DOKTORI ÉRTEKEZÉS

# A SZEMÉLYRE SZABOTT GYÓGYÁSZAT NYOMÁBAN

## DNS-MINTÁK TULAJDONSÁGAINAK VIZSGÁLATA

## ÚJ GENERÁCIÓS SZEKVENÁLÁSI ADATOK ALAPJÁN


PIPEK ORSOLYA ANNA

EÖTVÖS LORÁND TUDOMÁNYEGYETEM, TERMÉSZETTUDOMÁNYI KAR
FIZIKA DOKTORI ISKOLA
STATISZTIKUS FIZIKA, BIOLÓGIAI FIZIKA ÉS KVANTUMRENDSZEREK FIZIKÁJA PROGRAM

ISKOLAVEZETŐ: GUBICZA JENŐ, DSC, EGYETEMI TANÁR
PROGRAMVEZETŐ: KÜRTI JENŐ, DSC, EGYETEMI TANÁR




TÉMAVEZETŐ:

CSABAI ISTVÁN, DSC, EGYETEMI TANÁR

ELTE TTK, KOMPLEX RENDSZEREK FIZIKÁJA TANSZÉK


2019. OKTÓBER

# KÖSZÖNETNYILVÁNÍTÁS

Ezúton szeretnék köszönetet mondani témavezetőmnek, Csabai Istvánnak, aki mindenre nyitott lelkesedésével és széleskörű érdeklődésével olyan jól ismert fizikai megoldások használatára buzdított az egészségügyi témák feltérképezésénél, melyek új perspektívába helyezik a nagyon is aktuális problémákat. Egyéni látásmódja számomra is megkérdőjelezhetetlenné tette, hogy az ilyen komplex kérdésekre csak interdiszciplináris eszköztárral találhatjuk meg a választ.

Köszönöm továbbá minden orvos és biológus kutatótársunk, kiemelten Szüts Dávid, Moldvay Judit és Szállási Zoltán felbecsülhetetlen segítségét. Szaktudásuk, ötleteik és türelmes magyarázataik nélkül az eredmények interpretálása lehetetlen lett volna.

Végül, de semmiképp nem utolsósorban szeretném megköszönni Édesapámnak és Édesanyámnak a támogatást, továbbá hogy a tudományok, és főként a fizika és az orvostudomány iránti érdeklődést már kiskoromban felkeltették bennem.

A kutatást a FIEK_16-1-2016-0005, NVKP_16-1-2016-0004 és a Horizont 2020 innovációs program 643476 sz. pályázatok támogatták.

# AZ ÉRTEKEZÉS CÉLKITŰZÉSEI ÉS MOTIVÁCIÓI

Disszertációm elsődleges célja a manapság méltón népszerű koncepció, a személyre szabott gyógyászat módszereinek és jövőbeli lehetőségeinek áttekintése a teljesség igénye nélkül. Az utóbbi egy-két évtized alatt a DNS-szekvenálási technikák rohamos fejlődésen estek át, így hatalmas felhalmozódott adatmennyiség áll a genetikai témájú vizsgálódások rendelkezésére. Emellett az onkológia az egészségügy egyik legaktívabban kutatott ága, évente újabb és újabb gyógyszerek jelennek meg, melyeket a páciensek egyre specifikusabb csoportjaira optimalizálnak.

Mindez a rendkívüli mennyiségű adat feldolgozása azonban elengedhetetlenné teszi a különböző tudományágak összefogását. Hagyományos biológiai eszköztárral lehetetlen egy kb. $10^{11}$ karakterből álló, különféle zajokkal terhelt szövegből orvosilag releváns következtetéseket levonni. A fizikusi látásmód ezekben az esetekben nagy előnyt jelent, hiszen a statisztikus és valószínűségszámítási modellek, mint például a bayesi analízis, a Fisher-féle egzakt teszt, a $\chi^2$-teszt, az adatok információtartalmának és entrópiájának mérése, továbbá a dimenzióredukció mátrixok faktorizálása vagy főkomponens-analízis útján, mind olyan módszerek, melyek adattípustól függetlenül alkalmazhatóak tetszőleges „big data” analízis során.

Az értekezésben tehát főként a fizikusi módszertan bemutatására törekszünk egy látszólag távolinak tűnő tudományterület keretein belül, a legáltalánosabb biomarkerek vizsgálatán át haladva a konkrét genomi eltérések hatásainak feltérképezéséig. Egy fejezet erejéig emellett kitérünk a populáció-szintű genomikai statisztikák környezeti mintákból való kinyerésének lehetőségére és előnyeire is.

# TARTALOMJEGYZÉK

# A SZEMÉLYRE SZABOTT GYÓGYÁSZAT

## A SZEMÉLYRE SZABOTT GYÓGYÁSZAT FOGALMA ÉS JELENTŐSÉGE

A személyre szabott gyógyászat a hagyományos, kollektív („one-size-fits-all”) gyógymódok helyett olyan terápiás megoldásokat helyez előtérbe, melyek az adott páciens egyéni tulajdonságaihoz és igényeihez lettek optimalizálva, ezáltal elősegítve, hogy a betegeket a számukra biztonságos és hatásos kezeléseknek vessék csak alá, minimalizálva a kellemetlen mellékhatásokat. [1]

Egy egyén „tulajdonságai” alatt az őt jellemző genetikai, környezeti és klinikai információk összességét értjük, melyek együttes figyelembevételével a kezelések precízen testreszabhatóvá válnak. A páciensek ilyen jellegű stratifikációját elsősorban a DNS szekvenálás napjainkban történő rohamos fejlődése teszi lehetővé. [2] A szekvenálási módszerek és technikák a teljes humán genom legelső 2001-es publikálása [3] óta évről évre egyre olcsóbbá [4] és hatékonyabbá [5] válnak, így a genetikai tesztek napjainkra az orvosi rutin részét képezik [6]. Azok a jellegzetességek, melyek megléte (vagy hiánya) egy egyénben korrelál egy adott betegség megjelenésének kockázatával, a betegség lefolyásának súlyosságával, illetve a különböző kezelések hatékonyságával, felhasználhatóak ún. biomarkerként [7]. Ezek vizsgálatával egyrészt nagyon korai stádiumban detektálhatóak a kialakulni kezdő betegségek, mely jelentősen segíti a prevenciós törekvéseket, másrészt a már kialakult betegség esetében jóslás tehető annak hosszútávú következményeire és a különböző terápiás szerek eredményességére [1].

## EDDIGI EREDMÉNYEK

A jelenleg folyó biológiai és klinikai kutatások számottevő hányada különböző biomarkerek azonosításával és klinikai fontosságuk felderítésével foglalkozik. Ezek sokrétűségének szemléltetésére az alábbiakban néhány olyan példát mutatunk be, melyek a páciensekről elérhető különböző szintű információk kiaknázásán alapulnak.

Köztudott, hogy a tüdőrákok túlnyomó többsége a dohányzás következményének tudható be: a férfiak esetében a tüdődaganatos páciensek 90%-a, míg a nőknél 70-80%-a aktív vagy leszokott dohányos [8]. A nem dohányzókhoz képest a dohányzók esetében 30-szor nagyobb a tüdőrák kialakulásának kockázata. Ez az adat arra enged következtetni, hogy a dohányosok körében érdemes az átlagosnál gyakoribb tüdőszűrési vizsgálatokat végezni, hogy a korai diagnózissal a betegség kezelése minél hamarabb megkezdődhessen. Hagyományos értelemben véve a dohányzást éppen a kézenfekvősége miatt nem tekintjük valódi biomarkernek, de a példa jól illusztrálja, hogy a betegek egyéni körülményeinek figyelembevétele elengedhetetlen a diagnosztika és a terápia során.

Sokkal részletesebb információk nyerhetők a tumor és környékének immunhisztokémiai vizsgálatával. Az 1990-es évek kutatásai megmutatták, hogy azon sejtek, melyek bizonyos fehérjéket hordoznak a felszínükön, el tudják kerülni az immunrendszer támadásait

[9] [10]. A mechanizmus az immunrendszer normális működéséhez elengedhetetlen: ez óvja meg a szervezetet attól, hogy saját sejtjeit megtámadja, illetve elpusztítsa, azaz az autoimmun betegségektől. A 2000-es évek elejére felfedezték a PD-L1 fehérjét [11], melyet a daganatos sejtek jelentős hányada - gyakran ráktípustól függetlenül - a felületén hordoz, melynek segítségével kijátssza az őt potenciálisan megtámadó T-sejteket. A T-sejtek túlzott aktivitásának gátlásáért több útvonal is felel, ezek egyike a PD-1:PD-L1 tengely, mely a perifériás szövetben (így például a daganatban is) az immuntoleranciáért felel. Amennyiben a T-sejtek felületén megjelenő PD-1 fehérje összekapcsolódik PD-L1 ligandumával, a tumorsejtek apoptózisa gátlódik. Ennek a ténynek a felismerése ígéretes terápiás módszerek kifejlesztését tette lehetővé: az ún. immunellenőrzőpont-gátló szerek egy csoportja a T-, illetve egyéb immunsejteken megjelenő PD-1, vagy a tumorsejtek PD-L1 fehérjéjének blokkolásával megakadályozza a immunrendszer mesterséges gátlását, így újraindítja a tumorellenes immunválaszt. Az immunterápia a hagyományos, agresszív kemoterápiás kezelésekkel szemben jóval kevésbé toxikus és gyakran hatékonyabb is, elsősorban az előrehaladottabb daganatok esetében. Természetesen ahhoz, hogy a PD-1:PD-L1 tengely blokkolása valódi eredményre vezessen, elengedhetetlen, hogy a daganatsejtek ténylegesen ezt a mechanizmust használják a mesterséges immuntolerancia biztosításához, vagyis a tumorsejteknek jelentős mértékben expresszálniuk kell a PD-L1 fehérjét. Ezáltal a PD-L1 pozitivitás kvantitatív mérésével az immunterápia sikeressége jósolható lesz különböző páciensek esetén. A PD-L1 expresszió immunhisztokémiával meghatározott értéke, mint klinikai paraméter ún. prediktív biomarkere a PD-1 és PD-L1 gátló szerek eredményességének. Az immunellenőrzőpont-gátló szereket többféle szolid tumor (melanóma, nem kissejtes tüdőrák, urotelsejtes rák, vesesejtes rák, fej-nyaki laphámsejtes rák) és hematológiai daganat (klasszikus Hodgkin-limfóma) esetében sikerrel alkalmazzák [12].

Még árnyaltabb képet kaphatunk a daganatok genetikai sajátosságainak feltérképezésével. Bizonyos ráktípusokban (emlő, petefészek, prosztata) például jellemzően sérül a DNS kettősszálú töréseit javítani képes ún. homológ rekombinációs mechanizmus, így a tumor védtelenné válik az ilyen jellegű DNS-roncsolódásokkal szemben. Kettős száltörés a leggyakrabban akkor alakul ki, ha a DNS egyik szálán megjelenő hibát nem sikerül időben, még a replikáció előtt kijavítani [13]. Így az egyszálas javító mechanizmusok mesterséges gátlásával jelentősen megnövelhető a kettős száltörések gyakorisága, mely a tumorsejtek számára végzetes, míg az egészséges szövetben a hiba nélkül működő homológ rekombinációnak és a ritkább replikációnak köszönhetően nem okoz ennyire súlyos károsodást. Ezt az elvet használják ki az ún. PARP-inhibitor szerek, melyek az egyszálas javító mechanizmusokban fontos szerepet játszó PARP1 fehérje gátlásával mesterségesen kettős száltöréseket hoznak létre a DNS-ben. A tumorban a homológ rekombináció hibájára többféle jel utalhat, ilyenek például a javító mechanizmusban domináns szerepű BRCA1, BRCA2 és PALB2 fehérjéket kódoló gének mutációi, melyeket panelszekvenálásos diagnosztikai tesztek során célzottan vizsgálnak. Emellett a tumorokban előforduló,

a teljes genomra kiterjedő mutációs mintázatok [14] vizsgálatával azonosíthatók a daganatban aktív, különféle mutációs folyamatok „lenyomatai”. Mivel a DNS-javító mechanizmusok komplexitása miatt a homológ rekombináció hibáját nyilvánvalóan nem csak a fenti három génben keletkezett eltérés okozhatja, a specifikus mutációs mintázat megléte októl függetlenül képes jelezni a kettős szaltöréssel szembeni sérülékenységet. A BRCA1, BRCA2 és PALB2 mutációk, illetve a megfelelő mutációs mintázat megléte tehát prediktív biomarkerei a PARP-inhibitor terápia hatékonyságának.

### JÖVŐBELI LEHETŐSÉGEK ÉS PROBLÉMÁK

A személyre szabott gyógyászat a hagyományos reaktív terápiák túlsúlyát, vagyis mikor a betegség felfedezése után kell a megfelelő gyógymódot megtalálni, átbillenthetné a preventív stratégiák felé, mikor ismerve a genetikai hátteret, bizonyos magas kockázatú betegségekre gyakoribb szűréseket vagy akár megelőző terápiákat is végezhetünk. Ez egyrészről komoly életminőség javulást eredményezne, másrészről pedig agyagilag sokkal inkább kifizetődő lenne [15].

A diagnosztikus biomarkerek sztenderdizálásához olyan robusztus genomi jeleket kell találnunk, melyek akár már a vérből kimutatva is megbízhatóan utalnak egy adott betegség meglétére. Az egyik legelső ilyen törekvés a folyékony biopsziás minták használatára, a magzati Down-kór azonosítására irányult, állapotos nők véréből [16]. A TRACERx kutatás [17] keretein belül a vérbe kijutó, keringő tumor DNS-t analizálják, hogy a daganat fejlődését előrejelezzék. Ezek a megközelítések a továbbiakban nagyban hozzájárulnak majd ahhoz, hogy a pácienseknek adott kemoterápiás szereket azonnal egy másikra cseréljék, amint a rezisztencia jelei feltűnnek a vérből nyert genomban.

A személyre szabott, genomikai elváltozások nyomon követésén alapuló gyógyászatnak tehát számos előnye lehet mind prevenciós, mind diagnosztikai, mind pedig terapeutikus szempontból. Ahhoz azonban, hogy ténylegesen profitáljunk ezekből a módszerekből, elsőként feltérképező kutatások hosszú sorára van szükség, később pedig a genomikai adatok rutinszerű, klinikai begyűjtésére és tárolására. Ennek a gyakorlati megvalósítás nehézségein túl számos jogi és etikai akadálya van.

Az EU-ban például nincs központilag elfogadott szabályozás a genetikai információk kezelésére vonatkozóan [18]. Mivel a begyűjtött DNS nem csak a konkrét pácienst, hanem összes vérrokonát is valamilyen szinten jellemzi, pusztán a páciens beleegyezése az adatok kezelésébe valójában nem elegendő. Emellett felmerül a kérdés, hogy ha a teljes genetikai kód ismeretében meg tudjuk határozni a különféle, még nem diagnosztizált betegségek későbbi bekövetkezésének várható esélyét, erről pszichológiai szempontból hasznos-e az egyénnek tudnia [19]. Ugyanakkor a genetikai információk felhalmozódásával egyre valószínűbb, hogy ezekhez a munkáltatók és biztosítók is hozzáférést követelnek majd annak érdekében, hogy kiválogassák a genetikailag „alsóbbrendű” munkavállalókat és ügyfeleket. Ez a genetikai diszkrimináció a jövőbeli betegségek esélyére vonatkozóan az embe-

rek egy új, eddig nem látott hátrányos helyzetű csoportját hozza létre. Ezek a problémák rendkívül fontossá teszik, hogy megfelelő eszközökkel szabályozzák a genetikai adatokhoz való hozzáférést, elsősorban azért, hogy a nyilvánosság bizalma ne inogjon meg az újszerű orvosbiológiai és technológiai megoldásokban. Ez nem csak az egyének életszínvonalának javítása érdekében elengedhetetlen, de a tudományos látásmód és a kutatások elfogadottsága szempontjából is.

A személyre szabott gyógyászat tehát rendkívüli eredményekkel kecsegtet, így világszerte hatalmas anyagi és emberi erőforrásokat mozgósítanak a témában. Mindazonáltal, mint minden, az addigi rutinokat alapjaiban megváltoztató technológia esetén, óvatosan és körültekintően kell eljárnunk az alkalmazása során.

# KLINIKAI PARAMÉTEREK, MINT BIOMARKEREK A DAGANATOS BETEGSÉGEK GYÓGYÍTÁSÁBAN

## TÜDŐ ADENOKARCINÓMÁK METASZTÁZISAINAK STATISZTIKUS TULAJDONSÁGAI

Azoknak a klinikai paramétereknek a biomarkerként való használata, melyek egy adott betegség diagnózisa során automatikusan megállapításra és lejegyzésre kerülnek, talán szofisztikálatlan módszernek tűnhet, mégis megvan az az előnye, hogy további költséges vizsgálatok nélkül támpontot adnak abban, hogy a betegség lefolyása során milyen további szövődményekre lehet számítani, illetve milyen kezelési stratégiákat érdemes követni.

Annak a feltérképezésére, hogy a primer tüdő adenokarcinómában (LADC, lung adeno carcinoma) szenvedő páciensek esetében milyen gyakorisággal jelennek meg metasztázisok a különböző szervekben, illetve ezek milyen időközönként követik egymást, egy négy magyarországi intézet által 2001 és 2014 között gyűjtött adathalmazt elemeztünk [20]. Az ilyen témában korábban megjelent publikációk [21–26] gyakran ellentmondó eredményekről számoltak be azzal kapcsolatban, hogy a primer tumor elhelyezkedése a tüdőben milyen módon befolyásolja a későbbi áttétek megjelenésének helyét, illetve idejét. Ezért a kutatás során elsődlegesen azt vizsgáltuk, hogy a primer lokalizáció függvényében mennyire gyakoriak az egyes szervekben a korai, illetve késői áttétek. Ezen felül a metasztázisok megjelenésének egymástól való függetlenségét kívántuk tesztelni, továbbá a köztük eltelt idő hosszának és az érintett szervnek a kapcsolatát.

Az elemzett adathalmaz 1126 páciens adatait tartalmazta, akik különböző stádiumú LADC-től szenvedtek. A diagnosztizált metasztázisokat „korainak” tekintettük, amennyiben a tüdő daganat diagnózisától számított egy hónapon belül azonosították őket, egyébként „késői” metasztázisnak kategorizáltuk őket. A tüdő tumor elhelyezkedését minden páciensnél három, egymástól függetlennek tekintett bináris változóval jellemeztük: centrális/perifériás, felső/alsó régió, bal/jobb oldal. A felső régió a felső lebenyt és jobb oldalon a középső lebenyt foglalta magába, míg az alsó régió mindkét oldalon az alsó lebenyt jelentette.

Annak a vizsgálatára, hogy egy adott szervben megjelenő metasztázisok jellemzően inkább koraiak vagy későiek, meghatároztuk az összes olyan páciens számát, akiknek az adott szervben korai, illetve késői metasztázisa volt, majd azt a nullhipotézist alkalmazva, hogy a szerv korai és késői áttétei ugyanannyira valószínűek, egy $\chi$-négyzet tesztet használtunk. A szervenként kapott p-értékekre a többszörös tesztelés kompenzálására a Bonferroni-korrekciót alkalmaztuk [27]. Erre azért van szükség, mert minél több statisztikai tesztet végzünk el egy adathalmazon, annál valószínűbb, hogy egyszerűen a mérési hibákból adódóan legalább egynél találunk statisztikailag szignifikánsnak tűnő eredményt. Például egy hipotézis-vizsgálat esetén, $\alpha = 0,05$ szignifikancia szint mellett 5% annak a valószínűsége, hogy bár a nullhipotézis igaz, mi mégis elvetjük azt az adatok alapján. 100 ilyen vizsgálat elvégzése esetén azonban már a hibásan elvetett nullhipotézisek várható ér-

téke 5 lesz, illetve (független vizsgálatoknál) $1-0,95^{100} \simeq 99,4\%$-os valószínűséggel lesz legalább egy fals pozitív eredmény a 100-ból. Vagyis ha az összes eszünkbe ötlő tesztet elvégezzük egy adathalmazon, majd kiválasztjuk azokat, melyek szignifikáns eredményt adtak, az az adatok hibás interpretálásához vezet. A problémát különböző statisztikai módszerekkel orvosolhatjuk, például Bonferroni nyomán mesterségesen lecsökkenthetjük az $m$ db. tesztre alkalmazott $\alpha$ szignifikancia szint értékét $\alpha^* = \alpha/m$-re, így az $m$ tesztre nézve a fals pozitívok várható értéke $\alpha$ lesz. Ez egy viszonylag szigorú megközelítés, amivel a statisztikai erő ugyan lecsökken, ezért bizonyos esetekben érdemes megengedőbb módszerekhez (Holm-Bonferroni, Holm–Šidák korrekciók) folyamodni, de amikor a fals pozitív találatok minimalizálása a cél, a Bonferroni-módszer megbízható eredményeket ad.

Az LACD adathalmazon a különböző szervekre végzett tesztek azt mutatták, hogy a tüdőben, a mellhártyában, illetve a mellékvesében megjelenő áttétek tipikusan korán jelentkeznek, míg az agyi metasztázisok jellemzően későbbiek (A1. ábra (Melléklet)). A többi szerv esetében nem találtunk szignifikáns tendenciát a megjelenés idejére vonatkozóan.

A primer tumor elhelyezkedését tekintve azt találtuk, hogy a centrális tüdő tumorok még inkább hajlamosak a korai áttétek előidézésére, mint a perifériikusak (Fisher-féle egzakt teszt, $p = 0,02$), bár mindkét elhelyezkedés esetén gyakoribbak a korai metasztázisok, mint a későiek. Ilyen tendencia a jobb/bal oldali, illetve alsó/felső régiós elhelyezkedést illetően nem volt megfigyelhető, és megjegyezzük, hogy a fenti p-érték is szignifikanciáját veszti a többszörös tesztelésre tett korrekció során.

Annak a vizsgálatára, hogy a különböző metasztázisok mennyire gyakran fordulnak elő együttesen, az 1. ábra a) paneljén egy hőtérképen ábrázoltuk azoknak a pácienseknek a számát, akiknek mind az $i$, mind pedig a $j$ szervében találtak áttétet. Annak ellenére, hogy például a tüdő és csont áttétek láthatóan igen gyakran fordulnak elő együttesen, ebből az ábrából természetesen nem vonhatjuk le a következtetést, hogy a kétféle áttét valamilyen módon preferálná egymást, hiszen ezekből a metasztázisokból már önmagukban nagyon sokat azonosítottak a betegcsoportban. Így tehát teljesen véletlenül is előfordulhat, hogy egy páciensben mind csont, mind pedig tüdő áttétet is találnak. Annak érdekében, hogy képet kapjunk arról, hogy melyek azok a szerv-pár kategóriák, melyekbe a véletlenül várhatónál jelentősen több/kevesebb páciens került, elsőként minden $i$ szervre kiszámítottuk az áttét előfordulásának becsült $\hat{p}_i$ valószínűségét:

$$\hat{p}_i = \frac{N_i}{N}$$

ahol $N_i$ azoknak a betegeknek a száma, akiknél áttétet azonosítottak az $i$ szervben, $N$ pedig a teljes betegcsoport számossága. Ebből, amennyiben feltételezzük, hogy az áttétek egymástól függetlenül, véletlenül jelennek meg, kiszámítható annak a $\bar{p}_{ij}$ várható valószí-

nűsége, hogy egy páciensben az $i$ és $j$ szervekben is azonosítanak áttétet:

$$\bar{p}_{ij} = \hat{p}_i \cdot \hat{p}_j$$

Ezzel szemben az együttes $\hat{p}_{ij}$ valószínűségre az adatokból adható tényleges becslés

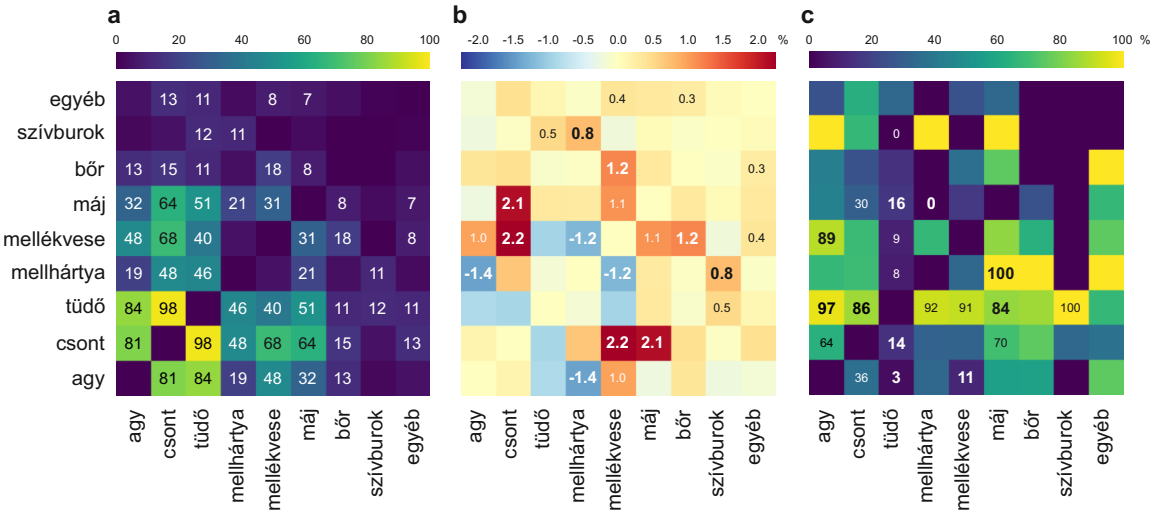
$$\hat{p}_{ij} = \frac{N_{ij}}{N}$$

ahol $N_{ij}$-k az 1. ábra a) paneljén található értékek. Arra vonatkozóan, hogy az elméleti $\bar{p}_{ij}$ és a tényleges $\hat{p}_{ij}$ értékek mennyire térnek el egymástól, a $d_{ij} = \hat{p}_{ij} - \bar{p}_{ij}$ előjeles mennyiséget definiáltuk, melyeknek százalékos értékeit az 1. ábra b) panelje mutatja. Az eredmények szerint a csont áttét a máj és mellékvese áttétekkel, a mellékvese pedig a bőr áttétekkel a vártnál gyakrabban jelenik meg együtt, míg a mellhártya metasztázisok az agyi és a mellékvesei metasztázisokkal a vártnál ritkábban esnek egybe. Annak a vizsgálatára, hogy amennyiben egy páciensnek az $i$ és $j$ szerveiben is találnak áttétet, van-e tendencia arra vonatkozóan, hogy az $i$ vagy a $j$ szervbeli áttét jelenik meg előbb, bevezettük az $i$ és $j$ áttétek közti $S_{ij}$ sorrend preferencia értékét:

$$S_{ij} = \frac{N_{i \to j}}{N_{ij}}$$

ahol $N_{i \to j}$ azoknak a pácienseknek a száma, akikben az $i$ szervbeli áttét időben megelőzte a $j$ szervbelit. Értelemszerűen $S_{ij} = 1 - S_{ji}$. Az így definiált sorrend preferenciák százalékos értékeit mutatja az 1. ábra c) panelje. (Azokat az eseteket, ahol az $i$ és a $j$ szervbeli áttétek közül mindkettő korai volt (a primer diagnózis után nem sokkal azonosították), nem vettük figyelembe, hiszen ilyenkor a sorrendiség nem pontosan meghatározható.) Láthatóan a tüdő áttétek rendszeresen megelőzik az agyi (97%), a csont (86%) és a máj (84%) metasztázisokat. Hasonlóan, a mellékvese áttét jellemzően korábban azonosítható, mint az agyi metasztázis (89%), a mellhártyai pedig korábban jelenik meg, mint az áttét a májban (100%).

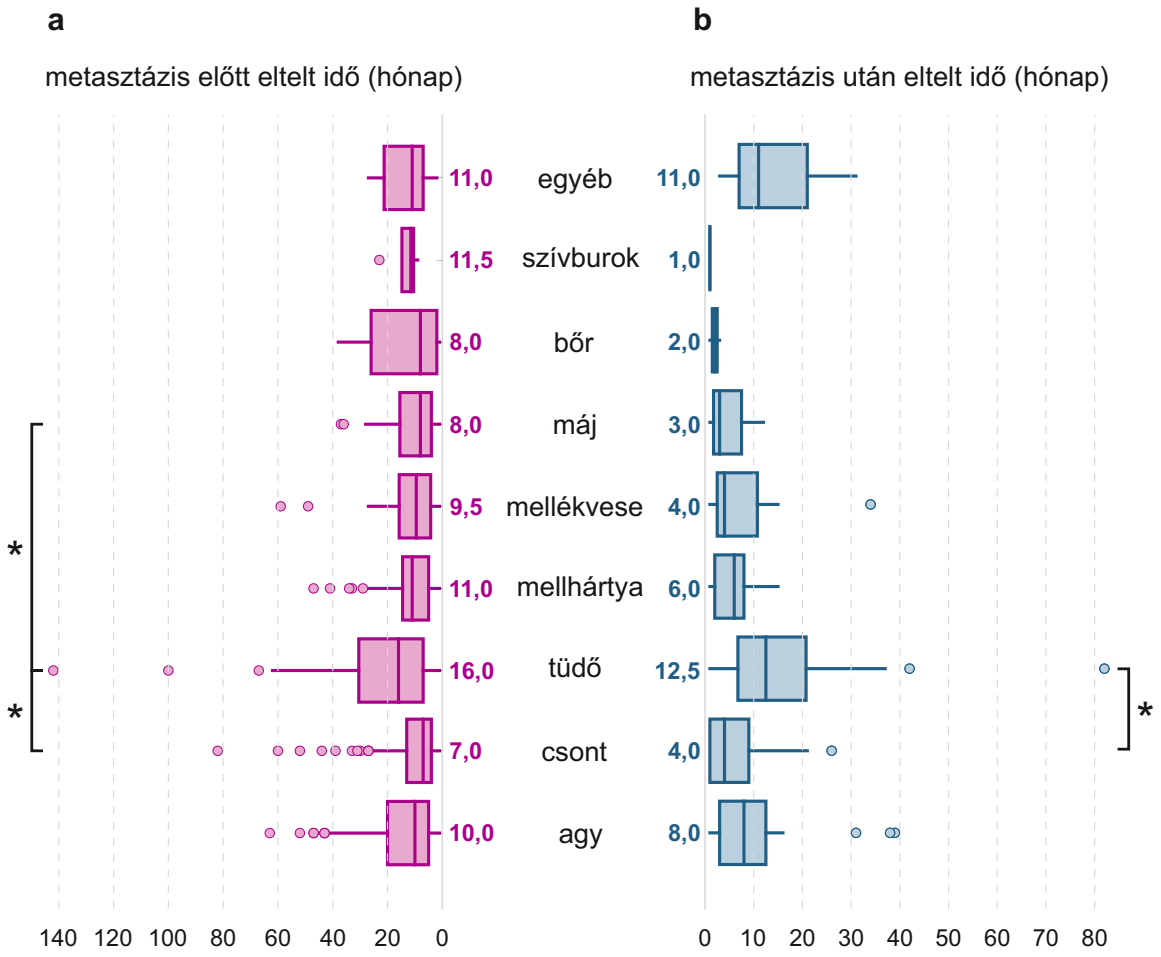
Ahhoz, hogy a metasztázisok között eltelő időről pontosabb képet kapjunk, megvizsgáltuk, hogy vannak-e olyan szervek, melyekben ha megjelenik egy áttét, akkor várhatóan a következő metasztázis rövid idő eltelte után szintén felbukkan, illetve, melyekben az áttét rövid idő után követi az azt megelőző metasztázist. Ehhez minden adott $i$ szervbeli áttét esetén megvizsgáltuk, hogy az érintett páciensben mennyi volt a metasztázis előtt eltelt $t_{i-} = t_i - t_m$ idő (ahol $t_m$ az $i$ szervbeli áttétet megelőző áttét diagnózisának ideje), illetve a metasztázis után eltelt $t_{i+} = t_n - t_i$ idő (ahol $t_n$ az $i$ szervbeli áttétet követő áttét diagnózisának ideje). A korai metasztázisokat kizártuk az elemzésből, mivel a megjelenésük pontos ideje bizonytalan. Azoknál a metasztázisoknál, melyek az adott páciens időben legkésőbb azonosított áttétei voltak, csak a $t_{i-}$ értékét számoltuk ki. A mért $t_{i-}$ és $t_{i+}$ értékek elosz-




**1. ábra. Metasztázisok megjelenése a különböző szerv-párokban. a.** Azoknak a páciensek a száma, akiknél áttétet azonosítottak az $i$ és $j$ szervekben is. A számok és a hőtérkép színei a betegek számát jelölik. Azokat a szerv-pár kategóriákat, melyekhez nem több, mint 5 páciens tartozott, nem jelöltük számokkal. A számok összege nem adja meg a teljes vizsgált kohorszban a betegek számát, hiszen egy páciensben gyakran nem pontosan két metasztázis jelenik meg. **b.** Az adott szerv-párokban megjelenő áttétek tényleges együttes valószínűségének és az elméleti együttes valószínűségnek a különbsége százalékban. A függetlenség tesztelését $\chi$-négyzet próbával végeztük, azzal a nullhipotézissel, hogy az áttét bekövetkezése a különböző szervekben egymástól függetlenül történik ($\alpha = 0,05$). Azokat a szerv-párokat jelöltük számmal, melyekre a hipotézisteszt eredménye szignifikáns volt, nagyobb és félkövér betűtípussal pedig azokat, melyekre a szignifikancia a Bonferroni-korrekció elvégzése után is érvényes maradt. **c.** Az $S_{ij}$ sorrend preferenciák százalékos értéke a szerv-párokra. A szignifikanciatesztelést $\chi$-négyzet próbával, azzal a nullhipotézissel végeztük, hogy a kétféle sorrendiség valószínűsége azonos ($\alpha = 0,05$). Azokat a szerv-párokat jelöltük számmal, melyekre a hipotézisteszt eredménye szignifikáns volt, nagyobb és félkövér betűtípussal pedig azokat, melyekre a szignifikancia a Bonferroni-korrekció elvégzése után is érvényes maradt. A fehér és fekete betűszínek az olvashatóságot segítik.

lásának jellemzőit mutatja rendre a 2. ábra a) és b) panelje. A tapasztaltak alapján ahhoz, hogy egy tüdő metasztázis kialakuljon az azt megelőző metasztázis után, tipikusan több idő kell (medián: 16,0 hónap), mint egy máj (medián: 8,0 hónap) vagy egy csont (medián: 7,0 hónap) áttét kialakulásához. Hasonlóan, az agyi áttétet követően a következő metasztázis általában később (medián: 12,5 hónap) azonosítható, mint egy csont (medián: 4,0 hónap) áttét esetén.

A vizsgálatok eredményei tehát azt mutatják, hogy a centrális primer tüdő tumorok a perifériku saknál agresszívebbek és hajlamosabbak a korai áttétképzésre. Ilyen tendencia a jobb/bal oldali és alsó/felső régiós elhelyezkedés esetén nem mutatkozott. Ezen felül azt tapasztaltuk, hogy bizonyos szerv-párok esetén a mindkettőben előforduló metasztázisok együttes bekövetkezésének a valószínűsége nem egyezik meg azzal, amit akkor várnánk, ha az áttétek véletlenszerűen követnék egymást. Emellett néhány szerv-pár esetén az egyik




**2. ábra. Metasztázisok között eltelt idő az adott szervbeli metasztázis előtt (a.) és után (b.).** A függőleges tengely mentén lévő színes számok az adott szervre mért értékek mediánját mutatják. A különböző szervekre jellemző időket a Kruskal–Wallis H-próba segítségével hasonlítottuk össze, azzal a nullhipotézissel, hogy a két szervre mért értékek mediánja megegyezik ($\alpha = 0,05$). Azokat a szerv-párokat, melyekre a nullhipotézis a többszörös tesztelésre való korrekció figyelembevételével együtt is elvethető volt, az ábrán csillaggal jelöltük.

szervben megjelenő metasztázis rendszeresen megelőzi a másik szervbelit, a metasztázisok között eltelt idő pedig erősen függ az érintett szervtől. Ezek a megfigyelések felhívják a figyelmet arra, hogy a különböző elhelyezkedésű primer tumorral diagnosztizált betegek kezelésénél különböző irányelveket kell követni, illetve, hogy a betegség lefolyása során kialakuló áttétek előre jelezhetik a továbbiakban esetlegesen érintett szerveket is.

### Vesefunkciók romlása tüdődaganatos, csontáttétes betegek kemoterápiás kezelése során

Ahogy a fentiek alapján is látható, a primer tüdő daganatban szenvedő páciensekben a leggyakrabban megjelenő metasztázisok egyike a csont áttét, mely a betegek 25-30%-ánál már a primer diagnózis során azonosítható, míg kb. 10%-uknál késői metasztázisként jelentkezik [28]. Az ilyen esetekben alkalmazott kezelések közül talán a biszfoszfonátok

használata a legelterjedtebb, melyek a kutatások szerint a csont metasztázisok kialakulásának megelőzésében is fontos szerepet töltenek be [29]. Ugyanakkor az olyan pácienseknél, akiknek a vesefunkcióik gyengültek, az ilyen típusú szerek ellenjavalltak [30–32]. Mivel a tüdőrák elsősorban az idős és jellemzően dohányzó emberekben jelentkezik, ezekben az esetekben a dohányzás okozta szív- és érrendszeri problémák is hozzájárulnak a vesefunkciók romlásához, továbbá a műtéti úton nem eltávolítható primer daganatok kezeléséhez gyakran használt platina-bázisú kemoterápiás szerek is nefrotoxikusak [33].

Annak a felmérésére, hogy a tüdő daganat lefolyása során a vesefunkciók miként változnak, illetve, hogy a romlásuk milyen mértékben gátolja a biszfoszfonát kezelés alkalmazását, egy 570 páciensből álló adathalmazt dolgoztunk fel [34]. A betegek mindegyikénél kialakult csont áttét a betegség során, a primer tumorok azonban különböző szövettani kategóriákba estek. A páciensek csaknem fele (kb. 41%) részesült platina-bázisú kemoterápiában, háromnegyedük (kb. 77%) pedig valamilyen típusú biszfoszfonát kezelésben a csont metasztázis megjelenését követően. A betegek fele magas vérnyomástól, 15%-uk cukorbetegségtől, harmaduk pedig COPD-től (krónikus obstruktív tüdőbetegség) szenvedett. A vesefunkciók degradálódásának elemzéséhez a betegek laborleletei alapján a kreatinin (normális tartomány: 36-106 $\mu$mol/L) és a BUN (blood urea nitrogen; karbamid) (normális tartomány: 1,7-8,3 mmol/L) szintjét rögzítettük a primer diagnóziskor, a csontáttét megjelenésének idején, illetve az utolsó elérhető adatok szerint. Az analízis során gyakorta mindkét mennyiséget egy bináris skálán (normális/kóros) jellemeztük.

Ahhoz, hogy a társbetegségek hatását megvizsgáljuk, minden lehetséges társbetegség csoportra meghatároztuk azoknak a pácienseknek az arányát a primer diagnózis idejekor, akiknek a kreatinin, illetve a BUN szintje a kóros tartományba esett (KTA: kóros tartomány arány, A1. táblázat). Mivel ekkor még egyik beteg sem részesült semmilyen kezelésben, így ezek hatását ezeknél a laborleleteknél nem kell vizsgálnunk. A különböző csoportokba eső páciensek száma alapján úgy tűnik, hogy a társbetegségek alapján történő 8 csoport létrehozása némely kategóriákban túl kevés esetszámot eredményez a későbbi statisztikai elemzéshez. Preferált lenne tehát kiválasztani azt a társbetegséget, melynek a hatása leginkább érződik a vesefunkciókat jellemző paraméterek értékein. Ennek a meghatározásához egymástól függetlenül összevetettük a kreatinin és a BUN értékek mediánját a három társbetegség által meghatározott, páronként diszjunkt (van/nincs) beteghalmazokon. Azt tapasztaltuk, hogy a magas vérnyomástól szenvedő betegekben mind a kreatinin szint (medián: 81,9 vs. 75,8 $\mu$mol/L; $p < 0,001$), mind pedig a BUN szintje (medián: 6,0 vs. 5,7 mmol/L; $p = 0,005$) szignifikánsan magasabb volt, mint azoknál, akiknek nem volt magas a vérnyomása. A mediánokat Mann–Whitney-próbával hasonlítottuk össze $\alpha = 0,05$ szignifikancia szint mellett. A fenti eredmények a többszörös tesztelés kompenzálására használt Bonferroni-korrekció mellett is szignifikánsak maradnak. Ehhez hasonló tendenciát nem találtunk akkor, mikor a betegeket a cukorbetegség megléte alapján kategorizáltuk, a COPD-től szenvedő páciensek esetében pedig csak a kreatinin szint volt szignifikánsan

magasabb (81,1 vs. 77,3 $\mu$mol/L; p=0,004), mint a COPD-mentes betegeknél.

Ezek alapján a 3. ábrán már csak a magas vérnyomás szerint csoportosított betegekre határoztuk meg a különböző időpontokban a KTA értékét, illetve a romlási rátát (RR), melyet a két időpont között újonnan kóros tartományba átkerülő betegek arányaként definiáltunk. (Ennek értelmében $\mathrm{KTA}_d + \mathrm{RR}_{d\to c} = \mathrm{KTA}_c$ és $\mathrm{KTA}_c + \mathrm{RR}_{c\to u} = \mathrm{KTA}_u$, ahol az alsó indexek a diagnózis idejére ($d$), a csont áttét megjelenési idejére ($c$) és a legutolsó elérhető adat idejére ($u$) utalnak.)

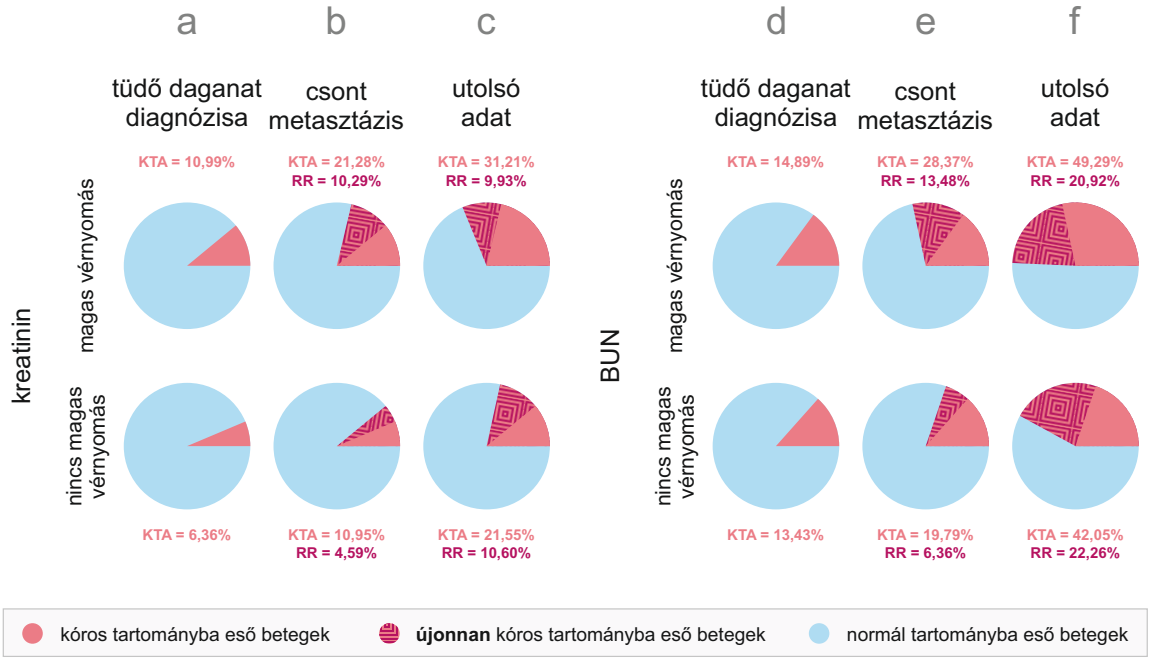
A teljes betegcsoportra nézve azt tapasztaltuk, hogy a csont metasztázis megjelenésénél mért kreatinin (medián: 77,0 $\mu$mol/L) és BUN (medián: 6,0 mmol/L) értékek jelentősen meghaladták a primer diagnóziskor felvett értékeket (kreatinin medián: 75,0 $\mu$mol/L; BUN medián: 5,4 mmol/L). A mediánok összevetéséhez használt Mann–Whitney-próba ($\alpha = 0,05$) szerint mindkét mennyiség értéke szignifikánsan nőtt ($p < 0,001$) az idő előrehaladtával. Hasonlóan, az utolsó mérési eredmények alapján (kreatinin medián: 83,0 $\mu$mol/L; BUN medián: 7,60 mmol/L) a csont áttétet követően is mindkét mennyiségre szignifikáns ($p < 0,001$) növekedés tapasztalható. (Ezek az eredmények Bonferroni-korrekció alkalmazása mellett is szignifikánsak maradtak.)

A 3. ábra b) és e) paneljeinek tanúsága szerint a diagnózistól a csont metasztázisig eltelt időben azok a páciensek, akiknek magas a vérnyomása, hajlamosabbak mind a kreatinin, mind a BUN szintjüket tekintve átlépni a kóros tartományba. A magas vérnyomásos csoportban az új esetek aránya 10,29% (kreatinin), illetve 13,48% (BUN) volt, míg ugyanezek az értékek a többi páciensre 4,59% (kreatinin) és 6,36%-nak (BUN) adódtak. A két csoport közti eltérés statisztikai szignifikanciáját Fisher-féle egzakt teszttel, $\alpha = 0,05$-os szignifikancia szint mellett becsültük, mely alapján mindkét mennyiségre szignifikáns eltérést tapasztaltunk (kreatinin: $p = 0,010$; BUN: $p = 0,005$) Bonferroni-korrekció alkalmazása mellett is.

A társbetegségtől függetlenül a csont metasztázis megjelenése után mintha felgyorsulna a vesefunkciók romlása: a nem magas vérnyomásos csoportban a kreatininre mért romlási ráta 4,59%-ról 10,60%-ra (3. ábra b) és c) panel) emelkedett, míg a BUN-ra ugyanez 6,36%-ról 22,26%-ra (3. ábra e) és f) panel). A magas vérnyomásos betegeknél pedig a BUN esetén a 13,48%-os romlási ráta a csontáttét után 20,92%-osra nőtt (3. ábra b) és c) panel).

Ezzel szemben a csontáttét megjelenése után az újonnan kóros esetek aránya a két betegcsoportban alig tér el egymástól (kreatinin: 9,93%, illetve 10,60%, BUN: 20,92% és 22,26%). Ebből arra lehet következtetni, hogy a csont metasztázis megjelenésével olyan erőteljesen nefrotoxikus folyamatok veszik kezdetüket a szervezetben, melyek a korábbi körülményektől (társbetegség) függetlenül felgyorsítják a veseműködés romlását. Nagyon fontos itt megjegyeznünk, hogy az ilyenkor beinduló nefrotoxikus folyamatok nem feltétlenül a csont áttét közvetlen biológiai következményei. A fentieknek megfelelően a vizsgált betegcsoport háromnegyede részesült biszfoszfonát kezelésben, mely köztudottan negatív

hatással van a veseműködésre. Tehát ebben az esetben nem tudjuk minden kétséget kizáróan elkülöníteni magának a csont metasztázis megjelenésének az inherens nefrotoxicitását a biszfoszfonát kezelés ismert következményeitől. Mindazonáltal azokra a betegekre korlátozva az elemzést, akik nem kaptak biszfoszfonát kezelést, azt tapasztaltuk, hogy a csont metasztázis után szintén nőtt a romlási ráta a csont áttét előtti értékhez képest. Ez az eredmény azt sejteti, hogy a csont metasztázis megjelenésével a vesefunkciók rohamos romlásnak indulnak, amit csak részben magyaráz meg a biszfoszfonát kezelés.




**3. ábra. A kreatinin és a BUN szintjének alakulása a tüdő daganat diagnózisától az utolsó laboratóriumi eredményekig a magas vérnyomástól érintett és nem érintett páciensekre.** KTA: az adott időpontban az adott betegcsoportban a kóros tartományba sorolt kreatinin/BUN szinttel rendelkező betegek aránya. RR: az adott betegcsoportban az előző időponthoz képest újonnan a kóros tartományba átlépő kreatinin/BUN szinttel rendelkező betegek aránya.

Mivel a fenti megközelítés nem volt megfelelő ahhoz, hogy tényleges különbséget tudjunk tenni a biszfoszfonát kezelés és a csont áttét megjelenésének nefrotoxikus hatása között, egy általánosított túlélés-elemzést végeztünk az adatokon.

Hagyományosan a túlélés vizsgálatánál arra a kérdésre keressük a választ, hogy bizonyos szempontok szerint csoportosítva a betegeket (például kezelés típusa, nem, stb.), megfigyelhető-e olyan jellegű tendencia, hogy az egyik csoportban a betegek „általában tovább élnek”, mint a másik csoportban. Kicsit konkrétabban fogalmazva a két csoportban egy adott esemény bekövetkezéséhez szükséges idők túlélési függvényét kívánjuk összehasonlítani. Legyen $T \geq 0$ egy folytonos véletlen valószínűségi változó, melynek eloszlásfüggvénye $F(t) := P(\{T \leq t\})$, valószínűségi sűrűsége pedig $f_T$. Ekkor a túlélési függvény

definíció szerint

$$S(t) = P(\{T > t\}) = \int_t^\infty f_T(u)du = 1 - F(t).$$

Egyszerűbben fogalmazva a túlélési függvény azt mondja meg, hogy mekkora annak a valószínűsége, hogy az adott esemény $t$-nél későbbi időpontban következik be. Hasznos még megadnunk a $h(t)$ kockázati függvény (hazard function) definícióját:

$$h(t) = \lim_{dt\to 0} \frac{P(t \le T < t + dt)}{dt \cdot S(t)} = \frac{f_T(t)}{S(t)} = \frac{-S'(t)}{S(t)},$$

ahol az utolsó egyenlőség igaz, mert $S'(t) = \frac{d}{dt}[1 - F(t)]$ és $f_T(t) = \frac{d}{dt}F(t)$. Vagyis $h(t)$ azt adja meg, hogy mekkora az esemény bekövetkezésének a valószínűsége a $t$ időpontban, feltéve, hogy addig még nem következett be.

A gyakorlatban a túlélési függvény konkrét alakját nem ismerjük, de a rendelkezésre álló adatokból becslést tehetünk rá, melyre az egyik legelterjedtebb módszer a Kaplan-Meier közelítés [35]. Jellemzően a számunkra elérhető információ a megfigyelt $N$ alanyról az, hogy az $n$. alany megfigyelésének teljes $c_n$ ideje alatt bekövetkezett-e az adott esemény ($\delta_n = \{0;1\}$) és ha igen, akkor pontosan mikor ($t_n$). Amennyiben az alanynál nem következett be az esemény a megfigyelés ideje alatt, cenzorált adatról beszélünk. Így minden $t_i$ diszkrét időpillanatot tekintve ismerjük az adott pillanatban bekövetkezett események $d_i$ számát, illetve azoknak az alanyoknak az $n_i$ számát, akiknél $t_i$ előtt még nem következett be az esemény. Ezek felhasználásával minden $t_i$-re kiszámítható az aktuális diszkrét, becsült $\bar{h}_i = d_i/n_i$ kockázata az esemény bekövetkezésének. Az összefüggés egyszerű meggondolások alapján adódik: annak a becsült valószínűsége az adatok alapján, hogy egy adott pillanatban bekövetkezik az esemény annyi, mint az adott pillanatban ténylegesen bekövetkezett események és az összes lehetséges érintett alany számának hányadosa. Ugyanezt természetesen például a maximum likelihood módszerrel is beláthattuk volna. Ennek felhasználásával a becsült túlélési függvény előáll, mint

$$\bar{S}(t) = \prod_{i:t_i \le t} (1 - \bar{h}_i) = \prod_{i:t_i \le t} (1 - \frac{d_i}{n_i})$$

Az így becsült túlélési függvényeket ábrázolva jellegzetes lépcsős görbéket kapunk, melyeket a két csoportra ábrázolva manuálisan összevethetünk, illetve a log-rank teszt segítségével eldönthetjük, hogy az adott szignifikancia szint mellett szignifikánsan eltérnek-e egymástól. Ahogy $N \to \infty$, a lépcsők egyre inkább kisimulnak és $\bar{S}(t) \to S(t)$.

Mindazonáltal a különböző paraméterek és körülmények túlélési függvényre gyakorolt hatásáról a becsült $\bar{S}(t)$ alapján még semmit nem mondhatunk. Az erre a feladatra az orvosi gyakorlatban leggyakrabban használt megoldás az ún. Cox-regressziós modell illesztése az adatokra, melyből a vizsgált paraméterek megváltozásának a kockázatra gyakorolt hatása

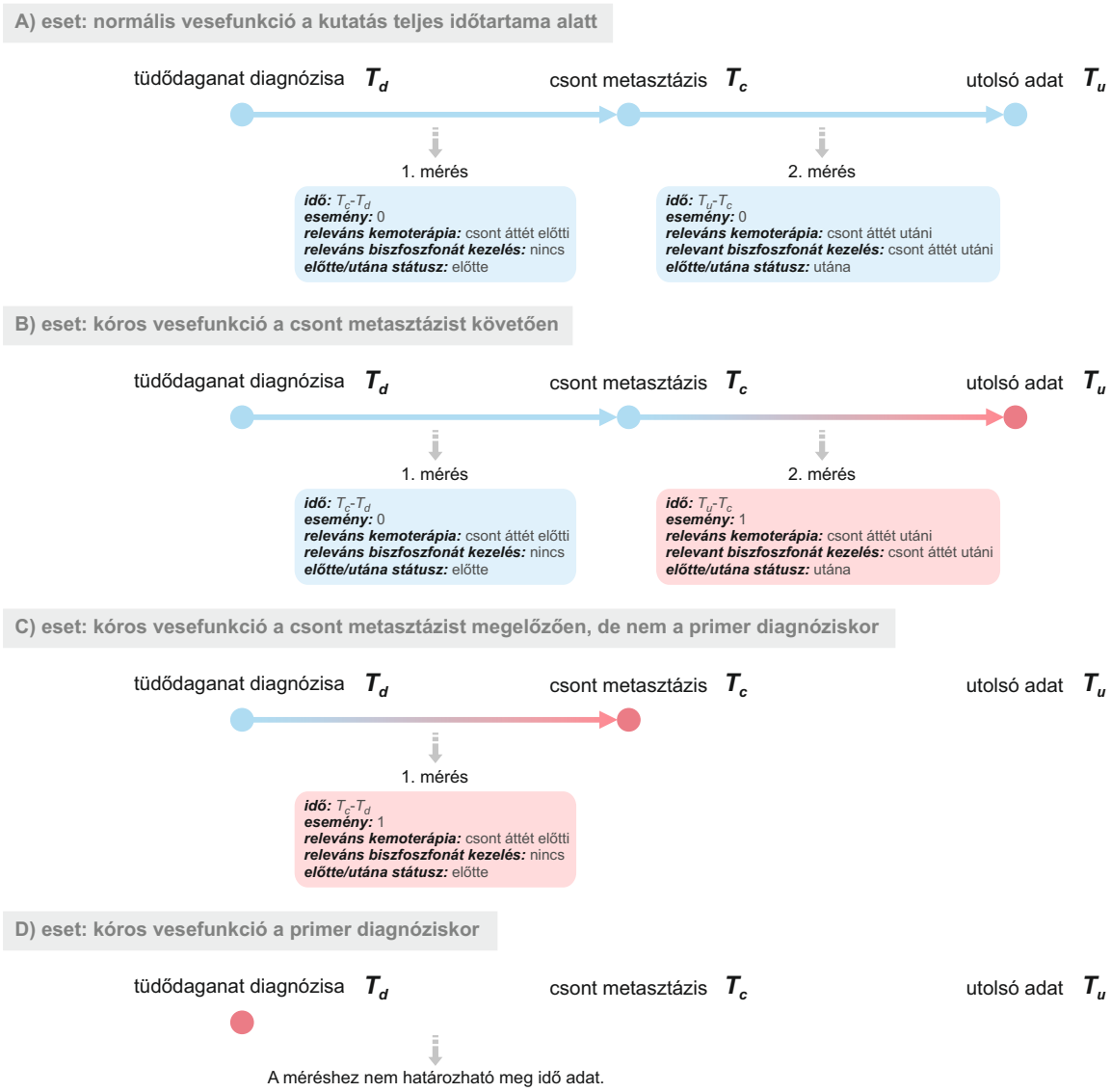
megbecsülhető. A modell legegyszerűbb verziója a kockázati függvényt az alábbi módon állítja elő:

$$h(t,X_i) = h_0(t)e^{(\beta_1 X_{i1}+\ldots+\beta_p X_{iK})},$$

ahol $X_i$ az $i$. betegre vonatkozó kovariánsok vektora, $h_0(t)$ az egyén alapkockázata („baseline") akkor, ha a kovariánsok mind nullák, $K$ pedig a kovariánsok száma. A modell másik elnevezése a Cox-féle arányos kockázati modell, hiszen a fenti képletből látható, hogy bináris kovariánsok esetén (pl. a beteg kapott-e biszfoszfonát kezelést vagy sem: $X_{i,bf} = \{0;1\}$) az adott kovariáns értéke által definiált két csoportban a kockázati függvények konstansszorosai ($e^{\beta_{bf}}$) egymásnak. Világos, hogy ez egy viszonylag szigorú feltevés, hiszen például előfordulhat, hogy egy bonyolult műtéti eljárás ugyan a beavatkozást követően jelentősen megnöveli a kockázatot, de hosszú távon lecsökkenti azt. A modellnek léteznek kiterjesztései, melyek időfüggő regressziós együtthatókat is képesek kezelni. Ugyanakkor az is elképzelhető, hogy egy bizonyos paraméter csak akkor játszik igazán jelentős szerepet a kockázatban, ha az alapvetően nem túl magas. A fenti modell alapján a biszfoszfonát kezelés a veseelégtelenség kockázatát például megduplázza, a többi kovariáns értékétől függetlenül. El tudunk azonban képzelni olyan pácienseket, akik mindhárom vizsgált társbetegségben szenvednek, így a számukra már önmagában magas kockázatot a biszfoszfonát kezelés ugyan kicsit megnöveli, de nem kétszerezi. Az ilyen esetek a fenti modellbe nem építhetőek bele.

