Multimorbid betegségek közös mechanizmusainak összgenomi felderítése hálózati-alapú meta-elemzéssel

Készítette

Manninger Miklós

Konzulens

Dr. Antal Péter

2018

TARTALOMJEGYZÉK

Áttekintés

Több betegség együttes jelenléte, a multimorbiditás és az ezzel gyakran együtt járó több gyógyszer egyidejű szedése, a polifarmácia egy egyre gyakoribb és nem csupán időskorban fellépő jelenség, amely mind személyes, mind társadalmi vonatkozásában egyre nagyobb terhet jelent. Kiemelkedő jelentőségű multimorbiditási betegség klasztert alkot a depresszió és komorbid betegségei, mint például az elhízás. Multimorbiditási betegségek megértésében ígéretes kutatási irány a közös hátterük felderítése, például közös molekuláris útvonalak, közös genetikai rizikófaktorok beazonosítása. A dolgozatban egy olyan módszertant és megvalósított rendszert mutatok be, amely betegségek közös genetikai hátterének vizsgálatát segíti automatizálva a genetikai asszociációs vizsgálatok lefuttatását, az eredmények génre aggregálását, a gén szintű eredmények hálózati terjesztését és a hálózati feldúsulási eredmények intelligens elemzését is.

Bevezetés

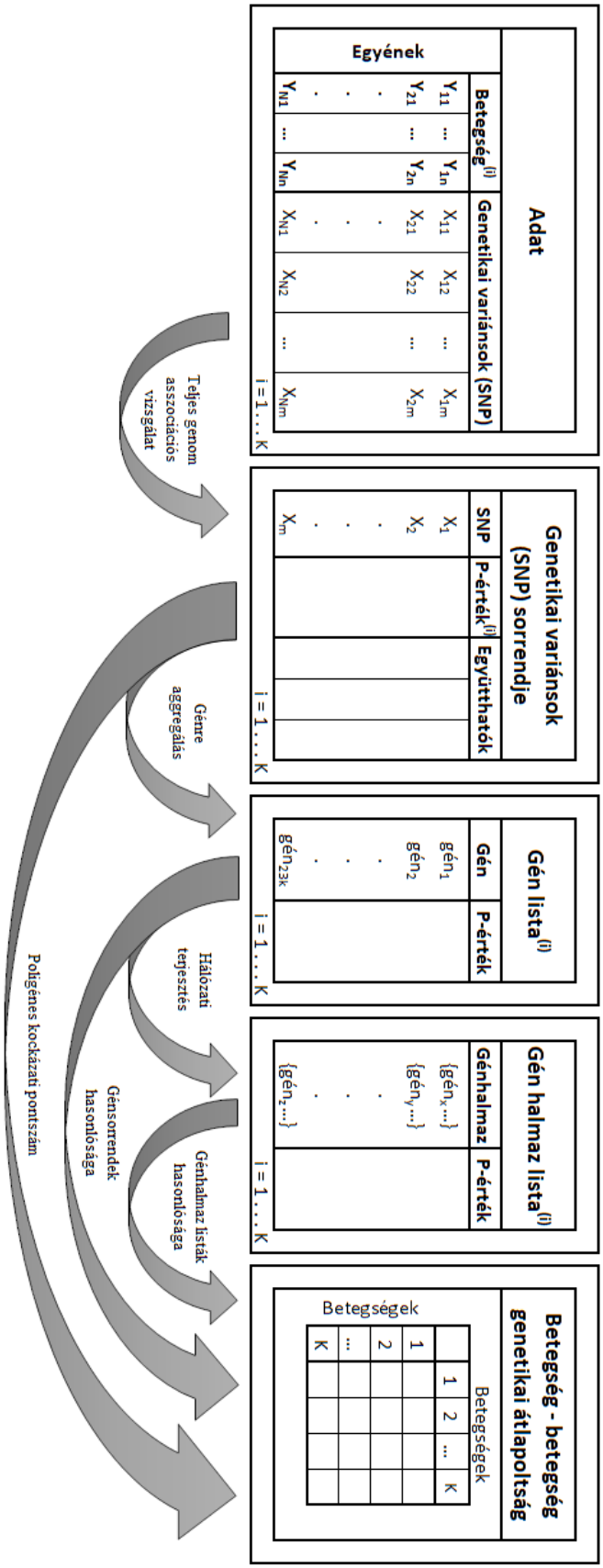
A betegségek molekuláris biológiai hátterének megismerése a megelőzés, diagnosztika, terápiaválasztás, gyógyszerek kifejlesztése szempontjából is meghatározó. Az orvosbiológiai mérési technikáknak és az egészségügyi adatok gyűjtésének az utóbbi évtizedekben bekövetkezett fejlődése forradalmi változást jelentett több betegség hátterének a megismerésében [3]. Azonban több betegség esetében a biológiai háttér felderítése elmaradt a remélt ütemtől, viszont az eddigi vizsgálatok negatív eredménye arra utal, hogy a biológiai hátteret nagyszámú genetikai és környezeti tényező komplex együttese alkotja. A molekuláris biológiai háttér átfogó vizsgálatai azonban azt is jelezték, hogy több betegség esetében a genetikai háttérük egy jelentős része közös [12,13,14]. A közös háttér felderítése viszont új módszereket igényel, amelyeket a kutatóorvosok hatékonyan tudnak felhasználni.

A molekuláris biológiai háttér megismerésében kitüntetett szerepet játszik a betegségek genetikai háttérének a megismerése, amely mind az érintett géneknek a megismerését jelenti, mind a géneket kódoló genetikai állománynak a releváns elváltozásainak a megismerését is [1]. A teljes genomi szélességű genetikai asszociációs vizsgálatok egy orvosi felhasználása a gyakori genetikai variánsokhoz származtat jellemzőket, amely egy adott betegséggel való statisztikai asszociáltságot írják le [2]. Az eredmények értelmezését nehezíti a genetikai variánsok génekhez rendelése, különösen a génszabályozási funkciókat ellátó variánsok esetében, amely a betegség szervi, szöveti vonatkozásának megfelelően változhat. További kihívás, a gének egymásra hatásának figyelembe vétele, a gének szabályozási hálózatának figyelembe vétele, amely szintén a betegség szervi, szöveti vonatkozásában más és más lehet.

A dolgozatban egy olyan általam megvalósított rendszert mutatok be, amely a következőket támogatja rendre több paraméterezés mellett [4,5]:

- genetikai asszociációs vizsgálatok lefuttatását [2],
- az eredmények génre aggregálását [7],
- a gén szintű eredmények hálózati terjesztését [5],
- és a több betegségre vonatkozó hálózati feldúsulási eredmények elemzését.

Dolgozatomban a hálózati feldúsulási eredmények szakértői elemzésének támogatására egy új elemzési szint felhasználását vizsgáltam meg, a génhalmazok sorrendjeinek a szintjét, amelyhez új, hasonlósági metrikák használatát dolgoztam ki.

