Tartu Ülikool

Loodus- ja täppisteaduste valdkond

Arvutiteaduste instituut

Kermo Saarse

Informaatika 4. aasta

Haiguste geneetiliste korrelatsioonide arvutamine LD-skoori regressiooni meetodiga

Informaatika eriala bakalaureuse töö (9 EAP)

Juhendajad: Kaur Alasoo

Jaanika Kronberg

Tartu 2021

Haiguste geneetiliste korrelatsioonide arvutamine LD-skoori regressiooni meetodiga

Bakalaureuse töö

Kermo Saarse

Lühikokkuvõte

Geneetiline korrelatsioon on suurus, mis võimaldab hinnata kahe või enama haiguse ühist geneetilist algpõhjust. Selles töös kasutatakse Tartu Ülikooli Geenivaramu andmete peal LD-skoori regressiooni meetodit, mis võimaldab geneetilisi korrelatsioone arvutada ülegenoomsetest assotsiatsiooniuuringutest saadud andmete põhjal. Töö käigus tuli välja, et suur osa andmetest ei sobinud meetodi jaoks ning edukalt arvutatud korrelatsioonide seast osutusid väga vähesed statistiliselt oluliseks. Kui neid geneetilisi korrelatsioone võrreldi suhteliste riskidega, ilmnes positiivne korrelatsioon. Seega on töö põhisõnum see, et edaspidi tuleb taolistes uuringutes kasutada suuremaid valimeid.

CERCS teaduseriala: B110 bioinformaatika

Calculating genetic correlations between diseases using LD-score regression

Bachelor thesis

Kermo Saarse

Abstract

Genetic correlation is a measure which estimates a common genetic cause for two or more diseases. In this thesis, data from Estonian BioBank is used as input to LD-score regression method, which allows to calculate genetic correlations using data from genome wide association studies. It turned out that a large part of the data was not suitable for the method and very few of the successfully calculated results were statistically significant. When genetic correlations were compared to relative risk, a positive correlation was detected. The main lesson from this thesis is that larger sample sizes must be used for similar studies in the future.

CERCS research specialization: B110 Bioinformatics

Sisukord

Sissejuhatus

Haigusi uurides on selgunud, et teatud haigused esinevad sagedamini koos kui teised. Teisisõnu, kui on teada, et inimesel esineb mingi haigus, võib varasemate teadmiste põhjal järeldada, et teatud tõenäosusega võib tal mõni teine haigus veel olla. Põhjuseid selleks on erinevaid. Näiteks võib üks haigus olla teist haigust soodustavaks teguriks või koguni teise haiguse põhjustajaks. Teine võimalus on see, et mõlemal haigusel on sama algpõhjus, kusjuures see põhjus võib olla nii geneetiline kui ka keskkonnast tingitud. Teades, et kaks haigust on omavahel tugevalt korreleeritud, võib sellest olla abi haigestumise riski prognoosimiseks isegi siis, kui korrelatsiooni põhjus ei ole teada.

Selle töö eesmärk on tuvastada statistilisi ja bioloogilisi seoseid erinevate haiguste vahel. Töös esimene osa annab ülevaate LD-skoori regressiooni meetodist ning annab selle jaoks vajalikud geneetika alased taustateadmised. Samuti tutvustatakse statistilisi meetodeid haiguste koosesinemise sageduse mõõtmiseks. Töö praktilises osas kirjeldatakse meetodi kasutamist Geenivaramu andmetega Tartu Ülikooli HPC klastris.

Ühenukleotiidised polümorfismid

Heinaru (2012) järgi on ühenukleotiidine polümorfism (ingl. *single nucleotide polymorphism* – SNP) nukleotiidi positsioon DNA's, mis populatsiooni erinevatel indiviididel sisaldab erinevat nukleotiidi (A, T, G või C), samas kui seda vahetult ümbritsevad nukleotiidid on kõigil samad. Erinevaid nukleotiide, mis SNP's võivad esineda, nimetatakse alleelideks.

Keharakkude DNA koosneb mitmest eraldi lõigust ehk kromosoomist. Kromosoomid moodustavad paare, mis sisaldavad samu pärilikke tunnuseid määravaid geene, kuid ei ole oma järjestuselt identsed. Neid paare nimetatakse homoloogilisteks kromosoomideks. Samamoodi on ka iga SNP esindatud kahe üksusena, kumbki erineval homoloogilisel kromosoomil, mis võivad mõlemad sisaldada sama või erinevat alleeli.

Sugurakkude moodustumiseks jagunevad eellasrakud selliselt, et igast kromosoomide paarist jääb valmis sugurakku ainult üks, kusjuures see võib olla identne ühega eellasraku homoloogidest või kombinatsioon mõlemast. Viimasel juhul vahetavad homoloogilised kromosoomid omavahel võrdse pikkusega DNA osi ehk toimub rekombinatsioon. Mida suurem on samal kromosoomil asuva kahe SNP vaheline kaugus, seda suurem on tõenäosus, et sugurakku sattuv kromosoom saab nende SNP'de alleelid erinevatelt homoloogidelt. See on ka põhjus, miks teatud SNP'de (või mõnede muude DNA järjestuste) alleelide kombinatsioonid esinevad sagedasti koos: nad asuvad samal kromosoomil üksteisele väga lähedal, nende vahel toimub rekombinatsioon väga harva ning seega päranduvad need järgmisesse põlvkonda alati koos (juhul kui nad päranduvad). Seda nähtust nimetatakse aheldustasakaalutuseks (ingl. *linkage disequilibrium* – LD) (Heinaru, 2012).

Biology Stack Exchange, 2015 järgi saab enamik geneetilisi variante alguse ühel inimesel toimunud ühe nukleotiidi mutatsioonist, mis on pärandunud tema järglastele. Tõenäosus, et kellegil teisel täpselt sama nukleotiid muteerub ja edasi pärandub, on kaduvväike. Seetõttu on igal tavalisel SNP'l kaks võimalikku alleeli. Suurema sagedusega alleel on see nukleotiid, mis

esines igal inimesel samas positsioonis enne SNP tekkimist. Madalama sagedusega alleel on tekkinud üksiku mutatsiooni teel selles positsioonis ja sellest on saanud uus SNP (Biology Stack Exchange, 2015).

Kuna erinevused inimeste DNA-s ehk genotüübis võivad selgitada erinevusi silmaga nähtavates või mõõdetavates tunnustes ehk fenotüübis, on kasulik teada, kus täpselt SNP'd asuvad ning millised võimalikud alleelid neil on. Selleks loodi rahvusvaheline HapMap projekt (Gibbs *et al*. 2003), mille eesmärk oli kaardistada inimese genoomi varieeruvaid DNA järjestusi (sealhulgas SNP'sid). Lisaks SNP asukohtade kindlaks tegemisele uuriti ka alleelide vahelist aheldustasakaalutust. Selle alusel jaotati inimese kromosoomid piirkondadeks, mida nimetatakse haploplokkideks. Need piirkonnad liiguvad suure tõenäosusega rekombinatsiooni käigus ühelt kromosoomilt teisele tervikuna. Samuti tehti iga haploploki puhul kindlaks võimalikud SNP alleelide kombinatsioonid ehk haplotüübid. See aitab edasistes uuringutes kaardistada inimeste SNP'sid. Selle asemel, et teha kindlaks iga teadaoleva SNP alleelid, piisab uurida igast haploplokist vaid mõnda, nn tag-SNP'd. Kasutades HapMap projektist saadud andmeid, saab kindlaks teha, millisesse haplotüüpi leitud alleelid kuuluvad ning selle põhal järeldada ülejäänud samasse haploplokki kuuluvate SNP'de alleelid.

