Odkrivanje zakonitosti iz literature kot pomoč pri interpretaciji podatkov pridobljenih z metodami visokozmogljivega sekvenciranja

Dimitar Hristovski[1] , Gaber Bergant[2], Andrej Kastrin[1], Borut Peterlin[2]

15. november 2018

[1]Univerza v Ljubljani, Medicinska fakulteta, Inštitut za biostatistiko in medicinsko informatiko
[2]Univerzitetni klinični center Ljubljana, Klinični inštitut za medicinsko genetiko

Motivacija

- Metode sekvenciranja nove generacije imajo velik potencial, toda ...

Motivacija

- Metode sekvenciranja nove generacije imajo velik potencial, toda ...
- Rezultati težki za interpretacijo (zlasti za diagnostične namene)

Motivacija

- Metode sekvenciranja nove generacije imajo velik potencial, toda ...
- Rezultati težki za interpretacijo (zlasti za diagnostične namene)
- Cilj: Razvoj bioinformatičnega orodja za podporo kliničnemu diagnostičnemu procesu

Slovarček izrazov

- Sekvenciranje naslednje generacije (NGS)

Slovarček izrazov

- Sekvenciranje naslednje generacije (NGS)
- Sekvenciranje celotnega eksoma

Slovarček izrazov

- Sekvenciranje naslednje generacije (NGS)
- Sekvenciranje celotnega eksoma
- Gen, mutacija, protein

Slovarček izrazov

- Sekvenciranje naslednje generacije (NGS)
- Sekvenciranje celotnega eksoma
- Gen, mutacija, protein
- Genetska variacija