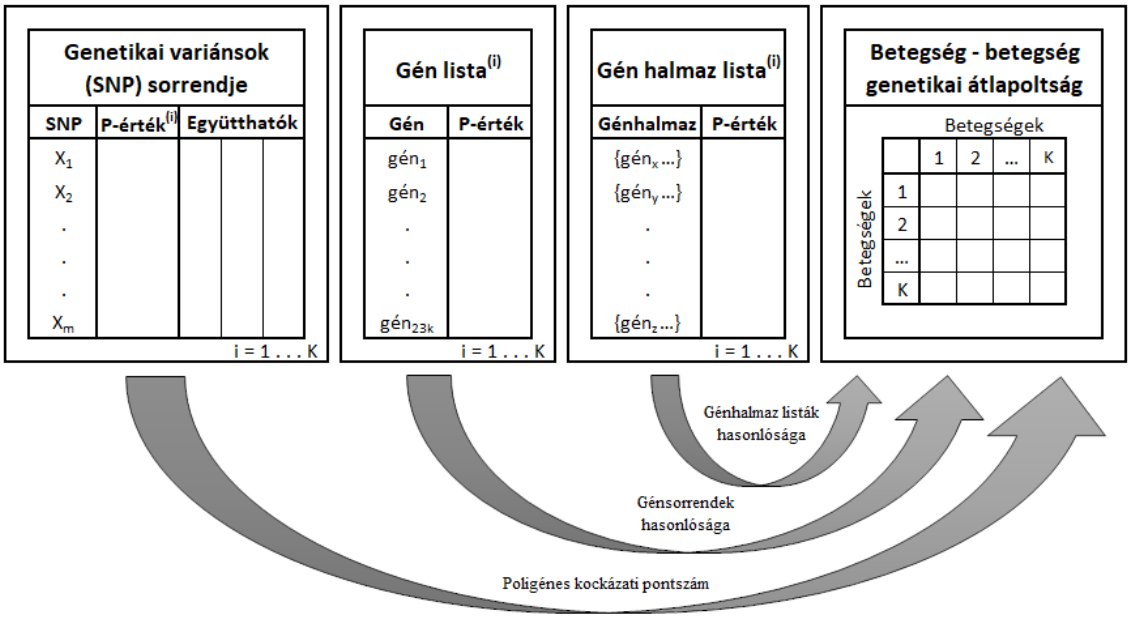


1. ábra: Betegségek közös genetikai hátterének vizsgálati módszerei és teljes munkafolyamata

Az *1. ábra* mutatja be ezt a rendszert, mely lépéseinek részletezését a dolgozatban fejezetekre bontottam.

A hálózati terjesztések eredménye egy gén halmazokból álló lista. Ha két betegséget szeretnénk megvizsgálni multimorbiditás szempontjából, akkor a hozzájuk tartozó elemzési eredmények hasonlóságát számítva, következtetni tudunk a két betegség kapcsolati viszonyára. A dolgozatban statisztikai genetikai elemzések eredményeinek három szinten történő összehasonlítását mutatom be. Ezek a szintek rendre a (1) genetikai variánsok sorrendje, melyben genetikai variánshoz rendelt egy pontszám a betegséggel való kapcsolatuk mértéke alapján, (2) gének sorrendje, melyben hasonló módon a génekhez rendelt egy értéket és (3) génhalmazok sorrendje, amely már gén csoportosulások együttes hatását jelenti az adott betegségre nézve. Két betegséghez ezeket az eredményeket összehasonlítva olyan géneket, génhalmazokat kereshetünk, melyek mindkét betegség vizsgálati eredményben befolyásoló tényezőknek számítottak, így a két vizsgált betegség között gén szintű kapcsolatot találhatunk.

Jelen dokumentumban ezt az eredményt egy táblázatban ábrázolom, mely minden betegség páros kapcsolatát hivatott szemléltetni. A *2. ábrán* látható, hogy mely vizsgálati eredmények mely adattáblákból származnak.

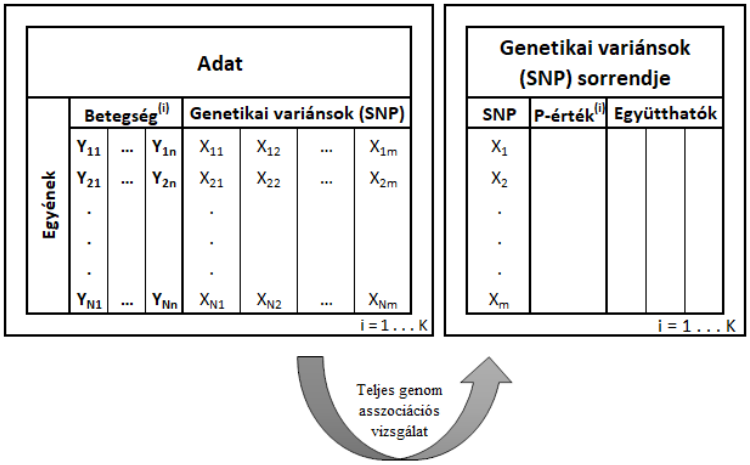


2. ábra: Multimorbiditás közös genetikai hátterének a vizsgálati folyamata

Dolgozatomban elsőként bemutatom a „genetikai variánsok sorrendje”, a „gén lista” és a „génhalmaz lista” eredményeket előállító munkafolyamatot (1-3. fejezet), majd ezt követően a háromféle eredmény összehasonlítását és kiértékelését. (4. fejezet).

1. Teljes genom asszociációs vizsgálatok

A teljes genom asszociációs vizsgálat (GWAS, Genome-wide association) egy genomi szintű statisztikai vizsgálat, mely a genetikai variánsok együttesét figyeli meg különböző egyénekben. Az egyes nukleotid polimorfizmusok (SNP, Single Nucleotide Polymorphism) és a főbb emberi betegségek közötti statisztikai asszociációt statisztikai tesztek sokaságával állapítja meg [1,2,3,4].



3. ábra
Első lépés: Teljes genom asszociációs vizsgálat

Eredményként minden SNP-hez hozzárendel egy úgynevezett p-értéket, mely meghatározza azt, hogy mely SNP-k kapcsolódása statisztikailag a leginkább plauzibilis és nem a véletlen következménye.

1.1. Bemenet

A kiindulási adattábla minden sora egy-egy emberhez tartozó genetikai variánsok szintjén lévő vizsgálatok eredményeit tárolja, valamint minden betegséghez tartalmaz külön-külön egy vagy több leíró, különböző célváltozót (Y), illetve az egyénhez tartozó genetikai variánsokat, esetünkben nevezetesen egynukleotidos polimorfimizmusokat (single nucleotide polymorphism, SNP-ket). Az asszociációs vizsgálat alapján azt szeretnénk megvizsgálni, hogy egy betegséghez mely SNP-k relevánsak, és melyek azok, amelyektől független a betegség.

1.2. Asszociációs elemzési eszközök

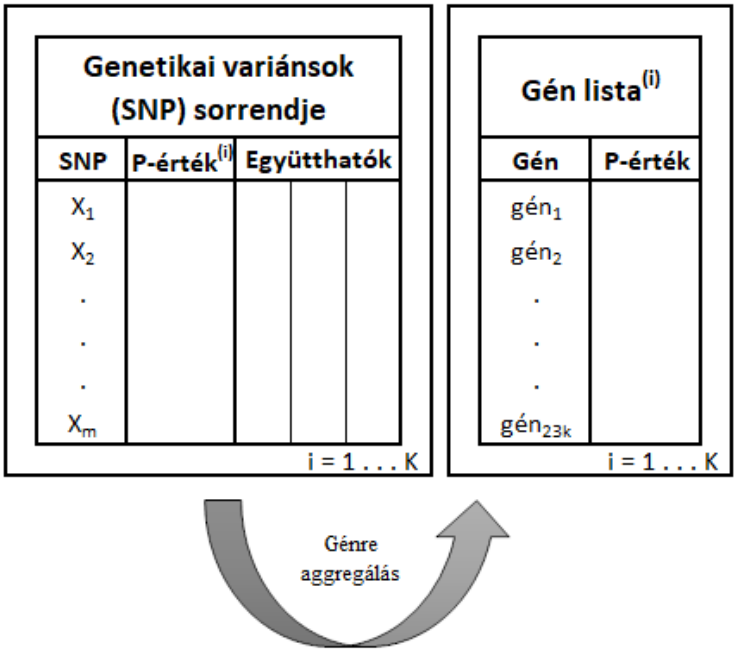
A leggyakrabban használt elemzési eszköz, amely statisztikai tesztek sokaságát kínálja a PLINK [9], illetve az általam implementált keretrendszer lehetővé teszi különböző lineáris kevert modellek futtatását is [11].