A modell illesztésének végső célja nem a kockázati függvény konkrét előállítása, mivel a $h_0(t)$ függvényről semmit nem tudunk és semmilyen feltételezést nem is teszünk rá. A feladat tehát a $\beta_k$ regressziós paraméterek meghatározása, vagyis az egyes kovariánsok relatív kockázatát kívánjuk megadni. Tehát végeredményben arról nem tudunk nyilatkozni, hogy adott $t$ időpillanatban az $i$. páciens esetében mekkora a valószínűsége a veseelégtelenség bekövetkezésének, de annyit állíthatunk, hogy ha a páciens kapott biszfoszfonát kezelést, akkor ez az ismeretlen valószínűség $e^{\beta_{bf}}$-szer akkora lesz, mintha nem kapott volna. Magukat a kovariánsokat az összes páciensre vett együttes likelihood függvény maximalizálásával lehet meghatározni. (Mivel $h_0(t)$ nem függ a kovariánsoktól, így a számítás során kiküszöbölhető.)

A konkrét adathalmazra vonatkozóan egy a hagyományostól kicsit eltérő, absztrakt túlélési analízist végeztünk a kreatinin és a BUN tekintetében külön-külön. Ehhez a méréseket a 4. ábra szerint definiáltuk: amennyiben a páciens veseműködését jelző paraméter értéke a kutatás teljes ideje alatt a normális tartományban maradt (A) eset), két cenzorált mérést jegyeztünk le. Ha a páciensnél az adott paraméter értéke a csont áttét diagnosztizálása után (vagyis mérhetően az utolsó adat szerint) váltott a kóros tartományba (B) eset), szintén két mérést könyveltünk, egy cenzoráltat a csont áttét előtt, illetve egy nem cenzoráltat utána. Ha már a csont metasztázis diagnózisánál kóros vesefunkciókat tapasztaltunk




**4. ábra. A túlélés analízishez használt mérések a különböző típusú betegek esetében**

(C) eset), egyetlen nem cenzorált mérést jegyeztünk. Amennyiben a kreatinin vagy BUN szintje már a primer diagnóziskor a kóros tartományba esett (D) eset), nem lehetett az átmenet idejét meghatározni, így innen nem származtak mérési adatok. Eseménynek a kreatinin vagy a BUN szintjének a kóros tartományba való átlépését tekintettük, a kétféle mennyiség esetén külön modelleket illesztettünk. A csont metasztázis előtti és utáni méréseket az „előtte/utána” bináris paraméter értékével különböztettük meg.

Természetesen ennek a modellnek több hátulütője is van. Egyrészt a méréseket egymástól függetlennek tekintettük, pedig a valóságban egyetlen páciens adataiból több mérést is generáltunk. Alapvetően nyilvánvaló, hogy ha már a csont metasztázis azonosításánál a kreatinin és/vagy BUN szintje erősen megközelíti a kóros tartomány határát, akkor később valószínűleg a csont metasztázis megjelenésétől, illetve egyéb kezelésektől függetlenül is könnyen átlép abba. Mindazonáltal a vesefunkciók kezdeti állapota bár mérésenként különböző, ami a kockázat konkrét értékét más-más mértékben befolyásolja, ezt a hatást be-

sűríthetjük a $h_0(t)$ függvénybe. Továbbá az adott méréshez rendelt kemoterápia értékének (kapott/nem kapott) mindig az adott időszakra meghatározott értéket tekintettük. Vagyis egy olyan páciensnél, aki a csont áttét előtt részesült kemoterápiában, de utána nem, az adataiból kinyert, áttétet követő mérési ponthoz a „nem kapott" értéket rendeltük. Ezzel tehát elhanyagoltuk a kemoterápiás szerek hosszútávú hatásait. Fontos továbbá megjegyezni, hogy a biszfoszfonát kezelés értéke (kapott/nem kapott) erősen korrelál az előtte/utána státusszal, hiszen a csont metasztázist megelőzően egyik beteg sem kapott ilyen típusú kezelést. Mivel azonban a betegek kb. negyede az áttétet követően sem részesült biszfoszfonát kezelésben, fontos a két paramétert megkülönböztetnünk.

A modellillesztés előtt minden vizsgált paraméter (A2. táblázat) szerinti csoportosításra legyártottuk az adott csoport túlélési görbéjének $\bar{S}(t)$ Kaplan-Meier becslését és grafikusan ellenőriztük, hogy fennáll-e az arányos kockázatok feltétele. Ehhez elsőként be kell vezetnünk az ún. kumulatív kockázati függvényt:

$$H(t) := \int_0^t h(u)du = \int_0^t \left(\frac{-S'(u)}{S(u)}\right) du = [-\log S(u)]_0^t = -\log S(t)$$

Mivel a Cox-modell feltételezése szerint $h(t) = h_0(t)e^{\sum_k \beta_k X_k}$, így

$$\begin{aligned} H(t) = \int_0^t h(u)du &= \left[\int_0^t h_0(u)du\right] e^{\sum_k \beta_k X_k} = H_0(t)e^{\sum_k \beta_k X_k} \\ S(t) = e^{-H(t)} &= \left(e^{-H_0(t)}\right)^{e^{\sum_k \beta_k X_k}} \\ \log S(t) &= -H_0(t)e^{\sum_k \beta_k X_k} \\ \log(-\log S(t)) &= \log H_0(t) + \sum_k \beta_k X_k \end{aligned}$$

Ha tehát csak egyetlen kovariánst tekintünk, és ennek az értéke (0 vagy 1) szerint választjuk szét a csoportokat, a két csoportban a becsült túlélési függvény $\log(-\log \bar{S}(t))$ transzformáltjai csak a kovariánshoz tartozó $\beta$ tagban fognak egymástól eltérni. Vagyis a két csoportra egy ábrán felrajzolva a $\log(-\log \bar{S}(t))$ görbéket a $t$ függvényében, párhuzamos, egymástól egy konstans értékkel eltolt egyeneseket kell kapnunk. Amennyiben tehát ez vizuálisan igazolható, az adott paraméterre valóban fennáll az arányossági feltétel és így az beválogatható a Cox-modell kovariánsai közé.

Miután ilyen módon kiválasztottuk azoknak a paramétereknek a listáját, melyek teljesítik az arányos kockázatok feltételét, megkezdtük a modellszelekciót. Ennek során különböző paraméter kombinációkkal illesztettünk modelleket, majd mindegyiknél meghatároztuk az AIC (Akaike-féle információs kritérium) [36] értékét. Végül azt a modellt választottuk ki, melynél az AIC értéke a legalacsonyabb volt (A3. táblázat). A jelentős mennyiségű cenzorált adatot tartalmazó adathalmazok esetén a hagyományosan a modell prediktív erejét jellemző $R^2$ paraméter értéke drasztikusan lecsökkenhet [37], így célszerű

más módon származtatni a jósági tényezőt. A szintén gyakran használt konkordancia [38] értéke a kreatinin esetében 0,65, a BUN esetében pedig 0,71 lett. A likelihood arány próba által kapott p-értékek mindkét modell esetén szignifikánsak voltak $\alpha = 0,05$ szignifikancia szint mellett. Alapvetően a likelihood arány próba abból a nullhipotézisből indul ki, hogy két tetszőlegesen megválasztott modell mellett a konkrét adatok megfigyelési valószínűségének (likelihoodjának) a hányadosa nem tér el jelentősen 1-től. Vagyis a két modell közül egyikről sem állíthatjuk, hogy jobban leírná a megfigyelt adatokat, mint a másik. Esetünkben az egyik modell a fenti modellszelekciós eljárás végső Cox-féle modellje, míg a másik modell a $\beta_i = 0 \forall i$. Ezek alapján tehát a kapott modellek bár láthatóan nem adnak tökéletes jóslatot a várható túlélési időre vonatkozóan, a kovariánsokat nem tartalmazó modelleknél azonban szignifikánsan jobbnak bizonyulnak.

A végső modellekben szereplő kovariánsok listájából (A3. táblázat) kitűnik, hogy a késői stádiumú primer tumorok esetén, illetve a magas vérnyomástól szenvedő pácienseknél a vesefunkciók romlása várhatóan hamarabb következik be. A biszfoszfonát kezelés és a kemoterápia nefrotoxikus hatására nem találtunk meggyőző bizonyítékot, ezzel szemben mind a kreatinin, mind pedig a BUN esetében szignifikánsan gyorsabban léptek át a vesefunkciók a kóros tartományba a csont metasztázist követően, mint előtte. Vélhetően a modellek fent diszkutált hiányosságai miatt ez az eredmény önmagában nem tekinthető bizonyító erejűnek arra vonatkozóan, hogy a csont metasztázis megjelenése inherensen felel a veseműködés romlásáért. Ennek a tényleges teszteléséhez egy olyan adathalmazra lenne szükség, ahol a primer tüdő tumorral rendelkező betegek egy részében kialakult csontáttét, a másik részében pedig nem. Ugyanakkor az eredmények feltétlenül felhívják arra a figyelmet, hogy a nefrotoxikus szerek esetében más körülményekkel és kockázati tényezőkkel kell számolni akkor, ha azokat preventív céllal (a csont metasztázis előtt) vagy reaktívan (a csont áttétet követően) alkalmazzák.

### Immunterápiás biomarkerek változásai metasztázisokban és platina-bázisú kemoterápia hatására

A hagyományos, diagnózis során felvett klinikai paraméterek és a laborvizsgálatokból nyerhető adatok vizsgálata mellett az utóbbi években elterjedni kezdő immunellenőrzőpontgátló szerek alkalmazásához elengedhetetlen az immunterápiás biomarkerek változásainak a feltérképezése is. A fent tárgyaltaknak megfelelően az ilyen típusú kezelés elsősorban azokban az esetekben vezet tényleges eredményre, amikor a tumorsejtek valóban a PD-L1 fehérje expresszálásával kerülik el a szervezetben normálisan fellépő immunválaszt [39]. Ezen kívül aktívan kutatják a terápiára adott válasz és a daganat környéki immunsejtek PD-L1 és PD-1 expressziója közti kapcsolatot is [40], illetve maguknak a daganat környéki és az azt infiltráló immunsejtek darabszámának és lokalizációjának hatását [41]. Tekintettel arra, hogy az immunterápia rendkívül költséges, fontos az olyan biomarkerek alkalmazása a várhatóan pozitívan reagáló betegcsoportok kiválasztása során, melyek valóban

megbízhatóan jósolják a kezelés kimenetelét. Ezért elengedhetetlen annak a felmérése is, hogy a különböző tényezők (egyéb kezelések, a mintavételezés helye, a primer tumor típusa, metasztázisok megjelenése, stb.) mennyiben befolyásolják a pácienseknél mérhető expressziós szinteket és a daganat környéki immunaktivitást. Mivel maga az immunhisztokémiai vizsgálat elvégzése, vagyis a biomarkerek értékeinek meghatározása is komoly anyagi terhet jelenthet, érdemes az értékeket hagyományos klinikai paraméterekkel is korreláltatni. Amennyiben fény derülne szignifikáns összefüggésekre az alapértelmezetten meghatározott klinikai változók és az expressziós szintek között, egy preszelekciós fázisban már a diagnózis során leszűkíthető lenne az immunterápiából várhatóan profitáló páciensek csoportja. A kezdeti biztató eredmények ellenére a betegek szelekciójára vonatkozóan egyelőre nincsenek bevett módszerek, és a PD-1/PD-L1 pozitivitás, mint biomarkerek tesztelésére vonatkozó feltételek sincsenek rögzítve [42]. Hasonlóan nem tisztázott, hogy azok a tumortípusok, melyeknél a PD-1/PD-L1 expresszió értéke prognosztizálja a várható túlélést, vajon jobban reagálnának-e az esetleges PD-1/PD-L1 gátló terápiákra, mint azok a daganattípusok, amiknél az expressziós szintek nem befolyásolják a túlélést [43, 44].

Itt fontos megjegyeznünk, hogy sajnos a PD-1 és PD-L1 expressziós szintek meghatározása a gyakorlatban sem egy jól kidolgozott, reprodukálható protokoll alapján történik. Az immunhisztokémia (IHC) során a vizsgált szövet egy vékony metszetére valamilyen módon megjelölt (jellemzően valamiféle enzimmel vagy fluoreszcens anyaggal) antitesteket juttatnak, melyek a szövet azon sejtjeihez kötnek, amiknek a felületén a komplementer antigén megtalálható. A metszetet ezután mikroszkóppal vizsgálva, az antigénnel rendelkező és nem rendelkező sejtek más-más színnel jelennek meg. Az expressziós szint számszerűsítését a metszetek mikroszkópos vizsgálatával patológusok végzik, többnyire vizuális besorolás útján valamiféle szemikvantitatív skála alapján. Gyakori például a tumorsejtek PD-L1 expresszióját a metszeten megfestett sejtek arányának a 0-1%, 1-5%, 5-10%, 10-50% és 50-100% intervallumok valamelyikébe való besorolásával meghatározni [45]. Világos, hogy a mérési módszer már önmagában rendkívül sok problémát felvet. Mivel a meghatározást manuálisan végzik, így a kétes esetekben a konkrét patológus szubjektív döntésén múlik, hogy melyik tartományba sorolja a kérdéses mintát. Ennek az aspektusnak az egységesítésére talán van remény azáltal, hogy a képek különböző szempontok szerinti klasszifikálását végző számítógépes algoritmusok az utóbbi időben soha nem látott fejlődésen mentek keresztül. A jelenleg leggyakrabban alkalmazott módszer azonban két független patológus egyéni döntéseinek összesítésén alapul. Az irodalmi tapasztalatok szerint a tumorsejtek PD-L1 expressziójának meghatározása tekintetében ez egy viszonylag megbízható technika [46–50], ezzel szemben azonban az immunsejtek PD-L1 pozitivitásának vizsgálata során az eredmények erősen függnek a használt antitesttől [46]. Emellett a teljes tumor hiteles reprezentálására metszetek sokaságát kellene a szövetből megvizsgálni, a daganatok közismert heterogenitása miatt. A gyakorlatban nem ritkán a teljes méréshez csak egy biopsziás minta áll rendelkezésre, így még ha annak többféle metszetét fel is térké-

pezik, akkor is fennáll a veszélye annak, hogy a teljes tumorra nézve tett következtetések tévesek lesznek [51]. Ugyanakkor nem szabad elfelejtenünk, hogy a daganatos betegségek gyógyítására alkalmazott immunterápiás módszerek és a várhatóan pozitívan reagáló betegek kiválasztásához szükséges feltételrendszer kialakítása még gyerekcipőben járnak. Így minden tudományos megfigyelés a biomarkerek tulajdonságait illetően fontos lépést jelenthet a technológia gyakorlatba történő átültetése során.

Annak a vizsgálatára, hogy a betegek várható túlélésének tekintetében valójában mennyire meghatározóak a különböző expressziós szintek mért értékei, illetve a daganat környékén található immunsejtek száma és eloszlása, egy olyan 208 pácienst tartalmazó adathalmazt vizsgáltunk [52], ahol a betegek mindegyike primer tüdő adenokarcinómában szenvedett, és a betegség lefolyása alatt agyi metasztázist diagnosztizáltak náluk. Mivel az agyi áttétek a tüdőrák mellett igen gyakran fordulnak elő, és az érintett páciensek túlélési statisztikái elszomorítóak [53], így feltétlenül relevanciája van egy ilyen speciális betegcsoport vizsgálatának. Bár az agyi metasztázisokat sokféle különböző kutatásban elemzik [54–56], ez az első olyan törekvés, mely kizárólag egy specifikus primer daganat agyi áttéteinek analízisét tűzte ki célul. A rendelkezésre álló szövetek metszetein a tumorba beszivárgó, kötőszöveti vázban lévő immunsejtek arányát hagyományos hematoxilin és eozin (H&E) festéssel két független patológus határozta meg a $< 20\%$ és $\geq 20\%$ kategóriákba történő besorolással. A daganat felületén vékony vagy vastag rétegben megjelenő immunsejtek ottléte alapján a metszeteket két csoportra osztották attól függően, hogy megjelent-e bármilyen vastagságú immunsejt réteg a tumor felületén vagy sem. A tumor- és immunsejtek PD-L1 és PD-1 pozitivitását a fent tárgyalt immunhisztokémiai eljárással a tumorsejtek esetén a 0-1%, 1-5%, 5-10%, 10-50% és 50-100% intervallumokba, az immunsejtek esetén pedig a 0-1%, 1-5%, 5-10% és 10-100% tartományokba sorolták a pozitivitást mutató sejtek arányától függően, az irodalmi standardnak megfelelően [44, 45]. A hagyományos klinikai paraméterek és a szövettani jellemzők figyelembevételével a fent ismertetett módon Cox-modellt illesztettünk a teljes túlélésre vonatkozó időadatokra (OS: overall survival; konkordancia: $0,75$). Az eredmények alapján a túlélést elsősorban befolyásoló tényezők a primer tumoron végzett műtét (a kockázat kevesebb, mint harmadára csökken, ha történt műtéti beavatkozás), a kemoterápia (a kockázat több, mint háromszorosára nő kemoterápia hiánya esetén), a páciens életkora (a 60 évnél fiatalabbak körében kb. feleakkora a kockázat), illetve a tumort körülvevő immunsejtréteg (ha nincs jelen, a kockázat több, mint másfélszeresre nő). Mivel az egyik fő faktor a primer tumor műtéti eltávolítása volt, egy klinikailag homogénebb alcsoportra ismételten elvégeztük a modellillesztést. Ebben a csoportban a páciensek mindegyike részesült primer műtétben, illetve csak egyetlen agyi metasztázist azonosítottak náluk. Az így definiált populációban a tumor körüli immunsejtréteg hiánya már háromszoros rizikófaktort eredményezett. Vagyis a tumorsejtek PD-L1 pozitivitásától függetlenül a várható túlélés azoknál a pácienseknél magasabb, akiknél a daganat környékén lévő immunsejtek száma magasabb. Ez arra utalhat,

hogy ezeknél a betegeknél magas PD-L1 expresszió esetén várhatóan pozitív hatása lehetne a fehérje blokkolásának, hiszen a immunsejtek jelenléte biztosítaná az immunválasz reaktiválását. Ezzel szemben az olyan páciensek, akiknél lokálisan nincsenek immunsejtek a daganat környékén, egy olyan kombinált terápiából profitálnának a legtöbbet, amelyben a PD-1/PD-L1 tengely blokkolásával párhuzamosan az immunsejtek sűrűsége a tumorban szintén növelhető [57]. Hozzá kell azonban tennünk, hogy önmagában a PD-1/PD-L1 pozitivitás prognosztikus jelentőségét nem tudtuk kimutatni a betegcsoporton, így előfordulhat, hogy az agyi áttéttel rendelkező páciensek esetén az anti-PD-1/PD-L1 terápiák nem feltétlenül hatásosak. Ezt a képet árnyalhatná a klinikailag sokkal homogénebb csoportok vizsgálata, hiszen nem lehetetlen, hogy a rendelkezésre álló viszonylag alacsony esetszám miatt a primer műtét és kemoterápia hatása elnyomja a molekuláris sajátosságok következményeit.

További kutatásaink során azokra a kérdésekre kerestük a választ, hogy megváltozik-e a tumorban és környékén található tumor- és immunsejtek PD-L1 és PD-1 expressziója neoadjuváns kemoterápia során [58], léteznek-e korrelációk a potenciálisan biomarkerként alkalmazható mennyiségek között egy adott páciensen belül, látunk-e összefüggést a páciens klinikai és a tumor szövettani jellemzői, illetve az expressziós szintek között [59], továbbá megfigyelhetőek-e különbségek az expressziós szintekben és a daganat környéki immunsejtek megjelenésében a primer tumor és annak metasztázisai között [60]. Az elemzések során hagyományos statisztikai módszereket használtunk az összefüggések feltárására. A korrelációkat a Spearman-féle $\rho$ rangkorrelációs együttható meghatározásával vizsgáltuk, aminek a hagyományos Pearson-korrelációhoz képest megvan az az előnye, hogy nem kifejezetten a két változó közti lineáris összefüggés azonosítására szolgál, hanem minden olyan esetben magas abszolút értéket vesz fel, mikor az egyik vizsgált mennyiség a másiknak monoton függvénye. A kapott $\rho$ érték tényleges szignifikanciája felmérhető egy olyan hipotézis-teszttel, mikor a nullhipotézis szerint a két mennyiség korrelálatlan (vagyis $\rho = 0$). A túlélési analízisek során a már fent tárgyalt Cox-féle többváltozós modellt használtuk az arányos kockázatok kritériumának változónkénti tesztelése, majd az AIC minimalizálására törekvő modellszelekciós lépések elvégzése után. Amennyiben a különböző paraméterek változásait vizsgáltuk a primer tumor és a metasztázis között, a pácienseket három csoportra osztottuk aszerint, hogy az adott mennyiség nőtt ($+1$), csökkent ($-1$) vagy nem változott (0). Az így számszerűsített változási irányok átlagán végzett t-próbával azt vizsgáltuk, hogy az átlag szignifikánsan eltért-e a nullától, vagyis látható-e bármilyen szisztematikus tendencia a változások irányában.

Különböző szövettani altípusokba tartozó primer tüdőtumorral diagnosztizált páciensek esetén azt vizsgáltuk meg, hogy a diagnóziskor gyűjtött biopszia és a neoadjuváns platina-bázisú kemoterápiát követően végzett műtéti eltávolításból származó minta között milyen eltérések voltak megfigyelhetők [58]. A tumort infiltráló immunsejtek arányát a fent leírt módon H&E festéssel, az expressziós szinteket pedig szintén a tárgyaltaknak

megfelelően immunhisztokémiai eljárással sorolták be a megfelelő szemikvantitatív skála intervallumaiba. Annak érdekében, hogy pontosabb képet nyerjünk a tumorsejtek PD-L1 expressziós szintjének változásáról, a metszeteket újrakategorizálták a nagyobb felbontású 0-1%, 1–5%, 6–10%, 11–20%, 21–30%, 31–40%, 41–50%, 51–60%, 61–70%, 71–80%, 81–90% és 91–100% skálán. Mivel összesen 41 páciens mintapárja állt csak rendelkezésünkre a kutatás során, a kapott összefüggéseket sokkal inkább előzetes tendenciákként, mint statisztikailag megerősített tényként kell interpretálnunk. A legszembetűnőbb eredmény talán az volt, hogy a neoadjuváns kemoterápia hatására a PD-L1-et kifejező tumorsejtek aránya az esetek kb. negyedében csökkent (24,4%), jelentős részében nem változott (68,3%) és csak elvétve nőtt (7,3%). Azoknál a pácienseknél, ahol csökkenést tapasztaltunk, a csökkenés gyakran igen számottevő volt (például 70%-os arányról <1%-ra). Ezek az eredmények a korábbi kutatások tapasztalataival is egybevágnak [61, 62]. Erre alapozva felmerülhet a gyanú, hogy a platina-bázisú kemoterápiás szerek célzottan a PD-L1 pozitív tumorsejteket pusztítják, így a globális expressziós szint csökkenését idézik elő. A rendelkezésünkre álló 10 olyan esetből, ahol ténylegesen csökkent a PD-L1 pozitív tumorsejtek aránya, 8-nál igazoltan csökkent a tumor mérete a kezelés hatására (1 esetben nem volt adatunk a terápiás válaszról). Annál a három betegnél azonban, ahol az expressziós szintben növekedést figyeltünk meg, a tumor mérete szintúgy lecsökkent, így a fenti elmélet önmagában még nem ad magyarázatot a látottakra. A tényleges orvosi relevanciával bíró tendenciák megfigyeléséhez természetesen jóval nagyobb adathalmazra lenne szükség. Mégis érdemes hangsúlyozni az eredmények jelentőségét: mivel az immunterápiát jelenleg tipikusan olyan esetekben alkalmazzák, ahol az elsődleges kemoterápiás kezelés, illetve a műtéti eltávolítás nem járt sikerrel, rendkívül fontos felmérni, hogy a kezdeti kezelések mennyiben befolyásolják azoknak a biomarkereknek az értékét, melyek alapján a páciens jogosulttá válik az immunterápiára. Amennyiben tehát a kemoterápia valóban megváltoztatja a PD-L1 expresszió szintjét, különös gondot kell fordítani arra, hogy a páciensek szelekcióját a kezelést követően gyűjtött mintákra alapozzák.

Bár a klinikai gyakorlatban egyelőre kizárólag a PD-L1 pozitív tumorsejtek arányát használják biomarkerként az immunterápiára való jogosultság meghatározásánál, felmerül az igény egyrészt precízebb, másrészt olcsóbban tesztelhető prediktorok azonosítására. Valójában csak a magas PD-L1 expresszióval rendelkező páciensek fele reagál jól a PD-1/PD-L1 inhibitor kezelésre, ezzel szemben pedig néhány esetben a PD-L1 negatív tumorok mérete is csökken az immunterápia hatására [63]. Mivel a diagnosztika során rutinszerűen alkalmazott H&E festésnél számos olyan szövettani paraméter meghatározásra kerül, melyeknek nincs közvetlen kihatása a kezelés kiválasztására [64, 65], praktikus lenne az ezek és a csak költségesen mérhető expressziós szintek közti összefüggéseket feltárni. A tüdő adenokarcinómától szenvedő betegek esetében eddig egyedül a lepidikus növekedési mintázat és a PD-L1 pozitív tumorsejtek aránya között igazoltak negatív korrelációt [66–68]. Laikus szempontból ez az összefüggés úgy magyarázható meg, hogy

lepidikus növekedéssel azok a típusú daganatok jellemezhetőek, melyeknél a tumorsejtek csak a meglévő tüdőhólyagocskák mentén helyezkednek el, és nem érték el a kötőszöveti vázat, az ereket, illetve a mellhártyát, vagyis kevésbé agresszívak. Ezzel szemben a PD-L1 pozitivitás arra utal, hogy a tumorsejtek sikeresen kijátszották a szervezet tumorellenes immunválaszát, vagyis hatékonyan tudnak növekedni. Így érezhető, hogy a kétféle jelenség várhatóan egymást kizáróan fordul elő. Ezzel együtt azonban világos, hogy az ilyen jellegű megérzéseket kétséget kizáróan igazolni kell, mielőtt a klinikai gyakorlatba átültethetőek lennének. A növekedési mintázat mellett a H&E festés során meghatározzák még a tumor rosszindulatúságának fokát (grade), az esetleges nekrózis (szövetelhalás) jelenlétét, az érrendszer érintettségét, illetve a kötőszöveti immunsejtek sűrűségét. Ezeknek a paramétereknek és expressziós szintek kapcsolatának vizsgálatához egy 268 tüdő adenokarcinómás páciensből álló betegcsoport műtéti mintáit elemeztük a fent tárgyalt H&E festési és immunhisztokémiai módszerekkel. A Spearman-korrelációs együtthatók meghatározása mellett kapott p-értékeket a többszörös tesztelésre a Holm-Bonferroni [69] módszerrel korrigáltuk és a továbbiakban csak azokat az eredményeket tárgyaljuk, melyek a korrekció után is szignifikánsak maradtak az $\alpha = 0,05$ szignifikancia-szint mellett.

Az eredményeink alapján a rosszindulatúság foka gyenge pozitív korrelációt mutat a nekrózis jelenlétével ($\rho = 0,325$), illetve gyengén negatívan korrelál a lepidikus növekedési mintázat megfigyelhetőségével ($\rho = -0,339$). Mindkét megfigyelés beleillik a naiv szemléletbe, miszerint az agresszívabb tumorok várhatóan szövetelhalással is járnak és nem a kevéssé invazív lepidikus növekedési mintázat szerint terjednek. Az expressziós szintek kapcsán a tumorsejtek és az immunsejtek PD-L1 pozitivitása közti pozitív korrelációt érdemes megemlítenünk ($\rho = 0,430$), melyet korábban kevert szövettani altípusú tüdődaganatos betegek csoportján már azonosítottak [70]. A vizsgált klinikai paraméterek (kor, COPD, dohányzás, nem) esetében egyedül az az összefüggés tűnt szignifikánsnak, hogy a dohányzók körében általában magasabb az immunsejtek PD-1 pozitivitása ($\rho = 0,275$). A nekrózis jelenléte és az immunsejtek PD-1, illetve a tumorsejtek PD-L1 pozitivitása között gyenge pozitív korrelációt figyeltünk meg (rendre $\rho = 0,290$ és $\rho = 0,283$), melyet korábbi kutatások nem demonstráltak. A fent tárgyalt szövettani paraméterek közti összefüggésekből logikusan következtetve, illetve korábbi irodalmi adatok alapján [66–68] a lepidikus növekedési mintázat, illetve a tumorsejtek PD-L1 expressziós szintje között negatív korrelációt vártunk, mely egyidejűleg mindhárom vizsgált expressziós paraméterre beigazolódott (immunsejtek PD-1: $\rho = -0,302$; immunsejtek PD-L1: $\rho = -0,306$; tumorsejtek PD-L1: $\rho = -0,329$). Ezek az eredmények bár önmagukban jelentéktelennek tűnhetnek, nem szabad elfelejtenünk, hogy a szervezetben fellépő antitumor immunválasz molekuláris mechanizmusai korántsem tisztázottak, így az ilyen heurisztikus megfigyelések a különböző paraméterek aggregált értékei között fontos ugródeszkát jelenthetnek a megértésükben.

Hasonlóan a kemoterápiás szerek hatásának vizsgálatához, felmerül a kérdés, hogy

amennyiben a primer tumor a betegség lefolyása során metasztatizál, elegendő-e pusztán a primer daganatból származó mintát elemezve levonni a következtetéseket az immunterápia alkalmazhatóságára vonatkozóan. Vagyis az expressziós szintek vajon változnak-e az eredetileg mért értékeikhez képest a különböző szövetekben megjelenő áttétekben. Mivel gyakran előfordul, hogy nincs lehetőség mind a primer tumor, mind pedig a metasztázis biopsziás mintavételezésére, így a két szövetben mért PD-1/PD-L1 pozitivitás közti erős korreláció arra utalna, hogy az egy szöveten meghatározott értékek jó becslések lehetnek a nem mintavételezhető szövet esetében is. Ennek a kérdésnek a feltérképezéséhez 61 páciens olyan primer tüdő adenokarcinóma mintáit használtuk fel, melyekhez agyi metasztázisból származó minta is tartozott [60]. A szövettani paraméterek meghatározása a kutatás során a fent leírtak alapján történt. Mivel a mennyiségek közti tényleges korrelációk megállapításához nem célszerű a mesterséges küszöbértékek bevezetése, ezért alapvetően az expressziós szintek fenti szemikvantitatív skálán meghatározott értékeit használtuk. Ugyanakkor mind korábbi tanulmányokban [44, 45], mind a klinikai gyakorlatban [71–74] sűrűn használják a különböző küszöbértékeket a páciensek kategorizálására. Például azokban az esetekben amikor a primer nem-kissejtes tüdőrák (NSCLC: non-small cell lung cancer) platina-bázisú kemoterápiás kezelése után a betegség előrehalad, az anti PD-L1/PD-1 terápiát másodlagos kezelésként azoknak a pácienseknek javasolják, akiknél a PD-L1 pozitivitás meghaladja az 1%-ot. Továbbá a pembrolizumab nevű PD-1 blokkoló gyógyszer használata elsődleges kezelésként is elfogadott olyan NSCLC-től szenvedő betegeknél, ahol a tumorsejtek PD-L1 expressziós szintje nagyobb 50%-nál. Különböző tanulmányok arra az eredményre jutottak, hogy az 5%-os küszöbbel definiált PD-L1-et expresszáló, illetve nem expresszáló tumorsejtekkel rendelkező betegek között azok profitáltak jobban a nivolumab nevű immunterápiás szer használatából, akiknek a tumorsejtjei PD-L1 pozitívak voltak. Így tehát megvizsgáltuk, hogy az ezen küszöbértékek alapján két csoportra osztva a betegeket, fennállnak-e a kérdéses korrelációk. Vagyis egy előzetes vizsgálat során kiválasztottuk azt a küszöbértéket (vagy annak hiányát), melynek használatával az adott korreláció a legszignifikánsabb volt a kérdéses paraméterek között és az így szűkített összehasonlítások listáján alkalmaztuk a Bonferroni-korrekciót a többszörös tesztelés kompenzálására. (Nem lenne indokolt a Bonferroni-korrekció használata az összes elvégzett vizsgálaton, hiszen ezek paraméter-páronként erősen függnek egymástól, mivel csak a küszöbérték konkrét megválasztásában különböznek.)

A vizsgálatok során nem találtunk szignifikáns korrelációt sem a daganatot infiltráló immunsejtek mennyisége, sem pedig a tumor felületén létrejövő immunsejtréteg megjelenésének tekintetében a primer tüdőtumor és az agyi metasztázis között. A két paraméter mennyisége között egyébként egy konkrét tumoron belül sem találtunk összefüggést. Ezzel szemben a tüdő daganatban és az agyi áttétben lévő tumorsejtek PD-L1 pozitivitása között erős pozitív korrelációt láttunk, küszöbérték használata nélkül és az összes standard küszöbértékkel is. Ez az összefüggés nem állt fenn sem az immunsejtek PD-1, sem pedig a

PD-L1 pozitivitására vonatkozóan. Ez arra enged következtetni, hogy bár az agyi áttét sejtjei rendkívül hasonlóak PD-L1 expresszió tekintetében a primer tumor sejtjeihez, a lokális immunkörnyezet jelentősen eltér a két szövetben. Ez azt jelenti, hogy ha kizárólagosan a tumorsejtek PD-L1 pozitivitása marad meg a klinikai gyakorlatban biomarkerként az immunterápia hatékonyságának predikciójára vonatkozóan, akkor a primer tumorból vett minta elemzése elegendő az agyi áttét kezeléséhez is. Ha azonban a tumorsejtek PD-L1 expressziós szintje mellett az immunsejtek lokális sűrűsége, illetve azok PD-1/PD-L1 expressziója is szerepet játszik majd a kezelés kiválasztásában, az agyi áttétek alapos vizsgálata szükséges a döntést megelőzően. Emellett minden fent vizsgált paraméterre meghatároztuk azoknak a pácienseknek a számát, akiknél az adott paraméter értéke a primer daganatról az áttétre nőtt/csökkent vagy nem változott. A betegek változás-szerinti eloszlásait összevetettük a különböző kezelést kapott páciensek csoportjaiban. Azt tapasztaltuk, hogy a változás irányának eloszlása egyik szövettani paraméter esetén sem függött szignifikánsan attól, hogy a betegek milyen terápiában részesültek a primer tumor diagnózisa és az agyi metasztázis műtétje között.

A fentiek alapján körvonalazódik a kép, miszerint az immunterápia bár rendkívül ígéretes eredményeket ad olyan esetekben, amikor a hagyományos gyógymódok mind kudarcot vallottak [57], mégis sok a tennivaló még, mire a használata ténylegesen elterjedhet és biztonsággal kiválasztható azoknak a pácienseknek a köre, akiknél valóban pozitív hatása lehet. Kutatásaink során arra törekedtünk, hogy a PD-1/PD-L1 tengely gátlását szolgáló immunterápiás szerek esetében a potenciálisan legjobban használható prediktív biomarkerek skáláját feltérképezzük. Ennek érdekében vizsgáltuk a tumor környéki immunsejtek (a daganatba beszivárgó stromális, illetve a daganat körüli immunsejtgyűrű) jelenlétének, az immunsejtek PD-1/PD-L1 pozitivitásának, illetve a tumorsejtek PD-L1 expressziós szintjének hatását a túlélésre vonatkozóan, ezek változásait különböző terápiás szerek hatására, illetve a primer tumorról a metasztázisra, továbbá korrelációikat az egyéb szövettani és klinikai paraméterekkel. Az elemzések során törekedtünk a klinikailag homogén betegcsoportok kiválasztására, hogy az eredményeket a lehető legkevésbé befolyásolják a kutatás hatáskörén kívül eső, nem kontrollált paraméterek.

# AZ ÚJ GENERÁCIÓS SZEKVENÁLÁSI TECHNOLÓGIÁK HÁTTERE ÉS A BENNÜK REJLŐ LEHETŐSÉGEK

## A DNS-SZEKVENÁLÁS RÖVID TÖRTÉNETE ÉS ÁLTALÁNOS CÉLJAI

A DNS-szekvenálás valójában alig több, mint 40 éves múltra tekinthet vissza [75], mégis ebben a néhány évtizedben gyökeresen átalakult mind a módszer technológiai háttere, mind pedig a potenciális felhasználási területek sora. Az 1950-60-as évek vívmányai lehetővé tették a fehérjékben az aminosavsorrend meghatározását, illetve az RNS nukleotidszekvenciájának megállapítását. Ez a forradalmi eredmény, az alanintranszfer RNS-ének 76 nukleotidos szekvenciájának azonosítása öt ember három éves munkájába került [76].

Végül 1976-ban kétféle DNS-szekvenálásra vonatkozó törekvés is sikerrel zárult. Mindkét módszer lényege az volt, hogy a DNS-t specifikus nukleotidoknál feltördelték, majd a keletkező darabok hosszának megállapításával a genomi pozíciók megfeleltethetőek voltak az adott nukleotidnak. Sanger és Coulson a meglévő DNS-szál mellé úgy szintetizált egy másikat, hogy a szokványos építőkövek és a szintézist végző enzim mellé alacsony koncentrációban fluoreszcensen megjelölt, egy specifikus nukleotidot tartalmazó, a szintézist blokkoló elemeket is hozzákevert [77]. Így azoknál a bázisoknál, melyek a blokkoló elemben jelenlévő nukleotid komplementerei voltak, időnként véletlenszerűen megszakadt a növekedés [78]. Maxam és Gilbert ezzel szemben egy kémiai módszert használt a DNS feldarabolásához [79]. A szekvenálni kívánt DNS-szál felszaporítása után az összes másolatot radioaktívan megjelölték, majd ezt követően olyan kémiai anyagokkal kezelték őket, melyek egy vagy kétféle specifikus bázist eltüntettek, végül pedig a DNS-t az abázikus helyek mentén feltördelték. Mindkét módszer esetén a kapott DNS-szakaszokat gélelektroforézissel [79] hossz szerint, egy bázisos felbontásban szétválogatták, az így kapott „létraszerű" képekből pedig a bázissorrend azonnal leolvasható volt. Néhány további újítás és automatizálás bevezetése után ezek a technikák széles körben elterjedtek, és az első generációs szekvenálási gépek egész sora jelent meg a piacon [80–84]. A szekvenált szakaszok („short read"-ek) hossza a kilobázisos (1 kb = $10^3$ bázis) nagyságrendbe esett, az ennél hosszabb genomok vizsgálata során pedig az ún. „shotgun" szekvenálási módszerrel, az egymással átfedő DNS-szakaszok szekvenálásával állították össze a teljes vizsgált szekvenciát. Ehhez nagy segítséget jelentett az ún. „paired-end sequencing" kifejlesztése, melynek során a vizsgált, fix hosszúságú DNS-szakasz két végéről kezdve olvasták le a bázisokat, így a keletkező short readek közti genomi távolság ismert volt, mely jelentősen könnyítette a genom rekonstruálását. Egy nagy genom rövid szekvenálási readekből történő összeillesztésére „de-novo assembly"-ként (de-novo illesztés) hivatkozunk. A nagyszabású, 1990-ben induló Humán Genom Projekt igáslova is ez a technológia lett, melynek során 2001-re megszületett a teljes emberi genom szekvenciájának első vázlata, majd 2004-re az első végleges verziója [3, 85].

Az 1980-90-es évek során folyamatos erőfeszítések irányultak a szekvenálási procedúra

gyorsabbá és olcsóbbá tételére. Bár a Humán Genom Projekt még nem profitált ezekből az újításokból, nagyjából egy évtizeddel a befejezése után az új- vagy másod-generációs szekvenálási technikák (NGS: next-generation sequencing) már messze túlszárnyalták az eredeti Sanger-féle módszert. Ezeknek a használata során a hagyományosan elemzett egyetlen DNS-szál helyett egyszerre szálak millióit lehet párhuzamosan szekvenálni [86], továbbá az utólagos, időigényes hossz szerinti szelektálást felváltotta az élőben zajló SBS (sequencing-by-synthesis; „szekvenálás szintézissel") eljárás, mikor a mintaként használt szálhoz egyesével hozzáépülő bázisokat azonnal a szintézis során leolvassák. Ezt részben a hídnövesztés (bridge amplification) módszere tette lehetővé, mellyel egy felületre ritkásan rögzített DNS-szálak közvetlen környezetében a szálak egzakt másolatait tartalmazó klaszterek növeszthetők. A szintézis során az egy klaszterbe tartozó kb. 1000 mintaszál mellé beépülő azonos bázis fluoreszcens fénye már detektálható jelet ad [86]. Ahhoz, hogy a szintézis ezek után folytatódhasson, fontos volt az olyan reverzibilisen blokkoló építőelemek kifejlesztése, melyeknek a szintézist gátló hatása visszafordítható volt. 2005-ben megjelent az első kereskedelmi forgalomban is kapható NGS készülék [75], majd hamarosan különböző cégek sokasága dobott a piacra hasonló eszközöket. 2007 és 2012 között az egy bázisra eső szekvenálási költség négy nagyságrenddel csökkent [4]. 2012 óta a versengés némileg leapadt, és az Illumina nevű cég gyakorlatilag monopol szerepet élvez [87], így a manapság legkönnyebben hozzáférhető szekvenálási adatok jelentős része valamilyen típusú Illumina platformon keletkezett. A szekvenált short readek hossza ugyan tipikusan csak néhány száz bázis, de a nagyon alacsony leolvasási hibaráta (kb. 0,1%) és a megfizethető költségek miatt a módszer töretlen népszerűségnek örvend.

Ugyanakkor, mivel az NGS technológiák majdnem mindegyike erősen támaszkodik a minta DNS-szálak amplifikációjára, mely gyakran szisztematikus, szekvencia-függő hibákhoz vezet, felmerült az igény a gyors, de sokszorosítás-mentes szekvenálási módszerek kifejlesztésére. Ezzel az ötlettel már az 1980-as évek óta aktívan foglalkoztak, de a törekvések nagy része kudarcba fulladt. Az utóbbi években azonban kétféle megoldás is ígéretes eredményeket hozott, ezek az ún. harmadik-generációs szekvenálási módszerek. A Pacific Biosciences (PacBio) nevű cég egy olyan módszert tökéletesített [88], melynél az SBS egyetlen DNS-szálon megy végbe, és a beépülő, egyetlen egy bázis fluoreszcens jelét úgy különítik el a szál körüli oldatban található többi építőelem zajától, hogy a reakciót egy zéró-módusú hullámvezető (ZMW: zero-mode waveguide) aljára korlátozzák, melynek a lineáris dimenziója kisebb, mint a megvilágító fény hullámhossza, így annak intenzitása exponenciálisan csökken a hullámvezető belseje felé [89]. Ezzel tehát a megfigyelt térfogat méretét a $10^{-21}$ liter nagyságrendre lehet szűkíteni. A PacBio szekvenálás előnye az amplifikációs lépés kiküszöbölése mellett, hogy rendkívül hosszú, 10 kb-t is meghaladó readek hozhatók létre vele, melyek a de-novo illesztést (egyszerűen kombinatorikai szempontból) nagyban megkönnyítik. Ezzel szemben a bázisok leolvasási hibája a 10%-os nagyságrendbe esik, de véletlen eloszlású, vagyis nem függ a szekvenciális környezettől.

A másik sikeresnek ígérkező sokszorosítás-mentes eljárás, a nanopore szekvenálás azt az elvet használja ki, hogy amennyiben egy DNS-szálat egy vékony ioncsatornán keresztülhúzunk, az áramló ionok áramában megfigyelhető mintázatok összefüggésben állnak az éppen keresztülhaladó nukleobázisok típusával. Az ágazat legfőbb képviselője az Oxford Nanopore Technologies (ONT) cég, mely 2014-re egy olyan szekvenáló eszközt fejlesztett (MinION) [90], ami méretre alig nagyobb egy hagyományos pendrive-nál, így korábban elképzelhetetlen előnyöket jelenthet a terepen történő mintavételezés és elemzés során. Bár a leolvasási pontosság egyelőre igen gyenge, a generált hosszú readek és a rendkívül könnyű hordozhatóság miatt nagy érdeklődés övezi a technika fejlődését.

Bár a harmadik-generációs szekvenálási módszerek nagyon reménytelinek tűnnek, a jelenleg elérhető adatok nagy része még főként NGS-alapú technikával keletkezett. Így a továbbiakban elsősorban az ilyen típusú szekvenálási adatok bioinformatikai elemzésére koncentrálunk.

### A szekvenálás céljai és szerepe a biomarker kutatásban

A kezdetekben, amikor a DNS-szekvenálás elsődleges célja a különböző genomok nagy skálás feltérképezése volt, a fő feladatot a repetitív szakaszokkal tarkított szekvenciák technikai hibáktól terhelt short readjeinek de-novo összeillesztése jelentette. A Humán Genom Projekt során a relatíve hosszú, paired-end readeket generáló, ámde költséges és lassú shotgun szekvenálást használták. Az NGS technológiák elterjedésével a szekvenált genomok száma jelentősen megugrott, de a rövid readeknek köszönhetően ezek minősége elmaradt a shotgun szekvenálás eredményeitől. A harmadik-generációs módszerek megjelenése azonban új lehetőséget jelent az így feltárt genomok finomhangolására és minőségük javítására [75].

A Humán Genom Projekt lezárásával a logikus következő lépcső a genomok újraszekvenálása és az emberek közti egyéni különbségek feltárása volt. A szekvenált short readeknek egy meglévő referenciagenomhoz történő újraillesztése merőben más és jóval egyszerűbb kihívás volt, mint a de-novo összeállítás. Az újonnan elterjedő számítógépes algoritmusoknak hála ez manapság rutinfeladatnak számít. A kellően redundánsan szekvenált genomok, vagyis ahol egyetlen genomi régióra átlagosan több short read is ráillik, lehetővé teszik az egyének közti eltérések, vagyis a mutációk feltérképezését. Az adott genomi pozícióra illesztett short readek számát lefedettségnek (coverage) nevezzük, a kb. 30-as átlaglefedettség már elegendő a mutációk megbízható detektálásához. Genomi mutáció természetesen bármilyen eltérés lehet a referenciaként használt genomhoz képest. Nagy skálán elképzelhető hosszú genomi régiók, akár teljes kromoszómák eltűnése (deléció) vagy többszöröződése (duplikáció), genomi szakaszok beillesztődése (inzerció) és cseréje (transzlokáció) különböző kromoszómák között, egy-egy régió irányának megváltozása (inverzió) és ezek tetszőleges kombinációja. Az ilyen jellegű nagy-skálás mutációk gyakran nem összeegyeztethetők az élettel, amennyiben pedig igen, tipikusan genetikai

betegségek okozói. Hasonlóan, kis-skálán, nukleobázisos felbontásban is beszélhetünk néhány bázis hiányáról (deléció), a referenciagenomban nem szereplő bázisok beszúródásáról (inzerció), vagy egy-egy bázis típusának megváltozásáról (pontmutáció; SNV: single nucleotide variation). A továbbiakban főként a kis-skálás mutációk elemzését tűzzük ki célul.

Alapvetően az egészséges egyének közti különbségeket (pl. szemszín, orr forma, stb.) kódoló genetikai eltérések is mutációként jelennek meg a genomok összehasonlítása során, ezeknek a vizsgálata azonban érthető okokból kevésbé hangsúlyos szerepet kapott, mint a betegségeket kódoló mutációk feltérképezése.

A DNS mutációját számos kiváltó tényező okozhatja, véletlen és célzott elváltozások egyaránt megjelenhetnek benne, melyek a teljes szervezet elpusztulását is okozhatják. A károsító hatások lehetnek külső, exogén tényezők (pl. UV-sugárzás, vírusok, stb.), de akár normális endogenetikus metabolikus folyamatok is, melyek naponta átlagosan ötszázezer molekuláris hibát generálnak a DNS szálon. Természetesen ez az érték az összesen kb. 3 milliárd bázisból álló láncon csekélynek tűnik (kb. 0,00017%), de ha a működéshez alapvetően szükséges gének károsodnak, úgy egyetlen bázison esett hiba is végzetes lehet. Továbbá, mivel a kettős hélix szerkezet megköveteli a két DNS szál egymáshoz való kapcsolódását, az egyik szálon lévő báziscsere a két lánc kettéválását is eredményezheti. Az esetek jelentős részében a károsult vagy mutálódott DNS többszöröződése éppen a megjelenő hiba miatt akadályoztatva van, így a hibás DNS-sel rendelkező sejtek hamar elpusztulnak, számuk elenyésző lesz. Néhány mutáció esetén azonban előfordulhat, hogy a módosult DNS-ű sejtek éppen a mutáció által valamiféle evolúciós előnyre szert téve gyorsabban szaporodnak, mint egészséges társaik, ezzel az egész szövetet veszélyeztetve. Különösen a gyorsan osztódó sejtek esetében jelent ez komoly problémát: a mutált sejtek korlátlan elszaporodása rákot okozhat. Fontos megjegyezni, hogy a DNS-t érintő hibák javarészt természetesek és a szervezetben működő számos javítási mechanizmusnak hála többségében korrigálódnak is. A valós problémát tehát nem közvetlenül a DNS meghibásodása okozza, hanem a javító folyamatok nem rendeltetésszerű lefolyása.

A DNS-javító mechanizmusok az észlelt hibától függően rendkívül sokszínűek lehetnek, az alábbiakban a teljesség igénye nélkül tekintünk át néhányat közülük. A legegyszerűbb esetekben a károsodás direkt módon visszafordítható, amennyiben teljesen egyértelmű, hogy a módosult DNS milyen eredeti alakból keletkezhetett. Ez a helyzet például az egy szálon kialakuló pirimidin dimerek esetén, melyek javítását a fotoliáz nevű enzim katalizálja, éppen az elnyelt UV fény aktivációjának hatására. Hasonlóképpen például a guanin metilációját a metilguanin metiltranszferáz (MGMT) enzim tudja visszafordítani. Ha a kettős hélixnek csak az egyik szálán sérül nukleotid, a hiba viszonylag egyszerűen javítható, hiszen a másik szál felhasználható mintaként. A korrigálás tipikusan két lépésből tevődik össze: elsőként a mutálódott nukleotid eltávolításra kerül, majd az üres helyre a szemközti szál megfelelő pozíciójának komplementer bázisát behelyezi egy enzim. Többféle

mechanizmus alkalmas ilyen jellegű javításra: a BER (base excision repair) elsősorban az oxidálódott, alkilizálódott vagy hidrolizálódott nukleotidokat cseréli ki, a NER (nucleotide excision repair) a hélixszerkezetet torzító mutációkat ismeri fel és javítja ki, az MMR (mismatch repair) pedig a nem összetartozó (tehát nem AT vagy CG) bázispárokat korrigálja. Lehetséges károsodási forma, amikor a DNS mindkét szála elszakad. Az ilyen hibák javítására két fő mechanizmus alkalmas: az NHEJ (non-homologous end joining), ami a DNS szál két végét gyakorlatilag közvetlenül „összeragasztja". A mechanizmus egyik altípusa (MMEJ: microhomology-mediated end joining) során valójában a javítást végző enzim keres egy rövid átfedő szakaszt a két szálon és amennyiben talál ilyet, a láncokat csatlakoztatja egymáshoz. Mivel azonban ekkor egyes bázisok elveszhetnek a láncvégekről, ez a mechanizmus óhatatlanul deléciókat eredményez. Fontos előnye azonban, hogy a javítás nem igényel egy érintetlen, mintának alkalmasan használható DNS részletet. Amennyiben mégis lehetőség van minta használatára (például a homológ pár másik tagja rendelkezésre áll, vagy a mitózis során a kromoszómakettőződés utáni fázisban a testvérkromoszóma használható e célra), a rekombinációs (HR: homológ rekombináció) eljárással sokkal pontosabb javításra van lehetőség, ekkor ugyanis a mintaként használt szakasz mintegy átmásolódik a sérült kromoszómára. Amennyiben nem sikerül a hibát kijavítani, a probléma kezelésére a sejt számos további opciót alkalmazhat. A hagyományos DNS-polimeráz enzimek (melyek a DNS másolásánál az új DNS összerakását végzik) nem képesek a hibás DNS-t mintaként használva újat legyártani, a károsodást elérve megakadnak és a szálon nem tudnak továbbhaladni. Lehetséges azonban ezeknek az enzimeknek a kicserélése olyanokra, melyek bizonyos toleranciával rendelkeznek a károsodásokkal szemben, ezzel lehetővé téve még a sérült DNS replikációját is. (Természetesen így az újonnan legyártott szálon is megmarad az eredetin észlelt hiba.) Az ilyen jellegű mechanizmusok gyakran hajlamosak pontmutációk beillesztésére, mivel azonban a DNS sokszorosítása feltétlenül szükséges a sejthalál elkerüléséhez, többnyire az ilyen „tévesztések" preferáltabbak, mint a másolás teljes megakasztása.

A DNS-javító mechanizmusok mindegyike rendkívül komplex, számos enzim és egyéb fehérje interakcióját teszi szükségessé. A továbbiakban a részletek pontos leírásától eltekintünk és csupán két fehérje DNS-javító szerepét emeljük ki illusztratív példaként. A PARP1 enzim és közvetve az azt kódoló PAPR1 gén elsősorban az egy szálas DNS-törések javításában játszik jelentős szerepet. Amennyiben a PARP1 valamilyen okból hiányzik a sejtből, és emiatt az egyszálú töréseket nem sikerül időben kijavítani, azok kettős szálú töréssé alakulnak, melyeket a HR gyakorlatilag hiba nélkül képes korrigálni, így egy nem HR-hiányos sejtben ez nem okoz valódi problémát [91]. A BRCA1 nevű fehérje (és az azt kódoló azonos nevű gén) más fehérjékkel együtt a kettősszálú törések javításáért felel. A BRCA1 fehérje a HR-t végző fehérjekomplexnek a tagja, ami megfelelő minta mellett a kettétört DNS-t hiba nélkül képes eredeti formájában reprodukálni. Emellett részt vesz a MMR folyamatokban is, ezzel jelentősen hozzájárulva a genom stabilizálásához.

A BRCA1 gén mutációja az emlő-, petefészek- és néha a prosztatadaganattal diagnosztizált betegek számottevő részénél megfigyelhető. Ebből tehát arra lehet következtetni, hogy a gén és az általa kódolt fehérje helyes működése elengedhetetlen a szervezet normális működésének fenntartásához. Ugyanakkor a homológ rekombináció hibájára utaló egyértelmű jel komoly fegyvertény a daganat gyógyításának tekintetében. Ha a rákos sejtekben ismerten rosszul működik a kettős száltörések javítása, érdemes lehet az ilyen jellegű hibákat célzottan előidézni. A fentiek alapján ezt legegyszerűbben az egyszálas törések javítatlanul hagyásával lehet elérni, vagyis a PARP1 blokkolásával, ami az egészséges sejtekben nem tesz komolyabb kárt, a HR-hiányos tumorsejtekben azonban végzetes hatása van. A PARP-inhibitorokat többféle daganattípus esetén sikerrel alkalmazzák [92]. A BRCA1 gén mutációja mellett a homológ rekombinációs mechanizmus hiányára utalhatnak a BRCA2 és PALB2 fehérjéket kódoló gének elváltozásai is. Vagyis egy daganatban a konkrét mutációk feltérképezése fontos információt jelenthet arra vonatkozóan, hogy a páciens várhatóan milyen kezelésre reagál majd pozitívan. Így egyrészt a potenciálisan nem működő kezelésekkel nem kell feleslegesen terhelni a beteg szervezetét, illetve a személyre szabott terápiákkal a kezelés költségei is csökkenthetők.

Ahhoz azonban, hogy a DNS mutációit, mint biomarkereket alkalmazhassuk a gyógyászatban, elsőként a megbízható mutáció detektálás lépéseit és a felmerülő bioinformatikai nehézségeket kell áttekintenünk.

### AZ ADATFELDOLGOZÁSI FOLYAMAT TIPIKUS LÉPÉSEI

A fentiek szerint az SBS során a különböző színű klaszterekről készített sorozatos képek kiértékelésével kapható meg a readenkénti bázissorrend. A gyakorlatban valójában négy szürkeárnyalatos kép készül, melyekről a klaszterek helyén a színintenzitás meghatározása útján dönthető el, hogy milyen típusú bázis épült be a növekvő DNS-szálakba az adott ciklusban. Ez bár koncepcionálisan nem tűnik bonyolult feladatnak, a gyakorlatban nem triviális a $mm^2$-enként nagyságrendileg egymillió klaszter végigkövetése. A leolvasás bizonytalanságát az ún. base quality értéke jelöli, melyet minden leolvasott bázishoz hozzárendelnek a szekvenálás során. Végül az NGS módszerek által meghatározott short read szekvenciák jellemzően szöveges formátumban válnak elérhetővé a további elemzések számára. Rendszerint ezek a fájlok az ún. FASTQ formátumot követik, ahol az egyes readek bázissorrendje egykarakteres kódolásban (tehát A/C/G/T), egy egysoros leírást követően jelenik meg, a base quality értékek pedig ez alatt, az ún. Phred-skálán, ASCII karakterekkel kódolva [93] találhatóak. A Phred-skálán mért $Q$ base quality egyszerű transzformáltja annak a $P_{\text{error}}$ valószínűségnek, hogy az adott bázist helytelenül azonosították a szekvenálás során: $Q = -10 \cdot \log_{10}(P_{\text{error}})$.

