TARTU ÜLIKOOL
Arvutiteaduse instituut
Informaatika õppekava

Marielle Lepson

Geeniontoloogia andmete muutus ajas g:Profiler'i näitel

Bakalaureusetöö (9 EAP)

Juhendaja(d): Liis Kolberg

Tartu 2021

Geeniontoloogia andmete muutus ajas g:Profiler'i näitel

Lühikokkuvõte:

Geeniontoloogia on geeniandmete kogum, mis on pidevas muutumises tulenevalt geeniandmete lisandumisest, muutumisest või eemaldamisest. g:Profiler on Tartu Ülikooli BIIT töörühma poolt loodud veebitööriist, mis kasutab geeniontoloogia andmeid, et paremini kirjeldada geeni nimekirju vastavalt nende funktsionaalsusele. Käesoleva töö eesmärk on kasutada andmekaeve meetodeid uurimaks geeniontoloogia andmete muutuseid ajas inimorganismi näitel. Samuti uuritakse elektrooniliste annotatsioonide esinemist ning hinnatakse nende kvaliteeti. Selleks kasutatakse g:Profiler'i andmete arhiive, mis ulatuvad aastasse 2009.

Võtmesõnad:

Geeniontoloogia, g:Profiler, andmekaeve

CERCS: P170 Arvutiteadus; P175 Informaatika, süsteemiteooria ; B110 Bioinformaatika

Changes in Gene Ontology over time on the example of g:Profiler

Abstract:

Gene Ontology is a set of knowledge that is in constant change, because of changes and updates made from gene data. g:Profiler is a popular web tool that uses Gene Ontology information to describe the gene lists according to their functionality. The aim of this Bachelor thesis is to use the data mining methods to analyse and describe the changes that have occured in Gene Ontology over time according to human data. Further, electronic annotations are being reviewed and evaluated using various data mining methods. For this purpose, g:Profiler data archives dating back to 2009 are used.

Keywords:

Gene Ontology, g:Profiler, data mining

CERCS: P170 Computer science; P175 Informatics, systems theory; B110 Bioinformatics

Sisukord

Sissejuhatus 4
Eelteadmised 6
1.1 Geeniontoloogia mõiste 6
1.2 Annotatsioonid ja tõendikoodid 8
1.3 Geeniontoloogia muutus ajas 10
1.4 Annotatsioonid g:Profiler'is 12
2. Metoodika 14
2.1 Andmed 14
2.2 Vahendid ja meetodid 14
3. g:Profiler'i andmete ülevaade 16
3.1 Ülesande püstitus 16
3.2 Analüüsi teostus 16
3.3 Tulemused 17
3.3.1 Ülevaade g:Profiler'i andmetest 17
3.3.2 Ülevaade elektroonilistest annotatsioonidest 20
4. Elektrooniliste annotatsioonide kvaliteedi hindamine 23
4.1 Ülesande püstitus 23
4.2 Analüüsi teostus 23
4.3 Elektrooniliste annotatsioonide kvaliteedi tulemused 26
4.3.1 g:Profiler'i tulemused 26
4.3.2 Võrdlus artikli tulemustega 27
5. Elektrooniliste annotatsioonide muutus eksperimentaalseteks annotatsioonideks 29
5.1 Ülesande püstitus 29
5.2 Analüüsi teostus 29
5.3 Tulemused 29
Kokkuvõte 32
Viidatud kirjandus 34
Lisad 38
I. Lisamaterjalid 38
II. Koodi repositoorium 42
Litsents 43

Sissejuhatus

Inimese arengu jaoks vajalikku informatsiooni sisaldavaid valke kodeerivad geenid [1]. Geenid osalevad organismi talitluseks ja arenemiseks olulistes protsessides. Selleks, et saada ülevaade, millistes protsessides geenid osalevad, loodi Geeniontoloogia Konsortsiumi (ingl k *The Gene Ontology Consortium*) poolt geeniontoloogia (ingl k *Gene Ontology*) [2, 3]. Geeniontoloogia on geenide funktsioonide kogu, mis sisaldab praeguseid teadmisi geenide funktsionaalsusest [2].

Geeni ja funktsiooni suhet nimetatakse annotatsiooniks. Sõltuvalt sellest, kuidas annotatsioon on tõestatud, jagunevad annotatsioonid gruppidesse. Annotatsioon võib olla leitud arvutuslike andmete põhjal. Sellist tõestust nimetatakse elektrooniliseks annotatsiooniks [4]. Geenide funktsionaalsust võidakse tõestada katseliselt laborites. Neid annotatsioone nimetatakse eksperimentaalseteks annotatsioonideks. Eksperimentaalset annotatsiooni peetakse kõige usaldusväärsemaks, kuid selle tõestamine on ajamahukas protsess [5]. Vastavalt statistikale, elektroonilisi annotatsioone esineb geeniontoloogias rohkem kui eksperimentaalseid, kuid neid peetakse vähem usaldusväärseks [4, 5].

Kuivõrd geeniontoloogia on pidevas muutumises, on olemas õigustatud huvi uurida geeniontoloogia andmete muutusi ajas. Geeniontoloogia on suur bioloogiline süsteem ning seetõttu tuleks uurimine läbi viia sobivate meetoditega. Üks võimalikest uurimismeetoditest on andmekaeve. Andmekaeve on uurimismeetod, mis hõlmab suurte andmemahtude analüüsimist eesmärgiga leida sealt seaduspärasusi, mille põhjal on võimalik teha kasulikke järeldusi [6]. Tulenevalt bioloogias ning meditsiiniteadustes käsitlevate andmete suurest mahust on andmekaeve meetodite kasutamine nendes valdkondades laialdaselt levinud.

Käesolevas töös uuritakse veebiserver g:Profiler'i geeniontoloogia andmeid andmekaeve meetoditel. g:Profiler'it kasutatakse mitmetes teaduslikes töödes [7, 8]. Kasutajatel võib tekkida küsimusi päringu tulemuste muutuste kohta, tehes g:Profiler'ist aja mööbudes korduspäringuid. Töö kirjutamise hetkel puudus süstemaatiline ülevaade g:Profiler'i versioonide andmete erinevustest. Seetõttu on ülevaatlikud joonised g:Profiler'i kasutajatele kasulikud, nähes nii andmete muutust. Lisaks hinnatakse elektrooniliste annotatsioonide kvaliteeti, vaadates automaatselt annoteeritud geeniandmete muutumist erinevate versioonide vahel.

Bakalaureusetöös on kasutatud veebiserveri g:Profiler 2009. – 2020. aastate versioonide andmeid inimese geeniandmete näitel, et uurida geeniontoloogia muutuseid ajas. Andmete analüüsimiseks kasutatakse programmeerimiskeelt Python ning andmegraafide ja jooniste tegemiseks teeke (Pandas, Numpy, Altair). Geeniontoloogia andmete uurimiseks on bakalaureusetöös püstitatud kolm eesmärki:

1) esitada ülevaade g:Profiler'i geeniontoloogia andmete ja elektrooniliste annotatsioonide muutustest ajas inimese näitel;
2) anda hinnang g:Profiler'i andmete näitel elektrooniliste annotatsioonide kvaliteedile ajas;
3) uurida elektrooniliste annotatsioonide muutumist eksperimentaalseteks annotatsioonideks.

Töö on jaotatud viieks peatükiks. Esimeses peatükis tutvustatakse lugejale tööks vajalikke eelteadmisi geeniontoloogia, andmete muutuse ja g:Profiler'i kohta. Teises peatükis on töö metoodika tutvustus. Järgnevad kolm eelpüstitatud eesmärke käsitlevat peatükki (3, 4, 5). Iga peatükk neist sisaldab sõnastatud ülesande püstitust, teostust ning tulemusi. Täpsemalt, kolmandas peatükis antakse ülevaade g:Profiler'i andmetest ja elektroonilistest annotatsioonidest. Neljandas peatükis hinnatakse g:Profiler'i andmetel elektrooniliste annotatsioonide

kvaliteeti varasemalt läbi viidud uuringu [4] meetodite alusel. Viiendas peatükis vaadeldakse elektrooniliste annotatsioonide muutust eksperimentaalseteks annotatsioonideks. Kokkuvõttes antakse ülevaade tehtud tööst ning tuuakse välja töös selgunud tähelepanekud ning olulisemad märkused. Lisade all on link GitLab repositooriumile, kus on saadaval analüüsi lähtekood.

Eelteadmised

Organismi ning selle tunnuste arenemises osalevad geenid [1]. Geenid on valke ja RNA-d kodeerivad järjestused, mis sisaldavad informatsiooni ühe või mitme tunnuse väljaarenemiseks ning neil on oma kindel funktsioon organismis. Geenide kodeerimisel valmivad geeni produktid (RNA, valgud). Geenide kohta on aja jooksul kogunenud rohkelt teadmisi. Näiteks on nüüdseks teada, et geenid WNT10A ja CUTC mõjutavad juuste morfoloogiat [9].

Selleks, et informatsioon oleks kõigile soovijatele kättesaadav, on andmed koondatud mitmetesse avalikesse andmebaasidesse. Üks tuntumaid andmebaase on Ensembl [10], mis sisaldab genoomide andmeid. Genoom on organismi kogu geneetiline materjal. Inimesel moodustab genoomi 24 kromosoomi ja mitokondri genoom. Geenid paiknevad kromosoomides kindlates piirkondades. Veel mõned levinumad andmebaasid on NCBI (ingl k *National Center for Biotechnology Information*) [11], mis on genoomi andmete analüüsiks, ning FlyBase [12], mis on äädikakärbeste spetsiifiline andmebaas. Geene uuritakse, et tuvastada nende funktsioone organismi toimimisel. Tänu sellistele uuringutele on lisaks eelnevatele andmebaasidele olemas ka andmebaasid, mis kirjeldavad tuvastatud geenide bioloogilisi funktsioone nagu KEGG [13], Reactome [14] ja Geeniontoloogia (GO) [2, 3].

1.1 Geeniontoloogia mõiste

Geeniontoloogia (GO) on bioloogiliste andmete kogum, mis sisaldab informatsiooni geenide ja geeni produktide funktsionaalsuse kohta. GO pakub bioloogilist mudelit, millesse on kogutud senised teadmised geenide funktsioonide kohta organismi tasemest molekulaarse tasemeni [2]. Peamine eesmärk on pakkuda nii inimesele kui ka masinale loetavat ning üheselt mõistetavat süsteemi, mis võimaldab kirjeldada geenide funktsioone.

GO andmebaas loodi 1998. aastal Geeniontoloogia Konsortsiumi (edaspidi GOC) poolt [3]. GOC vastab vajadusele uuendada geeniontoloogia struktuuri, pakkudes usaldusväärset arvutuslikku mudelit ning samuti laiendades bioloogilist kogumit. Kui algselt uuriti kolme organismi genoome (äädikakärbes, hiir, pagaripärm), siis praegu on geeniontoloogiast võimalik leida üle 4000 organismi andmeid (2020 oktoober seisuga) [2,15]. GO on üks teadmusbaasidest bioloogilistes ja meditsiinilistes uurimustes, mis põhineb arvutiteadusel [2, 3]. Näiteks, 2017. aastal analüüsiti geeni produkti IncRNA seost vähiga, kasutades funktsioonide leidmiseks geeniontoloogiat [16].

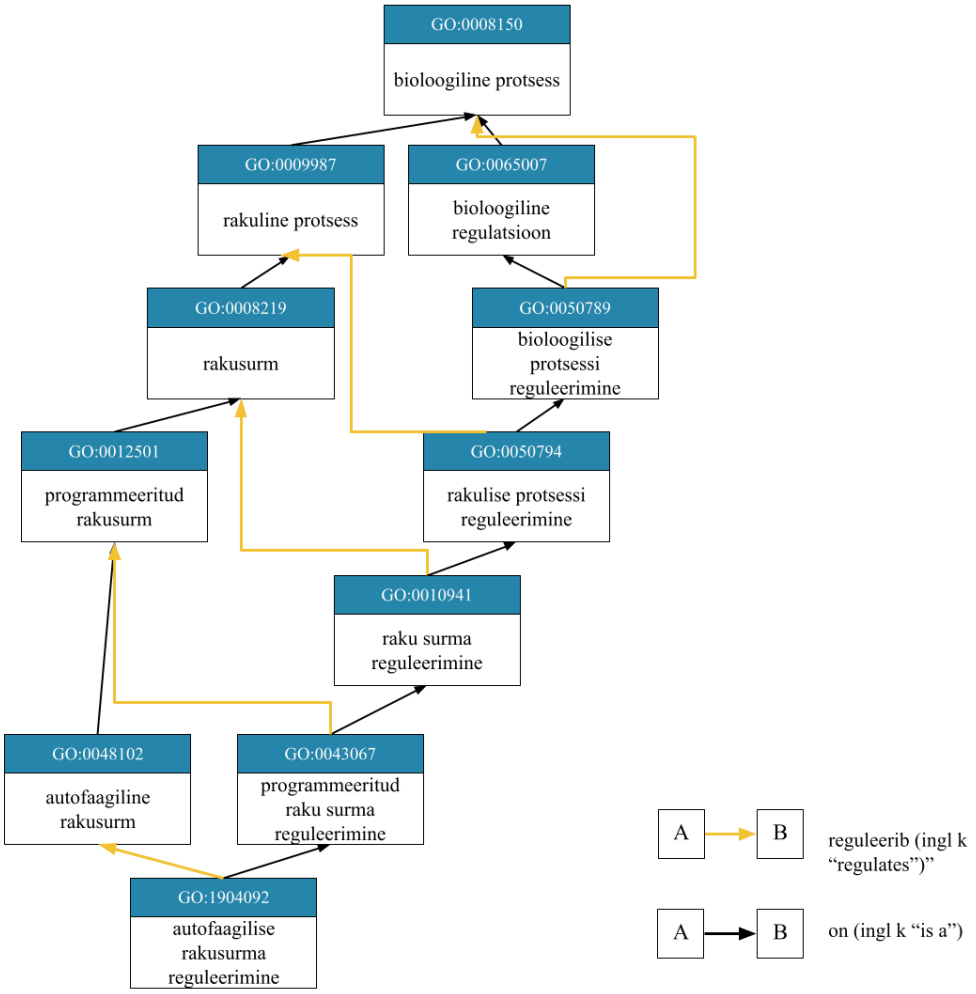
GO koosneb klassidest, mida nimetatakse termideks [3]. Term tähistab ühte bioloogilist funktsiooni või omadust. Geeniontoloogia statistika veebilehe[1] alusel esines geeniontoloogias 2020. aasta oktoobris 44 264 termi.

Struktuuri poolest on GO suunatud atsükliline graaf ehk graaf, mille servadel on suund ja millel puuduvad suunatud tsüklid (vt Joonis 1). Graafi tippudeks on termid ning termide vahelised semantilised seosed on kirjeldatud tippude vaheliste servadega. Terme kujutatakse hierarhilise struktuurina, kus ülevalpool olevad termid on üldisemad ning allpool on spetsiifilisemad termid. Käesolevas töös nimetatakse ülem-terme vanemaks ja alam-terme alluvateks. Kahe põhilise defineeritud semantilise seose inglise keelsed nimetused on *is-part* ja *is-a*, mis iseloomustavad termide vahelisi alluvussuhteid [17]. Seos *is-a* näitab, et alluv on spetsiifilisem funktsioon vanemast ning *is-part*, et alluv on osa enda vanemast [18]. Joonisel 1 on toodud ka seos *regulates*, mis näitab, kuidas A term mõjutab B termi. Igal termil võib olla rohkem kui üks vanem ja null kuni mitu alluvat.

[1] http://geneontology.org/stats.html Versioon : 2020-10

Kõikidel GO struktuuridel kehtib „tõese tee reegel" (ingl k *true path rule*). Viimane tähendab, et kui geen on seotud alluva funktsiooniga, siis geen täidab kindlasti ka vanema funktsiooni. Näiteks Joonisel 1, kui geen osaleb alamprotsessis „raku surma reguleerimine", siis osaleb see geen ka vanema protsessis „bioloogilise protsessi reguleerimine".

Joonis 1 on sinistes kastides näidatud terminikoodid. Igale geeniontoloogia termile on määratud vastav unikaalne terminikood. Terminikood algab alati tähtedega „GO" ning seejärel esineb unikaalne 7-numbriline numbrikood (näiteks GO:1904092). Igale geeniontoloogia terminikoodile vastab ka inimloetav nimetus. Näiteks terminikoodile GO:1904092 vastab protsess „autofaagilise rakusurma reguleerimine".



Joonis 1. Geeniontoloogia alamgraafi osa näidis alam-termi „autofaagilise rakusurma reguleerimine" (GO:1904092) põhjal (Joonis kohandatud tööriista QuickGO [19] tulemustest) [20]

Geeniontoloogia struktuuri jaotatakse omakorda kolmeks mittelõikuvaks alamgraafiks: bioloogiline protsess (BP), molekulaarne funktsioon (MF) ja rakuline komponent (CC). Neid alamgraafe nimetatakse domeenideks. Kõik kolm domeeni koosnevad erinevast hulgast termidest ja nende vahelistest seostest, kuid domeenide vahel puuduvad seosed. Kõikide alamtermide juurest on võimalik tippudest liikudes domeeni juurtipuni jõuda [21]. Kuna nendel kolmel alamgraafil ei ole ühist juurtippu, siis kutsutakse neid ka kolmeks eraldiseisvaks ontoloogiaks [21].

