2.félévi beszámoló

Szakállas Nikolett (szakallasn3@student.elte.hu)

Statisztikus Fizika, Biológiai Fizika és Kvantumrendszerek Fizikája PhD program

Témavezető: **Molnár Béla**, M.D., DSc
Semmelweis Egyetem
Belgyógyászati és Onkológiai Klinika

Belső konzulens: **Szabó Bálint**, PhD
Eötvös Loránd Tudományegyetem,
Természettudományi Kar
Biológiai Fizika Tanszék

Dolgozat címe: Új szövet alapú egyedi sejt izolációs rendszer kifejlesztése, annak biológiai alkalmazásai

1 Bevezetés, visszatekintés az előző félévi tevékenységre

Ahogy egyéb daganatos elváltozások kapcsán is ismeretes, úgy a vastagbél daganatok kialakulása is egy több lépcsős folyamat, mely során a normál vastagbél hám- vagy kötőszövetek neoplasztikus változásokon mennek keresztül mutációk felhalmozódása mellett. A normál és tumoros vastagbél szöveti eredetű minták teljes genomi és teljes exomi szekvenálása nyújthat betekintést az említett mutációs mintázatokra. Az elmúlt években a szekvenálási eljárások egyre gyorsabb terjedése és az ezzel párhuzamosan rohamosan fejlődő bioinformatikai módszerek tárháza egyre csak bővül, melynek hatására egyes daganat típusoknál fellelhető leggyakoribb génekről és azok mutációiról adatbázisok készültek. Ezen adatbázisokra alapozva a szekvenálási adatok kiértékelésének lezárultával összehasonlíthatóak a kísérletben részt vett minták mutációs tulajdonságai a már ismert és megalapozott mutációs adatbázissal (mely már referenciaként szolgál). Az előző féléves munkám folytatásaként ebben a félévben is tumor-normál (pontosabban tumor-tumor melleti normál szövet) mintapárok szekvenálási adatainak kiértékelésével foglalkoztam, azonban a Nanopore szekvenátor mellett immár Illumina szekvenátor adatait is feldolgoztam. Míg a Nanopore szekvenátorral WGS (teljes genomi), úgy az Illuminával WES (teljes exomi) szekvenálást végeztünk. A munkám célja a mutációs tulajdonságok és a különböző módszerek ezzel kapcsolatos jóságának meghatározása, illetve a módszerek egymáshoz hasonlított teljesítményét, tulajdonságait vizsgáltam.

A doktori kutatásom másik része olyan egyedi sejt izolációs rendszer fejlesztése, mellyel az egyedi sejt mérettartományban vagyunk képesek mintákat izolálni, gyűjteni. Az előzőleg ismertetett (előző féléves beszámoló) tulajdonságok átgondolásra kerültek, s új megalapozó kutatás indult, ennek köszönhető, hogy a doktori kutatásom második féléve főképp a bioinformatikai kiértékelésekre irányult. Hogy miért lényeges az egyes szekvenátorok összehasonlítása a kutatásom szempontjából, az úgy magyarázható, hogy a sejt izolációs rendszerrel gyűjtött mintákat később szekvenálni szeretnénk, így meghatározva a választ adott biológiai kérdésekre.

2 Az aktuális félévben elvégzett kutatások ismertetése

2.1 Minták szekvenálása, bioinformatikai kiértékelése

Az előző féléves munkámat kiegészítve már nemcsak az Oxford Nanopore Technologies PromethION24 típusú modellével végeztük a szekvenálási adatok gyűjtését, hanem Illumina NextSeq berendezése is használatra került. Az előbb említett eszközzel teljes genomi, míg a második berendezéssel teljes exomi szekvenálást végeztünk. A teljes genomi szekvenálás során egy adott organizmus (esetünkben humán) mintájában a teljes genomi (azaz teljes genetikai információ), míg teljes exom esetén az exomi (azaz a genomban lévő fehérje-kódoló régiók) bázissorendjét határozzuk meg. A két eszköz közötti további különbség, hogy míg a Nanopore szekvenátor kimeneti jelei áramsűrűségek, melyek függenek a nanopórusok nagyságától és melyeket nagyban meghatároz a bemenő minta DNS mennyisége, addig az Illumina esetén a readeket párhuzamos SBS-re (Sequencing by synthesis: szintézis alapú szekvenálás) alapozva szekvenáljuk, mely során DNS-polimerázokat és dNTP-ket használunk a DNS szálak replikálására.

