DOI:10.14753/SE.2017.1938

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

Doktori értekezés

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

DOI:10.14753/SE.2017.1938

Tartalomjegyzék

DOI:10.14753/SE.2017.1938

DOI:10.14753/SE.2017.1938

DOI:10.14753/SE.2017.1938

Rövidítések jegyzéke

Rövidítés	Angol elnevezés	Magyar elnevezés
ALL	acute lymphoid leukemia	akut limfoid leukémia
AUC, AUROC	area under the receiver operator characteristic curve	szenzitivitás-specificitás görbe alatti terület
AUPRC	area under the precision-recall curve	precizitás-szenzitivitás görbe alatti terület
BFM	Berlin, Frankfurt, Münster	-
CI	confidence interval	konfidencia-intervallum
CPU	central processing unit	központi számítási egység
CR	credibility region	hihetőségi tartomány
FDR	false discovery rate	hamis felfedezési arány
FN	false negative	hamis negatív
FP	false positive	hamis pozitív
FWER	familywise error rate	családi-szintű hiba
GATK	Genome Analysis Toolkit	-
HR	hazard ratio	hazárd arány
MAE	mean absolute error	átlagos abszolút hiba
MBS	Markov-blanket set	Markov-határ halmaz
NGS	next-generation sequencing	új generációs szekvenálás
OR	odds ratio	esélyhányados
PCR	polymerase chain reaction	polimeráz láncreakció
PO	posterior odds	posterior esélyhányados
RBF	radial basis function	radiális bázisfüggvény
SNP	single nucleotide polymorphism	egypontos nukleotid polimorfizmus
SVM	support vector machine	szupport vektor gép
TN	true negative	valódi negatív
TNR	true negative rate	valódi negatív arány
TP	true positive	valódi pozitív
TPR	true positive rate	valódi pozitív arány
UTR	untranslated region	nem transzlálódott régió
VQSR	variant quality score recalibration	variánsminőség-kalibráció

DOI:10.14753/SE.2017.1938

1. 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.

1.1. Általános alapfogalmak

A dolgozatban gyakran előkerülnek olyan fogalmak, amelyek az orvos vagy biológus olvasó számára nem feltétlenül ismertek. Ezért a bevezetőben szükségesnek tartom ezek rövid ismertetését, mely során nem a pontos, matematikai definíciók kimondása volt a

DOI:10.14753/SE.2017.1938

célom; hanem inkább intuitív magyarázatokat adni, hogy ezzel is segítsem a munkám és eredményeim megértését. A definíciókat a téma iránt mélyebben érdeklődő olvasó a vonatkozó irodalomban találja.

1.1.1. Valószínűség, valószínűségi változó

Az orvostudomány, de általánosságban minden tudományterület egyik legnagyobb problémája, hogy nem ismerjük a teljes igazságot. Ennek többféle oka is lehet, például az elméleti ismereteink hiánya, a gyakorlati tudatlanságunk (pl. egy adott beteg esetén nem ismerjük az összes klinikai vizsgálat eredményét) vagy az igazság megismerésének és áttekintésének irreálisan nagy anyag-, eszköz-, költség- vagy időigénye. Az ezekből eredő bizonytalanság kifejezésére *valószínűség*i állítások megfogalmazása ad lehetőséget. Ennek során egy állításhoz egy valószínűségi értéket rendelünk, amely az adott *esemény* bekövetkezésének valószínűségét jelenti. Ez a gyakorlatban azt a *hiedelmet* (belief) fejezi ki, hogy a pillanatnyi tudásunk birtokában, a jelenlegi helyzettől megkülönböztethetetlen esetek mekkora hányadában fog bekövetkezni az adott esemény. A valószínűségi érték mindig egy 0 és 1 közötti szám, ahol a 0 valószínűség annak a hiedelemnek felel meg, hogy az esemény biztosan nem fog bekövetkezni, míg az 1 érték azt a hiedelmet fejezi ki, hogy az esemény biztosan bekövetkezik.[1] Az előző megfogalmazásban is láttuk, hogy a valószínűség mértéke mindig függ a jelenlegi helyzettől, azaz a megfigyeléseinktől (tényektől, adatoktól). Mielőtt tények birtokába jutunk, előzetes vagy *a priori* valószínűségről beszélünk, a megfigyeléseink, tények birtokában pedig utólagos, *a posteriori* valószínűségről. A valószínűség jelölésére a P vagy Pr függvényt fogjuk használni, például annak az eseménynek a valószínűségét, hogy egy adott személy akut limfoid leukémiában (ALL) szenved, $Pr(ALL = igaz)$-al vagy egyszerűbben $Pr(ALL)$-el jelölhetjük.

Az események, amelyeknek a bekövetkezési valószínűségét meg szeretnénk állapítani, általában leírhatók ún. valószínűségi vagy *véletlen változók* formájában is. Ez valamilyen mérhető jellemzőt jelent, amelynek az értéke nem állandó, hanem a valószínűség törvényei szerint változik és a véletlentől függ. Valószínűségi változó lehet például az, hogy egy ember szenved-e egy adott betegségben, hány éves az illető a betegség diag-

[1] A valószínűség itt bemutatott értelmezése némileg tágabb, mint a klasszikus statisztikai értelmezés. Ez utóbbi alapján a valószínűség egy objektív, a megfigyelőtől független, a fizikai törvényszerűségekből következő érték, amely egy ismétlődő esemény relatív gyakoriságának határértéke.

DOI:10.14753/SE.2017.1938

nosztizálásakor vagy mi a genotípusa egy adott polimorfizmus esetén. Minden véletlen változóhoz tartozik egy tartomány, amely az általa felvehető lehetséges értékeket tartalmazza. Például egy adott egypontos nukleotid polimorfizmusra vonatkozó genotípus esetén a tartomány tipikusan három elemű: homozigóta vad, heterozigóta vagy homozigóta mutáns.[2] Az a függvény, amely egy adott változó esetén megadja a lehetséges értékek valószínűségi értékeit, az ún. *valószínűségi eloszlás*.

Amennyiben egy esemény valószínűségéről egy másik esemény bekövetkezésének ismerete alapján állítunk valamit, akkor *feltételes valószínűség*ről beszélünk. Például azt az eseményt, hogy egy adott személy ALL-ben szenved, ha tudjuk róla, hogy az rs1004474 polimorfizmus esetén a genotípusa GG értékű, feltételes valószínűséggel tudjuk kifejezni. Jelölése: $Pr(ALL|rs1004474 = GG)$. A feltételes valószínűségek megadhatók feltétel nélküliekkel, például: $Pr(A|B) = Pr(A,B)/Pr(B)$, ahol a $Pr(A,B)$ az A és B esemény együttes előfordulásának valószínűségét jelenti.

Fontos fogalom lesz a későbbiekben a *feltételes függetlenség* is. Akkor mondhatjuk hogy az A esemény feltételesen független a C eseménytől a B ismeretében, ha $Pr(A|B,C) = Pr(A|B)$. Intuitíven megfogalmazva: ha B-t ismerjük, akkor C ismerete már nem ad semmilyen plusz információt, ami befolyásolná a hiedelmünket A bekövetkezésére vonatkozóan.

1.1.2. Osztályozás, osztályozási feladat

Az osztályozás a statisztika (adatbányászat, gépi tanulás) egyik kiemelten fontos módszere. Az osztályozási feladat során az osztályozandó elemeket előre meghatározott csoportokba soroljuk, azaz az elemeket *osztálycímkékkel* látjuk el. Bináris osztályozási feladatok esetén a valóság és az osztályozó kimenete is kétféle lehet, ezeket *negatív* és *pozitív* osztálynak nevezhetjük.

Munkám során többek között genetikai variánsok kivonatolásának problémájával foglalkoztam. Ez a kérdés felfogható egy bináris osztályozási feladatként, ahol egy kivonatoló módszer a vizsgált genomi régióban felmerülő, referencia szekvenciától való eltéréseket osztályozza, és amennyiben a feltételezett variáns jellemzői megfelelnek a kivonatoló

[2] Ha a gén törlődésének vagy többszöröződésének lehetőségét is figyelembe vesszük, akkor több is lehet, pl. hemizigóta, nullizigóta stb. Illetve egyes polimorfizmusok esetén több alternatív allél is előfordulhat, amely szintén megnöveli a változó lehetséges értékeinek számát.

DOI:10.14753/SE.2017.1938

módszer által támasztott kritériumoknak, akkor az adott pozícióban egy variánst hív (a pozitív osztályba sorolja). Ha a bizonyítékok nem elég meggyőzőek, akkor a kivonatoló nem hív variánst (és ezzel a negatív osztályba sorolja). A dolgozatban emellett ún. szupport vektor gépet (support vector machine, SVM) használok a variánskivonatoló módszerek eredményének kombinálására. A következő alfejezetben röviden bemutatom ezt az algoritmust.

1.1.3. Osztályozás Szupport Vektor Gépekkel

A szupport vektor gép [1] egy olyan számítógépes (ún. gépi tanulási) algoritmus, amely tanítóminták alapján megtanul elemekhez címkéket rendelni, azaz osztályozni. A dolgozatban variánsok osztályozására (a valódi és a hamisan hívott variánsok megkülönböztetésére) használunk SVM-eket, de a biológiai problémákban rendkívül széles körben használhatók például génexpressziós profilok, fehérjeszekvenciák vagy DNS szekvenciák osztályozására a nagy pontosságuk, flexibilitásuk és nagydimenziós adatokra való alkalmazhatóságuk miatt [2, 3].

Az SVM általános esetben egy bináris osztályozási feladat megoldására képes: az elemeket pozitív vagy negatív kategóriába sorolja. Ehhez két alapvető koncepciót használ: az elemek lehető legszélesebb margóval történő szétválasztását és az ún. kernel függvényeket [4]. Az osztályozandó, illetve a tanítómintaként felhasznált elemek a tulajdonságaik (pl. a variánsok minőségét leíró jellemzők) alapján egy több dimenziós térben helyezkednek el, ahol a dimenziók száma a tulajdonságok számával egyezik meg. Az SVM ebben a térben egy olyan szeparáló hipersíkot[3] keres, amely a pozitív és a negatív tanítómintákat a lehető legnagyobb margóval választja szét. Ennek során az algoritmus implicit módon kiválasztja azokat a tanítómintákat (ún. szupport vektorokat), amelyek ténylegesen szerepet játszanak a hipersík meghatározásában.

A gyakorlati problémák legnagyobb részében azonban az elemek nem választhatók szét egy lineáris hipersíkkal. Ezt az SVM kétféle módon is képes megoldani: (1) ún. lágy margó (soft margin) használatával, amely megenged meghatározott mértékű hibákat is

[3]Hipersík: Az n-dimenziós euklideszi térben a térnek egy olyan lapos, n-1 dimenziós része, amely a teret két diszjunkt részre osztja. Például két dimenziós síkban a hipersík egy (egy dimenziós) egyenes, három dimenziós térben egy (két dimenziós) sík stb.

DOI:10.14753/SE.2017.1938

az osztályozásban, illetve (2) az elemek nemlineáris transzformációjával[4], ami lehetővé teszi, hogy a lineáris megoldást nem az eredeti térben, hanem az ún. jellemző térben[5] keressük meg, amely az eredeti térbe visszatranszformálva már egy nemlineáris hipersíkot eredményez.

A probléma matematikai formalizációja során valójában az elemeknek nem a pontos pozíciója, hanem az egymáshoz való hasonlósága számít. Ez a kernel függvények használatával lehetővé teszi, hogy az elemeket egy jóval nagyobb dimenziójú térbe transzformáljuk, amely, mint ahogy az előbb láttuk, az elemek nemlineáris szeparációját is megengedi. Az egyik leggyakrabban használt kernel függvény az ún. radiális bázisfüggvény (radial basis function, RBF), amely tulajdonképpen gaussi függvényeket illeszt a szupport vektorok köré.

1.1.4. Szenzitivitás, precizitás, hamis felfedezési arány

Az osztályozók teljesítményének mérése különféle mérőszámok, illetve mutatók kiszámításával lehetséges. Egy bináris osztályozási feladat (pl. variánskivonatolás) esetén egy döntés eredménye alapvetően négy féle lehet:

- Valódi pozitív (true positive, TP): Az adott elem a valóságban pozitív, és az osztályozás eredménye is pozitív (pl. a hívott variáns valóban létezik).

- Valódi negatív (true negative, TN): Az adott elem a valóságban negatív, és az osztályozás eredménye is negatív (pl. a kivonatoló nem hív egy nem létező variánst).

- Hamis pozitív (false positive, FP): Az adott elem a valóságban negatív, de az osztályozás eredménye pozitív (pl. a hívott variáns a valóságban nem létezik). Ezt a fajta tévedést elsőfajú hibának vagy I-es típusú hibának (Type I error) is szokás nevezni.

- Hamis negatív (false negative, FN): Az adott elem a valóságban pozitív, de az osztályozás eredménye negatív (pl. a kivonatoló nem hív egy létező variánst). Ezt a fajta tévedést másodfajú hibának vagy II-es típusú hibának (Type II error) is nevezik.

[4]Nemlineáris transzformáció: Egy olyan matematikai eljárás, amely úgy változtatja meg a változók skáláját, hogy az nem őrzi meg a változók közötti lineáris kapcsolatot. Például nemlineáris az a transzformáció, amely az x változóhoz annak négyzetgyökét vagy reciprokát rendeli.

[5]Jellemző tér: a változók transzformáció utáni magasabb dimenziós tere.

DOI:10.14753/SE.2017.1938

Egy adott, ismert valódi besorolású mintahalmazon végrehajtva az osztályozási feladatot, a döntések eredménye alapján a következő fontosabb teljesítménymutatókat tudjuk kiszámítani:

Szenzitivitás: Az osztályozó a valóságban pozitív elemek hányad részéről állította, hogy pozitív; azaz pl. egy variánskivonatoló a valódi variánsok hányad részét találta meg. Egyéb elnevezései: valódi pozitív arány (true positive rate, TPR), felidézés (recall).

Precizitiás: Az osztályozó által pozitívnak nyilvánított elemek hányad része pozitív a valóságban; azaz pl. a kivonatoló által hívott variánsok hányad része valódi variáns. Egyéb elnevezése: pozitív prediktív érték (positive predictive value, PPV)

Specificitás: Az osztályozó a valóságban negatív elemek hányad részéről állította, hogy negatív. Variánskivonatolók esetén nehezen értelmezhető, mert a nem valódi variánsok száma potenciálisan rendkívül nagy is lehet (indelek esetén potenciálisan végtelen). Egyéb elnevezése: valódi negatív arány (true negative rate, TNR)

Negatív prediktív érték: Az osztályozó által negatívnak nyilvánított elemek hányad része negatív a valóságban.

Hamis felfedezési arány: A precizitás ellentéte; az osztályozó által pozitívnak nyilvánított elemek hányad része negatív a valóságban, azaz hányad részről állítja tévesen, hogy pozitív. Angol elnevezése: false discovery rate (FDR).

		Valóság (feltétel)			
		Pozitív	Negatív		
Osztályozás eredménye (teszt)	Pozitív	Valós pozitív	Hamis pozitív (Elsőfajú hiba)	Precizitás = valós pozitív / teszt szerint pozitív	Hamis felfedezési arány (FDR) = hamis pozitív / teszt szerint pozitív
	Negatív	Hamis negatív (Másodfajú hiba)	Valós negatív	Negatív prediktív érték = valós negatív / teszt szerint negatív	
		Szenzitivitás = Valós pozitív / feltétel szerint pozitív	Specificitás = valós negatív / feltétel szerint negatív		

1. ábra. **A bináris osztályozók döntéseinek lehetséges kimenetelei, illetve a teljesítményük mérésére használható legfontosabb mérőszámok**

A négy lehetséges kimenetelt és az ezekből származtatott legfontosabb mutatókat az 1. ábrán láthatjuk.

1.1.5. Genetikai asszociációs vizsgálat

A betegségek genetikai hátterét tanulmányozó populációgenetikai asszociációs vizsgálatok célja az, hogy olyan variációkat (pl. egypontos nukleotid polimorfizmusokat, single nucleotide polymorphism, SNP) vagy ezeknek egy olyan mintázatát azonosítsuk, amely szisztematikusan eltér egy adott betegségben szenvedő és egészséges emberekben [5]. Ez ugyan elég egyszerűnek hangzik, de valójában fellép egy alapvető probléma: a genom olyan nagy méretű, hogy valódi oki tényezőnek tűnő polimorfizmusok, illetve eltérő mintázatok egyszerűen a véletlennek köszönhetően is jelentkezhetnek. Emiatt a valós és véletlen jelzések megkülönböztetésére rigorózus statisztikai módszereket, asszociációs teszteket végzünk. A munkámnak, így ennek a dolgozatnak sem volt célja a különféle asszociációs tesztek részletekbe menő összehasonlítása. Az érdeklődő olvasó számos összefoglaló közleményt találhat ebben a témában [5–9]. A munkám során a bayesi relevanciaelemzési módszertannal foglalkoztam, amely bayesi statisztikán és ún. valószínűségi hálózatokon alapul. Mivel a bayesi módszerek a megközelítésmódjukban alapvetően eltérnek a klasszikus statisztikai módszerektől, ezért a következő alfejezetekben röviden áttekintem ezeket, illetve rávilágítok a főbb különbségekre.

1.1.6. Frekventista statisztika

Hipotézistesztelés Genetikai asszociációs vizsgálatok esetén a fenotípussal (pl. betegséghajlam) asszociálódó változók (polimorfizmusok, klinikai paraméterek) meghatározására a leggyakrabban használt statisztikai technika a klasszikus *hipotézistesztelés* [5]. Ennek során minden egyes változóra teszteljük azt a hipotézist, hogy az *nem* asszociál a fenotípussal. Ez az ún. *null-hipotézis*, H_0. Amennyiben nincs elegendő bizonyítékunk arra, hogy ez a hipotézis nem igaz, akkor azt nem tudjuk elvetni; azaz nem tudjuk elfogadni az ún. *alternatív hipotézist*, H_1-et, amely szerint az adott változó és a fenotípus között asszociáció áll fenn. Azt a módszert, amivel összegezzük az adatainkban található bizonyítékokat (az ún. *teszt statisztika* kiszámításával) annak érdekében, hogy választani tudjunk a két hipotézis közül, *hipotézistesztelés*nek nevezzük. A teszt statisztika kiszámí-

tásának eredménye egy valószínűség (az ún. *p-érték*), ami a null-hipotézis abszurditásának mértékét jelzi. Más szóval, ha a p-érték kisebb mint egy előre definiált, nulla közeli α érték (ún. szignifikancia szint), az azt jelzi, hogy a null-hipotézis nagyon valószínűtlen, abszurd, így el kell vetnünk, és helyette el kell fogadnunk az alternatív hipotézist. A hipotézistesztelés folyamatát összefoglalva a 2. ábrán láthatjuk.

A leggyakrabban használt asszociációs tesztek például a Pearson-féle χ-négyzet teszt vagy a Fisher-féle egzakt teszt. A logisztikus regressziós modell alkalmazása szintén közkedvelt, ezzel ugyanis már komplexebb összefüggések tesztelésére is lehetőség van, mint például több SNP együttes hatásának vagy interakciójának elemzése, illetve környezeti változók, klinikai paraméterek (nem, életkor stb.) figyelembevétele.

Feltételezés	A null-hipotézis, H_0 igaz, azaz az v változó és a fenotípus között nem áll fenn asszociáció
Ezután	Kiszámítjuk a teszt statisztikát, z_v-t, és azt találjuk, hogy a p-érték (annak valószínűsége, hogy legalább z_v értéket figyelünk meg abban az esetben, ha a null-hipotézis H_0 igaz) kisebb mint α
De	Éppen az előbb figyeltük meg z_v-t
Tehát	A null-hipotézis hamis, és az alternatív hipotézis (majdnem biztosan) igaz, azaz a v változó asszociál a fenotípussal

2. ábra. **A frekventista hipotézistesztelés menete asszociációs vizsgálatokban**

Esélyhányados, konfidencia-intervallum Populációs genetikai asszociációs vizsgálatok esetén a leggyakrabban kiszámított asszociációs mérőszám az ún. esélyhányados (odds ratio, OR), amely azt mutatja meg, hogy mekkora a kimenetel esélyének aránya, ha valaki egy adott tényező hatásának ki van téve ahhoz képest, ha nincs kitéve. Például ha arra a kérdésre keressük a választ, hogy egy adott SNP domináns formája milyen mértékben emeli meg az ALL kialakulásának a kockázatát, akkor ezt az OR kiszámításával válaszolhatjuk meg. Ebben az esetben az OR azt jelenti, hogy mekkora az ALL kialakulásának esélye az SNP alléljának hordozásakor ahhoz képest, mint amekkora a betegség esélye, ha az allél nincs jelen. Az OR értéke mellett általában a becslés konfidencia-intervallumát (confidence interval, CI) is megadjuk (tipikusan 95%). Ez a tartomány azt adja meg, hogy ha a kísérletet végtelen sokszor elvégeznénk, akkor az esetek 95%-ában a kiszámított OR hova esne. Ha az OR szignifikánsan nagyobb mint 1, akkor a tényező növeli a kimenetel

rizikóját; ha kisebb, mint 1, akkor csökkenti.

Többszörös hipotézistesztelési probléma A genetikai asszociációs elemzéseknek egy súlyos problémával kell szembenéznie, amely akkor jelentkezik, ha egyszerre párhuzamosan több hipotézist is tesztelünk. Ez az ún. „többszörös hipotézistesztelési probléma" [5]. A választott statisztikai módszertől függetlenül minél nagyobb számú hipotézisünk van, annál nagyobb annak valószínűsége, hogy véletlenül extrém teszt statisztika-értékeket figyelünk meg, így egyre valószínűbb, hogy tévesen el fogjuk utasítani a null-hipotézist (és ezzel hamis pozitív kijelentést teszünk, azaz elsőfajú hibát követünk el). Sokféle megközelítés létezik ennek a problémának a kezelésére, amelyek abban különböznek, hogy milyen hibát próbálnak meg kontrollálni és mennyire konzervatívak.

A legkonzervatívabbnak tartott módszer az ún. *Bonferroni-eljárás*, amely a *családi-szintű hibát* (familywise error rate, FWER) kontrollálja. Ez annak a valószínűsége, hogy az összes vizsgált, nem valódi asszociáció közül legalább egyről tévesen azt állítjuk, hogy fennáll. A Bonferroni módszer során egyszerűen elosztjuk α-t (a megkívánt FWER szignifikanciaszintet) a hipotézisek számával. Például annak biztosítására, hogy 1000 statisztikai teszt elvégzése esetén is a családi-szintű hiba aránya kisebb legyen mint $0{,}05$, az elfogadási küszöböt 5×10^{-5}-re kell állítanunk. Ugyanakkor a Bonferroni-korrekció az SNP-k kapcsoltsága miatt a legtöbb vizsgálatban túlságosan konzervatív; használatával sok valódi asszociációt figyelmen kívül hagyunk (azaz hamis negatív kijelentést teszünk; másodfajú hibát követünk el). Ebben az esetben az egyik leggyakrabban használt módszer a *hamis felfedezési arány* (false discovery rate, FDR) kontrollálása. Az FDR, mint ahogyan az előző alfejezetben láttuk, a nem valódi asszociációk várható aránya azok között, amelyekről azt állítjuk, hogy fennállnak. Más szóval, ha a célunk az, hogy előálljunk hipotézisek egy olyan halmazával, amelynek a legnagyobb része igaz, akkor az FDR-t érdemes kontroll alatt tartani. Benjamini és Hochberg javasolt [10, 11] erre egy felfelé lépegető eljárást: az asszociációkat rendezzük sorba a p-értékük szerint, majd a legkisebbtől indulva viszonyítsuk azokat egy folyamatosan növekvő küszöbértékhez (a k-dik p-értéket viszonyítsuk $\frac{k}{m}\alpha$-hoz, ahol m a vizsgált hipotézisek száma), és utasítsuk el az összes nullhipotézist (azaz fogadjuk el az alternatív hipotézist) a legnagyobb olyan k-ig, amelyre a p-érték még kisebb, mint az adott küszöbérték. Ez egy kevésbé konzervatív

korrekciós eljárást eredményez, ami jobban illeszkedik a genetikai asszociációs tesztek felderítő jellegéhez.

1.1.7. Bayesi statisztika

A bayesi statisztikai módszerek az utóbbi időben számos tudományterületen rendkívüli népszerűségnek örvendenek [12], beleértve a genetikát [13–15] és a genetikai asszociációs vizsgálatokat is [8, 16–22]. A módszertan alapja az 1700-as évek közepén, Thomas Bayes tiszteletes és matematikus által megfogalmazott Bayes-tétel, amely az ok és okozat (vagy előzmény és következmény) felcserélhetőségét mondja ki (a tétel következik a feltételes valószínűség definíciójából):

$$Pr(B|A) = \frac{Pr(A|B) * Pr(B)}{Pr(A)} \propto Pr(A|B) * Pr(B), \tag{1}$$

ahol $\propto$ az arányosságot jelenti.

Orvosi példánál maradva, ha az A esemény a láz, a B esemény pedig az, hogy a beteg influenzás-e, akkor a $Pr(Láz|Influenza)$ feltételes valószínűség és a $Pr(Influenza)$, illetve $Pr(Láz)$ *a priori* valószínűségek segítségével meg tudjuk válaszolni azt a fordított ok-okozati relációban álló, diagnosztikai jellegű kérdést, hogy láz esetén mekkora az influenza valószínűsége, azaz mekkora a $Pr(Influenza|Láz)$ *a posteriori* valószínűség. Ez elsőre nem feltétlenül tűnik nagy jelentőségű eredménynek, hiszen a keresett valószínűség kiszámításához három másik valószínűségi értéket kell meghatároznunk. Azonban ezek megadása bizonyos esetekben (pl. tipikusan a diagnosztikai problémákban) jóval könnyebb: a $Pr(Láz|Influenza)$ feltételes valószínűség a betegség lefolyásából következik, az influenza patomechanizmusától függő állandó jellegű mennyiség, amely tipikusan nem változik és jól becsülhető (pl. az influenzás betegek hányad részénél tapasztalunk lázat); a $Pr(Influenza)$ valószínűség az adott időben influenzában szenvedő betegek arányát jelenti a teljes népességhez képest, szintén jól becsülhető és jól kezelhető például járványok kitörése esetén is, amikor ez az érték megugrik; a $Pr(Láz)$ valószínűség már kissé nehezebben kezelhető (az adott időben lázas emberek aránya a teljes populációhoz képest), azonban valójában nem szükséges meghatározni, mert egy normalizációs lépéssel kiküszöbölhető (emiatt szerepel a fenti képletben az arányosság).

A fenti példából az is látszik, hogy általános értelemben a Bayes-tétel a tudományos gondolkodás egyszerűsített modellje lehet [23, 24]. Azt mondhatjuk ugyanis, hogy rendelkezünk valamiféle tudással a világról (pl. influenza *a priori* valószínűsége), majd tapasztalatokat, adatokat gyűjtünk (pl. megmérjük a páciens testhőmérsékletét), ezt egybevetjük és súlyozzuk a kezdeti ismereteinkkel, és ezáltal a tudás egy magasabb szintjére jutunk el (pl. az influenza *a posteriori* valószínűsége).

A bayesi gondolkodást érintő leggyakoribb kritika az, hogy a priorok megfogalmazása gyakran szubjektív, a kísérletező hiedelmeitől, előzetes tudásától függ és emiatt a posteriorban keverednek objektív és szubjektív elemek [23]. Ezzel szemben a bayesi gondolkodók a klasszikus statisztikában azt kifogásolják, hogy az csak tömegjelenségekre, illetve elméletileg végtelen sokszor ismételhető kísérletekre használható, így a relatív gyakoriságokon alapuló valószínűségeket csak ilyen típusú mintákon lehet használni. A szubjektív priorok problémáját részben lehet kezelni az ún. neminformatív priorokkal, amelyek igyekeznek a lehető legjobban kifejezni az ismeretek hiányát, de ezek megfogalmazása (konstruálása) sokszor nehéz lehet, és a semlegességük sokszor vitára adhat okot [23].

A bayesi és frekventista statisztika legfontosabb különbségei A bayesi és frekventista statisztikai következtetések legfontosabb különbségei a kiinduló feltételezéseken alapulnak. Tekintsük az előző rész példáját: hogyan határozza meg a két statisztikai módszer az adott időpontban influenzás betegek arányát a teljes népességhez képest egy adott mintavétel alapján. A klasszikus statisztika felfogása szerint ez az arány egy jól meghatározott, rögzített, valós fizikai mennyiség, amely azonban nem ismert, de mintavételi kísérletekkel tetszőleges pontossággal meghatározható. Ezzel szemben a bayesi statisztikai felfogás szerint ez egy valószínűségi változó, azaz a lehetséges értékeinek egy valószínűségi eloszlása van, amelyet sok más tényező befolyásolhat (pl. van-e éppen járvány).

A két megközelítés másik nagy különbsége a minta értelmezésében van. A klasszikus statisztika ugyanúgy az egyetlen mintavétel alapján hozza meg a döntéseit, de ezt úgy teszi, hogy közben feltételezi a kísérlet ismételhetőségét. Így például ha egy statisztikai tesztet végzünk annak a kérdésnek az eldöntésére, hogy az influenzások adott időpontbeli aránya szignifikánsan nagyobb-e, mint $0{,}1$, akkor a statisztikai teszt eredményéül kapott

p-értéket úgy értelmezhetjük, hogy ha végtelenszer elvégezném a mintavételt és minden esetben kiszámítanám a teszt statisztikát, akkor mekkora valószínűséggel (az esetek hányad részében) kapnék a jelenlegi mintapopuláció alapján kiszámítottnál nagyobb teszt statisztika értéket. Ezzel szemben a bayesi felfogás szerint az ismételt mintavételre nincs szükség, a következtetéseinket (az influenza *a posteriori* eloszlását) az egyetlen mintapopuláció alapján határozzuk meg.

Végül különbség van a két megközelítés végeredményében is: a klasszikus esetben a korábban bemutatott pontbecslés és konfidencia-intervallum (ha a mintavételt végtelenszer ismételnénk, a keresett arány az esetek 95%-ában melyik tartományba esne); míg a bayesi esetben a végeredmény valójában a valószínűségi változó posterior eloszlása. Ez utóbbi alapján további eredmények is képezhetők, például a *hihetőségi tartomány* (melyik az a legszűkebb értéktartomány, amely a valószínűségi változó értékét 95% valószínűséggel tartalmazza) és a pontbecslés (melyik az az érték, amely a legkisebb hibával közelíti a valószínűségi változó eloszlását; azaz a lehető legjobban jellemzi a posterior eloszlást).

Hipotézisvizsgálat a bayesi megközelítésben A teljesség kedvéért röviden kitérünk a bayesi hipotézisvizsgálat módszerére is, de a dolgozat során használt bayesi megközelítés nem ezt a technikát fogja követni. A klasszikus statisztikához hasonlóan itt is van null-hipotézis és alternatív hipotézis, de szemléletbeli különbség van a két módszer között. A klasszikus esetben ugyanis a null-hipotézisnek kitüntetett szerepe van, és a fő kérdés, hogy a bizonyítékok ismeretében el tudjuk-e vetni vagy sem. A bayesi esetben a két hipotézis teljes mértékben egyenrangú, és arra a kérdésre keressük a választ, hogy a bizonyítékok (adatok, jelölése: D) tükrében melyiknek nagyobb a valószínűsége, azaz az ún. posterior esélyhányadost (posterior odds, PO) számítjuk ki:

$$PO = \frac{Pr(D|H_1)Pr(H_1)}{Pr(D|H_0)Pr(H_0)} \tag{2}$$

1.1.8. Túlélés-elemzés

Túlélés-elemzés alatt olyan statisztikai módszereket értünk, amelyek *túlélési adatok* elemzésére használhatók, azaz ahol a kimeneti (függő) változó egy adott típusú esemény bekövetkezésének ideje. Ezt általánosságban *túlélési idő*nek nevezzük, de a gyakorlatban

ezt az időt definiálhatjuk például rákos megbetegedések elemzésekor a teljes remissziótól a relapszusig eltelt időként, a diagnózistól a halálig eltelt időként, vagy műszaki példa esetén lehet ez az idő egy eszköz üzembe helyezésétől a meghibásodásáig eltelt idő is. Amennyiben minden minta esetén ismert lenne a túlélés ideje, akkor a klasszikus statisztika sok módszere bevethető lenne az adatok elemzésére. Azonban az elemzések többségében igaz, hogy nem minden egyén életében következik be esemény vagy az egyének kiesnek az elemző látóköréből (pl. költözés vagy a vizsgálttól eltérő halálok miatt), így az ő esetükben a túlélés ideje ismeretlen (ezek ún. cenzorált minták). Továbbá a túlélési adatok ritkán normális eloszlásúak, ellenben gyakran „eltoltak" és tipikusan sok korai és kevés késői eseményt tartalmaznak. Mindezek miatt a túlélési adatok elemzése rendszerint egyedi statisztikai módszerek használatát teszi szükségessé [25–28].

A közönséges regressziós modellekkel szemben a túlélés-elemzési módszerek képesek a cenzorált adatok kezelésre is. A módszerek által becsült két legfontosabb függvény a *túlélési* és *hazárd* függvény. A túlélési függvény azt adja meg, hogy egy adott időpontban mekkora az esemény túlélésének valószínűsége (azaz hogy az esemény nem következik be). A hazárd függvény egy adott időpontban annak a valószínűséget adja meg, hogy az esemény bekövetkezik, feltéve, hogy eddig még nem következett be. A túlélés-elemzés során általában a legfontosabb kérdés az, hogy egy adott faktor (pl. genetikai, környezeti változó) hogyan befolyásolja a túlélési időt.

Az egyik leggyakrabban használt nemparametrikus[6] teszt a Kaplan-Meier módszer [29], amely egy változó értéke alapján képzett csoportok különbségének összehasonlítására is használható (a χ-négyzet teszthez hasonló módon; ez az ún. log-rank teszt) [25]. Szintén nagyon gyakran használt módszer a Cox-regresszió [30], amely a hazárd függvényt (illetve annak logaritmusát) közelíti a független változók lineáris modelljével. A logisztikus regresszióhoz hasonlóan ebben a modellben is elemezhetünk kovariánsokat, illetve vizsgálhatjuk a prediktor változók interakcióit is, így bonyolultabb, többváltozós összefüggések tesztelésére is használható [26].

[6]Nemparametrikus teszt: Olyan statisztikai teszt, amely nem tételez fel semmilyen eloszlást az adatokban.

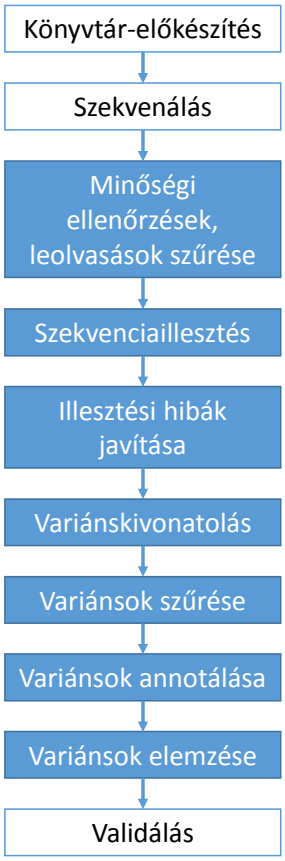
DOI:10.14753/SE.2017.1938

1.2. Genetikai variánsok meghatározása új generációs szekvenálással

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 [31]. 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 jelenlegi legnagyobb kapacitású készülék (Illumina HiSeq X) egyetlen futása során 1800 Gb méretű adat keletkezik, ami a vizsgált szekvencia 6 milliárd rövid (2×150 bp) leolvasását jelenti. Egy teljes genom szekvenálás során egyénenként átlagosan kb. 5 millió variánst (SNP-t és rövid inzerciót vagy törlődést, röviden: indelt) szoktak azonosítani, amelyből 144000 variáns új, azaz nem fordul elő a publikus adatbázisokban [32]. A teljes exom szekvenálások során a humán genom körülbelül 1%-nyi teljes kódoló szekvenciáját határozzák meg, amely során egyénenként átlagosan kb. 12000 variánst azonosítanak, amelynek 10%-a új [33, 34]. Ennek a hatalmas adatmennyiségnek az elemzése és értelmezése jelentős kihívásokat támaszt a kutatók számára. Az NGS projektek szűk keresztmetszete emiatt nem maga a DNS szekvenálása, hanem az adatmenedzsment és a kísérleti adatok szofisztikált elemzési munkafolyamatainak pontos kialakítása [35, 36], amely a jövőben várhatóan egyre nagyobb kihívást fog jelenteni [37].

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 [38–41] és (2) jelentős eltérés van a széles körben használt variánskivonatoló munkafolyamatok eredményei (azaz a hívott variánsok) között, még abban az esetben is, ha ugyanazokra a mérési adatokra alkalmazzák azokat [39, 40, 42, 43]. Ahhoz, hogy ezeket az eredményeket jobban megértsük, röviden áttekintjük egy tipikus elemzési munkafolyamat lépéseit (lásd 3. ábra). A továbbiakban ezeket a lépéseket rész-

letezzük (a teljesség kedvéért a munkafolyamat későbbi – az elemzésre kész variánsok előállításán túlmutató – elemeit is röviden bemutatjuk).



3. ábra. **Egy tipikus teljes genom vagy teljes exom szekvenálási projekt elemzési munkafolyamatának lépései.** A laboratóriumi előkészítés után a mintákat megszekvenálják, ami nagymennyiségű, rövid szekvencialeolvasásokat eredményez. A kísérlet minőségének ellenőrzése és a leolvasások minőségi szűrése után a szekvenciákat felillesztik a referenciagenomra, majd opcionálisan további minőségi javításokat végeznek. Az illesztések alapján megtörténik a variánsok hívása, majd minőségi szűrése. Ezután különböző adatbázisok és szoftverek felhasználásával a variánsokat funkcionálisan annotálják, végül elemzik (és szükség esetén tipikusan Sanger szekvenálással validálják). A bioinformatikai feladatok kék háttérrel vannak jelezve.

Szekvenálás Mivel a jelenlegi technológiák által megfelelő minőséggel leolvasható szekvenciák hossza viszonylag rövid, a DNS-t a könyvtár-előkészítés során fel kell darabolni, majd a szekvenálási platformtól függően a DNS darabokat PCR reakciókkal fel kell sokszorozni. Ezt követi a tényleges szekvenálás, amely során a DNS darabok szekvenciájának meghatározására kerül sor (leolvasás). A szekvenáló gépek minden egyes leolvasott bázishoz egy minőség pontszámot (ún. bázisminőségi mutatót) rendelnek, amely a ké-

sőbbi adatelemzési lépések esetén hasznos információként szolgál az adott bázis értékének megbízhatósága szempontjából. A bázisminőséget az ún. Phred-pontszámmal szokás megadni, amely a bázishiba valószínűségét fejezi ki (ha a hiba valószínűségét P-vel jelöljük, akkor $Q = -10log_{10}P$, lásd 1. táblázat).

1. táblázat. **Phred-pontszámok értelmezése**

Phred-pontszám	A hibás bázishívás aránya	A bázishívás pontosságának valószínűsége
10	1 a 10-ből	90%
20	1 a 100-ból	99%
30	1 az 1000-ből	99,9%
40	1 a 10 000-ből	99,99%
50	1 a 100 000-ből	99,999%
60	1 az 1 000 000-ból	99,9999%

Leolvasások szűrése A szekvenciák meghatározása után az első lépés a nyers leolvasások minőségének meghatározása és javítása. A szekvenáló gépek által kiadott leolvasások ugyanis többféle hibát tartalmazhatnak, például bázishívási (szubsztitúciós) vagy indel hibákat (pl. a homopolimer szakaszok hosszának tévesztése tipikusan Roche/454 és IonTorrent platformokon), alacsony minőségű leolvasásokat, kevert (ún. kiméra) szekvenciákat vagy adapter szekvenciák kontaminációját [44]. Mivel az ilyen típusú hibák kezelésére és kiszűrésére a munkafolyamat későbbi lépéseit megvalósító programok nincsenek teljes körűen felkészítve, ezért a hibás biológiai konklúziók elkerülése érdekében fontos, hogy kiszűrjük a felismerhető hibákat [36]. Ennek első lépése többek között a bázisminőségi pontszámok, a GC tartalom és a leolvasások hossz-eloszlásának ábrázolásából, illetve a feldúsult szekvenciarészletek és duplikált szekvenciák azonosításából áll [45]. Második lépésként pedig az azonosított hibák kiszűrése következik a szekvenciák nem megfelelő szakaszainak levágásával és a hibás vagy nem megfelelő hosszúságú szekvenciák eldobásával. Ezekre a feladatokra például a FASTQC [45], NGSQC [44], FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/index.html Hozzáférés: 2015.07.14.) és a PRINSEQ [46] szoftvereket használhatjuk.