1.3. Kimenet

Az asszociációs statisztikai vizsgálatok eredménye egy olyan tábla, mely minden SNP-hez egy értéket rendel hozzá, mely azt mutatja, hogy milyen mértékben áll fenn összefüggés az SNP és az adott betegség között. Ilyen érték a p-érték is, amely egy 0 és 1 közötti szám, mely minél közelebb van a 0-hoz, annál biztosabbnak jelzi az SNP és a betegség közötti összefüggés fennállását. Minden betegségre lefuttatva a genom asszociációs vizsgálatot, felállítható egy-egy SNP sorrend a p-értékek szerint növekvő módon.

2. Eredmények génre aggregálása

A GWAS számítás eredménye SNP-ek listáját tartalmazza. Ezzel a lépéssel szeretnénk egy olyan gén szintű jellemzőt származtatni, hogy egy gén összességében milyen statisztikai asszociáltságot mutat a betegséggel. A SNP-ekhez tartozó adatokból úgynevezett génre aggregáló függvényekkel számíthatjuk ki a génekhez tartozó értékeket. A statisztikai génre aggregáló függvények alapvető bemenete az SNP-k génekhez rendelése, amelyhez egy online platform, a FUMA segítségét használtam fel [7]. Ebből származtattam az általam használt hozzárendelő függvényt (a FUMA SNP2GENE függvényét). A génre aggregáláshoz a PEGASUS nevű módszert használta, fel a munkafolyamatomban [6].



4. ábra
Második lépés: Génre aggregálás

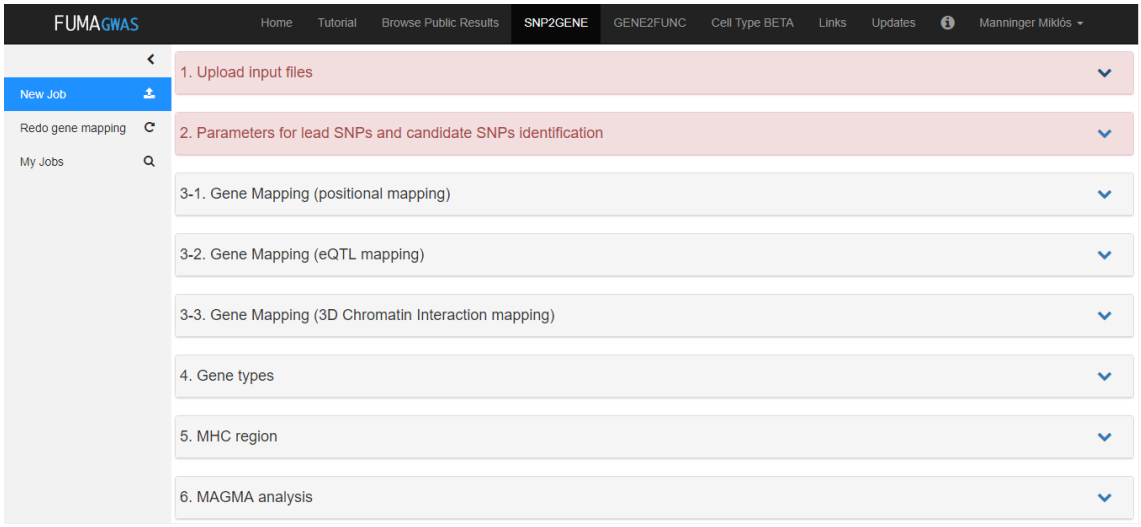
2.1. Bemenet

A génre aggregáláshoz szükség van a genetikai variánsok sorendjére. Minden betegséghez tartozik egy adattábla, mely az SNP-ekhez tartozó P-értékeket és opcionálisan egyéb együtthatókat is tartalmaz (például a genetikai variánsok betegségre vonatkozó rizikópontszámait). A táblázat első sorában az a gén variáns áll, mely a betegséget leginkább befolyásolja, azaz a legkisebb a p-értéke. A génre aggregálás folyamat segítségével a SNP-k sorrendjéből előállítjuk a gének sorrendjét, mely megmutatja, hogy az egyes gének milyen mértékben kapcsolódnak statisztikailag az egyes betegségekhez.

2.2. Gének definiálása a FUMA rendszerben

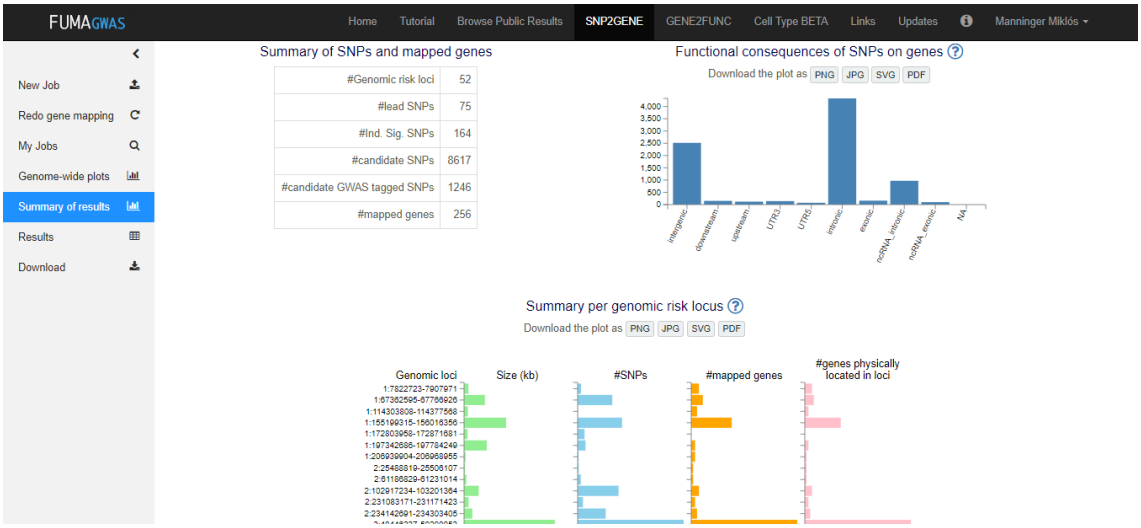
A FUMA GWAS (Functional Mapping and Annotation of Genome-Wide Association Studies) egy olyan platform, mely segít a GWAS eredményeket értelmezni, vizualizálni, vagy egy SNP2GENE funkció használatával elvégezni a génre aggregálást, így a SNP2GENE funkció segítségével az SNP-ekhez tartozó adatokat a génekhez köthetjük, minden génhez megkapva egy P-értéket, melyet a későbbi számításainkhoz használunk fel.

Az *5. ábra* mutatja, hogy milyen beállításokat lehet megadni egy FUMA futáshoz. A megfelelő paraméterezéssel olyan SNP2GENE függvény készíthető, mely a számunkra fontos követelményeknek tesz eleget, így minden futás személyre szabható, ami azért fontos, mert egyes betegségekhez más és más gének relevánsak, más és más szempontból kell megvizsgálni őket.



5. ábra: A FUMA online platform használata

A *6. ábra* egy futás eredményének egy részét mutatja be. A FUMA platform elterjedten használt teljes genomi asszociációs vizsgálatok elvégzésére, génre aggregálás számítására és az eredmények vizualizálására.