Molekulaarne funktsioon (juure terminikood GO:0003674) tähistab tegevusi, mis esinevad molekulaarsel tasandil nagu aktiivsus, katalüüs ja transport. Molekulaarne funktsioon nimetab, mida tehakse, aga mitte asukohta ega osalejat funktsioonis [2]. Selleks, et määrata molekulaarset funktsiooni on lisatud termidele juurde iseloomustav sõna, näiteks aktiivsus (näiteks „transferaasi aktiivsus“) [21].

Rakukomponent või rakuline komponent (juure terminikood GO:0005575) määrab, kus raku osas geeni funktsioon toimub, näiteks membraan või sünaps. Komponentide määratlus annab parema ülevaate keerulisemast eukarüootse ehk tuumaga raku ehitusest, kuna eukarüootne rakk sisaldab palju komponente, mida teised ei sisalda [2]. Rakukomponentide alla kuuluvad näiteks termid: „tuum“, „golgi kompleks“ ja „organell“.

Kõige suurem domeen, bioloogiline protsess (juure terminikood GO:0008150) tähistab tegevusi suuremal skaalal, mis toimuvad mitme kindla järjestusega molekulaarse protsessi tulemusena [2]. Bioloogilised protsessid on „rakusurm“ ja „DNA parandamine“. Üks geeniprodukt võib osaleda mitmes erinevas bioloogilises protsessis, mistõttu moodustuvad geeniontoloogias keerulised struktuurid. Näidis bioloogilise protsessi osast on Joonisel 1.

1.2 Annotatsioonid ja tõendikoodid

Vastavalt avaldatud artiklitele, andmebaasidele ja teistele allikatele on geeniontoloogia struktuuris iga term seotud geeni hulgaga, mis selles funktsioonis osalevad. See tähendab, et geeniontoloogia term ei ole spetsiifiline ainult ühte tüüpi geenile või geeni produktile, vaid sellega võib olla seotud rohkem geene. Termi ja geeni suhet koos tõendikoodiga nimetatakse annotatsiooniks ning geeni seostamist funktsiooniga annoteerimiseks. Näiteks, QuickGO [19] andmetel on termi „DNA sidumine“ ja geeni ZNF641 (ingl k *zinc finger protein 641*) vahel loodud annotatsioon. See tähendab, et geen ZNF641 osaleb DNA sidumise funktsioonis.

Kõikidele annotatsioonidele on määratud tõendikoodid (ingl k *evidence code*). Tõendikoodid näitavad meetodit, mille põhjal vastav annotatsioon on määratud. Annotatsioon viitab alati lisatud tõendikoodidele [2]. GOC poolt on kirjeldatud [22], et annotatsioone määratakse kas arvutuslike meetodite kaudu või manuaalselt. Nende sõnul on manuaalsed annotatsioonid eksperimentide ja teadusartiklite alusel kuraatorite poolt geeniontoloogia andmebaasi sisestatud. Automaatsed tõendikoodid põhinevad mitmetel arvutuslikel meetoditel, aga erinevalt manuaalsetest tõendikoodidest puudub neil kuraatoripoolne kontroll.

Annotatsiooni tõendikoode tähistatakse 1-3 suurtähelise lühendiga. Tõendikoodid on näiteks IDA (ingl k *Inferred from Direct Assay*) ja IEA (ingl k *Inferred from Electronic Annotation*). Kõik tõendikoodid on vastavalt geeniontoloogia lehel loodud jaotusele toodud välja tabelis 1. Esimeses veerus on kategooria nimi, teises sinna kuuluvad tõendikoodid ning viimases tulbas kategooria selgitus.

Ühel annotatsioonil võib esineda mitu tõendikoodi, näiteks 2021. jaanuar seisuga on geen PIGZ termiga „mannosüültransferaasi aktiivsus“ annoteeritud kolme[2] tõendikoodiga (IBA, IGI, EXP).

[2] https://www.ebi.ac.uk/QuickGO/annotations?taxonId=9606&taxonUsage=descendants&geneProductId=UniProtKB:Q86VD9&geneProductType=protein&goId=GO:0000030&goUsageRelationships=is_a,part_of,occurs_in&goUsage=descendants

Tabel 1. Geeniontoloogia tõendikoodid [23]

Kategooria	Tõendikoodid	Selgitus
Eksperimentaalsed tõendikoodid (ingl k *Experimental evidence codes*)	EXP tõendikoodid: EXP, IDA, IPI, IMP, IGI, IEP HTP tõendikoodid: HTP, HDA, HMP, HGI, HEP	EXP tõendikoodid: annotatsiooni tõestab eksperiment (näiteks immunofluorestsents). HTP tõendikoodid põhinevad kõrge läbilaskevõimega ehk rohkete andmetega eksperimentidel.
Fülogeneetiliselt järeldatud tõendikoodid (ingl k *Phylogenetically-inferred annotations*)	IBA, IBD, IKR, IRD	Tõestus geeni funktsiooni avastuse või puudumise tõttu fülogeneetiliste puude harude põhjal.
Arvutusanalüüsi tõendikoodid (ingl k *Computational analysis evidence codes*)	ISS, ISA, ISO, ISM, IGC, RCA	Tõendikood viitab sellele, et tõestamisel on kasutatud geenijärjestuse analüüsi või muud meetodit. Hõlmab mingil määral kuraatori abi.
Autori avaldatud tõendikoodid (ingl k *Author statement evidence codes*)	TAS, NAS	Neid tõendikoode kasutatakse siis, kui väide põhineb autori esitatud seisukohal. Jaguneb kaheks: jälgitav autori avaldus (TAS) ja jälgimatu autori avaldus (NAS).
Kuraatori avaldatud tõendikoodid (ingl k *Curator statement evidence codes*)	IC, ND	Annotatsioon on tehtud kuraatori otsuse põhjal, kuid ei sobinud teised tõendikoodid. IC (ingl *k Inferred by Curator*): puuduvad kindlad tõendid, kuid kuraatori poolt loogiliselt järeldatud. ND (ingl k *No biological Data available*): puudub informatsiooni MF, BP, CC kohta.
Elektrooniliselt tõestatud annotatsioonid (ingl k *Electronic annotation evidence code*)	IEA	Annotatsioonil puudub manuaalne kuraatori poolne ülevaatus. Põhineb homoloogial, järjestusel või teistel varasematel tõestustel.

Geeniontoloogias leidub rohkem informatsiooni nendest valdkondadest, mis antud hetkel on enim huvipakkuvad. Seetõttu leiab mõne funktsiooniga rohkem manuaalseid annotatsioone kui teisega. Korrektsete annotatsioonide valimine on keeruline protsess, sest alati ei ole kindlaid eksperimentaalseid tõendeid geeni osalemisest protsessis. GO tõendikoodide üheks omaduseks on pakkuda kasutajale võimalust hinnata annotatsiooni usaldusväärsust. Kui on ligipääs annotatsiooni tõestuse meetodile, on kasutajal kergem andmeid hinnata.

Hetkeseisuga ei ole veel välja toodud ametlikku tõendikoodide paremusjärjestust. Buza, McCarthy jt. [25] hindasid 2008. aastal tõendikoode viiepallisüsteemis. Viie punktiga hinnati eksperimentaalseid tõendikoode. Elektroonilisele tõendikoodile (IEA) hinnati tulemuseks kaks punkti. Null punkti sai ND (ingl k *No biological Data available*) tõendikood ja ühe punkti NR (ingl k *Not Recorded*), tähistamaks annotatsioone, mis on tehtud enne tõendikoodide loomist. Eksperimentaalseid tõendikoode peetakse kõige usaldusväärsemateks seetõttu, et on läbiviidud eksperiment, mis tõestab annotatsiooni. Automaatseid tõendikoode peetakse vähem usaldusväärseteks, kuna neil puudub kuraatoripoolne ülevaatus [4].

Annotatsioonidele kehtivad GOC poolt seatud geeniontoloogia struktuuri põhimõtted[3]. Järgnevalt tuuakse mõned neist välja. Esimeseks, iga geeni produkti võib annoteerida null kuni mitu korda erinevate termide juurde. Näiteks, geen LCT on annoteeritud Ensembl andmetel seitsme[4] molekulaarsesse funktsiooni kuuluva termiga. Teiseks, kõik geenid on seotud kõige täpsemini määratletud ja sõnastatud termiga, vastavalt tõendikoodidele. Vastasel juhul võidakse muuta termi nimetust või eemaldatakse term. Kolmandaks, iga annotatsioon peegeldab kõige uuemat informatsiooni, samas tuleb arvestada, et annotatsioonid võivad muutuda, kuna teadmised muutuvad ajas. See tähendab, et ühel hetkel võib automaatne annotatsioon olla tõestatud ka eksperimentaalselt, mis näiliselt lisab annotatsiooni tugevusele kindlust.

1.3 Geeniontoloogia muutus ajas

Levinud on valearusaam, et geeniontoloogia on muutumatu ja bioloogilised protsessid on püsivalt määratud, kuigi tegelikult on GO pidevalt muutuv struktuur [26]. Hetkel kasutatavad andmed katavad ainult senised teadmised. Andmetes toimuvad pidevad muutused, kuna ka meie teadmised muutuvad ajas. Statistika[5] andmetel on geeniontoloogias vaadeldavate organismide arv suurenenud. 2020. aasta oktoobris on 23 organismi rohkem (4666) kui 2020. aasta juulis (4643). Samuti, geeniontoloogia statistika veebilehe alusel 2018. aasta septembri kuni 2020. aasta oktoobrini kahanes termide arv andmebaasis 768 võrra.

Huntley jt. [26] tõdesid, et muutused võivad olla väikesed: nimetuse muutus termil, uus alluv term või siis ka teistsugused väiksemad muutused. Samas toimuvad mitme väiksema muudatuse koosmõjul suuremad struktuuri ümberkorraldused. Geeniontoloogias esinevateks põhilisteks muutusteks on:

- termiga seotud muutused: uus alluv, uus vanem, termi nimetuse muutus, termi eemaldamine, termile sünonüümide lisamine;
- annotatsioonide muutused: uus annotatsioon, annotatsiooni eemaldamine;
- tõendikoodidega seotud muutused: uus tõendikood, tõendikoodi eemaldamine.

[3] http://geneontology.org/docs/go-annotations/
[4] http://www.ensembl.org/Homo_sapiens/Gene/Ontologies/molecular_function?db=core;g=ENSG00000115850;r=2:135787850-135837184
[5] http://geneontology.org/stats.html

Geeniontoloogia muutuste rõhutamiseks tuuakse järgnevas loetelus artiklite ja andmebaaside alusel mõned näited. Esmalt tuuakse näiteid termide muutuste kohta (punktid 1-3). Seejärel on annotatsioonide muutused (punktid 4-7). Viimasena on toodud tõendikoodidega seotud muutused (punktid 8-9).

Eelpool kirjeldatud näited on toodud alljärgnevalt:

1. Esimene näide on termi eemaldamise kohta. Veebilehe QuickGO [19] andmetel on „ribosomaalse šaperooni aktiivsus“ (GO:0000005) 2008. aastal vananenud (ingl k *obsolete*) ning geeniontoloogiast kustutatud. Term kuulub domeeni molekulaarne funktsioon. Eemaldamise põhjuseks on toodud, et see viitab pigem geeni produktidele ja bioloogilisele protsessile kui molekulaarsele funktsioonile.
2. Kui term on mõnest versioonist kadunud, siis see ei tähenda, et term oleks vale olnud [26]. Term GO:0023067 (ingl k *obsolete signal transmission via lymphatic system*) asendati termiga GO:0023052 (ingl k *signaling*), sest esimene oli liiga mitmetähendusliku nimega.
3. Termidega võib toimuda ka teisi muudatusi, mis kajastuvad geeniontoloogias. QuickGO logi järgi on aastate jooksul molekulaarse funktsiooni domeeni kuuluva termiga „DNA sidumine“ (GO:0003677) seotud muudatusi olnud kokku 49. See on süsteemi sisestatud 2001. aasta märtsis. Kaks aastat hiljem, 2003. aasta märtsis, lisati nimetusele juurde sõna „aktiivsus“, mis hiljem uuesti eemaldati. Aastate jooksul on lisatud termile sünonüüme ning uuendatud on termi definitsiooni.
4. Muutused termidega mõjutavad nii geeniontoloogiat kui ka annotatsioone [26]. GOC avaldas 2019. aastal artikli [3], kus tehti katseprojekt annotatsioonide kvaliteedi hindamiseks, mille raames uuriti ~2500 manuaalset annotatsiooni. Eesmärkideks seati vähendada uute annotatsioonide arvu ning rohkem rõhku panna vanemate annotatsioonide parandamisele. 70-80% uuritavatest annotatsioonidest hiljem asendati või eemaldati. Projekti käigus lisati geeniontoloogiasse juurde uus term „transkriptsiooni regulaatori aktiivsus“ (GO:0140110), et grupeerida kokku kõiki terme, mis otseselt reguleerivad transkriptsiooni. Transkriptsioon on DNA ahelalt RNA ahela süntees. Uuel termil GO:0140110 esineb 12 otsest alluvat termi (2020. aasta novembri seisuga). Eelmine term eemaldati, sest term on vananenud ning juba teiste termide poolt geeniontoloogias kaetud.
5. Annotatsioonide arvud on aastate jooksul muutunud. Geeniontoloogia statistika[6] andmetel on 2020. aasta oktoobri seisuga kokku 8 049 377 annotatsiooni, neist 3 060 065 on bioloogilise protsessi annotatsioonid. 2018. aastal oli GO annotatsioone 7 288 273.
6. Annotatsioonide muutused võivad kajastuda rohkem huvipakkuvates alades ning muudatusi on vähem valdkondades, mida on vähem uuritud. Seoses 2020. aasta Covid-19 levikuga on geeniontoloogiat rohkelt kasutatud viiruse uurimisel [27-29]. Näiteks uuriti [27] ensüümi ja SARS-CoV-2 retseptori ACE2 üle-ekspressiooni rakkudes peale nakatumist. ACE2 on roll funktsioonis „tsütokiinide tootmise reguleerimine“ (GO:0001817). QuickGO andmetel on termi GO:0001817 2020. aasta jooksul uuendatud 11 korda ja termil on lisatud juurde 3 alluvtermi. Termi viimane muutus enne 2020. aastat oli 2008. aastal.

[6] http://geneontology.org/stats.html

7. Termide vahel esinevate semantiliste seoste muutused kajastuvad geeniontoloogias. Vastavalt tõese tee reeglile, iga termi vahelise seose eemaldamine mõjutab nendega seotud annotatsioone. Huntley jt [26] teatasid, et 2011. aastal vähenes manuaalsete ja automaatsete annotatsioonide arv ligi 2500 võrra, kuna eemaldati seos termide „transkriptsioon, DNA-malliga“ (GO:0006351) ja „DNA-d siduva transkriptsioonifaktori aktiivsus“ (GO:0003700) vahel.
8. Tõendikoodid mõjutavad annotatsioone. 2012. aastal uuendas Zebrafish Model Organism Database [30] oma identifikaatorite faili, mille tulemusena läks kaduma umbes 15 000 manuaalset annotatsiooni Uniprot GO failis [26]. Uniprot [31] pakub kasutajatele informatsiooni valgujärjestuste ja nende funktsionaalsuse kohta.
9. Annotatsiooni stabiilsus geeniontoloogias sõltub, kui usaldusväärne on annotatsiooni tõestus. Tõendikoodide hindamiseks on läbiviidud erinevaid uuringuid [4, 25]. Üks põhjus, miks annotatsioonid muutuvad: suur osa kõikidest annotatsioonidest on automaatselt tõestatud annotatsioonid. 2020. aasta oktoobri seisuga on 2 038 699 automaatsed annotatsioonid [15]. Positiivses võtmes tähendab IEA annotatsioonide küllus, et suur osa annotatsioone saab kiiresti tõestatud lühikese aja jooksul [26]. Teisalt automaatne annotatsioon ei ole kuraatorite poolt kontrollitud, mistõttu ei saa olla kindel nende tõepärasuses [2, 3]. Selleks, et säiliks annotatsioonide kvaliteet ning toimuks järjepidev annoteerimine kohtuvad geeniontoloogia kuraatorid regulaarselt, et analüüsida GO terme ja GO tõendikoode [32].

Nimetatud näited aitavad paremini mõista geeniontoloogias toimunud muutusi.

1.4 Annotatsioonid g:Profiler'is

g:Profiler (https://biit.cs.ut.ee/gprofiler) on Tartu Ülikooli bioinformaatika töörühma BIIT loodud veebitööriist bioloogidele, mis võimaldab arvutuslikel meetoditel analüüsida kasutaja sisestatud geeninimekirju. g:Profiler pakub kasutajale mitut töövahendit: g:GOSt, g:Convert, g:Orth ja g:SNPense. g:GOSt on g:Profiler'i tööriist geenide funktsionaalsuse analüüsi läbiviimiseks vastavalt kasutaja sisendile [33]. Tööriistas g:Convert saab teisendada kasutaja sisestatud geeninimesid erinevate identifitseerijate vahel. g:Orth leiab sisestatud geenidele ortoloogsed geenid ehk võimaldab geene teisendada erinevate organismide vahel. g:SNPense seostab SNP-ide koodid geenide nimedega ning annab vastavalt kromosoomidele SNP algus- ja lõpp-positsioonid. SNP (üksiknukleotiidsed polümorfismid) on DNA järjestuse muutused, kus järjestusest võib ära kaduda üks nukleotiid või hoopis juurde lisanduda. g:Profiler tugineb põhiliselt Ensembl [10] andmebaasile ja uuendab oma andmeid vastavalt sealsetele uuendustele [34]. g:Profiler'isse on võimalik sisestada ligikaudu sada erinevat identifitseerimise koodi tüüpi, mis annab kasutajale võimaluse kasutada eri formaatides geenide või valkude andmeid [33, 35]. Tööriist toetab lisaks inimesele veel mitmeid teisi organisme, kelle andmed on kättesaadavad Ensembl andmebaasidest (467 organismi 2019. aasta seisuga [35]) [33-35].