DOI:10.14753/SE.2017.1938

A leolvasások felillesztése a referencia szekvenciára A minőségi szűrések elvégzése után a leolvasásokat fel kell illeszteni a humán referencia szekvenciára. Az illesztés során egy adott pozíciót lefedő leolvasások számát *leolvasási mélységnek* vagy egyszerűen *lefedettségnek* nevezzük. Az utóbbi időben számos szoftver született az illesztési feladat megoldására [47], amelyek általában valamilyen kiegészítő adatszerkezetek (pl. indexek) felhasználásával oldják meg a rendkívül nagy mennyiségű szekvencia gyors illesztésének problémáját. Ezek alapján az illesztőprogramok két nagy csoportját különböztethetjük meg: (1) hash-tábla alapú illetve (2) szuffix fákon alapuló algoritmusok.

A hash-tábla alapú programok a BLAST [48] megoldását követik, amely a leolvasásokat rövid szakaszokra (k-merekre, azaz k hosszúságú szekvenciadarabokra) bontja, majd egy hash-tábla alapján megkeresi, hogy ezek hol találhatók a genomban. Ezt követően a találatok kiterjesztésével azonosítja azt a pozíciót, ahonnan a leolvasás nagy valószínűséggel származik, majd az optimális megoldást adó Smith-Waterman lokális szekvenciaillesztési algoritmussal meghatározza a végleges illeszkedést. A jelenleg használatos hash-tábla alapú programok a BLAST stratégiáját fejlesztették tovább valamilyen módon. Ilyen például a MAQ [49], amely a k-mereket nem egybefüggő szakaszokként definiálja, ami nagyobb szenzitivitást eredményez és a szekvenálási hibák kezelését is lehetővé teszi. A MAQ azonban nem képes „hézagok” beillesztésére (gapped alignment), így a referencia szekvenciához képest indeleket tartalmazó leolvasások felillesztésére sem. A szintén hash-tábla alapú MOSAIK [50] azonban már megoldja ezt a problémát.

Az illesztőprogramok másik nagy csoportja ún. szuffix fákat használ a leolvasások pozíciójának azonosítására. Ez egy olyan adatszerkezet, amely egy karaktersorozat összes utótagjának hatékony tárolására és ebből eredően karaktersorozatok rendkívül gyors keresésére használható. A szuffix fákon alapuló algoritmusok általában $10-20$-szor gyorsabbak a hash-tábla alapú programoknál, miközben a pontosságuk hasonló mértékű [51]. A leggyakrabban használt ilyen illesztőprogramok a BWA [51] és a Bowtie 2 [52], amelyek egyaránt képesek hézagos illesztésre, a bázishibák és a paired-end[7] leolvasások kezelésére. Emellett az illesztés minőségét leíró mutatókat állítanak elő, amely a későbbi variánskivonatolási és szűrési lépések során fontos információként szolgál a valódi és hamis

[7]Paired-end szekvenálás: a genomban egymástól meghatározott átlagos távolságra lévő szekvenciák szekvenálása, amely jóval nagyobb pontosságú illesztést tesz lehetővé az ismétlődő és alacsony komplexitású szakaszokra való könnyebb illeszthetőség miatt.

variánsok megkülönböztetéséhez.

Az illesztések hibáinak javítása Az illesztőprogramok elsődleges eredményei különböző típusú hibákat tartalmazhatnak. Például gyakran előfordul, hogy azok a leolvasások, amelyek végei indelek környékére esnek, hamis szubsztitúciós eltéréseket mutatnak a referencia szekvenciához képest. Ez a lokális szekvenciaillesztés algoritmusának működéséből fakad, ugyanis ebben az esetben valójában egy hézagot kellene nyitni, de ennek nagyobb büntetése van, mint a szubsztitúciós hibáknak. Ezekben a pozíciókban a variánskivonatoló programok tévesen SNP-ket hívhatnak, így célszerű az ilyen típusú hibákat kijavítani. A Genome Analysis Toolkit (GATK) programcsomag [53, 54] egyik eszköze az indelek környékére eső leolvasások újraillesztésével ezt a hibát próbálja kiküszöbölni. Egy másik gyakori probléma, hogy a szekvenáló platformok rosszul becsülik meg a bázisok minőségét. Ez szintén a későbbi variánskivonatolás hibájához vezethet, a variánsok hívása ugyanis nagymértékben a bázisminőségi mutatók pontosságán alapul. A GATK egy másik eszköze, a bázisminőségek újrakalibrálása (base quality score recalibration) publikus variánsadatbázisok felhasználásával empirikus hibamodelleket állít elő az illesztett leolvasások alapján, majd a hibamodellek segítségével pontosítja a bázisok minőségi pontszámait. Számos elemzési vizsgálatban azt találták, hogy mind az indelek környéki újraillesztés, mind pedig a bázisminőségek újrakalibrálása szignifikánsan javította a variánskivonatolás pontosságát [43, 55], bár az eredmények némileg ellentmondásosak, ugyanis más kutatócsoportok eredményei nem ezt igazolták [56].

Variánskivonatolás A variánsok megkeresése az elemzés legfontosabb lépése (amelyre a dolgozatban a variánshívás vagy variánskivonatolás elnevezéseket is használni fogjuk) [36]. A megfelelő minőségű variánshívás egyik legfontosabb tényezője a leolvasási mélység, ugyanis megfelelő lefedettség nélkül a valódi eltéréseket és a szekvenálási hibákat nem lehet megkülönböztetni [41]. A variánskivonatolási módszerek a hívott variánsok típusa alapján négy nagy csoportba oszthatók: (1) csíravonali variánshívók (SNP-k és rövid indelek hívására), (2) szomatikus variánshívók, (3) kópiaszám-változás detektáló programok és (4) strukturális variánsok (inverziók, transzlokációk, nagy indelek) meghatározására szolgáló módszerek. A továbbiakban röviden bemutatunk néhány gyakran használt csíravonali variánshívó programot.

SAMtools Az eredetileg Heng Li által fejlesztett, majd mások által továbbfejlesztett SAMtools [57] az egyik leggyakrabban használt NGS programcsomag, amely csíravonali variánsok kivonatolására is használható. A minták genotípusának megállapítása bayesi statisztikai módszereken alapul, amelyet más kivonatoló programok is átvettek és továbbfejlesztettek. Az algoritmus a referencia genom minden egyes pozícióján egyesével végiglépked (ahol van megfelelő mélységben illesztett szekvencia), és az adott pozícióban a leolvasott bázisok értékének és bázisminőségének figyelembevételével meghatározza a legnagyobb *a posteriori* valószínűségű genotípust. A nem homozigóta vad genotípus azt eredményezi, hogy a program az adott pozícióban egy variánst fog lejelenteni.

GATK UnifiedGenotyper A GATK [53] egy komplex programcsomag, amely NGS variánskivonatolásra, illetve ezzel összefüggő feladatok elvégzésére használható. A Broad Institute-ban fejlesztik, és rendkívül széleskörűen használják nagy genomi projektekben is (pl. 1000Genome Project, The Cancer Genome Atlas). A GATK két kivonatolót tartalmaz, amelyek közül a UnifiedGenotyper a régebbi, és jelenleg már nem fejlesztik tovább. Az algoritmus a SAMtools módszerének továbbfejlesztésén alapul, amely lehetővé teszi több minta együttes kivonatolását és a multiallélikus variánshívást is (a SAMtools újabb verziója is támogatja).

GATK HaplotypeCaller A GATK HaplotypeCaller algoritmusa szakított a genomi pozíciók egyesével történő bejárásával, és – szemben a korábban említett módszerekkel – az illesztéseket csak támpontként használja a variánskivonatolás során. Az algoritmus első lépésében meghatározza az ún. aktív régiókat, amelyek lényeges, a szekvenálási zajt meghaladó mértékű eltéréseket tartalmaznak a referencia szekvenciához képest. Ezután az aktív régióba eső leolvasásokat összeilleszti, és ennek segítségével meghatározza a régióba eső összes lehetséges haplotípust. A haplotípusokat az eredeti referencia szekvenciához illesztve a program megkapja a lehetséges variánsok tényleges genomi pozícióját. Ezután az algoritmus a leolvasásoknak a lehetséges haplotípusokra való visszaillesztésével a bázisminőségi pontszámok alapján meghatározza annak valószínűségét, hogy az adott leolvasást figyeltük meg, ha az adott haplotípus a valódi (ez az ún. likelihood). Végül a Bayes-tétel segítségével

kiszámítja minden egyes minta esetén a két legnagyobb *a posteriori* valószínűségű haplotípust, amely egyben a legvalószínűbb genotípus meghatározását is jelenti.

FreeBayes A FreeBayes bayesi statisztikai módszerek alapján SNP-k, indelek, több nukleotidot érintő polimorfizmusok (multi nucleotide polymorphisms) és komplex átrendeződések detektálására használható program [58]. A variánsok hívása a HaplotypeCallerhez hasonlóan haplotípusok rekonstruálásával történik.

Annotációs mutatók a variánsok minőségének jellemzésére A variánskivonatolás során az egyes módszerek számos mutatót, ún. annotációkat generálnak, amelyek a variánsok jóságát/valódiságát jellemzik a szekvenálási adatok alapján. A következőkben bemutatunk néhány fontosabb annotációs mutatót, illetve segítséget adunk az értelmezésükhöz (lásd pl. [39]).

Variáns minőség Minden kivonatoló módszer előállít egy központi jelentőségű annotációs mutatót, amely annak a valószínűségét adja meg Phred-pontszámmal kifejezve (lásd 1. táblázat), hogy az adott variáns legalább egy minta esetén nem homozigóta vad genotípusú (azaz valójában egy variábilis pozíció). Minél nagyobb ez az érték, annál biztosabbak lehetünk abban, hogy az adott variáns valójában létezik.

Szálirány-eltérés (strand bias) A szálirány-eltérés azt jelenti, hogy az alternatív allél és a referencia allél nem egyforma arányban fordul elő a pozitív és a negatív irányú szálakon. Ez az illesztés problémáját utalhat, és megkérdőjelezheti a variáns valódiságát, ugyanis a szekvenálás során elvileg megközelítőleg egyenlő arányban olvassa le a szekvenáló gép a szekvenciákat a pozitív és a negatív irányból. Eltérő lehet, hogy az egyes variánskivonatoló módszerek milyen tesztet használnak ennek a problémának a jelzésére, de a leggyakoribb a Fisher-féle egzakt teszt vagy a Wilcoxon-teszt használata.

Illesztési minőség eltérés Az illesztőprogramok minden leolvasáshoz megadnak egy – az illesztés minőségére utaló pontszámot. Amennyiben különbség van abban a tekintetben, hogy az alternatív és a referencia allélok inkább az alacsonyabb vagy magasabb illesztési pontszámmal rendelkező leolvasásokon fordulnak elő, az szintén

az illesztés problémájára hívhatja fel a figyelmet. Ezt általában Wilcoxon-teszttel számítják ki az egyes módszerek.

Pozíció-eltérés Akkor beszélünk pozíció-eltérésről, ha ahelyett, hogy a variáns a rá illeszkedő leolvasásokban egyenletesen elszórva fordulna elő, konzisztensen a leolvasások elején vagy végén található. Ezt általában szintén Wilcoxon-teszttel számítják ki az egyes variánskivonatoló módszerek.

Haplotípus-pontszám A GATK által kiszámított mutató, amely azt jelzi, hogy egy adott pozícióban kettőnél több haplotípus jelenik meg, ami illesztési problémákra utalhat. Minél nagyobb a mutató értéke, annál valószínűbb, hogy az adott variáns hamis.

Variánsok szűrése Általánosságban elmondható, hogy a variánskivonatolási módszerek – a precizitást másodlagos szempontnak tekintve – nagyfokú szenzitivitásra törekszenek, azaz „agresszíven" hívnak variánsokat, és a felhasználóra bízzák, hogy a variánsok minőségét jellemző annotációs mutatók segítségével az eredményekből válogassa ki a feltehetően valódi a variánsokat. A szűrések célja tehát a variánskivonatolási eredmények precizitásának növelése lehetőleg úgy, hogy a szenzitivitás mindeközben ne csökkenjen az elfogadhatónál nagyobb mértékben. Nem határozható meg azonban ezeknek a teljesítménymutatóknak egy – minden szekvenálási projektben egységesen elfogadható szintje, ugyanis a különböző célú projektekben eltérő lehet a hamis negatív és hamis pozitív hibák megítélése. Klinikai diagnosztikai esetekben (tipikusan célzott génpanelek, vagy egyes gének, pl. BRCA1/BRCA2 szekvenálása esetén) a hamis negatív hibáknak általában nagyobb jelentőséget tulajdonítanak. Ugyanis ha egy valódi oki variánst tévesen kiszűrünk, akkor a páciensről tévesen azt állíthatjuk, hogy nem hordoz veszélyes mutációt, ami akár a kezelés módját és kimenetelét is befolyásolhatja. A hamis pozitív találatok azonban a diagnosztikai esetben nem jelentenek ugyanekkora jelentőségű problémát, ugyanis komplementer mérési módszerekkel (pl. Sanger szekvenálással) az okozatinak tűnő variánsokat validálni lehet. Ezzel szemben egy kutatási projektben (pl. teljes genom szekvenálás esetén) a hamis pozitív variánsok nagyobb aránya már nagyobb problémát jelenthet, ugyanis az összes találat validálása már nem lenne költséghatékony, viszont az oki variánsokkal esetleg kapcsoltsági egyensúlytalanságban álló variánsok detektálá-

sa miatt nem jelent feltétlenül nagy problémát az oki variáns téves kiszűrése. Mindezek miatt a szekvenálási projektekről elmondható, hogy a variánsok szűrésének célja mindig az aktuális, alkalmazás-specifikus egyensúly megtalálása a szenzitivitás és precizitás elfogadható szintje között. Ennek alapján egy szűrő módszer nagyon hasznos tulajdonsága, ha a szűrést közvetlenül az elvárt precizitás értéke alapján tudjuk elvégezni. Az ilyen módszereket *precizitás alapú* szűrőnek nevezzük.

A variánsok manuális szűrése A variánsok szűrésének egyik lehetséges, gyakran használt módszere az ún. *manuális szűrők* (hard filtering) alkalmazása. Ez úgy történik, hogy (1) ki kell választani azokat az annotációs mutatókat, amelyek jól jellemzik a variánsok minőségét, majd (2) meg kell határozni azokat a küszöbértékeket, amelyek a lehető legjobban elválasztják a valódi variánsokat a hamisaktól. Ezt követően minden egyes variánsra ellenőrizni kell, hogy az megfelel-e a megadott feltételeknek. Ha nem, akkor a variánst el kell dobni. A manuális szűrők használatát több tényező is megnehezíti, többek között az annotációk komplex összefüggésrendszere [39, 43], az adott kísérleti beállításstól való függése, illetve a nehéz értelmezhetősége [38]. Mindezek miatt gyakran nem egyértelmű, hogy pontosan mi a megfelelő szűrőbeállítás. Léteznek ugyan általános javaslatok [53], de az elfogadható eredményt adó küszöbértékek megtalálása sok manuális kísérletezést és tesztelést igényel. A problémát tovább súlyosbítja, hogy a legtöbb annotációs mutató értéke függ az aktuális leolvasási mélységtől, így egy szűrőbeállítás, amely alacsony lefedettség esetén jól működik, nagyobb leolvasási mélység esetén már nem feltétlenül ad optimális megoldást. Ez az NGS vizsgálatokban gyakran tapasztalt nem egyenletes lefedettség miatt [59] még inkább megnehezíti a manuális szűrők használatát. Végül szintén hátrányos tulajdonságuk, hogy nem tudjuk megbecsülni az eredményül kapott variánslista precizitását.

A variánsok szűrése a variánsminőség újrakalibrálásával A variánskivonatolási eredmények precizitásának javítására, illetve a variánsok szűrésére használható a GATK által fejlesztett variánsminőség-kalibrációs (variant quality score recalibration, VQSR) algoritmus is. Ez a módszer egy gépi tanulási eljáráson alapul, és a felhasználó által megadott annotációk értéke, illetve nagy megbízhatóságú referencia variánsok felhasználásával megpróbálja megkülönböztetni a valódi és a hamis variánsokat. Ennek során a VQSR

megbecsüli a variánsok valódiságának valószínűségét. Ez egyrészt a variánsok szűrésére is használható, másrészt a valószínűségek alapján a módszer meg tudja jósolni egy adott variánshalmaz precizitását, így képes precizitás alapú szűrésre is. A módszer hátránya, hogy csak nagy adatmennyiségek esetén használható (legalább 30 teljes exom, vagy teljes genomok szekvenálása esetén) [53], illetve csak olyan organizmusokra, amelyekhez rendelkezésünkre állnak nagy megbízhatóságú referencia variáns készletek (pl. humán).

Variánskivonatolások kombinálása A variánskivonatolás szenzitivitásának növelésére több kutatócsoport is felvetette a különböző kivonatoló módszerek eredményének kombinációját [36, 39, 42, 60]. Ez azon a megfigyelésen alapul, hogy az egyes módszerek részben eltérő eredményeket adnak, és jellemzően minden kivonatoló talál olyan valódi variánsokat, amelyeket más módszer nem [39, 40, 42, 43]. Természetesen a különböző kivonatolók eredményének egyszerű uniója alacsonyabb precizitáshoz vezethet, így a kombináció során komplexebb megoldásokra van szükség. Cantarel és mtsai kifejlesztették a BAYSIC programot, amely nagy megbízhatóságú referencia variánsok felhasználása nélkül, egy bayesi statisztikai módszer segítségével képes a variánshalmazok kombinációjára, amely által az egyedi kivonatolóknál jobb teljesítmény érhető el [61]. A kombináció során a BAYSIC csak a konkrét variánspozíciókat használja fel, az annotációs információkat nem.

Variánsok annotálása A variánsok annotálása létfontosságú lépés a szekvenálási adatok elemzésében. Ennek során a variánsokhoz funkcionális információkat rendelünk, mint például jósolt funkció, hivatkozások különböző genomi adatbázisokra, konzerváltsági mutatók, allélfrekvencia információk különböző genomi projektekben, a variáns betegségokozó hatásának jóslása különböző predikciós algoritmusokkal, mikroRNS-t érintő variánsok esetén a jósolt mikroRNS-szerkezet megváltozása, funkcióvesztés jóslása, génszabályozás módosításának jóslása stb..

Az annotáció egyik legfontosabb lépése a variánsok funkcionális annotálása, azaz a variánsok potenciális hatásának jóslása a génekre, a transzkriptumokra, illetve a keletkezett fehérjetermékekre vonatkozóan. Ezt a feladatot számos programmal el tudjuk végezni (pl. ANNOVAR [62], VEP [63], SnpEff [64]). Fontos azonban megjegyezni, hogy a predikció alapjául kiválasztott transzkript halmaz (pl. ENSEMBL vagy REFSEQ) és a

szoftver megválasztása is nagyban befolyásolja a végeredményt [65].

A variánsok rs azonosítóval való ellátása, a különböző genomi adatbázisokba mutató hivatkozások és allélfrekvencia információk hasznos segítséget adhatnak az elemzéshez. Például gyakori lépés mendeli öröklődésű betegségek vizsgálata esetén azoknak a variánsoknak a kizárása az elemzésből, amelyek szerepelnek a dbSNP-ben, vagy bizonyos populációkban gyakoriak az 1000 Genom adatbázis adatai alapján. Ez a szűrés azon a feltételezésen alapul, hogy a ritka betegségekért a ritka variánsok a felelősek, azaz a populációban gyakran előforduló polimorfizmusok nem tehetők felelőssé a betegség kialakulásáért.

Fontos információ lehet a variáns környezetének evolúciós szekvencia konzerváltsága, amely mind a fehérjekódoló, mind a nem kódoló variánsok potenciálisan káros (deleterious) szerepére világíthat rá [66]. Számos predikciós szoftver született, amelyek biokémiai, evolúciós és strukturális információkat is felhasználva, általában gépi tanulási algoritmusok segítségével a variánsok károsságának jóslására használhatók (pl. SIFT [67], PolyPhen-2 [68], MutationTaster [69]). A dbNSFP adatbázis és annotációs szoftver [70] jelenlegi verziója (v3.0) 9 predikciós szoftver és számos egyéb adatbázis adatait foglalja magában, amely által a nemszinonim hatású variánsok széleskörű annotációját teszi lehetővé.

Variánsok elemzése A variánsok elemzési módszerének megválasztása nagyban függ a vizsgált jelleg fajtájától (pl. adott betegség kockázata, túlélés-elemzés, kvantitatív jelleg), gyakoriságától (pl. ritka vs. gyakori betegség), a vizsgált polimorfizmusok számától és gyakoriságától, az egyéb rendelkezésre álló fenotípusos információktól, a minták számától, illetve természetesen a megválaszolandó biológiai kérdéstől, csak hogy néhány fontosabb tényezőt említsek. Ezen módszerek bemutatása messze meghaladná a dolgozat kereteit. A bayesi relevanciaelemzési módszer többek között populációs asszociációs vizsgálatok és diszkretizált (pl. 5 éves túlélés) túlélés-elemzési vizsgálatok esetén, gyakori polimorfizmusok és diszkrét fenotípusos változók összefüggéseinek feltérképezésére használható, melyről részletesebben az 1.3. alfejezetben lesz szó.

DOI:10.14753/SE.2017.1938

1.3. 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. Ebben a fejezetben röviden áttekintjük a módszertan alapjait.

1.3.1. Bayes-hálók

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.

A valószínűségi változók együttes eloszlásának hatékony ábrázolására Bayes-hálókat (más néven valószínűségi hálózatokat) használhatunk. A Bayes-háló egy gráf, amelynek csomópontjai a modellezett tárgytartomány valószínűségi változóinak felelnek meg, a csomópontokat pedig irányított élek kötik össze. Egy X „szülő” csomópontból egy Y „gyermek” csomópontba futó él azt jelenti, hogy az X változó közvetlen befolyással van az Y változóra. A Bayes-hálóban minden csomóponthoz tartozik egy ún. feltételes valószínűségi tábla, amely azt írja le, hogy az adott változó értéke (eloszlása) hogyan függ a szülő változók értékétől. Egy irányított út[8] kezdőpontja az úton szereplő többi csomópont „őse”, míg az irányított út végpontja az út többi pontjának „leszármazottja”. Így egy irányított út azt jelenti, hogy az ős változó indirekt módon hatással van a leszármazott változókra. A valószínűségi hálók esetén fontos megkötés, hogy a gráf nem tartalmazhat irányított köröket, azaz egy csomópont nem lehet a saját leszármazottja vagy őse. A

[8]Irányított út: Egy adott csomópontból az élek irányultságának megfelelő út az élek mentén egy másik csomópontba.

DOI:10.14753/SE.2017.1938

továbbiakban a háló *struktúrá*ja alatt a csomópontok és az azokat összekötő élek összességét értjük, míg a háló *paraméterezés*e a feltételes valószínűségi táblák összessége.

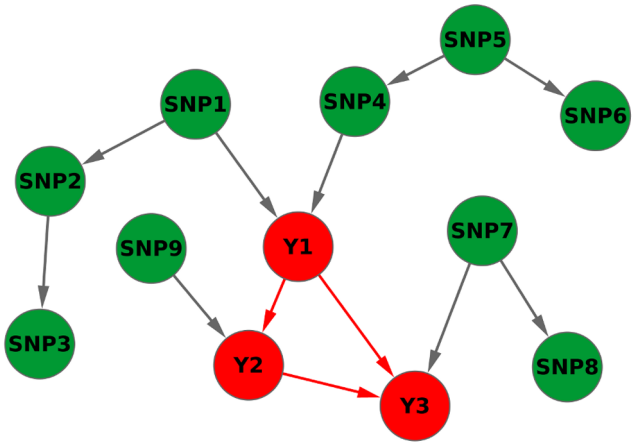
Egy valószínűségi háló akkor képes az együttes eloszlás *hatékony* ábrázolására, ha a háló struktúrája pontosan tükrözi az együttes eloszlásból kiolvasható feltételes függetlenségeket, azaz ha minden csomópontnak pontosan azok a szülei, amelyek az értékét közvetlenül befolyásolják.[9] Egy valószínűségi háló akkor tud megfelelően tömör lenni, ha a rendszer, amit modellez, lokálisan strukturált, azaz egy komponens csak korlátos számú másik komponenssel van közvetlen kapcsolatban. A genetikai asszociációs vizsgálatok tárgytartománya is tipikusan egy ilyen lokálisan strukturált rendszer, ugyanis a génekre is nagy általánosságban igaz, hogy is csak korlátos számú másik génnel állnak közvetlen kapcsolatban [71, 72].

1.3.2. A változók közötti kapcsolati típusok

Amennyiben rendelkezésünkre áll a tárgytartományt leíró valószínűségi háló, akkor annak struktúrájából kiolvashatók a változók között fennálló feltételes függetlenségi állítások. Közvetlenül adódik, hogy minden csomópont feltételesen független minden nem leszármazott csomóponttól az adott csomópont szüleinek ismeretében (azaz ha ismerjük a szülő csomópontoknak megfelelő valószínűségi változók konkrét értékét). Ezek a feltételes függetlenségi állítások továbbiakat vonnak maguk után, amelyeket az ún. *d-szeparáció* segítségével olvashatunk ki a gráf struktúrájából [73]. A d-szeparáció által megfogalmazható állítások közül számunkra legfontosabb az, hogy minden X változó feltételesen független az összes többi változótól a változót reprezentáló csomópont *Markov-határának* (az X csomópont szülei, gyermekei és gyermekeinek egyéb szülei) ismeretében (lásd 4. ábra). Intuitíven megfogalmazva egy változó Markov-határa azokat a változókat tartalmazza, amelyek bizonyos értelemben közvetlen módon korrelálnak az adott változóval, illetve a további modellbeli változók hatását elfedik a változó elől. Azaz ha ismerjük a Markov-határt alkotó változók értékét, akkor a többi változó értéke már lényegtelen; nem ad semmilyen plusz információt az adott csomópont értékére vonatkozóan. Bizonyos általános feltételek teljesülése esetén kimondhatjuk, hogy egy X változó pontosan

[9] Ez azt jelenti, hogy a változók minden olyan sorrendezésére, amely az élek irányítottságával konzisztens, teljesül, hogy egy adott változó a szülei ismeretében feltételesen független a sorrendben őt megelőző változóktól.

akkor *erősen releváns* egy Y változó szempontjából, ha X része az Y Markov-határának. Könnyű ellenőrizni, hogy a Markov-határbeliség (és így az erős relevancia is) szimmetrikus kapcsolat: Y pontosan akkor van X Markov-határában, ha van köztük él, vagy mindketten szülei egy másik változónak.



4. ábra. **A változók közötti különféle kapcsolati típusok ábrázolása egy Bayes-hálóban.** Az ábrán látható gráf egy Bayes-háló struktúráját ábrázolja. A csomópontok a modellezett tárgytartomány valószínűségi változóinak felelnek meg, piros színnel vannak jelölve a kitüntetett célváltozók, zölddel pedig a genetikai variánsoknak megfelelő csomópontok. A pontokat összekötő élek közvetlen kapcsolatokat reprezentálnak a megfelelő valószínűségi változók között. *Páronkénti kapcsolati típusok:* Közvetlen relevancia (pl. *SNP4* és *Y1* között, a köztük lévő közvetlen él miatt), Tiszta (főhatás nélküli) interakció (pl. *Y1* és *SNP9* között, a közös gyermekük *Y2* miatt), Erős relevancia (pl. *SNP1* és *Y1* között, a köztük lévő közvetlen relevancia miatt vagy pl. *SNP7* és *Y1* között, az *Y3* változón kersztül megvalósuló interakciós hatás miatt), Tranzitív relevancia (pl. *SNP5* és *Y3* között, a köztük lévő irányított út miatt), Zavart relevancia (pl. *SNP3* és *Y2* között, a köztük lévő irányítatlan út miatt), Asszociáció (pl. *SNP3* és *Y2* között vagy pl. *SNP5* és *Y3* között).*Variánshalmazok relevanciája:* Erős relevancia (más néven Markov-határ halmaz, pl. *Y1* változó Markov-határát az *Y1* szülei, gyermekei és a gyermekei egyéb szülei alkotják: {*SNP1*,*SNP4*,*Y2*,*Y3*,*SNP9*,*SNP7*} Amennyiben ezeknek a változóknak az értékét megismerjük, *Y1* függetlenné válik a háló többi csomópontjától.) Forrás: doi:10.1371/journal.pone.0033573.g001

A fentiek alapján adódik, hogy a genetikai asszociációs vizsgálatok során egy adott fenotípusos célváltozó (pl. betegség megléte) szempontjából erősen releváns változók jelentik azokat a genetikai és fenotípusos faktorokat, amelyeknek közvetlen hatása van a célváltozóra. Így a fő célunk az erősen releváns változók felderítése lesz (illetve ezen kapcsolatok valószínűségének becslése az adataink alapján). Egy adott faktor és a célváltozó között a következő kapcsolati típusokat definiáljuk a valószínűségi hálózatok struktúrája

DOI:10.14753/SE.2017.1938

alapján (lásd 4. ábra):

Közvetlen relevancia: A faktor és a célváltozó között közvetlen él fut. Ilyen kapcsolat van például az oki (kauzális) SNP és a fenotípus között.

Tiszta (főhatás nélküli) interakció: A faktornak és a célváltozónak van egy közös gyermeke. Ebben az esetben az adott faktornak nincs saját (fő) hatása, de a gyermek csomópont ismeretében már befolyásolja a célváltozó értékét, így interakciós hatást fejt ki a célváltozóra. Az ilyen típusú kapcsolat lehet például az, amikor egy episztatikus hatású SNP befolyásolja egy másik SNP hatását a fenotípus megjelenése szempontjából.

Erős relevancia: A faktor és a célváltozó között közvetlen relevancia vagy tiszta interakciós hatás van, azaz közvetlen él fut közöttük vagy van közös gyermekük.

Tranzitív relevancia: A faktor és a célváltozó között irányított út van. Ilyen kapcsolat lehet például az oki variánssal kapcsoltsági egyenlőtlenségben álló további variánsok és a célváltozó között.

Zavart relevancia: A faktor és a célváltozó között irányítatlan út van (azaz olyan út, amely során legalább egy csomópontban az élek egymás felé vagy egymástól elfelé néznek). A kapcsoltsági egyenlőtlenségek komplex összefüggésrendszere miatt ilyen kapcsolat lehet az oki variánssal kapcsolatban álló további variánsok és a célváltozó között.

Asszociáció: A faktor és a célváltozó között közvetlen, tranzitív vagy zavart relevancia van, azaz létezik irányított vagy irányítatlan út a két változó között, amely egyetlen élből is állhat. Ilyen kapcsolat lehet például az oki variáns és a célváltozó között, illetve az oki variánssal kapcsoltsági egyenlőtlenségben álló további variánsok és a célváltozó között.

Fontos hangsúlyozni, hogy a frekventista megközelítés szerinti asszociáció fogalma nem egyezik meg az erős relevancia fogalmával, ugyanis (1) az asszociáció nem tartalmazza a tiszta interakciós hatást, illetve (2) az erős relevancia a célváltozó szempontjából

közvetlen hatású variánsokat foglalja magában, azaz az asszociációval szemben nem tartalmazza a tranzitív és zavart relevanciát. Megjegyezzük továbbá, hogy két változó közötti erős relevancia által teremtett kapcsolat abban az értelemben mondható közvetlennek, hogy a *modellben szereplő* változók közül egyetlen további változó sem tehető felelőssé a közöttük lévő függőség kialakításáért. Létezhet ugyanakkor egy olyan, nem megfigyelt változó, amely közvetlennek tűnő hatást teremt, és amelyet ha megfigyelnénk, akkor ez a közvetlen hatás eltűnhetne. Emiatt a relevanciaelemzés eredményeit mindig a megfigyelt változók tükrében lehet csak értelmezni.

1.3.3. A változók közötti kapcsolati típusok valószínűségének meghatározása bayesi modell átlagolással

Az előző alfejezet alapján, ha meg akarjuk határozni a változók közötti függőségek és feltételes függetlenségek teljes rendszerét, az azzal lenne egyenértékű, hogy meghatározzuk a legvalószínűbb Bayes-háló struktúrát a mérési eredményeink (genotipizálási adataink és a fenotípusos jellemzők vizsgálata) alapján, majd a háló struktúrájából kiolvassuk a keresett összefüggéseket. A legtöbb esetben azonban a rendelkezésünkre álló adat mennyisége a változók számát tekintve nem elegendően nagy, így nagyon sok, nem elhanyagolható *a posteriori* valószínűségű modellstruktúra lehet. Ennek megfelelően nem célunk a teljes struktúra meghatározása, hanem a bayesi statisztika segítségével (ún. bayesi modell átlagolással) megbecsüljük az egyes kapcsolati típusok, mint strukturális jegyek fennállásának valószínűségét az adataink alapján:

$$Pr(f|D) = \sum_{G} f(G) * Pr(G|D), \tag{3}$$

ahol a G egy hálóstruktúrát, D pedig az adatainkat reprezentálja, az $f(G)$ pedig egy indikátorfüggvény, azaz értéke 1, ha az f strukturális tulajdonság (pl. két változó erős relevanciája) fennáll a G struktúrában, és 0, ha nem. Intuitíven megfogalmazva: egy strukturális tulajdonság *a posteriori* valószínűségét úgy számítjuk ki, hogy összeadjuk azon hálóstruktúrák *a posteriori* valószínűségét, amelyek tartalmazzák az adott strukturális tulajdonságot. Amennyiben az így kiszámított valószínűség közel van 1-hez, akkor majdnem minden legnagyobb valószínűségű modellben igaz, hogy az adott tulajdonság/kapcsolati

típus fennáll. Ha azonban a valószínűség alacsony, akkor az adott tulajdonság/kapcsolati típus nem áll fenn a legvalószínűbb modellekben.

A Bayes-tételt alkalmazva adódik, hogy $Pr(G|D) \propto Pr(D|G) * Pr(G)$. A $Pr(D|G)$ kifejezés az adat marginális likelihoodja a G struktúra mellett (azaz annak valószínűsége, hogy a D adatot figyeltük meg, ha a G struktúra valódi), a $Pr(G)$ érték pedig a G struktúra *a priori* valószínűsége. A kísérleteink során neminformatív priorokat használtunk. Általános feltételek teljesülése mellett a marginális likelihood könnyen számítható [74]. Mivel azonban a lehetséges hálóstruktúrák száma szuperexponenciális a változók számának függvényében, a 3. képletben szereplő egzakt összegzés helyett közelítő módszereket kell használni. A bayesi statisztikában leggyakrabban használt technika a Markov-lánc Monte Carlo módszerek használata [75], amely során egy ún. Markov-láncot[10] definiálunk a hálóstruktúrák felett úgy, hogy annak eloszlása a kívánt *a posteriori* $Pr(G|D)$ eloszlás legyen. Ezután egy tetszőlegesen választott (általában véletlen) struktúrából kiindulva a Markov-láncon véletlen sétákat teszünk, és a konvergencia elérése után mintavételezéssel hálóstruktúrákat választunk ki, amelyek alapján a 3. képlet értékét közelítjük. Egy lépés a Markov-láncban a struktúra lokális transzformációját jelenti; ilyen például egy él hozzáadása, kitörlése, illetve megfordítása (ügyelve arra, hogy ne jöjjön létre irányított kör a gráfban) [76, 77].

1.3.4. Interakciók és redundanciák meghatározása

A bayesi modell átlagolás során nemcsak az 1.3.2. alfejezetben definiált kapcsolati típusok, hanem egyéb tulajdonságok *a posteriori* eloszlását is kiszámíthatjuk. Az egyik legfontosabb ezek közül egy adott célváltozó szempontjából erősen releváns változók halmaza, azaz az ún. Markov-határ halmazok (Markov-blanket set, MBS) eloszlásának kiszámítása. Egy MBS halmaz poszteriorja azt fejezi ki, hogy az adataink tükrében mekkora annak valószínűsége, hogy csak és kizárólag a halmazban szereplő változók erősen relevánsak a célváltozó szempontjából. Az MBS halmazok eloszlása a legtöbbször szintén „lapos”, azaz nagyon sok nem elhanyagolható valószínűségű halmaz van. Lehetséges azonban, hogy kisebb részhalmazok (pl. egy, két, három stb. elemű részhalmazok) vi-

[10]Markov-lánc: Egy véletlen folyamat, amely egy adott állapottér elemei között lépked úgy, hogy a következő állapot valószínűségi eloszlása csak a jelenlegi állapottól függ, azaz nem függ attól az úttól, amelyen a jelenlegi állapotba eljutottunk.

szonylag stabilan részei a legvalószínűbb MBS halmazoknak. Ez lehetővé tesz egy köztes elemzési szintet: az ún. $k{-}MBS$ részhalmazok vizsgálatát, ahol k a részhalmaz méretét jelenti. Ezek alapján látható, hogy az $1{-}MBS$ megegyezik az 1.3.2. alfejezetben említett egyváltozós erős relevanciával (Markov-határbeliséggel).

A változók közötti összefüggések miatt egy adott célváltozó szempontjából erősen releváns változók halmazába nem feltétlenül egymástól függetlenül kerülnek be a változók. Előfordulhat ugyanis, hogy egy változó csak akkor kerül be, ha egy másik változó már bent van (interakció), illetve az is, hogy egy változó „kiszorít" egy másikat (redundancia). A $k{-}MBS$ halmazok valószínűségének becslése lehetővé teszi, hogy ezeket a hatásokat számszerűsítsük, azaz meghatározzuk a változók közötti redundanciákat és interakciókat. Ebben a kontextusban az interakció tehát azt jelenti, hogy két változó a vártnál gyakrabban fordul elő együtt az erősen releváns változók halmazaiban, a redundancia alatt pedig azt értjük, hogy két változó a vártnál ritkábban szerepel együtt az MBS halmazokban. Jelölje $\underline{Pr}(s|D)$ annak *a posteriori* valószínűségét, hogy az s halmaz egy erősen releváns részhalmaz ($k{-}MBS$), illetve legyen X_1 és X_2 két változó. Ekkor kiszámíthatjuk a következő mutatót:

$$R = \frac{\underline{Pr}(\{X_1,X_2\}|D)}{\underline{Pr}(\{X_1\}|D) * \underline{Pr}(\{X_2\}|D)}. \tag{4}$$

Ha R különbözik 1-től, akkor az interakcióra ($R > 1$) vagy redundanciára ($R < 1$) utal. Az interakciós arányt ekkor a $log(R)$, a redundancia arányt a $-log(R)$ kifejezéssel számíthatjuk ki. A képlet tetszőleges számú változóra kiterjeszthető, azaz lehetőség van szabadon választott számú változó közti interakció és redundancia kiszámítására is.

1.4. 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.) [78]. A gyermekkori malignus megbetegedések viszonylag ritkák (100000 gyermekre 15 új megbetegedés esik évente), de ezek közül a leukémia a leggyakoribb (kb. 35%), és a balesetek után ez jelenti a vezető gyermekkori halálokot. A különböző leukémia típusok közül az akut limfoid leukémia (ALL) a leggyakoribb (kb. 80%), ezután következik a gyakorisági

DOI:10.14753/SE.2017.1938

sorrendben az akut mieloid (kb. 15%), majd a krónikus mieloid leukémia (kb $0{,}5\%$) [79]. A magyarországi tapasztalatok azt mutatják, hogy az akut limfoid leukémiával kezelt betegek kb. 90%-ban remisszióba hozhatók, az öt éves túlélés pedig $80-85\%$ körüli, ami jónak tekinthető, bár – főleg a korai elhalálozások magasabb hányada miatt – elmarad a nyugat-európai átlagtól [79].

Magyarországon az ALL kezelését a Berlin, Frankfurt és Münster városok gyermekonkológiai klinikái által kezdeményezett nemzetközi munkacsoport (BFM) által kidolgozott protokollok alapján végzik. A betegeket a terápia elején a kockázatuk alapján alacsony (low risk, LR), közepes (medium risk, MR) vagy magas (high risk, HR) rizikójú csoportokba sorolják. A besorolás szempontjai az egyes protokollokban némileg eltérőek, de általában a következő kritériumokon alapulnak: (1) leukocitaszám a diagnózis felállításakor, (2) életkor, (3) prednizolonra adott válasz, (4) remissziós státusz a kezelés 33. napján, (5) t(9;22) (BCR/ABL1 fúzió, Philadelphia kromoszóma), t(4;11) (MLL/AF4 fúzió) kromoszomális transzlokációk léte és (6) immunfenotípus (T vagy B-sejtes leukémia). A kezelés során a rizikócsoport besorolástól függően a betegek vinkrisztint, daunorubicint, L-aszparaginázt, prednizolont, ciklofoszfamidot, citozinarabinozidot, 6-merkaptopurint, metotrexátot, adriamicint, dexametazont, doxorubicint és 6-tioguanint kapnak [79, 80].

Az ALL egy multifaktoriális betegség; genetikai és környezeti faktorok egyaránt befolyásolják a betegség kialakulását. Az eddig feltárt környezeti tényezők közé tartoznak bizonyos prenatális és posztnatális hatások, például vírusfertőzés, sugárhatás, bizonyos kemikáliák, drogok, szülői alkoholfogyasztás és dohányzás [81–88]. Számos vizsgálatban azt találták, hogy a betegség rizikóját befolyásoló környezeti tényezők hatását bizonyos génvariánsok módosítják [86, 89–91], így a genetikai háttér és a környezeti tényezők szerepe egyaránt jelentős az ALL-re való hajlam szempontjából.

