järjestusi. VNTRe on pikalt peetud rämps DNA-ks kuid järjest rohkem selgub nende mõju fenotüübile. Töös on lühiülevaade erinevatest programmidest, mida kasutatakse VNTR-ide leidmisel DNA järjestustest. Teise põlvkonna sekvneerimise levikuga ning täisgenoomide sekveneerimisega tekkis rida uusi võimalusi saada informatsiooni otse sekveneeritud andmetelt. Käesolevas bakalaureusetöös uurin *Per3* geeni VNTR-i koopiaarvu määramist teise põlvkonna sekveneerimisandmete põhjal, võrdlen erinevaid metoodikaid *k*-meri katvuse leidmiseks ning nende kasutamiseks VNTR-ide määramisel. Genotüüpide võrdlemisel võtsin aluseks PCR metoodikaga määratud *Per3* geeni polümorfismi rs57875989. Tulemustest lähtub, et GATK tööriista poolt määratud *Per3* genotüüp on ebatäpne võrreldes agaroosgeeli või *k*-mer metoodikal saadud tulemusega. Kuigi antud töös toimunud *k*-mer metoodikaga genotüpiseerimine ei ole piisavalt suure täpsusega, on see hetkel üks väheseid kui mitte ainukesi sekveneeritud täisgenoomi andmete põhjal VNTR arvude määramise meetodeid.

Märksõnad: VNTR, *k*-mer, teise põlvkonna sekveneerimine, genotüpiseerimine.
CERCS kood – B110, Bioinformaatika, meditsiinformaatika, biomatemaatika ja biomeetrika.

Abstract

Determing the number of tandem repeats in the human *Per3* gene from next generation sequencing data.

Variable number of tandem repeats are widespread throughout the human genome, consisting of many different-sized repeats. VNTRs are mostly thought as junk DNA, but many sources have shown that VNTR repeats affect the phenotype directly. In this paper I write about programs used for detecting tandem repeats from DNA sequence. NGS opened a lot of opportunities for getting data directly from sequenced reads. In this paper I tried to determine the number of tandem repeats in the *Per3* gene polymorphism rs57875989 from WGS data using three different *k*-mer methods. The Genotypes I compared with were verified via gel electrophoresis. In conclusion, the genotype determined by the GATK tool is not accurate compared to the gel electrophoresis or *k*-mer calculated results. K-mer methods that were used in this work, did not determine alleelse with the hoped accuracy, but using *k*-mer coverage is one of the only known ways to find an exact copy of tandem repeats from WGS data.

Keywords: VNTR, *k*-mer, next generation sequencing, genotyping

CERCS code – B110, Bioinformatics, medical informatics, biomathematics, biometrics.

Sisukord

Kasutatud lühendid

SSR – *simple sequence repeats* – lihtsad nukleotiidsed kordused

VNTR – *variable number of tandem repeats* – varieeruva arvuga tandeemsed kordused

INDEL – *insertions and deletions of bases* – nukleotiidide insertsioonid ja deletsioonid

STR – *short tandem repeats* – lühikesed tandeemsed kordused

DSB – *double strand break* – kaheahelalised katked

NGS – *next generation sequencing*- teise põlvkonna sekveneerimine

GWAS – *genome wide association study* – genoomi assotsiatsiooni uuring

SNP – *single nucleotide polymorphism* – üksiku nukleotiidi polümorfism

WGS – *whole genome sequence* – sekveneeritud täisgenoom

SSM – *single strand mispairing* – üksiku ahelda valestipaardumine

CNV – *copy number variation* – koopiaarvu variatsioon

TRF – *Tandem Repeat Finder* – tööriist tandeemsete korduste leidmiseks

Sissejuhatus

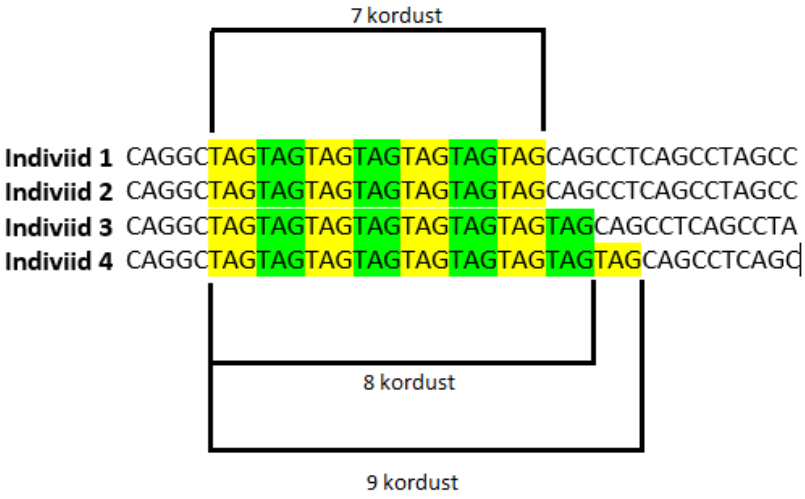
Viimase kümnendi jooksul on organismide täisgenoomide nukleotiidse järjestuste määramise hulk kasvanud väga kiiresti. Nn. teise põlvkonna sekveneerimisandmete hulk ühe indiviidi kohta on keskmiselt 20 korda iga kromosoomi lõigu kohta. Saadud sekveneerimisandmed on tavaliselt 100-150 aluspaari pikad ning asuvad kromosoomil paarikaupa 200 aluspaarise vahega. Sellised sekveneerimisandmed joondatakse referentsgenoomile, mille järel leitakse üles variandid järjestute ülesleidmise programme kasutades. Selline metoodika töötab hästi sekveneeritud lugemi sees unikaalse e. ühe korra genoomis oleva järjestuse olemasolul. Teistsuguste lugemite, mis sisaldavad valdavalt korduvaid järjestusi, genoomile paigutamine toimub paarilise asukoha info põhjal või jaotatakse referentsgenoomile ühtlaselt kõigi võimalike asukohtade vahel. Ühtlase jaotamise pärast võivad tekkida madala või liiga kõrge katvusega genoomiregioonid, kus variantide info on kallutatud või puudulik.
Käesolev bakalaureusetöö püüab leida võimaluse variatsioonide määramisel üht tüüpi, VNTR, korduste tarbeks, kasutades ära korduva motiivi lokalisatsiooni unikaalsust ja sekveneerimise katvust. VNTRe on peetud ebatähtsateks nende lokalisatsiooni tõttu mittekodeerivatel aladel. Töös tuleb ülevaade VNTR perekonna klassidest ning ka funktsionaalsetest tandeemsetest kordustest. Töö praktilise osa eesmärk on luua mudel, millega oleks võimalik *Per3* geeni näitel määrata ära tandeemselt korduva elemendi koopiaarvu ning võrrelda erinevate metoodikatega saadud k-mer katvuse kaudu leitud genotüübi vastavust geelelekotroforeesil saadud tulemustega.

Märksõnad: VNTR, *k*-mer, teise põlvkonna sekveneerimine, genotüpiseerimine.

1. Kirjanduse ülevaade

1.1 Tandeemsed kordused

Lihtsad korduvad järjestused on tandeemsed kordused, mis on inimese genoomis laialt levinud. Kordused koosnevad ideaalsetest või väheste erinevustega identsetest oligomeeridest. SSR-id(Simple sequence repeats), mis koosnevad lühikestest elementidest (1 – 13 aluspaari) nimetatakse mikrosatelliitideks. Pikemate korduvate üksustega (14 – 500 aluspaari) piirkondi nimetatakse minisatelliitideks. SSR-id moodustuvad inimese genoomist ligikaudu 3%, seejuures levinuim on dinukleotiidne kordus. Lihtsad korduvad järjestused on ülimalt olulised piirkonnad inimese geneetika uurimisel just nende varieeruvuse tõttu populatsioonis. Geneetilised markerid, mis põhinevad SSR-idel, on laialt kasutuses inimeste haiguste kaardistamisel(Lander ES *et al.*, 2001). Varasemalt arvati, et tandeemselt korduvad satelliit järjestused on rämps DNA ning neil ei ole organismis olulist tähtsust, käesolevaks ajaks on välja selgitatud et satelliit DNA osaleb heterokromatiini moodustumisel ning korrapärasel kromosoomide segregatsioonil.(Warburton *et al.*, 2008).