A nyers szekvenálási adatok további elemzése alapvetően kétféleképpen történhet. Egy ismeretlen genom feltérképezésénél a short readek de-novo illesztésével a teljes genom szekvenciája megkapható. Újraszekvenált genomok esetében a readeket elég a már meglé-

vő referenciagenomhoz illeszteni, ami egy jóval egyszerűbb feladat. Erre az utóbbi időben legelterjedtebben alkalmazott szoftveres megoldást a BWA [94] nevű eszköz adja. Az algoritmikus részletektől eltekintve a program minden szekvenált short readről eldönti, hogy melyik genomi szakaszra illeszkedik a legjobban, illetve, hogy mennyire tartja megbízhatónak az adott illesztést („mapping quality"). Az illesztés jóságát számos faktor befolyásolhatja, például a repetitív szakaszok a referenciagenomban, a read bázisainak minősége, az illesztési algoritmus érzékenysége vagy a paired end szekvenálás. Az eredményeket tipikusan ún. BAM fájlokban tárolják el az illesztéshez használt szoftverek. Ez a szokványosan felmerülő memórialimitációk miatt egy bináris formátum, melynek a szöveges, tartalmilag megegyező párja a SAM. Mindkét formátum manipulálására a legnépszerűbb programcsomag a samtools [95]. A BAM/SAM fájlok a readekről eredetileg ismert információk (azonosító, bázissorrend, base quality) mellett az illesztésre vonatkozó adatokat is tartalmazzák (referenciagenom, illesztés genomi koordinátája, mapping quality, esetleges másodlagos illeszkedés, eltérések a referenciagenomtól, stb.). A legtöbb mutációkat azonosító szoftver minden elemzett minta esetén egy-egy BAM fájlból indul ki.

Rendszeresen előfordul, hogy egy SAM fájl böngészése egyszerűen nem elég áttekinthető, ha egy pillantást vetnénk a nyers szekvenálási adatokra. Ilyenkor érdemes azt a samtools mpileup parancsával az ún. pileup formátumba [95] konvertálni, mely nem readenként csoportosítva, hanem a referenciagenom mentén lineárisan jeleníti meg az adott genomi pozícióra illeszkedő összes readből származó adatot. A pileup fájlból első ránézésre megállapítható, hogy egy adott pozícióra hány read illeszkedett (lefedettség), milyen ezeknek az irányultsága, továbbá, hogy illett-e a referenciával nem megegyező bázis a kérdéses helyre.

### Szisztematikus és véletlen hibák megjelenése, különböző elemzési szempontok

A különböző szekvenálási technológiák más és más szisztematikus és véletlen hibák megjelenését okozzák a nyers bázissorrendben. A tévesztések jelenléte és milyensége gyakran erősen függ a szekvenciális környezettől [96, 97], vagyis nem véletlenszerűen jelennek meg a short readekben. Ez a tévesen leolvasott bázisok akkumulálódását okozhatja egy-egy genomi pozícióban, amiből hibásan egy valójában biológiailag nem létező mutáció jelenlétére következtethetünk. Az ilyen esetekben érdemes kihasználni, hogy várhatóan az ilyen jellegű hibák minden egyszerre elemzett mintát érintenek, így az olyan genomi pozíciók, melyek a különböző mintákban gyakran „zajosak", nem lesznek megbízhatóak.

Emellett a base quality-k meghatározásához a szekvenátorok forgalmazói üzleti titoknak számító algoritmusokat használnak, melyek ugyancsak sűrűn szisztematikus hibákat vétenek. Bár ha egy bázis tévesen lett leolvasva, a ténylegesen ott lévő bázis utólagos meghatározása lehetetlen, mégis fontos információt jelent a base quality pontos értéke.

Amennyiben a base quality alapján jó okunk van feltételezni, hogy a kérdéses bázis hibás, a mutációk detektálásakor azt figyelmen kívül hagyhatjuk. Ezért fontos a bázisok minőségének lehető legprecízebb ismerete. Ezt a célt szolgálja a GATK programcsomag [98] BQSR (base quality score recalibration) [99] lépése, mely a rendelkezésre álló szekvenálási adatokat végigszkennelve feltérképezi, hogy milyen típusú genomi helyeken jelennek meg a base quality-kat érintő szisztematikus hibák. Az eszköz a szekvenciális kontextusra specifikus és a readbeli pozícióra jellemző statisztikákat vizsgálja (a readek végén jellemzően nagyobb a bázisleolvasás bizonytalansága), azzal a feltételezéssel, hogy egy előre definiált mutációs pozíciólistára illő bázisoktól eltekintve minden nem referencia bázis hibás. A kapott hibamodell segítségével pedig egy következő fázisban az összes mért base quality értéket korrigálja a megfigyelések alapján.

A mutációk azonosításakor a base quality-kből származó információk mellett érdemes továbbá kihasználni az illesztés jóságát jellemző mapping quality értékeket is. Egy olyan readben található nem-referencia bázisnak például nem feltétlenül szükséges nagy jelentőséget tulajdonítani, amelyről eleve nem tudjuk biztosan, hogy a megfelelő genomi régióra lett felillesztve. Mivel a referenciagenom alacsony komplexitású (ismétlődő) szakaszai (melyek általában GC-gazdagok) nagyban megnehezítik az illesztést, gyakran alkalmaznak az elemzések során ún. repeat maszkolást, melynek során az ilyen szakaszok bázisait az univerzális N-re cserélik, így ezeken a régiókon alapértelmezetten nem detektálnak mutációkat. Ez az illesztési hibák kiküszöbölése mellett azért is praktikus, mert az amplifikációs eljárás (PCR: polymerase chain reaction) hatékonysága is erősen függ a szekvenciák GC-tartalmától [100].

Gyakran praktikus továbbá az illesztést követően mesterségesen megszabadulni azoktól a readektől, amik tökéletesen megegyeznek egymással. Elvileg a szekvenálás során a DNS-t véletlenszerűen tördelik fel, így viszonylag kicsi annak az esélye, hogy két read tökéletesen azonos legyen. Ha ilyen mégis előfordul, az rendszerint valójában valamilyen szekvenálási műtermék, például az amplifikáció során egy adott DNS-szakasz aránytalan mennyiségben felszaporodott. Mivel ezek a readek nem tekinthetőek független megfigyeléseknek, érdemes csak egyet megtartani közülük. Ennek az utólagos szűrésnek az elvégzésére mind a samtools, mind a szintén népszerű GATK [101] programcsomag lehetőséget kínál. A további elemzések szempontjából szintén fontos a readek genomi koordináta szerinti sorba rendezése, ami hasonlóan a fenti két eszköz bármelyikével megvalósítható.

Magának a mutáció-detektálási módszernek a kiválasztása nagyban függ a kutatás konkrét céljaitól. Egyrészt megkülönböztetünk csíravonal (germline) és szomatikus mutációkat, melyek azonosítása során más-más elveket kell szem előtt tartanunk. Csíravonal mutációnak definíció szerint a csírasejtekben megtalálható DNS-variánsokat nevezzük, tehát kizárólag azokat, melyek továbböröklhetőek az utódra, akinek később minden sejtjében megjelennek [102]. Ezzel szemben a szomatikus mutációk a testi (vagy szomatikus) sejtek valamelyikében jönnek létre és annak utódsejtjeiben detektálhatók. Ezek az általános

megfogalmazások a rákkutatás területén némileg módosulnak: hagyományosan csíravonal mutációnak tekintjük azokat a genomi változásokat, melyek az adott páciens normál és tumoros szövetében is megtalálhatóak, szomatikusnak pedig azokat, melyek csak a tumorszövetben vannak jelen. A szomatikus variánsok detektálását általában erősen hátráltatja, hogy a klinikai biopsziás mintáknál gyakran előfordul, hogy a normál és a tumoros szövet keveredik, illetve egy tiszta tumor minta esetén sem ritka a daganatok ismert heterogenitása miatt, hogy különböző mutációkat felhalmozó sejtpopulációk egyvelegét kell elemeznünk. Emellett a tumorsejteknél sűrűn tapasztalhatunk kópiaszám változásokat (CNV: copy number variation) és aneuploiditást is, mivel a genomjuk jellemzően igen instabil. Ezek a tényezők erősen befolyásolják az adott genomi pozícióban mérhető allélfrekvenciák várható értékét. A germline variánsok esetében ezzel szemben markánsabb, több mintában is megbízhatóan azonosítható jelet keresünk.

Fontos továbbá minden tudományos kérdésfeltevés során eldöntenünk, hogy a detektáló módszer érzékenységét vagy specificitását szeretnénk maximalizálni. Vagyis melyik a preferáltabb számunkra: ha minden esetlegesen szóba jövő variánst azonosítunk, elfogadva, hogy ezzel a mutációk egy része nem bír valódi biológiai jelentőséggel (fals pozitív), vagy kizárólag a nagyon megbízható variánsokat találjuk meg, vállalva, hogy néhány valódi mutációt elveszítünk (fals negatív). Minden mutációt detektáló algoritmus az érzékenység és a specificitás együttes maximalizálására (vagyis a fals pozitív és fals negatív találatok együttes minimalizálására) törekszik, világos azonban, hogy végső soron valamilyen kompromisszumot kell kötnünk. A rákkutatás során hagyományosan inkább a specificitás maximalizálása mellett döntenek a szomatikus mutációk esetében.

Erősen függ a használt módszer hatékonysága a vizsgált minták típusától is. Heterogén klinikai minták esetén azt várjuk, hogy az alacsony allélfrekvenciával megjelenő szomatikus variánsokat is azonosítani tudjuk, ezzel szemben ha egyetlen sejt felszaporításából eredő, homogén populációkból mintavételezett DNS-t elemzünk, az alacsony allélfrekvenciájú pozíciókat egyszerűen zajnak tekinthetjük. Tehát a minták típusa merőben más problémák elé állítja a detektálást végző szoftvert.

### Ízelítő a mutáció-detektáló algoritmusok sorából

Ebben a fejezetben két, gyakran használt szomatikus mutációk detektálására (is) alkalmas szoftvert tekintünk át nagyvonalakban. Algoritmikus komplexitás szempontjából a skála két végén helyezkednek el, mégis mindkét eszköz viszonylag nagy népszerűségnek örvend.

A viszonylag egyszerű módszert alkalmazó VarScan 2 [103] páronként hasonlítja össze a tumor és a normál szövetből származó mintákat. Bemenetként pileup formátumú fájlokat vár, majd pozíciónként végigpásztázza a genomot. Mintánként egy Fisher-féle egzakt teszt segítségével eldönti, hogy az adott pozícióban található-e variáns a normál és a tumor mintában külön-külön. (A Fisher-féle egzakt teszt azt a nullhipotézist használja, hogy a

lokálisan leolvasott különféle bázisok számának eloszlása pusztán a szekvenálási hibákból származik és nem valódi variánsból.) A mintánkénti döntést bizonyos előre definiált szűrési feltételek segítik (pl. minimális lefedettség, minimális base quality, maximális Fisher-féle p-érték, stb.). Azokat a pozíciókat, melyeknél a tumor mintában variánst detektált az algoritmus, újra megvizsgálja egy Fisher-féle egzakt teszttel, most azzal a nullhipotézissel, hogy a normál és a tumor mintában a különféle bázisok valódi eloszlása megegyezik. Amennyiben a kapott p-érték kellően alacsony, az adott pozíciót szomatikus mutációnak tekinti. A fals pozitív találatok számának csökkentése érdekében emellett néhány empirikus szűrési feltételt is bevezetnek (pl. a variánst lefedő readek melyik szakaszán jelenik meg a mutáció, a variánst támogató readek irányultságának eloszlása, a referencia és a nem-referencia bázist támogató readek mapping quality-je közti különbség, stb.).

A manapság talán legelterjedtebben alkalmazott MuTect2 a GATK programcsomag [101] része és a VarScan 2 viszonylag naiv megközelítéséhez képest merőben más módszereket alkalmaz. Bemenetként egy normál és egy tumor minta BAM fájljait várja, de erősen javasolják, hogy emellett a felhasználó biztosítsa egy „normál panel” és egy csíravonal mutációkat tartalmazó adattábla hozzáférhetőségét. A normál panel normál minták sokaságából gyűjtött szekvenálási információk összességét tartalmazza. Ezt érdemes mindig olyan normál mintákból előállítani, melyek ugyanazokon a preparálási lépéseken (mintavételezés, szekvenálási protokoll, szekvenátor, előzetes szűrések) mentek keresztül, mint a vizsgálni kívánt mintapár, hiszen a szoftver ezt a forrást használja az esetleges szisztematikus szekvenálási és illesztési hibák kiszűrésére. A csíravonal mutációkat tartalmazó adattáblák [104] humán minták esetén online egyszerűen elérhetőek. Ezek a folyamatosan frissülő adatbázisok arra vonatkozó információkat tartalmaznak, hogy az egyes csíravonal variánsok az emberi populációban mennyire gyakran fordulnak elő. Ezt a MuTect2 elsősorban arra használja, hogy ha egy szomatikusnak tűnő variáns éppen egy helyre esik egy sűrűn előforduló csíravonal mutációval, akkor az nagy valószínűséggel ténylegesen csíravonal mutáció lehet.

Első lépésként a MuTect2 a vizsgált tumor minta szekvenálási adatainak végigszkennelésével ún. „aktív régiókat” azonosít a genom mentén [105]. Ezek olyan pozíciók környékén lévő szakaszok, melyekben a tumor mintában (viszonylag megengedő definícióval) „kellően megbízható” bizonyíték található egy variáns jelenlétére. Ezek után minden aktív régió esetén az algoritmus lokálisan újrailleszti az arra a szakaszra eredetileg illeszkedő short readeket és meghatározza a szekvenálási adatok által támogatott potenciális haplotípusok listáját. Ehhez elsőként létrehozza az adott régió referenciaszekvenciájának k-merjeiből ($k = \{10;25\}$) kapható irányított $B(4,k)$ De Bruijn-gráfot [106], melyben az élek súlya ebben a fázisban azonosan nulla. Ezek után a short readeket egyesével leképezi a kapott gráf útjaira és az érintett élek súlyát növeli eggyel, ha pedig korábban nem volt megfelelő él a gráfban, létrehozza azt és a súlyát 1-re állítja. Amennyiben a readben talál olyan k-mert, ami korábban nem volt a gráf része, új csúcsot ad a gráfhoz. Ahogy ezt a

procedúrát az összes short readdel elvégzi, a gráf bizonyos útjai egyre nagyobb súllyal rajzolódnak ki. Ezt követően zajszűrési célokból a kapott gráfból eltávolítja azokat a részleteket, melyeket csak néhány read adatai támogatnak. Végül a gráf összes útjának (vagyis a lehetséges haplotípusoknak) a végigjárásával kiszámolja azok likelihoodját az út által érintett élek átmeneti valószínűségeinek szorzataként. (Átmeneti valószínűségnek az adott élt támogató readek számának és az összes olyan élt támogató readek darabszámának hányadosát tekinti, melyek azonos csúcspontból indulnak ki, mint a vizsgált él.) Legutolsó lépésként kiválasztja a legmagasabb likelihoodal bíró 128 potenciális haplotípust, majd ezeket a hagyományos Smith-Waterman algoritmus [107] segítségével visszailleszti a régió referenciagenomjára, ezzel azonosítva az esetlegesen variánsokat tartalmazó pozíciókat [108].

A lehetséges haplotípusok körének leszűkítése után az összes tényleges short readet a kapott haplotípusokra illeszti a PairHMM [109] algoritmussal, ami képes a haplotípus-read párokhoz egy-egy likelihood értéket rendelni, ami azt mondja meg, hogy a vizsgált haplotípus mellett mekkora annak a valószínűsége, hogy tényleg az adott readet szekvenáljuk, figyelembe véve az adatok minőségére vonatkozó információkat (base quality). Ezek után a korábban meghatározott variánsok listáján végighaladva, minden variáns minden alléljére kiszámolja, hogy az adott allélt egy adott read mennyiben támogat. Ehhez az adott allélt tartalmazó haplotípusok közül kiválasztja azt, melyre az adott readdel vett likelihood a legmagasabb volt és ezt az értéket tekinti az allél-read likelihoodnak [110]. Ezekből az értékekből egy variációs Bayes modell segítségével az adott pozícióban minden megfigyelt allélre kiszámolja, hogy hányszor valószínűbb, hogy az adott allél ténylegesen létezik biológiailag, mint hogy nem. Az így felcímkézett mutációkat ezek után részben küszöbértékszerinti, részben pedig valószínűségi alapú szűrésekkel tovább kategorizálja, hogy a lehető legtöbb fals pozitív találattól megszabaduljon. Ezek során figyelembe veszi a variáns normál panelben való jelenlétét, a megfigyelt allélfrekvenciákat, a variánst támogató readek mapping quality-jét, irányultságát, a variáns read-beli pozícióját, az esetleges csíravonal státusz valószínűségét, a kontamináció hatásait és sok egyéb szempontot.

Láthatóan a MuTect2 rendkívül szofisztikált eszköztárral közelíti meg a mutációk detektálásának problémáját, ennek azonban megvan az ára. Még az erősen optimalizált algoritmusok ellenére is teljes genomok vizsgálata esetén gyakran rendkívül hosszú számítási időre kell felkészülnünk. Elődje, a 2013-as MuTect [111] hasonló, de koncepcionálisan kevésbé bonyolult módszerekkel határozza meg a szomatikus SNV-k listáját, inzerciókat és deléciókat (indeleket) viszont nem detektál. Futási idő tekintetében azonban négyszer gyorsabb utódjánál, habár még így is négyszer lassabb a sokkal egyszerűbb elveket követő VarScan 2-nél (2. táblázat). Mindkét szoftver alapvetően a klinikai, vagyis a szennyezett és heterogén minták nagy kapacitású számítógéppel történő elemzésére lett kifejlesztve, mely esetekben valóban megbízható eredményeket adnak. Az olyan kutatásoknál azonban, ahol lehetőség nyílik tiszta, homogén minták vizsgálatára, ellenben a számítási keretek szűkösek, érdemes más szempontok szerint optimalizált módszert választani.

### MUTÁCIÓS SPEKTRUMOK VIZSGÁLATA

Hagyományosan a rákkutatás genomikai ágazatai az ún. szomatikus driver mutációk azonosítására és ezek biomarkerként való alkalmazására törekednek. Driver mutációk azok a genomi elváltozások, melyek közvetlenül lehetővé teszik az érintett klón korlátlan felszaporodását. Egy olyan génben jelenlévő variáns, mely ismerten valamilyen DNS-javító vagy replikációs folyamatban játszik fontos szerepet, jelentősen megváltoztathatja az adott sejt és utódsejtjeinek szaporodási menetét. A BRCA1 és BRCA2 gének mutációi például tipikusan növelik bizonyos típusú daganatok kialakulásának kockázatát. Ugyanakkor éppen ezek a mutációk lehetővé teszik az érintett páciensek PARP-inhibitor terápiával történő kezelését. Így nem meglepő, hogy a kutatások számottevő része az ilyen nagy horderővel bíró genomi szakaszok azonosításával foglalkozik.

Ezzel szemben újabban a szomatikus mutációk statisztikus tulajdonságainak vizsgálata is elterjedni látszik. Erre az egyik leggyakrabban használt módszer az ún. mutációs spektrumok felrajzolása [112]. Spektrum alatt a szomatikus SNV-k számának eloszlását tekintjük a genomi környezettől függően. Konkrétabban a spektrumban az SNV-k bázisváltás (milyen bázisról milyenre változik) és az $r = 1$ sugarú genomi környezet (az SNV-t közvetlenül megelőző, illetve követő pozíció a genomban) alapján csoportosítódnak; így 96-féle kategóriát („mutációs triplettet") különböztetünk meg. ($4 \cdot 3 = 12$-féle bázisváltás, $4 \cdot 4 = 16$-féle környezet, melyek azonban a két DNS-szálra vonatkozóan páronként ekvivalensek, pl. A(C>A)G $\sim$ C(G>T)T.)

A mutációs spektrumok vizsgálatának alapvető célja, hogy azonosítsa az adott mintában operáló mutációs folyamatokat, melyekhez tartozó mutációs „szignatúrák" (konszenzus vektorok, melyek valójában egy diszkrét valószínűségi eloszlást írnak le) lineárkombinációjából alakul ki a konkrét mintában megfigyelt eloszlás. A feltételezés szerint a különböző daganattípusokban ugyanazok a mutációs folyamatok lenyomatai jelennek meg, de az adott betegségre specifikusan jellemző súlyfaktorokkal.

A mutációs folyamatok jellemző mintázatainak, a konszenzus szignatúráknak a meghatározásához a [112] tanulmány szerint minél többféle daganattípusból minél több páciens szomatikus mutációinak együttes vizsgálatára van szükség.

Az összesen $N$ db. mutációs folyamat mindegyikéhez rendelt $P_n = [p_n^1, p_n^2, ...p_n^K], 1 \leq n \leq N$ szignatúra (SNV-k esetén a fentiek alapján $K = 96$) egy adott $g$ páciens mintájában $e_g^n$ súllyal szerepel. Az ún. expozíciós vektor tehát egy adott páciensre $E_g = [e_g^1, e_g^2, ...e_g^n, ...e_g^N]$. A konkrét betegségtípusba sorolt $G$ db. mintában megfigyelhető triplettstatisztikát egy ún. $M$ „mutációs katalógusba" rendezve egy $K \times G$ méretű mátrixot kapunk, melynek $m_k^g$ elemei az $g$. páciensben, a $k$ típusú mutációs kategóriában megfigyelhető mutációk száma. Ez a mátrix a fenti feltételezések szerint előáll az $m_k^g \approx \sum_{n=1}^{N} p_n^k e_g^n$, vagyis $M \approx P \times E$ összefüggéssel. Mivel a három mátrix közül csak $M$-et ismerjük, $P$ és $E$ meghatározásához közelítő módszerek alkalmazására van szükség. A dekompozíciót a [112] kutatás a

nem-negatív mátrix faktorizációs (nnmf) eljárással végzi, melynek alapvető előnye, hogy az inherensen nem-negatív adatokból (mutációk darabszámai) úgy generálja le az azokat „leginkább jellemző" vektorokat, hogy azok elemei szintén nem-negatívok lesznek. Vagyis a kapott $P$ szignatúrák egyszerűen értelmezhetők és nem kell a „negatív mutáció" biológiai interpretációjával vesződni. Maga az algoritmus egy iteratív eljárást követ, melynek során alapértelmezetten az $M$ mátrix és a $P \times E$ szorzatmátrix különbségének Frobenius-normáját minimalizálja adott $k$ darab szignatúra mellett. Ezt követően [112] a betegségenként egyedien azonosított mutációs szignatúrákat egy felügyelet nélküli klaszterezési eljárással kondenzálja, vagyis az egymáshoz hasonló szignatúrákat helyettesíti az adott klaszter centroidjával, ezzel csökkentve a számukat. Az azonosított konszenzus szignatúrák listája online elérhető, és számos kutatás használja referenciaként. Bizonyos mutációs folyamatok esetén a szignatúra biológiai háttere is ismert, néhány esetben azonban nem sikerült ilyen összefüggést találni.

Nagyon fontos kihangsúlyozni, hogy attól függően, hogy milyen normalizálási eljárással preparáljuk az $M$ mátrix elemeit, jelentősen eltérő eredmények adódnak a nem-negatív mátrix faktorizáció után.

Természetesen lehetséges a normálás mellőzésével a nyers mutációs darabszámok használata is, ebben az esetben azonban fennáll annak a veszélye, hogy a spektrumok által kifeszített 96-dimenziós térben dominánsan az tesz különbséget a vektorok között, hogy mekkora az abszolút értékük. Az olyan daganattípusok spektrumai, melyek alapvetően kevés mutációt eredményeznek, az origó körül fognak csoportosulni, függetlenül a komponenseik konkrét értékeitől. Ezzel szemben a sok mutációval járó betegségekből származó minták már csak a vektorok normájának nagysága okán is elkülönülnek ettől a csoporttól.

Alternatívaként felmerül a spektrumok $\sum_{k=1}^{K} m_k^g = 1 \,\forall\, g$ normálása. Ez ugyan a fenti problémát kiküszöböli, de a biológiai interpretáció szempontjából megkérdőjelezhető, hiszen nem feltétlenül jogos mesterségesen összeskálázni a sok és a kevés mutációval járó eseteket. Elképzelhető, hogy bizonyos mutációs folyamatok biológiai hátterükből adódóan abszolút értékben jóval több mutációt eredményeznek, mint mások. Ekkor a fenti normálással a folyamat az egyik legalapvetőbb jellemzőjét veszíti el.

Különböző kísérleti elrendezésekben gyakran előfordul, hogy a genomi változások mutációs spektrumokra gyakorolt hatását nem humán genomokon vizsgálják. Ilyen esetekben is alapvető célkitűzés a mutációs folyamatok tettenérése a mintákban, azonban a kapott eredmények összevetése a referencia szignatúrákkal nem feltétlenül kézenfekvő. A megfigyelt mutációs spektrumok sajátosságait erősen befolyásolja a lokális genomi környezet, ezáltal pedig a genomban található összes típusú környezet gyakoriságának eloszlása. Például elképzelhető egy olyan genom, melyben az ACG triplett kiugróan gyakran megjelenik. Ilyenkor a spektrumon esetlegesen látható A(C>N)G csúcsok nem feltétlenül egy nagyon specifikus mutációs folyamatra utalnak, hanem pusztán arra a tényre, hogy egy ilyen genomon statisztikusan az ilyen típusú mutációk valószínűbbek. Szintén gyakori a

kísérletek során a teljes genom helyett csak az exom vizsgálata, ami ugyanilyen jellegű nehézségeket eredményezhet. Felmerül tehát a mutációs spektrumok triplett-gyakorisággal történő normálása is, amivel alapvetően különböző tulajdonságú genomokon mért eredmények összevethetővé válnak.

A fenti normalizációs eljárásokat külön-külön vagy együttesen használva egymástól teljesen eltérő eredményeket kapunk. Emellett, mivel az nnmf módszer során az algoritmusnak explicit módon meg kell adni a detektálni kívánt szignatúrák $k$ darabszámát, ennek az értéknek a kiválasztása is számos kérdést vet fel. Bár a [112] tanulmány az elemzés kezdetén szétválasztja a betegségtípusokat és szeparáltan vizsgálja azokat, majd utolsó lépésként klaszterezi az azonosított szignatúrákat, felmerül annak a lehetősége is, hogy az összes daganattípust együttesen elemezzük. A fent használt nnmf algoritmus sem kizárólagos opció a szignatúrák azonosítására, a főkomponens-analízis (PCA, principal component analysis) vagy a CUR-dekompozíció éppúgy használható a célra. A főkomponens-analízis tulajdonképpen egy ortogonális transzformáció, melynek során egy új koordinátarendszerbe képezzük le az adatokat. Az új tengelyeket hívjuk főkomponenseknek. Ezeket olyan módon definiáljuk, hogy az első komponensre történő vetítés eredményeként az adatoknak a lehető legnagyobb varianciája legyen, illetve minden további főkomponensre a megmaradó variancia lehető legnagyobb hányada jusson, azzal a megkötéssel, hogy az adott főkomponens minden korábban meghatározott főkomponensre ortogonális legyen. Szemléletesen úgy is tekinthetünk a folyamatra, mintha egy $K$-dimenziós (a szignatúrák esetében $K = 96$) ellipszoidot illesztenénk az adatpontokra a $K$-dimenziós térben és a főkomponenseket, mint az ellipszoid tengelyeit definiálnánk. Ezek közül aztán elhagyjuk azokat, melyek irányában az ellipszoid tengelye rövid, vagyis a variancia kis hányadát magyarázzák csak. A fennmaradó első $k$ főkomponens terére vetítve az adatokat, azt várjuk, hogy az eredeti adatpontok közti hasonlóságok és különbségek ebben a $k$-dimenziós altérben is viszonylag jól megmaradnak. Így a főkomponens-analízis elsődlegesen a sokdimenziós adatok vizsgálata során a dimenzió-redukcióra szolgál, csökkentve a komplexitást és segítve a vizualizációt. A CUR-dekompozíció egy vizsgált $A$ adatmátrixot a $C$, $U$ és $R$ mátrixok szorzatára bontja, ahol a $C$ mátrixban az $A$ mátrix oszlopainak részhalmaza, az $R$ mátrixban pedig az $A$ mátrix sorainak részhalmaza szerepel. Vagyis a szignatúrák azonosítása szempontjából a CUR-dekompozíció az összes vizsgált minta triplettspektruma közül megkeresi azt a $k$ darab erősen reprezentatív spektrumot, melyek a teljes mintahalmazt megfelelően jól tudják jellemezni. Bár az nnmf és a CUR-dekompozíció eredményeként kapott vektorok biológiai interpretációja alapvetően kézenfekvőbb, a PCA-nak ezzel szemben megvan az előnye, hogy a főkomponensek egymásra ortogonálisak. Emellett, mivel az nnmf során az algoritmus kezdetben véletlenszerűen inicializálja a $P$ és $E$ mátrixokat, előfordulhat, hogy a Frobenius-normának egy lokális minimumába való beragadás miatt nem találjuk meg a globális minimumot. Mindhárom megoldás lehetséges választ jelent a szignatúrák azonosítására, de minden esetben számolnunk kell az adott módszer előnyeivel

és hátrányaival is.

Annak a vizsgálatára, hogy a fenti elemzési opciók különböző megválasztásai miként befolyásolják a kapott eredményeket, egy interaktív jupyter notebookot hoztunk létre, mely a https://mybinder.org/v2/gh/pipekorsi/somaticSignatures/master címen kipróbálható. A notebook a [112] kutatás adatait elemzi, de lehetőséget biztosít az adattípus (csak teljes genom / csak exom / mindkettő), a daganattípusok csoportosításának (összes együtt / típusonként külön), a dekompozícióhoz használt algoritmus (nnmf, CUR, többféle PCA), a normalizálás módjának (nincs / 1-re normált / triplettgyakoriságra normált / mindkettő), illetve a mutációs szignatúrák számának a megválasztására. Az nnmf és PCA algoritmusokhoz az sklearn [113] Python csomag implementációját használtuk, a CUR-dekompozícióhoz pedig az rCUR [114] R csomag algoritmusát írtuk át Python kódra. A használat során látható, hogy a paraméterek bármelyikének megváltoztatásával az eredmények is teljesen különbözőek lesznek. Ez bár felvet némi kétséget a módszer megbízhatóságát illetően, mégis ez a jelenleg alkalmazott egyedüli olyan megközelítés, mely nem konkrét, a betegség kialakulását előidéző („driver”) mutációk keresésén alapul, hanem a teljes mutációs lista statisztikai tulajdonságait elemzi. Nyilvánvalóan a jövőben szükség lesz a konszenzus szignatúrák pontosítására, illetve pontos biológiai hátterük feltérképezésére. Ez leginkább olyan kísérletek során valósítható meg, melyeknél egy-egy specifikus genomikai változást előidézve vizsgálják, hogy annak milyen következményei vannak a mutációs spektrumok tekintetében, és ahol vélhetőleg nem keveredik többféle mutációs folyamat együttes hatása.

A fent hivatkozott referencia szignatúrák listája az utóbbi években egyre bővül [115], illetve már kiegészült az indelek és a közvetlen egymás mellett lévő „dupla SNV-k” (DNV, double nucleotide variation) csoportosításán alapuló spektrumokkal. A témában egyre szélesebb körű felhalmozódott adat és információ ellenére a különböző kísérletek során elemzett mintákban a mutációs folyamatok nyomainak felderítése korántsem egyszerű feladat: mivel a szignatúrák egymásra nem ortogonálisak, a mért spektrum triviális levetítése a konszenzus vektorokra nem célravezető megoldás. Így az egyes minták dekompozíciójához közelítő módszerekhez kell folyamodni, melyekre később térünk ki részletesen.

# MUTÁCIÓK GYORS ÉS MEGBÍZHATÓ DETEKTÁLÁSA

## KEVÉSSÉ ISMERT GENOMOK ELEMZÉSE SORÁN FELMERÜLŐ PROBLÉMÁK

Ahogy a korábbiak alapján láthattuk, a szomatikus mutációk detektálása nem minden esetben kézenfekvő feladat. Az alacsony komplexitású genomi régiók, a szekvenálási hibák [116, 117] és a genom különböző szakaszai közti homológia megnehezítik az illesztést és a valódi mutációt rejtő jelek zajtól való elkülönítését [96, 118–120]. Különösen problémás a DNS-ben történt elváltozások nyomon követése olyan esetekben, amikor nincs információnk arról, hogy melyek a gyakran mutálódó genomi pozíciók populációs szinten. Az ilyen adatok hiánya elsősorban a nem emberi minták, illetve ritkán szekvenált sejtvonalak vizsgálata esetén jelent gondot, amikor még a referenciának használt genom részletei sem teljesen kidolgozottak. Mivel emellett a legtöbb detektáló szoftver a humán genomra lett optimalizálva, valamint az esetek többségében daganatos genomok elemzését tűzik ki célul; nem meglepő, hogy másféle kísérleti elrendezésekben ezek az eszközök nem mindig megbízhatóak. Az így felmerülő nehézségekre a irodalomban egyik leggyakrabban látott megoldás a meglévő szoftverek alapértelmezett beállításokkal történő használata, majd további heurisztikus szűrési feltételek alkalmazása a fals pozitív találatok csökkentésének érdekében. Ennek a módszernek komoly hátránya, hogy a szűrési lépések jellemzően csak hiányosan vagy egyáltalán nem dokumentáltak, így a kapott eredmények gyakorlatilag reprodukálhatatlanok.

Egy specifikus, de gyakorlatias kísérleti elrendezés, amikor alapvetően egymással megegyező kezdeti sejtek populációjából kiindulva azt vizsgáljuk, hogy különböző mutagén kezelések milyen hatással vannak az egyes sejtekre. Ilyen jellegű kísérleteket rutinszerűen alkalmaznak olyan kutatásokban, melyek különféle gyógyszerek és környezeti tényezők mutációs hatásait [121, 122], a kezelésekre való rezisztencia kialakulását [123, 124], illetve különböző genetikai elváltozások okozta mutagén folyamatok feltérképezését [125] tűzik ki célul. Egy hasonló kísérletet elvégezve a DT40 bankivatyúk sejtvonalon, a teljes genom szekvenálási adatok vizsgálata során azt tapasztaltuk, hogy a hagyományos mutációdetektáló szoftverek, mint a VarScan 2 [103] vagy a MuTect [111], nem alkalmasak az adatok megfelelő elemzésére még akkor sem, ha a kontrollparamétereket a legoptimálisabbnak választjuk meg.

Az így adódó problémák áthidalására az IsoMutot, egy olyan mutáció-detektáló algoritmust dolgoztunk ki [126], mely rendkívül gyorsan és nagyon precízen képes a mutációk azonosítására az olyan esetekben, amikor több, alapvetően hasonló genomú (izogenikus), de különböző kezeléseknek alávetett minta áll rendelkezésünkre. A módszer teszteléséhez használt minták egyetlen sejt felszaporításából eredő, homogén populációkból (single cell clone) származtak, és feltételeztük, hogy a kezelések okozta mutációk egymástól függetlenül jelentek meg. A kidolgozott szoftver eredeti verziója tehát olyan pontmutációkat (SNV; single nucleotide variation) és indeleket (inzerciók és deléciók) keres, melyek csak

egyetlen mintában fordulnak elő. A több mintát is érintő mutációk kiszűrésével megszabadulunk a gyakran mutálódó genomi pozícióktól (SNP; single nucleotide polymorphism) és az illesztési hibák jelentős részétől, hiszen ezek gyakran minden mintában azonos helyen jelennek meg. A több minta egyidejű elemzése ilyen módon feloldja a pontatlan referenciagenom és a hiányos csíravonal mutációs adatbázis problémáját. Az IsoMut a mutációk azonosítása során egy nagyon egyszerű stratégiát követ: a legtöbb szűrési paraméter értékére fix határértékeket állít be, melyek a szekvenálási adatokkal egyértelmű összefüggésben vannak, így a felhasználó számára az eredmények interpretációja triviális feladat, nincs szükség a statisztikai modellek visszafejtésére. Amennyiben az adathalmaz kontroll mintákat is tartalmaz, melyekben nem várható egyedi mutáció, lehetőség nyílik az eredmények finomhangolására, mellyel a fals pozitívok száma tovább csökkenthető.

### Adatok és előkészítésük

A detektáló módszer teszteléséhez és optimalizálásához használt teljes genom szekvenálási adatszett mintái a bankivatyúk DT40 vérrákos sejtvonal [127] különböző kémiai anyagokkal kezelt klónjai voltak. A klónok egyik fele vad típusú (WT; wild type) volt, míg a többi mintában a BRCA1 génben létrehozott mesterséges, homozigóta mutáció miatt ez a DNS-javító mechanizmusokban kulcsszerepet betöltő gén funkcióját vesztette ($BRCA1^{-/-}$) [128]. A minták szekvenálásra történő előkészítését megelőzően a sejtpopulációkból egy-egy sejt elkülönítésével, majd felszaporításával olyan homogén genomú populációkat lehetett létrehozni, melyek szekvenálásával végeredményben egyetlen sejt genomját tudtuk elemezni. A kísérlet során használt kémiai anyagok átlagosan 50-5000 mutációt hoztak létre mintánként, ami nagyságrendileg összevethető a daganatos mintákra jellemző mutációs rátával [129]. Ez a viszonylag enyhe mutációs teher megköveteli, hogy az elemzés során a fals pozitív találatok számát a lehető legalacsonyabban tartsuk.

Összesen 30 mintát analizáltunk, melyek egymástól a genotípusukban (vad típus vagy $BRCA1^{-/-}$) és az őket ért kémiai kezelésben tértek el (B1. táblázat). Mivel a kezelés konkrét részletei és a genotípusok specifikus jellemzői a mutáció-detektáló módszer szempontjából irrelevánsak, így általánosan „WT” és „mutáns” klónokként hivatkozunk a mintákra, a kémiai kezeléseket pedig mutagén hatásuk erősségével jellemezzük. Ha két minta között a genotípus és a kezelés is megegyezett, a két minta genomja nem feltétlen volt azonos, hiszen az esetek többségében a megszekvenált DNS különböző kezdeti klónoktól eredt. Az egyedüli valóban megegyező mintapárok (S12, S15 és S27, S30) azonos DNS preparátum kétszeri megszekvenálásával lettek létrehozva. A duplikátumok használata nagyban segíti az elemzési módszerek megbízhatóságának tesztelését, hiszen két elvileg megegyező minta között az elemzési eredményekben sem várunk különbséget.

A minták szekvenálása az Illumina platformon történt két menetben, paired end, 125, illetve 150 bázispár hosszú short readeket létrehozva. A két szekvenálási alkalom közti technikai különbségek kiváló lehetőséget adtak arra, hogy az így fellépő műtermékek kompen-

zálásával olyan elemzési módszert fejlesszünk, mely a különböző módokon nyert adatok összevetésére is alkalmas. Az elemzett minták nyers szekvenálási adatai az ENA (European Nucleotide Archive; http://www.ebi.ac.uk/ena/) honlapjáról tölthetők le az ERP014915 azonosítóval.

Az analízis első lépéseként a nyers szekvenálási adatokból BAM, illetve pileup fájlokat hoztunk létre. Ehhez a short readeket a bankivatyúk (*Gallus gallus*) referencia genom Galgal4.73-as verziójához [130] illesztettük a bwa-mem paranccsal [94]. A duplikált readeket a samblaster program [131] segítségével távolítottuk el, továbbá a potenciális indelek közelében a readek újraillesztését a GATK IndelRealigner eszközzel [98] végeztük.

Az így elkészült BAM fájlokból a samtools mpileup parancsával [95] létrehoztunk egy közös pileup fájlt az összes vizsgált minta adatainak felhasználásával. Erre a lépésre elsősorban azért volt szükség, hogy a tesztelési fázisban, mikor ismételten vissza kellett nyúlni ezekhez az adatokhoz, lecsökkentsük a számításhoz szükséges időt. A végső szoftver használata során a pileup fájlokat azonban nem tároljuk el.

Az mpileup parancs futtatásakor bekapcsoltuk a „-B” és „-Q 30” kapcsolókat, ezek az opciók azonban az IsoMut szoftver használata során tetszőlegesen megválaszthatók. Mindazonáltal a „-Q 30” szűrési feltétel, mely minden olyan bázist figyelmen kívül hagy, melynek a bázisminősége 30 alá csökken, átlagos minőségű szekvenálási adatok mellett jogosan alkalmazható. Ezt egy olyan vizsgálattal igazoltuk, melynek során a bázisminőségre vonatkozó szűrési értéket folyamatosan növeltük és eközben vizsgáltuk azoknak a genomi pozícióknak az arányát, melyek megfelelően lefedettek maradtak (legalább 10-es lefedettség), illetve azoknak, melyek minden mintában teljesen „tiszták” voltak. Tisztaként azokat a genomi pozíciókat azonosítjuk, melyekben az összes leolvasott bázis megegyezik a referencia-bázissal. A B1. ábrán látható, hogy a 30-as határérték környékén a lefedett pozíciók száma drasztikusan csökkenni kezd, míg a tiszta pozíciók aránya ekkorra csaknem 90% körüli, így ezzel a szűréssel hatékonyan meg tudunk szabadulni a szekvenálási hibák okozta zaj jelentős részétől.

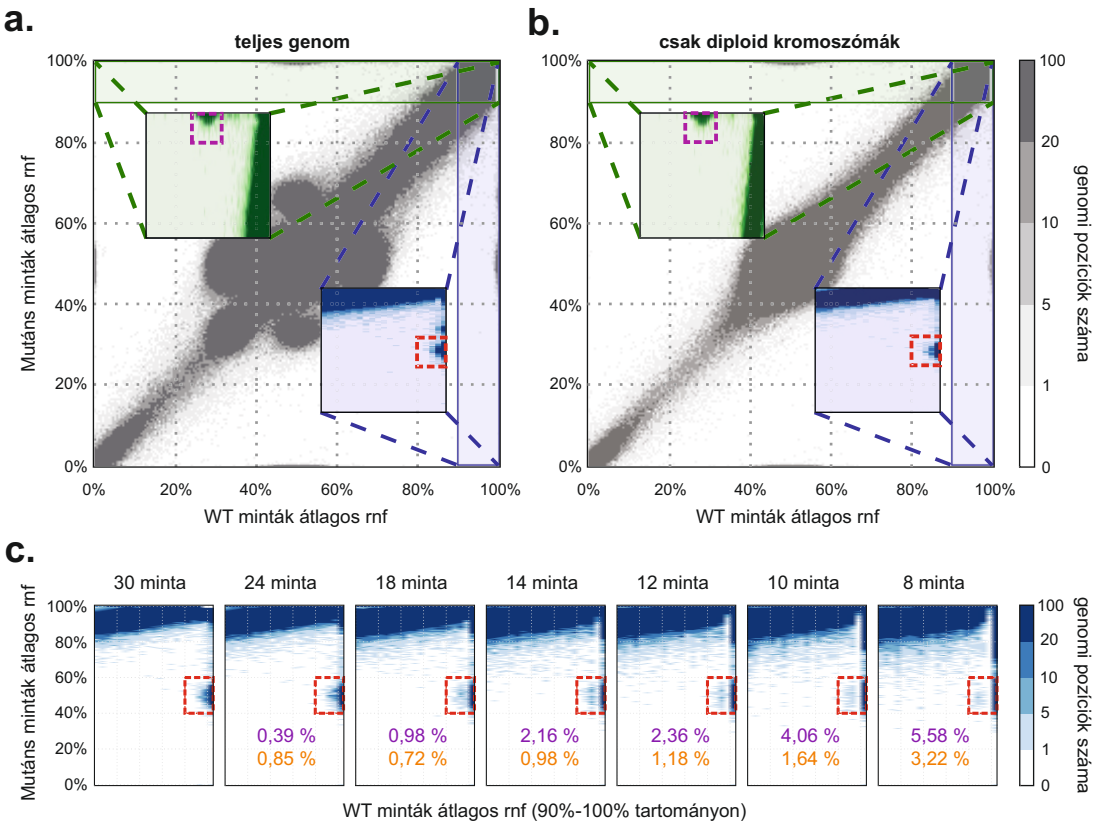
A tesztelés során használt fenti módon generált pileup fájlokból továbbá kiszűrtük azokat a mutáció-detektálás szempontjából „érdektelen” pozíciókat, melyekben egyik minta sem tért el legalább a leolvasott bázisok 10%-ában a referenciagenomtól. Ennek a lépésnek célja az idő- és tárhelynyerés volt, hiszen így az elemezni kívánt fájlok mérete mindössze 1%-ára csökkent. Mivel a módszer ennél a 10%-os küszöbértéknél szigorúbb határértékekkel operál, ez az előzetes szűrés a végső eredményeken nem változtat.

### Megbízható mutációs teszthalmazok létrehozása és a módszer megbízhatóságának tesztelése

A mutációkat detektáló eszköz algoritmikus részleteitől függetlenül a hatékonyság teszteléséhez szükség van egy olyan mutációs listára, melyeknek tagjairól kellően biztosan tudjuk, hogy valóban mutációk és nem műtermékek eredményéből adódtak. Elsőként tehát

ilyen megbízható mutációs referencia teszthalmazok létrehozására törekedtünk.

A vizsgált mintahalmaz kétféle genotípusú (WT és mutáns) klónokat tartalmazott, melyek különböző mutagén kezeléseknek voltak alávetve. Alapvetően tehát két típusú mutáció megjelenését várjuk a mintákban: a kezelések okozta, elsősorban heterozigóta, szomatikus mutációkét, melyek mintánként egyediek; illetve az adott genotípusra jellemző, homo- vagy heterozigóta csíravonal mutációkét, melyek genotípusonként minden mintában megjelennek. Az utóbbi kategória heterozigóta mutációit válogattuk ki a teszthalmazok generálása során.




**5. ábra. Mutációs teszthalmazok készítése a különböző genotípusú mintákban. a, b.** Az átlagos referencia bázis arányának (rnf: reference nucleotide frequency) alakulása a két mintacsoportban; **a.** a teljes genom mentén; **b.** csak a diploid régiókban. **c.** Ugyanilyen módon generált ábrák különböző mintaszámok esetén. A lilával jelölt értékek az eredeti teszthalmazból elvesztett pozíciók arányát, a narancssárgával jelöltek pedig az újonnan bekerült pozíciók arányát mutatják.

Ehhez első lépésként minden genomi pozícióban kiszámoltuk a referencia bázisok átlagos arányát a WT és a mutáns mintákban külön-külön. A teljes genomra így kapott eredményeket a 5. ábra a) paneljén ábrázoltuk. Láthatóan jól elkülönülő klaszterek jelzik a egyik genotípusban heterozigóta, a másikban pedig homozigóta referencia pozíciókat a [100, 50%] és az [50, 100%] koordináták környékén. Ezeknél kisebb további klaszterek jelennek meg a [100, 70%] és a [70, 100%] koordinátáknál a genom nem diploid régiói miatt, melyet alátámaszt a b) panel ábrája, ahol kizárólag a diploid szakaszokon kapott

eredményeket ábrázoltuk. Az ezen az ábrán már határozottan elkülönülő klaszterek pozícióit definiáltuk végül teszthalmazokként, így az algoritmust alapvetően a diploid genomokra optimalizáltuk, de a szoftver legújabb verziójában (lásd lent) a módszert kiterjesztettük aneuploid minták elemzésére is. Ezzel a módszerrel ugyan előfordulhat, hogy a teszthalmazokba kerülő pozíciók nem heterozigóta csíravonal mutációk az egyik genotípusban, hanem valójában a másik genotípus vesztette el a heterozigótaságát az adott szakaszon (LOH; loss of heterozygosity), ennek azonban nincs különösebb jelentősége, ugyanis a tesztelésre ezek a genomi pozíciók éppen úgy alkalmasak, mint a valódi csíravonal mutációk, biológiai eredetüktől függetlenül. A kapott teszthalmazok összesen megközelítőleg 4000 genomi pozíciót tartalmaznak, mely már egy elegendően nagy szám a valós pozitív és a fals pozitív találatok rátájának (TPR, true positive rate; FPR, false positive rate) megbízható becsléséhez.

A fenti módon felállított teszthalmazok segítségével a valós pozitív találatok számát úgy határoztuk meg, hogy mindkét genotípusú minták közül kiválasztva a kezdeti, kezelésnek nem alávetett mintát, majd ezt együtt elemezve a másik genotípus összes többi mintájával, a kezdeti klónban elvileg meg kell találnunk az adott teszthalmazban szereplő összes csíravonal mutációt, mint „egyedi" mutációkat. TPR-ként a kétféle genotípusú kezdőklónban valóban megtalált teszthalmazbeli mutációk arányának átlagát definiáltuk:

$$TPR = mean\left\{\frac{TP_{WT}}{N_{WT}}, \frac{TP_{mut}}{N_{mut}}\right\},$$

ahol $TP_X$ az $X$ genotípusú kezdőklónban talált $X$ teszthalmazbeli mutációk száma, $N_X$ pedig az $X$ genotípusú teszthalmazban található összes pozíció száma.

A fals pozitívok meghatározása számos különböző módszerrel történhet, ezek közül hármat együttesen használtuk az FPR becsléséhez. Egyrészt a fenti elrendezésben, mikor az egyik genotípusból csak a kezdőklónt, míg a másikból az összes mintát együtt elemezzük, minden olyan kezdőklónban talált mutáció fals pozitív, mely nem szerepel az adott teszthalmazban, hiszen ezekben a mintákban a kezelés okozta mutációk megjelenését nem várjuk. Ezek $FP_{1,X}$ darabszámát a valódi negatív találatok számával, tehát a vizsgált genomi régió $L$ hosszának és a teszthalmaz $N_X$ számosságának különbségével kell lenormálnunk:

$$FPR_{1,X} = \frac{FP_{1,X}}{L - N_X}$$

Továbbá a kétféle genotípus identikus mintáiban (S12, S15 és S27, S30) az elemzésben használt minták darabszámától függetlenül egyetlen egyedi mutációt sem lenne szabad azonosítanunk, hiszen minden elváltozásnak a testvér mintában is meg kell jelennie. Az identikus mintapárok tagjainak bármelyikében talált $FP_{2,S_i}$ mutációk számát a teljes vizs-

gált genomi régió $L$ hosszával kell leosztanunk:

$$FPR_{2,S_i} = \frac{FP_{2,S_i}}{L}$$

Ahhoz, hogy megvizsgáljuk, mennyire hatékony az algoritmus olyan esetekben, mikor az egyik genotípusú mintából csak nagyon kevés áll rendelkezésre, lefuttattuk az elemzést a minták olyan csoportjain, melyekben az egyik genotípusból az összes minta, míg a másikból csak a kezdőklón és egyetlen másik klón szerepelt. Ebben az esetben azt várjuk, hogy az alulreprezentált genotípus kezdőklónja semmilyen mutációt ne tartalmazzon, hiszen a csíravonal mutációk a genotípus másik mintájában is jelen vannak, kezelés okozta mutációk pedig a kezdőklónban nincsenek. Az $X$ genotípusú kezdőklónban mégis megtalált $FP_{3,X}$ mutációk számát ismét a teljes vizsgált genomi régió $L$ hosszával kell normálni:

$$FPR_{3,X} = \frac{FP_{3,X}}{L}$$

A végső FPR értékét a fenti részeredmények átlagaként definiáltuk:

$$FPR = mean(FPR_{1,WT}, FPR_{1,mut}, FPR_{2,S12}, FPR_{2,S15}, FPR_{2,S27}, FPR_{2,S30}, FPR_{3,WT}, FPR_{3,mut})$$

Fontos megjegyezni, hogy az így definiált FPR rendkívül szigorú optimalizálást tesz lehetővé, amire a gyakorlati esetek többségében ilyen formában nincs feltétlenül szükség. Mindazonáltal amikor a fals pozitívok minimalizálása a cél, a használt algoritmust érdemes a lehető legmostohább körülmények között tesztelni.

### Szűrési paraméterek definiálása

Az adatok elemzését szolgáló algoritmus az összes analizált mintáról összegyűjtött információ alapján minden genomi pozícióban ugyanazokat a szűrési feltételeket alkalmazva meghatározza, hogy az adott pozícióban bármelyik mintában található-e egyedi mutáció. Annak érdekében, hogy hatékonyan kiszűrjük a csíravonal mutációkat és azokat a fals pozitív találatokat, melyek az illesztési hibák következtében jelennek meg, három alapvető szűrési paramétert vezettünk be.

A szomatikus mutációt hordozó mintában a szekvenálás során leolvasott adatoknak meggyőzően alá kell támasztaniuk a mutáció létét. Ehhez egyrészt szükséges, hogy az adott genomi pozícióban elég sok adat álljon rendelkezésre, vagyis kellő mennyiségű short read illeszkedjen a referenciagenomra. Az ezt tesztelő szűrési paraméter a mutált mintában mérhető lefedettségre beállított alsó küszöbérték (*sample_cov_min*). Ha az elérhető adatok mennyisége a pozícióban megfelelő, azt is biztosítani kell, hogy ezek jelentős hányada a referenciagenomhoz képesti elváltozásról tanúskodjon. Ennek érdekében definiáltuk a leggyakoribb nem-referencia bázis gyakoriságának alsó határértékét (*sample_mut_freq_min*).

Mivel minden mintában csak az egyedi mutációk keresésére törekszünk, ezért elvárás, hogy az adott pozícióban a nem mutált minták „tiszták” legyenek, vagyis a lehető leginkább hasonlítsanak a referenciagenomra. Ennek a vizsgálatára vezettük be a legkevésbé tiszta (legzajosabb) nem mutált mintára vonatkozó, a referenciabázis gyakoriságát (*rnf*; reference nucleotide frequency) szabályozó alsó küszöbértéket (*other_rnf_min*).

Az ezekkel a szűrési feltételekkel azonosított potenciális SNV-k és indelek egy utóelemzési fázisba kerülnek, melynek során a samtools mpileup parancsát a „-B” opció nélkül is lefuttatjuk a kérdéses pozíciókra, majd a mutált mintára vonatkozó paraméterek (*sample_cov_min*, *sample_mut_freq_min*) értékét újra kiszámítjuk és ellenőrizzük, hogy továbbra is elérik a beállított határértékeket. A samtools által alapértelmezetten alkalmazott BAQ rekalibráció az indelek környezetében mesterségesen lecsökkenti a bázisminőséget, ugyanis ezeken a régiókon nagy a valószínűsége a helytelen illesztésnek. A „-B” kapcsoló ezt az alapértelmezett mechanizmust kapcsolja ki. Azért van szükség egyidejűleg a rekalibráció ki és bekapcsolására is, mert a mutált mintában a lehető legtisztább illesztést szeretnénk elérni, vagyis a lehető legtöbb zajtól mentesíteni kívánjuk az adatokat. Ha tehát a rekalibrációt bekapcsolva hagyjuk, az alapértelmezett bázisminőség-szűrés által az így lecsökkent minőségű bázisoktól megszabadulunk. Ezzel szemben az összes többi mintában megjelenő zajnak a maximális szintjéről igyekszünk információt szerezni, ehhez szükséges a BAQ rekalibráció kikapcsolása. Az indelek esetében emellett az utóelemzés során kiszűrjük azokat a potenciális mutációkat, melyek bármilyen másik „gyanús” pozíciónak a közvetlen közelében vannak. Gyanús pozíciónak itt azokat a genomi helyeket nevezzük, melyekben az indel gyakorisága legalább egy mintában meghaladta a 0,2-et, ezek „közvetlen közele” pedig a pozíció 10 bázisos környezete. Erre a lépésre azért van szükség, mert az indelt tartalmazó readek felillesztése gyakran problémákba ütközik és emiatt egy csíravonal indel gyakran egymástól némileg elcsúszva jelenik meg a különböző mintákban, ezzel azt a látszatot keltve, hogy szomatikus indelek sokasága tűnik fel több mintában egyszerre, egy rövid genomi szakaszra klaszterezve.

Az optimalizálás során az utóelemzés paraméterein nem változtattunk, és a három, fenti paraméternek az értékét változtatva kerestük azt a beállítást, mely az adott FPR követelmények mellett a lehető legmagasabb TPR-t biztosítja.

Mivel ez az optimalizálási eljárás egy meglehetősen specifikus kísérleti elrendezést, illetve kellően sok mintát igényel, ilyen formában gyakran nem megvalósítható a rendelkezésre álló szekvenálási adatok használatával. Annak érdekében, hogy egy gyorsabb és könnyebben testreszabható módszert kínáljunk az eredmények finomhangolására, a szoftver minden potenciális mutációhoz hozzárendel egy *S* értéket, mely annak a valószínűségével áll összefüggésben, hogy azt helytelenül kategorizáltuk egyedi mutációnak. Konkrétabban *S* annak a *p* valószínűségnek a negatív logaritmusa, hogy ha feltételezzük, hogy az adott pozícióban a két legzajosabb (legtöbb nem-referencia bázist tartalmazó) mintában a bázisok elméleti eloszlása azonos, akkor éppen a megfigyelt szekvenálási adatokat

kapnánk az eloszlásokból való véletlen mintavételezéssel. Tehát egy alacsony $p$ (magas $S$) érték azt jelenti, hogy igen valószínűtlen, hogy a két legzajosabb mintában valójában is azonos a bázisok eloszlása, vagyis a legzajosabb minta várhatóan ténylegesen egyedi mutációt tartalmaz az adott pozícióban. A $p$ valószínűség meghatározásához a Fisher-féle egzakt tesztet használjuk, melyhez egy $2 \times 2$-es kontingencia táblát definiálunk a két legzajosabb mintára az $n_R$ referencia és az $n_{NR}$ leggyakrabban előforduló nem-referencia bázisok számának meghatározásával:

| **Minta azonosító** | **1** | **2** |
|---|---|---|
| **referencia bázisok** | $n_R^1$ | $n_R^2$ |
| **leggyakoribb nem-referencia bázisok** | $n_{NR}^1$ | $n_{NR}^2$ |

**1. táblázat. Kontingencia táblázat a két legzajosabb mintára az adott genomi pozícióban.** A felső indexekben a minta azonosítók szerepelnek.