Ülegenoomsed assotsiatsiooniuuringud

Kui on teada, kus SNP'd asuvad ning millised alleelid neil on, on võimalik neid statistiliselt uurida. Wang *et al*. (2019) järgi on ülegenoomse assotsiatsiooniuuringu (ingl. *genome wide association study* – GWAS) eesmärk kindlaks teha, kas mõne SNP mingi alleel võiks olla seotud mõne fenotüübiga (tavaliselt haigusega). Selleks tehakse kindlaks iga uuringus osaleva inimese kõik tag-SNP'd ning jagatakse inimesed juhtude ja kontrollide gruppi vastavalt sellele, kas neil on uuritav haigus või ei ole. Kuna igal SNP'l on kaks võimalikku alleeli (olgu nendeks a ja A), võib homoloogiliste kromosoomide paaril olla kolm võimalikku kombinatsiooni ehk genotüüpi: aa, aA ja AA. Igale genotüübile seatakse vastavusse mingi arvuline väärtus.

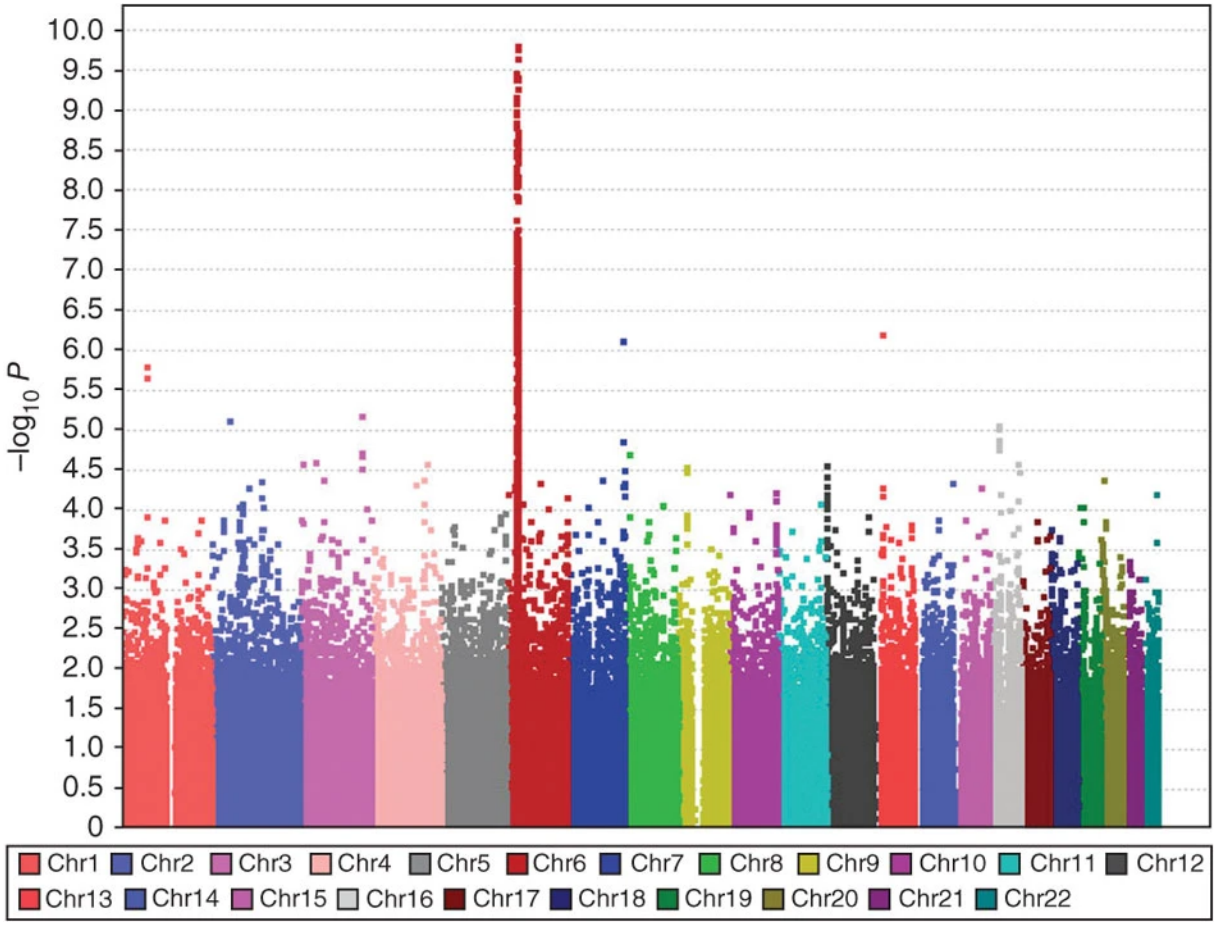
Nullhüpotees on seose puudumine genotüübi ja fenotüübi vahel ning alternatiivne hüpotees tähendab seose olemasolu. Tavaliselt on statistiliste testide olulisuse nivoo $\alpha = 0.05$, mis tähendab seda, et 5% testidest annavad väärpositiivse tulemuse. Kuna GWAS'is tehakse väga palju teste (testitakse LD'd arvesse võttes umbes miljonit sõltumatut SNP'd), siis tähendaks see väga palju valepositiivseid tulemusi. Seega seatakse olulisuse nivoo palju madalamaks, tavaliselt $\alpha = 5 \cdot 10^{-8}$.

Kui leitakse statistiliselt oluline seos mõne alleeli ja haiguse vahel, on alternatiivse hüpoteesi kehtimise korral kaks võimalust:

1. leitud alleel on haigust soodustav;
2. leitud alleel on haigust põhjustava DNA järjestusega (mis ei pruugi olla SNP) LD's.

Seega aitab GWAS kindlaks teha piirkonda genoomis, kus haigust põhjustav järjestus asub, kuid mitte seda järjestust ennast.

GWAS'is testitud SNP'd paigutatakse joonisele, mida nimetatakse *Manhattan plot*'iks (joonis 1). Joonise x-telg näitab positsiooni genoomis, kus erinevatele kromosoomidele vastavad alad on tähistatud eri värviga. X ja Y kromosoom on välja jäetud. Joonise y-teljel on p-väärtuse negatiivne kümnendlogaritm. Kuna see on p-väärtusega pöördvõrdeline, asuvad väikseima p-väärtusega (statistiliselt kõige olulisemad) SNP'd joonisel kõige kõrgemal. Üksteisega LD's olevad SNP'd on joonisel näha tulpadena.



Joonis 1. Manhattan plot, mis pärineb artiklist Bei JX. *et al*. (2010).

Üks esimesi GWAS-e on kirjeldatud artiklis Klein. *et al*. 2005, milles uuriti kollatähni kärbumise seotust SNP-dega. Uuriti 103611 SNP-d, millest statistiliselt oluliseks osutus ainult üks SNP tähisega rs380390. Sellest 1.8 kb kaugusel (1 kb tähendab 1000 nukleotiidi) asus kolmanda madalaima p-väärtusega SNP. HapMap projektist saadud andmete põhjal tuvastati, et mõlemad asuvad 41 kb pikkuses haploplokis, millel oli teada neli haplotüüpi. 99% uuritud kromosoomidest sisaldasid selles haploplokis neidsamu haplotüüpe. Neist kõige suurema riskiga haplotüüp kannab tähist N1, mis on neist ainuke, mis sisaldab rs380390 riskialleeli. Tuvastatud haploplokist 2 kb eemal asus SNP rs1061170, mis erinevalt kahest eelmisest oli mittesünonüümne. Mittesünonüümne SNP on selline, mis asub valku kodeerivas geenis ning mille erinevad alleelid põhjustavad erineva aminohappe lülitumist sünteesitavasse valku. Uuritud mittesünonüümsetest SNP-dest oli rs1061170 haigusega kõige tugevamini seotud. Lisaks sellele sisaldasid 97% haplotüübiga N1 kromosoomidest rs1061170 riskialleeli. Seega on võimalik, et rs1061170 on üks haigust soodustav tegur ning rs380390 riskialleel on

rs1061170 riskialleeliga tugevalt aheldunud. Artiklis toodi välja, et kõik nimetatud SNP-d asuvad ühe geeni sees, mida on ka varasemates uuringutes kollatähni kärbumisega seostatud.

