Genetikai variánskivonatoló munkafolyamatok automatikus fúziója és a bayesi relevanciaelemzés alkalmazása jelölt gén asszociációs vizsgálatokban

Doktori tézisek

Gézsi András

Semmelweis Egyetem
Molekuláris Orvostudományok Doktori Iskola

Témavezető:
Dr. Szalai Csaba, az MTA doktora, egyetemi tanár

Hivatalos bírálók:
Dr. Rónai Zsolt, PhD, egyetemi adjunktus
Dr. Maróti Zoltán, PhD, tudományos főmunkatárs

Szigorlati bizottság elnöke:
Dr. Dinya Elek, CSc, egyetemi tanár

Szigorlati bizottság tagjai:
Dr. Kiss András, PhD, egyetemi docens
Dr. Pataki Béla, PhD, egyetemi docens

Budapest
2016

Bevezetés

A genetikai és genomikai kutatások jelentősége egyre nagyobb az orvostudományban. A humán genom szekvenciájának teljes meghatározása, az egyre gyorsabb és olcsóbb szekvenálási technológiák rohamos fejlődése következtében a személyre szabott orvoslás bizonyos területeken már a klinikai rutin részévé vált. A szekvenálási adatok mennyiségének soha nem látott mértékű növekedése azonban jelentős kihívásokat támaszt az adatokat értelmezni és elemezni kívánó orvosok, biológusok és bioinformatikusok számára. A genetikai variánsok elemzése során az új generációs szekvenálási vizsgálatokkal meghatározott biológiai konklúziók nagy mértékben a hívott variánsok és genotípusok pontosságán alapulnak, amely azonban még nem minden esetben éri el a klinikai diagnosztikában való felhasználhatóság szintjét. Emiatt azok a bioinformatikai módszerek, amelyek javítani tudnak a variánshívások pontosságán, nagy mértékben hozzájárulhatnak a technológiák minél szélesebb körű használhatóságához. A munkám során kifejlesztettem egy szoftvert, amely különböző variánskivonatoló módszerek eredményének kombinálásával jobb teljesítményre képes, mint az egyedi módszerek.

A genetikai variánsok elemzése központi jelentőségű a betegségek patomechanizmusának feltárásában, a betegségre való hajlam, illetve a gyógyulást befolyásoló tényezők felderítésében és eredményesebb kezelési lehetőségek, terápiás protokollok kidolgozásában. A Bayes-statisztikán alapuló módszerek egyre nagyobb teret hódítanak a genetikai adatelemzésben is. A munkám során részt vettem a bayesi relevanciaelemzési módszertan kifejlesztésében, amely a genetikai variánsok és fenotípusos jellemzők komplex összefüggésrendszerének feltérképezésével a frekventista statisztikai módszerek hatékony alternatíváját nyújtja asszociációs vizsgálatok adatainak elemzésére. A bayesi módszertan használhatóságát és előnyeit a gyermekkori akut limfoid leukémia hajlamát és túlélését befolyásoló polimorfizmusok elemzésén keresztül mutatom be.

Új generációs szekvenálás

Az új generációs szekvenálási (next-generation sequencing, NGS) technológiák megjelenése forradalmasította többek között a humán genetikai és genomikai kutatásokat

is. A teljes genom, illetve teljes exom szekvenálás segítségével ritka és komplex betegségek genetikai háttere is felderíthető. A technológia folyamatos fejlődése és a gyártó cégek versenye miatt egyre nagyobb áteresztőképességű szekvenáló berendezések jelennek meg, amelyekkel egy bázis meghatározásának fajlagos költsége egyre olcsóbb.

A teljes NGS munkafolyamat meglehetősen komplex, sok elemzési lépésből áll, amely számos szoftver és adatbázis használatán alapul. Emiatt nem meglepő, hogy rengeteg bioinformatikai eszköz született az egyes elemi lépések, illetve akár a teljes folyamat elvégzésére, azonban a megfelelő eszközök kiválasztása és beállítása nem triviális. Számos kutatás kimutatta, hogy (1) nincs legjobb variánskivonatolási módszer vagy olyan konkrét munkafolyamat-beállítás, amelynek teljesítménye általános körülmények között, minden esetben felülmúlná a többiét és (2) jelentős eltérés van a széles körben használt variánskivonatoló munkafolyamatok eredményei között, még abban az esetben is, ha ugyanazokra a mérési adatokra alkalmazzák azokat.

Bayes-háló alapú relevanciaelemzés

A Budapesti Műszaki és Gazdaságtudományi Egyetem bioinformatikai munkacsoportjának tagjaként, dr. Antal Péter vezetésével, részt vettem egy statisztikai módszertan kidolgozásában, amely többek között genetikai asszociációs adatok elemzésére használható. A módszertan ún. Bayes-hálókat használ a tárgyterület változóinak modellezésére, illetve Bayes-statisztikai módszerekkel meghatározza a változók közötti komplex függőségek valószínűségét.