Nullhipotézisként tegyük fel, hogy a két mintában a bázisok számának eloszlása megegyezik, vagyis a legzajosabb minta nem tartalmaz egyedi mutációt. Az alábbi módon számolt $p$ ekkor azt a valószínűséget adja meg, hogy ebben az esetben éppen a ténylegesen megfigyelt adatokat kapjuk:

$$p = \frac{(n_R^1 + n_R^2)!(n_{NR}^1 + n_{NR}^2)!(n_R^1 + n_{NR}^1)!(n_R^2 + n_{NR}^2)!}{n_R^1! n_R^2! n_{NR}^1! n_{NR}^2! (n_R^1 + n_R^2 + n_{NR}^1 + n_{NR}^2)!},$$

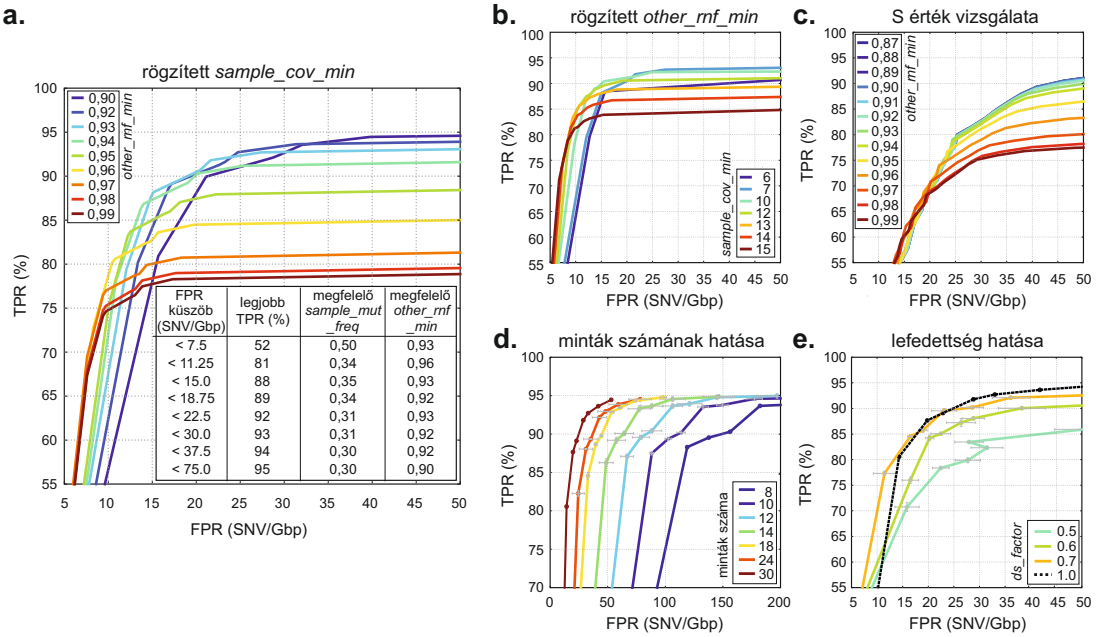
ahol ! a faktoriális operátor. A mutáció-detektáló szoftver ennek az értéknek az $S = -\log p$ transzformáltját rendeli a genomi pozíciókhoz.

Így tehát ha nincs is lehetőség a fenti három szűrési paraméter széleskörű optimalizálására, amennyiben az adathalmazban rendelkezésre állnak kontroll minták (melyekben nem számítunk egyedi mutációk megjelenésére), az $S$ értékre beállított küszöbérték segítségével a fals pozitívok száma minimalizálható.

### Optimális szűrési értékek, a minták számának és a lefedettségnek a hatása

Az optimális szűrési határértékek beállítása során tehát a fentiek szerint a lehető legszigorúbban határoztuk meg az aktuális FPR értékét. Ideálisabb körülmények mellett, amikor mindkét genotípusból kellően sok minta áll rendelkezésre, még jobb eredmények érhetőek el az $S$ érték megfelelő beállításával.

Elsőként fix *other_rnf_min*$= 0,93$ paraméter mellett vizsgáltuk meg a *sample_cov_min* és *sample_mut_freq_min* szűrők változtatásának együttes hatását (6. ábra, b) panel), mivel azonban a lefedettség gyakran már a szekvenálási adatok minősége és mennyisége által korlátozva van, így a későbbiekben egy rögzített, viszonylag megengedő *sample_cov_min*$=$ 7 értékre állítottuk be az ezt kontrolláló paramétert. Emellett az érték mellett vizsgáltuk a *sample_mut_freq_min* és *other_rnf_min* paraméterek változtatásának hatását a TPR és




| FPR küszöb (SNV/Gbp) | legjobb TPR (%) | megfelelő *sample_mut_freq* | megfelelő *other_rnf_min* |
|---|---|---|---|
| < 7.5 | 52 | 0,50 | 0,93 |
| < 11.25 | 81 | 0,34 | 0,96 |
| < 15.0 | 88 | 0,35 | 0,93 |
| < 18.75 | 89 | 0,34 | 0,92 |
| < 22.5 | 92 | 0,31 | 0,93 |
| < 30.0 | 93 | 0,31 | 0,92 |
| < 37.5 | 94 | 0,30 | 0,92 |
| < 75.0 | 95 | 0,30 | 0,90 |

**6. ábra. Különböző paraméterbeállítások mellett mért TPR és FPR értékek. a.** Az *other_rnf_min* (különböző görbék) és *sample_mut_freq_min* (görbék mentén) paraméterek változtatásának hatása konstans *sample_cov_min* $= 7$ mellett. A táblázat adatai az adott FPR mellett elérhető maximális TPR-t, illetve az ehhez szükséges paraméter-beállításokat tartalmazzák. **b.** A *sample_cov_min* (különböző görbék) és a *sample_mut_freq_min* (görbék mentén) paraméterek változtatásának hatása konstans *other_rnf_min* $= 0,93$ mellett. **c.** A *other_rnf_min* (különböző görbék) és az *S* (görbék mentén) paraméterek változtatásának hatása konstans *sample_cov_min* $= 5$ és *sample_mut_freq_min*$= 0,21$ mellett. **d.** A minták számának hatása. A mérési pontok megfelelnek az **a.** panel táblázatában látható paraméter-beállításoknak. A mérési pontok három véletlenszerűen kiválasztott mintahalmazra kapott eredmények átlagát, a hibavonalak pedig azok szórását ábrázolják. **e.** A lefedettség mesterséges csökkentésének hatása. A mérési pontok megfelelnek az **a.** panel táblázatában látható paraméter-beállításoknak. A mérési pontok három véletlenszerűen alulmintavételezett mérésre kapott eredmények átlagát, a hibavonalak pedig azok szórását ábrázolják. (*ds_factor*: down-sampling factor; alulmintavételezési ráta)

FPR értékekre (6. ábra, a) panel). Attól függően, hogy a kísérleti elrendezés várhatóan hány mutációt indukál a vizsgált mintákban, különböző mennyiségű fals pozitív találat tolerálható. A 6. ábra a) paneljén látható táblázat különböző FPR értékek mellett a maximális elérhető TPR értékét és az ehhez szükséges paraméter-beállításokat tartalmazza. A teszteléshez használt adathalmazon a viszonylag alacsonyra állított FPR mellett 92%-os TPR elérésére volt lehetőség a *sample_mut_freq_min*$= 0.31$, *other_rnf_min*$= 0.93$ és *sample_cov_min*$= 7$ beállításokkal.

Annak érdekében, hogy felmérjük, hogy a rendelkezésre álló minták száma mennyiben befolyásolja a kapott eredményeket, az eredeti 30 mintából különböző, $n$ számosságú részhalmazokat alkottunk, melyekben a mutáns és WT genotípusú minták száma megegyezett. A mutációs teszthalmazok létrehozását minden ilyen módon csökkentett adatszettre külön-külön elvégeztük, majd kiszámoltuk a 6. ábra a) paneljének táblázatában látható

beállítások mellett kapott TPR és FPR értékeket. A 6. ábra d) paneljének $8 < n < 30$ mérési pontjai három, a mintákból véletlenszerűen létrehozott, $n$ mintaszámú adathalmazra kapott eredmények átlagát, a hibavonalak pedig ezek szórását jelölik. Mivel az eredeti adatszettben 30 minta állt rendelkezésünkre, melyekből 8 az FPR és TPR értékek meghatározásához feltétlenül szükségesek voltak (például kezdőklónok, identikus mintapárok tagja), így az $n = 30$ és $n = 8$ mintahalmazokat csak egyféleképpen lehetett kiválasztani. Ezért az ezekhez tartozó mérési pontok az ábrán nem átlagokat, hanem egyetlen konkrét mérés eredményét reprezentálják.

Kevesebb minta vizsgálata során a teszthalmazok meghatározásához használt klaszterek egyre elmosódottabbakká válnak (5. ábra, c) panel), de a tesztszettből elvesző és újonnan besorolt pozíciók aránya viszonylag alacsony marad (kevesebb mint 6, illetve 4%) még mindössze 10 minta esetén is. A csökkentett mintaszám legszembetűnőbb következménye a megnövekedett FPR (6. ábra, d) panel), de nagyon szigorú paraméter-beállítások mellett helyenként a TPR is emelkedett. Ahhoz, hogy kb. 50/Gbp alatt tartsuk a fals pozitív találatok gyakoriságát, de legalább 85%-os TPR-t érjünk el, legalább 14 minta együttes elemzésére van szükség.

Az alacsonyabb szekvenálási mélység hatásának vizsgálatához a meglévő adatokat a mutáns genotípusú kezdőklón esetében alulmintavételeztük, ezzel különböző, az eredetinél alacsonyabb mesterséges lefedettségeket állítva elő. Ehhez minden genomi pozícióban az eredeti lefedettség 70, 60, illetve 50%-ának megfelelő darabszámú bázist véletlenszerűen kiválasztottunk. Időtakarékossági okokból ezt a eljárást csak az eredetileg azonosított SNV-k halmazán végeztük el, kiszűrve azokat a pozíciókat, melyek az alulmintavételezés hatására már nem teljesítették a *sample_cov_min* és *sample_mut_freq_min* paraméterekkel szabott feltételeket. A TPR és FPR értékeket a fentiekhez hasonlóan határoztuk meg a 6. ábra a) paneljének táblázatában szereplő beállításokkal, azzal a különbséggel, hogy az összes eredményt csak a mutáns genotípusra kapott mutációkból származtattuk. Mivel az alulmintavételezés véletlenszerűen történt, minden mérés során három, azonos mértékben alulmintavételezett adathalmazt használtunk. A 6. ábra e) paneljén látható ábra mérési pontjai ezek átlagaként, a hibavonalak pedig a szórásukként álltak elő.

Azt találtuk, hogy az eredeti lefedettség 70%-a mellett minimális különbség figyelhető meg az eredeti és az alulmintavételezett TPR és FPR értékek között, de a szekvenálási mélység további csökkentése alacsonyabb TPR-t és magasabb FPR-t eredményez. Mivel a teszteléshez használt mutáns kezdőklón átlagos lefedettsége 21 körüli volt, javasolt legalább 15-ös átlaglefedettségű mintákat használni a megbízható elemzéshez.

### INDELEK DETEKTÁLÁSI HATÉKONYSÁGA

Az indelek azonosítása alapjaiban véve problematikusabb a pontmutációk detektálásánál. Ennek egyik oka, hogy a referenciagenomra történő illesztés során a short readekben a hézagokat (gap) sokkal magasabb büntetőpont sújtja, mint az egy bázisos eltéréseket a

referenciától. Emiatt az illesztőprogramok a valódi indel azonosítása helyett hajlamosak inkább több pontmutáció együttes bevezetésével felillleszteni az indelt tartalmazó readet. Emellett sűrűn előfordul, hogy az egyébként csíravonal indelek a különböző mintákban néhány bázissal elcsúszva jelennek meg, melyek helyes értelmezése nem magától értetődő. Továbbá indelekből általában sokkal kevesebb van a genomban, mint SNV-ből, ezért elemzésük gyakran statisztikai problémákba is ütközik az alacsony mintaszám miatt.

Ahhoz, hogy képet kapjunk a kidolgozott módszer hatékonyságáról az indelek esetében, az SNV mutációs teszthalmazokhoz hasonlóan létrehoztunk két, indel teszthalmazt is. A 5. ábrával megfeleltethető eredmények az indelek esetén a B2. ábrán láthatóak. Az indel teszthalmazok összesen kb. 400 pozíciót tartalmaznak, tehát jóval kisebbek az pontmutációs változatuknál, így ez a részletes optimalizálást megnehezíti. Így a módszer hatékonyságának becslését csak a 6. ábra a) paneljén lévő táblázatban szereplő paraméterbeállításokkal végeztük el, melynek eredményei a B2. táblázatban láthatók. A kapott értékekből kitűnik, hogy az SNV detektáláshoz használt paraméterek megfelelően alacsonyan tartják az FPR-t indelek esetén is, de emellett a TPR értékek számottevően gyengülnek. Amennyiben az elemzés elsődleges célja kizárólagosan az egyedi indelek azonosítása, az SNV-kkel azonos paraméter-beállítások használata megalapozott. Ha viszont a detektálás érzékenysége a legfontosabb, célszerű megengedőbb paraméter értékeket választani.

### Összevetés a hagyományos algoritmusokkal

Az algoritmus fejlesztésének alapvető oka az volt, hogy a hagyományos mutáció detektáló eszközök a tapasztalataink szerint nem tudták az SNV-ket és indeleket a biológiai interpretációhoz szükséges pontossággal azonosítani anélkül, hogy a program lefuttatása után további szűréseket alkalmaztunk volna. A módszerünk hatékonyságát két népszerű szoftver, a VarScan 2 [103] és a MuTect [111] eredményeivel hasonlítottuk össze.

A VarScan 2-t a szomatikus mutációk detektálására alkalmas üzemmódjában futtattuk, melynek használatával tumor és normál minták összehasonlítására van lehetőség. Mind a „tumor”, mind pedig a „normál” mintának az identikus mintapárok egy-egy tagját választottuk, tehát minden így azonosított mutáció fals pozitív volt. Az összehasonlítást mintapáronként kétszer is elvégeztük, egyszer a pár egyik, egyszer pedig a másik tagját választva „tumor” mintának. A szűrési paramétereket és az utólagos szűrési lépéseket a [132] leírás szerint alkalmaztuk. Végül az egyedi mutációk száma 368, 410, 1264 és 922 lett rendre az S12, S15, S27 és S30 kontroll mintákban. Ezzel szemben az IsoMut által detektált fals pozitívok számra rendre 3, 1, 3 és 5 lett ugyanezekben a mintákban. A két szoftver hatékonyságának eltérését valószínűleg az okozhatja, hogy a VarScan 2 olyan szűrési módszerek kiaknázásán alapul, melyek a humán genomok elemzése során sikerrel használhatóak, de az aktuális adathalmazon nem voltak elérhetőek (pl. SNP adatbázisok).

A MuTectre elsősorban azért esett a választásunk, mert a tumor és normál minták összehasonlítása mellett olyan üzemmódja is létezik, melynek futtatása során egy normál min-

tákból kialakított referenciapanelt is fel lehet használni a fals pozitív találatok kiszűrésére. A MuTect ugyan nem detektál indeleket, de a legújabb változata, a MuTect2 már igen. Mivel azonban ez a verzió a kutatás idején még nem volt elérhető, így a régebbi MuTecttel vetettük össze eredményeinket. Bár az alapértelmezett beállításokat használva az adathalmazon, a MuTect nem teljesített kellően pontosan a biológiailag helyes interpretációhoz (B4a. ábra), a szoftver által a mutációkhoz rendelt, a detektálás megbízhatóságát jelző LOD paraméter küszöbértékének finomhangolásával azonban már jó eredményeket tudtunk elérni (B4b. ábra). Ehhez az optimalizációs folyamathoz szükség volt azonban a kontroll minták elérhetőségére, és egy kevésbé átfogó adathalmaz esetén nem lett volna megvalósítható.

A paraméterek hangolásához és a két módszer összehasonlításához is becsléseket kellett tennünk az FPR és TPR értékére. Egy ideális algoritmus a kontroll mintákban egyáltalán nem találna mutációkat, ezzel minimalizálva a fals pozitívok számát, miközben a kezelt mintákban minél magasabban tartaná az azonosított SNV-k mennyiségét. Természetesen a kezelt mintákban talált mutációk nem feltétlenül valós pozitívok, mégis ezeknek a számát használtuk a TPR becsléséhez. Azt találtuk, hogy nagyon alacsony FPR mellett az IsoMut és a MuTect teljesítménye nagyon hasonló (0,5/Gbp FPR mellett 0,7/Mbp valódi mutáció), megengedőbb FPR mellett azonban az IsoMut érzékenysége magasabb (1/Gbp FPR mellett 1/Mbp, illetve 0,75/Mbp valódi mutációt detektál rendre az IsoMut és a MuTect) (B3a. ábra). Továbbá kevés minta esetén, ugyanezen az adathalmazon az IsoMut jóval jobban teljesít a MuTectnél (B3b. és B3c. ábrák).

Annak érdekében, hogy becslést kaphassunk a különböző szoftverek futási idejére vonatkozóan, az IsoMutot, a MuTectet, a MuTect2-t és a VarScan 2-t lefuttattuk a bankivatyúk genomjának egyik legrövidebb kromoszómáján (28-as; 4,7 Mbp) az adatszettben található összes mintát használva. Egy szerény kapacitású asztali számítógépet használtunk, 23 GB memóriával és 12 maggal. A VarScan 2-t a szomatikus mutációkat detektáló „somatic” üzemmódban futtattuk, minden mintát egy „normál” mintával párban összehasonlítva, így összesen 30 összehasonlítást végeztünk el. A MuTect és a MuTect2 esetében követtük az általános használati útmutatókat a szoftverek internetes leírásában. Elsőként minden mintához egyedien létrehoztunk egy normál panelt az összes többi minta felhasználásával. Ezek után minden mintát egy „normál” mintával párban, a normál panellel együtt elemeztünk.

Az eredmények alapján az IsoMut megközelítőleg 170-szer gyorsabb a MuTect2-nél, több mint 40-szer a MuTectnél és több mint 10-szer a VarScan 2-nél (2. táblázat). Ezt a teljes bankivatyúk genomra extrapolálva, a 30 minta elemzése a használt számítógépen az IsoMuttal 5 órába, míg a MuTect2-vel több, mint 35 napba telne. Mivel az IsoMut kivételével mindegyik szoftver Java nyelven íródott, így a párhuzamosan futtatható folyamatok számát erősen limitálja a véges memória. Ezzel szemben az IsoMut esetében az egyetlen limitáló tényező a magok száma, és a futási idő fő korlátja az írási és olvasási sebesség.

A Java-ban írt eszközök teljesítménye jelentősen növelhető egy nagy kapacitású, 100-200 GB-os memóriával bíró számítógéppel. Mivel azonban az ilyen gépek nem minden kutatás során állnak rendelkezésre, az IsoMut egy jó alternatívát kínál az olyan esetekben, mikor a számítógépes kapacitás véges. Bár nem realisztikus bármelyik eszköz egy magon történő, párhuzamosítás nélküli futtatása, a 2. táblázatban egy ilyen oszlopot is feltüntettük a könnyebb összehasonlíthatóság kedvéért.

| **Eszköz** | **12 mag** | | | | **1 mag** | |
|---|---|---|---|---|---|---|
| | **Párhuzamos folyamatok száma** | **Futási idő** | **Futási idő 1 Gbp hosszú genomra extrapolálva** | **Futási idő az IsoMuthoz viszonyítva** | **Futási idő** | **Futási idő az IsoMuthoz viszonyítva** |
| IsoMut | 12 | 1 perc 24 s | 4 óra 56 perc | 1 | 7 perc | 1 |
| VarScan2 | 5-6 | 16 perc | 2 nap 8 óra | 11 | 1 óra 20 perc | 11 |
| MuTect | 6-7 | 1 óra 7 perc | 9 nap 20 óra | 48 | 4 óra 55 perc | 42 |
| MuTect2 | 4-5 | 4 óra | 35 nap 5 óra | 171 | 21 óra 6 perc | 178 |

**2. táblázat. Futási idő összehasonlítása a különböző mutáció detektáló szoftverek esetében.** A futtatáshoz használt számítógépnek 12 magja és 23 GB memóriája volt. A szoftvereket a 4,735 Mbp hosszú, 28-as bankivatyúk kromoszómán futtattuk, a használt adathalmaz összes (30) mintájának felhasználásával.

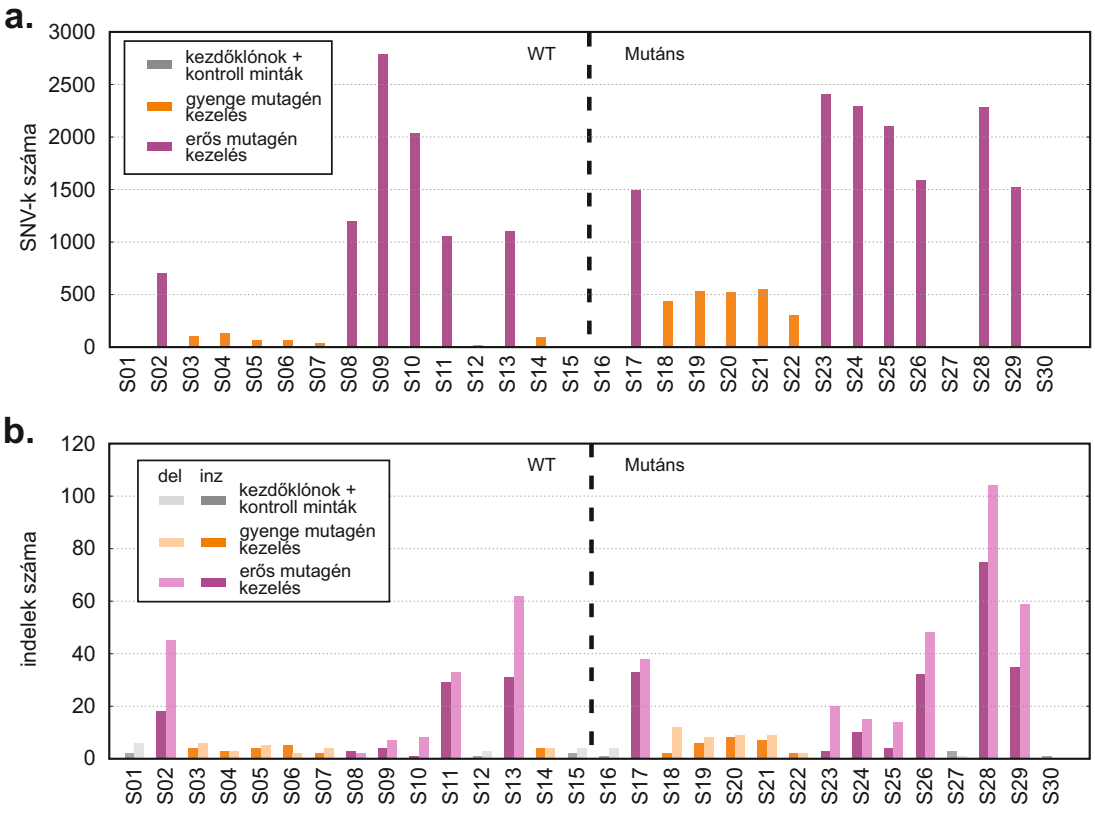
Bár az IsoMut által alkalmazott rögzített szűrési paraméterek használata a fent tesztelt további algoritmusokhoz képest szofisztikálatlannak tűnhet, ez az egyszerű megközelítés bizonyos szempontból előnyösebb, mint a komplikált statisztikai modellek használata, amellett, hogy éppen annyira hatékonynak bizonyult a vizsgált mintákon. A bioinformatikai szoftverek és mutáció detektáló algoritmusok rohamos fejlődése ellenére, az íratlan szabály megmaradt, miszerint a kétséges esetekben a legbiztosabb visszanyúlni az eredeti szekvenálási adatokhoz (például az IGV genom nézegető szoftverrel [133] vagy egy pileup fájlban). Az IsoMut által használt és mutációnként meghatározott paraméterek értékeiből a genomi pozícióra illett bázisok eloszlására egyértelmű következtetések tehetők, anélkül hogy különböző valószínűségszámítási modellek p-értékeit kellene visszafejtenünk.

### Szoftver implementáció

Az egyszerű alkalmazhatóság érdekében a fenti algoritmus számítás-intenzív részét C nyelven, a párhuzamosítást pedig Python-ban implementáltuk. A nyílt forráskódú eszköz eredeti verziója (IsoMut) a https://github.com/genomicshu/isomut oldalról tölthető le, a legújabb verzió (IsoMut2py) pedig a https://isomut2py.readthedocs.io/ oldalon található dokumentáció instrukcióit követve egyszerűen telepíthető.

Az eredeti verzió egyik fő eleme egy Python kód, melyet a megfelelő elérési útvonalak és paraméterértékek módosítása után a parancssorból futtathatunk. Bemenetként a referenciagenomra felillesztett szekvenálási adatokat BAM formátumban várja, illetve szükséges magának a referenciagenomnak az elérési útvonalát is megadni. Mindazonáltal a mutációkat nem a referenciától való eltérésként, hanem a minták egymástól való különbségeiként keressük. Az olyan kísérleti elrendezések esetében érdemes a szoftvert használni,

mikor több izogenikus minta áll rendelkezésre, és ezekben az egyedi mutációkat kívánjuk azonosítani. Minden esetben erősen javasolt a kontroll minták szekvenálása és használata. Ezek lehetnek a kezelést megelőző kezdőklónok vagy azonos DNS-preparátumok többszöri szekvenálási eredményei. Az IsoMut futtatása után ajánlott ezeknek a használatával az $S$ paraméter határértékének manuális beállítása úgy, hogy a kontroll mintákban talált mutációk száma a lehető legalacsonyabb, a többi mintában pedig a legmagasabb legyen. Az $S$ érték változtatásának a hatását a 6. ábra c) panelje szemlélteti rögzített *sample_mut_freq_min*$= 0,21$ és *sample_cov_min*$= 5$ paraméterek mellett. Az $S$ értékének utólagos beállítása azt is lehetővé teszi, hogy az SNV-kre, inzerciókra és deléciókra különkülön optimalizált értékeket használjunk. Amennyiben erre nincs lehetőség, érdemes a kívánt FPR értékéhez igazítani a szűrési paraméterek értékét a 6. ábra a) paneljének táblázata alapján. A vizsgált adathalmazra *sample_mut_freq_min*$= 0,31$, *other_rnf_min*$= 0,93$ és *sample_cov_min*$= 7$ paraméterekkel, az $S$ érték optimalizálása nélkül kapott eredményeket a 7. ábra demonstrálja. Kontroll mintánként átlagosan 6 mutációt detektáltunk (FPR $\approx 6 \cdot 10^{-9}$), a kezelt mintákban pedig számos esetben 2000-nél is többet, melyből egyértelműen látható, hogy az alacsony FPR nem a túl szigorú szűrés eredménye.




**7. ábra. Az IsoMut eredményei az $S$ paraméter optimalizálása nélkül. a.** A detektált SNV-k darabszáma mintánként. A függőleges szaggatott vonal a kétféle genotípusú mintákat választja el egymástól, a színkódok a kezelés típusát jelzik. **b.** A detektált indelek darabszáma mintánként. A függőleges szaggatott vonal a kétféle genotípusú mintákat választja el egymástól, a színkódok a kezelés típusát jelzik, a sötétebb oszlopok az inzerciókat, a világosabbak a deléciókat jelölik.

Az IsoMut2py egy Python modul, mely egyszerűen installálható a Python saját csomagkezelő szoftverével (pip). A csomag által használt C kód alapjai megegyeznek az eredeti IsoMut C kódjával, ám számos új funkcióval kibővítik azt. A megszokott paraméterek mellett a detektáló folyamat egyéb lépései is testreszabhatóvá váltak, illetve lehetőség nyílik az egyedi mutációk mellett közös mutációk azonosítására is. A modul emellett további függvényeket tartalmaz, melyek segítik a mutációs eredmények automatikus optimalizálását, ábrázolását, külső mutációs listák importálását, a minták kariotípusának összehasonlítását, illetve mutációs szignatúrák azonosítását. A következőkben az eredeti algoritmushoz képesti további elemzési lehetőségeket tekintjük át.

### KITERJESZTÉS ANEUPLOID MINTÁK ESETÉRE

Bár a humán genom alapvetően diploid (vagyis a testi kromoszómákból két példány található meg az emberi szervezetben), számos sejtvonal és más faj létezik, ahol nem ez a helyzet. Humán, leggyakrabban daganatos minták esetében is előfordul, hogy néhány kromoszómából elveszik egy példány, vagy éppen több, mint kettő van belőle. Bizonyos esetekben a ploiditás még az adott genomon belül sem állandó, hanem kromoszómáról kromoszómára, vagy akár genomi régióról genomi régióra változik. Az ilyen típusú mintákat aneuploidnak nevezzük.

Mivel a mutációk detektálása során használt *sample_mut_freq_min* paraméter az alapján válogatja ki a potenciális genomi pozíciókat, hogy az odailleszkedő szekvenált bázisok aránya megfelelő-e, fontos tudnunk, hogy milyen bázisgyakoriságokra reális számítani. Egy diploid genomnál ideális esetben a mutált bázis gyakorisága vagy 50%, vagy 100% (homozigóta mutációnál). Ezzel szemben egy olyan kromoszómán, amiből három példány is van, a gyakoriságot a 33%, 67% és 100% értékek körül várjuk. Így egy triploid genomon a diploid genomra beállított szűrési feltételek túlságosan szigorúak. Ennek a problémának az áthidalására az a megoldás, ha a lokális ploiditás ismeretében a *sample_mut_freq_min* paraméter értékét dinamikusan állítjuk be. Ehhez azonban elsőként szükség van a lokális ploiditás meghatározására a szekvenálási adatok alapján a teljes vizsgált genom mentén.

A szekvenálás alatt ideális esetben minden genomi szakaszból a mennyiségével arányos short read keletkezik, bár a PCR (polymerase chain reaction; polimeráz-láncreakció) során ezek az arányok módosulhatnak [134]. Ennek ellenére megalapozott abból a feltételezésből kiindulni, hogy a többszörösen megjelenő kromoszómák több short readet eredményeznek, így a lefedettségük is nagyobb lesz. A lokális lefedettségből tehát következtetések tehetők a lokális ploiditásra vonatkozóan. Amennyiben ismert az egyszeresen megjelenő (haploid) genomi régiók $c_{hap}$ lefedettsége, a $p$ lokális ploiditás a lokális $c$ lefedettség alapján elvileg egyszerűen meghatározható:

$$p \approx c/c_{hap}$$

Ehhez azonban elsődlegesen a $c_{hap}$ értékének meghatározására van szükség. Nyomatékosítjuk, hogy az alábbiakban leírt ploiditás-meghatározás csak szennyezetlen és homogén populációk adatait tartalmazó (single cell clone) minták esetén megbízható. Mivel azonban a szoftver eredeti funkciója éppen ezeknek a mintáknak az elemzése, így az aneuploid mintákra való kiterjesztésnél is ezt tartottuk szem előtt.

**Néhány gondolat a bayesi inferenciáról**

Mind a haploid lefedettség becslését, mind pedig a különböző ploiditású genomi komponensek súlyának és azok eloszlásainak a paramétereinek meghatározását bayesi keverékmodellek illesztésével végezzük el. A fogalmak tisztázása érdekében ezért elsőként a bayesi módszerek sajátosságait tekintjük át nagy vonalakban [135] bevezetője alapján.

A bayesi gondolkodás alapvetően eltér a statisztikában hagyományos frekventista felfogástól. Míg a frekventista elgondolás szerint egy esemény bekövetkezésének valószínűsége a bekövetkezések hosszútávú gyakorisága, addig a bayesi valószínűség sokkal közelebb áll az esemény bekövetkezésébe vetett hithez. Emiatt éppen a bayesi gondolatmenet tükrözi jobban a természetesen, intuitív módon becsült valószínűségeket. Nyilvánvalóan a „hit" fogalma teljesen szubjektív. Például ha két fél feldob egy dobókockát, melyet egyikük sem lát, és megtippelik, hogy melyik oldala került felülre, várhatóan mindketten $1/6$ valószínűséget rendelnének minden lehetséges számhoz. De ha az egyikük meglesné, hogy a kocka hogy ért földet, ő már biztosan 1 valószínűséggel azt az egyetlen számot hinné ténylegesen helyesnek, míg a másik fél „hite" nem változna. Tehát az egyének valamilyen esemény bekövetkezésébe vetett hite különböző lehet.

A bayesi felfogás szerint még mielőtt bármilyen adattal találkoznánk, egy $A$ esemény bekövetkezéséhez hozzárendelünk egy „kezdeti hit" értéket, ezt hívjuk $P(A)$ prior valószínűségnek. Ha például egy pénzérme feldobásakor kell a „fej" valószínűségét megbecsülnünk, kezdetben mindenképp jogos a $P(A) = 1/2$ választás. Ha később azt tapasztaljuk, az érme néhány feldobása után, hogy szinte mindig az „írás" kerül felülre, a korábbi hitünket ezekkel az $X$ tapasztalatokkal módosíthatjuk, megkapva a $P(A|X)$ poszterior valószínűség értékét, ami ebben az esetben $1/2$-nél kisebb lesz, de kevés dobás után nem annyira, mint a „fej"-ek gyakorisága lenne (tehát a frekventista valószínűség). Ahogy a megfigyelések száma növekszik, a kétféle valószínűség egyre közelebb kerül egymáshoz.

Maga a poszterior valószínűség számolása a Bayes-szabály alapján triviálisnak tűnhet:

$$P(A|X) = \frac{P(X|A)P(A)}{P(X)} \propto P(X|A)P(A)$$

ahol $P(X|A)$ a „likelihood", vagyis annak a valószínűsége, hogy ha az $A$ esemény ténylegesen bekövetkezik, akkor éppen az $X$ megfigyelést tesszük. Valójában a poszterior valószínűség meghatározása komoly számítási problémákba ütközik, de az eloszlásból történő

mintavételezés az esetek többségében megoldható. Ezt alkalmazzuk a lenti számítások során.

### A haploid lefedettség becslése

A haploid lefedettség becsléséhez az IsoMut2py egy bemeneti BAM fájlból elsőként egy ideiglenes pileup fájlt hoz létre, melyből a lokális átlag lefedettséget és a referencia-allél lokális gyakoriságát egy mozgóátlagolásos módszerrel határozza meg, mely értékeket az előre meghatározott tulajdonságú genomi pozíciókra (lásd: https://isomut2py.readthedocs.io/en/latest/code_ploidy_estimation.html) ideiglenes fájlokban tárolja el. A futási idő minimalizálása érdekében ez a folyamat erősen párhuzamosítottan, a genom rövid szakaszain egymástól függetlenül fut.

A következő lépésben a kapott eredményekből a vizsgált pozíciók véletlen mintavételezésével, 2000 olyan pontban, melyben a lefedettség az előre megadott $[c_{min}, c_{max}]$ intervallumba esik, meghatározzuk a lefedettség-eloszlást. Az extrém (alacsony vagy magas) lefedettségű pozíciók szűrésére azért van szükség, mert ezek gyakran csak a helytelenül felillesztett short readek járulékai. Az eloszlás véletlen mintavételezése, az adatok mennyiségének ily módon történő csökkentése a későbbi illesztést gyorsítja, és nem tapasztaltuk, hogy több megfigyelést használva pontosabb eredményeket tudnánk elérni. Az így kapott eloszlásra - annak standardizálása után, mely a priorok megválasztását teszi egyszerűbbé - a pymc3 Python modul [136] segítségével, MCMC (Markov chain Monte Carlo; Markov-lánc Monte Carlo) mintavételezéssel egy $K = 20$ komponensű bayesi Gauss keverék modellt illesztünk. Bár reálisan nem várjuk, hogy a lefedettség eloszlásban akár 20 különböző ploiditású régióból származó pozíciók is megjelenjenek, ebben a kezdeti lépésben azonban az a célunk, hogy a poszterior várható eloszlás a lehető legjobban illeszkedjen a megfigyelt adatokhoz, függetlenül a folyamat biológiai hátterétől. Ennek érdekében a 20 komponens használatával valójában a zajra is illesztünk. Arról, hogy az általunk választott $K$ értéke valóban elegendően magas ahhoz, hogy kellően jól tudjuk közelíteni az eredeti eloszlást, a gyakorlatban úgy bizonyosodhatunk meg, hogy a poszterior eloszlásban a komponensek egy részénél a súly várható értéke elhanyagolható kell, hogy legyen, ami tapasztalataink szerint a vizsgált esetekben teljesül. A poszterior eloszlás meghatározásához használt priorok a pymc3 modul útmutatása [137] szerint:

$$
\begin{aligned}
\alpha &\sim \text{Gamma}(1,1) \\
\beta_1, \ldots, \beta_K &\sim \text{Beta}(1,\alpha) \\
w_i &= \beta_i \prod_{j=i-1}^{i} (1-\beta_j) \\
\lambda_1, \ldots, \lambda_K &\sim U(0,5) \\
\tau_1, \ldots, \tau_K &\sim \text{Gamma}(1,1)
\end{aligned}
$$

$$\mu_i \mid \lambda_i, \tau_i \sim N\left(0, (\lambda_i \tau_i)^{-1}\right)$$
$$x \mid w_i, \lambda_i, \tau_i, \mu_i \sim \sum_{i=1}^{K} w_i \, N(\mu_i, (\lambda_i \tau_i)^{-1})$$

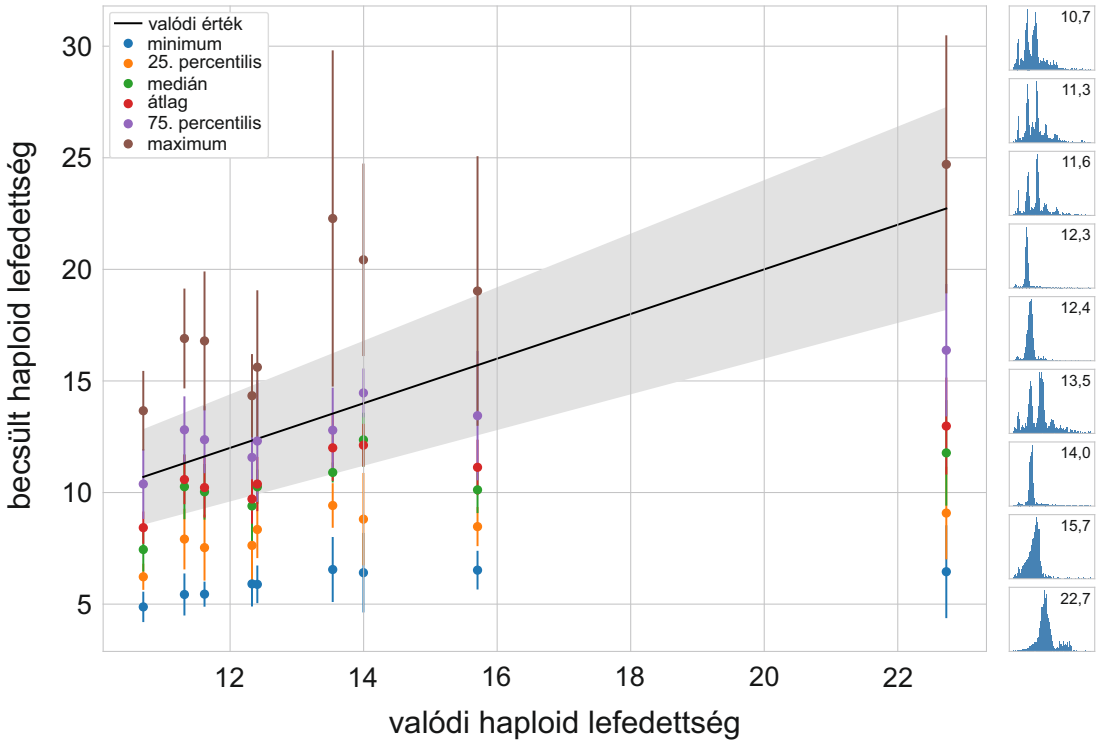
Az MCMC során a poszterior eloszlás mintavételezésével több ezer mintát generálunk, melyekből a modell paramétereinek poszterior várható értéke átlagolással meghatározható. A $K$ db. komponens súlyának várható értékét így kiszámítva, kiválasztjuk azt a $k^*$ komponenst, mely a legnagyobb súllyal szerepel a keverékben. Ebből a $\tilde{c}_{hap}$ értékét a következőképpen származtatjuk:

$$\tilde{c}_{hap} = \frac{\bar{\mu}_{k^*}}{\text{round}(\bar{\mu}_{k^*}/\bar{\mu}_{min})},$$

ahol $\bar{\mu_{k^*}}$ a legnagyobb súlyú komponenshez tartozó Gauss-eloszlás középértékének poszterior várható értékének visszatranszformált (destandardizált) értéke, $\bar{\mu}_{min}$ pedig a komponensekhez tartozó Gauss-eloszlások középértékeinek poszterior várható értékei közül a legkisebb olyannak a destandardizált értéke, ami még magasabb, mint az eredeti lefedettség-eloszlás minimuma. A $\bar{\mu}_k$-k minimumát azért nem érdemes szűrés nélkül használni a haploid lefedettség becslésére, mert a viszonylag nagy számú illesztett komponens miatt az esetek egy részében akár negatív értéke is lehet, ha a hozzá tartozó komponens súlya nem számottevő. Így azt a legkisebb $\bar{\mu}_k$-t választjuk ki helyette, melynek az értéke konzisztens az eredeti adatokkal. Még ennek ismeretében is meglepő a fenti bonyolult képlet használata. A haploid lefedettséget azért alapozzuk jelentős részben a legdominánsabb komponensre kapott várható középértékre, mert ennek az értéke a legbiztosabban becsülhető. Így elsőként megbecsüljük, hogy a legdominánsabb komponens melyik ploiditáshoz tartozhat (a fenti képlet nevezője), majd a hozzá tartozó középérték poszterior várható értékét osztjuk a kapott értékkel.

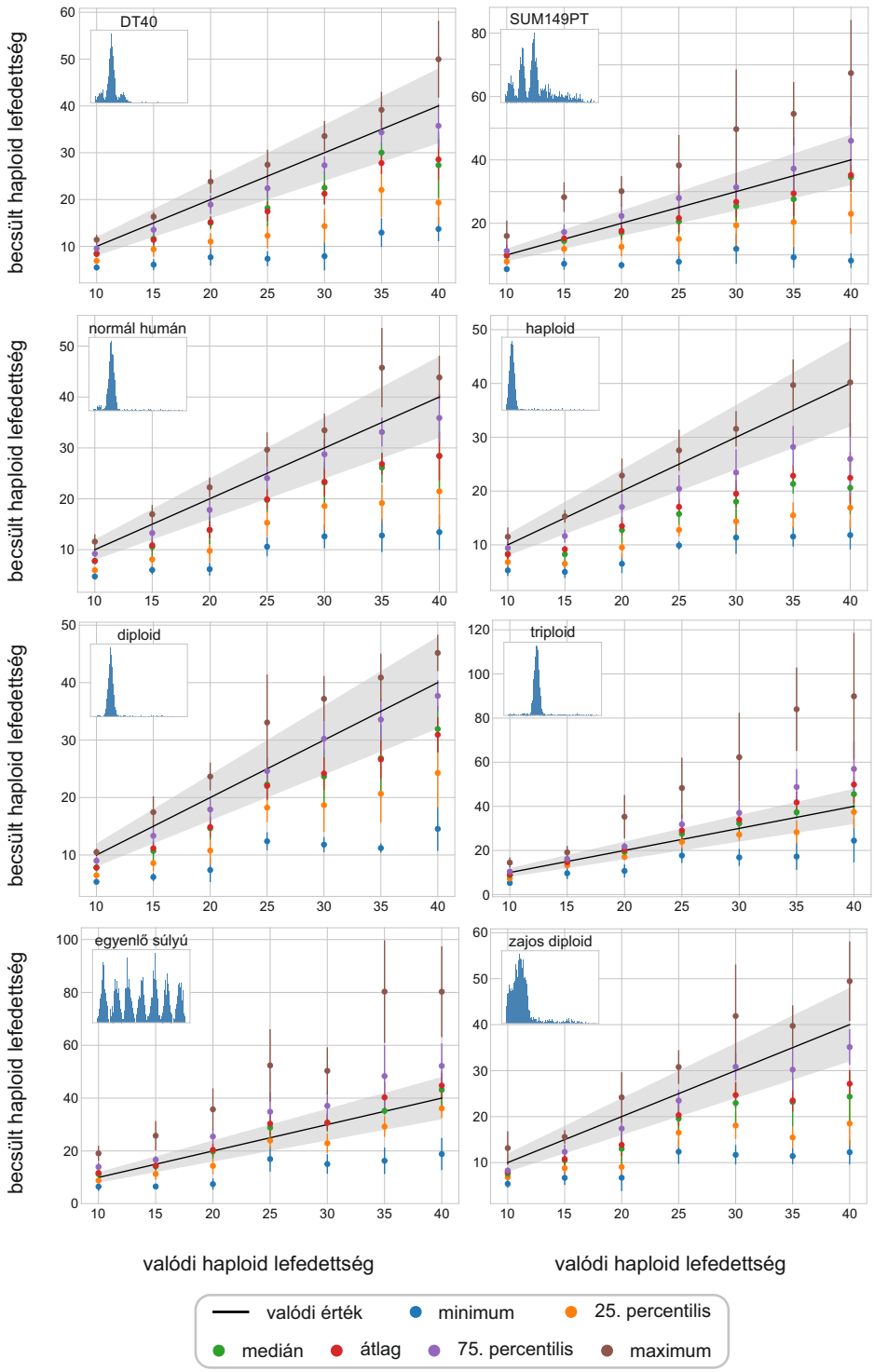
A teljes fenti illesztést és a $\tilde{c}_{hap}$ származtatását 10-szer egymás után elvégezzük, majd a végső becslést $c_{hap}$-ra a kapott $\tilde{c}_{hap}$-k $q$. percentiliseként határozzuk meg. Az, hogy $q$ értékét minek érdemes megválasztani, erősen függ az eredeti lefedettség-eloszlás tulajdonságaitól, így magától a vizsgált genomtól is. Ennek a demonstrálására kilenc olyan valódi minta szekvenálási adataiból határoztuk meg a haploid lefedettséget, melyeknek a lefedettség-eloszlása alapvetően eltér egymástól (8. ábra). Néhány kivétellel a $q = 75$ használatával a becsült érték az eredeti $c_{hap}$ 20%-os környezetébe esik, így alapértelmezett értékként az IsoMut2py ezt használja az elemzés során, de a megfelelő paraméter megváltoztatásával ez szabadon átállítható. Mivel a valós szekvenálási adatok használatával $c_{hap}$ tényleges értéke nem módosítható, 7-féle mesterségesen generált lefedettség-eloszlás esetén azt is teszteltük, hogy a haploid lefedettség értéke milyen hatással van a különböző $q$ értékek hatékonyságára. Minden eloszlás esetében hétféle $p_i, 1 \leq i \leq 7$ ploiditású, $w_i$ súlyú régiót szimuláltunk a genomon belül, mindegyik esetében feltételezve, hogy a

hozzájuk tartozó genomi pozíciók lefedettsége az $N\left(i \cdot c_{hap}, \sigma_i\right)$ Gauss-eloszlást követi. A mesterséges genomok a $w_i$ súlyfaktorokban és a $\sigma_i$ szórásokban különböztek egymástól (B3. táblázat). A 9. ábra különböző paneljei azt szemléltetik a különböző kezdeti eloszlásokra, hogy különböző $c_{hap}$ értékek esetén melyik $q$ érték adja a legjobb becslést. Ebben az esetben is azt találtuk, hogy a $q = 75$-tel kapott eredmények a legtöbb esetben elfogadhatóak, főként azoknál a genomoknál, melyek a gyakorlatban leginkább előfordulnak (DT40, aneuploid, normál humán, zajos diploid), mindazonáltal a tisztán triploid és haploid, illetve az „egyenlő súlyú” mesterséges genomok esetén nem ez a legjobb választás.




**8. ábra. A haploid lefedettség becslése különböző lefedettség-eloszlású genomok esetén.** A jobboldali panelek a különböző genomok lefedettség-eloszlásait szemléltetik, a számok a valódi haploid lefedettség értékét jelölik. A különböző színű pontok a különböző genomokon rögzített $q$-ra 10-szer elvégzett mérések átlagát, a hibavonalak ezek szórását mutatják. A szürke sáv a tényleges haploid lefedettség 20%-os környezetét jelöli.

Mivel a fenti becslés nem teljesen precíz eredményt kíván meghatározni, hanem pusztán arra szolgál, hogy az elemzés következő fázisában a haploid lefedettségre egy elfogadható ($\pm$20%-os intervallumon belüli) priort tudjunk felállítani, némi pontatlanság elfogadható. Ha azonban jól ismerjük a szekvenálási adatok minőségét, és a haploid lefedettségre manuálisan is tudunk becslést tenni (például egy normál humán genom esetén az átlagos lefedettség fele), a fenti lépés kihagyható, és az elemzés folytatható a manuálisan bevitt értékkel. Ezzel részben időt spórolhatunk, másrészt pedig ellenőrizhetjük, hogy biztosan reális eredményeket kapunk.




**9. ábra. A haploid lefedettség becslése különböző lefedettség-eloszlású mesterséges genomok esetén.** A különböző panelek a B3. táblázat szerint generált lefedettség-eloszlások méréseit ábrázolják. A különböző színű pontok a különböző genomokon rögzített $q$-ra 10-szer elvégzett mérések átlagát, a hibavonalak ezek szórását mutatják. A szürke sáv a tényleges haploid lefedettség 20%-os környezetét jelöli.

**Különböző ploiditású régiók súlyának és tulajdonságainak meghatározása a genom mentén**

Amint a haploid lefedettségre a fenti módon megkaptuk az első becslést, a lefedettség-eloszlásra egy újabb bayesi Gauss keverék modellt illesztünk, ezúttal azonban $K = 7$ komponenssel a [138] leírása szerint, a $c_{hap}$ értékét felhasználva a priorok beállításánál:

$$\begin{aligned} \mathbf{w} &\sim \text{Dirichlet}(), \sum_{i=1}^{7} w_i = 1 \\ c_1 &\sim U(c_{hap} \cdot 0,8, c_{hap} \cdot 1,2) \\ \mu_i &= c_1 \cdot i \\ \sigma_i &\sim U(0, c_{hap}/2) \\ t_{1 \leq k \leq N} &\sim \text{Categorical}(p = \mathbf{w}) \\ x_{1 \leq k \leq N} &\sim N(\mu_{t_k}, \sigma_{t_k}) \end{aligned}$$

Vagyis a modellt úgy állítjuk fel, hogy véletlenszerűen kiválasztjuk a 7 különböző ploiditást reprezentáló komponens súlyát, figyelembe véve, hogy a súlyok összegének 1-nek kell lennie. A tényleges $c_1$ haploid lefedettségre olyan priort adunk, ami a fenti módon becsült $c_{hap}$ 20%-os környezetéből azonos valószínűséggel választja ki az értékeket. (Azért fontos a $c_{hap}$ értékét legalább 20%-os pontossággal becsülni, mert ez a prior minden olyan értékhez, ami kívül esik ezen a tartományon, nulla valószínűséget rendel.) A Gauss-eloszlások középértékeit ebből a $c_1$ értékből determinisztikusan, annak egész számú többszöröseiként származtatjuk. Mindegyik Gauss szórására a priort a $[0, c_{hap}/2]$ intervallumon vett egyenletes eloszlásnak állítjuk be. A felső küszöbértéket azért érdemes a $c_{hap}$-hoz igazítani, mert így várhatóan kevésbé csúsznak össze, illetve cserélnek szerepet a különböző görbék az iterálás során. Ezek után minden $k$. mérési ponthoz ($1 \leq k \leq N$) véletlenszerűen, a komponensek súlyával arányosan választunk egy $t_k$ ($1 \leq t_k \leq 7$) kategóriát, vagyis hozzárendeljük valamelyik ploiditás-komponenshez. Annak a feltételes valószínűsége („likelihood”), hogy az $x_k$ mérési pont az adott modellparaméterek mellett épp olyan lefedettségű, amit valóban megfigyelünk, $N(\mu_{t_k}, \sigma_{t_k})$.

A fentiekhez hasonlóan, a kérdéses paraméterek végleges értékének azok poszterior várható értékét tekintjük, melyet az MCMC módszerrel történő mintavételezést követően egyszerű átlagképzéssel határozunk meg. Az illesztett 7 komponensű modell paraméterértékeit a későbbiekben a lokális ploiditás becslése során használjuk fel. Az IsoMut2py lehetőséget biztosít annak az ellenőrzésére, hogy az illesztett modell ténylegesen jól jellemzi-e az eredeti adatokat (B5. ábra). Amennyiben nagy eltérést tapasztalunk a modell és az eredeti lefedettség-eloszlás között, érdemes a 7 komponensű modell illesztését újra elvégezni egy manuálisan megbecsült $c_{hap}$ értékkel.

### Lokális ploiditás becslés

Miután a fenti módon megkaptuk a 7-féle ploiditáshoz tartozó Gauss-eloszlások paramétereit ($w_i$ súly, $\mu_i$ középérték, $\sigma_i$ szórás), ezeket oly módon használjuk fel a lokális ploiditás meghatározásához, hogy a genomon egy mozgóátlagolásos módszerrel minden átfedő régióra (az ablakméret és eltolás állítható paraméterek) kiszámoljuk a $c_{avg}$ átlagos lefedettséget. A poszterior valószínűsége annak, hogy az adott régió $i$ ploiditású:

$$P(i|c_{avg}) = w_i \cdot \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{(c_{avg}-\mu_i)^2}{2\sigma_i^2}}$$

Ezt az értéket minden $i$ ploiditásra meghatározzuk, majd az adott régió összes genomi pozíciójára lejegyezzük azt az $i$-t, melyre a maximális értéket kaptuk. Végül egy adott genomi pozícióban a becsült ploiditiás az összes olyan régióra kapott becslés átlagaként áll elő, mely a pozíciót tartalmazza.

### A heterozigótaság elvesztésének (LOH) lokális becslése

A heterozigótaság elvesztéséről (LOH; loss of heterozygosity) egy diploid genom esetén például akkor beszélhetünk, ha egy kromoszómából valamilyen oknál fogva törlődik az anyai vagy az apai példány, és így az eredetileg heterozigóta genomi pozíciókban is csak egy típusú bázist találunk, vagyis a referencia bázis gyakorisága a kb. 0,5 helyett 0 vagy 1 lesz. Ezt a koncepciót általánosítva az IsoMut2py azokon a genomi régiókon detektál LOH-t, ahol nem igaz, hogy a jelenlévő $i$ példányból pontosan 1-ben tér el a detektált bázis a maradék $i-1$ példányban találttól. Tehát amennyiben több példány is található egy adott genomi régióból (ploiditás $\geq 1$), arra számítunk, hogy az ott megjelenő mutációk mindig csak egy példányt érintenek ezek közül. Amennyiben ez a feltételezés nem helyes, a régiót LOH régióként azonosítjuk.

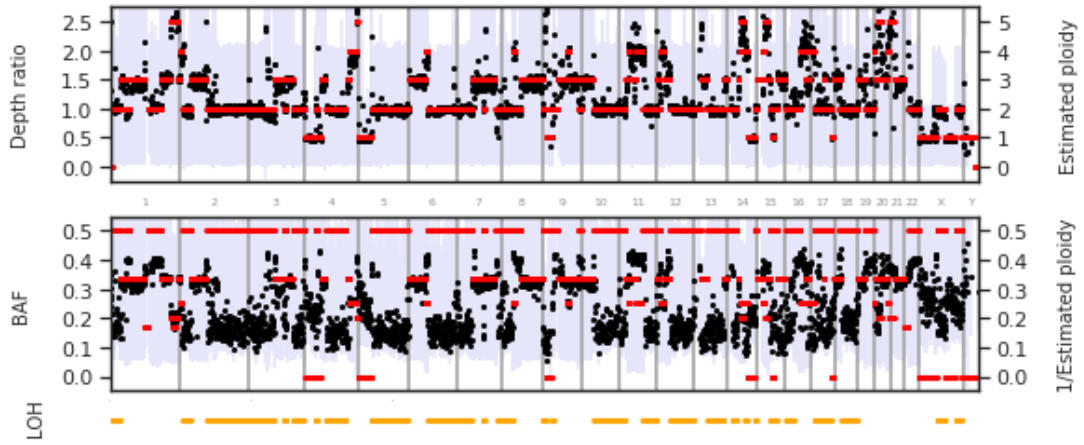
Mivel a ploiditás értéke feltétlenül egy egész szám, az elméletileg elképzelhető referencia allél frekvenciák is csak néhány diszkrét érték valamelyikét vehetik fel (a szimmetria miatt csak a legalább 1/2-et elérő értékeket tüntettük fel):

| **ploiditás** | **lehetséges allél frekvenciák ($AF_{elm} \geq 0,5$)** | | | |
|---|---|---|---|---|
| **1** | 1 | | | |
| **2** | 1/2 | 1 | | |
| **3** | 2/3 | 1 | | |
| **4** | 3/4 | 1/2 | 1 | |
| **5** | 4/5 | 3/5 | 1 | |
| **6** | 5/6 | 1/2 | 2/3 | 1 |
| **7** | 6/7 | 5/7 | 4/7 | 1 |

Miután a lokális ploiditást az adott genomi szakaszra megbecsültük, a nem haploid és nem triploid régiókon megvizsgáljuk, hogy a tapasztalt referencia allél gyakoriságok a

fenti $AF_{elm}$ elméleti allél frekvenciák közül melyikkel állnak a leginkább összhangban. Ez matematikailag azt jelenti, hogy a régióban tekintjük az összes olyan referencia allél frekvenciát, ami eléri az 1/2-es értéket, majd kiszámítjuk annak az együttes valószínűségét, hogy ezek mind az adott $AF_{elm}$-hez tartozó feltételezett eloszlásból ($\sigma = 2/c_{avg}$ normális, $AF_{elm} = 0,5$ és $AF_{elm} = 1$ esetén fél-normális eloszlás) származnak. Amennyiben nem a fenti táblázatban első helyen szereplő $AF_{elm}$ a legvalószínűbb, a régióban LOH-t detektálunk. A ploiditás becsléséhez hasonlóan a végső LOH értéke (igen/nem) az adott genomi pozíciót lefedő összes régióra becsült érték átlagaként áll elő.

