UNIVERZA V LJUBLJANI
BIOTEHNIŠKA FAKULTETA

Aleš MAVER

SINTEZA HETEROGENIH GENOMSKIH PODATKOV PRI ODKRIVANJU DEDNIH DEJAVNIKOV ZA MULTIFAKTORSKE BOLEZNI

DOKTORSKA DISERTACIJA

Ljubljana, 2016

UNIVERZA V LJUBLJANI
BIOTEHNIŠKA FAKULTETA
INTERDISCIPLINARNI DOKTORSKI ŠTUDIJSKI PROGRAM BIOMEDICINA
ZNANSTVENO PODROČJE GENETIKA

Aleš MAVER

SINTEZA HETEROGENIH GENOMSKIH PODATKOV PRI ODKRIVANJU DEDNIH DEJAVNIKOV ZA MULTIFAKTORSKE BOLEZNI

DOKTORSKA DISERTACIJA

SYNTHESIS OF HETEROGENEOUS GENOMIC DATA IN DISCOVERY OF GENETIC FACTORS FOR MULTIFACTORIAL DISEASES

DOCTORAL DISSERTATION

Ljubljana, 2016

Na podlagi Statuta Univerze v Ljubljani ter po sklepu Senata Biotehniške fakultete in sklepa Komisije za doktorski študij z dne 7.9.2013 je bilo potrjeno, da kandidat izpolnjuje pogoje za opravljanje doktorata znanosti na Interdisciplinarnem doktorskem študijskem programu Biomedicina, znanstveno področje genetika. Za mentorja je bil imenovan prof. dr. Borut Peterlin.

Doktorsko delo je zaključek Interdisciplinarnega doktorskega študijskega programa Biomedicina, znanstveno področje genetika. Raziskovalno delo je bilo opravljeno na Kliničnem inštitutu za medicinsko genetiko, Univerzitetni klinični center Ljubljana.

Komisija za oceno in zagovor:

Predsednik: Prof. Dr. Simon Horvat
Univerza v Ljubljani, Biotehniška fakulteta, Oddelek za zootehniko

Član: Prof. Dr. Damjana Rozman
Univerza v Ljubljani, Medicinska fakulteta, Inštitut za biokemijo

Član: Prof. Dr. Jernej Jakše
Univerza v Ljubljani, Biotehniška fakulteta, Oddelek za agronomijo

Datum zagovora: 07.06.2016

Podpisani izjavljam, da je disertacija rezultat lastnega raziskovalnega dela. Izjavljam, da je elektronski izvod identičen tiskanemu. Na univerzo neodplačno, neizključno, prostorsko in časovno neomejeno prenašam pravico shranitve avtorskega dela v elektronski obliki in reproduciranja ter pravico omogočanja javnega dostopa do avtorskega dela na svetovnem spletu preko Digitalne knjižnice Biotehniške fakultete.

Aleš Maver

KLJUČNA DOKUMENTACIJSKA INFORMACIJA (KDI)

ŠD Dd
DK UDK 575:616.8(043)=163.6
KG multifaktorske bolezni/omske študije/integracija podatkov/sekvenciranje nove generacije/multipla skleroza/Parkinsonova bolezen/molekularna genetika
AV MAVER, Aleš, dr. med.
SA PETERLIN, Borut (mentor)
KZ SI-1000 Ljubljana, Jamnikarjeva 101
ZA Univerza v Ljubljani, Biotehniška fakulteta, Interdisciplinarni doktorski študijski program Biomedicina, znanstveno področje genetika
LI 2016
IN SINTEZA HETEROGENIH OMSKIH PODATKOV PRI ODKRIVANJU DEDNIH DEJAVNIKOV ZA MULTIFAKTORSKE BOLEZNI
TD Doktorska disertacija
OP X, 93 str., 7 pregl., 20 sl., 185 vir.
IJ sl
JI sl/en
AI Tehnološki napredek na področju molekularne genetike je v zadnjem času omogočil nov vpogled v etiologijo multifaktorskih bolezni. Kljub bogatim informacijam, ki jih pridobimo z novimi omskimi pristopi, predstavlja interpretacija visoko dimenzionalnih podatkov pomemben izziv, tehnična in biološka ponovljivost rezultatov pa sta relativno nizki. Z namenom izboljšanja interpretacije rezultatov smo razvili nov pristop za sintezo heterogenih podatkov, ki temelji na ugotavljanju genomskih področij z zbiranjem signalov iz raznovrstnih omskih študij. Na primeru Parkinsonove bolezni (PB) smo pokazali zmogljivost pristopa pri identifikaciji vzročnih genov s poznano vlogo pri sporadični in družinski obliki PB (SNCA in UCHL1), prav tako pa smo lahko identificirali nove obetavne kandidatne gene za PB (YWHAE). Pristop smo uporabili tudi pri integraciji širokega nabora omskih študij pri multipli sklerozi (MS) in pokazali uporabnost pristopa pri interpretaciji podatkov eksomskega sekvenciranja pri bolnikih z MS. S sintezo podatkov različnih tipov omskih študij in rezultatov eksomskega sekvenciranja smo identificirali verjetno ali možno patogene redke različice genov pri 16,6 % bolnikov z družinsko obliko in pri 22,5 % bolnikov s sporadično obliko MS. Pričakujemo, da lahko z razvitim pristopom pomembno izboljšamo odkrivanje novih kandidatnih genov in dednih dejavnikov pri multifaktorskih bolezni.

KEY WORDS DOCUMENTATION (KWD)

DN Dn
DC UDC 575:616.8(043)=163.6
CX multifactorial diseases/omic studies/data integration/next generation sequencing/multiple sclerosis/Parkinson disease/molecular genetics
AU MAVER, Aleš
AA PETERLIN, Borut (supervisor)
PP SI-1000 Ljubljana, Jamnikarjeva 101
PB University of Ljubljana, Biotechnical faculty, Interdisciplinary Doctoral Programme in Biomedicine, Scientific Field Genetics
PY 2016
TI SYNTHESIS OF HETEROGENEOUS GENOMIC DATA IN DISCOVERY OF GENETIC FACTORS FOR MULTIFACTORIAL DISEASES
DT Doctoral dissertation
NO X, 93 p., 7 tab., 20 fig., 185 ref.
LA sl
AL sl/en
AB Technological advances in the field of molecular genetics have offered novel insight into the etiology of multifactorial diseases. Despite its rich information content, the interpretation of such high-dimensional results is challenging and the results are of limited technical and biological reproducibility. Aiming to improve the interpretation of such studies, we developed a novel approach for synthesis of heterogeneous omic data based on genomic localization of identified alterations. Using the example of Parkinsons disease (PD), we have shown the utility of this approach for identification of genes with established association with familial and sporadic PD (SNCA and UCHL1) and we also showed that the approach enabled identification of novel PD genes (YWHAE). We have used the approach for integration of a large body of omic studies in multiple sclerosis (MS) and have shown its utility in interpreting exome sequencing data in patients with MS. With the synthesis of data from omic studies, we were able to identify rare, likely or possibly pathogenic variant in 16,6 % of patients with familial MS and in 22,5 % patients with sporadic MS. We anticipate that presented strategy could significantly improve the detection of novel candidate genes and genetic risk factors for multifactorial diseases.

KAZALO VSEBINE

KAZALO PREGLEDNIC

KAZALO SLIK

OKRAJŠAVE IN SIMBOLI

CSF	cerebrospinalna tekočina (angl. *Cerebrospinal fluid*)
CŽS	centralni živčni sistem
EAE	eksperimentalni avtoimuni encefalitis (angl. *Experimental autoimmune encephalitis*)
ExAC	Konzorcij za združevanje eksomskih podatkov (angl. *Exome Aggregation Consortium*)
GEO	repozitorij ekspresijskih podatkov (angl. *Gene expression omnibus*)
GO	ontologija funkcij genov (angl. *GeneOntology*)
GWAS	asociacijske študije celotnega genoma (angl. *Genome-wide association studies*)
HGNC	Konzorcij za nomenklaturo humanih genov (angl. *Human gene nomenclature consortium*)
HLA	humani levkocitni antigen (angl. *Human leukocyte antigen*)
HPO	ontologija fenotipov človeka (angl. *Human Phenotype Ontology*)
KEGG	Kjotska enciklopedija genov in genomov (angl. *Kyoto Encyclopedia of Genes and Genomes*)
miRNA	mikroRNA
MS	multipla skleroza
NGS	sekvenciranje nove generacije (angl. *Next-generation sequencing*)
OMIM	baza podatkov o mendelskih boleznih človeka (angl. *Online Mendelian Inheritance in Man*), (OMIM, 2016)
PB	Parkinsonova bolezen
PFP	napovedana stopnja lažno pozitivnih rezultatov (angl. *Predicted false positive rate*)
SNP	polimorfizem enega nukleotida (angl. *Single nucleotide polymorphism*)

1 UVOD

1.1 OPREDELITEV PROBLEMA

Multifaktorske bolezni so poglavitni vzrok obolevnosti in umrljivosti v razvitem svetu in zajemajo kardiovaskularne, onkološke, avtoimune in druge pogoste skupine bolezni (Manolio in sod., 2009). Glede na trenutno poznavanje njihove etiologije nastanejo kot posledica skupnega prispevka dednih dejavnikov, dejavnikov okolja in njihovega medsebojnega vplivanja (Clayton in McKeigue, 2001; Manolio in sod., 2009). Rezultati epidemioloških in genetskih študij kažejo, da je število vzročnih dejavnikov pri nastanku tovrstnih bolezni veliko, prispevek posameznega dejavnika pa majhen (Clayton in McKeigue, 2001; Manolio in sod., 2009). Zaradi tega je odkrivanje posameznih dejavnikov tveganja za te bolezni zahtevno in zahteva senzitivne metode za odkrivanje majhnih prispevkov posamičnih vzročnih dejavnikov. Razkrivanje genetskih dejavnikov za multifaktorske bolezni predstavlja osnovo za poznavanje patogenetskih procesov, ki vodijo v nastanek bolezni, prav tako pa imajo ključen pomen pri napovedovanju, zgodnjem odkrivanju bolezni, poznavanju prognoze, ter napovedovanju odziva na morebitno zdravljenje. Ker je število genov, ki prispevajo k nastanku teh bolezni praviloma veliko, njihov posamičen učinek pa majhen, je napredek pri identifikaciji vzročnih genov počasen, klinična uporabnost odkritih sprememb pa majhna (Ioannidis in sod., 2006). Kljub obsežnim raziskovalnim prizadevanjem trenutno z znanimi genetskimi dejavniki tveganja še vedno ne moremo v celoti pojasniti prispevka dednosti, ki je bil ugotovljen v epidemioloških študijah večine multifaktorskih bolezni (Manolio in sod., 2009).

Hiter tehnološki razvoj visoko zmogljivih metod v genetiki hkrati z razkritjem zaporedja večjega dela človeškega genoma, je v zadnjih letih omogočil povsem nove možnosti vpogleda v genetsko etiologijo multifaktorskih bolezni. Nove omske metode preiskovanja s tehnologijo mikromrež in tehnologijo sekvenciranja nove generacije omogočajo sočasno preiskovanje velikega števila molekularnih bolezenskih sprememb, na nivoju celotnega genoma, epigenoma, transkriptoma, proteoma in drugih molekularno-bioloških nivojih (Cirulli in Goldstein, 2010; Grant in Hakonarson, 2008). Kljub prednostim tovrstnih metod, so znane tudi njihove pomembne omejitve, predvsem s plati statistične obravnave pridobljenih visoko-dimenzionalnih podatkov (Gregersen in Brehrens, 2003; Simon in sod., 2003). Preiskovanje povezave velikega števila sprememb z boleznijo predstavlja statistično preverjanje velikega števila statističnih hipotez, kar ima za posledico tudi povečano število lažno pozitivnih rezultatov. Po drugi strani pa se v teh študijah poveča možnost, da se resnično pozitivni rezultati izgubijo v statističnem šumu (Nadon in Shoemaker, 2002). Omenjeno predstavlja ključen problem pri odkrivanju diskretnih molekularnih sprememb pri multifaktorskih boleznih.

Dosedanji pristopi globalnega odkrivanja bolezenskih sprememb so bili v veliki meri osredotočeni na preiskovanje sprememb ločeno po posameznih molekularnih nivojih. Dejanski patogenetski in homeostatski procesi pa so na celičnem nivoju tesno prepleteni in niso omejeni le na posamičen omski nivo. Pri odkrivanju genov nam informacija o sočasnih spremembah na več molekularnih nivojih lahko pomaga pri razločevanju dejanskih bioloških sprememb od tistih, ki nastanejo zaradi tehničnih in statističnih vzrokov.

Enoznačnega in učinkovitega pristopa za tako integracijo heterogenih 'omskih' podatkov trenutno še ne poznamo. Prvi problem pri integraciji se pojavi že pri združevanju podatkov študij na istem molekularnem nivoju, saj meritve sprememb pogosto potekajo na različnih tehnoloških platformah (kot primer se pri merjenju sprememb v izražanju genov uporabljajo različne vrste mikromrež). Uspešnost združevanja je tu omejena že v tem koraku - zaradi nepopolnih in nestandardnih anotacij ugotovljenih sprememb (Cahan in sod., 2007). Problematika je še večja, ko želimo združiti podatke omskih študij na različnih molekularnih nivojih, kjer se uporabljajo popolnoma različni tehnološki pristopi. Večina predhodnih poskusov tovrstne integracije je bila omejena na prekrivanje skupin genov, ki so bili povezani z boleznijo v različnih vrstah študij. Ob tem pride do izgube informacij, saj so na eni strani številne odkrite spremembe v področjih med geni, po drugi strani pa dokazi iz nedavnih študij kažejo da narava ugotovljenih molekularnih sprememb ni omejena na področja genov. Zato se pojavlja potreba po novih pristopih k integraciji podatkov različnih tipov omskih študij za odkrivanje biološko pomembnih sprememb v kontekstu multifaktorskih bolezni. S tem razlogom smo se odločili za poskus razvoja novega pristopa za sintezo omskih podatkov z uporabo genomskega položaja kot skupnega imenovalca za integracijo podatkov – pristop *pozicijske integracije* genomskih podatkov. S takim pristopom se izognemo izgubi informacij pri preprosti uporabi gena kot skupnega imenovalca integracije. Uporabimo lahko podatke, ki niso vezani na gene (regije z dokazano genetsko vezavo z boleznijo, epigenetske spremembe, polimorfizme v medgenskih področjih), prav tako pa pridobimo tudi informacijo o sovplivu sprememb zaradi vezane epistaze (na primer sprememb, ki ležijo na genomu blizu, a v različnih genih).

Dodatni napredek pri odkrivanju genetskih dejavnikov za multifaktorske bolezni so v zadnjem času prinesle tudi nove možnosti sekvenciranja, ki so omogočile dostopno eksomsko in genomsko sekvenciranje in tako ponudile vpogled v prispevek redkih genetskih različic pri nastanku multifaktorskih bolezni (Cirulli in Goldstein, 2010). Pri sekvenciranju celotnega eksoma in genoma dobimo vpogled v velik del zaporedja humanega genoma, kar prinaša bistveno povečano množico možnih etiološko pomembnih sprememb v dednem zapisu. Ker do danes še ni znano, kakšen je dejanski model dedovanja

multifaktorskih bolezni, sta napoved patogenosti in ugotavljanje etiološkega pomena tako ugotovljenih različic zahtevna. Vključitev informacij iz omskih študij v interpretacijo podatkov pridobljenih z eksomskim in genomskim sekvenciranjem ponuja nove možnosti za interpretacijo teh podatkov in za odkrivanje vzročnih genetskih sprememb, ki prispevajo k razvoju multifaktorskih bolezni. Glede na to, da postaja eksomsko in genomsko sekvenciranje v zadnjem času poglavitni pristop za odkrivanje genetskih sprememb, povezanih z multifaktorskimi bolezni, bo izboljšanje interpretacije rezultatov s pristopom *pozicijske integracije* pomembno prispevalo poznavanju specifičnih dednih dejavnikov in izboljšalo naše poznavanje genetske osnove multifaktorskih bolezni.

1.2 NAMEN RAZISKAVE

Namen raziskave je razviti izviren pristop *integrativne pozicijske genomike* za sintezo heterogenih genomskih podatkov in s tem premostiti probleme, ki se pojavljajo pri dosedaj opisanih pristopih za tovrstno analizo omskih podatkov. Trenutno ne poznamo metode, s katero bi lahko združevali vse vrste podatkov, ki jih dobimo z novejšimi in raznolikimi visoko zmogljivimi metodami v molekularni biologiji. Namen našega dela je razvoj pristopa *pozicijske integracije*, s katerim bo mogoče na neposreden način in brez izgube informacij združiti večino vrst omskih podatkov, česar doslej opisane metode niso omogočale.

Pozicijska integracija predstavlja nov, z dokazi podprti informacijski vir za interpretacijo podatkov eksomskega in genomskega sekvenciranja, s katero lahko izboljšamo klinično uporabnosti in validnost te metode tudi v kontekstu multifaktorskih bolezni. Dosedanje metode za interpretacijo genetskih sprememb odkritih z eksomskim in genomskim sekvenciranjem temeljijo na populacijskih podatkih, podatkih o evolucijski ohranjenosti različic v genomu in na podlagi računske napovedi patogenosti. Pristop z integracijo heterogenih omskih podatkov predstavlja nov vir za interpretacijo odkritih genetskih različic in bo omogočal nov pristop pri selekciji genetskih sprememb različic za multifaktorske bolezni. S pristopom je možno bolj učinkovito identificirati dedne dejavnike za nastanek multifaktorskih bolezni in posledično prispevati k boljšemu razumevanju genetske arhitekture multifaktorskih bolezni.

1.3 RAZISKOVALNE HIPOTEZE

V okviru doktorske naloge smo postavili naslednji hipotezi:

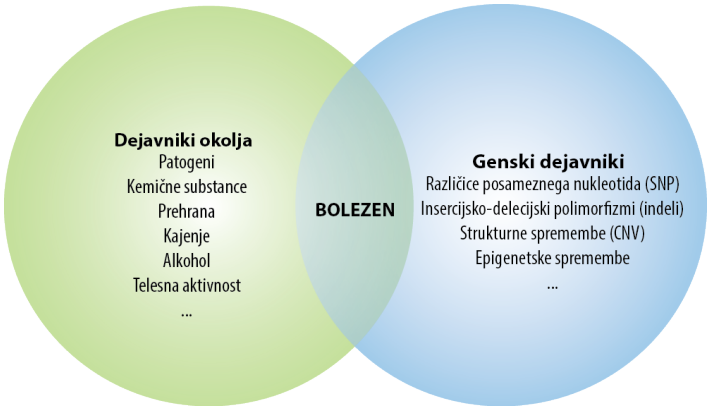
- Z izvirnim pristopom *pozicijske integracije* lahko pomembno izboljšamo učinkovitost sinteze heterogenih genomskih podatkov.

- Pristop *pozicijske integracije* heterogenih genomskih podatkov lahko pomembno izboljša interpretacijo rezultatov eksomskega in genomskega sekvenciranja v kontekstu multifaktorskih bolezni.

2 PREGLED OBJAV

2.1 MULTIFAKTORSKE BOLEZNI

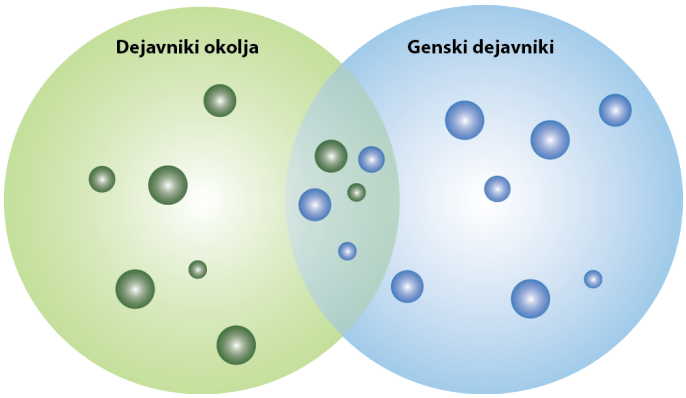
V primerjavi z mendelsko dedovanimi boleznimi, ki nastanejo zaradi okvare v posameznem genu, nastanejo multifaktorske bolezni kot posledica sočasnega prispevka več dednih dejavnikov v interakciji z dejavniki okolja (Slika 1). V primerjavi z monogenskimi boleznimi, ki so pretežno redke ali zelo redke, so multifaktorske bolezni pogoste in prizadenejo pomemben delež prebivalstva.



Slika 1: Trenutni model multifaktorskih bolezni (Oksenberg in Baranzini 2010).
Pri nastanku multifaktorskih bolezni sodeluje hkrati več dejavnikov okolja in več dednih dejavnikov, ki v medsebojnem sodelovanju privedejo do razvoja bolezenskega fenotipa.
Figure 1: Current model of etiology of multifactorial diseases (Oksenberg and Baranzini 2010).
Interaction of multiple environmental and genetic factors participate together in development of multifactorial diseases.

2.1.1 Klasični model nastanka multifaktorskih bolezni

Klasični modeli nastanka multifaktorskih bolezni predvidevajo, da k nastanku bolezni vodi medsebojni prispevek večjega števila posameznih dednih dejavnikov. Glede na to, da je predvideno število dejavnikov, ki pri posamezniku privede do bolezni veliko, je pri klasičnem pojmovanju nastanka multifaktorskih bolezni učinek posameznega dejavnika majhen (Slika 2).



Slika 2: Klasični model nastanka multifaktorskih bolezni z majhnim prispevkom številnih dednih dejavnikov in dejavnikov okolja.
Figure 2: In the classical model of etiology of multifactorial diseases, multiple genetic factors with minor effect sizes contribute to genesis of multifactorial diseases.

2.1.2 Dokazi za dedno pogojenost multifaktorskih bolezni

Dokaze za dedno pogojenost multifaktorskih bolezni črpamo iz epidemioloških študij, študij povečane pojavnosti bolezni v družinah, s študijami enojajčnih in dvojajčnih dvojčkov, prav tako pa tudi študije genetske vezave kažejo na pomembno vlogo specifičnih regij v človeškem genomu pri nastanku multifaktorskih bolezni.

S primerjavo tveganja v splošni populaciji s tveganjem pri enojajčnih dvojčkih, dvojajčnih dvojčkih ali drugih sorodnikih obolelih je mogoče za številne pogoste bolezni opredeliti mero relativnega tveganja (angl. *relative risk*) pri sorodnikih obolelih, ki odraža pomen dednih dejavnikov tveganja, na podlagi tega pa je mogoče oceniti tudi heritabiliteto bolezni (angl. *heritability*). Pri nekaterih boleznih (na primer epilepsija, avtizem in sladkorna bolezen tipa 1) je stopnja heritabilitete zelo visoka, pri drugih pa je ocena dedne komponente nižja v primerjavi s prispevki okolja (Parkinsonova bolezen, karcinom debelega črevesja in sladkorna bolezen tipa 2). V preglednici 1 je navedeno nekaj poglavitnih primerov multifaktorskih bolezni in njihove heritabilitete glede na podatke iz literature.

Preglednica 1: Primeri pomembnejših multifaktorskih bolezni in pripadajoče ocene heritabilitete glede na epidemiološke podatke v literaturi.
* Mera heritabilitete predstavlja delež predispozicije bolezni, ki jo lahko pripišemo dednim dejavnikom
Table 1: Examples of most important multifactorial diseases and heritability estimates for diseases, based on epidemiological evidence.
* Heritability is an estimated proportion of disease susceptibility that can be contributed to heritable factors.

Bolezen	**Heritabiliteta***	**Citat**
Avtizem	90 %	(Freitag, 2007)
Epilepsija	88 %	(Kjeldsen in sod., 2001)
Sladkorna bolezen tipa 1	88 %	(Hyttinen in sod., 2003)
Celiakija	87 %	(Nistico in sod., 2006)
Shizofrenija	81 %	(Sullivan in sod., 2003)
Alzheimerjeva bolezen	79 %	(Gatz in sod., 2006)
Starostna degeneracija makule	71 %	(Klaver in sod., 1998)
Bipolarna motnja	70 %	(Smoller in Finn, 2003)
Debelost	70 %	(Walley in sod., 2006)
Multipla skleroza	64 %	(Westerlind in sod., 2014)
Revmatoidni artritis	55 %	(Harney in sod., 2008)
Astma	30 %	(Tan in sod., 2005)
Hipertenzija	30 %	(Agarwal in sod., 2005)
Parkinsonova bolezen	30 %	(Do in sod., 2011)
Sladkorna bolezen tipa 2	26 %	(Poulsen in sod., 1999)
Karcinom debelega črevesja	13 %	(Czene in sod., 2002)

Klasični pristopi z genskim kartiranjem in pristop kandidatnih genov so v letu 1983 že omogočili lokalizacijo prvega vzročnega gena za monogensko bolezen pri človeku (za Huntingtonovo bolezen), v sledečih letih pa so s temi pristopi identificirali še številne druge gene za monogenske bolezni (Altshuler in sod., 2008; Gusella in sod., 1983). Kljub uspešnosti pri odkrivanju genov za monogenske bolezni, pa je bila moč teh pristopov pri odkrivanju majhnih prispevkov številnih dednih dejavnikov za multifaktorske bolezni omejena. Odkrivanje dednih dejavnikov za multifaktorske bolezni je tako v začetku potekalo počasi in z minimalnimi uspehi asociacijskih študij posameznih kandidatnih genov.

2.1.3 Napredek pri odkrivanju dednih dejavnikov za multifaktorske bolezni z omskimi pristopi

Pomemben napredek pri odkrivanju dednih dejavnikov pri nastanku je prinesel tehnološki razvoj visoko zmogljivih metod v genetiki hkrati z razkritjem zaporedja večjega dela človeškega genoma (International Human Genome Sequencing Consortium, 2004). Nove *omske* metode preiskovanja s tehnologijo mikromrež in tehnologijo sekvenciranja nove generacije omogočajo sočasno preiskovanje velikega števila molekularnih bolezenskih sprememb, tudi na nivoju celotnega genoma, epigenoma, transkriptoma, proteoma in drugih molekularno-bioloških nivojih (Cirulli in Goldstein, 2010; Grant in Hakonarson, 2008). Na področju multifaktorskih boleznih je tak pristop brez vnaprej postavljene (*a priori*) hipoteze o etiologiji preiskovane bolezni pomenil bistven napredek v primerjavi s študijami kandidatnih genov. Pristopi brez vnaprej postavljene hipoteze pomenijo prednost pri identifikaciji genov za multifaktorske bolezni, saj je trenutno znanje o funkciji številnih genov pomanjkljivo, kar otežuje izbor kandidatnih in identifikacijo vzročnih genov.

Pomembno mesto na področju identifikacije genov za multifaktorske bolezni z omskimi pristopi zasedajo asociacijske študije celotnega človeškega genoma. Tovrstne študije so po začetnem neuspehu pri identifikaciji dednih dejavnikov za multifaktorske bolezni, pomenile pomemben korak, saj so v primerjavi s študijami genetske vezave omogočile višjo ločljivost in večjo moč odkrivanja dednih dejavnikov z majhnim učinkom. S tovrstnimi študijami je bilo mogoče identificirati tudi spremembe v popolnoma novih genih, katerih biološka vloga je bila predhodno neznana.

2.1.4 Omejitve omskih pristopov

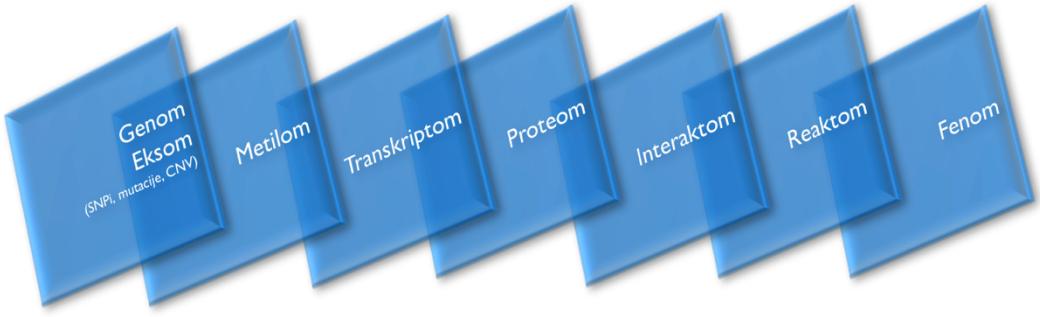
Kljub prednostim tovrstnih metod, so znane tudi njihove pomembne omejitve, predvsem s plati statistične obravnave na ta način pridobljenih visoko-dimenzionalnih podatkov (Gregersen in Brehrens, 2003; Simon in sod., 2003). Preiskovanje povezave velikega števila sprememb z boleznijo predstavlja statistično preverjanje velikega števila statističnih hipotez, ki ima za posledico povečano število lažno pozitivnih rezultatov. Po drugi strani pa se v teh študijah poveča možnost, da se resnično pozitivni rezultati izgubijo v statističnem šumu (Nadon in Shoemaker, 2002). Omenjeno predstavlja ključen problem pri odkrivanju diskretnih molekularnih sprememb pri multifaktorskih boleznih.

Novejše metode, ki temeljijo na sekvenciranju nove generacije so prinesle bistven napredek pri identifikaciji sprememb na raznovrstnih molekularnih nivojih. Novejše omske metode pomenijo bistven premik pri identifikaciji molekularnih sprememb in

biomarkerjev, kljub temu pa odpira bistveno povečan obseg identificiranih sprememb nov nivo analitičnih izzivov (Soneson in Delorenzi, 2013). V primeru merjenja transkriptomskih sprememb z metodami sekvenciranja RNA, se število preiskovanih sprememb razširi in obsega tudi morebitne redke in nove različice transkriptov (Wang in sod., 2009). Relativno majhno število bioloških replikatov pri eksperimentih RNA sekvenciranja ob širokem naborom pridobljenih rezultatov, pomeni povečano možnost identifikacije dodatno lažno pozitivnih rezultatov (Soneson in Delorenzi, 2013). Podobne izzive je moč srečati tudi na področju identifikacije dednih dejavnikov tveganja, kjer nove metode sekvenciranja razkrivajo prisotnost velikega števila novih različic in zelo redkih, katerih biološki pomen je neznan. Opredelitev pomena redkih različic pri nastanku multifaktorskih bolezni ostaja izziv, doseganje zadostne statistične moči za dokaz njihove prispevne vloge pa zahteva preiskovanje v zelo velikih populacijah bolnikov in zdravih preiskovancev (Zuk in sod., 2014).

2.2 INTEGRACIJA HETEROGENIH OMSKIH PODATKOV

Dosedanji pristopi globalnega odkrivanja bolezenskih sprememb so bili v veliki meri osredotočeni na preiskovanje sprememb ločeno po posameznih molekularnih nivojih. Dejanski patogenetski in homeostatski procesi pa so v organizmih na celičnem nivoju tesno prepleteni in niso omejeni le na posamičen omski nivo. Pri odkrivanju genov nam informacija o sočasnih spremembah na več molekularnih nivojih (Slika 3) lahko pomaga pri razločevanju dejanskih bioloških sprememb od tistih, ki nastanejo zaradi tehničnih in statističnih vzrokov.



Slika 3: Ilustracja koncepta integracije podatkov različnih bioloških nivojev z integrativnimi omskimi pristopi.
Figure 3: Illustration of the concept of overlapping multiple omic studies using the integrative approach.

V literaturi je bilo opisanih nekaj predhodnih poskusov sinteze heterogenih genomskih podatkov in informacijskih virov za namene prioritizacije kandidatnih genov pri multifaktorskih boleznih, ki so opisane v nadaljevanju.

Leta 2006 so Aerts in sod. predstavili pristop, pri katerem so uporabili učno množico genov, ki so bili že predhodno povezani s preiskovano boleznijo in na podlagi lastnosti te množice v različnih omskih virih (ekspresija, proteinske interakcije, baze genskih ontologij in drugih virov) izvršili globalno prioritizacijo ostalih genov glede na podobnost učni množici genov (Aerts in sod., 2006). Problem tega pristopa v kontekstu multifaktorskih bolezni je, da za prioritizacijo niso uporabili dejanskih eksperimentalnih podatkov za preiskovano bolezen, ampak so združevali referenčne podatke v raznih informacijskih virih. Integracija je potekala na nivoju genov, kar bi oteževalo implementacijo pristopa za uporabo določenih vrst omskih podatkov, pomembna pa je bila tudi izguba informacije pri redukciji podatkov na nivo genskih anotacij. Sorodni pristopi so bili kasneje uporabljeni tudi za razvoj drugih algoritmov za prioritizacijo kandidatnih genov (GeneProspector (Yu in sod., 2008), CANDID (Hutz in sod., 2008), ToppGene (Chen in sod., 2009), SUSPECTS (Adie in sod., 2005) in drugih).

Rasche in sod. so v letu 2008 predstavili pristop k sintezi omskih podatkov pri sladkorni bolezni tipa 2 (Rasche in sod., 2008). Kot podatkovne vire so uporabili empirične podatke študij merjenja globalnih sprememb v ekspresiji, študije pri mišjih modelih, podatke asociacijskih študij celotnega genoma in tudi podatke o povezavah v literaturi in bazah genskih ontologij. Pri integraciji so uporabljali statistično pomembnost, velikost spremembe, upoštevali pa so tudi mero entropije, ki je upoštevala prispevek posameznega molekularnega nivoja v končni oceni pomena gena pri končni prioritizaciji genov. Tudi v tem primeru so uporabili sintezo podatkov na osnovi prekrivajočih genov med študijami.

Podoben pristop so uporabili tudi Sun in sod. leta 2009, ki so uporabili eksperimentalne podatke iz asociacijskih študij celotnega genoma, globalnih razlik v ekspresiji genov v krvi in podatke iz literature za prioritizacijo kandidatnih genov pri shizofreniji (Sun in sod., 2009).