Samas artiklis toodi välja ka see, et juhtude ja kontrollide valimisel on oluline, et mõlemad valimid oleksid pärit samast populatsioonist, sest muidu võivad statistiliselt oluliseks osutuda ka sellised SNP-d, millel on mingi kindla geograafilise levikuga alleelid.

Mõisted ja meetodid

Järgnevad alapeatükid kirjeldavad statistilisi meetodeid ja olulisi mõisteid, mida selles töös on kasutatud.

Geneetiline korrelatsioon

Et paremini kirjeldada fenotüübi seotust genotüübiga, on kasutusele võetud päritavuse ja geneetilise korrelatsiooni mõisted. Heinaru (2012) järgi modelleeritakse fenotüüpe summana

$$x = \mu + g + e, \qquad (1)$$

kus x on fenotüübi väärtus, μon populatsiooni keskväärtus, g on geneetiline mõju ning e on keskkonna mõju. Kui eeldada, et geneetiline mõju ja keskkonna mõju on omavahel sõltumatud, siis kehtib seos

$$D(x) = D(g) + D(e), (2)$$

kus D tähistab vastava tunnuse dispersiooni. Tunnuse laiatähenduslikuks päritavuseks nimetatakse geneetilise dispersioonide suhet fenotüübi dispersiooni:

$$H^2 = \frac{D(g)}{D(x)} (3)$$

Geneetilise mõju saab omakorda jaotada mitmeks komponendiks, millest olulisim on geenide aditiivne mõju, milles iga tunnust mõjutav geen (või SNP) omab mingit mõju (mis sõltub omakorda selle alleelidest) ning erinevate geenide mõjud liituvad. Aditiivse geneetilise komponendi dispersiooni tähistatakse sümboliga D(a) ning selle kaudu defineeritakse kitsatähenduslik päritavus:

$$h^2 = \frac{D(a)}{D(x)} (4)$$

van Rheenen *et al*. (2019) järgi nimetatakse kahe tunnuse vaheliseks geneetiliseks korrelatsiooniks nende geneetiliste komponentide vahelist korrelatsiooni. Teisisõnu, kui kaks tunnust on modelleeritavad summadena

$$y_1 = g_1 + e_1 (5)$$

$$y_2 = g_2 + e_2 (6)$$

kus g ja e tähistavad vastavate tunnuste geneetilist ja keskkonnast põhjustatud komponenti, siis nende tunnuste vaheline geneetiline korrelatsioon on

$$\rho_g = \frac{cov(g_1, g_2)}{\sigma_{g_1}\sigma_{g_2}} (7)$$

kus $cov(g_1, g_1)$ on geneetiliste komponentide kovariatsioon ja σ_g on vastava geneetilise komponendi standardhälve.

Suhteline risk

Kui geneetiline korrelatsioon kirjeldab kahe tunnuse võimalikku geneetilist seotust, siis suhteline risk, mida kirjeldab Kim (2017), mõõdab seost ainult nende tunnuste esinemissageduste vahel. Selleks on vaja, et uuritavaid tunnuseid kirjeldatakse kahendväärtusega (näiteks kas on haigus või ei ole). Riskiks nimetatakse haiguse esinemise tõenäosust populatsioonis. Olgu meil kaks vaadeldavat haigust. Nende haigustega inimeste arvud valimis on näidatud tabelis 1.

	Haigus 1 olemas	**Haigus 1 puudub**	**Kokku**
Haigus 2 olemas	a	b	a + b
Haigus 2 puudub	c	d	c + d
Kokku	a + c	b + d	N = a + b + c + d

Tabel 1. Kahe haiguse esinemissagedused valimis.

Haiguse 1 risk nende inimeste seas, kellel haigus 2 on olemas, on arvutatav valemiga

$$r_{12+} = \frac{a}{a+b} (8)$$

Haiguse 1 risk nende inimeste seas, kellel haigus 2 puudub, on arvutatav valemiga

$$r_{12-} = \frac{c}{c+d} (9)$$

Haiguse 1 suhteline risk haiguse 2 suhtes defineeritakse kui

$$r_{12} = \frac{r_{12+}}{r_{12-}} = \frac{a(c+d)}{c(a+b)} (10)$$

Analoogselt saab arvutada haiguse 2 suhtelise riski haiguse 1 suhtes:

$$r_{21} = \frac{a(b+d)}{b(a+c)} (11)$$

Suhteline risk näitab mitu korda on ühe haiguse risk suurem siis, kui on teada, et esineb ka teine haigus võrreldes sellega kui teist haigust ei ole. Seda, kas leitud seos on ka statistiliselt oluline, saab leida Fisheri täpse testi abil. Nullhüpotees on see, et ühe haiguse esinemine ei mõjuta teise haiguse tõenäosust. Eeldades, et eespool toodud tabeli marginaalsagedused (s.t. veerg/rida "kokku") on fikseeritud, saab arvutada tõenäosuse, et tabeli neljas lahtris on just sellised väärtused:

$$P = \frac{(a+b)Ca * (c+d)Cc}{NC(a+c)} (12)$$

kus aCb tähendab b-elemendiliste kombinatsioonide arvu a elemendist. Nullhüpoteesi testimiseks tuleb genereerida kõikvõimalikud tabelid, kus on valimiga samad marginaalsagedused. Iga sellise tabeli jaoks tuleb arvutada tõenäosus eespool toodud valemiga. Kõik sellised tõenäosused, mis on väiksemad või võrdsed esialgse tabeli põhjal arvutatud tõenäosusega tuleb kokku liita. Saadud summa on p-väärtus. Kui see on väiksem kui olulisusnivoo, võib alternatiivse hüpoteesi vastu võtta.

LD skoori regressioon

Valemi (7) abil geneetilise korrelatsiooni arvutamiseks oleks vaja teada iga valimi indiviidi geneetilise mõju suurust. Kahes Bulik-Sullivan *et al*. (2015) artiklis kirjeldatud LD-skoori regressiooni meetodiga on võimalik geneetilist korrelatsiooni hinnata hoopis GWAS kokkuvõttestatistikute põhjal. Meetod on implementeeritud tarkvaras LDSC (Bulik-Sullivan 2015), mida selles töös kasutatakse haiguste paarikaupa geneetiliste korrelatsioonide arvutamiseks. Järgnevas kahes alapeatükis on meetodi kirjeldus.

Ühe tunnuse LD skoori regressioon

Meetodi kirjeldus on võetud artiklist Bulik-Sullivan *et al*. 2015a. Selle meetodiga hinnatakse mingi tunnuse kitsatähenduslikku päritavust h^2. Olgu valimis N inimest, kellel on mõõdetud mingi fenotüübi väärtus. Samuti on igal inimesel mõõdetud M erineva SNP genotüüp. Fenotüüpi modelleeritakse järgmiselt:

$$\phi = X\beta + \epsilon, (13)$$

kus ϕ on fenotüüpide vektor mõõtmetega $N\times1$, X on genotüüpide maatriks mõõtmetega $N\times M$, β on efektisuuruste vektor mõõtmetega $M\times1$ ja vektor ϵ mõõtmetega $N\times1$ sisaldab keskonnamõjusid ja mitteaditiivseid geenimõjusid. $N\times1$mõõtmetega vektor$X\beta$tähistab seega aditiivseid geenimõjusid, milles iga SNP mõju sõltub selle efektisuurusest ja genotüübist.