A teljes genomra kapott ploiditás és LOH értékeket az IsoMut2py ábrázoló függvényeinek segítségével könnyen vizualizálhatjuk (10. ábra).




**10. ábra. A teljes genomra kapott becsült ploiditás és LOH értékek az IsoMut2py modul ábrázolásában.** A fenti panelen feketével jelennek meg a különböző régiókra mozgóátlagolással kapott lefedettség értékek és a diploid lefedettség ($2 \cdot \mu_1$) hányadosa („depth ratio”), a kék sávok pedig az értékek 25. és 75. percentilisét jelölik. A piros pontok az IsoMut2py által becsült ploiditást mutatják. (Azokat a régiókat, ahol a becslés sikertelen volt, nulla ploiditással jelezzük.) A lenti panelen ugyanezzel a mozgóátlagos megközelítéssel a fekete pontok a nem domináns allél frekvenciáját („BAF”) jelölik, a kék sávok ezen értékek 25. és 75. percentilisét, a piros pontok pedig a becsült ploiditás reciprokát. Kényelmi szempontokból a haploid régiókon a piros pontok értékének nullát választottunk. A narancssárgával jelölt régiókon az IsoMut2py LOH-t detektált.

### Különböző minták kariotípusának összevetése

Bár hangsúlyozzuk, hogy az IsoMut2py ploiditás becslő funkciója nem a hagyományos CNV detektáló eszközök feladatát igyekszik ellátni, egy naiv kariotípus összehasonlítást mégis implementáltunk. Ezt főként a vizsgált genomok kezdeti feltérképezésére javasoljuk, mintsem pontos CNV azonosításra. Ennek elsődleges oka, hogy a CNV-k azonosításánál helytelen nem kiaknázni a két mintából származó összes szekvenálási információt. Ezzel szemben az IsoMut2py ploiditás becslő algoritmusait arra optimalizáltuk, hogy egyetlen minta short readjeit használva viszonylag megbízható eredményeket kapjunk a ploiditásra vonatkozóan.

Ha az összevetni kívánt két mintára a fenti módon megbecsültük a lokális ploiditást, az összehasonlítást a *compare_with_other()* függvény segítségével végezhetjük el. Elsőként összegyűjtjük az összes olyan genomi régiót, melyeknél a becsült ploiditás a két mintában eltérő ($i_1 \neq i_2$). Ezek után minden ilyen régióban legyártjuk a két minta közös pileup fájlját, melyből az adott régióra meghatározzuk mindkét mintában az átlagos lefedettséget ($c^1_{avg}$ és $c^2_{avg}$). Felhasználva a két mintára kapott 7 komponensű Gauss keverék modell paramétereit ($w^1_i$, $\mu^1_i$, $\sigma^1_i$ és $w^2_i$, $\mu^2_i$, $\sigma^2_i$), mindkét mintában meghatározzuk, hogy mi annak a poszterior valószínűsége, hogy az adott régió $i_1$, illetve $i_2$ ploiditású:

$$P(i_m|c^n_{avg}) = w^n_{i_m} \cdot \frac{1}{\sqrt{2\pi[\sigma^n_{i_m}]^2}} e^{-\frac{(c_{avg}-\mu^n_{i_m})^2}{2[\sigma^n_{i_m}]^2}},$$

ahol $m \in \{1,2\}$ és $n \in \{1,2\}$ a mintákat jelöli.

Ezekből annak a valószínűsége, hogy a kérdéses genomi szakasz ploiditása a két mintában megegyezik vagy különbözik:

$$P(\text{azonos}) = \max\left\{P(i_1|c^1_{avg}) \cdot P(i_1|c^2_{avg}), P(i_2|c^1_{avg}) \cdot P(i_2|c^2_{avg})\right\}$$
$$P(\text{különböző}) = P(i_1|c^1_{avg}) \cdot P(i_2|c^2_{avg}))$$

Az adott régiót minőségileg jellemző érték a $P(\text{különböző})/P(\text{azonos})$, melyre egy tetszőleges minimális értéket beállítva kiválogathatjuk azokat a szakaszokat, melyek nagy valószínűséggel ténylegesen különböző ploiditásúak a két mintában. A szakaszok hosszára vonatkozó szűrési paraméter értékét változtatva a legrövidebb ilyen szakasz hosszát állíthatjuk be.

Amennyiben a vizsgált mintát pusztán egy bed fájllal kívánjuk összevetni, melyhez nem tartozik a ploiditás becsléséből származó paraméterszett és BAM fájl, a függvény egyszerűen, a minőséget jellemző fenti érték nélkül kilistázza azokat a régiókat, melyekben eltérést tapasztal.

### A szűrési paraméterek ploiditásfüggő, dinamikus adaptálása

Amennyiben a mutáció detektálás során egy jól ismert, konstans ploiditású, de nem diploid genomot kívánunk elemezni, lehetőség van a *constant_ploidy* paraméter értékének manuális átállítására (alapértelmezetten $constant\_ploidy = 2$).

Ha egy bonyolultabb szerkezetű genomot vizsgálunk és a fenti módon határoztuk meg a lokális ploiditást a különböző pozíciókban, a kapott eredményeket az IsoMut2py *generate_ploidy_info_file()* függvényének a segítségével átalakíthatjuk bed formátumba, melyet a mutációk azonosítása során fel tudunk használni [139]. Ha az elemzett minták kariotípusa (csoportonként) megegyezik, a fenti ploiditás becslést elég (csoportonként) egyetlen mintán lefuttatni, majd a mutáció detektálás során jelezni, hogy melyik minta melyik

ploiditás-csoporthoz tartozik.

Az olyan esetekben, mikor bár a genom alapvetően konstans ploiditású, de néhány régióról tudjuk, hogy duplikáció vagy deléció érte, ezeket manuálisan felsorolhatjuk a később felhasználni kívánt bed fájlban.

A lokális ploiditásra vonatkozó információt a mutációk azonosítása során úgy használjuk fel, hogy az eredeti, diploid genomokon értelmezett *sample_mut_freq_min* szűrési paraméter értékét a lokális $i$ ploiditás függvényében a következőre módosítjuk:

$$sample\_mut\_freq\_min(i) = sample\_mut\_freq\_min \cdot 2/i$$

#### KÖZÖS MUTÁCIÓK KERESÉSE

Mivel az eredeti IsoMut szoftver az egyedi mutációk keresésére lett optimalizálva, az IsoMut2py is ezt tekinti alapértelmezett beállításnak. Ebben a verzióban azonban lehetőség nyílik az olyan mutációk detektálására is, melyek több, mint egy mintában megjelennek, amennyiben az *unique_mutations_only* paraméter értékét *False*-ra állítjuk. Ezzel a futási idő ugyan megnövekszik, de a minták közti leszármazási kapcsolatok feltérképezhetővé válnak.

A közös mutációk detektálása során az egyedi mutációk keresésénél megismert alapelveket követjük: elvárjuk, hogy a valóban mutált mintákban megjelenő alternatív bázis frekvenciája kellően magas legyen (*sample_mut_freq_min(i)*), míg a mutációt nem hordozó minták legyenek megfelelően tiszták (*other_rnf_min*). A detektáló algoritmus kimenetében megjelenő, a mutációkat jellemző paramétereket minden genomi pozícióban a „legkevésbé mutált” és a „legkevésbé tiszta” minták tulajdonságai alapján határozzuk meg, vagyis a lehető legszigorúbb értékeket tüntetjük fel. Az $S$ érték meghatározásához tehát nullhipotézisként feltesszük, hogy a „legkevésbé mutált” mintában és a „legzajosabb” mintában a bázisgyakoriságok elméleti eloszlása megegyezik. A Fisher-féle p-érték annak a valószínűsége, hogy ilyen nullhipotézis mellett éppen a megfigyelt bázisokat tapasztaljuk a két mintában. Mivel $S = -\log p$, minél nagyobb $S$ értéke, annál biztosabbak lehetünk benne, hogy a „legkevésbé mutált” minta valóban mutált, míg a „legzajosabb” valójában is csak zajos. Azokat a pozíciókat, melyekben találunk olyan mintát, ami nem elég tiszta, de nem is eléggé mutált, kiszűrjük, mint zajos helyeket.

#### TOVÁBBI ELEMZÉSI LEHETŐSÉGEK

**Az $S$ érték optimalizálása**

Az IsoMut esetében javasolt $S$ értékre történő utólagos optimalizálást az IsoMut2py képes automatikusan is elvégezni. Ehhez elsőként meg kell adnunk azoknak a kontroll mintáknak a listáját, melyekben nem számítunk egyedi mutációk megjelenésére. A fentiekben leírtaknak megfelelően ezek lehetnek a kísérletek kezdőklónjai vagy többszörösen megszekvenált DNS preparátumok. Amennyiben normál-tumor mintapárokkal dolgozunk,

a normál mintákat tekinthetjük kontrolloknak, hiszen ezekben szomatikus mutáció nem várható. Emellett egy hozzávetőleges becslést kell tennünk arra, hogy genomonként hány $FP_{max}$ fals pozitív találatot tartunk elfogadhatónak. Ezzel a beállítással $FP_{max}$ mennyiségű fals pozitív SNV-t, inzerciót és deléciót engedünk meg a vizsgált genomon. Ezt követően az *optimize_results()* függvény különböző $S$ értékek mellett meghatározza a kontroll mintákban és a többi mintában található egyedi mutációk számát, végül pedig egy olyan küszöbértéket választ $S$-re, mellyel a kontroll mintákban nem lépjük túl az előre definiált fals pozitívok számát, de a többi mintában a lehető legtöbb találatot kapjuk.

Az optimalizálást ploiditásonként és mutáció típusonként külön-külön végezzük el. Az adott ploiditásra jutó $FP_i$ fals pozitívok maximális számát alapvetően az adott ploiditású régiók teljes $L_i$ hosszának és a genom $L$ hosszának hányadosával arányosan kellene megválasztani. Annak érdekében azonban, hogy időt spóroljunk, az $L_i/L$ hányadost az adott $m$ mutáció típusból ($m \in \{\text{SNV, inzerció, deléció}\}$) az adott $i$ ploiditású régiókon detektáltak $N_{m,i}$ darabszámának és az adott mutáció típusból összesen azonosítottak $N_m$ darabszámának hányadosával közelítjük:

$$\mathrm{FP}_{i,m} \simeq \frac{\mathrm{FP}_{max} \cdot N_{m,i}}{N_m}$$

Ezzel az optimalizálást követően mutáció típusonként és ploiditásonként különböző $S^{\min}_{m,i}$ küszöbértékeket kapunk.

### Kérdéses mutációk további vizsgálata

Mivel még az optimalizálást követően is előfordulhat, hogy a mutációk végső listájában gyanús helyeket találunk, a *check_pileup()* függvénnyel manuálisan lekérdezhetjük az adott genomi pozícióban (vagy azok listájában) a szekvenálási adatokat, így végül ezekre alapozva akár új szűrési paraméterek bevezetésével újra válogathatjuk a mutációkat. Bár a pileup formátum gyakran nehezen átlátható, a fenti függvény kimenete a pileupban lévő információkat könnyen interpretálható, de adatvesztéssel nem járó formátumba kondenzálja.

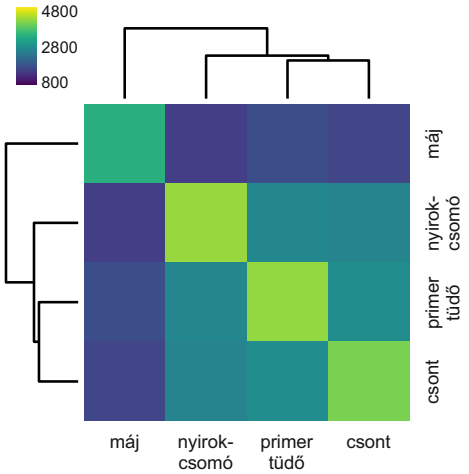
### Mutációk vizualizálása, szomatikus mutációk spektruma és dekompozíciója

Mihelyst a mutációk listáját a fenti módon véglegesítettük, a biológiai interpretációhoz fontos lehet az eredmények vizualizálása is. A minták összehasonlításához az egyik leghasznosabb ábrázolási lehetőség a mintákban talált mutációk darabszámának felrajzolása. Ezt az IsoMut2py mind az egyedi, mind az összes mutációra vonatkozóan egy oszlopdiagramon jeleníti meg a *plot_mutation_counts()* függvény segítségével.

A mutációk genomi pozíciójának vizsgálatára az egyik látványos módszer az ún. rainfall ábrák generálása, melyeken a vízszintes tengelyen a mutáció genomi pozíciója szerepel, a függőlegesen pedig az adott mutációnak a genomi távolsága az őt megelőző mu-

tációtól (11. ábra). Amennyiben a mutációk hajlamosak klaszterezőődni, az egy ilyen ábrán egyrészt láthatóvá válik a pontok vízszintes irányú besűrűsödéséből, másrészt pedig a függőleges tengelyen felvett alacsony értékekből. Emellett, ha a vizsgált mintában sok a dinukleotid mutáció (DNV), az a rainfall ábrán a függőleges (logaritmikus) tengely $10^0$ szintjénél rajzolódik ki.


**11. ábra. A detektált mutációk rainfall ábrája.** A különböző színek a mutációk típusát kódolják (INS: inzerció, DEL: deléció, SNV-k esetén a bázisváltozás alapján kategorizálva). A vízszintes tengelyen a mutáció abszolút genomi pozíciója látható (a megjelenített kromoszómasorrend mellett), a függőleges szürke vonalak a kromoszómák határát jelölik. A függőleges tengelyen logaritmikus skálán az adott mutációnak az előző variánstól számított genomi távolságát ábrázoltuk. A 9. kromoszóma közepe táján például az ábra lenti részére tömörülő sűrű ponthalmaz a mutációk lokális klaszterezőődését jelzi.




**12. ábra. A vizsgált minták hierarchikus klaszterezése a detektált mutációk száma alapján.** A hőtérkép azt ábrázolja, hogy hány olyan mutáció volt egy adott mintapár esetén, mely mindkét mintában megtalálható volt. Az átlóban az adott mintában talált összes mutáció darabszáma látható.

Az IsoMut2py modulba beépítettünk emellett egy naiv, hierarchikus klaszterezést folytató függvényt, mely arra alapozva, hogy a különböző minták között páronként hány közös mutáció volt, létrehozza a minták dendrogramját, melyből a köztük lévő leszármazási kapcsolatokra lehet következtetni (12. ábra).

A szomatikus mutációk statisztikus tulajdonságainak a vizsgálatára az egyik legelterjedtebb módszer a fent részletezett mutációs spektrumok felrajzolása. Az IsoMut2py modul a [115] tanulmány által közzétett referencia szignatúrákat használja és a *decompose_SNV_spectra()*, *decompose_DNV_spectra()* és *decompose_indel_spectra()* függvények ugyanezeken a bázisokon ábrázolják mintánként a detektált szomatikus mutációkat (13. ábra). (Lehetőség van az összes mutáció együttes spektrumának ábrázolására is, mivel azonban a referencia szignatúrákat szomatikus mutációs katalógusok alapján határozták meg, így az alapértelmezett beállítás is csak ezeket a mutációkat tekinti.) A DNV-spektrum esetében a hagyományos „lineáris” ábrázolás helyett az eredményeket egy mátrixon értelmezett hőtérképen is megjeleníthetjük (B6. ábra).


**13. ábra. A detektált mutációk spektrumai az IsoMut2py ábrázolásában. a.** Az egybázisos változások eloszlása a genomi környezettől függően (SNV-spektrum). **b.** Az egymás melletti bázisokat érintő dupla mutációk eloszlása az eredeti bázispár és a mutált bázispár szerint (DNV-spektrum). **c.** Az indelek eloszlása az indel hossza és a genomi környezet függvényében.

A mintákban megfigyelhető mutációs eloszlások esetén gyakran felmerül a kérdés, hogy hogy vezethetjük vissza az eredményeket a referencia szignatúrák különböző járu-

lékaira. Ehhez az IsoMut2py egy a [140] kutatás által inspirált EM (expectation maximization $\sim$ várható értéket maximalizáló) algoritmust implementál. Az elemzés során az adott minta spektrumának 1-re normált verziójával dolgozunk, elsősorban azért, mert ugyanez igaz a referencia szignatúrákra is.

Első lépésben a referencia szignatúrák kezdeti arányait inicializáljuk az adott mintában. Ehhez kiszámoljuk a minta $s$ spektruma és az összes $r_i$ szignatúra közti koszinusz hasonlóságot az alábbi módon:

$$S_i^{\cos} = 1 - \frac{s \cdot r_i}{|s| \cdot |r_i|}$$

A kezdeti $\theta_i$ szignatúra arányokat alapértelmezetten a koszinusz hasonlóságokkal arányosan vesszük fel, de lehetséges az egyenlő súlyfaktorok beállítása is. Mindkét esetben $\sum_i \theta_i = 1$.

Ezek után a spektrum összes $n$ mutációjára meghatározzuk, hogy mekkora $\delta_{n,i}$ valószínűségekkel származik egy adott $r_i$ referencia szignatúrától. Mivel a szignatúrák 1-re normáltak, ehhez egyszerűen meghatározzuk az adott mutáció $k_n$ típusát (bázisváltás és genomi környezet szerint) és megnézzük a szignatúrában az adott komponens súlyát:

$$\delta_{n,i} = r_{ik_n}$$

Az EM algoritmus iterációi során, az E-lépésben kiszámoljuk azt a $z_{n,i}$ feltételes valószínűséget, hogy az aktuális $\theta_i$ szignatúra súlyfaktorok mellett mekkora valószínűsége van annak, hogy az adott $n$ mutáció éppen az $i$. szignatúrából származik:

$$z_{n,i} = \frac{\theta_i \tilde{\delta}_{n,i}}{\sum_{i'} \theta_{i'} \tilde{\delta}_{n,i'}}$$

Ezt követően az M-lépésben frissítjük a korábbi $\theta_i$ értékeket a $z_{n,i}$-k alapján:

$$\theta_i = \frac{\sum_n z_{n,i}}{\sum_{i'} \sum_n z_{n,i'}}$$

Ezt az eljárást addig ismételjük, míg a $\theta_i$ értékekre kapott korrekció egy küszöbérték alá nem esik.

Miután megkaptuk a szignatúrák várható súlyát, egy utólagos szűrésnek vetjük alá őket. Elsőként minden mutációt besorolunk a $z_{n,i}$ mátrix alapján annak a szignatúrának a járulékába, amelyikből a legvalószínűbben származik. Ezt követően az összes szignatúra közül eldobjuk azokat, melyek a minta spektrumához nem járulnak hozzá legalább *filter_count* darab mutációval, illetve a mutációk legalább *filter_percent* százalékával. Egy további szűrési beállítással (*keep_top_n*) azt is megtehetjük, hogy csak az $n$ legdominánsabb (legtöbb mutációt adó) szignatúrát tartjuk meg. Ezek után a teljes EM-algoritmust újrafuttatjuk az

így megritkított szignatúrák halmazán.

Végül a különböző szignatúrák járulékait mintánként külön-külön oszlopdiagramon ábrázoljuk.

### Mutációs listák importálása

Előfordulhat, hogy az IsoMut2py optimalizációs, ábrázolási és dekompozíciós függvényeit egy meglévő mutációs halmazon szeretnénk használni, melyeket egy másik mutáció detektáló szoftverrel azonosítottunk. Amennyiben ezt az adathalmazt egy megfelelő formátumú pandas adattáblává tudjuk alakítani, az IsoMut2py összes függvénye alkalmazható lesz rá. A leggyakrabban a mutációk listáját VCF formátumban hozzák létre a különböző számítógépes eszközök, ezeket a pyvcf Python modul segítségével könnyen beolvashatjuk, majd néhány egyszerű átalakítással a megfelelő formátumra konvertálhatjuk (lásd https://isomut2py.readthedocs.io/en/latest/external_mutations.html).

# GENOMIKAI BIOMARKEREK: NGS TECHNIKÁVAL A DAGANATOS BETEGSÉGEK NYOMÁBAN

## A BRCA1 ÉS BRCA2 GÉNEK HIÁNYÁNAK HATÁSA A MUTÁCIÓS SPEKTRUMRA

A daganatos sejtekben igen gyakran előforduló genomi instabilitás fő okozója általában valamilyen DNS-javító mechanizmus hibája. A homológ rekombinációban aktívan résztvevő BRCA1 és BRCA2 fehérjéket kódoló gének öröklődő, csíravonal mutációi például növelik a petefészek- és az emlőrák kialakulásának kockázatát [141, 142]. A BRCA1/2-hiányos tumorokban számos kutatás során megfigyelhetők voltak a genomi instabilitás karakterisztikus, nagyskálás jegyei [143–145], ugyanakkor a pontmutációk spektrumában okozott specifikus változások feltérképezetlenek maradtak. Bár a mutációs szignatúrák megalkotásakor [112] sikerült azonosítani olyan mutációs folyamatok lenyomatát, melyek korreláltak a BRCA1/2 gének rendellenes működésével, mivel azonban egy daganaton belül számos genomi hiba halmozott hatása is megjelenhet, a ténylegesen BRCA1/2-specifikus mutációs spektrum vizsgálatához optimalizáltabb kísérleti elrendezésre van szükség.

Ennek érdekében az MTA Enzimológiai Intézetének Genom Stabilitás Kutatócsoportjával közösen végzett kutatásunk [146] során arra kerestük a választ, hogy mennyiben más a BRCA1/2-hibás sejtek mutációs spektruma, mint a vadtípusú sejteké normál körülmények között, illetve a DNS-mutációk számát erősen növelő metil-metánszulfonát (MMS) kezelés [147, 148] hatására. A szer elsősorban a DNS metilációját segíti elő, főként az adenin és guanin bázisoknál. A MeG nem akasztja meg a DNS-szintézist, és vele szemben ténylegesen a helyes citozin bázis épül be, de a MeA-nel szemben gyakran nem T szintetizálódik. Emellett a metiláció mindkét bázist igen instabillá teszi, melynek a következményeként gyakran abázikus helyek, majd esetlegesen kettős DNS-száltörések is kialakulhatnak. Ezeknek a javításában a homológ rekombináció - amennyiben működik - aktív szerepet kap. A HR hiánya okán bekapcsolódó NHEJ útvonal várhatóan növeli a rövid deléciók számát, de a pontmutációk tulajdonságaira vonatkozóan nem voltak korábbi eredmények.

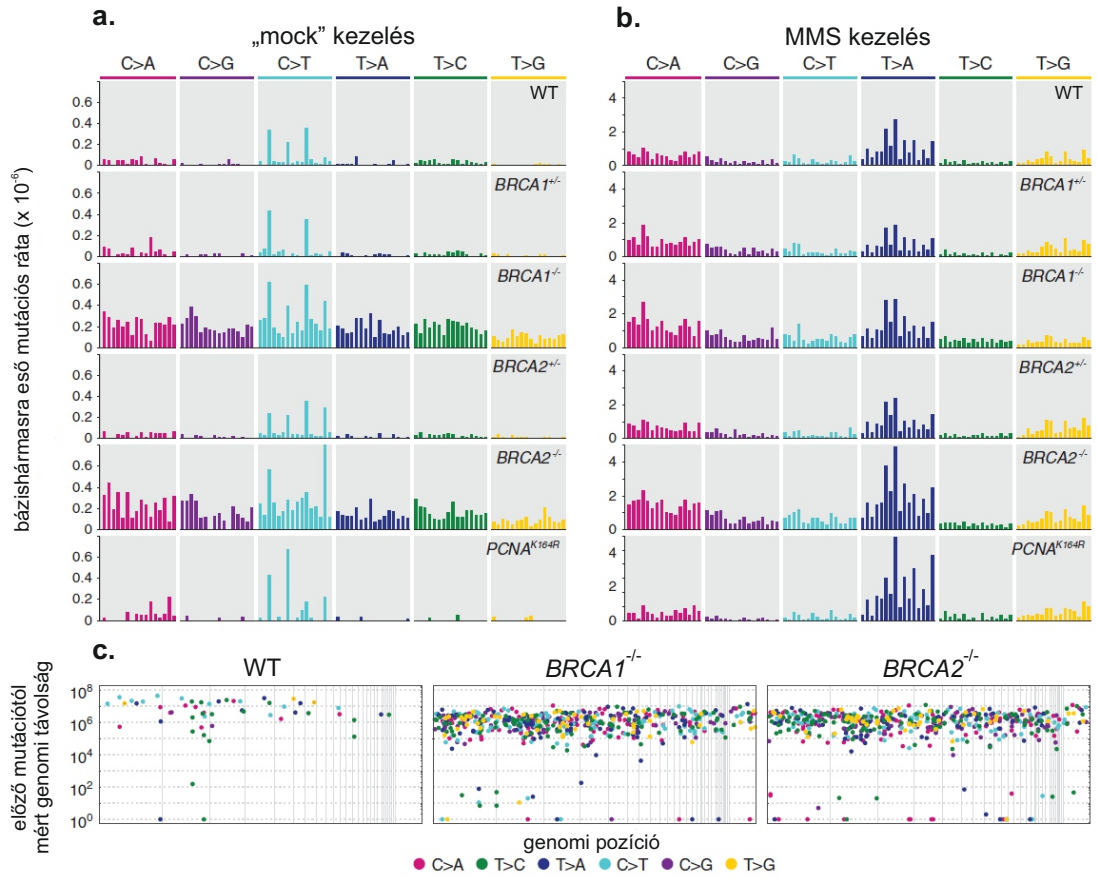
A kísérletek során az IsoMut optimalizálásához is használt DT40 sejtvonalat vizsgáltuk, részben mert genomja harmad olyan hosszú, mint a humán genom, illetve igen gyakran alkalmazzák DNS-javító mechanizmusok modellezése során [149], továbbá számos olyan izogenikus mutáns sejtvonala elérhető, melyekben célzottan valamilyen DNS-javításért felelős gént blokkoltak. A mérésekhez minden elemzett sejtvonalat (vadtípus (WT), BRCA1 génben heterozigóta mutációt hordozó ($BRCA1^{+/-}$), BRCA1 génben homozigóta mutációt hordozó ($BRCA1^{-/-}$), BRCA2 génben heterozigóta mutációt hordozó ($BRCA2^{+/-}$), BRCA2 génben homozigóta mutációt hordozó ($BRCA2^{-/-}$), mutáns PCNA gént hordozó ($PCNA^{K164R}$)) háromféle „kezelésnek” vetettük alá: kezelés- és szaporításmentes eset (kontroll kezdőklón), kezelésmentes, de szaporított eset („mock treatment”: álkezelés), il-

letve a tényleges MMS-kezelt eset. A kezdőklónokból egyet, az álkezelt és MMS-kezelt klónokból kettőt vagy hármat szekvenáltunk mutáns sejtvonalanként. Ez alól egyedül a $PCNA^{K164R}$ sejtvonal volt kivétel, melyből minden kezelést tekintve egy klónt szekvenáltunk (3. táblázat). A teljes genom szekvenálás előtt minden alkalommal a kezelt populáció egyetlen sejtjét szaporítottuk fel, ezzel homogén genomi állományt hoztunk létre.

| **Genotípus és kezelés** | $n$ | **SNV** | **inzerció** | **deléció** |
|---|---|---|---|---|
| **WT** | | | | |
| kezdőklón | 1 | 4 | 0 | 0 |
| „mock” kezelés | 3 | $72 \pm 5$ | $4{,}7 \pm 1{,}5$ | $1{,}7 \pm 0{,}6$ |
| MMS | 3 | $1489 \pm 620$ | $5{,}5 \pm 2{,}5$ | $6{,}0 \pm 2{,}1$ |
| $BRCA1^{+/-}$ | | | | |
| kezdőklón | 1 | 5 | 0 | 0 |
| „mock” kezelés | 2 | $63 \pm 14$ | $4{,}0 \pm 1{,}3$ | $2{,}5 \pm 1{,}5$ |
| MMS | 3 | $1582 \pm 840$ | $3{,}0 \pm 1{,}7$ | $9{,}3 \pm 3{,}5$ |
| $BRCA1^{-/-}$ | | | | |
| kezdőklón | 1 | 1 | 1 | 0 |
| „mock” kezelés | 3 | $562 \pm 75$ | $8{,}0 \pm 1{,}0$ | $12{,}7 \pm 1{,}2$ |
| MMS | 3 | $2414 \pm 201$ | $7{,}3 \pm 2{,}1$ | $24{,}7 \pm 6{,}5$ |
| $BRCA2^{+/-}$ | | | | |
| kezdőklón | 1 | 2 | 0 | 1 |
| „mock” kezelés | 3 | $79 \pm 13$ | $2{,}7 \pm 1{,}2$ | $2{,}0 \pm 2{,}0$ |
| MMS | 2 | $1629 \pm 88$ | $3{,}5 \pm 2{,}1$ | $9{,}0 \pm 4{,}2$ |
| $BRCA2^{-/-}$ | | | | |
| kezdőklón | 1 | 2 | 0 | 0 |
| „mock” kezelés | 3 | $511 \pm 21$ | $10{,}3 \pm 3{,}1$ | $33{,}0 \pm 5{,}0$ |
| MMS | 3 | $2986 \pm 324$ | $11{,}7 \pm 4{,}2$ | $40{,}3 \pm 2{,}1$ |
| $PCNA^{K164R}$ | | | | |
| kezdőklón | 1 | 1 | 0 | 0 |
| „mock” kezelés | 1 | 43 | 5 | 7 |
| MMS | 1 | 2286 | 3 | 6 |

**3. táblázat. A vizsgált mintákban detektált mutációk száma.** A kezdőklónokban talált mutációk fals pozitívok. ($n$: az azonos genotípusú és azonos kezelést kapott klónok száma; SNV: a mintákban detektált pontmutációk átlagos száma és szórása; inzerció: a mintákban detektált inzerciók átlagos száma és szórása; deléció: a mintákban detektált deléciók átlagos száma és szórása)

Az elemzés során a mintákra jellemző egyedi mutációk detektálására fektettük a hangsúlyt, melyeket az IsoMut eredeti verziójával detektáltunk. Az $S$ érték optimalizálásakor a kezdőklónokban maximálisan 5 fals pozitív SNV megjelenését engedélyeztük (3. táblázat). A kapott mutációs listából létrehoztuk az SNV-spektrumokat, illetve a rainfall ábrákat (14. ábra). A vad típusú mintáknál azt találtuk, hogy a „mock” kezelés mellett létrejövő spontán mutációk spektrumában leginkább a CpG bázispároknál megjelenő C>T bázisváltások a gyakoriak, ami a gerinces genomokban általában is gyakran megfigyelhető jelenség lenyomata, mikor a metilált citozinok deaminálódás során timinné alakulnak [150]. Vagyis a vizsgált DT40 sejtvonal hűen visszaadja a spontán mutációk esetében várt tendenciákat.




**14. ábra. A különböző sejtvonalak különböző kezelésnek alávetett klónjaiban detektált SNV-k spektruma és genomi eloszlása.** A detektált SNV-k triplett-spektrumai az álkezelésnek („mock" kezelés) (**a.**) és az MMS kezelésnek (**b.**) alávetett sejtvonalakra. Bár a színkódok eltérőek, a vízszintes tengely azonos a 13a. ábra vízszintes tengelyével. Annak érdekében, hogy a humán mintákkal összevethető eredményeket kapjunk, a mutációs ráta kiszámításához a mutációk darabszámát normáltuk az adott triplett előfordulásával a DT40 genomban. **c.** A mutációk genomi helyzetének és az őket megelőző mutációk genomi távolságának ábrázolása rainfall ábrán. A színkódok a bázisváltást jelölik. Minden ábrán a sejtvonal egy adott reprezentatív klónjából származó adatok szerepelnek. (WT: vad típus)

Ezzel szemben a BRCA1/2 homozigóta mutáns sejtvonalakban „mock" kezelés hatására hét-nyolcszorosára emelkedett mutációs rátát tapasztaltunk, de a spektrum alapvető sajátosságai nem változtak, vagyis az említett gének hiányában a spontán mutációk száma a lokális szekvenciális környezettől függetlenül, homogén módon, klasztereződés nélkül (14c. ábra) növekedett. Hasonló tendenciát a heterozigóta mutánsoknál nem figyeltünk meg.

Az MMS kezelés hatására az összes vizsgált genotípusnál jelentősen megugrott a detektált mutációk száma. A vad típusú klónok esetén megfigyelhető mutációs spektrum (14b. ábra) elsőként mutatja be az MMS konkrét DNS-rongáló hatását. A mutációk száma több mint hússzorosára nőtt a „mock" kezelésen átesett klónokhoz képest, elsősorban a T>A mutációs csúcsok erősödésével. Ebből arra lehet következtetni, hogy az MMS főként

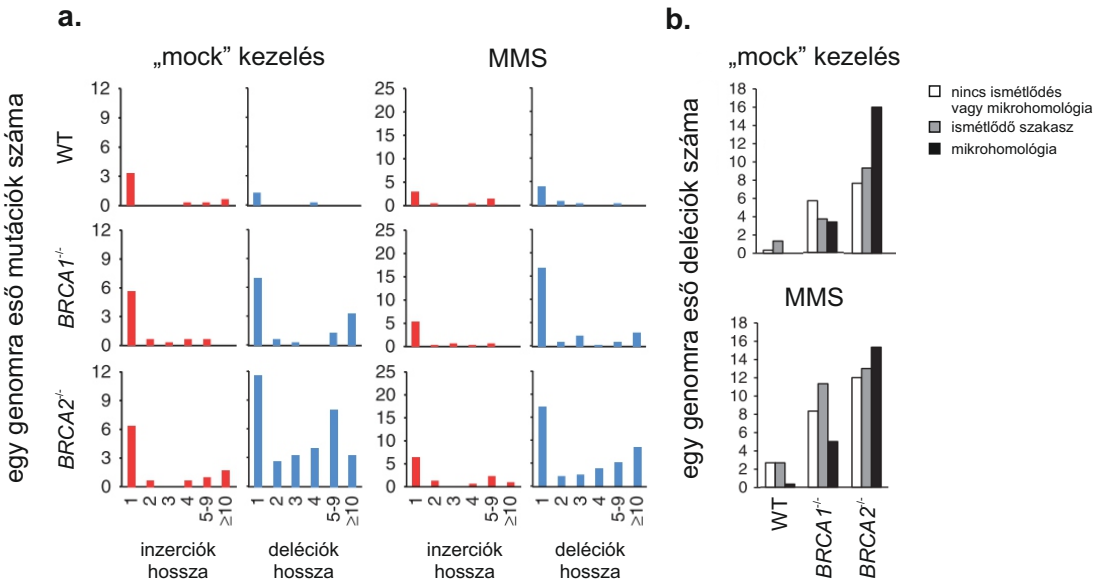
az adenin metilálása útján létrejövő abázikus helyeken fejti ki DNS-roncsoló hatását. A szerre továbbá rendkívül érzékenynek bizonyultak a homozigóta mutáns klónok, bennük ugyanis 2-3-szor annyi mutáció jelentkezett a kezelés hatására, mint a vad típusú vagy a heterozigóta mutáns mintákban. Érdekes módon a spektrumok alakja között azonban nem találtunk számottevő különbséget, ami arra utal, hogy a BRCA1/2 gének hiánya a mutációs folyamat alapvető mechanizmusát nem befolyásolja, csak azt, hogy milyen gyakran történik meg.

Annak a vizsgálatára, hogy az MMS kezelés okozta mutációs spektrumban milyen szerepet játszanak a különböző transzléziós DNS-javító útvonalak, a BRCA1/2 génkiütött minták mellett egy olyan genotípusú mintát is megszekvenáltunk, melyben a PCNA fehérje funkcionálisan roncsolódott. A PCNA mintegy aktiválófaktorként működik az egyik típusú tolerancia útvonal során. Amennyiben a PCNA mono-ubikvitinálódik, az ún. transzléziós útvonal (TLS: transzléziós szintézis) aktiválódik, mely egy olyan enzimre (DNS-polimeráz delta) cseréli a primer DNS-polimerázt, ami képes a hibás szakaszon is keresztülhaladni. Ha a PCNA poli-ubikvitinálódik, egy hasonló, de a DNS-károsodást kevesebb tévesztéssel javító mechanizmus lép életbe. A $PCNA^{K164R}$ mintákban a PCNA képtelen az ubikvitinációra, így minden PCNA-ubikvitinációfüggő transzléziós útvonal gátlódik, a szerepüket a REV1 fehérje által aktivált polimerázok veszik át. Az elemzés szerint ennek hatására a spontán mutációs ráta ugyan nem változik jelentősen (3. táblázat), de az MMS kezelést követő mutációs spektrumban a T>A mutációk szinte kizárólagosan dominálnak. Ennek és a WT, illetve BRCA1/2 mutáns mintákban megfigyelt spektrumnak a hasonlósága arra enged következtetni, hogy a TLS fontos szerepet játszik az MMS okozta mutációs mintázat kialakításában.

A BRCA gének mutációit korábban összefüggésbe hozták a tumorgenomokban megjelenő megnövekedett indel mutációs rátával [151]. Az elemzésünk során azt találtuk, hogy a spontán megjelenő inzerciók és deléciók száma szignifikánsan emelkedett a két homozigóta mutáns sejtvonalban a vad típushoz képest (3. táblázat): a $BRCA1^{-/-}$ klónokban négyszeres, a $BRCA2^{-/-}$ klónokban pedig nyolcszoros növekedést tapasztaltunk. Emellett a deléciók tekintetében a $BRCA2^{-/-}$ sejtvonal még a $BRCA1^{-/-}$ sejtvonalnál mérhetőnél is szignifikánsan több mutációval rendelkezett. Ugyanezek a tendenciák az MMS kezelés hatására is megmaradtak, bár összességében emelkedett mutációs számokkal. Az indelek hosszeloszlását vizsgálva (15a. ábra) arra lettünk figyelmesek, hogy bár a WT mintákra tipikusan a nagyon rövid (egy bázisos) indelek voltak jellemzőek, a homozigóta BRCA mutáns klónokban jóval hosszabb indelek is gyakran megjelentek. A $BRCA1^{-/-}$ mutánsoknál ez főként a delécióknál volt megfigyelhető, ahol az egybázisos deléciók mellett a 10-nél hosszabb deléciókat is sűrűn detektáltunk. Ezzel szemben a $BRCA2^{-/-}$ klónok esetében a deléciók hossz szerinti eloszlása kifejezetten szélessé vált. Ezek a jellegzetességek mind a „mock” kezelés, mind pedig az MMS kezelés hatására megmaradtak. A deléciókat továbbá a lokális szekvenciális környezet szerint három csoportra osztottuk. Ahol a

deléció egy repetitív, vagyis ismétlődő genomi szakaszra esett, szürkével jelöltük a 15b. ábrán. Ahol a deléció környezetében mikrohomológia (vagyis rövid, részlegesen átfedő szakaszok a deléció és a szomszédos bázisok között) volt megfigyelhető, azt az ábrán feketével tüntettük fel. Az egyik kategóriába sem illő deléciók pedig fehérrel szerepelnek. A vad típusú mintákhoz képest mindkét homozigóta mutánsban emelkedett számban jelentek meg a mikrohomológiát tartalmazó deléciók, különösen a *BRCA2*$^{-/-}$ klónok esetén. Az ilyen típusú mutációk rendkívül jellemzőek akkor, amikor a kettős DNS-száltöréseket az NHEJ DNS-javító mechanizmus, pontosabban ennek a mikrohomológia-közvetített verziója (MMEJ) korrigálja [152].

Mindezek a megfigyelések (a deléciók számának megnövekedése, a kiszélesedő hosszeloszlás, illetve a mikrohomológiát tartalmazó deléciók emelkedett száma) alapján arra következtethetünk, hogy a BRCA2 gén hiányában a kettős DNS-száltörések javítását elsődlegesen az NHEJ mechanizmus végzi, míg a BRCA1 roncsolása esetén más javítási folyamatok is aktívak lehetnek.




**15. ábra. A különböző sejtvonalak különböző kezelésnek alávetett klónjaiban detektált indelek hosszeloszlása. a.** A detektált indelek hossz szerinti eloszlása. **b.** A detektált deléciók eloszlása a szekvenciális környezettől függően. (WT: vad típus)

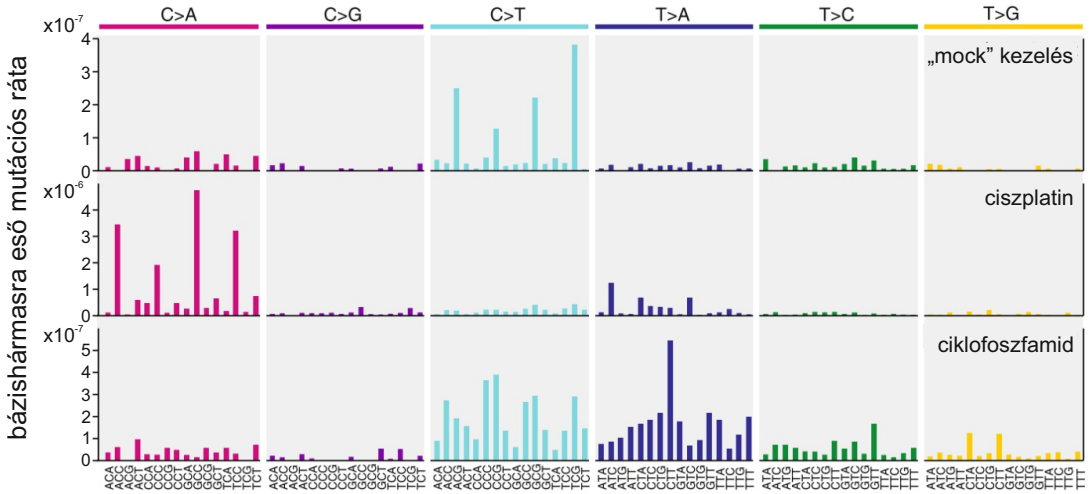
### A leggyakoribb kemoterápiás szerek mutagén hatásainak feltérképezése

A citotoxikus kemoterápiás szereket az 1950-es évek óta alkalmazzák a rákgyógyításban, és a mai napig a legtöbb daganat esetében ez az elsődleges kezelés. Ezek a szerek különböző mechanizmusok útján gátolják a sejtek szaporodását, például direkt DNS-károsodásokat okoznak, megzavarják a DNS-anyagcserét vagy a mitotikus apparátust. Ennek a folyamatnak a tumorsejtek károsításán túl az egészséges szövetre nézve is lehetnek mellékhatásai, illetve a kezelés okozta genomi mutációk rezisztencia kialakulásához, vagy

másodlagos tumor létrejöttéhez vezethetnek. Ezért nagyon fontos felmérni, hogy a különböző típusú gyógyszerek pontosan milyen módon károsítják a DNS-t, és számolni kell-e hosszútávú következményekkel. Ennek a kérdésnek a feltérképezésére az MTA Enzimológiai Intézetének Genom Stabilitás Kutatócsoportjával közösen végzett kutatásunkban [153] nyolc gyakran alkalmazott citotoxikus szer (4. táblázat) mutációs profilját vizsgáltuk a DT40 sejtvonalon.

| Kezelés | $n$ | SNV | inzerció | deléció |
|---|---|---|---|---|
| kezdőklón | 1 | 0 | 0 | 0 |
| „mock” kezelés | 4 | $47 \pm 20$ | $4{,}5 \pm 1{,}3$ | $3{,}0 \pm 0{,}8$ |
| ciszplatin | 3 | $812 \pm 193$ | $49{,}0 \pm 15{,}0$ | $83{,}0 \pm 21{,}2$ |
| ciklofoszfamid | 3 | $245 \pm 50$ | $3{,}0 \pm 1{,}7$ | $5{,}0 \pm 1{,}7$ |
| hidroxikarbamid | 3 | $74 \pm 9$ | $4{,}7 \pm 1{,}5$ | $3{,}7 \pm 1{,}2$ |
| gemcitabin | 3 | $57 \pm 31$ | $2{,}3 \pm 2{,}1$ | $3{,}0 \pm 1{,}7$ |
| 5-fluorouracil | 3 | $50 \pm 16$ | $3{,}0 \pm 1{,}0$ | $2{,}7 \pm 1{,}5$ |
| etopozid | 3 | $95 \pm 15$ | $3{,}7 \pm 1{,}2$ | $6{,}7 \pm 4{,}2$ |
| doxorubicin | 3 | $44 \pm 15$ | $3{,}3 \pm 0{,}6$ | $4{,}0 \pm 4{,}6$ |
| paklitaxel | 3 | $64 \pm 10$ | $1{,}0 \pm 1{,}0$ | $3{,}0 \pm 0{,}0$ |

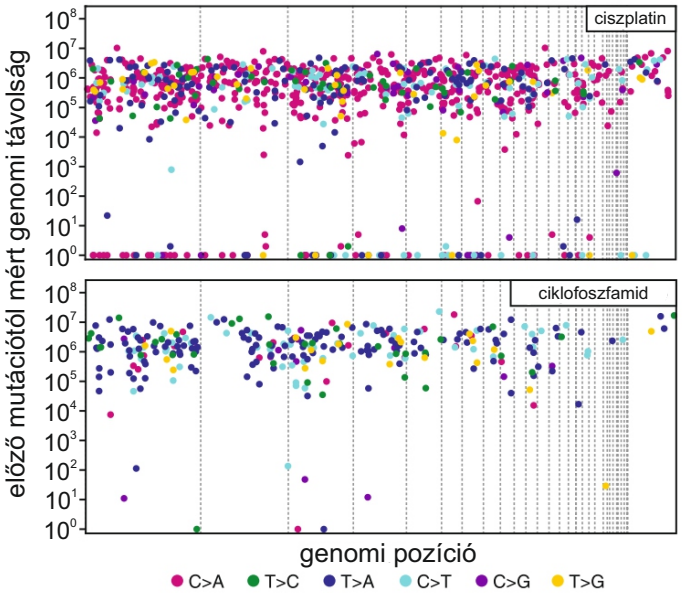
**4. táblázat. A nyolcféle kemoterápiás szerrel kezelt klónokban talált mutációk száma.** A kezdőklónban talált mutációk fals pozitívok. ($n$: az azonos kezelést kapott klónok száma; SNV: a mintákban detektált pontmutációk átlagos száma és szórása; inzerció: a mintákban detektált inzerciók átlagos száma és szórása; deléció: a mintákban detektált deléciók átlagos száma és szórása)



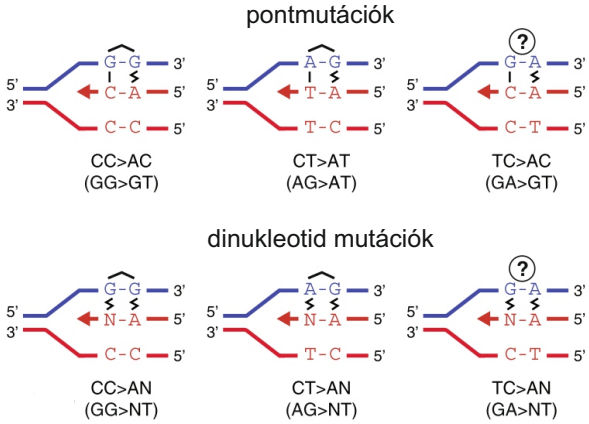

**16. ábra. A különböző kezelésnek alávetett klónokban detektált SNV-k spektruma.**

A minták preparálása és kezelése a BRCA génhiány hatásának vizsgálata során történtekkel azonosan zajlott. A klónokban található egyedi mutációk detektálásához az IsoMut eredeti verzióját használtuk. A „mock” kezeléshez képest a hidroxikarbamid, gemcitabin, 5-fluorouracil, doxorubicin és paklitaxel szerek nem okoztak szignifikáns növekedést

a mutációs rátában (4. táblázat). Az etopozid ugyan szignifikánsan megemelte a pontmutációk számát, de a spektrumukban nem okozott jelentős változást a „mock” kezeléshez képest. Ezzel szemben mind a ciszplatin, mind pedig a ciklofoszfamid hatására nagymértékben (rendre 17-szeresére és 5-szörösére) emelkedett az SNV-k száma és a spektrumban is jellegzetes, a „mock” kezeléstől eltérő csúcsok jelentek meg (16. ábra). A ciszplatin emellett az indelek számát is számottevően megnövelte (4. táblázat).




**17. ábra. A pontmutációk genomi helyzetének és az őket megelőző mutációk genomi távolságának ábrázolása rainfall ábrán.**




**18. ábra. A ciszplatin okozta szálon belüli kötések mentén létrejövő SNV-k és DNV-k.** A kérdőjel az eddig nem ismert GA dinukleotidoknál létrejövő kötést jelöli.

A ciszplatin esetében a spektrumon elsősorban a C>A, ezen belül pedig a főként a NCC>NAC, NCT>NAT és NTC>NAC csúcsok domináltak. Mivel a ciszplatin okozta genomi eltérések többsége az egy szálon lévő szomszédos purinbázisok között kialakított

kötésként realizálódik [154, 155], ezek az SNV-k olyan mutációkat reprezentálhatnak, melyek a kötött GG, AG és GA dinukleotidok 3'-as oldalon lévő bázisához rendelt helytelen komplementer bázis miatt jönnek létre (18. ábra). Mindazonáltal a ciszplatin okozta GA szálon belüli kötésekről korábbi kutatások nem számoltak be, ezért az NTC>NAC mutációkat tovább csoportosítottuk a szekvenciális környezetük szerint. Azt találtuk, hogy a mutációk jelentős része NTCC > NACC vagy NTCT > NACT típusú volt, ami arra enged következtetni, hogy a GG és AG kötött dinukleotiddal a 3' oldalon közvetlenül szomszédos bázis is hajlamos a mutációra (18. ábra). A maradék NTC>NAC mutációk többségénél azonban a genomi környezetet megvizsgálva az egyedüli potenciális bipurin kapcsolódás a GA szomszédos bázisok között jöhetett létre. Ebből arra következtettünk, hogy a ciszplatin ténylegesen létrehoz GA szálon belüli kötéseket is (18. ábra). Mindezek mellett megfigyelhető volt a CCA > CAA és CTN> CAN mutációk számának emelkedése is, ami arra utal, hogy a kötött GG és AG dinukleotidok 5' oldalú bázisánál is gyakori az adenozin téves beillesztése.

A fent tárgyalt SNV-ken kívül a ciszplatinnal kezelt mintákban sűrűn előfordultak dinukleotid mutációk is (18. ábra). Ez legszembetűnőbben a minták rainfall ábráin mutatkozik meg, mint a függőleges tengely 1-es szintjénél megjelenő adatpontok sokasága (17. ábra). Ezeknek a mutációknak a háromnegyede AG, GG vagy GA dinukleotidoknál jelent meg. Viszonylag gyakoriak voltak még a CA>AC dinukleotid mutációk is, melyeknek a jelentős része CCA szekvenciákra esett (CCA>CAC). Ez az ellentétes szálon tehát egy olyan TGG>GTG mutációként értelmezhető, melynek során egy GG kötött bipurin 5' oldali szomszédjánál is báziscsere történt. Összességében tehát azt találtuk, hogy a ciszplatin okozta GG, AG és GA szálon belüli kötések által összekapcsolt tetszőleges bázison és a mindkét oldalon közvetlenül szomszédos bázisokon is gyakran végbemennek pontmutációk.

A ciszplatin mindezek mellett sok rövid indel megjelenését is okozta a vizsgált mintákban (4. táblázat). Az inzerciók jelentős többsége egybázisos volt, melyek nagy része A vagy T. A timin inzerció szerinti szálon a mutációt megelőző két bázis az esetek négyötödében GG volt, feltételezhetően a szálon belül kialakult GG kötés miatt. Meglepő módon az inzerciót követő bázisok se véletlenszerűek voltak: az első bázis az esetek 84%-ában T, az első kettő pedig az esetek több mint felében TT volt. Vagyis a GG bipurin kötést tartalmazó szálat mintának használva, a DNS-szintézis során gyakran egy A bázis illesztődik be a GG dinukleotid után, ha a 3' oldali következő két bázis TT. A deléciók háromnegyede szintén egy bázis hosszú volt, ezek nagy része pedig hasonlóan a GG és AG mintázatokat érintette. Az egybázisos deléciók a bipurinok tetszőleges bázisán megjelentek. Hasonlóan, a két bázis hosszúságú deléciók gyakran mindkét bázist törölték.

Amennyiben a rákgyógyítás során alkalmazott kemoterápia erősen mutagén, felmerül annak az esélye, hogy az újonnan létrejövő mutációk olyan tumor szubklónokat tudnak létrehozni, melyek rezisztensek a használt szerre. Például aközül a vizsgált 7 mutáció közül,

melyek képesek voltak a nem működő BRCA2 gén funkcióját visszaállítani és a ciszplatin kezelésre való rezisztenciát kialakítani [124], kettő GGT>GGTT inzerció is szerepelt, mely messzemenően a leggyakoribb általunk megfigyelt ciszplatin-okozta inzerció. Vagyis maga a tumorellenes kezelés képes lehet egy még ellenállóbb daganat létrehozására, így a citotoxikus szerek mutagén hatásainak felmérése elsődleges fontosságú a tumor evolúciójának megértése szempontjából.

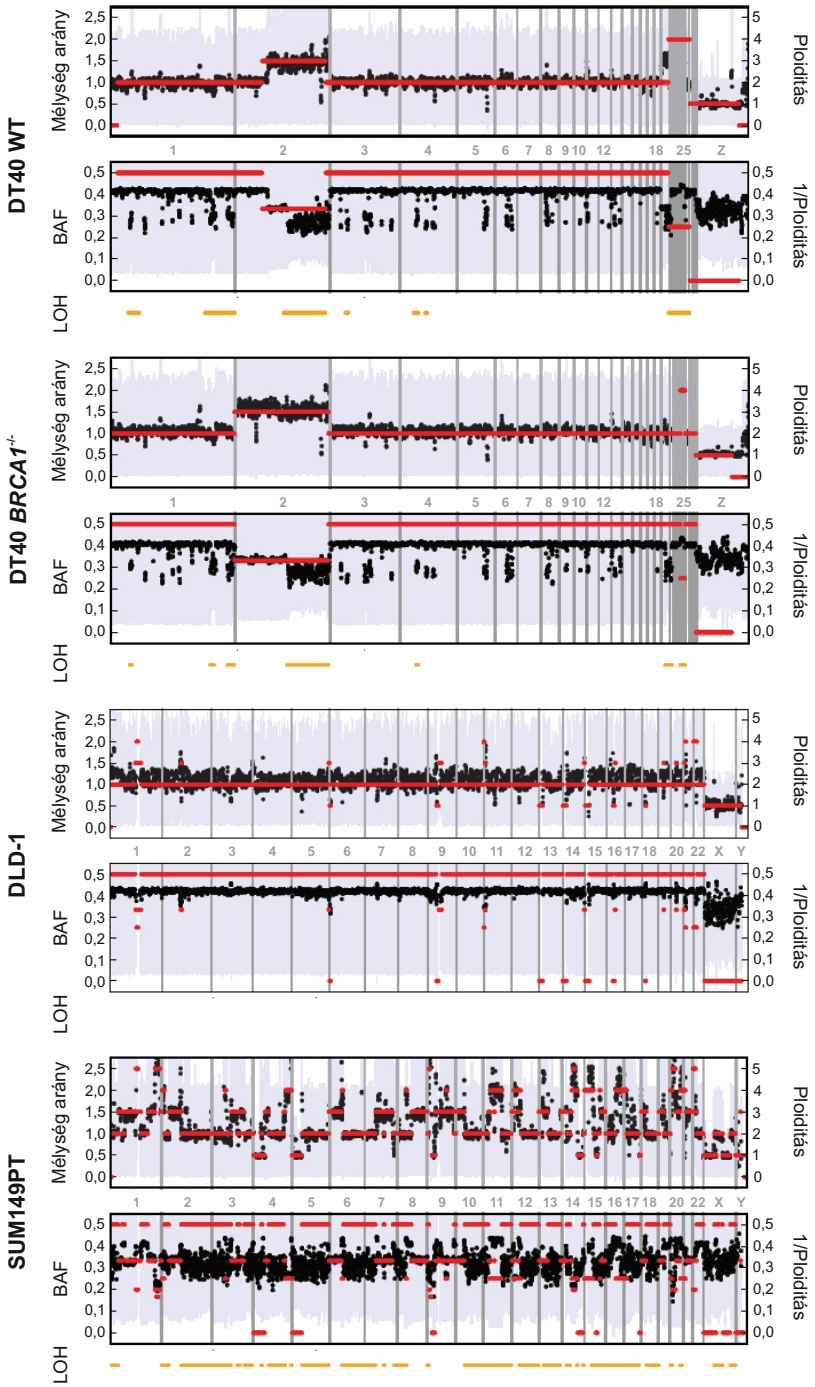
### Hosszútávú PARP-inhibitor kezelés mutációs következményei

A fenti citotoxikus szerekkel ellentétben a modern PARP-inhibitorokat specifikusan olyan esetekben alkalmazzák sikerrel, mikor a daganatos sejtekben a homológ rekombináció hibás működésére utaló jeleket találni. Ez legtipikusabban a BRCA gének mutációiként mutatkozik meg, de jelentős törekvések irányulnak a betegek olyan csoportjainak definiálására, ahol bár a BRCA génekben nem található mutáció, mégis várhatóan pozitívan reagálnának a PARP-inhibitor kezelésre [156–158]. Az ilyen típusú kezelés tehát az első olyan klinikai módszer, mely célzottan azokat a sejteket támadja, amikben egy konkrét DNS-javító mechanizmus nem működik megfelelően [159]. A biztató eredmények mellett azonban fontos felmérni, hogy a hosszútávú PARP-inhibitor kezelésnek milyen mutagén következményei lehetnek a megmaradó sejtekben. Ennek a vizsgálatára az MTA Enzimológiai Intézetének Genom Stabilitás Kutatócsoportjával végzett kutatásunk [160] során különböző BRCA mutáns és vad típusú sejtvonalak teljes genom szevenálási adatait elemeztük PARP-inhibitor kezelést követően.