A genetikai asszociációs vizsgálatok során a célunk az, hogy meghatározzuk azokat a genetikai variánsokat, amelyek befolyásolják egy adott fenotípus megjelenését (pl. egy betegség kialakulását), azaz tulajdonképpen a genotípus és a fenotípus komplex összefüggésrendszerét szeretnénk megismerni. Minden egyes megfigyelés (minta) tekinthető a megismerni kívánt rendszer egy adott állapotának, amit a minta konkrét genotípusa és fenotípusos jellemzői írnak le. Amennyiben ezeket valószínűségi változóknak tekintjük, akkor a célunkat úgy is megfogalmazhatjuk, hogy a tárgytartományt leíró együttes valószínűségi eloszlást, illetve annak struktúráját akarjuk feltérképezni. Ehhez a feladathoz használjuk az ún. Bayes-háló alapú relevanciaelemzést.

A gyermekkori akut limfoid leukémia

A leukémia a vérképző szervek rosszindulatú megbetegedését magában foglaló betegségcsoport, amelyben a kóros fehérvérsejtek burjánzása túlnövi a normális sejteket, és infiltrálja a különböző szerveket (pl. csontvelőt, idegrendszert, szemet stb.). A különböző leukémia típusok közül az akut limfoid leukémia (ALL) a leggyakoribb (kb. 80%), amely egy multifaktoriális betegség; genetikai és környezeti faktorok egyaránt befolyásolják a betegség kialakulását.

Az ALL ötéves túlélési aránya napjainkban körülbelül $80 - 90\%$. A jövőben új gyógyszercélpontok azonosításával, személyre szabott terápiával, a késői mellékhatások és a gyógyszertoxicitás kiszűrésével remélhetőleg ez az arány tovább javítható. Ennek sikeréhez nagyban hozzájárulhatnak a különböző farmakogenetikai vizsgálatok, amelyek a gyógyszermetabolizáló enzimeket, illetve gyógyszercélpontokat és transzportereket kódoló gének variánsainak hatásait elemzik.

Az emberi szervezetben a citokróm P450 enzimcsalád tagjai számos különböző típusú gyógyszer, valamint endogén anyag oxidatív metabolizmusában játszanak szerepet. A CYP3A4 a májban és a bélben legnagyobb mennyiségben előforduló citokróm P450 enzim, és ezáltal az egyik legfontosabb gyógyszermetabolizáló fehérje. Fontos szerepe van az ALL terápiájában alkalmazott gyógyszerek egy részének metabolizmusában (pl. vinkrisztin, ciklofoszfamid, dexametazon és doxorubicin). A CYP3A5 enzim szubsztrátspecificitása pedig nagy mértékben átfed a CYP3A4-el. Ennek ellenére gyakorlatilag még nem született olyan tanulmány, amelyben a *CYP3A4* vagy a *CYP3A5* polimorfizmusainak az ALL túlélésében betöltött szerepét vizsgálták volna. Ennek egyik oka az lehet, hogy az eddig talált funkcionális variánsok, amelyek befolyásolták a gének expresszióját, alacsony allélfrekvenciával rendelkeztek (maximum $4 - 5\%$) az egyébként kisméretű populációkban, így ezeknek az elemzéseknek a statisztikai ereje alacsony volt. A Genetikai, Sejt- és Immunbiológiai Intézet rendelkezésére álló biobank mérete azonban lehetővé teszi ezeknek a géneknek a vizsgálatát jóval nagyobb mintaszámon is.

Mivel a kemoterápiás kezelések során adott gyógyszerek hatásosságát és hatékonyságát számtalan gén, illetve azok bonyolult egymásra hatása befolyásolja, a farmakogenetikai vizsgálatokban fontos szerepe van azoknak a módszereknek, amelyek több változó együttes hatását, illetve az interakcióikat is képesek kimutatni.

Célkitűzés

A munkám során a következő célkitűzéseim voltak:

1. Különböző variánskivonatolási munkafolyamatok teljesítményének és együtt járásának (konkordanciájának) összehasonlítása, különös tekintettel az illesztőprogram megválasztásának és a leolvasási mélység hatásának elemzésére. A jelenleg javasolt manuális szűrési beállítások hatásának kiértékelése a munkafolyamatok szenzitivitására és precizitására.