Kõige populaarsem tööriist, g:GOSt võrdleb kasutaja sisestatud geeninimekirju oma andmebaasi vastu ja leiab selle põhjal neid geene kirjeldavad bioloogilised funktsioonid. Selleks viiakse läbi funktsionaalse rikastamise analüüs, milleks kasutatakse kumulatiivset hüpergeomeetrilist testi [35]. Tulemuseks tuvastab tööriist sisestatud geenidega statistiliselt oluliselt seotud funktsioonid, protsessid ja rajad. Geenide funktsionaalsuse informatsioon pärineb mitmetest andmebaasidest nagu GO, KEGG [13], Reactome [14], Wikipathways [36], TRANSFAC[37] jt[35]. Geeniontoloogia andmebaas on üks suurimaid ressursse tööriistas g:GOSt [35].

Tänu süstemaatilistele uuendustele pakub g:Profiler kasutajatele ajakohaseid GO andmeid ning mugavat kasutajaliidest. Tööriistas g:GOSt toetati 2019. aasta sügisel 178[7] ja 2020. aasta sügisel 197[8] erinevat Ensembl organismi. Andmete uuendused toimuvad aastas mitu korda. Teistes sarnastes tööriistades nagu DAVID [38, 39] kasutatakse andmeid, mida uuendati viimati aastaid tagasi (DAVID andmete uuendus 4 aastat tagasi 2020. aasta november seisuga) [33]. Käesoleval aastal on g:Profiler'i andmeid juba kolm korda uuendatud (2020. aasta september seisuga), mille tõttu g:Profiler sisaldab ajakohaseid andmeid.

Tööriista kõik versioonid on säilitatud g:Profiler'i veebilehel vaates „Arhiivid" (ingl k *Archives*[9]*).* Sealt leiab lingid eelmistele versioonidele, mida saab kasutada juhul kui on tarvis varasemaid tulemusi taasesitada, aga andmed on vahepeal muutunud.

[7] https://biit.cs.ut.ee/gprofiler_archive3/e97_eg44_p13/gost (2019-10-07)
[8] https://biit.cs.ut.ee/gprofiler_archive3/e100_eg47_p14/gost (2020-09-21)
[9] https://biit.cs.ut.ee/gprofiler/page/archives

2. Metoodika

2.1 Andmed

Käesolevas bakalaureusetöös on kasutatud g:Profiler'i arhiivi versioonide andmeid. Andmed pärinevad g:Profiler'i andmebaasist ning sisaldavad ainult inimesega seotud geeniontoloogia andmeid. Töö läbiviimise ajal oli analüüsitavaid faile ehk arhiivi versioone kokku 32. Kuivõrd g:Profiler'i andmeid uuendatakse aastas mitu korda, siis üks fail sisaldab andmeid, mida ühes versioonis kasutatakse. Lisa 1, Tabelis 2 on välja toodud g:Profiler'i versiooni failinimed ning nende avaldamise kuupäevad. Kõige vanem fail on nimetusega r0814_e56, mille andmed pärinevad vastavalt nimele Ensembl 56 versioonist. Selline versioon oli g:Profiler'is kasutusel 2009-02-02. Kõige uuem versioon analüüsitavatest failidest on r2007_e100_eg47, mis loodi 2020-07-07. See versioon põhineb Ensembl 100 ja Ensembl Genome 47 andmetel. Bakalaureusetöö läbiviimiseks valmistati failid ette BIIT töörühma poolt. Arusaadavuse mõttes viitab autor andmeid analüüsides versiooni avaldamise kuupäevale (vt Lisa 1, Tabel 2).

Üks fail koosneb andmeridadest, kus on kolm tabulaatoriga eraldatud veergu: terminikood, geeni identifikaator ja tõendikood. Esimesena esineb terminikood. Seejärel on geeni Ensembl identifikaator, mis on tähistatud tähtedega ENSG ning sellele järgneb unikaalne numbrikombinatsioon. Viimases veerus on vastavad tõendikoodide lühendid eraldatud püstkriipsuga.

Näide faili ridadest on toodud Joonisel 2. Ühe annotatsiooni moodustab versioonis r1760_e93_eg40: GO:0044238 ENSG00000130726 IEA|TAS|ISS|IDA|IMP. QuickGO lehe järgi on term bioloogilise protsessi domeenist ja termi nimetus on „esmane metaboolne protsess". Geeni nimi on TRIM28. Annotatsioonile on sellel ajahetkel määratud 5 tõendikoodi (IEA, TAS, ISS, IDA, IMP).

```
1792012  GO:0044238  ENSG00000130726 IEA|TAS|ISS|IDA|IMP
1792013  GO:0008150  ENSG00000145882 IBA|TAS|IEA
1792014  GO:0044425  ENSG00000096384 IEA
1792015  GO:0051128  ENSG00000136999 IMP|IEA
1792016  GO:0065007  ENSG00000169862 IMP|TAS|IEA
1792017  GO:0030182  ENSG00000125378 IEA
1792018  GO:0040007  ENSG00000186051 IEA
1792019  GO:1901576  ENSG00000176890 IEA|IDA|TAS|IC
```

Joonis 2. Kuvatõmmis faili r1760_e93_eg40 ridadest

Termide ontoloogiatesse määramiseks ja täiendava informatsiooni saamiseks kasutati käesolevas töös geeniontoloogia lehelt saadud ametlikult koostatud „go.obo" faili[10]. See on vabalt kättesaadav andmefail, mis sisaldab muu hulgas informatsiooni termi nimetuse, alternatiivsete ID-de, domeenide ja termi kasutuseloleku kohta. g:Profiler'i arhiivi failide ning „go.obo" faili andmete näitel viidi läbi lõputöös tehtud andmekaeve.

2.2 Vahendid ja meetodid

Kõikide andmete analüüsimiseks kasutati programmeerimiskeelt Python. Koodi kirjutamiseks kasutati tarkvara Jupyter Notebook [40], andmeanalüüsi ja – töötluse teeki Pandas [41] ja andmemassiivide teeki Numpy [42]. Joonised tehti teegi Altair[11] abil, mis võimaldab interaktiivseid jooniseid teha.

[10] http://geneontology.org/docs/download-ontology/#go_obo_and_owl
[11] https://altair-viz.github.io/

Andmete esmaseks analüüsimiseks loeti eelmainitud andmefailid arvuti mällu. Seejärel loodi Pythonis funktsioonid kirjeldava statistika teostuseks. Funktsioonid on koostati nii, et neid saaks andmete uuenemisel või versioonide juurde tekkimisel korduvalt kasutada. Funktsioonid on ka jooniste tegemiseks ning vajalike andmetabelite koostamiseks ning neid saab rakendada vastavalt sisestatud andmetele. Domeeni värvid on valitud vastavalt g:Profiler'i veebilehel esinevatele värvikoodidele.

Bakalaureusetööks loodud interaktiivseid jooniseid saab vaadata lingilt:

http://mlepson.gitlab.cs.ut.ee/l-put/.

3. g:Profiler'i andmete ülevaade

3.1 Ülesande püstitus

Andmed muutuvad ajas, sest teadmised muutuvad [26]. Geeniontoloogia uurimisel on esmalt vajalik teada, millised on kasutusel olevad andmed. g:Profiler on eksisteerinud veebiserverina üle kümne aasta, kuid kuni käesoleva hetkeni on puudunud süstemaatiline ülevaade seal olevatest geeniontoloogia andmetest läbi aja. See informatsioon võib olla kasulik ka g:Profiler'i kasutajatele, et rõhutada andmete muutust juhul kui kasutaja teeb andmetest korduspäringuid. Illustreerimaks geeniandmete muutumist ajas, koostas autor joonised, mille abil saab näidata kui palju inimesega seotud geeniontoloogia andmeid on g:Profiler'is olnud ja kuidas need on muutunud. Ülevaate andmiseks tehakse selgitavad joonised termide, annotatsioonide, geenide ja annotatsioonidega seotud tõendikoodide kohta.

g:Profiler'i andmetes kasutatud elektrooniliste annotatsioonide kohta puudub hetkel ülevaade. Vastavalt geeniontoloogia statistikale[12] on elektroonilisi annotatsioone 2020. aasta septembri seisuga 2 046 302. See on 65 081 annotatsiooni rohkem kui kuu varem. Eksperimentaalseid annotatsioone on 2020. aasta septembri seisuga üle 50% vähem kui elektroonilisi annotatsioone. g:Profiler'i andmete kirjelduse teises osas antakse ülevaade annotatsioonidest. Annotatsioonide tõendikoodide paremaks uurimiseks püstitas autor järgnevad uurimisküsimused:

1. Kui palju geene on eksperimentaalselt annoteeritud vähemalt ühe termiga?
2. Kui paljud termid on g:Profiler'i kõige uuemas versioonis ainult automaatselt tõestatud ehk IEA?
    a. Kui suured on ainult automaatsete annotatsioonidega termid?
3. Kui suur protsent erinevate termide annotatsioonidest moodustub IEA kõige uuemas versioonis?

3.2 Analüüsi teostus

Termide, geenide ja annotatsioonide andmetabelite kättesaamiseks on loodud kaks funktsiooni. Esimene funktsioon valib versiooni ning teine tagastab selle versiooni andmed. Geenide ja termide arvud moodustusid iga versiooni unikaalsete tulemuste arvu põhjal. Annotatsioonid loeti kokku vastavalt failis esinevatele ridadele. Termide ja annotatsioonide arvud esitatakse joondiagrammina ning geenide arv tulpdiagrammina. Annotatsioonide grupeerimiseks tõendikoodi alusel on esmalt tõendikoodid jagatud kuute rühma järgides geeniontoloogia statistika lehel esitatud jaotust[13]:

1. EXP ehk eksperimentaalselt tõestatud tõendikoodid (EXP, IDA, IPI, IMP, IGI, IEP);
2. HTP ehk kõrge läbilaskevõimega eksperimentaalselt tõestatud tõendikoodid (HTP, HDA,HMP, HEP, HTP, HGI);
3. IEA ehk automaatselt annoteeritud tõendikoodid;
4. ND ehk tõendikoodid, millel puuduvad bioloogilised andmed;
5. PHYLO ehk fülogeneetiliselt tõestatud (IBA, IBD, IKR, IRD) ;
6. Teised ehk kõik ülejäänud tõendikoodid, mis ei kuulu eelnevatesse kategooriatesse. Kõik tõendikoodid on toodud välja tabelis 1.

Jaotusi vaadeldakse ükshaaval ning loetakse kokku kõik annotatsioonid, mis sisaldavad sellesse kategooriasse kuuluvaid tõendikoode. Kuna tõendikoodidel puudub paremusjärjestus,

[12] http://geneontology.org/stats.html
[13] http://geneontology.org/stats.html

siis annotatsioon, mis sisaldab nii elektroonilist kui ka ND tõendikoodi, loetakse elektrooniliseks annotatsiooniks ja ND annotatsiooniks. Sama reegel kehtib kõikide teiste tõendikoodidega. Tulemusi kujutatakse virnastatud tulpdiagrammidena. Võrdluseks on värvikoodid valitud samad, mis geeniontoloogia statistika lehel[14] toodud tõendikoodide tabelis. Ülejäänud küsimuste jaoks on loodud funktsioonid, mis pärivad andmetabelitest andmeid vastavalt küsimusele.

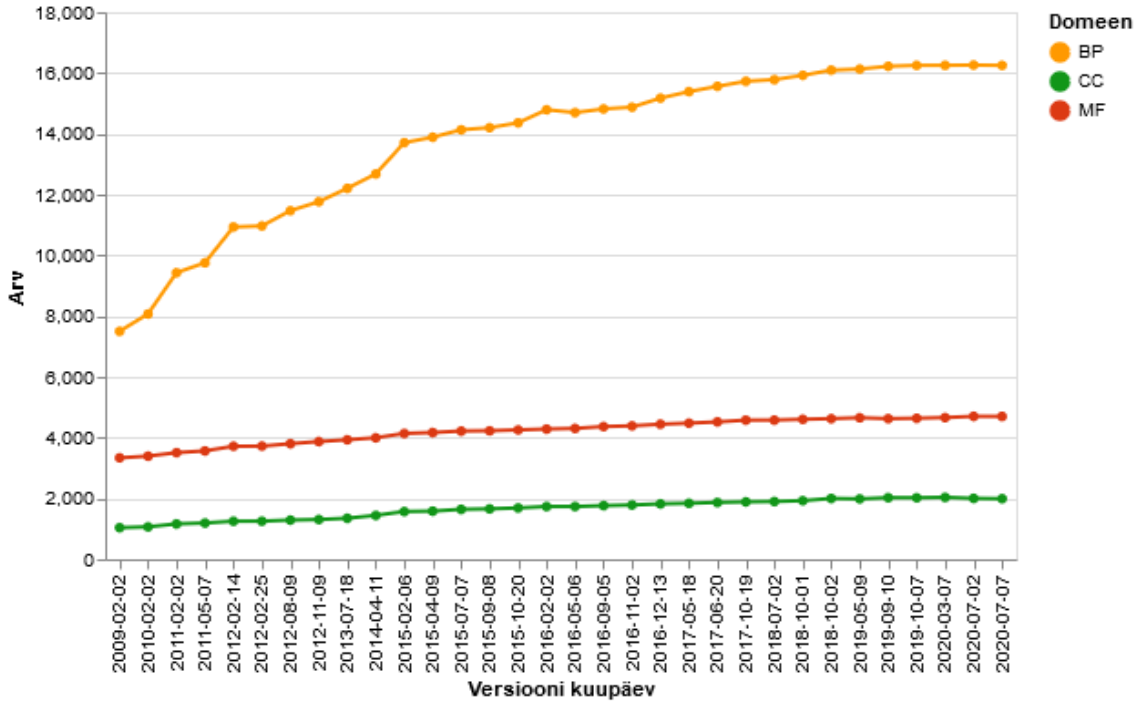
3.3 Tulemused

3.3.1 Ülevaade g:Profiler'i andmetest

Termide arvu muutused aja jooksul kajastavad geeniontoloogia muutumist. g:Profiler'i versioonide termide arv on iga domeeni puhul enamasti kasvanud (vt Joonis 3). Termide arvu kasvu on kõige paremini näha bioloogilise protsessi puhul. Kui 2009-02-02 avaldatud versioonis on bioloogilise protsessi terme 7502, siis 2015-02-06 esineb 13 709 termi ja kõige hilisemas 2020-07-07 uuenduses on 16 251 termi.

Märkimisväärne termide arvu muutus on kuupäevade 2014-04-11 ja 2015-02-06 vahel. Üle kõigi domeenide toimus muutus 18 124 termi pealt 19 422 termile. g:Profiler'i andmete tulemusena lisandus 2015-02-06 juurde 1404 uut termi. 2015-02-06 versioonist puuduvad 85 termi, mis esinesid eelnevas versioonis. Eeltoodud muutused võivad olla tingitud sellest, et Ensembl keskendus [43] 2014. aastal uue genoomi GRCh38 andmete uuendamisele. Uue inimese kogugenoomi kasutuselevõtu tõttu koguneb uusi andmeid [43].

Kõige uuem versioon 2020-07-07 sisaldab 25 termi, mis on „go.obo" faili põhjal märgitud „kadunud" termideks ning edaspidistes versioonides neid enam ei kajastata.



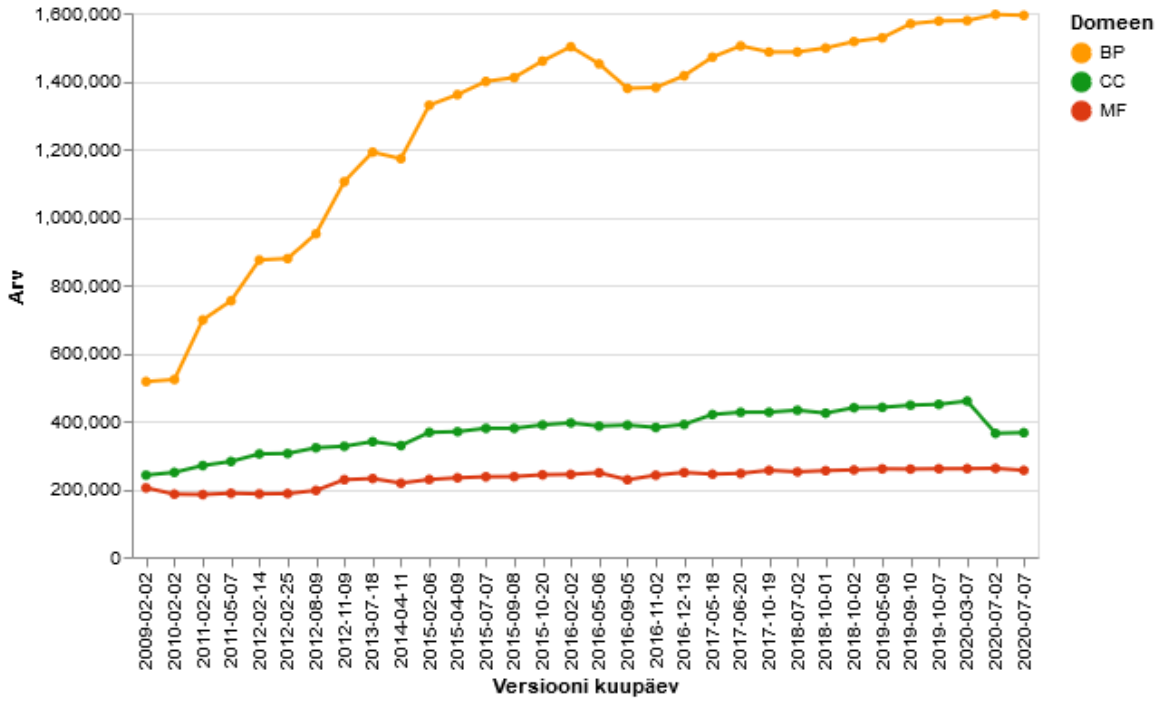
Joonis 3. Termide arvud GO domeenide kaupa g:Profiler'i versioonide lõikes.