Skupno vsem omenjenim pristopom je, da so v vseh primerih pri integraciji uporabljali le prekrivanje množic identificiranih genov, kar omejuje možnost vključitve številnih tipov študij pri katerih spremembe niso vezane zgolj na gene. Nadalje smo pri našem preliminarnem pregledu omskih podatkov za primer Parkinsonove bolezni (PB) ugotovili, da bi zaradi nepopolnih anotacij pri takem pristopu že v začetku izgubili informacije o pomembnih spremembah genske ekspresije v 7.5 % deležu. V primeru genetskih različic, ki so pomembno povezane s PB je prišlo pri pretvorbe do izgube informacij v deležu

50.6%. Glede na to, da se v zadnjem času pripisuje vse večji pomen tudi spremembam v regulatornih področjih med geni, lahko taka izguba informacij privede do bistveno nižje senzitivnosti integrativnih pristopov za odkrivanje novih genetskih dejavnikov multifaktorskih bolezni (Peltonen in sod., 2006).

2.3 NOVI MODELI NASTANKA MULTIFAKTORSKIH BOLEZNI - POMEN REDKIH RAZLIČIC Z VISOKIM UČINKOM

Omejen uspeh prizadevanj pojasniti dednost multifaktorskih bolezni je postavil pod vprašanje osnovni model nastanka multifaktorskih bolezni, po katerem naj bi nastale zaradi sodelovanja številnih dejavnikov dednosti in okolja z majhnim učinkom. Številni primeri monogenskih oblik multifaktorskih bolezni kažejo na možnost, da lahko nastanejo kompleksne bolezni tudi kot posledica vpliva posameznih redkih dednih dejavnikov z visokim učinkom. V takem primeru vzročnih dednih sprememb ne bi mogli identificirati z asociacijskimi študijami, tudi če bi te vključevale zelo veliko populacije in imele veliko statistično moč (Bodmer in Bonilla, 2008).

Po novejših modelih, so lahko za nastanek multifaktorskih bolezni odgovorne številne visoko patogene različice v istem genu, visoko patogene različice v različnih genih in visoko patogene različice genov v različnih bioloških poteh. Tovrstne spremembe bi bilo praktično nemogoče ugotoviti tako s klasičnimi analizami dedne vezave, prav tako ne z asociacijskimi študijami, ki uporabljajo pri analizi predvsem pogoste različice (Bodmer in Bonilla, 2008).

2.3.1 Monogenske in oligogenske oblike multifaktorskih bolezni

Danes poznamo številne oblike multifaktorskih bolezni, ki se obnašajo v skladu z zakoni mendelske genetike. V pomembnem deležu teh primerov poznamo tudi molekularne mehanizme in vzročno gensko spremembo, ki vodi do nastanka takih bolezni. Nekaj pomembnejših primerov multifaktorskih bolezni, pri katerih poznamo monogensko dedovane oblike bolezni je predstavljenih v Preglednici 2 (Peltonen in sod., 2006). V nekaterih primerih je oblike z monogensko predispozicijo moč prepoznati po nastopu v zgodnejši dobi, občasno pa je monogensko obliko nemogoče klinično ločiti od ti. sporadične oblike. Ker v pomembnem deležu patogene različice za monogensko pogojeno obliko bolezni lahko nastanejo na novo (*de novo*) ali pa je predispozicija recesivna, gre lahko za monogensko etiologijo tudi pri primerih, ki se v populaciji pojavljajo na videz sporadično.

Preglednica 2: Pregled multifaktorskih bolezni z monogensko dedovanimi oblikami.
Table 2: Overview of multifactorial diseases with monogenic forms.

Multifaktorska bolezen z monogensko obliko	**Vzročni geni**	**Citat**
Alzheimerjeva bolezen	*APOE*, *PSEN1*, *PSEN2*, *APP*	(Tanzi, 2012)
Avtizem	Preko 10 genov z opisanimi rekurentnimi de novo različicami pri otrocih z avtizmom	(Ronemus in sod., 2014)
Epilepsija	Preko 100 genov z opisanimi de novo različicami pri idiopatski epilepsiji	(Carvill in sod., 2013)
Parkinsonova bolezen	*PINK1*, *LRRK2*, *SCNA*, *PARK2*, *ATP13A2*, *DJ1*	(Crosiers in sod., 2011)
Amiotrofična lateralna skleroza	*C9orf72*, *SOD1*, *TARDBP*, *FUS*	(Boylan, 2015)
Sladkorna bolezen	Geni za sladkorno bolezen tipa MODY (sladkorna bolezen z nastopom v otroški dobi)	(Hattersley in sod., 2009)

Poznavanje dednega vzroka monogenskih oblik dednih bolezni predstavlja pomembno možnost za odkrivanje mehanizmov, ki vodijo v nastanek bolezni, prav tako pa predstavljajo možnost za usmeritev nadaljnjih raziskav v tako ugotovljene biološke poti (Peltonen in sod., 2006). Monogenskih oblik za številne multifaktorske bolezni trenutno še ne poznamo.

2.3.2 Družinske oblike multifaktorskih bolezni kot orodje za identifikacijo monogenskih ali oligogenskih oblik multifaktorskih bolezni

Za številne multifaktorske bolezni trenutno še ne poznamo monogenskih oblik, ki bi lahko služile kot izhodišče za nadaljnje raziskave mehanizmov njihovega nastanka in nadaljnjem odkrivanju dednih dejavnikov zanje. Kljub temu pa pogosto ugotavljamo v posameznih družinah izrazito kopičenje primerov z istimi bolezenskimi težavami. Kot primer so Kahana in sod. pokazali, da družinska oblika MS zajema kar 20 % vseh primerov bolnikov z MS (Kahana, 2000).

Dosedanje študije pri primerih z družinskimi oblikami multifaktorskih bolezni so bile opravljene predvsem s klasičnimi orodji za analizo genetske vezave, ki pa ne predvidevajo alelne in genske heterogenosti, ki jo srečujemo že pri večini monogenskih bolezni. Ta lastnost klasičnih pristopov odkrivanja genov za multifaktorske bolezni predstavlja resno omejitev pri odkrivanju novih vzročnih dednih sprememb, posebej v primeru, da

multifaktorske bolezni nastanejo kot posledica izjemno redkih in visoko patogenih različic v številnih različnih genih.

2.3.3 Sekvenciranje nove generacije in odkrivanje redkih, visoko patogenih različic pri bolnikih z multifaktorskimi boleznimi

Nov vpogled v etiologijo multifaktorskih bolezni omogočajo tehnologije sekvenciranja nove generacije, ki omogočajo dostopno in hitro pridobitev informacije o zaporedju vseh kodirajočih delov genov (eksomsko sekvenciranje) ali pa celotnega genoma (genomsko sekvenciranje) (Cirulli in Goldstein, 2010). Omenjene metode omogočajo identifikacijo možno patogenih redkih različic z visokim patogenim učinkom tudi v posameznih majhnih družinah.

Kljub hitremu razvoju te tehnologije je napredek na področju interpretacije in ovrednotenja ugotovljenih rezultatov potekal počasneje. Sicer že obstajajo podatki in orodja, ki jih uporabljamo za izbor pomembnih genetskih sprememb: podatki o pogostosti različic v populaciji (Siva, 2008), orodja za napoved patogenosti genetskih različic (Ng in Henikoff, 2003), mere za evolucijsko ohranjenost mesta z ugotovljeno različico (Davydov in sod., 2010). Nobeno od teh orodij pa ne zagotavlja zanesljive opredelitve patogenosti genetskih različic, zato so za funkcijsko opredelitev pogosto potrebne dolgotrajne študije celičnih ali živalskih modelov. Zato je velik izziv izpopolniti načine za uspešnejšo izbiro variant in napovedovanje vzročne povezanosti odkritih novih genetskih različic z multifaktorskimi boleznimi.

Glede na pregled literature, uporaba rezultatov integracije *omskih* podatkov za interpretacijo podatkov eksomskega in genomskega sekvenciranja še ni bila opisana. V posameznih primerih so opisovali uporabo podatkov asociacijskih študij celotnega genoma za izbor ključnih genetskih različic, odkritih s pristopi sekvenciranja nove generacije (Sanders in sod., 2012; Zhu in sod., 2011), vendar pa integrativni pristop z uporabo raznovrstnih empiričnih podatkov omskih študij še ni bil uporabljen pri interpretaciji rezultatov eksomskega in genomskega sekvenciranja pri multifaktorskih boleznih.

2.4 IZBOR MODELNIH MULTIFAKTORSKIH BOLEZNI

Za razvoj pristopa za integracijo heterogenih omskih podatkov in njegovo evalvacijo smo izbrali kot modelno bolezen Parkinsonovo bolezen (PB). PB predstavlja primer multifaktorske bolezni s širokim naborom omskih študij in podatkov, ki pa ima tudi dobro znane in opredeljene monogenske oblike bolezni. Zaradi tega predstavlja PB dober model

za uspešnost metode integracije pri identifikaciji kandidatnih genov, ki vsebujejo redke in visoko penetrantne različice za monogenske ali oligogenske oblike bolezni.

Za prikaz uporabnosti in zmogljivosti integrativnega pristopa pri odkrivanju novih kandidatnih genov na podlagi integracije heterogenih omskih podatkov, smo izbrali model multiple skleroze (MS), za katero trenutno monogenskih oblik še ne poznamo, prav tako pa je s trenutno znanimi dednimi dejavniki pojasnjen le manjši delež dednosti bolezni.

2.4.1 Parkinsonova bolezen

Za namene razvoja in ocene delovanja razvitega algoritma integracije smo uporabili PB kot model multifaktorske bolezni, pri kateri poznamo številne monogenske oblike, pri katerih so opisane tudi družine s tipičnimi mendelskimi vzorci dedovanja. PB je druga najpogostejša nevrodegenerativna bolezen, ki jo zaznamuje progresivno zmanjšanje števila dopaminergičnih nevronov v bazalnem jedru CŽS *substantia nigra* in se klinično kaže s progresivnimi simptomi tremorja, rigidnosti, bradikinezije in položajne nestabilnosti. PB običajno nastopi v poznejših življenjskih obdobjih, njena incidenca raste s starostjo, njena prevalenca pa znaša okoli 1.8 % pri posameznikih s starostjo nad 65 let (Mayeux, 2003). Dejavniki, ki privedejo do nastanka PB, so razen v nekaterih družinskih primerih, neznani. Glede na trenutne hipoteze, obravnavamo PB kot multifaktorsko bolezen, ki nastane zaradi kompleksne interakcije več dednih in okoljskih dejavnikov.

Pomemben del družinske PB in v redkih primerih sporadičnih oblik PB predstavljajo monogenske oblike zaradi patogenih dednih različic z visoko patogenim učinkom v preko 12 genih (Wider in sod., 2010). Tudi pri primerih PB brez ugotovljenega monogenskega vzroka je njena dedna komponenta prepričljiva in ocene heritabilitete PB znašajo do 30%. V prid dedni etiologiji PB govorijo predvsem epidemiološki dokazi različne frekvence bolezni v različnih populacijah in dokazi povečanega zbiranja PB v družinah (Steece-Collier in sod., 2002). Kljub navedenemu, v večini primerov - tudi ko se pojavlja bolezen v družinah, etiologija in vzročni dedni dejavniki še vedno niso znani. Pomemben napor je bil v zadnjem času vložen v analizo sprememb pri bolnikih s PB na omskem nivoju, zaradi česar obstaja v javno dostopnih bazah in repozitorijih obsežen nabor podatkov, ki predstavljajo dragocen vir za integrativno analizo.

Zaradi obilice omskih podatkov, klinično dobro definiranega fenotipa in dobro raziskanih monogenskih oblik PB z jasno opredeljenimi vzročnimi redkimi in visoko penetrantnimi različicami, predstavlja PB dober model bolezni za razvoj in evalvacijo algoritma za integracijo heterogenih omskih podatkov.

2.4.2 Multipla skleroza

Multipla skleroza (MS) je kronična vnetna in nevrodegenerativna bolezen centralnega živčnega sistema (CŽS), ki najpogosteje prizadene mlado odraslo populacijo med 20. in 40. letom starosti. V času napredovanja bolezni postanejo številni bolniki izrazito gibalno ovirani, kar povzroči pomembno breme tako za prizadete, kot tudi za njihove družine in širšo družbo. Etiologija MS trenutno še ni poznana - kot pri PB velja mnenje, da gre za multifaktorsko bolezen, do katere pride zaradi sovplivanja številni dednih in okoljskih dejavnikov. Epidemiološke študije kažejo na pomen virusnih okužb z Epstein-Barr virusom, kajenja in izpostavljenosti sončni svetlobi pri nastanku MS (Ascherio, 2013).

Kljub temu, da so dokazi za dedno komponento MS prepričljivi (Westerlind in sod., 2014), pa razen skromnih uspehov, ne poznamo jasnih in penetrantnih genetskih dejavnikov, ne za družinsko kot tudi ne za sporadično obliko MS (Dyment in sod., 2006).

Glede na navedeno, je možno, da bomo pojasnili nastanek MS šele z novimi modeli dednosti in dovzetnosti za nastanek multifaktorskih bolezni. Ob odsotnosti prepričljivih monogenskih oblik MS in pomanjkljivem znanju o dednih dejavnikih za nastanek MS, smo jo izbrali kot primer bolezni, kjer bi z integrativno analizo omskih podatkov in podatkov eksomskega sekvenciranja lahko identificirali nove, pomembne etiološke dejavnike pri njenem nastanku.

3 METODE IN MATERIALI

3.1 PRIDOBIVANJE IN ANALIZA PODATKOV

Raziskovalno delo smo pričeli s sistematičnim pregledom objavljenih študij o globalnih molekularnih spremembah pri izbranih primerih multifaktorskih nevrodegenerativnih bolezni ("Parkinsonova bolezen" in "Multipla skleroza") v bazah literature Medline (Wakeford in Roberts, 1993), EMBASE (Dunikowski, 2005) in podatkovnih repozitorijih GEO, (Edgar in sod., 2002), ArrayExpress (Parkinson in sod., 2007), Stanford Microarray database (Sherlock in sod., 2001) ter SRA (Leinonen in sod., 2011). Za vsako od izbranih modelnih multifaktorskih bolezni smo pripravili bazo podatkov, v katero smo vključili podatke asociacijskih študij celotnega genoma (GWAS), študij genetske vezave, študij merjenja globalnih sprememb ekspresije genov v raznolikih tkivih, rezultate proteomskih študij, študij o spremembah v metilaciji, spremembah v nivojih izražanja mikroRNA (miRNA) molekul in podatke o tarčah spremenjenih mikroRNA molekul (tarče miRNA). Poleg omenjenega smo pridobili tudi širok nabor podatkov iz referenčnih baz, ki vključujejo informacije o proteinskih interakcijah, ontologije funkcije genov, fenotipskih ontologij in bazah podatkov s povezavami v literaturi. Baza podatkov z zbranimi podatki je bila strukturirana v obliki tekstovnih datotek, kjer smo za vsako študijo shranili podatke v posamezni enostavni tekstovni datoteki s tabulatorsko ločenimi vnosi. V vsaki datoteki smo shranili podatek o imenu študije, tipu študije in tipu merjenega biološkega signala. Nadalje smo vključili podatke o rezultatih posamezne študije, vključno z dostopno številko spremembe (angl. *Accession number*) in podatek o statistični pomembnosti sprememb oziroma velikosti spremembe. V primerih, ko podatek o dostopni številki ni bi na voljo, smo uporabili genomske koordinate vključenih signalov (na primer pri vključevanju rezultatov o študijah genetske vezave). Baza je bila pripravljena v obliki, ki smo jo lahko uporabili za avtomatizirano in ponovljivo obdelavo v razvitem algoritmu, ki je predstavljen naknadno (v poglavju 3.3).

3.1.1 Programsko okolje za analize podatkov

Vsi zbrani podatki za omske spremembe so bili obdelani, shranjeni in vse analize opravljene v statističnem programskem jeziku R, različice 2.15.2, z orodji in knjižnicami paketa Bioconductor 2.11 (Gentleman in sod., 2004), razen kadar je v besedilu drugače navedeno. Računsko intenzivne postopke in razvoj spletnega orodja smo izvršili s pomočjo programskega jezika Python 2.7.3, natančneje z uporabo matematičnih funkcij v paketih SciPy in NumPy (Olivier in sod., 2002), spletno stran pa smo oblikovali s pomočjo programskega paketa za Python - CherryPy.

3.1.2 Strategija za zbiranje podatkov

Za namene integracije smo vključili le podatke študij, ki so neusmerjeno in celostno preiskovale biološke spremembe na različnih omskih nivojih. Študije, ki so bile usmerjene na omejen nabor genov ali genomskih področij, smo izključili iz nadaljnjih analiz. S tem smo se izognili pristranosti pri razvrščanju regij, ki so bile pogosteje preiskovane.

Za vsako vključeno študijo smo preučili objavljeno publikacijo ter morebitno dodatno gradivo ali surove izvorne podatke omskih analiz. Iz rezultatov študij smo pridobili podatke o statistični spremenjenosti (P vrednosti) ugotovljenih sprememb. V nabor zbranih signalov smo vključili le spremembe, ki so glede na uporabljene statistične teste pokazale nominalno P vrednost pod 0,05. Pri zbiranju podatkov smo uporabili P vrednosti pred korekcijo za večkratno testiranje.

V vključenih bioloških nivojih smo pri razvrščanju kot mero za velikost signala uporabili $-\log_{10}$P vrednosti, razen v primeru fenotipskih povezav in proteomskih študij, kjer P vrednosti posameznih sprememb niso bile dostopne. V primerih, ko P vrednosti niso bile na voljo in je bil na voljo le seznam pomembno spremenjenih genov, smo to množico sprememb vključili v nabor podatkov, vendar smo vsem pripisali arbitrarno vrednost 1. S tem smo zabeležili prisotnost spremembe, vendar pa so bile spremembe pri razvrščanju obravnavane kot enakovredne. Tako pridobljene vrednosti smo pri vseh kasnejših analitičnih korakih obravnavali enakovredno $-\log_{10}$P vrednostim.

3.1.3 Pridobivanje in priprava podatkov omskih študij za Parkinsonovo bolezen

Iskanje omskih študij pri PB smo pričeli z preiskovanjem baze Medline (Wakeford in Roberts, 1993) z iskalnim nizom:

(“Parkinson disease”[ti] OR "Parkinson's disease"[ti]) AND (transcriptom OR proteom* OR “genome-wide” OR “linkage scan” OR microarray OR profiling)*.

V nadaljnjem koraku smo opravili iskanje opravljenih omskih študij, predvsem s pristopom analize z mikromrežami in v iskanje vključili bazo GEO (Edgar in sod., 2002), Array Express (Rustici in sod., 2013), prav tako pa smo preiskali tudi Stanfordsko bazo podatkov o opravljenih eksperimentih z mikromrežami (Sherlock in sod., 2001). Za pridobivanje podatkov o rezultatih asociacijskih študij celotnega genoma smo v iskanje vključili tudi bazo dbGAP (Mailman in sod., 2007), preiskali pa smo tudi baze s podatki študij sekvenciranja nove generacije - Sequence read archive, SRA (Leinonen in sod., 2011).

Za primer PB smo pridobili in vključili podatke s šestih različnih bioloških nivojev in jih vključili v nadaljnjo integrativno analizo:

I. Asociacijske študije celotnega genoma pri bolnikih s PB,
II. Študije genetske vezave pri bolnikih s PB,
III. Transkriptomske študije možganskega tkiva pri bolnikih s PB,
IV. Transkriptomske študije periferne krvi pri bolnikih s PB,
V. Proteomske študije sprememb nivojev beljakovin v centralnem živčevju bolnikov s PB,
VI. Nabor fenotipsko kompatibilnih genov, povezanih s kliničnimi znaki pri PB.

Rezultati v raznolikih omskih študijah so bili podani v raznovrstnih anotacijah - asociacije polimorfizmov posameznega nukleotida, spremembe na nivoju sond na ekspresijskih mikromrežah, sprememb v nivojih transkriptov, spremembe na nivoju genov. Zaradi te raznolikosti v nadaljnjih delih za poimenovanje spremembe na kateremkoli omskem nivoju uporabljamo poenoteno poimenovanje: *signal*.

3.1.3.1 Asociacijske študije celotnega genoma

Podatke iz asociacijskih študij celotnega genoma smo pridobili iz prosto dostopnega vira baze asociacijskih študij celotnega genoma - *Open Access Database of Genome-wide Association Results project* (Johnson in O'Donnell, 2009), kjer smo pridobili celoten nabor polimorfizmov enega nukleotida (SNP), ki so bili statistično pomembno povezani s prisotnostjo PB. Za statistično pomembno povezane smo smatrali tiste polimorfizme, za katere so znašale nominalne P-vrednosti pod 0,05 (pred popravkom za večkratno testiranje statističnih hipotez). Vključili smo podatke dveh študij: prvo so opravili Marganore in sod. (Maraganore in sod., 2005) s 250.000 SNP-i na platformi Perlegen 250K and drugo, ki so jo opravili Fung in sod. na platformi Illumina Infinium 100K, z analiziranimi 100.000 polimorfizmi (Fung in sod., 2006). V integracijo smo vključili skupno 1604 SNP-ov z nominalnimi p-vrednostmi pod 0.05.

3.1.3.2 Podatki o genetski vezavi pri Parkinsonovi bolezni

Podatke o genetski vezavi PB s specifičnimi področji v človeškem genomu smo pridobili iz študije Foltynie in sod., ki so opravili študije vezavnega neravnovesja na nivoju celotnega genoma (Foltynie in sod., 2005). Študijo so opravili z 5.546 mikrosatelitnimi markerji, razporejenimi po celotnem genomu v skupni populaciji 374 bolnikov s PB v primerjavi z dvema ločenima populacijama 219 in 1490 zdravih preiskovancev brez znakov PB. Skupno so ugotovili 214 mikrosatelitnih markerjev, ki so bili v statistično pomembnem vezavnem neravnovesju s PB.

Na podlagi povprečne gostote v študiji uporabljenih markerjev smo izračunali resolucijo preiskave in ocenili velikost regij, ki smo jih smatrali kot povezane s PB. Omenjene regije vezavnega neravnovesja smo skupaj s pripadajočimi p-vrednostmi vključili v integracijo podatkov.

3.1.3.3 Podatki o transkriptomskih spremembah pri Parkinsonovi bolezni (spremembe v krvi in centralnem živčevju bolnikov)

Surove podatke o transkriptomskih spremembah v centralnem živčevju pri PB smo pridobili iz repozitorija GEO s pomočjo paketa *GEOquery* za R (Sean in Meltzer, 2007). Transkriptomske spremembe v centralnem živčevju in v krvi smo obravnavali kot dva ločena nabora podatkov in tako predvideli možne razlike v transkriptomskem profilu v teh dveh tkivih.

Za spremembe v centralnem živčevju smo pridobili podatke o treh podatkovnih setih (z GEO dostopnimi številkami GSE8397, GSE7621 and GSE7307), prav tako pa smo pridobili tudi en podatkovni set s transkriptomskimi spremembami v periferni krvi pri bolnikih s PB, z dostopno številko GSE6613. Za vse surove podatkovne sete smo s paketom *arrayQualityMetrics* opravili kontrolo kakovosti eksperimenta z mikromrežami, čemur je za vsak podatkovni set sledila ločena normalizacija s paketom *affyPLM* in filtracija signalov primerne kvalitete s paketom *genefilter* (v odvisnosti od platform uporabljenih mikromrež - Agilent ali Affymetrix), kadar je bilo to potrebno.

Za izboljšano reprezentiranost signalov v primerih ko je bilo za isti biološki nivo na voljo več podatkovnih setov (na primer, v primeru transkriptomskih študij sprememb v centralnem živčevju pri bolnikih s PB), smo napravili meta-analizo transkriptomskih podatkov, preden smo rezultate vključili v integracijo. V meta-analizo transkriptomskih sprememb v CŽS pri bolnikih s PB, smo vključili samo podatkovne sete pridobljene na platformi mikročipov Affymetrix U133 (ki vključuje nabore sond U133A, U133B in U133plus2). Na isto platformo smo se omejili z namenom maksimalnega izkoriščenja obstoječih podatkov in poenostavitve postopka meta-analize.

Meta-analizo transkriptomskih študij smo opravili s programskim paketom RankProd v okolju R, kjer smo z uporabo funkcije *RPadvance* lahko združeno analizirali in identificirali transkriptomske spremembe, ki so bile prisotne v treh podatkovnih setih, pridobljenih z analizo vzorcev CŽS pri bolnikih s PB. Analiza implementirana v paketu RankProd temelji na ne-parametrični statistiki in omogoča detekcijo genov oziroma transkriptov, ki so visoko na seznamu najbolj različno izraženih genov v različnih transkriptomskih študijah. Tako lahko s tem pristopom sočasno analiziramo študije,

opravljene v različnih eksperimentih, laboratorijih in na različnih tipih mikromrež (Hong in sod., 2006).

Da bi uporabili čimbolj enotne analize, smo s paketom *RankProd* ponovno analizirali tudi edini transkriptomski podatkovni set pridobljen z analizo celotne periferne krvi pri bolnikih s PB. Za ta namen smo uporabili funkcijo paketa RP z vključeno opcijo za analizo podatkov iz posameznega vira.

Za oba transkriptomska nivoja, smo uporabili napovedi stopnje lažno pozitivnih rezultatov (PFP), ki je mera za nenaključno visoko uvrščanje sprememb v vključenih študijah in predstavlja osrednji rezultat analize s programskim paketom RankProd.

3.1.3.4 Podatki o proteomskih spremembah pri Parkinsonovi bolezni

Za nivo proteomskih sprememb pri PB smo vključili podatke 7 študij, ki so preiskovale spremembe v proteomu CŽS pri bolnikih (Abdi in sod., 2006; Basso in sod., 2004; Choi in sod., 2004; Jin in sod., 2006; Sinha in sod., 2007; Sinha in sod., 2009; Werner in sod., 2008). Iz vsake študije smo vključili podatke o genih, ki kodirajo statistično pomembno spremenjenih beljakovin v vzorcih CŽS pri bolnikih, za katere smo ugotovili spremembe v nivojih s p-vrednostmi pod ali enako 0,05. Skupno smo vključili v nabor proteomskih sprememb podatke o 199 genih, ki kodirajo pomembno spremenjene beljakovine pri PB. V primeru, da podatki o signifikanci sprememb niso bili dostopni oziroma jih študije niso poročale, smo spremenjenim beljakovinam pripisali vrednost 1. S tem smo zabeležili prisotnost spremembe v količini te beljakovine, ne pa tudi izrazitosti spremembe.

3.1.3.5 Geni, fenotipsko povezani s kliničnimi simptomi in znaki pri Parkinsonovi bolezni

Dodatni nivo smo v integracijo vnesli z vključitvijo dodatnega sloja informacij o genih, ki so povezani z bolezenskimi znaki pri PB. Za ta namen smo vključili podatke o povezavi med fenotipi človeka in geni, ki so zbrani v bazi Ontologija človeških fenotipov (HPO). Vir podatkov temelji na povezavah med geni in kliničnimi simptomi pri monogenskih boleznih zaradi okvar v teh genih (Robinson in sod., 2008).

Za ta nivo podatkov, smo pridobili začetni nabor genov, ki so bili v predhodnih študijah kot monogenski vzrok že bili povezani z mendelsko dedovanimi oblikami PB - OMIM:168600 (OMIM, 2016). Zatem smo zbrali množico fenotipov in dostopne številke fenotipov, ki so s temi geni povezani glede na podatke baze HPO. Na ta način smo pridobili množico fenotipov v standardizirani nomenklaturi, ki so prisotni pri znanih monogenskih oblikah PB. V zadnjem koraku smo na podlagi povezav v bazi HPO

identificirali še vse druge gene, ki so s povezani z naborom kliničnih znakov pri monogenskih oblikah PB. S tem postopkom smo pridobili nabor genov, ki so sicer povezani s posameznimi simptomi in znaki pri PB, vendar pa s PB še niso bili neposredno povezani. Tem genom smo za vsak ujemajoč fenotip prišteli vrednost 1 enoto, s čemer smo dosegli, da so imeli bolj fenotipsko kompatibilni geni višjo vrednost. Tako dobljene vrednosti smo vključili v končni postopek integracije heterogenih omskih podatkov.

3.1.4 Pridobivanje in priprava podatkov omskih študij za multiplo sklerozo

Podatke omskih študij za MS smo pridobili s podobnih postopkom kot študije za PB, s tem da smo v primeru MS nabor zajetih študij razširili tudi na epigenomske študije (identificirali smo tudi podatke o spremembah v izražanju miRNA in spremembe v metilaciji genomskih regij pri MS).

Za re-analizo pridobljenih študij smo uporabili iste pristope kot pri obdelavi podatkov omskih študij pri PB. Za odkrivanje izvornih študij smo podobno kot pri PB opravili izčrpno iskanje objavljenih študij v literaturi in repozitorijih podatkov za omske študije.
Za iskanje po bazi Medline smo uporabili iskanje z iskalnim nizom:

[*"Multiple sclerosis"[ti] AND (transcriptom* OR proteom* OR "genome-wide" OR "linkage scan" OR microarray OR profiling)*].

V integracijo smo vključili podatke z naslednjih bioloških nivojev:

I. Asociacijske študije celotnega genoma pri bolnikih z MS,
II. Študije genetske vezave pri bolnikih z MS,
III. Transkriptomske študije možganskega tkiva pri bolnikih z MS,
IV. Transkriptomske študije periferne krvi pri bolnikih z MS,
V. Transkriptomske študije transkriptoma T celic pri bolnikih z MS,
VI. Proteomske študije sprememb nivojev beljakovin v krvi CŽS bolnikov z MS,
VII. Proteomske študije sprememb nivojev beljakovin v cerebrospinalni tekočini bolnikov z MS,
VIII. Proteomske študije sprememb nivojev beljakovin v tkivu CŽS bolnikov s MS,
IX. Spremembe v metilaciji genomskih regij pri bolnikih z MS,
X. Študije različnega izražanja miRNA pri bolnikih z MS,
XI. Geni, ki vsebujejo napovedane regulatorne tarče miRNA s spremenjenim izražanjem pri MS,

XII. Ortologi genov, pri katerih so pokazali spremenjeno izražanje na živalskih modelih z indukcijo eksperimentalnega automimunega encefalitisa.

V končnem koraku smo za MS vključili podatke z 12 različnih bioloških nivojev, vključitvenim kriterijem pa je v končnem koraku ustrezalo 52 študij. Za analizo smo uporabljali isto metodologijo kot je opisana za primer PB, razen v primerih bioloških nivojev, ki jih podrobneje opisujemo spodaj.

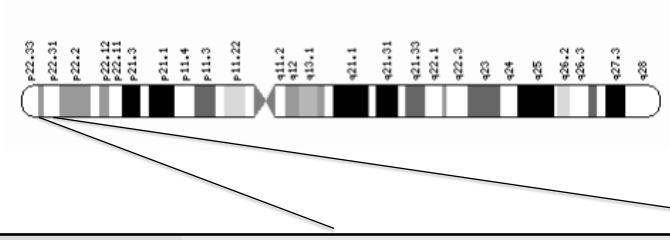
3.1.4.1 Podatki o spremembah miRNA in njihovih tarčah

Položaj genov miRNA smo pridobili s pomočjo anotacij o položaju miRNA v človeškem genomu v bazi genomskega brskalnika UCSC, z orodjem UCSC Table browser (Karolchik in sod., 2004). Tarčne gene, podvržene regulaciji miRNA molekulam s spremenjenim izražanjem pri MS, pa smo pridobili z orodjem BioMart z izbiro podatkovnega seta *Homo sapiens miRNA Target Regions* v genomskem sestavu b37 (Smedley in sod., 2009).

3.2 RAZVOJ IN UPORABA IZVIRNEGA PRISTOPA ZA INTEGRACIJO HETEROGENIH GENOMSKIH PODATKOV

3.2.1 Pozicijska integracija

Integracijo heterogenih omskih podatkov smo pričeli z umeščanjem statistično pomembnih signalov iz omskih študij na koordinatni sistem humanega genoma. Koordinatni sistem in dolžine posameznih kromosomov smo definirali na podlagi genomskega sestava UCSC, v različici hg19, ki je bila izdana februarja 2009 (ekvivalenta različici humanega genoma NCBI v37). Koncept pozicijske integracije shematsko prikazuje Slika 4.



		Regija							
Tip 'omske' študije	Tip signala	0-10kb	10-20kb	20-30kb	30-40kb	40-50kb	50-60kb	60-70kb	70-80kb
Transkriptomska (kri)	Sonda	0	0	3	1	0	0	0	0
Transkriptomska (možgani)	Sonda	1	0	2	1	0	0	0	1
Protemska	Gen	1	0	0	0	0	0	0	0
GWAS	SNP	1	0	3	0	0	0	0	0
Genetska vezava	Regija	0	0	1	1	0	0	1	1
Fenomski podatki (geni)	Gen	0	0	1	0	0	0	0	0
Skupni rezultat za regijo		3	0	10	3	0	0	1	2

Zbiranje signalov v regiji

Slika 4: Shematski prikaz pristopa pozicijske ali položajne integracije podatkov.
Na sliki prikazujemo podatke za 8 regij velikosti 10 kb. Vsaki regiji pripišemo vrednosti ($-\log_{10}$p vrednosti ali drugo mero, ki odraža izrazitost spremembe pri razvrščanju rezultatov, kot je opisano v poglavju 3.1.2). Vrednosti signalov ustrezajo meri za izrazitost spremembe pri preiskovani bolezni. Cilj integracije je identificirati tiste regije, kjer se zbirajo signali z več različnih bioloških nivojev, kot je na shemi prikazano za regijo 20-30 kb na kromosomu. V označeni regiji se pojavlja zbiranje podatkov tako na nivoju transkriptoma, asociacijskih študij, genetske vezave, v regiji pa so tudi geni s fenotipsko kompatibilnostjo preiskovani bolezni.
Figure 4: Schematic representation of the positional integrative approach of heterogeneous omic data.
Figure displays source data mapped to 8 regions 10 kb in size. Scores for features within each region are attributed to the region and the values summarized. The signals for integration represented $\log_{10}$p when available or another measure of scale as detailed in chapter 3.1.2. The aim of the integration process is to identify regions, where signals from multiple biological layes accumulate (as shown on the figure for 20-30 kb in the figure). Here, it is possible to detect presence of concurrent alterations on transcriptome, GWAS, linkage regions and presence of genes with phenotypic relation to the investigated disease.