2. Egy olyan szoftver kifejlesztése, amely az egyedi variánskivonatolási módszerek eredményét kombinálja, amely során felhasználja a variánsok minőségét leíró annotációs jellemzőket is. A program nagy megbízhatóságú referencia variánskészletek használata *nélkül*, kisebb genomi régiók vagy kevés minta esetén is tegye lehetővé a variánsok valószínűségének becslését és ezáltal a precizitás alapú szűrést. A kifejlesztett program teljesítményének összehasonlítása az egyedi variánskivonatolók, illetve egy alternatív kombinációs program (BAYSIC) eredményeivel.

3. A *CYP3A4* gén és a *CYP3A5* gyakori polimorfizmusainak a gyermekkori ALL túlélését befolyásoló hatásának vizsgálata, interakciók keresése a bayesi relevanciaelemzés segítségével.

4. A Bayes-háló alapú relevanciaelemzési módszertan tesztelése és összehasonlítása a frekventista statisztikával asszociációs vizsgálatokban. Ennek során a bayesi módszertan tesztelése, a frekventista vizsgálatok eredményeivel való összevetése, az eredmények metodológiai szempontból történő elemzése, illetve ezek alapján a módszertan továbbfejlesztése.

Módszerek

Variánskivonatoló programok teljesítményének és együtt járásának elemzése, a kivonatolási eredmények kombinálása

Az egyes variánskivonatoló programok, illetve az általam kifejlesztett VariantMetaCaller program teljesítményének összehasonlításához mesterséges szekvenciaadatokat állítottam elő. Az egyes variánskivonatolási módszerek szenzitivitásának és precizitásának mérése során során az általam generált mintákban szereplő valódi variánsokat használtam referenciaként, azaz az egyes módszerek eredményét ezekkel a referenciavariánsokkal vetettem össze. A szimulált adatok mellett valós szekvenálási adatokat is felhasználtam, melyet az Illumina BaseSpace honlapjáról töltöttem le. Ehhez elérhető egy nagy pontosságú, „platinum minőségű" referencia variánslista is, amit az összehasonlítások során referenciaként használtam.

A leolvasásokat először minőségi szűréseknek vetettem alá, majd a BWA–MEM és a Bowtie 2 programok segítségével felillesztettem a hg19 referenciagenomra. Ezt követően négy különböző variánskivonatoló programot futtattam le a két különböző illesztési eredményen az SNP-k és rövid indelek detektálására: a GATK UnifiedGenotyper és HaplotypeCaller programját, a FreeBayes-t és a SAMtools-BCFtools programok kombinációját. A GATK alapú kivonatolók eredményét manuális szűrők használatával leszűrtem a GATK ajánlásainak megfelelően. A FreeBayes és a SAMtools által hívott variánsokat szintén leszűrtem a variánsminőség minimális küszöbértékének meghatározásával.

Az általam kifejlesztett VariantMetaCaller program az egyedi szűretlen variánskivonatolási eredményeket kombinálja ún. szupport vektor gépek (SVM) használatával. Első lépésben a különböző variánskivonatolási módszerek által hívott variánsokat a program egyesíti, és az átlapolódó, esetlegesen eltérő variánsreprezentációkat egységesíti. Ezt követően minden egyes variánskivonatolóhoz és variánstípushoz (azaz külön az SNP-kre és külön az indelekre) a program egy SVM-et tanít. A tanítóminták kiválasztásához a következő heurisztikát használjuk: azok a variánsok lesznek a pozitív tanítóminták, amelyeket minden kivonatoló módszer megtalál, és azok lesznek a negatív tanítóminták, amelyeket csak egyetlen kivonatoló program talál meg. Végül mindegyik variánskivonatoló esetén minden egyes variánshoz ki-

számítjuk annak feltételes valószínűségét, hogy az adott variáns "valódi" (azaz a pozitív osztályba tartozik), majd ezt kiátlagoljuk a variánskivonatolók felett.

Az egyes módszerek teljesítményét precizitás-szenzitivitás görbék segítségével hasonlítottuk össze.

A CYP3A4 potenciális szerepének vizsgálata a gyermekkori akut limfoid leukémia farmakogenetikájában

A gyermekkori akut limfoid leukémia túlélését befolyásoló genetikai és környezeti faktorok elemzése során a Semmelweis Egyetem Genetikai Sejt- és Immunbiológiai Intézetében kialakított biobankot használtuk. A vizsgálatok során 1990 és 2010 között ALL-el diagnosztizált betegek adatait elemeztük (511 beteg). A vizsgálatokba bevont személyektől perifériás vérmintát gyűjtöttek.

A vizsgálat során a *CYP3A4* és *CYP3A5* gén egyes polimorfizmusait elemeztük.