Joonis 1. Lühikesed tandeemsed kordused. Antud joonisel on näidatud varieeruv TAG motiiv eri indiviidide vahel.

Mitmed omadused või haigused ei pärandu Mendeli seaduste järgi, kuigi uuringud kaksikutega tõestavad mõndade selliste haiguste pärilikkust. Enamik keerukamaid haiguseid ei ole põhjustatud mutatsioonist ühes geenis vaid mitmete geenide koosmõjul(Tabor *et al.*, 2002). GWAS(Genome wide

association study) tulekuga hakati haiguste päritavuse uuringutes rohkem tähelepanu pöörama geenisisestele polümorfismidele. Tehnoloogia arenguga muutus SNP markerite genotüpiseerimine kiireks ja odavaks. Saavutatud edu aga ei ole suudetud kasutada VNTR-ide korduvate elementide koopiaarvu määramisel. (Brookes, 2013) Tänu GWAS analüüsile on paljud SNP-d genotüpiseeritud ning neid seostatakse mitmete komplektsete pärilike haigustega. Nende efekt geeniekspressioonile tuleneb suure tõenäosusega mitmete teiste faktorite koosmõjust. Genoomi assotsiatsiooni uuringust selgus, et leitud SNP-d ei moodusta piisavalt suurt hulka haigustega seondatavaid polümorfseid piirkondi, mida on näidatud kvantitatiivsetes uuringutes(Hannan, 2010). Seega jääb järele veel VNTR-id, mis võivad olla mitmete keerukate pärilike omaduste taga(Hannan, 2010). Kõige tõenäolisemalt asuvad funktsionaalsed VNTR järjestused kodeerivatel aladel, seega mõjutavad otseselt transkriptsiooni. Üheks heaks sellekohaseks näiteks võib tuua dopamiini retseptor D4 kodeeriva geeni (DRD4) sees asuvat VNTRi. 48 aluspaarine tandeemselt korduv motiiv esineb geenis 2 kuni 11 koopiana. Tandeemne kordus antud geenis mõjutab putatiivset kolmandat tsütoplasmaatilist proteiini, lisades sellele 16 aminohappelise järjestuse iga kordusega(Chio *et al.*, 1994). See omakorda loob võimaluse, et retseptori töö on häiritud ning dopamiini seondumisel inhibeeritakse cAMP produktsiooni, mis viib omakorda adenülüül tsüklaasi aktiivsuse langemiseni(van-Tol *et al.*, 1991). Samuti on mitmeid tõendeid funktsionaalsete VNTR-ide paiknemisest geeni promootoraladel. Inimpopulatsioonis uuritud haigused mis on põhjustatud tandeemsete korduste muutlikkusest promootoraladel on Huntingtoni tõbi ja X-Fragiilsus sündroom(Usdin, 2008). VNTRe on tihtipeale aga keeruline seostada erinevate omaduste või haigustega. Põhilisteks probleemideks võib pidada proteiinide funktsioone erinevates rakkudes – VNTR järjestus valkudes võib olla funktsionaalne ainult mõnda tüüpi rakkudes ning mittefunktsionaalne teist tüüpi rakkudes. Samuti lisab uuringutele veelgi variatsiooni asjaolu, et tandeemsed järjestused ei ole tihtipeale omavahel identsed(Zarrin *et al.*, 1999).

1.1.1 Spetsiifilised VNTR perekonnad

Üheks levinuimaks VNTRiks genoomis on Alphoid DNA. Teadaolevalt on alfa satelliit DNA ainukene satelliit DNA, mis leidub kõikides inimese kromosoomi tsentromeerides. Korduse pikkus on enamasti 170 aluspaari ning korduste arv on varieeruv.(Lee *et al.*, 1997)

Sau3a perekonda kuuluvad korduvad üksused koosnevad viiest umbes 170 aluspaari pikkusest homoloogsest üksusest. Inimese haploidses genoomis leidub Sau3a klassi järjestusi ligikaudu 1000 koopiat.(Kiyama *et al.*, 1986)

1.1.2 Minisatelliit DNA

Minisatelliidid on genoomis paiknevad, enamasti mittekodeerivad alad, mille suurus jääb 10-100 aluspaari vahele(Haber and Louis, 1998). Minisatelliidseid lookuseid arvatakse olevat ligikaudu 1500 haploidse genoomi kohta. Paljud nendest lookustest omavad suurt varieeruvust, mis väljendub korduste arvus. Alleelide pikkuse ning mutatsioonisageduse vahel ei ole minisatelliitide puhul seost täheldatud. Mutatsioonid toimuvad enamasti lookuse äärealal. Minisatelliidid ei ole üle genoomi ühtlaselt jaotunud, vaid on koondunud põhiliselt kromosoomi otstesse. Minisatelliite ei seostata evolutsiooniliste eelistega, küll aga on teada mitmed haigused, mis on põhjustatud ministatelliitide koopiaarvust (Ramel, 1997). Hästi uuritud selline lookus on HRAS1 VNTR. Antud lookus on kõrgelt polümorfne, sisaldades nelja levinud alleeli ning tosinat haruldasemat alleeli, mis tekitavad mutatsioone vereloome rakkudes(Weitzel *et al.*, 2000).

1.1.3 Mikrosatelliit DNA

Mikrosatellidid, ehk SRT-d(Short tandem repeats) on korduvad üksused, mis koosnevad enamasti di-,tri-,tetra, ja pentanukleotiididest. Näiteks CACACA on levinud dinukleotiidne mikrosatelliit. Neid peetakse kõike suurema varieeruvusega piirkonnaks terves genoomis, erinevused tulenevad üksuste korduste arvust. Põhiline hulk mikrosatelliite leidub mitte-kodeerivatel aladel, intronites või geenide vahelistes piirkondades. Inimese genoomis on leitud üle miljoni mikrosatelliitide lookuste, mis moodustavad koguni 3% inimese genoomist. Selliste lookuste arv imetajates on positiivses korrelatsioonis genoomi suurusega. Taimerakkudest sekveneeritud DNA uurimine näitas, et mikrosatelliitide tihedus on negatiivses korrelatsioonis genoomi suurusega. Levinuimad on dinukleotiidsed kordused ning haruldasemad on kolmest nukleotiidist koosnevad tandeemsed kordused. On täheldatud, et inimese genoomis on mikrosatelliitide tihedus suurem kromosoomi otste lähedal, kui mujal piirkondades. (Ellegren, 2004)

Tabel 1. SSR sisaldus inimese genoomis. SSRid leiti kasutades Tandem Repeat Finder programmi kasutades parameetreid: sobivuse skoor 2, mittesobivuse skoor 3, Indel 5, minimaalne joondus 50, maksimaalne järjestuse pikkus 500 aluspaari ja minimaalne 1 aluspaar(Lander *et al.*, 2001).

SSR sisaldus inimese genoomis

Korduste arv	Aluspaare keskmiselt Mb-s	Keskmiselt SSR elemente Mb-s
1	1,660	36,7
2	5,046	43,1
3	1,013	11,8
4	3,383	32,5
5	2,686	17,6
6	1,376	15,2
7	906	8,4
8	1,139	11,1
9	900	8,6
10	1,576	8,6
11	770	8,7

Tabel 2. SSR korduste esinemine inimese genoomis. SSR nukleotiidsete korduste esinemine leiti samasuguste parameetritega nagu eelneval tabelil näidatud(Lander *et al.*, 2001).

SSR korduste esinemine

Korduv element	SSR arv Mb-s
AC	27,7
AT	19,4
AG	8,2
GC	0,1
AAT	4,1
AAC	2,6
AGG	1,5
AAG	1,4
ATG	0,7
CGG	0,6
ACC	0,4
AGC	0,3
ACT	0,2
ACG	0,0

1.2 Tandeemsete korduste tekkemehhanismid

Tandeemsed kordused on mutatsioonidele vastuvõtlikumad kui enamus teised piirkonnad genoomis, just ühesuguste järjest paiknevate blokkide tõttu.(Kovtun and McMurray, 2008) Tandeemsete korduste võimalikke tekkepõhjuseid võib olla palju, kirjanduse järgi jaotatakse VNTR tekkemehhanismid kahte rühma – replikatsioonist sõltuvad protsessid ja ahelate parandamisega seotud protsessid(Kovtun and McMurray, 2008).