6. ábra: FUMA GWAS eredmények vizualizálása

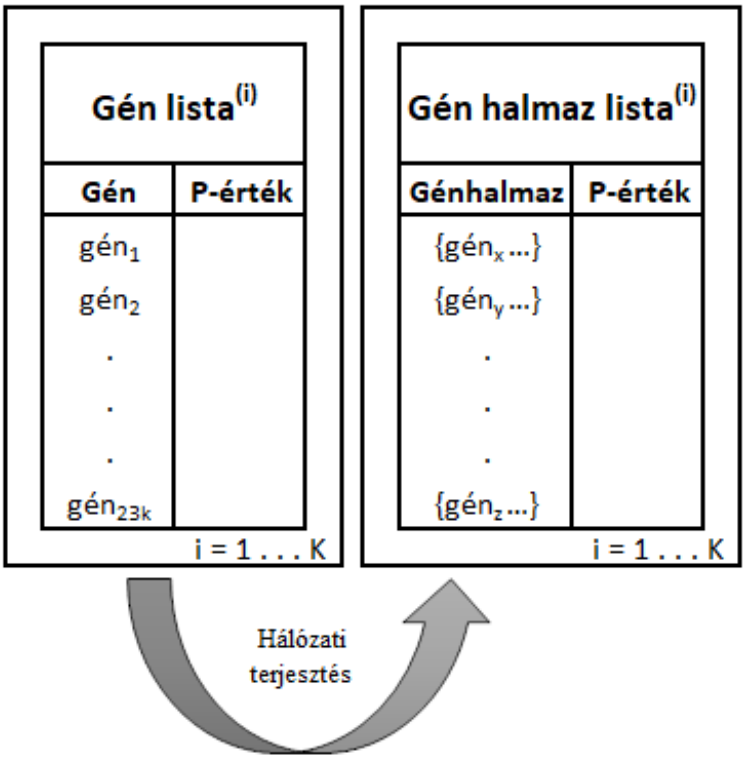
Ebben a dokumentumban nem foglalkozom az eredményül kapott diagrammokkal, mivel a fontos adatot számunkra a letölthető szövegfájlok tartalmazzák, melyekből előállíthatjuk a gének erősség szempontjából sorrendezett listáját.

2.3. Kimenet

A génre aggregálás után egy két oszlopos táblát kapunk minden betegséghez. A táblázat megmutatja azt, hogy az adott betegséghez az elemzésben felhasznált adatok alapján az adott gén kapcsolata statisztikailag mennyire alátámasztott. Azonban annak meghatározása, hogy egy betegség kialakulásában, megjelenésében milyen szerepet is játszik egy adott gén, további aggregálási módszerek és rendkívül nagy mennyiségű háttérismeret szükséges. Az utóbbi évtizedek forradalmi módszere a feldúsulás vizsgálat, amely génhalmazok elemeinek egyenletestől eltérő eloszlását vizsgálja meg. Egy további lehetőség a genetikai útvonalak és génszabályozási hálózatok felhasználása további aggregációra, amelyet a továbbiakban én is vizsgáltam.

3. Gén szintű eredmények hálózati terjesztése

Egy évtizede még a genomi vizsgálatok utolsó lépése a génlisták elemzése volt, feldúsulási vizsgálatokkal és úgynevezett génprioritizálási módszerekkel [3]. Még napjainkban is sokszor az elemzések csupán a gének sorrendjeinek összehasonlításáig terjednek. Kutatásomban ennek többváltozós kiterjesztését vizsgáltam, amikor nem csak a gének sorrendjeit értékeljük ki, hanem azt is, hogy léteznek-e olyan géncsoportosulások, melyek statisztikai kapcsolatban állnak akár több betegséggel is.



7. ábra
Harmadik lépés: Hálózati terjesztés

3.1. Bemenet

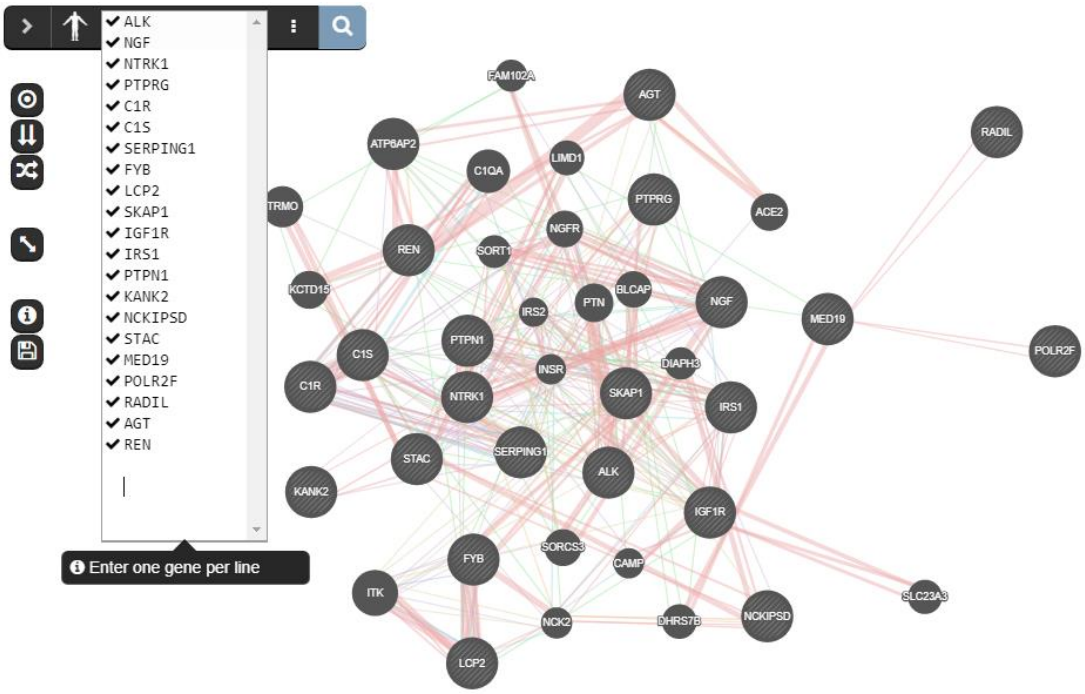
A munkafolyamat szempontjából ebben a lépésben egy génlistából egy génhalmaz lista készül. A hálózati terjesztés nem az egyes géneket vizsgálja, hanem egy megadott génhálózat segítségével génekből álló halmazokat, géncsoportokat. A hálózaton szétterjesztve a gének erősségét megvizsgálhatjuk, hogy az egymással szoros kapcsolatban álló gének együttesen is hatással vannak-e a betegségre. Az így kialakuló csomópontokon definiált aktivitás mintában rész hálózatokat (modulokat, algráfokat) keressünk, amelyek összpontszáma nagy, ezek definiálják majd a kimeneti génhalmazokat. Ezen halmazok hatását vizsgáljuk a betegség szempontjából, immáron nem a génekre külön-külön, hanem egy együttes értéket számítva.

3.2. Hotnet2: Hálózat elemző algoritmus

A hálózati terjesztési módszerek egyik népszerű eszköze a hotnet2 módszer [6], amely bemenetként egy hálózatot és egy csomóponti aktivációs értéket vár. A genetikai adatelemzési munkafolyamatban a hálózat egy génszabályozási hálózat vagy fehérje-fehérje interakciós hálózat, amelyben a csomópontok gének. Ennek megfelelően a csomóponti aktivációs értékek a gének statisztikai asszociáltsága az egyes betegségekkel, amely az előzőekben ismertetett génre aggregálásból származik. A hálózat származtatására egy népszerű külső eszközt használtam a munkafolyamatomban, a GeneMania-t [8].

3.3. GeneMania

A GeneMania egy online platform [8], mely egy génhálózatok összeállítására, megjelenítésére és lementésére alkalmas szolgáltatást biztosít. Használata rendkívül egyszerű, csak a hálózatban szereplő géneket kell megadni. Természetesen ez nem azt jelenti, hogy csupán a megadott gének közötti kapcsolatokat jeleníti meg, hanem olyan géneket is, melyek szorosan összefüggésben állnak a megadott gének egyikével. Így egy olyan hálózatot kapunk, mely jó reprezentálja a gének közötti kapcsolatokat. A *8. ábra* megmutatja, hogy a felsorolt génekhez milyen hálózat kapcsolódik.



8. ábra: A GeneMania platform

A génre aggregáláshoz hasonló módon, ebben a dolgozatban nem foglalkozunk a hálózat megjelenítésével, mivel a számunkra fontos hálózat letölthető az oldalról. A letöltött hálózatok munkafolyamatomban való használatának lehetővé tételéhez konvertert készítettem.