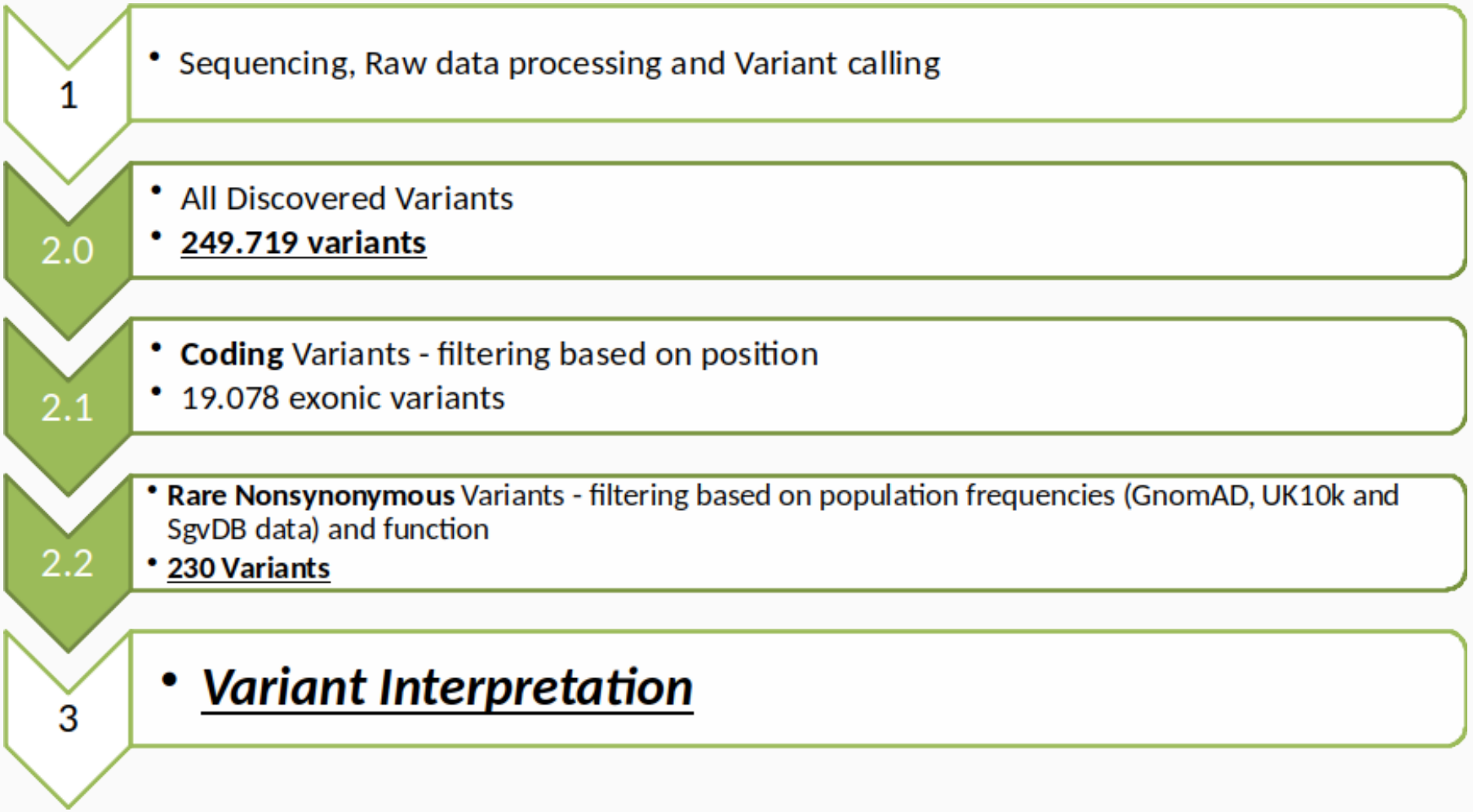
Slovarček izrazov

- Sekvenciranje naslednje generacije (NGS)
- Sekvenciranje celotnega eksoma
- Gen, mutacija, protein
- Genetska variacija
- Genotip: množica genov s specifičnimi mutacijami (operacionalna definicija)

Slovarček izrazov

- Sekvenciranje naslednje generacije (NGS)
- Sekvenciranje celotnega eksoma
- Gen, mutacija, protein
- Genetska variacija
- Genotip: množica genov s specifičnimi mutacijami (operacionalna definicija)
- Fenotip: množica kliničnih znakov (operacionalna definicija)

NGS delotok


1
• Sequencing, Raw data processing and Variant calling
2.0
• All Discovered Variants
• 249.719 variants
2.1
• Coding Variants - filtering based on position
• 19.078 exonic variants
2.2
• Rare Nonsynonymous Variants - filtering based on population frequencies (GnomAD, UK10k and SgvDB data) and function
• 230 Variants
3
• Variant Interpretation

Raziskovalna ideja

Uporaba odkrivanja zakonitosti iz literature (LBD) za izboljšanje procesa interpretacije rezultatov NGS.

Odkrivanje zakonitosti iz literature (LBD)

- Metoda za samodejno generiranje raziskovalnih domnev iz literature

Odkrivanje zakonitosti iz literature (LBD)

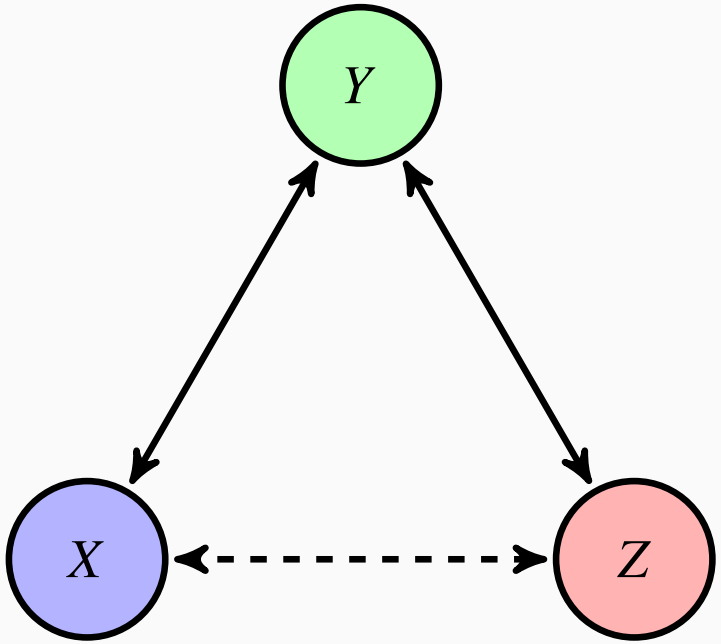
- Metoda za samodejno generiranje raziskovalnih domnev iz literature
- Vsaka domneva sledi vzorcu: `Koncept1` - `Relacija` - `Koncept2`