Üheks levinuimaks genoomse DNA kahjustuseks on kaksikahelate(DSB) katkemine. See on otseselt tingitud ioniseeriva kiirguse toimel või kokkupuutel kemikaalidega. Kaudselt põhjustab ahelate purunemist näiteks blokeerunud replikatsiooni kahvlid. Kaksikahelate õige kokku ligeerimine on genoomi säilimise jaoks ülioluline. Mittehomoloogiline rekombinatsioon ning mittehomoloogiliste katkete täitmine (NHJE) on arvatavasti põhilised mehhanismid, kuidas DSB parandatakse imetajate rakkudes. Homoloogiline rekombinatsioon kasutab vigastamata õdekromatiidi või homoloogilist kromosoomi näidisena. Kaheahelalise katke kleepuvate otse täitmine seevastu ei vaja homoloogilist järjestust(Takata *et al.*, 1998). On näidatud, et SSM(Single strand mispairing) mängib suurt rolli korduva DNA järjestuse evolutsioonis. Libiseva ahela mittepaardumine toimub DNA replikatsiooni käigus, kui mitu koopiat identset nukleotiidset järjestust on kõrvuti. Kui näidisahelas tekib aas, mis lõigatakse DNA-d parandavate ensüümide abil välja, toimub deletsioon. Aasa tekkimine tütarahelas kaasneb korduvate nukleotiidide lisamine ahelasse. SSM toimumise tagajärjel toimuvad insertsioonid või deletsioonid muudavad otseselt tandeemsete korduste koopiaarvu või koguni muudavad lihtsad ühest nukleotiidist koosnevad kordused keerulisemaks mitmealuspaariliseks tandeemseks korduseks(Levinson and Gutman, 1987). Üheks parandusega seotud protsessiks loetakse vananemisega ja oksidatiivsete kahjustuste tagajärjel tekkinud katkemist üheahelalises DNA-s, tekib juuksenõela struktuur, vastas ahela järgi sünteesitakse teine ahel tagasi asukohast kust aas tekkis. Juuksenõela struktuur aga parandatakse ja ligeeritakse samuti ahelasse, lisades aasa jagu kordusi tagasi järjestusse(Kovtun and McMurray, 2008).

1.3 Teise põlvkonna sekveneerimine

Teise põlvkonna sekveneerimine tegi võimalikuks paljude täisgenoomide sekveneerimise. Protsess muutus kiiremaks ja märgatavalt odavamaks võrreldes klassikalise Sangeri meetodiga. Sekveneeritud fragmentide pikkus on üldiselt lühike, 35bp kuni 400bp vahemikus, ning nende hulk suur. Segmendid katavad genoomi, ning ühte nukleotiidi sekveneeritakse mitmeid kordi. See tagab väikese vigade hulga ning kõrge kvaliteedilise sekveneerimise(Schatz *et al.*, 2010). Üheks populaarseks teise põlvkonna sekveneerimismasinate ja tehnoloogiate väljatöötajaks on Illumina, Inc. Nende sekveneerijad võtavad aluseks kaheahelalise DNA, transposoomid fragmenteeruvad ning jupitatud DNA ahelatele liidetakse otstesse adapterid, seejärel liidetakse fragmentidele komplementaarsed praimeri seondumissaidid ja indeksid. Seejärel DNA fragmendid hübridiseeruvad oligomeeridega ning neile sünteesitakse komplementaarne ahel, algne ahel pestakse välja. Tekkinud komplementaarset ahelat amplifitseeritakse nn ”sild amplifikatsiooni” meetodil. DNA polümeraas sünteesib tekkinud üheahelalisele ahelale komplementaarse ahela, moodustub kaheahelaline sild, ahelad denatureeritakse, mis tagab kaks ssDNA-d. Protsessi korratakse mitmeid kordi, kõiki fragmente amplifitseeritakse palju kordi. Sellele järgneb edaspidise ahela sekveneerimise etapp. Sekveneerimine toimub kus algselt liidetud adapteritle kinnitub komplementaarne praimer ning algab süntees, kus iga lisatud nukleotiid annab spetsiifilise valgussignaali ning see pildistatakse üles. Peale esimese fragmendi sekveneerimist pestakse sünteesitud ahel minema. Seejärel toimub sarnane protsess komplementaarse ahelaga. Selle tagajärjel pildistatakse üles miljonid lühikesed lugemid. Tekkinud lugemid sorteeritakse programmi abil ühte kobarasse, nii edaspidine kui vastaspidine. Tekkinud kobarad paigutatakse referentsgenoomi vastu.(Illumina, 2016) Teise põlvkonna sekveneerimise kasutusvõimalused on väga laialdased ning neid leitakse aina juurde. Praegusel hetkel kasutatakse NGSi põhiliselt transkriptsiooni analüüsiks, metagenoomika uuringud ja metülatsiooni analüüsid. Samuti on võimalik lugemitelt saada palju informatsiooni mõne haigusega seonduva geeni mutatsiooni kohta(Shendure and Ji, 2008).

1.4 Programmid korduste kirjeldamiseks

1.4.1 Repeatmasker

Repeatmasker on programm, mis loodi korduvate elementide identifitseerimiseks nukleotiidses järjestuses ning nende maskeerimiseks edasiseks analüüsiks. Programm on suuteline ka analüüsima valgu järjestusi. Repeatmasker otsib korduvaid järjestusi, võrreldes kasutaja poolt sisestatud FASTA failis sisalduvat genoomset järjestust ning teadaolevate korduste andmebaasi, nagu näiteks Repbase (Tarailo-Graovac and Chen, 2009). Repeatmaskerit saab kasutada veebipõhiselt kuni 100kb suuruste failide analüüsimiseks või lokaalselt käsurea Unix/Linux kaudu, kus failidel mahupiirangud puuduvad. Tööpõhimõte jaotatakse ülesannete järgi kahe programmi vahel, *cross_match* teostab joondamise ning Repeatmasker analüüsib ning edastab informatsiooni. Suurte failidega töötamisel on *cross_match* liialt ajakulukas ning on võimalik alternatiivina kasutada WU-BLAST-i, mis ei ole aga kahjuks nii tundlik. Programm koostab kolm vastuste faili, mis annavad ülevaate tulemustest: korduva elemendi nimetus, rühmitus, positsioon ja mitmesugused skoorid.(Tarailo-Graovac and Chen, 2009)

1.4.2 Tandem Repeat Finder

Tandem Repeat Finder(TRF) on algoritm, mis otsib soovitud nukleotiidsest järjestusest tandeemseid korduseid. Programm leiab kordused üles ilma korduvate elementide eelneva kirjeldamiseta. Erinevalt teistest algoritmidest, on TRF unikaalne, sest ta kasutab *k-tuple* paardumist. Eelnevalt pole tarvis sisestada korduva üksuse motiivi ega korduste arvu. Programmil puuduvad piirangud, kui suurt tandeemselt korduvat elementi on ta võimeline tuvastama. Töötleb deletsioone ning insertsioone eraldiseisvalt. Erinevalt teistest programmidest ei keskendu see algoritm kõige kõrgema skooriga homoloogilistele piirkondadele, vaid otsib tandeemseid korduseid, mis on tihtipeale varjul homoloogilistes piirkondades või jäävad paljudele programmidele kättesaamatuks. Detekteerimine sõltub stohhastilisest tandeemsete korduste mudelist, mida on täpsustatud samasugusus ning indelite protsent, erinevalt paljudel teistel juhtudel kasutuses olevast minimaalsest joonduse skoorist. Programm kõrvutab kaks korduvat elementi ning seejärel võrdleb nukleotiide Bernoulli jaotusega. pM, ehk kattuvuse tõenäosus antakse keskmise kattuvuse protsendina. Teine tõenäosusprotsent tähistatakse pI, mis näitab keskmist ühildamatuse protsenti. Algoritmil on kaks põhilist komponenti, tuvastav komponent ning analüüsiv komponent. Tuvastav komponent kasutab statistilisi kriteeriume, et avastada tandeemsete korduste kandidaate. Analüüsiv osa töötleb kandidaate edasi, joondades nad kõrvalasuvate

järjestustega. Kui vähemalt kaks koopiat samasugust järjestust on leitud, annab programm teate tandeemsest kordusest.(Benson, 1999)