Az analízis során a korábban is alkalmazott DT40 vad típusú és *BRCA*1$^{-/-}$ sejtvonalakat, illetve a DLD-1 (humán vastagbél-daganatos) és SUM149PT (humán, BRCA1 mutáns emlődaganatos) sejtvonalakat használtuk. A sejtek PARP-inhibitorral (Niraparib) történő kezelését és a teljes genom szekvenálást a fent részletezett kutatásokhoz hasonló módon végeztük. Minden sejtvonal esetén a Niraparib-kezelt klónok mellett „mock”-kezelt klónokat is vizsgáltunk.

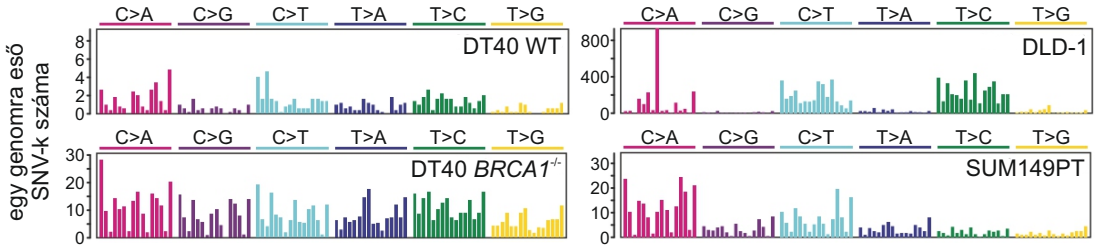
A genomi adatok elemzéséhez az IsoMut2py-t használtuk. Mivel a SUM149PT sejtvonal erős aneuploiditást mutat, elsőként minden sejtvonal kezdőklónján elvégeztük a ploiditás becslést (19. ábra). A többségében diploid DT40 és DLD-1 mintákon az így kapott eredmények a vártnak megfeleltek, a SUM149PT esetében pedig vizuálisan összevetve a kapott ábrát a [161, 162] kutatások eredményeivel, magas fokú egyezést tapasztaltunk.

Ezt követően a mutációdetektálás során a szűrési paramétereket a lokális ploiditásnak megfelelően adaptáltuk, illetve az $S$ értékre történő optimalizálási lépést a különböző ploiditású régiókon külön-külön végeztük el. Azt tapasztaltuk, hogy a „mock” kezelés során keletkezett spontán egyedi SNV-k száma (DT40 WT: $102 \pm 29$; DT40 *BRCA*1$^{-/-}$: $849 \pm 93$; DLD-1: $9799 \pm 1910$; SUM149PT: $608 \pm 146$) egyik sejtvonal esetén sem tért el szignifikánsan a Niraparibbal kezelt klónokban talált pontmutációk számától.




**19. ábra. A vizsgált sejtvonalak kezdőklónjainak kariotípusa.** A lokális ploiditás becslését az IsoMut2py modullal végeztük. A függőleges szürke vonalak és a panelek közti számok a kromoszómákat jelölik. **Felső panelek:** A mélység arány a lokális lefedettség és a becsült diploid lefedettség arányának átlagát mutatja (fekete pontok) a genom mentén, 1 Mbp széles és 50 kbp átfedéssel definiált szakaszokon. A szürke hibavonalak az átlagok mellett meghatározott interkvartilis terjedelmet szemléltetik. A piros pontok a lokálisan becsült ploiditást mutatják. **Középső panelek:** A BAF a nem-referencia („B") allél frekvenciájának átlagát mutatja (fekete pontok) a fenti csúszóablakos genomi régiókban. A szürke hibavonalak az interkvartilis terjedelmet jelölik. A piros pontok a lokálisan becsült ploiditás reciprokát mutatják. Haploid régiókon kényelmi szempontból definíció szerint nulla értéket vesznek fel. **Alsó panelek:** A narancssárga vonal a „loss of heterozygosity" (LOH; heterozigótaság elvesztése) eseményeket jelöli.

A kétféle DT40 sejtvonal közti különbség a spontán mutációs rátában megfelel a [146] kutatásunkban tapasztaltaknak. A „mock” és niraparibbal kezelt klónok SNV spektrumai között egyik sejtvonal esetében sem találtunk jelentős eltérést, a spontán kialakuló mutációs mintázat azonban a különböző sejtvonalak esetén más és más volt (20. ábra). A kapott spektrumokat és a rákgenomokban azonosított COSMIC szignatúrákat [163] közösen elemeztük a t-SNE (t-distributed Stochastic Neighbor Embedding) [164] módszerrel, melynek során az alapvetően sokdimenziós (esetünkben 96) vektorokat egy nemlineáris dimenzióredukciós eljárással egy kétdimenziós altérbe vetítjük a könnyebb vizualizálás céljából úgy, hogy az eredetileg „hasonló” pontok az alacsony dimenziós altérben is „hasonlóak” legyenek. (A „hasonlóság” konkrét értéke hagyományosan a pontok euklideszi távolságával van összefüggésben.) A kapott eredmény tehát egy kétdimenziós ábra, melyen minden pont egy referencia szignatúrát vagy a vizsgált sejtvonalak spontán mutációs spektrumait reprezentálja. A közeli pontok által jelölt eloszlások eredeti verziói is valamilyen módon egymáshoz hasonlóak. Az elemzés szerint a DT40 vad típusú klónok spontán mutációs mintázata leginkább az 1. szignatúrára hasonlít, mely hagyományosan az „öregedési szignatúra” nevet is viseli, utalva arra, hogy mutagén folyamatok nélkül idővel ilyen eloszlás szerint halmozódnak fel a genomi mutációk. A DT40 *BRCA*1$^{-/-}$ klónok spontán mutációi a 3. szignatúrára hasonlítottak, mely valóban jellemzően a BRCA1/2 mutáns tumorokban jelenik meg. A DLD-1 sejtvonalbeli „mock”-kezelt klónok mutációs spektrumai a 6., 15. és 20. szignatúrák eloszlásaira emlékeztettek, melyeket az MMR (mismatch repair) DNS-javító mechanizmus hibájával hoztak összefüggésbe [112]. A hasonlóság megmagyarázható a DLD-1 sejtvonalban található, az MSH6 nevű, MMR-ben aktívan résztvevő gén mutációjával. Meglepő módon az egyébként BRCA1-mutáns SUM149PT sejtvonal nem mutatott hasonlóságot a 3. szignatúrával. Mivel a sejtvonal klónjai emellett nem voltak különösebben érzékenyek a niraparib kezelésre sem, arra lehet következtetni, hogy további mutációk hatására a HR a sejtekben valamilyen módon visszanyerte funkcionalitását. Erre a feltételezésre azonban meggyőző bizonyítékot a genom vizsgálata során nem találtunk.




**20. ábra. A spontán mutációk eloszlása a vizsgált sejtvonalakban.** A vízszintes tengely kategóriái megegyeznek a 16. ábrán láthatókkal. A függőleges tengelyen a genomban talált mutációk darabszáma (és nem azok triplettgyakoriságokkal normált értéke) szerepel.

Az indelek elemzése szintén arra a konklúzióra vezetett, hogy a niraparib kezelés nem okoz jelentős emelkedést a genomi elváltozások számában a „mock”-kezelt klónokhoz képest.

Az eredményeink tehát alátámasztják, hogy PARP-inhibitor kezelés hatására nem növekszik számottevően a sejteket érő mutációs terhelés, így ezek a készítménynek biztonsággal és várhatóan hosszútávú mellékhatások nélkül alkalmazhatóak a rákgyógyításban. Ezzel szemben korábbi kutatásaink szerint a ciszplatin használata erősen megnöveli a mutációs rátát, ami másodlagos daganatokhoz vagy rezisztenciához vezethet. Jelenleg a klinikai gyakorlatban elsődlegesen a platinaszármazékokat alkalmazzák a petefészek- és bizonyos BRCA-mutáns emlődaganatok esetén is [165]. Az eredmények alapján, amennyiben lehetséges, felmerülhet a platina leváltása PARP-inhibitorokra, melyek csökkentik a normál sejtekre eső mutációs terhet, így alacsonyabb toxicitásúak.

### Metasztázisok törzsfájának meghatározása a genomi mutációk alapján

Annak illusztrálására, hogy klinikailag milyen jelentősége lehet a kemoterápiás szerek mutagén hatásának részletes leírása, az MTA Enzimológiai Intézetének Genom Stabilitás Kutatócsoportjával végzett kutatásunk [166] során egy fiatal tüdődaganatos férfi többszörös metasztázisainak törzsfáját határoztuk meg a ciszplatin genomi lenyomatának ismeretében. A DT40 sejtvonal ciszplatin mutációs spektrumának meghatározása után nem sokkal a ciszplatin humán sejtvonalakra, illetve daganatsejtekre gyakorolt hatását is feltérképezték [167]. Munkánk során ezt a humán sejtekre specifikus ciszplatin szignatúrát használtuk fel.

A klinikai képalkotó eljárások gyakran jóval később észlelik az áttétek jelenlétét, mint azok valójában kialakulnának. Az utólagos genomikai elemzések lehetővé teszik, hogy a metasztázisok feltűnésének idejét pontosabban meghatározzuk. Az így szerzett részletes információ a tipikus megjelenési időkre vonatkozóan segítségére lehet a klinikusoknak hasonló esetekben a kezelésekkel kapcsolatos döntések meghozatalában. A páciens tüdejében, csontjában, nyirokcsomójában és májában talált áttétekből, illetve a véréből post mortem mintavételezés után teljes genom szekvenálás készült, az így nyert adatokat elemeztük különféle szempontok szerint.

A tanulmány írásakor az IsoMut2py akkori verziója még nem volt képes a több mintában is megtalálható mutációk detektálására, így a szomatikus mutációk esetében a GATK MuTect2 [111] eszközét használtuk, számos utószűrési lépést alkalmazva a mutációs lista pontosítása érdekében, a csíravonal variánsokat pedig a GATK HaplotypeCaller segítségével azonosítottuk. A kutatás során használt elemzési folyamat nagyban inspirálta az IsoMut2py modul funkcióinak kialakítását, így a lent ismertetett elemzési lépéseknél külön kitérünk arra, hogy az IsoMut2py használatával az adott analízis miként lenne megvalósítható.

A különböző szervekből nyert szekvenálási adatokat tehát a vérmintával párba állítva

a MuTect2 szoftverrel elemeztük. Az így detektált (a vérhez képest szomatikusnak ítélt) mutációk listáit egybeöntve az összes potenciálisan mutálódott genomi pozícióban legeneráltuk az összes minta közös pileup fájlját. Ebből azokat a pozíciókat tartottuk meg, melyekben a vérmintában 2-nél kevesebb read támogatta az alternatív allélt. Az így megmaradó listát tovább szűrve azok a pozíciók maradtak meg, melyekben az összes mintára igaz volt, hogy az alternatív allélt vagy nulla (teljesen tiszta) vagy 2-nél több read (legalább szubklonális mutáció) támogatja. Ez a procedúra az IsoMut2py modul aktuális verziójával néhány lépésben elvégezhető. A mutációk detektálását a *unique_mutations_only = False* beállítással szükséges elindítani az összes vizsgált mintára. Annak érdekében, hogy ebben a lépésben még ne szűrjük meg túl szigorúan az eredményeket, érdemes a *min_sample_freq = 0,05* és *min_other_ref_freq = 0,7* paraméterválasztással élni. A kapott mutációs lista genomi pozícióiban a *get_details_for_mutations()* függvénnyel egy minden mintáról részletes információkat (lefedettség, referencia allél frekvencia, alternatív allél frekvencia) tartalmazó adattáblát kaphatunk a pileup fájl alapján. Ebből az adattáblából triviális műveletek segítségével kiszámítható mintánként az alternatív allélt támogató readek száma, illetve egy szintén kézenfekvő lépéssel ezek a fenti feltételeknek megfelelően szűrhetők. A mutációk detektálása tehát a korábban alkalmazott legalább három különböző szoftver helyett eggyel is elvégezhető.

Az így megszűrt mutációs listát használva a *plot_hierarchical_clustering()* függvénnyel egy lépésben ábrázolható a vizsgált minták dendrogramja. A kutatás során is elemzett négyféle szövet IsoMut2py modullal gyártott leszármazási fája épp a 12. ábrán látható, mely természetesen megegyezik a tanulmányban is szereplő konklúziókkal. A korábbiakhoz hasonlóan a mintákban talált egyedi és közös mutációk rainfall ábrái és mutációs spektrumai (SNV, DNV, indel) szintén egy-egy függvény segítségével ábrázolhatóak. A tumorszövetet tartalmazó minták triplettspektrumai között nem találtunk számottevő eltérést, sem az összes, sem a kizárólagosan egyedi mutációk tekintetében. Ez arra utalhat, hogy a tumorsejtekben végbemenő mutagén folyamatok nem függnek erősen a mikrokörnyezettől, vagyis az adott szervtől. Emellett az így kapott eloszlások és a referencia szignatúrák [163] t-SNE-vel [164] történő vizsgálata során sem találtunk egyik elfogadott mutagén folyamat lenyomatával sem jelentős hasonlóságot. Tehát a vizsgált tumorok mutációs mintázata nem hasonlított többek között a dohányzással összefüggésbe hozott 4. szignatúrára sem, ami konzisztens a páciens nem dohányzó státuszával.

A kutatás során az egyszerű leszármazási kapcsolatok vizsgálata mellett arra is törekedtünk, hogy a metasztázisok keletkezési idejét a betegség lefolyásának idővonalához társítsuk. Mivel a páciens kezelése során egyszeri alkalommal ciszplatint is kapott, a ciszplatin okozta mutációkban az alternatív allélfrekvenciából következtethetünk arra, hogy az adott metasztázis a kezelés előtt vagy után jött létre. A kezelés előtt is jelenlévő metasztázisok esetén várhatóan a ciszplatin-specifikus mutációk csak egy-egy klónt érintettek (szubklonálisak), így az allélfrekvencia ilyen esetekben alacsonyabb lesz. Ezzel szemben

egy olyan metasztázisnál, mely a ciszplatin kezelést követően egyetlen sejtből fejlődött ki, az allélfrekvenciák a kérdéses mutációkban robusztusabbak lesznek (klonális mutációk). A ciszplatin-okozta mutációk jelenlétének vizsgálatához a *decompose_SNV_spectra()* függvény segítségével megkaphatjuk a referencia szignatúrák járulékait az adott mintában. Ha a referencia spektrumok adatbázisához hozzáfűzzük a korábban azonosított ciszplatin-szignatúrát [167] is, a kapott eredmények alapján meghatározhatjuk azokat a metasztázisokat, melyekben a ciszplatin lenyomata jelentős súllyal hozzájárul a mutációs spektrumhoz. Esetünkben a primer tüdő tumorban és a csont-, illetve nyirokcsomó áttétekben találtuk ennek jelét. Annak érdekében, hogy biztosak lehessünk benne, hogy ciszplatinnal nem kezelt páciensek esetén valóban nem találunk hasonló nyomokat a mutációs mintázatban, a TCGA (The Cancer Genome Atlas) adatbázisból letöltött, ciszplatinnal nem kezelt tumorok szekvenálási adatain is elvégeztük a fenti kiértékelést. Azt találtuk, hogy ezekben a mintákban valóban nem jelenik meg a ciszplatin-szignatúra. Az SNV-spektrumok vizsgálatán túl a ciszplatin jellegzetes nyomait tetten érhetjük a DNV-kben is. A vizsgált mintákban a leggyakoribb DNV mutációk a CC>AA és a CT>AA voltak, melyek minden bizonnyal a ciszplatin okozta GG és AG szálon belüli kötések mentén alakultak ki. Vagyis az adatok alapján azt tapasztaltuk, hogy akár az egyszeri ciszplatin kezelés is detektálható nyomokat hagyhat a genomban.

A máj metasztázis esetében azt találtuk, hogy a ciszplatin-specifikus DNV-knél mérhető allélfrekvencia megegyezett azoknak a mutációknak az allélfrekvenciájával, melyek az összes mintában jelen voltak. Vagyis a máj mintában a ciszplatin mutációk klonálisak, így várhatóan a máj áttét a ciszplatin kezelést követően egyetlen sejtből fejlődött ki. Továbbá a májáttétben lévő egyedi SNV-k esetében nem találtuk meg a ciszplatin-szignatúra nyomát, de a nyirokcsomóban és a májban együttesen megjelenő SNV-knél igen. Vagyis a májban jelenlévő ciszplatin mutációk már a nyirokcsomóban is jelen voltak, így arra következtethetünk, hogy a májáttét a ciszplatin kezelés után, a nyirokcsomó áttétből alakult ki. Ezzel szemben a ciszplatin okozta DNV-k a csont- és nyirokcsomóáttétekben szubklonálisak, tehát ezek a metasztázisok a kezelést megelőzően jöttek létre, annak ellenére, hogy a képalkotó vizsgálatok során nem detektálták őket.

A daganatok klonalitásának vizsgálatával már korábban is bizonyították, hogy a szubklónok hozzájárulhatnak a metasztázisok képződéséhez és a különféle kezelésekre való rezisztencia kialakulásához [168, 169], így az egyedi szubklonális események feltérképezése és megértése nagyon fontos a személyre szabott gyógyászat szempontjából [170]. Kutatásunkban arra hívjuk fel a figyelmet, hogy a különféle kezelések pontos mutagén hatásának ismeretében a daganatok és metasztázisainak szubklonális fejlődése pontosabban nyomon követhető.

# POPULÁCIÓ-SZINTŰ GENETIKAI VIZSGÁLATOK KÖRNYEZETI MINTÁKBÓL

## AZ NGS TECHNOLÓGIÁK HASZNÁLATA SORÁN FELMERÜLŐ JOGI ÉS ETIKAI PROBLÉMÁK

A fentiek alapján úgy tűnhet, hogy arra, hogy felvegyük a versenyt a genetikai hátterű, agresszív daganatos megbetegedésekkel szemben, minden ember teljes genetikai kódjának a meghatározása és időbeli nyomon követése lehet az áhított megoldás. Egy ilyen adatbázis emellett számos demográfiai, bűnügyi és történelmi kutatás és alkalmazás során felbecsülhetetlen jelentőségű lenne. Mindazonáltal egy ilyen álom megvalósítása a tisztázatlan etikai és jogi következmények miatt egyelőre elérhetetlen.

Maga az adatgyűjtés is problematikus, hiszen a fenti elképzelés megkövetelné a világon minden ember DNS-ének megszerzését, illetve egyéb adatainak lekönyvelését, amihez mindenkivel személyes konzultációra lenne szükség. Emellett az adatok tárolására és elemzésére óriási számítógépes és emberi kapacitást kellene fordítani. A rendelkezésre álló egészségügyi adatok mennyiségének növekedésével egyre lehetetlenebb feladatnak tűnik az adatok bizalmas kezelésének megvalósítása, különösen amiatt, mert a személyes információkhoz való hozzáférés nincs kellően hatékonyan korlátozva [171]. Ezen a helyzeten ront az is, hogy a genetikai adatok nem csak arra a személyre vonatkozó információkat tartalmaznak, aki a genetikai tesztben aktuálisan részt vesz, hanem mindazokról, akik vele genetikai kapcsolatban állnak, így a legszélesebb körben értett családjáról is [172]. Továbbá a munkaadók, illetve biztosító cégek genetikai vizsgálatok eredményeire alapozott esetleges diszkriminatív stratégiái elriaszthatják az embereket a részvételtől [173].

Bár az etnikai adatok széleskörű elérhetősége elsődleges fontosságú a demográfiai tendenciák, munkavállalási szokások és lehetőségek, jövedelem eloszlások, végzettségi szintek, migrációs mintázatok, családfelépítés, szociális hálózatok tanulmányozásának szempontjából [174, 175], az ilyen információk begyűjtése jellemzően önbevallásos alapon, különféle nehezen egységesíthető módszerrel [176] és az adatok érzékenysége miatt szigorú jogszabályi keretek között történik [176].

## EGY LEHETSÉGES MEGOLDÁS: POPULÁCIÓGENOMIKAI KÖVETKEZTETÉSEK SZENNYVÍZMINTÁK VIZSGÁLATÁBÓL

A fenti problémákra egy lehetséges választ jelenthet az egyéni szekvenálási adatok összekeverése („poolozása”) és a lokális populáció közös genomikai jellemzőinek vizsgálata, az egyének azonosítása nélkül. Bár értelemszerűen a rendelkezésre álló genetikai felbontás ilyen jellegű mesterséges korlátozása egy kompromisszumos megoldás, az etikai problémák egy része ezzel feloldódna. Az egészségügyi rizikófaktorok populációs szintű nyomon követése tünetmentes egyénekben több tanulmány szerint is rendkívüli előnyökkel járna [177, 178], annak ellenére is, hogy az áhított személyre szabott gyógyászattal nem egyenértékű. Az etnikai kutatások ezzel szemben nem igényelnek konkrét személyes

azonosítást, így a közösségre vonatkozó aggregált eredmények releváns adatforrást jelentenének. Hasonló a helyzet bűnügyi szempontból: egy olyan adatbázis, mely világszinten tartalmazná a lokális populációkban a genetikai jellemzők eloszlását, felbecsülhetetlen jelentőségű lenne. Azokban az esetekben, amikor a bűnügy résztvevőinek genetikai azonosítása a cél, de a humán genomnak csak egy rövid szakasza nyerhető ki a rendelkezésre álló bizonyítékból, a lokális genetikai eloszlás felhasználható lenne priorként a DNS-egyezés valószínűségét meghatározó modellekben.

Bár ez a megoldás az etikai és jogi kérdések egy részére választ ad, az adatgyűjtés technikai megvalósításának praktikus nehézségein nem enyhít. Alternatívaként az egyénenkénti szekvenálási adatok összeöntése helyett felmerül az alapvetően kevert genetikai információt hordozó szennyvíztisztító telepekről való mintagyűjtés. Gyakorlati szempontból ez egy nagyon vonzó opció, hiszen a szennyvízben feltételezhetően egy viszonylag nagy és többnyire egészséges közösség genetikai anyaga keveredik, melyet egyébként nem lenne lehetséges monitorozni. Emellett az így gyűjtött minták inherensen kevertek, ezért nem szükséges egyéni hozzájárulást szerezni az elemzésükhöz, illetve az anonimizálási folyamat is kihagyható.

Annak az igazolására, hogy a szennyvízminták elemzésével valóban releváns eredmények kaphatók, a COMPARE Global Sewage Surveillance Project keretein belül 2016-tól kezdve világszerte gyűjtött szennyvízmintákat [179] vizsgáltuk [180]. Eredetileg a mintákat metagenomikai szempontból analizálták, az elsődleges cél az antibiotikum-rezisztencia gének és fertőző betegségeket terjesztő kórokozók eloszlásának a meghatározása volt az egészséges populációkban metagenomikai szekvenálás útján. Mindazonáltal a baktériumok és vírusok azonosítása mellett egy ilyen adathalmaz olyan további információk elemzését is lehetővé teszi, mely az eredeti kutatás keretein túlmutat. Mivel a [179] tanulmány eredményei szerint a szekvenálási minták átlagosan csak 0,2%-ban tartalmaztak emberi DNS-t, így a teljes genom genotipikus eloszlásának meghatározása nem tűnt megvalósíthatónak. Leszűkítve azonban a mitokondriumra a vizsgálatot, informatív eredményeket kaphatunk.

Korábbi kutatások már bizonyították, hogy a mitokondriális DNS elemzése elégséges arra, hogy szennyezett felszíni vizekben elkülöníthető legyen a emberi eredetű, és a szarvasmarhától, illetve a sertéstől származó ürülék [181]. A humán mitokondriális DNS (mtDNS) egy rövid (16 569 bázispár hosszú), cirkuláris DNS, ami minden emberi sejtben több példányban is megtalálható, emiatt még olyan mintákban is viszonylag könnyen azonosítható, melyek alapvetően alacsony koncentrációban tartalmaznak emberi DNS-t. Az emberi mitokondrium kizárólag anyai ágon öröklődik (bár friss kutatások ezt kétségbe vonják [182]) és korábbi eredmények szerint az öröklődés klonális, tehát az mtDNS az anyától az utódba rekombináció nélkül adódik át [183]. Ez azt jelenti, hogy a mitokondriumban található mutációk alapján nyomon lehet követni az evolúciós mintázatokat. A humán mitokondrium filogenetikai fájának levelei a mitokondriális haplotípusok, melyeket a köztük

lévő hasonlóságok alapján bővebb kategóriákba, ún. mitokondriális haplocsoportokba sorolunk, melyek a filogenetikai fa fő elágazási pontjai.

Évtizedekre visszanyúlóan rendkívül sok kutatás foglalkozik a különböző humán mtDNS haplocsoportok eloszlásának felderítésével a különböző népekre és földrajzi területekre vonatkozóan [184–192], elsősorban annak érdekében, hogy fény derüljön a különböző populációk eredetére és genetikai struktúrájára. Az így felgyűlt rengeteg információ miatt az eredetileg csak a kutatókat foglalkoztató genetikai eredet vizsgálat mára világszerte elérhető lett a nyilvánosság számára is. Emberek százezrei küldik el DNS-üket az ilyen teszteket végző több tucat kereskedelmi cég egyikének, hogy felfedjék családfájuk régen elfelejtett ágait, illetve őseik földrajzi eredetét. Sok kritika éri azonban az ilyen vállalatokat [193], legfőképp azért, mert olyan félrevezető információkat nyújtanak ügyfeleiknek, melyek alapjaiban véve befolyásolják személyes identitásukat. Az egyik fő negatívum, hogy a földrajzi eredetre tett következtetéseket olyan adatbázisokra alapozzák, melyek rendkívül kevés referencia mintát tartalmaznak. Amennyiben tehát megoldható lenne a különböző földrajzi helyeken a mitokondriális haplocsoportok eloszlásának időbeli, dinamikus monitorozása, az ilyen tesztek sokkal pontosabb információkkal tudnának szolgálni.

Kutatásunk során ezért arra törekedtünk, hogy a szennyvíz mintákban azonosítsuk az emberi mtDNS-t, majd az így nyert adatok alapján meghatározzuk a mtDNS haplocsoportok eloszlását a szennyvíztisztító telepek környékén.

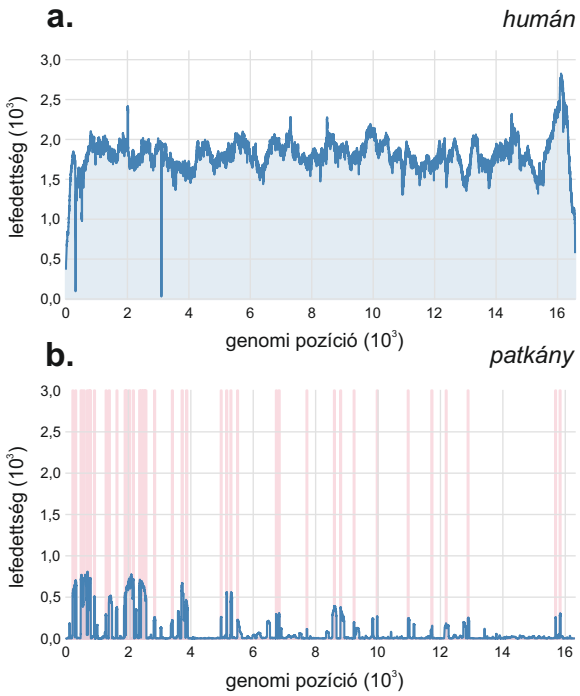
### Szennyvízminták gyűjtése és az emberi DNS azonosítása

A szennyvízmintákat 79 városi szennyvíztisztító telepről gyűjtötték, ezzel 74 várost és 60 országot lefedve világszerte [179], melyekből egy metagenomikára optimalizált protokollal nyerték ki a DNS-t [194], melyet további preparációs lépések után [179] az Illumina HiSeq platformon szekvenáltak. A nyers szekvenálási eredmények az ENA (European Nucleotide Archive; http://www.ebi.ac.uk/ena/) honlapjáról tölthetők le az ERP109094 azonosítóval.

A mintákból átlagosan 120 millió short read-et nyertek ki (8-tól 398 millióig terjedően az egyes mintákban), melyeket a humán mitokondrium referencia szekvenciájához (NCBI ID: NC_012920.1) [195] illesztettünk a BWA-MEM algoritmus [196] segítségével, az alapértelmezett beállításokat használva. Emellett elvégeztük az illesztést számos más gerinces (*Bos taurus*, *Sus scrofa*, *Danio rerio*, *Canis lupus familiaris*, *Gallus gallus*, *Ovis aries*, *Rattus norvegicus*) mtDNS szekvenciáját referenciának tekintve. Az illesztési eredményekben a samtools szoftverrel [197] azonosítottuk a PCR duplikátumokat, mivel azonban ilyeneket nem találtunk, a duplikátum eltávolítási lépésre nem volt szükség.

Annak érdekében, hogy megnöveljük a mitokondrium lefedettségét, azoknak a mintáknak az illesztési eredményeit egybe gyűjtöttük, melyek ugyanarról a szennyvíztisztító telepről származtak ugyan, de különböző időpontokban gyűjtötték őket. Azokban az esetekben, amikor egy városból több szennyvíztisztító telepről is sikerült mintát gyűjteni,

ezeket egymástól elkülönítve elemeztük. Csak azt a 44 mintát (C1. táblázat) elemeztük tovább, melyekben a humán mitokondriumra vonatkozó átlagos lefedettség elérte a 10-et (C1. ábra). Amennyiben több olyan mintában is megfelelő lefedettséggel megjelent az mtDNS, melyek azonos városból (El Paso mintái) vagy környékről (Kitwe és Lusaka mintái) származtak, lehetőség nyílt a módszer reprodukálhatóságának vizsgálatára is.




**21. ábra. Az ember és a vándorpatkány mtDNS-ének lefedettsége a vizsgált mintákban. a.** A humán mtDNS mentén megfigyelhető együttes lefedettség azokban a mintákban, melyekben az átlagos lefedettség elérte a 10-et. **b.** Együttes lefedettség a vándorpatkány mitokondriuma mentén (kék vonal). A piros függőleges sávok a vándorpatkány mitokondriumának azon részleteit jelölik, melyek a humán mtDNS-sel homológok. Ezt a mennyiséget minden genomi pozícióra vonatkozóan egy bináris skálán definiáltuk a következő módon: ha az adott genomi pozíció lefedhető volt a vándorpatkány mtDNS-ének egy olyan 19 bázis hosszú mozgó ablakával, ami a humán mtDNS-ben is megtalálható, a pozíciót homológnak tekintettük. Az ablakméretet 19-nek választottuk, mert az illesztéshez használt algoritmus alapértelmezetten azt követeli meg, hogy a felillesztett short readekből 19 szomszédos bázis hiba nélkül illeszkedjen a referenciagenomhoz.

Mivel a viszonylag magas átlagos lefedettség a humán mtDNS mentén esetleg adódhat abból is, hogy valójában nem emberi, de a humán mitokondrium egyes szakaszaival homológ (hasonló) szekvenciák lettek helytelenül felillesztve, ezért a lefedettséget a vizsgált 44 mintára összeöntve ábrázoltuk a 21. ábrán az ember és a vándorpatkány mitokondriuma mentén. A nagyságrendileg egyenletes eloszlás (a fluktuációk nem haladják meg az NGS adatoknál ismert szórást [198]) az emberi mtDNS mentén arra enged következtetni, hogy a magas átlagos lefedettséget nem csak a helyenként felszaporodó, rosszul felillesztett short readek adják. (Ugyanez a tendencia akkor is megfigyelhető, ha a mintákat egyesével vizsgáljuk, természetesen sokkal alacsonyabb átlagos lefedettséggel (C2. ábra) Ezzel szem-

ben a vándorpatkány esetében a megfigyelhető néhány csúcstól eltekintve a mitokondrium jobbára lefedetlen marad. A kiugró csúcsok emellett azokra a régiókra koncentrálódnak, melyek a humán mitokondriummal homológok, így várhatóan a valójában emberi eredetű short readek hibás felillesztése okozza őket.

A C1. ábrán emellett feltüntettük a kizárólag a humán, illetve a vándorpatkány mitokondriumra illett short readek számát, a mindkettőre illett readek számával együtt. A kizárólagosan humán readek darabszáma átlagosan 40-szerese azoknak a readeknek, melyek egyedien a vándorpatkányra illeszkedtek. A fent sorolt további gerinces fajok esetén is hasonló eredményeket kaptunk, mint a vándorpatkánynál, azonban még alacsonyabb átlagos lefedettséggel. A szennyvízminták mellett megszekvenált kontroll mintákban egyetlen humán mtDNS-re illeszkedő readet sem találtunk. Ezek az eredmények azt sejtetik, hogy az azonosított humán readek valóban emberi eredetűek és nem hibás illesztésből adódnak, illetve a kontroll minták tisztasága alapján a feldolgozás közben történt esetleges DNS-kontamináció is kizárható.

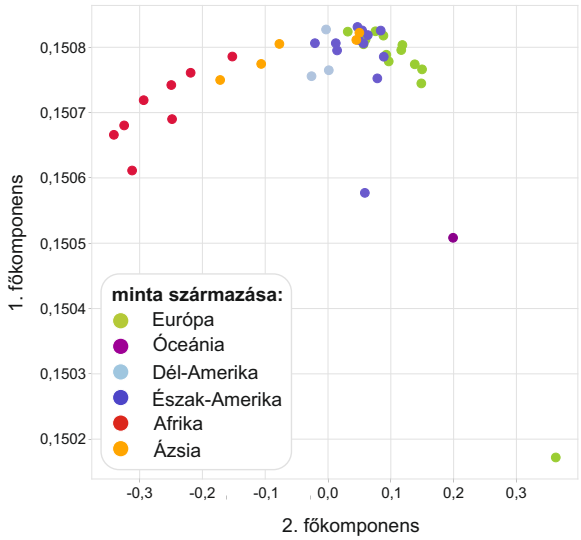
### Kezdeti elemzések: főkomponens-analízis, t-SNE, filogenetikai fa

Azért, hogy meggyőződjünk róla, hogy a mintákban detektált emberi mtDNS elegendő mennyiségű a hiteles tudományos következtetések levonására, elsőként különböző felügyelet nélküli klaszterezési algoritmusokkal igyekeztünk a mintákat eredetük szerint elkülöníteni egymástól. Korábbi kutatások során demonstrálták [199], hogy a humán mitokondriális genomon végzett főkomponens analízis hatékonyan el tudja különíteni az egyéneket attól függően, hogy milyen mitokondriális haplocsoportba tartoznak.

#### Főkomponens-analízis

Mivel a mitokondriális haplocsoportok eloszlása a különböző földrajzi területek között jelentősen eltér [188], ésszerű annak a feltételezése, hogy a különböző kontinensekről gyűjtött mintákban nyomokban talált mtDNS is eltéréseket mutat majd. Ennek az elméletnek az igazolására főkomponens-analízist (PCA - principal component analysis) végeztünk a vizsgált 44 mintán. Ennek során minden genomi pozícióban meghatároztuk az oda felillesztett readekben az adott helyen leggyakrabban előforduló bázist a samtools mpileup parancsának [197] segítségével. Az így kapott adatokat az ún. one-hot-encoding eljárással egy $(44, 4 \cdot 16\,569)$ dimenziójú mátrixszá alakítottuk úgy, hogy minden minta minden genomi pozíciójában a leggyakoribb bázishoz az 1, az összes többi bázishoz pedig a 0 értéket rendeltük. A főkomponens-analízist ezen a mátrixon a scikit-learn [113] Python csomaggal végeztük el, majd az adatokat levetítettük a kapott első két főkomponens által kifeszített altérre (22. ábra). Látható, hogy mind Ázsia, mind Afrika szennyvízmintái meggyőzően elszeparálódnak a többi mintától. Ezzel szemben az európai és amerikai minták némileg keverednek az ábrán, ami egybevág az intuíciónkkal és korábbi irodalmi bizonyítékokkal [199, 200] arra vonatkozóan, hogy ezeken a kontinenseken rendkívül diverz populációk él-

nek a hosszú migrációs folyamatoknak köszönhetően. Mindezek ellenére, néhány nem várt kiugró pont is megjelenik mind Európa, mind Észak-Amerika esetében. Ez valószínűleg annak tudható be, hogy a főkomponens-analízis során minden mintából származó információt egyetlen konszenzus szekvenciába tömörítettünk annak ellenére, hogy a mintavétel jellegéből adódóan valójában egy egész populáció kevert lenyomatát detektáljuk. Bár a leggyakoribb bázis kiválasztása egy gyakori módszer a konszenzus szekvencia generálásnál, ezzel mesterségesen kevert haplotípusokat hozunk létre egy mintán belül.




**22. ábra. A humán mtDNS-en 10-es átlagos lefedettséget elérő minták főkomponens-analízise.** A különböző kontinensekről származó mintákat különféle színű pontok jelölik.

### t-SNE

A fenti módon megalkotott mátrixot a főkomponens-analízis mellett a t-SNE [164] módszerrel is elemeztük. Ezzel a technikával is egy dimenzió-redukció vihető végbez az alapvetően sokdimenziós térben, de algoritmikusan más megközelítéssel, mint a főkomponens-analízis esetén. A vizsgálathoz ez esetben is a scikit-learn [113] csomagot használtuk. A dokumentáció javaslatát követve első lépésként főkomponens-analízissel kiválasztottuk az 50 legdominánsabb főkomponenst, majd a t-SNE-t ezen a csökkentett dimenziójú adatszetten futtattuk. A fent leírt klasztereket ezzel a módszerrel is sikerült reprodukálni.

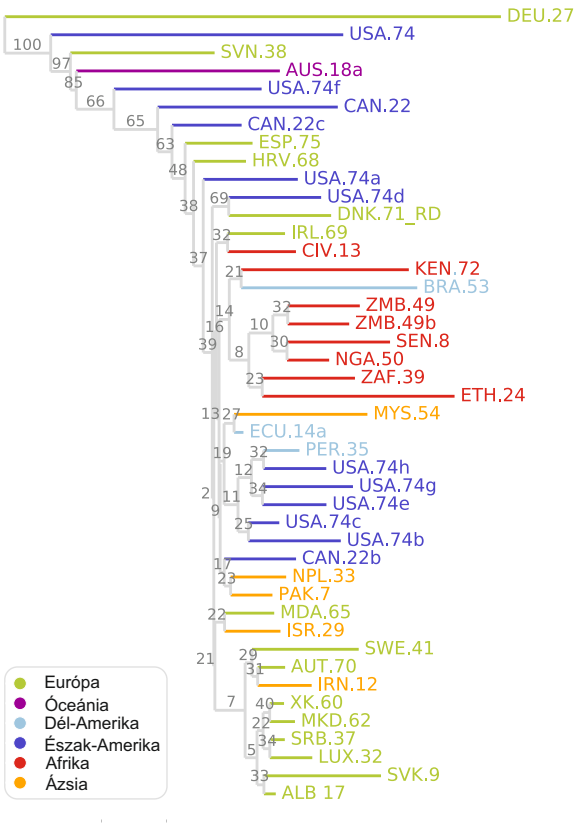
### Filogenetikai elemzés

Mivel az mtDNS haplocsoportok vizsgálata szorosan kapcsolódik az evolúciós mintázatok feltérképezéséhez [192, 201, 202], így magától értetődőnek tűnt egy filogenetikai elemzés lefuttatása az adatokon. A mtDNS haplotípusok filogenetikai fája és a tény, hogy a különböző haplocsoportok elterjedése földrajzilag igen eltérő, arra enged következtetni, hogy az egymáshoz közel gyűjtött minták az elemzett minták saját filogenetikai fáján is

feltételezhetően kládokat alkotnak majd. Első lépésként ismételten konszenzus szekvenciákat hoztunk létre minden mintához (a bcftools és vcfutils programcsomagokkal), melyek azonban a korábbi „többségi módszerhez" képest némileg kifinomultabbak, hiszen ha többféle bázist is detektálunk egy genomi pozícióban, a konszenzus szekvencia ezt képes a „pirimidin" vagy „purin" kategóriákkal jelezni. Az így létrehozott mintánkénti konszenzus szekvenciákat a Biopython ClustalW illesztőmodulja segítségével egymásra illesztettük, így egy olyan mátrixot létrehozva, melynek minden sorában egy konkrét minta szekvenciája szerepel. Azt találtuk, hogy a 16569 (a humán mtDNS hossza) oszlopot tartalmazó mátrixnak csak 1164 oszlopában volt bármilyen eltérés a minták között.

Ezt követően a Biopython Phylo modulja segítségével elsőként kiszámoltuk a minták távolságmátrixát, mely egy szimmetrikus, nemnegatív mátrix, melynek a $d_{ij}$ eleme azt mutatja meg, hogy az $i$. és $j$. minták szekvenciái mennyire térnek el egymástól. A távolságfogalmat többféleképpen definiálhatjuk, esetünkben a két szekvenciában nem megegyező karakterek és a szekvencia hosszának arányát tekintettük távolságnak. Ebből a távolságmátrixból a „neighbor-joining" [203] módszer segítségével egy egyszerű csillag-fából kiindulva, a leghasonlóbb (legkisebb távolságú) minták sorozatos összevonásával egy fát készítettünk. Ezt követően a maximális parszimónia [204] módszerével is rekonstruáltuk a filogenetikai fát. Ennek során elvileg az összes lehetséges fa közül keressük azt, amelyik a lehető legkevesebb szekvenciális helyettesítést igényli, vagyis a „legegyszerűbbet". Ez az elv [205] bár pusztán intuíción alapul, a tudományos gondolkodás az esetek nagy részében mégis szem előtt tartja: az adatokat jól magyarázó modellek közül válasszuk a lehető legkevésbé komplexet. Ez a feladat azonban egy sok mintát tartalmazó adathalmaz esetén korántsem egyszerű: már tíz minta esetén is több mint kétmillió gyökértelen fa felállítása lehetséges. Ezért a „legegyszerűbb" topológiájú fát valójában az összes lehetséges fának egy szűkített alterén érdemes keresnünk. Ebben van segítségünkre a korábban meghatározott neighbor-joining fa, melyet az algoritmusnak mint kiindulópontot adhatunk meg. Ezt követően a fák közti keresés a „nearest-neighbor interchange" (legközelebbi szomszédok cseréje) módszerrel történik, melynek során a kezdeti fához alapvetően hasonló fákra számoljuk ki az egyszerűséget jellemző mértéket, majd ezek közül választjuk ki a legnagyobbat. Mindezt a Biopython Phylo modulja segítségével végeztük el. A fák kládjainak megbízhatóságát az ún. bootstrap eljárással határoztuk meg: az illesztett fákat 1000-szer újrageneráltuk az eredeti adatokból való visszatevéses mintavételezést követően, majd kiszámoltuk azoknak a fáknak az arányát, melyekben az adott klád előfordult. Ezt az arányt százalékként tüntettük fel a 23. ábrán minden klád esetében. Az elvárásoknak megfelelően a maximális parszimónia módszer a neighbor-joining algoritmustól némileg eltérő szerkezetű fát eredményezett, de a legfontosabb jellemzőik azonosak voltak: az Afrikából származó minták egy többé-kevésbé jól szeparált kládot alkottak, míg az európai és amerikai minták összemosódtak. Fontos azonban megfigyelni, hogy a kládok megbízhatósága rendkívül alacsony volt, aminek az oka valószínűleg a főkomponens-analízisnél is említett

kevert haplotípusok mesterséges létrehozása lehetett.




**23. ábra. Az elemzett 44 mintára maximális parszimónia módszerrel illesztett fa 1000 bootstrappel.** A szürke számok a kládok bootrap módszerrel meghatározott százalékos megbízhatóságát jellemzik. A különböző kontinensekről származó mintákat külön színekkel jelöltük. A mintákra az azonosítóikkal hivatkozunk (C1. táblázat).

### A MINTÁK MTDNS HAPLOCSOPORT ÖSSZETÉTELE

A fenti módszerek eredményei jórészt összhangban vannak a naiv elvárásainkkal a minták klaszterezettségére vonatkozóan, így arra következtethetünk, hogy a szennyvízminták humán mtDNS tartalma elegendő ahhoz, hogy hihető tudományos konklúziókat vonhassunk le belőlük. Mindemellett azonban minden eddigi módszer erősen terhelt a mesterségesen létrehozott kevert haplotípusok problémájától. Ahhoz, hogy ezt a dilemmát feloldjuk, megkíséreltük a mintákban megtalálható mtDNS keverékben meghatározni a különböző haplotípusok járulékait.

#### A haplocsoport összetétel meghatározása mintánként

Ehhez a mixemt nevű szoftvert [206] használtuk, mely egy EM (expectation maximization $\sim$ várható értéket maximalizáló) algoritmus segítségével becsli meg a kevert DNS-ben a különböző haplotípusok arányát. Ehhez a Phylotree.org honlapon található adatbázist [207] veszi alapul, amelyben több mint 5000 humán mtDNS haplotípust definiáló mutáció szerepel. A program a referencia mitokondrium szekvenciára felillesztett short readeket

(BAM fájlok) tekinti bemenetnek, majd minden read és mtDNS haplotípus párhoz hozzárendel egy értéket, mely azt jellemzi, hogy a readben megfigyelhető mutációk mennyire konzisztensek az adott haplotípus variánsaival. A $j$. readre és a $g$ haplotípusra vonatkozóan ez a mennyiség az alábbi módon áll elő

$$\tilde{\delta}_{j,g} = \prod_{i,v \in V_j} \delta_{i,v,g},$$

ahol $v \in A,C,G,T$ a megfigyelt variáns az $i$. genomi pozícióban, $V_j$ pedig a $j$. readben található variánsok $(i,v)$ halmaza. A gyakorlatban $\delta_{i,v,g} = 1-\varepsilon$, ha $v$ éppen az a bázis, ami a $g$ haplotípusban megtalálható az $i$. genomi pozícióban, egyébként pedig $\varepsilon/3$, ahol $\varepsilon$ annak a valószínűsége, hogy a variáns hibásan lett detektálva (például szekvenálási hibák következtében). Mivel előfordulhat, hogy egy readen belül több mutáció is megjelenik, amik a haplotípusról ellentmondó információkat hordoznak, az $\varepsilon$ értékét pozíció-specifikusan választják meg $\varepsilon_i = \min\{0,5; \frac{m_i}{100}\}$-nak, ahol $m_i$ azt mondja meg, hogy az $i$. pozíció a Phylotree adatbázisa szerint hányszor mutálódott. Így egy olyan pozíció, ami gyakran mutálódik, kisebb súllyal szerepel a modellben. A readek $N$ halmazának, illetve a haplotípusok $G$ halmazának ismeretében a különböző haplotípusok $\{\theta_g\}$ járulékait a következő algoritmussal határozzák meg. A kezdeti $\{\theta_g\}$ értékeket véletlenszerűen, egy Dirichlet-eloszlásból húzva adják meg. Az E lépésben kiszámolják azt a feltételes valószínűséget, hogy éppen a $j \in N$ readben megfigyelt variánsokat látjuk a $g \in G$ haplotípusban, ha a $\{\theta_g\}$-k adottak:

$$z_{j,g} = \frac{\theta_g \tilde{\delta}_{j,g}}{\sum_{g' \in G} \theta_{g'} \tilde{\delta}_{j,g'}}$$

Az algoritmus M lépésében újraszámolják a $\{\theta_g\}$-k értékeit a már ismert $z_{j,g}$-k alapján:

$$\theta_g = \frac{\sum_{j \in N} z_{j,g}}{\sum_{g' \in G} \sum_{j \in N} z_{j,g'}}$$

A fenti folyamatot addig ismételik, míg a $\{\theta_g\}$ értékekre kapott korrekció egy bizonyos határérték alá nem kerül, vagyis az értékek konvergálnak. Ezek után két heurisztikus szűrési feltételt alkalmaznak annak érdekében, hogy csökkentsék a keveréket potenciálisan alkotó haplotípusok számát. Elsőként minden readhez meghatározzák azt a haplotípust, amihez a legnagyobb feltételes valószínűség tartozik az adott $\{\theta_g\}$-k mellett. Azokat a haplotípusokat, melyek kevesebb, mint $n$ (alapértelmezésben $n = 10$) read esetében jelentek meg domináns haplotípusként, a következő lépésekben elhagyják. Másodszorra minden haplotípus esetén megvizsgálják, hogy az őt definiáló összes egyedi variáns hányad részéről van tényleges információnk az adatok alapján. Azokat a haplotípusokat, melyeknél a megfigyelhető egyedi variánsok aránya nem ér el egy küszöbértéket (alapértelmezésben 0,5), szintén elhagyják. Ennek a szűrési feltételnek az alkalmazását alacsony lefedettség esetén

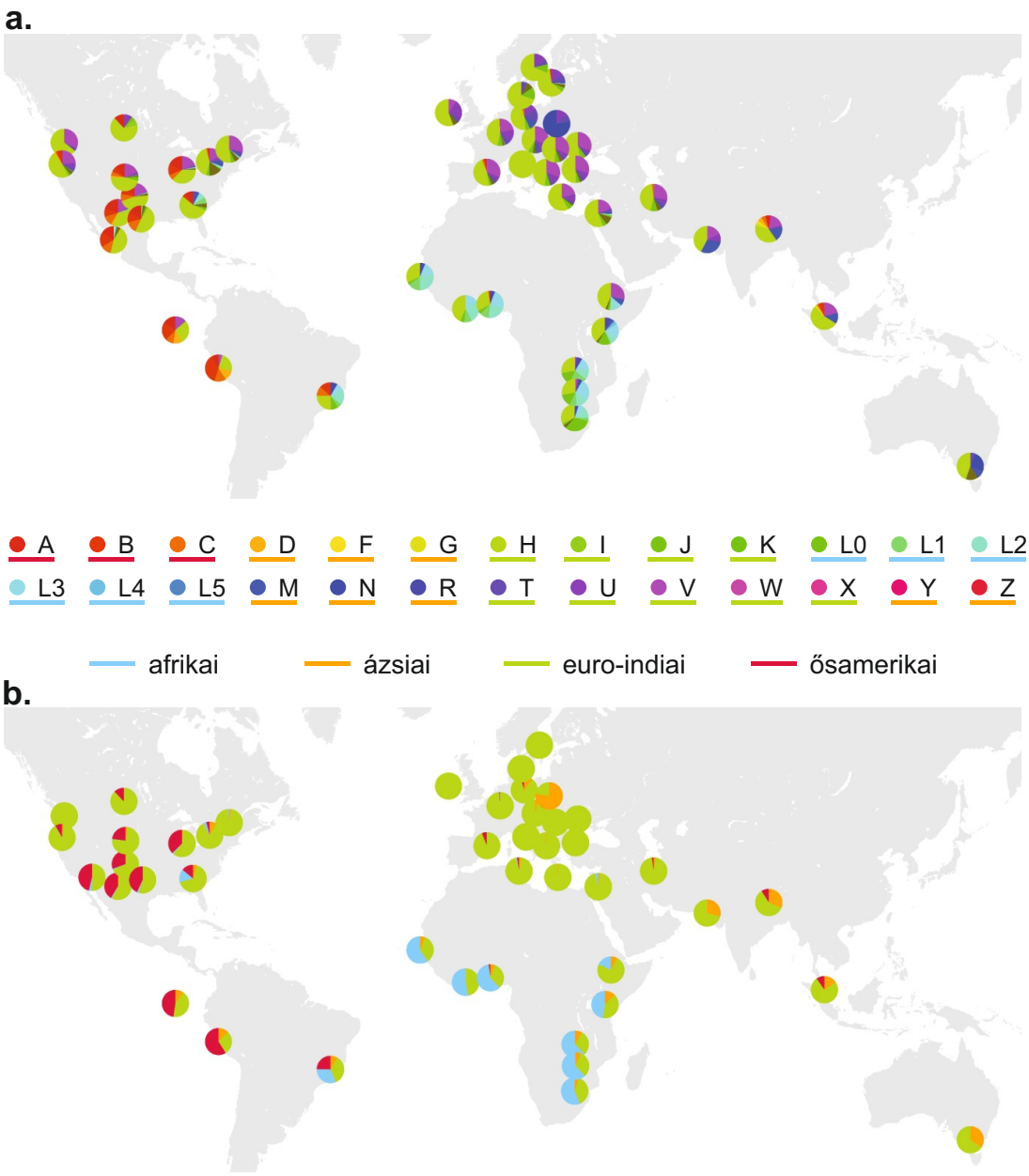
nem javasolják, így a szennyvízminták elemzése során ezt kikapcsoltuk a -V opció használatával. Ezek után megismételik a teljes EM-algoritmust az így leszűkített $G'$ haplotípus halmazzal. Végül a readeket a fennmaradó haplotípusokhoz az alapján rendelik hozzá, hogy melyikhez tartozott a legmagasabb feltételes valószínűség. Ha azonban egy read esetében a két legdominánsabb haplotípushoz tartozó feltételes valószínűségek aránya nem ér el egy előre definiált küszöbértéket (alapértelmezésben 2), a readet kategorizálatlannak tekintik. A végső haplotípus járulékokat a kategorizált readek arányai adják meg.

A kapott eredményeket a 24. ábrán kördiagramokon ábrázoltuk. Az irodalmi adatokkal való egyszerűbb összehasonlítás érdekében a haplotípusok járulékait haplocsoportonként összeadtuk, ezeket az ábrán külön színnel jelöltük. A lenti panelen a haplocsoportokat a még bővebb, földrajzi eredet szerinti kategóriákba vontuk össze, melyeket a haplocsoportokat jelző betűk alatt aláhúzással jelöltünk. Hogy segítsük a vizuális összehasonlítást a korábbi kutatások eredményeivel, a színeket azonosnak választottuk a [188] által a cikk 2. ábráján használt színekkel. Láthatóan a két ábra nagyon hasonló képet fest a lokális mtDNS haplocsoport eloszlásokra vonatkozóan, ami arra enged következtetni, hogy a szennyvízminták vizsgálatával hiteles eredmények kaphatók.

### Összevetés az irodalmi adatokkal

A mitokondriális haplocsoportok eloszlására vonatkozó korábbi kutatásokkal való részletesebb összevetéshez úgy igyekeztünk az adatokat gyűjteni, hogy mindig város-, de legalább régió-specifikus eredményekkel dolgozzunk. Amennyiben ez nem volt megvalósítható, az adott országra vonatkozó eloszlásokat ábrázoltuk az C3, C4, C5, C6. ábrákon. A függőleges szaggatott vonal bal oldalán a szennyvízmintákból származó, míg a jobb oldalon az irodalmi adatok találhatók. A szennyvízmintákat jellemző kördiagramok közepén a mixemt szoftver által sikeresen bekategorizált readek darabszáma, az irodalmi kördiagramok esetén pedig az adatok konkrét forrásának hivatkozása látható.

Az Egyesült Államokban a specifikus adatok hiánya miatt a városok etnikai összetételére vonatkozó cenzus adatokból indultunk ki. Ezeket [208] eredményei alapján átkonvertáltuk a többféle haplocsoportot magukba foglaló földrajzi eredet kategóriákba, melyeket a 24. ábrán az aláhúzások színei jelölnek. Mivel [208] nem tartalmazott arra vonatkozó információt, hogy az Amerikában élő ázsiai populáció milyen mitokondriális összetételű, ezért azzal a feltételezéssel éltünk, hogy minden önmagát ázsiai etnikumúnak valló ember az ázsiai földrajzi eredet csoportba sorolható az mtDNS haplocsoportja alapján is. Nyilvánvalóan ez a viszonylag megalapozatlan feltételezés némileg torzíthatja az eredményeket, de figyelembe véve, hogy az ázsiai populáció az amerikai városokban csak kis arányban van jelen, ezt a hatást elhanyagolhatónak tekintjük. A közvetlen összehasonlíthatóság kedvéért a szennyízminták haplocsoport összetételét is átkonvertáltuk a négy nagy földrajzi eredet kategóriára, ezek a C6. ábrán a szaggatott vonalról balra található kördiagramok belső sávján jelennek meg.




**24. ábra. Mitokondriális DNS haplocsoportok eloszlása az elemzett szennyvízmintákban. a.** Mitokondriális DNS haplocsoport eloszlás a különböző földrajzi területekről származó mintákban. A haplocsoportok mellett feltüntetett körök színei megegyeznek a kördiagramok színeivel. Az aláhúzások színei a bővebb földrajzi eredet kategóriákat jelzik. **b.** Mitokondriális DNS haplocsoport eloszlás a mintákban földrajzi eredet kategóriánként csoportosítva.