Az ALL hajlamával illetve farmakogenetikájával kapcsolatban az egyik gyakran vizsgált anyagcsere-útvonal a folát metabolizmus, amelynek jelentős szerepe van a DNS-metiláció, illetve a DNS-szintézis és javítás mechanizmusaiban.

Az utóbbi időben több teljes-genom asszociációs vizsgálat (genome-wide association study) azonosított olyan vérképzésben, a sejtek kialakulásában, proliferációjában, differenciálódásában vagy a sejtciklus szabályozásában szerepet játszó géneket, amelyek szignifikánsan befolyásolták az ALL kialakulásának kockázatát [78, 92–96].

DOI:10.14753/SE.2017.1938

1.5. A CYP3A4 potenciális szerepe a gyermekkori akut limfoid leukémia farmakogenetikájában

Az ALL öt éves 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. A gyermekkori ALL több okból is gyakori célpontja a farmakogenetikai kutatásoknak, ugyanis a környezetei tényezők hatása jóval körülhatárolhatóbb, mint a felnőttek esetén, másrészt jellemzően a betegek nem szenvednek egyéb társbetegségekben, amelyek befolyásolnák a terápiás készítmények hatását [97].

Az emberi szervezetben a citokróm P450 enzimcsalád tagjai számos különböző típusú gyógyszer, valamint endogén anyag oxidatív metabolizmusában játszanak szerepet. Az enzimek a szervezetbe került idegen anyagok (xenobiotikumok) I. fázisú metabolikus folyamataiban vesznek részt, ezáltal előkészítik az anyagokat a II. fázisú konjugációs reakciókra. 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 [98, 99], 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 CYP3A4 aktivitása és fehérje expressziója szignifikiánsan korrelál az mRNS expressziójával [100], több kutatásban is vizsgálták a gén expressziós szintjét befolyá-

DOI:10.14753/SE.2017.1938

soló tényezőket. Egy egészséges önkénteseket vizsgáló tanulmányban Ozdemir és mtsai megállapították, hogy az enzimszint nagyfokú variabilitásának örökölhetősége körülbelül 90%-os, ami a genetikai variációk jelentős szerepét mutatja a gén expressziós szintjének befolyásolásában [101]. Ez azonban egyes vizsgálatok szerint nem magyarázható önmagában a *CYP3A4* gyakori variánsainak hatásával [102]. Lamba és mtsai májdonorok vizsgálata során azt találták, hogy a páciens neme, illetve a *FoxA2*, *HNF4alpha*, *FoxA3*, *PXR*, *ABCB1* gének, és a *CYP3A4* promóterének egyes variánsai együttesen a májban mért *CYP3A4* expressziós szint variabilitásának $24{,}6\%$-át voltak képesek megmagyarázni [103].

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

DOI:10.14753/SE.2017.1938

2. Célkitűzések

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.

DOI:10.14753/SE.2017.1938

3. Módszerek

3.1. Mesterséges szekvenciaadatok előállítása

Az egyes variánskivonatoló programok, illetve az általam kifejlesztett VariantMetaCaller program teljesítményének összehasonlítását nagyban megkönnyíti mesterségesen előállított variánsok és szekvenciaadatok használata. Ekkor ugyanis (1) mivel pontosan ismerjük a mintákban szereplő variánsokat, így ki tudjuk számítani a módszerek különböző teljesítménymutatóit (pl. szenzitivitás, precizitás) is, és (2) számos olyan tényező hatását szisztematikusan ki tudjuk értékelni, amely az eredményeket befolyásolhatja (pl. illesztőprogram, leolvasási mélység).

Első lépésként mesterséges szekvenciaadatokat állítottam elő a következő módon: (1) Illusztrációs célokból kiválasztottam a 17-es kromoszómát, és létrehoztam 100 darab haploid kromoszómát a hg19 referencia szekvencia alapján. (2) Az egyes kromoszómák szekvenciáját módosítottam a következőképpen: egy általam fejlesztett szimulációs szoftver segítségével az Exome Aggregation Consortium (ExAC) publikusan elérhető variánslistájából (Exome Aggregation Consortium, Cambridge, MA, 0.3-as verzió, http://exac.broadinstitute.org Hozzáférés: 2015.01.20.) az alternatív allélok allélfrekvenciájával megegyező valószínűséggel véletlenszerűen kiválasztottam variánsokat, és a kromoszómák szekvenciájában a megfelelő pozíciókban a referencia allélt kicseréltem az alternatív allélra. Ezáltal olyan realisztikus mesterséges kromoszómákat kaptam, amelyekben lévő variánsok allélfrekvenciája megközelítőleg hasonló a valós variánsokéhoz. (3) A haploid kromoszómákat párosítottam, így összesen 50 diploid kromoszómával rendelkező mesterséges mintát kaptam. Ezeket öt, egyenként tíz tagot számláló csoportba soroltam. (4) Az előbbi szimulációs szoftver és az ART program [104] segítségével (verzió: VanillaIceCream 2014/11/03) egy paired-end szekvenálást szimuláltam (2×105 bp hosszú leolvasások, az átlagos inzert méret: 180, illetve a szórása: 10) a kromoszómák exonikus régióiról. Minden minta esetén több különböző lefedettségű adathalmazt hoztam létre ($4, 8, 12, 16, 20, 30, 40, 60, 100$ és 200) annak érdekében, hogy szisztematikusan ki tudjam értékelni a későbbiekben a leolvasási mélységnek hatását az egyes módszerek teljesítményére. A leolvasások véletlen hibákat is tartalmaztak, ehhez az ART program

Illumina szekvenáló specifikus hibamodelljét használtunk fel.

A csoportok számának meghatározása során a következő praktikussági szempontot tartottam szem előtt: mivel a variánskivonatolási munkafolyamatok lefuttatása a tízféle lefedettség szerint generált adatokon nagy számítási igénnyel rendelkezik, és a futási idő egyenesen arányosan nő a csoportok számával, a független csoportok számát lehetőség szerint minimalizáltam. Az így maghatározott öt csoport a későbbiekben elegendőnek bizonyult a statisztikai értékelésekhez.

Az egyes variánskivonatolási módszerek szenzitivitásának és precizitásának mérése során során a mintákban szereplő valódi variánsokat használtam referenciaként, azaz az egyes módszerek eredményét ezekkel a referenciavariánsokkal vetettem össze.

Megjegyzem, hogy a fenti generálási sémában az egyes variánsokat egymástól függetlenül választottuk ki, azaz az ilyen módon generált minták nem tükrözik a populációban megfigyelhető kapcsoltsági egyenlőtlenségeket. Ez azonban nem befolyásolja az eredményeket, ugyanis a leolvasások hossza kisebb, mint a variánsok átlagos távolsága, illetve a jelenlegi variánskivonatoló programok nem használnak fel referenciaként kapcsoltsági információkat.

A későbbiekben az egyes variánskivonatolási módszerek eredményét az exonikus célrégióra szűkítettem, ugyanis az ExAC adatbázis csak exonikus variánsokat tartalmaz.

Annak érdekében, hogy a módszerek teljesítményét különböző genomi célrégiók esetén is megvizsgálhassam, további szűréseket is végeztem: kiválasztottam $10-10$, egymást követő genomi régiót, amelyek összesített mérete 100, 200, 300, illetve 500 kb hosszúságú volt.

3.2. Valós szekvenciaadatok

Az Illumina BaseSpace honlapjáról (https://basespace.illumina.com/home/index Hozzáférés: 2015.02.10.), a „HiSeq 2000: TruSeq PCR-Free (Platinum Genomes)” projektből letöltöttem az NA12878-as egyén valós szekvenálási adatait. Ez a minta először a nemzetközi HapMap projektben szerepelt, és azóta számos kutatásban használták; mára minden bizonnyal ez lett a legtöbbször szekvenált emberi genom. Emiatt elérhető hozzá egy nagy pontosságú, „platinum minőségű” referencia variánslista is, amit az összehasonlítások során referenciaként használtam (Illumina Platinum Genomes, 6.0-ás verzió,

http://www.illumina.com/platinumgenomes/ Hozzáférés: 2015.02.10.).

3.3. Variánskivonatolási munkafolyamatok

A leolvasásokat először minőségi szűréseknek vetettem alá. Ehhez a picard-tools (Picard Tools, Broad Institute, 1.122-es verzió, http://broadinstitute.github.io/picard/ Hozzáférés: 2015.04.26.) programot használtam. A szűrések során eldobtam azokat a leolvasásokat, amelyek átlagos bázisminősége 25 Phred-pontszám alatti volt, vagy tartalmaztak nem egyértelmű nukleotidokat. Mivel a szekvenálás pontossága a leolvasás előrehaladtával általában fokozatosan romlik, előfordulhat, hogy a szekvencia vége rossz minőségű. Emiatt a leolvasások végéről a program adaptívan levágta azokat a bázisokat is, amelyek átlaga nem haladta meg a 20-as Phred-pontszámot. Amennyiben az így megmaradt szekvencia hossza kisebb volt mint 40 bp, a leolvasást a program eldobta.

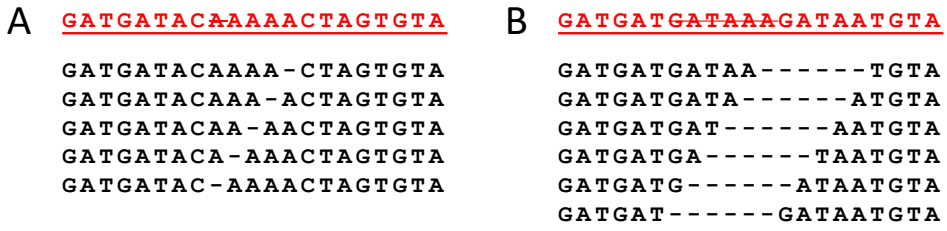
A minőségi szűrések után a leolvasásokat a BWA–MEM [51] és a Bowtie 2 [52] programok segítségével felillesztettem a hg19 referenciagenomra. A BWA esetén az alapértelmezett beállításokat használtam, a Bowtie 2 esetén pedig az ún. *very-sensitive*, azaz nagy szenzitivitású alapbeállítást.

A valós szekvenciaadatok esetén ezután lefuttattam a GATK bázisminőség-korrekcióját (base quality score recalibration) és az indel környéki szekvenciák újraillesztését (indel realignment).

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 [105] (3.3-0-ás verzió), a FreeBayes-t [58] (v0.9.20-17-g5f1bc44-dirty verzió) és a SAMtools-BCFtools programok [57] (1.1-22-gc61d8d1 verzió) kombinációját. Utóbbira a dolgozat során SAMtools-ként fogok hivatkozni. A programokat a FreeBayes kivételével alapértelmezett beállításokkal futtattam. A FreeBayes-t úgy állítottam be, hogy ne hívjon több nukleotidot érintő polimorfizmusokat (egymás melletti SNP-k összeolvasztásával) és komplex eseményeket (pl. azonos szálon lévő közeli indel és SNP egybeolvasztásával), mivel a többi módszer ilyen komplex variánsokat nem hív.

Általánosságban elmondható, hogy az indelek bizonyos esetekben a referencia szekvencia több lehetséges pozícióján is ábrázolhatóak (lásd 5. ábra), ez azonban megnehezíti

a variánskivonatolási eredmények összehasonlíthatóságát. Emiatt az indeleket a BCFtools program segítségével a lehetséges pozícióikon belül balra rendeztem.



5. ábra. **Törlődéses variánsok nem egyértelmű reprezentációjának illusztrálása.** Mindkét ábrán a felső, piros, aláhúzott szekvencia a referencia szekvencia része, amelyből az áthúzott részszekvencia törlődött. A feketével jelölt illesztések a törlődés lehetséges helyeit, illetve a lehetséges variánsreprezentációkat mutatják be, amelyek ekvivalensek egymással. A legalsó sor a lehetséges helyeken belül balra rendezett indelt ábrázolja, amely már az indel egyértelmű reprezentációját adja. **A**. A homopolimer szakaszból törlődött egyetlen nukleotid a homopolimer szakasz bármelyik pontján ábrázolható. **B** Példa egy bonyolultabb, nem ismétlődő, de átfedő szakasz törlődésére, amely szintén a szakasz bármelyik pontján ábrázolható.

A GATK alapú kivonatolók eredményét manuális szűrők használatával leszűrtem a GATK ajánlásainak [53] megfelelően. A FreeBayes és a SAMtools által hívott variánsokat szintén leszűrtem a variánsminőség minimális küszöbértékének meghatározásával. Ez az eljárás szintén megfelel a fejlesztők ajánlásainak. A szimulált adatok esetén 100-as Phred-pontszám, a valós adatok esetén 30-as Phred-pontszám volt a minimálisan elvárt variánsminőség.

3.4. A variánskivonatolók eredményeinek kombinálása a VariantMetaCallerrel

3.4.1. A VariantMetaCaller általános leírása

Az általam kifejlesztett VariantMetaCaller program az egyedi szűretlen variánskivonatolási eredményeket kombinálja SVM-ek használatával a következő módon:

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 pozitív tanítóminták

kiválasztásához a következő heurisztikát használjuk: kiindulunk egy t küszöbértékből, ami kezdetben a variánskivonatolók számával egyezik meg (azaz jelen esetben 4-el). Ezután meghatározzuk azokat a variánsokat, amelyeket legalább t módszer megtalált. Ha minden kivonatoló esetén van legalább egy ilyen variáns, akkor ezek lesznek a pozitív tanítóminták. Ellenkező esetben t-t addig csökkentjük, amíg nem találunk legalább egy, a feltételeknek megfelelő variánst. Ezután minden egyes kivonatoló esetén meghatározzuk azokat az egyedi variánsokat, amelyeket csak ez a módszer talált meg. Amennyiben minden módszer esetén találunk ilyen variánst, akkor ezek fogják alkotni a negatív tanítómintákat, és kétosztályos SVM-et használunk. Ha azonban van legalább egy olyan kivonatoló, amely esetén nincs egyedi, csak ezáltal hívott variáns, akkor egyosztályos SVM-et használunk, ahol csak a pozitív tanítómintákat vesszük figyelembe.

Az SVM funkcionalitás implementálásához a LIBSVM [106] fejlesztői könyvtárat használtam fel és módosítottam.

3.4.2. A Szupport Vektor Gépek paraméterezése

Az SVM tanítása során a program ún. radiális bázisfüggvényt (RBF) használ a kernelek kiszámításához. Ebből eredően az SVM tanításához két paraméter megadására van szükség. Ezek az osztályozás büntetési paramétere (C) és a gaussi függvények szélességét befolyásoló γ paraméter. A paraméterek konkrét értékét egy két szintű, rács-módszer típusú kereséssel állítjuk be. Az első szinten a C paraméter értékét 2^{-5}-től 2^{17}-ig növeljük 2^2-szoros lépésközzel (azaz 2^{-5}, 2^{-3}, 2^{-1}, $\ldots$), a γ értékét pedig 2^{-17}-től 2^3-ig, szintén 2^2-szoros lépésközzel, és minden pontban egy 5-szörös keresztkiértékelési sémával (lásd később) meghatározzuk az SVM pontosságát az adott pontnak megfelelő paraméterezés mellett. Ezt követően az első szinten legjobb eredményt adó térrészt a második szinten egy finomabb skálával ($2^{0.2}$-szeres lépésközzel) újra bejárjuk, és szintén keresztkiértékeléssel meghatározzuk a legnagyobb pontosságot eredményező paraméterezési beállítást. Legvégül a legjobbnak bizonyuló (C,γ) paraméterekkel kiszámítjuk a végső modellt a teljes adathalmaz felhasználásával.

A rács-módszer alapú keresést az OpenMP programkönyvtár felhasználásával párhuzamosítottam, így a program hatékonyan ki tudja használni a futtató számítógépen rendelkezésre álló számítási egységeket (CPU-k, CPU magok).

A keresztkiértékelés során az adathalmazt (a tanítómintákat) véletlenszerűen felosztjuk k egyenlő részre, majd minden egyes rész esetén a következőképpen járunk el: a kiválasztott $\frac{1}{k}$-ad rész lesz a *teszt* halmaz, a fennmaradó $\frac{k-1}{k}$-ad rész pedig a *tanító* halmaz. A tanító halmazzal megtanítjuk a modellt, majd annak felhasználásával megjósoljuk a teszt halmaz elemeinek hovatartozását. Mivel ez eleve ismert volt, ezért a teszt halmaz elemeinek valódi értéke alapján ki tudjuk számítani a modell pontosságát. A becsült végső pontosságot az k keresztkiértékelési lépés során elért pontosságok átlagaként kapjuk meg. A keresztkiértékelések során a felosztások számának (k) meghatározására érzékenységvizsgálatot végeztünk, mely során a $k = 5$ érték megfelelő kompromisszumnak bizonyult a szisztematikus eltérés mértéke (bias) és a futtatáshoz szükséges idő között.

3.4.3. A tanítás során felhasznált jellemzők, annotációk

A program minden egyes variánskivonatolóhoz előállít egy adathalmazt az adott kivonatoló által generált, illetve egyéb annotációs adatok alapján. Ez az adathalmaz definiálja azt a teret, amelyben az SVM a pozitív tanítópontként szereplő variánsokat elszeparálja a negatív tanítópontként szereplő variánsoktól. Az adathalmaz létrehozásához az alapvető annotációkon (pl. leolvasási mélység, variánsminőség, illesztési minőség stb.) túl további jellemzőket is felhasználunk: (1) az adott variánshoz legközelebbi variáns távolsága, (2) a genotípus-eloszlás entrópiájának átlaga illetve szórása az összes minta felett és (3) a referencia szekvencia entrópiája a variáns közvetlen közelében. Az összes felhasznált annotáció felsorolása és rövid magyarázata a Függelék: 1. táblázatban található.

3.4.4. Variánsok valószínűségének kiszámítása

Mindegyik variánskivonatoló esetén minden egyes variánshoz kiszámítjuk annak feltételes valószínűségét, hogy az adott variáns "valódi" (azaz a pozitív osztályba tartozik) [107]. Ezután a következő képlettel kiszámítjuk minden variáns valószínűségét a variánskivonatolók felett kiátlagolva (a módszerek között egyenlő súlyt feltételezve):

$$P_{SVM}(i) = \frac{1}{N} \sum_{j=1}^{N} Pr(y_{ij} = 1 | x_{ij}), \tag{5}$$

ahol N a variánskivonatolók száma, $Pr(y_{ij} = 1|x_{ij})$ pedig annak a valószínűsége, hogy az i-dik variáns a pozitív osztályba tartozik a j-dik variánskivonatoló esetén, az x_{ij} megfigyelések (annotációs mutatók értékei) alapján.

3.4.5. Várható precizitás kiszámítása

A variánsok valószínűségének kiszámítása lehetővé teszi, hogy megbecsüljük egy adott variánshalmaz várható precizitását, azaz a halmazba tartozó valódi variánsok várható arányát. A precizitás maximalizálásának érdekében a variánsokat valószínűségük szerint csökkenő sorrendbe rendezzük, és a sorrend mentén minden egyes i indexre kiszámítjuk a várható precizitást a következő képlettel:

$$\begin{aligned} E^{(i)}_{PREC} &= \frac{TP^{(i)}}{TP^{(i)} + FP^{(i)}} \\ &= \frac{\sum_{j=1}^{i} P^{(j)}_{SVM}}{\sum_{j=1}^{i} P^{(j)}_{SVM} + \sum_{j=1}^{i}(1 - P^{(j)}_{SVM})} \\ &= \frac{1}{i}\sum_{j=1}^{i} P^{(j)}_{SVM}, \end{aligned} \tag{6}$$

ahol $TP^{(i)}$ a valódi pozitívok, $FP^{(i)}$ pedig a hamis negatívok várható száma a sorrend mentén az i-dik indexű variánsra. Ekkor a várható hamis felfedezési arány (FDR) pedig nem más, mint $E^{(i)}_{FDR} = 1 - E^{(i)}_{PREC}$.

3.4.6. A módszerek összehasonlítása

Az egyes módszerek teljesítményét precizitás-szenzitivitás görbék segítségével hasonlítottuk össze. Egy ilyen görbe megalkotása egy adott módszer esetén a következőképpen történik: a variánsokat sorrendezzük variánsminőség szerint (a VariantMetaCaller esetén a variánsok valószínűsége szerint) csökkenő sorrendben, azonos értékek esetén leolvasási mélység szerint szintén csökkenő sorrendben. Ezután a sorrend mentén minden elemre kiszámítjuk a precizitást (a valódi variánsok aránya a sorrend elejétől az adott variánsig bezárólag) és a szenzitivitást (az összes valódi variáns hányad része található a sorrend elejétől az adott variánsig bezárólag). Végül a precizitást ábrázoljuk a szenzitivitás függvényében.

Ezen felül kiszámítottuk a precizitás-szenzitivitás görbék alatti terület is (area under the precision-recall curve, AUPRC) trapéz-szabály integrálással [108].

3.5. A CYP3A4 potenciális szerepének vizsgálata a gyermekkori akut limfoid leukémia farmakogenetikájában

3.5.1. Minták

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 mintagyűjtést a Magyar Etikai Bizottság (Egészségügyi Tudományos Tanács Tudományos és Kutatásetikai Bizottság, ETT TUKEB; Esetszám: 8-374/2009-1018EKU 914/PI/08) engedélyezte. A vizsgálatba bevont személyek, illetve kiskorúak esetén szüleik aláírták a beleegyező nyilatkozatot. A vizsgálatok során 1990 és 2010 között ALL-el diagnosztizált betegek adatait elemeztük (511 beteg). A kiválasztott minták a Magyar Gyermek Tumor Regiszter adatai alapján a megadott időtartamban Magyarországon regisztrált gyermekkori ALL-es megbetegedések több mint 40%-t lefedték. A betegeket az ALL Berlin-Frankfurt-Münster (BFM) 90 és 95-ös, vagy ALL IC 2002-es kemoterápiai protokollal kezelték, a rizikócsoport-besorolás a protokollokban meghatározott paraméterek szerint történt. A mintapopuláció jellemzőit a 2. táblázat mutatja be.

A vizsgálatokba bevont személyektől perifériás vérmintát gyűjtöttek. Mivel az elemzések célja csíravonali polimorfizmusok vizsgálata volt, a betegektől a mintát retrospektív módon a remissziós fázisban, illetve őssejt transzplantáción átesett betegek esetén a transzplantáció előtt vették.

3.5.2. A vizsgált gének és SNP-k kiválasztása

A vizsgálat során a *CYP3A4* és *CYP3A5* gént választottuk ki, amelyek ismert gyógyszermetabolizáló enzimeket kódolnak. A vizsgálandó SNP-k kiválasztásában a főbb szempontok a következők voltak: (1) a minor allél frekvencia a kaukázusi populációban meghaladja a 10%-ot, vagy (2) irodalmi adatok alapján feltételezhetően hatással van az gén által kódolt enzim expressziójára (azaz funkcionális SNP). A vizsgált SNP-ket és főbb

DOI:10.14753/SE.2017.1938

2. táblázat. **Az akut limfoid leukémia vizsgálat során elemzett mintapopuláció adatai**

	Betegek
Mintaszám	511
Nem (%)	
Férfi	290 (56,8)
Nő	221 (43,2)
Diagnóziskori életkor	
Átlag (±SD)	6,4 (±4,2)
Medián (Tartomány)	5,1 (1-18)
Rizikócsoport-besorolás (%)	
Alacsony rizikó, LR	94 (21,1)
Közepes rizikó, MR	299 (67,0)
Magas rizikó, HR	53 (11,9)
Sejtmorfológiai alcsoport (%)	
pre-B, B-ALL	372 (82,9)
pre-T, T-ALL	77 (17,1)
Citogenetika (%)	
Normál	127 (27,9)
Hiperdiploid	79 (17,4)
Egyéb	249 (54,7)

tulajdonságaikat a 3. táblázat foglalja össze.

3. táblázat. **Az akut limfoid leukémia vizsgálat során elemzett SNP-k főbb tulajdonságai** Rövidítések: MAF = minor allél frekvencia, UTR = nem transzlálódott régió

SNP (rs#)	Gén	Allél (1/2)[1]	Pozíció a genomban[2]	Genotípus gyakoriság (ALL)			MAF	Szerep a génben/ aminosavcsere v. alternatív név
				1/1[1]	1/2[1]	2/2[1]		
rs15524	CYP3A5	A/G	chr7:99245914	0,835	0,155	0,010	0,088	exon, UTR
rs776746		G/A	chr7:99270539	0,844	0,149	0,007	0,082	intron/ CYP3A5*3/*1
rs2404955	CYP3A4	G/A	chr7:99353279	0,768	0,214	0,018	0,124	near-gene-3
rs12333983		T/A	chr7:99354114	0,771	0,211	0,018	0,123	near-gene-3
rs2242480		C/T	chr7:99361466	0,773	0,217	0,010	0,119	intron
rs4646437		G/A	chr7:99365083	0,747	0,237	0,016	0,134	intron
rs2246709		A/G	chr7:99365719	0,538	0,395	0,067	0,264	intron
rs35599367		G/A	chr7:99366316	0,914	0,084	0,002	0,044	intron/ CYP3A4*22

[1]Az allélek a pozitív szálon; 1, gyakori allél; 2, rizikó allél

[2]Pozíciók a GRCh37.p10 v104.0 szerint

3.5.3. Genotipizálás

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).

DOI:10.14753/SE.2017.1938

3.5.4. Statisztikai elemzések

Frekventista statisztikai elemzések Az adatok frekventista elemzését az R statisztikai szoftverrel végeztem (R Foundation for Statistical Computing, Bécs, Ausztria; v3.0.3) [109]. A túlélési adatok egy- és többváltozós elemzésére Cox-regressziós modellt alkalmaztam, amelyhez a survival [110] csomagot használtam fel az R programban. A modellek szignifikanciájának kiszámításához a log-rank tesztet alkalmaztam, és egy eredményt akkor tekintettem szignifikánsnak, ha a kétoldalú p-érték kisebb volt, mint $0{,}05$. A Kaplan-Meier görbéket a survMisc [111] csomag felhasználásával állítottam elő. A statisztikai erő elemzéséhez a bootstrap módszert alkalmaztam. Ennek során az eredeti adathalmazt 10000-szer újramintavételeztem visszatevéses módon, majd minden így generált adathalmazra elvégeztem a statisztikai tesztet, és az erőt azzal az értékkel becsültem, hogy az esetek mekkora részében kellett elutasítanom a null-hipotézist (azaz az újramintavételezett esetek hányad részében kaptam szignifikáns eredményt).

A rizikócsoport-besorolási változók diszkriminatív teljesítményét a C-index [112] (konkordancia index) kiszámításával végeztem a pec [113] csomag felhasználásával. Ennek konfidencia-intervallumát bootstrap módszerrel számítottam ki. Ehhez az eredeti adatot 100-szor újramintavételeztem visszatevéses módon, majd minden adatsorra kiszámítottam a C-index értékét 75 időpontra ($0{,}2$ évtől kezdve a 15. év végéig $0{,}2$ évenként). Végül a 95%-os konfidencia-intervallumot minden egyes időpontra a C-index értékek $2{,}5$. és $97{,}5$. percentilisének meghatározásával számítottam ki. A kockázat-besorolási változók közötti különbséget t-teszttel számítottam ki minden egyes időpontra, majd a p-értékeket Benjamini-Hochberg módszerével [10] korrigáltam.

Bayesi relevanciaelemzés A bayesi relevanciaelemzés alapjait a 1.3. alfejezetben tárgyaltuk. Az elemzés futtatása során a következő beállításokat használtuk: a Bayes-háló struktúrák mintavételezéséhez (az 1.3. alfejezetben található 3. képlet közelítéséhez) az ún. MC^3 algoritmust (Metropolis Coupled Markov Chain Monte Carlo, [75]) használtuk. A Markov-láncokat a gráfstruktúrák felett definiáltuk; egy lépés a láncban a struktúra lokális megváltoztatását jelenti, amelyeket operátoroknak nevezünk [76, 77]. Egy véletlenszerű, körmentes gráfból kiindulva minden lépésben a következő három operátor közül választottunk egyet egyforma valószínűséggel: (1) él hozzáadása a gráfhoz két véletlen-

DOI:10.14753/SE.2017.1938

szerűen kiválasztott csomópont közé, (2) véletlenszerűen kiválasztott él megfordítása és (3) véletlenszerűen kiválasztott él törlése. Az első és második operátor esetén csak azokat a lépéslehetőségeket vettük figyelembe, amelyek végrehajtásával nem alakul ki irányított kör a gráfban. A láncokat 8×10^7 lépésig futtattuk, de az első 2×10^7 lépésben (ún. burn-in szakasz) nem mintavételeztünk, ekkor ugyanis a lánc feltehetően (konvergencia diagnosztikai mérések alapján) még nem a megfelelő ($Pr(G|D)$) stacionárius eloszlásban jár. A gráfokban egy csomópont szüleinek maximális számát 4-ben korlátoztuk, ugyanis (1) a rendelkezésre álló adat mennyisége általában nem elegendő nagyszámú szülő és a gyermekváltozó közötti kapcsolat kimutatására és (2) a lehetséges szülőhalmazok száma a binomiális koefficiens szerint ($\binom{n}{k}$, ahol k a szülőhalmaz mérete, n pedig a szülőként szóba jöhető változók száma) nő a szülőhalmazok méretével, ami a futási időt is ugyanilyen mértékben növeli meg. A tárgytartomány lokális strukturáltsága miatt ez a korlátozás azonban tipikusan nem jelent problémát (lásd 1.3. alfejezet).

A mintavételezett gráfstruktúrák alapján kiszámítottuk a változók közötti kapcsolati típusok és a célváltozó szempontjából erősen releváns változóhalmazok (MBS) *a posteriori* valószínűségét, illetve a változók közötti interakciókat és redundanciákat (lásd 1.3.2. alfejezet és 1.3.4. alfejezet).

DOI:10.14753/SE.2017.1938

4. Eredmények

4.1. Variánskivonatolási munkafolyamatok teljesítménye és konkordanciája

A munkám során szisztematikusan kiértékeltem egyes variánskivonatolási munkafolyamatok teljesítményét, különös tekintettel a leolvasási mélység hatásának elemzésére. Ehhez mesterséges szekvenálási adatokat állítottam elő, amelyek ismert eltéréseket tartalmaztak a referencia genomhoz képest. Illusztrációs célokból kiválasztottam a 17-es kromoszómát, és olyan diploid kromoszómákat generáltam, amelyek véletlenszerűen kiválasztott exonikus variánsokat tartalmaztak a publikusan elérhető Exome Aggregation Consortium allélfrekvencia adatainak megfelelően (lásd Módszerek). A célrégió a 17-es kromoszóma teljes exonikus tartománya volt, amelynek a mérete kb. 3,47 Mbp. Összesen 50 független mintát hoztam létre, amelyeket öt darab tízfős csoportba soroltam. Az SNP-k és indelek összes száma 14384, illetve 1852 volt, a polimorf SNP-k és indelek átlagos száma (és szórása) 3132 (29,5), illetve 455 (12,9) volt a generált mintákban. A mesterséges kromoszómákról különböző lefedettségi szinteket teljesítő paired-end szekvenálásokat szimuláltam, amelyek Illumina-specifikus szekvenálási hibákat is tartalmaztak.

A szimulált leolvasásokat két gyakran használt illesztőprogrammal, a BWA–MEM és a Bowtie 2 segítségével felillesztettem a humán referencia szekvenciára, majd az öt mintacsoporton variánskivonatolást végeztem négy különböző programmal (GATK HaplotypeCaller, GATK UnifiedGenotyper, FreeBayes és SAMtools).

Végül kiszámítottam a szűretlen variánshívási eredmények szenzitivitását és precizitását, illetve az egyes kivonatoló módszerek együtt járását.

4.1.1. Variánskivonatolási munkafolyamatok szenzitivitása és precizitása

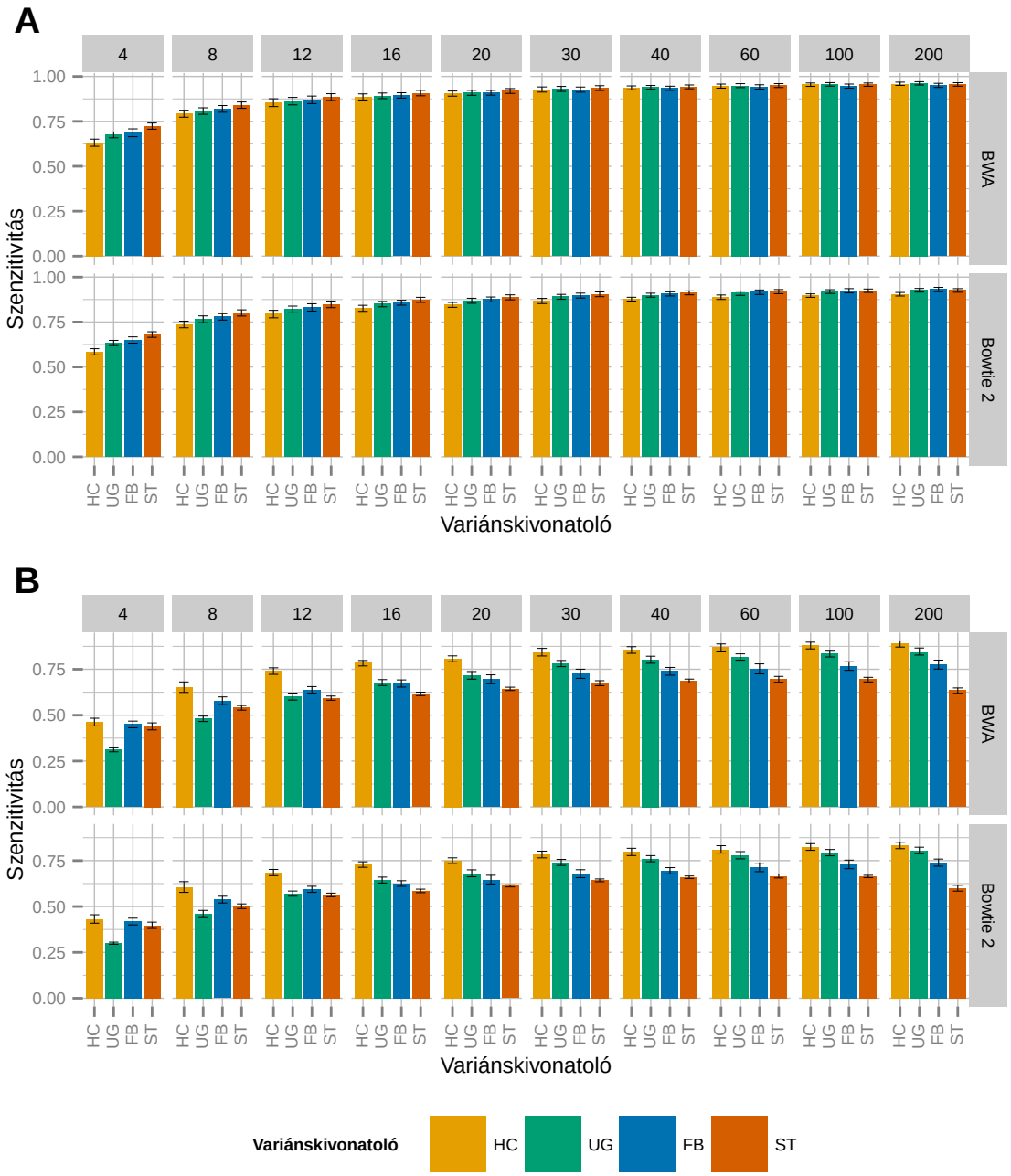
Általánosságban elmondható, hogy a variánskivonatolási módszerek szenzitivitása nőtt a leolvasási mélység növekedésével (lásd Függelék: 2. táblázat). Ez összhangban van az általános elvárással, mely szerint a variánsokat tartalmazó pozíciókról szerzett információ mennyiségének növekedése előnyös a variánshívás szempontjából [43]. Meglepő azonban, hogy a SAMtools szenzitivitása indelek és nagy leolvasási mélységek esetén

($> 60\times$ lefedettség) – az illesztőprogramtól függetlenül – csökkent a lefedettség növekedésével. SNP-k esetén kis és közepes lefedettségekre a SAMtools találta meg a legtöbb valódi variánst, függetlenül az illesztőprogramtól. Ezzel szemben nagyobb ($> 100\times$) leolvasási mélységek esetén a BWA illesztéseket használva a UnifiedGenotyper, míg a Bowtie 2 illesztéseket használva a FreeBayes bizonyult a legszenzitívebbnek (lásd Függelék: 2. táblázat és 6A. ábra). A második legtöbb valódi variánst megtaláló módszer szintén változott a lefedettség növekedésével, de általánosságban és főleg alacsony leolvasási mélység esetén elmondható, hogy a FreeBayes több valódi variánst talált meg, mint a GATK alapú variánskivonatolók. Indelek esetén a HaplotypeCaller szenzitivitása volt a legmagasabb a leolvasási mélységtől és a használt illesztőprogramtól függetlenül (lásd 2. táblázat és 6B. ábra). Alacsony lefedettség esetén a FreeBayes, míg $16\times$-nál nagyobb lefedettségek esetén a UnifiedGenotyper találta meg a második legtöbb valódi indelt.

Minden egyes variánskivonatoló módszer egy belső küszöbértéket használ arra, hogy eldöntse, hogy egy adott pozícióban jelentsen-e variánst vagy sem (az adott pozícióban megjelenő – a referencia szekvenciától való eltérések értékelésekor). Ebből eredően, ha egy módszer magasabb szenzitivitást ér el egy másiknál, akkor ez a precizitás csökkenésével is együtt járhat. Így azt fontos is megvizsgálni, hogy az egyes variánskivonatolóknak mekkora a precizitása, azaz hogy az általuk hívott variánsok hányad része volt valódi variáns. SNP-k esetén a HaplotypeCaller, indelek esetén pedig a UnifiedGenotyper precizitása volt a legmagasabb az illesztőprogramtól függetlenül (lásd Függelék: 2. táblázat). Az eredmények átláthatóbb megjelenítése céljából kiszámítottam a precizitás komplementerét: a hamis felfedezési arányt, azaz a hamisan hívott variánsok arányát az összes variánshoz képest (lásd 7. ábra). SNP-k esetén a leolvasási mélység növekedésével a SAMtools egyre kisebb, míg a UnifiedGenotyper egyre nagyobb arányban hívott hamis variánsokat. Ezzel szemben a HaplotypeCaller és a FreeBayes által hívott hamis variánsok aránya relatíve stabil volt. Indelek esetén a lefedettség növekedésével a SAMtools egyre nagyobb arányban hívott hamis variánsokat az illesztőprogramtól függetlenül, és jelentősen nagyobb, $2-4$-szeres hibaarányt mutatott a többi módszerhez képest.