Položaje genov in SNPov smo pretvorili v genomske položaje bodisi z uporabo orodja BioMart (Haider in sod., 2009), bodisi z uporabo anotacijskih paketov okolja Bioconductor v različici 2.8. Položaje sond, transkriptov, miRNA pa smo pretvorili v genomske položaje s prenosom celotnih anotacijskih setov in podatkov o položajih sprememb v repozitoriju genomskega brskalnika UCSC, z orodjem UCSC Table browser (Karolchik in sod., 2004). V primeru, da so bili podatki na voljo le v starejših genomskih sestavih (na primer pri podatkih o regijah v vezavnem neravnovesju z boleznijo) smo za pretvorbo položajev iz starejših v novejše koordinatne sisteme genoma uporabili orodje liftOver s spletišča genomskega brskalnika USCS (Kent in sod., 2002).

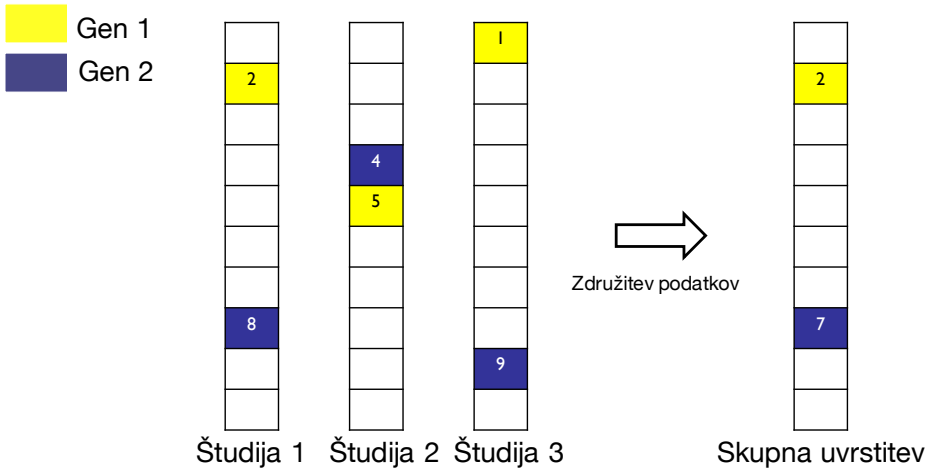
Referenčni genomski sestav (hg19, GRCh37) smo enakomerno razdelili na odseke definirane dolžine in regije izbrali tako da se je vsaka regija v 50 % prekrivala s predhodno regijo. Prekrivanje s sosednjimi regijami smo uvedli z namenom, da bi zajeli tudi morebitno zbiranje signalov na meji med dvema regijama, kar bi brez analize prekrivajočih regij lahko zgrešili. Pri analizi podatkov smo uporabljali genomske intervale velikosti 10 kb, 50 kb, 100 kb in 500 kb, kar je glede na velikost sestava hg19 (3.137 Mb) pomenilo, da smo preiskovali zbiranje signalov v 627.432, 125.486, 62.744 in 12.550 prekrivajočih regijah. Poleg položaja signalov v genomu smo zbrali tudi informacije o statistični pomembnosti ali velikost spremembe pri preiskovani bolezni. Podatek o izrazitosti ali pomembnosti spremembe nam je v nadaljnjih korakih omogočal prioritizacijo sprememb na posameznem biološkem nivoju.

Za namene umeščanja signalov v definirane genomske intervale smo pripravili skripto v jeziku R, s katero smo vsaki regiji oziroma intervalu pripisali vrednosti signalov iz različnih bioloških nivojev. V kolikor so bili vhodni signali podani v obliki P vrednosti, smo pripisali regiji vrednost $-\log_{10}p$, v kolikor je bil na voljo drug podatek smo uporabili le-tega, v kolikor smo imeli le dostop do seznama spremenjenih genov, smo privzeli vrednost 1.0 za vse vključene signale. V kolikor je bilo v isti regiji prisotnih več signalov z istega biološkega nivoja, smo vrednosti posameznih signalov v regiji sešteli. V kolikor genomska regija ni vsebovala nobenega pomembnega signala, smo ji pripisali vrednost 0.

Postopek integracije je potekal v dveh korakih. Najprej smo združili podatke iz več študij na istem molekularnem nivoju - združevanje v okviru posameznega biološkega nivoja smo opravili z različnimi metodami, v odvisnosti od biološkega tipa (na primer pri transkriptomskih podatkih z meta-analizo transkriptomskih podatkov). V drugem koraku pa smo podatke različnih nivojev združili v skupni, končni rezultat integracije. Vhodni podatek za ta korak je bila matrika s številom stolpcev, ki je ustrezalo številu vključenih bioloških nivojev in s številom vrstic, ki je ustrezalo številu intervalov na katerega smo

genom razdelili. V matriki so bile vsebovane vrednosti, pridobljene z uvrščanjem signalov omskih študij v ustrezne regije. Za vsak tip omske študije smo razvrstili regije glede na vsoto vrednost signalov v področju, od 1 za najvišje uvrščeno regijo do vrednosti N, ki predstavlja najnižje uvrščene regije. V kolikor sta imeli dve ali več regij enak rezultat smo izračunali povprečje uvrstitev.

Kot neparametričen pristop za združevanje podatkov heterogenih virov in heterogenih distribucij smo uporabili pristop uvrstitvene statistike. Razlog za uporabo pristopa uvrstitvene statistike je shematsko predstavljen na Sliki 5.



Slika 5: Pristop k integraciji podatkov z upoštevanjem razvrstitve regij v posameznem omskem nivoju.
Pristop z uvrstitveno statistiko omogoča odkrivanje genov ali sprememb, ki se konsistentno visoko uvrščajo na raznolikih bioloških nivojih. Tako je v prikazanem primeru Gen 1 v skupnem merilu bistveno višje uvrščen kot Gen 2, kljub temu da je v študiji 2 uvrščen nižje kot Gen 1. S tovrstnim pristopom lahko identificiramo konsistentno spremenjene signale, kljub visoki stopnji šuma in razlik v distribuciji vrednosti pri različnih virih podatkov.
Figure 5: The approach for integration using prioritization of regions for each included omic layer.
Ranking statistics allows for identification of genes that are consistently highly ranked across different biological layers. In the case shown, Gene 1 is ranked markedly higher on the list in comparison to Gene 2, despite attaining the lower score in Study 2. The approach allows for identification of consistently altered signals on multiple, different biological types and is robust in presence of technical noise and disparate distributions of included datasets.

Za integracijo smo uporabili prirejeno implementacijo produkta uvrstitev, ki so ga opisali Breitling in sod. v letu 2004 (Breitling in sod., 2004). Za vsako regijo smo tako izračunali vrednosti produkta uvrstitev (RP, *angl. rank product*) kot je opisano v enačbi 1:

$$RP_R = \prod_{i=1}^{k} \left(\frac{r_{i,R}}{n_i}\right) \quad ...(1)$$

V enačbi predstavljajo oznake naslednje: $r_{i,R}$ predstavlja uvrstitev regije R v biološkem nivoju i, n_i pa predstavlja število regij za biološki nivo i. k predstavlja število vseh bioloških nivojev, vključenih v integracijo.

Za oceno statistične pomembnosti zbiranja signalov iz različnih tipov omskih študij smo opravili simulacije, kjer smo vrednosti za regije naključno permutirali glede na položaj v genomu. Uvrstitve smo v vsakem ciklu permutacij primerjali z opaženimi uvrstitvami in izračunali permutirani produkt uvrstitev RP_{perm} in ga priperjali z opaženim produktom uvrstitev RP_{obs}. Na koncu smo izračunali pričakovani produkt uvrstitev RP_{perm} kot kaže enačba 2.

$$RP_{exp} \approx \frac{rank(RP_{perm})}{N_{perm}} \quad ...(2)$$

V končnem koraku smo izračunali mero, ki smo jo smatrali za ekvivalentno meri za stopnjo lažno pozitivnih najdb (q_R ali PFP), kot je bilo predlagano v objavljeni literaturi za algoritem izračuna neparametrične statistike na podlagi uvrstitev (Breitling in sod., 2004).

$$q_R = \frac{RP_{exp}}{rank(R)} \quad ...(3)$$

Ker je pomembnost posameznega nivoja sprememb pri sintezi podatkov lahko različna, smo implementirali tudi možnost obteževanja posameznih molekularnih nivojev glede na oceno njihove pomembnosti. Ker so podatki posameznih študij in posameznih bioloških nivojev glede na uporabljene aplikacije in kvaliteto vhodnih podatkov različni, smo implementirali tudi obteževanje vhodnih bioloških nivojev pri integraciji podatkov, kot prikazuje enačba 4.

$$RP_R = \left(\prod_{i=1}^{k} \left(\frac{r_{i,R}}{n_i} \right)^{w_i} \right)^{1/\sum_{i=1}^{k} w_i} \quad ...(4)$$

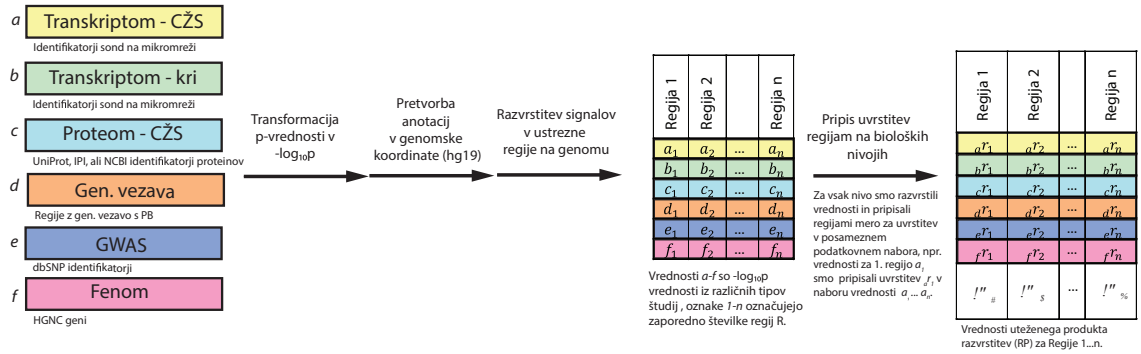
V enačbi znaka w_i predstavlja obtežitev posameznega vhodnega biološkega nivoja.

Ker je v številnih situacijah pomen posameznega biološkega nivoja pri končni integraciji neznan vnaprej, smo na primeru PB implementirali tudi način za kalibracijo optimalne konfiguracije obtežitev za vključene nivoje. To smo dosegli tako, da smo pridobili nabor genov z znano in prepričljivo povezavo s PB:

I. nabor genov v KEGG poti "Parkinsonova bolezen" (z dostopno številko *hsa05012*),
II. nabor genov, povezanih z monogenskimi boleznimi v bazi OMIM (OMIM, 2016),
III. nabor genov z znanimi asociacijami s PB v bazi genov za PB - PDgene *(Lill in sod., 2012)*.

Nato smo pripravili vse variacije konfiguracij uteži (vsakemu nivoju smo pripisali uteži v vrednosti 1-3) in preskusili 729 različnih konfiguracij uteži: $(p)V_3^6 = 3^6 = 729$. Za vsako konfiguracijo uteži smo izračunali rezultate integracije in preverili s katero konfiguracijo uteži dosežemo najvišje vrednosti za gene iz treh izbranih naborov genov z znanimi povezavami s PB.

Tako optimizirani nabor uteži smo uporabili tudi pri končni prioritizaciji regij za PB. Povzetek celotnega postopka prikazujemo na Sliki 6.



Slika 6: Podrobna shema postopka pozicijske integracije heterogenih omskih podatkov pri PB.
Figure 6: Detailed scheme of positional integrative approach for synthesis of heterogeneous omic data in Parkinson disease.

3.2.1.1 Korekcija pristranosti zaradi neenakomerne razporeditve genov

Ker je gostota genov v različnih regijah genoma neenakomerna, bi lahko zaradi razlik v gostoti prišlo do lažno povečanega ali lažno majhnega zbiranja signalov v določenih genomskih regijah. Za upoštevanje te možnosti smo določili gostoto genov v posameznih regijah genoma in permutacije zamejili na nabor regij s podobno gostoto genov. Gostoto genov smo določili z uporabo podatkov o položajih genov v repozitoriju Ensembl različice 54.

3.2.1.2 Dvostopenjska integracija podatkov v primerih, ko je za isti biološki nivo na voljo več raznolikih študij

Ker je v večini primerov pri integraciji podatkov na voljo večje število študij opravljenih na posameznem nivoju (na primer več študij tipa GWAS) in ker v številnih primerih nimamo dostopa do surovih podatkov iz omenjenih študij, meta-analize pogosto niso mogoče ali pa ni primernih orodij za združevanje podatkov. Za take primere smo vključili vmesno stopnjo integracije, kjer uporabimo uvrstitveno statistiko za izračun razporeditve regij - najprej za namen združevanja študij na posameznem biološkem nivoju, v drugi fazi pa ponovno integracijo za združevanje študij na različnih bioloških nivojih.

3.3 RAZVOJ SPLETNEGA ORODJA ZA UPORABO PRISTOPA POZICIJSKE INTEGRACIJE

Da bi omogočili dostop za uporabo pristopa pozicijske integracije tudi drugim skupinam in za analizo drugih primerov bolezni človeka, smo razvili spletno orodje, ki omogoča opravljanje lastne integrativne analize. Orodje smo zaradi lažje spletne implementacije in boljše izrabe računalniških sredstev implementirali v jeziku Python. Za namene lažje uporabe smo implementirali tudi samodejno pretvarjanje širšega nabora anotacij v položaje na genomu, samodejno umeščanje signalov na genomske koordinate in izračunavanje pomembnosti zbiranja signalov na regijah s permutacijami. Za izgradnjo spletnega vmesnika smo uporabili Python paket CherryPy 2.3.0 (paket python-cherrypy, cherrypy.org), večino ostalih funkcij smo implementirali sami s pomočjo paketov NumPy 1.6.2 (paket python-numpy, numpy.org) in SciPy 0.10.1 (paket python-scipy, scipy.org). Razvoj in postavitev platforme smo opravili v okolju Ubuntu Linux 12.10 (ubuntu.com), vse uporabljene pakete pa smo pridobili v repozitoriju paketov distribucije Ubuntu (packages.ubuntu.com).

3.4 EVALVACIJA PRISTOPA ZA INTEGRACIJO NA MODELU PARKINSONOVE BOLEZNI

Evalvacijo uspešnosti integrativnega pristopa za identifikacijo tako znanih povezav s preiskovano boleznijo, kot tudi novih kandidatnih genov, smo preverili s podrobno analizo funkcijskih lastnosti genov in podatkov v literaturi.

V prvem koraku smo vse gene, ki jih je identificiral pristop integrativne analize pri PB, preverili za direktne povezave v literaturi. Preiskali smo bazo Medline z iskalnimi nizom: "*Parkinson disease* AND *Gene*", kjer je "*Gene*" predstavljal kandidatni gen, identificiran v najviše uvrščenih regijah pri integrativni analizi (Dunikowski, 2005).

Da bi zajeli tudi morebitne indirektne povezave med odkritimi geni in PB, smo uporabili orodja za odkrivanje indirektnih povezav med pojmi v literaturi - BITOLA (Hristovski in sod., 2005). V načinu "*closed discovery*" smo izbrali za koncept X termin "Parkinsonova bolezen" (CUI:C0030567) za koncept Z pa smo uporabili gene v najvišje uvrščenih regijah po integraciji. Rezultate, ki jih je identificiralo orodje BITOLA predstavlja koncepte, ki so hkrati povezani s konceptoma X in Z in tako predstavljajo pojme Y, preko katerih se PB in kandidatni geni v literaturi indirektno povezujejo.

Nadalje smo opredelili tudi funkcijski profil identificiranih genov v najviše uvrščenih regijah, in sicer z analizo obogatenosti glede na anotacije funkcijskih kategorij ontologije genov - Gene Ontology (Ashburner in sod., 2000), v bazi funkcijskih poti KEGG (Kanehisa, 2002) in z uporabo baze metabolnih poti Reaktom (Croft in sod., 2011). Orodje *Reactome Skypainter* smo uporabili za opredelitev obogatenosti genov v reaktomskih poteh (Croft in sod., 2011).

Za statistično analizo obogatenosti smo uporabili hipergeometrični statistični test, implementiran v orodju GOstats za R (Falcon in Gentleman, 2007). Pri obogatitvenih analizah smo za primerjavo obogatitev uporabili za ozadje celoten nabor genov v bazi Ensembl. Dobljene P vrednosti smo popravili za večkratno preverjanje hipotez po protokolu Benjamini-Hochberg.

3.5 EKSOMSKO SEKVENCIRANJE PRI BOLNIKIH Z MULTIPLO SKLEROZO

Uporabnost rezultatov integrativnega pristopa pri interpretaciji eksomskega sekvenciranja smo preverili na skupini družinskih primerov z multiplo sklerozo, sporadičnih primerov multiple skleroze in posameznikov brez klinične slike MS.

3.5.1 Izbor bolnikov

V preiskavo z eksomskim sekvenciranjem smo vključili 48 bolnikov z družinsko obliko MS, 40 bolnikov s sporadično obliko MS in 92 kontrolnih primerov. Bolnike z MS smo obravnavali kot družinske primere, kadar je bil v ožjem sorodstvu vsaj še en obolel sorodnik. V večini primerov so bili v vključenih družinah oboleli vsaj trije posamezniki.

3.5.2 Izolacija nukleinskih kislin

Za izolacijo nukleinskih kislin (DNA) je bila, kri odvzeta v 3 mL EDTA epruvetah epicah, izolirana s sistemom FUJIFILM QuickGene-610L sistemom in kitom za izolacijo DNA iz polne periferne krvi. Sistem izolacije proizvajalca temelji na obarjanju DNA na mikrofilmski membrani, ki specifično veže DNA. Izolacija je bila v celoti opravljeni po navodilih proizvajalca in je vključevala centrifugiranje, obdelavo vzorca s proteinazo K, obarjanje z 99 % etanolom in elucijski postopek s 500 μL bidestilirane vode na napravi za avtomatsko izolacijo. Vzorci DNA so bili do molekularnih analiz shranjeni na -20 °C. Koncentracijo DNA smo merili s spektrofotometrično metodo, z napravo NanoDrop 1000 (Thermo Scientific, Wilmington, ZDA).

3.5.3 Eksomsko sekvenciranje pri bolnikih z multiplo sklerozo in zdravih kontrolah

Preiskavo usmerjenega sekvenciranja izbrane tarče za sekvenciranje smo opravili z metodologijo sekvenciranja nove generacije na vzorcu DNA. Na 50 ng vzorca DNA smo opravili encimsko fragmentacijo in obogatitev tarč po protokolu Illumina Nextera Coding Exome (Illumina, San Diego, ZDA) za zajem kodirajočih predelov v človeškem genomu s skupno velikostjo zajetih regij 45 Mb. Tarča za sekvenciranja je skupno zajemala področje 212,405 kodirajočih eksonov vseh genov človeškega genoma (kar zajema 97.8% kodirajočih regij ENSEMBL genov).

Vzorcem različnih preiskovancev smo dodali specifične indekse na 5' in 3' konce fragmentov, tako da je bilo mogoče v eni knjižnici ekvimolarno združiti vzorce 12-ih preiskovancev. Za kvantifikacijo in analizo kvalitete pripravljenih knjižnic smo uporabili meritev z napravo elektroforetsko analizo fragmentov Agilent Bioanalyzer 2100 (s kitom

Agilent High Sensitivity) in z uporabo napravo za fluorimetrično določanje koncentracije Life Technologies Qubit 2.0. Velikost pripravljenih knjižnic so v vseh primerih znašale od 300-1000 baznih parov, molarne koncentracije pa v vseh primerih nad 5 mM.

3.5.4 Sekvenciranje nove generacije

Pred sekvenciranjem smo knjižnico fragmentov dentaurirali z 0.1 M NaOH, jo z reagentom HT1 (proizvajalec Illumina) razredčili na 6.5 pM. Za interno kontrolo kvalitete smo uporabili fragmente virusa PhiX v deležu 1 %. Sekvenciranje vzorca smo opravili na napravi Illumina HiSeq 2500 na pretočni celici z dvema povezanima kanaloma po protokolu obojesmernega sekvenciranja v 2x100 ciklih, po hitrem protokolu (HiSeq RAPID Run mode). Pri vsakem zagonu smo pridobili vsaj 60 gigabaz (Gb) podatkov s kvaliteto nad Q30. Demultipleksiranje je bilo opravljeno s programsko opremo Bcl2Fastq (Illumina, San Diego, ZDA). Sekvenciranje smo opravili na dveh sekvenatorjih HiSeq 2500 v laboratoriju inštituta Institute of Applied Genomics (Udine, Italija).

3.5.5 Bioinformatska analiza podatkov sekvenciranja nove generacije

Za analizo podatkov eksomskega sekvenciranja smo razvili lastno analitično pot, ki je vsebovala elemente analize kakovosti sekvenciranja NGS podatkov, primarne analize (naleganje branj, pridobljenih z NGS), sekundarne analize (odkrivanje genetskih različic v podatkih NGS) in terciarne analize (anotacija in filtriranje ugotovljenih različic).

Kontrolo kvalitete podatkov NGS smo opravili z orodjem fastQC 0.10.1. Po odstranitvi podvojenih zaporedij smo opravili naleganje branj na referenčni genom človeka različice hg19 (v UCSC sestavu) z algoritmom BWA 0.6.3 (Li in Durbin, 2010). Branja smo v naslednjem koraku natančneje prilegali okrog ugotovljenih in predhodno znanih kratkih insercijsko-delecijskih polimorfizmov z orodjem GATKv2.8 IndelRealigner, informacije o kakovosti zapisa posameznih branih odsekov so bile popravljene z orodjem GATKv2.8 BaseRecalibrator (McKenna in sod., 2010). Za določanje nukleotidnih različic na podatkih posameznega preiskovanca smo uporabili sistem UnifiedGenotyper GATKv2.8, kjer smo upoštevali le različice na področjih s pokritostjo nad 5x in zanesljivostjo označevanja različic nad 30.0. Za opredelitev učinka variant smo uporabili orodji ANNOVAR in snpEff in izračunane napovedi patogenosti v bazi dbNSFPv2 (Liu in sod., 2013). Vsa zaporedja genov in referenčni transkripti temeljijo na zapisih v bazi RefSeq.

Pred interpretacijo ugotovljenih različic smo izločili različice s pogostnostjo nad 1 % v populacijah 1000genomes ali ESP6500, različice v nekodirajočih regijah, sinonimne kodirajoče različice in različice zunaj klinične tarče.

Za vsak vzorec smo ocenili mediano pokritost ciljanih regij in izračunali delež tarčnih regij z vsaj 10x globino sekvenciranja, kar na teh področjih omogoča nad 90 % senzitivnost ugotavljanja mutacij v heterozigotnem stanju in nad 99 % senzitivnost ugotavljanja mutacij v homozigotni obliki (Meynert in sod. 2011). Za vse eksome smo zahtevali minimalno mediano pokritost 30x in vsaj 85 % delež tarčnih regij s pokritostjo nad 10x.

Za namene analize bremena patogenih mutacij smo enake pogoje detekcije različic zagotovili s skupno ponovno analizo podatkov eksomskega sekvenciranja pri vseh vključenih preiskovancih. Za ta namen smo uporabili orodje GATK HaplotypeCaller v načinu za hkratno genotipizacijo večih vzorcev in tako pridobili matriko genotipov za vse vzorce hkrati. Pri pripravi rezultatov genotipizacije z GATK2.8 smo uporabili način "*EMIT_ALL_CONFIDENT_SITES*", kar je omogočilo, da imamo za vsako variantno mesto v genomu podatek o tem ali je različica tam prisotna, ali je različica zanesljivo odsotna ali pa je pokritost na mestu različice preslaba, da bi lahko zanesljivo ocenili prisotnost različice. Tak pristop je ključen za vse primerjalne analize na podlagi podatkov NGS, saj bi lahko zaradi neenakomerne pokritosti prišlo do tehnične pristranosti pri določanju različic.

3.5.5.1 Analiza različic v podatkih eksomskega sekvenciranja

Ugotovljene različice smo zbirali in anotirali s sistemom, ki temelji na programski opremi vtools in ANNOVAR. Za napovedovanje posledic genskih različic smo uporabili modele genov v bazi RefSeq, anotacije odkritih različic pa smo temeljili na bazi dbSNP v138. Podatke o pogostosti različic v populaciji smo črpali iz baze 1000 eksomov slovenske populacije na Kliničnem inštitutu za medicinsko genetiko, poleg tega pa smo uporabili tudi bazo 60.000 eksomov v konzorciju ExAC, podatke projekta UK10K in GoNL. Napovedi podatke smo pridobili iz baze že izračunanih napovedi za eksomske različice v bazi dbNSFP v2, poleg tega pa smo uporabili tudi napovedi SNPeff orodje za napovedovanje učinka na beljakovinsko zaporedje. Izračunane mere za ohranjenost zaporedja DNA smo pridobili iz baze GERP++. Prav tako smo uporabili tudi bazo ClinVar, kot vira podatkov o znanih kliničnih posledicah odkritih različic.

3.5.5.2 Filtriranje različic, pridobljenih z eksomskim sekvenciranjem

Vse različice, ki smo jih identificirali s pristopom eksomskega sekvenciranja smo razdelili v tri poglavitne razrede, glede na dokaze o patogenosti. V prvo kategorijo smo uvrstili visoko patogene različice (različice s prezgodnjo vstavitvijo stop kodona, s premikom bralnega okvirja s spremembo ključnega genskega zapisa za izrezovanje intronov). V drugo kategorijo smo uvrstili srednje patogene različice (različice z zamenjavo aminokislinskega zaporedja in insercijsko-delecijski polimorfizmi brez premika bralnega

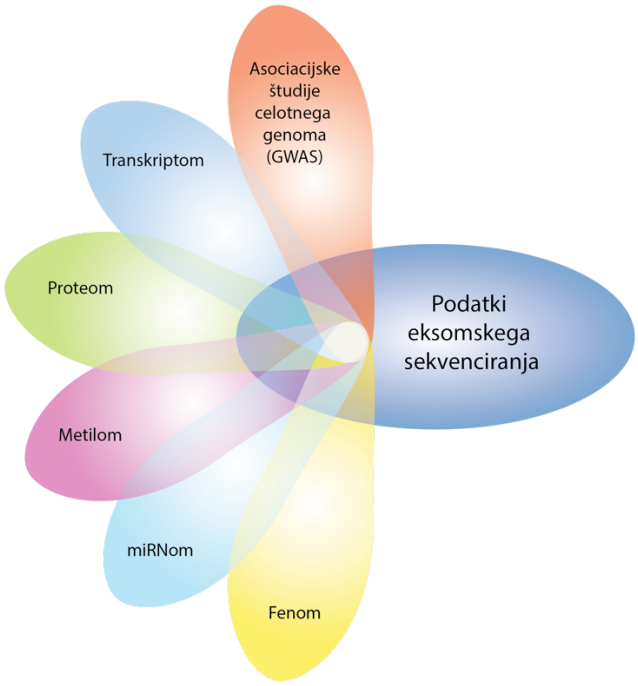
okvirja). V kategorijo z nizko pričakovano patogenostjo pa smo uvrstili vse preostale različice.

V kategoriji s srednjo stopnjo pričakovane patogenosti so bile najbolj zastopane različice z zamenjavo aminokislinskega zaporedja. Ker je biološki pomen tovrstnih različic v večini primerov neznan, smo njihov pričakovan učinek nadalje opredelili z algoritmi za napoved učinka na strukturo in delovanje beljakovinskega produkta. Za ta namen smo uporabili napovedne algoritme, ki uporabljajo naslednje parametre: evolucijsko ohranjenost mesta različice, evolucijsko ohranjenost domene z različico, sterični vpliv ugotovljene aminokislinske zamenjave, sprememba v naboju aminokisline, prisotnost morebitne disrupcije disulfidnih vezi v področju, pristotnost spremembe v aktivnem mestu beljakovine in številne druge parametre. Za namen te predikcije smo uporabili nabor algoritmov za napovedovanje patogenosti, ki glede na prej omenjene podatke podajo napoved patogenosti aminokislinske zamenjave: SIFT, PolyPhen-2, MutationTaster, CADD in Meta-SVM (Grimm in sod., 2015).

Z namenom odkrivanja različic, ki so napovedano visoko patogene in redke, smo se usmerili predvsem na različice s frekvenco v splošni populaciji manj kot 1 %, z vsaj srednjo napovedjo učinka (različice z zamenjavo aminokislinskega zaporedja, različice z verjetnim vplivom na izrezovanje intronov, različice s prezgodnjo vstavitvijo stop kodona in različice s premikom bralnega okvirja). Napovedi z zamenjavo aminokislinskega zaporedja smo obravnavali kot kandidatne, če so vsaj trije od petih uporabljenih algoritmov napovedali škodljiv učinek zaradi zamenjave aminokisline. Vse različice, ki smo jih ocenili kot potencialno pomembne pri bolnikih z MS, smo naknadno še usmerjeno preverili na nivoju izvornih NGS podatkov.

3.6 UPORABA PODATKOV POZICIJSKE INTEGRACIJE ZA INTERPRETACIJO REZULTATOV EKSOMSKEGA SEKVENCIRANJA PRI MULTIFAKTORSKIH BOLEZNIH

Za prioritizacijo in izbor različic, ugotovljenih pri eksomskem sekvenciranju pri bolnikih z MS, smo v nadaljnjem koraku uporabili rezultate integracije heterogenih omskih podatkov pri MS (shematsko je koncept prikazan na Sliki 7). Uporabili smo nabor genov v najvišje uvrščenih regijah, kjer smo ugotovili prekrivanje na vsaj treh bioloških nivojih in s statistično pomembnim zbiranjem signalov iz heterogenih omskih študij. Pozicijo odkritih sprememb smo prekrili s podatki pozicijske integracije in interpretirali variante glede na rezultat pozicijske integrativne analize. Preverili smo, ali prihaja do zbiranja genetskih različic z visokim potencialom za patogenost v regijah, ki so bile visoko uvrščene glede na rezultate pozicijske integracije. V končnem koraku smo izbrali nabor redkih, napovedano visoko patogenih različic v genih, ki so bili najvišje uvrščeni po integrativni analizi omskih podatkov pri MS.



Slika 7: Shema pristopa uporabe rezultatov integrativne analize pri multifaktorski bolezni za interpretacijo podatkov eksomskega ali genomskega sekvenciranja.
Seznam genov, ki smo jih identificirali na podlagi podatkov iz heterogenih omskih študij smo uporabili za zožitev nabora kandidatnih različic, odkritih s preiskavo eksomskega sekvenciranja.
Figure 7: Schematic presentation of utilization of positional integration results for interpretation of data generated by exome and genome sequencing.
The list of genes, identified on the basis of integrative analysis of omic studies has been used for focusing the list of candidate variants identified using exome sequencing.

4 REZULTATI

4.1 INTEGRACIJA HETEROGENIH OMSKIH PODATKOV PRI PARKINSONOVI BOLEZNI

4.1.1 Pregled zbranih podatkov iz vključenih študij za Parkinsonovo bolezen

S preiskovanjem literature in baz podatkov smo pridobili rezultate za 6 ločenih omskih nivojev sprememb pri PB, ki so bili zbrani v 15 študijah. Podatki o zbranih študijah, biološki nivoji vključenih študij in pripadajoči viri so navedeni v Preglednici 3.

Preglednica 3: Pregled študij, ki smo jih vključili v integrativno analizo omskih podatkov pri bolnikih s PB.
Table 3: Overview of studies included in the integrative analysis of omic data in Parkinson disease.

Zap. št. študije	Ime študije, letnica	Citat
Asociacijske študije celotnega genoma (GWAS)		
PB001	Maraganore 2005	(Maraganore in sod., 2005)
PB002	Fung 2006	(Fung in sod., 2006)
Študije genetske vezave		
PB003	Foltynie 2005	(Foltynie in sod., 2005)
Transkriptomske študije v krvi bolnikov s PB		
PB004	Scherzer 2007	(Scherzer in sod., 2007)
Transkriptomske študije v tkivu CŽS bolnikov s PB		
PB005	Moran 2006	(Moran in sod., 2006)
PB006	Lesnick 2007	(Lesnick in sod., 2007)
PB007	GSE7307	Neobjavljena študija, podatki so na voljo v repozitoriju GEO pod dostopno številko GSE7307
Proteomske študije sprememb nivojev beljakovin v bolnikov s PB		
PB008	Jin 2006	(Jin in sod., 2006)
PB009	Abdi 2006	(Abdi in sod., 2006)
PB010	Choi 2004	(Choi in sod., 2004)
PB011	Sinha 2007	(Sinha in sod., 2007)
PB012	Sinha 2009	(Sinha in sod., 2009)
PB013	Werner 2008	(Werner in sod., 2008)
PB014	Basso 2004	(Basso in sod., 2004)
Geni s fenotipskimi podobnostmi simptomov pri PB		
PB015	Baza Human Phenotype Ontology	(Robinson in sod., 2008)

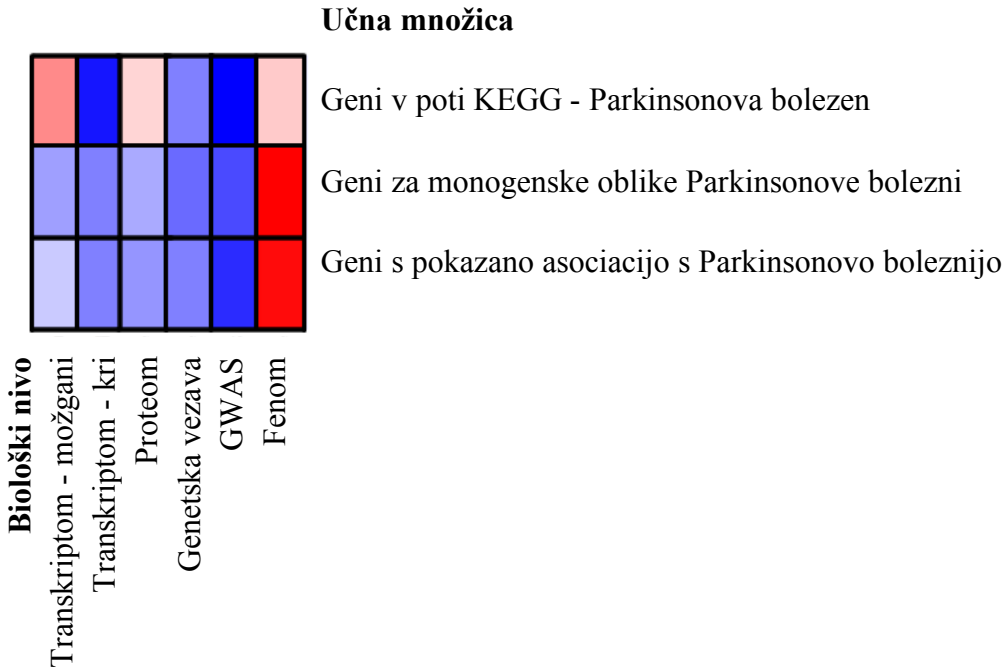
Zbrani podatki za integracijo so vključevali 4.700 genov s spremenjenim izražanjem po meta-analizi treh podatkovnih setov transkriptomskih študij na tkivu CŽS pri bolnikih s PB, 1.731 genov pa smo vkjlučili na podlagi študije o globalnem profilu razlik v izražanju genov v periferni krvi bolnikov s PB. V analizo smo vključili še 199 genov za beljakovine s spremenjenimi nivoji v različnih proteomskih študijah, vključili smo tudi 214 regij iz študij genetske vezave s PB, vključili smo podatke o 1.604 SNPih s statistično pomembno asociacijo s PB in 1.235 genov s fenotipsko kompatibilinostjo klinični sliki pri PB.