Termide lisamine või kustutamine mõjutab annotatsioone, sest uute termide puhul tekib vajadus luua uusi annotatsioone ning kustutamisel vajadus eemaldada. g:Profiler'i andmetes

[14] http://geneontology.org/stats.html

esinevate annotatsioonide arvud on esitatud Joonisel 4, mis näitab tulemusi vastavalt kolmele domeenile. Annotatsioonide arvud on alates 2009. aastast kasvanud, kuid leidub ka versioone, kus annotatsioonide arv on hoopis vähenenud. Selline tulemus esineb domeenil bioloogiline protsess (BP) vahemikus 2013-07-18 kuni 2014-04-11 ning 2016-02-02 kuni 2016-09-05.

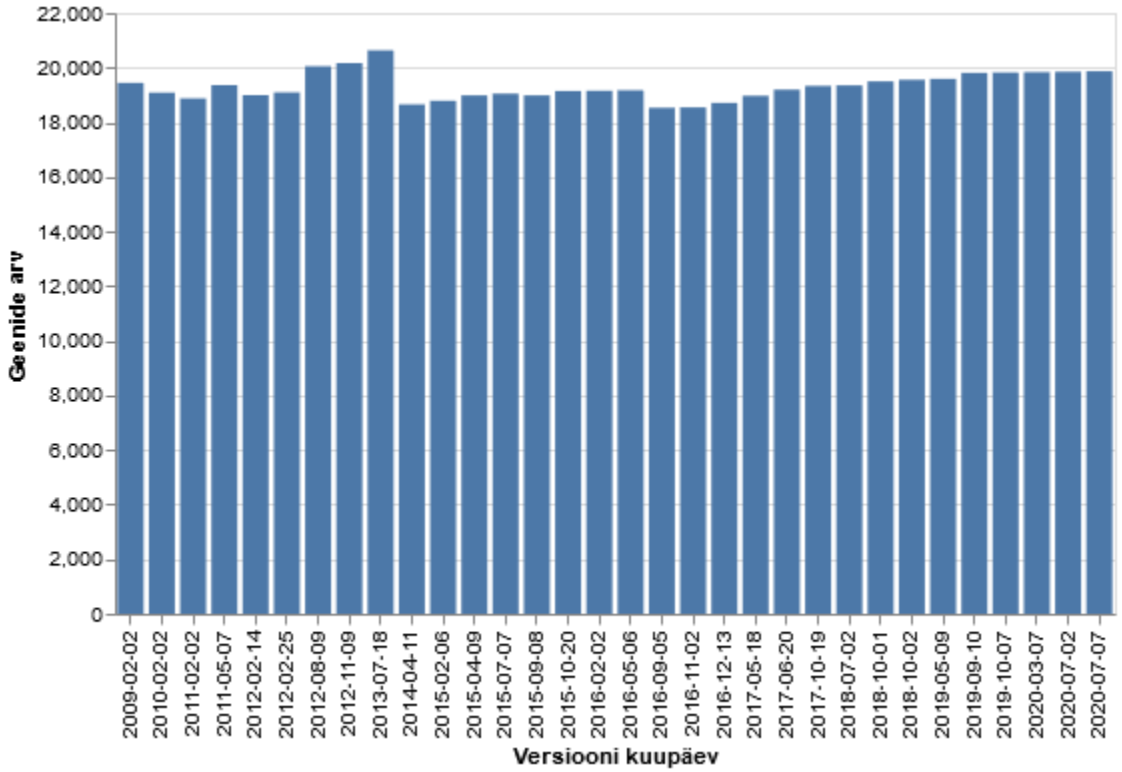
Kõige viimases versioonis on BP annotatsioone 1 593 463, CC annotatsioone 366 188 ja MF annotatsioone 255 388. Kõige rohkem annotatsioone esineb bioloogilisel protsessil ning ning molekulaarse funktsiooni (MF) annotatsioone on kõige vähem. Kõige suurem annotatsioonide arvu kasv on toimunud BP, kus 2009-02-02 versioonis on 516 457 annotatsiooni ning kõige uuemas versioonis on peaaegu miljon annotatsiooni rohkem (vt Joonis 4).



Joonis 4. Annotatsioonide statistika läbi versioonide g:Profiler'i andmetel

2016. aasta jooksul toimus BP annotatsioonidega mitmeid muutusi. g:Profiler'i andmetes tehti aasta jooksul kokku viis uuendust. Aasta alguses oli bioloogilise protsessi annotatsioone 1 501 220 ning aasta lõpuks 1 415 943 annotatsiooni. 2016. aastal avaldatud artiklis [32] kirjeldati annotatsioonide suurenemist. Artiklis toodi välja, et IntAct [44] andmebaasi lisamise tõttu on valkudega seotud annotatsioonide arvud rohkelt suurenenud. Kuivõrd 2016. aasta lõpul annotatsioonide arv suurenes, tuuakse järgnevalt näide termi lisamisest, millega kaasnesid uued annotatsioonid. 2016. aasta lõpupoole lisati g:Profiler'i andmebaasi molekulaarse funktsiooni domeeni kuuluv term GO:0005515 (ingl *k protein binding*) koos 11 066 annotatsiooniga.

Järgmisena vaadeldakse geenide arvu läbi versioonide (vt Joonis 5). Geenide arvud on aastate lõikes püsinud peamiselt muutumatuna. Vaadeldavate geenide arv versioonides jääb 18 000 ja 21 000 vahele. Inimese genoomis on valku kodeerivate geenide arv just selles vahemikus [45]. Joonisel on näha vähest „geenikadu" uuenduste 2013-07-18 ja 2014-04-11 vahel. Nimelt on 2014-04-11 versioonis 1981 geeni vähem kui eelmises versioonis. Samas ajavahemikus langes märkimisväärselt ka annotatsioonide arv (vt Joonis 4), mis on oodatav, aga termide arv kasvas (vt Joonis 3). Üheks põhjuseks võib olla, et 2013. aastal suurendati Ensembl's kvaliteedi kontrolli, mis sisaldas suuremahulist andmete ja andmebaasi kontrolli [46].



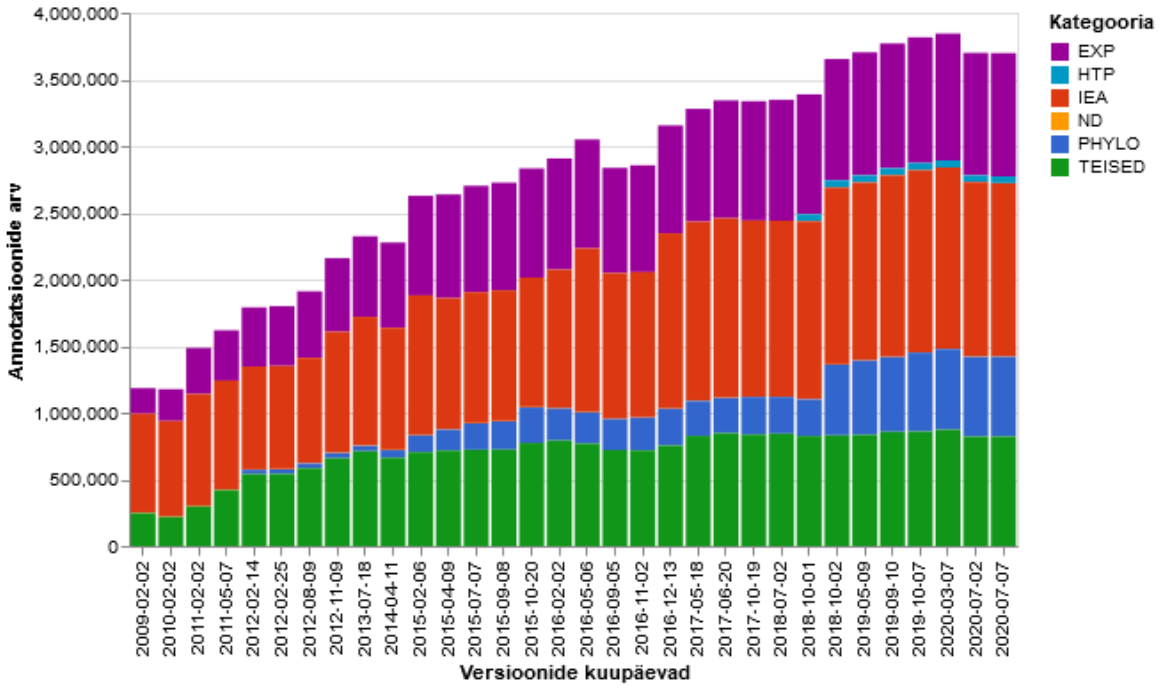
Joonis 5. Geenide arvud versioonide lõikes

g:Profiler'i annotatsioonid jaotatuna tõendikoodide kategooriate kaupa on esitatud Joonisel 6. Kuivõrd eksperimentaalseid tõendikoode sisaldavate annotatsioonide arv on rohkelt kasvanud, on ka elektrooniliste annotatsioonide (IEA) arv kasvanud. EXP kategooria tõendikoodiga on 2009-02-02 seisuga 192 279 annotatsiooni ja 2020-07-07 on 924 094 annotatsiooni. Vanimas versioonis on IEA tõendikoodiga 743 128 annotatsiooni ning uuemas 1 300 402 annotatsiooni inimese andmete põhjal. IEA annotatsioone on g:Profiler'i andmetel rohkem kui EXP tõendikoodiga annotatsioone.

Vaadeldes teisi kategooriaid Joonisel 6, esinevad HTP kategooriasse kuuluvad tõestatud annotatsioonid alates g:Profiler'i uuendusest r1760_e93_eg40, mis on avaldatud kuupäeval 2018-10-01. Tegemist on uute kõrge läbilaskevõimega eksperimentaalsete tõendikoodidega. 2018-10-01 versioonis on HTP kategooria tõendikoodidega 51 823 annotatsiooni. Järgnevalt vaadeldakse ühe sellise annotatsiooni tõendikoodide muutust:

- Avaldamise kuupäev: 2018-07-02 (r1750_e91_eg38):
  GO:0005623 ENSG00000117525 IBA|IDA|TAS|NAS|IC
- Avaldamise kuupäev: 2018-10-01 (r1760_e93_eg40):
  GO:0005623 ENSG00000117525 IDA|TAS|IC|IBA|HDA|NAS.

Annotatsioonil tekkis juurde uus HTP kategooria tõendikood (HDA) kuupäeval 2018-10-01. g:Profiler'i versioonides on annotatsioonidel kasutatud ainult kolme HTP tõendikoodi (HDA, HEP ja HMP).

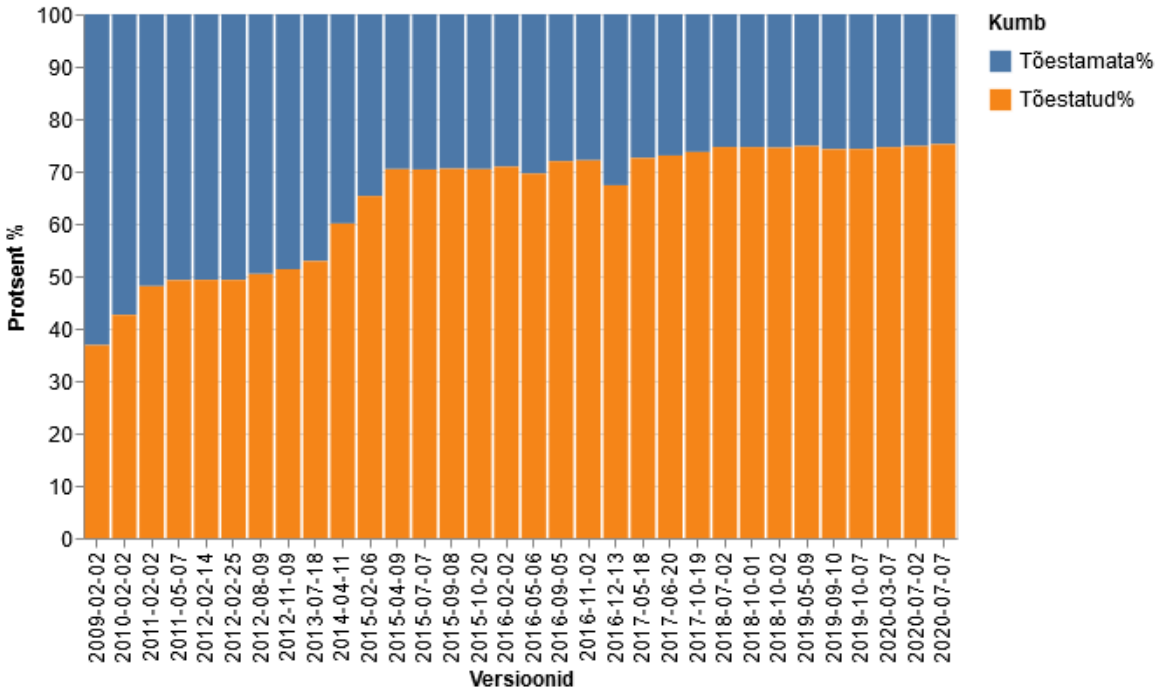


Joonis 6. Annotatsioonide arv tõendikoodi kategooria järgi

Annotatsioone, mis on tõestatud tõendikoodiga ND, on kõige uuemas versioonis ligikaudu 1600. Kategooria PHYLO esineb alates versioonist 2012-02-14. Fülogeneetiliselt tõestatud annotatsioonide arv on kasvanud. PHYLO tõendikoodidega annotatsioone on 2017-05-18 262 876 ja 2020-07-07 597 872. Fülogeneetilisi tõendikoode on hakatud rohkem kasutama, sest nende abil on võimalik efektiivselt kontrollida annotatsiooni kvaliteeti ning saada ülevaade uuritavast valgu perekonnast [3].

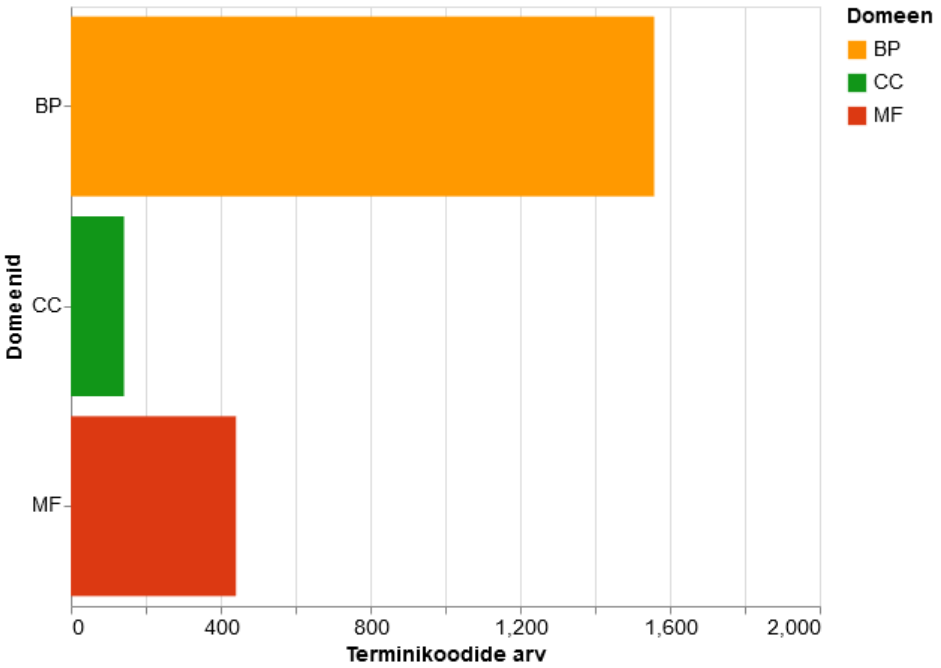
3.3.2 Ülevaade elektroonilistest annotatsioonidest

Üldise andmete kirjelduse järel uuritakse lähemalt annotatsioone. Esimesena, eksperimentaalsete annotatsioonidega geenide osakaalu erinevates versioonides. Siinkohal arvestatakse, et eksperimentaalsete alla kuuluvad kõik eksperimentaalsed tõendikoodid (tõestatud). Kõik ülejäänud tõendikoodid arvestatakse mitte eksperimentaalseteks tõendikoodideks (tõestamata). Joonisel 7 on näidatud, et eksperimentaalse annotatsiooniga geenide arv on aastatega kasvanud. Kui 2015. aasta alguses oli tõestatud geenide arvu osakaal 65,29% siis viimase g:Profiler'i uuenduse järgi on osakaal juba 75,19%.