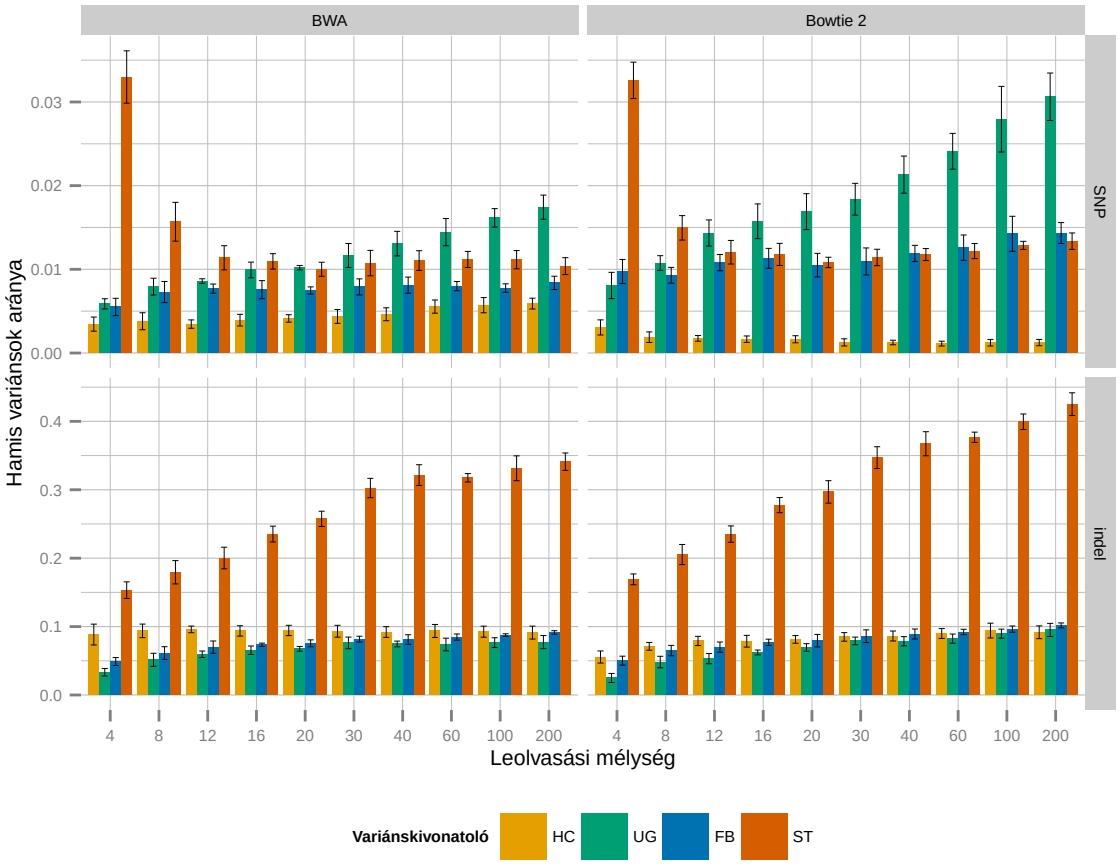
Az illesztőprogram hatása Általánosan elmondható, hogy az egyes variánskivonatoló módszerek több valódi variánst hívtak a BWA, mint a Bowtie 2 illesztések használatakor,

DOI:10.14753/SE.2017.1938



6. ábra. **Az egyedi variánskivonatolási módszerek szenzitivitása a szimulált adatokon.** Az oszlopdiagram az egyes variánskivonatolási módszerek szűretlen eredményeinek szenzitivitását mutatja be az SNP-k (**A**) és indelek (**B**) esetén. A felső sorban szereplő értékek a BWA, az alsó sorban szereplő értékek pedig a Bowtie 2 illesztéseken alapulnak. Az oszlopok a különböző lefedettségek esetén mért szenzitivitásértékeket jelenítik meg. A hibák a szenzitivitás becslésének 95%-os konfidencia-intervallumát mutatják. A szimulált adatok leírását lásd a Módszerek fejezetben. Rövidítések: FB = FreeBayes, HC = HaplotypeCaller, ST = SAMtools, UG = UnifiedGenotyper

DOI:10.14753/SE.2017.1938



7. ábra. **Az egyedi variánskivonatolási módszerek által hamisan hívott variánsok aránya a szimulált adatokon.** Az oszlopdiagram az egyedi variánskivonatolási módszerek által hamisan hívott variánsok arányát ábrázolja az SNP-k (felső sor) és indelek (alsó sor) esetén. A bal oldalon szereplő értékek a BWA, a jobb oldalon szereplő értékek pedig a Bowtie 2 illesztéseken alapulnak. Az oszlopok a különböző lefedettségek esetén mért értékeket jelenítik meg. A hibák a hibás felfedezési arány becslésének 95%-os konfidencia-intervallumát mutatják. A szimulált adatok leírását lásd a Módszerek fejezetben. Rövidítések: FB = FreeBayes, HC = HaplotypeCaller, ST = SAMtools, UG = UnifiedGenotyper

és a különbség statisztikailag is szignifikáns volt (lásd 4. táblázat). A legnagyobb különbséget a HaplotypeCaller esetén tapasztaltuk: az átlagos különbség SNP-k esetén $0{,}057$, indelek esetén pedig $0{,}053$ volt. Ezzel szemben a HaplotypeCaller precizitása a BWA illesztéseket használva szignifikánsan kisebb volt, mint a Bowtie 2 illesztések használatakor. A többi kivonatoló módszer esetén a szenzitivitás különbsége általában kisebb volt ($0{,}031$-$0{,}042$ között), viszont a mind a szenzitivitás, mind pedig a precizitás szignifikánsan nagyobb volt a BWA-t használva.

DOI:10.14753/SE.2017.1938

4. táblázat. **A BWA és a Bowtie 2 illesztőprogramok hatása az egyedi variánskivonatoló módszerek szenzitivitására és precizitására a szimulált adatok esetén.** Rövidítések: CI = konfidencia-intervallum

Variáns-típus	Variánskivonatoló módszer	Szenzitivitás			Precizitás		
		Átlagos különbség[1]	95% CI	p-érték[2]	Átlagos különbség[1]	95% CI	p-érték[2]
SNP	HaplotypeCaller	0,057	0,056-0,058	3,01E-57	-0,003	-0,003--0,002	2,19E-18
	UnifiedGenotyper	0,04	0,039-0,041	6,71E-57	0,007	0,006-0,008	1,40E-17
	FreeBayes	0,031	0,029-0,033	1,49E-33	0,004	0,003-0,005	2,29E-20
	SAMtools	0,034	0,032-0,035	7,45E-44	0,001	0-0,001	2,01E-04
Indel	HaplotypeCaller	0,053	0,05-0,055	5,55E-40	-0,012	-0,015--0,008	1,63E-08
	UnifiedGenotyper	0,034	0,03-0,037	3,30E-26	0,003	0-0,006	8,25E-02
	FreeBayes	0,042	0,04-0,044	2,30E-38	0,005	0,003-0,007	4,16E-07
	SAMtools	0,032	0,03-0,034	1,06E-36	0,046	0,04-0,052	8,89E-21

[1] A BWA használatával elért teljesítmény a Bowtie 2 illesztőprogramhoz viszonyítva

[2] Párosított t-teszt p-értéke

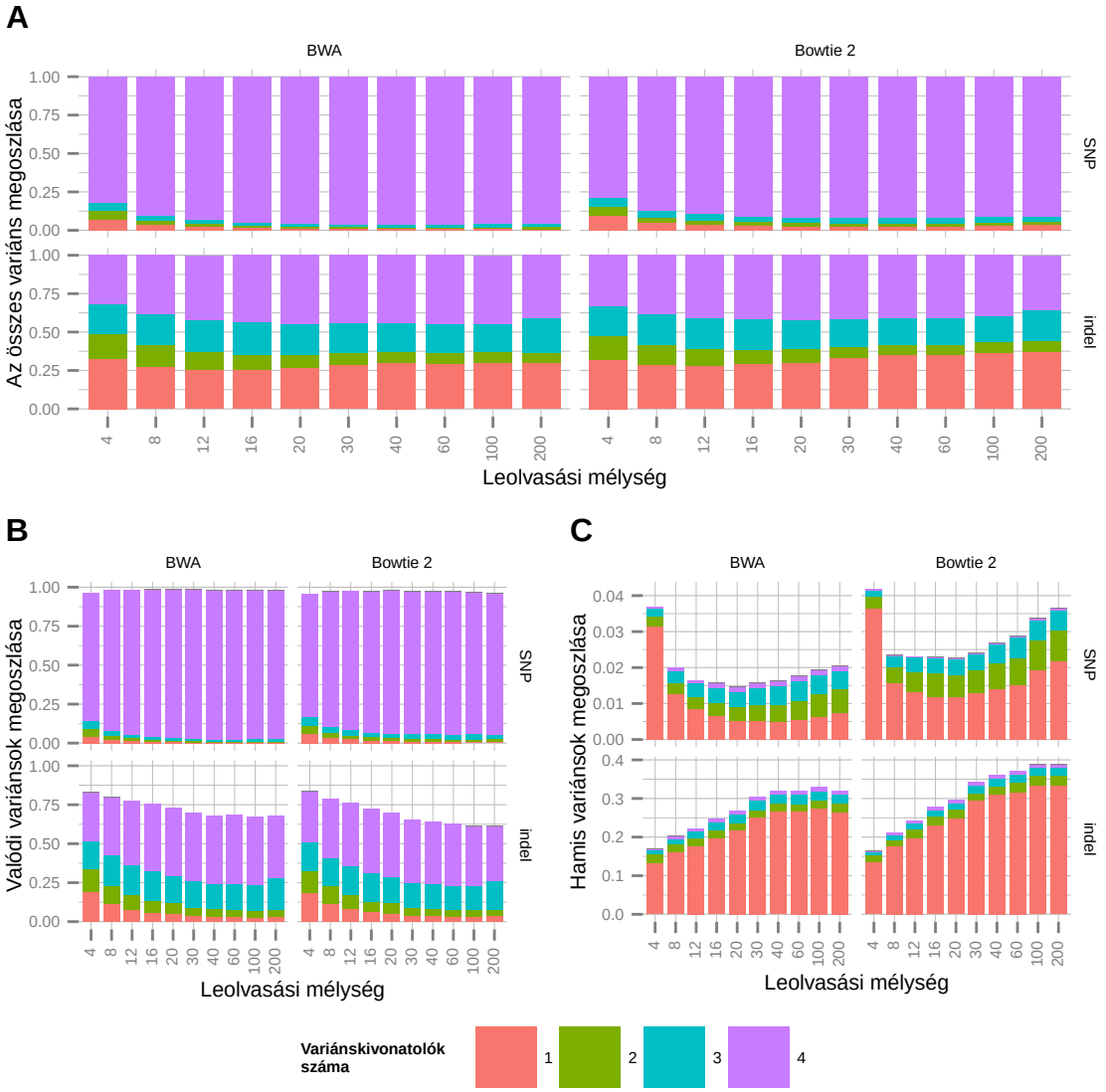
4.1.2. Variánskivonatolási módszerek konkordanciája

A variánskivonatolási módszerek teljesítményének összehasonlítása mellett az együtt járásukat (más néven konkordanciájukat) is kiértékeltük. Ez különösen fontos az újonnan kifejlesztett VariantMetaCaller módszer szempontjából, ugyanis ennek működése a kombinálandó módszerek konkordanciáján és komplementaritásán alapul.

Először kiszámítottuk, hogy az egyes variánsokat hány kivonatoló módszer találta meg. A mind a négy módszer szerint megtalált – teljesen konkordáns – variánsok aránya jóval nagyobb volt SNP-k, mint indelek esetén (lásd 8. ábra). A teljesen konkordáns variánsok aránya SNP-k esetén az illesztőprogramtól függően az alacsony lefedettségek esetén 78%-80% volt; ez a leolvasási mélység növekedésével 90%-95%-ra nőtt. Ezzel párhuzamosan az egyetlen módszerrel hívott variánsok aránya 7%-10%-ról 1%-2%-ra csökkent a lefedettség növekedésével. Alacsony leolvasási mélység esetén az egyetlen módszerrel hívott variánsok aránya volt a második legnagyobb, de a lefedettség növekedésével ennek az aránya lett a legkisebb.

Indelek esetén az egyes módszerek eredménye jóval nagyobb mértékben eltért egymástól, így a variánskivonatolók konkordanciája kisebb volt, mint SNP-k esetén. A leolvasási mélységtől függetlenül kevesebb mint a variánsok fele volt teljesen konkordáns, ugyanakkor az egy módszerrel hívott variánsok aránya minden esetben magasabb volt,

DOI:10.14753/SE.2017.1938



8. ábra. **A variánskivonatoló módszerek konkordanciája az összes hívott variáns, illetve a csak valódi vagy csak hamis variánsok esetén.** Az oszlopdiagram azt ábrázolja, hogy a négy egyedi variánskivonatoló módszer eredménye alapján – adott leolvasási mélység és illesztés mellett – hogyan oszlik meg a pontosan egy, kettő, három vagy négy módszerrel megtalált variánsok aránya az összes (**A**), a valódi (**B**), illetve a hamis (**C**) variánsok esetén. Mindhárom részábrán a felső sor az SNP-kre, az alsó sor az indelekre vonatkozik; a bal oldali oszlop eredményei a BWA illesztésen, a jobb oldali oszlop eredményei a Bowtie 2 illesztésen alapulnak.

mint 25%, és ez az arány volt a második legnagyobb.

A variánskivonatolók konkordanciája közepes lefedettség felett a leolvasási mélység növekedésével enyhén csökkent mind SNP-k, mind indelek esetén.

Ezután a konkordancia arányokat kiszámítottuk csak a valódi, illetve csak a hamis variánsokra szűkítve is. A 8B., illetve 8C. ábrán láthatjuk, hogy hogyan oszlik meg az egy, kettő, három vagy négy módszerrel megtalált valódi, illetve hamis variánsok aránya. A valódi variánsok esetén a teljesen konkordáns, azaz négy módszerrel is megtalált variánsok aránya általában a legmagasabb, míg a csak egy módszerrel hívott variánsok aránya a legalacsonyabb lefedettségek kivételével általában a legkisebb volt (lásd 8B. ábra). Ezzel párhuzamosan a hamis variánsok aránya az egy módszerrel hívott variánsok körében volt a legmagasabb, és elhanyagolható volt (SNP: $< 0{,}01\%$, indel: $< 0{,}1\%$) a négy módszerrel hívott variánsok esetén (lásd 8C. ábra).

A hamisan hívott variánsok aránya egy nagyságrenddel nagyobb volt indelek mint SNP-k esetén. Mindkét variánstípus esetén a hibás variánsok aránya emelkedett a leolvasási mélység növekedésével (SNP-k esetén $20\times$ lefedettség fölött). Nagy leolvasási mélységnél ez az arány az illesztőprogramtól függően kb. 2%-$3{,}6\%$ volt SNP-k és 30%-39% volt indelek esetén.

Végül kiszámítottuk a valódi variánsok arányát a pontosan egy, illetve négy módszer által hívott variánsok körében (lásd 9. ábra). A valódi variánsok aránya általában magas volt a mind a négy módszerrel megtalált variánsok körében mind az SNP-k ($> 99{,}83\%$ BWA és $> 99{,}94\%$ Bowtie 2 illesztéssel), mind az indelek ($> 97{,}6\%$) esetén. Ezzel szemben a valódi variánsok aránya az egyetlen módszerrel megtalált variánsok között jelentősen kisebb volt mindkét variánstípus esetén, és általában csökkent a lefedettség növekedésével (SNP-kre: $< 50\%$ $30\times$ lefedettség felett, indelek esetén: $< 15\%$ $30\times$ lefedettség felett).

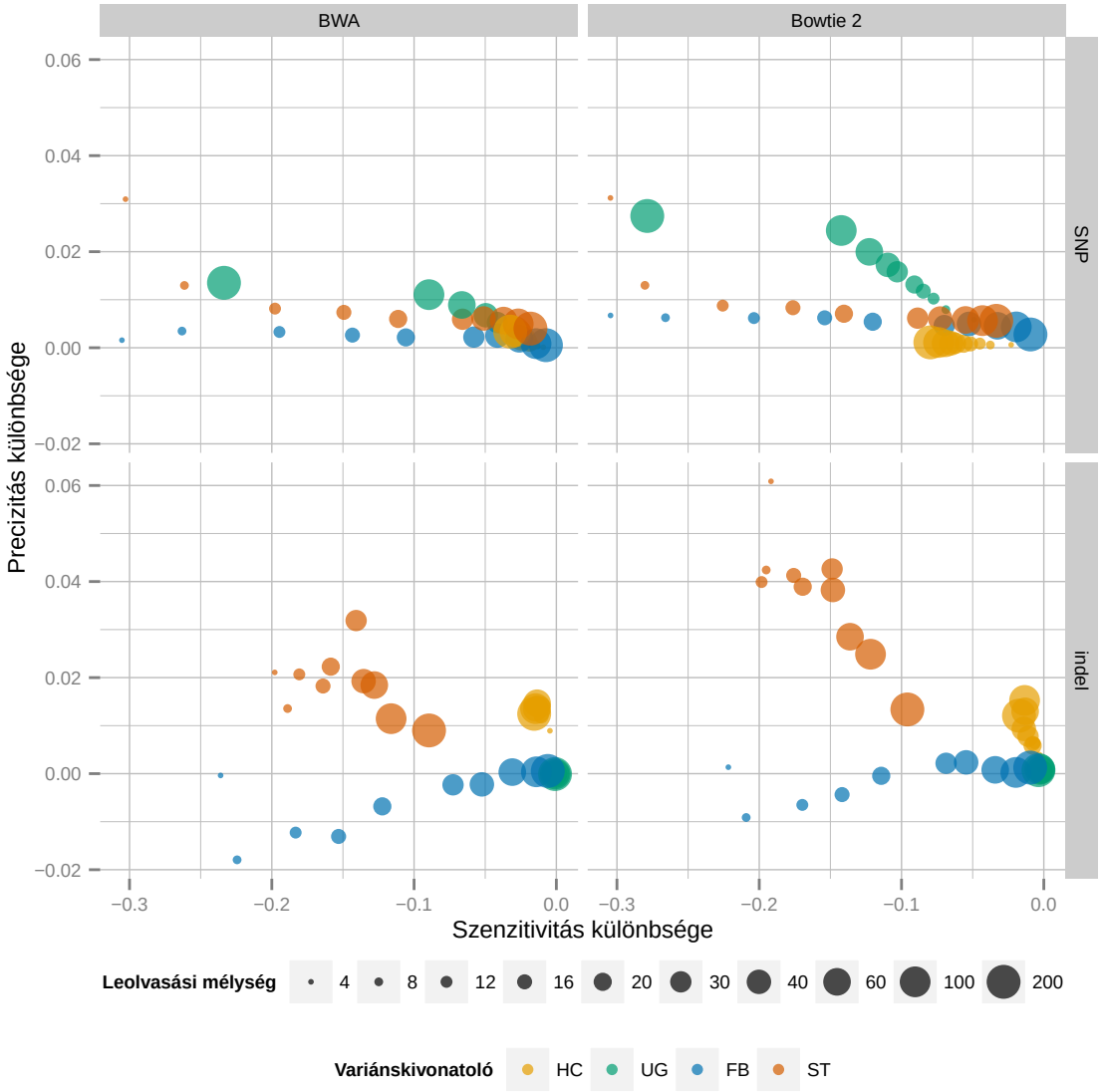
4.1.3. A manuális szűrők hatása a szenzitivitásra és a precizitásra

A variánskivonatolás precizitásának javítása érdekében a jelenlegi, széles körűen használt ajánlásoknak megfelelően elvégeztük a variánsok manuális szűrését (lásd Módszerek), észben tartva, hogy ezek nem feltétlenül jelentenek optimális megoldást. Megjegyezzük, hogy a szűretlen variánsok precizitása általában eleve magas volt (különösen SNP-k ese-

DOI:10.14753/SE.2017.1938

9. ábra. **A valódi variánsok aránya a pontosan egy, illetve négy módszer által hívott variánsok körében.** Az oszlopdiagram a valódi variánsok arányát mutatja a pontosan egy (piros), illetve négy (kék) módszer által hívott variánsok körében, különböző leolvasási mélységek és illesztőprogramok esetén. A hibák az arányok 95%-os konfidencia-intervallumát mutatják az öt mintacsoport eredményei alapján. Felső sor: SNP-k, alsó sor: indelek, bal oszlop: BWA illesztés, jobb oszlop: Bowtie 2 illesztés

tén), így a manuális szűrés várható precizitásnövelő hatása alacsony volt.



10. ábra. **A manuális szűrés hatása az egyedi variánskivonatolók eredményének szenzitivitására és specificitására.** A manuális szűréseket a jelenlegi ajánlásoknak megfelelően elvégeztük az egyedi variánskivonatolók eredményén. Az ábra pontjai a szenzitivitás és a precizitás átlagos megváltozását reprezentálják, ahol az átlagot az öt különböző mintacsoportok eredménye alapján számítottuk ki. A pontok mérete arányos a leolvasási mélységgel. Felső sor: SNP-k, alsó sor: indelek, bal oszlop: BWA illesztés, jobb oszlop: Bowtie 2 illesztés. Rövidítések: FB = FreeBayes, HC = HaplotypeCaller, ST = SAMtools, UG = UnifiedGenotyper

A manuális szűrés hatása jelentősen különbözött az egyes variánskivonatolási módszerek esetén (lásd 10. ábra). SNP-k esetén a szűrés hatása a leolvasási mélységtől függetlenül viszonylag alacsony volt a HaplotypeCaller eredményeire. A UnifiedGenotyper esetén a szenzitivitás jelentős mértékben csökkent a lefedettség növekedésével, míg a

precizitás csak kis mértékben emelkedett. A SAMtools és a FreeBayes esetén a manuális szűrés hatása hasonló karakterisztikát mutatott: a precizitás nagyon kis mértékben növekedett, a szenzitivitás pedig a leolvasási mélység növekedésével egyre nagyobb mértékben csökkent.

Indelek esetén a manuális szűrési stratégia közepesen jól teljesített a SAMtools és a HaplotypeCaller esetén, bár a szenzitivitás csökkenése általában felülmúlta a precizitás növekedését a SAMtools esetén. A szűrés a UnifiedGenotyper eredményeire csak nagyon kis hatással volt, a FreeBayes esetén pedig a hatás ellentétes volt az elvárttal, ugyanis a precizitás is csökkent.

4.2. Variánskivonatolók kombinálása: VariantMetaCaller

Az általam kifejlesztett VariantMetaCaller program több variánskivonatoló módszer eredményét kombinálja, mely során kihasználja az egyes módszerek konkordanciáját és komplementaritását. A szűretlen variánshívások egyesítése után a program minden egyes bemenetéül szolgáló módszerhez készít egy adathalmazt az adott kivonatoló, illetve a VariantMetaCaller által előállított annotációs adatok alapján. Ezeken az adathalmazokon a program variánstípusonként egy-egy SVM-t tanít, amely során a teljesen konkordáns (négy módszerrel is megtalált), illetve az egyetlen módszerrel hívott variánsok szolgálnak pozitív, illetve negatív tanítópéldaként. Ezáltal a VariantMetaCaller megpróbálja szétválasztani a valódi variánsokat a hamisan hívottaktól. A program minden egyes variánshoz kiszámít egy végső pontszámot, ami az adott variáns valódiságának a valószínűségét becsüli (lásd Módszerek).

A VariantMetaCallerrel a korábban bemutatott variánskivonatolási munkafolyamatokat kombináltam. Az illesztőprogramok hatását függetlenül értékeltem, azaz a kombináció során a négy variánskivonatoló program vagy a BWA vagy a Bowtie 2 illesztéseken alapult. A továbbiakban bemutatom a módszer teljesítőképességét először szimulált, majd valós adatok felhasználásával.

4.2.1. A VariantMetaCaller teljesítménye a szimulált adatokon

Az egyes módszerek teljesítményét ún. precizitás-szenzitivitás görbék segítségével jellemezzük. Ehhez először sorrendeztük az egyes variánskivonatolók eredményeit az általuk

jósolt variánsminőség szerint, illetve a VariantMetaCaller eredményét a variánsok becsült valószínűsége szerint. Ezután kiszámítottuk a szenzitivitást és a precizitást minden egyes variánskivonatoló esetén minden lehetséges variánsminőség-küszöbérték mellett, illetve a VariantMetaCaller esetén minden lehetséges variáns valószínűség mellett[11], végül ábrázoltuk az így kiszámított precizitást a szenzitivitás függvényében (lásd 11. ábra). A kiértékelések során a manuálisan leszűrt variánsokat vettük alapul. Amint az jól látható, a VariantMetaCaller a leolvasási mélységtől, az illesztőtől és a variánstípustól függetlenül felülmúlta az egyedi variánskivonatolókat a precizitás–szenzitivitás térben, azaz nagyobb precizitást ért el minden szenzitivitási szinten mint a bemenetéül szolgáló módszerek. A VariantMetaCaller maximális szenzitivitása ugyancsak magasabb volt, és a precizitás élesen csökkent a nagy szenzitivitás értékeknél.

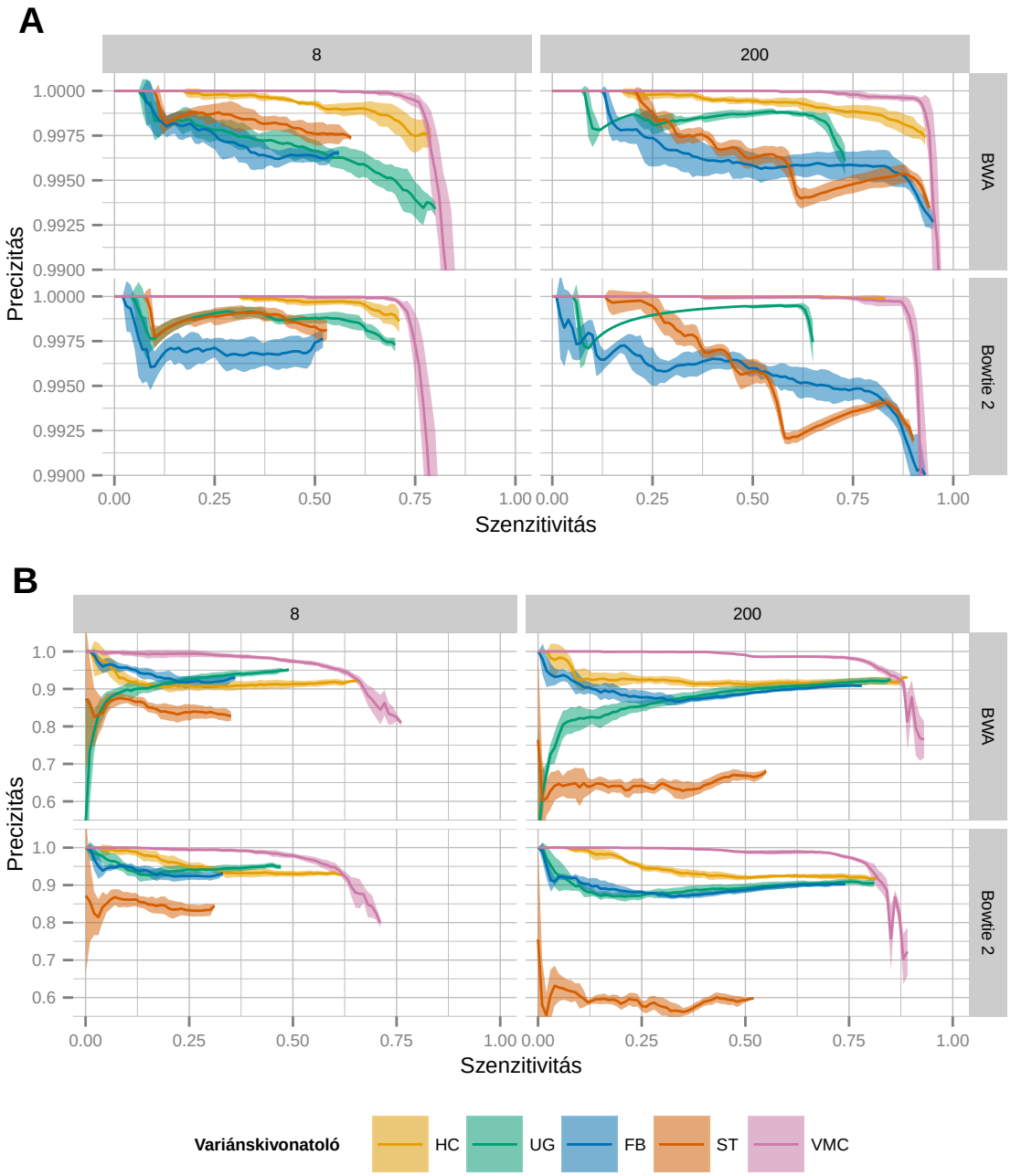
Ezt követően kiszámítottuk a precizitás-szenzitivitás görbe alatti területet (area under the precision–recall curves, AUPRC), amely azt mutatja meg, hogy egy adott mutató mennyire képes elkülöníteni a valódi variánsokat a hamisaktól. A VariantMetaCaller AUPRC pontszáma a lefedettségtől, az illesztőtől és a variánstípustól függetlenül minden esetben magasabb volt, mint az egyedi variánskivonatolók AUPRC értéke (lásd 12. ábra), és a különbség statisztikailag erősen szignifikáns volt (a Bonferroni-korrigált p-értékek maximuma SNP-k esetén: $0{,}002$, indelek esetén: $0{,}006$; párosított t-teszt a mintacsoportok között, lásd Függelék: 3. táblázat). A várakozásoknak megfelelően az AUPRC értéke a leolvasási mélység növekedésével nőtt, és SNP-k esetén magasabb volt mint indelek esetén.

SNP-k esetén a legmagasabb AUPRC pontszámmal rendelkező egyedi variánskivonatoló a lefedettségtől és az illesztőprogramtól függően változott (lásd 12A. ábra). Alacsony leolvasási mélységek esetén a UnifiedGenotyper esetén volt a legmagasabb az AUPRC, de a lefedettség növekedésével ez lett a legalacsonyabb. Indelek esetén a lefedettségtől és az illesztőprogramtól függetlenül a HaplotypeCaller nyújtotta a legmagasabb AUPRC értéket az egyedi módszerek közül (lásd 12B. ábra).

Az illesztőprogram hatása Az egyedi variánskivonatolók általában magasabb maximális szenzitivitást értek el a BWA, mint a Bowtie 2 illesztések használata esetén. Ebből

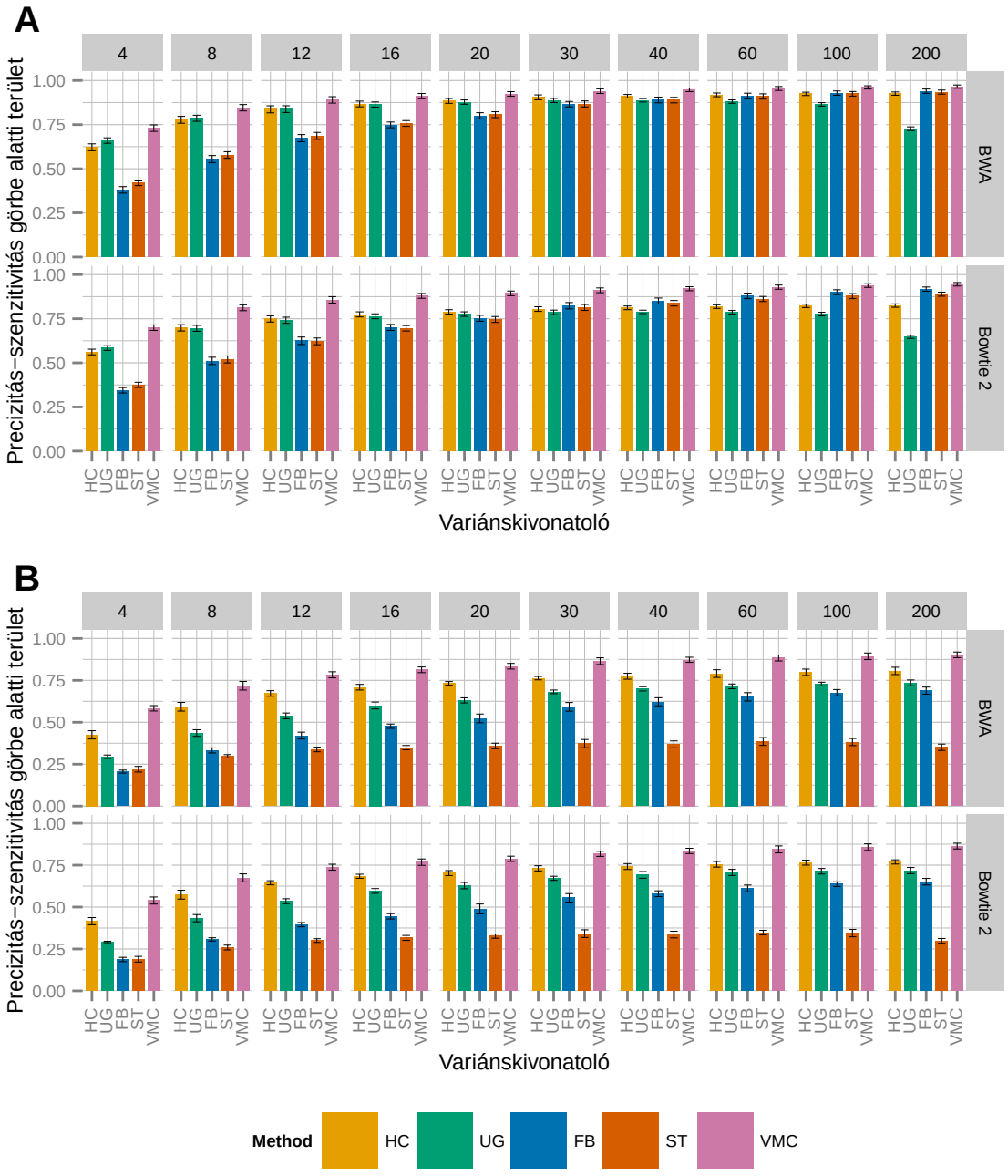
[11] Azaz minden küszöbértékre kiszámítottuk, hogy az afeletti variánsok hányad része valódi (precizitás), illetve hogy az összes valódi variáns hányad része van a küszöb felett (szenzitivitás).

DOI:10.14753/SE.2017.1938



11. ábra. **Precizitás-szenzitivitás görbék kiválasztott lefedettségek esetén a szimulált adatokon.** Az egyes manuálisan leszűrt variánskivonatolási módszerek, illetve a szűretlen bemeneteket kombináló VariantMetaCaller eredményére kiszámított precizitást ábrázoltuk a szenzitivitás függvényében minden lehetséges vágási küszöb mellett SNP-k (**A**) és indelek (**B**) esetén két kiválasztott leolvasási mélységre (bal oldal: $8\times$ és jobb oldal: $200\times$). Felső sor: BWA illesztés, alsó sor: Bowtie 2 illesztés. A sávok a becslés 95%-os konfidencia-intervallumát mutatják. Rövidítések: FB = FreeBayes, HC = HaplotypeCaller, ST = SAMtools, UG = UnifiedGenotyper, VMC = VariantMetaCaller

DOI:10.14753/SE.2017.1938



12. ábra. **Az egyes módszerek precizitás-szenzitivitás görbe alatti területe a szimulált adatokon.** Az oszlopdiagram az egyedi variánskivonatoló módszerek manuális szűrésének, illetve a VariantMetaCaller eredményének precizitás-szenzitivitás görbe alatti területét ábrázolja SNP-k (**A**) és indelek (**B**) esetén. A felső sorban szereplő értékek a BWA, az alsó sorban szereplő értékek pedig a Bowtie 2 illesztéseken alapulnak. Az oszlopok a különböző lefedettségek esetén kiszámított értékeket jelenítik meg. A hibák az AUPRC értékének 95%-os konfidencia-intervallumát mutatják. A szimulált adathalmaz leírását lásd a Módszerek fejezetben. Rövidítések: FB = FreeBayes, HC = HaplotypeCaller, ST = SAMtools, UG = UnifiedGenotyper, VMC = VariantMetaCaller

eredően a VariantMetaCaller által elért teljesítménymutatók a BWA illesztésekre alapozva szintén jobbak voltak (lásd 5. táblázat és Függelék: 4. táblázat). SNP-k esetén a VariantMetaCaller által elért maximális szenzitivitás átlagos különbsége a BWA, illetve a Bowtie 2 illesztésekre alapozott eredmények között $0{,}028$ volt (95%CI: $0{,}027$-$0{,}029$). Bár ez a különbség nem tűnik jelentősnek, a jelenlegi kísérleti beállítások mellett 1% szenzitivitásbeli különbség kb. 144-el több valódi variáns megtalálását jelenti. Indelek esetén a maximális szenzitivitás átlagos különbsége még nagyobb, $0{,}044$ volt (95%CI: $0{,}042$-$0{,}046$). A jelenlegi beállítások mellett 1% szenzitivitás különbség kb. 19-el több valódi indel hívását jelenti. A VariantMetaCaller által eredményezett teljes variánslista precizitása és az AUPRC értékek is magasabbak volt a BWA használatával minden leolvasási mélység és mindkét variánstípus esetén (lásd 5. táblázat).

5. táblázat. **Az illesztőprogram megválasztásának hatása a VariantMetaCaller különböző teljesítménymutatóira.** Rövidítések: CI = konfidencia-intervallum

	SNP			Indel		
Teljesítménymutató	Átlagos különbség[1]	95% CI	p-érték[2]	Átlagos különbség[1]	95% CI	p-érték[2]
Maximális szenzitivitás	0,028	0,027-0,029	$< 2{,}2*10^{-16}$	0,044	0,042-0,046	$< 2{,}2*10^{-16}$
Precizitás a teljes variánslistára	0,009	0,007-0,01	$< 2{,}2*10^{-16}$	0,039	0,025-0,053	$8{,}6*10^{-7}$
AUPRC	0,028	0,027-0,029	$< 2{,}2*10^{-16}$	0,042	0,040-0,044	$< 2{,}2*10^{-16}$

[1] A BWA használatával elért teljesítmény a Bowtie 2 illesztőprogramhoz viszonyítva

[2] Párosított t-teszt p-értéke

Különböző méretű genomi régiókon elért eredmények A VariantMetaCaller kisebb genomi célrégiókon való használhatóságának demonstrálása érdekében a teljes kromoszómát kisebb méretű régiókra szűkítettük. A célrégiók méretét úgy határoztuk meg, hogy hasonlóak legyenek a tipikusan jelenleg használatos célzott génpanelek méretéhez: a teljes exonikus tartomány 100 kb, 200 kb, 300 kb és 500 kb hosszúságú volt. Minden régióméretből tíz nem átfedő tartományt választottunk ki, és minden egyes régióra lefuttattuk az elemzést a korábbihoz hasonlóan. A variánsok átlagos számát az egyes régióméretek esetén a Függelék: 5. táblázatban láthatjuk. Ebben az esetben is elmondható, hogy a VariantMetaCaller nyújtotta a legmagasabb AUPRC értéket az egyes módszerek közül, a célrégió méretétől, a leolvasási mélységtől, az illesztőprogramtól és a variáns típusától függetlenül (lásd 13. ábra). A különbség a VariantMetaCaller és az egyedi vari-

ánskivonatolók között statisztikailag is szignifikáns volt (a Bonferroni-korrigált p-értékek maximuma SNP-k esetén: $0{,}041$, indelek esetén: $0{,}005$; párosított t-teszt a mintacsoportok és az adott régióméretbe tartozó különböző régiókon elért eredmények alapján). A teljes kromoszómán bemutatott eredményekhez hasonlóan az SNP-ken elért AUPRC pontszámok magasabbak voltak, mint indelek esetén.

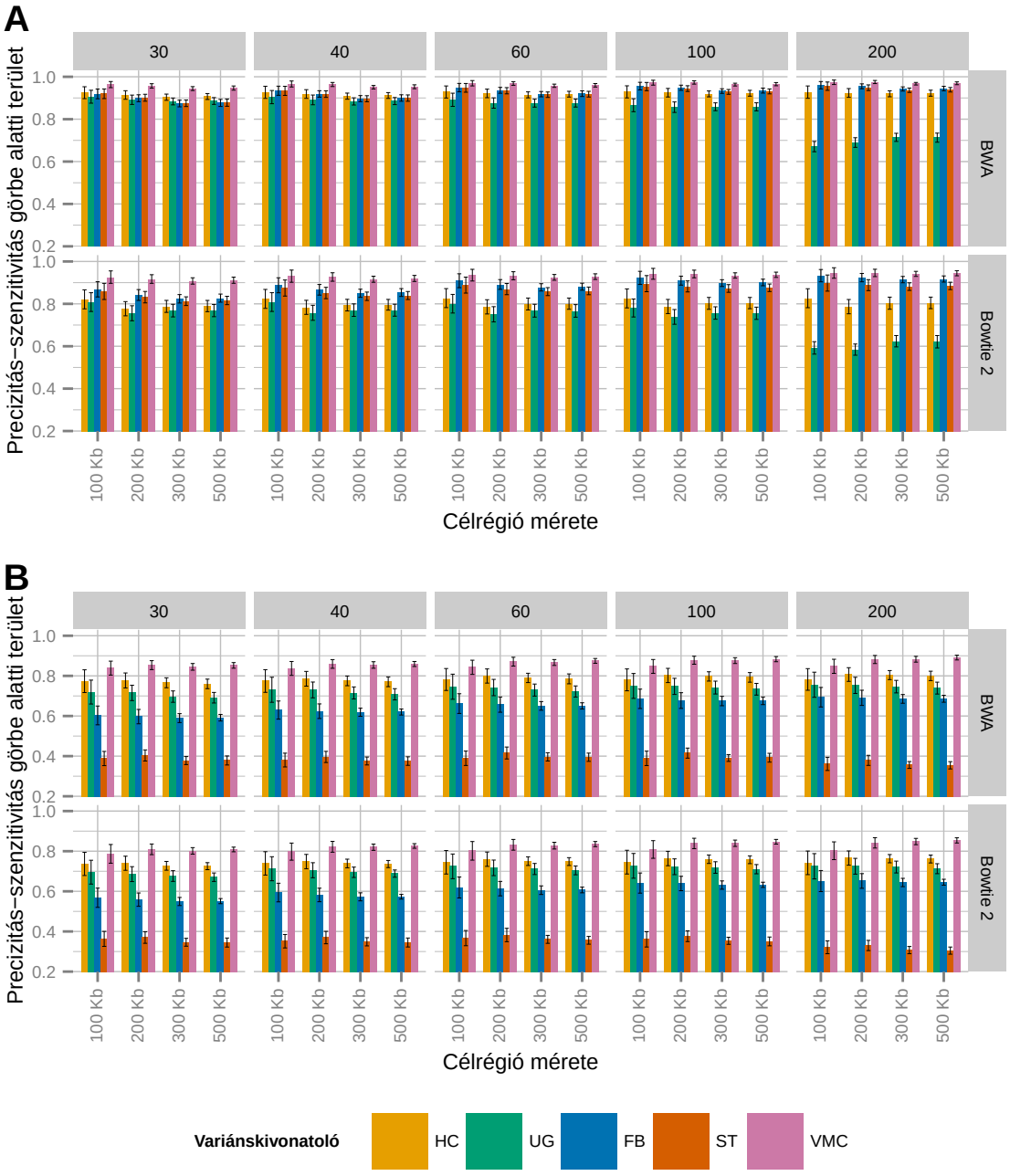
4.2.2. A VariantMetaCaller teljesítménye valós adatokon

A VariantMetaCaller teljesítményét valós szekvenálási adatokon is kiértékeltük. Ehhez egy publikusan elérhető szekvenálási adathalmazt használtunk, amely az NA12878-as kódszámú, rendkívül széleskörűen tanulmányozott emberi genom teljes szekvenálási adatait tartalmazza. Ehhez a genomhoz ugyanis rendelkezésre áll egy nagy megbízhatóságú, „platina" minőségű referencia variánslista is (*Platina referencia*), amelyet az *Illumina* cég állított elő. A minőségi szűrésen átesett leolvasásokat a korábbiakhoz hasonlóan a BWA–MEM és a Bowtie 2 programokkal felillesztettük a referencia szekvenciára, az illesztéseket leszűrtük a teljes exomra, majd bázisminőség-korrekciót végeztünk, és az indelek környékén újraillesztést végeztünk a GATK ajánlásainak megfelelően. Az átlagos leolvasási mélység mindkét illesztés esetén kb. $12\times$ volt. Végül, szintén a korábbiakhoz hasonlóan, a négy variánskivonatoló programmal meghatároztuk az SNP-ket és az indeleket (lásd Módszerek).