Odkrivanje zakonitosti iz literature (LBD)

- Metoda za samodejno generiranje raziskovalnih domnev iz literature
- Vsaka domneva sledi vzorcu: `Koncept1` - `Relacija` - `Koncept2`
- Primer: `Ribje olje` - `Zdravi` - `Raynaudjev sindrom`

Odkrivanje zakonitosti iz literature (LBD)

- Metoda za samodejno generiranje raziskovalnih domnev iz literature
- Vsaka domneva sledi vzorcu: `Koncept1` - `Relacija` - `Koncept2`
- Primer: `Ribje olje` - `Zdravi` - `Raynaudjev sindrom`
- Dva tipa iskanja: odprto ali zaprto



Podatkovni viri

- Klinični podatki o pacientih
  - genotip (iz NGS)
  - fenotip (od kliničnega genetika)

Podatkovni viri

- Klinični podatki o pacientih
  - genotip (iz NGS)
  - fenotip (od kliničnega genetika)
- Populacijski genetski podatki iz javnodostopnih zbirk
  - GnomAD projekt (140 000 eksomov iz populacije zdravih kontrol)
  - UK10k (10 000 zdravih kontrol iz Velike Britanije)
  - SgvDB (2500 kliničnih in celotnih eksomov iz Slovenije)

. . . podatkovni viri

- Napovedne vrednosti patogenosti (na osnovi prostodostopnih algoritmov, ki temeljijo na razliki med referenčnimi in alternativnimi biokemijskimi in prostorskimi lastnostmi, poziciji v proteinu itd.):
  - SIFT
  - Polyphen2
  - MutationTaster
  - PROVEAN.prediction
  - CADD.score
  - M.CAP.score

. . . podatkovni viri

- SemMedDB: distribucija semantičnih relacij izluščenih iz celotnega MEDLINE s pomočjo orodja SemRep (NLP orodje)

. . . podatkovni viri

- SemMedDB: distribucija semantičnih relacij izluščenih iz celotnega MEDLINE s pomočjo orodja SemRep (NLP orodje)
- Primer: iz stavka *dexamethasone is a potent inducer of multidrug resistance-associated protein expression in rat hepatocytes* SemRep izlušči:
  - Dexamethasone `STIMULATES` Multidrug Resistence-Associated Proteins
  - Multidrug Resistance-Associated Proteins `PART_OF` Rats
  - Hepatocytes `PART_OF` Rats

Metode

- Konstruiramo omrežje s katerim predstavimo:
  - genotip pacienta
  - fenotip pacienta
  - obstoječe (biomedicinsko) znanje

Metode

- Konstruiramo omrežje s katerim predstavimo:
  - genotip pacienta
  - fenotip pacienta
  - obstoječe (biomedicinsko) znanje
- Na osnovi omrežja bomo:
  - napovedovali nove povezave med genotipom in fenotipom
  - podali razlago za znane in nove povezave

Gradnja omrežja

Vozlišča

Gradnja omrežja

Vozlišča

- Pacienti

Gradnja omrežja

Vozlišča

- Pacienti
- Fenotipi (na osnovi *Human Phenotype Ontology*)

Gradnja omrežja

Vozlišča

- Pacienti
- Fenotipi (na osnovi *Human Phenotype Ontology*)
- Biomedicinski koncepti (s 126 podtipi):
  - argumenti semantičnih relacij izluščeni iz MEDLINE
    - Parkinsonova bolezen (`Disease or Syndrome`)
    - Levodopa (`Pharmacologic Substance`)
    - LRRK2 (`Gene or Genome`)

Povezave

Povezave

- PHENO: povezuje paciente s fenotipi

. . . gradnja omrežja

Povezave

- PHENO: povezuje paciente s fenotipi
- GENO: povezuje paciente z (mutiranimi) geni