Joonis 7. Eksperimentaalse annotatsiooniga geenide osakaal läbi versioonide

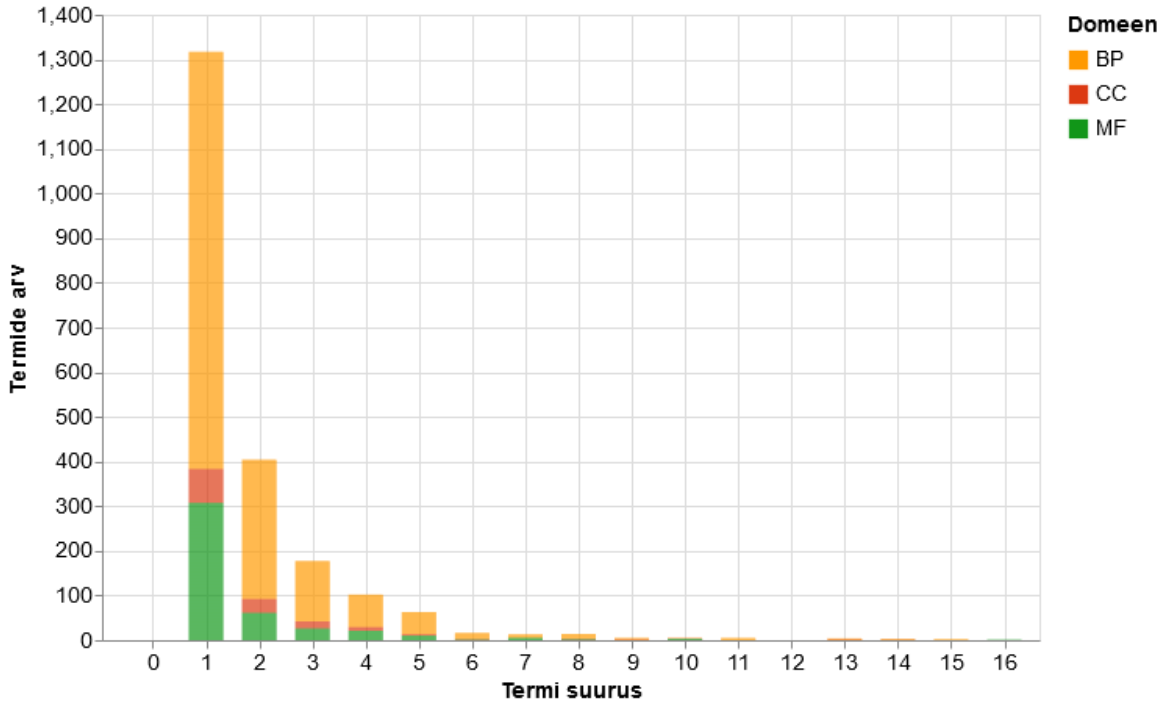
Kõige viimases andmeversioonis on kokku 4929 geeni, millel ei ole ühtegi eksperimentaalset annotatsiooni. Järgnevalt uuritakse kas leidub terme, millel on vaid elektroonilised annotatsioonid. Joonisel 8 on domeenide kaupa toodud, kui palju on versioonis 2020-07-07 terme, mille annotatsioonid on ainult automaatselt tõestatud. Kõige rohkem selliseid terme esineb bioloogilises protsessis (1558). Molekulaarses funktsioonis on 441 termi ning kõige vähem on rakukomponendis, kus on 142 sellist termi. Elektrooniliste annotatsioonidega termide arvud kokku moodustavad 9,33% kõikidest selles versioonis olnud termidest.



Joonis 8. Termide arv, mille annotatsioonid on ainult automaatselt tõestatud (IEA) tõendikoodid 2020-07-07 avaldatud versioonis

Üks sellistest termidest on „histidiini biosünteetiline protsess“ (GO:0000105). Termil on g:Profiler’i andmetel kaks annotatsiooni, mis on kõik 2009. aastast alates olnud ainult automaatselt tõestatud. Sellel termil on kolm alluvat ning tegemist on suhteliselt väikese termiga, mis viitab sellele, et tegu on spetsiifilise valdkonnaga, mida on vähe uuritud.

Kuna eelmainitud term sisaldas ainult kahte annotatsiooni, siis kui suured on teised ainult automaatselt tõestatud tõendikoodidega termid. Joonisel 9 on x-teljel esitatud termide suurused, mis näitab mahutavuse mõistes vaid esimest 15 suurusjärku. Jooniselt selgub, et termid, mille annotatsioonid on tõestatud ainult elektrooniliselt, on väikesed. Suuremas osas on need ühe annotatsiooniga ning selliseid terme on kõige rohkem bioloogilises protsessis. See viitab sellele, et erinevus selliste termide suuruste arvus võib tuleneda valdkonna uurimise populaarsusest. Kõige suuremal sellisel termil on 395 annotatsiooni, mis kõik on IEA tõendikoodiga tõestatud (term GO:0050911).



Joonis 9. 2020-07-07 versiooni ainult IEA-ga tõestatud termid suuruse järgi

Kui eelnevalt moodustus termi elektrooniliste annotatsioonide osakaal 100% siis edasi vaadeldakse elektrooniliste annotatsioonide osakaalu kõikides termides. Lisas 1, Joonisel 16 on näidatud, et terme, mille annotatsioonidest 0-24% moodustavad elektroonilised annotatsioonid, on 13 738. See tähendab, et suur osa termide annotatsioonidest on siiski peale IEA tõendikoodi tõestatud veel ka teise tõendikoodi poolt.

4. Elektrooniliste annotatsioonide kvaliteedi hindamine

4.1 Ülesande püstitus

g:Profiler'i andmete ülevaatest selgus, et automaatsete annotatsioonidega tõendikoodide arv ületab igas versioonis eksperimentaalsed tõendikoodid (vt Joonis 6). Automaatsete tõendikoodidega annotatsioone on palju ja need sisaldavad olulist informatsiooni. Paraku arvestatakse selliseid annotatsioone sageli vähem või jäetakse uurimustest välja [4, 25].

Kõige hilisem annotatsioonide hinnang tehti 2020. aasta septembris, mille tulemusena parandati 289 manuaalset annotatsiooni ning üle 52 700 elektroonilist annotatsiooni [46]. Kuivõrd elektroonilisi annotatsioone parandati rohkem kui manuaalseid, tekib küsimus, kui kvaliteetsed on elektroonilised annotatsioonid olnud läbi aja?

Varasemalt on teiste andmebaaside, nt UniProt [31] näitel uuritud automaatseid annotatsioone ja nende kvaliteeti [4], kuid g:Profiler'i andmete põhjal ei ole elektroonilisi annotatsioone hinnatud. Lisaks oli UniProt andmetel kvaliteedi hinnang tehtud aastavahemiku 2006-2011 põhjal, millest on nüüdseks juba mitu aastat möödunud. Elektrooniliste annotatsioonide kvaliteedi hinnang annab ülevaate nende usaldusväärsusest, sest paljud uued annotatsioonid põhinevad just IEA meetoditel. Tulenevalt IEA meetodil määratud annotatsioonide suurest osakaalust on nende kvaliteedi hindamine oluline ülesanne g:Profiler'i muutuste kirjeldamisel.

Bakalaureusetöö teine ülesanne on g:Profiler'i 2009.-2020. aasta andmete põhjal anda hinnang inimesega seotud GO elektrooniliste annotatsioonide (IEA) kvaliteedile võrreldes neid eksperimentaalsetega.

4.2 Analüüsi teostus

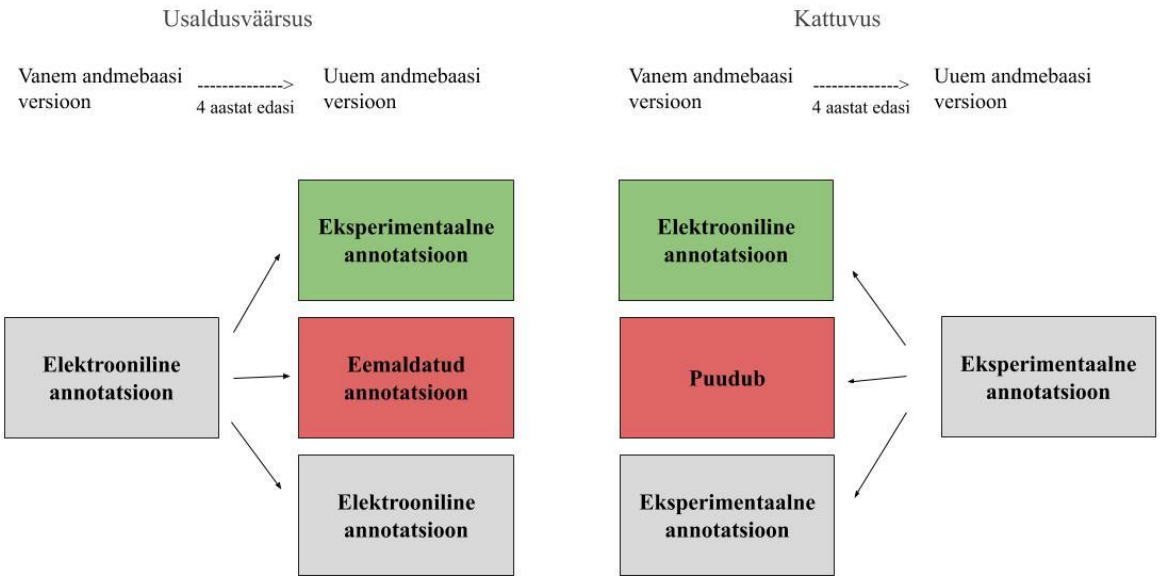
Elektrooniliste annotatsioonide kvaliteedi uurimine põhineb 2012. aastal avaldatud artiklil [4]. Artiklis võrreldi UniProt-GOA andmeid kattuvate nelja-aastaste intervallidega (2006-2009, 2007-2010, 2008-2011). Vaadeldi mitmete organismide geeniontoloogia andmeid. Erinevalt käesolevas tööst, artiklis kasutati termide ja ontoloogia struktuuri saamiseks 2011. aasta jaanuari OBO-XML faili ning annotatsioonid laeti alla Euroopa Bioinformaatika Instituudi (EBI) FTP leheküljelt. Töö eesmärgiks oli hinnata UniProt-GOA andmebaasi andmete alusel elektroonilisi annotatsioone.

Käesolevas bakalaureusetöös kasutatakse eelpoolmainitud artiklis kirjeldatud metoodikat, kasutades selleks g:Profiler'i andmeid, kohandades neid artiklis esitatud ning käesolevas töös püstitatud tingimustele. Vastavalt antud uuringus [4] esitatud meetoditele jaotatakse esmalt kõik annotatsioonide tõendikoodid kolme mitteülekattuvasse kategooriasse, kus sulgudesse on toodud sinna kategooriasse kuuluvad tõendikoodid:

1) exp - Eksperimentaalsed annotatsioonid (IDA, IPI, IMP, IGI, IEP, HTP, HMP, HDA, HGI, HEP);
2) non_exp - Mitteeksperimentaalselt tõestatud annotatsioonid (ISS; NAS, TAS, IC, RCA);
3) IEA - Elektroonilised annotatsioonid (IEA).

Artikkel [4] on avaldatud 2012. aastal, mil polnud veel HTP kategooria tõendikoode ning kõik eksperimentaalsed tõendikoodid olid ühiselt koos. Erinevalt artikli kategooriatele, tuleb g:Profiler'i andmete puhul eksperimentaalsete tõendikoodide kategooria alla arvestada ka 2018. aastal juurde tulnud HTP kategooria tõendikoodid (vt Joonis 6).

Elektrooniliste annotatsioonide kvaliteedi hindamiseks on uuringus [4] loodud kolm mõõtu, mis on tõlgitud eesti keelde vastavalt: usaldusväärsus (ingl k *reliability*), kattuvus (ingl k *coverage*) ja spetsiifilisus (ingl k *specificity*). Usaldusväärsus näitab, kui suur osa elektroonilistest annotatsioonidest on hiljem tõestatud eksperimentaalse tõendikoodi poolt. Kattuvus näitab, kui suures ulatuses suudavad automaatsed annotatsioonid hinnata ette eksperimentaalseid annotatsioone. Spetsiifilisus on vaadeldava termi suurus ja informatiivsus. Usaldusväärsuse ja kattuvuse põhimõte on näidatud Joonisel 10 ning valemid tähistatud vastavalt (1) ja (2).



Joonis 10. Usaldusväärsus ja kattuvus (Joonis kohandatud artikli joonise järgi [4])

Kuna vastavad mõõdud baseeruvad kahe versiooni võrdlusel, siis g:Profiler'i andmeid vaadeldakse nelja aasta kaupa, kus samm edasi toimub ühe aasta võrra. Näiteks võrreldakse 2009. aasta viimase versiooni (2009-02-02) annotatsioonide muutust 2012. aasta viimase versiooniga (2012-11-09). Seejärel liigutakse ühe aasta võrra edasi ning vaadeldakse 2010 ja 2013. aastate viimaseid versioone. Sama liikumist kasutati ka artiklis [4].

Joonisel 10 olevad värvid tähistavad annotatsiooni muutust. Kui elektrooniline annotatsioon ei muutu (hall värv), ei muutu ka tulemus ning seetõttu selliseid annotatsioone ei arvestata. Kui elektrooniline annotatsioon kinnitatakse eksperimentaalselt siis see lisab usaldusväärsust (roheline märgistus). Elektroonilise annotatsiooni puudumine mõjub negatiivselt väärtusele (punane). Joonisel 10 esinevat värviskeemi on kasutatud valemites (1) ja (2).

Usaldusväärsuse arvutamiseks loetakse failist kokku ühe termi (GO_i) kõik elektroonilised annotatsioonid, mis kinnitatakse 4 aastat hiljem uuemas versioonis eksperimentaalselt ning jagatakse kinnitatud annotatsioonide ja eemaldatud annotatsioonide summaga (vt valem 1) [4]. Usaldusväärsus esitatakse skaalal 0-1, kus null tähistab kõige madalamat ning üks kõige kõrgemat usaldusväärsust. g:Profiler'i andmete uurimisel ei arvestatud *NOT-qualifier* tähistusega terme, sest selliseid terme ei kaasata g:Profiler'i andmebaasi.

$$Usaldusväärsus\ (GO_i) = \frac{\sum kinnitatud}{\sum kinnitatud\ +\ \sum eemaldatud} \quad (1)$$

Kuna elektroonilisi annotatsioone esineb ühes versioonis rohkem kui eksperimentaalseid annotatsioone, siis kattuvus näitab eksperimentaalseid annotatsioone ennustanud elektrooniliste annotatsioonide arvu vanemas versioonis. Kattuvuse arvutamiseks loetakse kokku termi kõik annotatsioonid, mis on eksperimentaalsed uuemas versioonis ning jagatakse õigesti ennustatud annotatsioonide ja puuduvate annotatsioonide summaga (vt valem 2). Kattuvus näitab osakaalu skaalal 0-1, kuidas suudavad varem tehtud elektroonilised annotatsioonid ette ennustada eksperimentaalseid annotatsioone.

$$Kattuvus\ (GO_i) = \frac{\sum ennustatud}{\sum ennustatud\ +\ \sum puuduv} \quad (2)$$

Spetsiifilisuse arvutamine on näidatud valemis (3). Spetsiifilisus näitab kui suure osa ühe termi annotatsioonid moodustavad kõikidest annotatsioonidest. Spetsiifilisus jääb numbriskaalale, kus suurem number tähendab spetsiifilisemat termi. Näiteks termil GO:1990907 on 11 annotatsiooni ning spetsiifilisus 17,62. Termil GO:0035556 on 2886 annotatsiooni ning spetsiifilisus 9,58.

$$Spetsiifilisus\ (GO_i) = -log_2\left(\frac{GO_i\ annotatsioonide\ arv}{kõik\ annotatsioonid}\right) \quad (3)$$

Analüüsi tegemiseks on püstitatud tingimused [4]. Kolme mõõdu arvutamiseks vaadatakse ainult terme, mis esinevad mõlemas võrreldavas versioonis ehk nii vanemas kui ka uues. Usaldusväärsuse arvutamiseks võetakse arvesse termid, millel on vähemalt 10 eksperimentaalset annotatsiooni uuemas versioonis. Kattuvuse arvutamisel arvestatakse, et vanemas versioonis oleks vähemalt 10 automaatset annotatsiooni. Mõõtude arvutamise funktsioon töötab järgmiselt: esimesena tähistatakse kõik tõendikoodid vastavalt, kas nad on exp, non_exp või IEA. Seejärel rakendatakse funktsiooni, mis hakkab lugema usaldusväärsust, kattuvust ja spetsiifilisust nelja aasta kaupa, liikudes peale igat võrdlust ühe aasta võrra edasi.

Usaldusväärsusest, kattuvusest ja spetsiifilisusest esitatakse, nagu ka aluseks olevas artiklis [4], kolm ülevaatlikku karpdiagrammi joonist, mille tulemused kajastavad kolme valemi põhjal saadud tulemusi (vt valem 1, 2, 3). Karpdiagrammid on kirjeldava statistika teostuseks ning numbriliste andmete näitamiseks kvartiilide kaupa. Saadud Altair karpdiagrammid näitavad kokkuvõtlikult viite tulemust: miinimum ja maksimum väärtusi, mediaani ning tulemuste esimest ja kolmandat kvartiili. Kastidest väljaulatuvad jooned tähistavad kvartiile ning punktid erindeid, mis jäävad kvartiilidest välja.