3.4. Kimenet

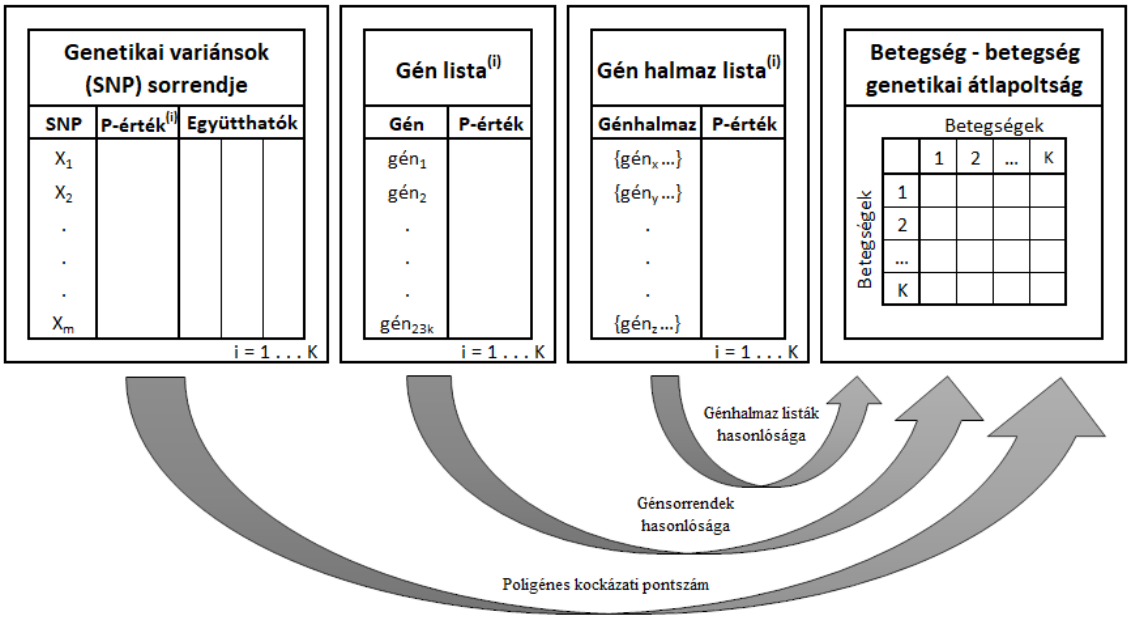
A hálózaton való terjesztéssel kialakulnak génhalmazok, alhálózatok, melyek közösen vannak hatással a betegségre. Ezen halmazok megtalálása jóval nehezebb, mint a génre aggregálás, de a génhalmazok megfelelő elemzésével még részletesebb eredményhez juthatunk. A betegség és egy génhalmaz kapcsolatának erősségét itt is egy p-érték jelzi.

A munkafolyamat szempontjából ennek a lépésnek az eredménye minden betegséghez egy olyan génhalmazokból álló lista, melyekhez egy p-érték tartozik.

4. Közös genetikai háttér elemzése

A kutatásomban összeállított munkafolyamat három szintet is kínál a betegségek közös genetikai hátterének vizsgálatára:

- Genetikai variánsok (SNP-ek) sorrendjei, együtthatói és p-értékei,
- gének sorrendjei,
- génhalmazok sorrendjei.



9. ábra

Utolsó lépés: Eredmények elemzése, összehasonlítása

4.1. SNP sorrend elemzés

A genetikai variánsok szintjén végzett (az úgynevezett poligénes kockázati pontszám alapú) vizsgálat a betegségek közös genetikai hátterének a vizsgálatában egy páronkénti elemzést tesz lehetővé. Alapja az egyes betegségekhez tartozó genetikai variánsokon definiált rizikó modellek hasonlósága. Alapesetben például a betegségekhez tartozó lineáris vagy logisztikus regressziós modelljeinek predikciós ereje más betegségekben. A predikciós erő mérésére bevett gyakorlat a megmagyarázott variancia, azaz a betegség egy folytonos leírójának a varianciája hányad részére csökken a modell felhasználásával. Gyakori

felhasználása, hogy egy B1 betegség alapvető leírójára illesztett modellnek a teljesítményét egy másik B2 betegség fontos leíróján értékeljük ki.

4.2. Génsorrend elemzés

A gének sorrendjének összevetésére megszokott a Spearman rang korreláció felhasználása, amely egy normált négyzetes eltérést mér a gének sorrendbeli eltérésére. A számításhoz megadható egy k paraméter, melyet a listák vágására használunk.

Vegyük az x listát és egy k természetes számot. A gén lista génneveket tartalmaz, és a hozzá tartalmazó p-értéket. A lista ez az érték alapján növekvő sorrendbe van rendezve. Jelölje $rg(x_i)$ az x lista i. elemének a rangját mely megegyezik az elem sorszámával, azaz i-vel.

Két betegség összehasonlításakor (A és B betegségek) minden g génhez két rang tartozik, $rg(g, A)$ és $rg(g, B)$.

Legyen az A betegséghez tartozó lista x és a B betegséghez tartozó lista y. Ekkor $d_i = rg(x_i) - rg(y_i)$ az i. párra kiszámolt különbség. Ekkor kiszámítható a két lista hasonlósága (r) az alábbi képletet alkalmazva [15]:

$$r_s = 1 - \frac{6 \sum_i {d_i}^2}{n * (n^2 - 1)}$$

ahol n az értékpárok száma.

Két betegség gén listájának a hasonlóságát ez a szám jelzi, mely egy 0 és 1 közötti érték. Minél közelebb van az 1-hez annál jobban hasonlít a két lista.

4.3. Gén halmaz sorrend elemzés

A hálózati analízis eredménye, hogy minden lefuttatott betegséghez eredményül kapunk egy sorrendezett génhalmaz listát. A gén lista összehasonlításhoz hasonló módon itt is célunk, hogy két vagy több betegség kapcsolatára következtethessünk a hozzájuk tartozó listák hasonlóságából. A gén listánál csak az elemek sorrendjét kellett vizsgálni, viszont a halmaz listáknál ez a módszer nem alkalmazható feltétlenül, hiszen a két vagy több lista elemei eltérhetnek egymástól. Éppen ezért először hasonló génhalmazokat kell keresni a két listában, majd ezekre megvizsgálni a sorrendi egyezést.

Egy közös genetikai háttérelem, esetünkben egy génhalmaz fontosságát a következők befolyásolják:

- hány darab betegséget érint, azaz hány betegség sorrendjében fordul elő a génhalmaz vagy hasonló génhalmaz,
- az adott sorrendekben milyen sorrenddel (hányadikként) fordul elő a génhalmaz vagy hasonló génhalmaz,
- ha csak hasonló génhalmaz fordul elő, akkor mekkora a hasonlóság.

Ezen komplex preferencia rendszer leírására egy kvantitatív pontszám megadásának lehetőségét dolgoztam ki, mely három lépésből áll: (1) motívumhalmaz számítása, (2) a génhalmaz elemeinek kicserélése és (3) a hasonlóság kiszámítása

4.3.1. Motívum halmaz számítása

Adott hasonlóságok és *B, H, E* számok mellett egy génhalmazt *m* motívumnak hívunk, ha *B* darab betegség génhalmaz listájának első *H* eleme közül van olyan génhalmaz, mely a motívumtól legfeljebb *E* eltérésre van. Számítási feladatunk, hogy keressük meg az összes lehetséges motívumot, azaz keressük meg az összes génhalmazt, mely megfelelően kicsi eltéréssel, megfelelő számú betegségben, megfelelő helyezést ért el.

Legyen *E(m)* az *m* motívum átlagos eltérése a megfelelő génhalmazoktól. Ekkor

$$E(m) = \delta * u(m) + \varepsilon * s(m), \text{ ahol}$$

- *u(m)* azon gének átlagos száma, melyek *m*-nek elemei, de a vizsgált génhalmaznak nem elemei.
- *s(m)* azon gének átlagos száma, melyek *m*-nek nem elemei, de a vizsgált génhalmaznak elemei.
- δ és ε állandók.