Povezave

- PHENO: povezuje paciente s fenotipi
- GENO: povezuje paciente z (mutiranimi) geni
- Semantične relacije (30 tipov):
  - predstavljajo (biomedicinsko) znanje
  - izluščene so iz celotnega MEDLINE s pomočjo SemRep
    - `TREATS`
    - `CAUSES`
    - `INHIBITS`
    - `STIMULATES`

Algoritem za podporo kliničnemu odločanju

- Vhod (za enega pacienta):

Algoritem za podporo kliničnemu odločanju

- Vhod (za enega pacienta):
  - množica genotipov (X)

Algoritem za podporo kliničnemu odločanju

- Vhod (za enega pacienta):
  - množica genotipov (X)
  - množica fenotipov (Z)

Algoritem za podporo kliničnemu odločanju

- Vhod (za enega pacienta):
  - množica genotipov (X)
  - množica fenotipov (Z)
- Filtriramo množico genotipov X

Algoritem za podporo kliničnemu odločanju

- Vhod (za enega pacienta):
  - množica genotipov (X)
  - množica fenotipov (Z)
- Filtriramo množico genotipov X
- Uporabimo LBD za:

Algoritem za podporo kliničnemu odločanju

- Vhod (za enega pacienta):
  - množica genotipov (X)
  - množica fenotipov (Z)
- Filtriramo množico genotipov X
- Uporabimo LBD za:
  - napovedovanje novih povezav med genotipom in fenotipom

Algoritem za podporo kliničnemu odločanju

- Vhod (za enega pacienta):
  - množica genotipov (X)
  - množica fenotipov (Z)
- Filtriramo množico genotipov X
- Uporabimo LBD za:
  - napovedovanje novih povezav med genotipom in fenotipom
  - razlago znanih in novih povezav

Algoritem za podporo kliničnemu odločanju

- Vhod (za enega pacienta):
  - množica genotipov (X)
  - množica fenotipov (Z)
- Filtriramo množico genotipov X
- Uporabimo LBD za:
  - napovedovanje novih povezav med genotipom in fenotipom
  - razlago znanih in novih povezav
  - rangiranje (prioritizacija) rezultatov

Algoritem za podporo kliničnemu odločanju

- Vhod (za enega pacienta):
  - množica genotipov (X)
  - množica fenotipov (Z)
- Filtriramo množico genotipov X
- Uporabimo LBD za:
  - napovedovanje novih povezav med genotipom in fenotipom
  - razlago znanih in novih povezav
  - rangiranje (prioritizacija) rezultatov
- Izhod:

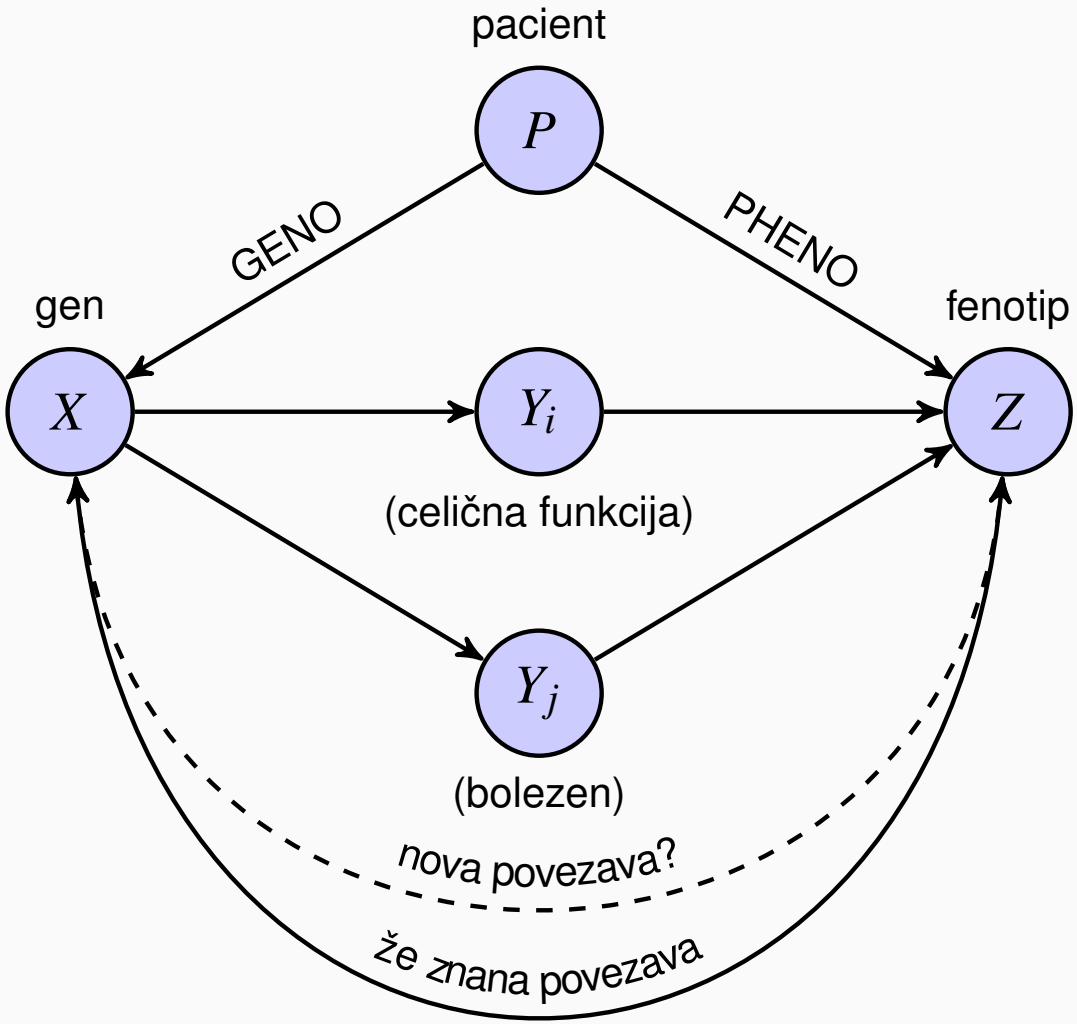
Algoritem za podporo kliničnemu odločanju

- Vhod (za enega pacienta):
  - množica genotipov (X)
  - množica fenotipov (Z)
- Filtriramo množico genotipov X
- Uporabimo LBD za:
  - napovedovanje novih povezav med genotipom in fenotipom
  - razlago znanih in novih povezav
  - rangiranje (prioritizacija) rezultatov
- Izhod:
  - napovedane (nove) povezave med genotipom in fenotipom

Algoritem za podporo kliničnemu odločanju

- Vhod (za enega pacienta):
  - množica genotipov (X)
  - množica fenotipov (Z)
- Filtriramo množico genotipov X
- Uporabimo LBD za:
  - napovedovanje novih povezav med genotipom in fenotipom
  - razlago znanih in novih povezav
  - rangiranje (prioritizacija) rezultatov
- Izhod:
  - napovedane (nove) povezave med genotipom in fenotipom
  - vmesni koncepti, ki povezujejo genotip in fenotip ter pojasnjujejo povezavo

Napovedovanje in pojasnjevanje novih kliničnih povezav



Implementacija

- Neo4j grafovska podatkovna zbirka
- Podpira grafovski podatkovni model
- Cypher kot deklarativni poizvedovalni jezik
- Zakaj smo izbrali Neo4j?
  - Ker se dobro prilega našim podatkom
  - Ker vsebuje algoritme za analizo omrežij (za nadaljnje delo)

Priprava podatkov, agregacija in nalaganje podatkov

- Agregacija z AWK skriptami
- Priprava vhodnih datotek z AWK skriptami in lupinskimi orodji (`join`, `sort`)
- Orodje Neo4j import uporabili za uvoz semantičnih relacij iz literature
- Za ostale podatke: `LOAD CSV ...FROM FILE ...`

Rezultati – konstrukcija podatkovne zbirke

- 1205 pacientov
- 262132 GENO povezav, ki povezujejo paciente z 15 294 vozlišči za gene (možne večkratne povezave)
- 4751 PHENO povezav med pacienti in 1450 vozlišči za fenotipe
- 27 263 265 procesiranih MEDLINE zapisov
- 91 567 597 instanc semantičnih relacij izluščenih s SemRep
- 20 818 782 semantičnih relacij med 277 160 biomedicinskimi koncepti