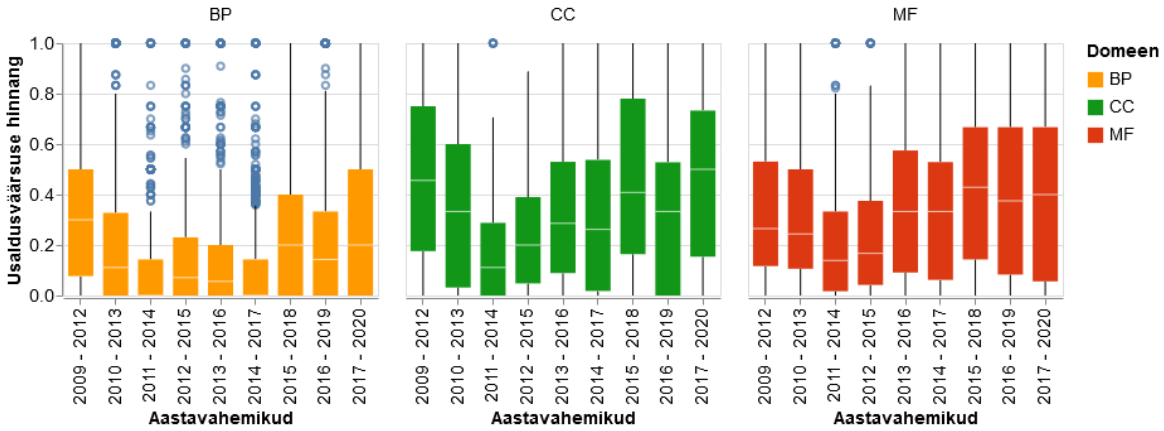
Järgnevalt analüüsitakse põhjalikumalt kõige värskema uuritava ajavahemiku (2017-2020) mõõte. Selleks võrreldakse termide spetsiifilisuse seost nende usaldusväärusega ning lisaks näidatakse usaldusväärsuse ja kattuvuse vahelisi seoseid hajuvusdiagrammiga. Kuivõrd g:Profiler'ist saadud andmed põhinevad ainult inimorganismil ning huviorbiidis on ainult elektroonilised annotatsioonid, ei olnud otstarbekas teisi artiklis esitatud analüüse järgi teha.

4.3 Elektrooniliste annotatsioonide kvaliteedi tulemused

4.3.1 g:Profiler'i tulemused

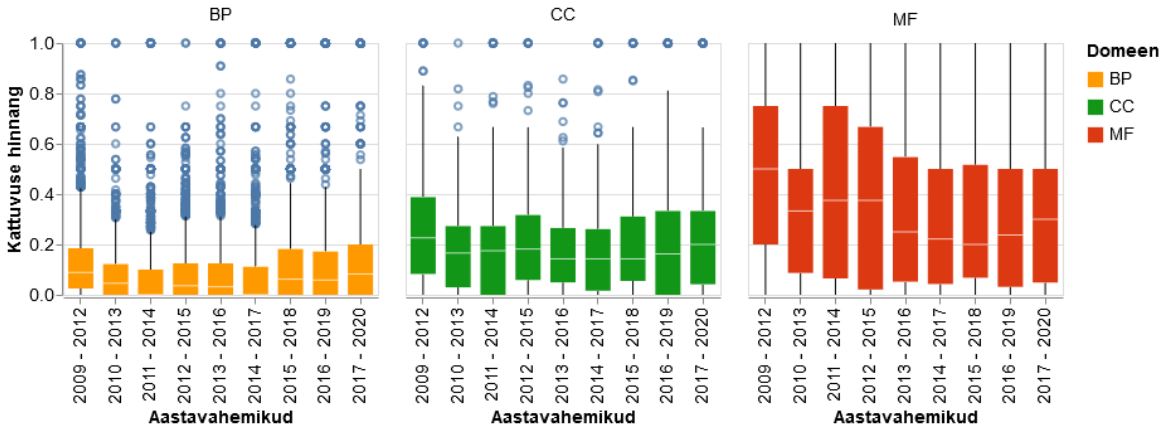
Erinevalt artiklis [4] tehtud joonisele, jaotatatakse g:Profiler'i tulemused ülevaate saamiseks kolme domeeni vahel. Esimesena vaadeldakse usaldusväärsust (vt Joonis 11). Jooniselt 11 selgub, et sõltumata domeenist hakkas 2009 – 2012 alates termide mediaan usaldusväärsus langema ning oli madalaim vahemikus 2011 – 2014. Seejärel elektrooniliste annotatsioonide usaldusväärsus kasvas mõnevõrra.

Järgnevalt vaadeldakse karpdiagrammide usaldusväärsust analüüsides ainult viimast kolme ajavahemikku (2015-2018, 2016-2019, 2017-2020). Bioloogilise protsessi (BP) elektroonilised annotatsioonid on g:Profiler'i andmete alusel kõige vähem usaldusväärsed ning esineb väiksem tõenäosus, et need eksperimentaalselt hiljem tõestatakse. Mediaan väärtuste alusel on rakulise komponendi (CC) ja molekulaarse funktsiooni (MF) usaldusväärsus parem, kuid siiski alla 0,5.



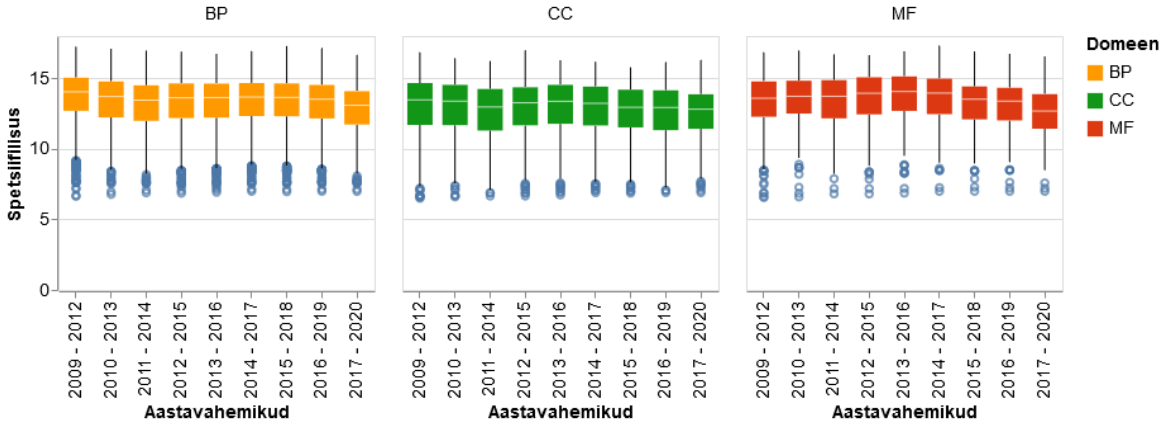
Joonis 11. Elektrooniliste annotatsioonide hinnang parameetri usaldusväärsus põhjal

Seejärel vaadatakse kattuvuse parameetri tulemusi (vt Joonis 12). Kattuvuse puhul ei mediaanväärtused suurel määral muutunud. Kõige madalama kattuvusega on bioloogiline protsess, mille mediaanväärtus on alati alla 0,2, kuid millel esineb palju erindeid. Mediaanväärtuse alusel, ennustavad eksperimentaalseid annotatsioone kõige paremini molekulaarsesse funktsiooni kuuluvad elektroonilised annotatsioonid, kuid ka nende puhul jääb mediaan kattuvus alla 0,5 (v.a 2009-2012).



Joonis 12. Elektrooniliste annotatsioonide hinnang parameetri kattuvus põhjal

Viimasena analüüsitakse elektrooniliste annotatsioonide spetsiifilisust (vt Joonis 13). Inimese andmete põhjal on termide mediaan-spetsiifilisus läbi aegade olnud stabiilselt sama. Mediaan termi spetsiifilisus on aastate jooksul olnud vahemikus 13-14 kõikide domeenide puhul. Näiteks bioloogilisse protsessi kuuluva termi „rakutsükli kontrollpunkt“ (GO:0000075) spetsiifilisus vahemikul 2017-2020 on 13,31. Termil on 229 annotatsiooni. Termi „molekulaarne funktsioon“ (GO:0003674) spetsiifilisus on 6,99. Termil on 17 277 annotatsiooni. Inimorganismi puhul pole termide spetsiifilisus märkimisväärselt muutunud.



Joonis 13. Elektrooniliste annotatsioonide hinnang parameetri spetsiifilisus põhjal

Samasugust kolme mõõdu abil tehtud analüüsi katsetati ka iga g:Profiler'i versiooni paari võrreldes, kuid kõigi parameetrite tulemused tulid väga madalad (nulli lähedased). See oli ka oodatav tulemus, sest nii lühikese ajaga (kohati paar kuud) ei jõuta teha piisavalt eksperimentaalseid katseid, et elektroonilisi annotatsioone tõestada.

Tulemuseks, kõige madalam elektrooniliste annotatsioonide kvaliteet on bioloogilisel protsessil ning kõige kõrgem molekulaarsel funktsioonil ja rakulisel komponendil. Põhjuseks võib olla, et bioloogilisi protsesse on keeruline laboris või eksperimentaalselt tõestada ja molekulaarseid funktsioone on kõige kergem tõestada [4, 48].

Järgmisena uuritakse ajavahemikul 2017-2020 olevate termide spetsiifilisuse ja usaldusväärsuse omavahelist seost (vt Lisa 1, Joonis 17). Väiksemad termid on üle terve joonise jaotunud. See tähendab, et väiksemal termil võib usaldusväärsus olla suurema tõenäosusega 1 või 0. Bioloogilises protsessis on üldiste termide usaldusväärsus madalam kui rakulises komponendis või molekulaarses funktsioonis. Ühegi üldisema termi (spetsiifilisus < 9,5) usaldusväärsus ei ole üle 0,6.

Viimasena võrreldakse 2017. aasta versiooni kvaliteeti 2020. aasta versiooni alusel (vt Lisa 1, Joonis 18). Joonis näitab elektrooniliste annotatsioonide, mis on tõestatud eksperimentaalselt, seost õigesti ennustatud annotatsioonidega. Domeenide usaldusväärsuse keskmine on sarnane, kuid kattuvuse keskmine varieerub domeenide vahel. Bioloogilise protsessi termidel on kõige madalam keskmine kattuvus ning molekulaarse funktsiooni termidel kõrgeim keskmine kattuvus, mis on kooskõlas g:Profiler'i karpdiagrammide tulemustega.

4.3.2 Võrdlus artikli tulemustega

Tulemusi võrreldes vaadeldakse ainult karpdiagrammidel olevaid mediaan väärtusi. Esmalt tutvustatakse UniProt-GOA analüüsi tulemusi ajavahemikel (2006-2009, 2007-2010, 2008-2011) [4]. Tulemuseks järeldati, et elektrooniliste annotatsioonide usaldusväärsus on tõusnud. Samas elektroonilise annotatsioonide kattuvus langes UniProt-GOA andmete alusel mõnevõrra. Spetsiifilisuse osas leiti, et uuemad UniProt-GOA versioonid sisaldavad natukene spetsiifilisemaid terme.

Spetsiifilisus tuli mõlemate andmebaaside tulemusena erinev, kus UniProt-GOA annotatsioonide [4] puhul saadi termide spetsiifilisuseks enamasti ~8. g:Profiler'i elektroonilistel annotatsioonidel on spetsiifilisus ~13 (vt Joonis 13). See võib olla vaadeldavate andmete tõttu kuna g:Profiler'is vaadeldi konkreetselt inimese andmeid.

Järgnevalt võrreldi termi spetsiifilisuse seost usaldusväärsusega. Artiklis [4] vaadeldi 2008-01-16 elektrooniliste annotatsioonide tulemusi ning leiti, et elektrooniliste annotatsioonide üldised termid on ühes vahemikus, kuid spetsiifilisemad termid varieeruvad üle skaala. Sarnastele tulemustele jõuti vaadeldes g:Profiler'i 2017-2020 vahemiku elektroonilisi annotatsioone (vt Lisa 1, Joonis 17). Joonisel on näha, et spetsiifilisemad termid esinevad rohkelt 0 või 1 usaldusväärsuse väärtustel.

Seejärel vaadeldi elektrooniliste annotatsioonide usaldusväärsuse ja kattuvuse seost kõigi kolme mõõdu näitel. Artiklis [4] võrreldi 2008-01-16 kvaliteeti 2011-01-11 versiooni tulemustega. Tulemuseks saadi, et elektrooniliste annotatsioonide usaldusväärsus on samalaadne, kuid erineb kattuvuse osas ning bioloogilisel protsessil on madalaim kattuvus ning molekulaarsetel funktsioonidel kõrgeim. Analoogsetele tulemustele jõuti ka 8 aastat hiljem, vaadeldes g:Profiler'i ajavahemikku 2017-2020 (vt Lisa 1, Joonis 18).

Kuivõrd g:Profiler'i andmed sisaldavad ainult inimese andmeid ning esimesed andmed pärinevad 2009. aastast ei saa otseselt ühiseid järeldusi teha UniProt-GOA ja g:Profiler'i tulemustest. Lisaks pärinevad andmed erinevatest andmebaasidest.

5. Elektrooniliste annotatsioonide muutus eksperimentaalseteks annotatsioonideks

5.1 Ülesande püstitus

Eelmises ülesandes vaadeldi elektrooniliste annotatsioonide kvaliteeti ajas. Samuti on olemas g:Profiler'i elektrooniliste annotatsioonide ülevaade. Elektrooniliste annotatsioonide kvaliteet kahe järjestikuse versiooni kaupa vaadeldes on väga madal (nulli lähedane). Hüpoteesina, eksperimentaalsete annotatsioonide tõestamine võtab keskmiselt rohkem aega, aga kahe g:Profiler'i versiooni vahel on ainult mõned kuud. Kuna g:Profiler'i versioone on 10 aasta jagu, siis on võimalik uurida elektrooniliste annotatsioonide muutumist eksperimentaalseteks.

Esimesena vaadeldakse, mitu versiooni võtab keskmiselt aega, et elektrooniline annotatsioon tõestataks eksperimentaalselt ning arvutatakse väärtus iga termi kohta. Tulemuseks peaks selguma termid, millega on keskmiselt läinud rohkem aega, et saada eksperimentaalseid annotatsioone ning termid, millel eksperimentaalne tõestus järgnes kiiremini. Seejärel vaadatakse, kui kaua on viimases versioonis olevad eksperimentaalsed annotatsioonid olnud eksperimentaalselt tõestatud.

5.2 Analüüsi teostus

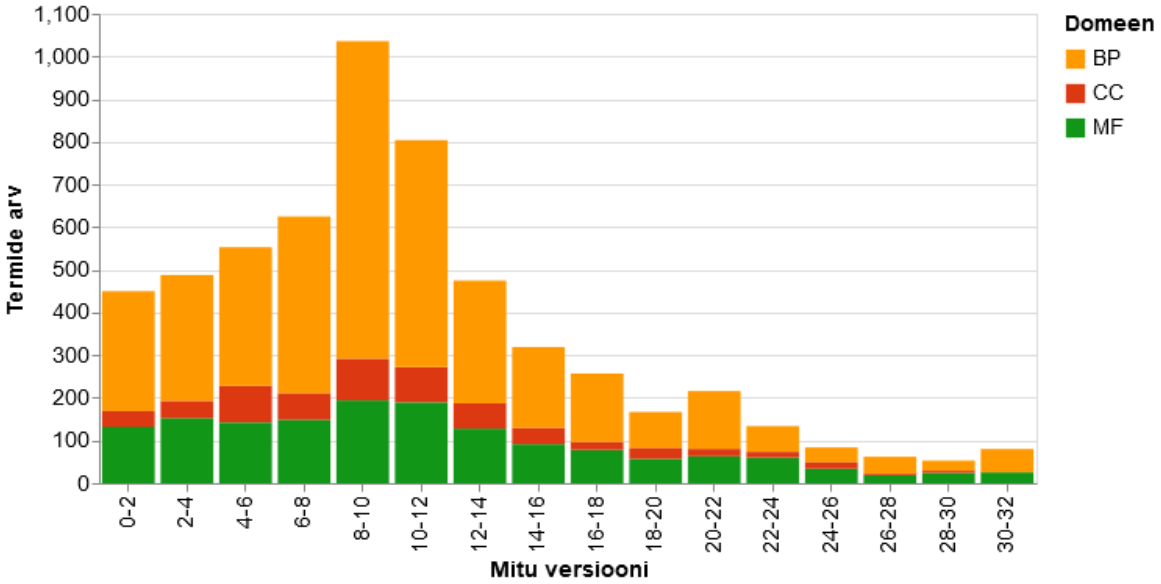
Esmalt uuritakse, mitu versiooni läheb keskmiselt aega, et annotatsiooni tõendikood muutuks IEAst eksperimentaalseks tõendikoodiks (vt tabel 1). Selleks arvutatakse keskmine versioonide arv, kui kaua võtab aega kuni automaatne annotatsioon tõestatakse eksperimentaalselt. Selline arvutus tehti vaid annotatsioonide kohta, mis on viimases versioonis juba eksperimentaalselt tõestatud. Arvestamata jäid annotatsioonid, millel IEA annotatsioon ei muutu eksperimentaalseks või on kohe eksperimentaalne. Versioonide lugemine algab iga annotatsiooni puhul esimese IEA tõendikoodiga, mis on annotatsioonil kuni esimese eksperimentaalse tõestuseni. Seejärel arvutatakse iga termi keskmine versioonide arv.