4.1.2 Rezultati integrativne analize omskih podatkov pri Parkinsonovi bolezni

Z izbranimi signali smo opravili analizo s pristopom pozicijske integracije, pri kateri smo analizirali zbiranje signalov v regijah velikosti 10 kilobaznih parov. Najprej smo preverili razporeditev signalov na nivoju celotnega genoma. Za zbiranje signalov iz vključenih omskih študij smo preiskali 616.108 genomskih regij - od teh v 476.810 (77.4 %) regij ni vsebovalo nobenega signala iz vključenih študij, v 121.011 regijah (19.6 %) smo zaznali signal le z enega biološkega nivoja, v 16.214 regijah smo ugotovili signale iz po 2 različnih tipov študij. V majhnem deležu regij 1.969 (0.32 %) smo ugotovili signale iz po treh tipov študij, v 103 regijah (0.017 %) smo ugotovili prekrivanje iz 4 tipov študij in v 1 primeru (0.00016 %) smo ugotovili prekrivanje signalov iz 5 različnih bioloških nivojev. Agregacije signalov na vseh šestih bioloških slojih nismo ugotovili za nobeno od preiskanih regij.

Najvišje uvrščene regije z največjo akumulacijo signalov (z ugotovljenimi spremembami na vsaj 4 bioloških nivojih), se je večina nahajala na kromosomu 17 (18 regij), kromomsomu 9 (17 regij), kromosomu 15 (14 regij), 18 (9 regij) in po 7 regij na kromosomih 4 in X. V številnih primerih identificiranih regij, smo ugotovili, da se bodisi prekrivajo ali pa sestavljajo daljše neprekinjene regije z zbiranjem signalov. Ko smo takšne sosednje regije združili, je preostalo 29 kontinuiranih regij različne velikosti, ki so vsebovale biološke spremembe na vsaj 4 bioloških nivojih.

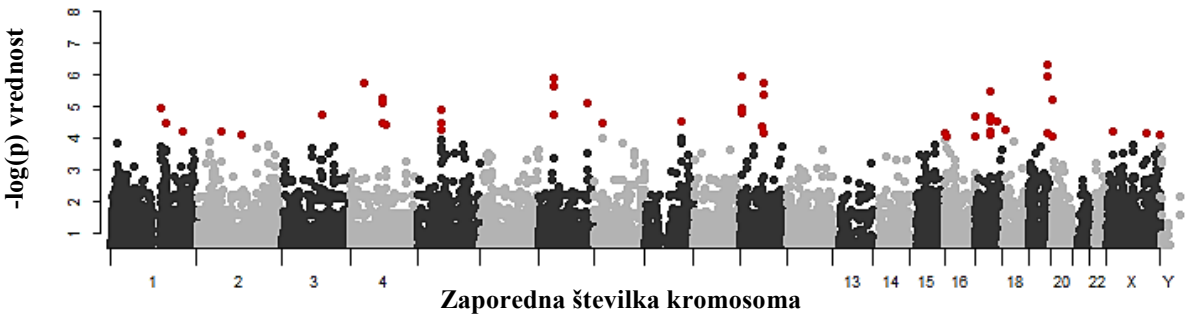
Pred končno integracijo podatkov smo poskusili identificirati optimalni nabor uteži za posamezni biološki sloj v končni integraciji. V primeru ko smo za nabor učnih genov uporabili KEGG pot z geni vključenimi v patogenezo PB, smo dosegli najvišjo prioritizacijo učnih genov, če smo najbolj obtežili transkriptomske študije v možganih, proteomske študije in deloma tudi fenomske študije (Slika 8). Ko pa smo za učno množico izbrali ozek nabor genov, povezanih z znanimi monogenskimi oblikami PB in geni v genetski asociaciji s PB pa se je integracija dala najvišje rezultate, če smo najvišje obtežili biološki sloj s fenomskimi podatki (Slika 8). Za končno prioritizacijo smo uporabili konfiguracijo uteži, kalibrirano na učni množici v KEGG poti za PB.



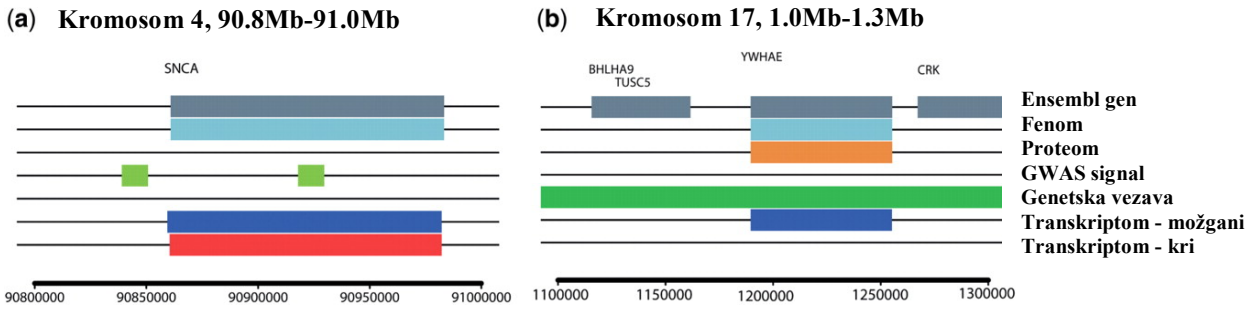
Slika 8: Ocena optimalne kombinacije uteži posameznih bioloških slojev pri integraciji omskih študij.
Optimalne uteži posameznih bioloških nivojev smo kalibrirali s tremi nabori učnih genov. Toplotna karta kaže s katero kombinacijo uteži posameznih bioloških slojev smo dosegli najvijšo prioritizacijo genov v učni množici - rdeča barva pomeni večjo obtežitev, modra barva pa manjšo obtežitev.
Figure 8: Calibration of best matrices of weights using different training gene sets.
Three sets of genes were utilized for optimization of weights attributed to included biological layers. The heatplot represents layers that were weighted more (red) and those that were given less weight (blue) in for attaining the highest ranking of genes in the training set.

V nadaljnjem koraku smo opravili obteženo pozicijsko integrativno analizo in analizo statistično pomembnega zbiranja signalov s heterogenih bioloških slojev z opravljanjem 1000 permutacij vrednosti za regije na posameznem sloju.

Končna razporeditev rezultatov integrativne analize je prikazana na Sliki 9. V skupnem merilu smo za 179 regij (0.029 %) ugotovili verjetnost verjetno nenaključno zbiranje signalov pod stopnjo lažno pozitivnih rezultatov enako 0.0001. Po združitvi blizu ležečih regij je rezultat pomenil v skupnem 33 regij, ki so vsebovale 29 genov. Vrhova najvišje statistične asociacije smo ugotovili v genih UCHL1 in SNCA na kromosomu 4, v regiji gena GFAP na kromosomu 17, APOE na kromosomu 19, poleg tega pa smo identificirali tudi regije s potencialnimi novimi kandidatnimi geni na kromosomih 17, 11 in 20.



Slika 9: Graf razporeditve rezultatov pozicijske integracije na področju celotnega genoma na primeru Parkinsonove bolezni.
X-vrednost prikazuje položaj na genomu, y-os pa višino signala na podlagi integrativne analize, torej informacijo o izrazitosti zbiranja signalov v genomu.
Figure 9: Genome-wide distribution of p values across the genome for integration in the example of Parkinson disease.
X-axis reflects genome position and y-axis represents significance estimates of integration values.



Slika 10: Dokazi, ki utemeljujejo visoko uvrstitev genov SNCA in YWHAE pri integrativni analizi za PB.
(**a**) Dokazi, ki utemeljujejo visoko uvrstitev gena SNCA in pripadajoče regije na kromosomu 4. Graf prikazuje podroben pregled položaja izvornih bioloških signalov iz večih omskih študij. (**b**) Dokazi, ki uvrščajo regijo gena YWHAE na kromosomu 17 med najvišje uvrščene regije.
Figure 10: Evidence for high ranking of SNCA and YWHAE genes in integrative analysis in Parkinson disease.
(**a**) Representation of co-location of biological signals withing SNCA gene region. Plot shows signals that contributed to high ranking of the SNCA gene region in the final priotiziation. (**b**) Evidence supporting prioritization of YWHAE region on chromosome 17

Če smo uporabili manj striktno mero za statistično pomembnost rezultatov (PFP<0.05) je to mero preseglo skupno 23.647 intervalov (3.84 %), ki so po združitvi sosednjih regij predstavljali 2.748 regij s skupno 2.183 geni. Zaradi velikega števila genov pri manj

restriktivni meri statistične pomembnost, smo uporabili regije, ki so presegle mejo PFP<0.0001 v vseh nadaljnjih analitičnih korakih.

Med regijami z najvišjimi vrednosti po integraciji, so nekatere regije vsebovale že znane gene, povezane s PB - bodisi v smislu monogenskih vzrokov, v meta-analizah ali posameznih asociacijskih študijah - med temi sta najbolj vidna primera gena SNCA in UCHL1. Na Sliki 10a predstavljamo empirične dokaze iz vključenih omskih študij, ki utemeljujejo visoko uvrstitev gena SNCA.

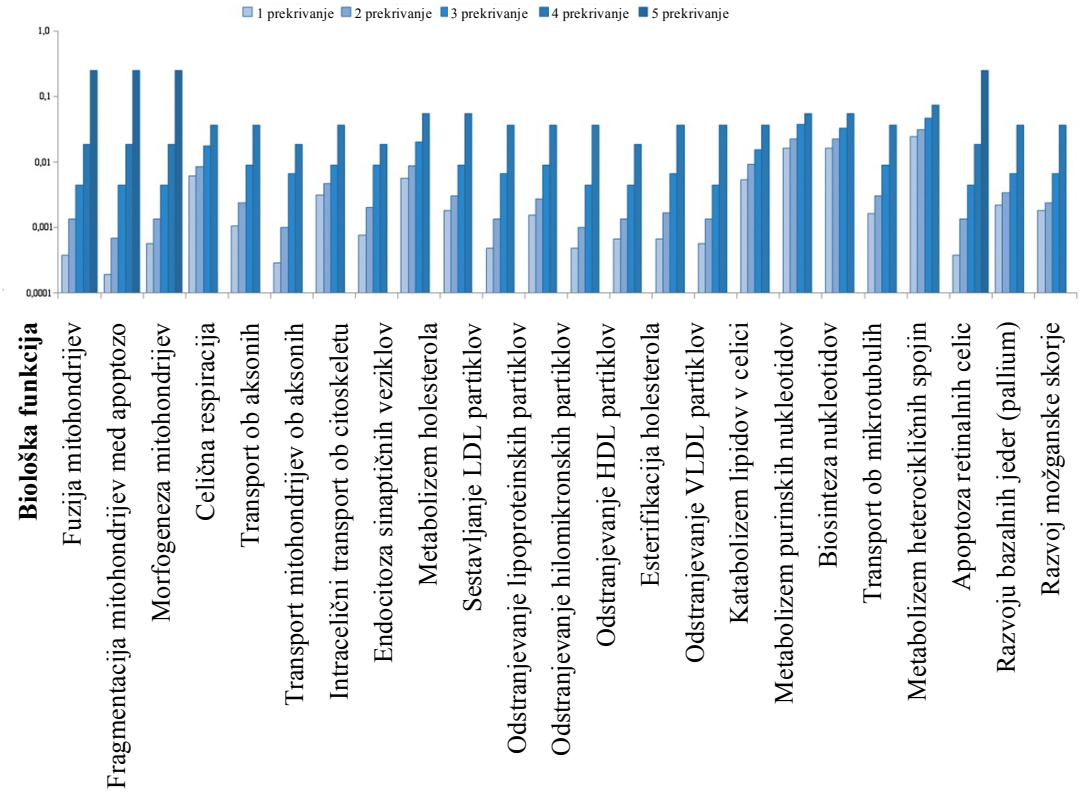
Poleg znano povezanih genov s PB, smo ugotovili tudi nove kandidatne gene, kjer smo ugotovili pomembno zbiranje signalov omskih študij, kljub temu pa geni predhodno še niso bili identificirani kot pomembni pri PB. Kot primer na Sliki 10b predstavljamo empirične dokaze iz vključenih omskih študij, ki utemeljujejo visoko uvrstitev gena YWHAE, ki dosedaj neposredno s PB še ni bil povezan, čeprav je bil v študijah na 4 omskih nivojih kazal statistične spremembe.

4.1.2.1 Evalvacija integrativnega pristopa pri odkrivanju novih kandidatnih regij in genov za multifaktorske bolezni (primer Parkinsonove bolezni)

Preiskava znanih asociacij med 29 geni v najvišje uvrščenih regijah po integraciji, je pokazala da je bilo od teh 15 genov že neposredno povezano s PB, vsaj 10 identificiranih genov pa se je pojavilo v literaturi skupaj s PB v vsaj 10 publikacijah (MAPT, BAX, APOE, GFAP, SNCA, PRNP and UCHL1).

Iskanje posrednih povezav v literaturi z orodjem BITOLA (Hristovski in sod., 2005) je pokazalo da so številni geni, ki dosedaj direktno še niso bili povezani s PB, povezani s PB indirektno: 11 preostalih genov preko njihove vloge v nevrodegenerativnih procesih, 3 geni preko regulatornih procesov v živčevju in 1 gen preko vpletenosti v regulacijo apoptoze nevronov, 4 geni pa preko povezav z drugimi boleznimi CŽS.

S hipergeometričnim testom smo tudi analizo obogatenosti funkcij genov v najvišje uvrščenih regijah glede na povezane funkcije v ontologiji GeneOntology (GO) (prikazano na sliki 11). Največji delež genov je bil povezan z GO pojmi za specifične celične procese v nevronih (predvsem s procesi za razvoj in delovanje aksonov), poleg tega pa tudi z onotološkimi termini za metabolizem malih molekul, metabolizmom lipidov in sterolov ter apoptotske procese.



Slika 11: Funkcijski profil genov v najvišje uvrščenih regijah, identificiranih s pristopom pozicijske integracije heterogenih omskih študij pri PB.
Z večjim prekrivanjem heterogenih signalov v regijah je obogatenost genov v povezavi s Parkinsonovo boleznijo čedalje večja.
Figure 11: Functional profiling of genes located in the top integrative regions, identified using the integrative omics approach in Parkinsons disease.
Increasing proportion of genes related to Parkinson disease may be observed when investigating genes in regions with greater number of overlaps.

Preverili smo tudi ali so geni, za katere smo ugotovili prekrivanje signalov na več bioloških nivojih (višja vrednost po integraciji heterogenih omskih podatkov) tudi bolj povezani z znanimi patogenetskimi procesi pri PB. Ugotovili smo da so v skupini genov z višjimi integrativnimi vrednostmi obogatene funkcije, povezane s patogenetskimi mehanizmi pri PB (vključujoč funkcijo mitohondrijev, metabolizmom lipidov/holesterola, malih molekul in nevrorazvojne procese, Slika 11).

Obogatitvene analize z anotacijami v bazi reaktomskih poti so pokazale, da so najvišje uvrščeni geni po integrativni analizi pri PB udeleženi pri reaktomskih poteh membranskega transporta (P=0.011), metabolizma lipidov in sinaptičnega prenosa (P=0.044).

4.2 RAZVOJ SAMOSTOJNEGA ORODJA ZA INTEGRATIVNO ANALIZO

Možnost integrativne analize smo implementirali tudi v obliki spletnega orodja za analizo drugih naborov podatkov, ki je na voljo na spletnem naslovu:
http://db.signgenetics.eu/integratomics/home .

Algoritem smo implementirali v obliki spletnega orodja s katerim lahko uporabniki sami opravijo svoje integrativne analize, tudi na drugih primerih bolezni človeka. Vhodne podatke je potrebno pripraviti v obliki tekstovnih datotek, ki vsebujejo sezname sprememb iz različnih omskih študij.

Tekstovne datoteke morajo biti pripravljene tako da je v prvi vrstici naveden omski tip študije (na primer GWAS, transkriptom, proteom - kri, idr.). V drugi vrstici mora biti navedeno ime študije, v tretji vrstici pa tip biološkega podatka. Uvodnim trem vrsticam sledijo podatki iz študije, ki v prvem stolpcu vsebujejo podatke o anotaciji spremembe, v drugem pa o meri za velikost spremembe, ki se bo uporabila pri razvrščanju med integracijo rezultatov (Slika 12).

GWAS	←	**Omski nivo, tip študije**
IMSGC 2007	←	**Ime študije**
SNP	←	**Tip podatka**
rs3129934	2,60E-40	← **Anotacija in mera za velikost spremembe)**
rs9270986	2,38E-39	
rs3129900	2,63E-37	
rs3129768	3,37E-35	
rs3131294	4,04E-32	
rs3130287	2,26E-28	
rs3129932	5,76E-28	
rs3135377	6,32E-28	

Slika 12: Prikaz pripravljenih vhodnih podatkov za uporabo razvitega programa za pozicijsko integracijo.
Figure 12: Excerpt from the file prepared for use with software for custom integrative analysis.

Pred zagonom integracije lahko definiramo tudi relativno obtežitev posameznega nivoja z vrednostmi 0-10. Izberemo lahko tudi število permutacij in velikost uporabljenih regij (priporočeno od 10 kb do 500 kb). Ko pripravimo datoteke za različne študije in različne tipe študij, datoteke z orodjem naložimo in poženemo algoritem za integracijo, rezultat pa bo podan v obliki razporeditve rezultatov integracije po celotnem genomu in seznam genov v najvišje uvrščenih regijah (Slika 13).

Algoritem bo v prvem koraku opravil integracijo študij z istega biološkega nivoja, v drugem koraku pa integracijo različnih bioloških nivojev.

Slika 13: Prikaz pregleda rezultata po zaključeni integrativni analizi z razvitim algoritmom za analizo lastnih omskih podatkov in spletnim vmesnikom.
Figure 13: Demonstration of results of the algorithm implemented in the web tool for custom analysis.

4.3 INTEGRACIJA HETEROGENIH OMSKIH PODATKOV PRI MULTIPLI SKLEROZI

4.3.1 Pregled zbranih podatkov iz vključenih študij za MS

S preiskovanjem literature in baz podatkov smo pridobili rezultate za 12 ločenih omskih nivojev sprememb pri MS, ki so bili pridobljeni v 52 študijah. Podatki o zbranih študijah, biološki nivoji vključenih študij in pripadajoči viri so navedeni v Preglednici 4.

Preglednica 4: Pregled študij, ki smo jih vključili v integrativno analizo omskih podatkov pri MS.
Table 4: Overview of studies included in the integrative analysis of omic data in multiple sclerosis.

Zap. št. študije	Ime študije, letnica	Citat
Asociacijske študije celotnega genoma (GWAS)		
MS001	IMSGC 2007	(Hafler in sod., 2007)
MS003	Baranzini 2009	(Baranzini in sod., 2009)
MS004	ANZgene, 2009	(Bahlo in sod., 2009)
MS005	Sanna 2010	(Sanna in sod., 2010)
MS006	Sawcer 2011	(Sawcer in sod., 2011)
MS007	Boneschi 2012	(Martinelli-Boneschi in sod., 2012)
Študije genetske vezave		
MS014	GAMES 2003	(Haines, 2003)
Transkriptomske študije v krvi bolnikov z MS		
MS082	Achiron 2010	(Achiron in sod., 2010)
MS083	Bomprezzi 2003	(Bomprezzi in sod., 2003)
MS084	Achiron 2004	(Achiron in sod., 2004)
MS085	Iglesias 2004	(Iglesias in sod., 2004)
MS086	Mandel 2004	(Mandel in sod., 2004)
MS087	Satoh 2005	(Satoh in sod., 2005)
MS088	Nickles 2013	(Nickles in sod., 2013)
MS089	Irizar 2014	(Irizar in sod., 2014)
MS090	Sarkijarvi 2006	(Sarkijarvi in sod., 2006)
MS091	Corvol 2008	(Corvol in sod., 2008)
MS093	Fossey 2007	(Fossey in sod., 2007)
MS095	Zhang 2011	(Zhang in sod., 2011)
MS097	Gandhi 2010	(Gandhi in sod., 2010)

se nadaljuje ...

Nadaljevanje **Preglednice 4:** Pregled študij, ki smo jih vključili v integrativno analizo omskih podatkov pri MS.

Zap. št. študije	**Ime študije, letnica**	**Citat**
Transkriptomske študije v tkivu CŽS bolnikov z MS		
MS070	Han 2012	(Han in sod., 2012)
MS071	Lindberg 2004	(Lindberg in sod., 2004)
MS072	Tajouri 2003	(Tajouri in sod., 2003)
MS073	Cunnea 2010	(Cunnea in sod., 2010)
MS074	Lock 2002	(Lock in sod., 2002)
MS075	Graumann 2003	(Graumann in sod., 2003)
MS076	Zeis 2015	(Zeis in sod., 2015)
MS077	Zeis 2008	(Zeis in sod., 2008)
Transkriptomske študije v CD4 celicah bolnikov z MS		
MS092	Zastepa 2014	(Zastepa in sod., 2014)
Proteomske študije sprememb nivojev beljakovin v periferni krvi bolnikov		
MS067	Ritchiedech 2009	(Rithidech in sod., 2009)
Proteomske študije sprememb nivojev beljakovin v cerebrospinalni tekočini (CSF) bolnikov		
MS063	Qin 2009	(Liu in sod., 2009)
MS064	Noben 2005	(Noben in sod., 2006)
MS065	Hammack 2004	(Hammack in sod., 2004)
MS066	Lehmeniscek 2007	(Lehmensiek in sod., 2007)
Proteomske študije sprememb nivojev beljakovin v CŽS bolnikov		
MS062	Han 2008	(Han in sod., 2008)
Študije globalne diferencialne ekspresije miRNA		
MS048	Boneschi 2012	(Martinelli-Boneschi in sod., 2012)
MS049	Siegel 2012	(Siegel in sod., 2012)
MS050	Keller 2009	(Keller in sod., 2009)
MS113	Junker 2009	(Junker in sod., 2009)
MS051	Cox 2010	(Cox in sod., 2010)
MS052	Lindberg 2010	(Lindberg in sod., 2010)
MS053	De Santis 2010	(De Santis in sod., 2010)
MS054	Keller 2014	(Keller in sod., 2014)
MS055	Jernas 2013	(Jernas in sod., 2013)
MS056	Fenoglio 2011	(Fenoglio in sod., 2011)
MS057	Sievers 2012	(Sievers in sod., 2012)
MS058	Noorbakhsh 2011	(Noorbakhsh in sod., 2011)
MS059	Gandhi 2013	(Gandhi in sod., 2013)

se nadaljuje ...

Nadaljevanje **Preglednice 4:** Pregled študij, ki smo jih vključili v integrativno analizo omskih podatkov pri MS.

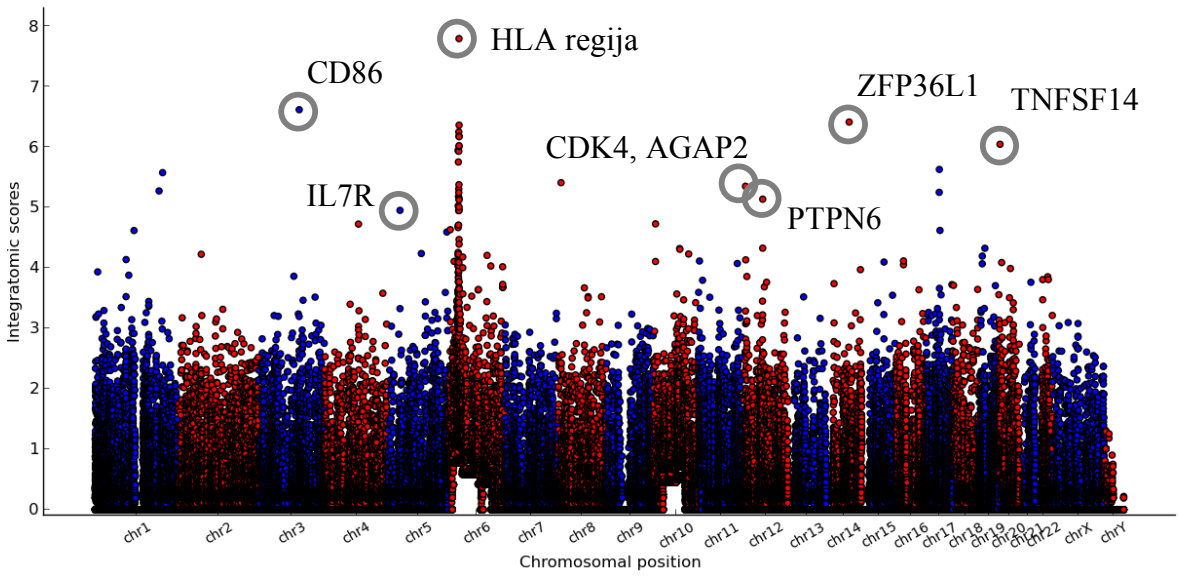
Zap. št. študije	Ime študije, letnica	Citat
Podatki o genskih tarčah miRNA s spremenjenim izražanjem		
MS110	Junker et al, 2009	(Junker in sod., 2009)
MS111	Keller 2009	(Keller in sod., 2009)
Regije različne metilacije pri bolnikihz MS		
MS069	Huynh 2014	(Huynh in sod., 2014)
Podatki o ortologih genov s spremenjenim izražanjem pri eksperimentalnim avtoimunem encefalitisu		
MS098	Nicot 2003	(Nicot in sod., 2003)
MS099	Brand 2005	(Brand-Schieber in sod., 2005)
MS100	Baranzini 2005	(Baranzini in sod., 2005)
MS101	Jonas 2014	(Jonas in sod., 2014)
MS102	Ibrahim 2001	(Ibrahim in sod., 2001)

4.3.2 Rezultati integrativne analize omskih podatkov pri MS

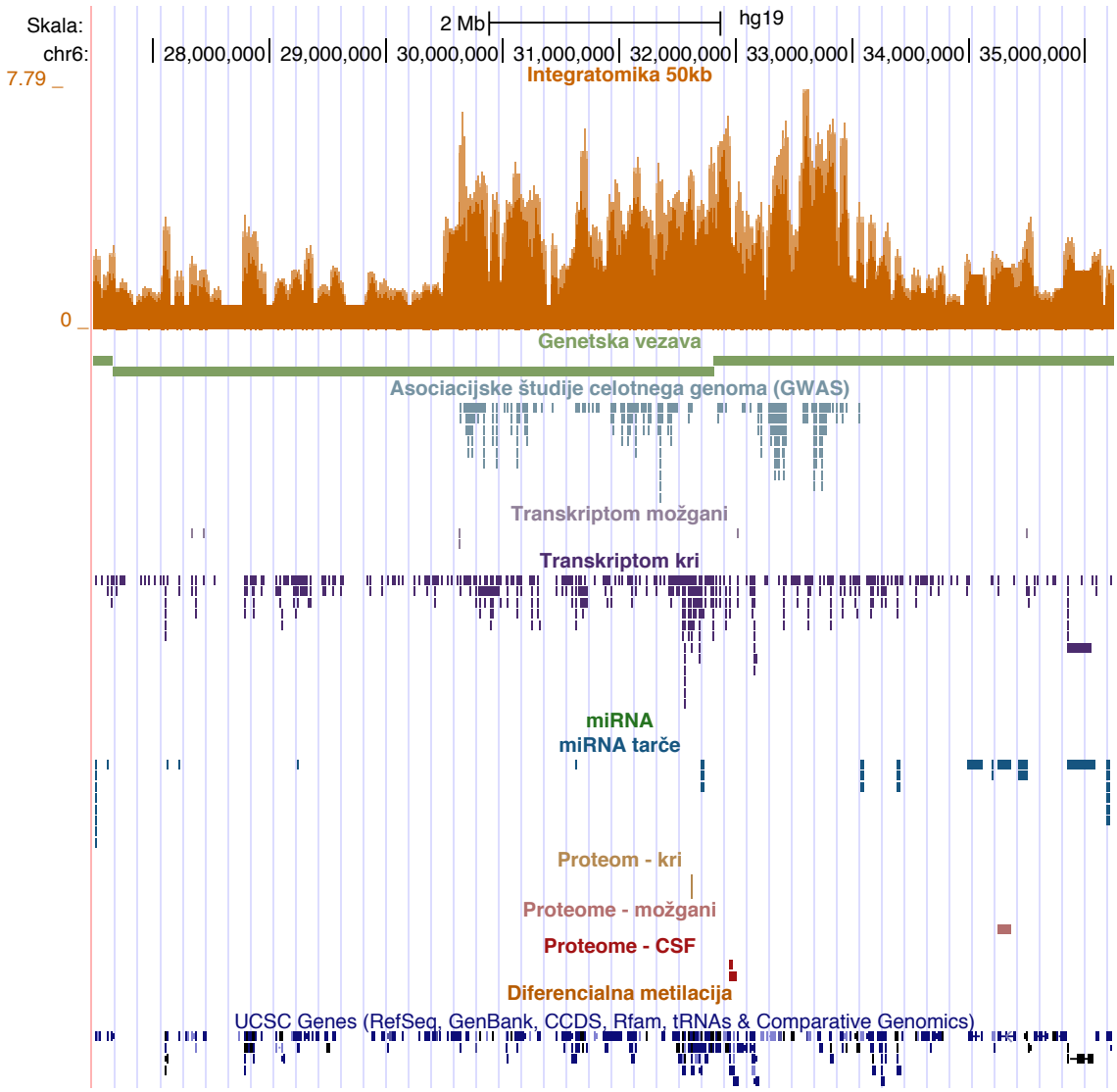
Zbrane omske študije pri MS smo vključili v integrativno analizo in z razvitim orodjem opravili integracijo na intervalih velikosti 100 kb. Rezultat razporeditev signalov na področjih celotnega genoma je prikazan na Sliki 14. Z integracijo heterogenih omskih podatkov pri MS smo identificirali 188 regij velikost 100 kb, kjer je prišlo do zbiranja signalov z ocenjeno statistično pomembnostjo PFP=0.001 (za stopnjo lažno pozitivnih rezultatov smo izbrali vrednost 0.001). Z namenom preverjanja izvornih podatkov v izbranih regijah, smo spremembe iz vključenih študij grafično prikazali in pregledali v genomskem brskalniku UCSC, s funkcijo prikaza lastnih podatkovnih virov ("*UCSC Custom tracks*"). Primer prikaza, s katerim smo ocenili zbiranje signalov in opravili izbor gena s prepričljivim zbiranjem signalov je prikazan za regijo HLA na Sliki 15.

Izrazito stopnjo zbiranja signalov smo ugotovili v področju kompleksa genov za humane levkocitne antigene HLA na kromosomu 6p21. Od 188 najvišje uvrščenih regij velikosti 100 kb smo jih 72 (38.3 %) ugotovili na kromosomu 6, večinoma v področju 6p21. Večina intervalov v področju 3 Mb področju HLA je kazala obogatenost signalov pri MS, ki so bili razporejeni po celotnem področju HLA. To opažanje je v skladu z znanimi podatki v literaturi, ki kaže na pomemben prispevek različnih elementov področja HLA pri nastanku MS. Podatke iz osnovnih študij, ki utemeljujejo visoko uvrstitev regije 6p21 pri integraciji prikazujemo na Sliki 15.

V najvišje uvrščenih regijah s PFP vrednostmi pod 0.001 se je nahajalo 435 genov. Funkcijske analize teh genov so v skladu po pričakovanjih pokazale najvišjo obogatenost pojmov v povezavi z vnetnimi in avtoimunimi procesi ter prezentacijo antigenov T celicam (p_{corr}<0.01), ki predstavljajo znane procese v povezavi z etiopatogenezo MS. Funkcijska analiza genov izven področja 6p21 je sicer pokazala manj izrazito obogatitev procesov povezanih z vnetjem, v ospredju pa so bili obogatene funkcije v povezavi s funkcijami citokinskih receptorjev, metabolizem lipidov in nespecifični vnetni odziv (p_{corr}<0.01), kar kaže na obogatenosti procesov v povezavi z MS tudi pri regijah identificiranih izven HLA področja. V regijah, ki obdajajo 6p21 so bile vrednosti po integraciji pomembno nižje, tudi stopnja zbiranja signalov je v bližnjih regijah izven HLA področja bistveno manjša.