2.1.1 Bővített bioinformatikai kiértékelés

A korábbi beszámolómban ismertetett kiértékelési lépések (bázisazonosítás + referencia genomra való illesztés, .bam fájlok rendezése, egyesítése, indexelése, alapstatisztika kinyerése, variánsazonosítás, methylációs mintázatok meghatározása és minőségellenőrzés) kiegészültek egy adatszűrési lépéssel és az azonosított mutációk referencia gén adatbázis alapján kerültek kiértékelésre.

Az újonnan bevezetett lépések rövid ismertetése:

- Szűrés: az alapstatisztikák meghatározásánál több minta esetén futottunk abba az anomáliába, hogy az ATGC arányok sem egymáshoz képest (AT-GC), sem különállóan (A-T-G-C) nem felelnek meg a Chargaff-szabályoknak. Ami azt jelenti, hogy AT $\neq$ GC és A,T $\neq$ 30%, G, C $\neq$ 20 %. Ennek okán az adatok részletesebb elemzésére volt szükség. A bázisazonosítás és a referencia genomra való illesztés után kapott .bam fájlok statisztikájának meghatározásának eredménye egy szövegfájl, mely tartalmazza a bázismennyiségeket readekre lebontva. Az adatok vizsgálata kapcsán felismertem azt az összefüggést, hogy bizonyos read hossz alatt (< 200 bázispár) az arányok elcsúsznak, így szükséges a rövidebb hosszak kiszűrése. Ezek a rövidebb hosszak feltételezhetően a DNS szál szekvenálás során való törése, sérülése kapcsán keletkeznek. Az adatok egyre szigorúbb szűréssel történő vizsgálata hosszabb átlagos read hosszat és megfelelő bázis arányokat eredményezett. Ennek kapcsán az összes további kiértékelés során ellenőriztem, hogy szükséges-e szűrés: a bázisok adatait hisztogramon ábrázoltam, s amennyiben azt tapasztaltam, hogy az adatok eloszlása erősen eltér a normál (vagy közel normál) eloszlástól, szűrést alkalmaztam.

- Referencia adatbázis használata a CRC tumorspecifikus gének kapcsán: a referencia CRC specifikus génmutációs lista forrása a TCGA (The Cancer Genome Atlas Program, [1]) által szolgáltatott gén adatbázis (`https://portal.gdc.cancer.gov/`) volt. Innen letöltve a 60 leggyakoribb mutált gén listáját, meg tudtam határozni ezen specifikus gének jelenlétét a szekvenált, mikrodisszektált, egyedi sejt nagyságú minták variánsadataiban és gyakoriságukat. Ez az összehasonlítás egyelőre a tumoros (CRC) mintákra készült el, és a várt eredményt kaptam: megtalálható a 12 CRC mintában a 60 leggyakoribb vastagbél rákra jellemző génmutáció változó előfordulással (a mennyiségek erősen függenek a szekvenálási lefedettségtől).

Az adatok elemzését általam írt és már meglévő, szabad felhasználású programok segítségével végeztem, az általam használt programnyelv Python volt, a felhasznált Ubuntu parancssori programok pedig a következőek voltak Nanopore szekvenálás esetén: *Guppy* - bázishívás + referencia genomra illesztés, *samtools sort/index/merge* - .bam fájlokon végzett műveletek végrehajtása, *modbam2bed* - methylációs vizsgálatok, *EPI2ME Labs Human Variation workflow* - variánsazonosítás, *bcftools annotate* - .vcf fájlok annotálása, *bcftools stats* - statisztikai adatok egy részének kinyerése + minőségellenőrzés: ezt többféleképpen elvégeztem több kiértékelési lépés után; Illumina szekvenálásnál: BaseSpace felületen fellelhető Dragen (germline és somatic pipeline) csomagok + minőségellenőrzés.