1.4.3 VNTRseek

Programm on loodud, et tuvastada efektiivselt tandeemseid korduseid üle kogu genoomi ning saada olulist informatsiooni korduste esinemise ja omaduste kohta. Programmi tööpõhi mõte jaguneb mitmeks erinevaks osaks. Esmalt kasutatakse Tandem Repeat Finder algoritmi, et tuvastada referentsist tandeemsed lookused ning tandeemsed kordused järjestuste fragmentidest. Lugemite kordused kaardistatakse referentsis olevate tandeemsete kordustega sarnasuse alusel. Õige kaardistamine kinnitatakse, kontrollides lugemite ning referentsi kõrvalasuvaid järjestusi. Lõpuks suudab teoreetiliselt programm välja selgitada ka genotüübi, võttes aluseks lugemitelt loetud koopiaarvud.(Gelfand *et al.*, 2014)

1.4.4 Computel ja Telseq

Telomeerid asuvad kromosoomide otstes, ning kaitsevad kromosoomi degradatsiooni ning otste kokku kleepumise eest, telomeeri pikkuste muutused on seotud mitmete kromosomaalsete hälvete ning haigustega(Aviv, 2004). Telomeeri primaarstruktuuri moodustavad lühikestest tandeemsed kordused(Inglehearn and Cooke, 1990).

Computel programm on kirjutatud keeles R ning on suuteline arvutama kogu telomeeri primaarstruktuuri keskmist pikkust WGS andmetest. Leidmaks vajalikke andmeid teostab Computel mitmeid järjestikke analüüse. Esmalt moodustatakse telomeeri spetsiifiline indeksjärjestus. Seejärel reastatakse lugemid vastava telomeeri indeksiga. Arvutatakse katvus telomeeri indeksi ja kogu genoomi lõikes. Nende tulemuste ja arvutuste põhjal annab programm arvatava telomeeri struktuuri pikkuse(Nersisyan and Arakelyan, 2015).

Teine populaarne programm WGS andmete põhjal telomeeri pikkuse arvutamiseks on Telseq. Telseq loeb Bam faililt kogu lugemite arvu, lugemite pikkused, genoomi pikkuse ja teostab selle põhjal arvutused ning annab samuti eeldatava indiviidi telomeeri pikkuse(Liu *et al.*, 2013).

2. Eksperimentaalne osa

2.1 Töö eesmark

Käesoleva bakalaureusetöö eesmärkideks on:

1. Määrata *k*-mer metoodikaga tandeemselt korduvate elementide koopiaarvu *Per3* geeni VNTR regioonis (rs57875989) teise põlvkonna sekveneerimise andmete põhjal.
2. Selgitada välja sobivaim *k*-meri katvuse arvutamise metoodika, leidmaks VNTR koopiaarvu, mis ühtub kõige paremini varasemalt PCR ja geelelektroforees meetodiga määratud rs57875989 genotüüpidega samadel indiviididel.
3. Hinnata GATK paketiga rs57875989 määratud genotüüpe samadel indiviididel.

2.2 Per3 geen

Per3 geen kuulub imetajate tsirkadiaansesse süsteemi(Nadkarni *et al.*, 2005). Antud geeni seostatakse mitmete käitumusharjumuste ning unehäiretega inimeste seas, nagu näiteks hilinenud unefaasi sündroom, päevane unisus ning üldine unestruktuur(González-Giraldo *et al.*, 2015). Geen asub esimeses kromosoomis (chr1), positsioonil (7844413..7905241), GRCh37.p13(NCBI, 2016). Kahealleelne VNTR polümorfism Per3 geenis sisaldab endas 4 või 5 kordust. Korduv element on 54 aluspaari pikk ning asub geeni fosforüleerimise saidis(Nadkarni *et al.*, 2005). Polümorfismi perioodilisuse geenis seostatakse otseselt unerütmiga. Uuringust selgus, et homosügootsete pikemate alleelidega (5/5) indiviidid eelistasid hommikul varem ärgata ning samuti tundsid nad päevasel ajal vähem väsimust, kui heterosügootsete alleelidega (4/5) või lühemate alleelidega (4/4) homosügootsed inimesed(Lazar *et al.*, 2012).

Per3 VNTR

Asukoht	Korduste arv	Identsusportsent kogu VNTRi ulatuses
Chr1: 7889934-7890186	4.685185	0.909449

CTACTACCGGTGCACTGTCCACGGGGTCACCTCCCAGGGAGAATCCATCCCATC

CTACTGCCAGCGCTCTGTCCACAGGATCG**CCTCCCATGAAGAATCCATCCCATC**

CTACTGCCAGCGCTCTGTCCACAGGATCG**CCTCCCATGAAGAATCCATCCCATC**

CTACTGCCAGCACACTGTCCATGGGATTGCCTCCCAGCAGGACTCCATCCCATC

CTACTGCCACTGTTCTGTCCACGGGGTCACCTCCCAG

Joonis 2. TRF programmiga leitud VNTR'i kirjeldus. Asukoht, motiiv, korduste arv, identsus protsent. 2. ja 3. motiiv on identsed ning polümorfsed ühe koopia puudumise osas. 25-mer järjestus, mida antud töös esinemissageduste arvutamisel kasutatakse, on märgitud tumedalt.