Slika 14: Razporeditev vrednosti po integraciji signalov različnih študij na nivoju celotnega genoma pri MS.
Uporabljena velikost intervalov pri predstavljeni analizi je 100 kilobaznih parov. Na sliki je razvidno, da najvišje vrednosti zbiranja signalov dosegajo regije na kromosomu 6p21 (HLA regija), ki je dosedaj najbolj znano povezan z dedno dovzetnostjo za MS. Poleg najvišjih integrativnih vrednostih na kromosomu pa je razvidno, da do pomembnega zbiranja signalov prihaja tudi na številnih ne-HLA regijah. Na sliki so na abscisni osi predstavljeni kromosomi in koordinate regij, na y osi pa so predstavljene -logPFP vrednosti. Zanimivejši geni v najvišje uvrščenih regijah so označeni na grafu.
Figure 14: Distribution of integration values after integrating signals from omic studies, based on 100 kb regions used for integration.
The figure shows the highest accumulation of signals on chromosome 6p21 on the HLA region, which is the region with most studied association with MS. In addition to high scores for 6p21 region, non-HLA regions also show considerable enrichemnt. Genome-wide distribution of p values across the genome. X-axis reflects genome position and y-axis represents significance estimates of integration values. Dominant genes of interest in selected regions are annotated on the plot.

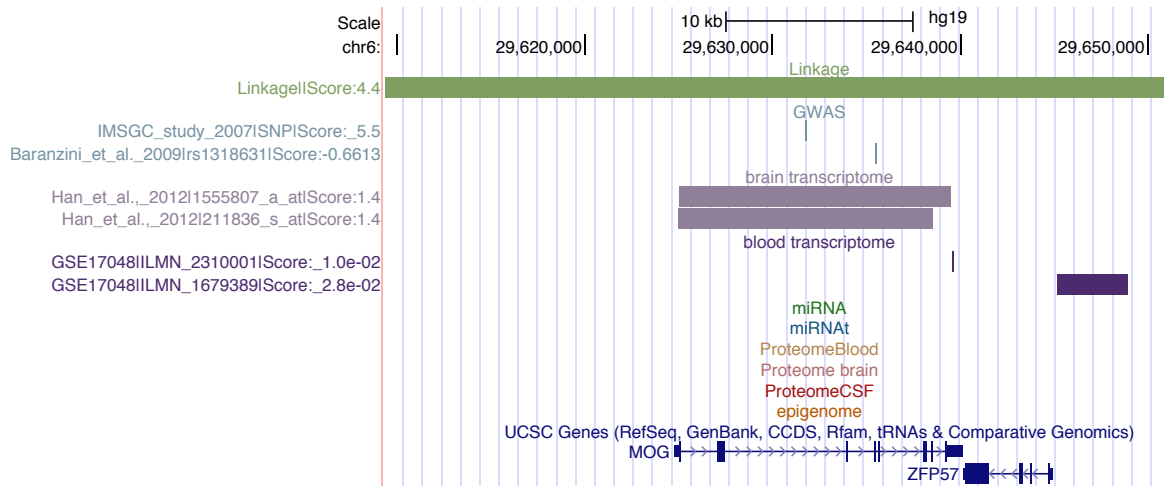


Slika 15: Pregled položajev signalov iz vključenih omskih študij pri MS - podroben pregled HLA regije na kromosomu 6p21.
Na področju kompleksa genov HLA je prisotno izrazito zbiranje signalov, podprto tako s podatki nivoju študij genetske vezave, GWAS študij, kot tudi na nivoju transkriptomski študij v krvi in na nivoju proteoma in miRNA tarč. Prikaz je bil pripravljen s pomočjo genomskega brskalnika UCSC.
Figure 15: Overview of signals contributing to the high integration scores for 6p21 region in MS. High score of 6p21 region stems from the co-occurence of signals form linkage, GWAS, transritpomic (blood and brain) and miRNA targets within HLA region. The figure was generating using UCSC genome browser.

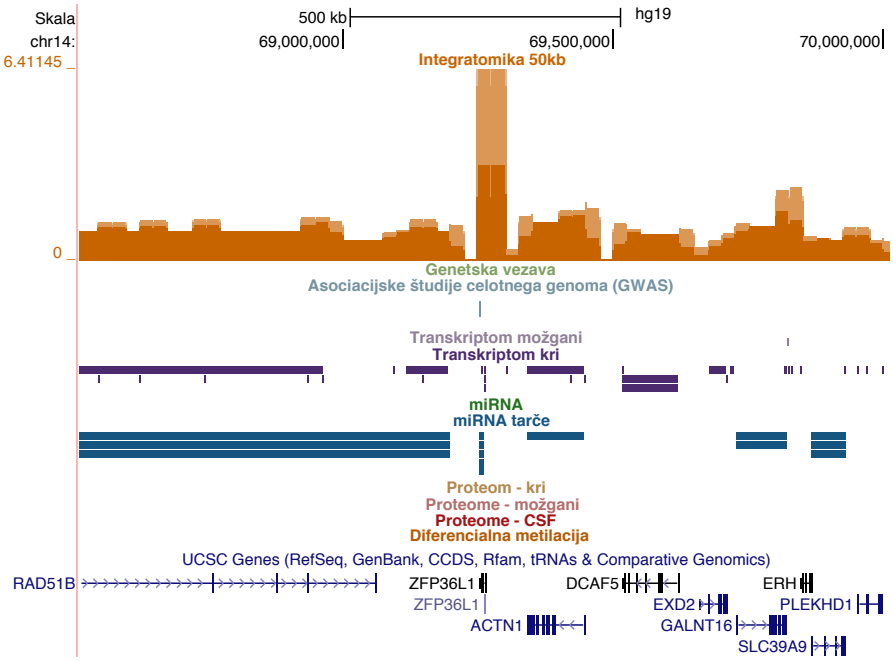
Za identifikacijo novih kandidatnih genov smo preiskali 20 regij z najvišjimi integrativnimi vrednostmi, ki so predstavljene tudi v preglednici 5. Od teh je bilo 15 regij na kromosomu 6 (vse v področju 6p21), ostale pa so se nahajale na kromosomih 12 (dve regiji), 3, 14 in 19. Med geni kandidati v HLA predelu kromosoma 6 je bil med najzanimivejšimi gen MOG, predstavljen na Sliki 16. Geni v najvišje uvrščenih ne-HLA regijah so bili naslednji: TNFSF14 (chr19), CD86 (chr3), AGAP2-TSPAN31-CDK4 (chr12), ZFP36L1 (chr14, Slika 17) in PTPN6 (chr12).

Preglednica 5: Seznam najvišje uvrščenih regij in pripadajočih genov pri integraciji podatkov za MS.
V tabeli so prikazane vrednosti posameznih omskih nivojev in skupen rezultat integracije. Najvišje uvrščene regije vsebujejo signale z vsaj štirih različnih bioloških nivojev, najvišje uvrščena regija pa celo signale s 6 različnih bioloških nivojev.
Table 5: The list of highest ranked regions after positional integration analysis of included omic studies for MS.
The analysis was performed using 100 kb size of the regions. The table also displays scores for each region on separate biological levels. The highest ranked regions contain signals from at least 4 distinct biological layers, with the higest ranked region containing co-occuring biological signals from 6 different study type.

Genomska regija (50kb)	Geni	logP vrednost (po integraciji)	Asociacijske študije celotnega genoma (GWAS)	Študije genetske vezave	Transkriptom - kri	Transkriptom - možgani	Transkriptom - T celice	Proteom - kri	Proteom - cerebrospinalna tekočina	Proteom - možgani	Diferencialna metilacija v krvi	miRNA spremembe	Tarčni transkripti za spremenjene miRNA	Transkriptomske spremembe pri EAE mišjih modelih
chr6-31900000-31999999	DOM3Z, MIR1236, RDBP, SKIV2L, STK19	**7,5**	2,5	4,4	2,5	0,0	7,1	0,0	5,6	0,0	0,0	0,0	0,0	5,7
chr6-29600000-29699999	MOG, SUMO2P1, ZFP57	**6,9**	4,7	4,4	0,7	3,5	1,5	0,0	0,0	0,0	0,0	0,0	0,0	3,1
chr6-32600000-32699999	HLA-DQB1, HLA-DQB1-AS1	**6,6**	4,6	4,4	2,2	0,0	1,5	0,0	0,0	0,0	0,0	0,0	0,0	6,6
chr6-31550000-31649999	AIF1, LST1, NCR3, SNORA38, UQCRHP1	**6,5**	2,7	4,4	6,4	0,0	5,5	0,0	0,0	0,0	0,0	0,0	0,0	2,2
chr6-31850000-31949999	EHMT2-AS1, ZBTB12	**6,3**	4,8	4,4	1,0	0,0	5,5	0,0	0,0	0,0	0,0	0,0	0,0	5,3
chr19-6650000-6749999	TNFSF14	**6,0**	3,0	0,0	2,3	0,0	2,1	0,0	5,4	0,0	0,0	0,0	0,0	5,4
chr6-32900000-32999999	BRD2, BRD2-IT1, HLA-DMA, HLA-DMB	**5,7**	2,8	4,4	3,4	0,0	0,8	0,0	0,0	0,0	0,0	0,0	0,0	4,0
chr6-31750000-31849999	HSPA1A, HSPA1B, HSPA1L, LSM2	**5,5**	0,0	4,4	7,5	1,9	6,9	0,0	0,0	0,0	0,0	0,0	0,0	6,1
chr6-30650000-30749999	MDC1, MDC1-AS1, NRM, RPL7P4, TUBB	**5,5**	3,1	4,4	3,7	0,0	4,3	0,0	0,0	0,0	0,0	0,0	0,0	1,2
chr6-32800000-32899999	PPP1R2P1, PSMB8, PSMB9, TAP1	**5,5**	2,6	4,4	1,2	0,0	1,4	0,0	0,0	0,0	0,0	0,0	0,0	7,5
chr6-30600000-30699999	C6orf136, DHX16, PTMAP1	**5,4**	4,6	4,4	1,2	0,0	2,8	0,0	0,0	0,0	0,0	0,0	0,5	1,7
chr6-32400000-32499999	HLA-DRA, HLA-DRB9	**5,4**	5,2	4,4	4,7	0,0	0,7	0,0	0,0	0,0	0,0	0,0	0,0	0,0
chr12-58100000-58199999	AGAP2, CDK4, TSPAN31	**5,4**	5,0	0,0	2,6	1,2	3,4	0,0	0,0	0,0	0,0	0,0	0,0	1,4
chr6-32700000-32799999	HLA-DQA2, HLA-DQB2, MIR3135B	**5,3**	3,4	4,4	0,4	0,0	0,2	0,0	0,0	0,0	0,0	0,0	0,0	6,6
chr14-69250000-69349999	ZFP36L1	**5,3**	5,6	0,0	2,0	1,7	0,4	0,0	0,0	0,0	0,0	0,0	2,1	0,0
chr6-30100000-30199999	TRIM10, TRIM15, TRIM40	**5,3**	5,8	4,4	1,2	0,0	3,4	0,0	0,0	0,0	0,0	0,0	0,0	0,6
chr6-31600000-31699999	APOM, BAG6, C6orf47, C6orf47-AS1,	**5,3**	2,6	4,4	2,8	0,0	4,8	1,0	0,0	0,0	0,0	0,0	0,0	0,0
chr6-31800000-31899999	C6orf48, NEU1, SLC44A4, SNORD48, SNORD52	**5,2**	2,4	8,8	2,0	0,0	3,0	0,0	0,0	0,0	0,0	0,0	0,0	0,7
chr6-32750000-32849999	HLA-DOB	**5,1**	3,2	4,4	0,9	0,0	0,9	0,0	0,0	0,0	0,0	0,0	0,0	4,7
chr12-7050000-7149999	C12orf57, C1S, EMG1, LPCAT3, MIR141,	**5,1**	0,0	0,0	4,5	3,6	2,8	0,0	0,0	2,0	0,0	0,0	0,7	3,0



Slika 16: Pregled zbiranja signalov na eni najvišje uvrščenih regij na kromosomu 6, v področju gena MOG.
Predstavljeni gen kodira za mielniski glikoprotein oligodendrocitov (angl. *myelin oligodendrocyte glycoprotein*). Slika prikazuje podporo za visoko uvrstitev regije na podlagi dokazov iz asociacijskih študij celotnega genoma, transkriptomskih študij na tkivu možganskega tkiva in transkriptomskih študij na periferni krvi. Prikaz je pripravljen z genomskim brskalnikom UCSC in prikazano na genomskem sestavu različica hg19.
Figure 16: Overview of signal aggregation in one of the highest ranked regions in MS, which contains MOG gene.
The presented gene codes for *myelin oligodendrocyte glycoprotein*. The figure demonstrates support for high ranking of the region, due to co-occurence of signals from linkage, GWAS and transcriptome (blood and brain). The figure was generating using UCSC genome browser with hg19 genome assembly.

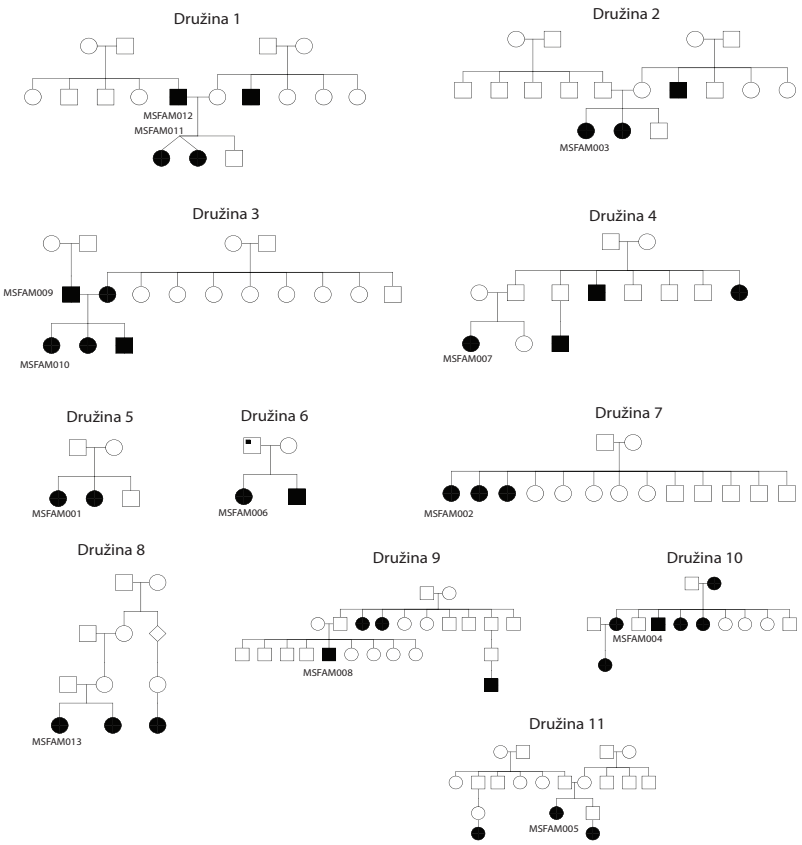


Slika 17: Pregled zbiranja signalov pri MS na eni najvišje uvrščenih ne-HLA regij v področju gena ZFP36L1.
Visoko uvrstitev omenjenega gena utemeljujejo prekrivajoči dokazi GWAS študij, transkriptomskih študij v krivi, geni pa je tudi podvržen regulaciji miRNA, ki imajo pri MS spremenjen profil izražanja.
Figure 17: Presentation of signal aggregation in one the the highest scoring non-HLA regions containing the ZFP36L1 gene.
The high ranking of the gene is supported by the evidence from GWAS studies, transcriptome studies in blood and miRNA level hits.

4.4 UPORABA ALGORITMA ZA INTEGRACIJO OMSKIH ŠTUDIJ PRI INTERPRETACIJI REZULTATOV EKSOMSKEGA SEKVENCIRANJA PRI MS

4.4.1 Družinski primeri bolnikov z MS

V preiskavo z eksomskim sekvenciranjem smo vključili 48 bolnikih z družinsko obliko MS. V nabor smo vključili bolnike iz 44 različnih družin - izbrane primere družin s posebej izrazitim kopičenjem primerov MS pa prikazujemo na sliki 18.



Slika 18: Primeri nekaterih zajetih z družinsko obliko MS, ki smo jih vključili v študijo s sekvenciranjem celotnega humanega eksoma.
Figure 18: Presentation of selected examples of MS families included in the exome sequencing study.

4.4.2 Različice odkrite z eksomskim sekvenciranjem

Z eksomskim sekvenciranjem smo v populaciji vključenih preiskovancev ugotovili skupno 205.574 genomskih različic, ki so dosegle zadostne kriterije kakovosti. Potem ko smo izločili podatke za 4 kontrolne primere, kjer je bilo število zanesljivih genotipov manjše od 20.000, smo v preiskani množici obdržali 48 bolnikov z družinsko obliko MS, 40 bolnikov s sporadično obliko MS in 88 kontrolnih primerov, pri katerih smo v povprečju imeli zanesljive podatke za 166.844 genomskih mest z različicami.

Ker je bil poglavitni cilj študije odkrivanje redkih različic z morebitnim vplivom na razvoj MS, smo v prvem koraku odfiltrirali različice s frekvenco preko 5 % v globalni populaciji 60.000 preiskovancev projekta ExAC, po čemer je v izboru preostalo še 143.070 različic. Med omenjenimi različicami so prevladovale različice z zamenjavo aminokislinskega zaporedja (56.895 različic, 39.7 %), sinonimne različice (31.109 različic, 21,7 %) in intronske različice (28.744, 20,0 %), v manjšem deležu pa smo ugotovili prisotnost tudi nekaterih visoko patogenih različic - 898 (0,64 %) različic s prezgodnjo vstavitvijo stop kodona, 792 (0,55 %) različic s premikom bralnega okvirja in 798 (0,56 %) različic z verjetnim vplivom na izrezovanje intronov. Različice so bile v globokih intronskih, drugih nekodirajočih ali intergenskih področjih smo zaradi slabše kakovosti v podatkih eksomskega sekvenciranja izločili iz nadaljnjih analiz.

4.4.3 Uporaba algoritma za integracijo omskih študij pri interpretaciji rezultatov eksomskega sekvenciranja pri MS

V naslednjem koraku smo za interpretacijo ugotovljenih različic uporabili rezultate integrativne analize. Različice, ki so se nahajale v najvišje uvrščenih regijah po integrativni omski analizi (PFP<0.001) smo preverili in izbrali različice, ki smo jih našli le pri bolnikih. Iskanje smo usmerili na različice za katere je bila napoved njihove patogenosti bodisi srednja (napovedano patogene različice z zamenjavo aminokislinskega zaporedja) ali visoka (različice s prezgodnjo vstavitvijo stop kodona, različice s premikom bralnega okvirja in različice z verjetnim vplivom na izrezovanje intronov).

Po filtraciji smo pri bolnikih z družinsko MS odkrili 7 različic z možnim ali verjetnim patogenim učinkom (Preglednica 6), pri bolnikih s sporadično obliko MS pa skupno 11 možno ali verjetno patogenih različic (Preglednica 7). Redke, visoko penetrantne različice smo identificirali pri 8 od 48 primerov (16.6 %) z družinsko obliko MS in pri 9 od 40 primerov (22.5 %) s sporadično obliko MS.

Napovedano patogene različice v nekaterih genih smo odkrili pri več neodvisnih bolnikih z MS. Pri enem bolniku z družinsko in drugem s sporadično obliko MS pa smo identificirali redke, napovedano patogene različice v genu ALPK2. V študiji tudi prvič opisujemo prisotnost izjemno redke visoko patogene različice s premikom bralnega okvirja v področju gena IL7R pri bolniku z MS.

Pri dveh bolnikih z družinsko MS (Glu438Gln) in bolniku s sporadično obliko MS (Glu438Gln) smo identificirali redke, napovedano patogene različice v genu AGAP2. Področje gena AGAP2 smo identificirali zaradi kopičenja heterogenih omskih signalov v širšem področju na kromosomu 12q13.3-12q14.1 ($P<1.10^{-5}$). Na nivoju posameznega gena je bilo prekrivanje bioloških signalov minimalno, v področju 100kb regije pa smo zaznali izrazito zbiranje signalov iz GWAS študij, študij transkriptoma v možganih in krvi, v področju pa se nahajajo tudi tarčni geni miRNA s spremenjenim izražanjem pri MS.

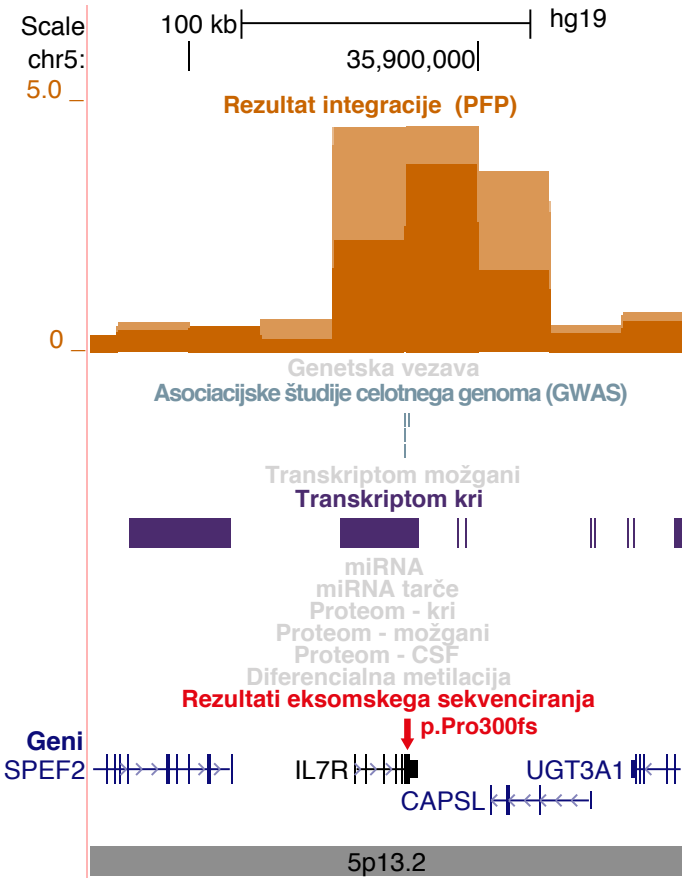
Preglednica 6: Seznam redkih visoko patogenih različic pri družinskih primerih z MS v genih, odkritih z integrativnim pristopom
(NP - ni prisotna, AK - aminokislina).
Table 6: The list of rare, likely or possibly pathogenic sequence variants in familial cases with MS. The variants are located in genes, identified using positional integrative approach
(NP - no present, AK - aminoacid).

Primer z družinsko obliko MS	Gen	Različica	Tip različice	Napovedana patogenost različice	Frekvenca v evropski populaciji (ExAC)
MS35, MSR002	AGAP2	p.Met446Thr	Zamenjava AK	SREDNJA	0,377 %
MS32	WNT9B	p.Arg222His	Zamenjava AK	SREDNJA	0,135 %
RD011	ALPK2	p.Tyr1864*	Vstavitev stop kodona	VISOKA	NP
MS23	CLEC16A	p.Asp877Asn	Zamenjava AK	SREDNJA	NP
MS36	MMEL1	p.Gln178*	Vstavitev stop kodona	VISOKA	NP
MS53	RREB1	p.Gly1627Arg	Zamenjava AK	SREDNJA	0,001 %
MS20	SORBS2	p.Arg310Ser	Zamenjava AK	SREDNJA	NP

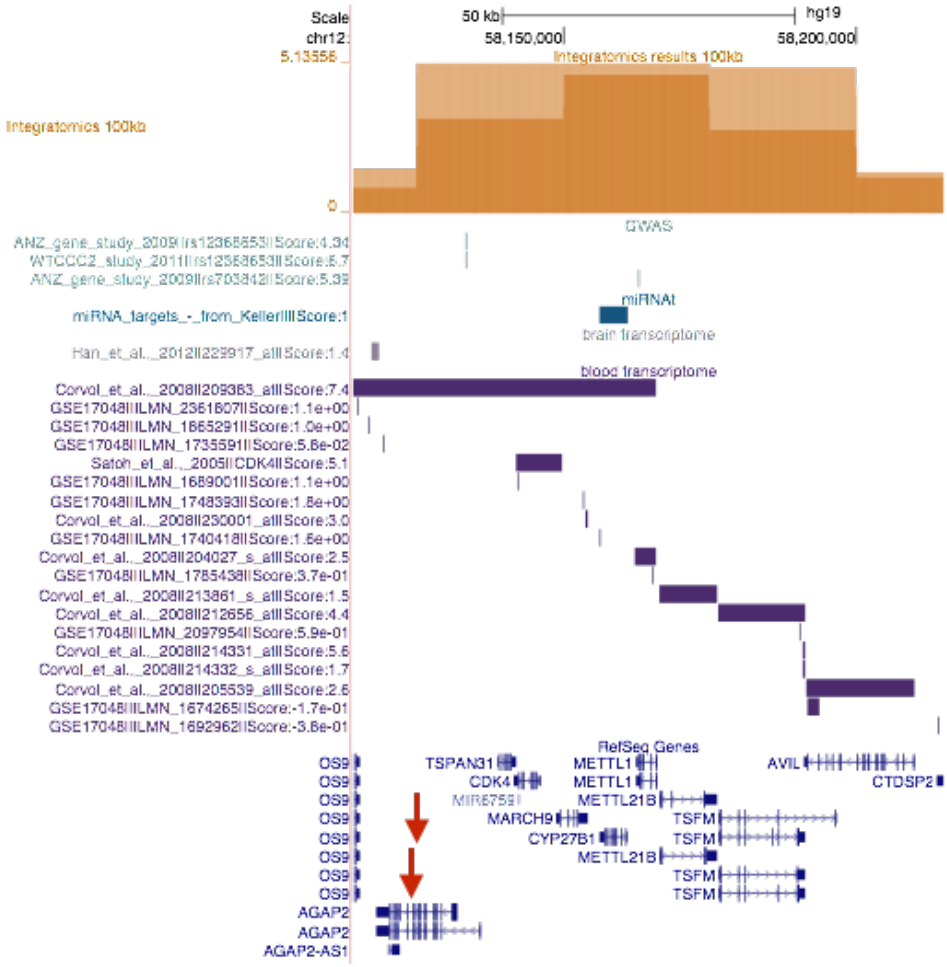
Preglednica 7: Seznam redkih visoko patogenih različic pri sporadičnih primerih z MS v genih, odkritih z integrativnim pristopom
(NP - ni prisotna, AK - aminokislina)
Table 7: The list of rare, likely or possibly pathogenic sequence variants in sporadic cases with MS. The variants are located in genes, identified using positional integrative approach
(NP - no present, AK - aminoacid).

Primer s sporadično obliko MS	Gen	Različica	Tip različice	Napovedana patogenost različice	Frekvenca v evropski populaciji (ExAC)
MSS020	AGAP2	p.Glu438Gln	Zamenjava AK	SREDNJA	0,389 %
MSS039	AHI1	p.Ser1011*	Vstavitev stop kodona	VISOKA	0,007 %
MSS003	ALPK2	p.Arg1913Cys	Zamenjava AK	SREDNJA	1,051 %
MSS010	DIAPH1	p.Arg340Ser	Zamenjava AK	SREDNJA	0,004 %
MSS030	IL7R	p.Pro300fs	Premik bralnega okvirja	VISOKA	0,003 %
MSS030	KCNMA1	p.Asn1181Ser	Zamenjava AK	SREDNJA	0,000 %
MSS027	KCNMA1	p.Asp789Glu	Zamenjava AK	SREDNJA	0,003 %
MSS029	MERTK	p.Glu823Gln	Zamenjava AK	SREDNJA	0,164 %
MSS004	MMEL1	p.Glu323Gln	Zamenjava AK	SREDNJA	0,267 %
MSS032	MYNN	p.Pro578Ser	Zamenjava AK	SREDNJA	NP
MSS029, MSS037	PLCL2	p.Gly400Ala	Zamenjava AK	SREDNJA	NP

Na Sliki 19 shematsko prikazujemo sintezo podatkov omskih študij in rezultata eksomskega sekvenciranja v področju gena IL7R, na podlagi katerega smo identificirali izredno redko visoko patogeno različico s premikom bralnega okvirja v genu IL7R pri bolniku z MS.



Slika 19: Prikaz rezultatov integracije za področje gena IL7R.
Identifikacija gena IL7R temelji na podatkih GWAS študij in transkriptomskih študij. V genu IL7R smo ob pomembnem zbiranjem podatkov identificirali tudi visoko patogeno različico s premikom bralnega okvirja v genu IL7R (p.Pro300fs).
Figure 19: Results of integration analysis for IL7R gene region.
Data points supporting high score for IL7R are presented and are based on data from GWAS studies and transcriptome studies. In addition to accumulation o scores from various omic studies, we have also identified a likely pathogenic IL7R sequence variant using whole exome sequencing (p.Pro300fs).



Slika 20: Prikaz rezultatov integracije za področje gena AGAP2.
V področju 12q13.3-12q14.1 smo s pristopom pozicijske integracije identificirali pomembno zbiranje signalov v področju genov AGAP2-CDK4-CYP27B1. V genu AGAP2 smo nadalje z eksomskim sekvenciranjem identificirali redke, možno patogene različice pri dveh bolnikih z družinsko MS in pri bolniku s sporadično obliko MS (rdeče puščice).
Figure 20: Results of integration analysis for AGAP2 gene region.
In the 12q13.3-12q14.1 region we identified increased accumulation of biological signals (in AGAP2-CDK4-CYP27B1 gene region). Combining this information with data from whole exome sequencing, we identified presence of rare, possibly pathogenic variants in AGAP2 gene in two patients with familial MS and in one patient with sporadic MS (red arrows).

5 RAZPRAVA

Tekom raziskave smo z namenom izboljšanja odkrivanja vzročnih dejavnikov za multifaktorske bolezni, razvili izviren algoritem pozicijske integracije podatkov heterogenih omskih študij. Z algoritmom smo implementirali metodo za odkrivanje omskih sprememb, ki se konsistentno odražajo na različnih bioloških nivojih in s tem poskušali izboljšati odkrivanje biološko pomembnih sprememb pri preiskovanih boleznih in jih uspešneje ločiti od statističnega šuma pri tovrstnih študijah. Pristop smo preverili in razvili na primeru multifaktorske bolezni - PB, zatem pa smo ga uporabili za integrativno analizo heterogenih genomskih podatkov na širokem naboru študij pri MS.

Pokazali smo uporabnost pristopa pri interpretaciji podatkov eksomskega in genomskega sekvenciranja, z namenom odkrivanja redkih in visoko patogenih različic pri multifaktorskih boleznih. Na modelu PB smo pokazali, da je z metodo integrativne omske analize mogoče prioritizirati gene, ki vsebujejo redke in visoko patogene različice za oblike multifaktorskih bolezni, ki nastanejo zaradi visoko patogene okvare v posameznih genih. Nadalje smo uporabili integracijo heterogenih omskih podatkov za prioritizacijo različic odkritih z eksomskim sekvenciranjem pri bolnikih z družinsko obliko MS in pri bolnikih s sporadično obliko MS v primerjavi z zdravimi kontrolami.

5.1 RAZVOJ ALGORITMA ZA INTEGRACIJO HETEROGENIH OMSKIH PODATKOV

Tehnološki napredek v molekularni biologiji in genetiki je v zadnjem času omogočil celostno molekularno preiskovanje sprememb pri boleznih človeka in bistveno pospešil razumevanje njihove etiologije in patogeneze, posebej v primeru multifaktorskih bolezni. Kljub številnim prednostim imajo omenjeni globalni oziroma omski pristopi pomanjkljivosti, ki izvirajo iz visoko dimenzionalne narave rezultatov tovrstnih analiz. Pri tovrstnih pristopih preiskujemo veliko število hipotez, najpogosteje na relativno majhni množici bioloških replikatov. To pomeni, da pri statistični analizi pa opravimo veliko število preskusov statističnih hipotez, kar ima za posledico povečano število lažno pozitivnih rezultatov. Zaradi tega je ponovljivost tako pridobljenih rezultatov relativno nizka, tako s tehnične, še posebej pa z biološke plati (Khan in sod., 1999; Kim in Park, 2004).

Z namenom odkrivanja novih vzročnih dejavnikov in morebitnih biomarkerjev, so bile pri večih multifaktorskih boleznih že opravljene številne omske študije, ki pa zaradi kompleksnih in heterogenih sprememb z majhnim učinkom pri multifaktorskih boleznih, niso podale enoznačnega rezultata (McCarthy in sod., 2008).

Ob navedenem smo postavili hipotezo, da bi lahko s postopkom združevanja več študij na istem biološkem nivoju in z združevanjem informacij o spremembah na različnih bioloških nivojih, lahko pomembno izboljšali identifikacijo bioloških sprememb in jih tako bolje ločili od tehničnega, statističnega in biološkega šuma (Maver in sod. 2009, Maver in Peterlin, 2011, Peterlin in Maver 2012). Glede na to, da so biološki nivoji med seboj povezani, v najočitnejšem primeru genom, transkriptom in proteom v centralni dogmi molekularne genetike, smo pričakovali, da se bodo biološke spremembe odražale konsistentno na več različnih bioloških nivojih, čeprav na posameznem biološkem nivoju z diskretnimi oziroma majhnimi učinki. Kot primer - različica posameznega nukleotida v promotorju gena vpliva na spremenjene nivoje izražanja bližnjega gena, kar se kasneje odrazi na spremenjenih nivojih izražanja beljakovinskega produkta gena v centralnem živčevju. Take spremembe lahko v blagi obliki identificiramo v asociacijskih študijah celotnega genoma, transkriptomskih študijah v tkivu CŽS in proteomskih študijah CŽS, vendar do izraza pride taka najdba šele, ko prekrijemo in hkrati preiskujemo različne biološke nivoje. Pri tem je pomembno, da se pri omskih študijah pogosto upošteva arbitrarna, umetno postavljena mera za statistično pomembnost odkritih sprememb, zato se spremembe z majhnim učinkom in mejno statistično povezanostjo pogosto izključijo iz nadaljnjih analiz kot statistični šum, čeprav so lahko odraz biološke spremembe pri bolezni. Ta problematika je posebej izrazita v primeru etiopatogenetsko zelo heterogenih bolezni, ki se predstavljajo z isto klinično sliko (kot fenokopije) in se zato pomembni učinki zaradi heterogenosti izgubijo. Pričakujemo, da ima delež takih najdb - čeprav z učinkom srednje ali majhne velikosti, lahko pomembno vlogo pri etiopatogenezi preiskovanih bolezni in tako biološko pomembne informacije, posebej v primeru, da spremembo ugotovimo hkrati na različnih bioloških nivojih. Pričakujemo, da integrativna analiza tako lahko izboljša odkrivanje vzročnih, predvsem dednih, dejavnikov, kot tudi morebitnih biomarkerjev za multifaktorske bolezni (Maver in sod. 2013).