Az egyedi módszerek szűretlen variánshívásainak konkordanciája alacsony volt (lásd 6. táblázat). A mind a négy egyedi variánskivonatoló által hívott, teljesen konkordáns SNP-k aránya $88{,}8\%$ volt a BWA és $84{,}27\%$ a Bowtie 2 illesztések használata esetén. A pontosan egy módszerrel megtalált SNP-k aránya $3{,}48\%$ volt a BWA, illetve ennél is magasabb, $8{,}83\%$ volt Bowtie 2 illesztéseket használva. A konkordancia arányok jóval alacsonyabbak voltak indelek esetén: kevesebb mint a variánsok felét hívta mind a négy kivonatoló, és az egyetlen módszerrel megtalált indelek aránya $21{,}36\%$ (BWA), illetve $20{,}43\%$ (Bowtie 2) volt.

A négy variánskivonatoló módszer szűretlen eredményeit kombináltuk a VariantMetaCaller segítségével. Az SVM tanítása során ebben az esetben is a teljesen konkordáns variánsok szolgáltak pozitív és az egyetlen módszer által megtalált variánsok szolgáltak negatív tanítómintaként. Az annotációs adatok fuzionálása után a VariantMetaCaller

DOI:10.14753/SE.2017.1938



13. ábra. **Az egyes módszerek precizitás-szenzitivitás görbe alatti területe különböző méretű genomi régiók esetén.** A teljes kromoszómát kisebb méretű régiókra szűkítettük, a célrégiók mérete 100 kb, 200 kb, 300 kb és 500 kb hosszúságú volt. Minden régióméretből tíz nem átfedő tartományt választottunk ki, és minden egyes régióra lefuttattuk az elemzést. Az oszlopdiagram az egyedi variánskivonatoló módszerek manuális szűrésének, illetve a VariantMetaCaller eredményének precizitás-szenzitivitás görbe alatti területét ábrázolja SNP-k (**A**) és indelek (**B**) esetén. (folytatás a következő oldalon)

13. ábra. (folytatás) A felső sorban szereplő értékek a BWA, az alsó sorban szereplő értékek pedig a Bowtie 2 illesztéseken alapulnak. Az oszlopok a különböző lefedettségek esetén kiszámított értékeket jelenítik meg. A hibák az AUPRC értékének 95%-os konfidencia-intervallumát mutatják. A szimulált adathalmaz leírását lásd a Módszerek fejezetben. Rövidítések: FB = FreeBayes, HC = HaplotypeCaller, ST = SAMtools, UG = UnifiedGenotyper, VMC = VariantMetaCaller

6. táblázat. **Az egyedi variánskivonatoló módszerek különféle kombinációi által hívott szűretlen variánsok száma a valós szekvenálási adatokon.** Rövidítések: FB = FreeBayes, HC = HaplotypeCaller, UG = UnifiedGenotyper, ST = SAMtools

Variánskivonatoló módszer				SNP				Indel			
				BWA		Bowtie 2		BWA		Bowtie 2	
HC	UG	FB	ST	Variánsok száma	Az összes variáns hány százaléka	Variánsok száma	Az összes variáns hány százaléka	Variánsok száma	Az összes variáns hány százaléka	Variánsok száma	Az összes variáns hány százaléka
+	+	+	+	48297	88,81%	43659	84,27%	2918	46,73%	2958	49,58%
+	+	+	-	440	0,81%	297	0,57%	532	8,52%	544	9,12%
+	+	-	+	432	0,79%	273	0,53%	76	1,22%	77	1,29%
+	-	+	+	296	0,54%	215	0,42%	638	10,22%	466	7,81%
-	+	+	+	1332	2,45%	1078	2,08%	53	0,85%	66	1,11%
+	+	-	-	222	0,41%	115	0,22%	18	0,29%	24	0,40%
+	-	+	-	82	0,15%	45	0,09%	254	4,07%	186	3,12%
+	-	-	+	57	0,10%	36	0,07%	235	3,76%	130	2,18%
-	+	+	-	367	0,67%	403	0,78%	56	0,90%	99	1,66%
-	+	-	+	164	0,30%	90	0,17%	23	0,37%	41	0,69%
-	-	+	+	798	1,47%	1021	1,97%	108	1,73%	156	2,61%
+	-	-	-	499	0,92%	178	0,34%	578	9,26%	342	5,73%
-	+	-	-	329	0,60%	330	0,64%	11	0,18%	40	0,67%
-	-	+	-	781	1,44%	3596	6,94%	256	4,10%	371	6,22%
-	-	-	+	285	0,52%	471	0,91%	489	7,83%	466	7,81%

megbecsülte minden variáns valódiságának valószínűségét. Az egyedi módszerek variánshívásait a BAYSIC [61] nevű programmal is kombináltuk, ami egy rejtett változós elemzést használva megbecsüli a variánsok valószínűségét. Ezen felül lefuttattuk a GATK variánsminőség-kalibrációs programját (VQSR) is a HaploytypeCaller és a UnifiedGenotyper által hívott variánsokra. A VQSR egy kevert Gauss-modellt[12] (Gaussian mixture model) illeszt a kvantitatív annotációs adatokra, és szintén variánsvalószínűségeket becsül, melyhez referenciaként nagy megbízhatóságú variánskészletet használ. Végezetül a variánshívásokat leszűkítettük a Platina referencia megbízható tartományára.

Kiszámítottuk az egyedi variánskivonatolók manuális szűréseinek, illetve a VQSR, a BAYSIC és a VariantMetaCaller eredményeinek szenzitivitását és precizitását. A Va-

[12]Kevert Gauss-modell: Egy olyan valószínűségi modell, amely azt feltételezi, hogy a mért adatok véges számú, különböző paraméterű Gauss (normális) eloszlások keverékéből származnak.

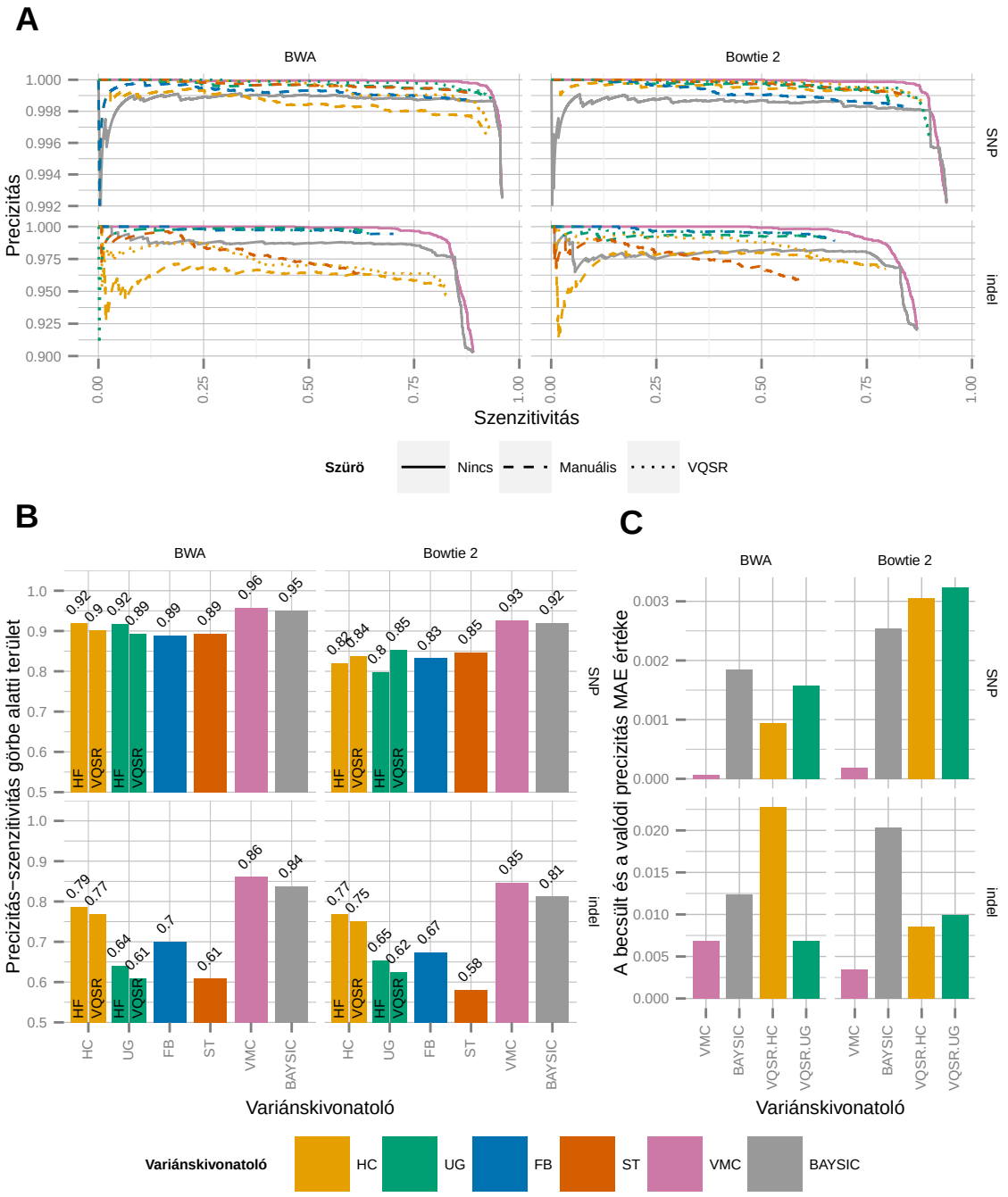
riantMetaCaller – a szimulált adatok esetén tapasztalt eredményekhez hasonlóan – az illesztőtől és a variánstípustól függetlenül általánosságban felülmúlta az egyedi variánskivonatolókat a precizitás–szenzitivitás térben, azaz nagyobb precizitást ért el a szenzitivitási szintek legnagyobb részében mint a többi módszer (lásd 14A. ábra). Ugyanezt az eredményt tükrözi, hogy a VariantMetaCaller érte el a legnagyobb AUPRC értéket (lásd 14B. ábra).

Az egyedi variánskivonatolók egymáshoz képesti teljesítőképessége hasonló volt a szintetikus adatok esetén tapasztaltakhoz. SNP-k esetén a BWA illesztéseket használva a manuálisan leszűrt HaploytypeCaller és UnifiedGenotyper egymáshoz hasonló eredményt ért el (AUPRC: $0{,}92$), és jobban teljesítettek, mint a szintén egymáshoz hasonló eredményt elérő FreeBayes és SAMtools (AUPRC: $0{,}89$). Ugyanakkor a Bowtie 2 illesztéseket használva a SAMtools teljesítménye jobb volt (AUPRC: $0{,}85$), mint a többi egyedi kivonatolóé, és a UnifiedGenotyper bizonyult a legrosszabbnak (AUPRC: $0{,}8$). Indelek esetén az eredmények minőségileg tükrözték a szintetikus adatok esetén látottakat azzal a különbséggel, hogy a UnifiedGenotyper és a FreeBayes egymáshoz képesti teljesítménye fordított volt. A VQSR csak a Bowtie 2 illesztéseket használva és csak SNP-k esetén bizonyult jobbnak, mint a manuális szűrési stratégia. Ez valószínűleg annak köszönhető, hogy a VQSR pontosabb működéséhez a jelenleg használtnál nagyobb adathalmazra lenne szükség.

A fúziós módszerek teljesítőképességét a kromoszómákra leszűkített célrégiókra is megvizsgáltuk. A VariantMetaCaller az esetek legnagyobb részében jobb teljesítményt nyújtott mint a BAYSIC, és a különbség statisztikailag is erősen szignifikáns volt (lásd 7. táblázat).

A VariantMetaCaller, a BAYSIC és a VQSR is megbecsüli a variánsok valódiságának valószínűségét, ami elméletileg lehetővé teszi azt, hogy az adott módszert alkalmazó kutató a variánsok valószínűsége alapján becsült precizitást használja fel a variánsok szűrésére. A precizitás pontos jóslása azonban csak a valószínűségek pontos becslésével lehetséges, ezért a módszereket összehasonlítottuk a becslés jóságának szempontjából is. Ez a következőképpen történt: mindhárom módszer esetén a variánsokat a becsült valószínűségek szerint csökkenő sorrendbe rendeztük, majd a sorrend mentén minden variánsra kiszámítottuk a várható precizitást (azaz a sorrend elejétől az adott variánsig bezárólag a

DOI:10.14753/SE.2017.1938



14. ábra. **A VariantMetaCaller és az egyedi kivonatoló módszerek eredményeinek összehasonlítása a valós szekvenálási adatokon.** Az NA12878 kódszámú minta teljes genomi szekvenálásából származó leolvasásokat felillesztettük a humán genomra a BWA és a Bowtie 2 illesztőprogramokkal, majd az illesztéseket leszűrtük a teljes exoni célrégióra. Lefuttattuk a GATK HaplotypeCaller, GATK UnifiedGenotyper, FreeBayes és SAMtools variánskivonatolókat, majd a szűretlen variánsokat kombináltuk a VariantMetaCaller és a BAYSIC programokkal. Az egyedi kivonatolók eredményeit leszűrtük az általános manuális szűrési javaslatoknak megfelelően, illetve a GATK alapú variánskivonatolásokat a VQSR használatával is leszűrtük. (folytatás a következő oldalon)

14. ábra. (folytatás) Ezt követően minden variánskivonatolási eredményt leszűkítettünk a referencia variánsok megbízható tartományára. **A**. Az egyes módszerek precizitás-szenzitivitás görbéje SNP-k (felső sor) és indelek (alsó sor) esetén a BWA (bal oszlop), illetve a Bowtie 2 (jobb oszlop) illesztőprogramok eredménye alapján. **B**. Az egyes módszerek precizitás-szenzitivitás görbe alatti területe SNP-k (felső sor) és indelek (alsó sor) esetén a BWA (bal oszlop), illetve a Bowtie 2 (jobb oszlop) illesztőprogramok eredménye alapján. **C**. A különböző variánsvalószínűséget is becslő módszerek által jósolt precizitás és a valódi precizitás átlagos abszolút hibája. Rövidítések: FB = FreeBayes, HC = HaplotypeCaller, HF = manuális szűrés (hard filter), MAE = átlagos abszolút hiba (mean absolute error), ST = SAMtools, UG = UnifiedGenotyper, VMC = VariantMetaCaller, VQSR = variánsminőség-kalibráció (variant quality score recalibration)

7. táblázat. **A VariantMetaCaller és a BAYSIC teljesítményének különbsége a humán kromoszómákra szűkített adatokon**

Variáns-típus	Illesztő-program	Kromoszómák száma[1]	p-érték[2]
SNP	BWA	21	$5{,}96*10^{-7}$
	Bowtie 2	22	$2{,}15*10^{-5}$
Indel	BWA	19	$5{,}33*10^{-5}$
	Bowtie 2	21	$3{,}93*10^{-6}$

[1] Hány kromoszóma esetén volt nagyobb az AUPRC értéke a VariantMetaCaller esetén a BAYSIC-hez képest

[2] Párosított Wilcoxon-teszt p-értéke

valódi variánsok várható arányát a variánsok becsült valószínűsége alapján, lásd 6. képlet). Ezután meghatároztuk a sorrend mentén a valódi precizitást is (azaz a sorrend elejétől az adott variánsig bezárólag a valódi variánsok pontos arányát a referenciavariánsok alapján), majd kiszámítottuk a becsült és a valódi precizitás átlagos abszolút hibáját (mean absolute error, MAE), azaz a különbségek abszolút értékének összegét.

Általánosságban elmondható, hogy a MAE alacsony volt minden esetben, tehát az egyes módszerek jól közelítették a valódi precizitást, de a VariantMetaCaller nyújtotta a legpontosabb becslést SNP-ek esetén az illesztőprogramtól függetlenül, illetve indelek esetén a Bowtie 2 illesztőprogram használatával (lásd 14C. ábra). Indelekre a BWA illesztések alapján a VariantMetaCaller és a UnifiedGenotyper esetén tapasztalt becslési pontosság hasonló volt, de a UnifiedGenotypernek jelentősen kisebb volt a szenzitivitása.

DOI:10.14753/SE.2017.1938

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

4.3.1. A polimorfizmusok önálló hatása a túlélésre

Az elemzés során azt vizsgáltuk, hogy az egyik legfőbb gyógyszer-metabolizáló enzim, a CYP3A4, illetve az ezzel átfedő szubsztrátspecificitással rendelkező CYP3A5 génjeiben lévő kiválasztott polimorfizmusok hogyan befolyásolják a gyermekkori akut limfoid leukémia túlélését. Az elemzést 511 ALL-ben szenvedő gyermek adatain végeztük. A mintapopuláció alapadatait a Módszerek fejezetben található 2. táblázatban láthatjuk. A polimorfizmusok hatását kétféle túlélés-típusra vonatkozóan vizsgáltuk. A *teljes túlélés* elemzésekor a vizsgált esemény a páciens halála, az *eseménymentes túlélés* elemzése során pedig a páciens halála vagy a betegségbe való visszaesés (recidíva). A populációban a teljes túlélési arány $85{,}5\%$, az eseménymentes túlélési arány pedig 81% volt.

A *CYP3A4* és *CYP3A5* gének kiválasztott polimorfizmusainak genotípus- és allélfrekvencia adatai hasonlóak a különböző genetikai adatbázisokban (dbSNP Home Page, http://www.ncbi.nlm.nih.gov/SNP/ Hozzáférés: 2014.05.29. és The Pharmacogenomics Knowledge Base, https://www.pharmgkb.org/ Hozzáférés: 2014.05.29.) és más európai/kaukázusi populáción végzett vizsgálatokban látottakhoz [98, 99, 101, 114] (lásd 3. táblázat).

Először a log-rank teszt használatával megvizsgáltuk, hogy a kiválasztott SNP-k befolyásolják-e a teljes vagy az eseménymentes túlélést (lásd 8. táblázat). A *CYP3A4* rs2246709 SNP-je szignifikánsan asszociált a teljes túléléssel (p-érték: $0{,}0028$) és az eseménymentes túléléssel is (p-érték: $0{,}014$). Ezeket az eredményeket a bayesi relevancia-elemzés is megerősítette, ugyanis az rs2246709 nagy valószínűséggel erősen relevánsnak bizonyult az ötéves teljes (Pr: $0{,}85$) és eseménymentes túlélés (Pr: $0{,}55$) szempontjából is. A túlélési arányok változását az eltelt idő függvényében a 15. ábrán láthatjuk. A *CYP3A5* rs15524 SNP-je szintén szignifikánsan asszociált mind a teljes túléléssel (p-érték: $0{,}031$) mind pedig az eseménymentes túléléssel (p-érték: $0{,}00058$). Ez az erős hatás annak a következménye, hogy a ritka homozigóta genotípusú betegek a vártnál nagyobb arányban fordulnak elő a mintapopulációban, azonban az ilyen genotípussal rendelkező betegek száma valójában alacsony (2, illetve 3 beteg; lásd 8. táblázat). Ez az asszociáció klinikai-

lag releváns lehet, de a betegek alacsony száma miatt ezeket az eredményeket szükséges lenne validálni nagyobb populáción is.

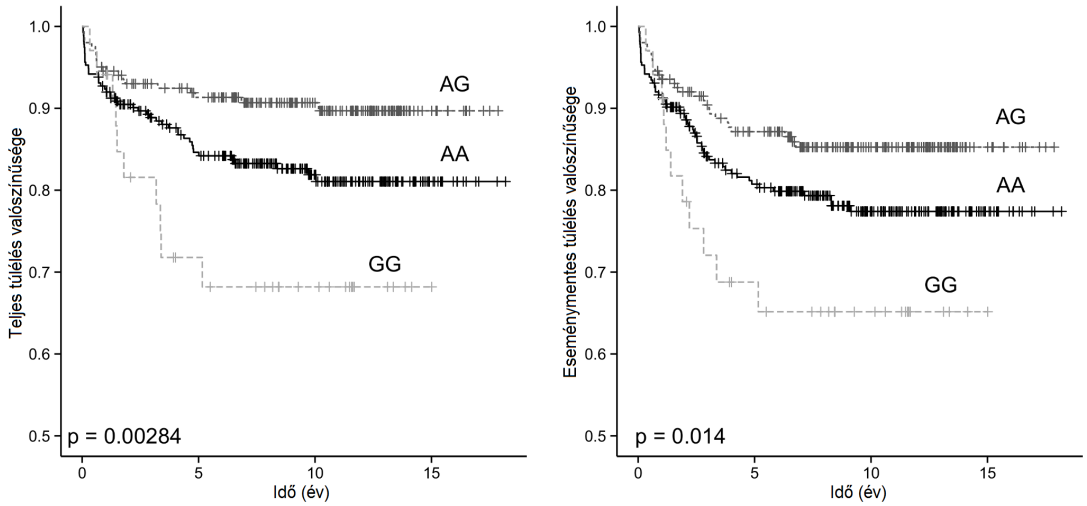
8. táblázat. **A log-rank teszt eredményei a *CYP3A4* és a *CYP3A5* polimorfizmusainak teljes- és eseménymentes túlélésre gyakorolt hatásának elemzése során.** Rövidítések: N = mintaszám

		Teljes túlélés				Eseménymentes túlélés			
Genotípus	**N**	**Megfigyelt**	**Elvárt**	**p-érték**	**Statisztikai erő**	**Megfigyelt**	**Elvárt**	**p-érték**	**Statisztikai erő**
rs15524 = AA	313	50	48,97			59	59,95		
rs15524 = AG	58	7	9,62	**0,031**	0,668	10	11,61	**0,00058**	0,838
rs15524 = GG	4	2	0,41			3	0,45		
rs776746 = GG	318	51	50,23			61	61,22		
rs776746 = GA	56	8	9,36	0,587	0,336	10	11,34	0,062	0,652
rs776746 = AA	3	1	0,41			2	0,45		
rs2404955 = GG	388	54	57,07			69	72,56		
rs2404955 = GA	108	18	16,7	0,247	0,505	23	20,93	0,396	0,43
rs2404955 = AA	9	3	1,23			3	1,51		
rs12333983 = TT	384	54	57,39			68	72,19		
rs12333983 = TA	105	18	16,35	0,249	0,506	23	20,29	0,359	0,451
rs12333983 = AA	9	3	1,26			3	1,52		
rs2242480 = CC	295	45	45,18			53	55,36		
rs2242480 = CT	83	14	13,12	0,684	0,086	19	15,8	0,45	0,201
rs2242480 = TT	4	0	0,7			0	0,85		
rs4646437 = GG	287	43	44,63			49	54,57		
rs4646437 = GA	91	16	14,44	0,89	0,194	23	17,31	0,294	0,454
rs4646437 = AA	6	1	0,93			1	1,11		
rs2246709 = AA	275	46	39,53			56	50,04		
rs2246709 = AG	202	19	30,81	**0,0028**	0,944	28	39,15	**0,014**	0,864
rs2246709 = GG	34	10	4,66			11	5,81		
rs35599367 = GG	459	65	63			82	81,31		
rs35599367 = GA	42	4	5,84	0,668	0,056	7	7,48	0,882	0,037
rs35599367 = AA	1	0	0,17			0	0,22		

Nem találtunk szignifikáns összefüggést a túlélés és a korábban leírt funkcionális SNP-k között (a *CYP3A5* rs776746 és a *CYP3A4* rs35599367 polimorfizmusa), bár az rs35599367 és az rs2246709 között a mintapopulációban kapcsoltsági egyenlőtlenség figyelhető meg (D': $0{,}88$, r^2: $0{,}1$). Ez azonban egyoldalú, ugyanis a két SNP allélfrekvenciája jelentősen eltér egymástól (rs35599367: $4{,}4\%$, rs2246709: $26{,}4\%$). Ugyanakkor a vizsgálatunkban ezeknek az SNP-knek a tesztelésekor a statisztikai erő is alacsonynak adódott a polimorfizmusok viszonylag alacsony allélfrekvenciája miatt (lásd 8. táblázat), így az asszociáció hiánya sem jelenthető ki teljes bizonyossággal.

Ezt követően egyváltozós és különböző klinikai paraméterekkel kiegészített többváltozós Cox-regressziós elemzéssel kiszámítottuk az rs2246709 hazárd hányadosait (hazard ratio, HR) a teljes és az eseménymentes túlélésre (lásd 9. táblázat és 10. táblázat). A

DOI:10.14753/SE.2017.1938



15. ábra. **A *CYP3A4* rs2246709 polimorfizmusának hatása a teljes- és eseménymentes túlélésre.** A Kaplan-Meier túlélési görbék a *CYP3A4* rs2246709 polimorfizmusának hatását szemléltetik a teljes túlélésre (bal oldal) és az eseménymentes túlélésre (jobb oldal). A görbék a túlélők arányát ábrázolják az idő függvényében. A p-értékek az SNP három lehetséges genotípusának a túlélésre gyakorolt hatása közötti különbség statisztikai szignifikanciáját mutatják.

többváltozós elemzés során a rizikócsoport-besorolást, az ALL sejtmorfológiai alcsoportját (B-, T-ALL), a citogenetikai abnormalitást, a kezelési protokollt és a betegek nemét vettük figyelembe. A Cox-regressziós elemzés eredménye megerősítette az rs2246709 teljes túlélésre gyakorolt szignifikáns hatását. A legkedvezőbb hatása az AG heterozigóta genotípusnak volt, amely 0,52 (95%CI: 0,28-0,98, p-érték: 0,04) HR értéket mutatott az AA genotípussal szemben (lásd 9. táblázat). Az rs2246709 polimorfizmus és az eseménymentes túlélés között nem találtunk szignifikáns asszociációt (lásd 10. táblázat).

4.3.2. A klinikai paraméterek és az rs2246709 polimorfizmus interakciójának hatása a túlélésre

A polimorfizmusok önálló hatásának vizsgálata után a bayesi relevanciaelemzés segítségével olyan interakciós hatásokat kerestünk, amelyek befolyásolhatják a betegek túlélését. A vizsgálandó interakciós hatások számának növelése érdekében az elemzésbe bevontuk további 34 gén 127 SNP-jét, amelyek a munkacsoport korábbi ALL vizsgálataiban szerepeltek. Az elemzés során csak a 10%-ot meghaladó allélfrekvenciájú polimorfizmusokat vizsgáltuk. Ahogyan az a Függelék: 6. táblázatában is látható, csak a *CYP3A4* rs2246709

DOI:10.14753/SE.2017.1938

9. táblázat. **A különböző klinikai paraméterek és az rs2246709 teljes túlélésre gyakorolt hatásának egyváltozós és többváltozós Cox-regressziós elemzésének eredménye.** Rövidítések: CI = konfidencia-intervallum, HR = hazárd hányados (hazard ratio), N = mintaszám, NE = események száma, Rizikócsoport esetén: LR = alacsony rizikó, MR = közepes rizikó, HR = magas rizikó

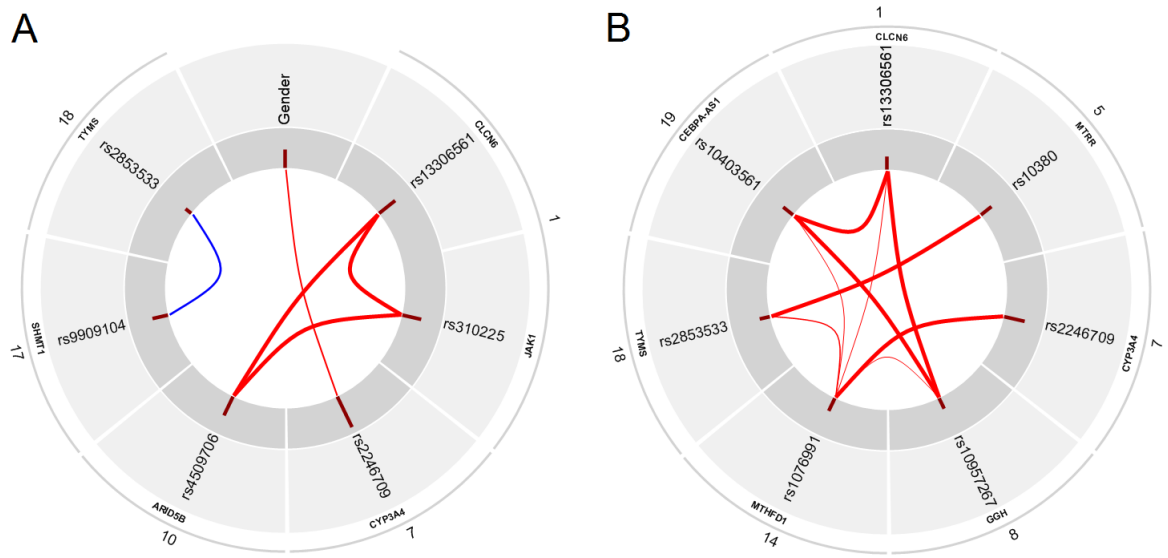
	Egyváltozós elemzés				**Többváltozós elemzés (N=373, NE=59, p=9,3*10^{-6})**			
Kovariáns	**N, NE**	**HR**	**95% CI**	**p-érték**	**N, NE**	**HR**	**95% CI**	**p-érték**
Rizikócsoport	LR=94, 7 MR=299, 39 HR=53, 20	2,87	1,87-4,40	**1,4*10^{-6}**	LR=90, 7 MR=244, 36 HR=39, 16	2,53	1,63-3,94	**3,8*10^{-5}**
Sejtmorfológia (Pre-B, B)	Pre-T, T=77, 15 Pre-B, B=372, 54	0,72	0,40-1,27	0,249	Pre-T, T=64, 13 Pre-B, B=309, 46	1,25	0,65-2,40	0,511
Citogenetika								
Normal	127, 20	1,00			104, 20	1,00		
Hiperdiploid	79, 8	0,63	0,28-1,43	0,268	66, 5	0,46	0,17-1,23	0,122
Egyéb	249, 44	1,14	0,67-1,93	0,638	203, 34	0,89	0,50-1,58	0,688
Protokoll								
‘88-‘90	123, 18	1,00			98, 15	1,00		
‘95	244, 45	1,34	0,78-2,33	0,289	235, 43	1,6	0,87-2,95	0,129
2002	117, 10	0,76	0,35-1,66	0,493	40, 1	0,22	0,03-1,70	0,148
Nem (Férfi)	Nő=221, 29 Férfi=290, 46	1,26	0,79-2,01	0,326	Nő=166, 22 Férfi=207, 37	1,15	0,66-2,00	0,629
rs2246709								
AA	275, 46	1,00			207, 39	1,00		
AG	202, 19	0,53	0,31-0,90	**0,02**	144, 14	0,52	0,28-0,98	**0,042**
GG	34, 10	1,85	0,93-3,66	0,079	22, 6	1,22	0,51-2,96	0,653

DOI:10.14753/SE.2017.1938

10. táblázat. **A különböző klinikai paraméterek és az rs2246709 eseménymentes túlélésre gyakorolt hatásának egyváltozós és többváltozós Cox-regressziós elemzésének eredménye.** Rövidítések: CI = konfidencia-intervallum, HR = hazárd hányados (hazard ratio), N = mintaszám, NE = események száma, Rizikócsoport esetén: LR = alacsony rizikó, MR = közepes rizikó, HR = magas rizikó

	Egyváltozós elemzés				**Többváltozós elemzés (N=373, NE=75, p=1,6*10^{-4})**			
Kovariáns	**N, NE**	**HR**	**95% CI**	**p-érték**	**N, NE**	**HR**	**95% CI**	**p-érték**
Rizikócsoport	LR=94, 9 MR=299, 52 HR=52, 21	2,52	1,71-3,70	**2,6*10^{-6}**	LR=90, 9 MR=244, 49 HR=39, 17	2,38	1,59-3,56	**2,7*10^{-5}**
Sejtmorfológia (Pre-B, B)	Pre-T, T=77, 18 Pre-B, B=372, 69	0,75	0,44-1,26	0,271	Pre-T, T=64, 15 Pre-B, B=309, 60	1,33	0,73-2,44	0,357
Citogenetika								
Normal	127, 25	1,00			104, 24	1,00		
Hiperdiploid	79, 12	0,76	0,38-1,51	0,426	66, 9	0,66	0,30-1,44	0,297
Egyéb	249, 53	1,11	0,69-1,79	0,665	203, 42	0,91	0,54-1,54	0,728
Protokoll								
‘88-‘90	123, 24	1,00			98, 21	1,00		
‘95	244, 52	1,14	0,70-1,85	0,597	235, 50	1,26	0,74-2,15	0,387
2002	117, 15	0,83	0,44-1,59	0,581	40, 4	0,6	0,20-1,77	0,357
Nem (Férfi)	Nő=221, 34 Férfi=290, 61	1,45	0,95-2,20	0,083	Nő=166, 26 Férfi=207, 49	1,41	0,86-2,32	0,174
rs2246709								
AA	275, 56	1,00			207, 47	1,00		
AG	202, 28	0,64	0,41-1,01	0,053	144, 21	0,64	0,38-1,08	0,094
GG	34, 11	1,69	0,89-3,23	0,111	22, 7	1,27	0,56-2,87	0,573

SNP-je bizonyult nagy valószínűséggel erősen relevánsnak az ötéves teljes és eseménymentes túlélés szempontjából. Ugyanakkor a bayesi elemzés feltárt néhány potenciális összefüggést a különböző változók (SNP-k és klinikai paraméterek) között (lásd 16. ábra). A továbbiakban csak a *CYP3A4* gént érintő interakciókat tárgyaljuk.



16. ábra. **Jelentősebb interakciók és redundanciák a vizsgált polimorfizmusok és klinikai paraméterek között a teljes és eseménymentes túlélés szempontjából.** Az ábra a változók közötti legfontosabb interakciókat és redundanciákat ábrázolja a teljes (**A**) és az eseménymentes (**B**) túlélés szempontjából. Az interakciókat a kör belsejében látható piros görbék mutatják, a redundanciák kék görbékkel vannak ábrázolva. A görbék vastagsága arányos a hatás mértékével. A belső szürke sávban látható bordó oszlopok az adott változó erős relevanciáját mutatják a célváltozó szempontjából. A külső körön a polimorfizmusokhoz tartozó gének neve és kromoszómája látható.

Az rs2246709 SNP és a betegek neme között teljes túlélés esetén interakciós hatást figyeltünk meg (interakciós arány = 0,22, 16A. ábra). Ezt a hatást Cox-regresszióval is megvizsgáltuk, mely során a két változó közötti interakciós tag módosító szerepét is elemeztük. A betegek neme erősen befolyásolta az rs2246709 teljes és eseménymentes túlélésre vonatkozó hatását is (teljes túlélés: 11. táblázat, eseménymentes túlélés: 12. táblázat). Érdekes módon ugyanis a heterozigóták, illetve a vad típusú homozigóták esetén a páciens nemétől függően fordított hatást figyeltünk meg: AG (heterozigóta) genotípus esetén a fiúk kockázata alacsonyabb volt mint a lányoké (HR: 0,26, 95%CI: 0,09-0,77, p-érték: 0,0158), AA (homozigóta vad) genotípus esetén viszont a fiúknak volt emelkedettebb a kockázata a lányokéhoz képest (HR: 2,39, 95%CI: 1,21-4,70, p-érték: 0,012).

Ugyanezt az interakciós hatást tapasztaltuk az eseménymentes túlélésre vonatkozóan is. Az eredményeink szerint a legkedvezőbb kilátásai a heterozigóta fiúknak voltak teljes és eseménymentes túlélés esetén is. A hatás szignifikáns maradt a rizikócsoporttal (lásd 11. táblázat és 12. táblázat) illetve az összes klinikai paraméterrel történt korrigálás után is (lásd Függelék: 7. táblázat).

11. táblázat. **A páciens neme és a *CYP3A4* rs2246709 polimorfizmusa közötti interakció hatása a teljes túlélésre.** Rövidítések: CI = konfidencia-intervallum, HR = hazárd hányados (hazard ratio), N = mintaszám, NE = események száma, Rizikócsoport esetén: LR = alacsony rizikó, MR = közepes rizikó, HR = magas rizikó

	Korrigálatlan (N=511, NE=75, p=2,2*10^{-4})				**Rizikócsoporttal korrigálva (N=446, NE=66, p=4,1*10^{-7})**			
Kovariáns	**N, NE**	**HR**	**95% CI**	**p-érték**	**N, NE**	**HR**	**95% CI**	**p-érték**
Rizikócsoport	-	-	-	-	LR=94, 7 MR=299, 39 HR=53, 20	2,48	1,61-3,82	**3,4*10^{-5}**
rs2246709, Nem								
AA, Nő	111, 11	1,00			100, 10	1,00		
AA, Férfi	164, 35	2,39	1,21-4,70	**0,012**	149, 33	2,1	1,03-4,28	**0,041**
AG, Nő	101, 15	1,53	0,70-3,32	0,287	87, 12	1,31	0,57-3,04	0,526
AG, Férfi	101, 4	0,39	0,12-1,23	0,109	83, 3	0,36	0,10-1,31	0,121
GG, Nő	9, 3	3,93	1,10-14,12	**0,036**	5, 2	4,16	0,91-19,05	0,066
GG, Férfi	25, 7	3,1	1,20-8,00	**0,019**	22, 6	2,31	0,83-6,40	0,108

12. táblázat. **A páciens neme és a *CYP3A4* rs2246709 polimorfizmusa közötti interakció hatása az eseménymentes túlélésre.** Rövidítések: CI = konfidencia-intervallum, HR = hazárd hányados (hazard ratio), N = mintaszám, NE = események száma, Rizikócsoport esetén: LR = alacsony rizikó, MR = közepes rizikó, HR = magas rizikó

	Korrigálatlan (N=511, NE=95, p=8,3*10^{-4})				**Rizikócsoporttal korrigálva (N=446, NE=82, p=2,6*10^{-6})**			
Kovariáns	**N, NE**	**HR**	**95% CI**	**p-érték**	**N, NE**	**HR**	**95% CI**	**p-érték**
Rizikócsoport	-	-	-	-	LR=94, 9 MR=299, 52 HR=53, 21	2,24	1,52-3,31	**4,7*10^{-5}**
rs2246709, Nem								
AA, Nő	111, 13	1,00			100, 12	1,00		
AA, Férfi	164, 43	2,55	1,37-4,74	**0,0031**	149, 39	2,18	1,13-4,17	**0,019**
AG, Nő	101, 18	1,57	0,77-3,21	0,214	87, 14	1,31	0,61-2,84	0,489
AG, Férfi	101, 10	0,84	0,37-1,91	0,673	83, 8	0,81	0,33-1,99	0,649
GG, Nő	9, 3	3,31	0,94-11,63	0,062	5, 2	3,73	0,83-16,71	0,085
GG, Férfi	25, 8	3,13	1,30-7,54	**0,011**	22, 7	2,44	0,96-6,24	0,062

A fentin kívül a bayesi elemzés egy másik interakcióra is rávilágított, méghozzá az *MTHFD1* rs1076991 SNP-je és az rs2246709 (*CYP3A4*) között eseménymentes túlélés

esetén (interakciós arány = $0{,}9$, 16B. ábra). A Cox-regressziós elemzés ezt a hatást is megerősítette, mind a teljes, mind az eseménymentes túlélés esetén az összes klinikai paraméterrel történt korrigálás után is (teljes túlélés: Függelék, 8. táblázat; eseménymentes túlélés: Függelék, 9. táblázat). Ahogyan az a táblázatokból is látható, a különböző genotípus kombinációk szignifikánsan különböző kockázatmódosító hatással rendelkeztek.

4.3.3. A rizikócsoport-besorolás módosítása a páciens neme és az rs2246709 genotípus alapján

A továbbiakban megvizsgáltuk, hogy a páciens nemének és az rs2246709 (*CYP3A4*) polimorfizmus genotípusának figyelembevételével lehet-e módosítani a betegek rizikócsoport-besorolását úgy, hogy az pontosabban jósolja a betegek halálának kockázatát.