A betegek DNS-ének izolálása QIAmp DNA Blood Midi/Maxi Kittel (Qiagen, Hilden, Németország) történt, a gyártó által előírt protokolloknak megfelelően. Az SNP-k genotipizálása Sequenom iPLEX Gold MassARRAY technológiával történt a kanadai McGill Egyetem és Génome Québec Innovációs Központban (Montreal, Kanada).

Az adatok frekventista elemzését az R statisztikai szoftverrel végeztem. A túlélési adatok egy- és többváltozós elemzésére Cox-regressziós modellt alkalmaztam. A statisztikai erő elemzéséhez a bootstrap módszert használtam.

A rizikócsoport-besorolási változók diszkriminatív teljesítményét a C-index (konkordancia index) kiszámításával végeztem. Ennek konfidencia-intervallumát bootstrap módszerrel számítottam ki. A kockázat-besorolási változók közötti különbséget t-teszttel számítottam ki minden egyes időpontra, majd a p-értékeket Benjamini-Hochberg módszerével korrigáltam.

Az adatokat bayesi relevanciaelemzéssel is vizsgáltuk. Ennek során a lehetséges Bayes-háló struktúrák mintavételezéséhez az ún. MC^3 algoritmust (Metropolis Coupled Markov Chain Monte Carlo) használtuk. A mintavételezett gráfstruktúrák alapján kiszámítottuk a változók közötti kapcsolati típusok és a célváltozó szempontjából erősen releváns változóhalmazok *a posteriori* valószínűségét, illetve a változók közötti interakciókat és redundanciákat.

Eredmények

Variánskivonatolási munkafolyamatok teljesítménye és konkordanciája

Az egyedi variánskivonatolási munkafolyamatok teljesítményét szimulált adatok segítségével hasonlítottuk össze. Az eredményeink más kutatócsoportokkal egyetértésben azt mutatták, hogy nem volt olyan általánosan legjobbnak mondható módszer, amelynek a szenzitivitása és precizitása is a lefedettségtől függetlenül felülmúlta volna a többiét. Általánosságban elmondható azonban, hogy a HaplotypeCaller jól teljesített: ez találta meg a legtöbb valódi indelt, és ez bizonyult a legprecízebbnek SNP-k kivonatolása esetén. Kimutattuk, hogy a leolvasási mélység növekedésével nőtt a variánskivonatoló módszerek szenzitivitása. Meglepő módon azonban a hamisan hívott variánsok száma egy bizonyos lefedettség fölött szintén nőtt, azaz a módszerek precizitása csökkent a lefedettség növekedésével.

Az egyes módszerek szenzitivitása és precizitása azonos lefedettség mellett jóval magasabb volt SNP-k, mint indelek esetén. Az eredményeink azt mutatják, hogy jelentős lefedettségbeli növekedésre van szükség ugyanakkora szenzitivitás eléréséhez (pl. az SNP-k esetén $16\times$ lefedettségnél tapasztalt szenzitivitást indelek esetén csak $200\times$ lefedettség mellett tudta elérni a HaplotypeCaller).

Az illesztőprogram megválasztása jelentősen befolyásolta a variánskivonatolás eredményét. A BWA használata általában szignifikánsan jobb eredményekre vezetett, mint a Bowtie 2 program használata.

Kimutattuk, hogy jelentős eltérés van a széles körben használt variánskivonatoló munkafolyamatok eredményei között. Az illesztőprogramok különbsége a variánskivonatolási módszerek konkordanciájára is hatással volt, a módszerek együtt járása ugyanis általában kisebb volt a Bowtie 2 illesztések használatakor.

A variánskivonatolás precizitásának javítása érdekében a bioinformatikai kiértékelések során gyakran alkalmaznak manuális variáns szűrést. Mivel azonban nem áll rendelkezésünkre olyan mutató vagy mutatók olyan kombinációja, amely egyértelműen megkülönböztetné a valódi és a hibásan hívott variánsokat, a precizitás és a szenzitivitás fordítottan viszonyulnak egymáshoz, azaz a precizitást csak a szenzitivitás csökkenése árán tudjuk növelni. A manuális szűréseket a jelenlegi ajánlá-

soknak megfelelően végeztük, észben tartva, hogy ezek nem feltétlenül jelentenek optimális megoldást. A manuális szűrők hatásának lefedettségtől való függése jelentős mértékben különbözött a GATK-, illetve nem GATK alapú variánskivonatoló módszerek esetén. A FreeBayes és a SAMtools esetén ugyanis a jelenlegi ajánlások szerint egyedül a becsült variánsminőség alapján, egy küszöbérték meghatározásával történt a variánsok szűrése. Mivel a variánsminőség mutató értéke a lefedettség növekedésével általában szintén nőtt egy adott variáns esetén, így a rögzített küszöbérték használata miatt egyre kevesebb variánst szűrtünk ki. A HaplotypeCaller és a UnifiedGenotyper esetén a szűrőfeltételek több mutató értékén alapulnak, így a lefedettségtől való függés is összetettebb.