Kahe SNP j ja k vahelist aheldustasakaalutlust mõõdetakse seose $r_{jk} = E(X_{ij}X_{ik})$ abil, mis ei sõltu i'st. Iga SNP j jaoks arvutatakse LD skoor

$$l_j = \sum_{k=1}^{M} r_{jk}^2, (14)$$

mis on seda suurem, mida rohkemate teiste SNP-dega on see SNP LD-s. Teisisõnu on see hinnang haploploki suurusele, milles see SNP asub. Samuti arvutatakse iga SNP jaoks efektisuuruse hinnang

$$\hat{\beta}_j = \frac{X_j^T \phi}{N}, (15)$$

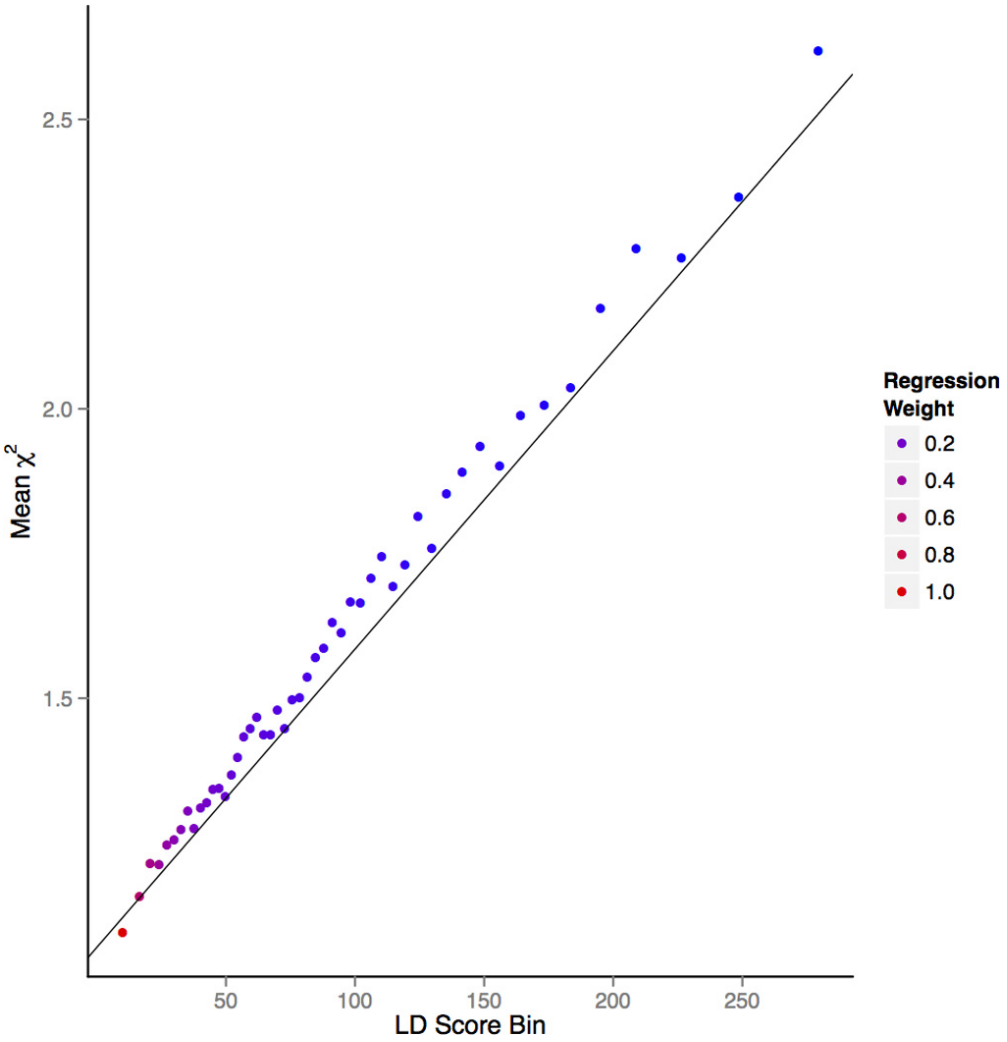
kus X_j tähistavab maatriksi X j'ndat veergu, ning χ^2-statistik

$$\chi_j^2 = N\hat{\beta}_j^2. (16)$$

χ^2-statistiku keskväärtus avaldub kui

$$E\left(\chi_j^2\right) = \frac{Nh_g^2}{M} l_j + 1 (17)$$

Järgmiseks leitakse lineaarne regressioonimudel, kus argumenttunnuseks on LD skoor l_j ja uuritavaks tunnuseks χ^2-statistik (vt. joonis 2). Mudelist saadakse regressioonikordaja $\alpha = \frac{Nh_g^2}{M}$, millest saab avaldada päritavuse hinnangu $h_g^2 = \frac{\alpha M}{N}$.



Joonis 2. LD-skoori regressiooni joonis. Iga värviline punkt tähistab LD-skoori kvantiili, punkti x-koordinaat on kvantiili keskmine LD-skoor ja y-koordinaat on kvantiili keskmine χ^2 -statistik

Tunnuste vaheline LD skoori regressioon

Meetodi kirjeldus on võetud artiklist Bulik-Sullivan *et al*. 2015b. Selle meetodiga hinnatakse kahe tunnuse vahelist geneetilist korrelatsiooni. Olgu meil kaks valimit suurustega N_1 ja N_2. Mõned inimesed võivad olla kaasatud mõlemasse valimisse, olgu nende arv N_s. Kummagis valimis on mõõdetud erinevate fenotüüpide väärtused, mida modelleeritakse samamoodi nagu eelmises peatükis:

$$y_1 = Y\beta + \delta, (18)$$

$$y_2 = Z\gamma + \epsilon, (19)$$

kus y_1 ja y_2 on valimite fenotüüpide vektorid mõõtmetega $N_1 \times 1$ ja $N_2 \times 1$, Y ja Z on genotüüpide maatriksid mõõtmetega $N_1 \times M$ ja $N_2 \times M$ (mõlemal valimil on mõõdetud samad SNP'd), β ja γ on efektisuuruste vektorid mõlemad suurusega $M \times 1$ ning δ ja ϵ on keskkonnamõjude ja mitteaditiivsete geenimõjude vektorid suurustega $N_1 \times 1$ ja $N_2 \times 1$.

Fenotüüpide päritavused defineeritakse kui $h_1^2 = \sum_{j=1}^{M} \beta_j^2$ ja $h_2^2 = \sum_{j=1}^{M} \gamma_j^2$ ning geneetiline kovariatsioon kui $\rho_g = \sum_{j=1}^{M} \beta_j \gamma_j$. Geneetiline korrelatsioon arvutatakse valemiga

$$r_g = \frac{\rho_g}{\sqrt{h_1^2 h_2^2}} (20)$$

Iga SNP *j* jaoks arvutatakse *z*-statistikud kummagi valimi jaoks:

$$z_{1j} = \frac{Y_j^T y_1}{\sqrt{N_1}}, (21)$$

$$z_{2j} = \frac{Z_j^T y_2}{\sqrt{N_2}} (22)$$

kus Y_j ja Z_j tähistavad maatriksite Y ja Z *j*'ndat veergu. *z*-statistikute korrutise keskväärtus avaldub valemiga

$$E\left(z_{1j} z_{2j}\right) = \frac{\sqrt{N_1 N_2}\rho_g}{M} l_j + \frac{N_s \rho}{\sqrt{N_1 N_2}}, (23)$$

kus ρ on fenotüüpide vaheline korrelatsioon nende N_s inimese seas, kes on mõlemas valimis. Hinnates lineaarse regressioonimudeli, kus argumenttunnuseks on LD skoor l_j ja uuritavaks

tunnuseks z-statistikute korrutis $z_{1j}z_{2j}$, leitakse regressiooni kordaja $\alpha = \frac{\sqrt{N_1 N_2}\rho_g}{M}$, millest saab avaldada geneetilise kovariatsiooni hinnangu $\rho_g = \frac{\alpha M}{\sqrt{N_1 N_2}}$. Geneetilise korrelatsiooni r_g leidmiseks on vaja leida ka päritavused h_1^2 ja h_1^2, millest kumbki leitakse eelmises peatükis kirjeldatud ühe tunnuse LD skoori regressiooni meetodiga.

ICD-10

ICD-10 on rahvusvaheline standard haiguste ja muude meditsiiniliste seisundite klassifitseerimiseks (World Health Organisation, 2020). Selles töös kasutatud andmestikes on fenotüübid kirjeldatud ICD-10 koodidega, millest igaüks koosneb tähest ja kahest numbrist. Koodide tähendused on saadud veebilehelt ICD10data.com, välja arvatud kood Z24, mis on saadud veebilehelt icd.who.int (World Health Organisation, 2019).