Ehhez először a Cox-regressziós modell alapján, amely az eredeti rizikócsoport-besorolást, a páciens nemét, az rs2246709 genotípusát és az utóbbi két változó interakcióját tartalmazta független változóként, kiszámítottuk minden beteg esetén az elhalálozás relatív rizikóját. A gyakorlatban ez azt jelenti, hogy a változók értékeinek összes lehetséges konfigurációjára, azaz a három rizikócsoport, a kétféle nem és a háromféle genotípusérték miatt $3 \times 2 \times 3 = 18$ lehetséges variációjára számítjuk ki a jósolt relatív rizikót. Ezek alapján kialakítottunk egy új rizikócsoport-besorolást (RiskGroupCoxModel). Az értékeket három csoportba osztottuk a következő szabály szerint: a $0{,}85$-nál kisebb relatív rizikóval rendelkező páciensek az alacsony; a $0{,}85$-nál nagyobb, de $3{,}0$-nál kisebb relatív rizikójú páciensek a közepes; a $3{,}0$-nál nagyobb relatív rizikójú betegek pedig a nagy rizikójú csoportba kerülnek. Ezeket a konkrét határértékeket a következőképpen állapítottuk meg: A 18 relatív rizikó értéket az összes lehetséges, konzisztens módon[13] három csoportra osztottuk. Minden ilyen csoportbeosztásra kiszámítottuk az átlagos C-indexet[14] (konkordancia index, lásd Módszerek fejezet) a vizsgált időtartományra, majd ezek közül kiválasztottuk azt a csoportbeosztást (illetve vágópont-variációt), amelyik a legmagasabb átlagos C-indexet eredményezte.

[13] Egy csoportosítást akkor nevezünk konzisztensnek, ha minden konfigurációpárra teljesül, hogy az alacsonyabb relatív rizikójú konfiguráció alacsonyabb, vagy ugyanabba a kockázatbesorolásba esik, mint a magasabb relatív rizikójú konfiguráció.

[14] A C-index az ún. konkordáns mintapárok (páciens párok) arányát méri az összes lehetséges mintapár között. Egy mintapár akkor konkordáns, ha a pár két tagja közül az, amelyik korábbi időpontban tapasztal eseményt, magasabb rizikójú csoportba tartozik, mint a pár másik tagja.

DOI:10.14753/SE.2017.1938

Ezután egy másik új rizikócsoport-besorolási mutatót is származtattunk (RiskGroupCTree) az új RiskGroupCoxModel besorolás alapján egy ún. döntési fa tanulásával. A célunk ezzel az volt, hogy meghatározzunk olyan viszonylag egyszerű szabályokat, amelyek képesek megragadni a Cox-regressziós modell „lényegét". A döntési fa a következő csoportbesorolási szabályokat állapította meg: (1) az AG genotípusú férfiak kerüljenek az alacsony kockázatú csoportba a korábbi rizikócsoport-besorolástól függetlenül, (2) azok a GG genotípusú betegek, akik korábban az alacsony rizikójú csoportba tartoztak, kerüljenek a közepes rizikójú csoportba a nemüktől függetlenül.

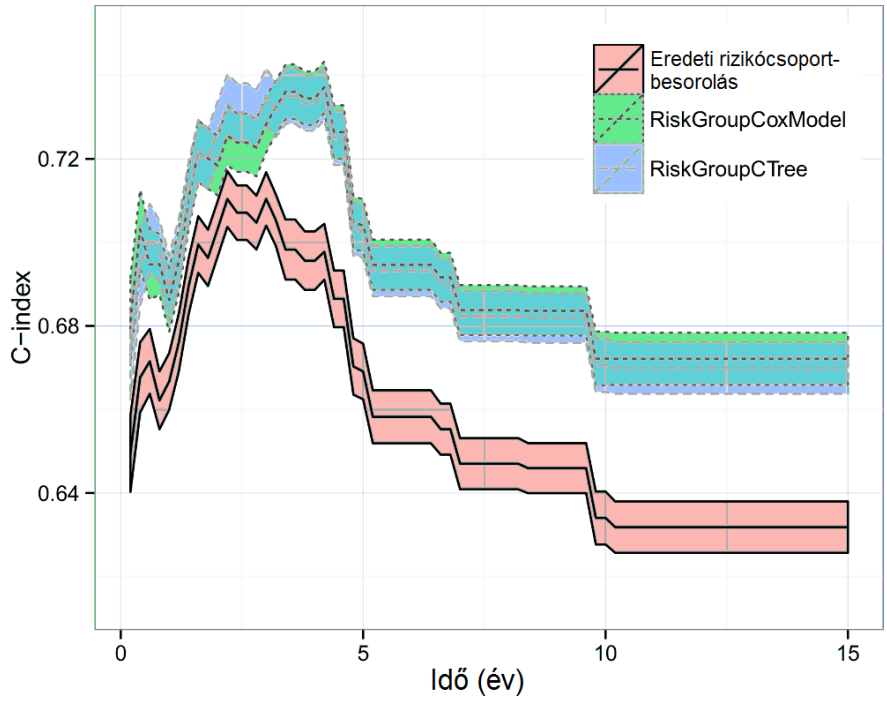
A különböző rizikócsoport-besorolások közötti különbségeket az egyes csoportokba besorolt páciensek számának tekintetében a 13. táblázat mutatja be.

13. táblázat. **A különböző rizikócsoport-besorolások közötti különbségek a besorolt páciensek számát tekintve**

		Eredeti rizikócsoport-besorolás			RiskGroupCTree		
		Alacsony	Közepes	Magas	Alacsony	Közepes	Magas
RiskGroupCoxModel	Alacsony	90	56	5	151	0	0
	Közepes	4	240	15	0	244	15
	Magas	0	3	33	0	3	33
RiskGroupCTree	Alacsony	90	56	5			
	Közepes	4	243	0			
	Magas	0	0	48			

Mindhárom rizikócsoport-besorolásra meghatároztuk azok diszkriminatív teljesítményét a C-index kiszámításával. Az eredményeket a teljes túlélés esetén a 17. ábrán láthatjuk. Mindkét új rizikócsoport-besorolás szignifikánsan felülmúlta az eredetit a vizsgált időtartam minden egyes pontjában ($0{,}2$ évtől 15 évig $0{,}2$ évente) teljes túlélés esetén (a Benjamini-Hochberg-korrigált p-értékek maximuma $< 0{,}001$). Az új besorolások C-index értékeinek átlagos különbsége az eredeti besoroláshoz képest $3{,}5$ százalékpont (RiskGroupCTree), illetve $4{,}21$ százalékpont (RiskGroupCoxModel) volt. Az eredmények azt is mutatják, hogy a RiskGroupCTree besorolást leíró szabályok jól megragadták a bonyolultabb Cox-modell lényegét, ugyanis a két új változó C-index értékei között nem volt szignifikáns a különbség (a Benjamini-Hochberg-korrigált p-értékek minimuma $> 0{,}08$).

DOI:10.14753/SE.2017.1938



17. ábra. **A rizikócsoport-besorolások teljesítményének összehasonlítása a teljes túlélés esetén.** A görbék a különböző rizikócsoport-besorolások becsült konkordancia indexét ábrázolják a vizsgált időszakban. A rózsaszínű görbe az eredeti rizikócsoport-besorolásnak felel meg, a zöld a Cox-modellen alapuló besorolást ábrázolja (RiskGroupCoxModel), a kék pedig a döntési fa alapú besorolás teljesítményét mutatja (RiskGroupCTree). A becsült konkordancia indexek 95%-os konfidencia-intervallumát a görbék körüli sávok jelzik.

DOI:10.14753/SE.2017.1938

4.4. A bayesi relevanciaelemzési módszertan alkalmazási lehetőségeinek vizsgálata asszociációs vizsgálatokban

A bayesi relevanciaelemzést munkacsoportunk több genetikai asszociációs vizsgálatban is alkalmazta. A vizsgálatok sikeres elvégzéséhez számos metodológiai fejlesztést végeztünk, amelyek magukban foglalják a változók közötti kapcsolati típusok definiálását (lásd 1.3.2. alfejezet) a változók közötti interakciók és redundanciák automatikus felderítését (lásd 1.3.4. alfejezet) és az eredmények intuitív megjelenítési lehetőségeinek kidolgozását. Munkám során a bayesi módszertant a 4.3. fejezetben bemutatott elemzésen kívül kettő, az akut limofid leukémia hajlamosítását és túlélését tanulmányozó jelölt gén asszociációs vizsgálatban alkalmaztam, illetve részt vettem egy - az asztma genetikai hátterét tanulmányozó parciális genomszűrési vizsgálat statisztikai kiértékelésében is. Ezek során fő feladatom a bayesi módszertan tesztelése, a frekventista vizsgálatok eredményeivel való összevetése, az eredmények metodológiai szempontból történő elemzése, illetve ezek alapján a módszertan továbbfejlesztése volt.

Jelen dolgozatban a célom nem a részletes eredmények, illetve az azokból levonható orvosbiológiai következtetések bemutatása, hanem a metodológiai eredmények ismertetése volt. Az előbbiek a munkacsoport tagjainak, Dr. Lautner-Csorba Orsolya és Dr. Ungvári Ildikó doktori értekezéseiben olvashatók [78, 115].

4.4.1. Releváns változók meghatározása

A bayesi relevanciaelemzés során a változók (polimorfizmusok és klinikai paraméterek) közötti összefüggések feltérképezésével arra a kérdésre keressük a választ, hogy egy adott célváltozó szempontjából melyek a releváns változók, amelyek közvetlen módon meghatározzák annak értékét. A releváns változók ismeretében a célváltozó értéke független lesz a többi változótól, azaz a releváns változók elfedik a többi változó hatását.

A Bayes-statisztikai módszertani keret alkalmazásának eredményképpen direkt valószínűségi állításokat kapunk egy adott változó relevanciájára vonatkozóan, például: „Az rs10821936 erős relevanciájának valószínűsége $0{,}76$ az ALL-re való hajlamosítás szempontjából.”. Ez az állítás azt jelenti, hogy a genotipizálási adataink és az elemzésbe bevont változók alapján $0{,}76$ (azaz 76%) annak a valószínűsége, hogy az rs10821936 SNP erősen

releváns, azaz meghatározó szerepet tölt be az ALL-re való hajlamban.

Az egyik fő különbség a bayesi relevanciaelemzés és a hagyományos statisztikai asszociációs tesztek között az, hogy az előbbi - az összes változó hatásának együttes modellezéséből eredően - képes megkülönböztetni a közvetlen hatásokat és a tranzitív (például a változók közötti kapcsoltsági egyenlőtlenségek miatt keletkező) asszociációkat. Az ALL irodalomból ismert hajlamosító génjeit vizsgáló genetikai asszociációs vizsgálatban például azt találtuk, hogy az *ARID5B* gén összes vizsgált SNP-je (rs4509706, rs4948487, rs10821936, rs4948496, rs4948502) nagy valószínűséggel (Pr: $> 0{,}8$) asszociált az ALL-hajlammal. Azonban csak az rs10821936 SNP-ről mondhatjuk el, hogy közvetlen relevanciával is bírt, ugyanis ennek posterior valószínűsége $0{,}76$, míg a többi SNP esetén $< 0{,}1$ volt. Így, bár ezek az SNP-k is asszociálnak az ALL-hajlammal, az is elmondható róluk, hogy nem oki, illetve nem funkcionális variánsok. Nagyon hasonló a helyzet az *IKZF1* gén polimorfizmusai esetén is, ahol az rs6954833, rs10235796 és rs4132601 asszociáltsága nagyon valószínű ($> 0{,}97$), de csak a rs4132601 esetén valószínű a közvetlen relevancia ($> 0{,}97$). Fontos hangsúlyozni azonban azt, hogy az eredményeket a modellbe bevont változók alapján szükséges értelmezni, azaz változók elhagyásával, illetve új (például valódi oki variánsok) hozzávételével a posteriorokra vonatkozó eredmények módosulhatnak.

4.4.2. Interakciók és redundanciák keresése

A bayesi módszer a változók közötti interakciók felderítését kétféle módon is támogatja. Az elsőt, az ún. strukturális interakciót, vagy tiszta (főhatás nélküli) interakciót az egyes kapcsolati típusok között definiáltuk (lásd 1.3.2. alfejezet). Ez abban az esetben jelenik meg, ha egy adott változónak önmagában nincs hatása az célváltozóra, azonban egy másik változóval együtt már igen. Ebben az esetben tipikusan módosítja a másik változó - célváltozót befolyásoló hatását. Ez a fajta interakciós hatás a Bayes-háló struktúrákban egy ún. v-struktúra formájában jelenik meg, azaz két, egymással nem összekötött csomópontból (a célváltozóból és az interaktáló változóból) egy-egy él fut egy közös gyermek csomópontba (amelynek célváltozóra vonatkozó hatását az interaktáló változó módosítja). A tiszta interakció hatás erősségének jellemzésére a hatás *a posteriori* valószínűségét használjuk.

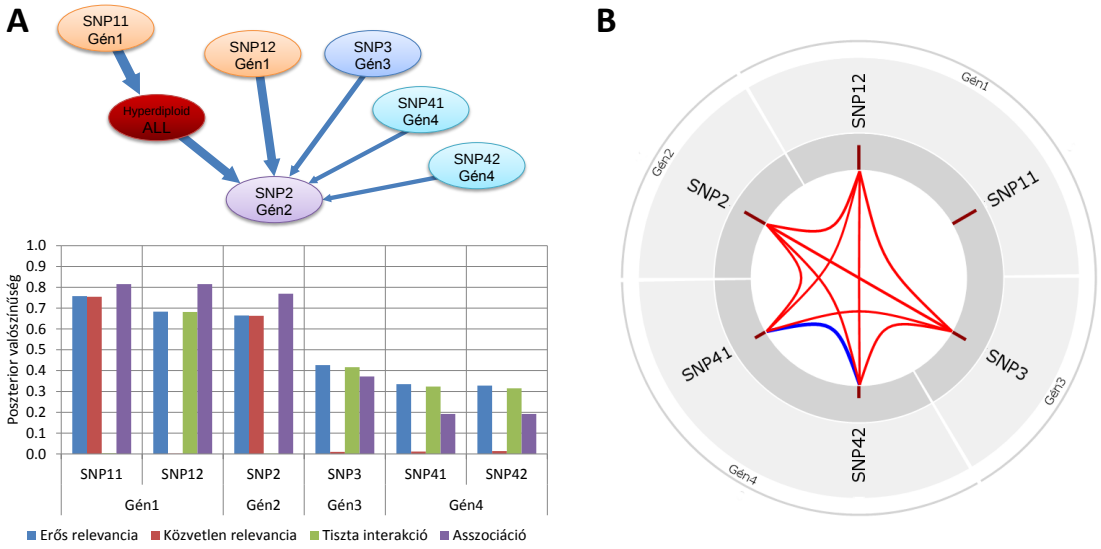
DOI:10.14753/SE.2017.1938

A másik fajta interakciós típust a jellegéből fakadóan statisztikai interakciónak nevezhetjük (lásd 1.3.4. alfejezet). Ezeket az interakciókat a célváltozó szempontjából erősen releváns változó*halmazok* vizsgálata alapján deríthetjük fel. Ekkor ugyanis, ha két (vagy több) változó gyakrabban fordul elő együtt a célváltozó szempontjából erősen releváns változóhalmazokban, mint ahogyan azt az egyváltozós marginális valószínűségek alapján várnánk, akkor ez a változók közötti interakciós hatásra utal. (Amennyiben ritkábban fordulnak elő együtt, akkor redundancia-hatásról beszélhetünk.) Ezeknek a hatásoknak az ábrázolására fejlesztettem ki a 4.3. fejezetben is látható, ún. interakciós körábrát (lásd 16. ábra), illetve egy szoftvert, amellyel az elemzések eredménye alapján ezeket az ábrákat egyszerűen előállíthatjuk. A statisztikai interakciós hatás erősségének jellemzésére azt az arányszámot használjuk, amely a változók tapasztalt, illetve elvárt együttes előfordulását jellemzi.

A fent bemutatott két interakció-típus a legtöbb esetben átfed egymással, azaz egy strukturális interakció általában megjelenik statisztikai interakció formájában is. Előfordulhatnak azonban eltérések is, ugyanis a statisztikai interakció nem feltétlenül tiszta interakcióként jelenik meg a modellekben. Az interakciótípusok átfedésének szemléltetésére a következőkben bemutatok egy valós vizsgálatból származó példát (a részletes eredmények, illetve az azokból levonható orvosbiológiai következtetések Dr. Lautner-Csorba Orsolya doktori értekezésében olvashatók [78]). A folát anyagcserében, transzportjában szerepet játszó gének ALL-re való hajlamosító hatását elemezve a hiperdiploid ALL alcsoportban számos interakciós hatást figyeltünk meg (lásd 18B. ábra). Ugyanezek az interakciós hatások a konszenzus modell struktúrája alapján is jól azonosíthatók (18A. ábra). A modellben ugyanis az SNP12, SNP3, SNP41 és SNP42 változók nem közvetlenül, hanem az SNP2 változón keresztül kötődnek a hiperdiploid ALL-hez, azaz legnagyobb valószínűséggel a hatásuk *tiszta interakciós*. Ez azt jelenti, hogy csak akkor válnak relevánssá a hiperdiploid ALL-hajlam szempontjából, ha az SNP2 polimorfizmus értéke is ismert lesz. Más szóval az SNP2 genotípusától függ, hogy a nem közvetlenül kötődő SNP-k genotípusa milyen mértékben befolyásolja a hiperdiploid ALL kialakulásának kockázatát. Zavaró lehet a biológus kutató számára, hogy a konszenzusos modellben a hiperdiploid ALL csomópontból irányított él mutat az SNP2 csomópontba. Az élek irányultsága azonban nem feltétlenül jelent ok-okozati összefüggést, hanem a füg-

DOI:10.14753/SE.2017.1938

gőségi/függetlenségi viszonyokat tükrözi. Ebben az esetben a modell azt fejezi ki, hogy az SNP2 csomópont egyéb „szülei” (azok a polimorfizmusok, amelyekből él mutat az SNP2 csomópontba) feltételesen függetlenek a hiperdiploid ALL-től, azonban az SNP2 polimorfizmusának genotípusa ismeretében már függővé válnak attól, azaz az együttes értékük befolyásolja a hiperdiploid ALL kialakulásának kockázatát.



18. ábra. **A strukturális és statisztikai interakciók ábrázolása egy valós genetikai asszociációs vizsgálat eredményei alapján.** **A**. A konszenzus modell (fent) a folát anyagcserében szerepet játszó gének vizsgálata során a hiperdiploid ALL kialakulásának kockázatát és annak legfontosabb befolyásoló tényezőit, illetve a köztük lévő összefüggéseket ábrázolja. A csomópontok a polimorfizmusoknak, illetve a célváltozónak felelnek meg, az élek a köztük lévő közvetlen hatást jelentik. Az élek vastagsága arányos a kapcsolat valószínűségével. Az oszlopdiagram (lent) a különféle kapcsolati típusok *a posteriori* valószínűségét ábrázolja a hiperdiploid ALL kialakulása célváltozó szempontjából. **B**. Az ábra a változók közötti legfontosabb interakciókat és redundanciákat ábrázolja a hiperdiploid ALL kialakulása szempontjából. Az interakciókat a kör belsejében látható piros görbék mutatják, a redundancia kék színnel van ábrázolva. A görbék vastagsága arányos a hatás mértékével. A belső szürke sávban látható bordó oszlopok az adott változó erős relevanciáját mutatják a célváltozó szempontjából. A külső körön a polimorfizmusokhoz tartozó gének neve látható. **Megjegyzés**: A gének és polimorfizmusok valódi nevei Dr. Lautner-Csorba Orsolya doktori értekezésében olvashatók [78].

Interakciós elemzések természetesen frekventista statisztikai módszerekkel is vizsgálhatók. Erre több lehetőség is kínálkozik, például változópáronkénti asszociációs tesztek formájában, vagy olyan logisztikus regressziós modell alkalmazásával, amely interakciós tagot is tartalmaz. Amennyiben azonban frekventista módszereket alkalmazunk, fellép a

többszörös hipotézistesztelés problémája, és az eredményeket az elvégzett tesztek számának megfelelően korrigálnunk kell. Mivel a tesztek száma azonos hatvánnyal nő, mint az interakció szempontjából vizsgált változóhalmazok mérete (azaz változópárok esetén négyzetesen, változóhármasok esetén köbösen), a korrekció a legtöbb esetben túlságosan konzervatív, azaz a valós interakciós hatások nem mutathatók ki szignifikánsan. A bayesi módszertan segítségével azonban automatikusan felderíthetjük tetszőleges méretű változóhalmazok interakciós hatásait, miközben a többszörös tesztelés problémáját a bayesi modellátlagolás használatával elkerüljük (illetve automatikusan kezeljük).

4.4.3. Több célváltozó kezelése

A bayesi relevanciaelemzés előnyei a frekventista módszerekkel szemben különösen azokban a helyzetekben nyilvánulnak meg, amelyekben több, egymással összefüggő fenotípusos célváltozó genetikai befolyásoló tényezőit vizsgáljuk. Ekkor a korábban ismertetett egyváltozós relevancia kiszámítása mellett a többváltozós, modell alapú megközelítés miatt további lehetőségek nyílnak a fenotípusos és genetikai változók összefüggésrendszerének feltérképezésére: a bayesi relevanciaelemzés ugyanis lehetővé teszi, hogy gyakorlatilag tetszőleges strukturális kérdés *a posteriori* valószínűségét kiszámítsuk.

Ezt egy parciális genomszűrési vizsgálatban mutattuk meg, amelynek során az asztma és a vele szoros összefüggésben lévő IgE szint, eozinofil szint és allergia genetikai hajlamosító tényezőit tanulmányoztuk. Minden egyes SNP-re és fenotípusos változóra kiszámítottuk annak *a posteriori* valószínűségét, (1) hogy az adott SNP erősen releváns az adott fenotípusos változó szempontjából, (2) hogy az adott SNP *csak* az adott fenotípusos változó szempontjából releváns, míg a többiéből nem, (3) hogy az adott SNP az adott fenotípusos változó szempontjából *nem* releváns, de bármely másik szempontjából igen. Végül minden egyes SNP-re kiszámítottuk azt is, (4) hogy bármely fenotípusos változó szempontjából releváns-e. Ez utóbbit az egyváltozós eredmények alapján is meg lehet becsülni az egyes változók függetlenségének feltételezésével, de az elemzés során megmutattuk, hogy a becsült értékek a fenotípusos változók összefüggésrendszere miatt eltérnek a többváltozós módszerrel kiszámított értéktől. A továbbiakban egy példával illusztrálom az előbbi számítások jelentőségét. A *PTGER2* gén rs12587410 SNP-jét vizsgálva azt kaptuk, hogy az erős relevancia *a posteriori* valószínűsége az IgE szint szempontjá-

ból $0{,}31$, az eozinofil szint szempontjából $0{,}38$, az allergia szempontjából $0{,}53$, illetve az asztma szempontjából pedig $0{,}81$ volt. Ez azt jelenti, hogy az rs12587410 erős relevanciája nem zárható ki egyik fenotípusos változó szempontjából sem, de legerősebb az asztma esetén volt. Amikor kiszámítottuk annak valószínűségét, hogy az SNP *csak* az adott fenotípusos változó szempontjából releváns, akkor az IgE szint és az eozinofil szint esetén $0{,}02$, az allergia esetén $0{,}04$ míg az asztma esetén $0{,}16$ valószínűséget kaptunk. Ez azt mutatja, hogy az SNP nem kapcsolható kizárólagosan egyetlen fenotípusos változóhoz sem. Végül amikor annak valószínűségét számítottuk ki, hogy az adott SNP az adott fenotípusos változó szempontjából *nem* releváns, de a többiéből igen, az IgE szint és az eozinofil szint esetén $0{,}5$-nél nagyobb, míg az allergia és az asztma esetén $0{,}5$-nél kisebb értéket kaptunk. Mindez azt jelzi, hogy az rs12587410 feltehetően az allergia és az asztma szempontjából is releváns, míg az IgE szint és eozinofil szint esetén ez a hatás csak az előbbieken keresztül jelenik meg.

DOI:10.14753/SE.2017.1938

5. Megbeszélés

5.1. Variánskivonatolási munkafolyamatok teljesítménye és konkordanciája

Az egyedi variánskivonatolási munkafolyamatok teljesítményét szimulált adatok segítségével hasonlítottuk össze. Az eredményeink más kutatócsoportokkal egyetértésben [38–41] 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. A HaplotypeCaller jó teljesítőképességét Pirooznia és mtsai kutatása is igazolta [43]. Liu és mtsai a miénkhez hasonló eredményre jutottak a SAMtools és a UnifiedGenotyper összehasonlító vizsgálatakor. Eredményeik szerint a SAMtools precizitása minden esetben felülmúlta a UnifiedGenotyperét, ugyanakkor a szenzitivitás tekintetében fordított volt a helyzet a leolvasási mélységtől függetlenül [40]. A mi eredményeink nagy leolvasási mélység esetén szintén ugyanezt igazolták, kis lefedettségek esetén azonban nem. A különbséget a variánskivonatoló algoritmusok fejlődése is okozhatja, ugyanis Liu és mtsai régebbi verziójú programokat használtak.

Korábbi kutatásokkal összhangban [43, 116] 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. Ez részben ellentmond más kutatócsoportok eredményeinek, de jól magyarázható a zaj (szekvenálási hibák) felszaporodásával. Mivel az egyes variánskivonatolók más módon kezelik a zajt, ezért eltérő mértékben és bizonyos esetekben eltérő trend szerint változott a precizitásuk. Tipikusan azok a módszerek hívtak nagyobb arányban hamis variánsokat, amelyek egyedi lókuszonként vizsgálják az eltérést a referencia szekvenciától, így valószínűsíthető, hogy ezek a módszerek jobban ki vannak téve az ilyen típusú hibáknak.

Az egyes módszerek szenzitivitása és precizitása azonos lefedettség mellett jóval magasabb volt SNP-k, mint indelek esetén (lásd 2. táblázat). Más kutatások is kimutat-

ták, hogy a jelenlegi indel-detektáló módszerek relatíve pontatlanabbak [42] mint az SNP kivonatolók, annak ellenére, hogy az indeleknek a géntermékre potenciálisan jóval nagyobb hatásuk lehet mint az SNP-knek [117]. Az indel-hívás nehézsége több tényezőből fakad: (1) azokat a leolvasásokat, amelyek átfednek egy beszúródott vagy törlődött genomi szakaszt, általában nehezebb felilleszteni a referencia szekvenciára, ugyanis a pontos illeszkedéshez ún. „hézagos" illesztésre van szükség [51, 118], (2) az indel referencia szekvenciához képesti pozíciója az esetek nagy részében nem egyértelmű, és elméletileg mindegyik helyes lehet [42]. Az első probléma általában jól kezelhető a lefedettség növelésével, illetve nagyméretű beszúródások vagy törlődések esetén a leolvasás hosszának növelésével [41]. Az eredményeink azonban 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). A nem egyértelmű variánsreprezentáció problémája részben kezelhető a lehetséges pozíciókon belüli balra rendezéssel (normalizálással), de ez nem fed le minden problémás esetet.

Más kutatásokhoz hasonlóan [119, 120] mi is azt találtuk, hogy az illesztőprogram megválasztása jelentősen befolyásolja a variánskivonatolás eredményét. A BWA használata általában szignifikánsan jobb eredményekre vezetett (lásd 4. táblázat). Highnam és mtsai a UnifiedGenotyper teljesítményét vizsgálva szintén kimutatták, hogy az magasabb szenzitivitást és precizitást ért el a BWA használatával a Bowtie 2-vel szemben [119].

Számos kutatás kimutatta, hogy jelentős eltérés van a széles körben használt variánskivonatoló munkafolyamatok eredményei között, még abban az esetben is, ha ugyanazokra a szekvencia adatokra alkalmazzák is azokat [39, 40, 42, 43]. A mi eredményeink is ezt a megfigyelést igazolták mindkét variánstípus esetén, de különösen indelekre vonatkozóan.

Érdekes módon a variánskivonatolók konkordanciája közepes lefedettség felett a leolvasási mélység növekedésével enyhén csökkent mind SNP-k, mind indelek esetén. Ez egybevág Yu és Sun [39], illetve O'Rawe és mtsai [42] eredményeivel, de ellentmond annak az általános elvárásnak, mely szerint a leolvasási mélység növekedésével a kivonatolók pontossága is nő, ami a módszerek együtt járásának növekedését is eredményezné. Korábban bemutattuk, hogy a leolvasási mélység változásával a szenzitivitás és a precizitás ellentétes irányban változik. A módszerek konkordanciájában megfigyelhető trendvál-

tozás azzal magyarázható, hogy alacsonyról közepes lefedettségek felé haladva a szenzitivitásnyereség felülmúlja a precizitásban bekövetkező veszteséget, amely összességében a pontosság növekedését eredményezi. Nagyobb leolvasási mélységek esetén azonban a szenzitivitás növekedése és a precizitás csökkenése már kiegyensúlyozottabb, vagy adott esetben fordított. Feltételezzük, hogy ez a jelenség a variánskivonatoló programok eltérő stratégiáiból ered, amellyel a különböző típusú statisztikai hibákat (pl. mintavételi hiba) és aszimptotikus hibákat (pl. szisztematikus eltérések) kezelik.

Az illesztőprogramok különbsége a variánskivonatolási módszerek konkordanciájára is hatással volt, a módszerek együtt járása ugyanis általában kisebb volt a Bowtie 2 illesztések használatakor. Ez részben megmagyarázható azzal, hogy a BWA használata esetén a kivonatolók általában nagyobb pontosságot értek el.

A variánskivonatolás precizitásának javítása érdekében a bioinformatikai kiértékelések során gyakran alkalmaznak manuális variáns szűrést [38, 39, 121]. 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 ajánlásoknak megfelelően végeztük (lásd Módszerek), 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.

5.2. 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 (lásd 11. ábra).

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 használtuk. Az AUPRC pontszám valószínűségként is értelmezhető: megmutatja, hogy mekkora a várható értéke a valódi variánsok arányának egy véletlenszerűen kiválasztott küszöbnél nagyobb mutatóval rendelkező variánsok között [122]. Az AUPRC-t gyakran használják az olyan erősen kiegyensúlyozatlan problémák teljesítményének a meghatározására, amelyekben a valódi negatívok száma nagy mértékben felülmúlja a valódi pozitívok számát. Ilyen például a dokumentumkeresés az interneten, de ugyanez fennáll a variánskivonatolás esetén is, hiszen a valódi negatív variánsok gyakorlatilag a teljes célrégiót lefedik. Az ilyen típusú problématerületeken az AUPRC pontszám sokkal informatívabb, mint például az általánosan ismert *receiver operating characteristic* görbe alatti terület (AUROC, AUC), ugyanis az AUPRC pontszámot nem nyomja el a valódi negatívok nagy száma.

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 VariantMetaCaller és az egyedi módszerek közötti különbség a leolvasási mélységtől, az illesztőprogramtól és a variáns típusától függően változó volt, amely több tényező együttes hatásának az eredménye: (1) az egyedi módszerek egymáshoz képesti szenzitivitásának változása (lásd 6. ábra), (2) az egyedi kivonatolók által hívott hamis variánsok arányának trendjellegű változása (lásd 7. ábra), (3) a manuális szűrések miatt változó mértékben megváltozott szenzitivitás és precizitás (lásd 10. ábra) és (4) a kivonatoló módszerek által nyújtott variánsminőség-becslés jóságának változása.

A VariantMetaCaller egyedi variánskivonatolókkal szemben tapasztalt jobb teljesítményét a szimulált adatokon végzett elemzések során egy illusztrációs célokból kiválasztott kromoszómán (17-es) mutattuk ki. Felvetődhet a kérdés, hogy mennyire általánosíthatók az eredmények más kromoszómákra, illetve a humán genom más részeire is. A VariantMetaCaller az egyedi variánskivonatolók együtt járását és komplementaritását használja ki. Ebből következően általánosságban elmondható, hogy minden olyan genomi régió esetén várhatóan jobban fog teljesíteni, mint a bemenetéül szolgáló módszerek, amelyre teljesül, hogy azon a régión a variánskivonatoló módszerek jellemzően együtt járnak (azaz a valódi variánsok nagy részét minden kivonatoló módszer megtalálja), de egymást ki is egészítik (azaz vannak olyan variánsok, amelyeket csak néhány kivonatoló talál meg). Mivel a variánskivonatoló módszereknek ez a tulajdonsága az előző fejezetben hivatkozott kutatási eredmények alapján a teljes genomra teljesül, a VariantMetaCaller teljesítménye várhatóan minden genomi régió esetén jobb lesz, mint az egyedi variánskivonatoló módszerek teljesítménye. Ezt támasztja alá az is, hogy a kombinációs módszer minden vizsgált genomi régióméret esetén szignifikánsan jobb eredményt ért el.

A VariantMetaCaller egy gépi tanulási eljárást használ a valódi és a hamis variánsok megkülönböztetésére, amelyhez pozitív és negatív tanítópéldák megadására van szükség. A VariantMetaCaller működése során azt megfigyelést használja ki, hogy a teljesen kon-

kordáns, minden egyedi kivonatoló módszer által megtalált variánsok legnagyobb része valós, míg a csak egyetlen módszer által megtalált variánsok nagyobb része nem valódi. Azonban, ahogyan azt a 9. ábra is mutatja, a VariantMetaCaller tanítására használt adat zajos, ugyanis a negatív tanítóminták egyes esetekben jelentős számban tartalmaznak valódi variánsokat. Ezért megvizsgáltuk ennek a zajnak a hatását, és a következőket találtuk: (1) Ha a pozitív tanítópéldákat felhasználtuk a tanításhoz, de a potenciálisan zajos negatív tanítópéldákat egyáltalán nem (egyosztályos SVM), akkor a program teljesítménye csökkent. (2) Ha összehasonlítottuk a VariantMetaCaller teljesítményét az eredeti és annak az idealisztikus tanítómintának a használatával, amely csak a valóban pozitív és valóban negatív tanítópéldákat tartalmazta, akkor az ideális esetben nőtt ugyan a program teljesítménye, de csak kis mértékben (átlagos AUPRC javulás: $< 0{,}007\%$ SNP-ek és $< 0{,}4\%$ indelek esetén). (3) Amennyiben csak három variánskivonatoló módszert kombináltunk a VariantMetaCaller segítségével, a program teljesítménye jellemzően kisebb volt, mint amikor mind a négy módszert használtuk (lásd Függelék: 10. táblázat). Így általánosságban is az várható, hogy további egyedi variánskivonatoló módszerek kombinálásával a zaj még tovább csökken, illetve a VariantMetaCaller teljesítménye tovább nő.

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 [6, 10, 11]. Ez a variánsok valószínűségének pontos becslésével válik elérhetővé: a valószínűségi értékekkel a variánsokat sorrendezhetjük, majd minden egyes küszöbér-

tékre 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.

5.3. 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.

A CYP3A4 az emberekben az egyik legfontosabb gyógyszer-metabolizáló enzim, de az ALL kemoterápiájában betöltött funkciója meglehetősen bonyolult. Fontos szerepe van a vinkrisztin, dexametazon és a doxorubicin szervezetből való kiürítésében (clearance), ugyanakkor a vinkrisztin és a dexametazon a CYP3A4 enzim működését gátolja (inhibitor), a doxorubicin viszont serkenti (inducer). Emellett a CYP3A4 nemcsak a ciklofoszfamid aktivációját, hanem a neurotoxikus hatású klóracetaldehid keletkezését is katalizálja, ami súlyos mellékhatásokhoz vezethet [123, 124]. Ezekből az ellentétes szerepekből eredően nehéz megjósolni a CYP3A4 enzim szintjében tapasztalható különbségek hatását az ALL-ben alkalmazott kemoterápia hatásosságára. Bár a CYP3A4 aktivitása akár 10 vagy 100-szoros különbségeket mutathat az egyének között, melynek örökölhetősége 90%-os, az enzimaktivitás genetikai háttere máig nem tisztázott. Nemrégiben azonosítottak egy funkcionális SNP-t a 6-os intronban (rs35599367; *CYP3A4*22*), amely csökkent CYP3A4 termelődéssel és aktivitással asszociált máj sejtekben, illetve korrelált statin dózisfüggéssel és tacrolimus farmakokinetikával [98, 99]. Ugyanebben a vizsgálatban az rs2246709 SNP gyengén asszociált allélikus mRNS expresszió egyensúlytalansággal (al-

lelic imbalance)[15] is, bár a tanulmány szerzői szerint ez az rs35599367 SNP-vel való kapcsoltsági egyenlőtlenségnek a következménye.

A mi vizsgálataink szerint az rs35599367 SNP nem befolyásolta a betegek túlélését, ezzel szemben az rs2246709 polimorfizmusnak jelentős hatása volt. A két SNP között kapcsoltsági egyensúlytalanságot állapítottunk meg, azonban az allélgyakoriságukban tapasztalt különbségek miatt ez a kapcsolat egyirányú volt, így valószínűsíthetően a kevésbé gyakori rs35599367 nem járulhatott jelentős mértékben hozzá az rs2246709 hatásához.

Az rs2246709 hatását tekintve nagy különbségeket mértünk a nemek között a túlélési arányokban. Több tanulmányban is kimutatták, hogy a nőkben szignifikánsan nagyobb a CYP3A4 aktivitása mint férfiakban [125–127]. Érdekes módon a legalacsonyabb rizikót a férfi heterozigótaság jelentette. Ezzel szemben a vad homozigóta AA genotípus férfiakban rosszabb túlélési aránnyal asszociált, mint nőkben. Jelenleg ezeket az eredményeket nem tudjuk megmagyarázni. Elméletileg elképzelhető, hogy az rs2246709 (vagy egy vele kapcsolt, ismeretlen variáns) befolyásolja az enzim expressziójának nemtől függő szabályozását. Továbbá, feltehetően a CYP3A4 kemoterápiában betöltött komplex szerepe miatt bármelyik homozigóta genotípus hordozása rosszabb kimenetellel jár férfiakban a terápiára adott válasz szempontjából.

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. Ez az eredmény arra is rávilágít, hogy a gén-gén interakciók vizsgálata és döntéstámogató rendszerekben való felhasználása bonyolult és nagy mennyiségű adatot igényel. Tudjuk ugyanis azt is, hogy a páciens neme is befolyásolja a rs2246709 polimorfizmus hatását, de a *nem*mel való interakció figyelembevétele már 18 alcsoportot eredményez, amelynek

[15] Allélikus egyensúlytalanság: Adott gén két alléljának egymáshoz képesti expressziós szintje ugyanabban a mintában mérve. Ezzel a technikával *cisz*-regulációs SNP-ket lehet azonosítani.

nagy része csak kevés pácienst foglal magában, és emiatt a statisztikai elemezhetősége problémákba ütközik.

Ezzel szemben 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 vizsgálataink erejét többek között az is adja, hogy szemben más kutatócsoportok elemzéseivel, amelyek 100 vagy annál kevesebb páciens adatai alapján vontak le következtetéseket [128–130], a mi elemzésünk jelentősen nagyobb méretű mintapopuláción alapult. Az elemzésnek emellett azonban van néhány gyenge pontja is. Bár a mintáink a teljes populációban megfigyelhetővel azonos arányban tartalmaztak relapszusos betegeket; azok a páciensek, akik a kemoterápia során terápia-rezisztens progresszív betegségben vagy a terápia miatti fertőzés vagy toxicitás fellépése miatt elhunytak, alulreprezentáltak voltak a minták között. Továbbá az adatok részleges hiánya miatt az egyes elemzések során bizonyos esetekben eltérő mintaszámokkal kellett dolgoznunk.

Fontos azonban megjegyezni, hogy az rs2246709 SNP nagyon gyakori a kaukázusi populációban, a heterozigóták aránya ugyanis 40% körüli [99] (illetve lásd a következő genomi adatbázisokban: ALFRED, http://alfred.med.yale.edu/alfred/recordinfo.asp?UNID=SI317310P Hozzáférés: 2014.05.29. és dbSNP Home Page, Reference SNP (refSNP) Cluster Report: rs2246709, http://www.ncbi.nlm.nih.gov/projects/SNP/snp_ref.cgi?rs=2246709 Hozzáférés: 2014.05.29.). Ez azt jelenti, hogy szemben a sokkal ritkábban előforduló funkcionális SNP-kkel (pl. rs35599367 minor allél frekvenciája 4,4% körüli), az SNP gyenge hatásának is jelentős hatása lehet a populáció szintjén. Amennyiben az eredményeink más populációkban vagy vizsgálatokban is megerősítésre kerülnek, akkor a *CYP3A4* gén rs2246709 polimorfizmusának genotipizálása akár a jövőbeli kockázatbecslő modellek és a személyre szabott ALL terápia részévé is válhat.

5.4. A bayesi relevanciaelemzési módszertan alkalmazási lehetőségeinek vizsgálata asszociációs vizsgálatokban

A bayesi relevanciaelemzést két jelölt gén asszociációs vizsgálatban alkalmaztuk a gyermekkori ALL hajlamosító tényezőinek felderítésére, szintén egy jelölt gén asszociációs

vizsgálatban a *CYP3A4* gén egyes gyakori polimorfizmusainak az ALL túlélését befolyásoló hatásainak elemzésére, illetve egy parciális genomszűrési vizsgálatban az asztma genetikai hajlamosító tényezőinek tanulmányozására.

Az elemzések során a bayesi relevanciaelemzési módszer több előnyös tulajdonságára 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.