Összességében az eredményeink azt mutatják, hogy a manuális szűrések haszna korlátozott volt: (1) a szenzitivitás általában nagyobb mértékben csökkent, mint amennyire a precizitás növekedett a szűrés hatására, illetve (2) ugyanaz a szűrőbeállítás nem volt megfelelő minden leolvasási mélység esetén.

Variánskivonatolók kombinálása: VariantMetaCaller

A VariantMetaCaller program egyedi variánskivonató módszerek eredményeit kombinálja, kihasználva azok erősségeit és komplementaritását.

Mivel minden kivonatoló módszer esetén vannak olyan valódi variánsok, amelyeket az nem talál meg, de egy vagy több másik módszer igen, ezért a VariantMetaCallerrel kombinált variánsok maximális szenzitivitása magasabb volt, mint bármelyik egyedi módszeré. A szenzitivitás növelésén túl a precizitás maximalizálása is alapvető fontosságú. Ezért azt is figyelembe kell venni, hogy egy adott módszer által kiszámított mutató mennyire képes megkülönböztetni a valódi és a hamis variánsokat. Az eredmények alapján a VariantMetaCaller által meghatározott variáns valószínűségi pontszám teljesíti ezt az elvárást: a variánsokat valószínűség szerint csökkenő sorrendbe állítva a precizitás a sorrend mentén a szenzitivitás növekedésével lassan csökkent, és csak a nagy szenzitivitás értékeknél kezdett élesen csökkenni.

Összességében elmondható, hogy a szimulált és a valós adatokon végzett elemzések eredménye alapján a VariantMetaCaller a leolvasási mélységtől, az illesztőtől és a variánstípustól függetlenül nagyobb precizitást ért el minden szenzitivitási szinten mint a bemenetéül szolgáló egyedi variánskivonató módszerek.

A variánsok sorrendezésére, illetve a valódi–hamis variánsok megkülönböztetésére használható mutatók teljesítményének számszerűsítésére a precizitás–szenzitivitás görbe alatti területet (AUPRC) használtuk. A VariantMetaCaller AUPRC pontszáma a lefedettségtől, az illesztőtől és a variánstípustól függetlenül minden esetben magasabb volt, mint az egyedi variánskivonatolók AUPRC értéke mind a szimulált, mind a valós adatokon végzett elemzések eredménye alapján. Szimulált adathalmazok használatával azt is megmutattuk, hogy a VariantMetaCaller kisebb méretű – tipikusan a célzott génpanelek méretéhez hasonló – célrégiók esetén is jobb teljesítményt nyújtott.

A dolgozat egyik célkitűzése az volt, hogy megmutassuk a köztes annotációs információ felhasználásának előnyét a variánshívások fuzionálásakor. Ennek érdekében a VariantMetaCaller által elért eredményeket összehasonlítottuk a BAYSIC-kel, amely ún. késői fúziót valósít meg, azaz a variánshívók kombinálásakor csak a konkrét variánshívásokat használja fel, annotációs adatokat nem. A teljes exomon elért eredményeket tekintve a VariantMetaCaller nyújtott jobb teljesítményt; az AUPRC értékek különbsége $1 - 4\%$ volt. Ez figyelemreméltó, ugyanis 1%-nyi különbség kb. 473 SNP és 49 indel pontosabb sorrendezését jelenti a jelenlegi kísérleti beállítások mellett. Ezen felül kiszámítottuk az AUPRC értékeket a két módszer esetén minden egyes kromoszómára szűkítve is, és azt találtuk, hogy a VariantMetaCaller az esetek legnagyobb részében jobb teljesítményt nyújtott mint a BAYSIC, és a különbség statisztikailag is erősen szignifikáns volt.

A munkám másik célkitűzése az volt, hogy egy rugalmas, könnyen értelmezhető megoldást adjak a variánsok szűrésére, a hamis felfedezési arányon alapuló paradigma analógiájára. Ez a variánsok valószínűségének pontos becslésével válik elérhetővé: a valószínűségi értékekkel a variánsokat sorrendezhetjük, majd minden egyes küszöbértékre ki tudjuk számítani a várható precizitást. Ezután a precizitás közvetlenül átfordítható a valódi, vagy ezzel ekvivalens módon a hamis variánsok várható számára. A VariantMetaCaller a variánsok valószínűségét pontosabban becsülte mint a többi módszer, így a program támogatja a kvantitatív, alkalmazás-specifikus szűrés lehetőségét.