Vrednotenje

- Za znane povezave genotip–fenotip:
  - s filtriranjem in rangiranjem dobimo:
    - dobro znane relacije na prvem mestu
    - ustrezne razlage (vmestne koncepte) za povezave

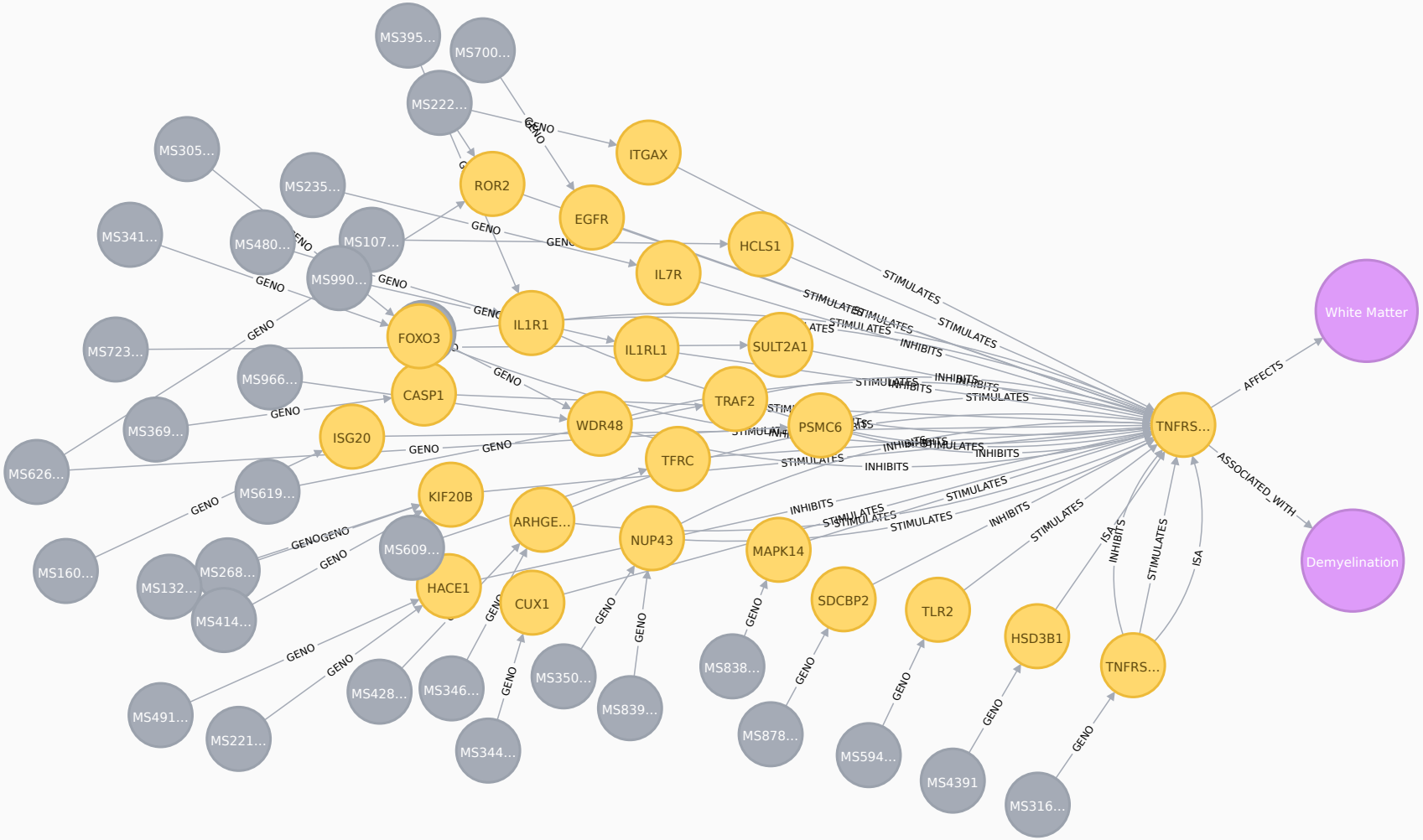
Vrednotenje

- Za znane povezave genotip–fenotip:
  - s filtriranjem in rangiranjem dobimo:
    - dobro znane relacije na prvem mestu
    - ustrezne razlage (vmestne koncepte) za povezave
- Za neznane povezave genotip–fenotip:
  - vrednotenje domenskih ekspertov kot nadaljnje delo
  - preliminarni rezulati zelo obetajoči