A módszernek mindezek mellett további előnyei is vannak, amelyeket ezekben a vizsgálatokban nem tudtunk kihasználni, de a következőkben röviden összefoglalom ezeket. A módszer a bayesi megközelítésből eredően lehetőséget nyújt *a priori* információk figyelembe vételére is, amelyek akár a kutató előzetes ismereteiből, tudásbázisokból, publikus adatbázisokból, korábbi kísérleti eredményekből stb. származhatnak. A bayesi hatáserősség-elemzés a hatáserősség *a posteriori* valószínűségi eloszlásának kiszámításán keresztül a hihetőségi tartományok ábrázolásával a hagyományos frekventista statisztikában megszokott konfidencia-intervallumnál jóval részletgazdagabb elemzésre ad lehetőséget. Végül a módszer alapját adó valószínűségi megközelítés annak a feladatnak a normatív megoldását is lehetővé teszi, hogy a polimorfizmusokra kiszámított eredményeket magasabb szintekre, például génekre, szabályozási útvonalakra, fehérjekomplexekre is kiterjesszük.

DOI:10.14753/SE.2017.1938

6. 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.

DOI:10.14753/SE.2017.1938

7. Összefoglalás

Az új generációs szekvenálási technológiák használhatósága nagy mértékben a hívott variánsok pontosságán alapul, amely a munkafolyamat komplexitása miatt nem minden esetben felel meg az elvárásoknak. Más kutatócsoportok eredményével összhangban megmutattuk, hogy a jelenleg gyakran használt variánskivonatoló módszerek eredménye sok esetben különbözik egymástól, és nincs olyan legjobb módszer, amely minden körülmények között felülmúlná a többit.

Kifejlesztettem egy új módszert (VariantMetaCaller), amely a variánsok minőségét jellemző annotációs adatok felhasználásával kombinálja a különböző variánshívási eredményeket, és megbecsüli a variánsok valódiságának valószínűségét. A VariantMetaCaller minden általam vizsgált genomi régió méret esetén szignifikánsan jobb teljesítményt nyújtott, mint a bemenetéül szolgáló variánskivonatolási módszerek, és felülmúlt egy alternatív fúziós szoftvert is, amely nem használt annotációs információkat a kombináció során. A variánsok valószínűségének pontos megbecslésével a VariantMetaCaller lehetőséget ad a variánsok kvantitatív, *várható precizitás* alapú szűrésére is, akár olyan alkalmazási területeken is, ahol eddig csak manuális szűrést lehetett használni.

A genetikai variánsok elemzése központi jelentőségű a betegségek patomechanizmusának feltárásában és eredményesebb kezelési eljárások kidolgozásában. A gyermekkori akut limfoid leukémia (ALL) hajlamát és túlélését befolyásoló genetikai és fenotípusos tényezőket három jelölt gén asszociációs vizsgálatban tanulmányoztuk. Ennek során megmutattuk, hogy az egyik legfontosabb gyógyszer-metabolizáló enzim, a CYP3A4 génjének rs2246709 polimorfizmusa szignifikánsan befolyásolta az ALL kemoterápiás kezelése túlélésének kockázatát. Ez utóbbi polimorfizmus és a páciens neme között egy erős interakciós hatást azonosítottunk. A hatást leíró egyszerű szabályok segítségével egy olyan új rizikócsoport-besorolást hoztunk létre, amely az eredetihez képest szignifikánsan jobb kockázatbecslést tett lehetővé. Az elemzések során megmutattam a bayesi relevanciaelemzési módszer használhatóságát és a frekventista módszerekkel szembeni előnyeit. A bayesi relevanciaelemzési módszertan a változók direkt és tranzitív hatásainak megkülönböztetésével, a különféle kapcsolati típusok definiálásával, illetve az interakciók és redundanciák feltérképezésével jóval részletgazdagabb elemzést tesz lehetővé.

DOI:10.14753/SE.2017.1938

8. Summary

The lower than expected accuracy of variant calling methods still poses a challenge for the wide-spread application of next-generation sequencing in research and clinical practice. Our results showed in line with several studies that (1) currently there is no single best general individual variant calling method at all circumstances, and (2) there are significant discrepancies between commonly used variant calling pipelines, even when applied to the same set of sequence data.

We developed a novel method, VariantMetaCaller, which combines annotation information from various variant callers and predicts the probability that a variant is a true genetic variant and not a sequencing artefact. VariantMetaCaller had significantly higher sensitivity and precision than the individual variant callers in all target region sizes, ranging from a few hundred kilobases to whole exomes. The novel method outperformed an alternative fusion method, BAYSIC, as well, which does not utilize annotation information. We also demonstrated that VariantMetaCaller supports a quantitative, precision based filtering of variants under wider conditions, therefore it can be a viable alternative to hard filtering.

Analyzing genetic variants is of prime importance in the dissection of the pathomechanism of diseases and in developing efficient therapies. We analyzed the genetic and phenotypic factors that influence the risk and the survival of acute lymphoid leukemia (ALL) in three candidate gene association studies. The rs2246709 polymorphism in the gene of an important drug metabolizing enzyme, CYP3A4, significantly influenced the survival rate of chemotherapy for ALL. The polymorphism and the gender of the patient had a strong interacting effect. We calculated new risk assessments involving this interaction and showed that it significantly outperformed the earlier risk-group assessments. During the analyses, we showed the superiority of Bayesian relevance analysis methodology over traditional frequentist methods. We showed that the Bayesian approach enables a more detailed analysis by the distinction of direct and transitive effects, by the definition of different dependency types, and by the automated discovery of interaction and redundancies between variables.

9. Hivatkozások

[1] Boser BE, Guyon IM, Vapnik VN., A Training Algorithm for Optimal Margin Classifiers. In: Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory, 144–152. 1992.

[2] Schölkopf B, Tsuda K, Vert JP., Kernel Methods in Computational Biology, The MIT Press, London, 2004, 71–92.

[3] Lodhi H. (2012) Computational biology perspective: Kernel methods and deep learning. WIREs Comput Stat, 4 (5): 455–465.

[4] Ben-Hur A, Ong CS, Sonnenburg S, Schölkopf B, Rätsch G. (2008) Support vector machines and kernels for computational biology. PLoS Comput Biol, 4 (10): e1000173.

[5] Balding DJ. (2006) A tutorial on statistical methods for population association studies. Nat Rev Genet, 7 (10): 781–791.

[6] Storey JD, Tibshirani R. (2003) Statistical significance for genomewide studies. Proc Natl Acad Sci U S A, 100 (16): 9440–5.

[7] Cordell HJ, Clayton DG. (2005) Genetic association studies. Lancet, 366 (9491): 1121–1131.

[8] Stephens M, Balding DJ. (2009) Bayesian statistical methods for genetic association studies. Nat Rev Genet, 10 (10): 681–690.

[9] Sham PC, Purcell SM. (2014) Statistical power and significance testing in large-scale genetic studies. Nat Rev Genet, 15 (5): 335–46.

[10] Benjamini Y, Hochberg Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Roy Statist Soc Ser B, 57 (1): 289–300.

[11] Benjamini Y. (2010) Discovering the false discovery rate. J Roy Statist Soc Ser B, 72 (4): 405–416.

[12] Ellison AM. (2004) Bayesian inference in ecology. Ecol Lett, 7 (6): 509–520.

[13] Shoemaker JS, Painter IS, Weir BS. (1999) Bayesian statistics in genetics: A guide

for the uninitiated. Trends Genet, 15 (9): 354–358.

[14] Beaumont MA, Rannala B. (2004) The Bayesian revolution in genetics. Nat Rev Genet, 5 (4): 251–261.

[15] O'Hara RB, Cano JM, Ovaskainen O, Teplitsky C, Alho JS. (2008) Bayesian approaches in evolutionary quantitative genetics. J Evol Biol, 21 (4): 949–957.

[16] Lunn DJ, Whittaker JC, Best N. (2006) A Bayesian toolkit for genetic association studies. Genet Epidemiol, 30 (3): 231–247.

[17] Guan Y, Stephens M. (2011) Bayesian variable selection regression for genome-wide association studies and other large-scale problems. Ann Appl Stat, 5 (3): 1780–1815.

[18] Li J, Das K, Fu G, Li R, Wu R. (2011) The Bayesian lasso for genome-wide association studies. Bioinformatics, 27 (4): 516–523.

[19] Marttinen P, Corander J. (2010) Efficient Bayesian approach for multilocus association mapping including gene-gene interactions. BMC Bioinformatics, 11: 443.

[20] Fridley BL. (2009) Bayesian Variable and model selection methods for genetic association studies. Genet Epidemiol, 33 (1): 27–37.

[21] Fridley BL, Serie D, Jenkins G, White K, Bamlet W, Potter JD, Goode EL. (2010) Bayesian mixture models for the incorporation of prior knowledge to inform genetic association studies. Genet Epidemiol, 34 (5): 418–426.

[22] Yi N, Zhi D. (2011) Bayesian analysis of rare variants in genetic association studies. Genet Epidemiol, 35 (1): 57–69.

[23] Hunyadi L. (2011) Bayes gondolkodás a statisztikában. Statisztikai szemle, 89 (10-11): 1150–1171.

[24] Efron B. (2013) Bayes' Theorem in the 21st Century. Science, 340 (6137): 1177–1178.

[25] Clark TG, Bradburn MJ, Love SB, Altman DG. (2003) Survival analysis part I: basic concepts and first analyses. Br J Cancer, 89 (2): 232–238.

[26] Bradburn MJ, Clark TG, Love SB, Altman DG. (2003) Survival Analysis Part II: Multivariate data analysis – an introduction to concepts and methods. Br J Cancer,

89 (3): 431–436.

[27] Bradburn MJ, Clark TG, Love SB, Altman DG. (2003) Survival analysis Part III: multivariate data analysis – choosing a model and assessing its adequacy and fit. Br J Cancer, 89 (4): 605–611.

[28] Clark TG, Bradburn MJ, Love SB, Altman DG. (2003) Survival analysis part IV: further concepts and methods in survival analysis. Br J Cancer, 89 (5): 781–786.

[29] Kaplan EL, Meier P. (1958) Nonparametric Estimation from Incomplete Observations. J Amer Statist Assoc, 53 (282): 457–481.

[30] Cox DR. (1972) Regression models and life tables. J Roy Statist Soc Ser B, 34 (2): 187–220.

[31] Gonzaga-Jauregui C, Lupski JR, Gibbs RA. (2012) Human Genome Sequencing in Health and Disease. Annu Rev Med, 63 (1): 35–61.

[32] Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, Boutell JM, Bryant J, Carter RJ, Keira Cheetham R, Cox AJ, Ellis DJ, Flatbush MR, Gormley NA, Humphray SJ, Irving LJ, Karbelashvili MS, Kirk SM, Li H, Liu X, Maisinger KS, Murray LJ, Obradovic B, Ost T, Parkinson ML, Pratt MR, Rasolonjatovo IMJ, Reed MT, Rigatti R, Rodighiero C, Ross MT, Sabot A, Sankar SV, Scally A, Schroth GP, Smith ME, Smith VP, Spiridou A, Torrance PE, Tzonev SS, Vermaas EH, Walter K, Wu X, Zhang L, Alam MD, Anastasi C, Aniebo IC, Bailey DMD, Bancarz IR, Banerjee S, Barbour SG, Baybayan PA, Benoit VA, Benson KF, Bevis C, Black PJ, Boodhun A, Brennan JS, Bridgham JA, Brown RC, Brown AA, Buermann DH, Bundu AA, Burrows JC, Carter NP, Castillo N, Chiara E Catenazzi M, Chang S, Neil Cooley R, Crake NR, Dada OO, Diakoumakos KD, Dominguez-Fernandez B, Earnshaw DJ, Egbujor UC, Elmore DW, Etchin SS, Ewan MR, Fedurco M, Fraser LJ, Fuentes Fajardo KV, Scott Furey W, George D, Gietzen KJ, Goddard CP, Golda GS, Granieri PA, Green DE, Gustafson DL, Hansen NF, Harnish K, Haudenschild CD, Heyer NI, Hims MM, Ho JT, Horgan AM, Hoschler K, Hurwitz S, Ivanov DV, Johnson MQ, James T, Huw Jones TA, Kang GD, Kerelska TH, Kersey AD, Khrebtukova I, Kindwall AP, Kingsbury Z, Kokko-Gonzales PI, Kumar A, Laurent MA, Lawley

CT, Lee SE, Lee X, Liao AK, Loch JA, Lok M, Luo S, Mammen RM, Martin JW, McCauley PG, McNitt P, Mehta P, Moon KW, Mullens JW, Newington T, Ning Z, Ling Ng B, Novo SM, O'Neill MJ, Osborne MA, Osnowski A, Ostadan O, Paraschos LL, Pickering L, Pike AC, Pike AC, Chris Pinkard D, Pliskin DP, Podhasky J, Quijano VJ, Raczy C, Rae VH, Rawlings SR, Chiva Rodriguez A, Roe PM, Rogers J, Rogert Bacigalupo MC, Romanov N, Romieu A, Roth RK, Rourke NJ, Ruediger ST, Rusman E, Sanches-Kuiper RM, Schenker MR, Seoane JM, Shaw RJ, Shiver MK, Short SW, Sizto NL, Sluis JP, Smith MA, Ernest Sohna Sohna J, Spence EJ, Stevens K, Sutton N, Szajkowski L, Tregidgo CL, Turcatti G, Vandevondele S, Verhovsky Y, Virk SM, Wakelin S, Walcott GC, Wang J, Worsley GJ, Yan J, Yau L, Zuerlein M, Rogers J, Mullikin JC, Hurles ME, McCooke NJ, West JS, Oaks FL, Lundberg PL, Klenerman D, Durbin R, Smith AJ. (2008) Accurate whole human genome sequencing using reversible terminator chemistry. Nature, 456 (7218): 53–59.

[33] Ng PC, Levy S, Huang J, Stockwell TB, Walenz BP, Li K, Axelrod N, Busam DA, Strausberg RL, Venter JC. (2008) Genetic variation in an individual human exome. PLoS Genet, 4 (8): e1000160.

[34] Robinson P, Krawitz P, Mundlos S. (2011) Strategies for exome and genome sequence data analysis in disease-gene discovery projects. Clin Genet, 80 (2): 127–132.

[35] Schadt EE, Linderman MD, Sorenson J, Lee L, Nolan GP. (2010) Computational solutions to large-scale data management and analysis. Nat Rev Genet, 11 (9): 647–657.

[36] Pabinger S, Dander A, Fischer M, Snajder R, Sperk M, Efremova M, Krabichler B, Speicher MR, Zschocke J, Trajanoski Z. (2014) A survey of tools for variant analysis of next-generation genome sequencing data. Brief Bioinform, 15 (2): 256–78.

[37] Sboner A, Mu X, Greenbaum D, Auerbach RK, Gerstein MB. (2011) The real cost of sequencing: higher than you think! Genome Biol, 12 (8): 125.

[38] Park MH, Rhee H, Park JH, Woo HM, Choi BO, Kim BY, Chung KW, Cho YB, Kim HJ, Jung JW, Koo SK. (2014) Comprehensive analysis to improve the vali-

dation rate for single nucleotide variants detected by next-generation sequencing. PloS One, 9 (1): e86664.

[39] Yu X, Sun S. (2013) Comparing a few SNP calling algorithms using low-coverage sequencing data. BMC Bioinformatics, 14 (1): 274.

[40] Liu X, Han S, Wang Z, Gelernter J, Yang BZ. (2013) Variant callers for next-generation sequencing data: a comparison study. PloS One, 8 (9): e75619.

[41] Neuman JA, Isakov O, Shomron N. (2013) Analysis of insertion-deletion from deep-sequencing data: software evaluation for optimal detection. Brief Bioinform, 14 (1): 46–55.

[42] O'Rawe J, Jiang T, Sun G, Wu Y, Wang W, Hu J, Bodily P, Tian L, Hakonarson H, Johnson WE, Wei Z, Wang K, Lyon GJ. (2013) Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. Genome Med, 5 (3): 28.

[43] Pirooznia M, Kramer M, Parla J, Goes FS, Potash JB, McCombie WR, Zandi PP. (2014) Validation and assessment of variant calling pipelines for next-generation sequencing. Hum Genomics, 8 (1): 14.

[44] Dai M, Thompson RC, Maher C, Contreras-Galindo R, Kaplan MH, Markovitz DM, Omenn G, Meng F. (2010) NGSQC: cross-platform quality analysis pipeline for deep sequencing data. BMC Genomics, 11 Suppl 4: S7.

[45] Andrews S. (2010) FastQC: A quality control tool for high throughput sequence data.
URL http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

[46] Schmieder R, Edwards R. (2011) Quality control and preprocessing of metagenomic datasets. Bioinformatics, 27 (6): 863–864.

[47] Li H, Homer N. (2010) A survey of sequence alignment algorithms for next-generation sequencing. Brief Bioinform, 11 (5): 473–83.

[48] McGinnis S, Madden TL. (2004) BLAST: At the core of a powerful and diverse set of sequence analysis tools. Nucleic Acids Res, 32 (Web Server issue): W20–25.

[49] Li H, Ruan J, Durbin R. (2008) Mapping short DNA sequencing reads and calling

variants using mapping quality scores. Genome Res, 18 (11): 1851–1858.

[50] Lee WP, Stromberg MP, Ward A, Stewart C, Garrison EP, Marth GT. (2014) MOSAIK: A hash-based algorithm for accurate next-generation sequencing short-read mapping. PLoS One, 9 (3): e90581.

[51] Li H, Durbin R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics, 25 (14): 1754–60.

[52] Langmead B, Salzberg SL. (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods, 9 (4): 357–9.

[53] McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res, 20 (9): 1297–303.

[54] Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, Banks E, Garimella KV, Altshuler D, Gabriel S, DePristo MA., From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. In: A Bateman, WR Pearson, LD Stein, GD Stormo, JR Yates (Szerk.), Current protocols in bioinformatics, John Wiley & Sons, Inc., Hoboken, NJ, USA, 2013, 11.10.1–11.10.33.

[55] Warden CD, Adamson AW, Neuhausen SL, Wu X. (2014) Detailed comparison of two popular variant calling packages for exome and targeted exon studies. PeerJ, 2: e600.

[56] Liu Q, Guo Y, Li J, Long J, Zhang B, Shyr Y. (2012) Steps to ensure accuracy in genotype and SNP calling from Illumina sequencing data. BMC Genomics, 13 Suppl 8: S8.

[57] Li H. (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics, 27 (21): 2987–93.

[58] Garrison E, Marth G. (2012) Haplotype-based variant detection from short-read sequencing. arXiv, 9. 1207.3907.

[59] Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, Bertoni A, Swerdlow HP, Gu Y. (2012) A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics, 13 (1): 341.

[60] Lam HYK, Pan C, Clark MJ, Lacroute P, Chen R, Haraksingh R, O'Huallachain M, Gerstein MB, Kidd JM, Bustamante CD, Snyder M. (2012) Detecting and annotating genetic variations using the HugeSeq pipeline. Nat Biotechnol, 30 (3): 226–9.

[61] Cantarel BL, Weaver D, McNeill N, Zhang J, Mackey AJ, Reese J. (2014) BAYSIC: a Bayesian method for combining sets of genome variants with improved specificity and sensitivity. BMC Bioinformatics, 15 (1): 104.

[62] Wang K, Li M, Hakonarson H. (2010) ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res, 38 (16): e164.

[63] McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F. (2010) Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics, 26 (16): 2069–70.

[64] Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly, 6 (2): 80–92.

[65] McCarthy DJ, Humburg P, Kanapin A, Rivas MA, Gaulton K, Cazier JB, Donnelly P. (2014) Choice of transcripts and software has a large effect on variant annotation. Genome Med, 6 (3): 26.

[66] Cooper GM, Shendure J. (2011) Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. Nat Rev Genet, 12 (9): 628–640.

[67] Ng PC, Henikoff S. (2003) SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res, 31 (13): 3812–3814.

[68] Adzhubei I, Jordan DM, Sunyaev SR. (2013) Predicting functional effect of hu-

man missense mutations using PolyPhen-2. Curr Protoc Hum Genet, Chapter 7: Unit7.20.

[69] Schwarz JM, Rödelsperger C, Schuelke M, Seelow D. (2010) MutationTaster evaluates disease-causing potential of sequence alterations. Nat Methods, 7 (8): 575–576.

[70] Liu X, Jian X, Boerwinkle E. (2011) dbNSFP: A lightweight database of human nonsynonymous SNPs and their functional predictions. Hum Mutat, 32 (8): 894–899.

[71] Dixon SJ, Costanzo M, Baryshnikova A, Andrews B, Boone C. (2009) Systematic mapping of genetic interaction networks. Annu Rev Genet, 43: 601–25.

[72] Costanzo M, Baryshnikova A, Bellay J, Kim Y, Spear ED, Sevier CS, Ding H, Koh JLY, Toufighi K, Mostafavi S, Prinz J, St Onge RP, VanderSluis B, Makhnevych T, Vizeacoumar FJ, Alizadeh S, Bahr S, Brost RL, Chen Y, Cokol M, Deshpande R, Li Z, Lin ZY, Liang W, Marback M, Paw J, San Luis BJ, Shuteriqi E, Tong AHY, van Dyk N, Wallace IM, Whitney JA, Weirauch MT, Zhong G, Zhu H, Houry WA, Brudno M, Ragibizadeh S, Papp B, Pál C, Roth FP, Giaever G, Nislow C, Troyanskaya OG, Bussey H, Bader GD, Gingras AC, Morris QD, Kim PM, Kaiser CA, Myers CL, Andrews BJ, Boone C. (2010) The genetic landscape of a cell. Science, 327 (5964): 425–31.

[73] Pearl J., Probabilistic Reasoning in Intelligent Systems, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1988, 117.

[74] Cooper GF, Herskovits E. (1992) A bayesian method for the induction of probabilistic networks from data. Machine Learning, 9: 309–347.

[75] Liu JS., Monte Carlo Strategies in Scientific Computing, Springer-Verlag, New York, USA, 2004.

[76] Madigan D, Andersson SA, Perlman M, Volinsky CT. (1996) Bayesian model averaging and model selection for markov equivalence classes of acyclic digraphs. Commun Stat Theory Methods, 25 (11): 2493–2520.

[77] Giudici P, Castelo R. (2003) Improving markov chain monte carlo model search

for data mining. Machine Learning, 50: 127–158.

[78] Lautner-Csorba O. (2013) A gyermekkori akut limfoid leukémiára való hajlam genetikai, valamint a kezelésre adott válasz farmakogenetikai vizsgálata. Ph.D. értekezés, Semmelweis Egyetem, Molekuláris Orvostudományok Doktori Iskola.

[79] Magyarosy E. (2000) A gyermekkori akut limfoblasztos leukémia kezelésében elért hazai eredmények. Magy Onkol, 44 (4): 255–259.

[80] Schrappe M. (2004) Evolution of BFM trials for childhood ALL. Ann Hematol, 83 Suppl 1: S121–S123.

[81] Greaves MF. (1997) Aetiology of acute leukaemia. Lancet, 349 (9048): 344–349.

[82] McNally RJQ, Eden TOB. (2004) An infectious aetiology for childhood acute leukaemia: A review of the evidence. Br J Haematol, 127 (3): 243–263.

[83] Eden T. (2010) Aetiology of childhood leukaemia. Cancer Treat Rev, 36 (4): 286–297.

[84] Greaves M. (2006) Infection, immune responses and the aetiology of childhood leukaemia. Nat Rev Cancer, 6 (3): 193–203.

[85] Lee KM, Ward MH, Han S, Ahn HS, Kang HJ, Choi HS, Shin HY, Koo HH, Seo JJ, Choi JE, Ahn YO, Kang D. (2009) Paternal smoking, genetic polymorphisms in CYP1A1 and childhood leukemia risk. Leuk Res, 33 (2): 250–258.

[86] Infante-Rivard C, Krajinovic M, Labuda D, Sinnett D. (2002) Childhood acute lymphoblastic leukemia associated with parental alcohol consumption and polymorphisms of carcinogen-metabolizing genes. Epidemiology, 13 (3): 277–281.

[87] MacArthur AC, McBride ML, Spinelli JJ, Tamaro S, Gallagher RP, Theriault G. (2008) Risk of childhood leukemia associated with parental smoking and alcohol consumption prior to conception and during pregnancy: the cross-Canada childhood leukemia study. Cancer Causes Control, 19 (3): 283–295.

[88] Milne E, Greenop KR, Scott RJ, Bailey HD, Attia J, Dalla-Pozza L, De Klerk NH, Armstrong BK. (2012) Parental prenatal smoking and risk of childhood acute lymphoblastic leukemia. Amer J Epidemiol, 175 (1): 43–53.

[89] Infante-Rivard C, Labuda D, Krajinovic M, Sinnett D. (1999) Risk of childhood

leukemia associated with exposure to pesticides and with gene polymorphisms. Epidemiology, 10 (5): 481–487.

[90] Infante-Rivard C. (2003) Diagnostic x rays, DNA repair genes and childhood acute lymphoblastic leukemia. Health Phys, 85 (1): 60–64.

[91] Chokkalingam AP, Bartley K, Wiemels JL, Metayer C, Barcellos LF, Hansen HM, Aldrich MC, Guha N, Urayama KY, Scélo G, Chang JS, Month SR, Wiencke JK, Buffler Pa. (2011) Haplotypes of DNA repair and cell cycle control genes, X-ray exposure, and risk of childhood acute lymphoblastic leukemia. Cancer Causes Control, 22 (12): 1721–30.

[92] Papaemmanuil E, Hosking FJ, Vijayakrishnan J, Price A, Olver B, Sheridan E, Kinsey SE, Lightfoot T, Roman E, Irving JAE, Allan JM, Tomlinson IP, Taylor M, Greaves M, Houlston RS. (2009) Loci on 7p122, 10q212 and 14q112 are associated with risk of childhood acute lymphoblastic leukemia. Nat Genet, 41 (9): 1006–1010.

[93] Sherborne AL, Hosking FJ, Prasad RB, Kumar R, Koehler R, Vijayakrishnan J, Papaemmanuil E, Bartram CR, Stanulla M, Schrappe M, Gast A, Dobbins SE, Ma Y, Sheridan E, Taylor M, Kinsey SE, Lightfoot T, Roman E, Irving JAE, Allan JM, Moorman AV, Harrison CJ, Tomlinson IP, Richards S, Zimmermann M, Szalai C, Semsei AF, Erdelyi DJ, Krajinovic M, Sinnett D, Healy J, Gonzalez Neira A, Kawamata N, Ogawa S, Koeffler HP, Hemminki K, Greaves M, Houlston RS. (2010) Variation in CDKN2A at 9p213 influences childhood acute lymphoblastic leukemia risk. Nat Genet, 42 (6): 492–494.

[94] Prasad RB, Hosking FJ, Vijayakrishnan J, Papaemmanuil E, Koehler R, Greaves M, Sheridan E, Gast A, Kinsey SE, Lightfoot T, Roman E, Taylor M, Pritchard-Jones K, Stanulla M, Schrappe M, Bartram CR, Houlston RS, Kumar R, Hemminki K. (2010) Verification of the susceptibility loci on 7p122, 10q212, and 14q112 in precursor B-cell acute lymphoblastic leukemia of childhood. Blood, 115 (9): 1765–1767.

[95] Vijayakrishnan J, Sherborne AL, Sawangpanich R, Hongeng S, Houlston RS, Pakakasama S. (2010) Variation at 7p122 and 10q212 influences childhood acute

lymphoblastic leukemia risk in the Thai population and may contribute to racial differences in leukemia incidence. Leuk Lymphoma, 51 (10): 1870–1874.

[96] Treviño LR, Yang W, French D, Hunger SP, Carroll WL, Devidas M, Willman C, Neale G, Downing J, Raimondi SC, Pui CH, Evans WE, Relling MV. (2009) Germline genomic variants associated with childhood acute lymphoblastic leukemia. Nat Genet, 41 (9): 1001–1005.

[97] Davidsen ML, Dalhoff K, Schmiegelow K. (2008) Pharmacogenetics Influence Treatment Efficacy in Childhood Acute Lymphoblastic Leukemia. J Pediatr Hematol Oncol, 30 (11): 831–849.

[98] Elens L, Bouamar R, Hesselink DA, Haufroid V, Van Der Heiden IP, Van Gelder T, Van Schaik RHN. (2011) A new functional CYP3A4 intron 6 polymorphism significantly affects tacrolimus pharmacokinetics in kidney transplant recipients. Clin Chem, 57 (11): 1574–1583.

[99] Wang D, Guo Y, Wrighton SA, Cooke GE, Sadee W. (2011) Intronic polymorphism in CYP3A4 affects hepatic expression and response to statin drugs. Pharmacogenomics J, 11 (4): 274–286.

[100] Watanabe M, Kumai T, Matsumoto N, Tanaka M, Suzuki S, Satoh T, Kobayashi S. (2004) Expression of CYP3A4 mRNA is correlated with CYP3A4 protein level and metabolic activity in human liver. J Pharmacol Sci, 94 (4): 459–62.

[101] Ozdemir V, Kalow W, Tang BK, Paterson AD, Walker SE, Endrenyi L, Kashuba AD. (2000) Evaluation of the genetic component of variability in CYP3A4 activity: a repeated drug administration method. Pharmacogenetics, 10 (5): 373–388.

[102] Lamba JK, Lin YS, Schuetz EG, Thummel KE. (2002) Genetic contribution to variable human CYP3A-mediated metabolism. Adv Drug Deliv Rev, 54 (10): 1271–94.

[103] Lamba V, Panetta JC, Strom S, Schuetz EG. (2010) Genetic predictors of interindividual variability in hepatic CYP3A4 expression. J Pharmacol Exp Ther, 332 (3): 1088–1099.

[104] Huang W, Li L, Myers JR, Marth GT. (2012) ART: a next-generation sequencing

read simulator. Bioinformatics, 28 (4): 593–4.

[105] DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet, 43 (5): 491–8.

[106] Chang CC, Lin CJ. (2011) LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems Technology, 2: 27:1–27:27. Szoftver elérhető: http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[107] Lin HT, Lin CJ, Weng RC. (2007) A note on Platt's probabilistic outputs for support vector machines. Machine Learning, 68 (3): 267–276.

[108] Tuszynski J. (2014) caTools: Tools: moving window statistics, GIF, Base64, ROC AUC, etc. R package version 1.17.1.
URL http://CRAN.R-project.org/package=caTools

[109] R Development Core Team R. (2011). R: A Language and Environment for Statistical Computing.
URL http://www.r-project.org

[110] Therneau T. (2012) A Package for Survival Analysis in S. R package version.
URL http://cran.r-project.org/package=survival

[111] Dardis C. (2013) survMisc: Miscellaneous functions for survival data.
URL http://cran.r-project.org/package=survMisc

[112] Harrell FE, Califf RM, Pryor DB, Lee KL, Rosati RA. (1982) Evaluating the yield of medical tests. JAMA, 247 (18): 2543–2546.

[113] Mogensen UB, Ishwaran H, Gerds TA. (2010) Evaluating random forests for survival analysis using prediction error curves. Technical report.
URL http://www.jstatsoft.org/v50/i11

[114] Elens L, van Gelder T, Hesselink DA, Haufroid V, van Schaik RHN. (2013) CYP3A4*22: promising newly identified CYP3A4 variant allele for personalizing pharmacotherapy. Pharmacogenomics, 14 (1): 47–62.

[115] Ungvári I. (2014) A gyermekkori asthma patomechanizmusát befolyásoló genetikai variációk és gén-környezet interakciók vizsgálata. Ph.D. értekezés, Semmelweis Egyetem, Molekuláris Orvostudományok Doktori Iskola.

[116] Meynert AM, Ansari M, FitzPatrick DR, Taylor MS. (2014) Variant detection sensitivity and biases in whole genome and exome sequencing. BMC Bioinformatics, 15 (1): 247.

[117] Zhao H, Yang Y, Lin H, Zhang X, Mort M, Cooper DN, Liu Y, Zhou Y. (2013) DDIG-in: discriminating between disease-associated and neutral non-frameshifting micro-indels. Genome Biol, 14 (3): R23.

[118] Krawitz P, Rödelsperger C, Jäger M, Jostins L, Bauer S, Robinson PN. (2010) Microindel detection in short-read sequence data. Bioinformatics, 26 (6): 722–9.

[119] Highnam G, Wang JJ, Kusler D, Zook J, Vijayan V, Leibovich N, Mittelman D. (2015) An analytical framework for optimizing variant discovery from personal genomes. Nat Commun, 6: 6275.

[120] Hatem A, Bozdağ D, Toland AE, Çatalyürek ÜV. (2013) Benchmarking short sequence mapping tools. BMC Bioinformatics, 14 (1): 184.

[121] Li H. (2014) Toward better understanding of artifacts in variant calling from high-coverage samples. Bioinformatics, 30 (20): 2843–2851.

[122] Boyd K, Eng KH, Page CD., Area under the Precision-Recall Curve: Point Estimates and Confidence Intervals. In: H Blockeel, K Kersting, S Nijssen, F Železný (Szerk.), Machine Learning and Knowledge Discovery in Databases, Springer Berlin Heidelberg, Berlin, Heidelberg, 2013, 451–466.

[123] Baumhäkel M, Kasel D, Rao-Schymanski RA, Böcker R, Beckurts KT, Zaigler M, Barthold D, Fuhr U. (2001) Screening for inhibitory effects of antineoplastic agents on CYP3A4 in human liver microsomes. Int J Clin Pharmacol Ther, 39 (12): 517–528.

[124] Lindley C, Hamilton G, McCune JS, Faucette S, Shord SS, Hawke RL, Wang H, Gilbert D, Jolley S, Yan B, Lecluyse EL. (2002) The effect of cyclophosphamide with and without dexamethasone on cytochrome P450 3A4 and 2B6 in human

hepatocytes. Drug Metab Dispos, 30 (7): 814–822.

[125] Wolbold R, Klein K, Burk O, Nüssler AK, Neuhaus P, Eichelbaum M, Schwab M, Zanger UM. (2003) Sex is a major determinant of CYP3A4 expression in human liver. Hepatology, 38 (4): 978–988.

[126] Waxman DJ, Holloway MG. (2009) Sex differences in the expression of hepatic drug metabolizing enzymes. Mol Pharmacol, 76 (2): 215–228.

[127] Thangavel C, Boopathi E, Shapiro BH. (2011) Intrinsic sexually dimorphic expression of the principal human CYP3A4 correlated with suboptimal activation of GH/glucocorticoid-dependent transcriptional pathways in men. Endocrinology, 152 (12): 4813–4824.

[128] Jaime-Perez JC, Colunga-Pedraza PR, Gomez-Almaguer D. (2011) Is the number of blood products transfused associated with lower survival in children with acute lymphoblastic leukemia? Pediatr Blood Cancer, 57 (2): 217–223.

[129] Borssén M, Palmqvist L, Karrman K, Abrahamsson J, Behrendtz M, Heldrup J, Forestier E, Roos G, Degerman S. (2013) Promoter DNA Methylation Pattern Identifies Prognostic Subgroups in Childhood T-Cell Acute Lymphoblastic Leukemia. PLoS One, 8 (6): e65373.

[130] Sharifi MJ, Bahoush G, Zaker F, Ansari S, Rafsanjani KA, Sharafi H. (2014) Association of -24CT, 1249GA, and 3972CT ABCC2 gene polymorphisms with methotrexate serum levels and toxic side effects in children with acute lymphoblastic leukemia. Pediatr Hematol Oncol, 31 (2): 169–77.

10. 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

DOI:10.14753/SE.2017.1938

11. Köszönetnyilvánítás

Köszönettel tartozom témavezetőmnek, Prof. Dr. Szalai Csabának, akitől doktori tanulmányaim és a kutatásom során rengeteg segítséget és biztatást kaptam. Köszönöm az Intézet volt és jelenlegi vezetőjének, Prof. Dr. Falus Andrásnak és Prof. Dr. Buzás Editnek, hogy biztosították a munkám elvégzéséhez szükséges feltételeket.

Hálásan köszönöm Dr. Antal Péternek, a Budapesti Műszaki és Gazdaságtudományi Egyetem Méréstechnika és Információs Rendszerek Tanszéke docensének fáradhatatlan munkáját és segítőkészségét, szakmai iránymutatását és baráti tanácsait. Nélküle nem jutottam volna el idáig. Köszönöm a BME bioinformatikai munkacsoport további tagjainak, Millinghoffer Andrásnak, Hullám Gábornak, Marx Péternek, Sárközy Péternek, Dr. Temesi Gergelynek, Dr. Bolgár Bencének, Huszár Balázsnak, Arany Ádámnak és Hajós Gergőnek az inspiráló szakmai együttműködést és a baráti beszélgetéseket. Köszönöm Dr. Gulyás-Kovács Attilának a tanácsokat és a VariantMetaCaller cikkben nyújtott szakmai segítségét.

Köszönettel tartozom Dr. Lautner-Csorba Orsolyának, akivel nagyon sokat dolgoztunk együtt a gyermekkori akut limfoid leukémiai vizsgálatok elemzésén. Köszönöm az Intézetben dolgozó további kollégáimnak, Dr. Félné Semsei Ágnesnek, Kutszegi Nórának és Fodor Lilinek a közös munkát.

Végül hálával tartozom a feleségemnek, Zsuzsinak, a családomnak és a barátaimnak, hogy eddigi életem során mindvégig támogattak.

A doktori tanulmányaim elvégzését és a cikkek megjelenését az NKTH (Nemzeti Kutatási és Technológiai Hivatal) TECH_08-A1/2-2008-0120 által támogatott Genagrid konzorcium és az OTKA 112872 számú pályázata tette lehetővé.

DOI:10.14753/SE.2017.1938

12. Függelék

A függelék 10 táblázatot tartalmaz, amelyek a fő eredményekhez szorosan nem kötődő részeredményeket mutatnak be.