A *CYP3A4* és a *CYP3A5* gének kiválasztott polimorfizmusainak hatása a gyermekkori ALL túlélésére

A munkám során a *CYP3A4* és a *CYP3A5* gének kiválasztott polimorfizmusainak a gyermekkori akut limfoid leukémia túlélését befolyásoló hatását vizsgáltam. Ennek során azt találtam, hogy a *CYP3A4* gén egy gyakori SNP-je (rs2246709) szignifikánsan befolyásolta az ALL-es betegek kemoterápia utáni túlélését, és ezt a hatást a páciens neme erősen befolyásolta. A nemek közötti különbség különösen jelentős volt az AG heterozigóta genotípusú betegek esetén; ebben az esetben a fiúknak szignifikánsan magasabb volt a túlélési aránya mint a lányoknak. Ezzel szemben a vad homozigóta AA genotípus fiúkban rosszabb túlélési aránnyal asszociált, mint lányokban.

A páciens neme és az rs2246709 polimorfizmus interakciója alapján egyszerű szabályok segítségével egy olyan új rizikócsoport-besorolást tudtunk megállapítani, amelynek a kockázatbecslési teljesítménye szignifikánsan felülmúlta a jelenlegiét.

A *CYP3A4* polimorfizmusai és további - az ALL hajlamosításban, illetve a folát-anyagcserében részt vevő kiválasztott 34 gén SNP-i között a bayesi relevanciaelemzés segítségével interakciókat kerestünk. Ennek során azt találtuk, hogy az *MTHFD1* gén egy SNP-je (rs1076991) szignifikánsan befolyásolta az rs2246709 hatását a túlélésre. Ez a gén a folát-anyagcsere útvonal része, amely a metotrexát kemoterápiás szer célpontja. Az rs1076991 polimorfizmus GG genotípusa megnövelte a B-sejtes ALL kialakulásának valószínűségét, de önmagában nem volt hatása a túlélésre. A CYP3A4 nem metabolizálja a metotrexátot, így feltehetően a két SNP hatása a különböző útvonalakon összeadódik.

A bayesi relevanciaelemzési módszertan alkalmazási lehetőségeinek vizsgálata asszociációs vizsgálatokban

A bayesi relevanciaelemzést (a *CYP3A4* gén elemzésén kívül) két jelölt gén asszociációs vizsgálatban alkalmaztuk a gyermekkori ALL hajlamosító tényezőinek felderítésére, illetve egy parciális genomszűrési vizsgálatban az asztma genetikai hajlamosító tényezőinek tanulmányozására.

Az elemzések során a bayesi relevanciaelemzési módszer több előnyös tulajdon-

ságára is sikerült rávilágítani, melyek a következők:

- A Bayes-statisztikai megközelítés miatt az egyes eredmények (hipotézisek) direkt valószínűségi állítások formájában fogalmazhatók meg, amelyek pontosan tükrözik az adatainkban rejlő információt. Ez, a frekventista megközelítéssel szemben, akár azt is lehetővé tenné, hogy az állításokból olyan valószínűségi adat- és tudásbázisokat építsünk, amelyek támogatják a komplex valószínűségi lekérdezéseket, a meta-analízisek könnyebb elvégezhetőségét, illetve az eredmények háttértudással való fúzióját.

- Mivel a módszer eredendően többváltozós, így egyszerre tudjuk elemezni az összes polimorfizmus, környezeti tényező és fenotípusos leíró függését a célváltozótól, illetve a változók bonyolult összefüggésrendszerét is. Ezáltal a változók direkt és tranzitív hatásai is megkülönböztethetők egymástól, illetve a módszer a különféle kapcsolati típusok definiálásával egy jóval gazdagabb nyelvet nyújt a célváltozót befolyásoló tényezők hatásának leírására és értelmezésére. Hagyományos statisztikai módszerekkel a változók közötti kapcsolatrendszer hasonló részletességgel csak korlátozott módon lenne felderíthető (pl. változópáronkénti asszociációs tesztekkel), melyet tovább súlyosbít a többszörös hipotézistesztelés miatt fellépő korrekció szükségessége.

- A többváltozós modellezés miatt lehetőség van a változók közötti interakciók és redundanciák feltérképezésére is. A bayesi relevanciaelemzés a folát anyagcserében szerepet játszó gének vizsgálatakor például egy komplex interakciós hatást mutatott ki a hiperdiploid ALL alcsoportban.

- A Bayes-statisztika továbbá egy automatikus és normatív megoldást ad a frekventista statisztikai módszereket sújtó többszörös hipotézistesztelési problémára, így az eredményeket nem kell korrigálni.