Pojasnjevanje znanih povezav genotip - fenotip

```
MATCH
(c1)<-[:PHENO_UMLS]-(p:Patient {id:"###"})-[r:GENO]->(g)
WHERE # filtering
(r.HMZinSLO < 2) AND (r.GnomADHMZ < 10) WITH c1,p,r,g
MATCH (c1)-[r2]-(g) WITH c1,r,p,r2,g
# ranking
ORDER BY r2.freq desc, r.functional_impact desc,
   r.GnomADAlleleCount asc, r.ExACGeneralMAF asc,
   r.HTZinSLO asc, r.UK10KAlleleCount asc,
   r.cadd_score desc, r.SIFT asc LIMIT 20
MATCH (c1)<-[r3]-(c2)<-[r4:ISA|STIMULATES|...]-(g)
WHERE ...exclude too general items... # more filtering
RETURN  distinct c1,r3,c2,r4,g
ORDER BY r3.freq*r4.freq desc;
```


TCEAL1
H3F3AP6
NSG1
BAX
CDKN1A
Cell Prolifer...
Apoptosis
TP53
Colonic Neopla...
Carcino...
Carcino...
Carcino...
Adenoc...
Tumorig...
Malignant
Growth
Polyp of large intesti...
STIMULATES
DISRUPTS
ASSOCIATED_WITH
AFFECTS
COEXISTS_WITH
CAUSES
OCCURS_IN


MS395...
MS700...
MS222...
ITGAX
ROR2
EGFR
MS305...
MS235...
MS341...
MS480...
MS107...
HCLS1
IL7R
MS990...
White Matter
FOXO3
IL1R1
IL1RL1
SULT2A1
MS723...
MS966...
CASP1
TRAF2
PSMC6
MS369...
ISG20
WDR48
TNFRS...
AFFECTS
MS626...
TFRC
MS619...
KIF20B
ARHGE...
NUP43
MAPK14
ASSOCIATED_WITH
Demyelination
MS160...
MS609...
MS268...
MS132...
HACE1
CUX1
SDCBP2
TLR2
HSD3B1
MS414...
MS838...
MS878...
MS594...
MS4391
MS316...
MS491...
MS221...
MS428...
MS346...
MS350...
MS839...
MS344...
GENO
STIMULATES
INHIBITS
ISA

Nadaljnje delo

- Evalvacija z biomedicinskega zornega kota

Nadaljnje delo

- Evalvacija z biomedicinskega zornega kota
- Razvoj spletne aplikacije z ustreznim
  - iskalnim modulom
  - vizualizacijskim modulom

Nadaljnje delo

- Evalvacija z biomedicinskega zornega kota
- Razvoj spletne aplikacije z ustreznim
  - iskalnim modulom
  - vizualizacijskim modulom
- Detekcija in filtriranje napačno pozitivnih rezultatov ter "preveč" splošnih konceptov in relacij

Nadaljnje delo

- Evalvacija z biomedicinskega zornega kota
- Razvoj spletne aplikacije z ustreznim
  - iskalnim modulom
  - vizualizacijskim modulom
- Detekcija in filtriranje napačno pozitivnih rezultatov ter "preveč" splošnih konceptov in relacij
- Vključitev v klinično-genetski diagnostični delotok

Zaključki

- Grafovska podatkovna zbirka Neo4j je ustrezna za hranjenje heterogenih genomskih podatkov, ki jih potrebujemo za diagnostično podporo v klinični genetiki

Zaključki

- Grafovska podatkovna zbirka Neo4j je ustrezna za hranjenje heterogenih genomskih podatkov, ki jih potrebujemo za diagnostično podporo v klinični genetiki
- LBD lahko uporabimo kot komplementarno metodo v klinični diagnostiki s poudarkom na novih povezavah gen–fenotip