2.3 Materjal ja metoodika

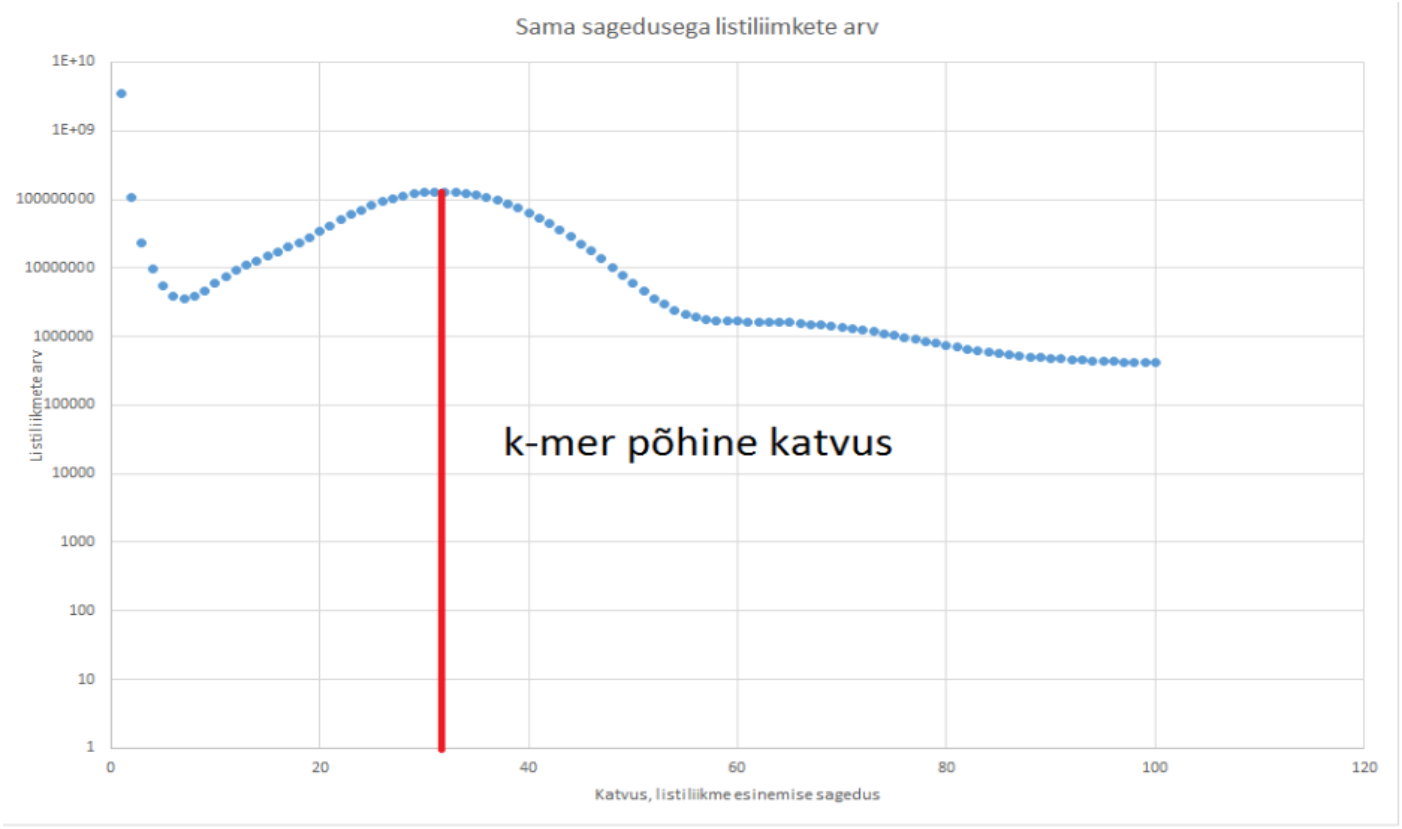
2.3.1 **Valim**

Valimisse valiti Tartu Ülikooli Eesti Geenivaramu geenidoonorid, kellel oli sekveneeritud täisgenoom ning varasemas uurimustöös genotüpseeritud geenis *Per3* asuv polümorfism rs57875989. Valimi suuruseks kujunes 64 indiviidi. Vajalikud andmed väljastas TÜ Eesti Geenivaramu. Genotüpiseerimine toimus Polümeraasi ahelreaktsioonil põhineval metoodikal ning hilisemal produktide eraldamisel agaroosgeelil(Moro, 2013). Valimisse kuuluvatel indiviididel võrreldi genotüüpe ka GATK tööriista abil saadud tulemustega.

2.3.2 **NGS sekveneerimisel katvuse arvutamine**

NGS sekveneerimise katvuse arvutamiseks on mitmeid erinevaid metoodikaid. Käesolevas töös kasutasin kolme erinevalt leitud *k*-meri katvust.

Esimese meetodi puhul leidsin sekveneeritud indiviidide *k*-mer keskmise katvuse sageduste histogrammi abil.



Joonis 3. X teljel on näidatud liikmete arv nimekirjas ning Y teljel on katvusega nimekirjas liikmete esinemise sagedus. Punane joon tähistab keskmist *k*-meri katvust.

Teise katvuse arvutamise metoodika on pärit GATK tööriistade seast ja see informatsioon saadi koos täisgenoomi järjestustega. GATK DepthOfCoverage on Broad Institute poolt välja töötatud programm, mis võtab sisendiks Bam failid ning selgitab välja selle põhjal katvuse. Katvust saab hinnata nii kogugenoomi lõikes või ainult konkreetse lookuse või geeni tarvis. Lisaks on võimalik ka sorteerida lugemeid ka nende kaardistamise või aluspaari kvaliteedi skoori alusel – sellega on võimalik tõsta tulemuste õigsust ning kvaliteeti.(Broad Institute, 2016)

Kolmanda ehk nn „Lihtsa“ meetodina katvuse leidmiseks kasutasin modifitseeritud valemit Liu, Yujian *et.al* tööst. Valem $C_k/C_n = N *(L-km+1)/C$, kus C_k tähistab kõikide *k*-meride arvu, C_n tähistab genoomi suurust, N tähistab lugemite arvu , L lugemite keskmist pikkust ning km *k*-meri pikkust. Valemi *k*-meri keskmise katvuse arvutamiseks võtsin $C_k = C*(L-km+1)/L$.(Liu *et al.*, 2013)

Leitud *k*-meri katvuste ning *k*-meri sageduse vahelise suhte leidmiseks kasutasin valemit S=Kc/(Ck/2), kus S tähistab *k*-meri sageduse ja katvuse suhet, K_c tähistab *k*-mer counti ning C_k tähistab keskmist

katvust. Samuti selgitasin välja kogugenoomi katvuse, selle sain teada sekveneerimisandmete tabelist.

2.3.3 K-mer listide tegemine

K-meri nimekirjad olid koostatud Glistmakeriga, mis kuulub tööriistapaketti GenomeTester4. Glistmaker koostab *k*-meride tabeli, kasutades sisendiks FASTA või FASTQ faili. Esmalt otsib programm välja kasutaja sisestatud failist välja kõik *k*-merid, ajutiselt leitakse ka *k*-meride pööratud komplementaarne järjestus. Leitud *k*-merid sorteeritakse ning reastatakse ja lõpuks leiab programm iga eelnevalt leitud oligomeeri esinemise arvu ehk sageduse.(Lepamets, 2014)

2.3.4 Päringu tegemine

K-meri listidest päringute tegemiseks kasutasin Glistquery programmi, mis on samuti GenomeTester4 paketi osa(Kaplinski *et al*. 2015). Glistquery on mõeldud eelnevalt koostatud k-meri listidest üles leidma *k*-meride sagedusi. Glistquery töötab binaarsel otsingul ning võimaldab otsida ka etteantud parameetrite piires mittekattuvaid *k*-mere. Varasemalt moodustatud *k*-meri listidest sooritasin päringud Glistquery-ga, et leida enda valitud k-meri esinemissagedus ehk *k-mer count* erinevatel indiviididel.
Andmete töötlemiseks sobilik *k*-mer sai valitud kasutades Samtools tööriista.(Li, 2011) Samtoolsi abil visualiseerisin teise põlvkonna sekveneerimise andmed *tview* käsuga (Joonis lisas). *K*-meri pikkuseks valisin 25 nukleotiidi, sest see tagab kindlasti piisava unikaalsuse ning ei ületa päringute tegemiseks etteantud piire. Pikema *k*-meriga töötamisel võivad osutuda probleemiks mutatsioonid ja sekvenerimisvead(Kaplinski *et al*. 2015). Uurimustöös kasutatavaks *k*-meriks valisin *Per3* geenis asuva korduva motiivi spetsiifilise järjestuse `5′ CCTCCCATGAAGAATCCATCCCATC 3′` (Joonis 2)**.**

2.3.5 Genotüüpide määramine

Indiviidide genotüüpide aluseks võtsin TÜ Geenivaramust väljastatud PCR metoodikaga saadud andmed. Samuti kasutasin genotüüpide võrdlemiseks ka GATK VCF failist saadud andmeid. Käesolevas bakalaureusetöös genotüübi väljaselgitamiseks kasutasin leitud *k*-meri counti ja katvuse suhteid ning võrdlesin antud tulemust lävenditega. Lävendid valisin visuaalse vaatluse teel, kuidas minu leitud tulemused ühtiksid geelelektroforeesil saadud tulemustega. Esimese, histogrammi põhjal saadud *k*-meri katvuse alumiseks lävendiks valisin 2,1 ja ülemiseks lävendiks 3,1. Teise,

DepthOfCover programmi kaudu saadud *k*-meri katvuse põhjal genotüübi ennustamise tarvis valisin alumiseks lävendiks 1,6 ja ülemiseks 3,1. Kolmanda valemipõhise meetodiga tegelemiseks valisin alumiseks lävendiks 1.5 ja ülemiseks lävendiks 2,5.