Az állandók megfelelő paraméterezésével beállítható, hogy az átlagos eltérés számítása milyen szempontok alapján történik. Minél nagyobb a δ, annál nagyobb mértékben büntetjük azt, ha a motívum felesleges elemeket tartalmaz. Hasonlóképpen minél nagyobb az ε, annál jobban büntetjük, ha a motívum nem tartalmaz szükséges elemeket. Tehát ezzel a két paraméterrel megszabható, hogy épp azt büntetjük, ha túl sok felesleges elem van a motívum halmazában, vagy éppen azt, ha hiányzik elem belőle.

Legyen *W(m)* az *m* motívum súlya, melyet a következő képlettel számíthatunk ki:

$$W(m) = \alpha * B(m) - \beta * H(m) - \gamma * E(m), \text{ ahol}$$

- *B(m)* azon betegségek száma, mely génhalmaz listájában szerepel *m* (a megfelelő helyezéssel és a megfelelően kis eltéréssel).
- *H(m)* az *m*-hez hasonlító génhalmazok átlagos helyezése.
- *E(m)* az *m* motívum átlagos eltérése a megfelelő génhalmazoktól.
- α, β és γ állandók.

Az állandók megfelelő paraméterezésével beállítható, hogy a motívum súlyának számítása milyen szempontok alapján történik, azaz mit jutalmazunk, illetve büntetünk.

Minél nagyobb az α, annál inkább jutalmazzuk azt, hogy milyen sok betegségben szerepel az adott motívum.

Minél nagyobb a β, annál nagyobb mértékben büntetjük az első helytől való távolságot.

Minél nagyobb a γ, annál jobban büntetjük a vizsgált halmazoktól való átlagos eltérést.

4.3.2. Génhalmaz lista elemeinek kicserélése

Két betegség génhalmaz lista elemeinek összehasonlításához szükséges, hogy a listák elemei közös értékkészlettel rendelkezzenek. A két lista összehasonlításakor egyszerű volt ezt elérni, hiszen egy lista elem egyetlen génből állt, mely vagy ugyan az volt, mint a másik listában lévő elem, vagy nem.

Ahhoz, hogy a két génhalmaz lista közös elemekkel rendelkezzen, ki kell cserélni az halmazokat. Vegyük az előző lépésben kiszámított motívum halmazt. A két hasonlítandó lista minden halmazát cseréljük le olyan motívumra, mely a motívum súlyát jutalmazva, az eltérésüket pedig büntetve a legkedvezőbbnek bizonyul. A jutalmazást és büntetést az eddigiekhez hasonló módon tehetjük meg.

4.3.3. kombinált hasonlóság kiszámítása

A két kapott génhalmaz lista már a génsorrend elemzés alfejezetben megismert módon összehasonlítható, hiszen az átalakítások után két génhalmaz lista már két listának tekinthető.

Legyen a két génhalmaz lista az elemek kicserélése után *x* és *y*.

Ekkor a két génhalmaz sorrend hasonlósága kiszámítható az alábbi képlettel:

$$S(x,y) = -\lambda * r(x,y) + \mu * \left(\sum_i w(x_i) + \sum_j w(y_j)\right), \text{ ahol}$$

- *r(x, y)* a két génhalmaz hasonlósága a génsorrend elemzés alfejezetben megismert módon kiszámítva.
- $w(x_i)$ az *x* lista *i.* eleméhez rendelt motívum súlya.
- λ és μ pedig állandók.

Minél nagyobb a kapott *S* szám, annál nagyobb a hasonlóság a két génhalmaz listának.

Az állandók megfelelő paraméterezésével beállítható, hogy a hasonlóság számításakor milyen szempontok szerint számolunk. Minél nagyobb a λ, annál inkább büntetjük a sorrendbeli eltéréseket, és minél nagyobb a μ, annál jobban büntetjük a halmazok közötti különbséget.

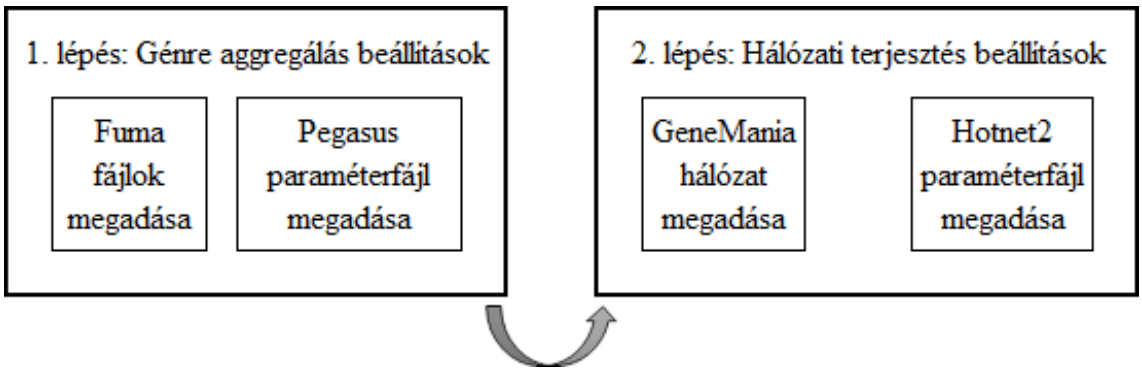
5. A közös genetikai háttér kutatását támogató munkafolyamat és rendszer

Ebben a fejezetben egy olyan munkafolyamat tervezetet mutatok be, mely segíti a betegségek vizsgálati eredményeinek hasonlítását.

5.1. A génre aggregálás és a hálózati terjesztés algoritmusainak paraméterezése

Ebben az ablakban megadható a génre aggregálás futásához szükséges paramétereket tartalmazó fájl, valamint a FUMA által generált SNP2GENE függvény, melyek segítségével a Pegasus elvégzi a génre aggregálást.

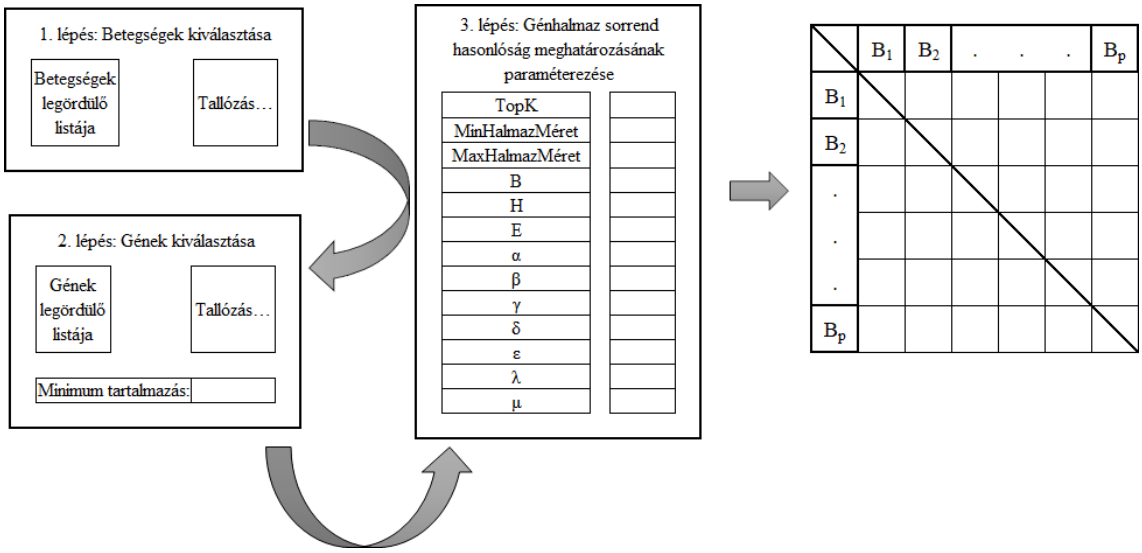
Ezt követően a hálózati terjesztés algoritmusának paraméterei adhatók meg, azaz a GeneMania-ból letöltött hálózati fájl, illetve az algoritmus saját paraméterezéséhez szükséges fájl.