- Lehetőség van továbbá egyszerre több célváltozó kezelésére is azáltal, hogy gyakorlatilag tetszőleges strukturális kérdés *a posteriori* valószínűsége kiszámítható a módszer segítségével. Ennek jelentőségét egy asztma parciális genomszűrési elemzés során mutattuk meg.

Következtetések

A munkám eredményeképpen az alábbi következtetéseket vonhatjuk le:

1. Kifejlesztettünk egy új módszert (VariantMetaCaller), amely új generációs szekvenálási variánskivonatoló programok variánshívási eredményeit kombinálja. A módszer a kombináció során (1) kihasználja az egyedi kivonatoló programok alacsony konkordanciáját illetve komplementaritását, (2) felhasználja az egyedi kivonatoló programok által generált nagy-dimenziós annotációs adatokat és (3) megbecsüli a variánsok valódiságának valószínűségét. Szimulált és valós szekvenálási adatok felhasználásával megmutattuk, hogy a VariantMetaCaller az általunk vizsgált genomi régió méretek esetén, néhány száz kilobázistól a teljes exomi méretig, szignifikánsan jobb teljesítményt nyújtott, mint a bemenetéül szolgáló variánskivonatolási módszerek.

2. Valós szekvenálási adatok használatával megvizsgáltuk a variánsok valódiságának valószínűségét becslő módszerek pontosságát. Az eredményeink szerint a VariantMetaCaller pontosabban becsülte a várható precizitást mint az alternatív módszerek. Ezáltal a VariantMetaCaller egy könnyen értelmezhető, kvantitatív, *várható precizitás* alapú szűrőt ad a felhasználó kutatók, biológusok, orvosok kezébe, amely lehetővé teszi, hogy megkeressük a variánskivonatolás szenzitivitásának és precizitásának alkalmazás-specifikus egyensúlyát. Az eredményeink alapján a VariantMetaCaller célzott génpanelek, illetve olyan organizmusok szekvenálása esetén is használható, amelyekhez jelenleg nem állnak rendelkezésre nagy megbízhatóságú referencia variánskészletek. Ezáltal a kvantitatív, precizitás alapú szűrés azokon az alkalmazási területeken is lehetővé válik, ahol eddig csak manuális szűréseket lehetett használni a variánshívások precizitásának növelésére.

3. Az egyik legfontosabb gyógyszer-metabolizáló enzim, a CYP3A4 génjének polimorfizmusait vizsgálva megmutattuk, hogy az rs2246709 SNP szignifikánsan befolyásolja az akut limfoid leukémia kemoterápiás kezelésének teljes és eseménymentes túlélésének kockázatát. A bayesi relevanciaelemzés segítségével kimutattuk, hogy a *CYP3A4* gén rs2246709 polimorfizmusának és a

folát-anyagcserében szerepet játszó *MTHFD1* gén egyik polimorfizmusának interakciója szignifikánsan befolyásolja a túlélési kockázatot.

4. A bayesi relevanciaelemzés segítségével kimutattuk, és frekventista statisztikai elemzéssel megerősítettük, hogy a *CYP3A4* gén rs2246709 polimorfizmusának és a páciens nemének interakciója szignifikánsan befolyásolja a gyermekkori akut limfoid leukémia kemoterápiás kezelésének túlélési kockázatát. Az interakciós hatást leíró egyszerű szabályok segítségével egy olyan új rizikócsoport-besorolási változót sikerült létrehozni, amely az eredeti besoroláshoz képest szignifikánsan jobb kockázatbecslést tett lehetővé.

5. Megmutattuk, hogy a bayesi relevanciaelemzési módszer a hagyományos frekventista statisztikai módszerekkel szemben a változók direkt és tranzitív hatásainak megkülönböztetésével, a különféle kapcsolati típusok definiálásával, illetve az interakciók és redundanciák automatikus feltérképezésével jóval részletgazdagabb elemzést tesz lehetővé genetikai asszociációs vizsgálatok adatainak elemzése során.

Saját publikációk jegyzéke

Az értekezésben felhasznált közlemények:

Gézsi A, Bolgár B, Marx P, Sarkozy P, Szalai C, Antal P. (2015) VariantMetaCaller: Automated fusion of variant calling pipelines for quantitative, precision-based filtering. BMC Genomics, 16 (1): 875. IF: 3,986

Gézsi A, Lautner-Csorba O, Erdélyi DJ, Hullám G, Antal P, Semsei ÁF, Kutszegi N, Hegyi M, Csordás K, Kovács G, Szalai C. (2015) In interaction with gender a common CYP3A4 polymorphism may influence the survival rate of chemotherapy for childhood acute lymphoblastic leukemia. Pharmacogenomics J, 15 (3): 241–247. IF: 4,229