Tabel 3. Lävendid k-mer genotüüpide määramise metoodikatele

Lävendid	Histogram	DepthOfCoverage	Lihtne
Ülemine	3,3	3	2,5
Alumine	2,1	1,7	1,5

2.4 Tulemused

1. Töös määrasin 64 indiviidi Per3 geeni polümorfismi rs57875989 genotüübid. Kolme erineva metoodikaga leitud k-mer katvuste abil ning teiste vahenditega leitud genotüüpide võrdlus on kokkuvõtva tabelina lisades (Lisa 2.).
2. Kõige suurema sama genotüüpide määramise protsendi andis DepthOfCover programmi poolt arvutatud katvuse ja genotüüpide lävendite kasutamine (Tabel 3 ja 4). Alumisest lävendist allapoole jäävad tulemused võrdsustasin homosügootsete deletsiooniga alleelidega. Lävendite vahel olevad tulemused lugesin heterosügootseteks variantideks ning ülemisest lävendist kõrgemad tulemused võrdsustasin homosügootsete pikkade alleelidega. Kõigi kolme erineva *k*-mer metoodikaga ennustatud genotüübid ennustasid praktiliselt samasugused tulemused, histogrammi abil saadud *k*-meri katvuse kaudu oli võimalik hinnata genotüüpi õigesti 73%-l juhtudel. DepthOfCover paketi määratud *k*-mer katvus kattus 78%-il juhtudel geelipildiga. Modifitseeritud valemipõhjal arvutatud nn ''Lihtne'' katvus määras ära dialleelse lookuse 75%-il indiviididel. Seega võib järeldada, et kõik k-mer metoodikad on sarnase efektiivsusega ning suutlikkusega antud probleemile lahendust leidma.
3. GATK tööriistaga määratud rs57875989 genotüübid vastasid 39% ulatuses geelil suurusi hinnatud tööde tulemustega. Sellest võib järeldada, et GATK VCF failist saadud vaadeldud polümorfismi genotüübi andmed on sisuliselt juhusliku jaotusega.

Tabel 4. Erinevate metoodikate täpsuse võrdlus.

	Histogram	DepthOfCover	Lihtne	GATK
Vaadatud indiviide	64	64	64	64
Geeli genotüübiga sama	47	50	48	25
Õige määramise protsent	73%	78%	75%	39%

2.5 Arutelu

Praegu maailmas laialt levinud variatsioonide püüdmise programmid ei võimalda saada täpset informatsiooni korduvate järjestuste variatsioonide kohta. GATK on spetsialiseerunud lühikeste variatsioonide leidmisele ja määramisele. Siinse töö põhjal võib väita, et 54 aluspaarine muutus korduvas järjestuses on GATK'ule juba kättesaamatus ulatuses. NGS andmetest on lihtsam üles leida lühemaid ja unikaalses regioonis paiknevaid variatsioone. Siinses töös vaadeldud *Per3* geeni VNTR'i saab klassifitseerida ka indel'ina (dialleelse esinemise tõttu), kuid sellegi poolest jäi tuntud GATK programmipakett hätta. VNTR-seek(Gelfand *et al.*, 2014) programmi täpsust käesoleva töö raames ei uuritud, kuna töö jätkuks on multialleelsete VNTR'ide summaarse koopiaarvu määramine lugemite pikkusest oluliselt suuremate korduvate DNA lõikude jaoks.

Siinses töös juurutatud metoodika pole ideaalne. Koopiaarvu mõõdetakse mõlema kromosoomi peale kokku ja seetõttu on igasugune lisainformatsioon alleelide esinemissagedusest ja iseloomust täpsust suurendav ja statistilist analüüsi lihtsustav. 80%-le lähenev määramistäpsus ja seda kahealleelse markeri juures muudab antud metoodika enama koopia- ja alleelivariantide arvu puhul veel ebatäpsemaks. Siiski tasub seda täiendada, kuna hetkel on ta parim (võib-olla ka ainuke) variant täisgenoomi andmetest VNTR-ide koopiaarvu määramisel, millede pikkus on suurem, kui lugemi pikkus.

Siinse töö tulemuste juures on huvitavaks leiuks ka väga madalad ja väga kõrged *Per3* VNTR koopiaarvud, mis võivad viidata hoopis CNV regioonile, kus ühe kromosoomi alleelel oleks nagu puudu või hoopis üle kas siis pikema DNA lõiguna või ainult motiivi ulatuses.

Kokkuvõte

Algselt rämps DNA-ks peetud VNTR-e uurides on leitud, et kodeerivates regioonides leiduvatel tandeemsetel kordustel on otsene mõju fenotüübile. Näiteks on näidatud, et tsirkadiaansesse süsteemi kuuluvas *Per3* geenis leiduvatel VNTR koopiaarv on seotud mitmete unehäiretega.

Hetkel puuduvad efektiivsed programmid või võimalused määrata VNTR koopiaarvu WGS lugemitelt. Käesolevas uurimustöös üritasin välja selgitada *k*-mer metoodikaga *Per3* geenis rs57875989 polümorfismi tandeemsete korduste koopiaarvu. Koopiaarvu määramiseks arvutasin kolme erineva metoodikaga *k*-mer katvused, võrdlesin katvuse ning *k*-mer sageduse suhet ja seejärel kõrvutasin nad lävenditega mille järgi määrasin alleelid. Võttes genotüübi aluseks TÜ Geenivaramust saadud geelipildid, selgus, et kõige täpsem tulemus sai, kui võtta aluseks GATK tööriista DepthOfCover leitud *k*-mer katvus ning sellepealt arvutada välja alleeli variandid. Genotüüp kattus ligikaudu 80%-l PCR metoodikal saadud tulemustega. Siiski, võrreldes teiste, histogrammi ja valemi abil leitud *k*-mer katvuse ning genotüübi ennustamisega, märkimisväärset erinevust ei leidnud. Seega võib öelda, et kõigi meetoditega on võimalik teatud täpsusega tulemused saada, aga sobilikem on VNTR genotüpiseerimise aluseks võtta DepthOfCover leitud *k*-mer katvus. 80%-line määramistäpsus pole piisavalt täpne ning seda metoodikat peab veel edasi arendama, et saada väärtuslikke tulemusi ning oleks võimalik töötada ka rohkemate alleelidega.

Üheks töö eesmärgiks oli ka GATK VCF failis oleva genotüübi võrdlemine teiste meetoditega saadud genotüüpidega ning selgus, et GATK suutis ainult 39%-lise täpsusega määrata ära alleelivariandid võrreldes geelipiltidega. See tähendab et GATK programm ei ole suuteline iseseisvalt *Per3* geenis asuvat 54 aluspaarilist VNTRi määrava täpsusega genotüpiseerida.

Summary

VNTRs have been initially thought of as junk DNA, but recent research has shown that tandemly repetitive elements located at coding regions have a direct impact for the phenotype. A good example is VNTR found in the *Per3* gene, which copy number affects sleeping patterns.

Currently there are no good tools for measuring VNTR copy number directly from WGS reads. In this work I have measured three different *k*-mer methods to determine the copy number of tandemly repetitive sequence in the *Per3* gene polymorphic region rs57875989. Firstly I found the *k*-mer coverage using a histogram, GATK DepthOfCoverage and formula based calculations. After that, I compared the k-mer count with previously found coverage and compute their ratio. Last step was comparing the determined ratio with the picked thresholds. Genotyping based on gel electrophoresis showed that the GATK tool DepthOfCoverage *k*-mer coverage was the most accurate for 80% success rate to check Per3 diallelic polymorphism. Although when compared to histogram and formulation based methods, there was statistically significant difference. Within a limited margin of error,good results can be achieved with all three methods but DepthOfCover *k*-mer would be the most suitable for VNTR genotyping from WGS data. An accuracy of 80% is not enough and it might be even less accurate when VNTR is multiallelic. Since this method is unique and probably one of few, it should be reasearched further.

Another goal was to see whether the genotype presented in GATK VCF file is correct, when compared to others. Results show that GATK could determine the right genotype with only a 39% success rate. In conclusion I can tell that the GATK program itself is not capable of finding the correct number of 54 basepair tandem repeats in *Per3* gene.

Kasutatud kirjandus

Aviv, A. (2004). Telomeres and human aging: facts and fibs. Sci. Aging Knowledge Environ. 2004: pe43.

Benson, G. (1999). Tandem repeats finder: A program to analyze DNA sequences. Nucleic Acids Res. 27: 573–580.

Brookes, K.J. (2013). The VNTR in complex disorders: The forgotten polymorphisms? A functional way forward? Genomics 101: 273–281.