10. ábra: Task paraméterezése

A megadott paraméterek által lefuttatott algoritmusok eredményét egy task-nak, azaz feladatnak nevezem.

5.2. Az összehasonlító algoritmus paramétereinek megadása



11. ábra: összehasonlításhoz felhasznált paraméterek megadása

5.2.1. Első lépés: Betegségek kiválasztása

Ebben a lépésben kiválasztható, hogy a generált task összes betegsége közül melyek eredményeit szeretnénk összehasonlítani. Lehetőség van egyesével kiválasztani a betegségeket egy listából, de pár tucat betegség vizsgálata esetén ez a megoldás már nem kényelmes. A másik lehetőség, hogy egy szöveges fájlt töltünk fel, melyben soronként egy betegség nevet tartalmaz. Abban az esetben, ha egy betegség nem szerepel a task által generált eredményhalmazban, a felhasználó értesítést kap erről, és a kiértékelés a többi betegséggel folytatódik. Ha egyik lehetőséget sem választva tovább lépünk, akkor a taskban szereplő összes betegség kiértékelésre kerül. Minden esetben a továbblépés előtt a rendszer jelzi, hogy hány darab betegség lett kiválasztva.

5.2.2. Második lépés: Gének kiválasztása

A harmadik lépés, az összehasonlításhoz kívánatos gének kiválasztása. Hasonlóan az előző lépéshez, itt is lehetőség van egy listából kiválasztani a géneket, vagy egy génneveket tartalmazó fájl feltöltésével megadni az összes vizsgálandó gént. Egy minimum számot is meg kell adni, mely meghatározza, hogy a génhalmazok esetében legalább hány génnek kell a halmazban szerepelnie a felsoroltak közül. Azokat a génhalmazokat, melyek nem tartalmazzák a minimum számnak megfelelő mennyiségű gént a felsoroltak közül, kivesszük a vizsgálandó génhalmaz listából. Ha egyik lehetőséget sem választjuk,

akkor az összes gén és génhalmaz kiértékelésre kerül. Minden esetben a továbblépés előtt a rendszer jelzi, hogy mely betegséghez hány darab génhalmaz maradt.

5.2.3. Harmadik lépés: Hasonlóság meghatározásának paraméterezése

Utolsó lépésként a megmaradt génlisták és génhalmaz listák hasonlóságának kiszámításának paraméterezése maradt. A felhasználó megadhatja a génhalmaz listák összehasonlítási algoritmusának a paramétereit:

Paraméterek	
MinHalmazMéret	Ennél kisebb elemszámú halmazok eldobása
MaxHalmazMéret	Ennél nagyobb elemszámú halmazok eldobása
TopK	A legjobb k darab vágása
B	Egy motívum betegséghez tartozásainak minimuma
H	Egy motívum átlagos helyezésének maximuma
E	Egy motívum átlagos eltérésének maximuma
α	Egy motívum betegség listához tartozásának jutalmazása
β	Egy motívum átlagos helyezésének büntetése
γ	Egy motívum átlagos eltérésének büntetése
δ	Egy motívum felesleges gének tartalmazásának büntetése
ε	Egy motívum szükséges gének hiányának büntetése
λ	Halmazok sorrendbeli eltéréseinek büntetése
μ	Halmazok közötti különbség büntetése

Ezeket a paramétereket a 4.3 Génhalmaz sorrend elemzés fejezetben már ismertetettem. Mindegyik értéknek van alapértelmezett értéke. Ha nem módosítjuk az értékeket, akkor minden adatot kiértékel és összehasonlít a rendszer.

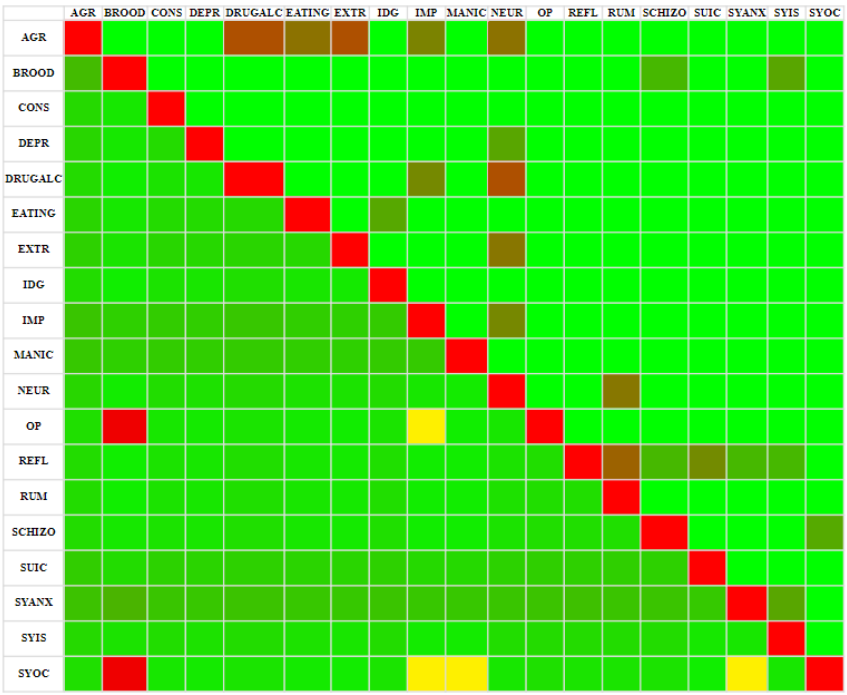
6. Alkalmazás

A létrehozott munkafolyamatot a Semmelweis Egyetem Gyógyszerhatástani Intézetének NewMood kutatásához kapcsolódva alkalmaztam az ottani adatokra Dr. Juhász Gabriella segítségével. Az előzetes eredmények a depresszió, migrén, ruminációs, szorongás és egyéb gyakran együtt jelentkező betegségek közös genetikai hátterének a felderítésére irányult [12,13,14].

A futtatáshoz a következő fenotípusos változóknak vizsgáltam a genetikai hátterét:

AGR = Agresszivitás, BROOD = Szomorkodás, CONS = Tudatosság, DEPR = Depresszió, DRUGALC = Drog vagy alkohol problémák, EATING = Evési zavar, EXTR = Kifelé forduló, IDG = Migrén, IMP = Impulzív, MANIC = Mániákus depresszió, NEUR = Neuroticizmus, OP = Nyíltság, REFL = Reflexió, reagálás, RUM = Rágódás, SCHIZO = Skrizofénia, SUIC = Öngyilkosság, SYANX = Aggodalmaskodás, SYIS = Törődés, SYOC = Kényszeres viselkedés.

A munkafolyamat korábban ismertetett alapparaméterei mellett a következő betegség-betegség hasonlóság adódott a közös, páronkénti genetikai átlapoltságok alapján,



12. ábra: két különböző paraméterezéssel számolt eredmény

Az alsó és felső háromszög ész két külön futás jelenít meg. A piros szín azt jelzi, hogy a vizsgálatok azt mutatják ki, hogy a két betegség között kapcsolat állhat fenn.

7. Összefoglalás

Kutató munkám során egy új területtel ismerkedtem meg, a bioinformatika, azon belül a statisztikai genetika [1], illetve részletesebben a teljes-genomi szélességű asszociációs vizsgálatokkal [2]. A genetikai változók milliós nagyságrendje és a rendkívül gazdag, többrétű háttértudás elérhetősége miatt új statisztikai módszertanok sokasága jött létre a hatékony következtetés és az eredmények értelmezésének támogatására [3,4]. A gének szintjén használt hálózati terjesztési módszerek sokféleségük ellenére egy univerzális kiegészítési lehetőséget kínálnak a genetikai adatelemzések kiterjesztésére [5].