Järgmisena uuritakse, millal kõige uuema versiooni eksperimentaalsed annotatsioonid olid viimati elektrooniliselt tõestatud. Selleks loetakse kokku viimases versioonis kõik eksperimentaalsete tõendikoodidega annotatsioonid. Siis kontrollitakse, mitu versiooni tagasi esines viimane IEA tõendikood. Väärtus arvutatakse iga termi kohta. Eksperimentaalsed tõendikoodid on toodud tabelis 1.

5.3 Tulemused

Esimesena uuris autor versioonide arvu, kaua võtab aega kuni annotatsioon tõestatakse ära eksperimentaalse tõendikoodi poolt. Joonisel 14 on toodud termide keskmised tulemused. Arvutuste tulemusena läheb ühel termil keskmiselt aega 10 versiooni. g:Profiler'is tehakse aastas 3-5 uuendust ehk annotatsiooni eksperimentaalseks tõestamiseks läheb umbes 2-3 aastat. Kõige sagedasem tulemus oli keskmiselt 8-10 versiooni (1036 termi) ning kõige rohkem on selliseid bioloogilise protsessi terme (745). See tähendab, et alates esimesest elektroonilisest annotatsioonist läks aega 8-10 versiooni, kuni elektrooniline annotatsioon tõestati eksperimentaalselt.

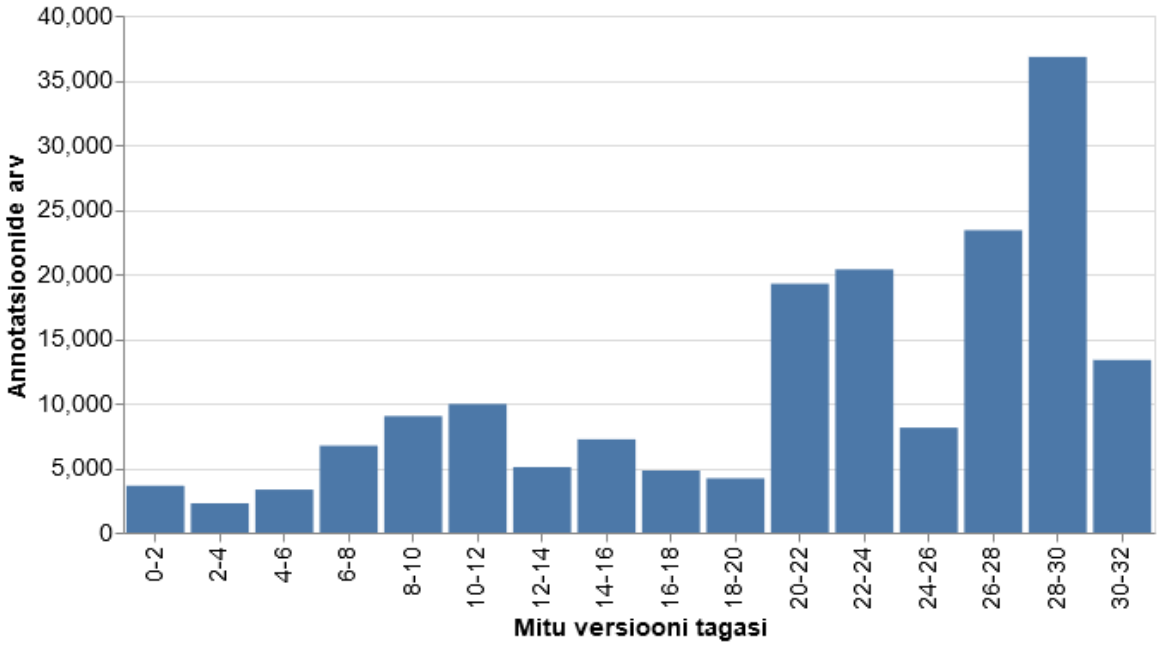
Sellise keskmisega termide alla kuuluvad näiteks bioloogilise protsessi term „nukleotiididega seondumine" (GO:0000166). Termi tõestamine eksperimentaalselt võttis g:Profiler'i andmetel keskmiselt aega ligi 11 versiooni. Termil esines esimene automaatne annotatsioon geeniga ENSG00000204389 versioonis 2009-02-02 ning 10 versiooni hiljem kuupäeval 2015-02-06 ka eksperimentaalselt (IDA tõendikood).



Joonis 14**.** Mitu versiooni läheb keskmiselt aega, et geeni annotatsiooni tõendikood muutuks IEA-st EXPiks (termi kohta)?

Eksperimentaalse annotatsioonini läks 450 termi puhul aega vaid 0-2 versiooni (vt Joonis 14). Keskmiselt 30-32 versiooni (8-10 aastat) läks aega 81 termil ning rakukomponendi terme sinna ei kuulunudki. Sellised termid on näiteks bioloogilisse protsessi kuuluv term „oksüdatiivne fosforüülimine“ (GO:0006119).

Järgnevalt vaadeldakse kõiki eksperimentaalseid annotatsioone versioonis 2020-07-07 ning uuritakse, mitu versiooni tagasi eksperimentaalsed tõendikoodid olid viimati automaatsed ehk kui uute eksperimentaalsete annotatsioonidega tegu on. Joonisel 15 on toodud tulemused viimase uuenduse seisuga.



Joonis 15. Millal 2020-07-07 versiooni eksperimentaalsed annotatsioonid olid viimati elektrooniliselt tõestatud?

Alates 2010. aastast on 36 811 annotatsiooni olnud eksperimentaalselt tõestatuna.

Versioonis 2010-02-02 esines annotatsioonil üks IEA tõendikood:

- GO:0090304 ENSG00000196584 IEA

Kõige uuemas versioonis 2020-07-07 on annotatsioon järgnev:

- GO:0090304 ENSG00000196584 IDA|IMP|IGI|IBA|TAS|IEA.

Viimase kahe uuendusega tekkis 3623 uut eksperimentaalset annotatsiooni, mis olid eelnevalt IEA. Mitmete selliste annotatsioonide juures on annotatsioon varasemalt olnud tõestatud ainult IEA tõendikoodidega. Teisel juhul on annotatsioon tekkinud juurde viimastel aastatel, näiteks term GO:0001401 annoteeriti geeniga ENSG00000128654 versioonis 2018-10-02.

Kokkuvõte

g:Profiler'i andmebaas sisaldab andmeid geeniontoloogia kohta. g:Profiler'it kasutatakse teadustöödes, sisestades g:Profiler'isse funktsionaalsuse analüüsiks geeniliste. Kuigi g:Profiler'i andmete uuendusi tehakse aastas mitu korda, puudub ülevaade kasutatavatest geenidest, funktsioonidest ning nende muutustest. Seetõttu on kasulik teha andmetest ülevaade. Samuti, kuna elektroonilisi annotatsioone on geeniontoloogia statistika alusel palju, tasub uurida nende esinemist ning hinnata kvaliteeti.

Käesoleva bakalaureusetöö eesmärkideks oli täita sissejuhatuses püstitatud kolm eesmärki. Töö käigus anti ülevaade g:Profiler'i veebiserveris esinenud andmetest ja elektroonilistest annotatsioonidest. Hinnati elektrooniliste annotatsioonide kvaliteeti ning uuriti elektrooniliste annotatsioonide muutumist eksperimentaalseteks. Töö autorile teadaolevalt ei oldud g:Profiler'i andmeid sellistel eesmärkidel varasemalt uuritud. Andmete analüüsimiseks kasutati andmekaeve meetodeid.

Töös leiti, et geeniontoloogia on g:Profiler'i andmetel ajas muutunud. Annotatsioonide ja termide arv on 2009. aastast alates kasvanud, eelkõige bioloogilise protsessi puhul. Geenide arv on peamiselt jäänud samaks. Jaotades annotatsioonid tõendikoodi kaupa leiti, et automaatseid tõendikoode sisaldavaid annotatsioone on rohkem kui eksperimentaalsete tõendikoodidega annotatsioone. Samuti on mõlema kategooria annotatsioonide arv kasvanud ajas. HTP kategooria tõendikoodidega annotatsioonid esinevad g:Profiler'i versioonides alates 2018. aastast.

Täpsem ülevaade annotatsioonidest näitas, et kõige uuema versiooni puhul esineb enim elektrooniliste annotatsioonidega terme bioloogilises protsessis. Selliste annotatsioonidega termide suurus näitas, et need termid on peamiselt ühe annotatsiooniga. Paljud termide annotatsioonid on siiski peale IEA tõestatud veel teise tõendikoodi poolt.

Elektrooniliste annotatsioonide kvaliteedi hinnang põhines meetoditel, kus kvaliteeti hinnati kolme parameetri põhjal: usaldusväärsus, kattuvus ja spetsiifilisus. Tulemusena, bioloogilise protsessi elektroonilistel annotatsioonidel on kõige madalam kvaliteet ning molekulaarse funktsiooni ja rakukomponendi termidel kõrgeim. Termide spetsiifilisus oli läbi g:Profiler'i versioonide ~13. Edasises töös võiks ajaperioodide erinevuste võrdlemiseks läbi viia ka statistilised testid.

Hüpoteesiks püstitati, et eksperimentaalsete annotatsioonide tõestamine võtab rohkem aega kui kahe g:Profiler'i versiooni vaheline aeg. Käesolevas töös läbiviidud analüüsist selgus, et keskmiselt läheb 2-3 aastat (10 g:Profiler'i versiooni), kuni automaatne annotatsioon muutub eksperimentaalseks annotatsiooniks. Samuti vaadeldi küsimust „Millal kõige uuema versiooni eksperimentaalsed annotatsioonid olid viimati elektrooniliselt tõestatud?". Üle 30 000 annotatsiooni on eksperimentaalselt tõestatuna olnud alates 2010. aastast. Viimase kahe g:Profiler'i versiooniga on tekkinud juurde ~3600 uut eksperimentaalset annotatsiooni.

Bakalaureusetöö tulemusena valmis Pythoni kood, mida on võimalik rakendada vastavalt sisestatud failidele. Loodud koodi abil on võimalik sisestada uuemaid versioonifaile ning saada selle tulemusena kirjeldava statistika teostuseks jooniseid. Järgmiste sammudena saab andmete ülevaate lisada ka g:Profiler'i kodulehele (tõlkides selle eelnevalt inglise keelde), andes kasutajatele võimaluse tutvuda andmetega ning vaadelda versioonide vahelisi erinevusi.

Lõpetuseks soovin tänada lõputöö juhendajat Liis Kolbergi tema suuniste ning abivalmiduse eest. Samuti tänan Tartu Ülikooli BIIT töörühma, kes valmistas ette tööks vajalikud failid

ning abistas omapoolsete nõuannete ja ettepanekutega lõputöö teostamisel. Lõputöö tegemine oli seetõttu arendav ja õpetlik kogemus.

Viidatud kirjandus

[1] Heinaru, A. Geneetika. Õpik kõrgkoolile. Tartu : Tartu Ülikooli Kirjastus, 2012

[2] M. Ashburner, C. A. Ball, J. Blake, D. Botstein, H. Butler, and J. Cherry, 'Gene ontology: Tool for the unification of biology', *The Gene Ontology Consortium. Nat Genet*, vol. 25, pp. 25–29, Jan. 2000.

[3] The Gene Ontology Consortium, 'The Gene Ontology Resource: 20 years and still GOing strong', *Nucleic Acids Res*, vol. 47, no. D1, pp. D330–D338, Jan. 2019, doi: 10.1093/nar/gky1055.

[4] N. Škunca, A. Altenhoff, and C. Dessimoz, 'Quality of Computationally Inferred Gene Ontology Annotations', *PLOS Computational Biology*, vol. 8, no. 5, p. e1002533, May 2012, doi: 10.1371/journal.pcbi.1002533.

[5] L. du Plessis, N. Škunca, and C. Dessimoz, 'The what, where, how and why of gene ontology—a primer for bioinformaticians', *Briefings in Bioinformatics*, vol. 12, no. 6, pp. 723–735, Nov. 2011, doi: 10.1093/bib/bbr002.

[6] T. Hastie, R. Tibshirani, ja J. Friedman, The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media, 2009.

[7] G. Wang *et al.*, 'The analysis of risk factors for diabetic nephropathy progression and the construction of a prognostic database for chronic kidney diseases', *Journal of Translational Medicine*, vol. 17, no. 1, p. 264, Aug. 2019, doi: 10.1186/s12967-019-2016-y.

[8] N. Vadgama, D. Lamont, J. Hardy, J. Nasir, and R. C. Lovering, 'Distinct proteomic profiles in monozygotic twins discordant for ischaemic stroke', *Mol Cell Biochem*, vol. 456, no. 1, pp. 157–165, Jun. 2019, doi: 10.1007/s11010-019-03501-2.

[9] G. E. Westgate, R. S. Ginger, and M. R. Green, 'The biology and genetics of curly hair', *Experimental Dermatology*, vol. 26, no. 6, pp. 483–490, 2017, doi: https://doi.org/10.1111/exd.13347.

[10] A. D. Yates *et al.*, 'Ensembl 2020', *Nucleic Acids Research*, vol. 48, no. D1, pp. D682–D688, Jan. 2020, doi: 10.1093/nar/gkz966.

[11] NCBI Resource Coordinators, 'Database resources of the National Center for Biotechnology Information', *Nucleic Acids Res*, vol. 46, no. D1, pp. D8–D13, 04 2018, doi: 10.1093/nar/gkx1095.

[12] J. Thurmond *et al.*, 'FlyBase 2.0: the next generation', *Nucleic Acids Research*, vol. 47, no. D1, pp. D759–D765, Jan. 2019, doi: 10.1093/nar/gky1003.

[13] M. Kanehisa and S. Goto, 'KEGG: kyoto encyclopedia of genes and genomes', *Nucleic Acids Res*, vol. 28, no. 1, pp. 27–30, Jan. 2000, doi: 10.1093/nar/28.1.27.

[14] B. Jassal *et al.*, 'The reactome pathway knowledgebase', *Nucleic Acids Res*, vol. 48, no. D1, pp. D498–D503, 08 2020, doi: 10.1093/nar/gkz1031.

[15] 'Gene Ontology Resource', *Gene Ontology Resource*. http://geneontology.org/stats.html (accessed Nov. 02, 2020). "10.5281/zenodo.2529950"

[16] L. Chen, Y.-H. Zhang, G. Lu, T. Huang, and Y.-D. Cai, 'Analysis of cancer-related lncRNAs using gene ontology and KEGG pathways', *Artificial Intelligence in Medicine*, vol. 76, pp. 27–36, Feb. 2017, doi: 10.1016/j.artmed.2017.02.001.

[17] The Gene Ontology Consortium, 'The Gene Ontology (GO) database and informatics resource', *Nucleic Acids Res*, vol. 32, no. suppl_1, pp. D258–D261, Jan. 2004, doi: 10.1093/nar/gkh036.

[18] B. Smith, J. Williams, and S.-K. Steffen, 'The Ontology of the Gene Ontology', *AMIA Annu Symp Proc*, vol. 2003, pp. 609–613, 2003.

[19] D. Binns, E. Dimmer, R. Huntley, D. Barrell, C. O'Donovan, and R. Apweiler, 'QuickGO: a web-based tool for Gene Ontology searching', *Bioinformatics*, vol. 25, no. 22, pp. 3045–3046, Nov. 2009, doi: 10.1093/bioinformatics/btp536.

[20] „QuickGO::Term GO:1904092". https://www.ebi.ac.uk/QuickGO/term/GO:1904092 (vaadatud okt 28, 2020).

[21] „Gene Ontology overview", Gene Ontology Resource. http://geneontology.org/docs/ontology-documentation/ (vaadatud okt 15, 2020).

[22] The Gene Ontology Consortium, 'Gene Ontology Annotations and Resources', *Nucleic Acids Res*, vol. 41, no. D1, pp. D530–D535, Jan. 2013, doi: 10.1093/nar/gks1050.

[23] „Guide to GO evidence codes", Gene Ontology Resource. http://geneontology.org/docs/guide-go-evidence-codes/ (vaadatud dets 01, 2020).

[24] H. Attrill *et al.*, 'Annotation of gene product function from high-throughput studies using the Gene Ontology', *Database (Oxford)*, vol. 2019, Jan. 2019, doi: 10.1093/database/baz007.

[25] T. J. Buza, F. M. McCarthy, N. Wang, S. M. Bridges, and S. C. Burgess, 'Gene Ontology annotation quality analysis in model eukaryotes', *Nucleic Acids Res*, vol. 36, no. 2, pp. e12–e12, Feb. 2008, doi: 10.1093/nar/gkm1167.

[26] R. P. Huntley, T. Sawford, M. J. Martin, and C. O'Donovan, 'Understanding how and why the Gene Ontology and its annotations evolve: the GO within UniProt', *Gigascience*, vol. 3, no. 1, Dec. 2014, doi: 10.1186/2047-217X-3-4.

[27] E. H. Chang *et al.*, 'Rhinovirus Infections in Individuals with Asthma Increase ACE2 Expression and Cytokine Pathways Implicated in COVID-19', *Am J Respir Crit Care Med*, vol. 202, no. 5, pp. 753–755, Jul. 2020, doi: 10.1164/rccm.202004-1343LE.