A szűrt adatok ATGC arányait a legtöbb minta esetén összehasonlítottam az adott mintán végzett genom assembly (Flye) eredményével. A genom assembly azt jelenti, hogy a szekvenálásból nyert szekvenciák (.fastq fájlok) alapján újra összeállítjuk a teljes genomi szekvenciát. Az assembly eredményét tekintve pedig meghatározhatóak az ATGC arányok a teljes genomra. Az eddigi tapasztalat az, hogy kellően erős adatszűrés visszaadja az általunk várt bázisarányokat, továbbá az

egyes szűrések eredményét ábrázolva azok tendálnak az assemblyből kapott arányok felé. Ez egy validációs lépés lehet az ATGC arányok helyességének meghatározására.

2.1.2 Különböző mérettartományú tumor-normál mintapárok szekvenálása

- Mikrodisszektált egyedi sejtek szekvenálása: 12 tumor-normál mintapár került összegyűjtésre, a minták lézermikrodisszekcióval lettek eltávolítva friss fagyasztott mintát tartalmazó lemezről. Ebben az esetben Illumina szekvenátorral nyertük ki az exomi DNS szekvenciákat. Az adatok kiértékelése egy külső felületen történt, melynek neve BaseSpace és kifejezetten az Illumina szekvenátorok adatainak kiértékeléséhez készült. Itt történt meg a bázisazonosítástól a variánsazonosításig az összes lépés, majd az azonosított variánsokat (annotálás után) összevetettem a referencia adatbázisból kinyert leggyakoribb génmutációkkal. A kapott eredmények jóságát egy korábbi kutatás eredményeire ([2]) alapozva ellenőrizzük, azonban ez már nem történik meg a beszámoló leadási határidejéig.

- Makrodisszektált minták szekvenálása: 7 tumor-normál mintapárt gyűjtöttünk össze műtéti mintákból. Ebben az esetben nagyobb volt a minták dimenziója (erre utal a makro előtag). A kísérlet nem csak a tumor-normál eltérések vizsgálatára összpontosított, hanem egyéb szempontok is figyelembe vételre kerültek: pl. hogyan hat a mintákra, ha a szekvenálásnál (Nanopore) az újrahasználható flow cell-eket valóban újra használjuk. Az eredményekből egyelőre az látszik, hogy használt flow cell esetén kevesebb bázist tudunk kinyerni (viszont itt figyelembe kell venni a bemeneti DNS mennyiséget is!).

Előző félévi beszámoló kiegészítése: adaptív-normál összehasonlítás. Az 1. féléves beszámolóm írásakor még nem készült el az adaptív szekvenálás kiértékelése. Azonban megkaptuk a várt eredményt: ezzel a módszerrel a ritkább, kisebb lefedettségű mutációk is kimutathatóak adott genomi szakaszon.

2.2 Egyedi sejt izolációs rendszer fejlesztése

Az ismertetett időszakban inkább a mintavételt követő biológiai vizsgálatokra, azok bioinformatikai kiértékelésére koncentráltam. A korábbi beszámolóban említett és röviden ismertetett rendszer alapjai azonban átgondolásra kerülnek, így az elkövetkezendő félévben ennek tudományos hátterére koncentrálok majd, illetve tervben van egy deszkamodell felállítása az átgondolt koncepció tesztelése és validálása kapcsán.

3 Tanulmányi tevékenység az aktuális félévben

A félév során az alábbi tárgyakat végeztem el:

- Sejtmozgás molekuláris és biofizikai mechanizmusai (FIZ/3/071E)
- Biostatisztika (Semmelweis Egyetem)
- Gráfok a bioinformatikában (FIZ/3/063E)

4 Konferenciák az aktuális félévben

Ebben a félévben 2023. május 17. és 19. között részt vettem a Nanopore London Calling 2023-as konferencián.

References

[1] The Cancer Gemone Atlas Program
`https://www.cancer.gov/tcga`

[2] A. Kalmar et. al.
Patterns of Somatic Variants in Colorectal Adenoma and Carcinoma Tissue and Matched Plasma Samples from the Hungarian Oncogenome Program
Cancers 2023, 15(3), 907;
doi: 10.3390/cancers15030907