Töö käik

Töö sisendandmeteks olid GWAS kokkuvõttestatistikud, mille on koostanud Geenivaramu genoomika-metaboloomika kaasprofessor Toomas Haller. GWAS'i käigus genotüübiti 13755 inimest ning andmed haiguste kohta saadi 8 erinevast andmebaasist.

Iga ICD-koodi kohta on kokkuvõttestatistiku fail, milles on iga SNP kohta järgmised tulbad (lisaks muudele tulpadele):

1. Kromosoomi number (1, 2, … , 22, X või Y).
2. Positsioon – mitmendal nukleotiidil SNP asub.
3. RS number – unikaalne tähis iga SNP jaoks.
4. Esimene alleel.
5. Teine alleel.
6. P-väärtus.
7. Z-skoor.
8. Uuringute arv, milles seda SNP-d genotüübiti.
9. Nende inimeste arv, kellel õnnestus see SNP genotüüpida (enamasti suurem osa valimist).

Kokku oli 597 erineva ICD koodiga andmefaili. Mingil põhjusel ei olnud RS numbri tulbas RS numbrit, vaid hoopis kromosoomi number ja positsioon alakriipsuga kokkukirjutatult. Kromosoomi numbri ja positsiooni tulbas oli aga igal real väärtus -9. Seega oli vaja andmeid töödelda. RS numbrid leiti andmebaasi dbSNP (Sherry *et al*. 1999) versioonist b151 kromosoomi numbri ja positsiooni järgi. Töödeldud andmetes kirjutati kõik väärtused õigetesse tulpadesse ning jäeti alles ainult need 1217311 SNP'd, mis on olemas failis w_hm3.snplist, sest nende jaoks on arvutatud LD skoorid, mis asuvad kataloogis eur_w_ld_chr (Broad Institute, 2016).

Et kokkuvõttestatistikuid saaks tarkvaraga LDSC kasutada, tuleb need esmalt teisendada programmi jaoks sobivasse formaati. Selle jaoks on tarkvaras olemas skript munge_sumstats.py. Skripti rakendati kõigile 597'le andmefailile. Neist 482 puhul andis skript hoiatuse, et χ^2-statistiku keskväärtus on liiga väike (alla 1.02), mis tähendab seda, et andmed ei ole LD skoori regressiooni jaoks sobivad. Tõenäoliselt olid põhjuseks liiga väikesed SNP efektisuurused. Alles jäi seega 115 kokkuvõttestatisku faili, mis olid meetodi jaoks sobivad. Nende vahel arvutati kõikvõimalike paaride kaupa geneetilised korrelatsioonid, kasutades skripti ldsc.py, mis viib läbi LD skoori regressiooni. Kokku oli 6555 sellist paari, kuna ICD koodide järjekord paaris ei ole oluline ning ei ole mõtet arvutada ICD koodi geneetilist korrelatsiooni iseendaga (tulemus oleks 1). Neist 4329 puhul oli logifailis kirjas geneetilise korrelatsiooni väärtuseks nan. 422 paari logifailis oli mingil põhjusel veateade. Ainult 1804 logifaili sisaldasid edukalt arvutatud geneetilist korrelatsiooni. Erinevaid ICD koode nende paaride hulgas oli 65.

Järgmisena võrreldi tunnustepaaride vahelisi geneetilisi korrelatsioone suhteliste riskidega. Selleks kasutati andmestikku, mis on koostatud Eesti Geenivaramu andmete põhjal 2019. aasta seisuga (Sepideh Sadegh *et al*.). Kasutades elektroonilisi terviseandmeid 52,000 geenidoonori kohta, koostati fail, kus on iga isik-haigus (ICD10 kood) kohta rida, kui inimesel on see haigus elu jooksul diagnoositud. Kirjutati skript, mis leidis, mitmel inimesel

esineb iga haigust ning mitmel inimesel 2 haigusepaari kombinatsiooni, kõigi haigusepaaride jaoks.

Valminud tabeli igal real on järgmised andmed:

1. Kahe haiguse ICD koodid.
2. Kummagi haigusega inimeste arvud valimis.
3. Selliste inimeste arv, kellel on diagnoositud mõlemad haigused.
4. Kummagi haiguse suhteline risk teise haiguse suhtes.
5. Suhteliste riskide geomeetriline keskmine.

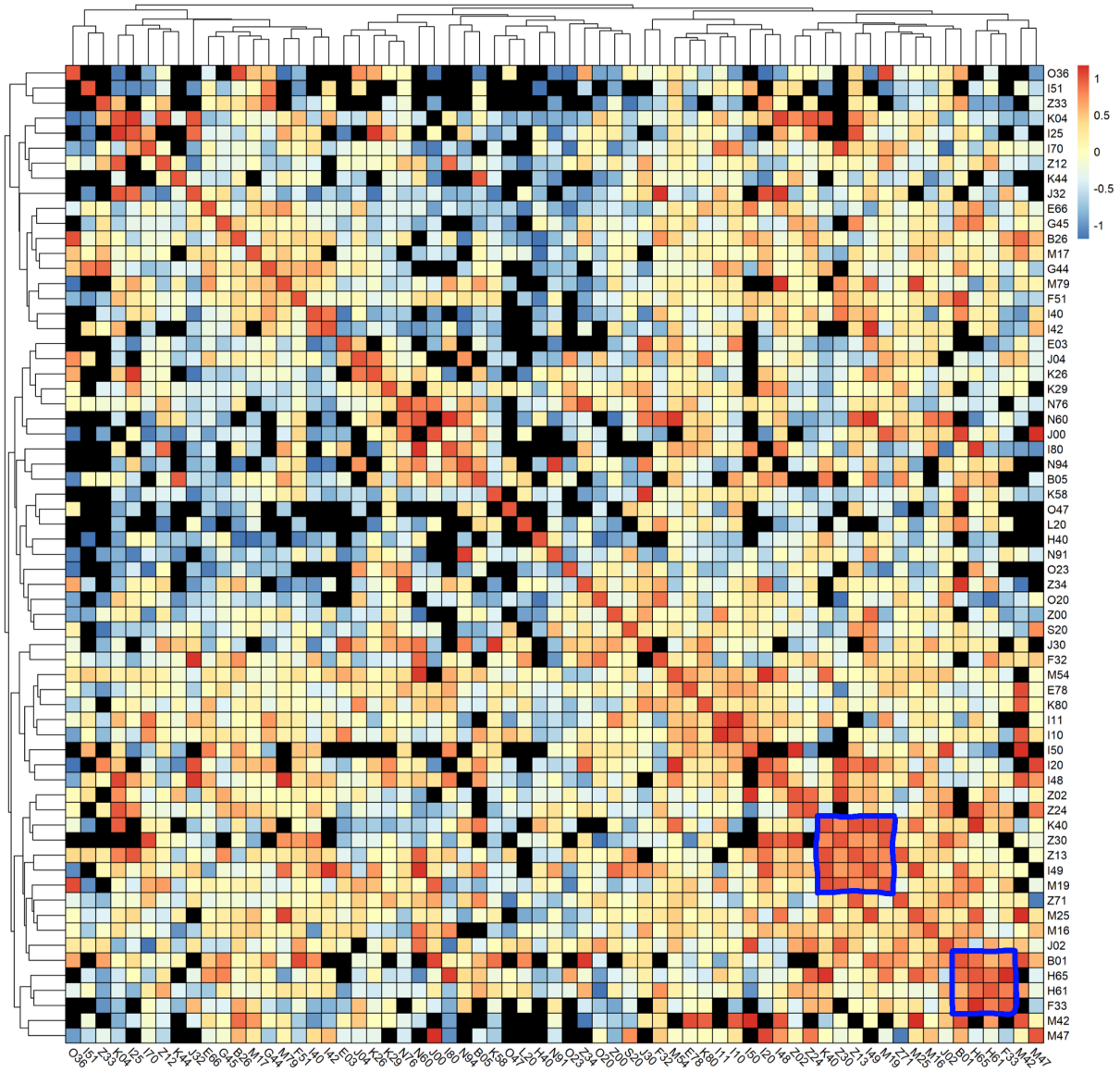
Tabelis oli kokku 304325 haiguste paari. Nendest kasutati ainult selliseid, kus mõlema haigusega inimeste arv oli vähemalt 100, kokku 47367 paari. Seejärel arvutati igaühe jaoks Fisheri täpse testi abil välja p-väärtus ja valiti välja need, millel p väärtus oli väiksem kui 0,05. Tabelis ei olnud nende inimeste arvu, kellel ei olnud kumbagi haigust diagnoositud. Küll aga olid seal olemas suhtelised riskid. Eeldades, et haiguse 1 suhteline risk haiguse 2 suhtes on arvutatud valemiga (10), saab sellest avaldada ilma haigusteta inimeste arvu:

$$r_{12} = \frac{a(c+d)}{c(a+b)} \Rightarrow d = \frac{r_{12}c(a+b)}{a} - c \quad (11)$$