Általánosságban véve a szennyvízmintákból nyert haplocsoport összetétel meglepően jó egyezést mutat a specifikus populációkra koncentráló, időigényes és gondos mintagyűjtést igénylő kutatások eredményeivel. Értelemszerűen akadnak különbségek is, de ez a mintagyűjtés merőben más természete miatt elkerülhetetlen. A legtöbb korábbi kutatás a vizsgált populáció evolúciós történelmére koncentrál, ezért már a minták beszerzésénél is szigorú szűrési feltételeket állít. Ezzel szemben a szennyvízben minden típusú humán mtDNS keveredhet. Mint minden statisztikai jellegű elemzésnél, az irodalmi eloszlások

meghatározásához használt viszonylag kevés minta, illetve a szennyvízmintákban található alacsony mtDNS koncentráció is hozzájárulhat a megfigyelt különbségekhez. Egy további bizonytalanságot adó faktor a keverék DNS dekompozíciójához használt mixemt szoftver hatékonyságának a romlása sok mtDNS haplocsoport keveredése esetén [206]. Az eredeti cikk alapján még három haplotípus esetén is az esetek többségében helyesen azonosítható a csupán nyomokban (5%-os arányban) megjelenő haplotípus is, és csak ritkán fordul elő, hogy azt hibásan egy hasonló haplotípusként azonosítja a program. Mivel az elemzés során a haplocsoportokra koncentráltunk, a haplotípusok ilyen esetleges felcserélése nem feltétlenül torzítja az eredményeket. Fontos azt is megjegyezni, hogy az Egyesült Államok városai esetén alkalmazott földrajzi eredet kategóriák létjogosultságát számos publikáció [209–213] kétségbe vonja, bár használatuk mind a tudományos irodalomban [214], mind a kereskedelmi eredet tesztelésben [215] széles körben elterjedt. A módszert érő kritikák ellenére az alapvető tendenciák konzisztensen megjelennek a szennyvízmintákban csak nyomokban megtalálható mtDNS elemzése során is.

**Reprodukálhatóság**

Annak érdekében, hogy képet kapjunk az eredmények reprodukálhatóságát illetően, az egy városból (El Paso, USA), de különböző szennyvíztisztító telepekről származó 4 mintából kapott eloszlásokat egymás mellett ábrázoltuk az C6. ábrán, az C5. ábrán pedig a két zambiai város, Kitwe és Lusaka eredményeit tüntettük fel. Az ábrák egyszerű manuális összehasonlításával látható, hogy az azonos környékről származó mintákban a haplocsoportok eloszlása rendkívül hasonló. Figyelembe véve, hogy a humán mitokondriális haplotípusok fáján [207] a H és V haplocsoportok feltűnően közel helyezkednek el egymáshoz, az El Pasonál feltüntetett kördiagramok még hasonlóbbá válnak.

Ez arra utal, hogy a mitokondriális haplocsoportok eloszlásának rekonstrukciója a szennyvízmintákból nem csak olyan eredményekre vezet, melyek az irodalmi adatokkal megegyeznek, hanem amelyek robusztusak és reprodukálhatóak is.

### Az eredmények jelentősége

A populációgenetikával foglalkozó kutatások jelentős része az eredetileg ott élő populációk feltérképezésére vagy ősi evolúciós és migrációs mintázatok felderítésére irányul, eközben pedig a lehető legjobban igyekszik a helyi kisebbségek, ideiglenesen ott élő külföldi munkások, bevándorlók és turisták okozta genetikai zaj kiküszöbölésére. Mindazonáltal nagy haszna lenne a lokális populációt valóban jellemző adatoknak is a demográfiai tendenciák, egészségügyi és az ehhez kapcsolódó szociális és gazdasági irányvonalak nyomon követéséhez.

Sok publikáció igazolta, hogy a különböző mitokondriális haplocsoportokban különböző gyakorisággal fordulnak elő bizonyos betegségek és egészségügyi kórképek, többek között a koszorúér-betegség, a diabéteszes retinopátia, a korai Alzheimer-kór, a frontotem-

porális demencia, az AIDS progresszió, illetve az emlő-, prosztata- és vesedaganat [216–220]. A haplocsoportok és a betegségek között megfigyelt korrelációkat már a klinikai gyakorlatban is biomarkerként vagy a páciensek kategorizálását segítő faktorként használják [221]. Ezek a tények azt sejtetik, hogy az mtDNS haplocsoportok széleskörű vizsgálata a különböző földrajzi területeken elősegíthetné a hatékonyabb, a helyi viszonyokhoz optimalizált prevenciós stratégiák kidolgozását.

Mivel a vizsgált minták preparálása a metagenomikai elemzésekhez optimalizáltan történt, felmerül annak a lehetősége, hogy ha specifikusan az emberi DNS kinyerését tűznénk ki célul, még gazdagabb adathalmazra tehetnénk szert. Számos genetikai betegség egyértelmű összefüggésben áll bizonyos pontmutációk, illetve inzerciók és deléciók meglétével, melyek nem azonos valószínűséggel jelennek meg a különböző populációkban [222–224]. Amennyiben lehetőség nyílna az ilyen variánsok gyakoriságának elemzésére a lokális közösségekben, egy hatalmas lépéssel közelebb kerülnénk az egyénekre optimalizált egészségügyi eljárások megvalósításához.

Mindezen előnyök ellenére is sok aggály övezi a genetikai vagy akár etnikai monitorozás bevezetését. Amellett, hogy sokan már magát az adatgyűjtést is diszkriminatívnak tartják, az ilyen érzékeny adatok tárolásához és kezeléséhez szükséges infrastruktúra még az esetek többségében nem áll rendelkezésre. A szennyvízminták elemzése ilyen szempontból egy nagyon ígéretes alternatíva: a mintagyűjtés nem jár invazív beavatkozással, nem igényli a résztvevők explicit beleegyezését, nem önbevallásos alapon történik, nem követeli meg a közösség aktív részvételét, és természeténél fogva egy erősen kevert mintát eredményez. Emellett, mivel a szennyvízgyűjtés nem igényel hosszas előkészületeket, illetve nagyösszegű beruházást, lehetőség nyílna a populációk időbeli, dinamikus nyomon követésére is.

# ÖSSZEGZÉS

A fentiekben áttekintettük a személyre szabott gyógyászat eszméje által támasztott kihívásokat mind az adatok megszerzése, mind pedig azok feldolgozása szempontjából.

Tüdő adenokarcinómában szenvedő betegek metasztázisainak vizsgálata során kimutattam, hogy a centrális primer tüdőtumorok a periférikusaknál agresszívebbek és hajlamosabbak a korai áttétképzésre. Emellett megmutattam, hogy bizonyos szervpárok esetén a mindkettőben megjelenő metasztázisok gyakorisága eltérő attól, mint ami a függetlenség feltételezéséből adódna. Továbbá megállapítottam, hogy a különböző szerveket érintő áttétek sorrendisége sem véletlenszerű.

A csontáttéttel rendelkező primer tüdődaganatos betegek csoportján felállított túlélési modell segítségével megállapítottam, hogy a késői stádiumú primer tumorok esetén, a magas vérnyomástól szenvedő pácienseknél, illetve a csontáttét megjelenését követően a kezdetben normális tartományba eső vesefunkciókat jellemző paraméterek kóros tartományba lépése várhatóan rövidebb idő elteltével következik be.

Az immunterápiás biomarkerek elemzése során kimutattam, hogy a daganat környéki immunsejtek jelenlétének prognosztikus hatása van primer tüdődaganatos betegeknél. Megállapítottam továbbá, hogy a PD-L1 fehérjét kifejező tumorsejtek aránya neoadjuváns, platina-bázisú kemoterápia hatására csak elvétve nő és leggyakrabban nem változik. A lepidikus tumornövekedési mintázat és az immunsejtek PD-1 és PD-L1, illetve a tumorsejtek PD-L1 expressziós szintjei között demonstráltam a negatív korrelációt. A tüdőtumor és annak agyi metasztázisa között a tumorsejtek PD-L1 expressziójára vonatkozóan erős pozitív korrelációt találtam.

Bemutattam egy új, NGS adatok elemzésére alkalmas, a legnépszerűbb módszereknél nagyságrendekkel gyorsabb és sok esetben precízebb mutáció-detektáló algoritmust, mellyel a pontmutációk, inzerciók és deléciók azonosítása mellett lehetőség nyílik a minták kariotípusának vizsgálatára, optimalizációs lépések elvégzésére, illetve számos utólagos elemzési és ábrázolási opció közül választhatunk.

Többféle sejtvonalon végzett kísérlet során szemléltettem a módszer hatékonyságát és segítségével kimutattuk a BRCA1 és BRCA2 gének mutációs spektrumra gyakorolt hatását, különböző citotoxikus kezelések és a PARP-inhibitor terápia mutagén következményeit, és demonstráltuk, hogy a genomi elváltozások alapján miként lehet a tumor és metasztázisainak evolúciójára következtetni.

Végül a személyre szabott gyógyászat irányába tett köztes lépésként bemutattam, hogy környezeti minták vizsgálata során miként lehet populáció-szintű genetikai következtetéseket levonni. Az különböző városokból származó szennyvízmintákból reprodukálhatóan, és a korábbi adatokkal összevethetően megállapítottam a humán mitokondriális haplocsoportok lokális eloszlását.

# SUMMARY

In our work, we have explored the challenges of data acquisition and processing set by the idealistic, but nonetheless promising concept of personalised medicine.

By examining the metastases of patients suffering from lung adenocarcinoma, I have shown that primary tumours with a central localisation are more aggressive than peripheral ones and more likely to promote early metastases. Moreover, within a single patient, the frequency of metastases in specific organ pairs tend to differ from what would be expected by assuming independence. The order in which metastases of various organs appear was also not random.

For patients with primary lung cancer and bone metastases, I have defined an approximate survival model and shown that initially normal renal function parameters are likely to deteriorate into the pathological range sooner in cases of primary tumours in later stages, hypertension and after the appearance of the bone metastasis.

While investigating immunotherapeutic biomarkers, I have established the prognostic role of the presence of peritumoral immune cells. I have also shown that the ratio of PD-L1 expressing tumour cells only rarely increases and mostly stays stable during neoadjuvant, platinum-based chemotherapy. I have demonstrated the negative correlation between lepidic growth pattern and the PD-1 and PD-L1 expression of immune cells, moreover the PD-L1 expression of tumour cells. I have discovered a strong positive correlation in the PD-L1 expression levels of tumour cells between the primary lung tumour and its brain metastasis.

I have designed a new mutation detection algorithm to analyse NGS data, which is substantially faster and in many cases more precise than state of art tools. Besides recovering SNVs and indels, the software can estimate the karyotypes of the samples, uses automatic optimization steps to finetune the results and offers a wide range of post-processing and visualization options.

I have demonstrated the easy usability of the tool on the results of multiple cell line experiments and showed the effect of BRCA1 and BRCA2 gene loss on mutation spectra, determined the mutagenic consequences of many cytotoxic and PARP inhibitor therapies and illustrated how genomic aberrations can be used to uncover the evolution of a tumour and its metastases.

Finally, as an intermediate step towards personalised medicine, I have shown how to use environmental samples to draw population-level genetic conclusions. From wastewater samples collected from different cities, I was able to determine the local mitochondrial haplogroup composition with good reproducibility and reasonable agreement with previous literary data.

# SAJÁT PUBLIKÁCIÓK

## AZ ÉRTEKEZÉS ALAPJÁUL SZOLGÁLÓ KÖZLEMÉNYEK

Klikovits T., Lohinai Z., Fábián K., Gyulai M., Szilasi M., Varga J., Baranya E., Pipek O., Csabai I., Szállási Z., Tímár J., Hoda MA., Laszlo V., Hegedűs B., Renyi-Vamos F., Klepetko W., Ostoros G., Döme B., Moldvay J.: New insights into the impact of primary lung adenocarcinoma location on metastatic sites and sequence: A multicenter cohort study. *Lung Cancer* **126,** 139–148. ISSN: 18728332 (2018). (IF: 4,599; idézők: 2 (független: 2))

Fábián K., Puskás R., Kakuk T., Prés L., Fejes D., Szegedi Z., Rojkó L., Szállási Z., Döme B., Pipek O., Moldvay J.: Renal Impairment Hampers Bisphosphonate Treatment in a Quarter of Lung Cancer Patients with Bone Metastasis. *Basic and Clinical Pharmacology and Toxicology* **122,** 126–132. ISSN: 17427843 (2018). (IF: 2,452)

Téglási V., Reiniger L., Fábián K., Pipek O., Csala I., Bagó AG., Várallyai P., Vízkeleti L., Rojkó L., Tímár J., Döme B., Szállási Z., Swanton C., Moldvay J.: Evaluating the significance of density, localization, and PD-1/PD-L1 immunopositivity of mononuclear cells in the clinical course of lung adenocarcinoma patients with brain metastasis. *Neuro-Oncology* **19,** 1058–1067. ISSN: 15235866 (2017). (IF: 10,091; idézők: 8 (független: 6))

Rojkó L., Reiniger L., Téglási V., Fábián K., Pipek O., Vágvölgyi A., Agócs L., Fillinger J., Kajdácsi Z., Tímár J., Döme B., Szállási Z., Moldvay J.: Chemotherapy treatment is associated with altered PD-L1 expression in lung cancer patients. *Journal of Cancer Research and Clinical Oncology* **144,** 1219–1226. ISSN: 14321335 (2018). (IF: 3,332; idézők: 6 (független: 5))

Reiniger L., Téglási V., Pipek O., Rojkó L., Glasz T., Vágvölgyi A., Kovalszky I., Gyulai M., Lohinai Z., Rásó E., Tímár J., Döme B., Szállási Z., Moldvay J.: Tumor necrosis correlates with PD-L1 and PD-1 expression in lung adenocarcinoma. *Acta Oncologica* **58,** 1087–1094. ISSN: 1651226X (2019). (IF: 3,298; idézők: 1 (független: 1))

Téglási V., Pipek O., Lózsa R., Berta K., Szüts D., Harkó T., Vadász P., Rojkó L., Döme B., Bagó AG., Tímár J., Moldvay J., Szállási Z., Reiniger L.: PD-L1 expression of lung cancer cells, unlike infiltrating immune cells, is stable and unaffected by therapy during brain metastasis. *Clinical Lung Cancer* **20,** 363–369. ISSN: 19380690 (2019). (IF: 4,117)

Pipek O., Ribli D., Molnár J., Póti Á., Krzystanek M., Bodor A., Tusnády GE., Szallasi Z., Csabai I., Szüts D.: Fast and accurate mutation detection in whole genome sequences of multiple isogenic samples with IsoMut. *BMC Bioinformatics* **18,** 1–11. ISSN: 1471-2105 (2017). (IF: 2,511; idézők: 7 (független: 3))

Németh E., Krzystanek M., Reiniger L., Ribli D., Pipek O., Sztupinszki Z., Glasz T.,

Csabai I., Moldvay J., Szallasi Z., Szüts D.: The genomic imprint of cancer therapies helps timing the formation of metastases. *International Journal of Cancer* **145,** 694–704. ISSN: 10970215 (2019). (IF: 4,982)

Zámborszky J., Szikriszt B., Gervai JZ., Pipek O., Póti Á., Krzystanek M., Ribli D., Szalai-Gindl JM., Csabai I., Szallasi Z., Swanton C., Richardson AL., Szüts D.: Loss of BRCA1 or BRCA2 markedly increases the rate of base substitution mutagenesis and has distinct effects on genomic deletions. *Oncogene* **36(6),** 746-755. ISSN: 1476-5594. (2016). (IF: 6,634; idézők: 23 (független: 19))

Szikriszt B., Póti Á., Pipek O., Krzystanek M., Kanu N., Molnár J., Ribli D., Szeltner Z., Tusnády GE., Csabai I., Szallasi Z., Swanton C., Szüts D.: A comprehensive survey of the mutagenic impact of common cancer cytotoxics. *Genome Biology* **17:** 99. (2016). (IF: 14,028; idézők: 59 (független: 54))

Póti Á., Berta K., Xiao Y., Pipek O., Klus GT., Ried T., Csabai I., Wilcoxen K., Mikule K., Szallasi Z., Szüts D.: Long-term treatment with the PARP inhibitor niraparib does not increase the mutation load in cell line models and tumour xenografts. *British Journal of Cancer* **119,** 1392–1400. ISSN:15321827 (2018). (IF: 5,416; idézők: 1 (független: 1))

Pipek OA., Medgyes-Horváth A., Dobos L., Stéger J., Szalai-Gindl J., Visontai D., Kaas RS., Koopmans M., Hendriksen RS., Aarestrup FM., Csabai I.: Worldwide human mitochondrial haplogroup distribution from urban sewage. *Scientific Reports* **9,** 11624. ISSN: 2045-2322 (2019). (IF: 4,011)

### Egyéb közlemények

Molnár J., Póti Á., Pipek O., Krzystanek M., Kanu N., Swanton C., Tusnády GE., Szallasi Z., Csabai I., Szüts D.: The Genome of the Chicken DT40 Bursal Lymphoma Cell Line. *G3: GENES, GENOMES, GENETICS* **4(11),** 2231-2240. (2014). (IF: 2,742; idézők: 11 (független: 6))

Turajlic S., Xu H., Litchfield K., Rowan A., Horswell S., Chambers T., O'Brien T., Lopez JI., Watkins TBK., Nicol D., Stares M., Challacombe B., Hazell S., Chandra A., Mitchell TJ., Au L., Eichler-Jonsson C., Jabbar F., Soultati A., Chowdhury S., Rudman S., Lynch J., Fernando A., Stamp G., Nye E., Stewart A., Xing W., Smith JC., Escudero M., Huffman A., Matthews N., Elgar G., Phillimore B., Costa M., Begum S., Ward S., Salm M., Boeing S., Fisher R., Spain L., Navas C., Grönroos E., Hobor S., Sharma S., Aurangzeb I., Lall S., Polson A., Varia M., Horsfield C., Fotiadis N., Pickering L., Schwarz RF., Silva B., Herrero J., Luscombe NM., Jamal-Hanjani M., Rosenthal R., Birkbak NJ., Wilson GA., Pipek O., Ribli D., Krzystanek M., Csabai I., Szallasi Z., Gore M., McGranahan N., Van Loo P., Campbell P., Larkin J., Swanton C., the TRACERx Renal Consortium.: Deterministic Evolutionary Trajectories Influence Primary Tumor Growth: TRACERx Renal. *Cell* **173(3),** 595-610.e11 (2018). (IF: 36,216; idézők: 43 (független: 40))

# HIVATKOZÁSOK


1. Mathur, S., Sutton, J., Sutton, J. & Sutton, J. Personalized medicine could transform healthcare. *Biomedical Reports* **7,** 3–5. ISSN: 2049-9434 (July 2017).
2. Tremblay, J. & Hamet, P. Role of genomics on the path to personalized medicine. *Metabolism: Clinical and Experimental* **62,** S2–S5. ISSN: 00260495 (2013).
3. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921. ISSN: 00280836 (Feb. 2001).
4. National Human Genome Research Institute. *DNA Sequencing Costs: Data | NHGRI* 2018. https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data (2019).
5. Heather, J. M. & Chain, B. The sequence of sequencers: The history of sequencing DNA. *Genomics* **107,** 1–8. ISSN: 1089-8646 (Jan. 2016).
6. McPherson, E. Genetic diagnosis and testing in clinical practice. *Clinical medicine & research* **4,** 123–9. ISSN: 1539-4182 (June 2006).
7. Atkinson, A. J. *et al. Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework* Mar. 2001. doi:10.1067/mcp.2001.113989. http://doi.wiley.com/10.1067/mcp.2001.113989.
8. Walser, T. *et al.* Smoking and lung cancer: the role of inflammation. *Proceedings of the American Thoracic Society* **5,** 811–5. ISSN: 1546-3222 (Dec. 2008).
9. Agata, Y. *et al.* Expression of the PD-1 antigen on the surface of stimulated mouse T and B lymphocytes. *International Immunology* **8,** 765–772. ISSN: 0953-8178 (May 1996).
10. Ishida, Y., Agata, Y., Shibahara, K. & Honjo, T. Induced expression of PD-1, a novel member of the immunoglobulin gene superfamily, upon programmed cell death. *The EMBO journal* **11,** 3887–95. ISSN: 0261-4189 (Nov. 1992).
11. Freeman, G. J. *et al.* Engagement of the PD-1 immunoinhibitory receptor by a novel B7 family member leads to negative regulation of lymphocyte activation. *The Journal of experimental medicine* **192,** 1027–34. ISSN: 0022-1007 (Oct. 2000).
12. Fésüs, V. Az immunonkológia újdonságai a szolid tumorok és a hematológiai daganatok kezelésében – az immunellenőrzőpont-gátlók. *Magyar Onkológia,* 116–125 (2017).
13. McGlynn, P. & Lloyd, R. G. Recombinational repair and restart of damaged replication forks. *Nature Reviews Molecular Cell Biology* **3,** 859–870. ISSN: 1471-0072 (Nov. 2002).
14. Alexandrov, L. B., Nik-zainal, S., Wedge, D. C. & Aparicio, S. A. J. R. Signatures of mutational processes in human cancer. *Nature* **500,** 415–421 (2014).
15. Chan, I. S. & Ginsburg, G. S. Personalized Medicine: Progress and Promise. *Annual Review of Genomics and Human Genetics* **12,** 217–244. ISSN: 1527-8204 (2011).
16. Wan, J. C. *et al. Liquid biopsies come of age: Towards implementation of circulating tumour DNA* Apr. 2017. doi:10.1038/nrc.2017.7. http://www.nature.com/articles/nrc.2017.7.
17. Jamal-Hanjani, M. *et al.* Tracking Genomic Cancer Evolution for Precision Medicine: The Lung TRACERx Study. *PLoS Biology* **12,** 1–7. ISSN: 15457885 (July 2014).
18. De Paor, A. Genetic Discrimination: A Case for a European Legislative Response? *European Journal of Health Law* **24,** 135–159. ISSN: 15718093 (2017).
19. Almqvist, E. W., Bloch, M., Brinkman, R., Craufurd, D. & Hayden, M. R. A worldwide assessment of the frequency of suicide, suicide attempts, or psychiatric hospitalization after predictive testing for Huntington disease. *American Journal of Human Genetics* **64,** 1293–1304. ISSN: 00029297 (May 1999).
20. Klikovits, T. *et al.* New insights into the impact of primary lung adenocarcinoma location on metastatic sites and sequence: A multicenter cohort study. *Lung Cancer* **126,** 139–148. ISSN: 18728332 (2018).
21. Kinsey, C. M. *et al.* Invasive adenocarcinoma of the lung is associated with the upper lung regions. *Lung Cancer* **84,** 145–150. ISSN: 18728332 (May 2014).

22. Tseng, C. H. *et al.* EGFR mutation and lobar location of lung adenocarcinoma. *Carcinogenesis* **37,** 157–162. ISSN: 14602180 (Feb. 2015).

23. Byers, T. E., Vena, J. E. & Rzepka, T. F. Predilection of lung cancer for the upper lobes: An epidemiologic inquiry. *Journal of the National Cancer Institute* **72,** 1271–1275. ISSN: 14602105 (June 1984).

24. Sun, W., Yang, X., Liu, Y., Yuan, Y. & Lin, D. Primary Tumor Location Is a Useful Predictor for Lymph Node Metastasis and Prognosis in Lung Adenocarcinoma. *Clinical Lung Cancer* **18,** e49–e55. ISSN: 19380690 (Jan. 2017).

25. Mujoomdar, A. *et al.* Clinical predictors of metastatic disease to the brain from non-small cell lung carcinoma: Primary tumor size, cell type, and lymph node metastases. *Radiology* **242,** 882–888. ISSN: 00338419 (Mar. 2007).

26. Fábián, K. *et al.* Significance of primary tumor location and histology for brain metastasis development and peritumoral brain edema in lung cancer. *Oncology (Switzerland)* **91,** 237–242. ISSN: 14230232 (2016).

27. Bondar, J. & Putter, J. Simultaneous Statistical Inference. *Technometrics* **10,** 415–416. ISSN: 15372723 (1968).

28. Tai, K. H. & Foroudi, F. in *Prostate Cancer: A Comprehensive Perspective* 6, 1055–1063 (Dec. 2013). ISBN: 9781447128649. doi:10.1007/978-1-4471-2864-9{_}87, http://www.ncbi.nlm.nih.gov/pubmed/11110597%20http://theoncologist.alphamedpress.org/cgi/doi/10.1634/theoncologist.5-6-463.

29. Gnant, M., Dubsky, P. & Hadji, P. in *Recent results in cancer research. Fortschritte der Krebsforschung. Progres dans les recherches sur le cancer* 65–91 (2012). doi:10.1007/978-3-642-21892-7{_}3, http://www.ncbi.nlm.nih.gov/pubmed/22307370%20http://link.springer.com/10.1007/978-3-642-21892-7_3.

30. Chang, J. T. *et al. Renal Failure with the Use of Zoledronic Acid [7] (multiple letters)* Oct. 2003. doi:10.1056/NEJM200310233491721, http://www.ncbi.nlm.nih.gov/pubmed/14573746%20http://www.nejm.org/doi/abs/10.1056/NEJM200310233491721.

31. Perazella, M. A. & Markowitz, G. S. *Bisphosphonate nephrotoxicity* Dec. 2008. doi:10.1038/ki.2008.356, http://www.ncbi.nlm.nih.gov/pubmed/18685574%20https://linkinghub.elsevier.com/retrieve/pii/S0085253815532006.

32. Scagliotti, G. V. *et al.* Overall survival improvement in patients with lung cancer and bone metastases treated with denosumab versus zoledronic acid: Subgroup analysis from a randomized phase 3 study. *Journal of Thoracic Oncology* **7,** 1823–1829. ISSN: 15560864 (Dec. 2012).

33. Toffart, A. C., Belaiche, S., Moro-Sibilot, D., Couraud, S. & Sakhri, L. *Impact of lung cancer treatments on renal function* Dec. 2014. doi:10.1016/j.rmr.2014.03.008, http://www.ncbi.nlm.nih.gov/pubmed/25496793.

34. Fábián, K. *et al.* Renal Impairment Hampers Bisphosphonate Treatment in a Quarter of Lung Cancer Patients with Bone Metastasis. *Basic and Clinical Pharmacology and Toxicology* **122,** 126–132. ISSN: 17427843 (Jan. 2018).

35. Kaplan, E. L. & Meier, P. Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association* **53,** 457–481. ISSN: 1537274X (June 1958).

36. Akaike, H. A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control* **19,** 716–723. ISSN: 15582523 (Dec. 1974).

37. Zhang, Z. Semi-parametric regression model for survival data: Graphical visualization with R. *Annals of Translational Medicine* **4,** 461. ISSN: 23055847 (Dec. 2016).

38. Harrell, F. E., Lee, K. L. & Mark, D. B. Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine* **15,** 361–387. ISSN: 02776715 (Feb. 1996).

39. Reck, M. *et al.* Pembrolizumab versus Chemotherapy for PD-L1-Positive Non-Small-Cell Lung Cancer. *New England Journal of Medicine* **375,** 1823–1833. ISSN: 15334406 (Nov. 2016).

40. Spira, A. I. *et al.* Efficacy, safety and predictive biomarker results from a randomized phase II study comparing MPDL3280A vs docetaxel in 2L/3L NSCLC (POPLAR). *Journal of Clinical Oncology* **33,** 8010–8010. ISSN: 0732-183X (May 2015).

41. Menon, S., Shin, S. & Dy, G. *Advances in cancer immunotherapy in solid tumors* Nov. 2016. doi:10.3390/cancers8120106. http://www.mdpi.com/2072-6694/8/12/106.

42. Kerr, K. M. & Nicolson, M. C. Non-small cell lung cancer, PD-L1, and the pathologist. *Archives of Pathology and Laboratory Medicine* **140,** 249–254. ISSN: 15432165 (Mar. 2016).

43. Herbst, R. S. *et al.* Predictive correlates of response to the anti-PD-L1 antibody MPDL3280A in cancer patients. *Nature* **515,** 563–567. ISSN: 14764687 (Nov. 2014).

44. Yang, C. Y., Lin, M. W., Chang, Y. L., Wu, C. T. & Yang, P. C. Programmed cell death-ligand 1 expression is associated with a favourable immune microenvironment and better overall survival in stage i pulmonary squamous cell carcinoma. *European Journal of Cancer* **57,** 91–103. ISSN: 18790852 (Apr. 2016).

45. Festino, L. *et al. Cancer Treatment with Anti-PD-1/PD-L1 Agents: Is PD-L1 Expression a Biomarker for Patient Selection?* June 2016. doi:10.1007/s40265-016-0588-x. http://link.springer.com/10.1007/s40265-016-0588-x.

46. Scheel, A. H. *et al.* Harmonized PD-L1 immunohistochemistry for pulmonary squamous-cell and adenocarcinomas. *Modern Pathology* **29,** 1165–1172. ISSN: 15300285 (Oct. 2016).

47. Huynh, T. G. *et al.* Programmed cell death ligand 1 expression in resected lung adenocarcinomas: Association with immune microenvironment. *Journal of Thoracic Oncology* **11,** 1869–1878. ISSN: 15561380 (Nov. 2016).

48. Rebelatto, M. C. *et al.* Development of a programmed cell death ligand-1 immunohistochemical assay validated for analysis of non-small cell lung cancer and head and neck squamous cell carcinoma. *Diagnostic Pathology* **11,** 95. ISSN: 17461596 (Dec. 2016).

49. Roach, C. *et al.* Development of a Companion Diagnostic PD-L1 Immunohistochemistry Assay for Pembrolizumab Therapy in Non-Small-cell Lung Cancer. *Applied Immunohistochemistry and Molecular Morphology* **24,** 392–397. ISSN: 15334058 (July 2016).

50. Vennapusa, B. *et al.* Development of a PD-L1 Complementary Diagnostic Immunohistochemistry Assay (SP142) for Atezolizumab. *Applied immunohistochemistry & molecular morphology : AIMM* **27,** 92–100. ISSN: 1533-4058 (Feb. 2019).

51. Munari, E. *et al.* PD-L1 expression heterogeneity in non-small cell lung cancer: Evaluation of small biopsies reliability. *Oncotarget* **8,** 90123–90131. ISSN: 19492553 (Oct. 2017).

52. Téglási, V. *et al.* Evaluating the significance of density, localization, and PD-1/PD-L1 immunopositivity of mononuclear cells in the clinical course of lung adenocarcinoma patients with brain metastasis. *Neuro-Oncology* **19,** 1058–1067. ISSN: 15235866 (2017).

53. Peters, S., Bexelius, C., Munk, V. & Leighl, N. *The impact of brain metastasis on quality of life, resource utilization and survival in patients with non-small-cell lung cancer* Apr. 2016. doi:10.1016/j.ctrv.2016.03.009. http://www.ncbi.nlm.nih.gov/pubmed/27019457%20https://linkinghub.elsevier.com/retrieve/pii/S0305737216000426.

54. Berghoff, A. S., Lassmann, H., Preusser, M. & Höftberger, R. Characterization of the inflammatory response to solid cancer metastases in the human brain. *Clinical and Experimental Metastasis* **30,** 69–81. ISSN: 02620898 (Jan. 2013).

55. Berghoff, A. S. *et al.* Density of tumor-infiltrating lymphocytes correlates with extent of brain edema and overall survival time in patients with brain metastases. *OncoImmunology* **5,** e1057388. ISSN: 2162402X (Jan. 2016).

56. Harter, P. N. *et al.* Distribution and prognostic relevance of tumor-infiltrating lymphocytes (TILs) and PD-1/PD-L1 immune checkpoints in human brain metastases. *Oncotarget* **6,** 40836–40849. ISSN: 19492553 (Dec. 2015).

57. Antonia, S. J., Vansteenkiste, J. F. & Moon, E. Immunotherapy: Beyond Anti–PD-1 and Anti–PD-L1 Therapies. *American Society of Clinical Oncology Educational Book* **35,** e450–e458. ISSN: 1548-8756 (May 2016).

58. Rojkó, L. *et al.* Chemotherapy treatment is associated with altered PD-L1 expression in lung cancer patients. *Journal of Cancer Research and Clinical Oncology* **144,** 1219–1226. ISSN: 14321335 (2018).

59. Reiniger, L. *et al.* Tumor necrosis correlates with PD-L1 and PD-1 expression in lung adenocarcinoma. *Acta Oncologica* **58,** 1087–1094. ISSN: 1651226X (2019).

60. Téglási, V. *et al.* PD-L1 Expression of Lung Cancer Cells, Unlike Infiltrating Immune Cells, Is Stable and Unaffected by Therapy During Brain Metastasis. *Clinical Lung Cancer.* ISSN: 19380690. doi:10.1016/j.cllc.2019.05.008, https://doi.org/10.1016/j.cllc.2019.05.008 (2019).

61. Sheng, J. *et al.* Expression of programmed death ligand-1 on tumor cells varies pre and post chemotherapy in non-small cell lung cancer. *Scientific Reports* **6,** 20090. ISSN: 20452322 (Apr. 2016).

62. Zhang, P. *et al.* Upregulation of programmed cell death ligand 1 promotes resistance response in non-small-cell lung cancer patients treated with neo-adjuvant chemotherapy. *Cancer Science* **107,** 1563–1571. ISSN: 13497006 (Nov. 2016).

63. Hirsch, F. R. *et al.* PD-L1 Immunohistochemistry Assays for Lung Cancer: Results from Phase 1 of the Blueprint PD-L1 IHC Assay Comparison Project. *Journal of Thoracic Oncology* **12,** 208–222. ISSN: 15561380 (Feb. 2017).

64. Araki, K. *et al.* Excellent prognosis of lepidic-predominant lung adenocarcinoma: low incidence of lymphatic vessel invasion as a key factor. *Anticancer research* **34,** 3153–6. ISSN: 1791-7530 (June 2014).

65. Tsao, M. S. *et al.* Subtype classification of lung adenocarcinoma predicts benefit from adjuvant chemotherapy in patients undergoing complete resection. *Journal of Clinical Oncology* **33,** 3439–3446. ISSN: 15277755 (Oct. 2015).

66. Saruwatari, K. *et al.* Aggressive tumor microenvironment of solid predominant lung adenocarcinoma subtype harboring with epidermal growth factor receptor mutations. *Lung Cancer* **91,** 7–14. ISSN: 18728332 (Jan. 2016).

67. Yeo, M. K. *et al.* Association of PD-L1 expression and PD-L1 gene polymorphism with poor prognosis in lung adenocarcinoma and squamous cell carcinoma. *Human Pathology* **68,** 103–111. ISSN: 15328392 (Oct. 2017).

68. Toyokawa, G. *et al.* Relevance Between Programmed Death Ligand 1 and Radiologic Invasiveness in Pathologic Stage I Lung Adenocarcinoma. *Annals of Thoracic Surgery* **103,** 1750–1757. ISSN: 15526259 (June 2017).

69. Holm, S. A simple sequential rejective method procedure. *Scandinavian Journal of Statistics* **6,** 65–70 (1979).

70. Jiang, L. *et al.* PD-L1 expression and its relationship with oncogenic drivers in non-small cell lung cancer (NSCLC). *Oncotarget* **8,** 26845–26857. ISSN: 19492553 (Apr. 2017).

71. Kim, S. *et al.* Comparative analysis of PD-L1 expression between primary and metastatic pulmonary adenocarcinomas. *European Journal of Cancer* **75,** 141–149. ISSN: 18790852 (Apr. 2017).

72. Mansfield, A. S. *et al.* Temporal and spatial discordance of programmed cell death-ligand 1 expression and lymphocyte tumor infiltration between paired primary lesions and brain metastases in lung cancer. *Annals of Oncology* **27,** 1953–1958. ISSN: 15698041 (Oct. 2016).

73. Takamori, S. *et al.* Discrepancy in programmed cell death-ligand 1 between primary and metastatic non-small cell lung cancer. *Anticancer Research* **37,** 4223–4228. ISSN: 17917530 (Aug. 2017).

74. Malhotra, J., Jabbour, S. K. & Aisner, J. *Current state of immunotherapy for non-small cell lung cancer* Apr. 2017. doi:10.21037/tlcr.2017.03.01, http://www.ncbi.nlm.nih.gov/pubmed/28529902%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5420529%20http://tlcr.amegroups.com/article/view/13072/10996.

75. Shendure, J. *et al.* DNA sequencing at 40: Past, present and future. *Nature* **550,** 345–353. ISSN: 14764687 (Oct. 2017).

76. Holley, R. W. *et al.* Structure of a ribonucleic acid. *Science* **147,** 1462–1465. ISSN: 00368075 (Mar. 1965).

77. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America* **74,** 5463–5467. ISSN: 00278424 (Dec. 1977).

78. Heather, J. M. & Chain, B. *The sequence of sequencers: The history of sequencing DNA* Jan. 2016. doi:10.1016/j.ygeno.2015.11.003. https://www.sciencedirect.com/science/article/pii/S0888754315300410.

79. Maxam, A. M. & Gilbert, W. A new method for sequencing DNA. *Proceedings of the National Academy of Sciences of the United States of America* **74,** 560–564. ISSN: 00278424 (Feb. 1977).

80. Smith, L. M., Fung, S., Hunkapiller, M. W., Hunkapiller, T. J. & Hood, L. E. The synthesis of oligonucleotides containing an aliphatic amino group at the 5 terminus: Synthesis of fluorescent DNA primers for use in DNA sequence analysis. *Nucleic Acids Research* **13,** 2399–2412. ISSN: 03051048 (1985).

81. Ansorge, W., Sproat, B. S., Stegemann, J. & Schwager, C. A non-radioactive automated method for DNA sequence determination. *Journal of Biochemical and Biophysical Methods* **13,** 315–323. ISSN: 0165022X (Dec. 1986).

82. Ansorge, W., Sproat, B., Stegemann, J., Schwager, C. & Zenke, M. Automated DNA sequencing: Ultrasensitive detection of fluorescent bands during electrophoresis. *Nucleic Acids Research* **15,** 4593–4602. ISSN: 03051048 (June 1987).

83. Swerdlow, H. & Gesteland, R. Capillary gel electrophoresis for rapid, high resolution DNA sequencing. *Nucleic Acids Research* **18,** 1415–1419. ISSN: 03051048 (Mar. 1990).

84. Luckey, J. A. *et al.* High speed DNA sequencing by capillary electrophoresis. *Nucleic Acids Research* **18,** 4417–4421. ISSN: 03051048 (Aug. 1990).

85. Collins, F. S., Lander, E. S., Rogers, J. & Waterson, R. H. Finishing the euchromatic sequence of the human genome. *Nature* **431,** 931–945. ISSN: 00280836 (Oct. 2004).

86. Illumina Inc. *An introduction to Next-Generation Sequencing Technology (Pub. No. 770-2012-008-B)* tech. rep. (2017). www.illumina.com/technology/next-generation-sequencing.html.

87. Greenleaf, W. J. & Sidow, A. The future of sequencing: Convergence of intelligent design and market Darwinism. *Genome Biology* **15,** 303. ISSN: 1474760X (Mar. 2014).

88. Eid, J. *et al.* Real-time DNA sequencing from single polymerase molecules. *Science* **323,** 133–138. ISSN: 0036-8075 (Jan. 2009).

89. Levene, H. J. *et al.* Zero-mode waveguides for single-molecule analysis at high concentrations. *Science* **299,** 682–686. ISSN: 00368075 (Jan. 2003).

90. Loman, N. J. & Quinlan, A. R. Poretools: A toolkit for analyzing nanopore sequence data. *Bioinformatics* **30,** 3399–3401. ISSN: 14602059 (Dec. 2014).

91. Wang, Z. Q. *et al.* Mice lacking ADPRT and poly(ADP-ribosyl)ation develop normally but are susceptible to skin disease. *Genes and Development* **9,** 509–520. ISSN: 08909369 (Mar. 1995).

92. Jagtap, P. & Szabo, C. *Poly(ADP-ribose) polymerase and the therapeutic effects of its inhibitors* May 2005. doi:10.1038/nrd1718. http://www.nature.com/articles/nrd1718.

93. Illumina Inc. Quality Score. https://support.illumina.com/help/BaseSpace_OLH_009008/Content/Source/Informatics/BS/QualityScoreEncoding_swBS.htm (2011).

94. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760. ISSN: 1367-4803 (July 2009).

95. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079. ISSN: 13674803 (Aug. 2009).

96. Meacham, F. *et al. Identification and correction of systematic error in high-throughput sequence data* tech. rep. (2011), 451. doi:10.1186/1471-2105-12-451.

97. Schirmer, M., D'Amore, R., Ijaz, U. Z., Hall, N. & Quince, C. Illumina error profiles: Resolving fine-scale variation in metagenomic sequencing data. *BMC Bioinformatics* **17,** 125. ISSN: 14712105 (Dec. 2016).

98. McKenna, A. *et al.* The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* **20,** 1297–1303. ISSN: 10889051 (Sept. 2010).

99. GATK. *Base Quality Score Recalibration (BQSR)* 2018. https://software.broadinstitute.org/gatk/documentation/article?id=11081.

100. Aird, D. *et al.* Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biology* **12,** R18. ISSN: 14747596 (2011).

101. Van der Auwera, G. A. *et al.* From fastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. *Current Protocols in Bioinformatics* **43,** 1–33. ISSN: 1934340X (Oct. 2013).

102. Griffiths, A. J. F. *An introduction to genetic analysis* 860. ISBN: 0716735202 (W.H. Freeman, 2000).

103. Koboldt, D. C. *et al.* VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research* **22,** 568–576. ISSN: 10889051 (Mar. 2012).

104. Sherry, S. T. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research* **29,** 308–311. ISSN: 13624962 (Jan. 2001).

105. GATK. *ActiveRegion determination (HaplotypeCaller & Mutect2)* https://software.broadinstitute.org/gatk/documentation/article?id=11077.

106. Bruijn de, N. G. A combinatorial problem. *Proceedings of the Section of Sciences of the Koninklijke Nederlandse Akademie van Wetenschappen te Amsterdam* **49,** 758–764. ISSN: 00465755 (1946).

107. Smith, T. F. & Waterman, M. S. Identification of common molecular subsequences. *Journal of Molecular Biology* **147,** 195–197. ISSN: 00222836 (Mar. 1981).

108. GATK. *Local re-assembly and haplotype determination (HaplotypeCaller & Mutect2)* https://software.broadinstitute.org/gatk/documentation/article?id=11076.

109. Durbin, R., Eddy, S., Krobh, A. & Mitchison, G. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids.* ISBN: 9780511790492. doi:10.1017/CBO9780511790492. http://ebooks.cambridge.org/ref/id/CBO9780511790492%20http://dx.doi.org/10.1017/CBO9780511790492 (Cambridge University Press, Cambridge, 1998).

110. GATK. *Evaluating the evidence for haplotypes and variant alleles (HaplotypeCaller & Mutect2)* https://software.broadinstitute.org/gatk/documentation/article?id=11078.

111. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature Biotechnology* **31,** 213–219. ISSN: 10870156 (Mar. 2013).

112. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500,** 415–421. ISSN: 14764687 (Aug. 2013).

113. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. arXiv: 1201.0490. http://arxiv.org/abs/1201.0490 (Jan. 2012).

114. Bodor, A., Csabai, I., Mahoney, M. W. & Solymosi, N. RCUR: An R package for CUR matrix decomposition. *BMC Bioinformatics* **13,** 103. ISSN: 14712105 (May 2012).

115. Alexandrov, L. B. *et al.* The Repertoire of Mutational Signatures in Human Cancer. *bioRxiv,* 322859 (July 2019).

116. Kinde, I., Wu, J., Papadopoulos, N., Kinzler, K. W. & Vogelstein, B. Detection and quantification of rare mutations with massively parallel sequencing. *Proceedings of the National Academy of Sciences* **108,** 9530–9535. ISSN: 0027-8424 (June 2011).

117. Campbell, P. J. *et al.* Subclonal phylogenetic structures in cancer revealed by ultra-deep sequencing. *Proceedings of the National Academy of Sciences* **105,** 13081–13086. ISSN: 0027-8424 (Sept. 2008).

118. Forster, M. *et al.* From next-generation sequencing alignments to accurate comparison and validation of single-nucleotide variants: The pibase software. *Nucleic Acids Research* **41,** e16–e16. ISSN: 03051048 (Jan. 2013).

119. Nakamura, K. *et al.* Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Research* **39,** e90–e90. ISSN: 03051048 (July 2011).

120. Nielsen, R., Paul, J. S., Albrechtsen, A. & Song, Y. S. *Genotype and SNP calling from next-generation sequencing data* June 2011. doi:10.1038/nrg2986. http://www.ncbi.nlm.nih.gov/pubmed/21587300%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3593722%20http://www.nature.com/articles/nrg2986.

121. Johnson, G. E. Mammalian cell HPRT gene mutation assay: Test methods. *Methods in Molecular Biology* **817,** 55–67. ISSN: 10643745 (2012).

122. Mortelmans, K. & Zeiger, E. The Ames Salmonella/microsome mutagenicity assay. *Mutation Research - Fundamental and Molecular Mechanisms of Mutagenesis* **455,** 29–60. ISSN: 00275107 (Nov. 2000).

123. Lázár, V. *et al.* Bacterial evolution of antibiotic hypersensitivity. *Molecular Systems Biology* **9,** 700. ISSN: 17444292 (Jan. 2013).

124. Sakai, W. *et al.* Secondary mutations as a mechanism of cisplatin resistance in BRCA2-mutated cancers. *Nature* **451,** 1116–1120. ISSN: 14764687 (Feb. 2008).

125. Lagerqvist, A. *et al.* DNA repair and replication influence the number of mutations per adduct of polycyclic aromatic hydrocarbons in mammalian cells. *DNA Repair* **10,** 877–886. ISSN: 15687864 (Aug. 2011).

126. Pipek, O. *et al.* Fast and accurate mutation detection in whole genome sequences of multiple isogenic samples with IsoMut. *BMC Bioinformatics* **18,** 1–11. ISSN: 14712105 (2017).

127. Molnár, J. *et al.* The genome of the chicken DT40 bursal lymphoma cell line. *G3: Genes, Genomes, Genetics* **4,** 2231–2240. ISSN: 21601836 (Nov. 2014).

128. Zámborszky, J. *et al.* Loss of BRCA1 or BRCA2 markedly increases the rate of base substitution mutagenesis and has distinct effects on genomic deletions. *Oncogene* **36,** 746–755. ISSN: 14765594 (Feb. 2017).

129. Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499,** 214–218. ISSN: 00280836 (July 2013).

130. Cunningham, F. *et al.* Ensembl 2015. *Nucleic Acids Research* **43,** D662–D669. ISSN: 13624962 (Jan. 2015).

131. Faust, G. G. & Hall, I. M. *SAMBLASTER: Fast duplicate marking and structural variant read extraction* in *Bioinformatics* **30** (Sept. 2014), 2503–2505. doi:10.1093/bioinformatics/btu314. http://www.ncbi.nlm.nih.gov/pubmed/24812344%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4147885%20https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btu314.

132. Koboldt, D. C., Larson, D. E. & Wilson, R. K. Using varscan 2 for germline variant calling and somatic mutation detection. *Current Protocols in Bioinformatics* **44,** 15.4.1–15.4.17. ISSN: 1934340X (Dec. 2013).

133. Robinson, J. T. *et al. Integrative genomics viewer* Jan. 2011. doi:10.1038/nbt.1754. http://www.ncbi.nlm.nih.gov/pubmed/21221095%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3346182%20http://www.nature.com/articles/nbt.1754.

134. Kanagawa, T. *Bias and Artifacts in Multitemplate Polymerase Chain Reactions (PCR)* Jan. 2003. doi:10.1263/jbb.96.317. https://www.sciencedirect.com/science/article/pii/S1389172303901307.

135. Davidson-Pilon, C. *Bayesian methods for hackers. Probabilistic programming and Bayesian inference* 226. ISBN: 9780133902839 (2016).

136. Salvatier, J., Wiecki, T. V. & Fonnesbeck, C. Probabilistic programming in Python using PyMC3. *PeerJ Computer Science* **2016,** e55. ISSN: 23765992 (Apr. 2016).

137. Rochford, A. *Dirichlet process mixtures for density estimation* 2016. https://docs.pymc.io/notebooks/dp%7B%5C_%7Dmix.html.

138. Flaxman, A. & Wiecki, T. *Gaussian Mixture Model — PyMC3 3.6 documentation* https://docs.pymc.io/notebooks/gaussian%7B%5C_%7Dmixture%7B%5C_%7Dmodel.html (2019).

139. Pipek, O. *Advanced ploidy estimation — isomut2py 2.0.1 documentation* https://isomut2py.readthedocs.io/en/latest/PE%7B%5C_%7Dadvanced.html (2019).

140. Vohr, S. H. *et al.* A phylogenetic approach for haplotype analysis of sequence data from complex mitochondrial mixtures. *Forensic Science International: Genetics* **30,** 93–105. ISSN: 18780326 (Sept. 2017).

141. Fackenthal, J. D. & Olopade, O. I. *Breast cancer risk associated with BRCA1 and BRCA2 in diverse populations* Dec. 2007. doi:10.1038/nrc2054. http://www.ncbi.nlm.nih.gov/pubmed/18034184%20http://www.nature.com/articles/nrc2054.

142. Alsop, K. *et al.* BRCA mutation frequency and patterns of treatment response in BRCA mutation-positive women with ovarian cancer: A report from the Australian ovarian cancer study group. *Journal of Clinical Oncology* **30,** 2654–2663. ISSN: 0732183X (July 2012).

143. Abkevich, V. *et al.* Patterns of genomic loss of heterozygosity predict homologous recombination repair defects in epithelial ovarian cancer. *British Journal of Cancer* **107,** 1776–1782. ISSN: 00070920 (Nov. 2012).

144. Birkbak, N. J. *et al.* Telomeric allelic imbalance indicates defective DNA repair and sensitivity to DNA-damaging agents. *Cancer Discovery* **2,** 366–375. ISSN: 21598274 (Apr. 2012).

145. Popova, T. *et al.* Ploidy and large-scale genomic instability consistently identify basal-like breast carcinomas with BRCA1/2 inactivation. *Cancer Research* **72,** 5454–5462. ISSN: 00085472 (Nov. 2012).

146. Zámborszky, J. *et al.* Loss of BRCA1 or BRCA2 markedly increases the rate of base substitution mutagenesis and has distinct effects on genomic deletions. *Oncogene* **36.** ISSN: 1476-5594. doi:10.1038/onc.2016.243. http://www.ncbi.nlm.nih.gov/pubmed/27452521%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5096687 (July 2016).

147. Wyatt, M. D. & Pittman, D. L. *Methylating agents and DNA repair responses: Methylated bases and sources of strand breaks* Dec. 2006. doi:10.1021/tx060164e. http://www.ncbi.nlm.nih.gov/pubmed/17173371%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2542901%20https://pubs.acs.org/doi/10.1021/tx060164e.

148. Rydberg, B. & Lindahl, T. Nonenzymatic methylation of DNA by the intracellular methyl group donor S-adenosyl-L-methionine is a potentially mutagenic reaction. *The EMBO Journal* **1,** 211–216. ISSN: 0261-4189 (1982).

149. Yamazoe, M., Sonoda, E., Hochegger, H. & Takeda, S. *Reverse genetic studies of the DNA damage response in the chicken B lymphocyte line DT40* Aug. 2004. doi:10.1016/j.dnarep.2004.03.039. http://www.ncbi.nlm.nih.gov/pubmed/15279806%20https://linkinghub.elsevier.com/retrieve/pii/S1568786404001107.

150. Pfeifer, G. P. *Mutagenesis at methylated CpG sequences* in *Current Topics in Microbiology and Immunology* **301** (Springer-Verlag, Berlin/Heidelberg, 2006), 259–281. doi:10.1007/3-540-31390-7{_}10. http://www.ncbi.nlm.nih.gov/pubmed/16570852%20http://link.springer.com/10.1007/3-540-31390-7_10.

151. Nik-Zainal, S. *et al.* Mutational processes molding the genomes of 21 breast cancers. *Cell* **149,** 979–993. ISSN: 00928674 (May 2012).

152. Ottaviani, D., LeCain, M. & Sheer, D. *The role of microhomology in genomic structural variation* Mar. 2014. doi:10.1016/j.tig.2014.01.001. http://www.ncbi.nlm.nih.gov/pubmed/24503142%20https://linkinghub.elsevier.com/retrieve/pii/S016895251400002X.

153. Szikriszt, B. *et al.* A comprehensive survey of the mutagenic impact of common cancer cytotoxics. doi:10.1186/s13059-016-0963-7.

154. Eastman, A. Reevaluation of Interaction of cis-Dichloro(ethylenediamine)platinum(II) with DNA. *Biochemistry* **25,** 3912–3915. ISSN: 15204995 (July 1986).

155. Lemaire, M. A., Schwartz, A., Rahmouni, A. R. & Leng, M. Interstrand cross-links are preferentially formed at the d(GC) sites in the reaction between cis-diamminedichloroplatinum (II) and DNA. *Proceedings of the National Academy of Sciences of the United States of America* **88,** 1982–1985. ISSN: 00278424 (Mar. 1991).

156. Livraghi, L. & Garber, J. E. *PARP inhibitors in the management of breast cancer: Current data and future prospects* Dec. 2015. doi:10.1186/s12916-015-0425-1. http://www.ncbi.nlm.nih.gov/pubmed/26268938%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4535298%20http://bmcmedicine.biomedcentral.com/articles/10.1186/s12916-015-0425-1.

157. Sonnenblick, A., De Azambuja, E., Azim, H. A. & Piccart, M. *An update on PARP inhibitors - Moving to the adjuvant setting* Jan. 2015. doi:10.1038/nrclinonc.2014.163. http://www.ncbi.nlm.nih.gov/pubmed/25286972%20http://www.nature.com/articles/nrclinonc.2014.163.

158. Ang, Y. L. & Tan, D. S. *Development of PARP inhibitors in gynecological malignancies* July 2017. doi:10.1016/j.currproblcancer.2017.02.008. http://www.ncbi.nlm.nih.gov/pubmed/28583748%20https://linkinghub.elsevier.com/retrieve/pii/S0147027216301015.

159. O'Connor, M. J. *Targeting the DNA Damage Response in Cancer* Nov. 2015. doi:10.1016/j.molcel.2015.10.040. http://www.ncbi.nlm.nih.gov/pubmed/26590714%20https://linkinghub.elsevier.com/retrieve/pii/S109727651500831X.

160. Póti, Á. *et al.* Long-term treatment with the PARP inhibitor niraparib does not increase the mutation load in cell line models and tumour xenografts. *British Journal of Cancer* **119,** 1392–1400. ISSN: 15321827 (Nov. 2018).

161. Forozan, F. *et al.* Molecular cytogenetic analysis of 11 new breast cancer cell lines. *British Journal of Cancer* **81,** 1328–1334. ISSN: 00070920 (1999).

162. Grigoriadis, A. *et al.* Molecular characterisation of cell line models for triple-negative breast cancers. *BMC Genomics* **13,** 1–14. ISSN: 14712164 (2012).

163. Tate, J. G. *et al.* COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Research* **47,** D941–D947. ISSN: 0305-1048 (Jan. 2019).

164. Van Der Maaten, L. & Hinton, G. *Visualizing Data using t-SNE* tech. rep. (2008), 2579–2605. http://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf.

165. Ledermann, J. A. *et al.* Newly diagnosed and relapsed epithelial ovarian carcinoma: ESMO clinical practice guidelines for diagnosis, treatment and follow-up. *Annals of Oncology* **24,** vi24–vi32. ISSN: 15698041 (Oct. 2013).

166. Németh, E. *et al.* The genomic imprint of cancer therapies helps timing the formation of metastases. *International Journal of Cancer* **145,** 694–704. ISSN: 10970215 (Aug. 2019).

167. Boot, A. *et al.* In-depth characterization of the cisplatin mutational signature in human cell lines and in esophageal and liver tumors. *Genome Research* **28,** 654–665. ISSN: 15495469 (Apr. 2018).

168. Patch, A. M. *et al.* Whole-genome characterization of chemoresistant ovarian cancer. *Nature* **521,** 489–494. ISSN: 14764687 (May 2015).

169. Burrell, R. A. & Swanton, C. *Tumour heterogeneity and the evolution of polyclonal drug resistance* Sept. 2014. doi:10.1016/j.molonc.2014.06.005. http://www.ncbi.nlm.nih.gov/pubmed/25087573%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5528620%20http://doi.wiley.com/10.1016/j.molonc.2014.06.005.

170. Favero, F. *et al.* Glioblastoma adaptation traced through decline of an IDH1 clonal driver and macro-evolution of a double-minute chromosome. *Annals of Oncology* **26,** 880–887. ISSN: 15698041 (May 2015).

171. Niemiec, E. & Howard, H. C. Ethical issues in consumer genome sequencing: Use of consumers' samples and data. *Applied & translational genomics* **8,** 23–30. ISSN: 2212-0661 (Mar. 2016).

172. Alzu'bi, A., Zhou, L. & Watzlaf, V. Personal genomic information management and personalized medicine: challenges, current solutions, and roles of HIM professionals. *Perspectives in health information management* **11,** 1c. ISSN: 1559-4122 (2014).

173. Brothers, K. B. & Rothstein, M. A. Ethical, legal and social implications of incorporating personalized medicine into healthcare. *Personalized medicine* **12,** 43–51. ISSN: 1741-0541 (2015).