Integracijo podatkov iz različnih omskih virov so predhodno že implementirali za odkrivanje genov s konsisteno visokimi rezultati na različnih bioloških nivojih (Aerts in sod., 2006; Sun in sod., 2009). Opisani poskusi integracije so uporabljali predvsem integracijo z uporabo gena kot skupnega imenovalca za združevanje informacij - v tem primeru so vse podatke prevedli na nivo imen genov oziroma njihovih dostopnih številk in nato združili informacije iz različnih študij oziroma omskih nivojev. Tak pristop ima številne pomanjkljivosti, posebej glede na kompleksnost podatkov, ki jih pridobivamo v zadnjem času.

Prva pomanjkljivost združevanje podatkov na nivoju anotacij genov izhaja iz izrazite neenotnosti v poročanju sprememb, identificiranih v različnih tipih omskih študij. Informacije iz omenjenih študij, posebej pred letom 2010, so poročane z raznolikimi in

nestandardnimi anotacijami - v nekaterih študijah je mogoče pridobiti podatke o spremembah na nivoju genov, v drugih o spremembah na nivoju sond z mikročipov ali na nivoju transkriptov oziroma drugih anotacij. Pretvorba anotacij na nivo gena je pogosto nepopolna in predstavlja izgubo informacij. Problem je bil predhodno že identificiran na nivoju integracije študij istega tipa, pri opravljanju meta-analiz podatkov transkriptomskih študij (Cahan in sod., 2007). Pri preliminarni oceni vhodnih podatkov smo ugotovili, da 306 (7.5 %) sond na platformi Affymetrix s spremenjenim izražanjem v centralnem živčevju pri bolnikih s PB ni bilo mogoče pretvoriti na nivo anotacij genov HGNC, ko smo uporabili orodje BioMart. Problematika je še večja pri pretvorbi genomskih različic v povezavi s PB, kjer smo uspeli pretvoriti le 50.6 % polimorfizmov enega nukleotida (SNP) v anotacije HGNC. Pričakujemo, da vključevanje vse bolj raznolikega nabora študij in bioloških nivojev verjetno pomeni čedalje večjo izgubo informacij zaradi problematike pretvorbe anotacij.

Drugi problem integracije, vezane na nivo genov, izhaja iz dejstva, da pomemben del omskih sprememb ni omejen na področja genov. Številni tipi bioloških sprememb se nahajajo in vplivajo na dovzetnost za bolezni izven področij genov in znanih je več primerov ko različice dednega materiala več kilobaznih parov stran od vzročnega gena vplivajo na funkcijo ali izražanje in s tem na bolezenski fenotip (Kleinjan in van Heyningen, 2005). Opisani so primeri, ko so vzročne spremembe ugotovili tudi en megabazni par stran od gena, ki je neposredno povezan z boleznijo (Kimura-Yoshida in sod., 2004; Kleinjan in sod., 2001; Lettice in sod., 2002; Pfeifer in sod., 1999). Pristopi integracije, ki temeljijo na združevanju podatkov na nivoju genov, tovrstne medsebojne učinke lahko prezrejo. Ta problem ilustrirajo primeri, ko funkcijski polimorfizem ni v neposredni bližini vzročnega gena in je v bazah zaradi bližine pripisan sosednjemu genu, ki ni povezan z boleznijo. Ker taka informacija lahko pomembno prispeva h kompleksnosti integrativne analize v multifaktorskih bolezni, predstavlja ta omejitev pomemben dejavnik pri klasični pristopih,vezanih na genske anotacije (Steidl in sod., 2007).

Tretja omejitev pristopa integracije na nivoju genov, pa je v dejstvu, da obstajajo številni tipi podatkov, ki zajemajo večje regije z več geni ali pa se znotraj posameznega gena izrazitost sprememb različno obnaša. Takšen tip študij so na primer študije genetske vezave, kjer z dovzetnostjo za bolezen niso neposredno povezani geni, temveč regije, kot je bila v primeru PB študija Foltynie et al. (Foltynie in sod., 2005). Rezultati študij genetske vezave so daljši intervali, ki jih ni mogoče enostavno povezati ali reducirati na nivo genskih anotacij, zato je tudi tukaj pretvorba na skupni imenovalec gena problematična. Podobno problematiko opazujemo tudi pri študijah razlik v metilaciji, kjer so spremembe glede na položaj v genomu zvezni in niso jasno zamejene na področje

posameznega gena in kot take tudi predstavljajo oviro pri integraciji (Urdinguio in sod., 2009).

Na koncu je potrebno poudariti tudi pomen nekodirajočih sprememb pri multifaktorskih bolenznih. V številnih primerih so bili dedni dejavniki za bolezni človeka odkriti globoko v področjih brez genov, kot je bilo na pokazano na primeru multifaktorske bolezni - nesindromskega razcepa neba in ustnice (Birnbaum in sod., 2009). Poleg tega pa je v zadnjem času vse bolj v ospredju pomen nekodirajočih, regulatornih mest, ki se lahko nahajajo tudi izven genskih anotacij. Trenutno je s pristopi za integracijo, omejenimi na genske anotacije, takšne nekodirajoče spremembe nemogoče preiskovati. Kljub temu, da je trenutno sekvenciranje celotnih genomov pri velikih skupinah preiskovancev cenovno še nedostopno, v prihodnosti pričakujemo porast študij in podatkov sekvenciranja celotnih genomov človeka tudi na področju raziskav multifaktorskih bolezni. Zato je ključnega pomena predvideti pristope, ki bodo omogočali interpretacijo podatkov tudi v nekodirajočih regijah človeškega genoma.

Za rešitev predstavljene problematike, smo razvili pristop za integracijo heterogenih omskih podatkov na osnovi genomskega položaja sprememb, identificiranih v omskih študijah. Probleme s pretvarjanjem anotacij na nivo genskih anotacij smo rešili z neposrednim pretvarjanjem sprememb v ustrezne položaje na referenčnem genomu človeka. V primerih ko pretvorba v položaje na genomu neposredno ni mogoče, je vedno mogoče uporabiti neposredno iskanje zaporedij sond proizvajalca mikromrež z algoritmom za identifikacijo položaja spremembe v človeškem genomu. S pozicijskim pristopom smo v odvisnosti od velikosti definiranih regij poleg tega zajeli podatke o morebitni genetskih interakcijah bližnjih sprememb in zaznali tudi primere, ko so se spremembe zbirale v ločenih, a bližnjih genih. Pri implementaciji pristopa za analizo pri modelnima boleznima PB in MS, smo ugotovili, da je pristop uporaben za širok nabor tipov omskih študij, saj smo neposredno vključevali študije o spremembah nivojev miRNA, položajev diferencialno metiliranih CpG otočkov in variacij v številu kopij specifičnih področij genoma.

Ker so bile distribucije vhodnih podatkov zelo raznolike in smo pri izboru podatkov občasno pridobili le delež rezultatov, direktna združitev podatkov različnih študij in različnih tipov študij ni bila mogoča. Da bi podatke lahko primerljivo obravnavali in izračunali mero za nenaključno zbiranje heterogenih signalov v področjih genoma, smo uporabili ne-parametričen pristop s statistiko uvrstitvnega produkta, kjer smo primerjali dejansko uvrstitev regije na posameznih bioloških nivojih s pričakovano razporeditvijo signalov po permutacijah izvornega nabora podatkov. Končna uvrstitev regije je tako hkrati odražala vsebnost signalov iz različnih tipov bioloških študij, kot tudi izrazitost

sprememb na posameznem biološkem nivoju. Regije, ki so vsebovale izrazite spremembe na večih bioloških nivojih, so bile po integraciji uvrščene najvišje. V primeru vključitve več študij istega biološkega tipa smo vključili tudi rešitev za združevanje več študij na posameznem biološkem nivoju, kjer smo najprej združili študije istega tipa (na primer različne asociacijske študije celotnega genoma) in šele nato združili študije različnega tipa.

Kljub predstavljenim prednostim, ima pristop pozicijske integracije tudi slabosti. Izbor velikosti fiksnih regij v nekaterih primerih ni enoznačen in vnaprej ni mogoče napovedati, katera velikost regij je najbolj primerna za identifikacijo kandidatnih genov. Izbor premajhne regije lahko pomeni izgubo interakcij med oddaljenimi signali, izbor prevelikih regij pa lahko poveča delež lažno pozitivnih najdb. V kolikor je cilj integracije identifikacija povprečno enega gena na regijo, je smiselno izbrati velikost regij 10 kb, saj pri tej velikosti regije vsebujejo povprečno 0.5 gena na regijo, kar zmanjša potrebo po preverjanju zbiranja signalov v odkritih regijah. Po drugi strani pa uporaba tako malih regij lahko pomeni da zgrešimo pomembne interakcije med signali, ki so oddaljeni več kot 10 kb. Pri uporabi algoritma smo ugotovili, da je smiselno rezultate integrativne analize za izbrane, zanimive regije preveriti tudi na nivoju izvornih podatkov in oceniti, kateri gen v regiji predstavlja najzanimivejšega kandidata za nadaljnje študije.

5.2 EVALVACIJA INTEGRATIVNEGA PRISTOPA NA MODELNI MULTIFAKTORSKI BOLEZNI - PARKINSONOVI BOLEZNI

PB smo uporabili kot model multifaktorske bolezni in s pristopom pozicijske integracije preiskali nabor 15 omskih študij oziroma virov podatkov na 6 različnih bioloških nivojih. Identificirali smo 179 regij s pomembno povečanim zbiranjem signalov v vključenih študijah.

Uspešnost pristopa smo ocenili z analizo genov, ki smo jih identificirali z omenjenim pristopom. Preiskali smo obseg direktnih in indirektnih povezav identificiranih genov s PB v literaturi. Glede na podatke v literaturi, je bilo 51.7 % genov v najviše uvrščenih regijah v 51.1 % že preiskovanih ali povezanih s PB. Med temi geni so bili uvrščeni tudi nekateri najbolj konsistentno povezani geni z dovzetnostjo za PB, kot sta *UCHL1* in *SNCA*. Preostali geni so bili v pomembnem deležu povezani s PB preko posrednih povezav v literaturi, ali pa so bili vpleteni v patogenetske poti pri PB.

Obogativene analize s termini ontologije genov (GO) so pokazale, da so bili identificirani geni funkcijsko povezani s procesi, ki so bili že vpleteni v nastanek PB in vključujejo mitohondrijsko disfunkcijo, ubikvitinacijo in metabolizem lipidov in sterolov. Ko smo

velikost preiskanih regij zmanjšali na 10 kb, se je obogatenost teh terminov še dodatno povečala.

S pristopom smo lahko identificirali tudi gene v najvišje uvrščenih regijah, ki pa v preteklosti še niso bili povezani s PB. Na primer, gen *YWHAE* je bil sicer v času, ko smo opravili integrativno analizo, le v eni študiji posredno opisan v povezavi s PB, čeprav empirični dokazi iz omskih študij prepričljivo govorijo v prid njegovega pomena pri PB (spremememnjen ekspresijski profil v CŽS, spremenjeni nivoji beljakovinskega produkta, lokacija v regiji v vezavnem neravnovesju pri PB, v njegov prid pa govori tudi fenotipska sorodnost gena s klinično sliko pri PB). Glede na podatke baze GO, je bil gen povezan s funkcijami nevronske migracije, razvoja CŽS, intracelularne signalizacije in regulacije apoptotskih procesov, kar ustreza trenutnemu videnju patogeneze PB (Gandhi in Wood, 2005). Za proteinski produkt gena *YWHAE* je bilo pokazano, da je vključen v interakcijo s proteinom Parkin, ki je E3 ubikvitin-protein ligaza, mutacije v genu, ki ga kodira, pa so bile v recesivni obliki povezane s PB z zgodnjim nastopom (Davison in sod., 2009).

Poleg predstavljenih, so imeli tudi drugi identificirani geni, funkcije, ki ustrezajo znanim etiopatogenetskim procesom pri PB: sinaptični prenos med nevroni, transport mitohondrijev ob mikrotubulih, razvoj CŽS metabolizem amiloidnih beljakovin in v procese mielinacije, kar ugotovljene gene uvršča med možne nove kandidatne gene za PB.

5.3 UPORABA RAZVITEGA PRISTOPA ZA INTEGRACIJO OMSKIH ŠTUDIJ PRI MULTIPLI SKLEROZI

Za primer MS smo zbrali podatke 52 različnih študij, opravljenih na 12 različnih omskih nivojih in opravili integrativno analizo. Z analizo smo identificirali 188 regij velikosti 100 kb, kjer je prišlo do pomembnega zbiranja sprememb iz različnih omskih študij. V odkritih regijah, se je nahajalo 435 genov, funkcijske študije pa so pokazale da pri teh genih dominirajo biološke funkcije, v povezavi z vnetnimi in avtoimunimi procesi ter prezentacijo antigenov T celicam.

Izrazito stopnjo zbiranja signalov, za kar 38.3 % najvišje uvrščenih regij, smo ugotovili v področju kompleksa genov za humane levkocitne antigene HLA na kromosomu 6p21, kar ustreza dosedaj znanim študijam, ki kažejo na osrednji pomen regije HLA pri dovzetnosti za MS (Lincoln in sod., 2005). Ugotovili smo tudi, da se zbiranje sprememb omskih študij pojavlja v celotnem razponu regije HLA, kar je ravno tako v skladu s študijami, ki kažejo, da je dovzetnost, za MS povezana s predispozicijo v HLA kompleksu in odvisna od številnih različic v večih predelih te genomske regije (Patsopoulos in sod., 2013). Izrazito povečano zbiranje signalov smo med HLA signali ugotovili v področju gena *MOG*, ki nosi

zapis za Mielinski oligodendrocitni glikoprotein (OMIM:159465, OMIM, 2016). Ugotovili smo, da se v področju gena nahaja polimorfizem, ki so ga v dveh ločenih GWAS študijah (Baranzini in sod., 2009; Hafler in sod., 2007) povezali z multiplo sklerozo, njegovo spremenjeno izražanje pa so ugotovili tako v tkivu centralnega živčevja (Han in sod., 2012), kot tudi v periferni krvi (Gandhi in sod., 2010). S provokacijskimi študijami pri živalskih modelih so pokazali, da avtoimunost na MOG oponaša spekter nevropatoloških sprememb pri MS (Storch in sod., 1998), pokazali pa so tudi, da je MOG osrednji mielinski glikoprotein na katerega se odzivajo periferni monociti v krvi (Kerlero de Rosbo in sod., 1993).

V področju genov izven HLA regij smo zaznali izrazito zbiranje signalov v področju naslednjih genov: *TNFSF14*, *ZFP36L1*, *PTPN6* in v področju nabora genov *AGAP2-TSPAN31-CDK4*.

Gen *TNFSF14* predstavlja kandidatni gen, ki je bil šele nedavno identificiran kot pomemben pri dovzetnosti za MS. Nedavne študije so pokazale da predstavlja enega ključnih regulatorjev preživetja CD4+ T-spominskih celic (Soroosh in sod., 2011), povečano izražanje pa vodi do avtoimunosti pri mišjih modelih (Shaikh in sod., 2001). Nabor rezultatov omskih študij, ki podpira visoko uvrstitev gena, zajema tako rezultate GWAS študij, (Sawcer in sod., 2011,), transkriptomskih študij v krvi pri bolnikih z MS (Gandhi in sod., 2010), proteomskih študij (Liu in sod., 2009; Noben in sod., 2006), mišjih modelov (Whitney in sod., 2001); gen pa je tudi podvržen regulatorni kontroli miRNA, ki so pri MS spremenjeno izražene (Jernas in sod., 2013).

Gen *ZFP36L1* predstavlja enega novejših genov v povezavi z MS in nosi zapis za EGF-odzivni dejavnik oziroma gen zgodnjega odziva B-celic, v zadnjem času pa ga povezujejo s protivnetnim delovanjem (Perga in sod., 2015). Integracija je pokazala konsistentno prisotnost sprememb v vseh vključenih trankriptomskih študijah (možgani, kri in CD4+ celice pri bolnikih), prav tako pa je bil identificiran tudi v največji GWAS študiji doslej, poleg tega je gen podvržen tudi regulatorni kontroli na nivoju miRNA, ki so pri MS spremenjeno izražene. Posebnost tega gena pri integrativni analizi je bila v tem, da so bile spremembe na posameznem nivoju relativno diskretne, vendar je bil gen visoko uvrščen prav zaradi visoke konsistence rezultatov med primerljivimi študijami in tudi na različnih omskih nivojih, kar je razvidno iz preglednice 5, v sekciji z Rezultati.

Podobno smo ugotovili tudi visoko uvrščenost genske regije, ki vsebuje zapis gena *PTPN6*, ki je povezan s kronično avtoimunostjo in vnetjem, ki je podobno tistemu pri nevtrofilnih dermatozah (Hendriks in Pulido, 2013).

Identificirali smo tudi povečano zbiranje signalov v širšem področju regije na kromosomu 12q13.3-12q14.1, ki vsebuje več genov, ki so bili predhodno že povezani z MS (*AGAP2*, *CDK4*, *CYP27B1*). V tem primeru smo ugotovili zbiranje signalov tako GWAS študij kot transkriptomskih študij v več genih hkrati. Pregled literature je pokazal da so bile spremembe v različnih genih tega področja predhodno že povezane z različnimi imunskimi boleznimi (vključno s sladkorno boleznijo tipa 1, revmatoidnim artritisom, celiakijo in tudi z multiplo sklerozo). V letu 2013 so Alcina in sod. pokazali da so ugotovljene spremembe posledica deregulacije širokega področja 12q13.3-12q14.1 zaradi različice v intergenskem ojačevalnem zaporedju tega področja (Alcina in sod., 2013). Primer tega področja kaže na pomen pozicijske integracije sprememb iz omskih študij, saj bi predstavljeno regijo z integracijo na nivoju posameznih genov zaradi fragmentiranosti signalov v tem področju verjetno izločili kot manj pomembno pri MS.

5.3.1 Uporaba pristopa integrativne omike za interpretacijo podatkov eksomskega sekvenciranja

Nove možnosti, ki jih je prinesla tehnologija sekvenciranja nove generacije, so omogočile identifikacijo redkih in visoko patogenih različic tudi pri multifaktorskih boleznih. Pri nekaterih primerih multifaktorskih bolezni so monogenske oblike in vzročni geni že poznani, za številne druge primere pa tovrstni vzročni geni še niso poznani (Peltonen in sod., 2006). Med multifaktorske bolezni, pri katerih vzročni geni še niso poznani, sodi tudi MS. Poglavitna problematika pri interpretaciji genetskih sprememb, ki jih identificiramo pri eksomskem in genomskem sekvenciranju je v tem, da odkrivamo izjemno število različic, od katerih za veliko večino nimamo točnega podatka o njihovi patogenosti (Ormond in sod., 2010). Pri multifaktorskih boleznih je problematika še izrazitejša, saj so morebitno vzročne različice s srednjim do velikim učinkom lahko v nizkem deležu prisotne tudi v splošni populaciji, kar bistveno otežuje odkrivanje patogenih različic. Poleg tega trenutno kljub številnim razvitim orodjem še nimamo zadosti natančnih algoritmov, ki bi omogočali ločevanje patogenih od nepatogenih različic v podatkih eksomskega in genomskega sekvenciranja.

Z namenom izboljšanega odkrivanja različic smo pri identifikaciji potencialno vzročnih različic pri multifaktorskih boleznih razvili strategijo, s katero interpretiramo podatke eksomskega sekvenciranja na podlagi rezultatov integracije širokega nabora omskih študij pri izbrani bolezni - v našem primeru MS. Da smo lahko omske študije neposredno uporabili v procesu interpretacije, smo podatke s pristopom pozicijske integracije združili v enoten rezultat, ki je odražal širok nabor informacij iz vseh vključenih študij. Na primeru PB smo že predhodno pokazali, da je mogoče s takim pristopom integracije identificirati kandidatne gene, ki so povezni z monogenskimi oblikami sicer multifaktorskih bolezni.

Prav tako smo pokazali, da z večanjem števila virov, na podlagi katerih prioritiziramo kandidatne gene oziroma regije, progresivno izboljšamo zmogljivost algoritma za identifikacijo z boleznijo že povezanih ali pa posredno - funkcionalno povezanih genov.

V naši študiji smo opravili eksomsko sekvenciranje pri 48 bolnikih z družinskimi oblikami MS, kot tudi pri 40 primerih s sporadičnimi oblikami MS, za primerjavo pa smo uporabili primerljivo število (92) kontrolnih primerov. Družinske primere smo v študiji v velikem deležu vključili namenoma, saj smo pričakovali, da bo verjetnost identifikacije redkih in visoko patogenih različic pri teh preiskovancih višja. Pri odkrivanju možno vzročnih različic smo se usmerili na gene, pridobljene z integrativno analizo pri MS. S sintezo omskih podatkov in podatkov eksomskega sekvenciranja smo redke, možno patogene različice identificirali pri 16,6 % bolnikov z družinsko obliko MS in pri 22,5 % bolnikov s sporadično obliko MS. Pri bolnikih z družinsko obliko MS smo v genih identificiranih z integrativnim pristopom identificirali redke visoko patogene različice v genih *AGAP2* (pri dveh bolnikih) in *MMEL1*, srednje patogene pa v genih *WNT9B*, *ALPK2*, *CLECL16A*, *RREB1* in *SORBS2*. Pri bolnikih s sporadično obliko MS smo identificirali visoko patogene redke različice v genih *IL7R* in *AHI1*, srednje patogene pa še v genih *AGAP2*, *ALPK2*, *DIAPH1*, *KCNMA1*, *MERTK*, *MMEL1*, *MYNN* in *PLCL2*. Patogene različice v nekaterih genih smo ugotovili pri več bolnikih, pri treh bolnikih v genu *AGAP2*, pri dveh bolnikih v genu *ALPK2* in pri dveh v *KCNMA1*, različice v drugih genih pa smo našli pri posameznih bolnikih.

V naši študiji prvič opisujemo prisotnost visoko patogene redke različice v genu *IL7R* pri bolniku z MS. *IL7R* je eden izmed prvih genov, ki so ga v povezavi z MS identificirali poleg lokusa HLA in predstavlja enega najbolj konsistentno povezanih genov z dovzetnostjo za MS, tako pri družinskih oblikah, kot pri sporadičnih oblikah MS (Gregory in sod., 2007). Visoko patogene različice smo nadalje identificirali tudi v genih *MMEL1*, *AHI1* in *AGAP2*, ki so bili v asociacijskih študijah povezani z dovzetnostjo za razvoj MS (Sawcer in sod., 2011). Poleg tega smo nekatere različice istega gena odkrili pri več bolnikih z MS (*AGAP2*, *KCNMA1* in *ALPK2*).

Z integracijo pristopa pozicijske integracije in eksomskega sekvenciranja smo identificirali prisotnost možno patogenih in redkih različic pri več primerih tako družinske, kot sporadične oblike MS. Kljub temu, da prekrivanje signalov na nivoju posameznega gena v področju gena AGAP2 ni pokazalo pomembnega prekrivanja rezultatov z različnih bioloških nivojev, smo identificirali statistično pomembno zbiranje signalov v področju 100 kb regije na kromosomu 12q13.3-12q14.1. Glede na rezultate pozicijske integracije smo usmerili analizo rezultatov eksomskega sekvenciranja pri bolnikih z MS tudi na gene v tem področju in identificirali prisotnost možno patogenih redkih različic v genu *AGAP2*

pri dveh bolnikih z družinsko MS in bolniku s sporadično obliko MS. V prikazanem primeru je izpostavljen pomen pozicijske integracije, saj področja gena *AGAP2* s klasičnimi pristopi integracije omskih podatkov ne bi identificirali.

V trenutni študiji smo imeli možnost preveriti uporabnost pristopa pozicijske integracije predvsem na podatkih eksomskega sekvenciranja, s katerim smo pridobili podatke o genetskih različicah v kodirajočih regijah genoma pri bolnikih z MS. Pričakovati je, da bomo v bližnji prihodnosti pridobivali vse več podatkov o genetskih različicah pri MS tudi na nivoju celotnega genoma. Pristop pozicijske integracije z integracijo vezano na genomska področja omogoča integracijo podatkov, ki ni omejena le na področja trenutno znanih genov, ampak omogoča vključitev podatkov in nepristrano analizo vseh področij genoma. Zato pričakujemo, da bo mogoče pristop uporabiti tudi za interpretacijo podatkov genomskega sekvenciranja, ki bo v bližnji prihodnosti verjetno eden poglavitnih pristopov za identifikacijo dednih dejavnikov multifaktorskih bolezni.

6 SKLEPI

- Razvili smo algoritem za sintezo heterogenih omskih podatkov, ki temelji na ugotovljanju genomskih področij z zbiranjem rezultatov iz heterogenih omskih študij (pristop *pozicijske integracije*).
- Pokazali smo, da je mogoče z uporabo pristopa pozicijske integracije združiti raznovrstne tipe bioloških podatkov, pridobljenih z omskimi študijami različnih omskih nivojev.
- Na primeru Parkinsonove bolezni (PB), ki predstavlja eno najpogostejših multifaktorsko pogojenih nevroloških bolezni, smo pokazali zmogljivost pristopa pri identifikaciji tako vzročnih genov s poznano vlogo pri PB (primer gena *SNCA* in *UCHL1*), kot tudi novih kandidatnih genov (primer gena *YWHAE*).
- Na primeru PB smo pokazali, da vključitev večjega števila omskih nivojev omogoča boljšo identifikacijo genov, ki so bodisi neposredno bodisi funkcijsko povezani s preiskovano boleznijo.
- Na primeru multiple skleroze (MS) smo pokazali, da je z uporabo algoritma pozicijske integracije mogoče združiti obsežen in zelo heterogen nabor podatkov (ki je zajemal 52 študij na 12 različnih bioloških nivojih).
- Tudi na primeru MS smo pokazali uspešnost identifikacije znanih vzročnih genov (področje HLA, *MOG*, *IL7R*) in nove genske povezave , ki bi bile sicer spregledane v šumu rezultatov omskih študij (*ZFP36L1*, *PTPN6*, *TNFSF14*, *AGAP2*).

S tem smo potrdili hipotezo: "Na podlagi izvirnega pristopa *pozicijske integracije* lahko pomembno izboljšamo učinkovitost sinteze heterogenih genomskih podatkov. S pristopom lahko uspešneje in na podlagi empirični dokazov iz omskih študij identificiramo nove gene kandidate in mehanizme pri boleznih človeka."

- Na primeru PB smo pokazali, da je z integrativnim omskim pristopom mogoče identificirati gene, ki vsebujejo redke, visoko penetrantne različice za monogenske oblike multifaktorskih bolezni (gena *SNCA* in *UCHL1*).
- Pri MS smo pokazali, da je mogoče rezultate integrativne analize omskih študij pri multifaktorskih boleznih vključiti v interpretacijo podatkov eksomskega in genomskega sekvenciranja.
- Pokazali smo, kako lahko integrativen pristop bolje usmeri analizo širokega nabora različic, ugotovljenih pri eksomskem sekvenciranju in usmeri interpretacijo različic v genih, za katere je na voljo empirična podpora iz heterogenih omskih študij.
- Na primeru integrativne analize področja gena *AGAP2* pri MS smo pokazali, da je s sintezo podatkov omskih študij in rezultatov eksomskega ali genomskega sekvenciranja mogoče uspešneje identificirati potencialno vzročne redke in visoko patogene različice pri multifaktorskih bolezni. S sintezo podatkov heterogenih

omskih študij smo uspeli pri 16.6 % bolnikov s družinsko MS in 22.5 % bolnikov s sporadično MS identificirati redke in potencialno patogene mutacije.

S tem smo potrdili hipotezo: "Interpretacijo podatkov eksomskega in genomskega sekvenciranja v kontekstu multifaktorskih bolezni lahko pomembno izboljšamo z uporabo pozicijske integracije podatkov heterogenih omskih študij."

7 POVZETEK (SUMMARY)

7.1 POVZETEK

Multifaktorske bolezni so poglavitni vzrok obolevnosti v razvitem svetu in zajemajo kardiovaskularne, onkološke, avtoimune in druge pogoste skupine bolezni. Glede na trenutno razumevanje nastanejo kot posledica skupnega prispevka številnih dednih dejavnikov, dejavnikov okolja in njihovega medsebojnega sovplivanja. Odkrivanje in razumevanje posameznih dejavnikov tveganja za te bolezni sta kljub velikim vložkom v raziskave in več desetletjem naporov, dala le borne rezultate. Pomemben napredek pri razumevanju etiologije multifaktorskih bolezni so prinesle nove metode v molekularni biologiji, vendar pa je interpretacija rezultatov teh študij zahtevna, tehnična in biološka ponovljivost rezultatov pa je pogosto omejena.

V prvem delu doktorske naloge smo z namenom izboljšanja odkrivanja novih kandidatnih genov in bolezenskih mehanizmom razvili nov algoritem za sintezo heterogenih podatkov omskih študij. Razvit pristop temelji na identifikaciji genomskih področij, kjer sovpadajo spremembe ugotovljene v večih ločenih študijah in na različnih bioloških nivojih. Predhodno razviti algoritmi za sintezo heterogenih omskih sprememb so bili omejeni na združevanje podatkov preko pretvarjanja različnih tipov bioloških sprememb v anotacije genov. V primerjavi s temi pristopi ima pristop pozicijske integracije, kjer biološke spremembe združujemo na podlagi njihovega genomskega položaja, več prednosti. Z razvitim algoritmom je mogoče podatke učinkoviteje in popolneje pretvoriti na nivo skupnega imenovalca, z uporabo pristopa zajamemo interakcije med spremembami v bližnjih genskih področjih. Poleg tega pristop omogoča vključitev sprememb, ki niso zamejene zgolj na področja posameznih genov (na primer metilacijske spremembe), možno pa je združevati tudi podatke o intergenskih spremembah (na primer genomske različice v intergenskih področjih).

Primer PB smo uporabili kot model multifaktorske bolezni na katerem smo preverili zmogljivost in prednosti razvitega opisanega pristopa. V integrativno analizo smo vključili nabor 15 omskih študij oziroma virov podatkov na 6 različnih bioloških nivojih in identificirali 179 genomskih regij s pomembno povečanim zbiranjem sprememb iz vključenih študij. Analiza genov zajetih v omenjenih regijah je pokazala, da je bilo 51.7 % genov v odkritih regijah že preiskovanih ali povezanih s PB, vključno z *UCHL1* in *SNCA*, ki sta glede na podatke v literaturi prepričljivo povezana z etiopatogenezo PB. Preostali geni so bili povezani s PB preko posrednih povezav v literaturi, ali pa so bili funkcijsko vpleteni v znane patogenetske poti pri PB.

Pri PB smo uspeli identificirati tudi nove kandidatne gene, ki predhodno še niso bili povezani s PB, kljub temu da rezultati omskih študij prepričljivo kažejo v prid njihovega pomena pri PB. Za kandidatni gen *YWHAE* smo na podlagi integrativne analize ugotovili, da empirični dokazi iz omskih študij prepričljivo govorijo v prid njegovega pomena pri PB (spremenjen ekspresijski profil v CŽS, spremenjeni nivoji beljakovinskega produktav CŽS, nahajanje v regiji v vezavnem neravnovesju pri PB, fenomska kompatibilnost s PB). Poleg tega je nadaljnja funkcijska karakterizacija *YWHAE* pokazala tudi, da je vključen v interakcijo s proteinom Parkin, ki je E3 ubikvitin-protein ligaza in ga kodira gen *PARK2*. Mutacije v slednjem genu v recesivni obliki predstavljajo znan vzrok za PB z zgodnjim nastopom.

Nadalje smo na modelu PB ugotovili, da geni pri katerih smo ugotovili prekrivanje signalov na več bioloških nivojih, kažejo izrazitejšo obogatitev patogenetskih procesov pri PB). Poleg tega smo z identifikacijo gena *SCNA* pri integrativni analizi pokazali, da je z razvitim pristopom mogoče identificirati tudi gene, ki vsebujejo redke, visoko penetrantne različice za monogenske oblike multifaktorskih bolezni.

V drugem delu raziskovalnega dela smo razvit pristop uporabili za integracijo širokega nabora 52 obstoječih omskih študij pri MS, ki so bile opravljene na 12 različnih bioloških nivojih. Pokazali smo, da je mogoče s pristopom združiti raznolik in obsežen nabor bioloških podatkov. Identificirali smo 188 genomskih regij (s 435 geni) z izrazitim kopičenjem sprememb, izmerjenih v izvornih študijah pri bolnikih z MS. Med regijami z najvišjimi vrednostmi smo v skladu s pričakovanju ugotovili regijo 6p21 s področjem humanega levkocitnega antigena (HLA). Poleg te regije pa smo ugotovili tudi pomembne kandidatne gene v številnih ne-HLA regijah, vključno z geni *TNFSF14*, *CD86*, *ZFP36L1*, in *PTPN6* ter gensko regijo, ki jo definirajo geni *AGAP2*-*TSPAN31*-*CDK4*.