Selliselt oli võimalik täita Fisheri testi jaoks vajalik 2x2 tabel. Tulemuseks oli 41577 paari, millel oli p-väärtus väiksem kui 0,05. Nende hulgast tuli omakorda välja valida need paarid, mille puhul oli geneetiline korrelatsioon edukalt arvutatud ja p-väärtus väiksem kui 0,05. Selliseid paare leiti 9.

Tulemused

Joonisel 3 on kujutatud programmiga LDSC arvutatud geneetilised korrelatsioonid maatriksi kujul. Selles on sinise joonega ümbritsetud kaks kõrge positiivse geneetilise korrelatsiooniga klastrit. Nendes olevad ICD koodid on loetletud tabelites 2 ja 3. Tabelites 4 ja 5 on näha vastavalt 20 kõige tugevamat positiivset ja 20 kõige tugevamat negatiivset geneetilist korrelatsiooni koos vastavate ICD koodide kirjelduste ja p-väärtustega. Tabelis 6 on näidatud kõik need 15 paari, mille puhul andis LDSC p-väärtuse, mis on väiksem kui 0,05. Kui p-väärtustele rakendati FDR korrektsiooni, osutusid kõik p-väärtused 1 lähedaseks. FDR (ingl. *False Discovery Rate*) korrektsioon on meetod, mis suurendab kõiki p-väärtuseid selliselt, et osad neist muutuvad suuremaks kui 0,05 ning seega väheneb statistiliselt oluliste tulemuste arv ning ühtlasi ka valepositiivsete tulemuste arv.



Joonis 3. 65 ICD10 koodi paarikaupa arvutatud geneetilised korrelatsioonid. Maatriks on diagonaali suhtes sümmeetriline. Musta värvi lahtrid tähendavad seda, et geneetilise korrelatsiooni arvutamine nende paaride puhul ei õnnestunud.

ICD kood	nimetus
K40	Kubemesong
Z30	Arstivisiit seoses rasestumisvastaste vahenditega
Z13	Osalemine sõeluuringul
I49	Südame rütmihäired
M19	Osteoartriit

Tabel 2. Üks kahest klastrist, kus kõikidel ICD koodidel on üksteisega kõrge positiivne geneetiline korrelatsioon.

ICD kood	nimetus
B01	Tuulerõuged
H65	Keskkõrvapõletik
H61	Väliskõrvahaigused
F33	Depressioon

Tabel 3. Üks kahest klastrist, kus kõikidel ICD koodidel on üksteisega kõrge positiivne geneetiline korrelatsioon.

ICD kood 1	ICD kood 2	Geneetiline korrelatsioon	p-väärtus
M42 (osteokondroos)	I50 (südamepuudulikkus)	1,1748	0,7537
M42 (osteokondroos)	M25 (liigesehaigused)	1,1708	0,5237
F33 (depressioon)	H65 (keskkõrvapõletik)	1,1656	0,3103
J00 (nina-neelupõletik)	M47 (lülisamba artroos)	1,1631	0,4155
K40 (kubemesong)	H65 (keskkõrvapõletik)	1,1509	0,4681
N60 (rindade fibroplastilised muutused)	I49 (südame rütmihäired)	1,1508	0.3067
J32 (ninakõrvalurgete põletik)	F32 (ühekordne depressioonihoog)	1,1409	0.2827
I11 (kõrge vererõhuga seotud südamehaigused)	I10 (Kõrgvererõhktõbi)	1,1402	0,0015
M79 (Other and unspecified soft tissue disorders)	I48 (kodade virvendusarütmia)	1,1392	0,1803
M54 (alaseljavalu)	I20 (rinnaangiin)	1,1381	0,0341
I42 (kardiomüopaatia)	I49 (südame rütmihäired)	1,1365	0,1068
Z02 (Encounter for administrative examination)	I50 (südamepuudulikkus)	1,1356	0,739
I80 (veenipõletik)	N60 (ridade fibroplastilised muutused)	1,1319	0,5039
M54 (alaseljavalu)	N60 (rindade fibroplastilised muutused)	1,1222	0,2424
K58 (soole	J30 (allergiline nohu)	1,1171	0,3065

ärritussündroom)			
I48 (kodade virvendusarütmia)	J32 (ninakõrvalurgete põletik)	1,1148	0,1887
Z34 (arstivisiit raseduse ajal)	B01 (tuulerõuged)	1,1083	0,4027
K04 (hambasäsi ja hambajuurt ümbritsevate kudede haigused)	I25 (südame isheemiatõbi)	1,1045	0,4708
K26 (kaksteistsõrmiksoole haavand)	I25 (südame isheemiatõbi)	1,1042	0,3639
N91 (puuduv või harv menstruatsioon)	N94 (naissuguorganite ja menstruatsiooniga seotud valu)	1,0991	0,3489

Tabel 4. 20 kõige tugevamat positiivset geneetilist korrelatsiooni. Koodile Z02 ei õnnestunud sobivat eestikeelset tõlget leida.

ICD kood 1	ICD kood 2	Geneetiline korrelatsioon	p-väärtus
O36 (imikuga seotud probleemid)	I20 (rinnaangiin)	-1,1894	0,565
B26 (mumps)	H40 (glaukoom)	-1,1887	0,1872
I40 (südamelihase põletik)	N94 (naissuguorganite ja menstruatsiooniga seotud valu)	-1,1832	0,3818
H61 (väliskõrvahaigused)	O20 (veritsus raseduse algstaaadiumis)	-1,1761	0,1141
J02 (neelupõletik)	I70 (ateroskleroos)	-1,1759	0,211
O36 (imikuga seotud probleemid)	M79 (Other and unspecified soft tissue disorders)	-1,1729	0,429
I51 (südamehaigus)	H40 (glaukoom)	-1,1722	0,4798
M17 (põlve osteoartroos)	H40 (glaukoom)	-1,1702	0,2936
Z30 (arstivisiit seoses rasestumisvastaste vahenditega)	O20 (veritsus raseduse algstaaadiumis)	-1,1674	0,6569
O36 (imikuga seotud probleemid)	O23 (kuseteede infektsioon raseduse ajal)	-1,1622	0,5852
I70 (ateroskleroos)	O23 (kuseteede infektsioon raseduse ajal)	-1,1596	0,2507
N60 (rindade fibroplastilised muutused)	Z00 (arstlik läbivaatus ilma eelneva kaebuse või diagnoosita)	-1,1583	0,3202
M42 (osteokondroos)	Z34 (arstivisiit raseduse ajal)	-1,1559	0,6269
K26 (kaksteistsõrmiksoole haavand)	J32 (ninakõrvalurgete põletik)	-1,1548	0,3056

S20 (rindkere vigastus)	Z33 (rasedus)	-1,1532	0,2343
H40 (glaukoom)	O20 (veritsus raseduse algstaaadiumis)	-1,145	0,3307
I40 (südamelihase põletik)	E03 (kilpnäärme alatalitlus)	-1,1397	0,1561
O23 (kuseteede infektsioon raseduse ajal)	E66 (ülekaal ja rasvumine)	-1,1335	0,134
O36 (imikuga seotud probleemid)	K04 (hambasäsi ja hambajuurt ümbritsevate kudede haigused)	-1,1318	0,4973
O36 (imikuga seotud probleemid)	J00 (nina-neelupõletik)	-1,1307	0,5169