174. Gill, P. S. & Johnson, M. Ethnic monitoring and equity. *Bmj* **310,** 890. ISSN: 14685833 (1995).

175. Liao, Y. *et al.* Surveillance of health status in minority communities - Racial and Ethnic Approaches to Community Health Across the U.S. (REACH U.S.) Risk Factor Survey, United States, 2009. *MMWR Surveill Summ* **60,** 1–44. ISSN: 1545-8636 (Electronic) 0892-3787 (Linking) (2011).

176. Farkas, L. *Analysis and comparative review of equality data collection practices in the European Union Data: Data collection in the field of ethnicity* ISBN: 9789279660849. doi:10.2838/447194. http://ec.europa.eu/newsroom/just/document.cfm?action=display%7B%5C&%7Ddoc%7B%5C_%7Did=45791 (2017).

177. Gabai-Kapara, E. *et al.* Population-based screening for breast and ovarian cancer risk due to BRCA1 and BRCA2. *Proceedings of the National Academy of Sciences of the United States of America* **111,** 14205–10. ISSN: 1091-6490 (Sept. 2014).

178. Perkins, B. A. *et al.* Precision medicine screening using whole-genome sequencing and advanced imaging to identify disease risk in adults. *Proceedings of the National Academy of Sciences* **115,** 3686–3691. ISSN: 0027-8424 (Apr. 2018).

179. Hendriksen, R. S. *et al.* Global monitoring of antimicrobial resistance based on metagenomics analyses of urban sewage. *Nature Communications* **10,** 1124. ISSN: 2041-1723 (Dec. 2019).

180. Pipek, O. A. *et al.* Worldwide human mitochondrial haplogroup distribution from urban sewage. *Scientific Reports* **9,** 11624. ISSN: 2045-2322 (Dec. 2019).

181. Martellini, A., Payment, P. & Villemur, R. Use of eukaryotic mitochondrial DNA to differentiate human, bovine, porcine and ovine sources in fecally contaminated surface water. *Water Research* **39,** 541–548. ISSN: 00431354 (2005).

182. Luo, S. *et al.* Biparental Inheritance of Mitochondrial DNA in Humans. *Proceedings of the National Academy of Sciences of the United States of America* **115,** 13039–13044. ISSN: 1091-6490 (Dec. 2018).

183. Hagström, E., Freyer, C., Battersby, B. J., Stewart, J. B. & Larsson, N. G. No recombination of mtDNA after heteroplasmy for 50 generations in the mouse maternal germline. *Nucleic Acids Research* **42,** 1111–1116. ISSN: 03051048 (Jan. 2014).

184. Torroni, A. *et al.* Classification of European mtDNAs From an Analysis of Three European Populations. *Genetics* **144,** 1835–1850 (1996).

185. Comas, D. *et al.* Admixture, migrations, and dispersals in Central Asia: evidence from maternal DNA lineages. *European Journal of Human Genetics* **12,** 495–504 (2004).

186. Chen, Y. S., Torroni, A., Excoffier, L., Santachiara-Benerecetti, A. S. & Wallace, D. C. Analysis of mtDNA variation in African populations reveals the most ancient of all human continent-specific haplogroups. *American journal of human genetics* **57,** 133–49. ISSN: 0002-9297 (1995).

187. Cann, R. L., Stoneking, M. & Wilson, A. C. *Mitochondrial DNA and human evolution* 1987. doi:10.1038/325031a0.

188. Rishishwar, L. & Jordan, I. K. Implications of human evolution and admixture for mitochondrial replacement therapy. *BMC Genomics* **18,** 140. ISSN: 14712164 (Dec. 2017).

189. Underhill, P. A. & Kivisild, T. Use of Y Chromosome and Mitochondrial DNA Population Structure in Tracing Human Migrations. *Annual Review of Genetics* **41,** 539–564. ISSN: 0066-4197 (2007).

190. Cavalli-Sforza, L. L. & Feldman, M. W. The application of molecular genetic approaches to the study of human evolution. *Nature Genetics* **33,** 266–275. ISSN: 15461718 (2003).

191. Torroni, A. *et al.* Asian affinities and continental radiation of the four founding Native American mtDNAs. *American journal of human genetics* **53,** 563–90. ISSN: 0002-9297 (Sept. 1993).

192. Van Oven, M. & Kayser, M. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Human mutation* **30,** 386–394. ISSN: 10981004 (2009).

193. Deborah A. Bolnick *et al.* The Science and Business of Genetic Ancestry Testing. *Science* **318,** 399–400 (2007).

194. Knudsen, B. E. *et al.* Impact of Sample Type and DNA Isolation Procedure on Genomic Inference of Microbiome Composition. *mSystems* **1.** ISSN: 2379-5077. doi:10.1128/mSystems.00095-16, http://www.ncbi.nlm.nih.gov/pubmed/27822556%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5080404 (2016).

195. Andrews, R. M. *et al.* Reanalysis and revision of the cambridge reference sequence for human mitochondrial DNA [5]. *Nature Genetics* **23,** 147. ISSN: 10614036 (1999).

196. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. **25,** 1754–176010 (2009).

197. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079. ISSN: 13674803 (2009).

198. Ekblom, R., Smeds, L. & Ellegren, H. Patterns of sequencing coverage bias revealed by ultra-deep sequencing of vertebrate mitochondria. *BMC genomics* **15,** 467. ISSN: 1471-2164 (2014).

199. Biffi, A. *et al.* Principal-Component Analysis for Assessment of Population Stratification in Mitochondrial Medical Genetics. *American Journal of Human Genetics* **86,** 904–917. ISSN: 00029297 (June 2010).

200. Simoni, L., Calafell, F., Pettener, D., Bertranpetit, J. & Barbujani, G. Geographic Patterns of mtDNA Diversity in Europe. *The American Journal of Human Genetics* **66,** 262–278. ISSN: 00029297 (Jan. 2002).

201. Ingman, M., Kaessmann, H., Pääbo, S. & Gyllensten, U. Mitochondrial genome variation and the origin of modem humans. *Nature* **408,** 708–713. ISSN: 00280836 (2000).

202. Maca-Meyer, N., González, A. M., Larruga, J. M., Flores, C. & Cabrera, V. M. Major genomic mitochondrial lineages delineate early human expansions. *BMC Genetics* **2,** 13. ISSN: 14712156 (2001).

203. Saitou, N. & Nei, M. THE NEIGHBOR-JOINING METHOD - A NEW METHOD FOR RECONSTRUCTING PHYLOGENETIC TREES. *Molecular Biology and Evolution* **4,** 406–425 (1987).

204. Langley, C. H. & Fitch, W. M. An examination of the constancy of the rate of molecular evolution. *Journal of Molecular Evolution* **3,** 161–177. ISSN: 14321432 (1974).

205. Schaffer, J. What Not to Multiply Without Necessity. *Australasian Journal of Philosophy* **93,** 644–664. ISSN: 00048402 (Oct. 2015).

206. Vohr, S. H. *et al.* A phylogenetic approach for haplotype analysis of sequence data from complex mitochondrial mixtures. *Forensic Science International: Genetics* **30,** 93–105. ISSN: 18780326 (2017).

207. Van Oven, M. & Kayser, M. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Human mutation* **30,** 386–394. ISSN: 10981004 (2009).

208. Just, R. S. *et al.* Full mtGenome reference data: Development and characterization of 588 forensic-quality haplotypes representing three U.S. populations. *Forensic Science International: Genetics* **14,** 141–155. ISSN: 18780326 (2014).

209. Emery, L. S., Magnaye, K. M., Bigham, A. W., Akey, J. M. & Bamshad, M. J. Estimates of continental ancestry vary widely among individuals with the same mtDNA haplogroup. *American Journal of Human Genetics* **96,** 183–193. ISSN: 15376605 (Feb. 2015).

210. Watkins, W. S. *et al.* Genetic analysis of ancestry, admixture and selection in Bolivian and Totonac populations of the New World. *BMC Genetics* **13,** 39. ISSN: 14712156 (May 2012).

211. Cardena, M. M. *et al.* Assessment of the Relationship between Self-Declared Ethnicity, Mitochondrial Haplogroups and Genomic Ancestry in Brazilian Individuals. *PLoS ONE* **8,** e62005. ISSN: 19326203 (2013).

212. Poetsch, M. *et al.* Determination of population origin: A comparison of autosomal SNPs, Y-chromosomal and mtDNA haplogroups using a Malagasy population as example. *European Journal of Human Genetics* **21,** 1423–1428. ISSN: 10184813 (Dec. 2013).

213. Salas, A. *et al.* The mtDNA ancestry of admixed Colombian populations. *American Journal of Human Biology* **20,** 584–591. ISSN: 10420533 (2008).

214. Bamshad, M., Wooding, S., Salisbury, B. A. & Stephens, J. C. *Deconstructing the relationship between genetics and race* Aug. 2004. doi:10.1038/nrg1401. http://www.nature.com/articles/nrg1401.

215. Royal, C. D. *et al. Inferring Genetic Ancestry: Opportunities, Challenges, and Implications* 2010. doi:10.1016/j.ajhg.2010.03.011. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2869013/.

216. Kofler, B. *et al.* Mitochondrial DNA haplogroup T is associated with coronary artery disease and diabetic retinopathy: a case control study. *BMC Medical Genetics* **10,** 35. ISSN: 1471-2350 (Dec. 2009).

217. Krüger, J., Hinttala, R., Majamaa, K. & Remes, A. M. Mitochondrial DNA haplogroups in early-onset Alzheimer's disease and frontotemporal lobar degeneration. *Molecular Neurodegeneration* **5,** 8. ISSN: 1750-1326 (Feb. 2010).

218. Hendrickson, S. L. *et al.* Mitochondrial DNA haplogroups influence AIDS progression. *AIDS* **22,** 2429–2439. ISSN: 02699370 (Nov. 2008).

219. Darvishi, K., Sharma, S., Bhat, A. K., Rai, E. & Bamezai, R. N. K. Mitochondrial DNA G10398A polymorphism imparts maternal Haplogroup N a risk for breast and esophageal cancer. *Cancer Letters* **249,** 249–255. ISSN: 03043835 (May 2007).

220. Booker, L. M. *et al.* North American white mitochondrial haplogroups in prostate and renal cancer. *Journal of Urology* **175,** 468–473. ISSN: 00225347 (Feb. 2006).

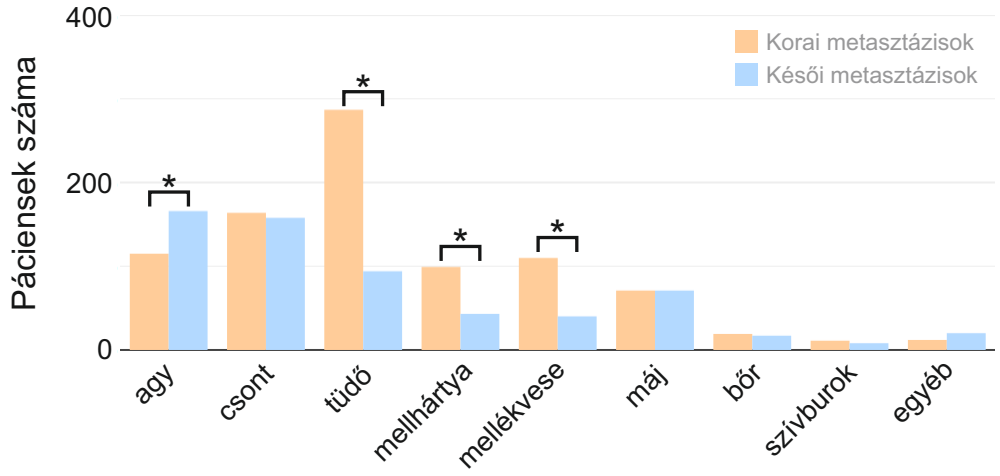
221. Urzúa-Traslaviña, C. G. *et al.* Relationship of Mitochondrial DNA Haplogroups with Complex Diseases. *Journal of Genetics and Genome Research* **1,** 1–5. ISSN: 23783648 (Dec. 2014).

222. Van Beek, E. J. *et al.* Rates of TP53 Mutation are Significantly Elevated in African American Patients with Gastric Cancer. *Annals of Surgical Oncology* **25,** 2027–2033. ISSN: 15344681 (July 2018).

223. Bollig-Fischer, A. *et al.* Racial diversity of actionable mutations in non-small cell lung cancer. *Journal of Thoracic Oncology* **10,** 250–255. ISSN: 15561380 (Feb. 2015).

224. Kurian, A. W. BRCA1 and BRCA2 mutations across race and ethnicity: distribution and clinical implications. *Current Opinion in Obstetrics and Gynecology* **22,** 72–78. ISSN: 1040-872X (Feb. 2010).

# MELLÉKLET - KLINIKAI PARAMÉTEREK, MINT BIOMARKEREK




**A1. ábra. A különböző szervek metasztázisainak időbeli eloszlása.** A csillagok azokat a szerveket jelölik, melyeknél a korai és késői metasztázisok gyakorisága között szignifikáns eltérés mutatkozott $\chi$-négyzet teszttel, Bonferroni-korrekció alkalmazásával ($\alpha = 0,05$).

| **Társbetegségek** | **Esetek száma** | **Kreatinin KTA (%)** | **BUN KTA (%)** |
|---|---|---|---|
| nincs | 180 | 7,78 | 15,56 |
| csak magas vérnyomás | 144 | 11,81 | 11,81 |
| csak cukorbetegség | 12 | 8,33 | 25,00 |
| csak COPD | 84 | 2,38 | 7,14 |
| magas vérnyomás + cukorbetegség | 41 | 12,20 | 17,07 |
| magas vérnyomás + COPD | 72 | 8,33 | 14,29 |
| cukorbetegség + COPD | 7 | 14,29 | 14,29 |
| mindhárom | 25 | 12,00 | 16,00 |

**A1. táblázat. A kreatinin és BUN szintjük alapján kóros tartományba eső páciensek aránya (KTA) a különböző társbetegség csoportokban a primer tumor diagnózisakor.**

| Paraméter | Lehetséges értékek |
|---|---|
| biszfoszfonát kezelés | igen/nem |
| primer tumor műtéti eltávolítása | igen/nem |
| stádium | korai/késői |
| dohányzás | igen/nem |
| nem | férfi/nő |
| magas vérnyomás | igen/nem |
| cukorbetegség | igen/nem |
| COPD | igen/nem |
| kemoterápia | igen/nem |
| csont metasztázis előtt/után | előtte/utána |

**A2. táblázat. A Cox-modell illesztése során felmerülő paraméterek és lehetséges értékeik.**

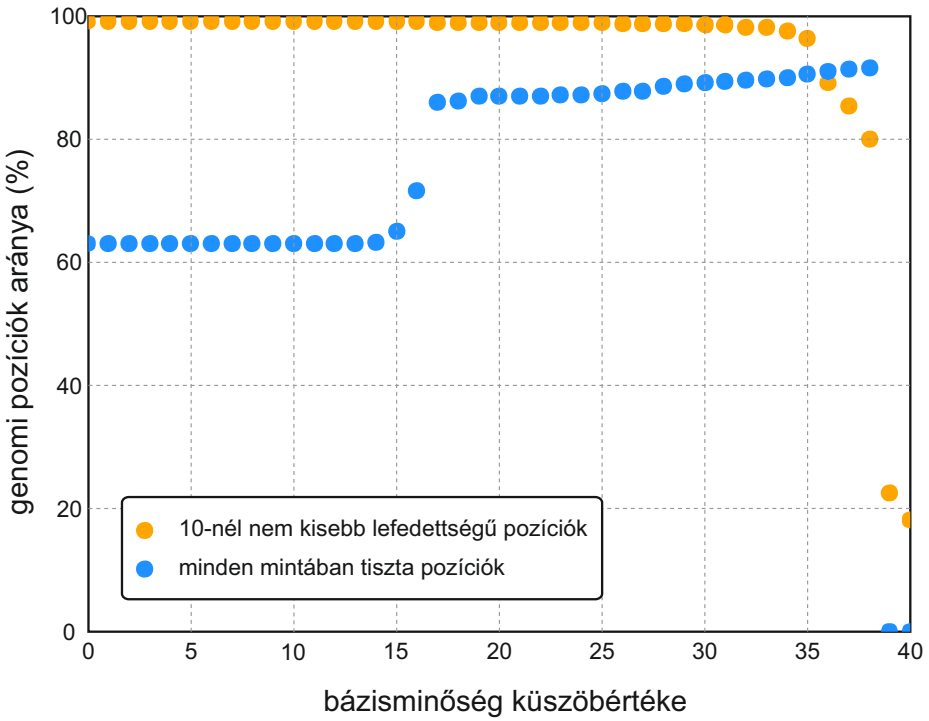
| Paraméter értéke | p-érték | relatív kockázat | 95%-os konfidencia intervallum |
|---|---|---|---|
| **kreatinin** | | | |
| **késői stádium** | 0,030 | 1,651 | 1,050-2,597 |
| **magas vérnyomás** | 0,020 | 1,639 | 1,081-2,485 |
| **biszfoszfonát kezelés** | 0,045 | 0,533 | 0,288-0,987 |
| **dohányzás** | 0,861 | 0,955 | 0,572-1,595 |
| **csont metasztázis után** | <0,001 | 3,022 | 1,591-5,740 |
| **BUN** | | | |
| **késői stádium** | 0,042 | 1,532 | 1,015-2,313 |
| **magas vérnyomás** | 0,033 | 1,372 | 1,025-1,836 |
| **biszfoszfonát kezelés** | 0,015 | 0,538 | 0,327-0,885 |
| **primer tumor műtéti eltávolítása** | <0,001 | 0,392 | 0,270-0,569 |
| **csont metasztázis után** | <0,001 | 2,655 | 1,582-4,457 |
| **kemoterápia** | 0,022 | 0,687 | 0,498-0,947 |

**A3. táblázat. A vesefunkciók elromlási idejére illesztett Cox-modellek paramétereihez tartozó relatív kockázatok értékei.**

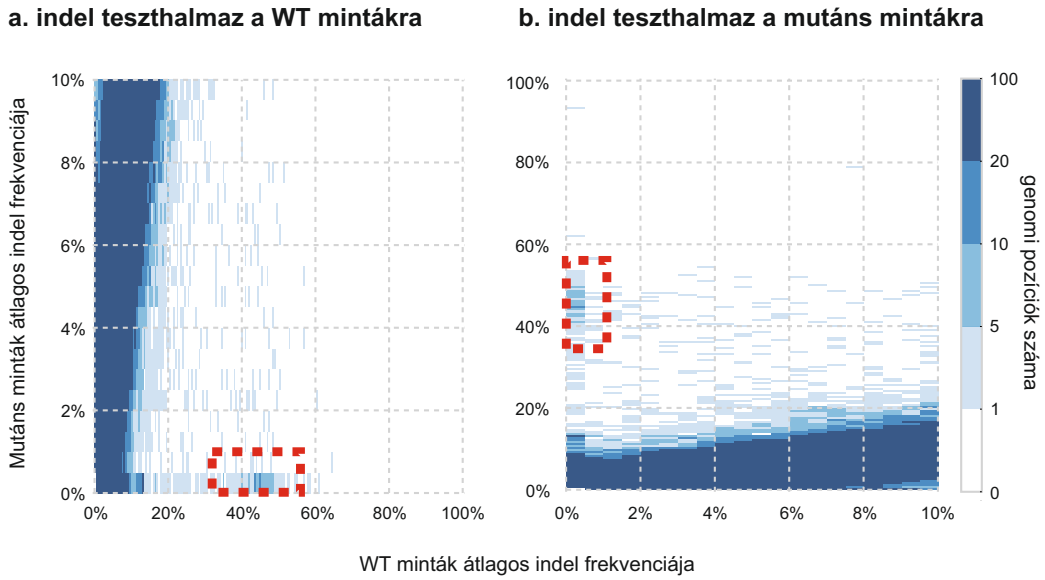
# MELLÉKLET - MUTÁCIÓK GYORS ÉS MEGBÍZHATÓ DETEKTÁLÁSA

| Mutáns minták | | WT minták | |
|---|---|---|---|
| **minta azonosító** | **kezelés** | **kezelés** | **minta azonosító** |
| S16 | kezdőklón | kezdőklón | S01 |
| S17 | erős mutagén kezelés | erős mutagén kezelés | S02 |
| S18 | gyenge mutagén kezelés | gyenge mutagén kezelés | S03 |
| S19 | gyenge mutagén kezelés | gyenge mutagén kezelés | S04 |
| S20 | gyenge mutagén kezelés | gyenge mutagén kezelés | S05 |
| S21 | gyenge mutagén kezelés | gyenge mutagén kezelés | S06 |
| S22 | gyenge mutagén kezelés | gyenge mutagén kezelés | S07 |
| S23 | erős mutagén kezelés | erős mutagén kezelés | S08 |
| S24 | erős mutagén kezelés | erős mutagén kezelés | S09 |
| S25 | erős mutagén kezelés | erős mutagén kezelés | S10 |
| S26 | erős mutagén kezelés | erős mutagén kezelés | S11 |
| S27 | duplikált kontroll minta | duplikált kontroll minta | S12 |
| S28 | erős mutagén kezelés | erős mutagén kezelés | S13 |
| S29 | erős mutagén kezelés | gyenge mutagén kezelés | S14 |
| S30 | duplikált kontroll minta | duplikált kontroll minta | S15 |

**B1. táblázat. A mutáció detektáló algoritmus teszteléséhez használt minták.**



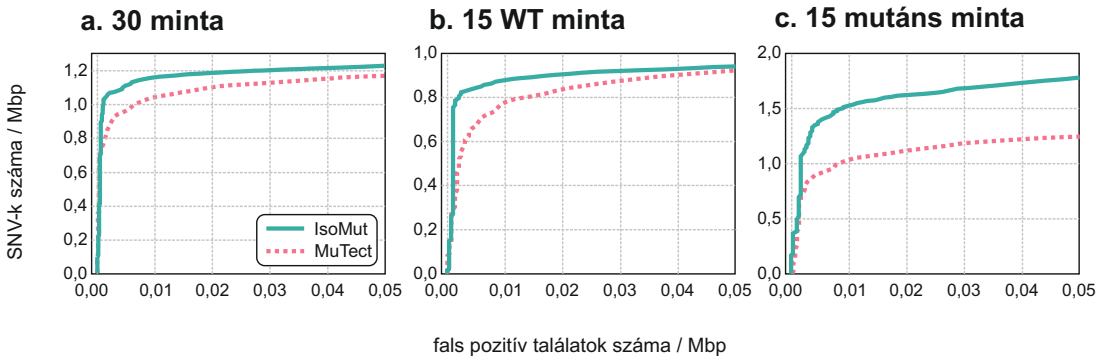

**B1. ábra. A bázisminőség szűrési küszöbértékének beállítása.** A megfelelő mértékben lefedett ($\geq 10$) pozíciók (narancssárga) és a minden mintában tiszta (kizárólag referencia) pozíciók (kék) rátájának változása a bázisminőségre beállított szűrési érték változtatásával.



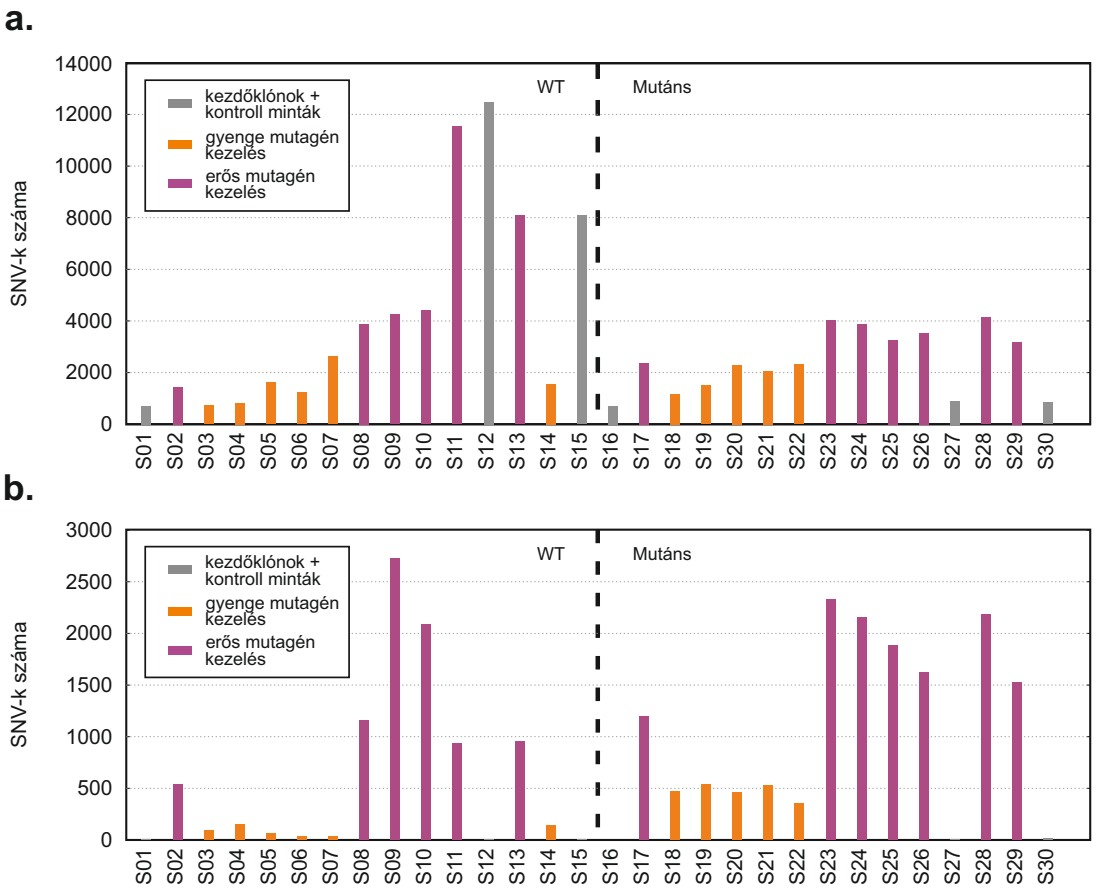

**B2. ábra. Indel teszthalmazok a kétféle genotípusra.**

| *sample_mut_freq_min* | *other_rnf_min* | *sample_cov_min* | **FPR** ($10^{-9}$) | **TPR** (%) |
|---|---|---|---|---|
| 0,5 | 0,93 | 7 | 14,30 | 46,82 |
| 0,34 | 0,96 | 7 | 21,56 | 64,35 |
| 0,35 | 0,93 | 7 | 27,06 | 68,77 |
| 0,34 | 0,92 | 7 | 31,68 | 70,63 |
| 0,31 | 0,93 | 7 | 29,04 | 75,90 |
| 0,31 | 0,92 | 7 | 33,00 | 76,87 |
| 0,30 | 0,92 | 7 | 35,86 | 78,03 |
| 0,30 | 0,90 | 7 | 42,46 | 79,17 |

**B2. táblázat. Indelek detektálásának hatékonysága különböző paraméter-beállításoknál.** A feltüntetett paraméter-beállítások megegyeznek a 6. ábra a) paneljének táblázatában található értékekkel.



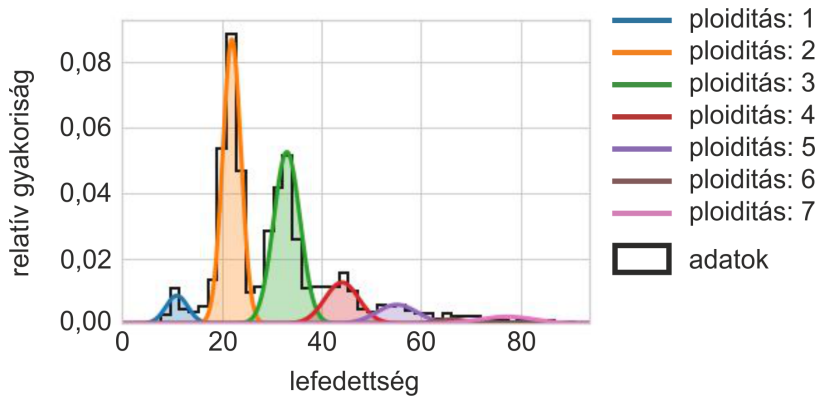

**B3. ábra. Kvázi-ROC görbék az IsoMut és MuTect teljesítményének összevetéséhez.** A vízszintes tengelyen a kontroll mintákban talált fals pozitív találatok 1 Mbp-re jutó száma, a függőlegesen pedig a kezelt mintákban azonosított SNV-k 1 Mbp genomra jutó száma szerepel. **a.** A teljes adatszett 30 mintájára futtatva. **b.** A 15 WT mintán futtatva. **c.** A 15 mutáns mintán futtatva.



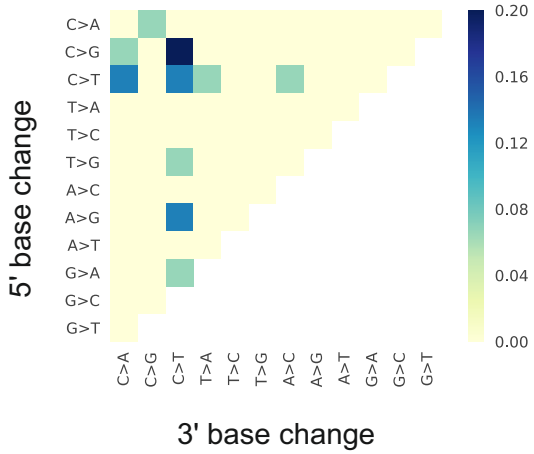

**B4. ábra. A MuTect által detektált egyedi SNV-k száma a vizsgált mintákban. a.** Alapértelmezett beállításokkal. **b.** Az LOD paraméter finomhangolásával. Az LOD paraméter határértékét úgy állítottuk be, hogy a kontrollmintákban azonosított mutációk darabszámát a lehető legalacsonyabban, míg a többi mintában detektáltakét a legmagasabban tartsuk. Az ábrázolt mutációkra LOD $\geq 20$.

| mesterséges genom | $w_i$ | | | | | | | $\sigma_i$ | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $w_1$ | $w_2$ | $w_3$ | $w_4$ | $w_5$ | $w_6$ | $w_7$ | $\sigma_1$ | $\sigma_2$ | $\sigma_3$ | $\sigma_4$ | $\sigma_5$ | $\sigma_6$ | $\sigma_7$ |
| **DT40** | 0,15 | 0,66 | 0,15 | 0,01 | 0,01 | 0,01 | 0,01 | 3 | 2 | 3 | 5 | 8 | 8 | 8 |
| **SUM149PT** | 0,15 | 0,25 | 0,30 | 0,12 | 0,09 | 0,06 | 0,03 | 2,5 | 2 | 2 | 4 | 7 | 8 | 8 |
| **normál humán** | 0,06 | 0,90 | 0,01 | 0,01 | 0,01 | 0,005 | 0,005 | 2,5 | 2 | 4 | 4 | 7 | 8 | 8 |
| **haploid** | 0,92 | 0,02 | 0,02 | 0,01 | 0,01 | 0,01 | 0,01 | 2 | 3 | 4 | 4 | 7 | 8 | 8 |
| **diploid** | 0,02 | 0,92 | 0,02 | 0,01 | 0,01 | 0,01 | 0,01 | 2 | 3 | 4 | 4 | 7 | 8 | 8 |
| **triploid** | 0,02 | 0,02 | 0,92 | 0,01 | 0,01 | 0,01 | 0,01 | 3 | 3 | 2 | 4 | 7 | 8 | 8 |
| **egyenlő súlyú** | 1/7 | 1/7 | 1/7 | 1/7 | 1/7 | 1/7 | 1/7 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| **zajos diploid** | 0,40 | 0,50 | 0,02 | 0,02 | 0,02 | 0,02 | 0,02 | 6 | 4 | 4 | 4 | 7 | 8 | 8 |

**B3. táblázat. Mesterséges genomok lefedettség-eloszlásainak paraméterei.** A mesterséges DT40 és a SUM149PT genomok a megfelelő sejtvonalak tapasztalati lefedettség-eloszlása alapján készültek. Előbbi egy többségében diploid, de haploid és triploid régiókat is tartalmazó genom, míg utóbbi erősen aneuploid. A lefedettség-eloszlások a fenti paraméterekkel 7 Gauss-eloszlás különböző $w_i$ súlyokkal vett keverékeként álltak elő. A Gauss-eloszlások középértékeinek a választott haploid lefedettség egész számú többszöröseit tekintettük.




**B5. ábra. A tényleges lefedettség-eloszlás a 7 komponensű keverék modellel együtt ábrázolva.**

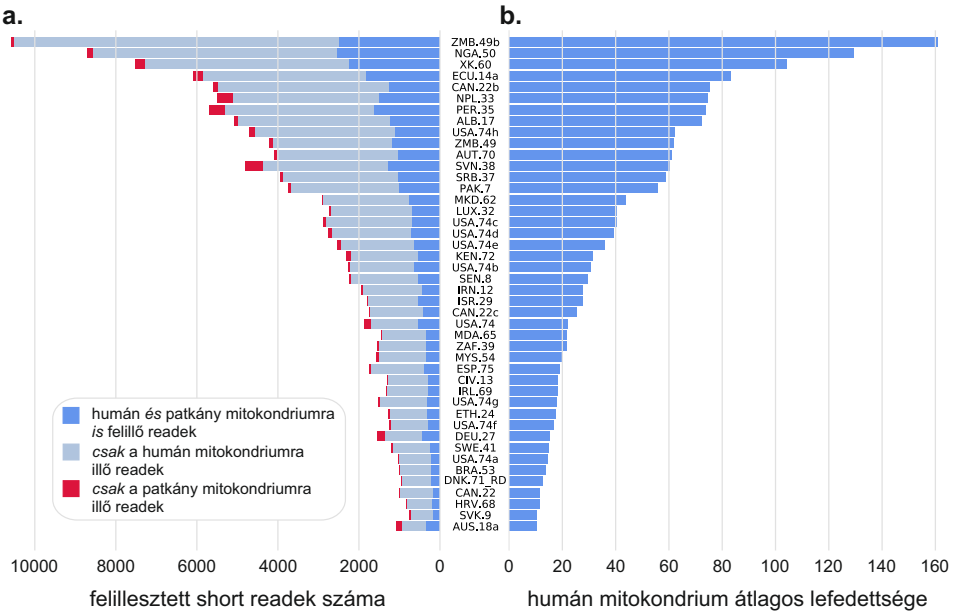



**B6. ábra. A DNV-spektrum hőtérképen az IsoMut2py ábrázolásában.** A színskála az érintett genomi pozíciópárok arányát jelzi. A vízszintes tengely mentén a DNS 3’ végéhez közelebbi, a függőlegesen az 5’ véghez közelebbi eredeti és mutált bázis szerinti kategóriák láthatók.

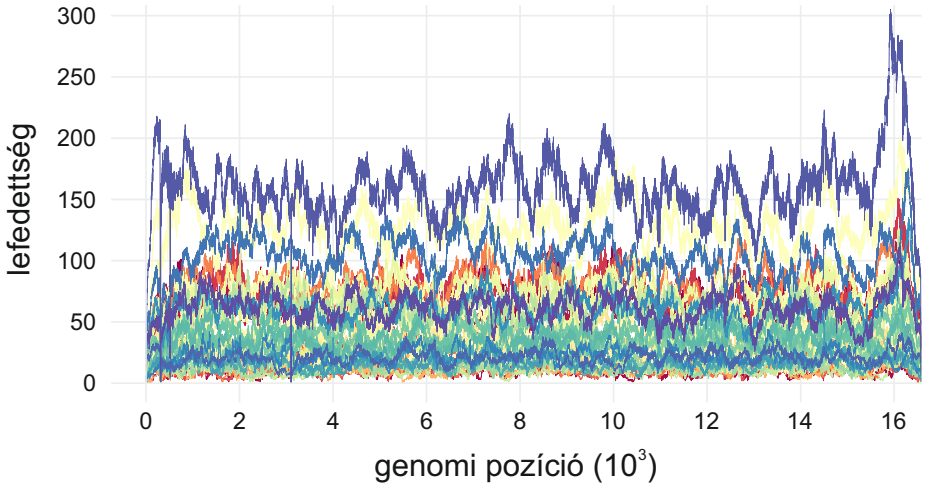
# MELLÉKLET - POPULÁCIÓ-SZINTŰ GENETIKAI VIZSGÁLATOK KÖRNYEZETI MINTÁKBÓL

| Minta azonosító | Szennyvíz gyűjtőhely (város, ország) |
|---|---|
| ALB.17 | Tirana, Albánia |
| AUS.18a | Melbourne, Ausztrália |
| AUT.70 | Bécs, Ausztria |
| BRA.53 | Belo Horizonte, Brazília |
| CAN.22b | Toronto, Kanada |
| CAN.22c | Ottawa, Kanada |
| CAN.22 | Regina, Kanada |
| CIV.13 | Abidjan, Elefántcsontpart |
| DEU.27 | Berlin, Németország |
| DNK.71_RD | Koppenhága, Dánia |
| ECU.14a | San Cristóbal, Ecuador |
| ESP.75 | Barcelona, Spanyolország |
| ETH.24 | Addisz-Abeba, Etiópia |
| HRV.68 | Zágráb, Horvátország |
| IRL.69 | Galway, Írország |
| IRN.12 | Teherán, Irán |
| ISR.29 | Jeruzsálem, Izrael |
| KEN.72 | Nairobi, Kenya |
| LUX.32 | Luxembourg, Luxemburg |
| MDA.65 | Chișinău, Moldova |
| MKD.62 | Szkopje, Észak-Macedónia |
| MYS.54 | Kuala Lumpur, Malajzia |
| NGA.50 | Lagos, Nigéria |
| NPL.33 | Katmandu, Nepál |
| PAK.7 | Karacsi, Pakisztán |
| PER.35 | Lima, Peru |
| SEN.8 | Dakar, Szenegál |
| SRB.37 | Belgrád, Szerbia |
| SVK.9 | Pozsony, Szlovákia |
| SVN.38 | Ljubljana, Szlovénia |
| SWE.41 | Uppsala, Svédország |
| USA.74a | Seattle, USA |
| USA.74b | Chicago, USA |
| USA.74c | El Paso, USA |
| USA.74d | Portland, USA |
| USA.74e | El Paso, USA |
| USA.74f | El Paso, USA |
| USA.74g | El Paso, USA |
| USA.74h | Denver, USA |
| USA.74 | Atlanta, USA |
| XK.60 | Pristina, Koszovó |
| ZAF.39 | Pretoria, Dél-afrikai Köztársaság |
| ZMB.49b | Kitwe, Zambia |
| ZMB.49 | Lusaka, Zambia |

**C1. táblázat. Minta azonosítók és szennyvízgyűjtőhelyek a 44 mintára, melyekben a humán mitokondrium lefedettsége elérte a 10-et.**

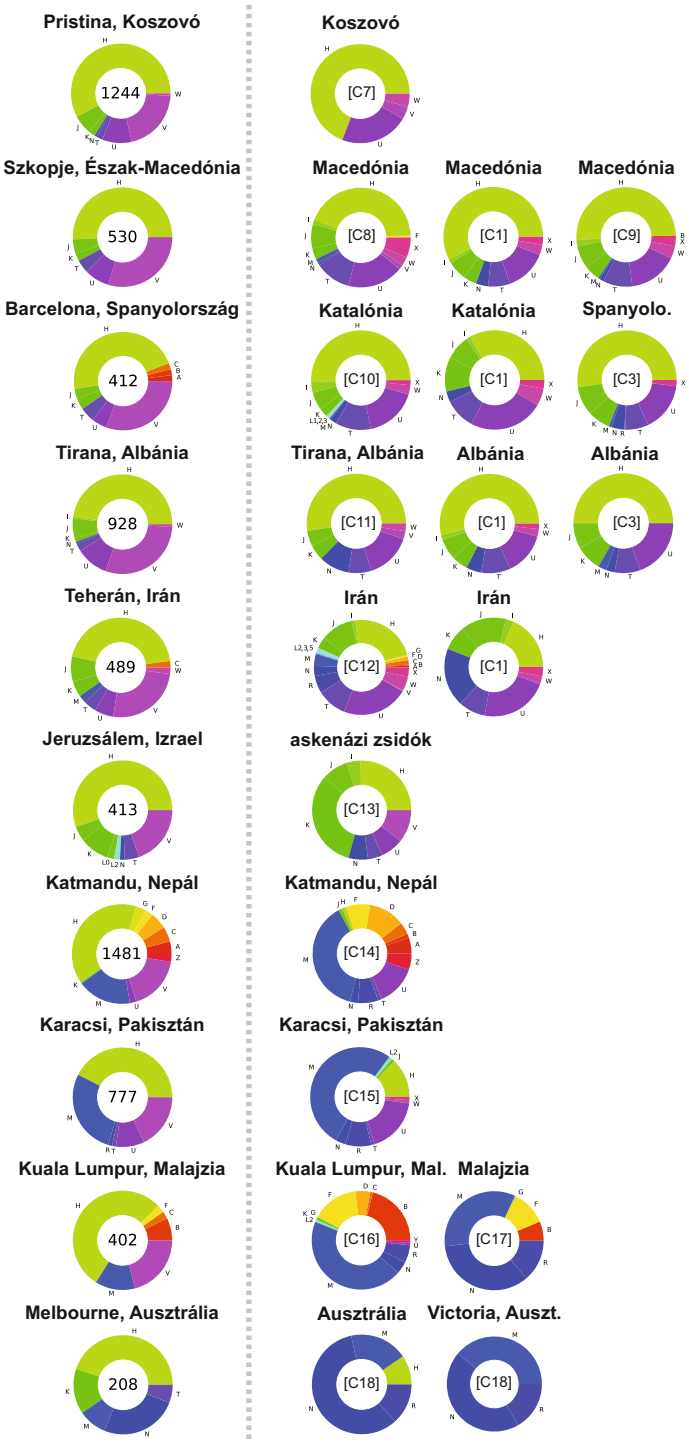



**C1. ábra. A referenciagenomra történt illesztés eredménye a szennyvízmintákban.** **a.** A humán és a vándorpatkány mitokondriumra illett short readek darabszáma azokban a mintákban, ahol az emberi mtDNS átlagos lefedettsége elérte a 10-et. **b.** A humán mtDNS átlagos lefedettsége. A függőleges tengelyen a minták azonosítói szerepelnek (C1. táblázat).

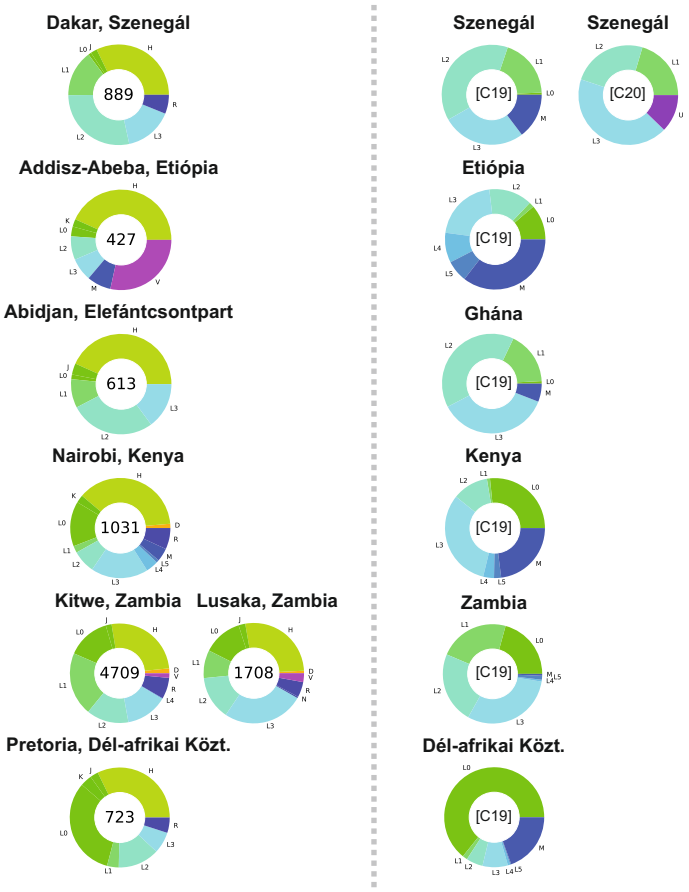



**C2. ábra. A humán mtDNS lefedettsége a vizsgált mintákban.** Az emberi mitokondrium lefedettsége azokban a mintákban, ahol az átlagos lefedettsége elérte a 10-et. A színes vonalak egy-egy mintára vonatkozó értékeket ábrázolnak. A mintákban megfigyelhető hozzávetőlegesen egyenletes eloszlás arra utal, hogy a hibásan felillesztett, nem humán readek száma elenyésző.

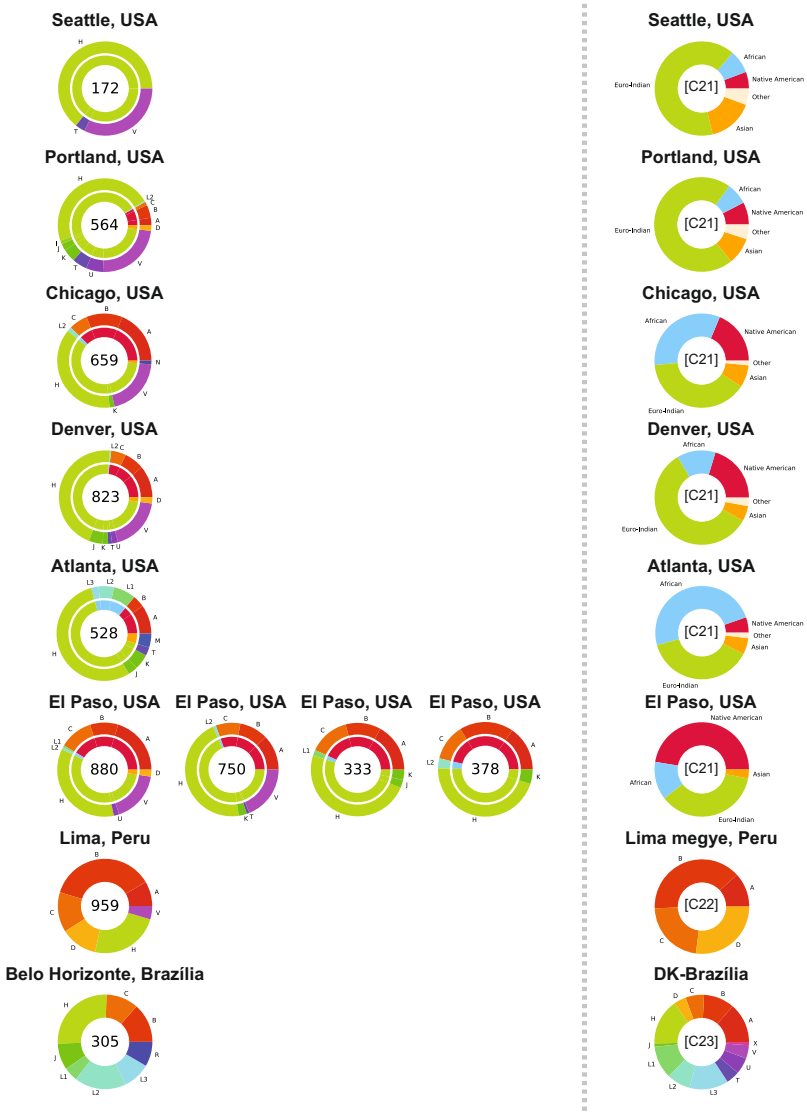

**C3. ábra. Az eurázsiai városok humán mtDNS haplocsoport összetétele.** A szaggatott függőleges vonal bal oldalán lévő eredmények a szennyvízminták elemzéséből, a jobb oldalon lévők irodalmi adatokból származnak. A színkódok megegyeznek a 24. ábrán láthatóakkal. A kördiagramok közepén lévő számok a szaggatott vonal bal oldalán a sikeresen bekategorizált readek darabszámát, a jobb oldalán pedig az adatok forrására vonatkozó hivatkozás sorszámát jelölik [C1–C6]. (ÉK: Északkelet)




**C4. ábra. Az eurázsiai városok és Melbourne humán mtDNS haplocsoport összetétele.** A szaggatott függőleges vonal bal oldalán lévő eredmények a szennyvízminták elemzéséből, a jobb oldalon lévők irodalmi adatokból származnak. A színkódok megegyeznek a 24. ábrán láthatóakkal. A kördiagramok közepén lévő számok a szaggatott vonal bal oldalán a sikeresen bekategorizált readek darabszámát, a jobb oldalán pedig az adatok forrására vonatkozó hivatkozás sorszámát jelölik [C1, C3, C7–C18]. (Mal.: Malajzia; Auszt.: Ausztrália)




**C5. ábra. Az afrikai városok humán mtDNS haplocsoport összetétele.** A szaggatott függőleges vonal bal oldalán lévő eredmények a szennyvízminták elemzéséből, a jobb oldalon lévők irodalmi adatokból származnak. A színkódok megegyeznek a 24. ábrán láthatóakkal. A kördiagramok közepén lévő számok a szaggatott vonal bal oldalán a sikeresen bekategorizált readek darabszámát, a jobb oldalán pedig az adatok forrására vonatkozó hivatkozás sorszámát jelölik [C19, C20]. (Közt.: Köztársaság)




**C6. ábra. Az észak- és dél-amerikai városok humán mtDNS haplocsoport összetétele.** A szaggatott függőleges vonal bal oldalán lévő eredmények a szennyvízminták elemzéséből, a jobb oldalon lévők irodalmi adatokból származnak. A színkódok megegyeznek a 24. ábrán láthatóakkal. A kördiagramok közepén lévő számok a szaggatott vonal bal oldalán a sikeresen bekategorizált readek darabszámát, a jobb oldalán pedig az adatok forrására vonatkozó hivatkozás sorszámát jelölik. Az Egyesült Államok városai esetén az irodalmi adatokat az etnikumra vonatkozó cenzus adatokból származtattuk, ezek a bal oldali kördiagramok belső paneljeivel közvetlenül összevethetők [C21–C23]. (DK: Délkelet)

## MITOKONDRIÁLIS HAPLOCSOPORT ADATFORRÁSOK


C1. Maciamo. Eupedia. https://www.eupedia.com/europe/european%7B%5C_%7Dmtdna%7B%5C_%7Dhaplogroups%7B%5C_%7Dfrequency.shtml.

C2. Lappalainen, T. *et al.* Population structure in contemporary Sweden - A Y-chromosomal and mitochondrial DNA analysis. *Annals of Human Genetics* **73,** 61–73. ISSN: 00034800 (2009).

C3. Cocoş, R. *et al.* Genetic affinities among the historical provinces of Romania and Central Europe as revealed by an mtDNA analysis. doi:10.1186/s12863-017-0487-5. https://bmcgenet.biomedcentral.com/track/pdf/10.1186/s12863-017-0487-5.

C4. Helgason, A. *et al.* mtDNA and the Islands of the North Atlantic: Estimating the Proportions of Norse and Gaelic Ancestry. *Am. J. Hum. Genet* **68,** 723–737 (2001).

C5. Poetsch, M., Wittig, H., Krause, D. & Lignitz, E. Mitochondrial diversity of a northeast German population sample. *Forensic Science International* **137,** 125–132. ISSN: 03790738 (2003).

C6. Šarac, J. *et al.* Maternal genetic heritage of southeastern europe reveals a new croatian isolate and a novel, local sub-branching in the X2 haplogroup. *Annals of Human Genetics* **78,** 178–194. ISSN: 14691809 (2014).

C7. Bosch, E. *et al.* Paternal and maternal lineages in the Balkans show a homogeneous landscape over linguistic barriers, except for the isolated Aromuns. *Annals of Human Genetics* **70,** 459–487. ISSN: 00034800 (2006).

C8. Derenko, M. *et al.* Complete Mitochondrial DNA Diversity in Iranians. doi:10.1371/journal.pone.0080673. http://journals.plos.org/plosone/article/file?id=10.1371/journal.pone.0080673%7B%5C&%7Dtype=printable (2013).

C9. Feder, J. *et al.* Differences in mtDNA haplogroup distribution among 3 Jewish populations alter susceptibility to T2DM complications. *BMC Genomics* **9.** doi:10.1186/1471-2164-9-198. http://www.biomedcentral.com/1471-2164/9/198 (2008).

C10. Gayden, T. *Genetic Diversity in the Himalayan Populations of Nepal and Tibet* PhD thesis (Florida International University, Mar. 2012). doi:10.25148/etd.FI12042312. http://digitalcommons.fiu.edu/etd/580.

C11. Quintana-Murci, L. *et al.* Where West Meets East: The Complex mtDNA Landscape of the Southwest and Central Asian Corridor. *Am. J. Hum. Genet* **74,** 827–845 (2004).

C12. Jinam, T. A. *et al.* Evolutionary history of continental southeast asians: Early train hypothesis based on genetic analysis of mitochondrial and autosomal DNA data. *Molecular Biology and Evolution* **29,** 3513–3527. ISSN: 07374038 (Nov. 2012).

C13. Nagle, N. *et al.* Mitochondrial DNA diversity of present-day Aboriginal Australians and implications for human evolution in Oceania. *Journal of Human Genetics advance online publication.* doi:10.1038/jhg.2016.147. https://www.kullillaart.com.au/assets/files/Journal%20of%20Human%20Genetics.pdf (2016).

C14. Maruyama, S., Nohira-Koike, C., Minaguchi, K. & Nambiar, P. MtDNA control region sequence polymorphisms and phylogenetic analysis of Malay population living in or around Kuala Lumpur in Malaysia. https://link.springer.com/content/pdf/10.1007%7B%5C%%7D2Fs00414-009-0355-6.pdf.

C15. Čoklo, M. *et al.* Diversity of Y-chromosomal and mtDNA Markers Included in Mediscope Chip within Two Albanian Subpopulations from Croatia and Kosovo: Preliminary Data. *Collegium antropologicum* **40,** 195–8. ISSN: 0350-6134 (2016).

C16. Zimmermann, B. *et al.* Mitochondrial DNA control region population data from Macedonia. *Forensic Science International: Genetics* **1,** e4–e9. ISSN: 18724973 (Dec. 2007).

C17. Santos, C. *et al.* Mitochondrial DNA and Y-chromosome structure at the mediterranean and Atlantic façades of the Iberian Peninsula. *American Journal of Human Biology* **26,** 130–141. ISSN: 10420533 (2014).

C18. Cvjetan, S. *et al.* Frequencies of mtDNA Haplogroups in Southeastern Europe. *Coll. Antropol* **28913,** 193–198 (2004).

C19. Silva, M. *et al.* 60,000 years of interactions between Central and Eastern Africa documented by major African mitochondrial haplogroup L2. doi:10.1038/srep12526. https://www.nature.com/articles/srep12526.pdf (2015).

C20. Stefflova, K., Dulik, M. C., Pai, A. A., Walker, A. H. & Zeigler-Johnson, C. M. Evaluation of Group Genetic Ancestry of Populations From Philadelphia and Dakar in the Context of Sex-Biased Admixture in the Americas. *Americas. PLoS ONE* **4,** 1–10 (2009).

C21. U.S. Census Bureau. *The Demographic Statistical Atlas of the United States - Statistical Atlas* 2015. https://statisticalatlas.com/ (2018).

C22. Sandoval, J. R. *et al.* The Genetic History of Peruvian Quechua-Lamistas and Chankas: Uniparental DNA Patterns among Autochthonous Amazonian and Andean Populations. *Annals of Human Genetics* **80,** 88–101. ISSN: 14691809 (2016).

C23. Alves-Silva, J. *et al.* The Ancestry of Brazilian mtDNA Lineages. *The American Journal of Human Genetics* **67,** 444–461. ISSN: 00029297 (2002).

# ADATLAP

**a doktori értekezés nyilvánosságra hozatalához***

## I. A doktori értekezés adatai

A szerző neve: Pipek Orsolya Anna

MTMT-azonosító: 10060440

A doktori értekezés címe és alcíme: A személyre szabott gyógyászat nyomában - DNS-minták tulajdonságainak vizsgálata új generációs szekvenálási adatok alapján

DOI-azonosító: 10.15476/ELTE.2019.243

A doktori iskola neve: ELTE TTK Fizika Doktori Iskola

A doktori iskolán belüli doktori program neve: Statisztikus fizika, biológiai fizika és kvantumrendszerek fizikája program

A témavezető neve és tudományos fokozata: Csabai István, DSc

A témavezető munkahelye: ELTE TTK Komplex Rendszerek Fizikája Tanszék

## II. Nyilatkozatok

**1.** A doktori értekezés szerzőjeként
a) hozzájárulok, hogy a doktori fokozat megszerzését követően a doktori értekezésem és a tézisek nyilvánosságra kerüljenek az ELTE Digitális Intézményi Tudástárban. Felhatalmazom a Természettudományi kar Dékáni Hivatal Doktori, Habilitációs és Nemzetközi Ügyek Csoportjának ügyintézőjét, hogy az értekezést és a téziseket feltöltse az ELTE Digitális Intézményi Tudástárba, és ennek során kitöltse a feltöltéshez szükséges nyilatkozatokat.

**2.** A doktori értekezés szerzőjeként kijelentem, hogy
a) az ELTE Digitális Intézményi Tudástárba feltöltendő doktori értekezés és a tézisek saját eredeti, önálló szellemi munkám és legjobb tudomásom szerint nem sértem vele senki szerzői jogait;
b) a doktori értekezés és a tézisek nyomtatott változatai és az elektronikus adathordozón benyújtott tartalmak (szöveg és ábrák) mindenben megegyeznek.

**3.** A doktori értekezés szerzőjeként hozzájárulok a doktori értekezés és a tézisek szövegének plágiumkereső adatbázisba helyezéséhez és plágiumellenőrző vizsgálatok lefuttatásához.

Kelt: Budapest, 2019. 09. 19.

Pipek Orsolya

……

**a doktori értekezés szerzőjének aláírása**

**ELTE SZMSZ SZMR 12. sz. melléklet*