DOI:10.14753/SE.2017.1938

1. táblázat. **A VariantMetaCaller működése során felhasznált annotációk csoportosítása és rövid leírása.** Rövidítések: FB = FreeBayes, HC = HaplotypeCaller, ST = SAMtools, UG = UnifiedGenotyper

Annotáció	Leírás	SNP-k				Indelek			
		HC	UG	FB	ST	HC	UG	FB	ST
AB	Allél egyensúly a heterozigóta genotípusok esetén: 0 és 1 közötti szám a referens és alternatív allélokat tartalmazó leolvasások arányát mutatja (ref/(ref+alt))			+				+	
ABHet	Allél egyensúly a heterozigóta genotípusok esetén: (ref/(ref+alt))		+						
ABHom	Allél egyensúly a homozigóta genotípusok esetén: (A/(A+O)) ahol A az allél (ref vagy alt) és O minden más		+						
ABP	A megfigyelt allél egyensúly valószínűsége Phred-pontszámmal kifejezve Hoeffding egyenlőtlensége alapján számítva			+				+	
BaseQRankSum	Az alternatív vs. referencia allélokat tartalmazó bázisok minőségi pontszámai közötti eltérés Wilcoxon-tesztjének eredménye (Z-score)	+	+			+	+		
BasesToClosestVariant	A legközelebbi variáns távolsága	+	+	+	+	+	+	+	+
BQB	Az alternatív vs. referencia allélokat tartalmazó bázisok minőségi pontszámai közötti eltérés Wilcoxon-tesztjének eredménye				+				
ClippingRankSum	Az alternatív vs. referencia allélokat tartalmazó leolvasások végeinek levágása közötti eltérés Wilcoxon-tesztjének eredménye (Z-score)	+				+			
DP	Leolvasási mélység	+	+	+	+	+	+	+	+
EntropyCenter_15	A referencia szekvencia Shannon entrópiája a variáns 15 bázis széles környezetében	+	+	+	+	+	+	+	+
EntropyCenter_7	A referencia szekvencia Shannon entrópiája a variáns 7 bázis széles környezetében	+	+	+	+	+	+	+	+
EntropyLeft_15	A referencia szekvencia Shannon entrópiája a variáns 15 bázis széles környezetében a variánstól balra	+	+	+	+	+	+	+	+
EntropyLeft_7	A referencia szekvencia Shannon entrópiája a variáns 7 bázis széles környezetében a variánstól balra	+	+	+	+	+	+	+	+
EntropyRight_15	A referencia szekvencia Shannon entrópiája a variáns 15 bázis széles környezetében a variánstól jobbra	+	+	+	+	+	+	+	+
EntropyRight_7	A referencia szekvencia Shannon entrópiája a variáns 7 bázis széles környezetében a variánstól balra	+	+	+	+	+	+	+	+
FS	A szálirány eltérést tesztelő Fisher egzakt teszt p-értéke Phred-pontszámmal kifejezve	+	+			+	+		
GC	A variáns környéki referencia szekvencia GC tartalma		+				+		
GenotypeEntropyMean	A minták közötti genotípuseloszlások entrópiájának átlaga	+	+	+	+	+	+	+	+
GenotypeEntropySD	A minták közötti genotípuseloszlások entrópiájának szórása	+	+	+	+	+	+	+	+
HaplotypeScore	A variánspozíció konzisztenciája legfeljebb két szegregáló haplotípussal		+						
HOB	A HOM annotáció szisztematikus eltérésének mértéke				+				+
HRun	A leghosszabb homopolimer szakasz bármelyik irányban		+				+		
ICB	Inbreeding koefficiens binomiális teszt				+				+
IDV	Legfeljebb hány leolvasás támogat egy indelt								+
IMF	Legfeljebb a leolvasások hányad része támogat egy indelt								+
InbreedingCoeff	Inbreeding koefficiens binomiális teszt	+	+			+	+		
LikelihoodRankSum	Az alternatív vs. referencia allélokat tartalmazó haplotípus valószínűségek közötti eltérés Wilcoxon-tesztjének eredménye (Z-score)	+				+	+		
MQ	Illesztési minőségi pontszámok négyzetes közepe	+	+		+	+	+		+
MQB	Illesztési minőségi pontszámok eltérésének Wilcoxon-tesztje				+				
MQM	Az alternatív allélokat tartalmazó leolvasások átlagos minősége			+				+	
MQMR	A referencia allélokat tartalmazó leolvasások átlagos minősége			+				+	
MQRankSum	Illesztési minőségi pontszámok eltérésének Wilcoxon-tesztje	+	+			+	+		
MQSB	Illesztési minőségi pontszámok eltérésének vs. a száleltérés Wilcoxon-tesztje				+				+
ODDS	A legvalószínűbb vs. a második legvalószínűbb genotípuskombináció esélyhányadosának logaritmusa			+				+	
QD	Variáns megbízhatóság/Minőség per leolvasási mélység	+	+			+	+		
QUAL	Annak a valószínűsége Phred-pontszámmal kifejezve, hogy az adott variáns legalább egy minta esetén nem homozigóta vad genotípusú	+	+	+	+	+	+	+	+
SAP	Szálirány eltérés az alternatív allélokra			+				+	
SGB[1]	Szegregációs mutató				+				+
SOR	Szálirány eltérést mutató szimmetrikus esélyhányados	+	+			+	+		
SRP	Szálirány eltérés a referencia allélokra			+				+	

További referenciák

[1] http://samtools.github.io/bcftools/rd-SegBias.pdf

DOI:10.14753/SE.2017.1938

2. táblázat. **Az egyedi variánskivonatolási módszerek szenzitivitása és precizitása a szimulált adatokon.** A legjobb módszer félkövér betűtípussal van jelölve. Rövidítések: CI = konfidencia-intervallum

Variáns-típus	Illesztő-program	Leolvasási mélység	Szenzitivitás (95% CI)				Precizitás (95% CI)			
			HaplotypeCaller	UnifiedGenotyper	FreeBayes	SAMtools	HaplotypeCaller	UnifiedGenotyper	FreeBayes	SAMtools
SNP	BWA	4	0,631 (0,611-0,651)	0,675 (0,658-0,691)	0,686 (0,665-0,708)	**0,724 (0,706-0,741)**	**0,997 (0,996-0,998)**	0,994 (0,993-0,995)	0,994 (0,993-0,996)	0,967 (0,963-0,971)
		8	0,792 (0,773-0,812)	0,807 (0,789-0,825)	0,819 (0,801-0,838)	**0,84 (0,822-0,858)**	**0,996 (0,995-0,997)**	0,992 (0,991-0,993)	0,993 (0,991-0,994)	0,984 (0,981-0,987)
		12	0,854 (0,832-0,876)	0,862 (0,842-0,883)	0,87 (0,849-0,891)	**0,885 (0,866-0,904)**	**0,997 (0,996-0,997)**	0,991 (0,991-0,992)	0,992 (0,992-0,993)	0,989 (0,987-0,99)
		16	0,886 (0,869-0,903)	0,892 (0,876-0,908)	0,894 (0,879-0,909)	**0,907 (0,892-0,922)**	**0,996 (0,995-0,997)**	0,99 (0,989-0,991)	0,992 (0,991-0,994)	0,989 (0,988-0,99)
		20	0,904 (0,89-0,919)	0,909 (0,895-0,923)	0,909 (0,896-0,922)	**0,919 (0,906-0,932)**	**0,996 (0,995-0,996)**	0,99 (0,989-0,99)	0,993 (0,992-0,993)	0,99 (0,989-0,991)
		30	0,927 (0,913-0,941)	0,93 (0,916-0,944)	0,927 (0,912-0,941)	**0,935 (0,922-0,948)**	**0,996 (0,995-0,997)**	0,988 (0,987-0,99)	0,992 (0,991-0,993)	0,989 (0,987-0,991)
		40	0,936 (0,926-0,947)	0,939 (0,929-0,95)	0,934 (0,922-0,946)	**0,942 (0,931-0,953)**	**0,995 (0,994-0,996)**	0,987 (0,985-0,989)	0,992 (0,991-0,993)	0,989 (0,987-0,99)
		60	0,946 (0,935-0,957)	**0,949 (0,938-0,96)**	0,941 (0,929-0,954)	**0,949 (0,938-0,96)**	**0,994 (0,993-0,995)**	0,986 (0,984-0,988)	0,992 (0,991-0,993)	0,989 (0,988-0,99)
		100	0,954 (0,945-0,963)	**0,956 (0,946-0,966)**	0,946 (0,934-0,958)	0,954 (0,945-0,963)	**0,994 (0,993-0,995)**	0,984 (0,982-0,985)	0,992 (0,992-0,993)	0,989 (0,987-0,99)
		200	0,959 (0,95-0,968)	**0,961 (0,952-0,971)**	0,95 (0,939-0,962)	0,955 (0,946-0,965)	**0,994 (0,993-0,995)**	0,983 (0,981-0,984)	0,992 (0,991-0,993)	0,99 (0,988-0,991)
	Bowtie 2	4	0,585 (0,567-0,602)	0,634 (0,619-0,648)	0,65 (0,633-0,668)	**0,68 (0,665-0,696)**	**0,997 (0,996-0,998)**	0,992 (0,99-0,994)	0,99 (0,988-0,992)	0,967 (0,965-0,97)
		8	0,736 (0,718-0,755)	0,765 (0,746-0,784)	0,779 (0,761-0,797)	**0,8 (0,783-0,818)**	**0,998 (0,997-0,999)**	0,989 (0,988-0,99)	0,991 (0,99-0,992)	0,985 (0,983-0,987)
		12	0,794 (0,773-0,815)	0,82 (0,801-0,839)	0,831 (0,811-0,852)	**0,848 (0,83-0,867)**	**0,998 (0,998-0,999)**	0,986 (0,984-0,988)	0,989 (0,988-0,99)	0,988 (0,986-0,99)
		16	0,826 (0,809-0,843)	0,85 (0,835-0,865)	0,857 (0,844-0,871)	**0,873 (0,859-0,887)**	**0,998 (0,998-0,999)**	0,984 (0,982-0,987)	0,989 (0,987-0,99)	0,988 (0,987-0,99)
		20	0,845 (0,832-0,859)	0,868 (0,854-0,882)	0,876 (0,862-0,889)	**0,887 (0,874-0,901)**	**0,998 (0,998-0,999)**	0,983 (0,98-0,986)	0,99 (0,988-0,991)	0,989 (0,988-0,99)
		30	0,867 (0,853-0,881)	0,89 (0,876-0,904)	0,897 (0,883-0,911)	**0,904 (0,891-0,917)**	**0,999 (0,998-0,999)**	0,982 (0,979-0,984)	0,989 (0,987-0,991)	0,989 (0,987-0,99)
		40	0,877 (0,866-0,888)	0,899 (0,888-0,91)	0,906 (0,895-0,918)	**0,912 (0,901-0,923)**	**0,999 (0,998-0,999)**	0,979 (0,976-0,981)	0,988 (0,987-0,989)	0,988 (0,987-0,989)
		60	0,888 (0,875-0,9)	0,91 (0,899-0,922)	0,916 (0,903-0,928)	**0,919 (0,907-0,931)**	**0,999 (0,999-0,999)**	0,976 (0,973-0,979)	0,987 (0,986-0,989)	0,988 (0,987-0,989)
		100	0,897 (0,887-0,906)	0,919 (0,909-0,93)	**0,924 (0,912-0,936)**	**0,924 (0,915-0,934)**	**0,999 (0,998-0,999)**	0,972 (0,967-0,977)	0,986 (0,983-0,988)	0,987 (0,987-0,988)
		200	0,904 (0,895-0,914)	0,927 (0,917-0,937)	**0,931 (0,919-0,943)**	0,926 (0,916-0,936)	**0,999 (0,998-0,999)**	0,969 (0,966-0,973)	0,986 (0,984-0,987)	0,987 (0,985-0,988)
Indel	BWA	4	**0,463 (0,442-0,484)**	0,311 (0,3-0,322)	0,45 (0,432-0,468)	0,439 (0,42-0,458)	0,912 (0,893-0,931)	**0,967 (0,96-0,974)**	0,951 (0,944-0,958)	0,847 (0,832-0,862)
		8	**0,653 (0,624-0,681)**	0,48 (0,465-0,495)	0,578 (0,557-0,6)	0,54 (0,528-0,553)	0,906 (0,894-0,919)	**0,949 (0,937-0,96)**	0,939 (0,927-0,95)	0,821 (0,799-0,842)
		12	**0,74 (0,722-0,758)**	0,601 (0,582-0,62)	0,638 (0,62-0,656)	0,593 (0,581-0,604)	0,904 (0,898-0,91)	**0,94 (0,935-0,946)**	0,93 (0,919-0,941)	0,8 (0,78-0,819)
		16	**0,783 (0,768-0,798)**	0,678 (0,662-0,694)	0,673 (0,653-0,692)	0,616 (0,607-0,625)	0,906 (0,897-0,916)	**0,935 (0,927-0,942)**	0,927 (0,924-0,93)	0,765 (0,75-0,779)
		20	**0,807 (0,791-0,823)**	0,717 (0,696-0,738)	0,696 (0,672-0,72)	0,643 (0,634-0,652)	0,906 (0,896-0,915)	**0,933 (0,928-0,937)**	0,924 (0,918-0,931)	0,742 (0,729-0,756)
		30	**0,843 (0,823-0,864)**	0,781 (0,764-0,798)	0,726 (0,701-0,75)	0,674 (0,661-0,688)	0,907 (0,896-0,918)	**0,924 (0,913-0,934)**	0,918 (0,913-0,924)	0,697 (0,68-0,715)
		40	**0,855 (0,837-0,873)**	0,802 (0,783-0,821)	0,739 (0,718-0,76)	0,686 (0,676-0,697)	0,908 (0,898-0,918)	**0,925 (0,92-0,931)**	0,919 (0,91-0,928)	0,679 (0,66-0,697)
		60	**0,869 (0,85-0,888)**	0,817 (0,799-0,834)	0,752 (0,725-0,779)	0,695 (0,679-0,711)	0,906 (0,895-0,918)	**0,926 (0,915-0,938)**	0,915 (0,91-0,921)	0,683 (0,675-0,69)
		100	**0,879 (0,861-0,898)**	0,835 (0,817-0,853)	0,768 (0,744-0,791)	0,694 (0,681-0,706)	0,907 (0,897-0,917)	**0,923 (0,915-0,932)**	0,912 (0,91-0,915)	0,669 (0,646-0,691)
		200	**0,888 (0,871-0,904)**	0,847 (0,827-0,866)	0,776 (0,751-0,8)	0,634 (0,619-0,648)	0,909 (0,897-0,92)	**0,923 (0,911-0,935)**	0,909 (0,905-0,912)	0,659 (0,643-0,675)
	Bowtie 2	4	**0,432 (0,409-0,455)**	0,3 (0,294-0,306)	0,418 (0,4-0,437)	0,398 (0,38-0,415)	0,944 (0,933-0,956)	**0,975 (0,967-0,983)**	0,95 (0,942-0,958)	0,831 (0,821-0,841)
		8	**0,606 (0,577-0,636)**	0,459 (0,439-0,479)	0,538 (0,519-0,556)	0,502 (0,489-0,514)	0,929 (0,922-0,936)	**0,952 (0,941-0,962)**	0,935 (0,926-0,944)	0,795 (0,776-0,813)
		12	**0,685 (0,668-0,702)**	0,57 (0,556-0,584)	0,594 (0,577-0,61)	0,563 (0,553-0,573)	0,921 (0,913-0,929)	**0,947 (0,938-0,956)**	0,93 (0,921-0,939)	0,765 (0,75-0,78)
		16	**0,729 (0,714-0,744)**	0,644 (0,628-0,66)	0,625 (0,61-0,641)	0,585 (0,576-0,595)	0,921 (0,911-0,932)	**0,938 (0,933-0,942)**	0,923 (0,917-0,929)	0,722 (0,709-0,736)
		20	**0,75 (0,735-0,765)**	0,681 (0,662-0,7)	0,647 (0,623-0,67)	0,612 (0,607-0,618)	0,919 (0,912-0,926)	**0,931 (0,924-0,937)**	0,921 (0,909-0,932)	0,703 (0,683-0,724)
		30	**0,783 (0,765-0,802)**	0,74 (0,724-0,757)	0,679 (0,657-0,701)	0,643 (0,635-0,65)	0,915 (0,907-0,923)	**0,921 (0,914-0,928)**	0,914 (0,903-0,926)	0,653 (0,633-0,673)
		40	**0,797 (0,778-0,817)**	0,761 (0,744-0,777)	0,695 (0,678-0,712)	0,659 (0,652-0,666)	0,914 (0,905-0,923)	**0,921 (0,913-0,93)**	0,911 (0,902-0,92)	0,633 (0,611-0,655)
		60	**0,811 (0,791-0,831)**	0,779 (0,76-0,799)	0,713 (0,69-0,736)	0,666 (0,655-0,677)	0,91 (0,901-0,919)	**0,918 (0,909-0,926)**	0,908 (0,903-0,913)	0,623 (0,614-0,633)
		100	**0,824 (0,806-0,843)**	0,793 (0,776-0,811)	0,729 (0,706-0,753)	0,664 (0,657-0,67)	0,906 (0,892-0,919)	**0,91 (0,902-0,918)**	0,904 (0,898-0,909)	0,601 (0,587-0,615)
		200	**0,833 (0,815-0,85)**	0,806 (0,788-0,823)	0,739 (0,719-0,758)	0,6 (0,584-0,616)	0,908 (0,897-0,92)	**0,905 (0,893-0,917)**	0,898 (0,894-0,903)	0,575 (0,554-0,595)

DOI:10.14753/SE.2017.1938

3. táblázat. **AUPRC értékek, illetve a VariantMetaCaller és az egyedi variánskivonatoló módszerek által elért AUPRC értékek különbségei a szimulált adatokon.** Rövidítések: AUPRC = precizitás-szenzitivitás görbe alatti terület (area under the precision-recall curve), VMC = VariantMetaCaller

Variáns-típus	Illesztő-program	Leolvasási mélység	A VMC által elért AUPRC	HaplotypeCaller			UnifiedGenotyper			FreeBayes			SAMtools		
				AUPRC	Az AUPRC értékek különbsége	p-érték[1]	AUPRC	Az AUPRC értékek különbsége	p-érték[1]	AUPRC	Az AUPRC értékek különbsége	p-érték[1]	AUPRC	Az AUPRC értékek különbsége	p-érték[1]
SNP	BWA	4	0,730	0,622	0,109	6,11E-07	0,660	0,071	3,34E-06	0,380	0,350	1,33E-08	0,420	0,310	4,43E-09
		8	0,846	0,777	0,068	3,70E-06	0,786	0,060	5,35E-07	0,555	0,291	8,09E-08	0,578	0,268	1,10E-07
		12	0,890	0,837	0,054	1,73E-06	0,838	0,052	2,30E-06	0,673	0,217	3,90E-07	0,686	0,204	3,17E-07
		16	0,912	0,867	0,045	3,42E-07	0,864	0,048	1,01E-07	0,748	0,163	1,77E-07	0,757	0,155	1,80E-07
		20	0,924	0,884	0,039	7,64E-06	0,878	0,046	9,47E-07	0,800	0,123	4,54E-06	0,806	0,118	3,78E-06
		30	0,940	0,905	0,035	2,18E-08	0,887	0,053	1,53E-06	0,866	0,074	3,54E-06	0,867	0,073	7,38E-06
		40	0,947	0,911	0,036	1,26E-06	0,888	0,059	5,43E-07	0,890	0,057	2,13E-05	0,889	0,058	1,52E-05
		60	0,955	0,919	0,036	3,38E-07	0,881	0,074	1,54E-06	0,912	0,042	2,58E-05	0,910	0,045	2,18E-05
		100	0,961	0,924	0,037	6,80E-07	0,865	0,096	3,10E-07	0,928	0,032	1,02E-05	0,924	0,036	1,91E-05
		200	0,965	0,926	0,040	8,30E-07	0,727	0,239	2,74E-06	0,940	0,026	2,21E-05	0,934	0,031	1,40E-05
	Bowtie 2	4	0,699	0,561	0,138	9,65E-08	0,584	0,115	7,77E-08	0,345	0,355	3,42E-09	0,375	0,324	1,81E-09
		8	0,812	0,698	0,114	4,02E-07	0,695	0,117	3,92E-08	0,511	0,301	8,71E-08	0,519	0,293	7,42E-08
		12	0,856	0,749	0,108	3,18E-08	0,741	0,115	3,32E-09	0,625	0,231	1,16E-07	0,622	0,235	7,54E-08
		16	0,879	0,775	0,105	1,44E-09	0,764	0,116	8,17E-11	0,701	0,179	1,76E-07	0,696	0,184	7,49E-08
		20	0,894	0,789	0,105	5,04E-08	0,776	0,117	1,70E-08	0,753	0,141	2,18E-06	0,746	0,148	3,56E-07
		30	0,911	0,805	0,106	9,34E-09	0,786	0,125	2,81E-08	0,824	0,087	3,44E-06	0,814	0,097	1,12E-06
		40	0,920	0,812	0,108	5,20E-08	0,789	0,131	1,58E-08	0,850	0,070	1,22E-05	0,837	0,082	4,68E-06
		60	0,929	0,819	0,110	2,13E-08	0,787	0,142	3,04E-08	0,880	0,049	1,02E-05	0,862	0,067	1,85E-06
		100	0,938	0,823	0,115	1,22E-07	0,776	0,162	5,70E-08	0,901	0,037	9,14E-06	0,878	0,060	5,10E-06
		200	0,945	0,825	0,121	1,09E-08	0,647	0,298	7,49E-07	0,917	0,028	1,80E-05	0,889	0,056	4,89E-06
Indel	BWA	4	0,583	0,425	0,158	4,90E-06	0,294	0,289	3,20E-07	0,205	0,378	4,06E-07	0,220	0,364	4,70E-08
		8	0,718	0,592	0,125	1,90E-06	0,436	0,282	4,72E-06	0,332	0,386	2,27E-06	0,297	0,421	1,99E-06
		12	0,783	0,672	0,111	4,25E-05	0,538	0,245	4,88E-06	0,421	0,363	2,81E-06	0,338	0,445	5,31E-07
		16	0,813	0,709	0,104	1,67E-05	0,600	0,213	7,42E-07	0,477	0,336	5,59E-07	0,349	0,464	7,69E-08
		20	0,834	0,732	0,102	5,86E-05	0,631	0,204	1,68E-07	0,523	0,311	1,59E-07	0,359	0,476	5,70E-07
		30	0,865	0,763	0,102	1,16E-05	0,681	0,183	3,19E-06	0,592	0,273	3,78E-06	0,373	0,492	4,18E-07
		40	0,873	0,775	0,098	2,33E-05	0,700	0,173	2,11E-08	0,622	0,250	1,90E-06	0,368	0,504	4,83E-07
		60	0,884	0,790	0,093	2,85E-05	0,714	0,169	8,19E-07	0,652	0,232	5,74E-06	0,385	0,498	2,89E-07
		100	0,894	0,798	0,095	7,01E-05	0,728	0,166	8,92E-06	0,676	0,217	2,14E-07	0,382	0,512	6,35E-07
		200	0,902	0,806	0,096	2,59E-05	0,735	0,167	5,03E-06	0,689	0,213	7,70E-07	0,351	0,551	2,44E-07
	Bowtie 2	4	0,539	0,416	0,123	1,16E-06	0,292	0,247	4,18E-06	0,187	0,352	6,56E-07	0,189	0,350	1,52E-08
		8	0,674	0,574	0,100	1,35E-06	0,434	0,240	8,81E-06	0,307	0,367	1,00E-06	0,259	0,415	1,45E-06
		12	0,738	0,645	0,093	9,01E-06	0,535	0,203	9,24E-06	0,395	0,343	2,03E-06	0,300	0,438	9,33E-08
		16	0,768	0,684	0,084	2,02E-05	0,596	0,172	3,77E-07	0,445	0,323	2,03E-06	0,316	0,452	3,59E-07
		20	0,788	0,703	0,085	3,41E-07	0,628	0,160	7,89E-07	0,489	0,299	1,70E-06	0,326	0,461	2,22E-07
		30	0,817	0,731	0,087	2,51E-06	0,672	0,146	5,83E-06	0,555	0,262	4,83E-06	0,341	0,476	8,40E-07
		40	0,835	0,742	0,093	9,45E-07	0,692	0,143	2,04E-05	0,580	0,255	1,03E-07	0,336	0,499	7,32E-07
		60	0,844	0,755	0,089	1,71E-05	0,706	0,138	3,07E-06	0,611	0,233	3,33E-07	0,348	0,496	1,92E-07
		100	0,858	0,765	0,093	1,51E-05	0,714	0,144	1,60E-05	0,637	0,221	9,86E-07	0,345	0,513	8,73E-07
		200	0,864	0,769	0,095	6,82E-06	0,718	0,146	5,34E-06	0,651	0,213	4,62E-07	0,298	0,566	4,13E-07

[1] Párosított t-teszt p-értéke

DOI:10.14753/SE.2017.1938

4. táblázat. **A BWA illetve Bowtie 2 illesztőprogramok hatása a VariantMetaCaller maximális szenzitivitására a szimulált szekvenálási adatokon.** Rövidítések: CI = konfidencia-intervallum

Variáns-típus	Leolvasási mélység	A VariantMetaCaller maximális szenzitivitása				
		Illesztőprogram		Átlagos különbség (BWA vs. Bowtie 2)	95% CI	p-érték[1]
		BWA	Bowtie 2			
SNP	4	0,731	0,700	0,031	0,027-0,034	0,000015
	8	0,846	0,813	0,033	0,032-0,035	0,000001
	12	0,891	0,857	0,034	0,031-0,036	0,000005
	16	0,912	0,880	0,032	0,03-0,034	0,000001
	20	0,924	0,894	0,030	0,028-0,032	0,000001
	30	0,940	0,911	0,029	0,027-0,031	0,000002
	40	0,947	0,920	0,027	0,025-0,029	0,000001
	60	0,955	0,929	0,026	0,023-0,028	0,000007
	100	0,961	0,939	0,023	0,021-0,024	0,000001
	200	0,966	0,946	0,020	0,019-0,021	0,000001
Indel	4	0,605	0,555	0,050	0,041-0,058	0,000094
	8	0,741	0,690	0,051	0,046-0,055	0,000008
	12	0,803	0,754	0,049	0,044-0,054	0,000012
	16	0,830	0,782	0,047	0,038-0,057	0,000162
	20	0,850	0,804	0,047	0,04-0,053	0,000048
	30	0,880	0,831	0,048	0,044-0,053	0,000009
	40	0,888	0,849	0,039	0,035-0,042	0,000009
	60	0,899	0,858	0,041	0,033-0,048	0,000102
	100	0,908	0,872	0,036	0,032-0,04	0,000016
	200	0,917	0,880	0,037	0,032-0,042	0,000038

[1] Párosított t-teszt p-értéke

5. táblázat. **A különböző típusú variánsok száma a kisebb méretű szimulált genomi régiókban**

Célrégió mérete	SNP-k átlagos száma (szórás)	Indelek átlagos száma (szórás)	A polimorfikus SNP-k átlagos száma mintánként (szórás)	A polimorfikus indelek átlagos száma mintánként (szórás)
100 kb	450 (73,8)	60 (22,0)	98 (17,1)	14 (4,9)
200 kb	814 (159,3)	111 (30,6)	181 (44,3)	26 (7,1)
300 kb	1185 (192,8)	160 (31,7)	262 (55,4)	40 (10,5)
500 kb	1976 (238,5)	267 (30,8)	436 (53,2)	67 (9,3)

DOI:10.14753/SE.2017.1938

6. táblázat. **A változók erős relevanciájának a posteriori valószínűsége az öt éves teljes és eseménymentes túlélés szempontjából.** Egy változót relevánsnak nevezünk, ha relevanciájának valószínűsége nagyobb vagy egyenlő mint 0,5 (félkövér betűvel jelölve). A közepesen releváns változók posterior értékei ($\Pr \geq 0{,}4$ és $\Pr < 0{,}5$) aláhúzással vannak jelölve.

Változó (SNP, klinikai paraméter)	Gén	Teljes túlélés	Esemény-mentes túlélés	Változó (SNP, klinikai paraméter)	Gén	Teljes túlélés	Esemény-mentes túlélés
rs1202179	*ABCB1*	0.01	0.00	rs1950902	*MTHFD1*	0.20	0.12
rs2235013	*ABCB1*	0.02	0.01	rs2236225	*MTHFD1*	0.14	0.03
rs9282564	*ABCB1*	0.09	0.07	rs1801131	*MTHFR*	0.06	0.06
rs2066853	*AHR*	0.05	0.03	rs1801133	*MTHFR*	0.00	0.00
rs2282883	*AHR*	0.00	0.00	rs12759827	*MTR*	0.00	0.00
rs2282885	*AHR*	0.00	0.00	rs1805087	*MTR*	0.02	0.01
rs2088102	*AKR1A1*	0.04	0.02	rs3768142	*MTR*	0.01	0.00
rs4506592	*ARID5B*	0.05	0.03	rs4659724	*MTR*	0.03	0.01
rs4509706	*ARID5B*	<u>0.49</u>	0.11	rs10380	*MTRR*	0.23	0.32
rs4948487	*ARID5B*	0.00	0.00	rs1532268	*MTRR*	0.00	0.00
rs4948496	*ARID5B*	0.02	0.03	rs162036	*MTRR*	0.18	0.14
rs4948502	*ARID5B*	0.01	0.00	rs1801394	*MTRR*	0.00	0.00
rs3817074	*BAX*	0.07	0.05	rs2966952	*MTRR*	0.03	0.02
rs11876772	*BCL2*	0.01	0.00	rs3776455	*MTRR*	0.00	0.00
rs12457893	*BCL2*	0.00	0.01	rs3024979	*NAB2*	0.15	0.07
rs1801018	*BCL2*	0.01	0.00	rs703817	*NAB2*	0.00	0.01
rs1893806	*BCL2*	0.01	0.01	rs2229974	*NOTCH1*	0.05	0.01
rs2850761	*BCL2*	0.00	0.00	rs3124596	*NOTCH1*	0.00	0.00
rs4987845	*BCL2*	0.14	0.08	rs3124603	*NOTCH1*	0.01	0.00
rs8092560	*BCL2*	0.13	0.03	rs3124999	*NOTCH1*	0.02	0.02
rs1005695	*CBR1*	0.25	0.07	rs1469908	*NQO1*	0.01	0.01
rs20572	*CBR1*	0.02	0.02	rs1800566	*NQO1*	0.22	0.10
rs998383	*CBR1*	0.24	0.05	rs1143684	*NQO2*	0.01	0.01
rs1056892	*CBR3*	0.02	0.00	rs2070999	*NQO2*	0.02	0.01
rs11575815	*CCR5*	0.00	0.00	rs643333	*SHMT1*	0.01	0.01
rs1799988	*CCR5*	0.09	0.02	rs9909104	*SHMT1*	<u>0.42</u>	0.10
rs3087253	*CCR5*	0.00	0.00	rs1051266	*SLC19A1*	0.01	0.00
rs10403561	*CEBPA-AS1*	0.03	0.33	rs7499	*SLC19A1*	0.01	0.00
rs12434881	*CEBPE*	0.02	0.01	rs2276299	*SLC22A8*	0.06	0.03
rs8015478	*CEBPE*	0.01	0.00	rs3809069	*SLC22A8*	0.07	0.06
rs13306561	*CLCN6*	<u>0.48</u>	0.35	rs4149183	*SLC22A8*	0.01	0.00
rs2470893	*CYP1A1*	0.02	0.01	rs10841769	*SLCO1B1*	0.00	0.00
rs2246709	***CYP3A4***	**0.85**	**0.57**	rs11045818	*SLCO1B1*	0.06	0.07
rs2404955	*CYP3A4*	0.10	0.10	rs17328763	*SLCO1B1*	0.01	0.01
rs11742668	*DHFR*	0.12	0.14	rs4149056	*SLCO1B1*	0.07	0.03
rs12517451	*DHFR*	0.06	0.01	rs10208033	*STAT1*	0.02	0.01
rs1650723	*DHFR*	0.04	0.01	rs2030171	*STAT1*	0.26	0.02
rs2612100	*ENOSF1*	0.00	0.00	rs3088307	*STAT1*	0.01	0.00
rs10957267	*GGH*	0.03	0.31	rs12949918	*STAT3*	0.00	0.00
rs11545078	*GGH*	0.29	0.15	rs17405722	*STAT3*	0.16	0.09
rs719235	*GGH*	0.00	0.00	rs3198502	*STAT5A*	0.02	0.01
rs1695	*GSTP1*	0.02	0.12	rs4029774	*STAT5B*	0.01	0.01
rs7941395	*GSTP1*	0.04	0.09	rs2518463	*TPMT*	0.01	0.01
rs4132601	*IKZF1*	0.04	0.05	rs2842951	*TPMT*	0.02	0.02
rs12063205	*JAK1*	0.03	0.06	rs1004474	*TYMS*	0.00	0.00
rs310225	*JAK1*	<u>0.45</u>	0.02	rs2853533	*TYMS*	0.23	0.23
rs11888	*JAK3*	0.14	0.01	rs2853741	*TYMS*	0.00	0.01
rs3212713	*JAK3*	0.01	0.02	rs9967368	*TYMS*	0.00	0.00
rs1677626	*MSH3*	0.02	0.03	rs745686	*ZBTB25*	0.04	0.01
rs1076991	*MTHFD1*	0.07	<u>0.40</u>	Nem		<u>0.47</u>	0.21

DOI:10.14753/SE.2017.1938

7. táblázat. **A páciens neme és a *CYP3A4* rs2246709 polimorfizmusa közötti interakció klinikai paraméterekkel korrigált hatása a teljes és az eseménymentes túlélésre.** Rövidítések: CI = konfidencia-intervallum, HR = hazárd hányados (hazard ratio), N = mintaszám, NE = események száma, Rizikócsoport esetén: LR = alacsony rizikó, MR = közepes rizikó, HR = magas rizikó

	Teljes túlélés (N=373, NE=59, p=4,84*10^{-6})				**Eseménymentes túlélés (N=373, NE=75, p=1,2*10^{-4})**			
Kovariáns	**N, NE**	**HR**	**95% CI**	**p-érték**	**N, NE**	**HR**	**95% CI**	**p-érték**
Rizikócsoport	LR=90, 7 MR=244, 36 HR=39, 16	2,45	1,57-3,82	**7,67*10^{-5}**	LR=90, 9 MR=244, 49 HR=39, 17	2,31	1,54-3,47	**5,1*10^{-5}**
Sejtmorfológia (Pre-B, B)	Pre-T, T=64, 13 Pre-B, B=309, 46	1,38	0,72-2,67	0,332	Pre-T, T=64, 15 Pre-B, B=309, 60	1,43	0,78-2,63	0,247
Citogenetika								
Normal	104, 20	1,00			104, 24	1,00		
Hiperdiploid	66, 5	0,48	0,18-1,29	0,145	66, 9	0,68	0,31-1,47	0,324
Egyéb	203, 34	0,95	0,53-1,70	0,868	203, 42	0,95	0,56-1,61	0,86
Protokoll								
‘88-‘90	98, 15	1,00			98, 21	1,00		
‘95	235, 43	1,62	0,88-2,98	0,123	235, 50	1,28	0,75-2,17	0,37
2002	40, 1	0,24	0,03-1,84	0,17	40, 4	0,64	0,22-1,87	0,412
rs2246709, Nem								
AA, Nő	88, 10	1,00			88, 12	1,00		
AA, Férfi	119, 29	1,99	0,94-4,19	0,071	119, 35	2,17	1,10-4,27	**0,025**
AG, Nő	74, 11	1,2	0,51-2,85	0,676	74, 13	1,22	0,55-2,68	0,626
AG, Férfi	70, 3	0,37	0,10-1,34	0,13	70, 8	0,82	0,33-2,01	0,663
GG, Nő	4, 1	2,03	0,26-16,1	0,501	4, 1	1,89	0,24-14,7	0,544
GG, Férfi	18, 5	2,07	0,68-6,29	0,2	18, 6	2,44	0,88-6,77	0,086

DOI:10.14753/SE.2017.1938

8. táblázat. **Az *MTHFD1* (rs1076991) és a *CYP3A4* (rs2246709) genotípus interakciójának korrigálatlan és rizikócsoporttal korrigált hatása a teljes túlélésre.** Rövidítések: CI = konfidencia-intervallum, HR = hazárd hányados (hazard ratio), N = mintaszám, NE = események száma, Rizikócsoport esetén: LR = alacsony rizikó, MR = közepes rizikó, HR = magas rizikó

	Korrigálatlan (N=511, NE=75, p=0,0029)				**Klinikai paraméterekkel korrigált (N=373, NE=59, p=2,9*10^{-6})**			
Kovariáns	**N, NE**	**HR**	**95% CI**	**p-érték**	**N, NE**	**HR**	**95% CI**	**p-érték**
Rizikócsoport	-	-	-	-	LR=90, 7 MR=244, 36 HR=39, 16	2,69	1,73-4,19	**1,23*10^{-5}**
Sejtmorfológia (Pre-B, B)	-	-	-	-	Pre-T, T=64, 13 Pre-B, B=309, 46	1,03	0,52-2,04	0,928
Citogenetika								
Normal	-	-	-	-	104, 20	1,00		
Hiperdiploid	-	-	-	-	66, 5	0,44	0,16-1,20	0,11
Egyéb	-	-	-	-	203, 34	0,89	0,49-1,59	0,683
Protokoll								
‘88-‘90	-	-	-	-	98, 15	1,00		
‘95	-	-	-	-	235, 43	1,77	0,94-3,32	0,078
2002	-	-	-	-	40, 1	0,24	0,03-1,85	0,171
Nem (Férfi)	-	-	-	-	Nő=166, 22 Férfi=207, 37	1,16	0,66-2,04	0,6
rs1076991 : rs2246709								
AA, AA	66, 5	1,00			46, 3	1,00		
AA, AG	48, 7	1,91	0,61-6,02	0,269	32, 4	2,72	0,59-12,58	0,2
AA, GG	6, 2	5,22	1,01-26,92	**0,048**	2, 1	23,74	2,22-254,2	**0,009**
AG, AA	157, 28	2,46	0,95-6,37	0,064	122, 25	3,74	1,11-12,53	**0,033**
AG, AG	96, 6	0,81	0,25-2,64	0,723	69, 5	1,47	0,34-6,26	0,606
AG, GG	15, 3	2,64	0,63-11,0	0,184	12, 2	2,32	0,38-13,9	0,359
GG, AA	52, 13	3,53	1,26-9,89	**0,017**	39, 11	6,11	1,67-22,36	**0,006**
GG, AG	58, 6	1,31	0,40-4,31	0,651	43, 5	1,85	0,43-7,96	0,406
GG, GG	13, 5	5,94	1,72-20,5	**0,005**	8, 3	4,99	0,96-25,9	0,056

DOI:10.14753/SE.2017.1938

9. táblázat. **Az *MTHFD1* (rs1076991) és a *CYP3A4* (rs2246709) genotípus interakciójának korrigálatlan és rizikócsoporttal korrigált hatása az eseménymentes túlélésre.** Rövidítések: CI = konfidencia-intervallum, HR = hazárd hányados (hazard ratio), N = mintaszám, NE = események száma, Rizikócsoport esetén: LR = alacsony rizikó, MR = közepes rizikó, HR = magas rizikó

	Korrigálatlan (N=511, NE=95, p=0,001)				**Klinikai paraméterekkel korrigált (N=373, NE=75, p=7,5*10^{-6})**			
Kovariáns	**N, NE**	**HR**	**95% CI**	**p-érték**	**N, NE**	**HR**	**95% CI**	**p-érték**
Rizikócsoport	-	-	-	-	LR=90, 9 MR=244, 49 HR=39, 17	2,45	1,63-3,67	**1,4*10^{-5}**
Sejtmorfológia (Pre-B, B)	-	-	-	-	Pre-T, T=64, 15 Pre-B, B=309, 60	1,15	0,61-2,17	0,665
Citogenetika								
Normal	-	-	-	-	104, 24	1,00		
Hiperdiploid	-	-	-	-	66, 9	0,61	0,28-1,33	0,21
Egyéb	-	-	-	-	203, 42	0,89	0,53-1,52	0,674
Protokoll								
‘88-‘90	-	-	-	-	98, 21	1,00		
‘95	-	-	-	-	235, 50	1,37	0,79-2,38	0,257
2002	-	-	-	-	40, 4	0,64	0,21-1,90	0,419
Nem (Férfi)	-	-	-	-	Nő=166, 26 Férfi=207, 49	1,44	0,87-2,37	0,156
rs1076991 : rs2246709								
AA, AA	66, 7	1,00			46, 5	1,00		
AA, AG	48, 12	2,44	0,96-6,20	0,061	32, 8	3,13	1,00-9,82	0,05
AA, GG	6, 2	3,76	0,78-18,12	0,098	2, 1	11,36	1,22-105,3	**0,032**
AG, AA	157, 33	2,07	0,92-4,68	0,08	122, 28	2,53	0,97-6,60	0,058
AG, AG	96, 8	0,76	0,27-2,09	0,591	69, 6	0,98	0,29-3,25	0,972
AG, GG	15, 3	1,9	0,49-7,36	0,351	12, 2	1,51	0,29-7,84	0,623
GG, AA	52, 16	3,18	1,31-7,72	**0,011**	39, 14	4,67	1,65-13,22	**0,004**
GG, AG	58, 8	1,24	0,45-3,43	0,673	43, 7	1,62	0,50-5,20	0,419
GG, GG	13, 6	5,28	1,77-15,7	**0,003**	8, 4	4,32	1,10-16,9	**0,036**

DOI:10.14753/SE.2017.1938

10. táblázat. **A VariantMetaCaller által elért AUPRC értékek 3 illetve 4 egyedi variánskivonatoló kombinálásával.** A legjobb eredményt adó beállítás félkövér betűtípussal van jelölve. Rövidítések: FB = FreeBayes, HC = HaplotypeCaller, ST = SAMtools, UG = UnifiedGenotyper

		BWA					Bowtie 2				
Variáns-típus	Leolvasási mélység	HC UG FB	HC FB ST	HC UG ST	UG FB ST	HC UG FB ST	HC UG FB	HC FB ST	HC UG ST	UG FB ST	HC UG FB ST
SNP	4	0,7046	0,7295	0,7272	0,7296	**0,7302**	0,6701	0,6981	0,6847	0,6989	**0,6992**
	8	0,8292	0,8448	0,8436	0,8443	**0,8456**	0,7922	0,8108	0,8042	0,8113	**0,8121**
	12	0,8781	0,8896	0,8892	0,8893	**0,8902**	0,8414	0,8557	0,8517	0,8559	**0,8565**
	16	0,9029	0,9110	0,9109	0,9107	**0,9115**	0,8677	0,8783	0,8763	0,8788	**0,8794**
	20	0,9178	0,9231	0,9232	0,9229	**0,9238**	0,8854	0,8925	0,8903	0,8930	**0,8935**
	30	0,9365	0,9390	0,9394	0,9388	**0,9398**	0,9065	0,9096	0,9078	0,9103	**0,9110**
	40	0,9442	0,9461	0,9467	0,9459	**0,9470**	0,9160	0,9185	0,9159	0,9195	**0,9199**
	60	0,9529	0,9540	0,9545	0,9539	**0,9546**	0,9264	0,9271	0,9236	0,9284	**0,9289**
	100	0,9596	0,9600	0,9607	0,9602	**0,9608**	0,9366	0,9357	0,9303	0,9376	**0,9379**
	200	0,9644	0,9641	0,9654	0,9648	**0,9654**	0,9443	0,9426	0,9359	0,9449	**0,9451**
Indel	4	0,5213	**0,5859**	0,5385	0,5579	0,5833	0,4923	**0,5407**	0,4996	0,5112	0,5392
	8	0,6864	**0,7228**	0,6898	0,6818	0,7181	0,6447	0,6718	0,6452	0,6349	**0,6737**
	12	0,7608	**0,7837**	0,7672	0,7440	0,7832	0,7121	0,7356	0,7233	0,7022	**0,7377**
	16	0,8008	0,8130	0,8011	0,7712	**0,8132**	0,7533	0,7661	0,7593	0,7288	**0,7680**
	20	0,8218	0,8321	0,8240	0,7921	**0,8332**	0,7724	0,7847	0,7821	0,7519	**0,7879**
	30	0,8558	0,8633	0,8592	0,8270	**0,8642**	0,8066	0,8131	0,8141	0,7831	**0,8179**
	40	0,8686	0,8700	0,8680	0,8405	**0,8737**	0,8239	0,8312	0,8320	0,8020	**0,8351**
	60	0,8806	0,8831	0,8814	0,8480	**0,8835**	0,8378	0,8390	0,8421	0,8091	**0,8440**
	100	0,8912	0,8919	0,8910	0,8596	**0,8937**	0,8508	0,8519	0,8537	0,8216	**0,8575**
	200	0,8994	0,8980	0,8998	0,8620	**0,9022**	0,8604	0,8587	0,8599	0,8250	**0,8638**