Chio, C.L., Drong, R.F., Riley, D.T., Gill, G.S., Slightom, J.L., and Huff, R.M. (1994). D4 dopamine receptor-mediated signaling events determined in transfected Chinese hamster ovary cells. J. Biol. Chem. 269: 11813–11819.

Consortium., I.H.G.S. (2001). Initial sequencing and analysis of the human genome. Nature 412: 860–921.

Ellegren, H. (2004). Microsatellites: simple sequences with complex evolution. Nat. Rev. Genet. 5: 435–445.

Gelfand, Y., Hernandez, Y., Loving, J., and Benson, G. (2014). VNTRseek-a computational tool to detect tandem repeat variants in high-throughput sequencing data. Nucleic Acids Res. 42: 8884–8894.

González-Giraldo, Y., González-Reyes, R.E., Mueller, S.T., Piper, B.J., Adan, A., and Forero, D.A. (2015). Differences in planning performance, a neurocognitive endophenotype, are associated with a functional variant in PER3 gene. Chronobiol. Int. 1–5.

Haber, J.E., and Louis, E.J. (1998). Minisatellite origins in yeast and humans. Genomics 48: 132–135.

Hannan, A.J. (2010). Tandem repeat polymorphisms: modulators of disease susceptibility and candidates for “missing heritability.” Trends Genet. 26: 59–65.

Inglehearn, C.F., and Cooke, H.J. (1990). A VNTR immediately adjacent to the human pseudoautosomal telomere. Nucleic Acids Res 18: 471–476.

Kiyama, R., Matsui, H., and Oishi, M. (1986). A repetitive DNA family (Sau3A family) in human chromosomes: extrachromosomal DNA and DNA polymorphism. Proc. Natl. Acad. Sci. U. S. A. 83: 4665–4669.

Kovtun, I. V, and McMurray, C.T. (2008). Features of trinucleotide repeat instability in vivo. Cell Res. 18: 198–213.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., *et al.* (2001). Initial sequencing and analysis of the human genome. Nature 409: 860–921.

Lazar, A.S., Slak, A., Lo, J.C., Santhi, N., von Schantz, M., Archer, S.N., Groeger, J.A., and Dijk, D.J. (2012). Sleep, diurnal preference, health, and psychological well-being: a prospective single-allelic-variation study. Chronobiol Int 29: 131–146.

Lee, C., Wevrick, R., Fisher, R.B., Ferguson-Smith, M.A., and Lin, C.C. (1997). Human centromeric DNAs. Hum. Genet. 100: 291–304.

Levinson, G., and Gutman, G. a (1987). Slipped-strand mispairing: a major mechanism for DNA sequence evolution. Mol. Biol. Evol. 4: 203–221.

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27: 2987–2993.

Liu, B., Shi, Y., Yuan, J., Hu, X., Zhang, H., Li, N., Li, Z., Chen, Y., Mu, D., and Fan, W. (2013). Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects. arXiv 1308.2012.

Nadkarni, N.A., Weale, M.E., von Schantz, M., and Thomas, M.G. (2005). Evolution of a length

polymorphism in the human PER3 gene, a component of the circadian system. J Biol Rhythm. 20: 490–499.

Nersisyan, L., and Arakelyan, A. (2015). Computel: Computation of mean telomere length from whole-genome next-generation sequencing data. PLoS One 10.

Ramel, C. (1997). Mini- and microsatellites. In Environmental Health Perspectives, pp. 781–789.
Schatz, M.C., Delcher, A.L., and Salzberg, S.L. (2010). Assembly of large genomes using second-generation sequencing. Genome Res. 20: 1165–1173.

Shendure, J., and Ji, H. (2008). Next-generation DNA sequencing. Nat Biotechnol 26: 1135–1145.

Zarrin, A.A., Malkin, L., Fong, I., Luk, K.D., Ghose, A., and Berinstein, N.L. (1999). Comparison of CMV, RSV, SV40 viral and V??1 cellular promoters in B and T lymphoid and non-lymphoid cell lines. Biochim. Biophys. Acta - Gene Struct. Expr. 1446: 135–139.

Tabor, H.K., Risch, N.J., and Myers, R.M. (2002). Candidate-gene approaches for studying complex genetic traits: practical considerations. Nat. Rev. Genet. 3: 391–397.

Takata, M., Sasaki, M.S., Sonoda, E., Morrison, C., Hashimoto, M., Utsumi, H., Yamaguchi-Iwai, Y., Shinohara, A., and Takeda, S. (1998). Homologous recombination and non-homologous end-joining pathways of DNA double-strand break repair have overlapping roles in the maintenance of chromosomal integrity in vertebrate cells. EMBO J. 17: 5497–5508.

Tarailo-Graovac, M., and Chen, N. (2009). Using RepeatMasker to identify repetitive elements in genomic sequences. Curr. Protoc. Bioinforma.

Usdin, K. (2008). The biological effects of simple tandem repeats: Lessons from the repeat expansion diseases. Genome Res. 18: 1011–1019.

van-Tol, H.H., Bunzow, J.R., Guan, H.C., Sunahara, R.K., Seeman, P., Niznik, H.B., and Civelli, O. (1991). Cloning of the gene for a human dopamine D4 receptor with high affinity for the antipsychotic

clozapine. Nature 350: 610–614.

Warburton, P.E., Hasson, D., Guillem, F., Lescale, C., Jin, X., and Abrusan, G. (2008). Analysis of the largest tandemly repeated DNA families in the human genome. BMC Genomics 9: 533.

Weitzel, J.N., Ding, S., Larson, G.P., Nelson, R.A., Goodman, A., Grendys, E.C., Ball, H.G., and Krontiris, T.G. (2000). The HRAS1 minisatellite locus and risk of ovarian cancer. Cancer Res. 60: 259–261.

Kasutatud veebiaadressid

BroadInstitute GATK DepthOfCoverage (20. mai, 2016)
https://www.broadinstitute.org/gatk/guide/tooldocs/org_broadinstitute_gatk_tools_walkers_coverage_DepthOfCoverage.php

NCBI Gene andmebaas (10. mai, 2016) http://www.ncbi.nlm.nih.gov/gene/

Illumina Inc. Next generation sequencing (21.mai, 2016)
http://www.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf

Lisad

Lisa 1. Indiviidi nr. 32 (genotüübiga 4/5) samtools tview väljatrükk rs57875989 ümbruses, kuhu on lisatud punktidega ja siin töös kasutatad *k*-mer järjestus ja asukoht VNTR-i regioonis. Siinses väljatrükis ei näidata deletsiooni (tähistatud tärnidena lugemite sees) asukohta nn. VNTR'i raamis. Joonisel on esimesel real chr1 asukohad, siis referentsgenoomi järjestus, siis k-mer asukohta näitav rida. Järgmine allpool on indiviidid konsensusjärjestus ja edasi allpool lugemite järjestuse. "." ja "," tähendavad sama nukleotiidi, mis on samal positsioonil refrentsgenoomis. Lugemitel olevad tähed tähendavad seevastu erinevusi.

Lisa 2. Kokkuvõttev tabel tulemustega. Vasakpoolsest veerust vaadatuna on esimeses tulbas indiviidide koodid, seejärel on näidatud k-mer count ehk sagedus. Katvuse lahtrites on vasakpoolseimas tulbas *k*-mer katvus arvutatuna histogrammi põhjal, seejärel GATK tööriista kaudu leitud *k*-mer katvus ning parempoolseimas katvuse tulbas on nn „Lihtne“ ehk üldise valemiga arvutatud katvus. Genotüübi lahtrites on näidatud erinevatel moodustel saadud genotüüpe, 44 on lühikeste alleelidega homosügoot, 45 heterosügoot ning 55 tähistab pikemate alleelidega homosügooti. Genotüübi tulpadest vasakpoolseim on agaroosigeeli pildi põhjal määratud genotüüp, seejärel genotüüp, mis on leitud histogrammi kaudu saadud k-mer katvuse põhjal. Järgnevas veerus on GATK DepthOfCoverage leitud *k*-meri katvuse põhjal saadud genotüüp ning parempoolseimas genotüübi tulbas on valemi põhjal saadud genotüüp. Suhtarvude lahtrites vasakpoolseim on histogrammi põhjal saadud katvuse ning *k*-mer counti suhe, keskmine veerg on DepthOfCoverage katvuse ja *k*-mer counti suhtarv ning parempoolseim on valemiga leitud *k*-mer katvuse ning sageduse suhe. VCF veerg kujutab endast GATK VCF failis kuvatud alleele võrreldes referentsgenoomiga. Viimane ning parempoolseim blokk näitab geelipildi kokkulangevust teiste meetoditega. 0 tähistab mittesobivust ja 1 näitab kokkulangevust.