Nadalje smo želeli pokazati, da je s sintezo heterogenih omskih podatkov in podatkov eksomskega sekvenciranja mogoče izboljšati odkrivanje redkih, visoko patogenih različic pri multifaktorskih boleznih. Pri skupinah bolnikov z družinsko obliko MS, s sporadično obliko MS in skupino kontrolnih preiskovancev, smo opravili sekvenciranje celotnega humanega eksoma. Interpretacijo pridobljenega nabora genetskih različic smo omejili na regije identificirane z integracijo omskih študij pri MS. S tem pristopom smo uspeli identificirati redke, visoko patogene različice pri 16,6 % bolnikov z družinsko obliko MS in pri 22,5 % bolnikov s sporadično obliko MS. Pri bolnikih z družinsko obliko MS smo ugotovili prisotnost visoko patogenih različic v genih *AGAP2* in *MMEL1*, srednje patogene pa v genih *WNT9B*, *ALPK2*, *CLECL16A*, *RREB1* in *SORBS2*. Pri bolnikih s sporadično obliko MS smo identificirali visoko patogene redke različice v genih *IL7R* in AHI1, srednje patogene pa še v genih *AGAP2*, *ALPK2*, *DIAPH1*, *KCNMA1*, *MERTK*, *MMEL1*,

MYNN in *PLCL2*. Identificirani geni predstavljajo zanimive kandidatne gene za nadaljnje, usmerjene študije za odkrivanje redkih, visoko penetrantnih različici pri MS.

V predstavljeni raziskavi predstavljamo izviren pristop k interpretaciji in razumevanju kompleksnih sprememb, ki jih ugotavljamo pri multifaktorskih boleznih. Pokazali smo, da lahko z razvitim pristopom uspešno integriramo podatke omskih študij in rezultate uporabimo za karkterizacijo znanih in identifikacijo novih genov oz. mehanizmov pri multifaktorskih boleznih. Pristop smo uporabili na obsežnem naboru omskih študij pri MS in pokazali, da je mogoče z omenjenim pristopom identificirati nove kandidatne gene, genomske regije in nove mehanizme, ki bi jih z analizo na nivoju posameznih študij zaradi tehničnega, statističnega in biološkega šuma lahko spregledali. Za namene identifikacije novih dednih dejavnikov pri MS smo opravili tudi prvo študijo z eksomskim sekvenciranjem pri družinskih in sporadičnih primerih te bolezni. Pokazali smo, da je mogoče podatke omskih študij vključiti v interpretacijo rezultatov eksomskega in genomskega sekvenciranja. S sintezo pristopa integracije omskih podatkov in rezultatov eksomskega sekvenciranja smo pri 16,6 % bolnikov z družinsko MS in pri 22,5 % bolnikov s sporadično obliko MS identificirali redke in potencialno patogene različice.

7.2 SUMMARY

Multifactorial diseases represent the leading cause of morbidity in the developed world and encompass cardiovascular diseases, cancer, autoimmune disorders and other common diseases. According to current understanding, these diseases arise due to contribution of multiple genetic and environmental factors and their interactions. Discovery and understanding of specific risk factors for these diseases has provided only partial results, despite significant investments and decades of research. An important advancement in understanding these disorders has been provided by novel technical advancements in molecular biology, however, the interpretation of results of these studies is difficult and the results are commonly not replicated in follow-up studies.

In the first part of the present dissertation, we developed a novel algorithm for synthesis of heterogeneous omic data with aim of improving the interpretation and identification of novel candidate genes or disease mechanisms. The approach we developed is based upon integration using genomic positions as the commond denominator of datasets form included studies. Previously developed approaches have based the integration step on conversion of annotations to gene-level. Positional integration offers several advantages in comparison with such gene-centric approach. Using the developed algorithm, it is possible to more completely and fully convert the annotations to a universal common denominator and at the same time the developed approach also enables detection of regional interactions in the regions studies. In addition to this, the approach enables inclusion of alterations that

are not limited to regions of specific genes (ie., methylation alterations) and it is also possible to merge data for changes that are identified in the intergenic regions (ie., regulatory SNPs)

We employed the example of Parkinson disease (PD) as a model of multifactorial disease, where we have tested performance and benefits of the developed approach. We included 15 omic studies performed on 6 biological layers and we were able to identify 179 genomic regions with increased accumulation of signals form included studuies. Analysis of genes in the included regions has shown that 51.7 % of the identified genes have previously been associated or investigated in PD, including UCHL1 and SNCA, which are one of the most convincingly associated genes with PD. The remaining genes have been, in a notable proportion, associated with PD through indirect associations in the literature, or have been implicated in the functional pathways in PD.

We were also able to identify new candidate genes for PD with some notable examples of genes that have not yet been directly associated with PD, despite the convincing body of evidence of their relevance evident in omic stuides. In the case of YWHAE candidate gene, we were able to show using integrative analysis of omic data, that empiric data consistently support its role in PD (altered brain transcriptional levels, altered protein levels in brain, location in genomic region, linked to PD and phenomic compatibility with PD). In addition, further functional characterization of the identified gene has shown that it is also interacting with Parkin protein, which is coded by a gene PARK2, associated with autosomal recessive early-onset PD.

We have further determined that with increasing overlap of signals from various biological layers in PD, functional enrichment of genes in those regions tends to converge towards processes implicated in PD. With the identification of SNCA gene in PD, we have also shown we can identify genes associated with monogenic forms of multifactorial diseases.

In the second part of our research, we employed the developed approach to integrate a comprehensive set of 52 studies in multiple sclerosis (MS) that originate from 12 different biological levels. We have shown that it is possible to integrate a heterogeneous and large body of biological data using the approach developed. Performing the integrative analysis in MS, we identified 188 genomic regions (with 435 genes) with increased accumulation of alterations, detected in original omic studies in MS. In accordance with exptectations, the identified regions covered 6p21 regon with human leukocyte antigen - HLA gene complex. We have however, also identified significant candidate genes in non-HLA regions, including the following genes: TNFSF14, CD86, ZFP36L1, PTPN6 and AGAP2-TSPAN31-CDK4 gene region.

We further aimed to show that syntesis of heterogeneous omic data with the data from exome sequencing can facilitate identification of rare, highly penetrant variants in multifactorial diseases. For this reason we performed whole exome sequencing in patients with familial MS, in patients with sporadic MS and in a group of healthy controls. The interpretaion of sequencing results was focused on the regions identified using positional integrative approach. Using this strategy, we were able to identify rare, potentially pathogenic variants in 16,6 % of patients with familial MS and in 22,5 % of patients with sporadic MS. In patients with familial MS, we identified presence of highly pathogenic variants in AGAP2 and MMEL1 genes, and presence of moderately pathogenic variants in WNT9B, ALPK2, CLECL16A, RREB1 and SORBS2 genes. In patients with sporadic MS, we identified highly pathogenic variants in IL7R and AHI1 genes, and presence of moderately pathogenic variants in AGAP2, ALPK2, DIAPH1, KCNMA1, MERTK, MMEL1, MYNN and PLCL2 genes. The identified candidate genes represent potential targets for further, focused studies aiming to detect presence of potentially high impact variants in MS.

In the present research, we developed an innovative approach to interpretation and understanding of complex alterations in multifactorial diseases. We have shown that using the developed approach, it is possible to efficiently integrate data drom omic studies and that it is possible to use the results for characterization and identification of novel genes or mechanisms in multifactorial diseases. We have utilized the approach to integrate a comprehensive body of evidence in MS and have shown that it is possible to identify novel, plausible candidate genes, regions or mechanisms that would be missed without the integrative approach due to technical, statistical or biological noise. Aiming to identify novel genetic factors in MS, we also performed the first study that utilized exome sequencing in familial and sporadic cases with MS. We have shown that omic data may be utilized in interpretation of exome sequencing data. With the synthesis of omic data with exome sequencing data we were ultimately able to identify rare, possibly pathogenic variants in 16,6 % of familial MS cases and in 22,5 % sporadic cases of MS.

8 VIRI

Abdi F., Quinn J.F., Jankovic J., McIntosh M., Leverenz J.B., Peskind E., Nixon R., Nutt J., Chung K., Zabetian C., Samii A., Lin M., Hattan S., Pan C., Wang Y., Jin J., Zhu D., Li G.J., Liu Y., Waichunas D., Montine T.J., Zhang J. 2006. Detection of biomarkers with a multiplex quantitative proteomic platform in cerebrospinal fluid of patients with neurodegenerative disorders. Journal of Alzheimer's Disease, 9, 3: 293-348

Achiron A., Grotto I., Balicer R., Magalashvili D., Feldman A., Gurevich M. 2010. Microarray analysis identifies altered regulation of nuclear receptor family members in the pre-disease state of multiple sclerosis. Neurobiology of Disease, 38, 2: 201-209

Achiron A., Gurevich M., Magalashvili D., Kishner I., Dolev M., Mandel M. 2004. Understanding autoimmune mechanisms in multiple sclerosis using gene expression microarrays: treatment effect and cytokine-related pathways. Clinical & Developmental Immunology, 11, 3-4: 299-305

Adie E.A., Adams R.R., Evans K.L., Porteous D.J., Pickard B.S. 2005. Speeding disease gene discovery by sequence based candidate prioritization. BMC Bioinformatics, 6, 55, DOI:10.1186/1471-2105-6-55: 13 str.

Aerts S., Lambrechts D., Maity S., Van Loo P., Coessens B., De Smet F., Tranchevent L.C., De Moor B., Marynen P., Hassan B., Carmeliet P., Moreau Y. 2006. Gene prioritization through genomic data fusion. Nature Biotechnology, 24, 5: 537-544

Agarwal A., Williams G.H., Fisher N.D. 2005. Genetics of human hypertension. Trends in Endocrinology and Metabolism, 16, 3: 127-133

Alcina A., Fedetz M., Fernandez O., Saiz A., Izquierdo G., Lucas M., Leyva L., Garcia-Leon J.A., Abad-Grau Mdel M., Alloza I., Antiguedad A., Garcia-Barcina M.J., Vandenbroeck K., Varade J., de la Hera B., Arroyo R., Comabella M., Montalban X., Petit-Marty N., Navarro A., Otaegui D., Olascoaga J., Blanco Y., Urcelay E., Matesanz F. 2013. Identification of a functional variant in the KIF5A-CYP27B1-METTL1-FAM119B locus associated with multiple sclerosis. Journal of Medical Genetics, 50, 1: 25-33

Altshuler D., Daly M.J., Lander E.S. 2008. Genetic mapping in human disease. Science, 322, 5903: 881-888

Ascherio A. 2013. Environmental factors in multiple sclerosis. Expert Review of Neurotherapeutics, 13, 12: 3-9

Ashburner M., Ball C.A., Blake J.A., Botstein D., Butler H., Cherry J.M., Davis A.P., Dolinski K., Dwight S.S., Eppig J.T., Harris M.A., Hill D.P., Issel-Tarver L., Kasarskis A., Lewis S., Matese J.C., Richardson J.E., Ringwald M., Rubin G.M.,

Sherlock G. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature Genetics, 25, 1: 25-29

Bahlo M., Wang J., Gibson R.A., Galwey N., Naegelin Y., Barkhof F., Radue E.W., Lindberg R.L., Uitdehaag B.M. 2009. Genome-wide association study identifies new multiple sclerosis susceptibility loci on chromosomes 12 and 20. Nature Genetics, 41, 7: 824-828

Baranzini S.E., Bernard C.C., Oksenberg J.R. 2005. Modular transcriptional activity characterizes the initiation and progression of autoimmune encephalomyelitis. Journal of Immunology, 174, 11: 7412-7422

Baranzini S.E., Wang J., Gibson R.A., Galwey N., Naegelin Y., Barkhof F., Radue E.W., Lindberg R.L., Uitdehaag B.M., Johnson M.R., Angelakopoulou A., Hall L., Richardson J.C., Prinjha R.K., Gass A., Geurts J.J., Kragt J., Sombekke M., Vrenken H., Qualley P., Lincoln R.R., Gomez R., Caillier S.J., George M.F., Mousavi H., Guerrero R., Okuda D.T., Cree B.A., Green A.J., Waubant E., Goodin D.S., Pelletier D., Matthews P.M., Hauser S.L., Kappos L., Polman C.H., Oksenberg J.R. 2009. Genome-wide association analysis of susceptibility and clinical phenotype in multiple sclerosis. Human Molecular Genetics, 18, 4: 767-778

Basso M., Giraudo S., Corpillo D., Bergamasco B., Lopiano L., Fasano M. 2004. Proteome analysis of human substantia nigra in Parkinson's disease. Proteomics, 4, 12: 3943-3952

Birnbaum S., Ludwig K.U., Reutter H., Herms S., Steffens M., Rubini M., Baluardo C., Ferrian M., Almeida de Assis N., Alblas M.A., Barth S., Freudenberg J., Lauster C., Schmidt G., Scheer M., Braumann B., Berge S.J., Reich R.H., Schiefke F., Hemprich A., Potzsch S., Steegers-Theunissen R.P., Potzsch B., Moebus S., Horsthemke B., Kramer F.J., Wienker T.F., Mossey P.A., Propping P., Cichon S., Hoffmann P., Knapp M., Nothen M.M., Mangold E. 2009. Key susceptibility locus for nonsyndromic cleft lip with or without cleft palate on chromosome 8q24. Nature Genetics, 41, 4: 473-477

Bodmer W., Bonilla C. 2008. Common and rare variants in multifactorial susceptibility to common diseases. Nature Genetics, 40, 6: 695-701

Bomprezzi R., Ringner M., Kim S., Bittner M.L., Khan J., Chen Y., Elkahloun A., Yu A., Bielekova B., Meltzer P.S., Martin R., McFarland H.F., Trent J.M. 2003. Gene expression profile in multiple sclerosis patients and healthy controls: identifying pathways relevant to disease. Human Molecular Genetics, 12, 17: 2191-2199

Boylan K. 2015. Familial Amyotrophic Lateral Sclerosis. Neurologic Clinics, 33, 4: 807-830

Brand-Schieber E., Werner P., Iacobas D.A., Iacobas S., Beelitz M., Lowery S.L., Spray D.C., Scemes E. 2005. Connexin43, the major gap junction protein of astrocytes, is

down-regulated in inflamed white matter in an animal model of multiple sclerosis. Journal of Neuroscience Research, 80, 6: 798-808

Breitling R., Armengaud P., Amtmann A., Herzyk P. 2004. Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. FEBS Letters, 573, 1-3: 83-92

Cahan P., Rovegno F., Mooney D., Newman J.C., St Laurent G., 3rd, McCaffrey T.A. 2007. Meta-analysis of microarray results: challenges, opportunities, and recommendations for standardization. Gene, 401, 1-2: 12-18

Carvill G.L., Heavin S.B., Yendle S.C., McMahon J.M., O'Roak B.J., Cook J., Khan A., Dorschner M.O., Weaver M., Calvert S., Malone S., Wallace G., Stanley T., Bye A.M., Bleasel A., Howell K.B., Kivity S., Mackay M.T., Rodriguez-Casero V., Webster R., Korczyn A., Afawi Z., Zelnick N., Lerman-Sagie T., Lev D., Moller R.S., Gill D., Andrade D.M., Freeman J.L., Sadleir L.G., Shendure J., Berkovic S.F., Scheffer I.E., Mefford H.C. 2013. Targeted resequencing in epileptic encephalopathies identifies de novo mutations in CHD2 and SYNGAP1. Nature Genetics, 45, 7: 825-830

Chen J., Bardes E.E., Aronow B.J., Jegga A.G. 2009. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. Nucleic Acids Res, 37, Web Server issue: W305-W311

Choi J., Levey A.I., Weintraub S.T., Rees H.D., Gearing M., Chin L.S., Li L. 2004. Oxidative modifications and down-regulation of ubiquitin carboxyl-terminal hydrolase L1 associated with idiopathic Parkinson's and Alzheimer's diseases. Journal of Biological Chemistry, 279, 13: 13256-13264

Cirulli E.T., Goldstein D.B. 2010. Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nature Reviews Genetics, 11, 6: 415-425

Clayton D., McKeigue P.M. 2001. Epidemiological methods for studying genes and environmental factors in complex diseases. Lancet, 358, 9290: 1356-1360

Corvol J.C., Pelletier D., Henry R.G., Caillier S.J., Wang J., Pappas D., Casazza S., Okuda D.T., Hauser S.L., Oksenberg J.R., Baranzini S.E. 2008. Abrogation of T cell quiescence characterizes patients at high risk for multiple sclerosis after the initial neurological event. Proceedings of the National Academy of Sciences of the United States of America, 105, 33: 11839-11844

Cox M.B., Cairns M.J., Gandhi K.S., Carroll A.P., Moscovis S., Stewart G.J., Broadley S., Scott R.J., Booth D.R., Lechner-Scott J., Consortium A.N.M.S.G. 2010. MicroRNAs miR-17 and miR-20a inhibit T cell activation genes and are under-expressed in MS whole blood. PloS One, 5, 8: e12132, DOI: 10.1371/journal.pone.0012132: 7 str.

Croft D., O'Kelly G., Wu G., Haw R., Gillespie M., Matthews L., Caudy M., Garapati P., Gopinath G., Jassal B., Jupe S., Kalatskaya I., Mahajan S., May B., Ndegwa N., Schmidt E., Shamovsky V., Yung C., Birney E., Hermjakob H., D'Eustachio P., Stein L. 2011. Reactome: a database of reactions, pathways and biological processes. Nucleic Acids Research, 39, Database issue: D691-D697

Crosiers D., Theuns J., Cras P., Van Broeckhoven C. 2011. Parkinson disease: insights in clinical, genetic and pathological features of monogenic disease subtypes. Journal of Chemical Neuroanatomy, 42, 2: 131-141

Cunnea P., McMahon J., O'Connell E., Mashayekhi K., Fitzgerald U., McQuaid S. 2010. Gene expression analysis of the microvascular compartment in multiple sclerosis using laser microdissected blood vessels. Acta Neuropathologica, 119, 5: 601-615

Czene K., Lichtenstein P., Hemminki K. 2002. Environmental and heritable causes of cancer among 9.6 million individuals in the Swedish Family-Cancer Database. International Journal of Cancer, 99, 2: 260-266

Davison E.J., Pennington K., Hung C.C., Peng J., Rafiq R., Ostareck-Lederer A., Ostareck D.H., Ardley H.C., Banks R.E., Robinson P.A. 2009. Proteomic analysis of increased Parkin expression and its interactants provides evidence for a role in modulation of mitochondrial function. Proteomics, 9, 18: 4284-4297

Davydov E.V., Goode D.L., Sirota M., Cooper G.M., Sidow A., Batzoglou S. 2010. Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Computational Biology, 6, 12: e1001025, DOI: 10.1371/journal.pcbi.1001025: 13 str.

De Santis G., Ferracin M., Biondani A., Caniatti L., Rosaria Tola M., Castellazzi M., Zagatti B., Battistini L., Borsellino G., Fainardi E., Gavioli R., Negrini M., Furlan R., Granieri E. 2010. Altered miRNA expression in T regulatory cells in course of multiple sclerosis. Journal of Neuroimmunology, 226, 1-2: 165-171

Do C.B., Tung J.Y., Dorfman E., Kiefer A.K., Drabant E.M., Francke U., Mountain J.L., Goldman S.M., Tanner C.M., Langston J.W., Wojcicki A., Eriksson N. 2011. Web-based genome-wide association study identifies two novel loci and a substantial genetic component for Parkinson's disease. PLoS Genetics, 7, 6: e1002141, DOI: 10.1371/journal.pgen.1002141: 14 str.

Dunikowski L.G. 2005. EMBASE and MEDLINE searches. Canadian Family Physician, 51, 1191-1191

Dyment D.A., Yee I.M., Ebers G.C., Sadovnick A.D., Canadian Collaborative Study G. 2006. Multiple sclerosis in stepsiblings: recurrence risk and ascertainment. Journal of Neurology, Neurosurgery and Psychiatry, 77, 2: 258-259

Edgar R., Domrachev M., Lash A.E. 2002. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Research, 30, 1: 207-210

Falcon S., Gentleman R. 2007. Using GOstats to test gene lists for GO term association. Bioinformatics, 23, 2: 257-258

Fenoglio C., Cantoni C., De Riz M., Ridolfi E., Cortini F., Serpente M., Villa C., Comi C., Monaco F., Mellesi L., Valzelli S., Bresolin N., Galimberti D., Scarpini E. 2011. Expression and genetic analysis of miRNAs involved in CD4+ cell activation in patients with multiple sclerosis. Neuroscience Letters, 504, 1: 9-12

Foltynie T., Hicks A., Sawcer S., Jonasdottir A., Setakis E., Maranian M., Yeo T., Lewis S., Brayne C., Stefansson K., Compston A., Gulcher J., Barker R.A. 2005. A genome wide linkage disequilibrium screen in Parkinson's disease. Journal of Neurology, 252, 5: 597-602

Fossey S.C., Vnencak-Jones C.L., Olsen N.J., Sriram S., Garrison G., Deng X., Crooke P.S., 3rd, Aune T.M. 2007. Identification of molecular biomarkers for multiple sclerosis. Journal of Molecular Diagnostics, 9, 2: 197-204

Freitag C.M. 2007. The genetics of autistic disorders and its clinical relevance: a review of the literature. Molecular Psychiatry, 12, 1: 2-22

Fung H.C., Scholz S., Matarin M., Simon-Sanchez J., Hernandez D., Britton A., Gibbs J.R., Langefeld C., Stiegert M.L., Schymick J., Okun M.S., Mandel R.J., Fernandez H.H., Foote K.D., Rodriguez R.L., Peckham E., De Vrieze F.W., Gwinn-Hardy K., Hardy J.A., Singleton A. 2006. Genome-wide genotyping in Parkinson's disease and neurologically normal controls: first stage analysis and public release of data. Lancet Neurology, 5, 11: 911-916

Gandhi K.S., McKay F.C., Cox M., Riveros C., Armstrong N., Heard R.N., Vucic S., Williams D.W., Stankovich J., Brown M., Danoy P., Stewart G.J., Broadley S., Moscato P., Lechner-Scott J., Scott R.J., Booth D.R., Consortium A.N.M.S.G. 2010. The multiple sclerosis whole blood mRNA transcriptome and genetic associations indicate dysregulation of specific T cell pathways in pathogenesis. Human Molecular Genetics, 19, 11: 2134-2143

Gandhi R., Healy B., Gholipour T., Egorova S., Musallam A., Hussain M.S., Nejad P., Patel B., Hei H., Khoury S., Quintana F., Kivisakk P., Chitnis T., Weiner H.L. 2013. Circulating microRNAs as biomarkers for disease staging in multiple sclerosis. Annals of Neurology, 73, 6: 729-740

Gandhi S., Wood N.W. 2005. Molecular pathogenesis of Parkinson's disease. Human Molecular Genetics, 14 18: 2749-2755

Gatz M., Reynolds C.A., Fratiglioni L., Johansson B., Mortimer J.A., Berg S., Fiske A., Pedersen N.L. 2006. Role of genes and environments for explaining Alzheimer disease. Archives of General Psychiatry, 63, 2: 168-174

Gentleman R.C., Carey V.J., Bates D.M., Bolstad B., Dettling M., Dudoit S., Ellis B., Gautier L., Ge Y., Gentry J., Hornik K., Hothorn T., Huber W., Iacus S., Irizarry R., Leisch F., Li C., Maechler M., Rossini A.J., Sawitzki G., Smith C., Smyth G.,

Tierney L., Yang J.Y., Zhang J. 2004. Bioconductor: open software development for computational biology and bioinformatics. Genome Biology, 5, 10: R80, DOI: DOI**:** 10.1186/gb-2004-5-10-r80: 16 str.

Grant S.F., Hakonarson H. 2008. Microarray technology and applications in the arena of genome-wide association. Clinical Chemistry, 54, 7: 1116-1124

Graumann U., Reynolds R., Steck A.J., Schaeren-Wiemers N. 2003. Molecular changes in normal appearing white matter in multiple sclerosis are characteristic of neuroprotective mechanisms against hypoxic insult. Brain Pathology, 13, 4: 554-573

Gregersen P.K., Brehrens T.W. 2003. Fine mapping the phenotype in autoimmune disease: the promise and pitfalls of DNA microarray technologies. Genes and Immunity, 4, 3: 175-176

Gregory S.G., Schmidt S., Seth P., Oksenberg J.R., Hart J., Prokop A., Caillier S.J., Ban M., Goris A., Barcellos L.F., Lincoln R., McCauley J.L., Sawcer S.J., Compston D.A., Dubois B., Hauser S.L., Garcia-Blanco M.A., Pericak-Vance M.A., Haines J.L. 2007. Interleukin 7 receptor alpha chain (IL7R) shows allelic and functional association with multiple sclerosis. Nature Genetics, 39, 9: 1083-1091

Grimm D.G., Azencott C.A., Aicheler F., Gieraths U., MacArthur D.G., Samocha K.E., Cooper D.N., Stenson P.D., Daly M.J., Smoller J.W., Duncan L.E., Borgwardt K.M. 2015. The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. Human Mutation, 36, 5: 513-23

Gusella J.F., Wexler N.S., Conneally P.M., Naylor S.L., Anderson M.A., Tanzi R.E., Watkins P.C., Ottina K., Wallace M.R., Sakaguchi A.Y. 1983. A polymorphic DNA marker genetically linked to Huntington's disease. Nature, 306, 5940: 234-238

Hafler D.A., Compston A., Sawcer S., Lander E.S., Daly M.J., De Jager P.L., de Bakker P.I., Gabriel S.B., Mirel D.B., Ivinson A.J., Pericak-Vance M.A., Gregory S.G., Rioux J.D., McCauley J.L., Haines J.L., Barcellos L.F., Cree B., Oksenberg J.R., Hauser S.L. 2007. Risk alleles for multiple sclerosis identified by a genomewide study. New England Journal of Medicine, 357, 9: 851-862

Haider S., Ballester B., Smedley D., Zhang J., Rice P., Kasprzyk A. 2009. BioMart Central Portal--unified access to biological data. Nucleic Acids Research, 37, Web Server issue: W23-W27

Haines J.L. 2003. A meta-analysis of whole genome linkage screens in multiple sclerosis. Journal of Neuroimmunology, 143, 1-2: 39-46

Hammack B.N., Fung K.Y., Hunsucker S.W., Duncan M.W., Burgoon M.P., Owens G.P., Gilden D.H. 2004. Proteomic analysis of multiple sclerosis cerebrospinal fluid. Multiple Sclerosis, 10, 3: 245-260

Han M.H., Hwang S.I., Roy D.B., Lundgren D.H., Price J.V., Ousman S.S., Fernald G.H., Gerlitz B., Robinson W.H., Baranzini S.E., Grinnell B.W., Raine C.S., Sobel R.A., Han D.K., Steinman L. 2008. Proteomic analysis of active multiple sclerosis lesions reveals therapeutic targets. Nature, 451, 7182: 1076-1081

Han M.H., Lundgren D.H., Jaiswal S., Chao M., Graham K.L., Garris C.S., Axtell R.C., Ho P.P., Lock C.B., Woodard J.I., Brownell S.E., Zoudilova M., Hunt J.F., Baranzini S.E., Butcher E.C., Raine C.S., Sobel R.A., Han D.K., Weissman I., Steinman L. 2012. Janus-like opposing roles of CD47 in autoimmune brain inflammation in humans and mice. Journal of Experimental Medicine, 209, 7: 1325-1334

Harney S.M., Vilarino-Guell C., Adamopoulos I.E., Sims A.M., Lawrence R.W., Cardon L.R., Newton J.L., Meisel C., Pointon J.J., Darke C., Athanasou N., Wordsworth B.P., Brown M.A. 2008. Fine mapping of the MHC Class III region demonstrates association of AIF1 and rheumatoid arthritis. Rheumatology (Oxford, England), 47, 12: 1761-1767

Hattersley A., Bruining J., Shield J., Njolstad P., Donaghue K.C. 2009. The diagnosis and management of monogenic diabetes in children and adolescents. Pediatric Diabetes, 10, 1: 33-42

Hendriks W.J., Pulido R. 2013. Protein tyrosine phosphatase variants in human hereditary disorders and disease susceptibilities. Biochimica et Biophysica Acta, 1832, 10: 1673-1696

Hong F., Breitling R., McEntee C.W., Wittner B.S., Nemhauser J.L., Chory J. 2006. RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis. Bioinformatics, 22, 22: 2825-2827

Hristovski D., Peterlin B., Mitchell J.A., Humphrey S.M. 2005. Using literature-based discovery to identify disease candidate genes. International Journal of Medical Informatics, 74, 2-4: 289-298

Hutz J.E., Kraja A.T., McLeod H.L., Province M.A. 2008. CANDID: a flexible method for prioritizing candidate genes for complex human traits. Genetic Epidemiology, 32, 8: 779-790

Huynh J.L., Garg P., Thin T.H., Yoo S., Dutta R., Trapp B.D., Haroutunian V., Zhu J., Donovan M.J., Sharp A.J., Casaccia P. 2014. Epigenome-wide differences in pathology-free regions of multiple sclerosis-affected brains. Nature Neuroscience, 17, 1: 121-130

Hyttinen V., Kaprio J., Kinnunen L., Koskenvuo M., Tuomilehto J. 2003. Genetic liability of type 1 diabetes and the onset age among 22,650 young Finnish twin pairs: a nationwide follow-up study. Diabetes, 52, 4: 1052-1055

Ibrahim S.M., Mix E., Bottcher T., Koczan D., Gold R., Rolfs A., Thiesen H.J. 2001. Gene expression profiling of the nervous system in murine experimental autoimmune encephalomyelitis. Brain, 124, 10: 1927-1938

Iglesias A.H., Camelo S., Hwang D., Villanueva R., Stephanopoulos G., Dangond F. 2004. Microarray detection of E2F pathway activation and other targets in multiple sclerosis peripheral blood mononuclear cells. Journal of Neuroimmunology, 150, 1-2: 163-177

International Human Genome Sequencing Consortium. 2004. Finishing the euchromatic sequence of the human genome. Nature, 431, 7011: 931-945

Ioannidis J.P., Trikalinos T.A., Khoury M.J. 2006. Implications of small effect sizes of individual genetic variants on the design and interpretation of genetic association studies of complex diseases. American Journal of Epidemiology, 164, 7: 609-614

Irizar H., Munoz-Culla M., Sepulveda L., Saenz-Cuesta M., Prada A., Castillo-Trivino T., Zamora-Lopez G., Lopez de Munain A., Olascoaga J., Otaegui D. 2014. Transcriptomic profile reveals gender-specific molecular mechanisms driving multiple sclerosis progression. PloS One, 9, 2: e90482, DOI: 10.1371/journal.pone.0090482: 18 str.

Jernas M., Malmestrom C., Axelsson M., Nookaew I., Wadenvik H., Lycke J., Olsson B. 2013. MicroRNA regulate immune pathways in T-cells in multiple sclerosis (MS). BMC Immunology, 14, 32, DOI**:**10.1186/1471-2172-14-32: 11 str.

Jin J., Hulette C., Wang Y., Zhang T., Pan C., Wadhwa R., Zhang J. 2006. Proteomic identification of a stress protein, mortalin/mthsp70/GRP75: relevance to Parkinson disease. Molecular and cellular proteomics, 5, 7: 1193-1204

Johnson A.D., O'Donnell C.J. 2009. An open access database of genome-wide association results. BMC Medical Genetics, 10, 6, DOI:10.1186/1471-2350-10-6: 17 str.

Jonas A., Thiem S., Kuhlmann T., Wagener R., Aszodi A., Nowell C., Hagemeier K., Laverick L., Perreau V., Jokubaitis V., Emery B., Kilpatrick T., Butzkueven H., Gresle M. 2014. Axonally derived matrilin-2 induces proinflammatory responses that exacerbate autoimmune neuroinflammation. Journal of Clinical Investigation, 124, 11: 5042-5056

Junker A., Krumbholz M., Eisele S., Mohan H., Augstein F., Bittner R., Lassmann H., Wekerle H., Hohlfeld R., Meinl E. 2009. MicroRNA profiling of multiple sclerosis lesions identifies modulators of the regulatory protein CD47. Brain, 132, 12: 3342-3352

Kahana E. 2000. Epidemiologic studies of multiple sclerosis: a review. Biomedicine and Pharmacotherapy, 54, 2: 100-102

Kanehisa M. 2002. The KEGG database. Novartis Foundation Symposium, 247, 91-101; discussion 101-103, 119-128, 244-152

Karolchik D., Hinrichs A.S., Furey T.S., Roskin K.M., Sugnet C.W., Haussler D., Kent W.J. 2004. The UCSC Table Browser data retrieval tool. Nucleic Acids Research, 32, Database issue: D493-D496

Keller A., Leidinger P., Lange J., Borries A., Schroers H., Scheffler M., Lenhof H.P., Ruprecht K., Meese E. 2009. Multiple sclerosis: microRNA expression profiles accurately differentiate patients with relapsing-remitting disease from healthy controls. PloS One, 4, 10: e7440, DOI:10.1371/journal.pone.0007440: 7 str.

Keller A., Leidinger P., Steinmeyer F., Stahler C., Franke A., Hemmrich-Stanisak G., Kappel A., Wright I., Dorr J., Paul F., Diem R., Tocariu-Krick B., Meder B., Backes C., Meese E., Ruprecht K. 2014. Comprehensive analysis of microRNA profiles in multiple sclerosis including next-generation sequencing. Multiple Sclerosis, 20, 3: 295-303

Kent W.J., Sugnet C.W., Furey T.S., Roskin K.M., Pringle T.H., Zahler A.M., Haussler D. 2002. The human genome browser at UCSC. Genome Research, 12, 6: 996-1006

Kerlero de Rosbo N., Milo R., Lees M.B., Burger D., Bernard C.C., Ben-Nun A. 1993. Reactivity to myelin antigens in multiple sclerosis. Peripheral blood lymphocytes respond predominantly to myelin oligodendrocyte glycoprotein. Journal of Clinical Investigation, 92, 6: 2602-2608

Khan J., Bittner M.L., Chen Y., Meltzer P.S., Trent J.M. 1999. DNA microarray technology: the anticipated impact on the study of human disease. Biochimica et Biophysica Acta, 1423, 2: M17-M28

Kim R.D., Park P.J. 2004. Improving identification of differentially expressed genes in microarray studies using information from public databases. Genome Biology, 5, 9: R70, DOI:10.1186/gb-2004-5-9-r70: 20 str.