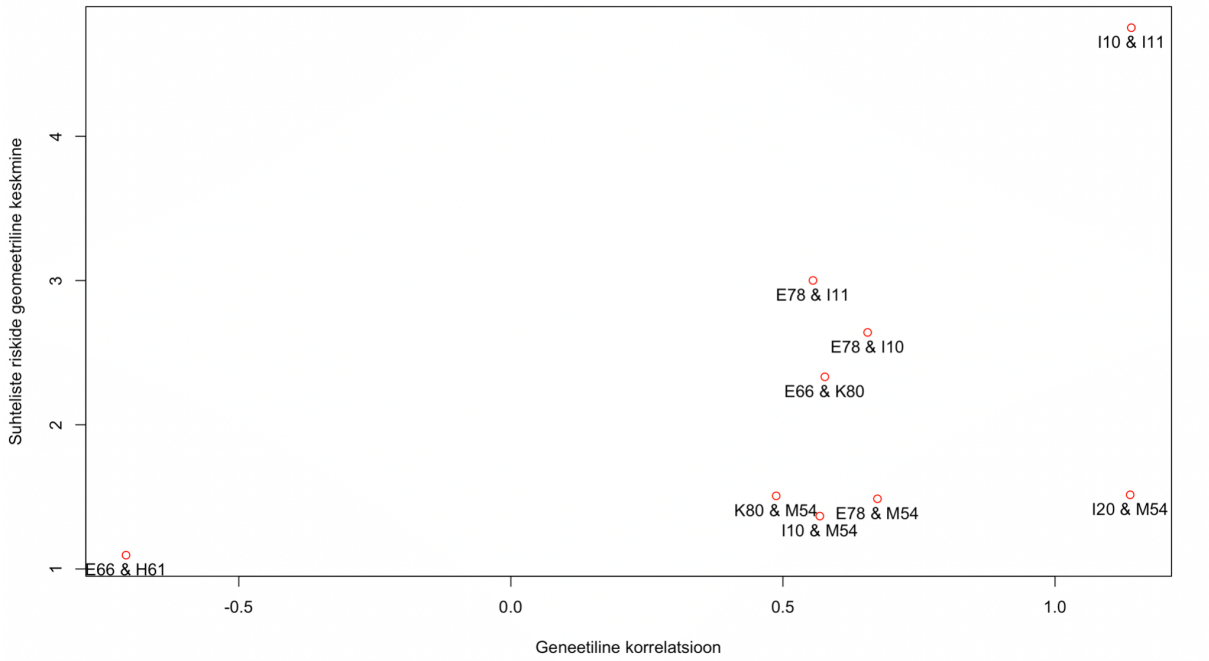
Tabel 5. 20 kõige tugevamat negatiivset geneetilist korrelatsiooni

ICD kood 1	**ICD kood 2**	**Geneetiline korrelatsioon**	**p-väärtus**
Z71 (tervisealase küsimusega arsti poole pöördumine)	Z02 (Encounter for administrative examination)	-0.7991	0,0363
H61(väliskõrvahaigused)	E66 (ülekaal ja rasvumine)	-0.7072	0,02
E78 (lipoproteiinide ainevahetuse häired)	Z24 (viirushaiguse vastu vaktsineerimine)	-0.4076	0,0448
K80 (sapikivitõbi)	M54 (alaseljavalu)	0.4876	0,0324
E78 (lipoproteiinide ainevahetuse häired)	I11(kõrge vererõhuga seotud südamehaigused)	0.5552	0,0444
M54 (alaseljavalu)	I10 (Kõrgvererõhktõbi)	0.5678	0,0357
N76 (põletik vagiinas ja vulvas)	E78 (lipoproteiinide ainevahetuse häired)	0.5682	0,0149

K80 (sapikivitõbi)	E66 (ülekaal ja rasvumine)	0.577	0,0185
H61 (väliskõrvahaigused)	Z24 (viirushaiguse vastu vaktsineerimine)	0.5899	0,0447
H61 (väliskõrvahaigused)	Z02 (Encounter for administrative examination)	0.5926	0,0363
E78 (lipoproteiinide ainevahetuse häired)	I10 (Kõrgvererõhktõbi)	0.6556	0,0021
E78 (lipoproteiinide ainevahetuse häired)	M54 (alaseljavalu)	0.6737	0,0021
Z24 (viirushaiguse vastu vaktsineerimine)	Z02 (Encounter for administrative examination)	0.9078	0,0016
I20 (rinnaangiin)	M54 (alaseljavalu)	1.1381	0,0341
I11 (kõrge vererõhuga seotud südamehaigused)	I10 (Kõrgvererõhktõbi)	1.1402	0,0015

Tabel 6. ICD-koodide paarid, mille nominaalne p-väärtus on väiksem kui 0,05.

Joonisel 3 on hajuvusdiagramm, mis saadi geneetilisi korrelatsioone ja suhtelisi riske omavahel võrreldes. Selles on kõik sellised paarid, millel nii suhtelise riski kui ka LDSC arvutatud p-väärtused on väiksemad kui 0,05. Üks paar (E66 ja H61, joonise alumises vasakus nurgas) on seal selline, kus suhteliste riskide geomeetriline keskmine on positiivne, kuid geneetiline korrelatsioon on negatiivne.



Joonis 3. Hajuvusdiagrammi iga punkt on ICD-koodide paar, horisontaalteljel on geneetiline korrelatsioon (arvutatud tarkvaraga LDSC) ning vertikaalteljel on suhteliste riskide geomeetriline keskmine.

Arutelu

Mõningad geneetilised korrelatsioonid leiti väga lähedaste või osaliselt kattuvate tunnuste vahel.

Kõrgvererõhktõbi (I10) ja kõrge vererõhuga seotud südamehaigused (I11) - teine eeldab esimese olemasolu.

Osteokondroos (M42) ja liigesehaigused (M25). Neist esimene on selgroolülide vaheketaste haigus. Seega on see lähedalt seotud liigesehaigustega või siis kuulub nende hulka.

Puuduv või harv menstruatsioon (N91) ning naissuguorganite ja menstruatsiooniga seotud valu (N94) - mõlemad on seotud menstruatsiooniga.

Leiti ka selliseid seoseid, mis on kinnitust leidnud teistes statistilistes uuringutes.

Kood E78 ehk lipoproteiinide ainevahetuse häired oli positiivses geneetilises korrelatsioonis koodidega I11 ehk kõrge vererõhuga seotud südamehaigused ja I10 ehk kõrgvererõhktõbi. E78 on üldine kood mitmete erinevate haiguste ja sümptomite kohta, millest üks on düslipideemia ehk kolesterooli ja triglütseriidide liiga kõrge või liiga madal tase veres. Besral et. al. 2019 läbi viidud uuringus selgus, et see soodustas südame pärgarterite lubjastumist. Kui patsientidel ei olnud varasemalt probleeme kõrge vererõhuga, suurendas düslipideemia pärgarterite lubjastumise riski 2,5 korda. Kõrge vererõhu olemasolu korral suurendas düslipideemia pärgarterite lubjastumise riski 18 korda.

E78 seos kõrge vererõhuga on ilmnenud ka Liu *et al.* (2016) uuringus, milles uuriti kõrge vererõhu komorbiidsust teiste haigustega, kasutades andmeid erinevatest Hiina haiglatest. Haiguse komorbiidsus kõrge vererõhuga defineeriti kui haiguse risk kõrge vererõhuga patsientide seas ehk selles töös valemi (8) järgi. 20 kõige tugevamast komorbiidsusest kolmandal kohal oli hüperlipideemia (13,81%). Hüperlipideemia tähendab normist kõrgemat triglütseriidide taset veres ja liigitub koodi E78 alla.