A dolgozatban egy munkafolyamat automatizáló rendszert mutattam be és annak kiterjesztését a betegségek közös genetikai hátterének a felderítésére. Ennek megfelelően a dolgozat elsőként bemutatta a genetikai adatelemzési munkafolyamat lépéseit [4,5,6], annak támogatására létrehozott eszközeimet, (1) a genetikai variánsok génekhez rendelésének exportálását a FUMA rendszerből [7], (2) genetikai hálózatok exportálását a GeneMania rendszerből [8], illetve (3) a teljes adatelemzési lánc futtatását. Ez magában foglalja az asszociációs elemzést végző Bolt-LMM [10] és PLINK [9] futtatását, a génre aggregálást végző Pegasus [6] futtatását, és a hálózati terjesztést végző Hotnet2 futtatását [6,10].

A dolgozat második részében az automatizált futtatásokból származó párhuzamos eredmények elemzésére mutattam be módszereket és azok alkalmazását. A javasolt módszer kapcsolódó betegségek többváltozós, közös hátterének a felderítését segíti, amely a hálózati elemzések eredményeként adódó génhalmaz sorrendeken alapul. Nevezeten az automatizáló rendszer a munkafolyamatokat lefuttatva az egyes betegségekhez rendre egy génhalmaz sorrendet származtat, amelyben a génhalmazok sorrendjét és pontszámát a betegséggel való statisztikai kapcsolatuk határozza meg. A javasolt elemzési módszer pedig a betegségpárok, -hármasok, sokaságok génhalmaz sorrendjeiben keres közös motívumokat, lehetőséget biztosítva az orvosszakértői preferenciák specifikálására. A kidolgozott felületen megadható és keresésnél figyelembe vett pontszámmal súlyozható szempontok lehetőséget adnak annak figyelembe vételére, hogy a motívum (1) milyen és hány darab betegségben vesz részt, (2) a betegségek génhalmaz sorrendjeiben milyen helyezéshez (helyezésekhez) rendelhető egy halmaz-halmaz hasonlósággal, és (3) milyen halmaz-halmaz hasonlóság használt az elemzésben, amely gének elmulasztott felfedezésének, illetve gének hibás felfedezéséhez költségéhez kötődik.

A kidolgozott rendszert a Semmelweis Egyetem Gyógyszerhatástani Intézetének NewMood kutatásához kapcsolódva alkalmaztam az ottani adatokra Dr. Juhász Gabriella segítségével. Az előzetes eredmények a depresszió, migrén, rumináció, szorongás és egyéb gyakran együtt jelentkező betegségek közös genetikai hátterének a felderítésére irányult [12,13,14].

A jövőben ígéretes lehetőségnek tűnik a p-értékek helyett bayesi *a posteriori* valószínűségek használatát megvizsgálni, a gének fontosságának meghatározásához felhasználni úgynevezett génprioritizáló rendszereket, illetve a kombinatorikusan definiált job-ok kimerítő jellegű teljes lefuttatása helyett azok adaptív, részleges futtatásának szabályozását megvizsgálni.

Irodalomjegyzék

[1] Balding, D. J., Bishop, M., & Cannings, C. (Eds.). (2008). *Handbook of statistical genetics*. John Wiley & Sons.

[2] Stram, D. O. (2014). Design, analysis, and interpretation of genome-wide association scans. New York: Springer.

[3] Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., & Yang, J. (2017). 10 years of GWAS discovery: biology, function, and translation. *The American Journal of Human Genetics*, *101*(1), 5-22.

[4] Maier, R. M., Visscher, P. M., Robinson, M. R., & Wray, N. R. (2018). Embracing polygenicity: a review of methods and tools for psychiatric genetics research. *Psychological medicine*, *48*(7), 1055-1067.

[5] Cowen, L., Ideker, T., Raphael, B. J., & Sharan, R. (2017). Network propagation: a universal amplifier of genetic associations. *Nature Reviews Genetics*, *18*(9), 551.

[6] Nakka, P., Raphael, B. J., & Ramachandran, S. (2016). Gene and network analysis of common variants reveals novel associations in multiple complex diseases. *Genetics*, genetics-116.

[7] Watanabe, K., Taskesen, E., Bochoven, A., & Posthuma, D. (2017). Functional mapping and annotation of genetic associations with FUMA. *Nature communications*, *8*(1), 1826.

[8] Zuberi, K., Franz, M., Rodriguez, H., Montojo, J., Lopes, C. T., Bader, G. D., & Morris, Q. (2013). GeneMANIA prediction server 2013 update. *Nucleic acids research*, *41*(W1), W115-W122.

[9] Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., ... & Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, *81*(3), 559-575.

[10] Leiserson, M. D., Vandin, F., Wu, H. T., Dobson, J. R., Eldridge, J. V., Thomas, J. L., ... & Lawrence, M. S. (2015). Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nature genetics*, *47*(2), 106.

[11] Loh, P. R., Tucker, G., Bulik-Sullivan, B. K., Vilhjalmsson, B. J., Finucane, H. K., Salem, R. M., ... & Patterson, N. (2015). Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nature genetics*, *47*(3), 284.

[12] Marx, P., Antal, P., Bolgar, B., Bagdy, G., Deakin, B., & Juhasz, G. (2017). Comorbidities in the diseasome are more apparent than real: What Bayesian filtering reveals about the comorbidities of depression. *PLoS computational biology*, *13*(6), e1005487.

[13] Wang, K., Gaitsch, H., Poon, H., Cox, N. J., & Rzhetsky, A. (2017). Classification of common human diseases derived from shared genetic and environmental determinants. *Nature genetics*, *49*(9), 1319.

[14] Anttila, V., Bulik-Sullivan, B., Finucane, H. K., Walters, R. K., Bras, J., Duncan, L., ... & Patsopoulos, N. A. (2018). Analysis of shared heritability in common disorders of the brain. *Science*, *360*(6395), eaap8757.

[15] https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient, 2018, október, 29

Függelék

A munkafolyamatot támogató egyes modulok

- P-értékekből heat fájlok generálása Hotnet2 bemenetnek

  A Hotnet2 úgynevezett heat fájlokkal dolgozik, ez a génre nézve egy vektorizáció. Egy script segítségével és a Hotnet2 egy lépésével a génnév és p-érték párosból egy génnév - heat érték párost generáltam.

- Genemania kimenetből Hotnet2 hálózati bemenet készítése

  - Genema kimenetből éllista és indexlista előállítása

    A Genemania kimenete egy szöveg fájl, mely tabulátorokkal elválasztott oszlopokból áll. Az első két oszlopban gén nevek találhatóak. Ezek reprezentálják a hálózatot úgy, hogy az első és a második oszlopban szereplő gén között van kapcsolat. A Hotnet2 bemenetként egy éllistát és egy indexlistát vár. Ez a két fájl tulajdonképpen a Genemania kimenetének szétbontása úgy, hogy a géneket egy számmal helyettesítjük. Tehát az index fájlban egy génhez rendelünk egy indexet, az éllistában pedig a génnevek helyére a hozzájuk tartozó index kerül.

  - Éllista és indexlista felhasználása Hotnet2 bemenetként

    A Hotnet2 paraméterezéséhez szükséges a megadott éllista - indexlista páros megadása, melyet az algoritmus futása előtt feldolgozásra kerül, permutálja a megadott konfigurációnak megfelelően a hálózatot, később ezeket használja fel a számításaiban.

- Előre megadott heat-, hálózat- és konfigurációs fájlok felhasználása a Hotnet2 futásához

  A futáshoz felhasznált heat fájlokat, hálózatokat és parametrizációkat mind egy - egy fájlba kell beírni, majd ezeknek a behelyettesítésével, kiszámolásával lefut a megfelelő számú számítás, mely minden aktivizációra, minden hálózatban és minden parametrizációnak felhasználásával kiszámítható. Az összes futási eredményt összegyűjti és eltárolja későbbi felhasználásra, például vizualizációra.