[28] K.-M. McLaughlin *et al.*, 'COVID-19-Related Coagulopathy—Is Transferrin a Missing Link?', *Diagnostics*, vol. 10, no. 8, Art. no. 8, Aug. 2020, doi: 10.3390/diagnostics10080539.

[29] C. Cava, G. Bertoli, and I. Castiglioni, 'In Silico Discovery of Candidate Drugs against Covid-19', *Viruses*, vol. 12, no. 4, Art. no. 4, Apr. 2020, doi: 10.3390/v12040404.

[30] L. Ruzicka *et al.*, 'The Zebrafish Information Network: new support for non-coding genes, richer Gene Ontology annotations and the Alliance of Genome Resources', *Nucleic Acids Research*, vol. 47, no. D1, pp. D867–D873, Jan. 2019, doi: 10.1093/nar/gky1090.

[31] The UniProt Consortium, 'UniProt: a worldwide hub of protein knowledge', *Nucleic Acids Research*, vol. 47, no. D1, pp. D506–D515, Jan. 2019, doi: 10.1093/nar/gky1049.

[32] The Gene Ontology Consortium, 'Expansion of the Gene Ontology knowledgebase and resources', *Nucleic Acids Res*, vol. 45, no. D1, pp. D331–D338, Jan. 2017, doi: 10.1093/nar/gkw1108.

[33] J. Reimand *et al.*, 'g:Profiler—a web server for functional interpretation of gene lists (2016 update)', *Nucleic Acids Res*, vol. 44, no. W1, pp. W83–W89, Jul. 2016, doi: 10.1093/nar/gkw199.

[34] J. Reimand, M. Kull, H. Peterson, J. Hansen, and J. Vilo, 'g:Profiler—a web-based toolset for functional profiling of gene lists from large-scale experiments', *Nucleic Acids Res*, vol. 35, no. suppl_2, pp. W193–W200, Jul. 2007, doi: 10.1093/nar/gkm226.

[35] U. Raudvere *et al.*, 'g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update)', *Nucleic Acids Res*, vol. 47, no. W1, pp. W191–W198, Jul. 2019, doi: 10.1093/nar/gkz369.

[36] D. N. Slenter *et al.*, 'WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research', *Nucleic Acids Research*, vol. 46, no. D1, pp. D661–D667, Jan. 2018, doi: 10.1093/nar/gkx1064.

[37] V. Matys *et al.*, 'TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes', *Nucleic Acids Res*, vol. 34, no. Database issue, pp. D108-110, Jan. 2006, doi: 10.1093/nar/gkj143.

[38] D. W. Huang, B. T. Sherman, and R. A. Lempicki, 'Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources', *Nat Protoc*, vol. 4, no. 1, pp. 44–57, 2009, doi: 10.1038/nprot.2008.211.

[39] D. W. Huang, B. T. Sherman, and R. A. Lempicki, 'Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists', *Nucleic Acids Res*, vol. 37, no. 1, pp. 1–13, Jan. 2009, doi: 10.1093/nar/gkn923.

[40] T. Kluyver *et al.*, 'Jupyter Notebooks – a publishing format for reproducible computational workflows', in *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, 2016, pp. 87–90, doi: 10.3233/978-1-61499-649-1-87.

[41] W. McKinney, 'Data Structures for Statistical Computing in Python', Austin, Texas, 2010, pp. 56–61, doi: 10.25080/Majora-92bf1922-00a.

[42] C. R. Harris *et al.*, 'Array programming with NumPy', *Nature*, vol. 585, no. 7825, pp. 357–362, Sep. 2020, doi: 10.1038/s41586-020-2649-2.

[43] F. Cunningham *et al.*, 'Ensembl 2015', *Nucleic Acids Res*, vol. 43, no. D1, pp. D662–D669, Jan. 2015, doi: 10.1093/nar/gku1010.

[44] B. H. M. Meldal *et al.*, 'The complex portal - an encyclopaedia of macromolecular complexes', *Nucleic Acids Res*, vol. 43, no. Database issue, pp. D479–D484, Jan. 2015, doi: 10.1093/nar/gku975.

[45] I. Ezkurdia *et al.*, 'Multiple evidence strands suggest that there may be as few as 19 000 human protein-coding genes', *Human Molecular Genetics*, vol. 23, no. 22, pp. 5866–5878, Nov. 2014, doi: 10.1093/hmg/ddu309.

[46] P. Flicek *et al.*, 'Ensembl 2014', *Nucleic Acids Res*, vol. 42, no. Database issue, pp. D749–D755, Jan. 2014, doi: 10.1093/nar/gkt1196.

[47] V. Wood *et al.*, 'Term Matrix: a novel Gene Ontology annotation quality control system based on ontology term co-annotation patterns', *Open Biology*, vol. 10, no. 9, p. 200149, doi: 10.1098/rsob.200149.

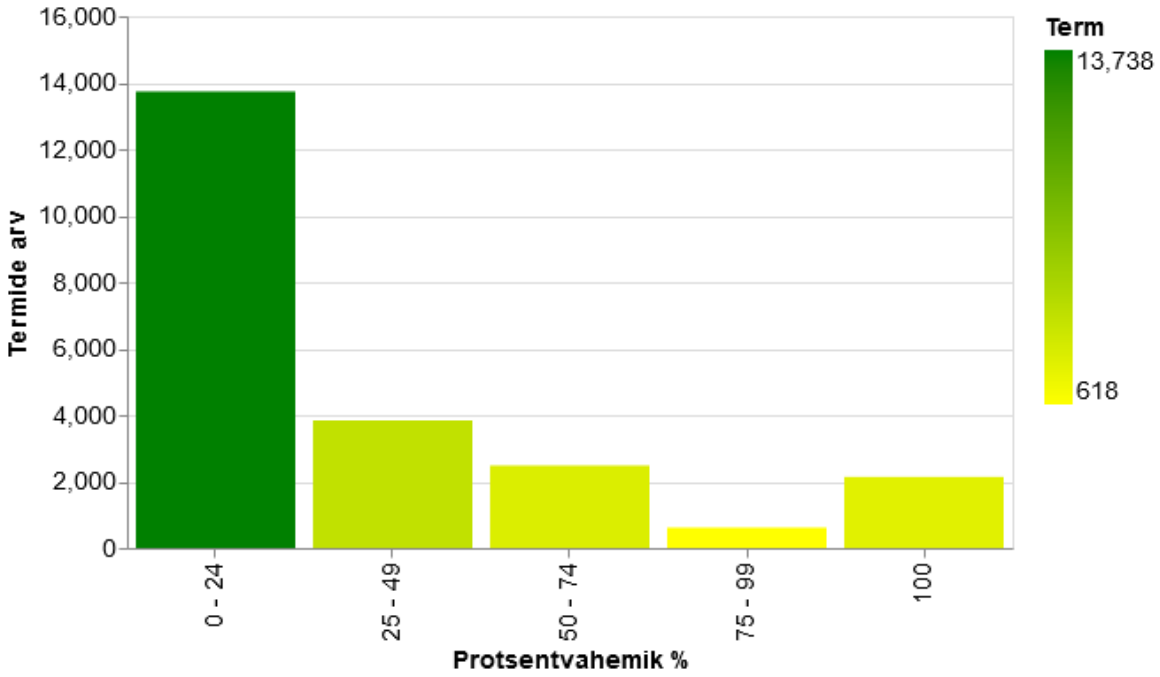
[48] P. Gaudet, M. S. Livstone, S. E. Lewis, and P. D. Thomas, 'Phylogenetic-based propagation of functional annotations within the Gene Ontology consortium', *Brief Bioinform*, vol. 12, no. 5, pp. 449–462, Sep. 2011, doi: 10.1093/bib/bbr042.

Lisad

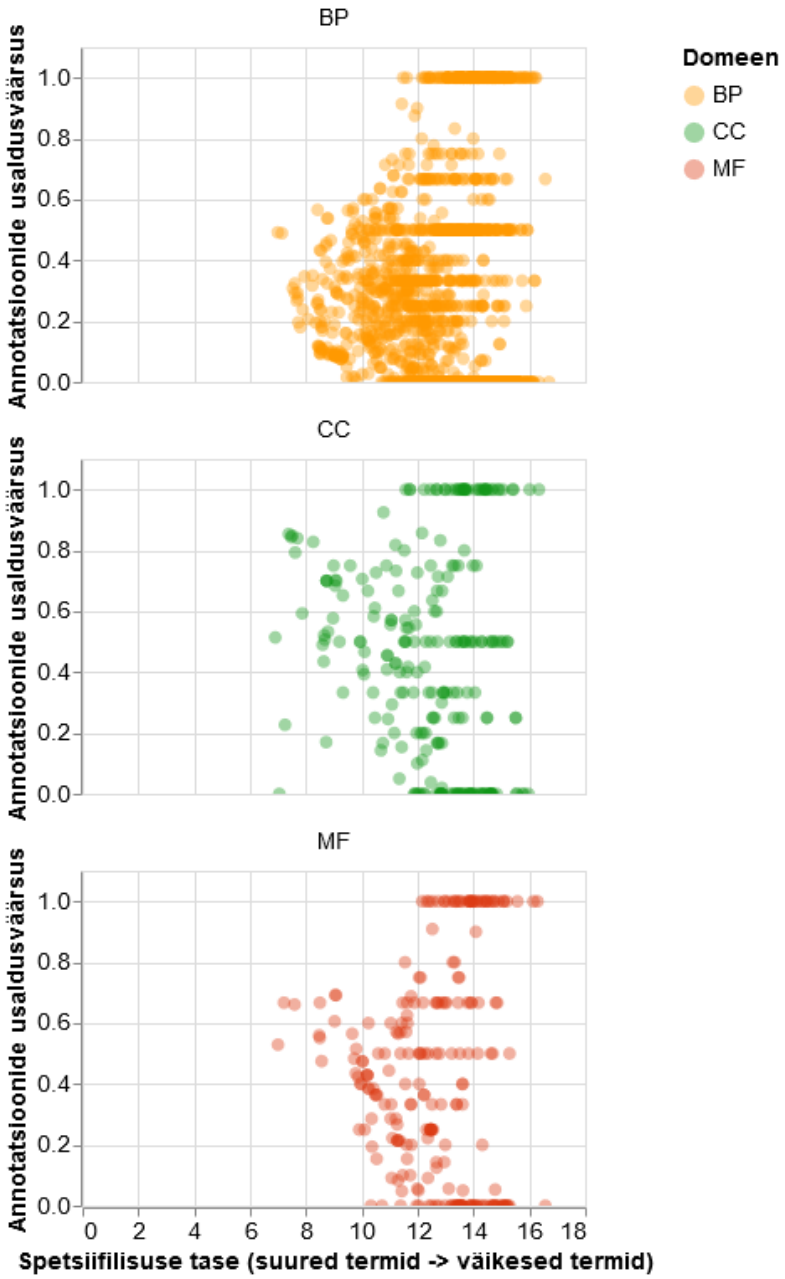
I. Lisamaterjalid

Versiooni faili nimetus	Avaldamise kuupäev (aasta-kuu-päev)
r0814_e56	2009-02-02
r0818_e57_eg4	2010-02-02
r0981_e61_eg8	2011-02-02
r0998_e62_eg9	2011-05-07
r1070_e64_eg11	2012-02-14
r1075_e65_eg12	2012-02-25
r1177_e68_eg15	2012-08-09
r1185_e69_eg16	2012-11-09
r1227_e72_eg19	2013-07-18
r1270_e75_eg22	2014-04-11
r1353_e78_eg25	2015-02-06
r1395_e79_eg26	2015-04-09
r1435_e80_eg27	2015-07-07
r1440_e81_eg28	2015-09-08
r1477_e82_eg29	2015-10-20
r1536_e83_eg30	2016-02-02
r1622_e84_eg31	2016-05-06
r1665_e85_eg32	2016-09-05
r1705_e86_eg33	2016-11-02
r1709_e87_eg34	2016-12-13
r1730_e88_eg35	2017-05-18
r1732_e89_eg36	2017-06-20
r1741_e90_eg37	2017-10-19
r1750_e91_eg38	2018-07-02
r1760_e93_eg40	2018-10-01
r2001_e94_eg41	2018-10-02
r2002_e95_eg42	2019-05-09
r2003_e96_eg43	2019-09-10
r2004_e97_eg44	2019-10-07
r2005_e98_eg45	2020-03-07
r2006_e99_eg46	2020-07-02
r2007_e100_eg47	2020-07-07

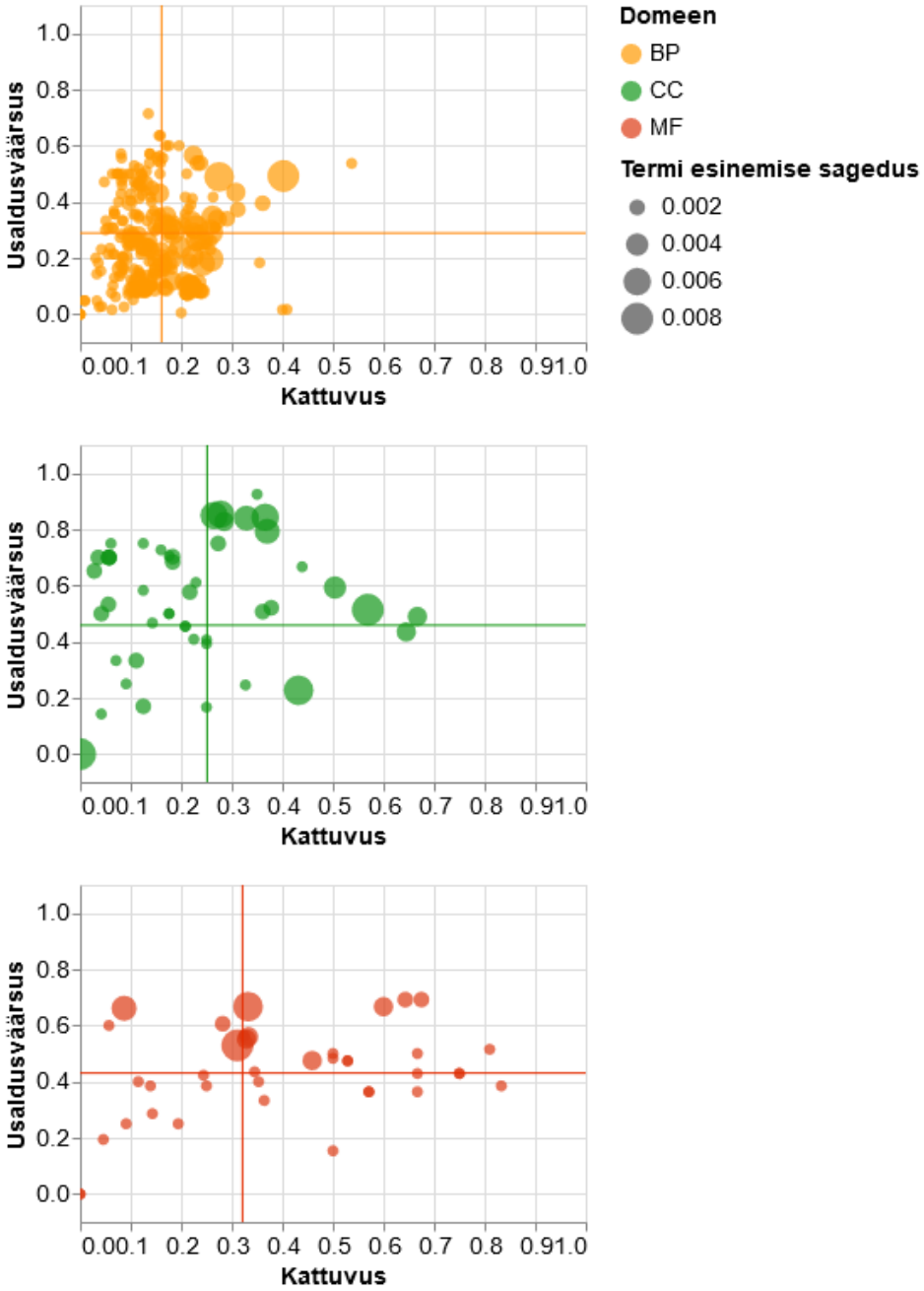
Tabel 2. g:Profiler'i versioonide faili nimetused ja avaldamise kuupäevad



Joonis 16. Kui suur protsent termi annotatsioonidest moodustab IEA



Joonis 17. Spetsiifilisuse seos usaldusväärsusega ajavahemikul 2017-2020. Iga punkt joonisel tähistab ühte GO termi [4]



Joonis 18. Kattuvuse võrdlus usaldusväärsusega ajavahemikul 2017-2020. Keskmise tähistamiseks on joonisel horisontaalsed ja vertikaalsed jooned. Ringi suurused näitavad termi suurust.

II. Koodi repositoorium

Analüüsi lähtekood asub GitLabi repositooriumis aadressil https://gitlab.cs.ut.ee/Mlepson/l-put.

Analüüsi interaktiivsetele joonistele pääseb ligi lingil http://mlepson.gitlab.cs.ut.ee/l-put/.

Litsents

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Marielle Lepson,

1. Annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose „Geeniontoloogia andmete muutus ajas g:Profiler'i näitel", mille juhendaja on Liis kolberg, reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada digitaalarhiivi DSpace kuni autoriõiguse kehtivuse lõppemiseni.

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

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

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

Marielle Lepson

14.01.2021