Seos oli ka sapikivide (E78) ning ülekaalu ja rasvumise (E66) vahel. Stender et. al. (2013) uuris sapikivide seotust kõrge kehamassiindeksiga. Leiti statistiliselt oluline seos kõrge kehamassiindeksi (KMI) ning sapikivide esinemise vahel. Lisaks genotüübiti igal uuringus osalenud patsiendil kolm geneetilist varianti, mida on varasemates uuringutes kõrge KMI-ga seostatud. Ka selles uuringus tuvastati nende geneetiliste variantide teatud alleelide seos kõrge KMI-ga. Tuvastati ka seos nende alleelide ja sapikivide vahel, kuigi see seos oli statistiliselt oluline ainult naiste seas. Seega on võimalik, et ülekaal (mida saab mõõta KMI abil) on sapikivide põhjustaja või soodustav tegur, eriti naiste hulgas.

Huizar et al. 2019 kirjeldab meditsiinilist diagnoosi nimetusega vatsakeste enneaegsetest kokkutõmmetest tingitud kardiomüopaatia. See tähendab, et vahel võivad südame vatsakesed teha normaalse südamerütmiväliseid lööke (kood I49 ehk südame rütmihäired). Paljudel juhtudel on see seisund ohutu, kuid mõningal juhul võib see põhjustada vasaku vatsakese puudulikkust, mis liigitub koodi I42 ehk kardiomüopaatia alla. Kui ravida südame rütmihäiret, paraneb ka sellest põhjustatud kardiomüopaatia. Selline olukord võis tingida positiivse geneetilise korrelatsiooni koodide I42 ja I49 vahel.

Tabelis 5, kus on 20 kõige tugevamat negatiivset korrelatsiooni, leidub palju ainult naistele rakenduvaid ICD koode, mis on enamasti seotud rasedusega. Võimalik, et GWAS ei võtnud arvesse seda, et need tunnused on mõõdetud ainult naistel ning seega võivad nende tunnuste geneetilised korrelatsioonid olla kallutatud.

Vaadates joonist 3, ilmneb tunnustepaaride seas mõningane positiivne korrelatsioon suhtelise riski ning geneetilise korrelatsiooni vahel. Kuna aga joonisel on ainult 9 paari, on selle põhjal siiski raske järeldusi teha.

Kokkuvõte

Töö sisendandmeteks olid GWAS kokkuvõttestatistiku failid, millest igaüks kirjeldab ühte ICD koodi. Töö eesmärk oli LDSC tarkvara kasutades leida geneetilisi korrelatsioone ICD koodide paaride vahel ning võrrelda neid suhteliste riskidega.

Kuna erinevaid ICD koode oli 597, võinuks erinevaid paare nende vahel olla 177906. Kuid geneetiline korrelatsioon õnnestus arvutada vaid 1804 paari kohta, millest algselt osutusid statistiliselt oluliseks 15 paari. Peale FDR korrektsiooni statistiliselt olulisi paare ei olnud. Nende seast leiti mõned sellised seosed, mida ka teised teadusartiklid kinnitavad. Nende 15 paari seast 9-le leiti ka suhtelised riskid. Geneetiliste korrelatsioonide ning suhteliste riskide seas ilmnes positiivne korrelatsioon.

Kuna nii vähesed paarid osutusid statistiliselt oluliseks, on põhisõnum see, et edaspidistes uuringutes tuleb kasutada oluliselt suuremaid valimeid. LD-skoori regressioon võib anda loogilisi tulemusi, kuid selles töös kasutatud ~14,000 indiviidiga valim on meetodi jaoks liiga väike. Õnneks on Tartu Ülikooli geenivaramu geenidoonorite arv viimaste aastate jooksul kasvanud 200,000-ni, mis muudab ka perspektiivikaks ka LD-skoori regressiooni rakendamise nendele andmetele.

Viited

Heinaru, A. (2012). Geneetika. Tartu Ülikooli Kirjastus.

Biology Stack Exchange. (2015) https://biology.stackexchange.com/questions/37020/why-do-almost-all-snps-have-two-alleles (vaadatud 07.05.2021).

Gibbs, R., Belmont, J., Hardenbol, P. *et al*. (2003). The International HapMap Project. *Nature*, 426, 789–796.

Wang, M.H, Cordell, H.J, Steen, K.V. (2019). Statistical methods for genome-wide association studies. *Seminars in Cancer Biology*, 55, 53–60

Bei JX. *et al*. (2010). A genome-wide association study of nasopharyngeal carcinoma identifies three new susceptibility loci. *Nature Genetics*, 42, 599–603.

Klein, R.J. *et al*. (2005). Complement Factor H Polymorphism in Age-Related Macular Degeneration. *Science*, 308, 385–389.

van Rheenen, W., Peyrot, W.J., Schork, A.J. *et al*. (2019). Genetic correlations of polygenic disease traits: from theory to practice. *Nature Review Genetics*, 20, 567–581

Kim, H.Y. (2017). Statistical notes for clinical researchers: Risk difference, risk ratio, and odds ratio. *Restorative Dentistry and Endodontics*. 42, 72-76

Bulik-Sullivan, B. *et al*. (2015a). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics*, 47, 291–295.

Bulik-Sullivan, B. *et al*. (2015b). An atlas of genetic correlations across human diseases and traits. *Nature Genetics*, 47, 1236–1241.

Bulik-Sullivan, B. (2015). LDSC [tarkvara]. https://github.com/bulik/ldsc (vaadatud 17.04.2020).

World Health Organisation. (2020). International Statistical Classification of Diseases and Related Health Problems (ICD). https://www.who.int/standards/classifications/classification-of-diseases (vaadatud 05.05.2021).

Sherry, S.T., Ward, M., Sirotkin, K. (1999) dbSNP – Database for Single Nucleotide Polymorphisms and Other Classes of Minor. *Genetic Variation. Genome Res.*, 9, 677–679.

Broad Institute (2016). https://data.broadinstitute.org/alkesgroup/LDSCORE/ (vaadatud 17.04.2020).

(Sepideh Sadegh *et al*.) Personal communication with Sepideh Sadegh, Jaanika Kronberg and Toomas Haller.

2021 ICD-10-CM Codes (2020). https://www.icd10data.com/ICD10CM/Codes (vaadatud 17.04.2020).

World Health Organisation. (2019). ICD-10 Version:2019. https://icd.who.int/browse10/2019/en (vaadatud 05.05.2021).

Ariyanti, R., Besral, B. (2019). Dyslipidemia Associated with Hypertension Increases the Risks for Coronary Heart Disease: A Case-Control Study in Harapan Kita Hospital, National Cardiovascular Center, Jakarta. *Journal of Lipids*, 2517013.

Liu, J., Ma J., Wang, J. et al. (2016). Comorbidity Analysis According to Sex and Age in Hypertension Patients in China. *International Journal of Medical Sciences*, 13, 99-107.

Stender, S., Nordestgaard, B.G, Tybjærg-Hansen, A. (2013). Elevated body mass index as a causal risk factor for symptomatic gallstone disease: A Mendelian randomization study. *Hepatology*, 58, 2133-2141.

Huizar, J.F., Ellenbogen, K.A, Tan, A.Y., Kaszala, K. (2019). Arrhythmia-Induced Cardiomyopathy: JACC State-of-the-Art Review. *Journal of the American College of Cardiology*, 73, 2328-2344.

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Kermo Saarse,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose "Haiguste geneetiliste korrelatsioonide arvutamine LD-skoori regressiooni meetodiga", mille juhendaja on Kaur Alasoo ja kaasjuhendaja on Jaanika Kronberg, reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada digitaalarhiivi DSpace kuni autoriõiguse kehtivuse lõppemiseni.

2. Annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi DSpace kaudu Creative Commonsi litsentsiga CC BY NC ND 3.0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, kuni autoriõiguse kehtivuse lõppemiseni.

3. Olen teadlik, et punktides 1 ja 2 nimetatud õigused jäävad alles ka autorile.

4. Kinnitan, et lihtlitsentsi andmisega ei riku ma teiste isikute intellektuaalomandi ega isikuandmete kaitse õigusaktidest tulenevaid õigusi.

Kermo Saarse
07.05.2021