Indiviidid	Sagedus	Katvus			GENOTÜÜP					Suhtarvud			VCF	Geeli genotüübiga sama			
	k-mer	Histogram	GATKDOC	Lihtne	Geel	Histogram	GATKDOC	Lihtne	GATK	Histogram	GATKDOC	Lihtne	GATK	Histogram	GATKDOC	Lihtne	GATK
1	54	34	40	44	55	45	45	45	55	3,2	2,7	2,5	0/0	0	0	0	1
2	32	29	35	42	45	45	45	45	55	2,2	1,8	1,5	0/0	1	1	1	0
3	15	24	29	33	44	44	44	44	45	1,3	1,0	0,9	0/1	1	1	1	0
4	22	24	29	34	44	44	44	44	44	1,8	1,5	1,3	1/1	1	1	1	1
5	20	22	26	30	44	44	44	44	45	1,8	1,5	1,3	0/1	1	1	1	0
6	44	33	39	45	55	45	45	45	55	2,7	2,3	2,0	0/0	0	0	0	1
7	7	24	29	35	44	44	44	44	44	0,6	0,5	0,4	1/1	1	1	1	1
8	35	30	36	40	45	45	45	45	55	2,3	1,9	1,8	0/0	1	1	1	0
9	29	28	33	37	45	44	45	45	55	2,1	1,8	1,6	0/0	0	1	1	0

10	16	28	33	41	44	44	44	44	45	1,1	1,0	0,8	0/1	1	1	1	0
11	21	25	30	35	44	44	44	44	55	1,7	1,4	1,2	0/0	1	1	1	0
12	25	20	24	28	45	45	45	45	55	2,5	2,1	1,8	0/0	1	1	1	0
13	26	17	20	25	45	45	45	45	45	3,1	2,6	2,1	0/1	1	1	1	1
14	35	21	25	32	45	55	45	45	55	3,3	2,8	2,2	0/0	0	1	1	0
15	39	29	35	40	45	45	45	45	55	2,7	2,2	2,0	0/0	1	1	1	0
16	12	17	20	24	45	44	44	44	55	1,4	1,2	1,0	0/0	0	0	0	0
17	23	23	27	34	44	44	45	44	44	2,0	1,7	1,4	1/1	1	0	1	1
18	23	29	34	40	44	44	44	44	45	1,6	1,4	1,2	0/1	1	1	1	0
19	27	29	35	43	44	44	44	44	55	1,9	1,5	1,3	0/0	1	1	1	0
20	33	25	30	35	45	45	45	45	45	2,6	2,2	1,9	0/1	1	1	1	1
21	11	19	22	26	45	44	44	44	44	1,2	1,0	0,8	1/1	0	0	0	0
22	15	22	26	33	45	44	44	44	55	1,4	1,2	0,9	0/0	0	0	0	0
23	35	26	31	36	45	45	45	45	55	2,7	2,3	1,9	0/0	1	1	1	0
24	35	26	31	37	45	45	45	45	55	2,7	2,3	1,9	0/0	1	1	1	0
25	73	31	37	55	55	55	55	55	55	4,7	3,9	2,7	0/0	1	1	1	1
26	9	18	21	25	44	44	44	44	55	1,0	0,9	0,7	0/0	1	1	1	0
27	51	26	31	39	55	55	55	55	55	3,9	3,3	2,6	0/0	1	1	1	1
28	40	26	31	37	55	45	45	45	55	3,1	2,6	2,2	0/0	0	0	0	1
29	18	18	21	25	45	44	45	44	55	2,0	1,7	1,4	0/0	0	1	0	0
30	39	25	30	38	45	45	45	45	55	3,1	2,6	2,1	0/0	1	1	1	0
31	24	26	31	38	44	44	44	44	45	1,8	1,5	1,3	0/1	1	1	1	0
32	27	17	20	25	45	45	45	45	45	3,2	2,7	2,2	0/1	1	1	1	1
33	41	28	33	38	45	45	45	45	55	2,9	2,5	2,2	0/0	1	1	1	0
34	15	19	23	28	44	44	44	44	55	1,6	1,3	1,1	0/0	1	1	1	0
35	14	18	21	25	45	44	44	44	55	1,6	1,3	1,1	0/0	0	0	0	0
36	16	17	20	23	45	44	44	44	55	1,9	1,6	1,4	0/0	0	0	0	0
37	35	15	18	33	55	55	55	45	55	4,7	3,9	2,1	0/0	1	1	0	1
38	14	20	24	29	44	44	44	44	44	1,4	1,2	1,0	1/1	1	1	1	1
39	21	23	27	33	44	44	44	44	55	1,8	1,6	1,3	0/0	1	1	1	0
40	19	26	31	36	44	44	44	44	44	1,5	1,2	1,1	1/1	1	1	1	1
41	9	19	22	27	44	44	44	44	55	0,9	0,8	0,7	0/0	1	1	1	0
42	41	14	17	26	55	55	55	55	55	5,9	4,8	3,2	0/0	1	1	1	1
43	19	19	22	24	45	44	45	45	55	2,0	1,7	1,6	0/0	0	1	1	0

44	27	25	30	35	45	45	45	45	55	2,2	1,8	1,5	0/0	1	1	1	0
45	17	25	30	36	44	44	44	44	44	1,4	1,1	0,9	1/1	1	1	1	1
46	32	24	28	33	45	45	45	45	55	2,7	2,3	1,9	0/0	1	1	1	0
47	25	25	30	33	44	44	44	45	55	2,0	1,7	1,5	0/0	1	1	0	0
48	27	29	35	42	44	44	44	44	44	1,9	1,5	1,3	1/1	1	1	1	1
49	27	22	26	30	45	45	45	45	45	2,5	2,1	1,8	0/1	1	1	1	1
50	24	31	37	44	44	44	44	44	45	1,5	1,3	1,1	0/1	1	1	1	0
51	17	24	28	31	45	44	44	44	55	1,4	1,2	1,1	0/0	0	0	0	0
52	34	26	31	35	55	45	45	45	55	2,6	2,2	1,9	0/0	0	0	0	1
53	41	29	35	40	55	45	45	45	55	2,8	2,3	2,1	0/0	0	0	0	1
54	31	18	21	25	45	55	45	45	45	3,4	3,0	2,5	0/1	0	1	1	1
55	19	29	35	40	44	44	44	44	44	1,3	1,1	1,0	1/1	1	1	1	1
56	26	18	21	24	45	45	45	45	55	2,9	2,5	2,2	0/0	1	1	1	0
57	26	20	24	28	44	45	45	45	44	2,6	2,2	1,9	1/1	0	0	0	1
58	36	27	32	35	45	45	45	45	55	2,7	2,3	2,1	0/0	1	1	1	0
59	24	29	34	39	44	44	44	44	44	1,7	1,4	1,2	1/1	1	1	1	1
60	27	24	28	32	45	45	45	45	55	2,3	1,9	1,7	0/0	1	1	1	0
61	27	24	28	34	45	45	45	45	55	2,3	1,9	1,6	0/0	1	1	1	0
62	25	34	41	46	44	44	44	44	55	1,5	1,2	1,1	0/0	1	1	1	0
63	22	21	25	28	44	44	45	45	45	2,1	1,8	1,6	0/1	1	0	0	0
64	24	29	34	42	44	44	44	44	44	1,7	1,4	1,1	1/1	1	1	1	1

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina Madis Sarapuu, sünnikuupäev 17.09.1993

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose

Inimese *Per3* geeni tandeemse korduse koopiaarvu määramine teise põlvkonna sekveneerimisandmetest.

mille juhendajateks on Tarmo Puurand, Maris Teder-Laving

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, 24.05.2016