Lautner-Csorba O, **Gézsi A**, Erdélyi DJ, Hullám G, Antal P, Semsei ÁF, Kutszegi N, Kovács G, Falus A, Szalai C. (2013) Roles of Genetic Polymorphisms in the Folate Pathway in Childhood Acute Lymphoblastic Leukemia Evaluated by Bayesian Relevance and Effect Size Analysis. PLoS One, 8 (8): e69843. IF: 3,534

Lautner-Csorba O, **Gézsi A**, Semsei AF, Antal P, Erdélyi DJ, Schermann G, Kutszegi N, Csordás K, Hegyi M, Kovács G, Falus A, Szalai C. (2012) Candidate gene association study in pediatric acute lymphoblastic leukemia evaluated by Bayesian network based Bayesian multilevel analysis of relevance. BMC Med Genomics, 5: 42. IF: 3,466

Ungvári I, Hullám G, Antal P, Kiszel PS, **Gézsi A**, Hadadi E, Virág V, Hajós G, Millinghoffer A, Nagy A, Kiss A, Semsei AF, Temesi G, Melegh B, Kisfali P, Széll M, Bikov A, Gálffy G, Tamási L, Falus A, Szalai C. (2012) Evaluation of a partial genome screening of two asthma susceptibility regions using bayesian network based bayesian multilevel analysis of relevance. PLoS One, 7 (3): e33573. IF: 3,730

Antal P, Hullam G, **Gézsi A**, Millinghoffer A. (2006) Learning complex bayesian network features for classification. Proc. of third European Workshop on Probabilistic Graphical Models. Prague, Czech Republic; 9–16.

Az értekezésben felhasznált közlemények kumulatív impakt faktora: 18,945

Az értekezésben felhasznált könyvfejezetek:

Hullám G, **Gézsi A**, Millinghoffer A, Sárközy P, Bolgár B, Srivastava SK, Pál Z, Buzás EI, Antal P. Bayesian systems-based genetic association analysis with effect strength estimation and omic wide interpretation: a case study in rheumatoid arthritis. Methods Mol Biol 2014, 1142: 143–176.

Antal P, Millinghoffer A, Hullam G, Hajos G, **Gezsi A**, Szalai C, Falus A. Bayesian, Systems-based, Multilevel Analysis of Associations for Complex Phenotypes: from Interpretation to Decision. In: Christine Sinoquet, Raphael Mourad (szerk.) Probabilistic graphical models for genetics. Oxford University Press, New York, 2014: 319-360.

Gézsi A. Génexpressziós adatok standard asszociációs elemzése. In: Antal P (szerk.), Bioinformatika: Molekuláris méréstechnikától az orvosi döntéstámogatásig. Typotex Kiadó, Budapest, 2014: 107-120.

Egyéb – az értekezésben fel nem használt – eredeti közlemények:

Kutszegi N, Semsei AF, **Gézsi A**, Sági JC, Nagy V, Csordás K, Jakab Z, Lautner-Csorba O, Gábor KM, Kovács GT, Erdélyi DJ, Szalai C. (2015) Subgroups of Paediatric Acute Lymphoblastic Leukaemia Might Differ Significantly in Genetic Predisposition to Asparaginase Hypersensitivity. PLoS One, 10 (10): e0140136. IF: 3,234

Temesi G, Virág V, Hadadi E, Ungvári I, Fodor LE, Bikov A, Nagy A, Gálffy G, Tamási L, Horváth I, Kiss A, Hullám G, **Gézsi A**, Sárközy P, Antal P, Buzás E, Szalai C. (2014) Novel genes in Human Asthma Based on a Mouse Model of Allergic Airway Inflammation and Human Investigations. Allergy Asthma Immunol Res, 6 (6): 496–503. IF: 2,160

Béres A, Lelovics Z, Antal P, Hajós G, **Gézsi A**, Czéh A, Lantos E, Major T. (2011)

„Does happiness help healing?" Immune response of hospitalized children may change during visits of the Smiling Hospital Foundation's Artists. Orv Hetil, 152 (43): 1739–1744.

Gézsi A, Budde U, Deák I, Nagy E, Mohl A, Schlammadinger Á, Boda Z, Masszi T, Sadler JE, Bodó I. (2010) Accelerated clearance alone explains ultra-large multimers in von Willebrand disease Vicenza. J Thromb Haemost, 8 (6): 1273–1280. IF: 5,439

Egyéb – az értekezésben fel nem használt – könyvfejezetek:

Gézsi A. Metagenomika. In: Antal P (szerk.), Bioinformatika: Molekuláris méréstechnikától az orvosi döntéstámogatásig. Typotex Kiadó, Budapest, 2014: 264-273.

Az összes publikáció kumulatív impakt faktora: 29,778