Kimura-Yoshida C., Kitajima K., Oda-Ishii I., Tian E., Suzuki M., Yamamoto M., Suzuki T., Kobayashi M., Aizawa S., Matsuo I. 2004. Characterization of the pufferfish Otx2 cis-regulators reveals evolutionarily conserved genetic mechanisms for vertebrate head specification. Development, 131, 1: 57-71

Kjeldsen M.J., Kyvik K.O., Christensen K., Friis M.L. 2001. Genetic and environmental factors in epilepsy: a population-based study of 11900 Danish twin pairs. Epilepsy Research, 44, 2-3: 167-178

Klaver C.C., Wolfs R.C., Assink J.J., van Duijn C.M., Hofman A., de Jong P.T. 1998. Genetic risk of age-related maculopathy. Population-based familial aggregation study. Archives of Ophthalmology, 116, 12: 1646-1651

Kleinjan D.A., Seawright A., Schedl A., Quinlan R.A., Danes S., van Heyningen V. 2001. Aniridia-associated translocations, DNase hypersensitivity, sequence comparison and transgenic analysis redefine the functional domain of PAX6. Human Molecular Genetics, 10, 19: 2049-2059

Kleinjan D.A., van Heyningen V. 2005. Long-range control of gene expression: emerging mechanisms and disruption in disease. American Journal of Human Genetics, 76, 1: 8-32

Lehmensiek V., Sussmuth S.D., Tauscher G., Brettschneider J., Felk S., Gillardon F., Tumani H. 2007. Cerebrospinal fluid proteome profile in multiple sclerosis. Multiple Sclerosis, 13, 7: 840-849

Leinonen R., Akhtar R., Birney E., Bower L., Cerdeno-Tarraga A., Cheng Y., Cleland I., Faruque N., Goodgame N., Gibson R., Hoad G., Jang M., Pakseresht N., Plaister S., Radhakrishnan R., Reddy K., Sobhany S., Ten Hoopen P., Vaughan R., Zalunin V., Cochrane G. 2011. The European Nucleotide Archive. Nucleic Acids Research, 39, Database issue: D28-D31

Leinonen R., Sugawara H., Shumway M., International Nucleotide Sequence Database C. 2011. The sequence read archive. Nucleic Acids Research, 39, Database issue: D19-D21

Lesnick T.G., Papapetropoulos S., Mash D.C., Ffrench-Mullen J., Shehadeh L., de Andrade M., Henley J.R., Rocca W.A., Ahlskog J.E., Maraganore D.M. 2007. A genomic pathway approach to a complex disease: axon guidance and Parkinson disease. PLoS Genetics, 3, 6: e98, DOI:10.1371/journal.pgen.0030098: 8 str.

Lettice L.A., Horikoshi T., Heaney S.J., van Baren M.J., van der Linde H.C., Breedveld G.J., Joosse M., Akarsu N., Oostra B.A., Endo N., Shibata M., Suzuki M., Takahashi E., Shinka T., Nakahori Y., Ayusawa D., Nakabayashi K., Scherer S.W., Heutink P., Hill R.E., Noji S. 2002. Disruption of a long-range cis-acting regulator for Shh causes preaxial polydactyly. Proceedings of the National Academy of Sciences of the United States of America, 99, 11: 7548-7553

Li H., Durbin R. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics, 26, 5: 589-595

Lill C.M., Roehr J.T., McQueen M.B., Kavvoura F.K., Bagade S., Schjeide B.M., Schjeide L.M., Meissner E., Zauft U., Allen N.C., Liu T., Schilling M., Anderson K.J., Beecham G., Berg D., Biernacka J.M., Brice A., DeStefano A.L., Do C.B., Eriksson N., Factor S.A., Farrer M.J., Foroud T., Gasser T., Hamza T., Hardy J.A., Heutink P., Hill-Burns E.M., Klein C., Latourelle J.C., Maraganore D.M., Martin E.R., Martinez M., Myers R.H., Nalls M.A., Pankratz N., Payami H., Satake W., Scott W.K., Sharma M., Singleton A.B., Stefansson K., Toda T., Tung J.Y., Vance J., Wood N.W., Zabetian C.P., andMe Genetic Epidemiology of Parkinson's Disease C., International Parkinson's Disease Genomics C., Parkinson's Disease G.C., Wellcome Trust Case Control C., Young P., Tanzi R.E., Khoury M.J., Zipp F., Lehrach H., Ioannidis J.P., Bertram L. 2012. Comprehensive research synopsis and systematic meta-analyses in Parkinson's disease genetics: The PDGene

database. PLoS Genetics, 8, 3: e1002548, DOI:10.1371/journal.pgen.1002548: 10 str.

Lincoln M.R., Montpetit A., Cader M.Z., Saarela J., Dyment D.A., Tiislar M., Ferretti V., Tienari P.J., Sadovnick A.D., Peltonen L., Ebers G.C., Hudson T.J. 2005. A predominant role for the HLA class II region in the association of the MHC region with multiple sclerosis. Nature Genetics, 37, 10: 1108-1112

Lindberg R.L., De Groot C.J., Certa U., Ravid R., Hoffmann F., Kappos L., Leppert D. 2004. Multiple sclerosis as a generalized CNS disease--comparative microarray analysis of normal appearing white matter and lesions in secondary progressive MS. Journal of Neuroimmunology, 152, 1-2: 154-167

Lindberg R.L., Hoffmann F., Mehling M., Kuhle J., Kappos L. 2010. Altered expression of miR-17-5p in CD4+ lymphocytes of relapsing-remitting multiple sclerosis patients. European Journal of Immunology, 40, 3: 888-898

Liu S., Bai S., Qin Z., Yang Y., Cui Y., Qin Y. 2009. Quantitative proteomic analysis of the cerebrospinal fluid of patients with multiple sclerosis. Journal of Cellular and Molecular Medicine, 13, 8A: 1586-1603

Liu X., Jian X., Boerwinkle E. 2013. dbNSFP v2.0: a database of human non-synonymous SNVs and their functional predictions and annotations. Human Mutation, 34, 9: E2393-E2402

Lock C., Hermans G., Pedotti R., Brendolan A., Schadt E., Garren H., Langer-Gould A., Strober S., Cannella B., Allard J., Klonowski P., Austin A., Lad N., Kaminski N., Galli S.J., Oksenberg J.R., Raine C.S., Heller R., Steinman L. 2002. Gene-microarray analysis of multiple sclerosis lesions yields new targets validated in autoimmune encephalomyelitis. Nature Medicine, 8, 5: 500-508

Mailman M.D., Feolo M., Jin Y., Kimura M., Tryka K., Bagoutdinov R., Hao L., Kiang A., Paschall J., Phan L., Popova N., Pretel S., Ziyabari L., Lee M., Shao Y., Wang Z.Y., Sirotkin K., Ward M., Kholodov M., Zbicz K., Beck J., Kimelman M., Shevelev S., Preuss D., Yaschenko E., Graeff A., Ostell J., Sherry S.T. 2007. The NCBI dbGaP database of genotypes and phenotypes. Nature Genetics, 39, 10: 1181-1186

Mandel M., Gurevich M., Pauzner R., Kaminski N., Achiron A. 2004. Autoimmunity gene expression portrait: specific signature that intersects or differentiates between multiple sclerosis and systemic lupus erythematosus. Clinical and Experimental Immunology, 138, 1: 164-170

Manolio T.A., Collins F.S., Cox N.J., Goldstein D.B., Hindorff L.A., Hunter D.J., McCarthy M.I., Ramos E.M., Cardon L.R., Chakravarti A., Cho J.H., Guttmacher A.E., Kong A., Kruglyak L., Mardis E., Rotimi C.N., Slatkin M., Valle D., Whittemore A.S., Boehnke M., Clark A.G., Eichler E.E., Gibson G., Haines J.L.,

Mackay T.F., McCarroll S.A., Visscher P.M. 2009. Finding the missing heritability of complex diseases. Nature, 461, 7265: 747-753

Maraganore D.M., de Andrade M., Lesnick T.G., Strain K.J., Farrer M.J., Rocca W.A., Pant P.V., Frazer K.A., Cox D.R., Ballinger D.G. 2005. High-resolution whole-genome association study of Parkinson disease. American Journal of Human Genetics, 77, 5: 685-693

Martinelli-Boneschi F., Esposito F., Brambilla P., Lindstrom E., Lavorgna G., Stankovich J., Rodegher M., Capra R., Ghezzi A., Coniglio G., Colombo B., Sorosina M., Martinelli V., Booth D., Oturai A.B., Stewart G., Harbo H.F., Kilpatrick T.J., Hillert J., Rubio J.P., Abderrahim H., Wojcik J., Comi G. 2012. A genome-wide association study in progressive multiple sclerosis. Multiple Sclerosis, 18, 10: 1384-1394

Martinelli-Boneschi F., Fenoglio C., Brambilla P., Sorosina M., Giacalone G., Esposito F., Serpente M., Cantoni C., Ridolfi E., Rodegher M., Moiola L., Colombo B., De Riz M., Martinelli V., Scarpini E., Comi G., Galimberti D. 2012. MicroRNA and mRNA expression profile screening in multiple sclerosis patients to unravel novel pathogenic steps and identify potential biomarkers. Neuroscience Letters, 508, 1: 4-8

Maver A., Hristovski D., Rindflesch T.C., Peterlin B. 2013. Integration of data from omic studies with the literature-based discovery towards identification of novel treatments for neovascularization in diabetic retinopathy. BioMed Research International, 2013: 848952, DOI: 10.1155/2013/848952: 7 str.

Maver A., Peterlin B. 2011. Positional integratomic approach in identification of genomic candidate regions for Parkinson's disease. Bioinformatics, 27, 14: 1971-1978

Maver A., Medica I., Peterlin B. 2009. Search for sarcoidosis candidate genes by integration of data from genomic, transcriptomic and proteomic studies. Medical Science Monitor, 15, 12: 22-28

Mayeux R. 2003. Epidemiology of neurodegeneration. Annual Review of Neuroscience, 26, 81-104

McCarthy M.I., Abecasis G.R., Cardon L.R., Goldstein D.B., Little J., Ioannidis J.P., Hirschhorn J.N. 2008. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nature Reviews Genetics, 9, 5: 356-369

McKenna A., Hanna M., Banks E., Sivachenko A., Cibulskis K., Kernytsky A., Garimella K., Altshuler D., Gabriel S., Daly M., DePristo M.A. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research, 20, 9: 1297-1303

Meynert A.M., Ansari M., FitzPatrick D.R., Taylor M.S. 2014. Variant detection sensitivity and biases in whole genome and exome sequencing. BMC Bioinformatics, 15, 247, DOI:10.1186/1471-2105-15-247: 11 str.

Moran L.B., Duke D.C., Deprez M., Dexter D.T., Pearce R.K., Graeber M.B. 2006. Whole genome expression profiling of the medial and lateral substantia nigra in Parkinson's disease. Neurogenetics, 7, 1: 1-11

Nadon R., Shoemaker J. 2002. Statistical issues with microarrays: processing and analysis. Trends in Genetics, 18, 5: 265-271

Ng P.C., Henikoff S. 2003. SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Research, 31, 13: 3812-3814

Nickles D., Chen H.P., Li M.M., Khankhanian P., Madireddy L., Caillier S.J., Santaniello A., Cree B.A., Pelletier D., Hauser S.L., Oksenberg J.R., Baranzini S.E. 2013. Blood RNA profiling in a large cohort of multiple sclerosis patients and healthy controls. Human Molecular Genetics, 22, 20: 4194-4205

Nicot A., Ratnakar P.V., Ron Y., Chen C.C., Elkabes S. 2003. Regulation of gene expression in experimental autoimmune encephalomyelitis indicates early neuronal dysfunction. Brain, 126, 2: 398-412

Nistico L., Fagnani C., Coto I., Percopo S., Cotichini R., Limongelli M.G., Paparo F., D'Alfonso S., Giordano M., Sferlazzas C., Magazzu G., Momigliano-Richiardi P., Greco L., Stazi M.A. 2006. Concordance, disease progression, and heritability of coeliac disease in Italian twins. Gut, 55, 6: 803-808

Noben J.P., Dumont D., Kwasnikowska N., Verhaert P., Somers V., Hupperts R., Stinissen P., Robben J. 2006. Lumbar cerebrospinal fluid proteome in multiple sclerosis: characterization by ultrafiltration, liquid chromatography, and mass spectrometry. Journal of Proteome Research, 5, 7: 1647-1657

Noorbakhsh F., Ellestad K.K., Maingat F., Warren K.G., Han M.H., Steinman L., Baker G.B., Power C. 2011. Impaired neurosteroid synthesis in multiple sclerosis. Brain, 134, 9: 2703-2721

Oksenberg J.R., Baranzini S.E. 2010. Multiple sclerosis genetics--is the glass half full, or half empty? Nature Reviews Neurology, 6, 8: 429-37

Olivier B.G., Rohwer J.M., Hofmeyr J.H. 2002. Modelling cellular processes with Python and Scipy. Molecular Biology Reports, 29, 1-2: 249-254

OMIM, 2016. Baltimore, Johns Hopkins University School of Medicine. http://www.omim.org/ (28.2.2016)

Ormond K.E., Wheeler M.T., Hudgins L., Klein T.E., Butte A.J., Altman R.B., Ashley E.A., Greely H.T. 2010. Challenges in the clinical application of whole-genome sequencing. Lancet, 375, 9727: 1749-1751

Parkinson H., Kapushesky M., Shojatalab M., Abeygunawardena N., Coulson R., Farne A., Holloway E., Kolesnykov N., Lilja P., Lukk M., Mani R., Rayner T., Sharma A., William E., Sarkans U., Brazma A. 2007. ArrayExpress--a public database of microarray experiments and gene expression profiles. Nucleic Acids Research, 35, Database issue: D747-D750

Patsopoulos N.A., Barcellos L.F., Hintzen R.Q., Schaefer C., van Duijn C.M., Noble J.A., Raj T., Imsgc, Anzgene, Gourraud P.A., Stranger B.E., Oksenberg J., Olsson T., Taylor B.V., Sawcer S., Hafler D.A., Carrington M., De Jager P.L., de Bakker P.I. 2013. Fine-mapping the genetic association of the major histocompatibility complex in multiple sclerosis: HLA and non-HLA effects. PLoS Genetics, 9, 11: e1003926, DOI:10.1371/journal.pgen.1003926: 10 str.

Peltonen L., Perola M., Naukkarinen J., Palotie A. 2006. Lessons from studying monogenic disease for common disease. Human Molecular Genetics, 15, supp. 1: R67-R74

Perga S., Montarolo F., Martire S., Berchialla P., Malucchi S., Bertolotto A. 2015. Anti-inflammatory genes associated with multiple sclerosis: a gene expression study. Journal of Neuroimmunology, 279: 75-78

Peterlin B., Maver A. 2012. Integrative 'omic' approach towards understanding the nature of human diseases. Balkan Journal of Medical Genetics, 15, Suppl: 45-50

Pfeifer D., Kist R., Dewar K., Devon K., Lander E.S., Birren B., Korniszewski L., Back E., Scherer G. 1999. Campomelic dysplasia translocation breakpoints are scattered over 1 Mb proximal to SOX9: evidence for an extended control region. American Journal of Human Genetics, 65, 1: 111-124

Poulsen P., Kyvik K.O., Vaag A., Beck-Nielsen H. 1999. Heritability of type II (non-insulin-dependent) diabetes mellitus and abnormal glucose tolerance--a population-based twin study. Diabetologia, 42, 2: 139-145

Rasche A., Al-Hasani H., Herwig R. 2008. Meta-analysis approach identifies candidate genes and associated molecular networks for type-2 diabetes mellitus. BMC Genomics, 9, 310, DOI:10.1186/1471-2164-9-310: 17 str.

Rithidech K.N., Honikel L., Milazzo M., Madigan D., Troxell R., Krupp L.B. 2009. Protein expression profiles in pediatric multiple sclerosis: potential biomarkers. Multiple Sclerosis, 15, 4: 455-464

Robinson P.N., Kohler S., Bauer S., Seelow D., Horn D., Mundlos S. 2008. The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. American Journal of Human Genetics, 83, 5: 610-615

Ronemus M., Iossifov I., Levy D., Wigler M. 2014. The role of de novo mutations in the genetics of autism spectrum disorders. Nature Reviews Genetics, 15, 2: 133-141

Rustici G., Kolesnikov N., Brandizi M., Burdett T., Dylag M., Emam I., Farne A., Hastings E., Ison J., Keays M., Kurbatova N., Malone J., Mani R., Mupo A., Pedro Pereira R., Pilicheva E., Rung J., Sharma A., Tang Y.A., Ternent T., Tikhonov A., Welter D., Williams E., Brazma A., Parkinson H., Sarkans U. 2013. ArrayExpress update--trends in database growth and links to data analysis tools. Nucleic Acids Research, 41, Database issue: D987-D990

Sanders S.J., Murtha M.T., Gupta A.R., Murdoch J.D., Raubeson M.J., Willsey A.J., Ercan-Sencicek A.G., DiLullo N.M., Parikshak N.N., Stein J.L., Walker M.F., Ober G.T., Teran N.A., Song Y., El-Fishawy P., Murtha R.C., Choi M., Overton J.D., Bjornson R.D., Carriero N.J., Meyer K.A., Bilguvar K., Mane S.M., Sestan N., Lifton R.P., Gunel M., Roeder K., Geschwind D.H., Devlin B., State M.W. 2012. De novo mutations revealed by whole-exome sequencing are strongly associated with autism. Nature, 485, 7397: 237-241

Sanna S., Pitzalis M., Zoledziewska M., Zara I., Sidore C., Murru R., Whalen M.B., Busonero F., Maschio A., Costa G., Melis M.C., Deidda F., Poddie F., Morelli L., Farina G., Li Y., Dei M., Lai S., Mulas A., Cuccuru G., Porcu E., Liang L., Zavattari P., Moi L., Deriu E., Urru M.F., Bajorek M., Satta M.A., Cocco E., Ferrigno P., Sotgiu S., Pugliatti M., Traccis S., Angius A., Melis M., Rosati G., Abecasis G.R., Uda M., Marrosu M.G., Schlessinger D., Cucca F. 2010. Variants within the immunoregulatory CBLB gene are associated with multiple sclerosis. Nature Genetics, 42, 6: 495-497

Sarkijarvi S., Kuusisto H., Paalavuo R., Levula M., Airla N., Lehtimaki T., Kaprio J., Koskenvuo M., Elovaara I. 2006. Gene expression profiles in Finnish twins with multiple sclerosis. BMC Medical Genetics, 7, 11, DOI:10.1186/1471-2350-7-11: 10 str.

Satoh J., Nakanishi M., Koike F., Miyake S., Yamamoto T., Kawai M., Kikuchi S., Nomura K., Yokoyama K., Ota K., Kanda T., Fukazawa T., Yamamura T. 2005. Microarray analysis identifies an aberrant expression of apoptosis and DNA damage-regulatory genes in multiple sclerosis. Neurobiology of Disease, 18, 3: 537-550

Sawcer S., Hellenthal G., Pirinen M., Spencer C.C., Patsopoulos N.A., Moutsianas L., Dilthey A., Su Z., Freeman C., Hunt S.E., Edkins S., Gray E., Booth D.R., Potter S.C., Goris A., Band G., Oturai A.B., Strange A., Saarela J., Bellenguez C., Fontaine B., Gillman M., Hemmer B., Gwilliam R., Zipp F., Jayakumar A., Martin R., Leslie S., Hawkins S., Giannoulatou E., D'Alfonso S., Blackburn H., Martinelli Boneschi F., Liddle J., Harbo H.F., Perez M.L., Spurkland A., Waller M.J., Mycko M.P., Ricketts M., Comabella M., Hammond N., Kockum I., McCann O.T., Ban M., Whittaker P., Kemppinen A., Weston P., Hawkins C., Widaa S., Zajicek J., Dronov S., Robertson N., Bumpstead S.J., Barcellos L.F., Ravindrarajah R., Abraham R., Alfredsson L., Ardlie K., Aubin C., Baker A., Baker K., Baranzini S.E., Bergamaschi L., Bergamaschi R., Bernstein A., Berthele A., Boggild M., Bradfield J.P., Brassat D., Broadley S.A., Buck D., Butzkueven H., Capra R., Carroll W.M., Cavalla P., Celius E.G., Cepok S., Chiavacci R., Clerget-Darpoux F., Clysters K., Comi G., Cossburn M., Cournu-Rebeix I., Cox M.B., Cozen W., Cree B.A., Cross A.H., Cusi D., Daly M.J., Davis E., de Bakker P.I., Debouverie M.,

D'Hooghe M B., Dixon K., Dobosi R., Dubois B., Ellinghaus D., Elovaara I., Esposito F., Fontenille C., Foote S., Franke A., Galimberti D., Ghezzi A., Glessner J., Gomez R., Gout O., Graham C., Grant S.F., Guerini F.R., Hakonarson H., Hall P., Hamsten A., Hartung H.P., Heard R.N., Heath S., Hobart J., Hoshi M., Infante-Duarte C., Ingram G., Ingram W., Islam T., Jagodic M., Kabesch M., Kermode A.G., Kilpatrick T.J., Kim C., Klopp N., Koivisto K., Larsson M., Lathrop M., Lechner-Scott J.S., Leone M.A., Leppa V., Liljedahl U., Bomfim I.L., Lincoln R.R., Link J., Liu J., Lorentzen A.R., Lupoli S., Macciardi F., Mack T., Marriott M., Martinelli V., Mason D., McCauley J.L., Mentch F., Mero I.L., Mihalova T., Montalban X., Mottershead J., Myhr K.M., Naldi P., Ollier W., Page A., Palotie A., Pelletier J., Piccio L., Pickersgill T., Piehl F., Pobywajlo S., Quach H.L., Ramsay P.P., Reunanen M., Reynolds R., Rioux J.D., Rodegher M., Roesner S., Rubio J.P., Ruckert I.M., Salvetti M., Salvi E., Santaniello A., Schaefer C.A., Schreiber S., Schulze C., Scott R.J., Sellebjerg F., Selmaj K.W., Sexton D., Shen L., Simms-Acuna B., Skidmore S., Sleiman P.M., Smestad C., Sorensen P.S., Sondergaard H.B., Stankovich J., Strange R.C., Sulonen A.M., Sundqvist E., Syvanen A.C., Taddeo F., Taylor B., Blackwell J.M., Tienari P., Bramon E., Tourbah A., Brown M.A., Tronczynska E., Casas J.P., Tubridy N., Corvin A., Vickery J., Jankowski J., Villoslada P., Markus H.S., Wang K., Mathew C.G., Wason J., Palmer C.N., Wichmann H.E., Plomin R., Willoughby E., Rautanen A., Winkelmann J., Wittig M., Trembath R.C., Yaouanq J., Viswanathan A.C., Zhang H., Wood N.W., Zuvich R., Deloukas P., Langford C., Duncanson A., Oksenberg J.R., Pericak-Vance M.A., Haines J.L., Olsson T., Hillert J., Ivinson A.J., De Jager P.L., Peltonen L., Stewart G.J., Hafler D.A., Hauser S.L., McVean G., Donnelly P., Compston A. 2011. Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. Nature, 476, 7359: 214-219

Scherzer C.R., Eklund A.C., Morse L.J., Liao Z., Locascio J.J., Fefer D., Schwarzschild M.A., Schlossmacher M.G., Hauser M.A., Vance J.M., Sudarsky L.R., Standaert D.G., Growdon J.H., Jensen R.V., Gullans S.R. 2007. Molecular markers of early Parkinson's disease based on gene expression in blood. Proceedings of the National Academy of Sciences of the United States of America, 104, 3: 955-960

Sean D., Meltzer P.S. 2007. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. Bioinformatics, 23, 14: 1846-1847

Sherlock G., Hernandez-Boussard T., Kasarskis A., Binkley G., Matese J.C., Dwight S.S., Kaloper M., Weng S., Jin H., Ball C.A., Eisen M.B., Spellman P.T., Brown P.O., Botstein D., Cherry J.M. 2001. The Stanford Microarray Database. Nucleic Acids Research, 29, 1: 152-155

Siegel S.R., Mackenzie J., Chaplin G., Jablonski N.G., Griffiths L. 2012. Circulating microRNAs involved in multiple sclerosis. Molecular Biology Reports, 39, 5: 6219-6225

Sievers C., Meira M., Hoffmann F., Fontoura P., Kappos L., Lindberg R.L. 2012. Altered microRNA expression in B lymphocytes in multiple sclerosis: towards a better understanding of treatment effects. Clinical Immunology, 144, 1: 70-79

Simon R., Radmacher M.D., Dobbin K., McShane L.M. 2003. Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. Journal of the National Cancer Institute, 95, 1: 14-18

Sinha A., Patel S., Singh M.P., Shukla R. 2007. Blood proteome profiling in case controls and Parkinson's disease patients in Indian population. Clinica Chimica Acta, 380, 1-2: 232-234

Sinha A., Srivastava N., Singh S., Singh A.K., Bhushan S., Shukla R., Singh M.P. 2009. Identification of differentially displayed proteins in cerebrospinal fluid of Parkinson's disease patients: a proteomic approach. Clinica Chimica Acta, 400, 1-2: 14-20

Siva N. 2008. 1000 Genomes project. Nature Biotechnology, 26, 3: 256

Smedley D., Haider S., Ballester B., Holland R., London D., Thorisson G., Kasprzyk A. 2009. BioMart--biological queries made easy. BMC Genomics, 10, 22, DOI:10.1186/1471-2164-10-22: 12 str.

Smoller J.W., Finn C.T. 2003. Family, twin, and adoption studies of bipolar disorder. American Journal of Medical Genetics. Part C: Seminars in Medical Genetics, 123C, 1: 48-58

Soneson C., Delorenzi M. 2013. A comparison of methods for differential expression analysis of RNA-seq data. BMC Bioinformatics, 14, 91, DOI: 10.1186/1471-2105-14-91: 18 str.

Soroosh P., Doherty T.A., So T., Mehta A.K., Khorram N., Norris P.S., Scheu S., Pfeffer K., Ware C., Croft M. 2011. Herpesvirus entry mediator (TNFRSF14) regulates the persistence of T helper memory cell populations. Journal of Experimental Medicine, 208, 4: 797-809

Steece-Collier K., Maries E., Kordower J.H. 2002. Etiology of Parkinson's disease: Genetics and environment revisited. Proceedings of the National Academy of Sciences of the United States of America, 99, 22: 13972-13974

Steidl U., Steidl C., Ebralidze A., Chapuy B., Han H.J., Will B., Rosenbauer F., Becker A., Wagner K., Koschmieder S., Kobayashi S., Costa D.B., Schulz T., O'Brien K.B., Verhaak R.G., Delwel R., Haase D., Trumper L., Krauter J., Kohwi-Shigematsu T., Griesinger F., Tenen D.G. 2007. A distal single nucleotide polymorphism alters long-range regulation of the PU.1 gene in acute myeloid leukemia. Journal of Clinical Investigation, 117, 9: 2611-2620

Storch M.K., Stefferl A., Brehm U., Weissert R., Wallstrom E., Kerschensteiner M., Olsson T., Linington C., Lassmann H. 1998. Autoimmunity to myelin oligodendrocyte glycoprotein in rats mimics the spectrum of multiple sclerosis pathology. Brain Pathology, 8, 4: 681-694

Sullivan P.F., Kendler K.S., Neale M.C. 2003. Schizophrenia as a complex trait: evidence from a meta-analysis of twin studies. Archives of General Psychiatry, 60, 12: 1187-1192

Sun J., Jia P., Fanous A.H., Webb B.T., van den Oord E.J., Chen X., Bukszar J., Kendler K.S., Zhao Z. 2009. A multi-dimensional evidence-based candidate gene prioritization approach for complex diseases-schizophrenia as a case. Bioinformatics, 25, 19: 2595-6602

Tajouri L., Mellick A.S., Ashton K.J., Tannenberg A.E., Nagra R.M., Tourtellotte W.W., Griffiths L.R. 2003. Quantitative and qualitative changes in gene expression patterns characterize the activity of plaques in multiple sclerosis. Brain Research: Molecular Brain Research, 119, 2: 170-183

Tan H., Walker M., Gagnon F., Wen S.W. 2005. The estimation of heritability for twin data based on concordances of sex and disease. Chronic Diseases in Canada, 26, 1: 9-12

Tanzi R.E. 2012. The genetics of Alzheimer disease. Cold Spring Harbor Perspectives in Medicine, 2, 10: 1-10

Urdinguio R.G., Sanchez-Mut J.V., Esteller M. 2009. Epigenetic mechanisms in neurological diseases: genes, syndromes, and therapies. Lancet Neurology, 8, 11: 1056-1072

Wakeford R., Roberts W. 1993. Using Medline for comprehensive searches. BMJ, 306, 6889: 1415

Walley A.J., Blakemore A.I., Froguel P. 2006. Genetics of obesity and the prediction of risk for health. Human Molecular Genetics, 15, 124-130

Wang Z., Gerstein M., Snyder M. 2009. RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics, 10, 1: 57-63

Werner C.J., Heyny-von Haussen R., Mall G., Wolf S. 2008. Proteome analysis of human substantia nigra in Parkinson's disease. Proteome Science, 6, 8, DOI:10.1186/1477-5956-6-8: 14 str.

Westerlind H., Ramanujam R., Uvehag D., Kuja-Halkola R., Boman M., Bottai M., Lichtenstein P., Hillert J. 2014. Modest familial risks for multiple sclerosis: a registry-based study of the population of Sweden. Brain, 137, 3: 770-778

Whitney L.W., Ludwin S.K., McFarland H.F., Biddison W.E. 2001. Microarray analysis of gene expression in multiple sclerosis and EAE identifies 5-lipoxygenase as a component of inflammatory lesions. Journal of Neuroimmunology, 121, 1-2: 40-48

Wider C., Ross O.A., Wszolek Z.K. 2010. Genetics of Parkinson disease and essential tremor. Current Opinion in Neurology, 23, 4: 388-393

Yu W., Wulf A., Liu T., Khoury M.J., Gwinn M. 2008. Gene Prospector: an evidence gateway for evaluating potential susceptibility genes and interacting risk factors for human diseases. BMC Bioinformatics, 9, 528, DOI:10.1186/1471-2105-9-528: 8 str.

Zastepa E., Fitz-Gerald L., Hallett M., Antel J., Bar-Or A., Baranzini S., Lapierre Y., Haegert D.G. 2014. Naive CD4 T-cell activation identifies MS patients having rapid transition to progressive MS. Neurology, 82, 8: 681-690

Zeis T., Allaman I., Gentner M., Schroder K., Tschopp J., Magistretti P.J., Schaeren-Wiemers N. 2015. Metabolic gene expression changes in astrocytes in Multiple Sclerosis cerebral cortex are indicative of immune-mediated signaling. Brain, Behavior, and Immunity, 48, 313-325

Zeis T., Kinter J., Herrero-Herranz E., Weissert R., Schaeren-Wiemers N. 2008. Gene expression analysis of normal appearing brain tissue in an animal model for multiple sclerosis revealed grey matter alterations, but only minor white matter changes. Journal of Neuroimmunology, 205, 1-2: 10-19

Zhang H., Jarjour A.A., Boyd A., Williams A. 2011. Central nervous system remyelination in culture--a tool for multiple sclerosis research. Experimental Neurology, 230, 1: 138-148

Zhu C., Li X., Yu J. 2011. Integrating Rare-Variant Testing, Function Prediction, and Gene Network in Composite Resequencing-Based Genome-Wide Association Studies (CR-GWAS). G3: Genes, Genomes, Genetics, 1, 3: 233-243

Zuk O., Schaffner S.F., Samocha K., Do R., Hechter E., Kathiresan S., Daly M.J., Neale B.M., Sunyaev S.R., Lander E.S. 2014. Searching for missing heritability: designing rare variant association studies. Proceedings of the National Academy of Sciences of the United States of America, 111, 4: 455-464

ZAHVALA

Na prvem mestu se zahvaljujem svojemu dolgoletnemu mentorju, profesorju Borutu Peterlinu. Hvala Vam za skrbno mentorstvo, neštete napotke, ure dragocenih diskusij in za vsak dan polno mero ustvarjalnosti in navdušenja za znanost. Za to se Vam zahvaljujem tako v okviru tega doktorskega dela kot za raziskovalno in strokovno delo sicer.

Zahvaljujem se vsem dolgoletnim sodelavcem in prijateljem na Kliničnem inštitutu za medicinsko genetiko - hvala ker ste mi od samega začetka pomagali stopiti na pot genetike in ker ste si vedno bili pripravljeni vzeti čas za pomoč pri delu. Delo na inštitutu je zaradi vas res edinstvena izkušnja. Hvala Igorju Medici - čeprav Vas ni več med nami, ste pustili neizbrisljiv in nepozaben pečat.

Posebej sem hvaležen Alenki Hodžić za pomoč pri laboratorijskem delu, za pomoč pri zaključevanju dela, za vse spodbude in podporo takrat, ko sem to najbolj potreboval. Hvala Mariji Volk - ker mi brez izjeme vedno stojiš ob strani, tudi tokrat s strokovnim pregledom dela v zadnjih minutah. Hvala Lovru Vidmarju za pregled dela in vso podporo pri zaključevanju dela. Hvala tudi Poloni Lavtar, Ireni Jurman in vsem drugih sodelavcem, brez katerih ne bi bilo mogoče uresničiti raziskovalnih ciljev.

Posebej se zahvaljujem tudi najdražjim prijateljem - Urški in Urošu, ker sta bila vedno ob strani in ker sta vedno razumela in počakala, tudi ko se je raziskovalno delo zavleklo pozno v noč. Hvala Milošu in Darini, za gostoljubnost v Ljubljani in ker sta nase prevzela številne moje skrbi in bila vedno pripravljena pomagati pri stvareh, zaradi katerih mi je tolikokrat zmanjkalo časa.

Na koncu pa še hvala najdražjim - staršem in bratu Mateju. Hvala ker ste me naučili pomena poštenosti, vztrajnosti in dobronamernosti. Brez vaše podpore in odrekanja mi ne bi uspelo.